Heart Disease Prediction Using Machine Learning

S Muskan, R.Sethu madhavi
Student, Assistant Professor Computer Science and Engineering, Computer Science and Engineering Reva University, Bengaluru, India, Reva University, Bengaluru, India.

Abstract: The number of heart disease cases is rising quickly every day, making it crucial and worrisome to anticipate any such illnesses in advance. This diagnosis is a challenging task that requires accuracy and efficiency. The study article primarily focuses on identifying patients who, given a variety of medical characteristics, are more likely to suffer heart disease. Using the patient's medical history, we developed a heart disease prediction algorithm to determine the likelihood of a heart disease diagnosis or not. Utilizing various machine learning methods, including logistic regression and KNN, we were able to predict and categorize the patient with heart disease. A very useful method was employed to control how the model can be applied to increase the precision of a person's heart attack prediction. The suggested model's strength was quite pleasing; it could accurately detect signs of heart illness in a specific person using KNN and Logistic Regression, outperforming other classifiers like Naïve Bayes, among others, with a good degree of accuracy. Thus, by utilizing the provided model to determine the likelihood that the classifier can correctly and precisely diagnose cardiac illness, a sizable amount of pressure has been released. The Given heart disease prediction system lowers costs and improves medical care. We get important information from this experiment that will aid in the prediction of heart disease patients.

I. INTRODUCTION

Heart disease refers to various conditions affecting the heart and is the leading cause of death globally, accounting for 17.9 million deaths annually, according to the WHO. Key risk factors include high cholesterol, obesity, high triglycerides, and hypertension. The American Heart Association lists symptoms like irregular heartbeat, swollen legs, and rapid weight gain, which complicate diagnosis, especially in older adults, as they can mimic other conditions. Advances in research and access to patient data have enabled the use of machine learning (ML) and artificial intelligence (AI) to improve heart disease diagnosis. Different ML and deep learning models have been used to classify and predict heart disease. For example, Melillo et al. applied the CART algorithm, achieving 93.3% sensitivity and 63.5% specificity in detecting congestive heart failure. Rahhal et al. enhanced performance using deep neural networks with ECG data, while Guidi et al. developed a clinical decision support system using multiple ML models, achieving 87.6% accuracy with random forest and CART. Zhang et al. combined natural language processing with a rule-based method to achieve 93.37% accuracy in classifying heart failure from clinical notes. Parthiban and Srivatsa applied SVM techniques to predict heart disease in diabetic patients with 94.60% accuracy. Challenges like high data dimensionality are tackled through feature engineering and selection, boosting classification and prediction accuracy. Dun et al.'s research found neural networks achieved 78.3% accuracy in detecting heart disease, with other models like logistic regression, SVM, and ensemble techniques performing well. Singh et al. used generalized discriminant analysis and extreme learning machines to reach 100% accuracy in detecting coronary heart disease. Yaghouby et al. achieved similar success in classifying arrhythmias with multilayer perceptron neural networks. Dimensionality reduction techniques like PCA are key to managing high-dimensional data, improving model performance and processing time. Rajagopal and Ranganathan's study utilizing PCA and neural networks for cardiac arrhythmia classification achieved a 99.83% F1 score. Many studies use a 13-feature Cleveland dataset, consistently showing high accuracy across models. Heart disease is more common in men, who are twice as likely to experience a heart attack. The Cleveland dataset, first established in 1988, remains a benchmark for heart disease prediction studies, showing promising results across different databases. Researchers continue to refine ML and AI models to improve heart disease diagnosis, aiming for better patient care and reduced healthcare costs.

1.1 Problem statement and Objectives of statement

i. There is ample related work in the fields directly related to this paper. ANN has been introduced to produce the highest accuracy prediction in the medical field. The back propagation multilayer perception (MLP) of ANN is used to predict heart disease. The obtained results are compared with the results of existing models within the same domain and found to be improved

ii. Objectives Of statement: For healthcare providers and medical researchers, our heart disease prediction system leverages advanced machine learning algorithms to accurately predict the likelihood of heart disease in patients based on their medical history and clinical data. Unlike traditional diagnostic methods, our solution uses cutting-edge techniques such as logistic regression and K-Nearest Neighbours to provide precise, efficient, and early detection of heart disease. This empowers healthcare professionals to make informed decisions, improve patient outcomes, and reduce healthcare costs through proactive and preventive care.

