Improved K-Nearest Neighbor for Predicting Heart Disease

A.Ruhinaaz
PG Student, Department of CSE, Jawaharlal Nehru Technological University College of Engineering (Autonomous) Anantapuramu

***Abstract*— Data and information have become major assets for most businesses. knowledge discovery distinct procedure and information removal an necessary pace in health record. Records are set of information with a exact fine distinct arrangement and reason The existing system uses classification algorithms Logistic regression, artificial neural network, k-nearest neighbor, to check on selected attributes in a training data. Logistic regression algorithm applies on the binary data. Artificial neural is the process of learning to separate samples into different classes by finding common features between samples of known classes. The proposed system uses improved k-nearest neighbor algorithm with Euclidean distance measure. The experimental results show that the proposed system gives better results than existing system in terms of accuracy.**

***Index Terms*—Accuracy, Heart Disease, KNN. Machine Learning, Recall.**

I. INTRODUCTION

According to recent survey by WHO (World health organization) 17.9 million people die each year because of heart related diseases and it is increasing rapidly. With the increasing population and disease, it is become a challenge to diagnosing disease and providing the appropriate treatment at the right time. But there is a light of hope that recent advances in technology have accelerated the public health sector by developing advanced functional biomedical solutions. This paper aims at analyzing the various data mining techniques namely Naïve Bayes, Random Forest Classification, Decision tree and Support Vector Machine by using a qualified dataset for Heart disease prediction which is consist of various attributes li e gender, age, chest pain type, blood pressure, blood sugar etc. The research includes finding the correlations between the various attributes of the dataset by utilizing the standard data mining techniques and hence using the attributes suitably to predict the chances of a heart disease. These machine learning techniques take less time for the prediction of the disease with more accuracy which will reduce the dispose of valuable lives all over the world.

Health is one amid the globe challenges for civilization. World health organization has mentioned that for an Individual proper health is the fundamental right. So, to keep people fit and healthy proper health care services should be provided. 31 percentage of all deaths worldwide are because of heart related problems [6]. Diagnosis and treatment of heart disease is very complex, particularly in developing countries, due to the lack of diagnostic devices and a shortage of physicians and other resources affecting proper prediction and treatment of cardiac patients. With this concern in the recent times computer technology and machine learning techniques are being used to develop software to assist doctors in making decision of heart disease in the preliminary stage. Early stage detection of the disease and predicting the probability of a person to be at risk of heart disease can reduce the death rate. Medical data mining techniques are used in medical data to extract meaningful patterns and knowledge. Medical information has redundancy, multi-attribution, incompleteness and a close relationship with time. The problem of using the massive volumes of data effectively becomes a major problem for the health sector. Data mining provides the tactic and s ill to change these information mounds into helpful executive information. This predication method for heart disease would help cardiologists in taking quicker decisions so that more patients can receive treatments within a shorter period of time, resulting in saving millions of life.

