Predicting the survival of patients with Heart Failure-Machine Learning Methods

T.Hemalatha[1], G.Mounika[2], K.Prasanthi[3], G.Hemavani[4]
Student, Computer Science and Engineering

Narasaraopeta Engineering College, Narsaraopeta

hemathatiparthi01@gmail.com, mounikaguntupalli99@gmail.com,
kollaprasanthi8@gmail.com, gorantlahemavavi77@gmail.com

ABSTRACT

Heart failures are the number 1 cause of death globally, taking an estimated 17.9 millions of lives each year. CVDs are a group of disorders of the heart and blood vessels and include coronary heart disease, cerebrovascular disease, rheumatic heart disease and other conditions. In four out of 5 CVD deaths are mostly occurred due to heart attacks and strokes, and one third of these deaths are occurring prematurely in people who were under 70 years of age. Mostly these cardiovascular diseases can be prevented by addressing the behavioral risk factors such as tobacco use like cigretes, unhealthy diet and also obesity, physical inactivity and harmful use of alcohol using population-wide types of strategies. Individuals at risk of CVD may demonstrate by raised blood pressure, glucose, and lipids as well as overweight and obesity. These can be easily measured in the primary care facilities. Identifying those at highest risk of CVDs and ensuring that they receive appropriate treatment so it can prevent premature deaths. And there is access to essential non communicable disease medicines and basic health technologies in all primary health care facilities is also very essential to ensure that those in need receive treatment and also counseling.

Keywords:

Heart failure, Serum creatinine, Ejection fraction

1. INTRODUCTION:

Heart failure is a term covering any disorder of the heart. Heart failure have become a major concern to deal with as studies show that the number of deaths due to heart failure have increased significantly over the past few decades in India, in fact it has become the leading cause of death in India.

A study shows that from 1990 to 4016 the death rate due to heart failures have increased around 34 per cent from 155.7 to 409.1 deaths per one lakh population in India.

Thus preventing Heart failures has become more than necessary. Good data-driven systems for predicting heart failures can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the Heart failures, and the predictions made are quite accurate.

Classification algorithms are one of the most important category of supervised machine learning algorithms. These algorithms require a very large training set. These training data sets are consisting of different features and also different attributes which describe the individual sample. Since we are doing different supervised learning algorithm. All the training set are labelled correctly. The classification algorithms such as decision trees, Gaussian Naïve Bayes, Random Forest, K- nearest neighbours, K-Means and support vector machines (SVM), develop model with these data with many different parameters. When we have a new unlabeled sample, then we can use the model to predict the label of new sample. These techniques are used for predicting whether a person is died or not.

2. LITERATURE SURVEY

CVD are disorders of the heart and blood vessels including, coronary heart disease (heart attacks), cerebrovascular diseases (strokes), heart failure (HF), and other types of pathology [1]. Altogether, Heart diseases cause the death of approximately 17 million people worldwide annually, with the fatalities figures on the rise for the first time in the 50 years in the United Kingdom [4]. In particular, heart failure occurs when the heart is unable to pump enough blood to the body, and it is usually caused by diabetes, high blood pressure, or other heart conditions or diseases [3]. The clinical community groups the heart failure into two different types based on the ejection fraction value, that is the proportion of blood which is pumped out of the heart during a single contraction, given in the form of percentage with certain values ranging between 50% and 75%. The former is heart failure due to reduced ejection fraction, previously known as heart failure due to left ventricular systolic dysfunction or systolic heart failure and characterized by an ejection fraction smaller than 40% [4]. The latter is heart failure with preserved ejection fraction, formerly called diastolic heart failure or heart failure with normal ejection fraction. In this case, left ventricle contracts normally during systole, but the ventricle is stiff and fails to relax normally during diastole, thus impairing filling [5–10]. For the quantitative evaluation of the disease progression, clinicians rely on the New York Heart Association functional classification, including four classes ranging from no symptoms from ordinary activities (Class I) to a stage where any physical activity brings on discomfort and symptoms occur at rest (Class IV)[12-19]. Despite it is in widespread use, there is no any other consistent method of assessing the NYHA score, and this classification fails to reliably predict the basic features, such as walking distance or exercise tolerance on formal testing [11].

