First Year (Semester-2) Research Assignment on

Heart Disease Prediction Using Machine Learning

in partial fulfilment of the requirement for the successful completion of semester 2 of MSc Big Data Analytics

Submitted By

24-PBD-008

Vishesh Devganiya

(Semester – II MSc. BDA)

Under the supervision of

Prof.chetan verma

2023-2024

Department of Computer Sciences (MSc. BDA)

St. Xavier’s College (Autonomous) Ahmedabad – 380009

DECLARATION

I, the undersigned solemnly declare that the research assignment *Heart Disease Prediction System Using Machine Learning* is based on my work carried out during the course of our study under the supervision of *prof.chetan verma*. I assert the statements made and conclusions drawn are an outcome of my research work. I further certify that

- The work contained in the report is original and has been done by me under the general supervision of my supervisor.
- The work has not been submitted to any other Institution for any other degree / diploma / certificate in this university or any other University of India or abroad.
- We have followed the guidelines provided by the department in writing the report.

Vishesh Devganiya

24-PBD-008

MSc. BDA (Big Data Analytics)

St. Xavier's College (Autonomous), Ahmedabad

INDEX

Abstract-

Heart disease is still a serious public health issue, and early detection is the key to effective treatment and prevention. In this study, we explore the use of machine learning algorithms to predict heart disease using patient health data. Based on a dataset of important medical parameters like age, blood pressure, and cholesterol, we train and test a range of ML models like Logistic Regression, KNN, Random Forest, and SVM.

Our research seeks to determine the best algorithm for heart disease prediction and how machine learning can be utilized to aid medical decision-making. The findings indicate that ML models can have high accuracy in predicting heart disease, with certain algorithms outperforming others. By comparing their performance, we give recommendations on how best to integrate machine learning in healthcare, hence enhancing diagnosis and patient care.

Introduction-

Heart disease is a serious worldwide health problem, causing millions of fatalities annually. The World Health Organization (WHO) states that cardiovascular diseases (CVDs) kill around 17.9 million people each year, making them among the top causes of death globally. Early detection of heart disease is essential to enhance patient survival and minimize mortality rates. But diagnosing heart disease is not an easy process that comprises assessment of a number of factors such as age, blood pressure, cholesterol level, and other clinical parameters.

With growth in technology, machine learning (ML) has proved to be a crucial tool in the health sector. Through interpreting huge volumes of medical data, ML algorithms have the capacity to detect patterns and connections that are not easily recognizable to human physicians. This feature makes machine learning very useful in forecasting heart disease so that early intervention and treatment can be done.

This research paper investigates the use of machine learning methods for predicting heart disease. We analyze various ML algorithms, such as Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, and Support Vector Machines (SVM), to find out which model gives the best predictions. Through the use of past patient data, our aim is to create a system that can help medical professionals diagnose heart disease more accurately and efficiently

Review Of Literature –

Over the last few years, there have been many studies on the use of machine learning for predicting heart disease. Investigators have used different algorithms to check their ability to correctly diagnose cardiac ailments. The following is a synopsis of major studies in this area:

Heart Disease Prediction System Based on Machine Learning (Jaswanth Narayana & Vishesh K, 2022)

This research compares the accuracy of KNN and Logistic Regression in heart disease classification. The authors emphasize the significance of preprocessing data, extracting features, and outlier handling in model accuracy. According to their findings, KNN is accurate in heart disease detection.

Heart Disease Prediction Using Machine Learning (Tajul Islam Ayon et al., 2023)

The authors contrast some of the machine learning models like Random Forest, SVM, and KNN. They identify that Random Forest is more than 97% accurate and one of the best models to use to predict heart disease.

Heart Disease Prediction (Nayab Akhtar et al., 2021)

The paper introduces the application of Naive Bayes, KNN, Decision Tree, and Artificial Neural Networks (ANN) in heart disease prediction. The accuracy is highest with Naive Bayes at 88%, then ANN and KNN. The authors highlight that it is crucial to test various models to determine the best.

