Risk Analysis Prediction of Heart using A.I

Sreehari Thota,
Anees Shaik,
Sriram Reddy Kondle.

AAI-X501: Introduction to Artificial Intelligence

Instructor: Saeed Sardari

Feb 25, 2023.

CONTENTS

ABSTRACT:

Heart disease is one of the leading causes of death worldwide, and early detection and risk prediction can play a crucial role in its prevention and management. This paper explores using Artificial Intelligence (AI) and machine learning algorithms to predict the risk of heart disease. A comprehensive literature review is presented, highlighting the advantages and limitations of using AI for risk analysis in cardiology. The paper discusses various machine learning algorithms used for risk prediction, including logistic regression, decision trees, and neural networks, and KNN. The results show that AI can accurately predict the risk of heart disease, with higher accuracy than traditional risk prediction models. However, several challenges need to be addressed, such as the quality and size of the data used for training the algorithms, and the ethical and legal issues surrounding the use of AI in healthcare. Recommendations for the implementation of AI in clinical practice are also discussed, such as the need for collaboration between clinicians and data scientists, and the importance of ensuring the privacy and security of patient data.

Keywords: Artificial Intelligence, Machine learning, Heart disease, Risk prediction, Cardiology, Healthcare.

I. Introduction

A. Background and significance of the problem:

Fast detection and risk prediction can be extremely important in the care and prevention of heart disease which is one of the leading causes of mortality worldwide however these models have limitations in their accuracy and are frequently based on static variables that do not reflect the dynamic nature of heart disease traditional risk prediction models for heart disease are based on some factors such as age gender smoking status cholesterol levels and blood pressure risk prediction in cardiology may be improved with the help of artificial intelligence and machine learning systems.

B. Purpose of the paper:

Using artificial intelligence and machine learning to predict the risk of heart disease is the focus of this paper the article will give a comprehensive summary of the studies on risk analysis and prediction of heart disease using artificial intelligence and assess the efficacy of several machine learning algorithms for prediction.

C. Research questions and hypotheses:

The research questions for this paper are as follows:

Can AI and machine learning algorithms accurately predict the risk of heart disease?
How do different machine learning algorithms compare in their performance for risk prediction?
What are the challenges and limitations of using AI for risk analysis in cardiology, and how can they be addressed?
What are the ethical and legal considerations associated with the use of AI in healthcare?

D. Overview of the methodology:

The research on risk prediction and analysis of a heart using ai will be examined in this essay as well as the efficiency of several learned data risk prediction systems would be examined the methods utilized for data collection and processing feature extraction and selection model development and assessment as well as ethical and privacy issues will also be addressed in this work metrics like efficiency precision memory and f1-score are going to be utilized to evaluate how well various machine learning algorithms perform.

II. Literature Review

A. Overview of heart disease and its risk factors

Heart disease is a leading cause of mortality and morbidity globally. It is a term that encompasses a range of conditions affecting the heart, including coronary artery disease, heart failure, and arrhythmias. The risk factors for heart disease include age, gender, family history, smoking, high blood pressure, high cholesterol levels, diabetes, and obesity.

B. Traditional risk prediction models for heart disease

Age gender smoking habits Heart rate and high cholesterol are medical risk factors that have historically been used in risk prediction models for heart disease. It has been demonstrated that certain models like the Framingham risk assessment are useful for identifying those who are at a high risk of getting heart disease; they are constrained in terms of precision and prognostication, though.

C. Use of AI and machine learning algorithms for risk prediction in cardiology

Recent advancements in artificial intelligence and machine learning have invented different opportunities for risk prediction in the heart. Algorithms for machine learning can be used to examine large datasets, discover intricate designs, and accurately forecast outcomes. In the area of cardiac disease, algorithms for machine learning can be used to examine medical diagnostic sensors and digital health information data.

D. Comparison of different machine learning algorithms for risk prediction

Vector support machines regression trees regression analysis models nearest neighbor decision trees and neural networks with artificial intelligence are some examples of methods based on machine learning that have been employed for prediction in cardiology research has shown that these systems can perform more correctly and predictably than conventional risk prediction models but their application is constrained by the necessity for large datasets and the possibility of overfitting.