3.PROPOSED SYSTEM:

3.1 Experimental setup

This experiment was conducted on Intel® Core™ i3 Processors with 64 bit Windows10. Anaconda 5.1.0 Python distribution is used in this experiment. The dataset used in this project is heart_failure_clinical_records dataset which is obtained from Kaggle Machine Learning repository. The dataset contains 13 attributes which are used to predict the failure. **Fig. 1.**shows the dataset **Fig. 2.** shows the process of the proposed system of heart failure prediction **fig.3** shows attribute discription.

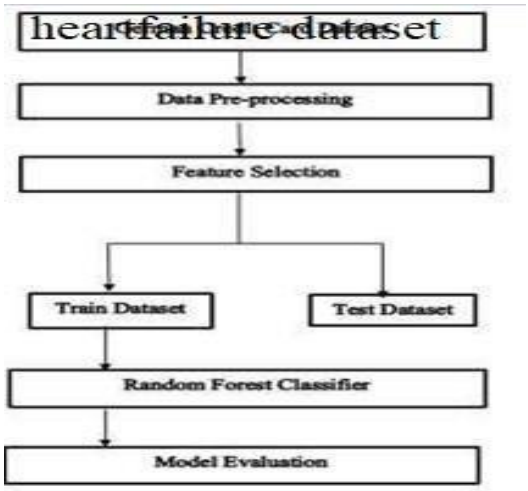


Fig 2. Proposed system of heart failure prediction

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	anaemia	creatinine	diabetes	ejection_f	high_bloo	platelets	serum_cre	serum_so	sex	smoking	time	DEATH_EVENT	
2	75	0	582	0	20	1	265000	1.9	130	1	0	4	1	
3	55	0	7861	0	38	0	263358	1.1	136	1	0	6	1	
4	65	0	146	0	20	0	162000	1.3	129	1	1	7	1	
5	50	1	111	0	20	0	210000	1.9		1	0	7	1	
6	65	1	160	1	20	0	327000	2.7	116	0	0	8	1	
7	90	1	47	0	40	1	204000	2.1	132	1	1	8	1	
8	75	1	246	0	15	0	127000	1.2	137	1	0	10	1	
9	60	1	315	1	60	0	454000	1.1	131	1	1	10	1	
10	65	0	157		65	0	263358	1.5	138	0	0	10	1	
11	80	1	123	0	35	1	388000	9.4	133	1	1	10	1	
12	75	1	81	0	38	1	368000	4	131	1	1	10	1	
13	62	0	231	0	25	1	253000	0.9	140	1	1	10	1	
14	45	1	981	0	30	0		1.1	137	1	0	11	1	
15	50	1	168	0	38	1	276000	1.1	137	1	0	11	1	
16	49	1	80	0	30	1	427000	1	138	0	0	12	0	
17	82	1	379	0	50	0	47000	1.3	136	1	0	13	1	
18	87	1	149	0	38	0	262000	0.9	140	1	0	14	1	
19	45	0	582	0	14	0	166000	0.8	127	1	0	14	1	

Fig1:Dataset

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40,..., 95]
Anaemia	Decrease of red blood cells or hemoglobin	Boolean	0, 1
High blood pressure	If a patient has hypertension	Boolean	0, 1
Creatinine phosphokinase (CPK)	Level of the CPK enzyme in the blood	mcg/L	[23,..., 7861]
Diabetes	If the patient has diabetes	Boolean	0, 1
Ejection fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14,..., 80]
Sex	Woman or man	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01,..., 850.00]
Serum creatinine	Level of creatinine in the blood	mg/dL	[0.50,..., 9.40]
Serum sodium	Level of sodium in the blood	mEq/L	[114,..., 148]
Smoking	If the patient smokes	Boolean	0, 1
Time	Follow-up period	Days	[4,...,285]
DEATH EVENT (TARGET)	If the patient died during the follow-up period	Boolean	0, 1

NOTE: mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per litre

Fig3 Attribute description

3.2. Outlier Identification:

Fig. 4. is the boxplot of all attributes of dataset in which we can observe the distribution of different attributes.

Fig. 5. is the boxplot of dataset after removing outliers.

