Tanvi Sharma et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 4, Issue 3, Sept 2017, pp. 17-21

Prediction of Heart Disease Using Cleveland Dataset: A Machine Learning Approach

Tanvi Sharma, Sahil Verma, Kavita
Kurukshetra University, Kurukshetra (Haryana)

Abstract: **A large amount of data is accumulated by the health-care industry. This data contains effective patterns which enable efficient decision-making. These patterns often go unexplored. Various machine learning methods can be incorporated in such situation. This work uses various machine learning methods such as Decision Tree, MARS, Random Forest and TMGA to realize the data mining goals. Results show that the Decision Tree method predicts the diagnosis of heart disease most effectively from patient's data. The dataset used to carry on this research work is taken from the popular UCI repository and is known as the Cleveland Dataset. It is implemented on the R platform.**

Keywords: **Machine Learning, Prediction, Heart Disease, Decision Tree**

1. Introduction

Data mining is an amalgamation of various fields such as machine learning, image processing, pattern recognition, statistical manipulations [1]. It is basically concerned with the data manipulation and data processing. It provides an insight into the data. The data contains knowledge about the structure, models and useful patterns. These patterns are utilized to have in-depth knowledge about the data [2].

Healthcare is a data intensive process. Many processes run simultaneously producing new data every second [3]. This is a field in which a lot of work has been carried out now-a-days. With the help of medical field data and computer aids, new algorithms can be derived which are very effective in predicting the occurrence and non-occurrence of a disease. The combined field of knowledge is termed as health informatics [4].

The data collection about different diseases is very important. Medical and health areas are among the most important sections in industrial societies [5]. The extraction of knowledge from a massive volume of data related to diseases and medical records using the data mining process can lead to identifying the laws governing the creation, the development of epidemic diseases.

Some medical applications of data mining are [5, 6]:

- Prediction of cost incurred in health-care.
- To determine the treatment of a particular disease.
- Diagnosis, prediction of diseases of most kind etc.

Health informatics is defined as an evolving scientific discipline that deals with the collection, storage, retrieval, communication and optimal use of health related data, information and knowledge [6]. It is the field of study applied to clinical care, nursing, public health and biomedical research all dedicated to the improvement of patient care and population health.

Data mining for healthcare is useful in evaluating the effectiveness of medical treatments. Through comparing and contrasting various causes, symptoms, and treatment methodologies, data mining can produce an analysis of treatments that can correct specific symptoms most effectively [7]. It is widely used in healthcare fields due to its descriptive and predictive power. It can predict health insurance fraud, healthcare cost, disease prognosis, disease diagnosis, and length of stay needed in a hospital. It also obtains frequent patterns from biomedical and healthcare databases such as relationships between health conditions and a disease, relationships among diseases and relationships among drugs etc.

The quality healthcare facility targets at [7]:

- Provision of the healthcare treatments which are safe.
- Use of computerized knowledge in determining and predicting health-related issues.
- Provision of various treatments to be given to patients on the basis of collected and analyzed patient's information.
- Effective treatment to be made available in lesser time.
- Timely and efficient prediction of a disease.

2. Materials and Methods

Risk of heart disease increases due to a number of factors including age, family history, smoking, poor diet, high blood pressure, high blood cholesterol and obesity. Cleveland Heart Disease The dataset is available for the sake of prediction of heart disease at the UCI Repository. The attributes used in the course of this work is given below in Table 1:

1. Analysis of Heart Disease Prediction Methods

Data Mining was developed to extract the knowledge and experience in the software used. R is an open source software program that was developed to be an independent data mining tool which provides many different algorithms

for data mining and machine learning has been used in the present study.