This product positioning statement clearly defines the target market (healthcare providers and medical researchers), the product (heart disease prediction system using machine learning), its unique value (accurate, early detection, and efficient diagnostics), and how it differs from traditional methods (advanced machine learning techniques). It ensures that all teams within the company are aligned on the product's value proposition when communicating with potential users or customers.

1.2 Contribution to the project

Using machine learning models, we predict heart disease with Python's Scikit-Learn. We secure data with certificate authorities and elliptical curve cryptography. Hyper ledger Fabric forms our block chain base, with Node.js smart contracts managing data integration. A block chain explorer visualizes data transparency, while a Node.js app verifies system functionality. This version distils the key elements of the project into a succinct description, focusing on the technologies used and their contributions to the overall system.

1.3 Organization of Report

Introduction

Heart disease is one of the leading causes of death globally, making its early detection crucial. This project aims to develop a predictive model for heart disease using machine learning techniques. The main objectives are to analyze patient data, identify significant features, and build a model that can predict the likelihood of heart disease.

Literature Review

Numerous studies have been conducted on heart disease prediction using various machine learning algorithms. Commonly used techniques include logistic regression, decision trees, support vector machines, and neural networks. These studies highlight the importance of feature selection and data preprocessing in improving model accuracy. Additionally, ensemble methods like Random Forest and Gradient Boosting have shown promising results.

Methodology

Data Collection:

- The dataset used in this project was obtained from [source, e.g., UCI Machine Learning Repository]. It contains [number] instances and [number] attributes related to patient health metrics.

Data Preprocessing:

- Handling missing values, normalizing numerical features, and encoding categorical variables.
- Splitting the dataset into training and testing sets (e.g., 80/20 split).

Feature Selection:

- Identifying the most relevant features using techniques like correlation analysis and feature importance scores.

Model Selection:

- Experimenting with various algorithms such as Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines.

Evaluation Metrics:

- Using metrics like accuracy, precision, recall, F1-score, and AUC-ROC to evaluate model performance.

Experiments and Results:

- Initial model performance using default parameters.
- Optimizing model parameters using techniques like Grid Search or Random Search.
- Comparing the performance of different models.

Results:

- Presenting the results in the form of tables and graphs.

- Discussing the model's performance in terms of accuracy, precision, recall, F1-score

II. LITERATURE SURVEY

The book *"Machine Learning with Python: Design and Develop Machine Learning and Deep Learning Techniques using Real World Code Examples"* [1] by Abhishek Vijayvargia, published in 2019, offers practical insights into applying machine learning techniques with Python, specifically focusing on cardiovascular datasets commonly used for disease prediction. Another study, *"Prediction of Heart Disease using Machine Learning Algorithms"* [2] by Mr. Santhana Krishnan J. and Dr. Geetha S., published in 2019, highlights the use of decision tree and naive Bayes algorithms. These methods were applied to predict heart disease using features like age, sex, chest pain type, blood pressure, cholesterol levels, ECG results, maximum heart rate, exercise-induced angina, ST depression, and fluoroscopy results.

In the book *"Machine Learning for Beginners: The Definitive Guide to Neural Networks, Random Forests, and Decision Trees"* [3] by Jennifer Grange, published in 2017, fundamental machine learning algorithms such as neural networks, random forests, and decision trees are discussed, particularly for classification and regression tasks. These techniques are essential for building models that predict disease outcomes based on clinical data.

The paper *"A Data Mining Model for Predicting Coronary Heart Disease using Random Forest Classifier"* [4] by A. S. Abdullah and R. R. Rajalaxmi, presented at an international conference in 2012, used the random forest classifier to develop a predictive model for coronary heart disease. This model employed data mining techniques to analyze relevant health indicators and make accurate predictions.

Additionally, the research *"Using PSO Algorithm for Producing Best Rules in Diagnosis of Heart Disease"* [5] by A. H. Alkeshuosh et al., presented at the International Conference on Computer Applications (ICCA) in September, utilized the Particle Swarm Optimization (PSO) algorithm. The PSO algorithm was employed to generate optimal diagnostic rules for heart disease, improving the accuracy and efficiency of predictions by intelligently selecting the best set of rules.

