How to Cite:
Thakur, S. K., & Sriramulu, S. (2022). Prognosticating of cardiovascular disease using machine learning algorithms. *International Journal of Health Sciences*, *6*(S5), 5174–5186. https://doi.org/10.53730/ijhs.v6nS5.9756

Prognosticating of cardiovascular disease using machine learning algorithms

Shridhar Kumar Thakur
School of Computing Science and Engineering, Galgotias University, Delhi NCR
Email: thakur.sridhar1@gmail.com

Srinivasan Sriramulu
School of Computing Science and Engineering, Galgotias University, Delhi NCR
Email: s.srinivasan@galgotiasuniversity.edu.in

***Abstract*---**Machine learning (ML) is such an application of the artificial intelligence, which has been outstanding in the field of research in modern science and technologies. In this paper it is explained how a machine can determine if a person is suffering from the cardiovascular disease or not. Heart disease has become so common that many people suffer from heart disease across the world, and even lose their lives. By considering certain factors such as chest pain, cholesterol level, a person's age and other factors a machine can determine whether the person has cardiovascular disease or not. Different supervised learning algorithms has been used in this paper in subject to make diagnosis of heart disease uncomplicated.

***Keywords*---**K Neighbors Classifier algorithm, Support Vector Classifier algorithm, Decision Tree algorithm, Random Forest Classifier algorithm, Heart Disease.

Introduction

The human body is made up of many different parts, each with its own function. The heart is one of such an organ. Heart pumps blood in order to supply the blood throughout the body. Human body can be in dangerous conditions if the heart doesn't performs it's function. Cardiovascular disease is one of the prominent causes of death [1]. So, it is necessary to make sure that any organ in our body such as our heart organ system or any different system present in human body should always remain healthy and in proper condition. Unfortunately, Due to the busy day to day routine, most people around the world have been plagued by heart disease as they don't give proper time on health issues.

If there is any technology that is able to help people with the diagnosis of such kind diseases before they become fatal and cause serious damage, will be proven useful in saving person's life and also their hard earned money. The methods of Data mining has proved themselves to be useful in the prediction of cardiovascular diseases in human body. By discovering patterns and trends that were previously unknown database and use of acquired information can make guessing of the models easy[2]. Data mining means extracting useful information from huge datasets [3]. Machine learning is a technique that can be used to diagnose the heart disease before it becomes fatal and make major damages to a person. In the modern era, machine learning has become an emerging field of computer science and technology and the methods can be used to differentiate if a person is healthy or is effected with heart disease.

Literature Review: Many researches have been performed on this sector and researchers have come up with many different procedures to predict heart diseases using supervised algorithms from machine-readable learning algorithms. After so much of research, many research articles have also been written after publishing explaining this particular issue. A survey conducted on heart disease was also presented in the form of data that came up as analysis result for the functionality of the several models of supervised machine learning algorithms and methods [4]. After analyzing all papers, in one of those papers, rigorous work was seen in the paper and to predict whether a person is suffering from heart disease or not Graphical User Interface (GUI) was created, by the help of the a classifier commonly known as Weighted Association rule-based Classifier [5]. A new approach was introduced in some other paper, focusing on a co-active neuro-fuzzy interference (CANFIS). It is a system mostly used for predicting the cardiac infraction [6]. The summary of the most familiarly used methods of predicting heart issues and its severity is provided in one of the research papers [7]. One of the papers introduced how to differentiate the non identical cardiovascular disease and also describes how the Naive Bayes classifier can be used for the motive of segregation [8]. In one of the papers, research was done which includes the results of many different articles in which one or more algorithms of data mining were used for the cardiovascular disease prognosis which gave accuracy. [9]

