Heart Disease Detection

By Sagar Patel

Abstract

This research paper analyses data on heart patients across U.S. states using data analysis and machine learning classification algorithms to predict whether a patient has heart disease based on their health reports. Various Factors such as blood pressure, cholesterol, body mass index (BMI), and glucose levels are considered. Additionally, exploratory data analysis (EDA) is applied using visual charts to derive meaningful insights from the data. Python programming is utilized for data analysis and machine learning model implementation.

- **Dataset & Code Files**

This dataset contains 10,000 records of patient data related to heart disease. It aims to aid in the exploration and modeling of cardiovascular health risks and treatment outcomes. The data includes various medical and demographic features, making it an excellent resource for developing predictive models and conducting health analytics.

- **Key Features:**

- **Blood Pressure (mmHg):** Hypertension ranges.
- **Cholesterol (mg/dL)**
- **Body Mass Index (BMI)**
- **Glucose Level (mg/dL)**
- **Heart Disease:** Binary indicator (0 = No, 1 = Yes).
- **State:** Contains only from the USA.
- **Age**
- **Hospital**
- **Treatment**

The study focuses on how we can predict the likelihood of someone developing or already suffering from heart disease based on these factors. Furthermore, the research highlights how machine learning & numbers can aid in the identification of such diseases, contributing to advancements in medical science and improving patient care.

- **Literature Review:**

1. The first step in my research, conducted using Python programming in Jupyter Notebook, was to import necessary libraries. For Mathematical calculations, use the **NumPy** library. To handle and manipulate the dataset, utilized **Pandas**. For creating charts and visuals, I relied on **Matplotlib** and **Seaborn**. This libraries from the foundation for data analysis and visualization in Python.

2. The second step involved data cleaning and validation. This process included checking for any missing (null) or duplicate values to ensure the data integrity. Additionally, examined the unique values in each column to understand the data's structure. The describe() function was used to generate a statistical overview of numerical columns, providing key metrics such as min and max value, mean, and standard deviation. This step helps in identifying potential issues and understanding the overall distribution of the data.

3. The third step involved creating new features from the values in columns such as blood pressure, cholesterol, body mass index (BMI), and glucose level. Based on medical science guidelines, these values were categorized into meaningful indicators like "Low", "Normal", and "High". By transforming raw numerical data into categorized labels, making it easier to derive

insights and create more informative charts and visuals. This step plays a key role in simplifying complex data and improving the clarity of the analysis.

4. The fourth step was exploratory data analysis (EDA), where I began with several visuals to gain insights from the data:
    1. Distribution Plot: To understand the spread of numeric columns like blood pressure, cholesterol, BMI, and glucose level by visualizing the range from minimum to maximum values and their frequency counts.
    2. Pair Plot: This visual helped analyze how different columns relate to each other and how the dependent variable heart disease is affected by each feature through various scatterplots.
    3. Box plot: Mostly used for detecting outliers, this revealed some outliers in the blood pressure, cholesterol, and BMI column. In our case, I choose not to remove them, as having only high BP, cholesterol, and BMI does not necessarily, a person suffering from heart disease.
    4. Line Chart for New Features: Based on categorized features, create a line chart to show which category of people have the highest chance of heart disease.
    5. Count Plot of Heart Disease by Hospital: Helped to identify which hospitals have the highest number of heart disease patients, offering useful insights into healthcare trends.
    6. Geo Map Using Folium: Visualised which of the states have the highest number of heart patients, providing a geographic perspective on heart disease distribution.
    7. Correlation Heat Map: Illustrated how different features related to each other, providing valuable insights for feature selection to make predictions for the later stages.
5. The fifth step, perform label encoding on the categorical features created in 3rd step, This transformation will convert categorical levels into numeric values, making them easier to analyze and include in the correlation heat map for feature selection.
6. In the sixth step, using a correlation heat map choose the best features, such as blood pressure, cholesterol, BMI, and glucose level. Then, split the dataset into training and testing sets, with a 20% test size. Setting a random_state for the shuffling process and applying StandardScaler will standardize the features, enhancing model performance by ensuring the data's consistency.
7. The seventh step is, to apply **Machine Learning Classification Algorithms** to predict the Dependent variable heart disease using four features (Support value = 2000):

Machine Learning Classifiers	Accuracy	True +ve	False +ve	False -ve	True -ve
Logistic Regression	77%	776	252	215	757
Decision Tree Classifier	92%	942	86	78	894
Random Forest Classifier	94%	953	75	50	922
Support Vector Machine Classifier	93%	936	92	47	925
K-Nearest Neighbour Classifier	95%	964	64	45	927
XG Boost Classifier	94%	954	74	49	923

- **Conclusion:** K-Nearest Neighbour Classifier gives the highest predictive accuracy of 95%. This high accuracy indicates that KNN is particularly adept at recognizing patterns within the dataset & effectively differentiating between patients with and without heart disease. The result demonstrates the significant potential of machine learning in medical diagnostics, where such predictive tools can provide valuable support to healthcare professionals. This study highlights the role of feature selection, data processing, and classification methods in developing accurate prediction models that could contribute to preventive measures and improved patient outcomes in the field of cardiology.