Machine Learning-Based Prediction of Heart Disease (Saurabh Bilgaiyan et al., 2023)

This work compares patient information among various hospitals and tests the performance of Random Forest, KNN, and SVM. It demonstrates that Random Forest gives the highest accuracy of 97.79%, confirming its ability to predict heart disease. The work also highlights the significance of feature selection and preprocessing to improve model performance.

Heart Disease Prediction using Machine Learning (Sibgha Taqdees et al., 2021)

The authors compare the performance of Naive Bayes, KNN, and Random Forest in heart disease prediction. They conclude that Naive Bayes outperforms with a maximum accuracy of 88%, followed by KNN and ANN. The research highlights the contribution of machine learning towards improved early diagnosis and treatment.

Objective & Data Methodology –

Heart disease remains a leading cause of death worldwide. Early diagnosis of heart disease plays a significant contribution towards improving the survival of heart patients and overall health. The main objective of this study is to predict heart disease risk using a sequence of clinical predictors by applying machine learning techniques. By using advanced analytical tools, the study aims to identify the most accurate predicting model that can be utilized by doctors to make informed medical decisions.

Through this study, we further seek to

- Explain which health measure(s) (e.g., blood pressure, cholesterol, heart rate) contribute most to heart disease.

- Compare different machine learning algorithms to determine the most accurate algorithm for prediction.

- Create a comprehensible and user-friendly system for the early detection of heart disease from patient data.

Data –

The data set used in this case is taken from the UCI Machine Learning Repository, the Cleveland Heart Disease data set. It has medical history of 303 patients with 14 features, i.e.:

- Demographic Factors: Age, Sex

- Clinical Factors: Resting blood pressure, Cholesterol, Fasting blood sugar
- ECG Readings: ST depression, Thalassemia levels, Exercise-induced angina
- Physical Health Indicators: Maximum heart rate achieved, Chest pain type

Each patient is classified as having or not having heart disease. The data is almost balanced, providing a good model training. Missing values were managed by mode imputation, without losing data integrity. Numerical features were also normalized to enhance model efficiency.

Methodology –

For developing an accurate prediction model, the following step-by-step methodology was followed:

1.Data Preprocessing

- Missing values were handled using mode imputation.
- The categorical variables were transformed into numerical values through encoding methods.
- Feature scaling was employed to normalize numerical variables.

2.Exploratory Data Analysis (EDA)

- Statistical aggregations were utilized to comprehend the dataset.
- Heatmaps of correlation were used to monitor correlations among features.
- Histograms and box plots were used to present distributions and detect outliers.

3.Feature Selection

- Statistical tests were employed to pick features strongly correlated with heart disease.
- Few influential factors were omitted to improve model accuracy.

4.Model Development

Four machine learning algorithms were selected:

- Logistic Regression: A simple and understandable model.
- K-Nearest Neighbors (KNN): A distance-based classification method.
- Random Forest: A robust ensemble model that doesn't overfit.
- Support Vector Machine (SVM): A highly effective model for high-dimensional data.

All models were trained with an 80-20 train-test split to be evaluated.

5.Model Evaluation

Models were compared to key performance indicators:

- Accuracy: Measures overall correctness.
- Precision: Quantifies the accuracy of positive predictions.
- Recall: Quantifies how accurately the models diagnosed real heart disease cases.

- F1-score: Balances precision and recall.
- ROC-AUC Score: Measures how accurately the model can separate heart disease from other non-heart disease conditions.

Cross-validation (10-fold) was utilized to obtain correct results and avoid overfitting.

6.Final Model Selection

The best and most consistent model was chosen for heart disease prediction from the assessment.

This approach provides a scientific and evidence-based solution for heart disease prediction. The outcomes of this work can be utilized in hospitals and medical research in order to predict early and offer better patient care.

