Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220031

Heart Disease Prediction Using Decision Tree in Comparison with k-Nearest Neighbor to Improve Accuracy

Maria Pratyusha[a,1] and K.V.Kanimozhi[b]
[a] *Research Scholar, Dept. of CSE, Saveetha School of Engineering,*
[b] *Asst Prof, Dept. of CSE, Saveetha School of Engineering,*
SIMATS, Chennai, India

Abstract. The target of the task is to foresee the coronary illness by Novel Decision Tree (DT) in examination with k-Nearest Neighbor (KNN) utilizing Cleveland dataset. Coronary Disease forecasting is performed by applying Decision Tree (N=20) and k-Nearest Neighbor (N=20) algorithms. Decision Tree algorithm uses the tree structure to make decisions. K-nearest neighbor is an easy approach to solve regression and classification problems. Cleveland heart dataset is utilized for identification and prediction. The data consists of 76 attributes however, only 14 features are selected that help in diagnosing a patient healthy or affected. Accuracy of cardiovascular risk prediction using k-NN is 68.9% & using decision tree is 81.9%. There exists a statistical significant difference between DT and k-NN with 0.035($p<0.05$). Decision Tree algorithm appears to perform significantly better than k- Nearest Neighbor algorithm for heart disease prediction.

Keywords. Heart Disease, K-Nearest Neighbor, Machine Learning, Novel Decision Tree, Prediction, Supervised Algorithm.

1. Introduction

Coronary illness prediction is the most common way of foreseeing the quantity of individuals impacted with heart illness or not [1].The purpose of this study is to predict heart disease using a decision tree algorithm [2]. Heart is an indispensable organ in the human body and siphons blood through the organization of conduits and veins called as cardiovascular framework.It is also prone to many diseases and it's important to prevent that from happening [3]. Many real time applications also include Health Care System, Medical Image processing [4].There are around 108 articles published in IEEE and 158 articles published in Science Direct for the past 5 years. These researchers have compared various algorithms. like NB, Bagging, the best result was obtained after WEKA resampling[5]. Theyhave proposedresearch on supervised machine learning algorithms, using the histogram, the accuracy, precision and f1 score [1].

Chi square statistics is used to test the data. Theyhave conducted a comparative study for various algorithms like Gaussian Naive Bayes, KNN classifier, SVM

[1]Maria Pratyusha, Department of CSE, Saveetha School of Engineering (Deemed to be), SIMATS, Chennai, TamilNadu, India - 602105. E-mail: mariapratyusha17@saveetha.com.

classifier and extra trees and concluded that Gaussian Naïve Bayes gives higher accuracy in contrast to the rest[6], [7]. Some researchers implemented a framework to discover coronary illness utilizing FCM and SEM. The model joins an additional of one numerical and a factual device to discover the connections and client relapse system for investigating the reason for coronary illness.This is the best suited paper among all the research papers [8].Our wide portfolio in research has translated into publications in numerous interdisciplinary projects [9-12].

The limitation of the existing literature is that it lacks accuracy while obtaining the result. Hence aim is to refine the performance using the novel decision tree algorithm in comparison with k-nearest neighbors by utilizing the recommended number of attributes.

2. Materials and Methods

Two groups are used for the research. The first one is k-Nearest Neighbors & the second being Decision Tree. Sample size was calculated by using previous study results; G Power software is used to calculate the sample size. The alpha value was set to 0.05, G power 80% and beta value was 0.02 [13].

Cleveland dataset UCI Repository is used for diagnosing the patient with or without a heart disease [14]. This dataset contains 14 (recommended) attributes and 303 entries. For both the groups 138 samples each are taken. The dataset is parted 2 ways one for testing & the other for preparation. For training data, 60% samples are taken and the rest for test data. After splitting the dataset, the algorithm is fit into train and test sets to predict the accuracy values.

2.1 K-Nearest Neighbour

K-Nearest Neighbors (KNN) is perhaps the most straightforward calculation utilized in Machine Learning for relapse and characterization issues.KNN algorithms use information and order new information in view of comparability measures(e.g. distance function). Categorization is performed by a major support to its neighbors [15].

2.2 Decision Tree

DT is a supervised algorithm that creates a classification model by constructing a decision tree. Every hub in the tree determines a test on a trait, each branch sliding from that hub relates to one of the potential qualities for that property.

A system with Windows OS and Hard disk capacity of 50GB is used. RAM of 8GB and Language used is Python, implemented in Jupyter (Anaconda).The processor used is intel i5, 7th Gen.

2.3 Statistical Analysis

Statistical Software used for our study is the IBM SPSS version 26. Using SPSS's descriptive and group statistics for accuracy are evaluated. Independent variables for Heart Disease Prediction are age, cholesterol, thalach, trestbps, ca and the dependent

variable is accuracy. The independent sample t test was performed to find the μ, σ and the σ_M statistical significance among groups.

