Comparative Study of KNN and Decision Tree in Prediction of Heart Disease

Aditya Ramesh
Dept. of Computer Science Engineering
PES University
Bengaluru, India
ramesh.aditya12@gmail.com

Dandu Satvik
Dept. of Computer Science Engineering
PES University
Bengaluru, India
satvikdandu6@gmail.com

Koushik Varma Mandapati
Dept. of Computer Science Engineering
PES University
Bengaluru, India
mkoushikvarma@gmail.com

Yamajala Siddhardha
Dept. of Computer Science Engineering
PES University
Bengaluru, India
yamajalasidhhardha1@gmail.com

Abstract—**Heart disease is one of the most common causes of death these days, and accurate prediction of heart disease is considered to be one of the most important topics in the domain of clinical data analysis. Complex tests to check the patient's health condition can be performed in hospitals with the help of doctors and physicians, but it is vital to keep monitoring the patient's health parameters when they are at home. For this purpose, there is a need for a simple and easy to implement algorithm which is capable of accurately predicting heart disease using some key parameters relating to the patient's health. This paper explores the effectiveness of KNN algorithm and Decision Tree in predicting heart disease on data obtained from the UCI Machine Learning Repository. To optimize results, the best K value is found using an elbow plot for KNN and pruning is performed on the Decision Tree to prevent overfitting.**

Keywords—Heart disease prediction, Data Analysis, Machine Learning, k-Nearest Neighbors, Decision Tree

I. INTRODUCTION

Heart disease is considered to be one of the deadliest chronic diseases in the world, accounting for an estimated 32% of all global deaths [1]. With advances in medical treatment and the identification of important cardiovascular risk factors, the contribution of heart disease to overall mortality began declining [2]. However, in the past decade, overall heart disease mortality rates have begun to stagnate and more worryingly, heart disease mortality rates among young adults has started to rise [3]. By making significant efforts to re-establish the downward trajectory of premature heart disease mortality rate, the health of the world could be improved [4].

Thus, in the modern day world, detection of heart disease is one of the most important tasks to improve the health of the world. However, accurate detection of heart disease is a complicated task and one of the biggest challenges in the healthcare sector. A survey conducted by the World Health Organization (WHO) found that medical professionals were only able to predict two-thirds of heart disease, making it very important to use other means so as to attain a higher success rate in predicting heart disease [5].

The healthcare sector has an abundance of data, but the existing data is not being utilized to its full capacity, and there is an absence of successful data analysis methods to find dominant patterns in healthcare data, especially when it comes to heart disease. To offset this, various data mining techniques could be used for the prediction of heart disease [6]. There are a variety of data mining techniques available for regression and classification of data. Under supervised learning, Naive Bayes, Logistic Regression, Decision Trees and k-Nearest Neighbors are some of the algorithms available. Under unsupervised learning, K-means Clustering, Principal Component Analysis and Independent Component Analysis are some of the techniques present.

The k-nearest neighbors (KNN) algorithm is a popular supervised machine learning algorithm that can be used to solve both classification and regression problems. It is very simple, easy to implement and the prediction is performed in a reasonable amount of time [7]. Decision Tree is another highly popular supervised machine learning algorithm that can solve classification problems by converting the data to a tree representation. It is easy to interpret, does not require scaling of data and thus performs relatively well even for larger datasets [8].

This paper intends to compare the effectiveness of the KNN algorithm and Decision Tree in predicting heart disease. These algorithms were chosen for the problem statement at hand since they are two of the most popular supervised machine learning algorithms around, and they are simple, easy to implement, understand and visualize. The choice of the k-value plays an important role in determining the accuracy of the KNN model, and so the optimal k-value will be found with the use of elbow plots. Decision Trees are prone to overfitting, thereby affecting the model's accuracy when dealing with untrained data. As a result, pruning will be performed on the Decision Tree wherein some decision nodes will be removed from the tree to generalize it and produce better results for untrained data.

