International Conference on Intelligent Technologies & Science - 2021
(ICITS-2021)

Prediction of Cardiac Arrest and Recommending Lifestyle Changes to Prevent It using Machine Learning

Vicky Singh[1] , Brijesh Pandey[2]
Computer Science and Engineering,
Goel Institute of Technology & Management, Lucknow, India
vicky.singh10a@gmail.com

Abstract: The keys issues with the previous research works were improper pre-processing of the data due to which the data was skewed and contained larger numbers. Hence the model would be biased due to the skewness of the data and it would take more time to process the data due to not scaling the data and it makes the model more complex to learn the linear and non-linear relation between the features. The data were directly used to train the models without even hyper tuning the models, without hyper tuning the model would be build the models on the default parameters hence makes it more difficult to respond according to the data used for training purpose. Feature extraction is most important portion of any learning process. In this research work, we have applied machine learning algorithms to the heart disease dataset and proposed a recommendation system based on basic information like age, sex, cholesterol level, etc. The system will also recommend some lifestyle changes in terms of exercises and food to cure the patient.

Keywords-Heart Attack, Machine Learning, KNN, Random Forest, Artificial Neural Networks.

1. Introduction

India has recorded a 4 fold rise in mortality and morbidity due to non-communicable disease and incredible down fall in the communicable diseases past 40 years. In the non-communicable diseases mortality and morbidity has raised 4 times linked to heart disease. The Lancet and its associated journals have revealed the detailed estimates of cardiovascular diseases shows that their prevalence has gone up in every Indian state between 1990 and 2016, but there is vast variation among states. The prevalence of heart disease has increased by over 50% from 1990 to 2016 in India, with an increase observed in every state. Heart disease and stroke together contributed to 28.1% of total deaths in India in 2016-compared with 15.2% in 1990.Heart disease contributed 17.8% of total deaths and stroke contributed 7.1% of deaths and disability from heart disease was significantly higher in men than in women, but was similar among men and women for stroke.

Deaths due to cardiovascular diseases rose from 13 lakh in 1990 to 28 lakh in 2016. The number of prevalent cases of cardiovascular diseases has increased from 2.57 crore in 1990 to 5.45 crore in 2016. The prevalence was the highest in Kerala, Punjab and Tamil Nadu, followed by Andhra Pradesh, Himachal Pradesh, Maharashtra, Goa and West Bengal. More than half of the total cardiovascular disease deaths in India in 2016 were in people younger than 70 years. This proportion was the highest in less developed states, which is a major cause for concern with respect to the challenges posed to the health systems.

Reducing premature deaths from cardiovascular diseases in the economically productive age groups requires urgent action across all states of India, "the researchers have observed.

The study shows that the response has to be appropriate to the content of each state. By shining the torchlight on specific disease burdens that each state must prioritize, this study will help direct health system resource to maximize impact through early prevention and effective treatment. Explained Professor K Srinath Reddy, President, Public Health Foundation of India, and one of the joint senior authors of the series. Other joint senior authors are Dr Sourya Awaminathan and Dr Lalit Dandora.

2. Related Work:

Many research have been done for the classification of Coronary heart disease by using the machine learning techniques. For training the machine learning model the data set is available on the uci machine learning repository. Mostly used data set is the Cleveland dataset. The Cleveland data set contains 303 records with 14 attributes namely age, sex, chest pain type (4 values), resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar > 120 mg/dl, resting electrocardiographic results (values 0,1,2), maximum heart rate achieved, exercise induced angina, old peak = ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels (0-3) coloured by fluoroscopy , thal: 3 = normal; 6 = fixed defect; 7 = reversible defect, num (the predicted attribute). Originally the data set contains 72 attributes out of which these 14 attributes are commonly used by the researchers.

