iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53752

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Prognosis of Heart Disease using Machine Learning Algorithms

Shivam Chauhan[1], Shraddha Pandey[2], Simmi Singh Panwar[3], Vaishnavi Ojha[4], Prof. Yashi Bhardwaj[5]
[1, 2, 3, 4]*B.Tech IT Department Students, IMS Engineering College Ghaziabad, Uttar Pradesh India*
[5]*Assistant Professor, IMS Engineering College, India*

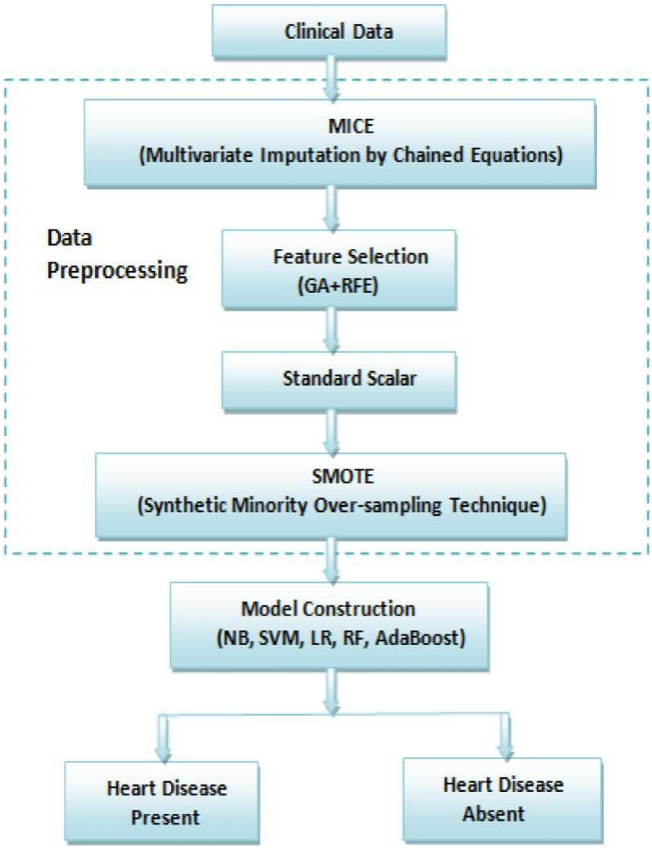
Abstract: ***This article presents the prediction of the heart diseases by using the machine learning algorithm. One of the major causes of morbidity in the world's population is the prediction of heart attacks.***
Cardiovascular disease is a very essential disease that is included in the clinical data analysis as one of the most crucial sections for the prediction. In this study, we describe a technique to heart attack prediction that uses machine learning to analyse several risk factors and make predictions about heart attacks.
Heart disease cases are rising quickly every day, thus it's crucial and worrisome to predict any potential illnesses in advance. This diagnosis is a challenging task that requires accuracy and efficiency.
The primary focus of the research paper is on which patients, given certain medical characteristics, are more likely to suffer heart disease. Using the patient's medical history, we developed a system to determine if a heart disease diagnosis is likely or not for the patient.
Keywords: Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, Heart Disease Prediction.

I. INTRODUCTION

Heart disease is the leading cause of death worldwide. Cardiovascular disease kills more people each year than any other cause, with approximately 12 million deaths from heart disease each year.
In the United States, someone dies of a heart attack every 34 seconds. A heart attack is usually a tragic event caused by a blockage of blood flow to the heart or brain. People at risk for heart disease may experience increased blood pressure, high blood sugar and lipid levels, and stress.
All of these parameters can be easily measured at home with basic sanitation. Coronary heart disease, cardiomyopathy, and cardiovascular disease are the categories of heart disease.
The term "heart disease" covers a variety of conditions that affect the heart and blood vessels, and how fluid enters the blood and circulates throughout the body. Cardiovascular disease (CVD) is responsible for much morbidity, disability and death. The diagnosis of diseases is an important and complex task in medicine.
Medical diagnosis is considered an important but difficult task that needs to be done effectively and efficiently. It is very useful to automate this task. Unfortunately, not all physicians are experts in a discipline and there are places where resources are scarce. Data mining can be used to uncover hidden patterns and knowledge that contribute to successful decision making.
It plays a key role in helping healthcare professionals make informed decisions and provide quality services to the public. The approach that health organizations offer to professionals who no longer have the knowledge and skills is also very important.
One of the main limitations of existing methods is the inability to draw precise conclusions when needed. In our approach, we use different data mining techniques and machine learning algorithms, Naive Bayes, k-Nearest Neighbours (KNN), Decision Trees, Artificial Neural Networks (ANN), Random Forests to predict heart disease based on certain health parameters regression is used, which provides the higher accuracy as compared to other algorithm of machine learning. Once a prediction rate of the diabetes has been achieved by the medical department, then further safety measures can be achieved.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

