8th International Conference on Fusion of Science and Technology
January 6-10, 2020; JCBUST, YMCA, Faridabad (NCR), India
ISFT/2020/P/130

Heart Disease Prediction model using Machine Learning and Data Mining

Mudita Garg
Maharaja Agrasen Institute of Technology, GGSIPU

INTRODUCTION

Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health.
The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyse by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on Cleveland heart diseases dataset from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques.

A. DATA COLLECTION

Two datasets from the UCI Machine Learning Repository [1] were analyzed, for a total of 760 subjects, of which 348 had the diagnosis of heart disease Features analyzed included:
• Number of Major Vessels Colored by Fluoroscopy
• ST Depression Induced by Exercise Relative to Rest
•Chest Pain Type
•Maximum Heart Rate (beats/min)
• Resting Blood Pressure
• Cholesterol
•Gender/Sex
• Resting electro cardio graphic results
• Exercise Induced Angina
• Fasting Blood Sugar
•Slope of the Peak Exercise ST Segment
•Age
• Presence of Defect, Reversible or Fixed

Fig1 shows the correlation matrix of all the features of the dataset. There is a strong correlation between Age , Blood Sugar and Blood pressure.

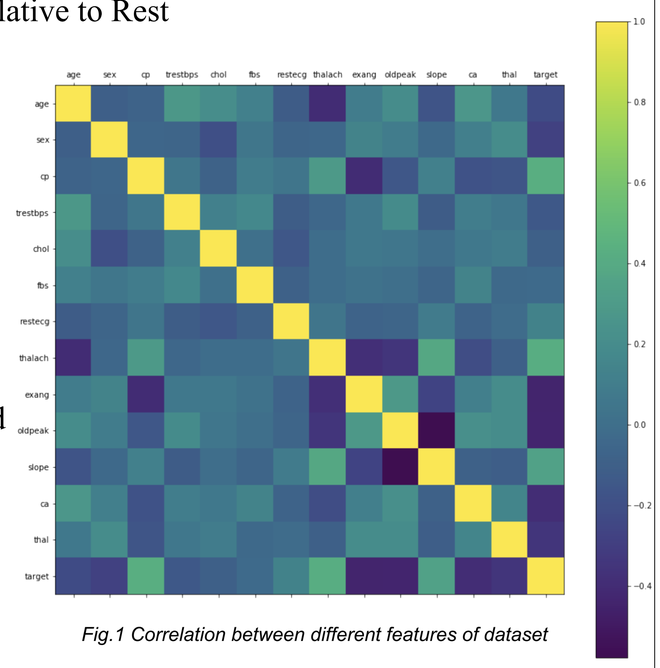


Fig.1 Correlation between different features of dataset

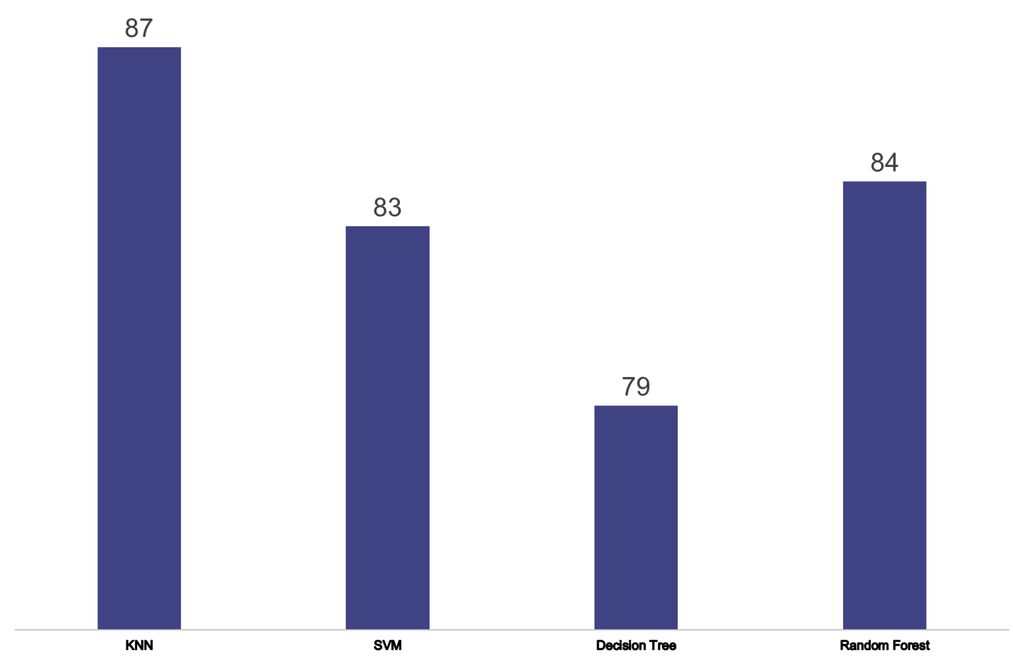
B. DATA CLEANING

The collected data were used to create a structured database system. The fields were identified, duplications were extracted, missing values were completed, and the data was coded according to attribute domain value. After data cleaning the number of cases was reduced mainly due to unavailability of clinical result. Attribute domain values are provided by practicing cardiologist.

MODELLING

• **Logistic regression** is a nonlinear transformation of the linear regression model between the input variables and the binary class assignment. It maps this linear regression using a function like the sigmoid (S-shape) so that all probabilities are mapped between 0 and 1.
• In **KNN**, the K nearest neighbours of a given data point are analysed and the majority class of these K neighbours is assigned to the data point.
• **Support vector machines** (SVM) is a machine learning technique that looks at separations in data when mapped to higher-dimensional space and determines which side of the split a data point is in. This algorithm is used to classify the patients into groups according to the risk posed to them based on the parameters provided.