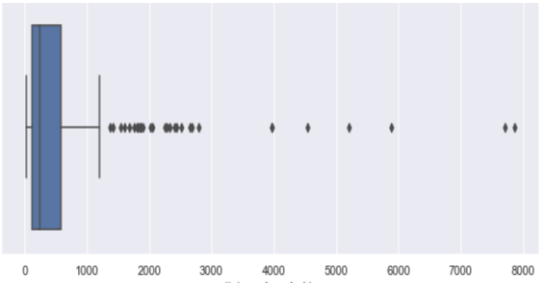


Fig. 4. Boxplot of dataset before removing outliers

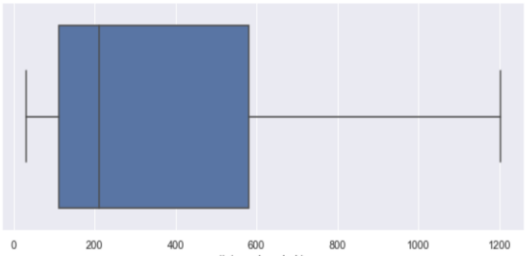


Fig. 5. Boxplot of dataset after removing outliers

3.3 Preprocessing

Preprocessing is the first step while creating the machine learning model. It is the process of converting raw dataset into cleaned dataset. Raw data contains noise, missing values, duplicate values which is not suitable for machine learning model. So, pre-processing is required for cleaning the data and making it suitable for machine learning model.

Handling Missing Values

Handling missing values is important because Machine learning algorithms do not support missing values in the data.. The fig6 below is a heatmap representing the missing values. In this graph missing values are present in Saving account, Checking Account features.

Fig. 6. shows the dataset before filling missing values, there are missing values in Savings accounts and Checking account. **Fig. 7.** shows the dataset after filling missing values..

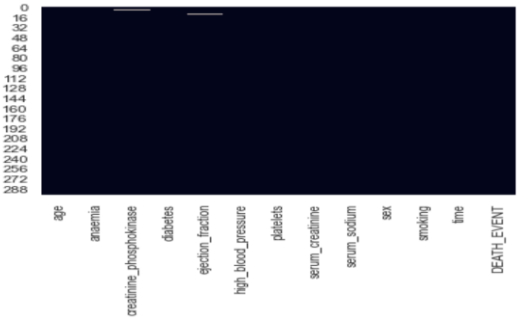


Fig. 6. Dataset before filling missing values

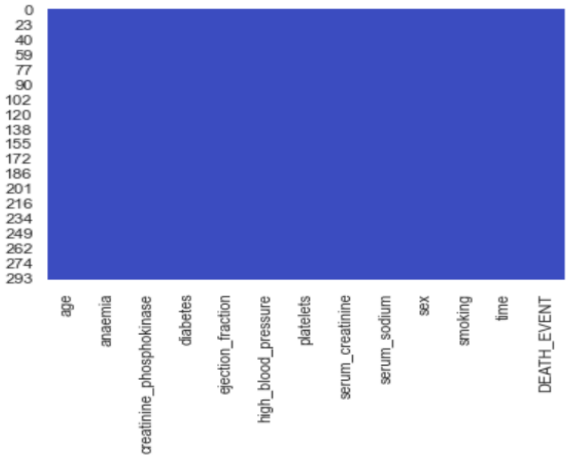


Fig. 7. Dataset after filling missing values

3.4 SMOTE

SMOTE (Synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem.

	precision	recall	f1-score	support
0	0.91	0.89	0.90	36
1	0.56	0.62	0.59	8
accuracy			0.84	44
macro avg	0.73	0.76	0.74	44
weighted avg	0.85	0.84	0.84	44

```
Before OverSampling, counts of label '1': 50
Before OverSampling, counts of label '0': 125
```

Fig. 8. Class distribution Before SMOTE technique

```
After OverSampling, the shape of train_X: (250, 12)
After OverSampling, the shape of train_y: (250,)

After OverSampling, counts of label '1': 125
After OverSampling, counts of label '0': 125
```

Fig.9. Class distribution After SMOTE technique

3.5 Feature Selection:

It is process of reducing the number of input attributes or variables when developing a model in order to reduce the computation cost of model and in some cases to increase or improve the performance or accuracy of this model.

Correlation

Correlation is the statistical measure that indicates the extent to which two or more variables that vary together. A positive correlation indicates that the extent to which those variables can be increase or decrease in parallel. A negative correlation indicates that the extent to which one variable can be increases as the other decreases. Fig 10 shows correlation for Heart Failure Dataset.

After applying correlation it is observed that no two

columns are linear related that is there is no dependency between attributes. So, there is no need to drop any column from dataset.