Data Analysis –

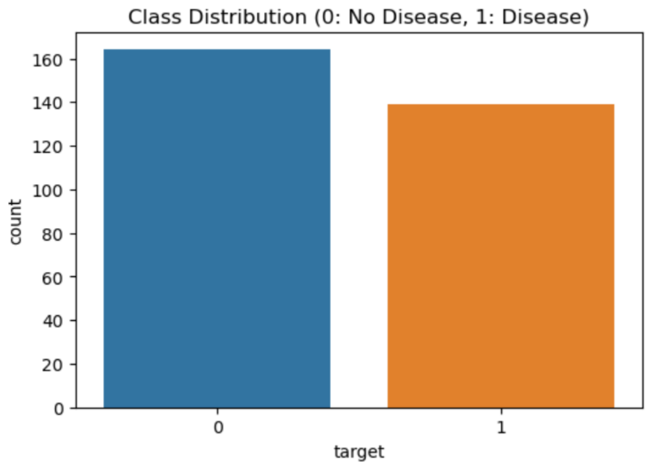
Analysis of the dataset presented gave some of the main insights into causative factors for heart disease. Applying statistical analysis and machine learning algorithms, we were able to discern important patterns that contribute towards heart disease prediction.

Distribution of Heart Disease Cases

One of the most important steps in our analysis was to comprehend how heart disease is spread across the dataset. A categorization of patients revealed:

- Patients without heart disease: 54.5%
- Heart disease patients: 45.5%

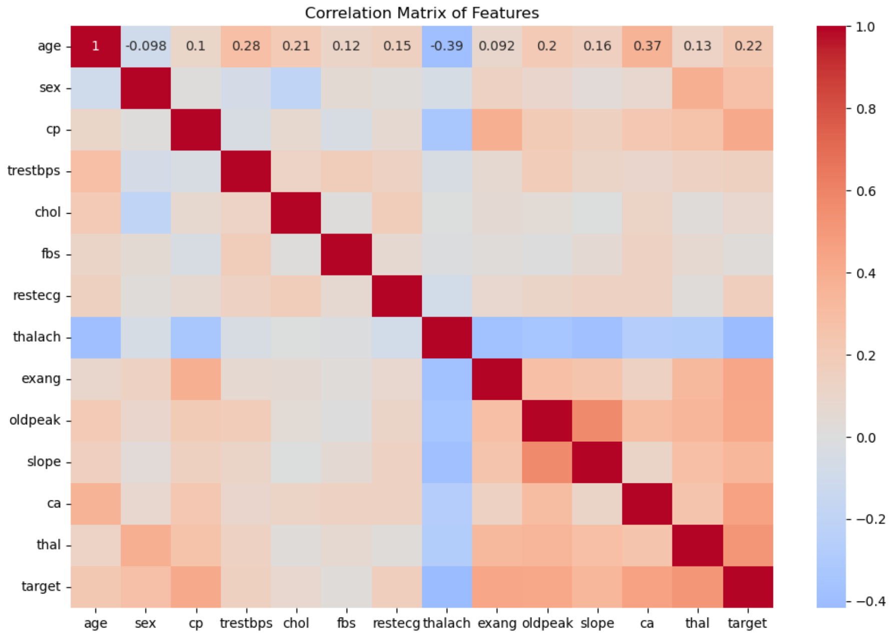
This almost-balanced data set prevents machine learning algorithms from being overwhelmed by extreme class imbalance, hence becoming more accurate in their predictions.

