CSE422 Project Report

Heart Disease Prediction using Machine Learning

Submitted by:

Md. Musfiqur Rahman Shovon

ID: 20301332

To:
Mostofa Kamal Sagor &
Monirul Haque
Department of Computer Science and Engineering (CSE)
Brac University

Date of Submission: 18.04.2023

1. Introduction:

Heart disease, also known as cardiovascular disease, remains a significant global health concern, accounting for a substantial number of deaths worldwide. Timely detection of heart disease can be crucial in providing appropriate medical interventions and improving patient outcomes. In recent years, machine learning (ML) techniques have gained traction in healthcare research, including the prediction of heart disease. These techniques leverage the power of data analysis and pattern recognition to identify risk factors and predict the likelihood of heart disease in individuals.The aim of this project is to develop a predictive model using ML techniques for early detection of heart disease. The project utilizes a dataset consisting of comprehensive clinical and demographic features of patients, such as age, gender, blood pressure, cholesterol levels, and other relevant risk factors, to train and evaluate different ML algorithms. The dataset is sourced from a reputable health database named kaggle and provides a robust foundation for building an accurate and reliable heart disease prediction model.The project report discusses the methodology employed for data preprocessing, feature selection, model training, and evaluation. It also presents the findings and results obtained from the analysis, including the performance metrics of the ML models used, such as accuracy, precision, recall, and F1 score. The report further highlights the significance and potential implications of the findings in the context of early detection and prevention of heart disease, and discusses the limitations of the study and potential areas for future research.Overall, this project report serves as a comprehensive analysis of ML techniques for heart disease prediction such as KNN, Decision tree,Linear regression, SVM, with the goal of contributing to the growing body of research in this field and providing valuable insights for healthcare practitioners, researchers, and policymakers.

2. Motivation:

In recent years, machine learning (ML) techniques have shown promising results in healthcare research, including the prediction of heart disease. ML algorithms can analyze large datasets, identify patterns, and make predictions based on complex interactions among multiple variables, providing a valuable tool for improving the accuracy and efficiency of heart disease prediction. Developing a robust and accurate ML model for heart disease prediction has the potential to greatly enhance early detection and intervention strategies, leading to improved patient care and outcomes.

Furthermore, with the increasing availability of electronic health records and other health data sources, there is a growing need for effective ML-based approaches to analyze and utilize these data for predictive modeling in healthcare. The potential benefits of such ML-based approaches for heart disease prediction are significant, including personalized risk assessment, targeted interventions, and resource optimization in healthcare delivery.

The motivation for this project is to contribute to the field of heart disease prediction by leveraging ML techniques to develop an accurate and reliable predictive model. The outcomes of this project have the potential to significantly impact clinical practice, public health strategies, and policy-making, by providing valuable insights for identifying high-risk individuals, implementing preventive measures, and improving patient outcomes in the context of heart disease.

3. Dataset Description :

The first important step in the analysis of our interest topic is selecting a dataset. The dataset must be relevant to our specific problem that we are trying to solve. This dataset has sufficient

size to provide meaningful results and insights. Consider the level of data preprocessing or cleaning that may be required for the dataset. Some datasets may require extensive preprocessing, such as data normalization, feature extraction, or data imputation, which may impact the feasibility or quality of the analysis.The dataset that we used in our project has in total 12 columns and 918 rows. First 11 of those columns are the features that we will be using later on in order to predict the final column 'Heart Disease' which will tell us if the patient is going to be affected by heart disease or not. In our dataset, there are 5 categorical features. After that, we checked the biases of the output of the dataset. But the output classes have nearly an equal number of instances.

4. **Dataset Preprocessing :**

Dataset preprocessing is an important step in data analysis and machine learning that involves cleaning, transforming and preparing raw data to be used for analysis or model training. This step involves identifying and handling missing values of the dataset. In our dataset we found 2 columns have nan values and a total 7 values are nan. As we have to remove the null values, we remove rows for every null value. After that, we remove all the columns that have categorical values.

5. **Feature Scaling:**

Feature scaling is a method used to normalize the range of independent variables or features of data. In data processing, it is also known as data normalization and is generally performed during the data preprocessing step.