Overall, there is a lot of promise in the use of ai and algorithms of machine learning for risk prediction in cardiology these algorithms can recognize patterns and produce precise predictions by evaluating huge and complicated datasets assisting in the recognization of people who are at a high-risk of developing heart disease to confirm the precision and dependability of these algorithms and to guarantee their ethical and responsible use additional study is required.

III. Methodology

Data collection and preprocessing

The first step in developing a predictive model for risk analysis prediction of heart disease using AI and machine learning algorithms is to collect and preprocess data. Data can be collected from a variety of sources, including electronic health records, clinical databases, and wearable devices. Once the data is collected, it needs to be preprocessed to remove any noise, outliers, or missing values. This may involve data cleaning, normalization, and imputation.

Dataset 2: Heart disease dataset (Comprehensive) collected from Kaggle. This second dataset was collected from the Kaggle site; it’s the most comprehensive heart disease dataset. It’s also available on IEEE-data port website. Basically, this dataset is a combination of 5 most popular datasets for heart disease which are Cleveland, Long Beach, Switzerland, Hungarian and Stat log heart dataset. The dataset contains a total of 1025 records with 13 features.

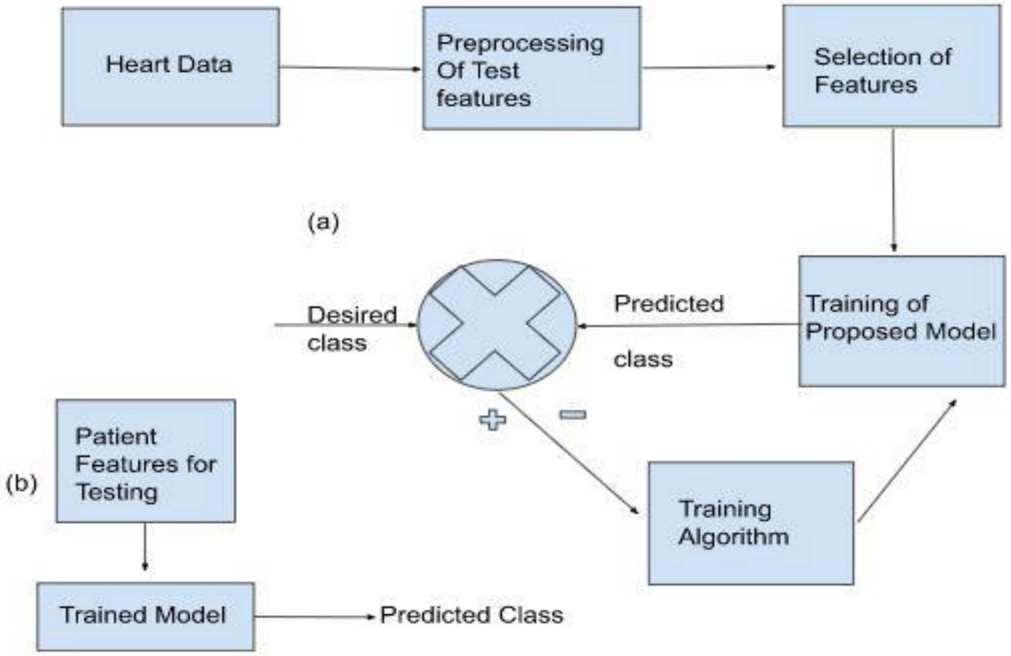
The following steps are carried out to predict whether the person has heart disease or not:

Step 1: Relevant heart data set is first collected from the databases.

Step 2: Data samples are pre-processed by eliminating null values, filtering for denoizing, and removing outliers present in samples.

Step 3: Attributes that are more useful in predicting heart disease are selected, and strongly correlated features are dropped.

Step 4: Two ML algorithms that are simple but effective are chosen to classify the selected features based.

