**Heart Disease Prediction**

Madhuri Basava

Bellevue University

Applied Data Science DSC680

Amirfarrokh Iranitalab

**Milestone 3**

**Topic:** The project I chose is Heart Disease Prediction.

This project focuses on improving Heart health by creating models that can predict the likelihood of the individual suffering from heart disease based on the given features.

**Business Problem**:

Heart disease is any condition that impairs the heart’s capacity to function normally. In recent years, it has become the leading cause of death in the world. Congestive heart failure (CHF) prevalence is expected to rise by 46% by 2030 compared to 2012 rates.

Given a dataset containing various features related to an individual's health and lifestyle (e.g., age, sex, cholesterol levels, blood pressure, etc.), the task is to build a machine-learning model to predict the presence or absence of heart disease. The model will be trained on historical data and evaluated for its accuracy and reliability in predicting heart disease in unseen data.

**Background/History:**

Heart disease is one of the leading causes of death globally. Early detection and management are crucial for reducing mortality rates and improving the quality of life for individuals at risk. Traditional diagnostic methods often involve extensive medical tests, which can be time-consuming and expensive. Machine learning offers a promising approach to developing predictive models that can assist in the early identification of heart disease using readily available health data.

**Research Questions:**

1. Can we predict who is more prone to heart disease with the values obtained from our dataset?
2. What aged people are more susceptible to heart health problems?
3. Is gender related to heart health issues? Which gender is more affected?
4. How is the individual’s blood pressure related to heart health?
5. Is Heart disease risk more for individuals with high cholesterol problems?
6. How is the individual’s blood sugar level related to heart health?
7. Does chest pain type influence the risk of heart disease?
8. Do electrocardiographic test results suggest the likelihood of heart disease?
9. How is heart disease related to Max heart rate?
10. Is Exercise-induced angina related to heart disease?

**Dataset:**

The dataset is taken from the Kaggle website.

[Heart Disease | EDA and Classification | ACC:94.5% (kaggle.com)](https://www.kaggle.com/code/michaelxyjonathan/heart-disease-eda-and-classification-acc-94-5/input?select=heart_statlog_cleveland_hungary_final.csv)

**Data Description**: There are a total of 12 fields.

Age: The age of the individual.

Sex: Gender of the individual.

Chest pain type: type of chest pain experienced.

Resting bp s: resting Blood Pressure in mm Hg

Cholesterol: Serum Cholesterol in mg/dl

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

resting ecg: Results of the resting electrocardiographic test.

max heart rate: Maximum heart rate achieved during exercise.

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

old peak: ST depression induced by exercise relative to rest.

stslope: Slope of the peak exercise ST segment.

Target: Presence of heart disease (1 = yes, 0 = no).

**Methods and analysis:**

I chose to do the prediction analysis with three different models as below.

1. K-Nearest Neighbors Model, and
2. Naïve Bayes Model
3. Logistic Regression

I chose these models to improve accuracy and reduce false positives and false negatives (Inaccurate predictions).

I plan to evaluate the results by calculating the Area under the ROC curve.

It measures the overall performance of the binary classification model. As both TPR (True Positive Rate) and FPR (False Positive Rate) range between 0 to 1, the area will always lie between 0 and 1, and a greater value of AUC denotes better model performance.

I plan to follow the below steps to evaluate these results.

1. Load the data set into a data frame.
2. Perform the Exploratory Data Analysis to understand the characteristics of the data set.
3. Clean the data set. Remove the unnecessary features.
4. Evaluate the correlation between the variables in the dataset.
5. A screenshot of a graph

   Description automatically generated

A screenshot of a graph

Description automatically generated

1. Divide the data set into a train and test data set and apply various models.
2. Create a confusion matrix to show the performance of each model to evaluate the predicted values from the model vs. the actual values from the test dataset.
3. Calculate the AUC-ROC for each model and choose the best model.

Confusion Matrix for the 3 Models:

1. K-Nearest Neighbors Model

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

A graph showing the true positive rate

Description automatically generated

1. Naive Bayes Model

A screenshot of a computer

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

A graph of a positive curve

Description automatically generated

1. Logistic Regression

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated

A graph showing the value of a positive rate

Description automatically generated

Conclusion:

From the above results, we can infer that the accuracy of the Logistic Regression is around 0.92 which is greater than the accuracy of the K-Nearest Neighbors Model, which is 0.89, and the Naïve Bayes Model which is 0.84.

Also, the Area Under the Curve (AUC) value for KNN is 0.9430, the AUC for Naïve Bayes Model is 0.9321, and the AUC of Logistic regression is 0.9711.

So, we can choose Logistic Regression for modeling.

Also, by analyzing the above Heat Map and its correlation with the Heart Attack chart, we can answer most of our above questions.