The rest of the paper is organized as follows. Section II presents the related work in the field of applying machine learning algorithms for heart disease prediction. Section III describes the dataset being used, the pre-processing techniques followed to clean the data, the models built to approach the problem at hand and the evaluation metrics used to assess the models. In Section IV, the results of the models are presented, and optimization techniques are applied on both algorithms to observe better results. Section V concludes the paper, following which the individual contributions of each member of the project team are specified.

II. RELATED WORK

Javeed et al. (2020) proposed a feature selection method for feature refinement and subsequently used neural networks for classification to predict heart disease. A classification accuracy of 91.11% was achieved with the

ANN-based system while an accuracy of 93.33% was obtained with the DNN-based diagnostic system. Additionally, the authors found that the diagnostic system showed better performance and outcome than other state-of-the-art machine learning models. It was concluded that the proposed system could be used to help physicians make accurate decisions while diagnosing heart disease. [9]

However, the proposed system produces such a high accuracy of prediction only for small datasets. For larger datasets, there will be a higher degree of noise which would substantially reduce the accuracy of the system and accordingly, other approaches to normalize the data would have to be followed and deep learning with more optimization would need to be used to achieve more promising results. Decision Trees are robust to outliers or noisy data since these extreme values never cause much reduction in the Residual Sum of Squares (RSS), since they are never involved in the split.

Enriko et al. (2016) proposed the usage of the KNN algorithm to predict heart disease, and made use of simplifying parameters, so that the system could be used in machine-to-machine (M2M) remote monitoring of patients. To improve the accuracy, KNN algorithm was used with parameter weighting method and out of the 13 parameters recommended, only 8 parameters were used since they were simple and instantaneous parameters that could be measured at home. It was concluded that the model gave better results in quicker time in comparison with other data mining algorithms like Naive Bayes. Thus, the proposed system could be used in the future in machine-to-machine (M2M) technology to treat patients at home. [10]

While the heart disease evaluation is performed in a lesser amount of time, as instant parameters were used, the accuracy of the system will not be as high as it could be if all 13 parameters were utilized, since all the attributes or health parameters in the dataset were of some relevance in determining whether the patient had heart disease or not. The proposed KNN model in this paper will make use of all available parameters to predict heart disease with a significantly higher accuracy, as accuracy is more important than the time taken when it comes to predicting heart disease.

Saw et al. (2020) used a logistic regression model to predict whether the patients were suffering from heart disease or not. The authors pinpointed the most relevant factors that contributed to heart disease and analyzed these factors to predict the overall risk for a specific individual using logistic regression. The random search technique was utilized wherein random combinations of the hyper-parameters are tried in a trial and error manner to find the best solution for the built model. In the model proposed, a sigmoid function was used to help with the graphical representation of classified data. It was concluded that men were more susceptible to heart disease than women, and the model was evaluated to have an accuracy of 87% [11].

The proposed model will not be able to deliver the same accuracy over time since the symptoms involved will evolve over time thereby leading to some of the old training data becoming irrelevant and not as important. Thus, instead of using a model-based learning approach for evolving parameters, an instance-based approach could be followed wherein there will be no training period and all the training instances will be memorized by the model. Thus any change in the symptoms could easily be reflected onto the model, would cost relatively less and could be achieved in a short period of time.

III. Methodology

The approach used in this paper is depicted in Fig. 1 and is based on the following steps. Initially, the required data was collected from the UCI Machine Learning Repository. Subsequently, pre-processing techniques were applied to get rid of incomplete, inconsistent and noisy data. The KNN and Decision Tree models were then built using some assumptions after which the relevant evaluation metrics were used to assess the models. Each of these steps is described in detail in the following subsections.