Table 2: Machine Learning Models

Model Name	Accuracy	Time Taken
Decision Tree [8, 9]	93.24	112.36
Multivariate Adaptive Regression Splines (MARS) [10]	91.04	112.42
Random Forest [11]	89.95	112.36
Tree-Model from Genetic Algorithm (TMGA) [12]	88.85	112.39

Decision trees are frequently used in classification problem. It is simpler method where each feature is represented by an internal node and further branches are build upon the test conditions and their outcome. Multivariate adaptive regression splines (MARS) is a dual use model. This model is applied to both the regression and classification problems. Basically, the reliability can be calculated using this model. Random forest is based upon the ensembling technique. It works efficiently on large voluminous dataset. TMGA is a tree-model based upon the evolutionary algorithm. It is also a dual-use model which can handle both regression and classification problems optimally

Table 1: Attributes and their Values

Attributes	Values
age	In years
sex	0 = female and 1 = male
cp	Chest pain type: → Value 1: typical angina → Value 2: atypical angina → Value 3: non-anginal pain → Value 4: asymptomatic
trestbps	Resting blood pressure
chol	Cholesterol
fbs	Fasting blood sugar
restecg	Resting cardiographic results
thalach	Maximum heart rate achieved
exang	Exercise induced angina (1 = Yes and 0 = No)
oldpeak	ST depression induced by exercise relative to rest
slope	The slope of the peak exercise ST segment
ca	Number of major vessels (0-3) colored by flourosopy
thal	3 = Normal; 6 = Fixed Defect; 7 = Reversable Defect
Num	Diagnosis of heart disease (angiographic disease status) → Value 0: < 50% diameter narrowing → Value 1: > 50% diameter narrowing

The proposed system is shown below:

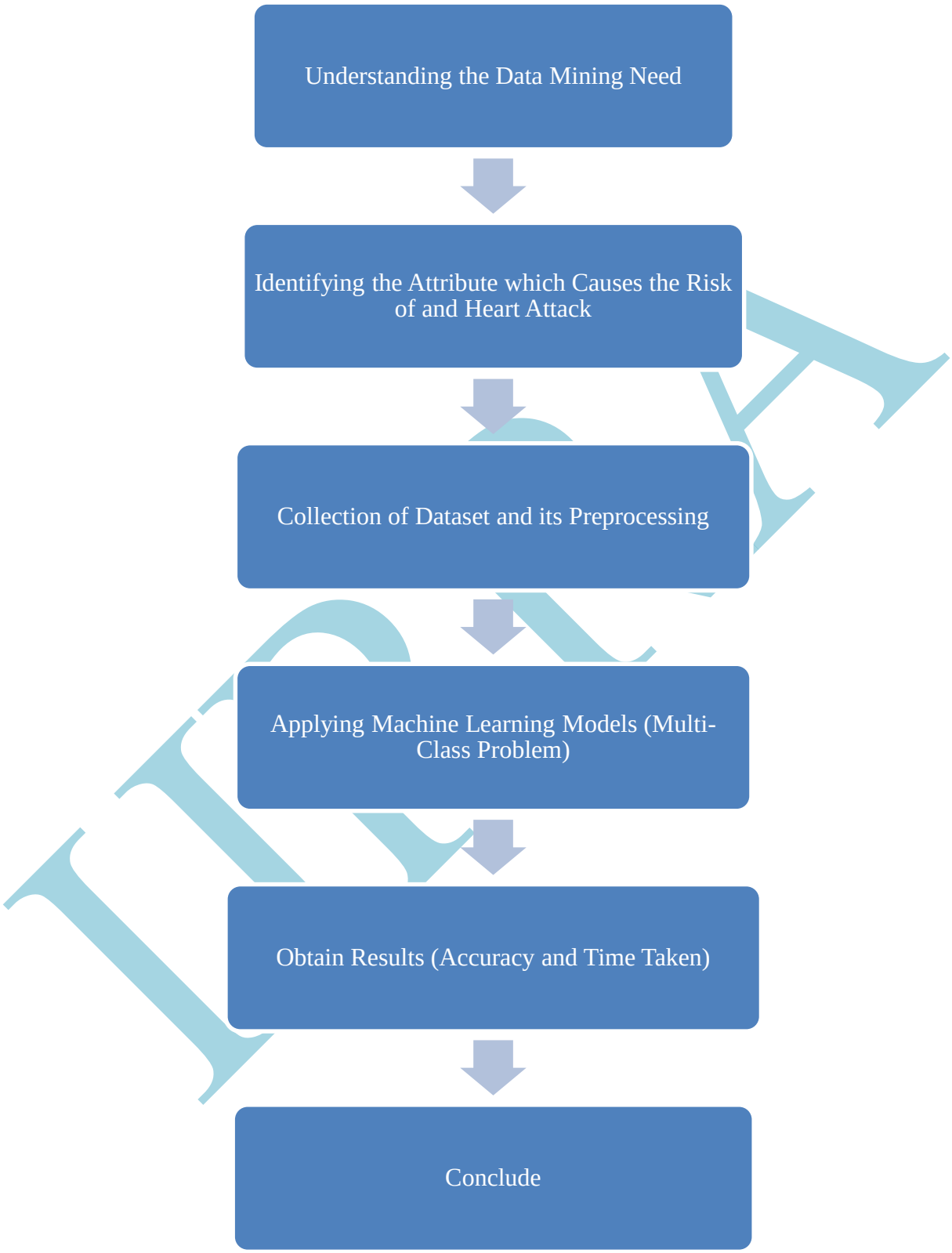

Figure 1: The Proposed Workflow

.

