HEALTH PROBLEMS PREDICTION PROJECT

Submitted by:
Aji Santhosh Kumar Nair (19BCS1903), Anil Kumar Pradhan (19BCS1898), Keshav Mishra (19BCS1887),
Sahil Lamba (19BCS1922) & Sahil Mallick (19BCS1073)

Role of Machine Learning in our project

Machine learning has been proven to be successful in assisting in decision making and prediction from the huge amount of data provided by the health care sector. This project aims to predict future heart problems by analyzing patient data and determining whether or not they have heart disease using a machine-learning algorithm. Machine Learning techniques can be extremely useful in this regard. Even though heart disease can manifest itself in various ways, there is a common set of key risk factors that impact whether someone is at risk for heart disease or not. We may say that this technique can be extremely well fitted to predict heart disease as it gathers data from numerous sources, classifies it under appropriate headings, and then analyses it to extract the needed data.

Why do we need a Heath Problems Prediction System?

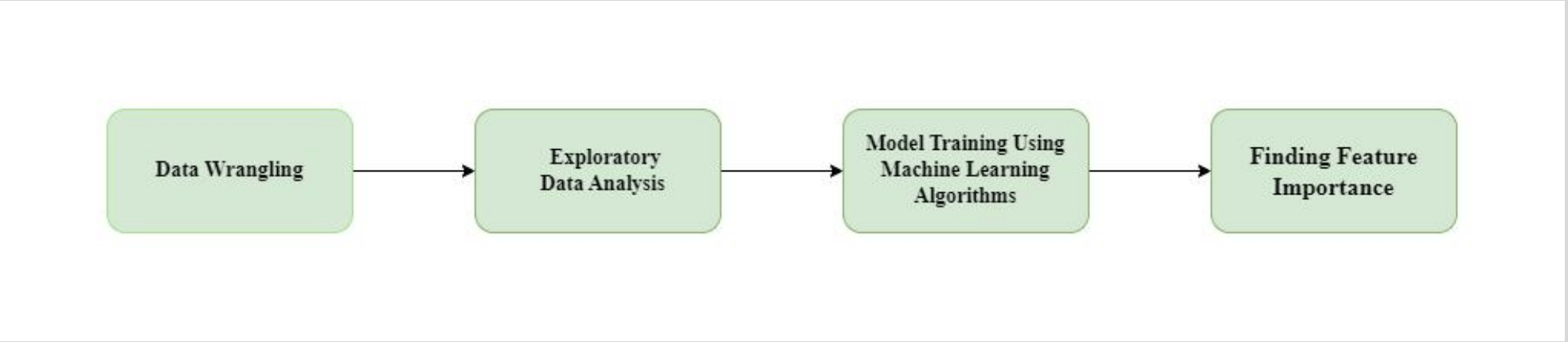
One of the leading causes of morbidity and mortality among the global population is heart disease. The World Health Organization estimates that heart disease causes 12 million deaths worldwide each year. In the last years ago, the prevalence of heart disease has been rising quickly throughout the world. Numerous studies have been carried out in an effort to identify the most important risk factors for heart disease and to precisely estimate the overall risk. Heart disease is also referred to as a silent killer because it causes a person to pass away without any evident signs. Early detection of heart disease is critical in making decisions about lifestyle adjustments in high-risk patients, which minimizes complications.

Modules

This project is implement using the following modules:

- Data Wrangling
- Exploratory Data Analysis
- Model Training Using Machine Learning Algorithms
- Finding Feature Importance.