In modern heart disease prediction, the integration of advanced techniques such as deep learning models, including convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has also gained traction. These models can effectively handle time-series data, such as ECG signals, and extract meaningful patterns for more accurate predictions. Furthermore, the use of ensemble learning, which combines multiple models to boost performance, has proven beneficial in improving prediction accuracy. Finally, feature engineering and the application of dimensionality reduction techniques like t-SNE and PCA have become crucial in managing large, complex datasets, enhancing the performance of models while reducing processing time.

III. PROPOSED WORK

This work outlines the methodology and performance evaluation metrics used in heart disease prediction. It provides a summary of the key components.

3.1 Methodology

In this project, a machine learning model was developed to enhance the accuracy of heart disease prediction. Initially, datasets were collected and divided into training and testing sets. Various algorithms, including SVM, KNN, Naïve Bayes, logistic regression, and Random Forest, were applied to analyze the data. Key features were identified to improve prediction accuracy. The performance of these algorithms was assessed using metrics like accuracy, sensitivity, specificity, and precision. Data visualization techniques were employed to present the results visually. Among the algorithms, Random Forest achieved the highest accuracy at 80%. This machine learning approach aims to detect heart disease at an early stage, improving patient outcomes.

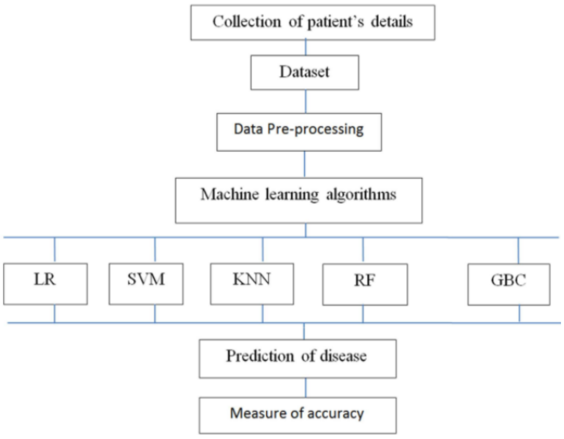


Figure 1. Architecture of prediction models.

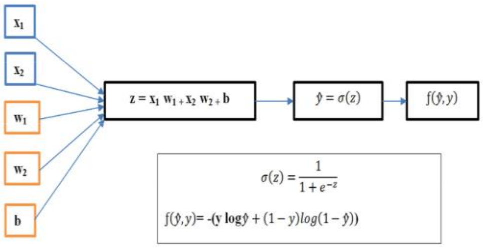

x_1
x_2
w_1
w_2
b
$z = x_1 w_1 + x_2 w_2 + b$
$\hat{y} = \sigma(z)$
$f(\hat{y}, y)$
$\sigma(z) = \frac{1}{1+e^{-z}}$
$f(\hat{y}, y) = -(y \log\hat{y} + (1-y)log(1-\hat{y}))$

Figure 2. Logistic Regression model.

Logistic regression, often used for classification and predictive analysis, is applied to estimate binary outcomes (e.g., an event occurring or not, represented by 1 or 0) from independent variables. Below is the process of how the logistic regression model operates.

3.2 Performance Evaluation Techniques

Several key metrics are vital in evaluating the performance of a heart disease prediction model using machine learning. The confusion matrix summarizes true positives, true negatives, false positives, and false negatives. Accuracy represents the proportion of correct classifications, while precision (positive predictive value) measures the accuracy of positive predictions. Recall (sensitivity) evaluates the model's effectiveness in identifying actual positives, and the F1 score balances precision and recall. Specificity (true negative rate) assesses the proportion of correctly identified negatives.

The ROC curve and the area under the ROC curve (AUC-ROC) illustrate the trade-off between true positive and false positive rates across different thresholds. Cross-validation, such as k-fold, ensures the robustness of the model by evaluating its performance on different data subsets. The Precision-Recall curve is particularly useful for imbalanced datasets, while the Matthews Correlation Coefficient (MCC) provides a balanced assessment by considering all confusion matrix components.