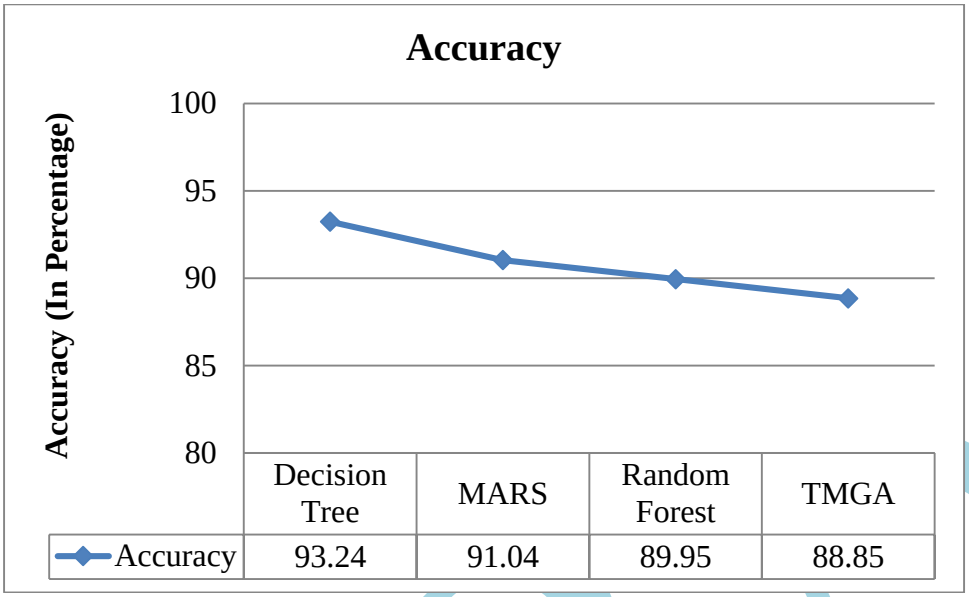

Figure 2: Accuracy Plot

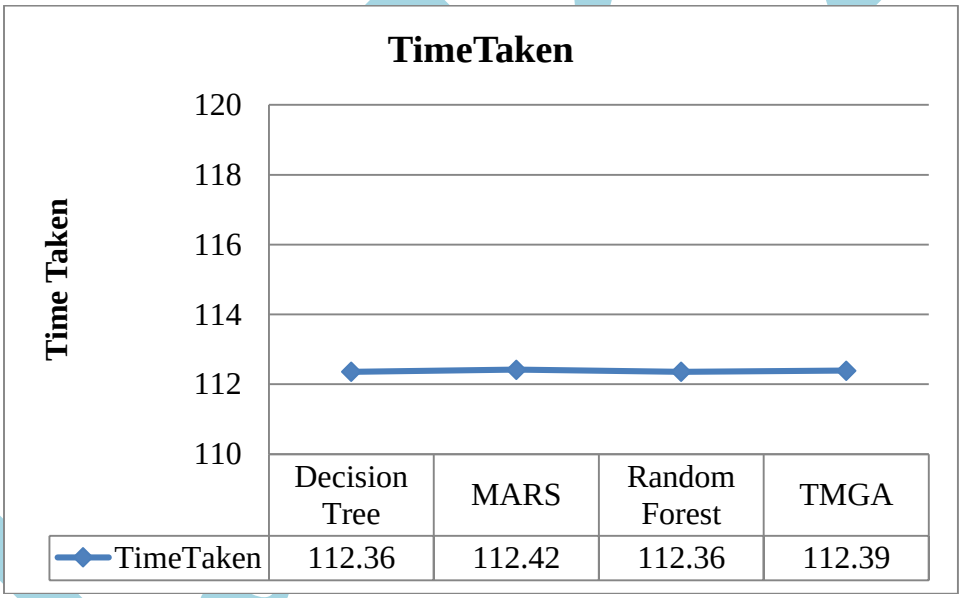

Figure 3: Time Taken Plot

The above Figure 2 and 3 clearly shows that Decision Tree model best predicts upon the Cleveland Dataset. It yields the best accuracy and in minimum time span. The standard dataset partition for training-testing dataset is 70-30. This can be further scaled to yield much better results.

2. Conclusion

In this work, we explore four machine learning methods with fourteen properties for predicting the heart disease according to the Cleveland dataset. The absolute quality of the results is calculated from the accuracy and the time taken by the machine learning models. From the above discussion, it is concluded that the Decision Tree performs the best out of MARS, Random Forest and TMGA in terms of accuracy as well as time taken.

References:

[1] Roiger, R.J., 2017. *Data mining: a tutorial-based primer*. CRC Press.

[2] Shmueli, G. and Lichtendahl Jr, K.C., 2017. *Data Mining for Business Analytics: Concepts, Techniques, and Applications in R*. John Wiley & Sons.

[3] Wang, Y., Kung, L., Wang, W.Y.C. and Cegielski, C.G., 2017. An integrated big data analytics-enabled transformation model: Application to health care. *Information & Management*.

[4] Pradhan, M., 2017. Status and Challenges on Adaptation for Indian Healthcare Services with Data Mining Technique. In *Innovative Healthcare Systems*

for the 21st Century (pp. 285-298). Springer International Publishing.

[5] Vatsalan, D., Sehili, Z., Christen, P. and Rahm, E., 2017. Privacy-Preserving Record Linkage for Big Data: Current Approaches and Research Challenges. In *Handbook of Big Data Technologies* (pp. 851-895). Springer International Publishing.

[6] Zhang, Y., Qiu, M., Tsai, C.W., Hassan, M.M. and Alamri, A., 2017. Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. *IEEE Systems Journal*, *11*(1), pp.88-95.

[7] Chen, J.H. and Asch, S.M., 2017. Machine learning and prediction in medicine—beyond the peak of inflated expectations. *N Engl J Med*, *376*(26), pp.2507-2509.

[8] Veronesi, F., Korfiati, A., Buffat, R. and Raubal, M., 2017, May. Assessing Accuracy and Geographical Transferability of Machine Learning Algorithms for Wind Speed Modelling. In *International Conference on Geographic Information Science* (pp. 297-310). Springer, Cham.

[9] Le-Dong, N.N., Hua-Huy, T., Nguyen Ngoc, H.M., Martinot, J.B. and Dinh Xuan, A.T., 2017. Detection Of Interstitial Lung Disease In Systemic Sclerosis Using A Machine Learning Approach Based On Pulmonary Function Tests. In *A76. WHAT'S IN THE TOOL BOX TO ASSESS LUNG FUNCTION* (pp. A2531-A2531). American Thoracic Society.

[10] Roy, S.S., Roy, R. and Balas, V.E., 2017. Estimating heating load in buildings using multivariate adaptive regression splines, extreme learning machine, a hybrid model of MARS and ELM. *Renewable and Sustainable Energy Reviews*.

[11] Subasi, A., Alickovic, E. and Kevric, J., 2017. Diagnosis of Chronic Kidney Disease by Using Random Forest. In *CMBEBIH 2017* (pp. 589-594). Springer, Singapore.

[12] Naghibi, S.A., Ahmadi, K. and Daneshi, A., 2017. Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping. *Water Resources Management*, pp.1-15.

