**PREDICTION OF CARDIOVASCULAR DISEASES**

**USING MACHINE LEARNING**

Dr K P Manikandan, manikandankp@mits.ac.in

K. Harshitha, 19691A0546@mits.ac.in

B. Anusha, 19691A0509@mits.ac.in

S. Girija, 19691A0540@mits.ac.in

M. Keerthana Royal, 19691A0562@mits.ac.in

Department of Computer Science and Engineering

Madanapalle Institute of Technology and Science,

Angallu, 517325

**ABSTRACT**

Recently, the prediction of cardiovascular disease has been growing expeditiously all over the world. Diagnosing patients with CVD could be a challenging task. With the avail of digital health information, identifying CVD could be made more facile by establishing efficient and precise early stage CVD presage algorithms. It may be more benign for the medical business to utilize machine learning and data mining techniques. Several data mining algorithms, including decision trees, desultory forests, machine learning, etc., were utilized in this research. They were compared predicated on their performance metrics like precision, precision, specificity and sensitivity. This study includes a dataset cognate to the detection of CVD cases with 12 attributes like smoke, active, cholesterol and so on. Feature paramount scores were estimated for each attribute to find the more impacted attributes on CVD detection. Primarily, the data was preprocessed and later algorithms were implemented and compared. An optimal algorithm was culled predicated on the obtained results to be utilized for CVD presage.

**KEYWORDS**

Cardiovascular diseases, Decision Tree, Random Forest, Data Preprocessing, Performance Metrics, cross Validation, Split Validation

**LITERATURE SURVEY**

M. Marimuthu and colleagues [7] proposed One of the most frequent disorders that might shorten people's life nowadays is heart disease. Heart disease claims the lives of 17.5 million people each year. Because the heart is such an important part of our bodies, its proper operation is critical to our survival. Heart disease is a disorder that inhibits heart function. A person's risk for coronary heart disease is used in several aspects of health promotion and clinical practise. Using multivariate regression analysis, a risk prediction model may be created from longitudinal research.

Datamining, as defined by Anooj and colleagues [1], integrates statistical analysis, machine learning, and database technology to reveal hidden patterns and connections in large databases. Machine learning techniques have been utilised in a variety of medical service industries, including the prediction of surgical treatment success, medical testing, medicine, and the discovery of connections between clinical and diagnostic data. Modern medical diagnosis is a very sophisticated operation that needs precise patient data as well as conceptual understanding of medical literature.

Dey and colleagues [4], According to a recent study conducted by the Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), heart disorders account for around 25% of deaths among those aged 25 to 69. It is the top cause of death in the globe. Some academics identify heart illness using statistical and data mining tools. There are several complicated data mining approaches and algorithms that are utilized for prediction in various fields. Data mining is a critical stage in the discovery of knowledge. To uncover hidden patterns, it integrates statistical analysis, machine learning, and database technology.

Liu X and colleagues [6], The diagnosis of cardiac disease is complicated by the presence of several variables. To assist physicians in making speedy judgements and minimizing diagnostic mistakes, categorization systems allow clinicians to immediately evaluate medical data in great depth. These systems are built by creating a model that can categorize existing records based on sample data. Several classification algorithms have been created and used as classifiers to help clinicians diagnose patients with heart disease.

Paul A K and colleagues [10], Expert expertise is used to make most medical diagnostic decisions. In many circumstances, not all of the expert's experiences contribute to an accurate illness diagnosis. Researchers have used a variety of methodologies, such as attribute reduction, rule extraction, fuzzy model optimization, and so on. Yet, noisy data in datasets, irrelevant qualities, and a lack of good fuzzy rules are significant impediments to providing the optimal judgement. In this paper, we offer a fuzzy decision support system based on genetic algorithms for forecasting the risk level of heart disease. The following is how our suggested fuzzy decision support system (FDSS) works: I Preprocess the dataset, ii) Choose effective characteristics using various approaches, iii) Generate weighted fuzzy rules based on selected attributes using GA, iv) Create the FDSS using the created fuzzy knowledge base, and v) Predict heart disease.

Coronary artery disease (CAD) is caused by atherosclerosis in the coronary arteries and leads in cardiac arrest and heart attack, according to Verma L et al [19]. Angiography, an expensive, time-consuming, and highly technical invasive procedure, is used to diagnose CAD. As a result, researchers are being pushed to develop alternative methodologies, such as machine learning algorithms that might employ noninvasive clinical data for illness diagnosis and severity assessment. In this paper, we describe a novel hybrid technique for CAD diagnosis that combines risk factor identification with correlation-based feature subset (CFS) selection, particle swam optimization (PSO), and K-means clustering methods. Multi-layer perceptron learning algorithms are examples of supervised learning algorithms (MLP) , multinomial logistic regression MLR and C4.5 are then utilized to model CAD scenarios.