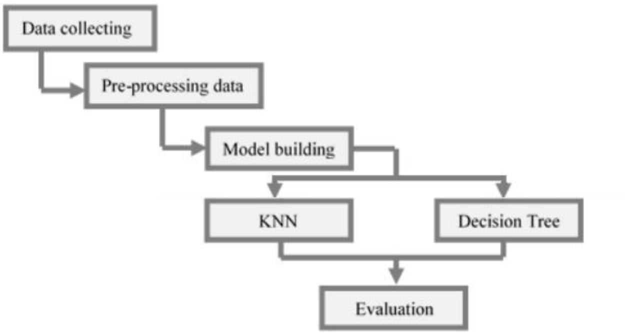
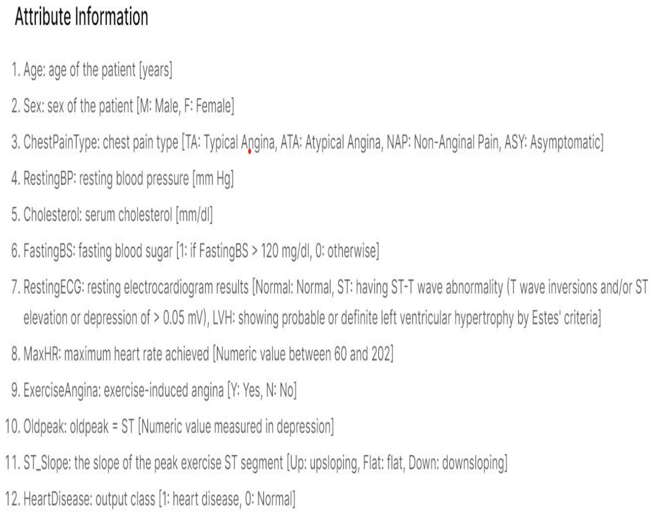


Fig. 1. Methodology Steps

A. Dataset Details

The relevant data required for analysis was obtained from the *Index of Heart Disease* datasets from the UCI machine learning repository via Kaggle [12]. 5 heart datasets were combined over 11 common features which made it the largest heart disease dataset available for research purposes. The original dataset consisted of 1190 observations and after removing the duplicate observations, this was shortened to 918 observations. There are a total of 12 attributes, all of them being health parameters relevant to heart disease prediction as shown in Fig. 2. The output class is binary in nature, with 1 representing the presence of heart disease and 0 representing the absence of heart disease factors.



Attribute Information

1. Age: age of the patient [years]
2. Sex: sex of the patient [M: Male, F: Female]
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. RestingBP: resting blood pressure [mm Hg]
5. Cholesterol: serum cholesterol [mm/dl]
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202]
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No]
10. Oldpeak: oldpeak = ST [Numeric value measured in depression]
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. HeartDisease: output class [1: heart disease, 0: Normal]

Fig. 2. Attribute information for the dataset

B. *Pre-Processing Techniques*

Data pre-processing is a data mining technique that is used to convert raw data into more meaningful, cleaner information that can be used for analysis. In the real world, data is full of redundancies, missing values and inconsistencies. This can compromise the integrity of the data being analyzed and as a result, pre-processing the data is an essential step before building models. The following techniques were applied on the obtained dataset.

- Removing missing values and duplicate rows
- Visualizing highly independent features and checking for multi-collinearity
- Finding and removing outliers or noisy data using box-plots
- Visualizing data using pie charts and scatter plots to make inferences regarding most likely values of attributes leading to heart disease
- Standardizing features and scaling to unit variance

It was observed that heart disease was most prominent in people whose age lies between 55 and 65. Additionally, it was observed that an unusual amount of men with heart disease suffered from asymptomatic chest pain when compared with typical angina, atypical angina and non-anginal pain.