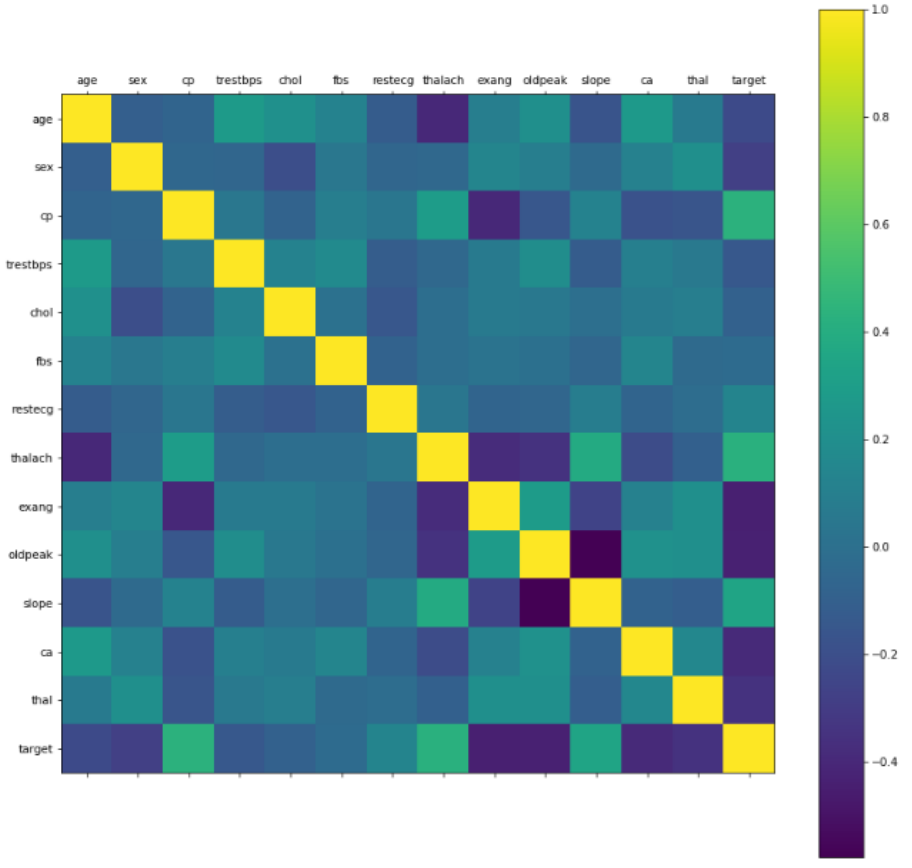
Data Collection: The first step in setting the model is to find a set of data containing the characteristics and findings of a human suffering from a cardiovascular disease and people who does not agree with the outcome that even if he/she is affected from the cardiovascular disease or not. The set of data which has been used in this test was taken from a website (https://archive.ics.uci.edu/ml/datasets/Heart+Disease). The programming language used to make the test is Python. The thirteen attributes found in the data set are used. Qualifications information for the dataset are also available on the website of Kaggle. The next step in carrying the research forward is analyzing the data. In this case, data set info is required. Collecting DataFrame summary, the info () function has been used in the dataset provided Pandas Library.

The explanatory function () provided by one of the python's library commonly known as Pandas Library and it is used to retrieve mathematical information form the dataset as a description of the number of attributes used. Targeted

attribute value mentioned as 1 value, if the patient has cardiovascular disease, or 0 for the person who does not have cardiovascular disease. Further the dataset will be verified whether it is a balanced dataset or not. This process is conducted using a countplot, function and it is provided by the Seaborn library (another machine learning library), in the target attribute.

Dataset details:

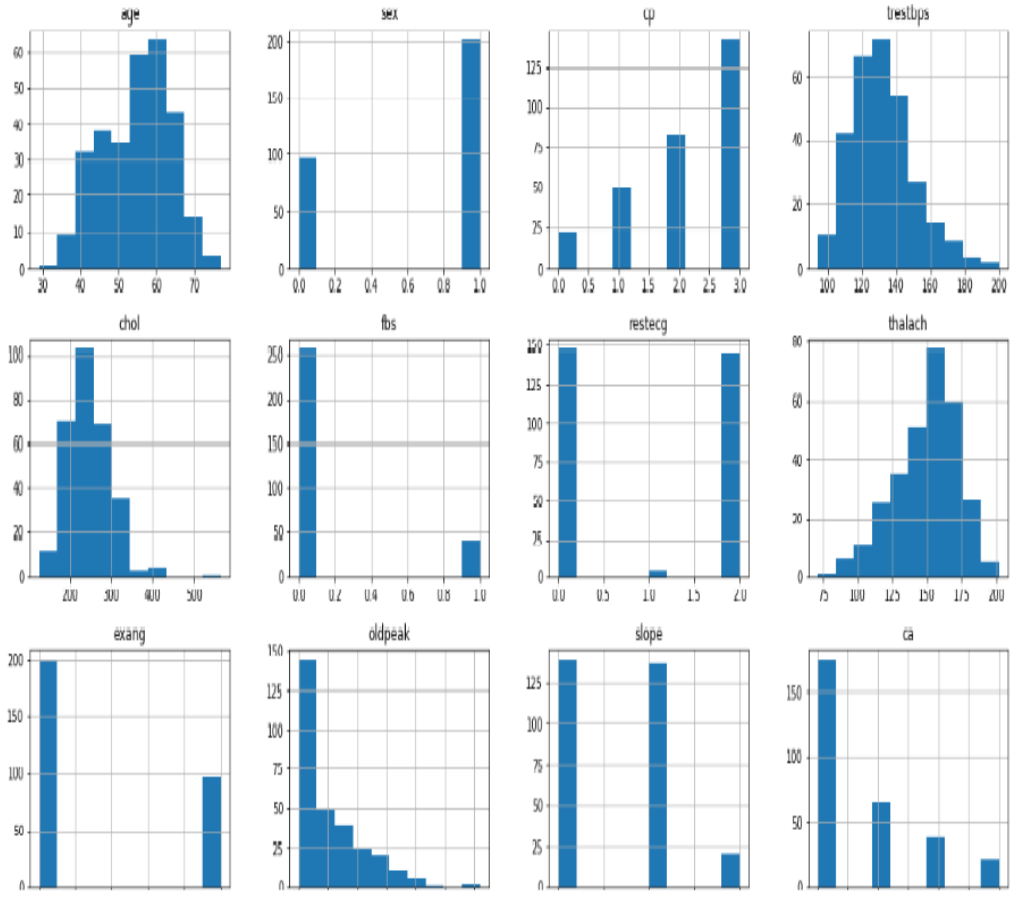
To get the depth of dataset, Correlation matrix is obtained.