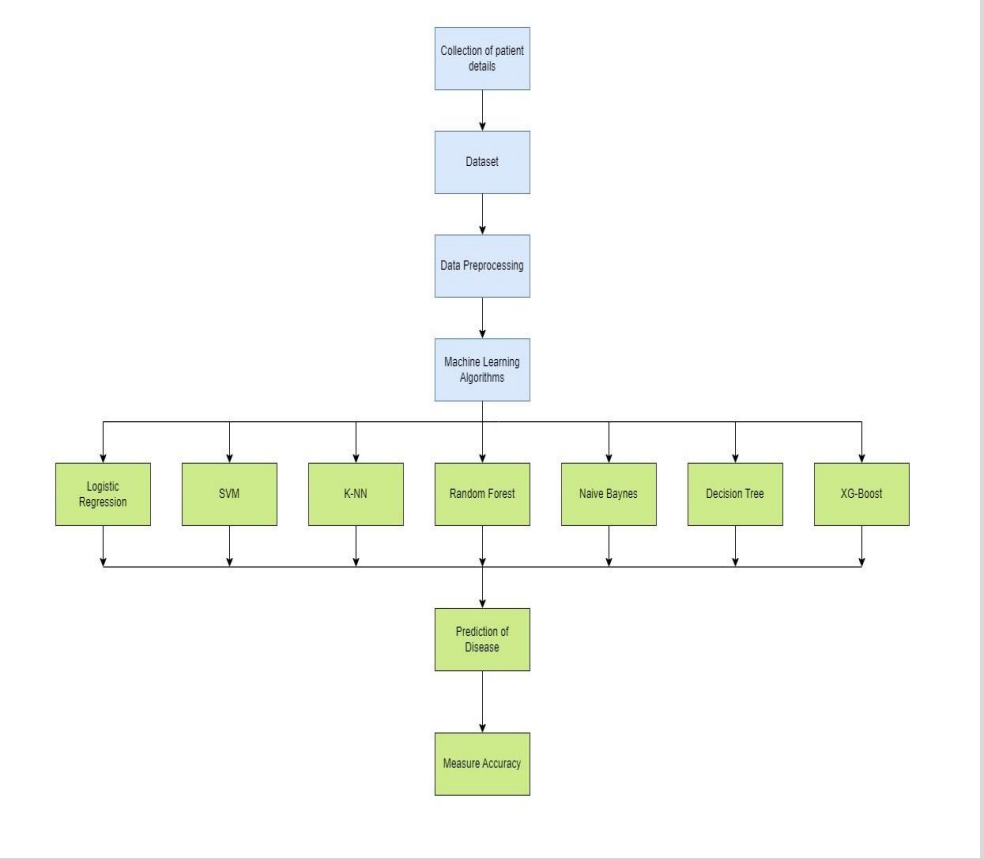
Proposed System



Block Diagram of the Proposed System

Design Flow

Our heart disease prediction system is comparable to this Fig, which also describes the research methods for creating a classification model needed for the prediction of heart problems in patients. The model serves as the foundation for predicting heart disease using any machine learning technique. To make predictions, a classifier must be trained using the dataset, and then a classification model must be created, which is then fed a new unknown record, and the forecast is formed.

