ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse185922020.pdf
https://doi.org/10.30534/ijatcse/2020/185922020

Classification Model for Prediction of Heart Disease using Correlation Coefficient Technique

Sireesha Moturi [1], Dr. Srikanth Vemuru [2], Dr.S.N.Tirumala Rao [3]
[1] Research Scholar, KLEF, Vaddeswaram,India, sireeshamoturi@gmail.com
Assoc. Prof., Narasaraopeta Engineering College, Narasaraopet, Inida,
[2] Professor, KLEF, vaddeswaram,India, vsrikanth@kluniversity.in
[3] Professor& HOD, Narasaraopeta Engineering College, Narasaraopet, Inida, nagatirumalarao@gmail.com

ABSTRACT

Today health care services have come up with an advanced way to treat patients having different diseases. Among all, one of the harmful diseases is the cardiovascular disease that can't be visible with a unadorned eye and comes right away when its limitations are reached. With rise in population, there is a rise in heart disease rate. Today, diagnosing patients in an effective manner have become a challenging task. The healthcare industry picks up large quantity of healthcare data but, rarely that is used to extract hidden patterns for efficient decision making purpose. Thus, we proposed to develop an approach which will help practitioners to diagnosis heart related disease. So, there is a necessity to develop a decision making system which will helps practitioners to predict heart diseases in an easier way and will offer automated predictions about the condition of the patient's heart so that further treatment can be done effectively. This proposed system will not only make accurate predictions about heart disease but also brings down cost & time. The Machine Learning algorithms have determined to be most accurate & reliable and hence used in this paper.

Key words : Machine Learning, Decision Tree, Navie Bayes, KNN, Random Forest

1. INTRODUCTION

Machine Learning(ML) is one among the applications of Artificial Intelligence (AI) which gives computers, the capacity to learn automatically from experience rather than explicitly programmed. It focuses on developing programs which can access facts and use those to learn. The major intention is to allow computers to learn themselves automatically without the need of human intervention and also alter actions accordingly.

1.1 Machine Learning Techniques

Machine learning techniques are basically categorized into Supervised Learning, Unsupervised Learning, Semi-Supervised Learning and Reinforcement Learning. **Supervised Machine Learning** system trains the machine using labelled data which is useful for future data processing. To develop predictive system Supervised Learning has been proved as a powerful on classification problems [23].**Unsupervised Machine Learning** system trains the machine using neither classified nor labelled data. **Semi-Supervised Machine Learning** system trains the machine using the combinations of labelled and unlabeled data. **Reinforcement Machine Learning** is a crucial sort of ML technique in which an agent learns interactions with the surrounding agents via performing actions and seeing results.

1.2 Importance of Machine Learning in HealthCare

Now a day hospitals are more, patients are more and the data is also generating more and more. Most of the hospitals are maintaining Hospital Information system to maintain patient data and health care. But, unfortunately this data is rarely used for decision making purpose. Sometimes the doctor's time is not enough in order to analyze this increasing availability of data or it may take more time for analysis. If we process this data with the help of Artificial Intelligence with in the fraction of seconds it will be completed and machine is also not susceptible to information overloaded and short term memory loss. Communication through web is becoming increasingly, so it makes difficult for extracting information on the web [6]. The importance of healthcare in Machine Learning is increasing significantly due to its ability to process huge datasets (structured and unstructured) efficaciously beyond the range of human capability, then consequently convert evaluation of that statistical information into clinical insights that assist physicians in making plans and presenting care, which ultimately results in better outcomes, reduces the prices of care, and increases patients satisfaction. Using these types of advanced analytics, we can provide better information to phycians at the point of patient care. So, there is a need of developing an automated

disease prediction system by using ML techniques to detect the disease early [24]. Content extraction and identity of biological datasets is gaining prominence in the today's world [1].

2. RELATED WORK

Data Mining performs a critical role in the information extraction process by identifying hidden sequences, performing regression and classification, constructing analytical models, performing clustering and representing the mined results with the help of different visualization and presentation techniques. So the information mining system should be able to find patterns at different levels of granularities to meet different user expectations or applications [19]. Information on the web now-a-days has structured and unstructured kind of records, homogeneous, heterogeneous and mixed varieties of data and current websites present a larger wide variety of difficulties and complexities than traditional ones[14]. The availability of information on the web today is more, so in order to deal this huge information, there is a need of challenging tools [11].

Machine Learning is majorly used in the medical field such as predicting Breast Cancer, diagnosis of Heart Disease etc. since it is a multidisciplinary field. For better diagnosis decision support system is more useful[22]. By using this technique researchers are developing various methods in-order to predict the heart diseases with high accuracy.

"Coalesce-based binary table: an enhanced algorithm for mining frequent patterns" was introduced by Sireesha Moturi [17]. In this method they used Coalesce function to find Frequent Patterns by representing the data in the Binary Table format without generating candidate relations.

"Adaptive Weighted Fuzzy Rule-Based System for the risk level assessment of Heart Disease" was proposed by Animesh Kumar Paul et al. [2]. In this work, a fuzzy framework is used to find the risk assessment of heart disease by utilizing modified DMSPSO and ensemble technique. They used different statistical methods to select the vital attributes which can help to generate Weighted Fuzzy Rules to diagnosis the coronary illness.

"An Integrated Decision Support System Based on ANN and Fuzzy_AHP for Heart Failure Risk Prediction" was introduced by Oluwarotimi Williams Samuela [4]. In this paper they used Fuzzy Analytic Hierarchy Process (Fuzzy_AHP) to calculate the global weight of every attribute. They performed classification based on the attribute global weight.

Ali. Adeli et al., introduced "A Fuzzy Expert System for Heart Disease Diagnosis"[5]. They used Long beach and Cleveland database ("UCI Machine Learning Repository: Heart Disease Data Set") to analyse the proposed system. They designed Fuzzy Expert System (FES) by using Mamdani inference method, fuzzification and defuzzification techniques.

"Optimized Feature Extraction and Hybrid Classification Model for Heart Disease and Breast Cancer Prediction" was introduced by Sireesha Moturi [18]. In this method they used New Levy Dragonfly technique to select the features and hybrid classification technique with the combination of Support Vector Machine and Deep Belief Network for classification.

"Prediction of Heart Disease using Neural Network" was proposed by Dangare et al. in [9]. To improve the diagnosis process they used multilayer neural network along with back propagation algorithm.

"Prediction of Heart Disease using Classification Algorithm" was introduced by Mosima Anna et al. [10]. In their model they used Decision Tree for Disease Diagnosis. Authors proved that their model is better than other classification techniques.

3. CLASSIFICATION

It is a process of categorizing data into given classes. Its primary goal is to identify the class of our new data. In this paper we have applied classification algorithms like **K-Nearest Neighbours(KNN)**, **Random Forest, Decision Tree**, **and Naive Bayes.**

3.1 K-Nearest Neighbours: KNN is one of the major supervised Machine Learning algorithms and it is suitable for regression as well as classification problems. In this algorithm it uses labelled data for training and new records are to be classified based on the similarity measures (e.g. distance function, proximity measure, error rate etc.). Classification is done based on the concept of majority vote system to its neighbours. The test data is labelled with the class which is having highest nearest neighbours. In regression problem it considers mean of K labels and in classification problem it considers mode of K labels.

3.2 Decision Tree: It is one of the famous methods for supervised learning problems. It is represented as a tree which contains set of nodes like root, internal and external nodes. Except leaf node remaining all other nodes are used to test some condition and leaf nodes are used to represent class label information. Always our classification starts from root and moving down to the leaf for classification. Finally it produces set of rules which are used to assign the class label to unlabeled samples. It is very simple to understand and represent. It can handle both numerical and categorical data.

3.3 Random Forest: It can be used for classification as well as regression problems. It generates multiple decision trees and finally combine them in-order to obtain more accurate predictions.

3.4 Naive Bayes: These are the set of probabilistic classification algorithms based on the concept of Bayes' Theorem for predictive modelling. It contains set of algorithms where all of them are based on one principle, i.e. independence assumption between every pair of features is being classified. It is a simple but more powerful algorithm

for predictive modelling. It is suitable for two class as well as multi-class classification problems.

4. EXPERIMENTAL SETUP

* Intel(R) Core™2 i7-5500U CPU @ 2.40GHz, 4MB(Megabyte) Cache memory, 8 gigabyte (GB) RAM(Random Access Memory) is used in our study.

* Process base frequency is 2.40 GHz , Bus Speed is 5 GT/s DBI2

Windows 10 Home with 64 bit Operating System and x64-based processor is used to study the performance and comparing the accuracy of Naive Bayes, Decision tree, Random forest and KNN algorithms. Algorithms were implemented in Annaconda3-5.1.0-Windows-x86_64. It is a scientific distribution of Python, which includes set of libraries and other useful tools. It does include the Spyder IDE, Jupyter notebook etc. Here we have used jupyter notebook. It allows us to perform data pre-processing, Machine Learning, Data Visualization, and much more[12].

5. ANALYSIS

The Cleveland Heart Disease dataset is used in this work which is extracted from UCI ('University of California, Irvine') repository. Totally it contains 72 attributes among those only 14 attributes are used in the diagnosis process as shown in Table-1.

Table 1: List of attributes to predict Heart Disease

1. Age: in years
2. Sex
3. CP: Chest Pain type
4. trestbps: Resting Blood Pressure
5. Chol: serum cholestoral in mg/dl
6. FBS: (Fasting Blood Sugar > 120 mg/dl)
7. RestEcg: Resting Electrocardiographic results
8. thalach: maximum heart rate achieved
9. Exang: Exercise induced Angina
10. oldpeak = ST depression induced by exercise relative to rest
11. slope: the slope of the peak exercise ST segment
12. CA: number of major vessels (0-3) colored
13. Tthal
14. num: diagnosis of heart disease (angiographic disease status)
    It is integer valued from 0 (no presence) to 4

The dataset contains 303 instances and 5 classes i.e., 0,1,2,3,4.

The dataset is converted into a CSV (Comma Separated Values) file.

5.1 Data Pre-processing

In-order to get better results by using a model in Machine Learning, data format has to be in a proper manner. One of the major step in any information extraction or mining process is pre-processing of the data [3]. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. Those null values have to be managed using raw data. So, before feeding data to an algorithm we have to apply different kinds of transformations to our data which is referred as pre-processing. By performing pre-processing the raw data which is not feasible for analysis is to be converted into clean data [13].

5.1.1Missing values

Filling missing values is one of the pre-processing techniques to fill missing values. The missing values in the dataset are represented as '?' but it a non-standard missing value and it has to be converted into a standard missing value as NaN. So that pandas can be used to detect the missing values. The Fig1 indicates a heatmap which is used to represent the missing values. In this graph missing values are present in ca, thal features. We have filled that missing values using the median of the features. After filling the missing values our heat map looks like as shown in Fig-2.

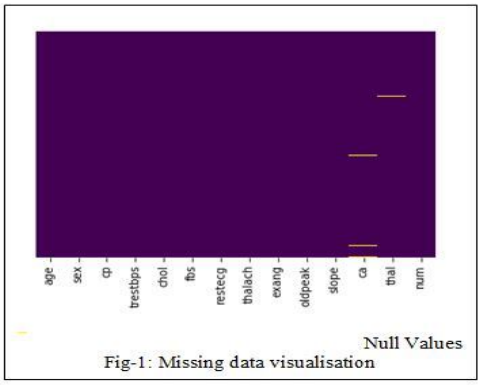

Fig-1: Missing data visualisation

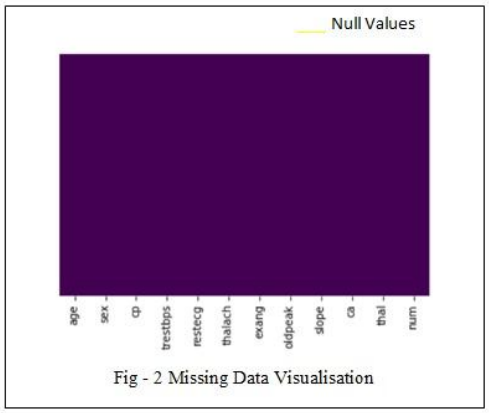

Fig - 2 Missing Data Visualisation

5.1.2Discretization

It is the process of converting continuous-values (infinite set of values) to a discrete value (finite set of values) by specifying a set of continuous intervals which falls in the range of the variable values. We have categorized the data as shown in Table-2. A classification algorithm such as Random Forests (RF) is chosen for its ability to handle high-dimensional data, which benefits from discretization in the analysis of genetic data. It can handle both continuous data and categorical data for classification. However, it uses discretization technique to handle the continuous data. A simple probabilistic classification algorithm i.e., Naive Bayes (NB) benefits from discretization of data in-order to perform well in many domains. But whenever we are using KNN we should use

dummies since our data contains both continuous and categorical attributes.

If we consider the class labels of the disease dataset Class label 0 is in 164 instances, Class label 1 is in 55 instances, Class label 2 is in 36 instances, Class label 3 is in 35 instances and Class label 4 is in 13 instances. By observing the Class Label information we identified that the dataset is suffering from class imbalence problem. So we have to balance the class labels by transforming 1,2,3,4 class labels into single Class label as 1.

Table 2: Categerization of input data

Attribute	Range	Linguistic variable	Attribute	Range	Linguistic variable
Age	16-38	Low	Thalach	50-141	Low
	33-45	Medium		111-194	Medium
	40-58	High		152-250	High
	52-80	Very High	Exang	1	yes
Sex	0	female		0	no
	1	male	Oldpeak	0.00-2.00	Low
Cp	1	typical angina		1.50-4.20	Risk
	2	atypical angina		2.55-7.00	Terrible
	3	non-anginal pain	Slope	1	upsloping
	4	asymptomatic		2	flat
Trestbps	80-134	Low		3	downsloping
	127-153	Medium	Ca	0	Vessels0
	142-172	High		1	Vessels1
	154-200	Very High		2	Vessels2
Chol	50-197	Low		3	Vessels3
	188-250	Medium	Thal	3	normal
	217-307	High		6	fixed defect
	281-700	Very High		7	reversible defect
Fbs	1	yes	Num	0	Healthy
	0	no		1-4	Sick(1-4)
Resteeg	0	Normal			
	1	ST-T abnormal			
	2	Hypertrophy			

5.1.3 Feature scaling

By performing data normalization we can standardize the range of every independent variable in data processing. This feature is very useful in data pre-processing step. The major purpose of StandardScaler is to transform each attribute in to the specific range like mean value is of 0 and standard deviation is of 1 [15].The feature scaling is performing on every attribute and the mean and standard deviation of attributes are calculated as shown in eq-1.

$$y_i - \text{mean}(y)/\text{stdev}(y) \quad \text{eq-1}$$

In eq-1 yi represents the values of attribute y.

After applying Standarscalar the data set will be looks like as shown in Table-3.

Table 3: Result of StandardScalar

```
df_feat.head()
```

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal
0	0.359948	0.686202	-2.251775	0.004197	-0.096315	2.394438	1.016684	-0.026506	-0.696631	-0.057544	2.274579	-0.711131	0.658017
1	0.359948	0.686202	0.877985	0.131375	0.004318	-0.417635	1.016684	-0.128170	1.435481	-0.057544	0.649113	2.504881	-0.895586
2	0.359948	0.686202	0.877985	-0.122980	-0.096315	-0.417635	1.016684	-0.128170	1.435481	-0.057544	0.649113	1.432877	1.175885
3	-1.032361	0.686202	-0.165268	-0.122980	-0.096315	-0.417635	-0.996749	-0.026506	-0.696631	-0.057544	2.274579	-0.711131	-0.895586
4	-0.568258	-1.457296	-1.208521	-0.122980	-0.096315	-0.417635	1.016684	-0.026506	-0.696631	-0.057544	-0.976352	-0.711131	-0.895586

5.1.4 Correlation Coefficient Method

Eq-2 is used to find the relation between two attributes p and q by using Correlation Coefficient Method.

$$r_{p,q} = \sum(p_i - \bar{p})(q_i - \bar{q})/n\sigma_p\sigma_q \quad \text{eq-2}$$
$$= \sum(p_i\, q_i) - n\bar{p}\,\bar{q}/\, n\sigma_p\sigma_q$$

In the above equation n indicated no. of samples, p_i and q_i are the corresponding values of p and q attributes in samples, $\bar{p}$ represents mean of attribute P and $\bar{q}$ indicates mean of attribute q, σp denotes standard deviation of p and σq indicates standard deviation of q. Normally, $-1 \leq r_{p,q} \leq +1$.

If rp,q < 0, then p and q are negatively correlated.

If rp,q =0, then there is no correlation between p and q i.e p and q are independent attributes

If rp,q > 0, then p and q are positively correlated.

We can drop the attributes that are having correlation coefficient value as 0 as it indicates that the variables are independent with respect to the prediction attribute.

6. RESULT ANALYSIS

In-order to choose n value for KNN algorithm we used error rate as metric. We choose the K value where the error rate is low. Figure 3 represents error rate w.r.t to K value before applying Correlation Coefficient. Figure 4 represents error rate w.r.t to k value after applying correlation coefficient [16].

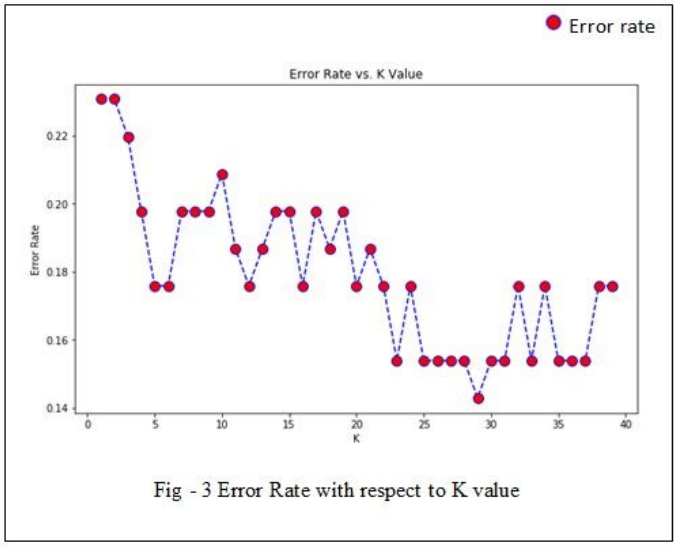

Fig - 3 Error Rate with respect to K value

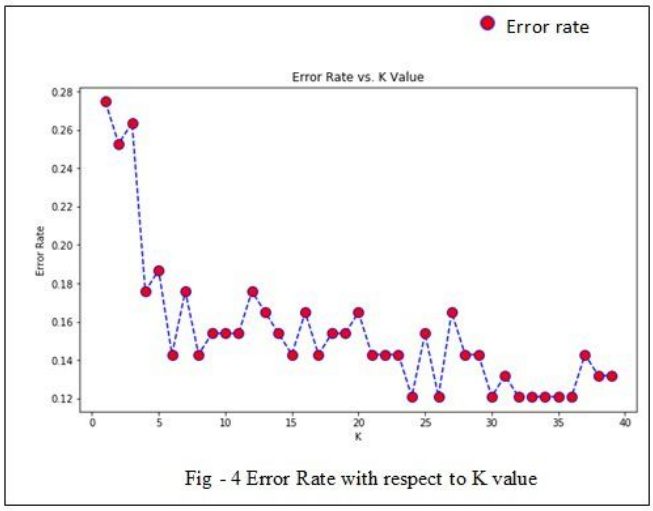

Fig - 4 Error Rate with respect to K value

Confusion Matrix*:* It is a comparison table which is used to evaluate classifier performance based on the test dataset for which actual class labels are known.

		Predicted class: P	Predicted class: N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

The Positive Class will predict the model correctly if it's result is TP. Correspondingly, The Negative Class will predict the model correctly if it's result is TN. The Positive Class will predict the model wrongly if its result is FP. Correspondingly, The Negative Class will predict the model wrongly if its result is FN. The following are the different parameters which are used to evaluate the model.

True Positive Rate (TPR)

It is the ratio of individuals who actually have the disease were identified as having the disease.

TPR= TP / (TP + FN)

True Negative rate (TNR)

It is the ratio of individuals who actually do not have the disease were identified as not having the disease.

TNR= TN / (TN + FP) =1-FPR

Positive Predictive Value (PPV)

If the test end result is positive what's the chance that the patient really has the problem.

PPV= TP / (TP + FP)

Negative Predictive Value (NPV)

If the test end result is negative what's the chance that the patient does not have disease.

NPV= TN / (TN + FN)

Miss Rate or False Negative Rate (FNR)

It is the percentage of the people who are actually having the disease will be recognised as not having the disease.

sFNR= FN / (FP + TN)

Fall-out or False Positive Rate (FPR)

It is the proportion of the people who actually who do not have the disease, but they will be identified as having the disease.

FPR=FP/ (FP + TN)

False Discovery Rate (FDR)

It shows the percentage of the people recognized as having the disease actually who do not have the disease.

FDR= FP / FP + TP

False Omission Rate (FOR)

It shows the percentage of a negative test result for which the true condition is positive.

FOR= FN / (FN + TN)

Accuracy

The accuracy reflects the total proportion of individuals which are correctly classified.

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

F1 score

F1 Score indicates harmonic mean of precision and sensitivity

F1=2TP/ (2TP+ FP + FN)

After building the model we can check the effectiveness of model with the help of Confusion Matrix. We compare the performance of Navie Bayes, Decision Tree, KNN and Random Forest by considering different parameters like Accuracy, Error Rate, sensitivity and specificity etc. These algorithms are evaluated based on the results of confusion matrix with Feature Selection and without Feature Selection. The results of Decision Tree before Correlation and after

Correlation are showed in Fig-5. The results of Random Forest before Correlation and after Correlation are showed in Fig-6. Fig-7 is used to show the results of Navie Bayes before Correlation and after Correlation. Fig-8 shows the results of KNN before Correlation and after Correlation. Fig- 9 and Fig-10 shows comparison parameters of different algorithms before and after Correlation. By observing Fig.11 we identified that KNN gives better accuracy when compared to other algorithms like Decision Tree, Random Forest and Navie Bayes. Table 1. Shows comparison of our proposed model with existing methods.

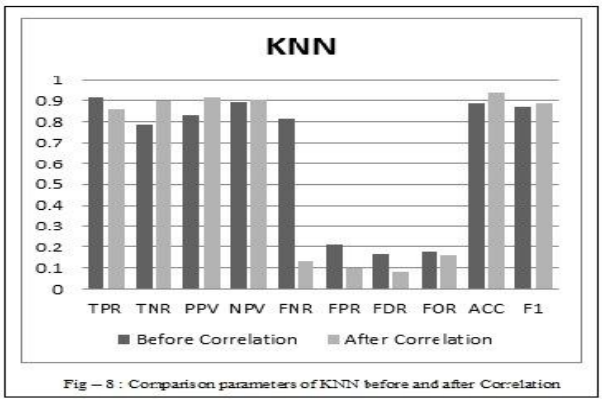

Fig – 8 : Comparison parameters of KNN before and after Correlation

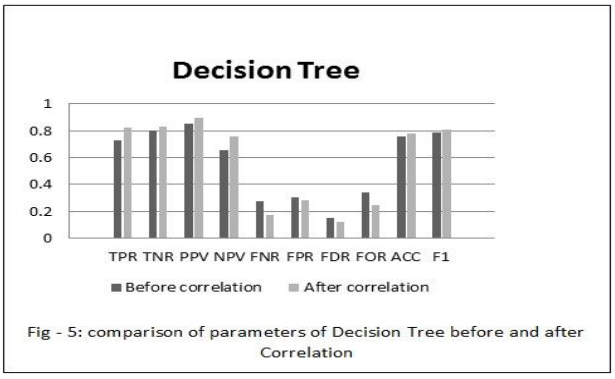

Fig - 5: comparison of parameters of Decision Tree before and after Correlation

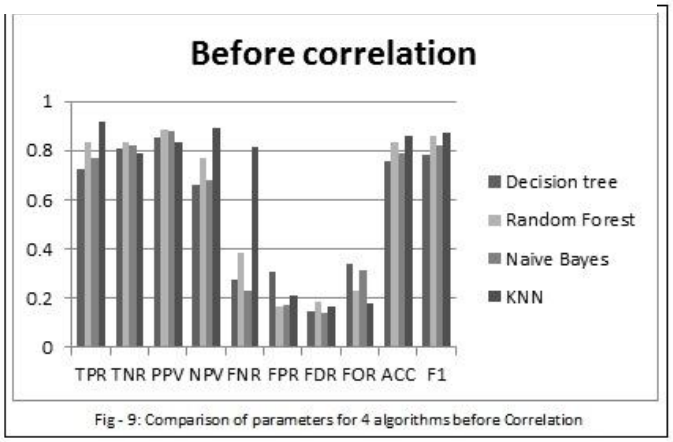

Fig - 9: Comparison of parameters for 4 algorithms before Correlation

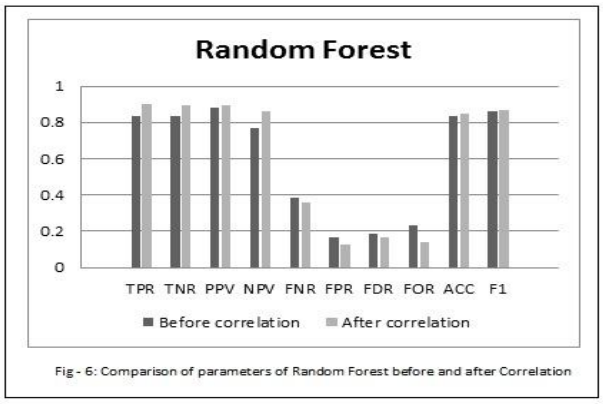

Fig - 6: Comparison of parameters of Random Forest before and after Correlation

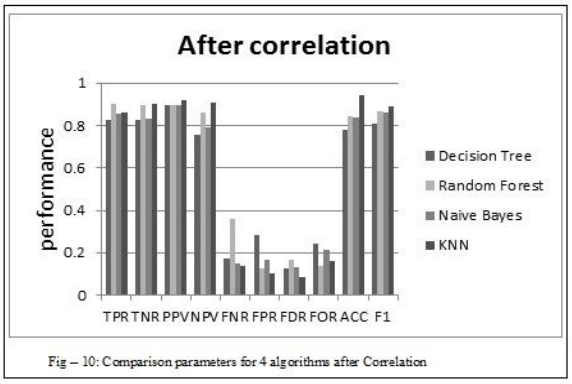

Fig – 10: Comparison parameters for 4 algorithms after Correlation

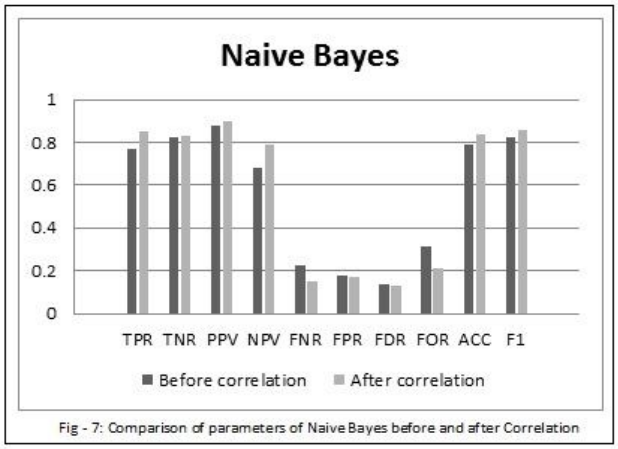

Fig - 7: Comparison of parameters of Naive Bayes before and after Correlation

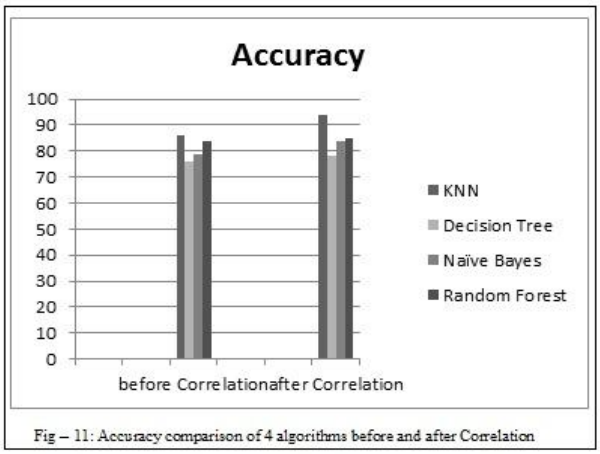

Fig – 11: Accuracy comparison of 4 algorithms before and after Correlation

Table 4: Comparison of results

Author	Method	Accuracy
Animesh Kumar et al. [2]	Adaptive Weighted Fuzzy Rule Based System	92.31
L Ali et al.[20]	Chi-Square Statistical Model with DNN	91.57
	Logistic Regression before Feature Selection	84
	Logistic Regression after Feature Selection	89
	SVM before Feature Selection	86
	SVM after Feature Selection	88
Youness Khourdifi et al.[21]	SVM with Feature Selection	83.55
	Navie Bayes with Feature selection	86.15
Resul Das et al. [8]	Neural Network with Ensemble method	89.01
Oluwarotimi Williams et al. [4]	ANN with Fuzzy AHP	91.10
M.Sireesha et al. [18]	NL-DA	83
Our study	Decision Tree without Correlation	75.8
	Decision Tree with Correlation	78
	Random Forest without Correlation	83.5
	Random Forest with Correlation	85.6
	Navie Bayes without Correlation	79.1
	Navie Bayes with Correlation	83.5
	KNN without Correlation	87.7
	KNN with Correlation	**93.4**

7. CONCLUSION

The major purpose of this work is the comparative study among different classification Algorithms based on different performance measures in-order to find whether the person is having the disease or not. We have implemented 4 algorithms, all these algorithms were applied on the Cleveland Heart Disease dataset. The accuracy varies for different classification algorithms. The highest accuracy is given when we have used KNN algorithm with Correlation factor which is nearly 94%. According to our analysis KNN gives better classification results when compared to other methods. In future, it is also possible to apply the proposed technique to predict other diseases.

REFERENCES

1. Prakash K.B., Rangaswamy M.A.D. "Content extraction of biological datasets using soft computing techniques.", Journal of Medical Imaging and Health Informatics, 2016: 932-936. https://doi.org/10.1166/jmihi.2016.1931
2. Paul, Animesh Kumar & Chandra Shill, Pintu & Rabin, Md. Rafiqul Islam & Murase, Kazuyuki., "Adaptive weighted fuzzy rule-based system for the risk level assessment of heart disease", in Applied Intelligence.2017, 48. DOI:10.1007/s10489-017-1037-6.
3. Kolla B.P., Raman A.R. "Data Engineered Content Extraction Studies for Indian Web Pages." Advances in Intelligent Systems and Computing, 2019: 505-512. https://doi.org/10.1007/978-981-10-8055-5_45
4. Oluwarotimi Williams Samuel, Grace Mojisola Asogbon , Arun Kumar Sangaiah , Fang Peng , Guanglin Li , "An Integrated Decision Support System Basedon ANN and Fuzzy_AHP for Heart Failure Risk Prediction" , in Expert Systems With Applications (2016), DOI: 10.1016/j.eswa.2016.10.020.
5. Adeli A, Neshat M, "A Fuzzy Expert System for Heart Disease Diagnosis", In: Proceedings of the international multi-conference of engineers and computer scientists, vol I, 2010, pp: 1–6.
6. Prakash K.B., Dorai Rangaswamy M.A., Raman A.R. "Text studies towards multi-lingual content mining for

web communication." Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, 2010: 28-31. https://doi.org/10.1109/TISC.2010.5714601
7. Cheung N, " Machine learning techniques for medical analysis.School of Information Technology and Electrical Engineering", BSc Thesis, University of Queenland
8. Das R, Turkoglu I, Sengur A, "Effective diagnosis of heart disease through neural networks ensembles", in Expert Systems with Applcations, 2009, 36(4), pp:7675–7680, DOI:10.1016/j.eswa.2008.09.013
9. Dangare A, "Data mining approach for prediction of heart disease using neural network", in IJCET 2012; 3: 30-40.
10. Hlaudi Daniel Masethe, Mosima Anna, "Prediction of heart disease using classification algorithm", in Wcess 2014.
11. Prakash K.B., Dorai Rangaswamy M.A. "Content extraction studies using neural network and attribute generation." Indian Journal of Science and Technology, 2016: 1-10
12. https://en.wikipedia.org/wiki/Anaconda_(Python_distribution)
13. https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/
14. K.B., Prakash. "Information extraction in current Indian web documents." International Journal of Engineering and Technology (UAE), 2018: 68-71. https://doi.org/10.14419/ijet.v7i2.8.10332
15. https://en.wikipedia.org/wiki/Feature_scaling#Standardization
16. https://www.udemy.com/python-for-data-science-and-machine-learning-bootcamp
17. M. Sireesha, Srikanth Vemuru and S. N. TirumalaRao, " Coalesce based binary table: an enhanced algorithm for mining frequent patterns", International Journal of Engineering & Technology, vol. 7, no. 1.5, pp. 51-55, 2018
18. 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
19. 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
20. Amin UI Haq et al. "A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms" Mobile Information Systems, Volume 2018, Article ID 3860146, 21 pages https://doi.org/10.1155/2018/3860146
21. Youness Khourdifi et al. "Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization", International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019. https://doi.org/10.22266/ijies2019.0228.24
22. Sri Hari Nallamala et al. "Qualitative Metrics on Breast Cancer Diagnosis with Neuro Fuzzy Inference Systems", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No.2, March - April 2019. https://doi.org/10.30534/ijatcse/2019/26822019
23. Oguntimilehin A et al. "A Clinical Diagnostic Model Based on Supervised Learning", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No.3, May - June 2019. https://doi.org/10.30534/ijatcse/2019/94832019
24. Jide E et al, "Breast Cancer Predictive Analytics Using Supervised Machine Learning Techniques", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 8, No.6, November - December 2019. https://doi.org/10.30534/ijatcse/2019/70862019