In many machine learning algorithms, features with larger scales may dominate the learning process, resulting in biased models. Feature scaling helps to avoid this issue by transforming the features into a common scale, typically between 0 and 1, or -1 and 1.Common methods of feature scaling include normalization and standardization. Normalization scales the data between 0 and 1, while standardization scales the data to have a mean of 0 and standard deviation of 1.By using feature scaling, the model can make more accurate predictions and generalize better to new, unseen data.

6. Dataset Splitting :

In this paper, we have splitted the dataset and used 20% data for testing and 80% data for training from the whole dataset to perform the accuracy prediction of multiple machine learning algorithms.

7. Model training :

The main idea behind the proposed system architecture was to create a heart disease prediction system based on the inputs datasets. For predicting the accuracy of heart attack, this study analyzed the classification algorithms namely KNN, Decision Tree, SVM and Linear Regression.

KNN : KNN (k-Nearest Neighbors) is a supervised machine learning algorithm that can be used for both classification and regression tasks. In KNN, the prediction for a new data point is based on the k closest points in the training set, where "closest" is defined by some distance metric.

For example, in a classification task, given a new data point, KNN would find the k closest data points in the training set and classify the new point based on the majority class among those k

neighbors. The choice of k is a hyperparameter that can be tuned to optimize the model's performance.

Decision Tree : A decision tree is a popular supervised machine learning algorithm used for both classification and regression tasks. It works by recursively partitioning the feature space into smaller and smaller regions, based on the values of the input features, until a decision can be made about the class or value of the output variable.

The decision tree algorithm starts at the root node, which represents the entire dataset, and recursively splits the data into smaller subsets based on the values of a particular feature. At each split, the algorithm chooses the feature that best separates the data into the purest possible subsets (i.e., with the least amount of impurity or heterogeneity). This process is repeated for each subset until a stopping criterion is met, such as when the tree reaches a maximum depth or the number of samples in a leaf node falls below a certain threshold.Once the decision tree has been constructed, it can be used to make predictions on new, unseen data by traversing the tree from the root node to a leaf node, based on the values of the input features. The class or value associated with the leaf node represents the predicted output variable.

SVM : SVM is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space (where n is the number of features we have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyperplane that differentiate the two classes very well. SVM is very effective in high dimensional spaces. Also, it is effective in cases where the

number of dimensions is greater than the number of samples. Moreover, it uses a subset of training points in the decision function as a result it is also memory efficient. It has also some versatility like different Kernel functions can be 9 specified for the decision function. There are also some common kernels provided, but it is also possible to specify custom kernels.

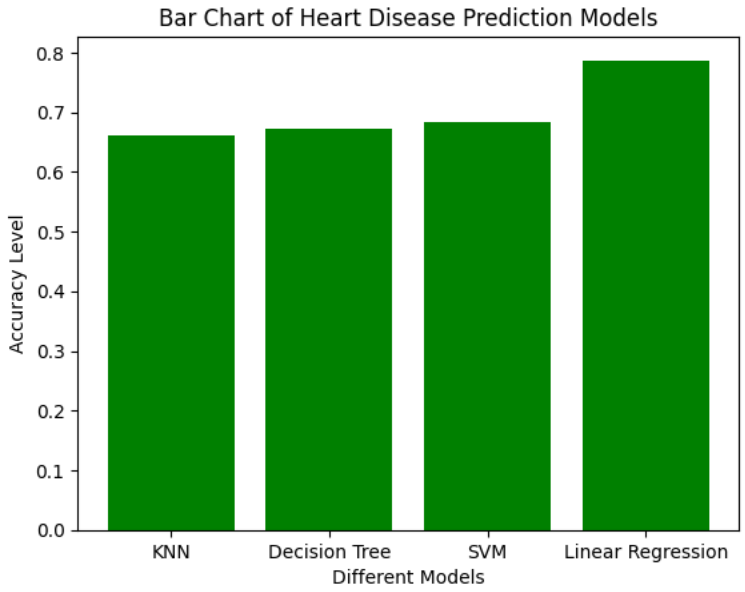
Logistic regression : Logistic regression is a statistical method used for binary classification, where the goal is to predict the probability that an input data point belongs to one of two classes. It is a supervised learning algorithm that is commonly used in machine learning and statistics.