II. FLOWCHART



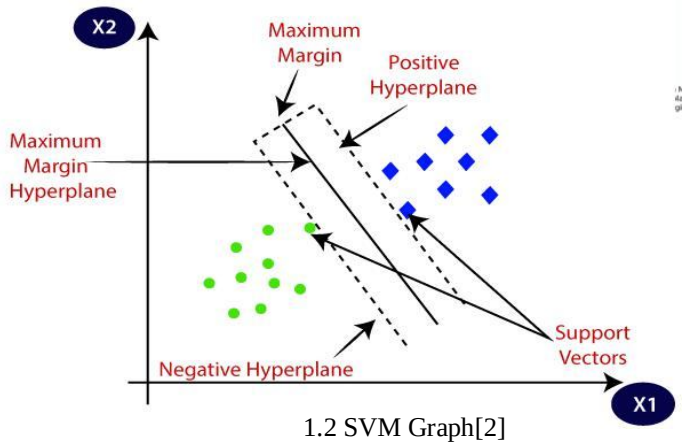
1.1 Generic Model Predicting Heart Disease[1]

III. MODELS USED

A. SVM

Finding the best line to divide the data into the two categories is how SVM operate[3]. This is accomplished through the use of an optimization procedure that only takes into account the training dataset's data examples that are most closely related to the line that best demarcates the classes. The technique's name derives from the fact that the examples are known as support vectors.

Lastly, only a small number of records can be divided by a simple straight line. There are times when a straight line or even an area made of polygons needs to be marked out. Through the use of SVM, this is accomplished by transforming the data into a higher dimensional space before drawing the conclusions and making the forecasts.



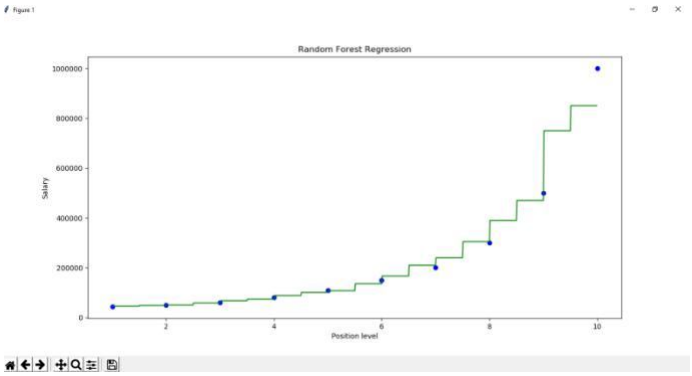
1.2 SVM Graph[2]

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

B. Random Forest

In order to increase the dataset's predictive accuracy, a classifier called Random Forest uses many decision trees on different subsets of the input data. It is a very well-liked supervised machine learning technique.

Learning techniques called Random Decision Forests and Random Forests are similar. This approach is used for jobs like classification and regression, among others. The concept behind these techniques is to construct numerous decision trees during training time and output either the mean prediction or the class of each individual tree. (regression).

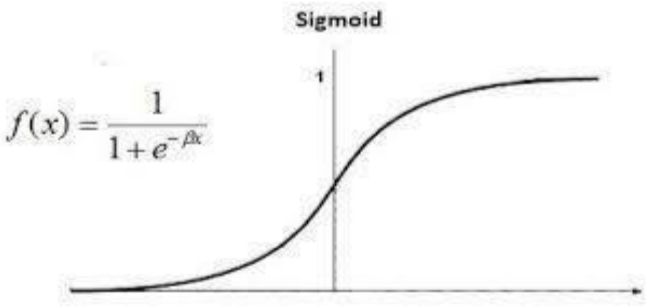


1.2 Random Forest Graph[1]

1.3

C. Logistic Regression Model

A regularization technique[3] called a ridge estimator is used in the application. By minimizing the coefficients the model learns during training, this technique aims to make the model simpler. How much pressure to apply to the method to shrink the size of the coefficients is determined by the ridge parameter. This regularization will be disabled if this is set to 0. Based on some dependent variables, the machine learning classification process known as logistic regression is used to forecast the likelihood of a given class. In essence, the logistic regression model generates the logistic of the outcome after computing the sum of the input features (in most cases, there is a bias term). It forecasts a categorical dependent variable's result. As a result, the result must be a discrete or categorical value.