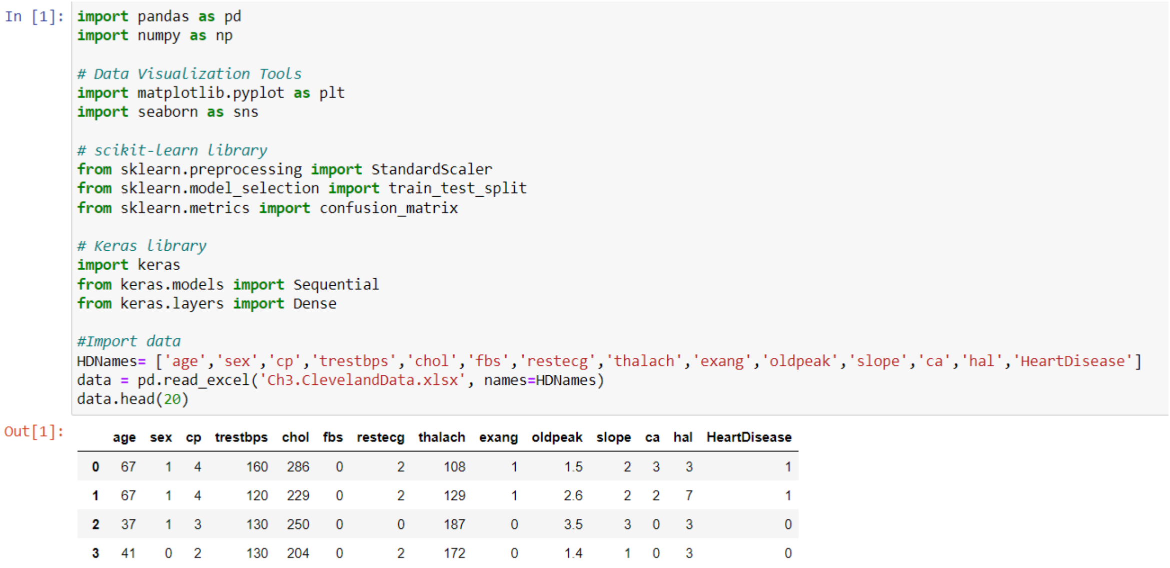
Additionally, Cohen's Kappa measures agreement while accounting for chance, and Logarithmic Loss evaluates the model's probability-based predictions. These metrics provide a comprehensive evaluation, ensuring the model's accuracy, reliability, and overall effectiveness in predicting heart disease.

IV. RESULTS AND DISCUSSION

- The dataset consists of 1,025 observations, with 54.3% of patients diagnosed with heart disease and 45.7% not affected. Males are more prevalent in the dataset and exhibit a higher rate of heart disease than females.
- Key features influencing heart disease prediction include age, cholesterol levels, chest pain type (cp), maximum heart rate (thalach), and the slope of the peak exercise ST segment. Males aged 55-68 with cholesterol levels between 200-300 mg/dl have a higher likelihood of heart disease.
- Positive correlations with heart disease were observed for factors like chest pain type (cp), maximum heart rate (thalach), and slope, with specific ranges of these features linked to the presence or absence of heart disease.
- Five machine learning algorithms were tested: Logistic Regression (LR), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), and Gradient Boosting Classifier (GBC). Logistic Regression outperformed others with an accuracy of 95%.
- Performance metrics, including the confusion matrix, showed that Logistic Regression achieved the highest true positive and true negative rates. The F1-score, recall, and precision were also 95%, making it the most reliable model for predicting heart disease in this study.

4.1 Implementation

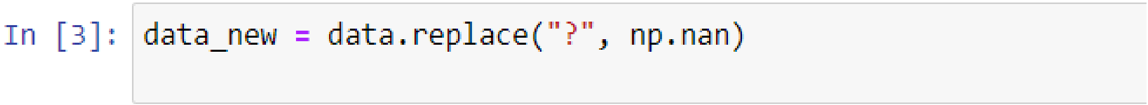
- Collect heart disease datasets from trusted sources like the UCI Machine Learning Repository.
- Pre-process the data by addressing missing values, normalizing numerical features, encoding categorical variables, and splitting the data into training and testing sets.
- Identify and select features that have a significant impact on heart disease prediction.
- Create or transform features to enhance the model's predictive performance.
- Select appropriate machine learning algorithms such as Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting.
- Train the models on the training dataset and fine-tune hyper parameters using grid search or randomized search techniques.
- Evaluate the models on the testing dataset.
- Use performance metrics like accuracy, precision, recall, F1 score, ROC-AUC, and the confusion matrix to assess the effectiveness of the models.