The heat map above clearly shows features such as cp (chest pain) and thalack (high frequency in heart rate) have a positive relationship with the aimed attribute.

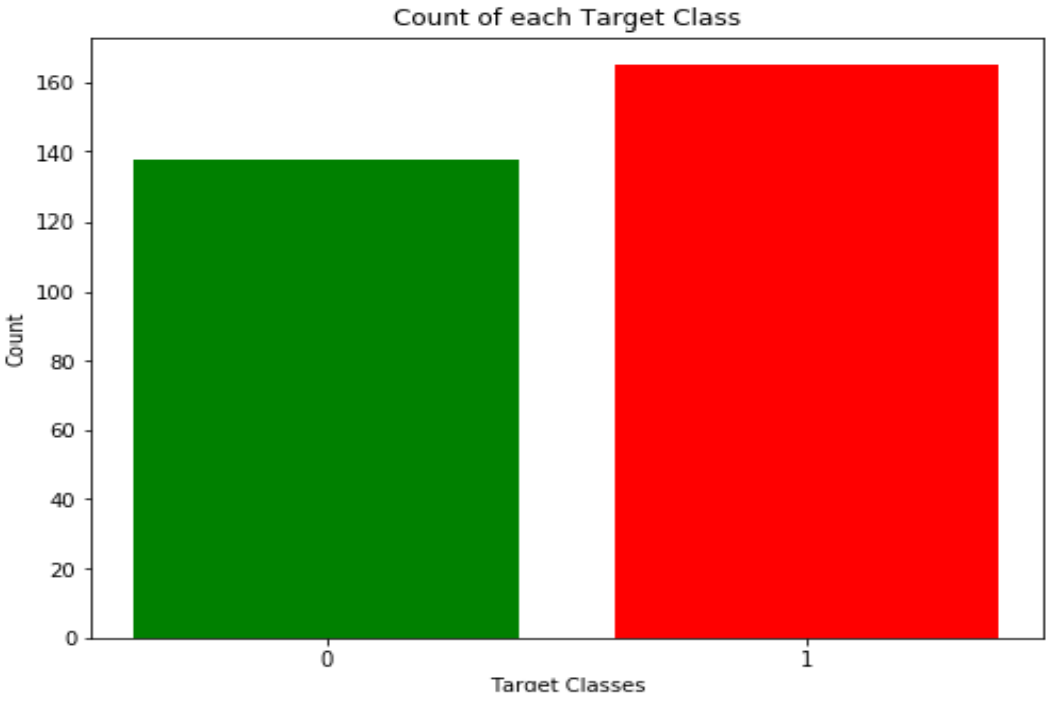
Now when the relationship has been tested, changes in the need to change the dynamics of a category like fbs, restecg, sex, cp, exang, slope, ca and thal is needed to be change into a dummy flexible. And doing so can be done using the get_dummies method which is part of Pandas library. After creation of a dummy variable, column data such as age, trestbps, chol, thalach and oldpeak needed standard measurement, as they had many different kind of values and units. To proceed with this, Scikit library in Python was taken in use.

In order to get the accuracy the set of information is divided into two parts, 66.6% of data was taken as training data from the entire data set and testing 33.3% data for the entire dataset. After the preparation of data, algorithms are used along with confusion matrix created. The results were obtained during algorithm accuracy. Accuracy has been achieved by using a confusion matrix.



After analyzing the histograms obtained above, it can be observed that the range of distribution is different in each feature. Thus, scaling before the predictions will give best outcome. Also, the categorical features does stand out.