1.4 Logistic Regression Graph[2]

D. Software Requirement Proprieties

Python 3.6.0 is a dynamic object-oriented programming language that may be applied to a variety of software development projects. It comes with substantial standard libraries, has excellent support for integration with other languages and technologies, and can be learnt in a fewdays. Many Python developers claim to have experienced significant productivity increases and believe the language promotes the creation of higher quality, more maintainable code.

Jupyter Notebook: A server-client programme that enables editing and running notebook papers through a web browser is known as the Jupyter Notebook App. The Jupyter Notebook App can be deployed on a remote server and accessible via the internet, or it can be run locally on a desktop without the need for an internet connection (as detailed in this paper). The Jupyter Notebook App contains a "Dashboard" (Notebook Dashboard), a "control panel" exposing local files and letting to open notebook documents or shutting down their kernel in addition to displaying, editing, and running notebook documents.

II. LITERATURE SURVEY

A quiet Significant amount of work related to the diagnosis of Cardiovascular Heart disease using Machine Learning algorithms has motivated this work. We propose random forest method for the prediction of heart disease in order to create a machine learning system that can more accurately predict a patient's risk of occuring heart attack early on.The results indicated that the prediction system is able to forecast the heart disease effectively, efficiently, most significantly,and quickly.

In the world of medicine, ML algorithms are well- known for their ability to predict disease. In an effort to get the best and most accurate findings, numerous researchers have used ML approaches to predict heart disease including SVM, logistic regression, decision tree, K-Nearest Neighbors (KNN), and Random Forest.A brief literature survey is presented here.

Avinash Golande and et al. studies various different ML algorithms that can be used for classification of heartdisease. Research was carried out to study Decision Tree, KNN and K-Means algorithms that can be used for classification and their accuracy were compared[1]. This research concludes that accuracy obtained by Decision Tree was highest further it was inferred that it can be made efficient by combination of different techniques and parameter tuning.

Theresa Princy. R, et al. executed a survey including different classification algorithm used for predicting heart disease. The classification techniques used were Naive Bayes, KNN (K Nearest Neighbour), Decision tree, Neural network and accuracy of the classifiers was analyzed for different number of attributes [5].

Nagaraj M Lutimath, et al. has performed the heart disease prediction usinh naive bayes classification and SVM(Support Vector Machine). The performance measures used in analysis are Mean Absolute Error and Root Mean Squared Error, it is established that SVM was emerged as superior algorithm in terms of accuracy over Naive Bayes.

Authors were able to get a prediction accuracy of 80% using the configuration provided. Authors developed a prediction model based on a single machine learning algorithm in the relevant study displayed above. It is evident that a single machine learning algorithm approach did not produce outstanding prediction results, but that the outcomes may be improved by combining different machine learning techniques into an ensemble.

III. METHODS USED

First, we preprocess two separate data sets. In the preprocessing step, correlations between the attributes of the dataset are analyzed to find useful features for diagnosing diabetes. The data is then split into two sets, training and testing. The training set is used to develop predictive machine learning models using various machine learning algorithms. It then evaluates the performance of the proposal on various metrics. Finally, we use Flask to deploy our best machine learning models to our web application. Then we outline the workflow for each part.

A. Data Collection

To ensure the robustness of the model, two alternative data sets were collected, each containing a different number of elements or features. The data set was compiled from variety of sources, including diabetes statistics and health profiles from people around the world and from various medical Institutions.

B. Data Cleaning

The first step in handling a data set is to clean it by following a systematic procedure to remove unnecessary records and attributes. First, the dataset consists of several categorical values that need to be removed for privacy reasons, such as hospital number, episode date, and episode description. The data set alsoconsisted of missing diabetes type values for some patients, which is important information for our study because we studied diabetes complications in diabetic patients. Therefore, all 26 instances experiencing this issue were removed. Anothernecessary step in this study is checking the total number of missing values per record (or patient). By testing different percentages using all the classifiers, it was found that removing all the records with 60% of missing values achieved better performance compared to other experiments where this problem was ignored. Following the approach in [19], the missing values were also investigated per attribute. Based on several experiments, a threshold of 40% was set for this step, meaning that any attribute with missing values larger than or equal to 40% should be dropped from the dataset. Since this dataset haslarge number of numerical attributes, it was found that 16numerical attributes have missing values of more than 40%. More precisely, most of these properties have more than 90% missing values. The whole database is split into training and testing database. The 80% data is taken for training while remaining 20% data is used for testing.