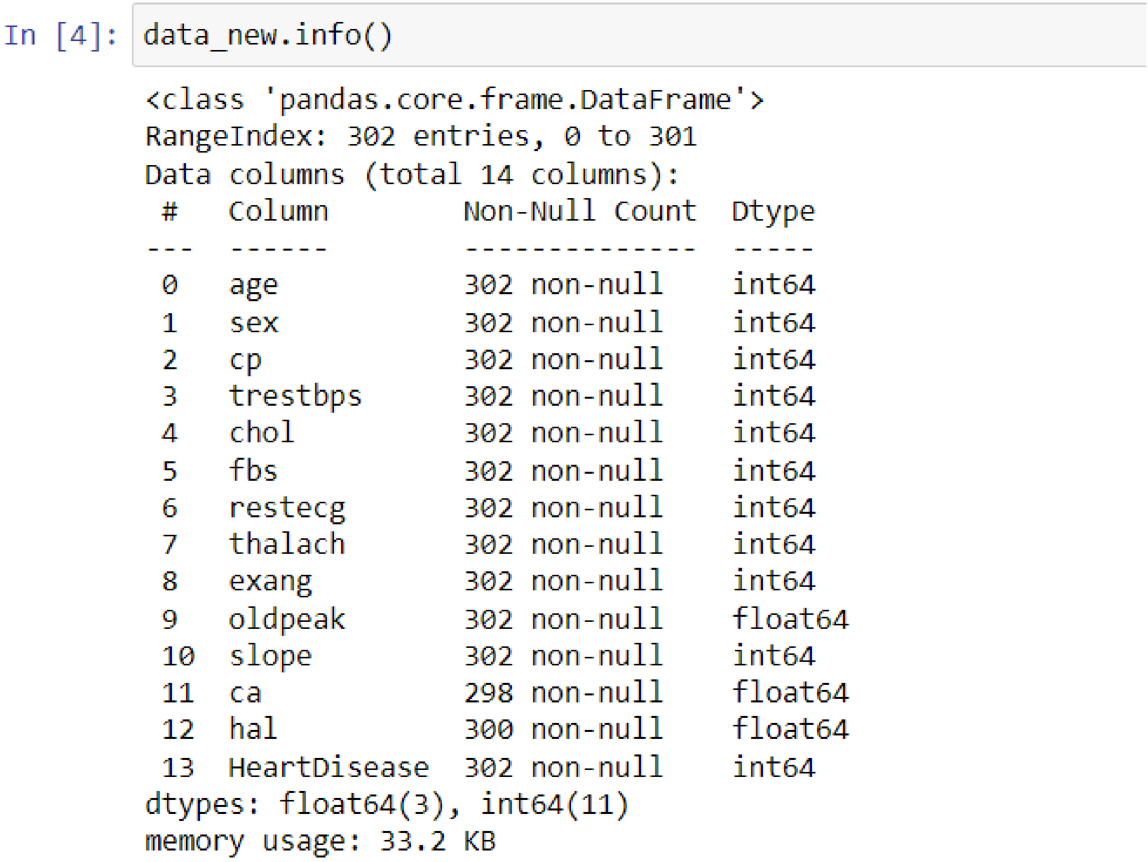
```
In [1]: import pandas as pd
        import numpy as np

        # Data Visualization Tools
        import matplotlib.pyplot as plt
        import seaborn as sns

        # scikit-learn library
        from sklearn.preprocessing import StandardScaler
        from sklearn.model_selection import train_test_split
        from sklearn.metrics import confusion_matrix

        # Keras library
        import keras
        from keras.models import Sequential
        from keras.layers import Dense

        #Import data
        HDNames= ['age','sex','cp','trestbps','chol','fbs','restecg','thalach','exang','oldpeak','slope','ca','hal','HeartDisease']
        data = pd.read_excel('Ch3.ClevelandData.xlsx', names=HDNames)
        data.head(20)
```

Out[1]:

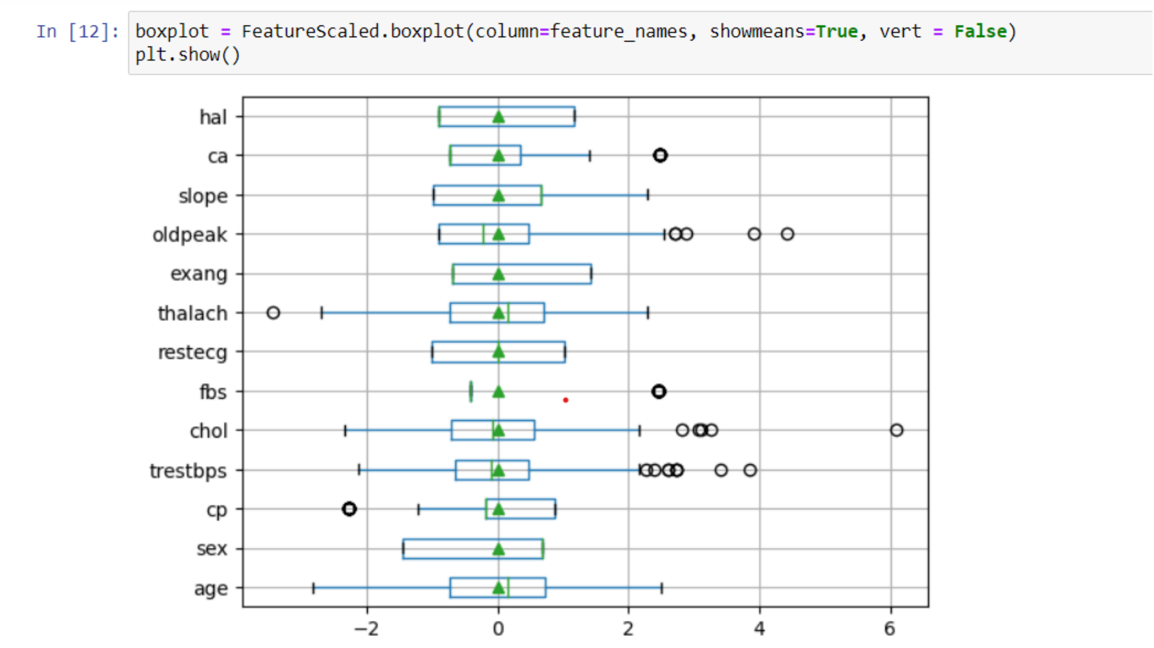
	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	hal	HeartDisease
0	67	1	4	160	286	0	2	108	1	1.5	2	3	3	1
1	67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
2	37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
3	41	0	2	130	204	0	2	172	0	1.4	1	0	3	0