II. LITERATURE SURVEY

As in[1] researchers designed a trained tool to analyse a heart disease using a technique WAC(weighted associative classifier) and Naïve Bayes, the proposed model is used to train medical students and nurses. As in [2] proposed a framework to observe frequent occurring diseases by using a data mining techniques such as Apriority. Visualization model is designed to present the trends graphically. As in [3] compare and evaluated the various performance measures by implementing ten Classification data mining algorithm. The results are evaluated from the patient data base. As in [4]. The author discovered many approach to diagnosis the heart disease .The various classifier techniques are the Decision Tree, Naive Bayes(NB),J48 K-Nearest Neighbours (KNN) and SMO are used to classify data set. As in [5] formulated classification techniques to diagnosis the heart disease. Naive Bayesian integrates with Classification to diagnosis Heart Disease As in [6] designed integrated neural network architecture such as Multilayer Neural Network integrated with Back propagation Learning Algorithm to diagnose the Heart Disease and done an experiment with Heart Disease dataset. An analysis conducted by a research team to predict the risd of Acute Coronary Syndrome (ACS) [7]. They have analyzed the dataset in three seeds and the best Correctly Classified Instances achieved for AdaBoost is 75.49%, 76.28% for Bagging, 72.33% for -NN, 75.30% for Random Forest, 72.72% for SVM [8]. Another analysis was conducted to determine the performances of several algorithms named Decision tree (J48), Naive Bayes, Random Forest, Adaboost, Bagging, Multilayer Perceptron, Simple Logistic to predict diseases using data mining techniques [9]. Around 768 instances with 500 tested negatives and 268 tested positives were used one of the experiments and the function Replace Missing Values in WE A tool was used to handle missing data [10]. A survey showed that the accuracy of different data mining algorithms such as Naive Bayes, Decision Tree, Neural Network , K - Nearest Neighbour and Logistic Regression to predict heart disease depends on the number of risk factors and their types [11]. A smart phone based risk prediction application has developed to predict Heart Attack and different types of risk factors like hypertension, diabetes, dyslipidemia, smoking, family history, obesity, stress, etc. were collected from 506 different patients with three categories: low, medium and high [12]. The app was tested on 89 participants with having Acute Coronary Syndrome (ACS). About 83.9% of patients with high category had ischemic heart disease (IHD) while only 12.5% are in class low and on the other hand, those who have ACS 86.69% of them had high scores [12].

III. PROPOSED SYSTEM

K Nearest neighbor (KNN) is a simple algorithm, which stores all cases and classify new cases based on similarity measure. KNN algorithm also called as 1) case based reasoning63. 2) nearest neighbor 3)example-based reasoning 4) instance based learning 5) memory based reasoning. NN algorithms have been used since 1970 in many applications like statistical estimation and pattern recognition etc.

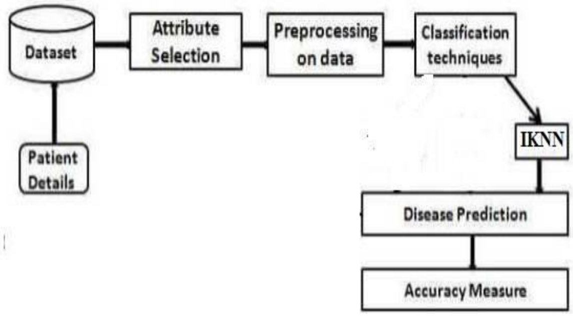
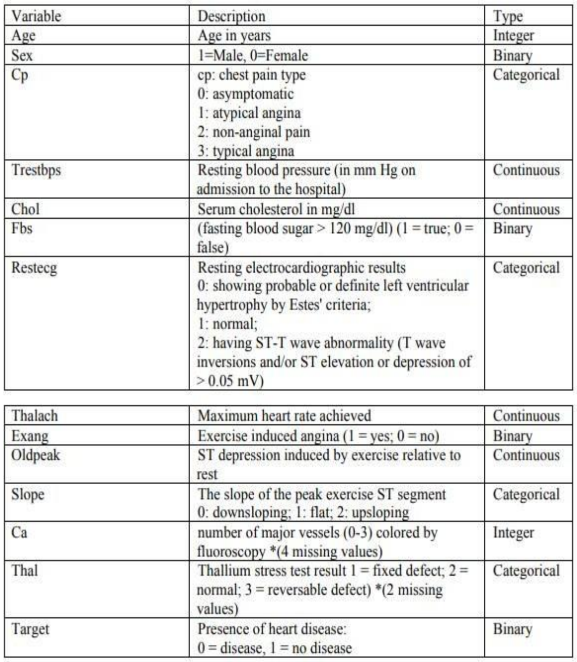


Fig. 1: Proposed system architecture

Here first we need to collect patient details. Then we select attribute selection. After that we preprocess the data. Then we apply classification technique like improved KNN. Finally we predict the risk of disease. The proposed system is as given in fig1.

The dataset applied in this revise is the Cleveland found in the UCI device knowledge warehouse. It consists of 13 attributes estimated on 303 persons. The 14th variable, called target, is a binary variable that signals the presence of heart disease or not. The variables and their descriptions are discussed in fig2.