Wiharto et al. [20], The use of trending classification algorithms and feature selection has been widely used in the development of coronary heart disease detection systems. A completed diagnostic system may be classified into two types: those that used the feature selection procedure and those that did not. Feature selection is a preprocessing step that determines which features impact and which do not in any activity modelling or data analysis. Ranking selection and subset selection are the two types of feature selection methods. Rating mismatched precisely supplied the rank on every current feature and overriding functionality that does not satisfy specified requirements. The process of selection used while looking for a subset is called subset selection. Subset selection is a way of selecting a set of features that are thought to be the best features.

**EXISTING SYSTEM**

* Using machine learning techniques, Methaila et al predicted cardiac disease. KNN algorithms, decision trees such as CART, C4.5, CHAID, J48, ID3 algorithms, and naive bayes approaches are commonly employed for prediction. This system takes 13 medical characteristics as input, processes them using machine learning algorithms, and displays the most correct one.
* Rairikar et al. employed three major machine learning approaches in their research: decision trees, neural networks, and the Nave Bayes classifier. These three strategies are used to perform the core task of data prediction.

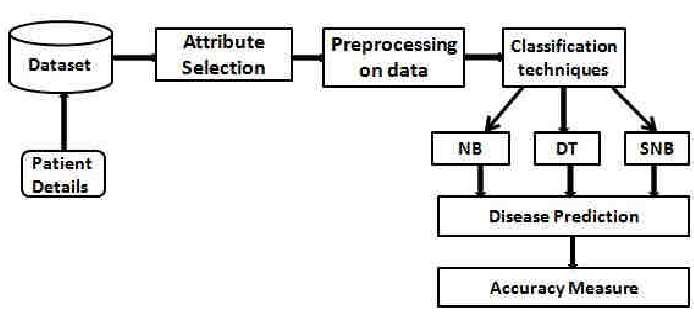
**INTRODUCTION**

Cardiovascular disease is the leading cause of mortality in the twenty-first century (CVD). This study will attempt to predict a person's risk of cardiovascular disease by analysing various factors using various data mining algorithms such as optimised decision tree, deep learning, random forest, decision tree, and so on, based on several performance metrics such as sensitivity, accuracy, specificity, and precision.

Early detection of CVD reduces treatment costs and reduces mortality rates. Cardiovascular disease is caused by a variety of factors, including excessive blood pressure, low glucose levels, and high cholesterol.

The majority of CVDs may be avoided by identifying behavioural risk factors such as smoking, drinking, and drug use. Overweight and obesity are symptoms of cardiovascular disease patients. These findings aid in the development of a more accurate method of forecasting cardiac disease, therefore assisting every patient.

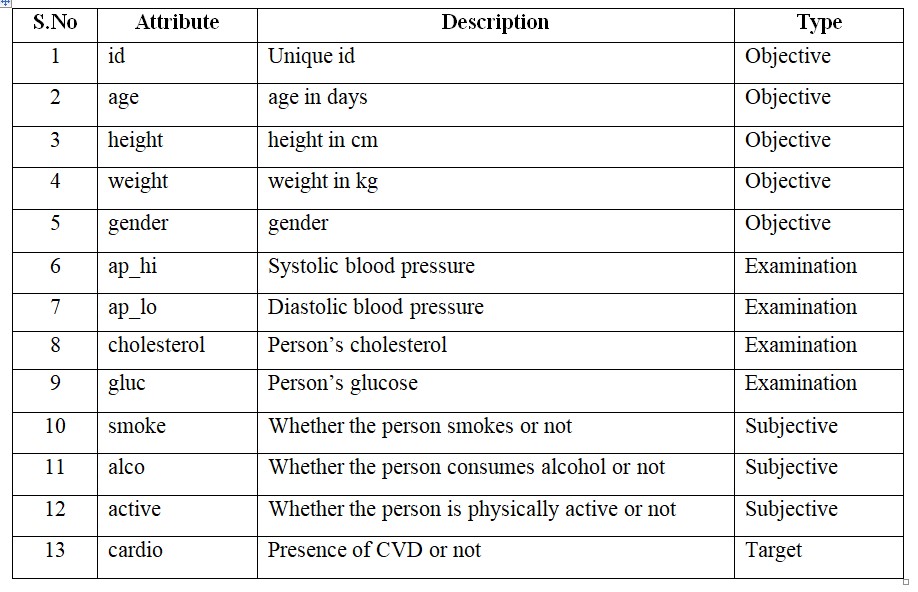
**PROPOSED MODEL**



A data set was gathered via kaggle for this investigation. The data collection has 12 characteristics and 70000 records.