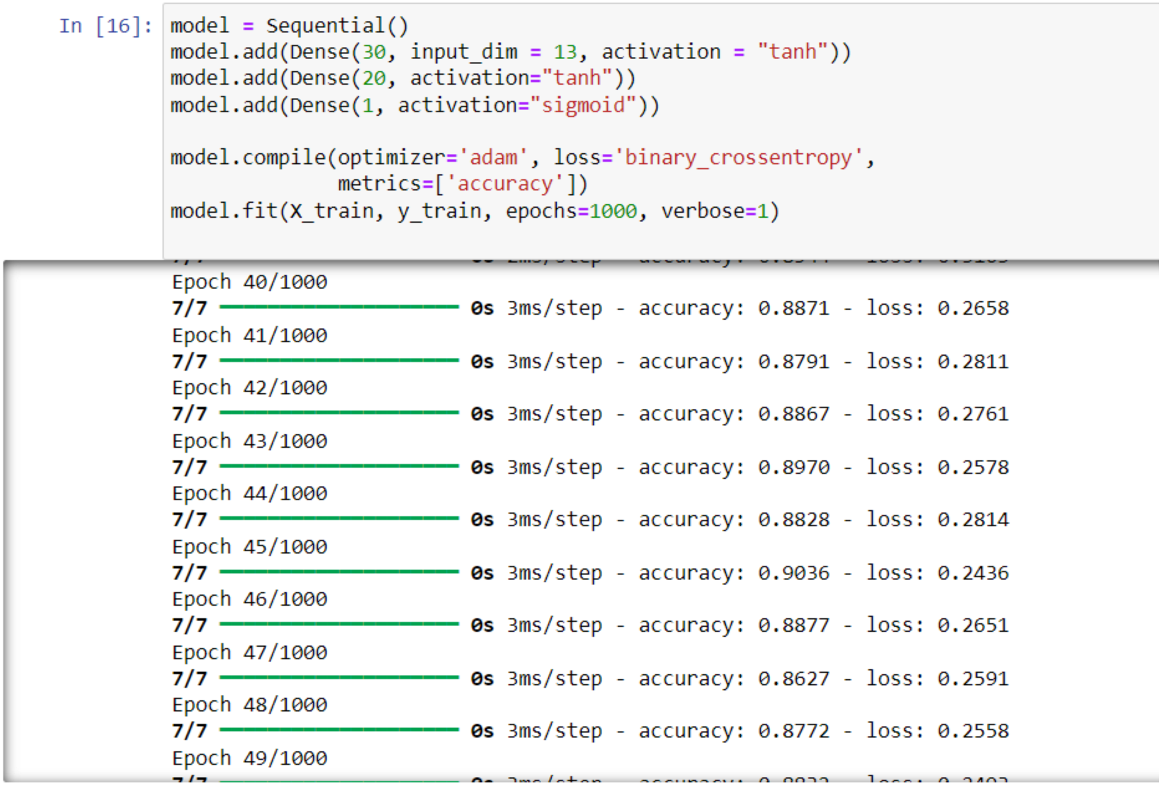
```
In [3]: data_new = data.replace("?", np.nan)
```



```
In [4]: data_new.info()

        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 302 entries, 0 to 301
        Data columns (total 14 columns):
         #   Column        Non-Null Count  Dtype  
        ---  ------        --------------  -----  
         0   age           302 non-null    int64  
         1   sex           302 non-null    int64  
         2   cp            302 non-null    int64  
         3   trestbps      302 non-null    int64  
         4   chol          302 non-null    int64  
         5   fbs           302 non-null    int64  
         6   restecg       302 non-null    int64  
         7   thalach       302 non-null    int64  
         8   exang         302 non-null    int64  
         9   oldpeak       302 non-null    float64
         10  slope         302 non-null    int64  
         11  ca            298 non-null    float64
         12  hal           300 non-null    float64
         13  HeartDisease  302 non-null    int64  
        dtypes: float64(3), int64(11)
        memory usage: 33.2 KB
```

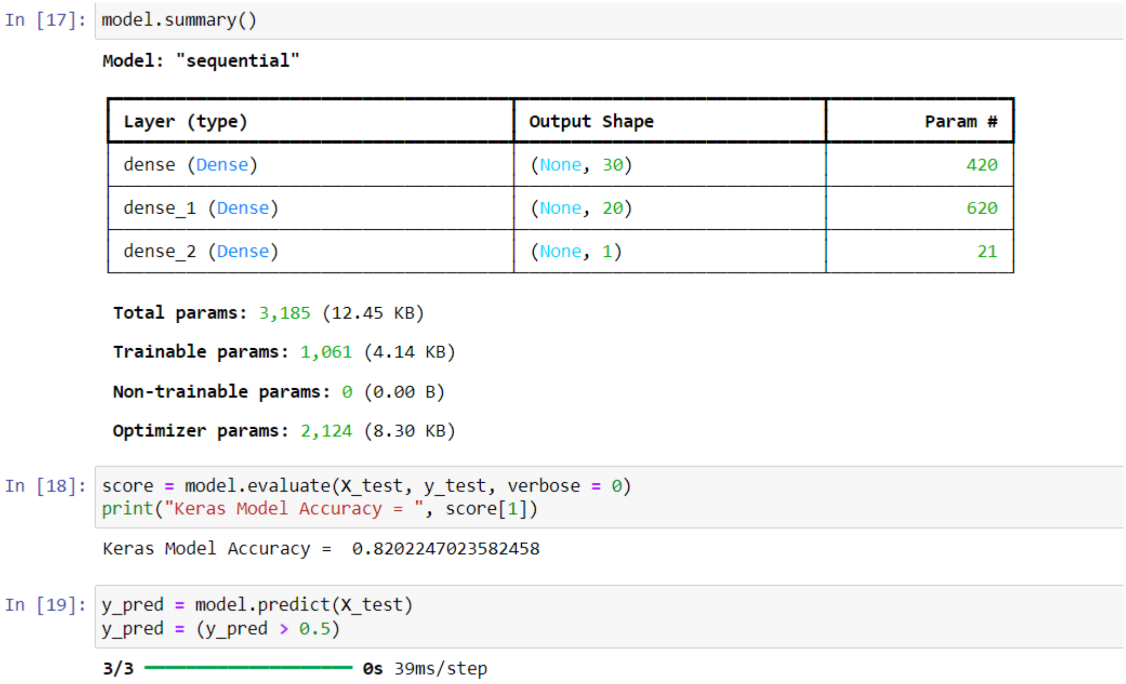


```
In [12]: boxplot = FeatureScaled.boxplot(column=feature_names, showmeans=True, vert = False)
         plt.show()
```



```
In [16]: model = Sequential()
         model.add(Dense(30, input_dim = 13, activation = "tanh"))
         model.add(Dense(20, activation="tanh"))
         model.add(Dense(1, activation="sigmoid"))

         model.compile(optimizer='adam', loss='binary_crossentropy',
                       metrics=['accuracy'])
         model.fit(X_train, y_train, epochs=1000, verbose=1)
```

```
Epoch 40/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8871 - loss: 0.2658
Epoch 41/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8791 - loss: 0.2811
Epoch 42/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8867 - loss: 0.2761
Epoch 43/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8970 - loss: 0.2578
Epoch 44/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8828 - loss: 0.2814
Epoch 45/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.9036 - loss: 0.2436
Epoch 46/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8877 - loss: 0.2651
Epoch 47/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8627 - loss: 0.2591
Epoch 48/1000
7/7 ━━━━━━━━━━━━━━━━━━━━ 0s 3ms/step - accuracy: 0.8772 - loss: 0.2558
Epoch 49/1000
```