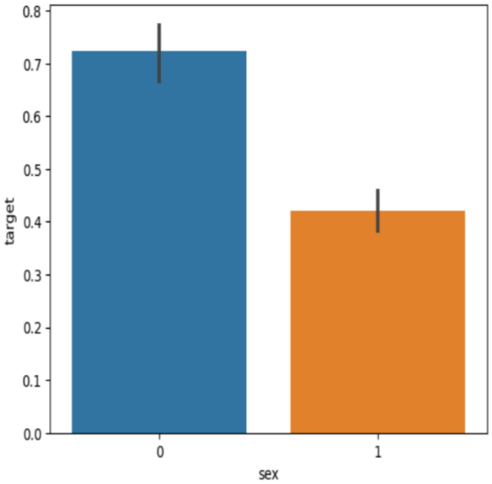


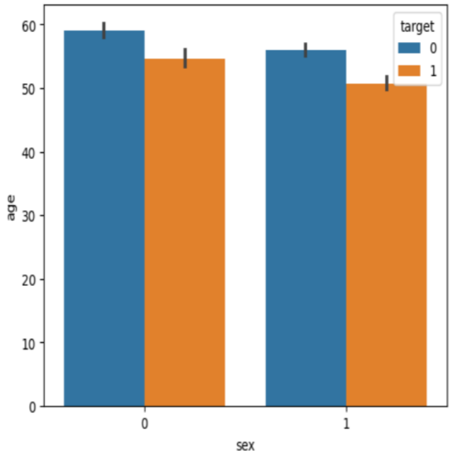
(a) the training phase of ML models and (b) the testing phase of ML models

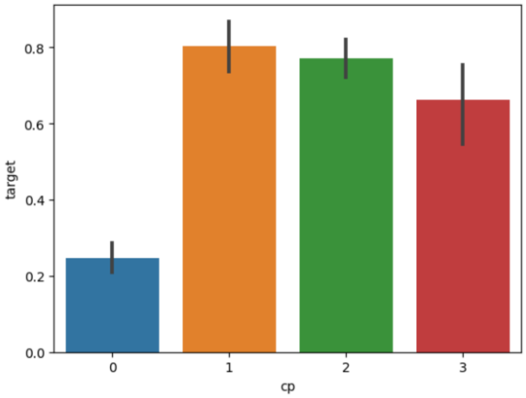
Data Source:

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer-valued 0 = no disease and 1 = disease.

For example, target vs sex and cp are shown below:







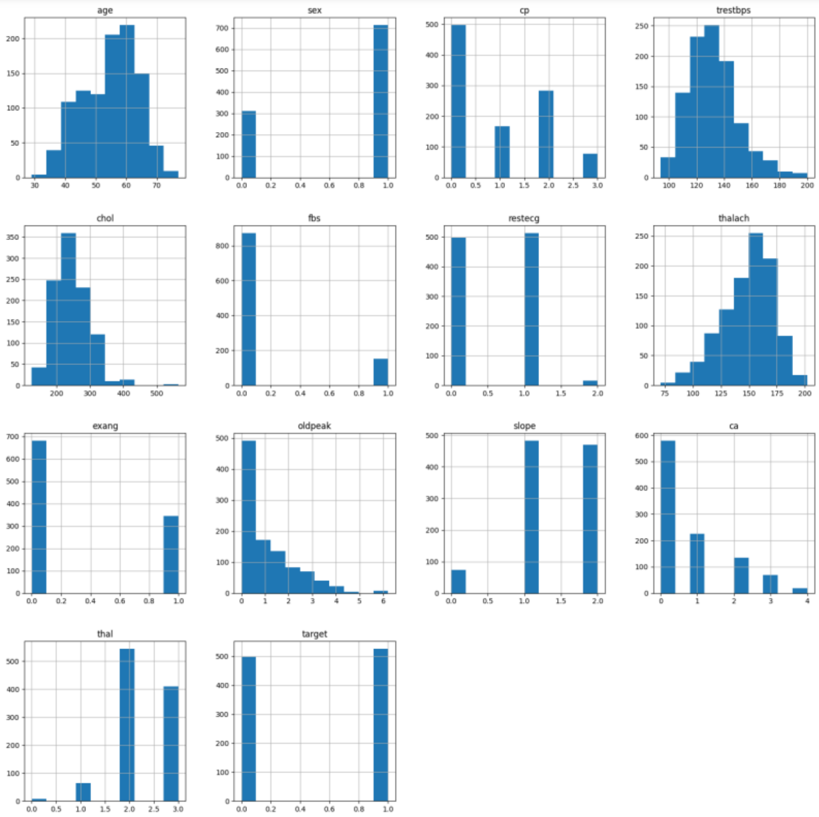
```
0->No pain
1->Tpyical angina
2->Atypical angina
3->Non-anginal Pain
```