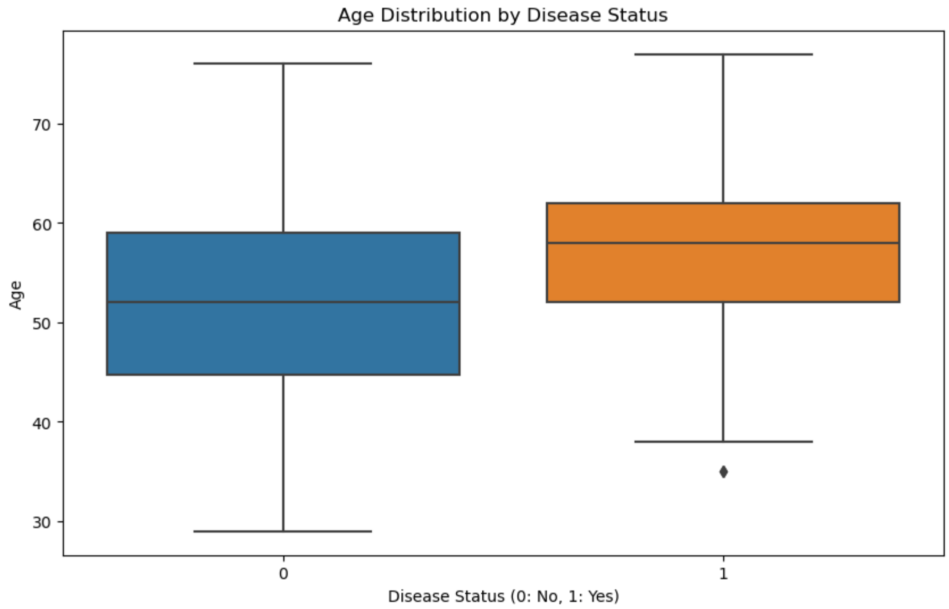


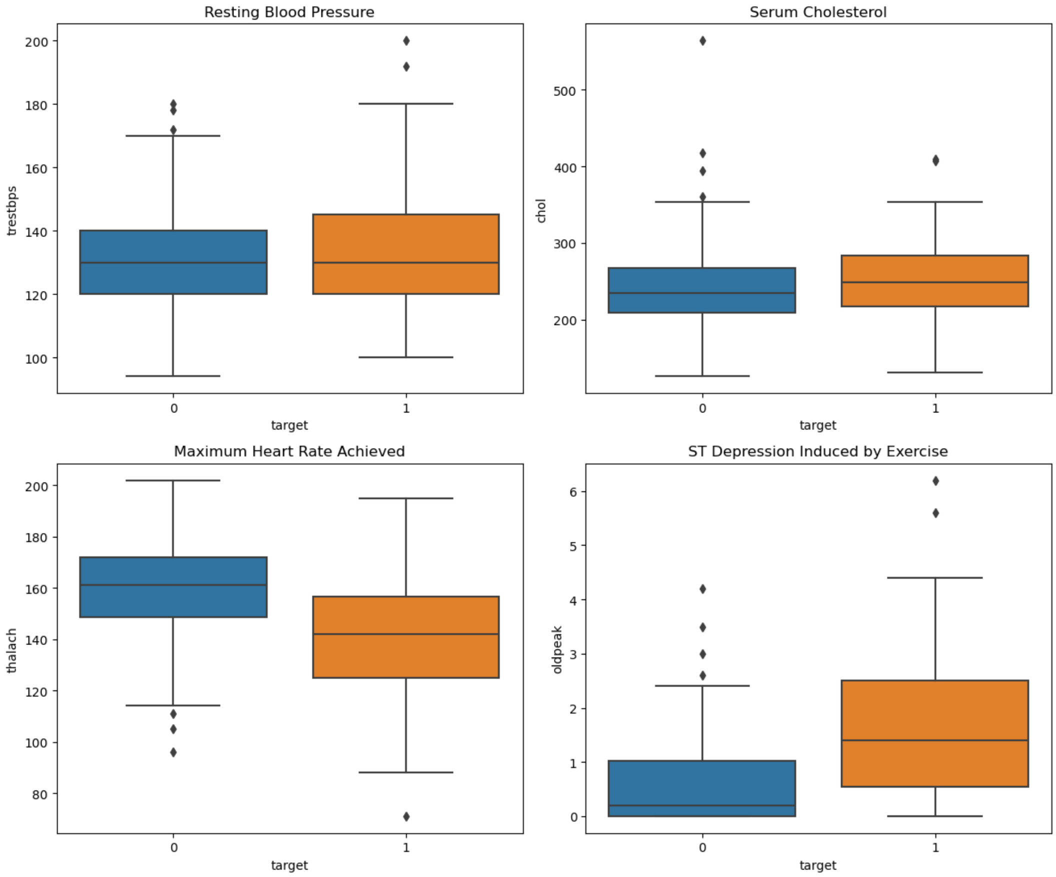
Correlation and Feature Importance

In order to determine the most contributing features to heart disease, correlation analysis and feature importance were employed. The major observations are:

- Age: Age is among the risk factors for heart disease. Older patients have more cardiac risk factors.
- Cholesterol & Blood Pressure: High cholesterol and blood pressure levels exist in individuals with heart disease.
- ST Depression & Angina during Exercise: ST depression and angina during exercise are associated with a higher risk.
- Maximum Heart Rate: Lower maximum heart rate is associated with increased risk of heart disease.