C. Choosing a Model

After simulation data is generated Using four Machine learning algorithms we will implement predictions of diabetes.

1) *Random Forest:* It is an ensemble teaching method for Classification and regression and other tasks. It works by constructing a set. Decision trees during training and inference [1] A class that is a class mode, or Average prediction of individual trees. First The algorithm for a random decision forest is Generated by Tin Kam Ho using random subspace method. Ho set to get it Accuracy He should overtrain if possible. Randomly limit sensitive selection function given data.
2) *K-nearest Neighbors:* Using the WEKA[1] tool and the K-nearest Neighbor classification technique, the training dataset, input, and output are selected in the first step.Factors must originate from. The second part of this course is standardizing the data, which ensures that the distance degree assigns the same weight to each variable. k-nearest neighbors are nonparametric. Methods used for classification and regression. The input consists of k-nearest training. An example of a feature space. Distance from point of interest to point. The training data set you use. In classification,The value of k in the method is always positive. The k-NN algorithm is local data structure.
3) *Decision Tree:* Classification and Regression Trees is a more modern name for decision trees. (CART). To assess a data instance, they build a tree, starting at the root and moving to the leaves (roots) until a prediction can be made. The method for building a decision tree involves repeatedly choosing the best split point to use as a prediction tool until the tree reaches a set depth. The tree is built, and then it is trimmed to increase the model's generalizability to new data.
4) *Logistic Regression:* Logistic regression is machine learning Algorithm used for classification problem, this is a predictive analytics algorithm. It is based on the concept of probability. You can call logistic regression linear. Regression model but logistic regression Use a more complex cost function. It is called the logistic function. theory Logistic regression tends to limit the cost. It is a function between 0 and 1. So linear A function cannot show it as it can the result of this function 1 if the hypothesis is large, 1 if it is small Then 0.

D. Model Training

After processing the dataset and selecting the machine learning algorithms to be used, the next step was to build the actual models by training each algorithm using the processed dataset. Extensive experiments were conducted both to train and fine-tune the models. The rest of this section will discuss the detailed steps.

E. Testing

We often test our trained model in this.

IV. FUTURE SCOPE

Further research needs to be conducted to improve classification accuracy using advanced algorithms such as bagging, support vector machine or decision table, etc.

We can determine the predictive performance of each algorithm and apply the proposed system to the region of interest. We can add other necessary features to improve the accuracy of the algorithm. Stakeholders should use it as a dedicated tool to make better decisions. We have not changed the parameters of our implementation.

It can be improved and adjusted by changing the experimental settings in the future. In the future, much more could be done using more heart disease data and using different data reduction techniques. For better outcomes and predictions of heart disease, high quality datasets without inconsistencies can be used.

V. CONCLUSION

Our research focuses on the application of data mining techniques in healthcare, particularly in the detection of heart disease. Heart disease is a deadly disease that can lead to death. Data mining techniques are implemented using the following algorithms, KNN, Neural Networks, Decision

Trees, Naive Bayes, and Random Forests. We measure performance in terms of accuracy, TN, FP, FN and TP rates, and some algorithms. We conducted five experiments with the same data set to predict heart disease.

The results of all implemented algorithms are displayed in tabular form for better understanding and comparison. Experiments show that Naive Bayes has the highest accuracy rate of 88%, followed by ANN and KNN with an accuracy rate of 87%. Our results suggest that data mining can be used and applied in the healthcare industry to predict and diagnose diseases at an early stage,