3. Result

Table 1 shows group statistics which gives the accuracy mean of 81.9 for novel Decision Tree that appears to be more when compared with k-Nearest Neighbor, which has the accuracy mean of 68.85. Standard deviation and mean errors are calculated (Standard error mean for DT is 1.15 and kNN is 1.62).

Table 1.Group Statistics analysis for both algorithms based on Accuracy (Mean of DT 81.9 is more compared to kNN 68.9 and Standard Error Mean for DT is 1.393 and kNN is 1.522).

		N	μ	σ	σM
Accuracy	k-Nearest Neighbour	10	68.9	4.813	1.522
	Decision Tree	10	81.9	4.404	1.393

μ – Mean σ – Standard Deviation σ_M – Standard Error Mean

Table 2. Independent Sample Test for significance and standard error determination. P value is 0.035 ($p<0.05$) considered to be statistically significant and 95% confidence intervals were calculated.

	Levene's test for Equality of Variances		**T-test for equality of means**						
	F	**Sig.**	**t**	**df**	**Sig. (2tailed)**	**Mean Difference**	**Std. Error Difference**	**95% Confidence interval of the diffrence**	
								Lower	**Upper**
Equal variances assumed	.12	.035	-5.6	18	.000	-11.64	2.063	-15.97	-7.308
Equal variances not assumed	0.0	.00	-5.6	17	.000	-11.64	2.063	-15.97	-7.305

Table 2 shows the mean accuracy between k-NN and DT. Mean accuracy obtained in case of Decision Tree appears to be better than k- Nearest Neighbors and the standard deviation is slightly better for Decision Tree. From the results it's evident that Decision Tree has better accuracy in comparison to k-Nearest Neighbour.

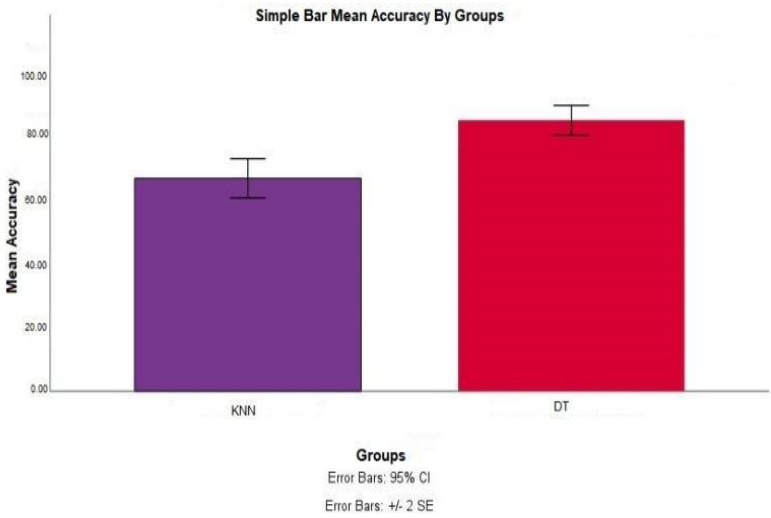

Figure 1. Bar Chart depicting the difference between Mean Accuracy of kNN and DT algorithms.

Mean accuracy of DT is better than kNN and standard deviation is slightly better for DT than kNN. DT appears to have variable results with standard deviation ranging from lower 60's to upper 80's. X axis represents K-Nearest Neighbor and Decision Tree algorithms and Y axis represents the mean accuracy ± 2 SE.

4. Discussion

In this study the Decision Tree algorithm appears to have better accuracy than k-Nearest Neighbour with p = 0.035 (which is less than 0.05), independent test sample. Improved mean accuracy score for Decision Tree is 0.82 while the same for k - nearest neighbors is 0.69. The sample attributes are tested statistically using the SPSS Tool. Researchershave concluded that DT surpassed all the other implementations. In this research,DT is better than NB and kNN [16]. Decision tree algorithms have shown better performance with the limited attributes in the research done by [17]. Theycompared all the supervised machine learning algorithms and concluded that the decision tree performs better. Theyhave used Naive Bayes, Decision tree and k-NN and have concluded that k-NN in this case performs better with accuracy[5]. This research has also conducted some algorithm analysis on the Cleveland dataset and has culminated that it has an accuracy of about 80%, which is better than the rest [18]. In our study of 6 works, 4 works are similar findings and 2 works are dissimilar findings. Based on the above discussion it proves that Decision Tree appears to have better accuracy and performance than k-Nearest Neighbors.

The limitations of Decision Tree are that the versions are not compatible to execute all required modules for the algorithm. There are also many violations and non-violations errors which have to be rectified using different parameters. In future we strive to improve this model with better and fewer attributes and try to get precise

results. As supervised learning helps a lot in solving real time problems like that of prediction we can implement detection for uncovered diseases.