It is always good to gain results if the target class in the dataset are approximately of equal size. And here in this case target class is not exactly same but it is close to the 50% target.



Checking for Skewed data

Skewed data is used to check the distribution of the data whether the given data is asymmetric or symmetric. This can be done by plotting the distribution plots so that some elucidation of the data can be observed. To get an overview of the analysis of data Different plots are presented. The distribution of gender and age, the distribution of cardiovascular ache and rest bp, the distribution of ecg resting electrode and thalach, the dividing of oldpeak and exang, the division of ca and slope, and the dealing out with target and thal, the division of cholesterol and fasting blood all are taken into account and analysis is done.

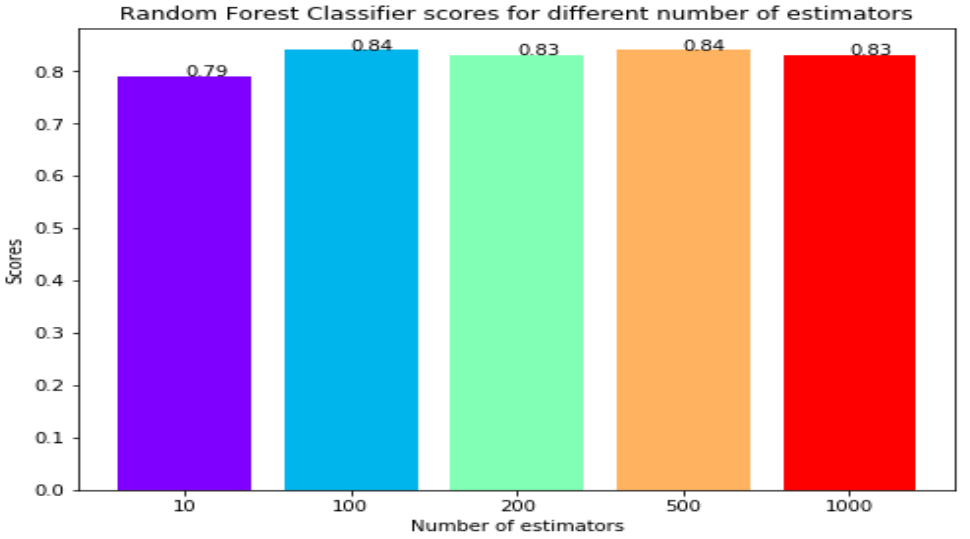
Splitting data

We've separated 66.6% of data from the available data set to train our model and rest 33.3% we've preserved to test our model [6]. The main reason to split the data into 66.6 : 33.3ratio is because if we train our model with maximum amount of the data then the accuracy of the data is best obtained.

Methodology

Machine learning techniques used by us in this reseaech paper and the experiment to obtain results are K Neighbors Classifier algorithm, Support Vector Classifier algorithm, Decision Tree Classifier algorithm, Random Forest Classifier algorithm.

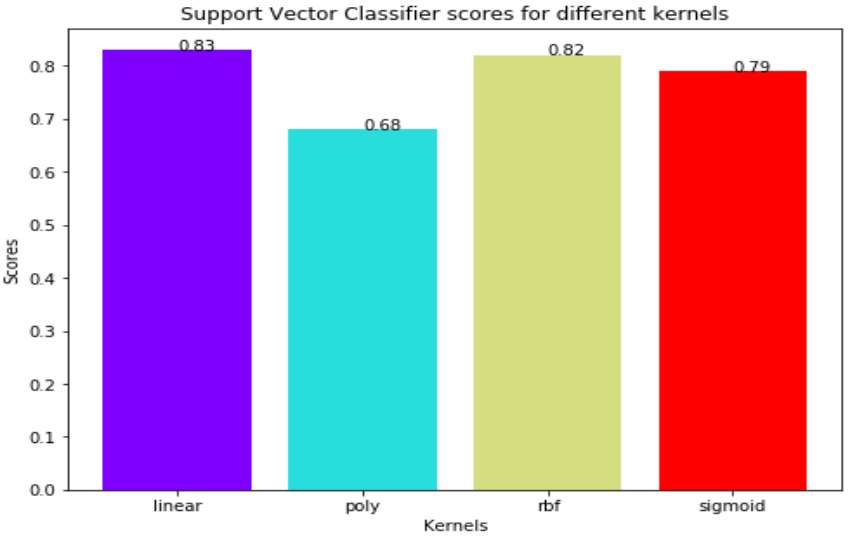
Random Forest regressor: Random Forest Regression algorithm is a supervised machine learning algorithm, it uses an integrated learning approach to decode. In this Ensemble learning method is a technique that merges the data obtained from various machine learning models which results a more accurate prediction than obtained from single model prediction. A random forest model which is a meta estimator that also controls overfitting and Fit a set of classification decision tree to different subsamples of the dataset and use averaging to improve prediction accuracy. The accuracy obtained in Random forest algorithm is 84.0% with [100,500] estimators.