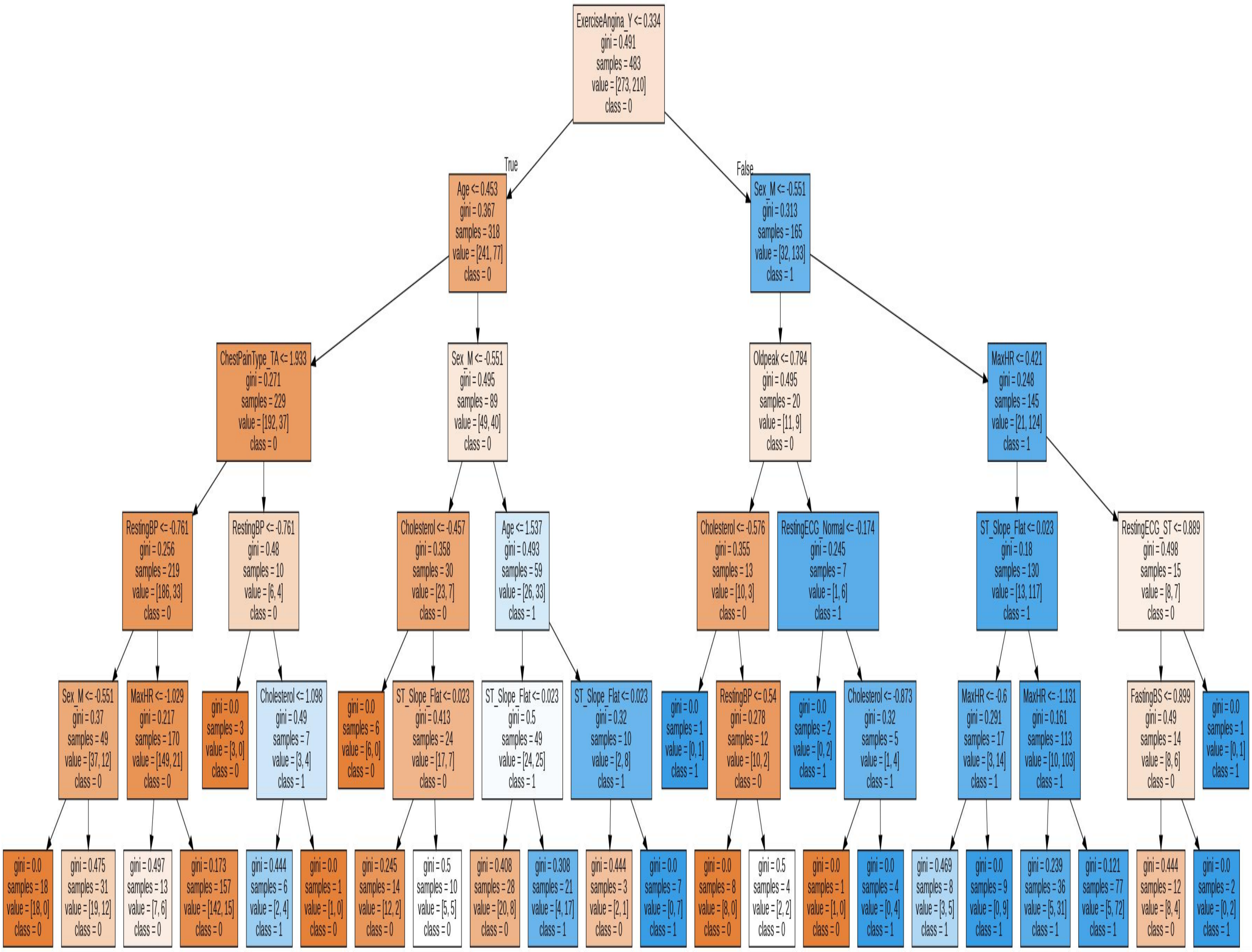
C. *Models Used*

For the purpose of heart disease prediction, there is a need for models that are highly accurate and relatively simple at the same time so as to obtain reliable results in a short amount of time. While there are a variety of supervised algorithms to choose from, the simplest and easiest to implement algorithms are the k-Nearest Neighbors (KNN) and Decision Tree algorithms. The details of these models and their pros and cons is described below.