In [17]:
```
model.summary()
```

```
Model: "sequential"

┏━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━━━━━━━━━━┳━━━━━━━━━━━━━━━┓
┃ Layer (type)                    ┃ Output Shape           ┃       Param # ┃
┡━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━━━━━━━━━━╇━━━━━━━━━━━━━━━┩
│ dense (Dense)                   │ (None, 30)             │           420 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_1 (Dense)                 │ (None, 20)             │           620 │
├─────────────────────────────────┼────────────────────────┼───────────────┤
│ dense_2 (Dense)                 │ (None, 1)              │            21 │
└─────────────────────────────────┴────────────────────────┴───────────────┘

 Total params: 3,185 (12.45 KB)

 Trainable params: 1,061 (4.14 KB)

 Non-trainable params: 0 (0.00 B)

 Optimizer params: 2,124 (8.30 KB)
```

In [18]:
```
score = model.evaluate(X_test, y_test, verbose = 0)
print("Keras Model Accuracy = ", score[1])
```

```
Keras Model Accuracy =  0.8202247023582458
```

In [19]:
```
y_pred = model.predict(X_test)
y_pred = (y_pred > 0.5)
```

```
3/3 ━━━━━━━━━━━━━━━━━━━━ 0s 39ms/step
```

V. CONCLUSION

Heart disease is an increasing global health issue, making early and accurate prediction essential for effective treatment and prevention. This study aims to build a machine learning model that reliably predicts heart disease, with a focus on accuracy metrics from the confusion matrix. Five algorithms were tested: Logistic Regression, Support Vector Machine, K-Nearest Neighbors, Random Forest, and Gradient Boosting Classifier. Logistic Regression proved to be the most accurate, achieving a 95% accuracy rate. This approach can be further applied to predict other conditions, such as cardiovascular disease, diabetes, breast cancer, tumors, and more, by utilizing historical data and incorporating additional machine learning techniques.

REFERENCES

[1] Victor Chang, Vallabhanent Rupa Bhavani, Ariel Qianwen Xu, MA Hossain. An artificial intellegence model for heart disease detection using machine learning. Healthcare Analytics, volume 2, November 2022, 100016.

[2] Ghumbre, S. U., & Ghatol, A. A. (2012). Heart disease diagnosis using machine learning algorithm. In Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012 (pp. 217-225). Springer, Berlin, Heidelberg.

[3] Khaled Mohamed Almustafa. Prediction of heart disease and classifiers sensitivity analysis. Almustafa BMC Bioinformatics (2020) 21: 278.

[4] Ordonez C (2006). Associate rule discovery with the train and test approach for heart disease prediction. IEEE Transaction on Information Technology in Biomedicine, 10 (2), 334-43.