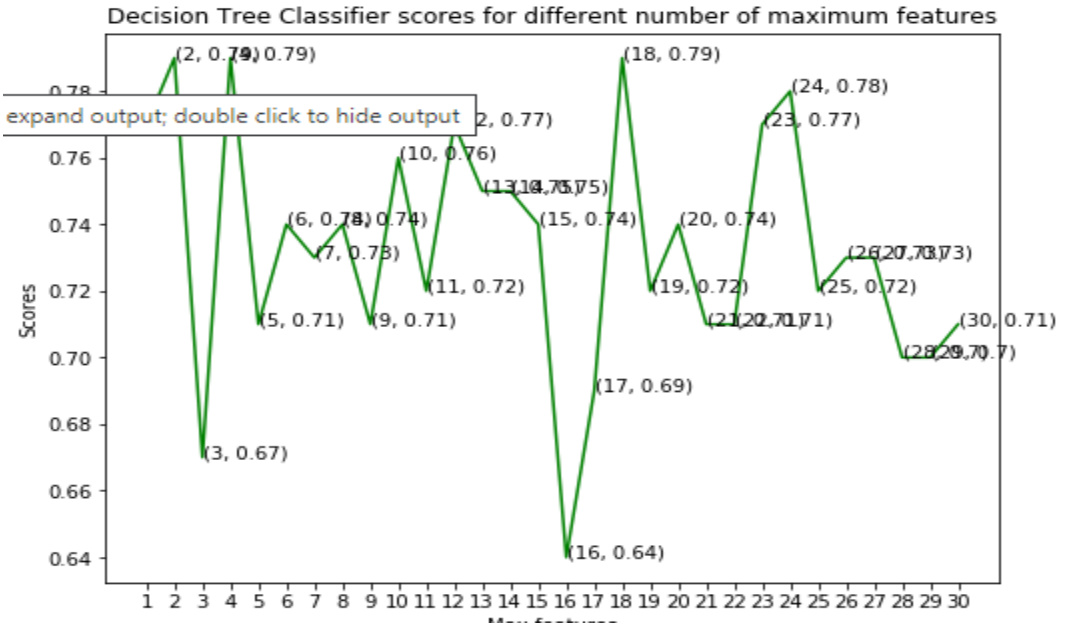
The diagram above shows the formation of the Random Forest.

2. Support Vector Machine:

SVM, It is most commonly used Supervised machine learning algorithm, which is commonly used for Scheduling and retrofitting problems. However, The main use of SVM done for Distribution issues in supervised Machine Learning. The main aim of the Support Vector Machine is to form a decision line between the plots which can divide the n-dimensional space in different classes to make all the data point placed easily in the best fitted category in the upcoming experiments. This decision making space is known as the hyperplane. SVM selects the extra vectors or points that helps in generating the hyperplane. These extra vectors are called supporting vectors, which is why the algorithm is called Vector Support Machine. The accuracy obtained in SVM model is 83.0% with linear kernel.