5. Conclusion

In this work, we have implemented a heart illness forecast environment utilizing two methods kNN and DT. The former appears to have lesser accuracy than the latter. Decision Tree appears to perform significantly better than k -Nearest Neighbour for Heart Disease Prediction.

References

[1] Palaniappan, Sellappan, and RafiahAwang. Intelligent Heart Disease Prediction System Using Data Mining Techniques. In 2008 IEEE/ACS International Conference on Computer Systems and Applications. 2008;108–15.

[2] Hoffman, Julien I. E., and Samuel Kaplan. The Incidence of Congenital Heart Disease. Journal of the American College of Cardiology 2002;39 (12): 1890–1900.

[3] Karthiga, A. S., M. S. Mary, and M. Yogasini. Early Prediction of Heart Disease Using Decision Tree Algorithm."International journal ofhttps://www.researchgate.net/profile/Safish-Mary/publication/315023624_Early_Prediction_of_Heart_Disease_Using_Decision_Tree_Algorithm/links/58c84b57aca2723ab16eba60/Early-Prediction-of-Heart-Disease-Using-Decision-Tree-Algorithm.pdf.2017.

[4] Günal, Ö., Ç. E. Öztürk, G. Aksan, and G. Erdoğan. Evaluation of Electrocardiographic Ventricular Repolarization Variables in Patients with Newly Diagnosed COVID-19. Journal of. https://www.sciencedirect.com/science/article/pii/S0022073620304969. 2020

[5] Khateeb, Nida, and Muhammad Usman. Efficient Heart Disease Prediction System Using K-Nearest Neighbor Classification Technique. In The International Conference, 2017: 21–26. unknown.

[6] Chandra Shekar, K., Priti Chandra, and K. Venugopala Rao. An Ensemble Classifier Characterized by Genetic Algorithm with Decision Tree for the Prophecy of Heart Disease. In Innovations in Computer Science and Engineering, 2019;9–15. Springer Singapore.

[7] Panda, Debjani, and Satya Ranjan Dash. Predictive System: Comparison of Classification Techniques for Effective Prediction of Heart Disease. In Smart Intelligent Computing and Applications,.2020;203–13. Springer Singapore.

[8] KetutAgungEnriko, I., Muhammad Suryanegara, and DadangGunawan. 2016. "Heart Disease Prediction System Using K-Nearest Neighbor Algorithm with Simplified Patient's Health Parameters" 8 (12): 59–65

[9] Johnson, Jayapriya, Ganesh Lakshmanan, Biruntha M, Vidhyavathi R M, KohilaKalimuthu, and DurairajSekar. 2020. "Computational Identification of MiRNA-7110 from Pulmonary Arterial Hypertension (PAH) ESTs: A New microRNA That Links Diabetes and PAH." Hypertension Research: Official Journal of the Japanese Society of Hypertension 43 (4): 360–62.

[10] Keerthana, B., and M. S. Thenmozhi. 2016. "Occurrence of Foramen of Huschke and Its Clinical Significance." Journal of Advanced Pharmaceutical Technology & Research 9 (11): 1835.

[11] Krishna, R. Nivesh, and K. YuvarajBabu. Estimation of Stature from Physiognomic Facial Length and Morphological Facial Length. Journal of Advanced Pharmaceutical Technology & Research.2016; 9 (11): 2071.

[12] Kannan, Roghith, and M. S. Thenmozhi. Morphometric Study of Styloid Process and Its Clinical Importance on Eagle's Syndrome." Journal of Advanced Pharmaceutical Technology & Research.2016; 9 (8): 1137

[13] Gabriel Khan, M. Heart Disease Diagnosis and Therapy: A Practical Approach. Springer Science & Business Media.2005.

[14] Nahar, Jesmin, Tasadduq Imam, Kevin S. Tickle, and Yi-Ping Phoebe Chen. Association Rule Mining to Detect Factors Which Contribute to Heart Disease in Males and Females. Expert Systems with Applications 2013;40 (4): 1086–93.

[15] Laaksonen, J., and E. Oja. Classification with Learning K-Nearest Neighbors." In Proceedings of International Conference on Neural Networks (ICNN'96).1996; 3:1480–83 vol.3.

[16] Kim, Jaekwon, Jongsik Lee, and Youngho Lee. Data-Mining-Based Coronary Heart Disease Risk Prediction Model Using Fuzzy Logic and Decision Tree. Healthcare Informatics Research 2015; 21 (3): 167–74.
[17] Kumar, Abhishek, Pardeep Kumar, Ashutosh Srivastava, V. D. Ambeth Kumar, K. Vengatesan, and AchintyaSinghal. Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients. In Advances in Computing and Data Sciences, 2020; 507–18. Springer Singapore.
[18] Jabbar, M. Akhil, B. L. Deekshatulu, and Priti Chandra. Classification of Heart Disease Using K-Nearest Neighbor and Genetic Algorithm. Procedia Technology. 2013Jan: 85–94.