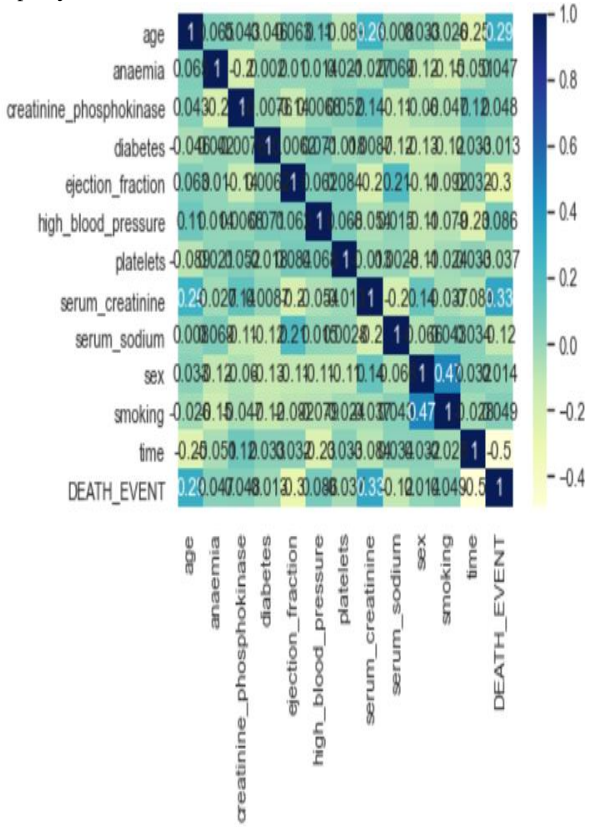


Fig.10 correlation for heart failure dataset

4.CLASSIFICATION TECHNIQUES

In classification techniques, classification is one of the data mining function that assigns items in the collection to target categories or classes. Classification is a supervised learning and can be performed on both structured and unstructured data. The goal of classification is that to accurately predict the target class for the each case in the data. Classes are called as targets or labels or categories. Classification predictive modelling is task of mapping function (f) from input variables (X) to discrete output variables
(y). There are many classification algorithms available in machine learning. Compared to previous work mentioned above, our paper presents many approaches that are used for predicting credit card fraud they are feature selection, k-fold cross validation, Chi-square test, Recursive feature elimination (RFE), Information Gain and six different machine learning algorithms. Three approaches gives comparison between 5 classification algorithms: Logistic Regression, Random Forest classifier, Gaussian Naïve Bayes, K- Nearest Neighbors, Decision Tree.

4.1Holdout cross-validation:

The **Holdout method** is the simplest kind of cross validation.

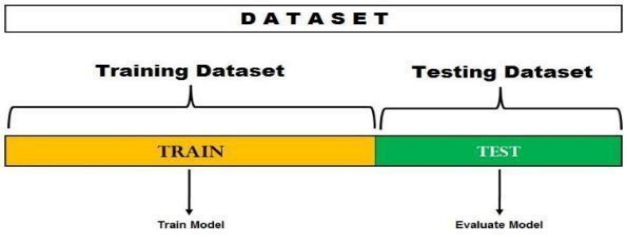
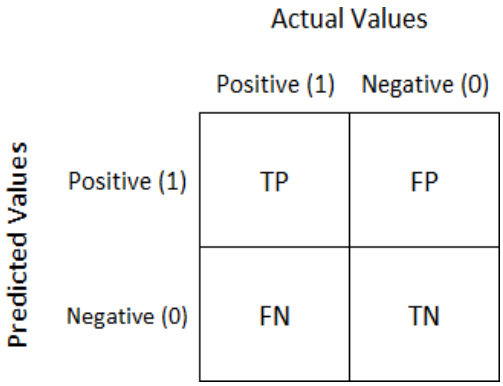


Fig.11.Holdout Cross Validation

4.2 Performance Evaluation

Performance Evaluation of classification algorithm is calculated by using confusion matrix as shown in **Fig. 15**. Confusion matrix is a performance measurement for the machine learning classification problem where the output can be two or more classes. Confusion matrix is a table with 4 different types of combinations for the predicted and the actual values. Performance is calculated by considering actual and predicted class. A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which true values are known.