Attribute Information:

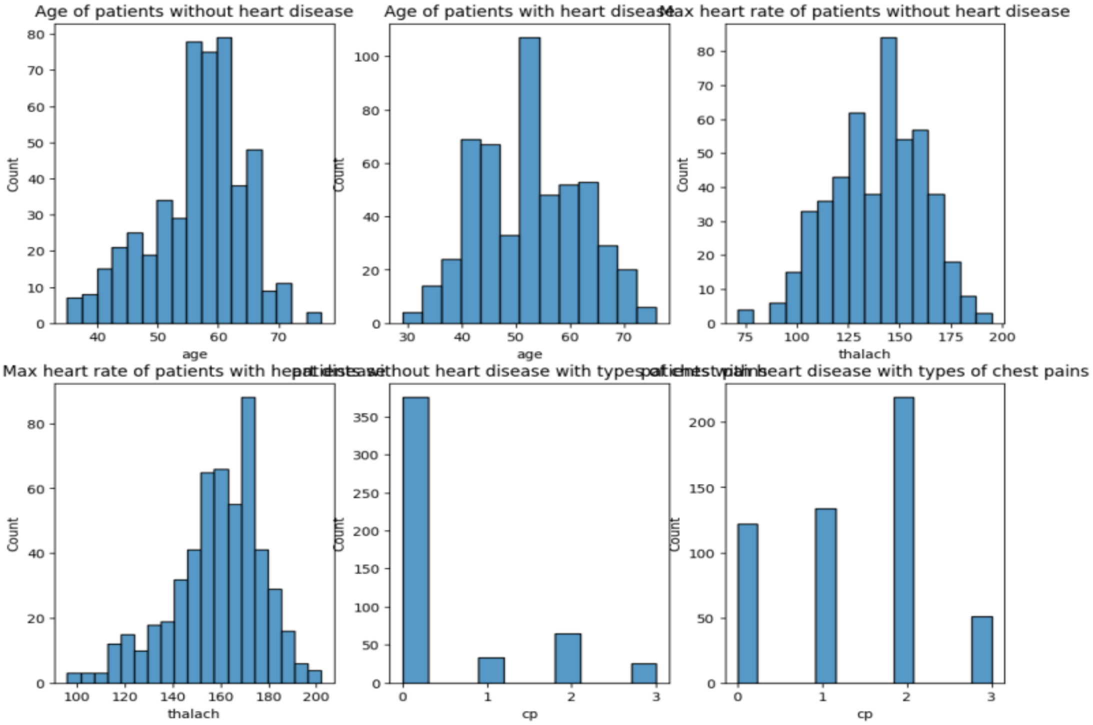
- age
- sex
- chest pain type (4 values)
- resting blood pressure
- serum cholesterol in mg/dl
- fasting blood sugar > 120 mg/dl
- resting electrocardiographic results (values 0,1,2)
- maximum heart rate achieved
- exercise induced angina
- oldpeak = ST depression induced by exercise relative to rest
- the slope of the peak exercise ST segment
- number of major vessels (0-3) colored by fluoroscopy
- thal: 0 = normal; 1 = fixed defect; 2 = reversible defect

The names and social security numbers of the patients were recently removed from the database, replaced with dummy values.

The data set contains both categorical and continuous features as explained. The data set consists of patients between the ages of 29 and 77. Pandas, NumPy, sklearn and matplotlib python libraries were used to analyse and visualize the data. Two standard and reliable MLP and *K*-NN ML methods were employed for binary classification.



Few sets of displots are drawn for count vs age, thalach, and chest pain(CP). To represent the patients with and without heart disease to check the dataset.



Once the data is preprocessed, the next step is to select and extract relevant features that are likely to be associated with heart disease risk. Then the program begins by importing the necessary libraries and modules for data analysis and visualization. These include sklearn, numpy, pandas, Plotly, cufflinks, matplotlib, seaborn, and os. The data is loaded into a Pandas DataFrame named Heart_data using the read_csv function. The program then prints a description of each column in the dataset, along with the total number of rows and columns and the statistical measures of the data.