Information preprocessing prepares rough data for extra taking care of. Data preprocessing is used in informational collection driven applications, for instance, customer relationship the board and rule-based applications. For our proposal we are utilizing standard scaler from the sk-learn library for preprocessing our information. We pick this one over the numerous different ones since it suits very well with our framework.



Variable	Description	Type
Age	Age in years	Integer
Sex	1=Male, 0=Female	Binary
Cp	cp: chest pain type 0: asymptomatic 1: atypical angina 2: non-anginal pain 3: typical angina	Categorical
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)	Continuous
Chol	Serum cholesterol in mg/dl	Continuous
Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	Binary
Restecg	Resting electrocardiographic results 0: showing probable or definite left ventricular hypertrophy by Estes' criteria; 1: normal; 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)	Categorical
Thalach	Maximum heart rate achieved	Continuous
Exang	Exercise induced angina (1 = yes; 0 = no)	Binary
Oldpeak	ST depression induced by exercise relative to rest	Continuous
Slope	The slope of the peak exercise ST segment 0: downsloping; 1: flat; 2: upsloping	Categorical
Ca	number of major vessels (0-3) colored by fluoroscopy *(4 missing values)	Integer
Thal	Thallium stress test result 1 = fixed defect; 2 = normal; 3 = reversable defect) *(2 missing values)	Categorical
Target	Presence of heart disease: 0 = disease, 1 = no disease	Binary

Fig 2: Cleveland Dataset Features Description

Algorithm

Step 1: Locate the K training cases which are neighboring to unidentified cases.

Step 2: Find the distances.

Step 3: Rank them using standard ranking.

Step 4: Select attributes based on ranking.

Step 5: Pick the most commonly occurring classification for these instances.

Step 6: Compute accuracy of the classifier, which calculates the capability of the classifier to correctly classify unknown model.

IV. EXPERIMENTAL RESULTS

The experiments are carried on Cleveland heart disease dataset. Accuracy is termed as ratio of the number of correctly classified instances to the total number of instances.

Accuracy = (TP+TN) / (TP+TN+FP+FN)

TP – True Positive

TN – True Negative

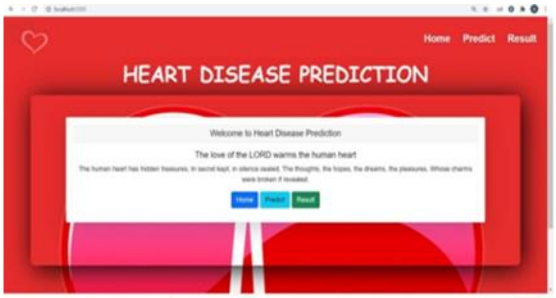
FP – False Positive

FN – False Negative

Recall is the ratio of actual true instances out of all the items which are true. Recall = TP/(TP+FN)

F - Measure is the harmonic mean of both precision and recall.

F - Measure = 2*(Precision*Recall)/(Precision + Recall)



HEART DISEASE PREDICTION

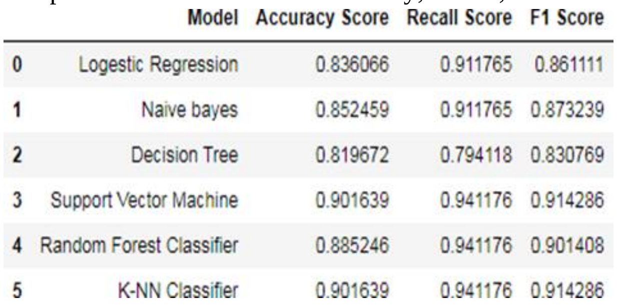
Welcome to Heart Disease Prediction

The love of the LORD warms the human heart

The human heart has hidden treasures, in secret kept, in silence sealed. The thoughts, the hopes, the dreams, the pleasures, Whose charms were broken if revealed.

Fig 3. Performance of different classifiers

As shown in fig3 the IKNN performs better when compared to other models in accuracy, Recall, F1.