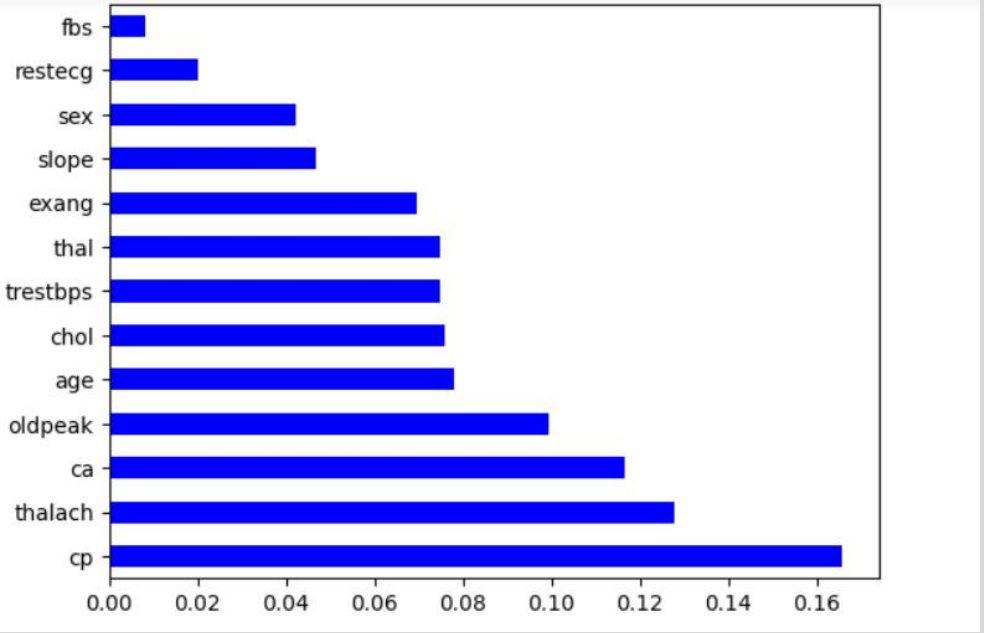


Attributes of the Heart disease dataset

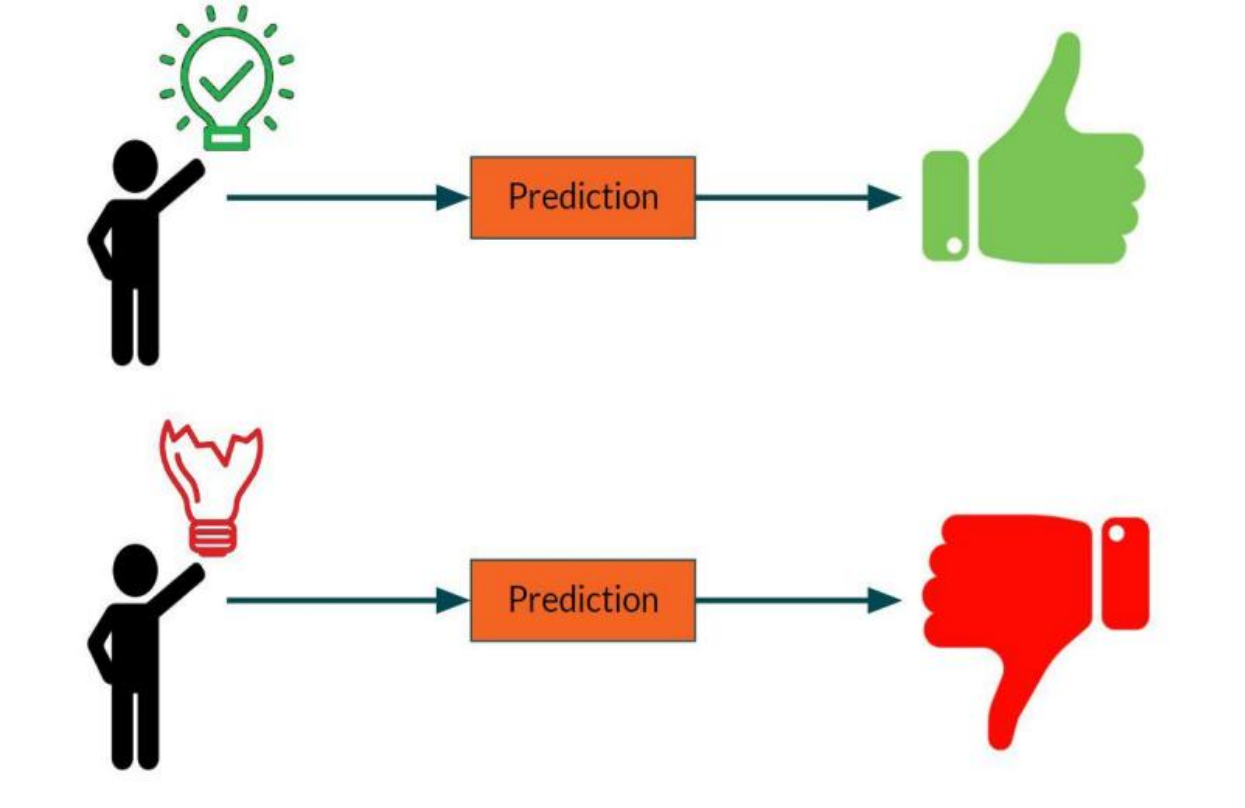
Attribute	Representation	Information Attribute	Description
Age	Age	Integer	Age in years (29 to 77)
Sex	Sex	Integer	Gender instance (0 = Female, 1 = Male)
ChestPainType	Cp	Integer	Chest pain type (1: typical angina, 2: atypical angina, 3: non- anginal pain, 4: asymptomatic)
RestBloodPressure	Trestbps	Integer	Resting blood pressure in mm Hg[94, 200]
SerumCholestoral	Chol	Integer	Serum cholesterol in mg/dl[126, 564]
FastingBloodSugar	Fbs	Integer	Fasting blood sugar > 120 mg/dl (0 = False, 1= True)
ResElectrocardiographic	Restecg	Integer	Resting ECG results (0: normal, 1: ST-T wave abnormality, 2: LV hypertrophy)
MaxHeartRate	Thalach	Integer	Maximum heart rate achieved[71, 202]
ExerciseInduced	Exang	Integer	Exercise induced angina (0: No, 1: Yes)
Oldpeak	Oldpeak	Real	ST depression induced by exercise relative to rest[0.0, 62.0]
Slope	Slope	Integer	Slope of the peak exercise ST segment (1: up-sloping, 2: flat, 3: down-sloping)
MajorVessels	Ca	Integer	Number of major vessels coloured by fluoroscopy (values 0 - 3)
Thal	Thal	Integer	Defect types: value 3: normal, 6: fixed defect, 7: irreversible defect
Class	Class	Integer	Diagnosis of heart disease (1: Unhealthy, 2: Healthy)