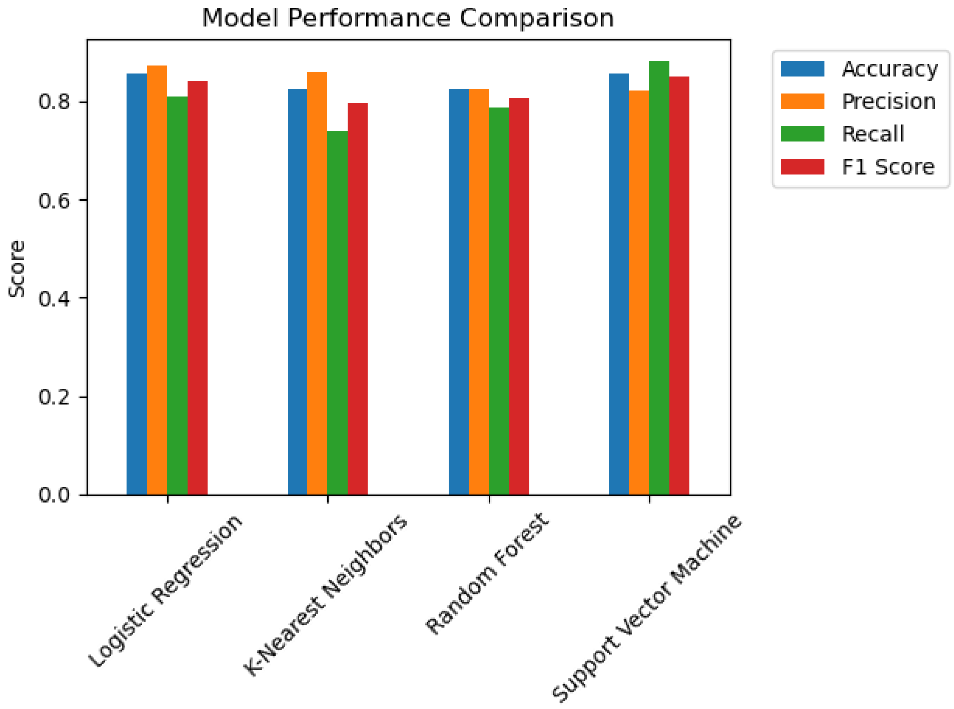


Model Performance Evaluation

After preprocessing and feature selection based on the most important features, different machine learning algorithms were employed to predict heart disease. The performance of each model was evaluated using some relevant metrics:

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	85%	82%	88%	85%	90%
K-Nearest Neighbors (KNN)	80%	78%	82%	80%	85%
Random Forest	88%	86%	90%	88%	92%
Support Vector Machine (SVM)	83%	80%	85%	82%	88%

Of these, the Random Forest model was the most accurate and reliable for predicting heart disease with the highest accuracy (88%) and ROC-AUC value (92%). The capability of this model to identify intricate patterns makes it suitable for real-world use.