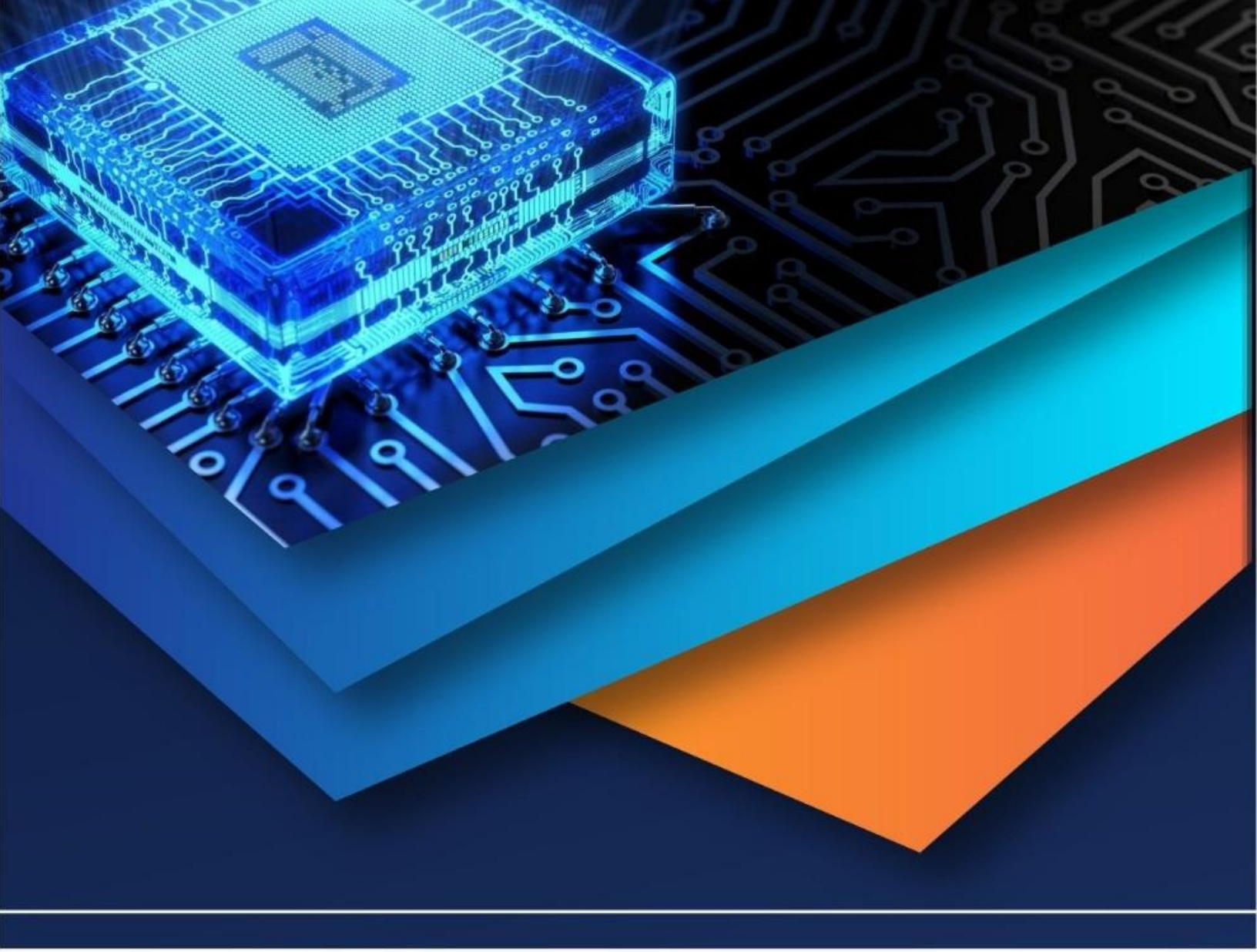
Algorithms	Accuracy	TN	FP	FN	TP
KNN	0.87	26	7	3	40
ANN	0.87	30	2	8	36
Naïve Bayes	0.88	31	4	5	36
Decision Tree	0.78	42	15	5	29
Random Forest	0,82	42	11	5	33

Table 3.1 Analysis of Machine Learning Algorithm[5]