	Model	Accuracy Score	Recall Score	F1 Score
0	Logestic Regression	0.836066	0.911765	0.861111
1	Naive bayes	0.852459	0.911765	0.873239
2	Decision Tree	0.819672	0.794118	0.830769
3	Support Vector Machine	0.901639	0.941176	0.914286
4	Random Forest Classifier	0.885246	0.941176	0.901408
5	K-NN Classifier	0.901639	0.941176	0.914286

Fig 4. Accuracy Graph

The accuracy of the proposed IKNN performs better when compared to existing models as shown in fig4 and heart disease prediction system is as shown in fig5.

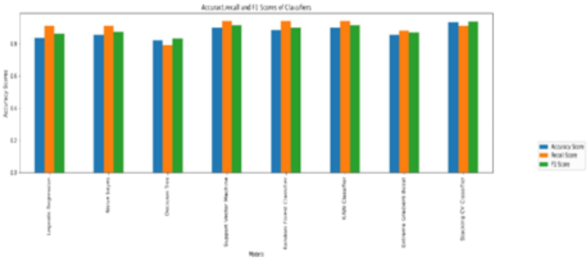


Fig 5. Home page of Heart Disease prediction system

V. CONCLUSIONS

The recommended work shows that the expectation of hazard from cardiovascular infections gives best outcomes dependent on actual elements. The result generated by this system has been evaluated and validated on data of patients with the Doctor's diagnosis (predictions). This proposed system will help the doctors to plan for a better medication and provide the patient with early diagnosis as it performs reasonably well even without retraining. The Improved K-Nearest neighbor method is proposed for heart disease prediction. It compared with all other machine learning techniques .The results showed a great accuracy standard for producing a better estimation result.

REFERENCES

[1]N. AdityaSundar, P. Pushpa Latha, M. Rama Chandra, “Performance analysis of classification data mining techniques over heart disease data base”, International journal of engineering science & advanced technology Volume2, Issue-3, 470 – 478.

[2]Mohammed Abdul Khaleel, Sateesh Kumar Pradhan, Finding Locally Frequent Diseases Using Modified Apriori Algorithm,” International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 10, October 2013.

[3] Chris Ding, Xiaofeng He,K -means Clustering via Principal Component Analysis, Chris”, Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720.

[4]R. Lakshmi, M.Veera Krishna, S. Prem Kumar, “Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability”, International Journal of Scientific and Research Publications.

[5]Boshra Bahrami, Mirsaeid Hosseini Shirvani, “Prediction and Diagnosis of Heart Disease by Data Mining Techniques”, Journal of Multidisciplinary Engineering Science and Technology (JMEST) ISSN: 3159-0040 Vol. 2 Issue 2, February–2015.

[6]Carlos Ordonez, Edward Omincenski and Levien de Braal,“Mining Constraint Association Rules to Predict Heart Disease”, IEEE International Conference on Data Mining, IEEE Computer Society, ISBN-0-7695-1119- 8, pp: 433-440, 2001.

[7]M. Raihan, M. Islam, P. Ghosh, S. Shaj, M. Chowdhury, S. Mondal and A. More, "A Comprehensive Analysis on Risk Prediction of Acute Coronary Syndrome Using Machine Learning Approaches", in 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1 - 6.

[8]D. Kinge and S. Gaikwad, "Survey on data mining techniques for disease prediction", International Research Journal of Engineering and Technology (IRJET), vol. 05, no. 01, pp. 630-636, 2018.

[9]N. Kaur, V. Gupta, S. Kataria, "An Efficient Hybrid Classifier to Improve the Health Prediction using Data Mining", International Journal of Research in Electronics and Computer Engineering, vol. 06, no. 03, pp. 1171- 1174, 2019.

[10]B. Rathnayakc and G. Ganegoda, "Heart diseases prediction with Data Mining and Neural Network Techniques", in 2018 3rd International Conference for Convergence in Technology (I2CT), Pune, India, 2018, pp. 1 - 6.

[11]More, M. Raihan, A. More, S. Padule and S. Mondal, "Smart phone based “heart attack” risk prediction; innovation of clinical and social approach for preventive cardiac health", Journal of Hypertension, vol. 36, p. e321, 2018

[12]M. Marcus, "Some jobs seem riskier when it comes to heart health", CBS NEWS, 2016.