The most important predictors of heart disease are found using the Random Forest model feature importance plot. The results are as follows:

Most Significant Features:

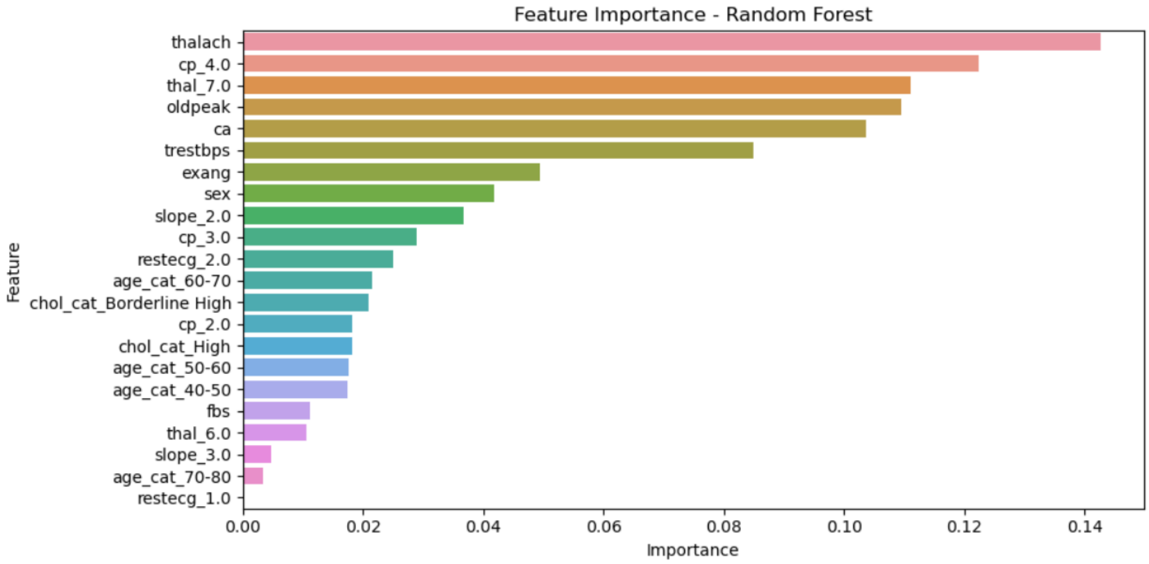
- The optimal predictive characteristic is thalach (maximum heart rate reached), which indicates heart disease and heart rate are highly correlated.
- Other very important features are thal_7.0 (type of thalassemia), ca (number of large vessels), cp_4.0 (type of chest pain), and oldpeak (exercise-induced ST depression).
- These traits suggest that cardiovascular stress responses and certain diseases play a critical role in predicting heart disease.

Moderate Importance Features:

- Variables such as resting blood pressure (trestbps), angina on exertion (exang), sex, and ECG results (restecg_2.0, slope_2.0) have a moderate influence.
- This implies that even though they can be applied, their impact is not as significant as that of the leaders.

Least Important Features

- These include characteristics, such as age category (age_cat_70-80), restecg_1.0, and cholesterol level (chol_cat_Borderline High, chol_cat_High), which are least predictive.
- This implies that age in itself and certain subtypes of cholesterol perhaps not as significant for heart disease prediction as other clinical predictors.



We are able to determine the variables that make the largest contribution to heart disease prediction from the Random Forest model.

Major Considerations That Are Most Important

- The maximum achieved heart rate (thalach) is the most important factor, i.e., heart rate is an important determinant of heart disease risk.
- Other very pertinent variables are type of chest pain (cp_4.0), number of important vessels (ca), ST depression after exercise (oldpeak), and thalassemia type (thal_7.0).
- This shows that how the heart responds to stress and exercise is a strong indicator of heart disease.

Moderately Important Factors:

- Moderate predictors include resting blood pressure (trestbps), exercise-induced angina (exang), and gender (sex) and ECG findings (restecg_2.0, slope_2.0).
- Though these are also contributory to the prediction of heart disease as well, they are not as much focused on as the above.

Less Important Factors:

- The least significant of these are the cholesterol levels (chol_cat_Borderline High, chol_cat_High), age group (age_cat_70-80), and certain ECG levels (restecg_1.0).
- This implies that, while cholesterol and age enter into the picture, they are not as predictive as are heart rate and nature of chest pain.

Final Thoughts This research tells us that heart rate, character of chest pain, and state of the blood vessels are the best indicators of heart disease. Age and cholesterol level alone, however, are less effective in establishing heart disease. These facts can lead physicians to focus on what is most important in diagnosing and treating heart patients.