REFERENCES

[1] Avinash Golande, Pavan Kumar T, "Heart Disease Prediction Using Effective Machine Learning Techniques", International Journal of Recent Technology and Engineering, Vol 8, pp.944- 950,2019.
[2] T.Nagamani, S.Logeswari, B.Gomathy," Heart Disease Prediction using Data Mining with Mapreduce Algorithm", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-3,January 2019.
[3] Fahd Saleh Alotaibi," Implementation of Machine Learning Model to Predict Heart Failure Disease", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 10, No. 6, 2019.
[4] Anjan Nikhil Repaka, Sai Deepak Ravikanti, Ramya G Franklin, "Design And Implementation Heart Disease Prediction Using Naives Bayesian", International Conference on Trends in Electronics and Information(ICOEI 2019).
[5] Theresa Princy R,J. Thomas,'Human heart Disease Prediction System using Data Mining Techniques', International Conference on Circuit Power and Computing Technologies,Bangalore,2016.
[6] Nagaraj M Lutimath, Chethan C,Basavaraj S Pol.,'Prediction Of Heart Disease using Machine Learning', International journal Of Recent technology and Engineering,8,(2S10), pp 474-477, 2019.
[7] UCI, —Heart Disease Data Set.[Online]. Available(Accessed on May 12020):https://www.kaggle.com/ronitf/heart-disease-uci.
[8] Sayali Ambekar, Rashmi Phalnikar,"Disease RiskPrediction by Using Convolutional Neural Network",2018 Fourth International Conference on Computing Communication Control and Automation.
[9] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres,—Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field,‖ in MachineLearning Paradigms, 2019, pp. 71–99.
[10] Jafar Alzubi, Anand Nayyar, Akshi Kumar. "Machine Learning from Theory to Algorithms: An Overview", Journal of Physics: Conference Series, 2018.
[11] Fajr Ibrahem Alarsan., and Mamoon Younes 'Analysis and classification of heart diseases using heartbeat features and machine learning algorithms',Journal Of Big Data,2019;6:81.
[12] Internet source [Online].Available (Accessed on May 1 2020):http://acadpubl.eu/ap
[13] Jee S H, Jang Y, Oh D J, Oh B H, Lee S H, Park S W & Yun Y D (2019). A coronary heart disease prediction model: the Korean Heart Study. BMJ open, 4(5), e005025.
[14] Panna A, Magnusson P K, Pedersen N L, de Faire U, Reilly M, Ärnlöv J & Ingelsson E (2017). Multilocus genetic risk scores for coronary heart disease prediction. Arteriosclerosis, thrombosis, and vascular biology, 33(9), 2267-72.
[15] Brown N, Young T, Gray D, Skene A M & Hampton J R (2019). Inpatient deathsfrom acute myocardial infarction, 2019-20: analysis of data in the Nottingham heart attack register. BMJ, 315(7101), 159-64.
[16] Folsom A R, Prineas R J, Kaye S A & Soler J T (2018). Body fat distribution and self-reported prevalence of hypertension, heart attack, and other heart disease in older women. International journal of epidemiologyy, 18(2), 361-7.
[17] Shinde R, Arjun S, Patil P & Waghmare J (2015). An intelligent heart disease prediction system using k-means clustering Soni J, Ansari U, Sharma D & Soni S (2011). Predictive data mining for medical diagnosis: an overview of heart disease prediction. International Journal of Computer Applications, 17(8), 43-8
[18] Babu, S., Vivek, E., Famina, K., Fida, K., Aswathi, P., Shanid, M., Hena, M.: Heart disease diagnosis using data mining technique. In: 2017 International conference Using Machine Learning for Heart Disease Prediction 13 of Electronics, Communication and Aerospace Technology (ICECA). vol. 1, pp. 750–753. IEEE (2017)
[19] Cai, J., Luo, J., Wang, S., Yang, S.: Feature selection in machine learning: A new perspective. Neurocomputing 300, 70–79 (2018)
[20] Dangare, C.S., Apte, S.S.: Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications 47(10), 44–48 (2012)
[21] Fang, X., Hodge, B.M., Du, E., Zhang, N., Li, F.: Modelling wind power spatial-temporal correlation in multi-interval optimal power flow: A sparse correlation matrix approach. Applied energy 230, 531–539 (2018)
[22] Gavhane, A., Kokkula, G., Pandya, I., Devadkar, K.: Prediction of heart disease using machine learning. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA). pp. 1275–1278. IEEE (2018)
[23] Hasan, S., Mamun, M., Uddin, M., Hossain, M.: Comparative analysis of classification approaches for heart disease prediction. In: 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2). pp. 1–4. IEEE (2018)
[24] Jenzi, I., Priyanka, P., Alli, P.: A reliable classifier model using data mining approach for heart disease prediction. International Journal of Advanced Research in Computer Science and Software Engineering 3(3) (2013)

[25] Kalaiselvi, C.: Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom). pp. 3099– 3103. IEEE (2016).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY