PROJECT REPORT

Heart Disease Prediction with Logistic Regression

Devtern

PATEL VISHRUTKUMAR DINESHBHAI

DECEMBER 30, 2023

INDEX

Heart Disease Prediction with Logistic Regression

1. Introduction

Cardiovascular diseases, with heart disease at the forefront, stand as the primary cause of mortality worldwide. As the global burden of heart-related illnesses persists, the need for early identification of individuals at high risk becomes increasingly crucial. Timely preventive measures and interventions play a pivotal role in mitigating the impact of cardiovascular diseases. In response to this imperative, this study leverages logistic regression to develop a sophisticated model for predicting heart disease risk based on comprehensive patient data.

2. Objective

The overarching objective of this study is to construct a robust logistic regression model tailored to accurately predict the presence or absence of heart disease in patients. Recognizing the multifaceted nature of cardiovascular health, our focus extends to 14 key risk factors, each contributing uniquely to the overall risk profile. In pursuit of this objective, the study encompasses a comprehensive exploration of the dataset, encompassing thorough data acquisition and preprocessing steps. The subsequent model development and evaluation phases are designed to illuminate the intricate relationships between the selected risk factors and heart disease. By systematically assessing model performance, we aim to derive actionable insights into the predictive capabilities of the logistic regression approach. This objective aligns with the broader mission of advancing predictive analytics in the realm of cardiovascular health. By harnessing the power of logistic regression, we endeavor to refine risk prediction models, fostering a deeper understanding of the dynamics that underlie heart disease development. Through the successful achievement of this objective, we anticipate contributing valuable tools to the medical community, empowering healthcare professionals with enhanced decision-support systems for identifying and managing heart disease risks in diverse patient populations.

3. Methods

3.1 Data Acquisition and Preprocessing

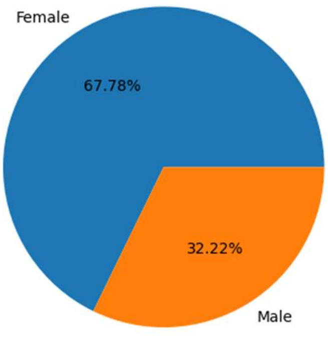
The dataset utilized in this study comprised 14 features and a binary target variable representing heart disease diagnosis.

- 3.1.1 Initial exploration of the data revealed no missing values or duplicates.
- 3.1.2 Data types were appropriately identified, ensuring compatibility with the analysis.
- 3.1.3 Categorical features, such as "Heart Disease" were encoded using LabelEncoder for model compatibility.

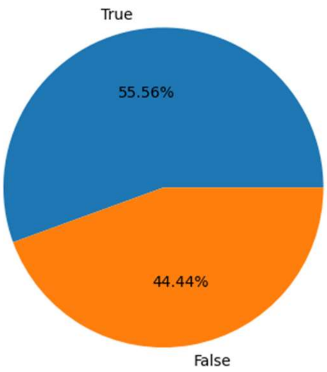
3.2 Exploratory Data Analysis (EDA)

- Descriptive statistics, including mean, median, and standard deviation, were calculated for important variables to understand their distributions.

- Visualizations, such as Pie charts, were employed to explore relationships between features and uncover potential patterns.



[Sex : Male vs Female]



[Heart Disease : True vs False]

4. Model Development and Evaluation

The dataset was divided into a training set (80%) and a test set (20%) using a random state of 42 for reproducibility.

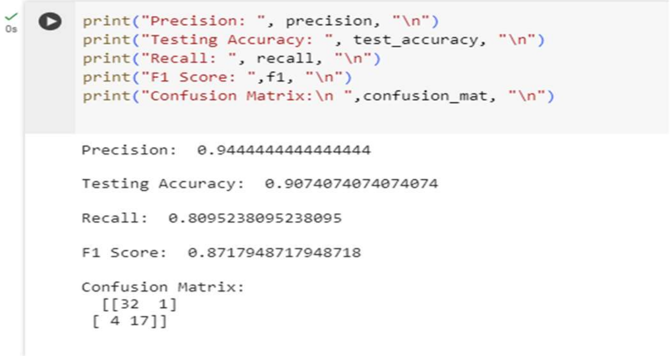
- 4.1 A logistic regression model was trained on the training set to establish the relationship between features and heart disease risk.
- 4.2 The trained model was subsequently evaluated on the independent test set to assess its performance in predicting unseen data.
- 4.3 Evaluation metrics included accuracy, precision, recall, F1-score, and a confusion matrix.

4.4 Results

The logistic regression model achieved an overall accuracy of 90.74% on the test set.

- 4.4.1 Precision: 94.44% (low rate of false positives)
- 4.4.2 Recall: 80.95% (modest potential for false negatives)
- 4.4.3 F1-score: 87.18% (balanced assessment of precision and recall)

The confusion matrix illustrates a correct identification of 32 out of 33 patients without heart disease and 17 out of 21 patients with heart disease.



```
print("Precision: ", precision, "\n")
print("Testing Accuracy: ", test_accuracy, "\n")
print("Recall: ", recall, "\n")
print("F1 Score: ",f1, "\n")
print("Confusion Matrix:\n ",confusion_mat, "\n")
```

```
Precision:  0.9444444444444444

Testing Accuracy:  0.9074074074074074

Recall:  0.8095238095238095

F1 Score:  0.8717948717948718

Confusion Matrix:
  [[32  1]
 [ 4 17]]
```

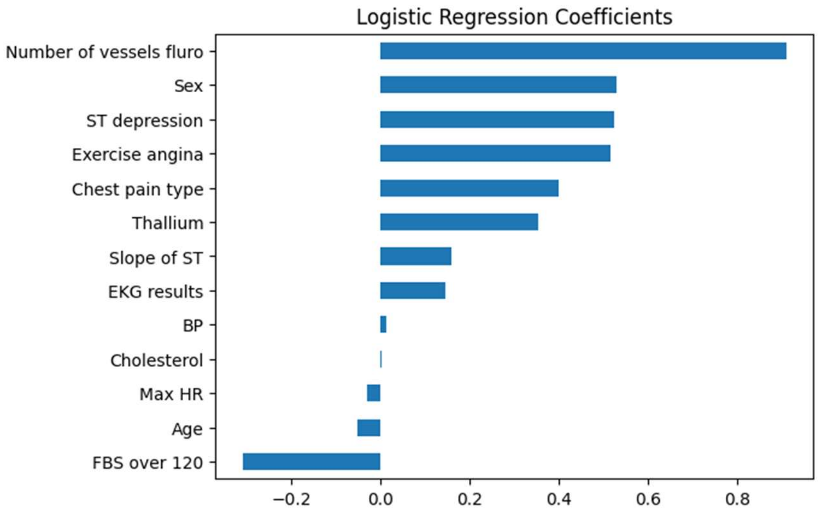
[Performance Metrices]

4.5 Model Coefficients and Feature Importance

Analysis of model coefficients revealed the most influential features in predicting heart disease risk.

- 4.5.1 Key risk factors identified included (list features with highest positive coefficients).

- 4.5.2 Visualization of coefficients using a horizontal bar chart provided a clear interpretation of feature importance.



[Logistic Regression Coefficients]

5. Conclusion

The logistic regression model demonstrated promising results in predicting heart disease risk, suggesting its potential as a valuable clinical decision-support tool. The model's high accuracy, precision, and balanced F1-score showcase its capability in identifying individuals at risk.