1. Increasing age is a major risk factor for heart disease. The arteries tend to narrow and harden over time, increasing the risk of coronary artery disease, heart attack, and stroke.
2. Downsloping Peak Exercise ST Segment (ST slope 0.55): The presence of a downsloping peak exercise ST segment in an ECG report is associated with a higher probability of heart disease. This feature might be indicative of ischemia, which reduces blood flow to the heart.
3. Exercise-Induced Angina (exercise angina, 0.49): Exercise-induced angina is associated with a lower likelihood of heart disease. Similar to typical angina, individuals experiencing chest pain during exercise are more likely to seek early medical intervention, reducing the risk of advanced heart disease.
4. Non-anginal Chest Pain (chest pain, 0.47): Non-anginal chest pain also shows a significant positive correlation with heart disease. This type of chest pain is often mistaken for indigestion or muscle pain, possibly leading to delayed diagnosis and treatment.
5. ST Depression Induced by Exercise Relative to Rest (oldpeak, 0.40): ST depression induced by exercise, a sign of possible heart stress, shows a negative correlation with heart disease. This could suggest effective treatment and management of patients with this symptom, decreasing the likelihood of severe heart disease.
6. Maximum Heart Rate Achieved (-0.40): A high maximum heart rate achieved during testing is associated with a higher likelihood of heart disease. A high heart rate during exercise could reflect an underlying stress on the heart, which might indicate some form of cardiovascular disease.

**Assumptions:**

It is assumed that the data is representative of the population in the data set and is correctly recorded. Handling of missing data through imputation or other techniques is assumed to be correct. Also, assumed that the selected features have good predictive power.

The selected models do not introduce any biases and are enough to achieve accurate results

**Limitations:**

The dataset might not be representative of the entire population due to sampling bias

The dataset may contain errors. The models may perform well on training data but may be poorly on unseen data. Handling sensitive data involves strict privacy requirements. Healthcare professionals should understand the model and use it effectively.

**Challenges/Issues:**

* A common problem in medical datasets is where the number of patients with heart disease may be significantly lower than those without, leading to an imbalanced dataset.
* High risk of overfitting due to the small size of this dataset.
* Ensuring that the model does not propagate biases present in the training data, which can lead to unfair treatment of certain groups, is challenging.

**Future Uses/Additional Applications:**

This project has the potential to extend beyond simple risk prediction, offering valuable applications in preventive healthcare, personalized treatment, public health research, remote monitoring, employee wellness programs, education, and pharmaceutical research. By leveraging the insights and capabilities of the model, various stakeholders can improve outcomes, and also improve the quality of care and services provided.

**Recommendations:**

The recommendations for this project are to maintain high-quality, representative data through rigorous cleaning and augmentation, and to focus on relevant feature selection and engineering. Experiment with multiple models and ensemble methods. Develop a user-friendly interface for healthcare providers, integrating real-time monitoring and a continuous retraining pipeline. Protect patient privacy by adhering to regulations and implementing robust security measures, while regularly auditing the model for biases to ensure fairness and transparency. This comprehensive approach will enhance the model’s reliability, utility, and ethical compliance, ultimately supporting better clinical decision-making and patient outcomes.

**Implementation Plan:**

To implement the heart disease prediction project, start by collecting and cleaning high-quality, representative data, ensuring it is anonymized and compliant with privacy regulations. Perform feature selection and engineering based on domain knowledge and statistical techniques, followed by experimenting with various machine learning models and ensemble methods, using cross-validation to ensure performance and avoid overfitting. Develop an interpretable model and deploy it using scalable cloud services, creating a user-friendly interface for healthcare providers to input patient data and receive predictions with explanations. Implement continuous monitoring and a retraining pipeline to maintain model accuracy, and regularly audit the model for biases to ensure fairness. Document the process thoroughly and provide transparency regarding the model's limitations and ethical considerations to build trust and support in clinical settings.

**Ethical Considerations:**

Limitations in accuracy involving false positives and false negatives can lead to incorrect predictions, which may cause unnecessary stress for families or overlook cases that need attention. Issues such as consent, data privacy, and the rights of individuals and families may also be a risk. privacy concerns, especially if the data used for predictions is not adequately protected. Protecting sensitive information is important to prevent unauthorized access or misuse. Acquiring informed consent becomes a crucial ethical consideration. Individuals and families need to be notified about the limitations, implications, and potential consequences of the prediction.

**References**

* *Heart Health prediction dataset is retrieved from the Kaggle website:*

[*Heart Disease | EDA and Classification | ACC:94.5% (kaggle.com)*](https://www.kaggle.com/code/michaelxyjonathan/heart-disease-eda-and-classification-acc-94-5/input?select=heart_statlog_cleveland_hungary_final.csv)

* *Article on heart disease:* [*Enhancing heart disease prediction using a self-attention-based transformer model | Scientific Reports (nature.com)*](https://www.nature.com/articles/s41598-024-51184-7)
* [Heart Disease Prediction using Machine Learning | Aman Kharwal (thecleverprogrammer.com)](https://thecleverprogrammer.com/2020/11/10/heart-disease-prediction-using-machine-learning/)