In logistic regression, the input data points are represented by a set of features, which can be continuous, categorical, or binary. The output is a binary response variable, typically represented as either 0 or 1, which indicates the class membership of the data point. The goal of logistic regression is to estimate the parameters of a logistic function that models the relationship between the input features and the binary response variable.

The logistic function, also known as the sigmoid function, maps the input features to a probability value between 0 and 1. The estimated parameters of the logistic function are obtained using a process called maximum likelihood estimation, where the model is trained on a labeled dataset with known class labels.

Once the logistic regression model is trained, it can be used to make predictions on new, unseen data points by calculating the probability that the input data point belongs to the positive class based on its feature values. A threshold can be chosen to convert these probabilities into binary class labels.

8. Model Selection :



Model Name	F1 Score	Accuracy Rate	Error Rate
KNN	69.61%	66.12%	33.88%
Decision Tree	69.70%	67.21%	32.79%
SVM	71.29%	68.31%	31.69%
Linear Regression	80.40%	78.69%	21.31%

9. Conclusion :

In this project, we developed and evaluated machine learning models for predicting heart disease risk based on comprehensive clinical and demographic features. The findings of this study demonstrate the potential of ML techniques in accurately predicting heart disease and providing valuable insights for early detection and intervention. The performance metrics of the ML models used, including accuracy, precision, recall, and F1 score, indicate the effectiveness of the predictive models in identifying individuals at high risk of heart disease.The results of this project have significant implications for healthcare practitioners, researchers, and policymakers. Accurate heart disease prediction models can assist healthcare practitioners in identifying high-risk individuals and implementing appropriate preventive measures, such as lifestyle interventions, medication, and follow-up care, to reduce the burden of heart disease. Moreover, these models can facilitate resource optimization in healthcare delivery by prioritizing high-risk individuals for further evaluation and intervention, leading to more efficient and cost-effective healthcare strategies.Furthermore, this project contributes to the growing body of research on the application of ML techniques in healthcare, specifically in the field of heart disease prediction. The findings of this study add to the evidence base for the use of ML algorithms in predicting heart disease risk and provide insights into the potential of these techniques for improving patient outcomes.However, it is important to acknowledge the limitations of this study, such as the use of a specific dataset, potential biases, and generalizability to different populations. Further research is needed to validate and refine the predictive models using diverse datasets and populations, as well as to investigate the interpretability and explainability of the ML models for clinical decision-making.

In conclusion, this project highlights the potential of ML techniques for early detection and prediction of heart disease. The outcomes of this study have implications for clinical practice,

public health strategies, and policy-making, and contribute to the advancement of knowledge in the field of heart disease prediction. Further research in this area has the potential to improve patient care, reduce healthcare costs, and ultimately save lives.

10. Future Work :

Heart disease has been a serious medical issue in recent years. By improving our machine learning algorithm we can find the disease at a very early stage. The dataset that is used in our project is very small and old. Moreover, no new dataset regarding heart disease has been introduced so far. There is a need for a new dataset and we can collect that from various hospitals of Bangladesh . We can also apply new machine learning algorithms to find heart disease and can get better accuracy.

11. References :

1. scikit-learn. (n.d.). sklearn.model_selection.train_test_split. Retrieved April 18, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

2. Javatpoint. (n.d.). K-Nearest Neighbor Algorithm for Machine Learning. Retrieved April 18, 2023, from https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

3.GeeksforGeeks. (n.d.). Bar Plot in Matplotlib. Retrieved April 18, 2023, from https://www.geeksforgeeks.org/bar-plot-in-matplotlib/

4. Kaggle (n.d.). Heart Disease Retrieved April 18,2023 from https://www.kaggle.com/search?q=heart+disease

5. Analytics Vidhya. (2022, February 1). Heart Disease Prediction using Machine Learning. Retrieved April 18,2023,from https://www.analyticsvidhya.com/blog/2022/02/heart-disease-prediction-using-machine-learning/