K-Nearest Neighbor (KNN) is a supervised machine learning algorithm that determines the output class of a new instance based on the class of its K nearest neighbors in terms of distance. For classification purposes, the mode of the neighbors' classes is taken to be the output class of the new instance. While there are many distance metrics available, this paper uses the Euclidean distance which is calculated using (1).

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \qquad (1)$$

Since KNN is a lazy learner that follows instance-based learning, there is no training needed and new data can be added seamlessly without affecting the accuracy of the



ExerciseAngina_Y <= 0.334 gini = 0.491 samples = 483 value = [273, 210] class = 0
True
False
Age <= 0.453 gini = 0.367 samples = 318 value = [241, 77] class = 0
Sex_M <= -0.551 gini = 0.313 samples = 165 value = [32, 133] class = 1
ChestPainType_TA <= 1.933 gini = 0.271 samples = 229 value = [192, 37] class = 0
Sex_M <= -0.551 gini = 0.495 samples = 89 value = [49, 40] class = 0
Oldpeak <= 0.784 gini = 0.495 samples = 20 value = [11, 9] class = 0
MaxHR <= 0.421 gini = 0.248 samples = 145 value = [21, 124] class = 1
RestingBP <= -0.761 gini = 0.256 samples = 219 value = [186, 33] class = 0
RestingBP <= -0.761 gini = 0.48 samples = 10 value = [6, 4] class = 0
Cholesterol <= -0.457 gini = 0.358 samples = 30 value = [23, 7] class = 0
Age <= 1.537 gini = 0.493 samples = 59 value = [26, 33] class = 1
Cholesterol <= -0.576 gini = 0.355 samples = 13 value = [10, 3] class = 0
RestingECG_Normal <= -0.174 gini = 0.245 samples = 7 value = [1, 6] class = 1
ST_Slope_Flat <= 0.023 gini = 0.18 samples = 130 value = [13, 117] class = 1
RestingECG_ST <= 0.889 gini = 0.498 samples = 15 value = [8, 7] class = 0
Sex_M <= -0.551 gini = 0.37 samples = 49 value = [37, 12] class = 0
MaxHR <= -1.029 gini = 0.217 samples = 170 value = [149, 21] class = 0
gini = 0.0 samples = 3 value = [3, 0] class = 0
Cholesterol <= 1.098 gini = 0.49 samples = 7 value = [3, 4] class = 1
gini = 0.0 samples = 6 value = [6, 0] class = 0
ST_Slope_Flat <= 0.023 gini = 0.413 samples = 24 value = [17, 7] class = 0
ST_Slope_Flat <= 0.023 gini = 0.5 samples = 49 value = [24, 25] class = 1
ST_Slope_Flat <= 0.023 gini = 0.32 samples = 10 value = [2, 8] class = 1
gini = 0.0 samples = 1 value = [0, 1] class = 1
RestingBP <= 0.54 gini = 0.278 samples = 12 value = [10, 2] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1
Cholesterol <= -0.873 gini = 0.32 samples = 5 value = [1, 4] class = 1
MaxHR <= -0.6 gini = 0.291 samples = 17 value = [3, 14] class = 1
MaxHR <= -1.131 gini = 0.161 samples = 113 value = [10, 103] class = 1
FastingBS <= 0.899 gini = 0.49 samples = 14 value = [8, 6] class = 0
gini = 0.0 samples = 1 value = [0, 1] class = 1
gini = 0.0 samples = 18 value = [18, 0] class = 0
gini = 0.475 samples = 31 value = [19, 12] class = 0
gini = 0.497 samples = 13 value = [7, 6] class = 0
gini = 0.173 samples = 157 value = [142, 15] class = 0
gini = 0.444 samples = 6 value = [2, 4] class = 1
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.245 samples = 14 value = [12, 2] class = 0
gini = 0.5 samples = 10 value = [5, 5] class = 0
gini = 0.408 samples = 28 value = [20, 8] class = 0
gini = 0.308 samples = 21 value = [4, 17] class = 1
gini = 0.444 samples = 3 value = [2, 1] class = 0
gini = 0.0 samples = 7 value = [0, 7] class = 1
gini = 0.0 samples = 8 value = [8, 0] class = 0
gini = 0.5 samples = 4 value = [2, 2] class = 0
gini = 0.0 samples = 1 value = [1, 0] class = 0
gini = 0.0 samples = 4 value = [0, 4] class = 1
gini = 0.469 samples = 8 value = [3, 5] class = 1
gini = 0.0 samples = 9 value = [0, 9] class = 1
gini = 0.239 samples = 36 value = [5, 31] class = 1
gini = 0.121 samples = 77 value = [5, 72] class = 1
gini = 0.444 samples = 12 value = [8, 4] class = 0
gini = 0.0 samples = 2 value = [0, 2] class = 1