Sharma Purushottam et al, [2] proposed a c45 rule (decision tree) and partial tree methodology for heart disease prediction .the dataset used for this paper is the Cleveland Clinic dataset which is available on the UCI .many rules were discussed in this paper for building a decision tree with the help of a hill climbing algorithm and exhaustive algorithm. The parameters

used for hill climbing algorithm were confidence (minimal confidence by a node to be considered as a node in the tree, value used was 0.25), minItemsets (minimum number of leaf under a node, value used was 2) and threshold (it was used for finding best subset under a node if the value is less than threshold then an exhaustive algorithm is used or else hill climb algorithm is used, value is 10). The paper was implemented using KEEL and WEKA tools. The result for the proposed methodology showed accuracy of 86.7% compared to the accuracy of decision tree which was 73.5.limitation of the research is that the data used for training the model was not treated for skewness and missing values. The data with class imbalance (skewness) would result in over fitted or under fitted model. For result and analysis K-Fold cross validation was not used. The missing values were replaced by using the mean of the values.

Youness Khourdifi et al, [3] proposed a Particle Swarm Optimization (PSO) combined with Ant Colony Optimization (ACO) and Fast Correlation-Based Feature Selection (FCBF) based optimized machine learning classifier. The data set used was the UCI heart disease data. WEKA tool was used for the implementation. The classification models used were K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Random Forest and a Multilayer Perception | Artificial Neural Network optimized by PCO and ACO. The best model were K-Nearest Neighbors and Random Forest after the optimization. Fast Correlation-Based Feature Selection was used for preprocessing of the dataset which included dimension reduction and removing the redundant data. the accuracy achieved by each classifier after the optimization were K-Nearest Neighbors of 99.6%, SVM 83%, Random forest 99.7%,NB 85%,MLP 90%. Limitation of the research is that it used uci Cleveland dataset which is skewed and contains the missing values. The skewed dataset is used for the feature selection process using ant colony optimization, no such work is done on the part of the machine learning algorithm. For selecting the features the ant colony optimization divides the data in to multiple smaller units and then using the correlation groups are formed. Due to class imbalance there is a greater chance of overfitting of machine learning models that were used including SVM, KNN, Logistic regression and mlp.

Costa W.L et al, [4] proposed a Decision support system using artificial neural network (ANN).the ANN used consist of a single hidden layer. Hidden layer consisted of 6 neurons and sigmoid as the activation function with a learning rate of 0.28, the optimization was done by using Nesterov's momentum optimizer to find the global minimum. The dataset used for this paper is the Cleveland Clinic dataset which is available on the UCI Machine Learning repository. The model has an accuracy of 90.76% and precision, recall and f1score to be 0.91. The Amazon EC2 cloud server was used for the hyper parameter tuning of the model. Limitation of the research is that the skewness of the data was not removed and the missing values were replaced by using the median of the values. The model was built by hyper tuning the parameters. The sigmoid function was used with a learning rate of 0.28. The sigmoid function would get struck in a local minimum and hence cannot reach an optimum value. Further due to hyper parameter tuning the model could not work better for new data as it is trained for the specific pattern.

3. Methodology:

The proposed model includes 2 stages in which the first stage is to generate an optimized model for prediction of the presence of heart disease in the patient. The block diagram is as shown in the figure 1. For building the optimized model we use the dataset discussed above. The dataset is treated for the missing values by using the NB (naïve Bayes) technique. Naïve Bayes is a probabilistic model based on the Bayes' theorem [13].

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Where A and B are events and B!=0. After treating the missing values we use the Pearson correlation coefficient to determine the correlation between the output and the input columns so that we could wisely choose the important columns in the feature set. It is always necessary that the input features must be correlated with the output feature also known as label .the label is always dependent on the input features so that there should be some linear or non-linear relation among them. The correlation coefficient is visualized by using a heat map. The heat map uses various colour codes to represent the relation among the various features columns. Then the data is divided into training and test set. For splitting the data into training and testing set we used the train test split method which firstly shuffles the indexes of data and then the data is split into training and testing sets. The training set is then oversamples to remove the skewness of the data that is the class imbalance so that the training set would not develop a biased model that is either under fitted or over fitted model. The training data is skewed that is imbalanced .hence we use the ADASY over sampler for resampling of the data. In oversampling we try to add more data in the training set for the minority class and in the under-sampling majority class data is removed to balance the data set. ADASYN oversampling involves the process of creating the row of the minority class dataset so that a balance can be brought. The oversampling will provide a better understand ability of the model regarding the entire population of the data set .rather than taking a portion of data from the population for building the model. The resampled training data is used for building the model. The machine learning algorithm used is random forest is hyper tuned for the parameters like number of estimators, max depth. The number of estimators is the bag size used for building the decision tree and the max depth is the length of the tree constructed over each of the bags formed by using the bootstrapping technique. For hyper tuning we have used the grid search cv which will use all the possible combinations of the number of estimators and the max depth provided to it via

the parameters list. The grid search cv results in the best fit parameters and the best score which was obtained. The best parameters is obtained according to the max score, the min time for fitting the data and the min time taken for predicting the test data. Then we use the best fit parameters to build our random forest classifier model. The result analysis is done by dividing the dataset in 80:20 ratio in which 80 percent of data is to be used for training purpose and rest 20 percent for testing. Traditionally the rest of 20 percent is predicted by the model and compared with the actual outcome. There are greater chances that the testing data would be biased towards any outcome. Hence you will have a statistical approach i.e. K-Fold Cross Validation. We have used K- Fold of 5 for the result analysis. The k-fold of 5 will form five sub dataset out of the original data sets where each subset contains 20 percent of original dataset. By doing so we have equal chance of selecting randomly dataset with presence and absence as an outcome.

Four of the sub datasets are used for training the model and left out dataset is predicted based on which the accuracy is calculated. The process is repeated for five times and the accuracy obtained in five times is the final accuracy of the model.

Further many tests are conducted during the K-Fold cross validation which includes the confusion matrix which contains the number of correctly predicted patients and falsely predicted patients with heart disease. Based on the attributes of the confusion matrix, further the precision of model, recall of the model accuracy of the model, AUC and F1 scores are calculated, by using the formula discussed in the multi-dimensional result analysis segment.
The results of the optimized model is compared with other models which are SVC (support vector machine classifier), Decision tree classifier, and Adaboost classifier. It would give a glimpse of the optimized model that whether it is out performing all the models or not.

Further in the second phase of the model development the optimized model that we obtained is used for the recommendation system with the prediction of heart disease in new patients. The 13 different attributes are collected from the user for the sake of prediction. Based on the outcome the user will have recommendation for the change in life style to reverse the heart conditions if possible or else the preventions to avoid any fatalities. We have implemented the recommendation system based on the vital information like bp, cholesterol levels and the exercise patterns the data used for the recommendation system is taken from the various medical websites and university. Further the model uses the simple conditional statements for the comparison of the vitals like bp, triglyceride, and cholesterol from the range shown in the table below. The block diagram of the recommendation model is shown the figure 2.

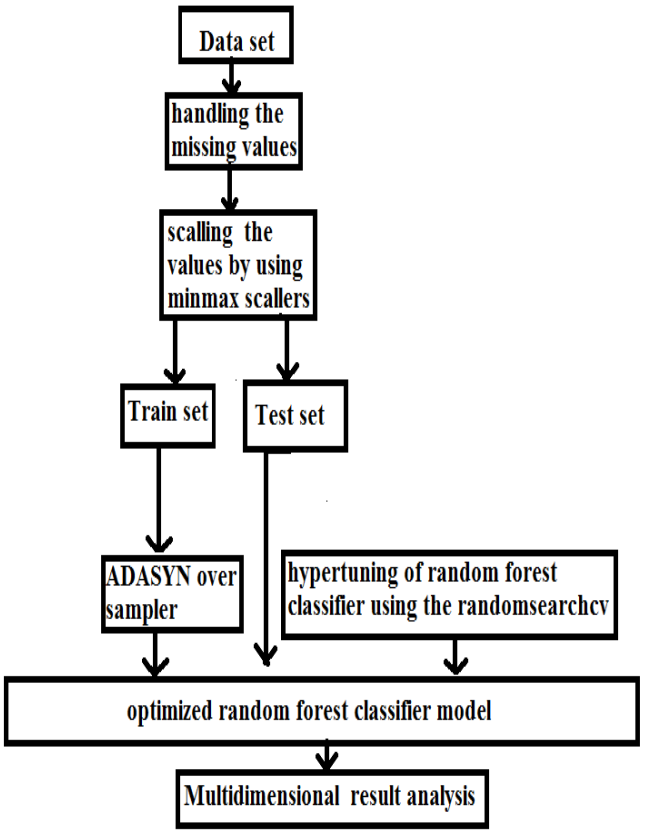


Fig 1. The proposed model.

Table 1: Triglyceride reference range.

S.No.	Levels	Measurement
1.	Normal	Under 150 mg/dl
2.	Borderline High	151-200 mg/dl
3.	High	201-499 mg/dl
4.	Very High	500 mg/dl or higher

Table 2: cholesterol reference range.

	Desirable	Borderline High	High
Total Cholesterol	Less than 200	200-239	240 and above
LDL Cholesterol(*bad cholesterol***)**	Less than 130	130-159	160 and higher
HDL Cholesterol(*good cholesterol***)**	50 and higher	40-49	Less than 40

Table 3: Blood pressure reference range

Blood Pressure Category	Systolic mm Hg (upper number)	Diastolic mm Hg (lower number)

Normal	Less than 120	Less than 80
Elevated	120-129	Less than 80
High Blood Pressure (Hypertension) Stage 1	130-139	80-89
High Blood Pressure (Hypertension) Stage 2	140 or Higher	90 or Higher
Hypertensive Crisis	Higher than 180	Higher than 120

.

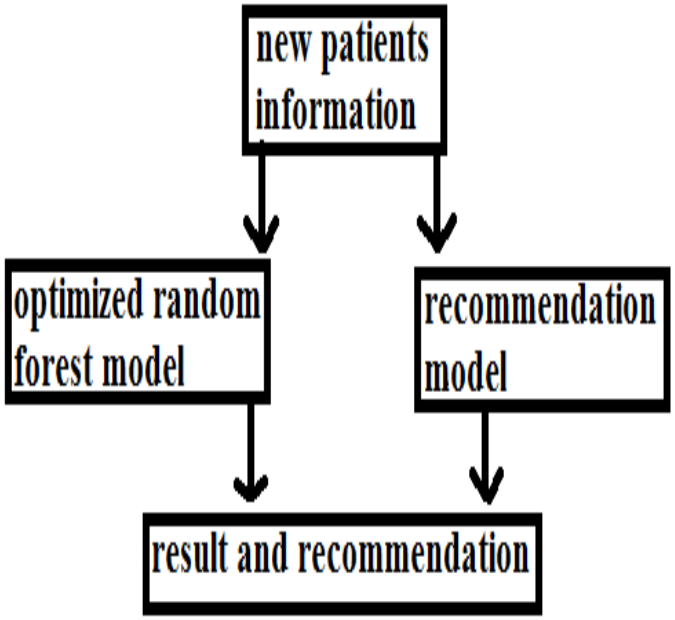


Fig 2: recommendation model.

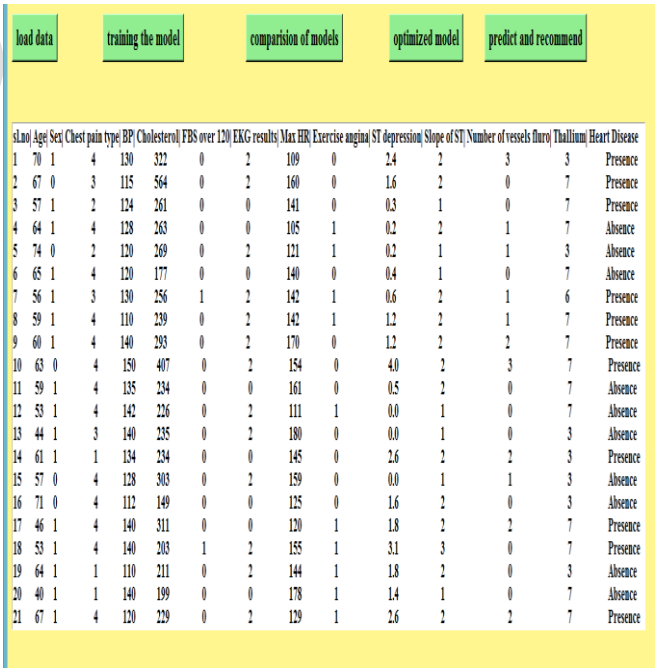
4. Result and Discussion:

For implementation of the proposed model python programming language is used. Python is a popular technology due its versatility of been used in the field of web site development, app development, IoT (internet of things) and in the field of research and analytics. For implementation of our model we have used the various tools of python like scikit-learn, pandas, imblearn, matplotlib, seaborn, jupyter notebook. For the frontend is developed by using the tkinter tool available in python. The dataset is obtained from the www.kaggle.com website. Kaggle is a website which provides a platform for researchers and data scientist, it offers various dataset for research purpose.

The GUI (graphical user interface) is divided into 5 different buttons the first button is for loading the data, second button is for training the model, third button is for comparisons of various models with our optimized random forest model, forth button is for the loading the optimized model and fifth button is for the prediction and recommendation of a new patient. The GUI is shown in the fig 3.

Fig 3: The default GUI page

When we click the load data button it prompts to select the dataset file available in our local directory. The data set will be loaded and the first 21 rows with all the fourteen columns are shown as shown in the figure 4.



sLno	Age	Sex	Chest pain type	BP	Cholesterol	FBS over 120	EKG results	Max HR	Exercise angina	ST depression	Slope of ST	Number of vessels fluro	Thallium	Heart Disease
1	70	1	4	130	322	0	2	109	0	2.4	2	3	3	Presence
2	67	0	3	115	564	0	2	160	0	1.6	2	0	7	Presence
3	57	1	2	124	261	0	0	141	0	0.3	1	0	7	Presence
4	64	1	4	128	263	0	0	105	1	0.2	2	1	7	Absence
5	74	0	2	120	269	0	2	121	1	0.2	1	1	3	Absence
6	65	1	4	120	177	0	0	140	0	0.4	1	0	7	Absence
7	56	1	3	130	256	1	2	142	1	0.6	2	1	6	Presence
8	59	1	4	110	239	0	2	142	1	1.2	2	1	7	Presence
9	60	1	4	140	293	0	2	170	0	1.2	2	2	7	Presence
10	63	0	4	150	407	0	2	154	0	4.0	2	3	7	Presence
11	59	1	4	135	234	0	0	161	0	0.5	2	0	7	Absence
12	53	1	4	142	226	0	2	111	1	0.0	1	0	7	Absence
13	44	1	3	140	235	0	2	180	0	0.0	1	0	3	Absence
14	61	1	1	134	234	0	0	145	0	2.6	2	2	3	Presence
15	57	0	4	128	303	0	2	159	0	0.0	1	1	3	Absence
16	71	0	4	112	149	0	0	125	0	1.6	2	0	3	Absence
17	46	1	4	140	311	0	0	120	1	1.8	2	2	7	Presence
18	53	1	4	140	203	1	2	155	1	3.1	3	0	7	Presence
19	64	1	1	110	211	0	2	144	1	1.8	2	0	3	Absence
20	40	1	1	140	199	0	0	178	1	1.4	1	0	7	Absence
21	67	1	4	120	229	0	2	129	1	2.6	2	2	7	Presence

Fig 4: first 21 rows of the heart disease dataset

The multi-dimensional result is visualized by using the bar graph plotted with the help of matplotlib library. The graph is shown in the figure 5.

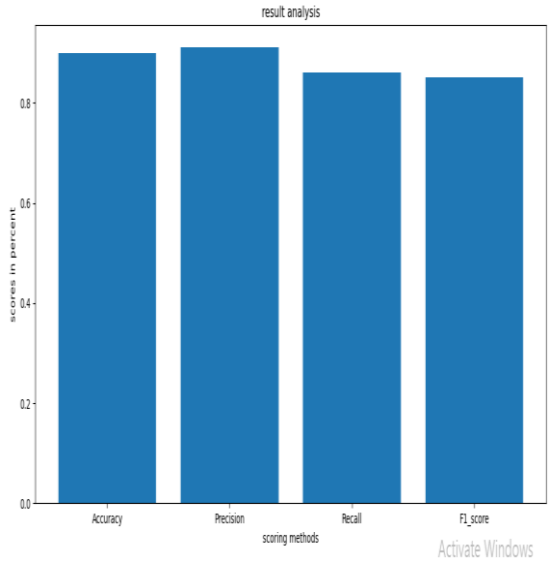


Fig 5. Graphical representation of result for the optimized model

The same dataset is used to training the other machine learning models like SVC, Decision tree classifier, Adaboost classifier. The result is obtained for the trained machine learning models. The result shows that the optimized random forest classifier has outperformed all the other models. The accuracy of the optimized machine learning model is 90% where the accuracy of SVC is 69% that is it is a under fitted model, the decision tree classifier resulted an accuracy of 76 % which is better in comparison of the SVC but under fitted model when compared to the optimized random forest classifier, and the accuracy of Adaboost classifier is 78% which is better than bot SVC and Decision tree but not better than the proposed model. The result is shown in the figure 6.

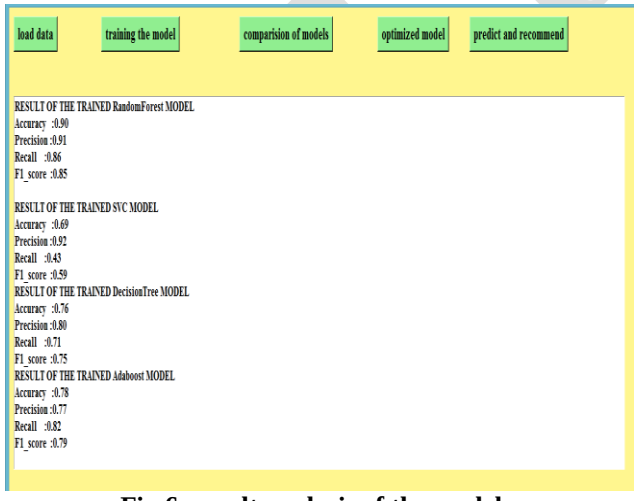


Fig 6: result analysis of the models

The result of every model is visualized by using the bar graph plotted with the help of matplotlib library. The graph is shown in the figure 7.

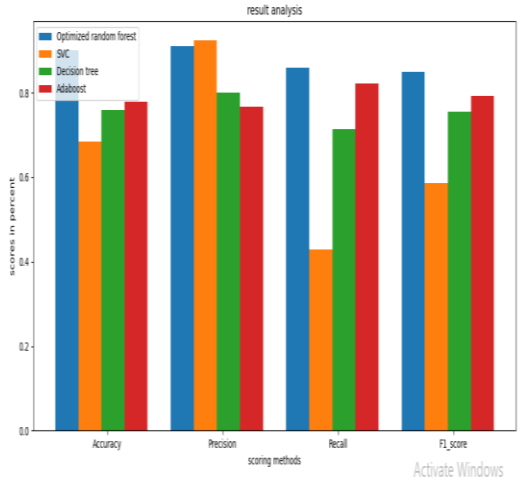


Fig 7. Graphical representation of result of the models.

The figure 8 shows the user interface for the prediction system. The user needs to fill the details required in the columns available in the left side of the GUI. After filling the details in the GUI one needs to click the predict button firstly which make sure that all the details are filled then the result is shown in the left side of the GUI. The result includes both the presence and absence result along with the recommendations based on the irregularities. It will recommend food and lifestyle changes to the patient in order to cure the patient.