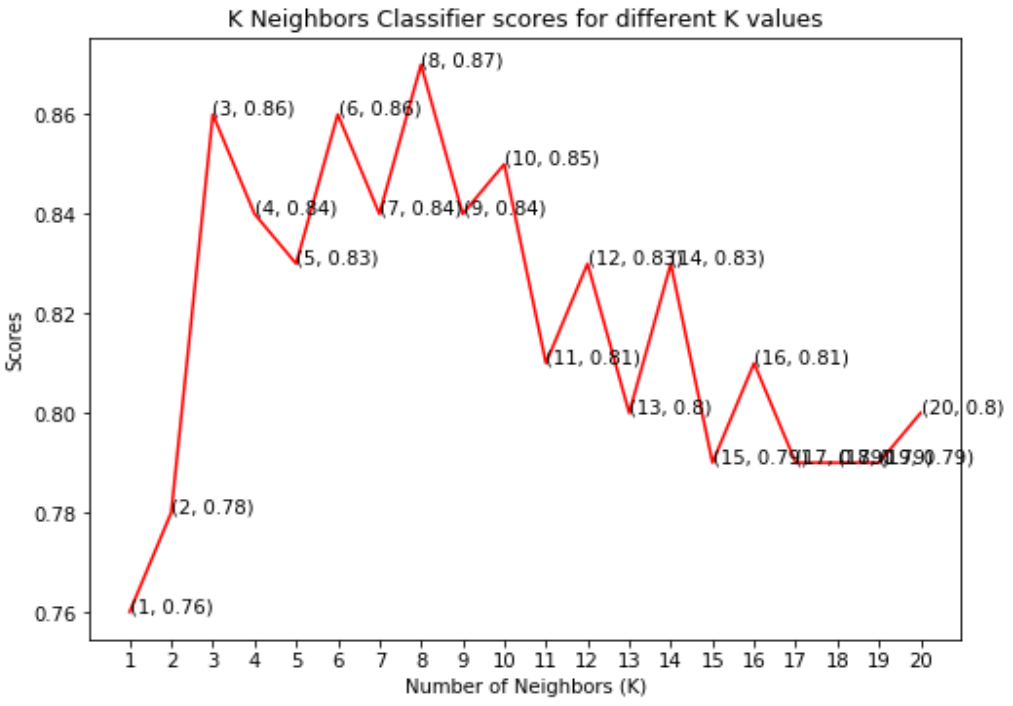


3. Decision Tree regressor: Decision Tree algorithm is also a supervised learning method that is being used for planning and retrospective problems, but is often preferred in solving planning problems. It is a divider which is tree-shaped, where each part represents different functions, such as the nodes inside represents the elements of the database, the branche nodes represents the rules of decision and the results are represented by each leaf node.
In the Decision Tree, there are only two main nodes, which are named as Leaf Node and Resolution Node. Resolution Nodes are mainly used to make any decision and have many branches, and Leaf nodes are the results of the decisions and have no other branches.
Decisions or tests are made on the basis of the given data.
It is metaphorical representation of finding all possible answers to any problem / decision-based situation.
It is called a trunk tree because, like commonly seen trees, it also starts with a root node, stretches out to more branches and forms a structure similar as tree.
Using decision tree the accuracy we obtained is around 79.0% with [2, 4, 18] maximum features.



4. K-Nearest Neighbour: A very simple supervised Learning algorithm contigent on Supervised Learning technique is K-Nearest Neighbour.
The main function of K-NN algorithm is to find the similarity in between the newly generated case and previously available cases and after identifying similarity, to select and put the new input into the most similar category in the available listed categories.
K-NN is an algorithm which saves all the available information and on basis of that it classifies a new milestone of any input points based on its similarity. It means whenever any new input is created then its category can be effortlessly recognized and classified into its well suited category by the help of K- NN algorithm.
K-NN algorithm is a technique which can be used for Regression problems and Classification problems as well. But commonly it is used for Classification type problems.
K-NN doesn't make any assumptions on underlying data. So it is also known as non-parametric algorithm.

K-NN algorithm is also known as a lazylearner algorithm, because it stores the dataset rather than learning and during classification, instead of performing any action on the dataset it doesn't learn from the training dataset immediately
.
KNN algorithm just stores the data whenever it gets the new data set in the training phase, after that it classifies the new data into a category that is replica to the new data.



The accuracy using KNN model was obtained around 87.0% with 8 neighbors.

Conclusion

In the paper, Four machine learning methods are proposed and analysis was also performed and the achieved results also were promising. The conclusion on which we reached is that the machine learning algorithm's performance are better in this analytic process. Many research scholars have also previously also suggested that the machine learning algorithm would be used when the amount of information in the available data set shouldn't be that large, which is also shown in this paper. Comparison were done in between the methods of comparing are precision, confusion matrix, specificity, sensitivity, and F1 score. Also, for the available features in the data set, K-Neighbors classifier algorithm performed better when the preprocessing was applied on data set in the ML approach.

Future Work

The time taken in the computation of deploying the model was also reduced. In the process it was also derived that the dataset the should be normalized. If it is not done the the model gets overfitted and the result obtained gets biased sometimes and the accuracy achieved from the biased dataset is not enought when working on a model for real-world live data problems. And this issue can lead to the variation of the model on different datasets. Statistical analysis of data is also equally important when the data set is being analyzed and preprocessed and also it should contain a Gaussian distribution, and after that the seller

discovery is equally important and for that a strategy which is commonly called as the Isolation Forest and is used to manage the same. Due to not so large dataset, we faced difficulty with the sample size. If a large amount of data set is provided, the results and the accuracy can be increased much in the deep learning, and Machine Learning too and model creation will also be accurate. In the ANN architecture the method applied by us has given the increase in the accuracy obtained which we had compared with the different research scholars. The amount of data in the dataset can be increased after that the process of deep learning with many other optimizing methods could be used after that we can obtained the most promising result. Many different optimization techniques including machine learning techniques can also be used in order to get the increased evaluation result. Still improving the estimation results using machine learning techniques and several other optimizing techniques can be done. More different methods can be used to normalize the data and compare the results. And for the convenience of patients and physicians, you will find more ways to integrate ML and DL models trained on heart disease with specific media.