As a result, we tested with three techniques: neural networks, decision trees, and naive bayes.

**DETAILS OF FEATURES**



⮚ Objective data : true information

⮚ Examination data : results from medical exams

⮚ Subjective data : information collected from patient

⮚ Gender values : 1 = women, 2 = men;

⮚ Cholesterol values : 1 = normal, 2 = beyond normal, 3 = exceptionally high

⮚ Smoke values : 0 = no, 1 = yes;

⮚ Alco values : 0 = no, 1 = yes;

⮚Active values : 0 = no, 1 = yes;

⮚ CVD values : 0 = no, 1 = yes.

**MERITS AND DEMERITS**

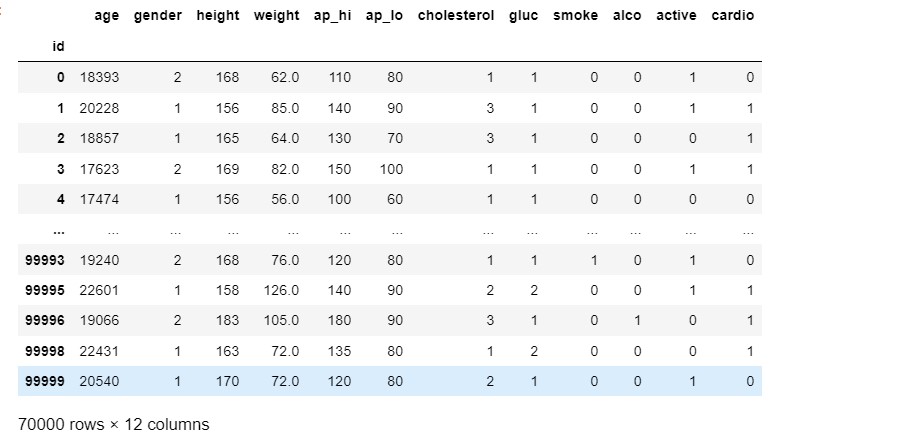
**Merits:**

* For improved analysis and understanding, eight supervised ML algorithms were tested, and feature importance ratings were created, allowing us to better care for our health
* Prediction of cardiac illness might be important for early detection, increasing mortality rate;
* demonstrates that basic supervised algorithms can have a big influence in real life.

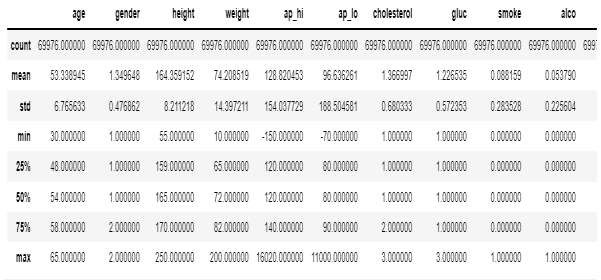
**Demerits:**

* The ensemble technique can be used to improve accuracy;
* larger datasets produced moderate results.