Fig 8: User interface for the prediction and recommendation

5. Conclusion:

In our proposed work we have developed a better prediction model for prediction of patient's heart disease. The model was compared to other machine learning models which included SVC, Decision tree classifier. The proposed model outperformed all the other machine learning models. This model has a prediction accuracy of 90%.and we have implemented the recommendation system one of its kind. Which would guide the user regarding the steps to be taken for fighting the reversible heart diseases. For the development of model the Cleveland dataset is used. A proper feature

extraction mechanism was developed by using the combination of Pearson correlation, ADASYN over sampler and the naïve Bayes technique.
The proposed model also recommend lifestyle changes to the patients in order to cure the patient.

References:

[1]. Coronary heart disease and risk factors in India – On the brink of an epidemic?, M.N. Krishnan Indian Heart J. 2012 Jul; 64(4): 364–367. Doi: 10.1016/j.ihj.2012.07.001
[2]. Sharma Purushottam, Dr. Kanak Saxena, Richa Sharma, "Heart Disease Prediction System Evaluation using C4.5 Rules and Partial Tree", Springer, Computational Intelligence in Data Mining,vol.2, 2015, pp.285-294
[3]. Khourdifi, Youness and M. Bahaj. "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 12 (2019): 242-252.
[4]. Costa W.L., Figueiredo L.S., Alves E.T.A. (2019) "Application of an Artificial Neural Network for Heart Disease Diagnosis". In: Costa-Felix R., Machado J., Alvarenga A. (eds) XXVI Brazilian Congress on Biomedical Engineering.[online] IFMBE Proceedings, vol 70/2. Springer, Singapore. Availabe: https://doi.org/10.1007/978-981-13-2517-5_115
[5]. Amarbayasgalan T., Lee J.Y., Kim K.R., Ryu K.H. (2019) "Deep Autoencoder Based Neural Networks for Coronary Heart Disease Risk Prediction". In: Gadepally V. et al. (eds) Heterogeneous Data Management, Polystores, and Analytics for Healthcare. DMAH 2019, Poly 2019. Lecture Notes in Computer Science, vol 11721. Springer, Cham[online].Available: https://doi.org/10.1007/978-3-030-33752-0_17
[6]. K. G. Dinesh, K. Arumugaraj, K. D. Santhosh and V. Mareeswari, "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), Coimbatore, 2018, pp. 1-7, doi: 10.1109/ICCTCT.2018.8550857.
[7]. S. Mohan, C. Thirumalai and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," in IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[8]. N. L. Fitriyani, M. Syafrudin, G. Alfian and J. Rhee, "HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System," in IEEE Access, vol. 8, pp. 133034-133050, 2020, doi: 10.1109/ACCESS.2020.3010511.
[9]. J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan and A. Saboor, "Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare," in IEEE Access, vol. 8, pp. 107562-107582, 2020, doi: 10.1109/ACCESS.2020.3001149.
[10]. Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, 2020, Heart Disease Prediction using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 09, Issue 04 (April 2020)
[11]. Shah, Devansh & Patel, Samir & Bharti, Drsantosh. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science. 1. 10.1007/s42979-020-00365-y.
[12]. RAJESH, N et al. Prediction of Heart Disease Using Machine Learning Algorithms. International Journal of Engineering & Technology, [S.l.], v. 7, n. 2.32, p. 363-366, may 2018. ISSN 2227-524X.
[13]. Wikipedia, Bayes' theorem [online]. Available: https://en.wikipedia.org/wiki/Bayes%27_theorem
[14]. Kim, Y.J., Saqlian, M. & Lee, J.Y. "Deep learning–based prediction model of occurrences of major adverse cardiac events during 1-year follow-up after hospital discharge in patients with AMI using knowledge mining". Pers Ubiquit Comput (2019)[online].Available: https://doi.org/10.1007/s00779-019-01248-7
[15]. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.