4.2 Classification

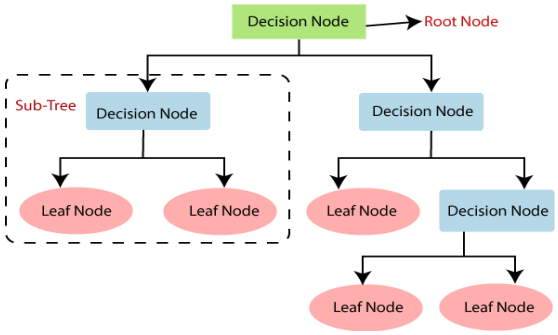
Classification is a technique which is used to classify data into some definite number of classes. The main goal of classification is to accurately predict the target class for each case in the data.

Classification Algorithms in Machine Learning:

A common job of machine learning algorithms is to acknowledge objects and having the ability to separate them into categories. This process is named classification, and it helps us segregate vast quantities of knowledge into discrete values. Frequently used data mining algorithms are Decision tree, Naïve Bayes, Logistic Regression, Random Forest Classifier, Support Vector Machine, K-Nearest Neighbours.

1. Decision tree classifier

Decision Tree is a supervised learning method used in data mining for classification and regression methods.



.

fig 13. Decission tree classifier

2.Random Forest Classifier:

Random forest is a type of supervised machine learning algorithm based on ensemble learning.

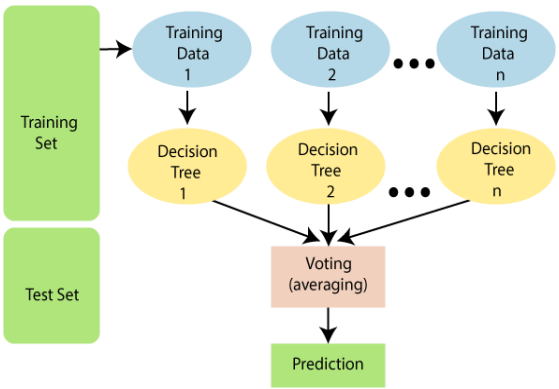


Fig. 14. Random Forest Classifier

3.Naïve Bayes:

Naïve Bayes is a classification algorithm that relies on strong assumptions of the independence of covariates in applying Bayes Theorem. Bayes theorem is P(C|X) = P(X|C) * P(C)/P(X), where X is the data tuple and C is the class such that P(X) is constant for all classes.

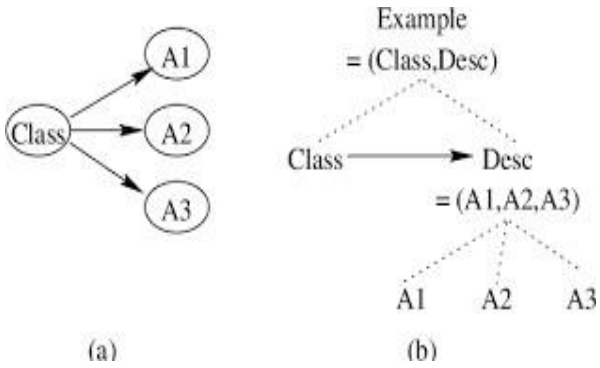


Fig .15 Navie Bayes

4 .K-Nearest Neighbours:

K-Nearest Neighbour also known as KNN is a supervised learning algorithm that can be used for regression as well as classification problems.

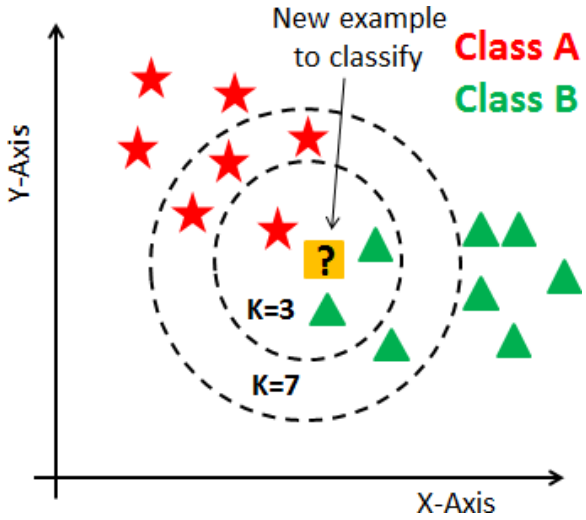


Fig 16. K-Nearest Neighbours

5.Logistic Regression: Logistic regression is the process of modeling probability of a discrete outcome given as an input variable.