Reference

World Health Organization, Cardiovascular Diseases, WHO, Geneva, Switzerland, 2020, https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1.

American Heart Association, Classes of Heart Failure, American Heart Association, Chicago, IL, USA, 2020, https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/classes-of-heart-failure.

American Heart Association, Heart Failure, American Heart Association, Chicago, IL, USA, 2020, https://www.heart.org/en/health-topics/heart-failure.

S. Shalev-Shwartz and S. Ben-David, "Understanding machine learning," From Theory to Algorithms, Cambridge University Press, Cambridge, UK, 2020.View at: Google Scholar

T. Hastie, R. Tibshirani, and J. Friedman, "The elements of statistical learning," Data Mining, Inference, and Prediction, Springer, Cham, Switzerland, 2020.View at: Google Scholar

S. Marsland, "Machine learning," An Algorithmic Perspective, CRC Press, Boca Raton, FL, USA, 2020.View at: Google Scholar

7. P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability," IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 3, pp. 727–733, 2013.View at: Publisher Site | Google Scholar

M. M. A. Rahhal, Y. Bazi, H. Alhichri, N. Alajlan, F. Melgani, and R. R. Yager, "Deep learning approach for active classification of electrocardiogram signals," Information Sciences, vol. 345, pp. 340–354, 2016.View at: Publisher Site | Google Scholar

G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," IEEE Journal of Biomedical and Health Informatics, vol. 18, no. 6, pp. 1750–1756, 2014.View at: Publisher Site | Google Scholar

R. Zhang, S. Ma, L. Shanahan, J. Munroe, S. Horn, and S. Speedie, "Automatic methods to extract New York heart association classification from clinical notes," in Proceedings of the 2017 IEEE International Conference on

Bioinformatics and Biomedicine (BIBM), pp. 1296–1299, IEEE, Kansas City, MO, USA, November 2017.View at: Publisher Site | Google Scholar
G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," International Journal of Applied Information Systems, vol. 3, no. 7, pp. 25–30, 2012.View at: Publisher Site | Google Scholar
E. Keogh and A. Mueen, "Curse of dimensionality," in Encyclopedia of Machine Learning and Data Mining, C. Sammut and G. I. Webb, Eds., Springer, Cham, Switzerland, 2017.View at: Publisher Site | Google Scholar
D. Wettschereck and T. G. Dietterich, "An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms," Machine Learning, vol. 19, no. 1, pp. 5–27, 1995.View at: Publisher Site | Google Scholar
D. Wettschereck, D. W. Aha, and T. Mohri, "A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms," Lazy Learning, vol. 11, no. 1/5, pp. 273–314, 1997.View at: Publisher Site | Google Scholar
M.-S. Yang and Y. Nataliani, "A feature-reduction fuzzy clustering algorithm based on feature-weighted entropy," IEEE Transactions on Fuzzy Systems, vol. 26, no. 2, pp. 817–835, 2018.View at: Publisher Site | Google Scholar
R. Chen, N. Sun, X. Chen, M. Yang, and Q. Wu, "Supervised feature selection with a stratified feature weighting method," IEEE Access, vol. 6, pp. 15087–15098, 2018.View at: Publisher Site | Google Scholar
M. Imani and H. Ghassemian, "Feature extraction using weighted training samples," IEEE Geoscience and Remote Sensing Letters, vol. 12, no. 7, pp. 1387–1391, 2015.View at: Publisher Site | Google Scholar
H. Liu and H. Motoda, Feature Extraction, Construction and Selection, Springer, Cham, Switzerland, 1998.
B. Dun, E. Wang, and S. Majumder, "Heart disease diagnosis on medical data using ensemble learning," 2016.View at: Google Scholar
R. S. Singh, B. S. Saini, and R. K. Sunkaria, "Detection of coronary artery disease by reduced features and extreme learning machine," Medicine and Pharmacy Reports, vol. 91, no. 2, pp. 166–175, 2018.View at: Publisher Site | Google Scholar
F. Yaghouby, F. Yaghouby, A. Ayatollahi, and R. Soleimani, "Classification of cardiac abnormalities using reduced features of heart rate variability signal," World Applied Sciences Journal, vol. 6, no. 11, pp. 1547–1554, 2009.View at: Google Scholar
B. M. Asl, S. K. Setarehdan, and M. Mohebbi, "Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal," Artificial Intelligence in Medicine, vol. 44, no. 1, pp. 51–64, 2008.View at: Publisher Site | Google Scholar
I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Feature Extraction: Foundations and Applications, Springer, Cham, Switzerland, 2008.
R. Rajagopal and V. Ranganathan, "Evaluation of effect of unsupervised dimensionality reduction techniques on automated arrhythmia classification," Biomedical Signal Processing and Control, vol. 34, pp. 1–8, 2017.View at: Publisher Site | Google Scholar
D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast

cancer," IEEE Access, vol. 6, pp. 28936–28944, 2018.View at: Publisher Site | Google Scholar