Fig. 3. Initial Decision Tree before pruning

algorithm. It requires only two parameters, the K-value and the distance metric. However, it is sensitive to noisy data and does not work well with larger datasets and higher dimensions. The choice of the k-value plays an important role in determining the accuracy of the model, as a small k-value could result in overfitting and a large k-value could result in underfitting. In this paper, the k-value is initially set to be 13.

Decision Tree is another popular supervised machine learning algorithm that represents the data in the form of a tree, with each node representing the class of the problem and each edge indicating the choice made based on the evaluated results. This paper utilizes the *Iterative Dichotomiser 3* (ID3) algorithm, where attributes having the least entropy or the maximum information gain are chosen. The complete decision tree generated for the dataset at hand is illustrated in Fig. 3.

The generated Decision Tree has a depth of 5 and a total of 53 nodes, with 27 different leaf nodes. While decision trees are intuitive with no requirement for either scaling or normalization, it takes longer to train the model and any small change in the data can cause instability in the decision tree. As a result, decision trees are prone to overfitting and they can be generalized by pruning the tree, which will be seen in Section IV.

D. *Evaluation Metrics*

There are a wide variety of evaluation metrics that can be applied to determine the efficiency of the models. This paper utilizes accuracy, precision, recall and the F1-score, all of which can be calculated using the elements of the confusion matrix. The four main factors involved in the confusion matrix are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

The accuracy of the model represents how well it is able to correctly identify high risk of getting heart disease and is given by (2).

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{2}$$

The precision of the model represents the proportion of patients diagnosed with heart disease who actually had high risk and is given by (3).

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{3}$$

The recall or sensitivity of the model represents how many of the patients with heart disease were positively identified and is given by (4).

$$\text{recall or sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{4}$$

The F1-score is a measure of the model's accuracy which is calculated by combining the precision and recall of the model, as shown in (5).

$$\text{F} - \text{measure} = \frac{2\ \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \tag{5}$$

IV. Results and Discussion

The results obtained for both the K-Nearest Neighbors (KNN) algorithm and the Decision Tree are analyzed in this section, with an emphasis on optimizing the results in both cases. The KNN result can be optimized by choosing an appropriate k-value and the Decision Tree accuracy could be improved by reducing overfitting with the help of pruning techniques.

A. *K-Nearest Neighbor (KNN)*

The KNN model generated results with an accuracy of 86.4% with an initial k-value of 13, which is one more than the number of attributes. The choice of the k-value determines the accuracy of the model. A smaller k-value could result in overfitting and a larger k-value could result in underfitting. As a result, finding the ideal k-value can prevent both underfitting and overfitting, thereby maintaining the optimal balance.

To determine the optimal k-value, plots of k-value vs F1-score and k-value vs error (elbow curve) are mapped as depicted in Fig. 4 and Fig. 5 respectively. It can be seen that for k = 7, the F1-score is maximum and the error is minimum, meaning the optimal k-value for the dataset is 7. The confusion matrix for the KNN model with k = 7 is shown in Table I. The improved accuracy of the model with a k-value of 7 was found to be 89.37%, an increase of almost 3%.

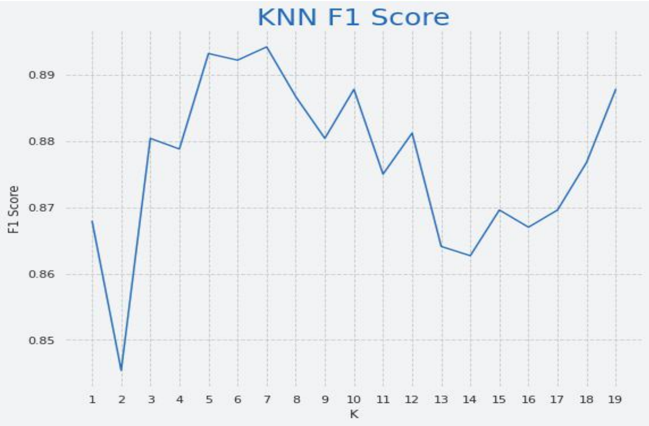


Fig. 4. Plot of K-value vs F1-score

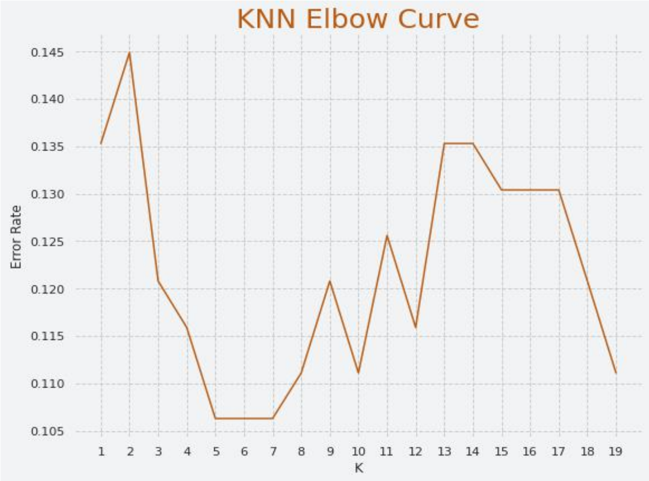


Fig. 5. Plot of K-value vs Error (Elbow Curve)

TABLE I. CONFUSION MATRIX FOR KNN (K=7)

Prediction	**Reference**	
	Heart Disease	**Normal**
Heart Disease	92	11
Normal	11	93

B. *Decision Tree*

The initial decision tree generated for the dataset had a depth of 5 and 27 different leaf nodes, as shown in Fig. 3. It generated an accuracy of 87.78% in the training phase and an accuracy of 77.78% in the testing phase. The significant difference in accuracy clearly indicates overfitting of the model to the training data. Decision Trees are prone to overfitting which could lead to alarmingly low accuracy rates for testing data. To avoid this, the decision tree could be pruned and thus generalized.

Pruning is a data compression technique that reduces the size and depth of the decision tree by eliminating sections of the tree that are non-essential to classify instances. The decision tree obtained after pruning the original tree is shown in Fig. 6, with the depth being reduced from 5 to 2 and the leaf nodes being reduced from 27 to 2. The confusion matrix for the pruned decision tree is shown in Table II. The pruned tree produced a training accuracy of 81.37% and a testing accuracy of 84.06%, an increase of over 6% from the original testing accuracy of 77.78%.