Feature Importance

The top four significant features, according to the Feature Importance graph, were chest pain type (cp), maximal heart rate achieved (thalach), number of main vessels (ca), and ST depression generated by exercise relative to rest (oldpeak).



Prediction of Heart Disease



Conclusion

Heart disease is a major killer in India and around the world; applying promising technology such as machine learning to the early prediction of heart disease would have a significant influence on society. Early detection of heart disease can help high-risk patients make lifestyle changes and, as a result, prevent complications, which can be an important achievement in the area of medicine. Every year, more people are diagnosed with heart disease. This requires an early diagnosis and therapy. The use of appropriate technology support in this area can be extremely valuable to the medical community and patients. Logistic Regression, Nave Bayes, SVM, Decision Tree, K-NN, Random Forest and XGBoost are the seven machine learning algorithms used to in this project.

The projected attributes leading to heart disease in patients are present in the dataset, which contains 76 features, 13 of which are crucial for evaluating the system. If all of the features are taken into account, our system will have a lower efficiency. Attribute selection is done to improve efficiency. In this case, n features must be chosen in order to evaluate the model with greater accuracy. Because the correlation of several features in the dataset is almost equal, they are discarded. When all of the attributes in the dataset are included, the efficiency drops dramatically.

The accuracies of all seven machine learning methods are compared, and one prediction model is developed as a result. Hence, the goal is to use multiple evaluation measures such as confusion matrix, accuracy, precision, recall, and f1-score to efficiently predict the disease. When all seven are compared, the extreme Random Forest algorithm has the greatest accuracy of 80%.

Future Work

The future work in this area can be made to produce an impact on the accuracy of the various classification algorithms. For additional improvement applying genetic algorithm in order to reduce the actual data for acquiring the optimal subset of attribute that is perfectly adequate for heart disease prediction. The automated prediction of heart disease using actual real-time data from health care organizations that can be produced using big data. They can be fed as streaming data, and using the data, a real-time study of the patients can be prepared.

Thanks!