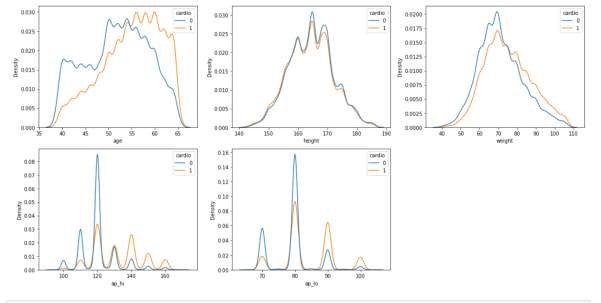
**Dataset**



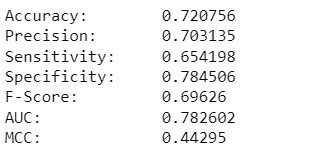
**Statistical Measure**



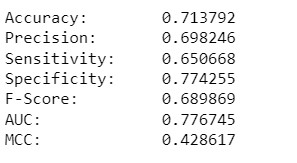
## **KDE Plot**



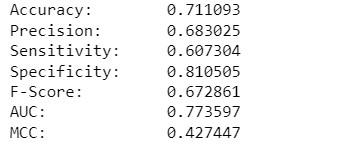
### **Result of Logistic Regression**



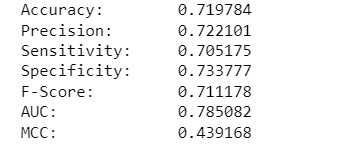
### **FResults for KNN**



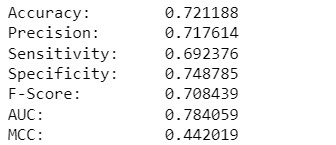
### **Results for Gaussian NB**



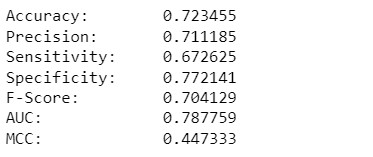
### **Results for DecisionTree**

****

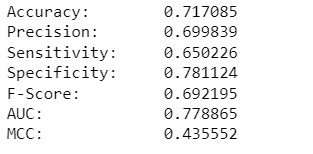
### **Results for Optimised Decision Tree**

****

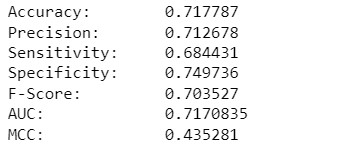
### **Result for Gradient Boosting**



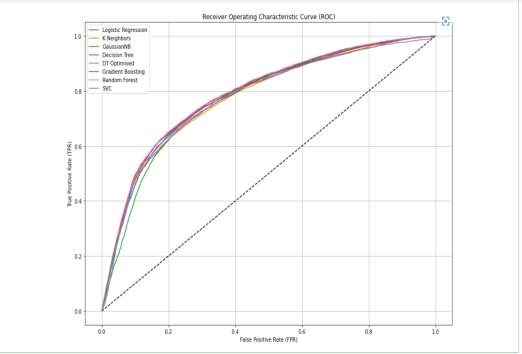
### **Result for Randomforest**



### **Result for DeepLearning**



**ROC CURVE**



# CONCLUSION AND FUTURE WORKS

When all of these aspects are known through various data mining methodologies, patients suffering from heart failure can receive better care, and mortality and morbidity can be reduced. A large number of parameters were utilised to compare various supervised machine learning approaches. For each feature, feature significance scores were computed. The best induced approach used the best DT technique and the split validation method, which were both compatible with the first attribute situation. To boost performance, two or more classifiers might be combined. Dataset training can be used to increase accuracy. To improve performance, stronger algorithms, such as neural networks, might be used.

**REFERENCES**

Saba Bashir, Z. S. (2019). Improving Heart Disease Prediction Using Feature Selection Approach. 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) (pp. 619 - 623). Islamabad: IEEE.

Senthilkumar Mohan, C. T. (2019). Effective Heart Disease Prediction Using Hybrid

Machine Learning Techniques. Special Section on Smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet of Things. IEEE Access.

Aditi Gavhane, G. K. (2018). Prediction of heart disease using machine learning. 2nd International Conference on Electronics, Communication, and Aerospace Technology (pp. 1275 - 1278). IEEE

Peter, T. J., & Somasundaram, K. (2012, March). An empirical study on prediction of heart disease using classification data mining techniques. In IEEE-International conference on advances in engineering, science and management (ICAESM-2012) (pp. 514-518). IEEE.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, 81542-81554.

Srinivas, K., Rao, G. R., & Govardhan, A. (2010, August). Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques. In 2010 5th International Conference on Computer Science & Education (pp. 1344-1349). IEEE.

Shouman, M., Turner, T., & Stocker, R. (2012, March). Using data mining techniques in heart disease diagnosis and treatment. In 2012 Japan-Egypt Conference on Electronics, Communications and Computers (pp. 173-177). IEEE.

Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018, March). Prediction of cardiovascular disease using machine learning algorithms. In 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT) (pp. 1-7). IEEE.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021).

Improving the prediction of heart failure patients’ survival using SMOTE and effective data mining techniques. IEEE access, 9, 39707-39716.

Sivagowry, S., Durairaj, M., & Persia, A. (2013, February). An empirical study on applying data mining techniques for the analysis and prediction of heart disease. In 2013 international conference on information communication and embedded systems (ICICES) (pp. 265-270). IEEE.

Gandhi, M., & Singh, S. N. (2015, February). Predictions in heart disease using techniques of data mining. In 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE) (pp. 520-525). IEEE.