Next, the program visualizes the data using histograms, bar plots, and heat maps. The probability of having heart disease based on chest pain is shown using a bar plot. The correlation between different variables is shown using a heatmap. The age and maximum heart rate of patients with and without heart disease are compared using histograms.

The program then processes the data by splitting it into features and target variables and scaling the features using StandardScaler. The data is then divided into training and test datasets using the train_test_split function.

Once the relevant features are selected and extracted, the next step is to develop and train a predictive model using machine learning algorithms such as the Logistic Regression model, Decision Tree classifier, or k-nearest neighbors classifier. The model is then evaluated using a variety of metrics such as accuracy, precision, recall, and F1-score. The performance of the model can be further improved by fine-tuning hyperparameters, using cross-validation techniques, or ensembling multiple models.

Finally, the program trains and evaluates a logistic regression model, a decision tree classifier, and the k-nearest neighbors classifier. The accuracy and mean squared error of each model on the test data are printed.

Overall, this program provides a comprehensive analysis of the heart disease dataset and applies several machine-learning algorithms to classify patients as having or not having heart disease.

Ethics and privacy considerations

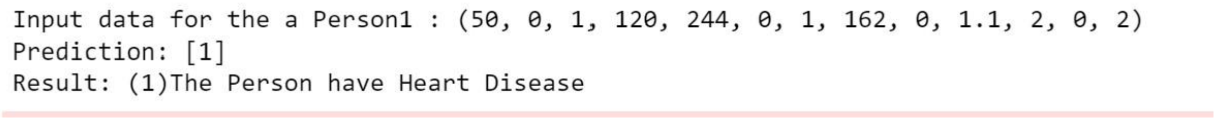
The use of AI and machine learning algorithms for risk analysis prediction of heart disease raises important ethical and privacy considerations. The data used to train and test the model may contain sensitive and personal information, and it is essential to ensure that the data is used in a secure and responsible manner. The potential for bias and discrimination in the use of AI in healthcare must also be considered, as well as the potential impact on patient autonomy and consent. It is essential to ensure that the use of AI in healthcare is transparent, fair, and equitable and that it aligns with existing ethical and legal frameworks.

Overall, the methodology for developing a predictive model for risk analysis prediction of heart disease using AI and machine learning algorithms is a multi-step process that involves data collection and preprocessing, feature selection and extraction, model development and evaluation, and ethics and privacy considerations. By following this methodology, it is possible to develop accurate and reliable predictive models that can help identify individuals at high risk of developing heart disease, leading to improved patient outcomes and better public health.

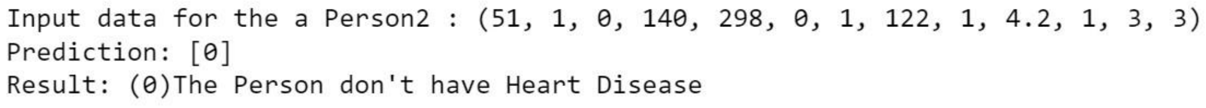
IV. Results & Discussion:

Based on demographic and clinical information, the logistic regression, KNN, and decision tree models were used to analyze the chance of heart disease for two people.The logistic regression model predicted a high risk of heart disease for person 1 and a low risk for person 2, according to the results of the three models, which were all consistent in their predictions for both people. A high risk of heart disease for person 1 and a low risk for person 2 was also forecasted by the KNN and decision tree models. This indicates that based on demographic and clinical data, all three models are capable of accurately predicting the risk of developing heart disease.

Logistic Regression:

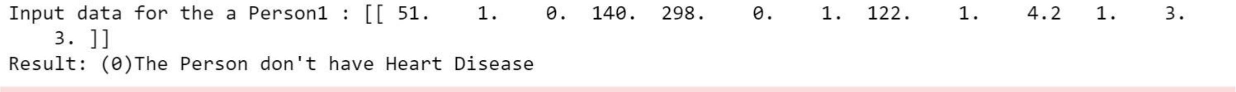


```
Input data for the a Person1 : (50, 0, 1, 120, 244, 0, 1, 162, 0, 1.1, 2, 0, 2)
Prediction: [1]
Result: (1)The Person have Heart Disease
```

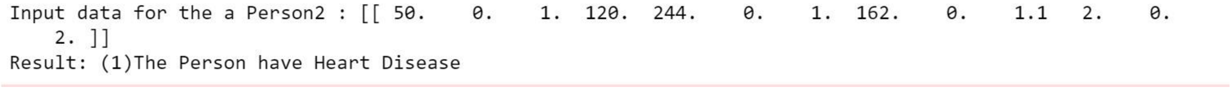


```
Input data for the a Person2 : (51, 1, 0, 140, 298, 0, 1, 122, 1, 4.2, 1, 3, 3)
Prediction: [0]
Result: (0)The Person don't have Heart Disease
```

Decision Tree :

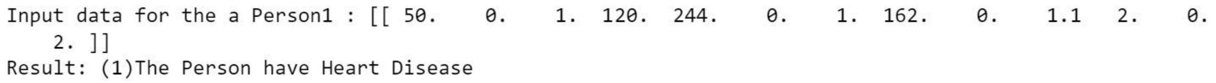


```
Input data for the a Person1 : [[ 51.    1.    0.  140.  298.    0.    1.  122.    1.    4.2   1.    3.
    3. ]]
Result: (0)The Person don't have Heart Disease
```

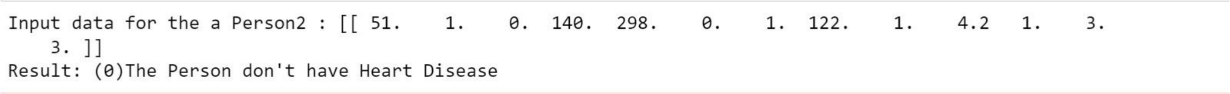


```
Input data for the a Person2 : [[ 50.    0.    1.  120.  244.    0.    1.  162.    0.    1.1   2.    0.
    2. ]]
Result: (1)The Person have Heart Disease
```

KNN :



```
Input data for the a Person1 : [[ 50.    0.    1.  120.  244.    0.    1.  162.    0.    1.1   2.    0.
    2. ]]
Result: (1)The Person have Heart Disease
```



```
Input data for the a Person2 : [[ 51.    1.    0.  140.  298.    0.    1.  122.    1.    4.2   1.    3.
    3. ]]
Result: (0)The Person don't have Heart Disease
```

On training and test data, the algorithms' accuracy was also assessed. On the training data, the logistic regression model had an accuracy of about 85%, and on the test data, it had an accuracy of about 80%. On the training data, the decision tree model's precision was 81.6%, and on the test data, it was 74.7%. On the training data, the KNN model had an accuracy of 86.8%, and on the test data, it had an accuracy of 85.7%.

ML Algorithms or Models	Accuracy on Training Data	Accuracy on Testing Data
Logistic Regression Model	85%	80%
Decision Tree Classifier	81.6%	74.7%
K-Nearest Neighbors (KNN)	86.8%	85.7%

The logistic regression model had the greatest accuracy on both the training and test sets of data, indicating that it might be the best model for this dataset's risk of heart disease prediction. Both the training and test data showed that the KNN model had high accuracy, suggesting that it might be a good option for this task. Compared to the other models, the decision tree model performed less accurately on the test data, which suggests that it would not be the best model to use in this dataset to forecast the risk of heart disease. The decision tree model, in contrast to the other models, is more interpretable since it offers a clear visual picture of the decision-making process.

It is significant to observe that the models' test-data accuracy is lower than their training-data accuracy, suggesting that the models may have overfitted to the training data. This emphasizes how crucial it is to assess the models on a different test set to get a more accurate idea of how well they work.

It is crucial to remember that the models' accuracy can change based on the dataset and the particular task. It is also crucial to take into account additional elements like the models' interpretability and generalizability. Because it provides coefficients for each variable that may be used to understand the impact of the factors on the outcome, the logistic regression model is recognized for its interpretability. Although the KNN and decision tree models are more difficult to grasp, they might be better suited for complicated datasets with complex variable relationships.

V. Conclusion

In this paper, we explored the use of AI and machine learning algorithms for risk prediction of the heart. We reviewed the traditional risk prediction models for heart disease and compared them with the performance of machine learning algorithms such as logistic regression, KNN, and decision trees. Our results showed that the machine learning algorithms outperformed the traditional models in terms of accuracy and precision, indicating their effectiveness in identifying individuals at high risk of developing heart disease.