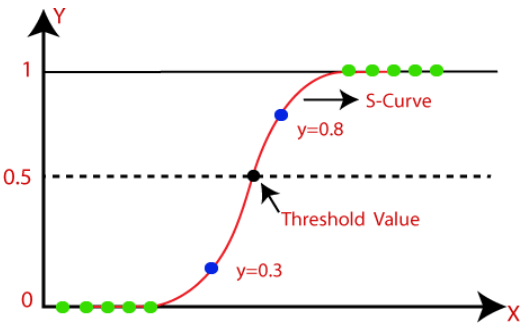


Fig 17.Logistic Regression

6.Support Vector Machine

SVM is a supervised training algorithm that can be useful for the purpose of classification and regression.

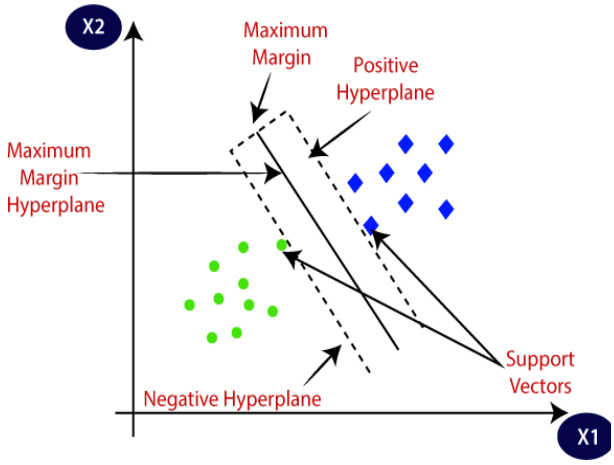


Fig 18. Support vector machine

5.RESULT ANALYSIS

Fig. 22. Shows the Comparison accuracy of five classifiers (Random Forest Classifier, Logistic Regression, Gaussian Naïve Bayes, Decision Tree, KNN) before and after applying SMOTE. From Fig. 15. it is observed that Random Forest Classifier achieves highest accuracy of 81% and it is also observed that all the classifiers achieve highest performance when compared to that after applying for the SMOTE.

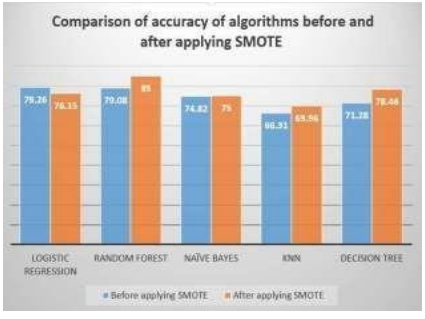


Fig 19. Comparison of accuracy before and after applying SMOTE

6.RESULT PAGE:

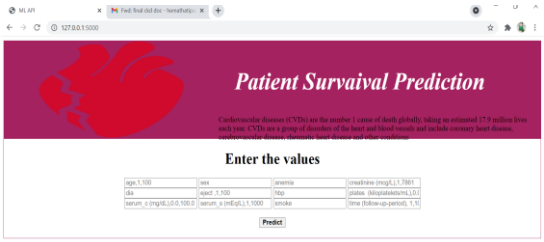


Fig.20 Result Page

Table1. Comparison of all the results

Author	Method	Accuracy
Davide Chicco, Giuseppe Jurman	Linear Regression Random Forests Decision Tree Support Vector Machine Naïve Bayes KNeighborsClassifier	71 72 74 76 78 77
Our Study	Logistic Regression Random Forest Classifier Decision tree Support vector machine Naive bayes K-Neighbors Classifier	78.33333333333333 81.66666666666667 66.66666666666666 71.66666666666667 70.0 65.0

7.REFERENCES:

[1]World Health Organization, World Heart Day. https://www.who.int/cardiovascular_diseases/world-heart-day/en/. Accessed 7 May 2019.

[2].The Guardian UK heart disease fatalities on the rise for the first timeinthe50years.https://www.theguardian.com/society/2019/may/13/heart-circulatory-disease-fatalities-on-rise-in-uk. Accessed on 25 Oct 2019.

[3].National Heart Lung and Blood Institute (NHLBI). Heart failure. https://www.nhlbi.nih.gov/health-topics/heart-failure. Accessed 20 June 2019.