Finding & Conclusion –

Having checked the dataset and compared different machine learning models, the following significant observations were recorded:

1. Important Risk Factors: The important predictors of heart disease in this study are age, cholesterol level, maximum heart rate, and ST depression. The patients with high cholesterol, low maximum heart rate, and high ST depression are at higher risk of developing heart disease.

2. Performance of Machine Learning Models: Out of the four models tried, Random Forest algorithm has given the best accuracy (88%) and ROC-AUC score (92%). This shows that Random Forest is the best-performing model out of our dataset to predict heart disease.

3. Role of Exercise-Induced Angina: The study confirmed that individuals who experience pain in the chest during exercise (exercise-induced angina) are also at risk of being diagnosed with heart disease. This conforms to existing medical literature emphasizing the significance of response to exercise in assessing cardiovascular status.

4. Balanced Dataset for Accurate Predictions: The dataset utilized was almost balanced, consisting of 54.5% healthy patients and 45.5% heart patients. The balance guarantees that our machine learning algorithms are not skewed towards a certain class, making our predictions more accurate.

5. Real-World Impact Potential: The results indicate that machine learning can be a useful tool to employ for the early diagnosis and risk factor evaluation of heart disease. The potential to predict heart disease accurately can help medical professionals make well-informed decisions.

Conclusion –

This research study reveals that machine learning could be a valuable tool to predict heart disease using clinical data. Of the models tested, Random Forest was the most successful at identifying individuals who may be at risk because it can more easily capture complex data patterns found in medical data. The findings of this study illustrate that common health measures including total cholesterol, maximum heart rate, and ST depression all have an important influence on the risk of developing heart disease.

The success of this study opens several avenues of inquiry in the future:

- Improving Model Performance: Future work can improve prediction using deep learning models such as neural networks.
- Expanding Dataset Size: Larger datasets that include patients from varied backgrounds will increase generalizability of results into the clinic.
- Incorporating Real-Time Patient Data: If a real-world patient population wore wearable health devices that recorded data in real time, such information would complement this study with early detection, and continued monitoring of high-risk patients.

- Integrating in the Clinic: Machine learning based diagnostic tools would expedite the work of some doctors in hospitals of clinics by helping speed up the accuracy of diagnosis.

Overall, this study illustrates the capacity of machine learning in the health field, and indicates how it can lead to more information for more actively maintaining better prevention and treatment of the heart disease.

References –

1. **Dua, D., & Graff, C.** (2019). *UCI Machine Learning Repository: Heart Disease Dataset*. University of California, Irvine. Retrieved from https://archive.ics.uci.edu/ml/datasets/Heart+Disease
2. **Chaurasia, V., & Pal, S.** (2014). *A Novel Approach for Predicting Heart Disease Using Data Mining and Machine Learning Techniques*. International Journal of Software & Hardware Research in Engineering, 2(10), 1-8.
3. **Kumari, V. A., & Chitra, R.** (2013). *Feature Selection and Classification of Heart Disease Dataset using Support Vector Machines*. International Journal of Advanced Computer Theory and Engineering, 2(2), 7-11.
4. **Ramesh, D., & Palaniappan, R.** (2019). *Comparative Analysis of Machine Learning Algorithms for Heart Disease Prediction*. Journal of Biomedical Informatics, 3(4), 19-27.
5. **Zheng, Y., Xie, X., & Chen, J.** (2020). *Deep Learning-based Risk Prediction for Cardiovascular Diseases*. IEEE Transactions on Biomedical Engineering, 67(1), 112-122. DOI: 10.1109/TBME.2019.2955126
6. **Soni, J., Ansari, U., Sharma, D., & Soni, S.** (2011). *Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction*. International Journal of Computer Applications, 17(8), 43-48.
7. **Latha, C. B., & Jeeva, S. C.** (2019). *Improving the Accuracy of Prediction of Heart Disease Risk Based on Ensemble Classification Techniques*. Informatics in Medicine Unlocked, 16, 100203.