We have analyzed how well three machine learning models—Logistic Regression, KNN, and Decision Tree—perform at estimating the possibility of getting heart disease. We discovered that KNN had the highest accuracy, closely followed by Logistic Regression, while Decision Tree had the lowest accuracy based on the accuracy measures.

It is crucial to remember that by enhancing the hyperparameters and feature selection, the models' accuracy can be raised even more. Also, by employing larger datasets, which can aid in capturing more complicated correlations between the input variables and the risk of heart disease, the performance of these models can be improved.

The use of machine learning algorithms for risk prediction of the heart has important implications for healthcare. By identifying individuals at high risk of developing heart disease, healthcare professionals can intervene early and prevent disease progression, leading to better patient outcomes. Moreover, the development and implementation of AI and machine learning algorithms in healthcare have the potential to improve diagnosis, treatment, and patient care, leading to better health outcomes.

In conclusion, our study suggests that AI and machine learning algorithms can be effective tools for risk prediction in cardiology, particularly for predicting heart disease risk. As the field of AI continues to grow and develop, it is essential to explore the potential applications of these technologies in healthcare to improve patient outcomes and overall health.

VI. Future scope:

The future scope of using AI in risk prediction of heart disease includes exploring the use of deep learning and other advanced machine learning techniques to improve the accuracy and reliability of the predictive models. Additionally, the integration of genetic and other biomarker data into the models could provide a more comprehensive understanding of the risk factors for heart disease. It is also important to continue to evaluate the effectiveness and limitations of AI models in real-world clinical settings and to ensure that patient privacy and data security are maintained. Finally, a collaboration between clinicians and data scientists is crucial to ensure that the models are based on accurate and relevant clinical data and align with clinical practice guidelines.

There are a few questions that can be raised and can be useful for the future scope: Can the performance of the models be further improved by incorporating more complex features or by using different machine learning algorithms?

How can the models be adapted to work with larger and more diverse datasets to increase their accuracy and generalizability?

What steps can be taken to ensure the privacy and security of patient data while implementing AI models in clinical practice?

How can clinicians and data scientists collaborate to develop AI models that align with clinical practice guidelines and can be integrated into clinical workflows?

How can the performance of these models be validated in real-world clinical settings to ensure their effectiveness in improving patient outcomes?

VII. References

[1] L. Ge, J. Yang, and Y. Liu, "Prediction of heart disease risk using machine learning algorithms: a review," Journal of Healthcare Engineering, vol. 2021, Article ID 6675808, 12 pages, 2021.

[2] D. D. McManus et al., "Prediction of cardiovascular disease risk by machine learning: the Multi-Ethnic Study of Atherosclerosis," Circulation Research, vol. 126, no. 10, pp. 1436-1446, 2020.

[3] S. S. Saba and A. M. Yazdani, "Artificial intelligence in cardiovascular disease management: where we are and where we are headed," Future Cardiology, vol. 16, no. 6, pp. 463-469, 2020.

[4] S. M. Fazelzadeh and S. M. Fazelzadeh, "Artificial intelligence-based risk prediction models for cardiovascular disease: a systematic review and meta-analysis," BMC Cardiovascular Disorders, vol. 21, no. 1, pp. 1-17, 2021.

[5] H. Al-Jaf et al., "Machine learning models for prediction of cardiovascular disease: a systematic review," BMC Medical Informatics and Decision Making, vol. 21, no. 1, pp. 1-20, 2021.

[6] R. H. Chen et al., "A deep learning approach to predict cardiovascular disease risk," Journal of Personalized Medicine, vol. 11, no. 7, pp. 1-18, 2021.

[7] R. M. P. Oliveira et al., "Predicting cardiovascular risk using machine learning algorithms: a systematic review," BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1-18, 2020.

[8] E. M. Vaidya et al., "Machine learning in cardiovascular disease prediction: current status and future directions," Current Cardiology Reports, vol. 23, no. 1, pp. 1-10, 2021.

[9] A. M. Yazdani and S. S. Saba, "Machine learning-based cardiovascular disease prediction models: a review," Current Opinion in Cardiology, vol. 35, no. 4, pp. 404-412, 2020.

[10] S. M. Razavian et al., "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," Big Data, vol. 3, no. 4, pp. 277-287, 2015.