[4].Meng F, Zhang Z, Hou X, Qian Z, Wang Y, Chen Y, Wang Y, Zhou Y, Chen Z, Zhang X, Yang J, Zhang J, Guo J, Li K, Chen L, Zhuang R, Jiang H, Zhou W, Tang S, Wei Y, Zou J. Machine learning for the prediction of sudden cardiac death in the heart failure,patients with low left ventricular ejection fraction: study protocol for a retro-prospective multicenter registry in China. Br Med J (BMJ) Open. 2019; 9(5):023724.

[5].Nauta JF, Jin X, Hummel YM, Voors AA. Markers of left ventricular systolic dysfunction when the left ventricular ejection fraction is normal. Eur J Heart Fail. 2018; 20:1636–8.

[6].Pfeffer MA, Braunwald E. Treatment of the heart failure with preserved ejection fraction. Reflections on its treatment with one of the aldosterone antagonist. J Am Med Assoc (JAMA) Cardiol. 2016; 1(1):7–8.
[7].Mesquita ET, Grion DC, Kubrusly MC, Silva BBFF, Santos ÉAR. Phenotype mapping of heart failure with preserved ejection fraction. Int J Cardiovasc Sci. 2018; 31(6):652–61.
[8].Nanayakkara S, Kaye DM. Targets for the heart failure with the preserved ejection fraction. Clin Pharmacol Ther. 2017; 102:228–37.
[9].Katz DH, Deo RC, Aguilar FG, Selvaraj S, Martinez EE, Beussink-Nelson L, Kim K-YA, Peng J, Irvin MR, Tiwari H, Rao DC, Arnett DK, Shah SJ. Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction. J Cardiovasc Transl Res. 2017; 10(3):275–84.
[10]. M.Sireesha, S. N. TirumalaRao, Srikanth Vemuru, Frequent Itemset Mining Algorithms: A Survey Journal of Theoretical and Applied Information Technology Vol - 96, No .3, Feb - 2018 ISSN - 1992-8645, Pages – 744 – 755.
[11].M. Sireesha, Srikanth Vemuru and S. N. Tirumala Rao, ”Coalesce based binary table: an enhanced algorithm for mining frequent patterns", International Journal of Engineering and Technology, vol. 7, no. 1.5, pp. 51-55, 2018.
[12]. M.Sireesha, S. N. TirumalaRao, Srikanth Vemuru, Optimized Feature Extraction and Hybrid Classification Model for Heart Disease and Breast Cancer Prediction International Journal of Recent Technology and Engineering Vol - 7, No 6, Mar - 2019 ISSN - 2277-3878, Pages – 1754 – 1772
[13] .M.Sireesha, Srikanth Vemuru, S.N.Tirumala Rao “Classification Model for Prediction Of Heart Disease Using Correlation Coefficient Technique” International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9, No. 2, March - April 2020, Pages- 2116 – 2123.
[14] . Sireesha Moturi , Dr. S. N. Tirumala Rao, Dr. Srikanth Vemuru,. (2020). Predictive Analysis of Imbalanced Cardiovascular Disease Using SMOTE. International Journal of Advanced Science and Technology, 29(05), 6301 - 6311.
[15]. Moturi S., Tirumala Rao S.N., Vemuru S. (2021) Risk Prediction-Based Breast Cancer Diagnosis Using Personal Health Records and Machine Learning Models. In: Bhattacharyya D., Thirupathi Rao N. (eds) Machine Intelligence and also Soft Computing. Advances in the Intelligent Systems and also in Computing, vol 1280. Springer, Singapore. https://doi.org/10.1007/978-981-15-9516-5_37
[16]. Moturi S., Srikanth Vamuru, Tirumala Rao S.N. (2021) ECG based Decision Support System for Clinical Management using Machine Learning Techniques. IOP Conference Series: Materials for Science and Engineering. Volume 1085, Annual International Conference on Emerging Research Areas on "COMPUTING & COMMUNICATION SYSTEMS FOR A FOURTH INDUSTRIAL REVOLUTION" (AICERA 2020) 14th-16th December 2020, Kanjirapally, India
[17]. Introduction to KNN, K-Nearest Neighbors : Simplified, analyticsvidhya.com
[18].Quinlan JR. C4.5: Programs for Machine Learning.; 2014:302. https://books.google.com/books?hl=fr&lr=&id=b3ujBQAAQBAJ&pgis=1. Accessed January 5, 2016.
[19]. 3.1 Cross-validation: evaluating estimator performance – scikit-learn 0.22.2
[20.] J. Han and M. Kamber,”Data Mining Concepts and Techniques”, Morgan Kauffman Publishers, 200