S. Negi, Y. Kumar, and V. M. Mishra, "Feature extraction and classification for EMG signals using linear discriminant analysis," in Proceedings of the 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall), IEEE, Bareilly, India, September 2016.View at: Publisher Site | Google Scholar

D. Avendaño-Valencia, F. Martinez-Tabares, D. Acosta-Medina, I. Godino-Llorente, and G. Castellanos-Dominguez, "TFR-based feature extraction using PCA approaches for discrimination of heart murmurs," in Proceedings of the 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 5665–5668, IEEE, Minneapolis, MN, USA, September 2009.View at: Publisher Site | Google Scholar

P. Kamencay, R. Hudec, M. Benco, and M. Zachariasova, "Feature extraction for object recognition using PCA-KNN with application to medical image analysis," in Proceedings of the 2013 36th International Conference on Telecommunications and Signal Processing (TSP), pp. 830–834, IEEE, Rome, Italy, July 2013.View at: Publisher Site | Google Scholar

N. R. Ratnasari, A. Susanto, I. Soesanti, and Maesadji, "Thoracic X-ray features extraction using thresholding-based ROI template and PCA-based features selection for lung TB classification purposes," in Proceedings of the 2013 3rd International Conference on Instrumentation, Communications, Information Technology and Biomedical Engineering (ICICI-BME), pp. 65–69, IEEE, Bandung, Indonesia, November 2013.View at: Publisher Site | Google Scholar

UCI Machine Learning Repository, "Heart disease data set," 2020, http://archive.ics.uci.edu/ml/datasets/heart+disease.View at: Google Scholar

S. S. Khan and S. M. K. Quadri, "Prediction of angiographic disease status using rule based data mining techniques," Biological Forum—An International Journal, vol. 8, no. 2, pp. 103–107, 2016.View at: Google Scholar

S. Kumar, "Predicting and diagnosing of heart disease using machine learning algorithms," International Journal of Engineering and Computer Science, vol. 6, no. 6, pp. 2319–7242, 2017.View at: Publisher Site | Google Scholar

R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, vol. 36, no. 4, pp. 7675–7680, 2009.View at: Publisher Site | Google Scholar

Khidoyatova, M. R., Kayumov, U. K., Inoyatova, F. K., Fozilov, K. G., Khamidullaeva, G. A., & Eshpulatov, A. S. (2022). Clinical status of patients with coronary artery disease post COVID-19. International Journal of Health & Medical Sciences, 5(1), 137-144. https://doi.org/10.21744/ijhms.v5n1.1858

K. Srinivas, G. Raghavendra Rao, and A. Govardhan, "Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in Proceedings of 2010 5th International Conference on Computer Science & Education, pp. 1344–1349, IEEE, Hefei, China, August 2010.View at: Publisher Site | Google Scholar

Widana, I.K., Sumetri, N.W., Sutapa, I.K., Suryasa, W. (2021). Anthropometric measures for better cardiovascular and musculoskeletal health. Computer Applications in Engineering Education, 29(3), 550–561. https://doi.org/10.1002/cae.22202

N. G. B. Amma, “Cardiovascular disease prediction system using genetic algorithm and neural network,” in Proceedings of the 2012 International Conference on Computing, Communication and Applications, IEEE, Dindigul, India, February 2012.View at: Publisher Site | Google Scholar

T. Santhanam and E. P. Ephzibah, “Heart disease classification using PCA and feed forward neural networks,” Mining Intelligence and Knowledge Exploration, Springer, Cham, Switzerland, 2013.View at: Publisher Site | Google Scholar

Harvard Medical School, “Throughout life, heart attacks are twice as common in men than women,” 2020, https://www.health.harvard.edu/heart-health/throughout-life-heart-attacks-are-twice-as-common-in-men-than-women.View at: Google Scholar

A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, “Classification models for heart disease prediction using feature selection and PCA,” Informatics in Medicine Unlocked, vol. 19, Article ID 100330, 2020.View at: Publisher Site | Google Scholar

P. Ramprakash, R. Sarumathi, R. Mowriya, and S. Nithyavishnupriya, “Heart disease prediction using deep neural network,” in Proceedings of the 2020 International Conference on Inventive Computation Technologies (ICICT), pp. 666–670, IEEE, Coimbatore, India, February 2020.View at: Publisher Site | Google Scholar