• A **decision tree** is a flowchart-like structure in which each internal node represents a "test" on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.
In decision analysis, a decision tree and the closely related influence diagram are used as a visual and analytical decision support tool, where the expected values (or expected utility) of competing alternatives are calculated.

RESULTS



CONCLUSIONS

This project analyses the accuracy of prediction of heart disease using an ensemble of classifiers. The Cleveland heart dataset from the UCI machine learning repository was utilized for training and testing purposes.
The project involved analysis of the heart disease patient dataset with proper data processing. Then, 4 models were trained and tested with maximum scores as follows:
K Neighbours Classifier: 87%
Support Vector Classifier: 83%
Decision Tree Classifier: 79%
Random Forest Classifier: 84%

FUTURE WORK

• Obtain and analyse data with a greater feature space.
• Family history of heart disease can also be a reason for developing a heart disease as mentioned earlier. So, this data of the patient can also be included for further increasing the accuracy of the model. This work will be useful in identifying the possible patients who may suffer from heart disease in the next 10 years. This may help in taking preventive measures and hence try to avoid the possibility of heart disease for the patient. So when a patient is predicted as positive for heart disease, then the medical data for the patient can be closely analysed by the doctors. An example would be - suppose the patient has diabetes which may be the cause for heart disease in future and then the patient can be given treatment to have diabetes in control which in turn may prevent the heart disease.

REFERENCES

1. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.
2. Weng, S.F., Reps, J., Kai, J.,Garibaldi, J.M.,Qureshi, N. (2017) Can Machine-learning improve cardiovascular risk prediction using routine clinical data? PLoSONE, 12(4),art. no. 0174944.
3. https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset
4. M.Nikhil Kumar, K.V.S Koushik, K.DeepakDepartment of CSE, VR Siddhartha EngineeringCollege, Vijayawada, Andhra Pradesh, IndiaPrediction Heart Diseases using Data mining andmachine learning algorithms and tools.(2018)
5. J Thomas, R Theresa Princy Human heartdisease prediction system using data miningtechniques (2016)