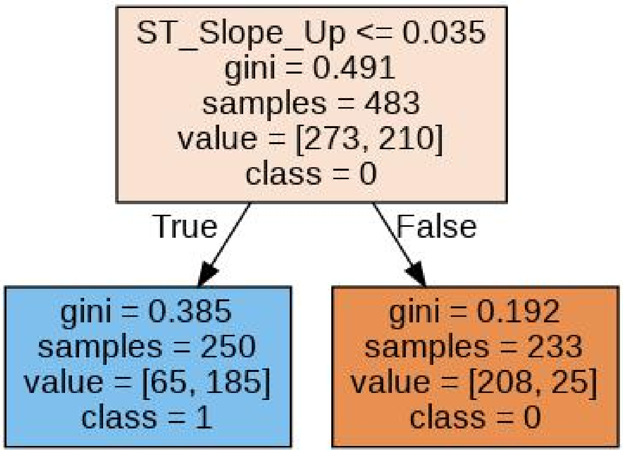


Fig. 6. New Decision Tree after pruning

TABLE II. CONFUSION MATRIX FOR DECISION TREE AFTER PRUNING

Prediction	**Reference**	
	Heart Disease	**Normal**
Heart Disease	93	10
Normal	13	91

V. CONCLUSIONS

In this paper, a comparison of the effectiveness of the k-Nearest Neighbors (KNN) algorithm and the Decision Tree in predicting heart disease was performed. The data was obtained from a UCI machine learning repository via Kaggle, following which pre-processing was performed to clean the data. The KNN and Decision Tree models were then built with some initial assumptions, and were evaluated using accuracy, precision, recall and F1-score. The results were optimized by choosing the appropriate k-value for KNN and through pruning for the Decision Tree. The optimized KNN model gave an accuracy of 89.37% and the optimized Decision Tree gave an accuracy of 84.06%. Thus, it can be concluded that KNN is more efficient than Decision Tree in predicting heart disease.

INDIVIDUAL CONTRIBUTIONS

- *Aditya Ramesh (PES1UG19CS033)* - Optimization of the models and their evaluation
- *Dandu Satvik (PES1UG19CS128)* - Building the Decision Tree
- *Koushik Varma Mandapati (PES1UG19CS230)* - Building the KNN model
- *Yamajala Siddhardha (PES1UG19CS588)* - Data Pre-Processing and Visualization

REFERENCES

[1] World Health Organization, June 2021, “Cardiovascular diseases (CVDs),” [online] retrieved on 12 November 2021 from https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[2] L. Goldman and E.F. Cook, "The decline in ischemic heart disease mortality rates. An analysis of the comparative effects of medical interventions and changes in lifestyle," Ann Intern Med, vol. 101, no. 6, pp. 825-836, 1984.

[3] E.S. Ford and S. Capewell, "Coronary heart disease mortality among young adults in the U.S. from 1980 through 2002: concealed leveling of mortality rates," J Am Coll Cardiol, vol. 50, no. 22, pp. 2128-2132, 2007.

[4] M. D. Ritchey, H. K. Wall, M. G. George and J. S. Wright, "US trends in premature heart disease mortality over the past 50 years: Where do we go from here?," Trends in Cardiovascular Medicine, vol. 30, no. 6, pp. 364-374, 2020.

[5] H. Sharma and M. Rizvi, "Prediction of Heart Disease using Machine Learning Algorithms: A Survey," unpublished.

[6] M. Gandhi and S. N. Singh, "Predictions in heart disease using techniques of data mining," 2015 International Conference on Futuristic Trends on Computational Analysis and Knowledge Management (ABLAZE), pp. 520-525, 2015.

[7] X. Wu, V. Kumar, R. Quinlan and J. Ghosh, "Top 10 algorithms in data mining," Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, 2008.

[8] H. Patel and P. Prajapati, “Study and Analysis of Decision Tree Based Classification Algorithms,” International Journal of Computer Sciences and Engineering, vol. 6, pp. 74-78, 2018.

[9] A. Javeed, S. S. Rizvi, S. Zhou, R. Riaz, S. U. Khan and S. J. Kwon, "Heart Risk Failure Prediction Using a Novel Feature Selection Method for Feature Refinement and Neural Network for Classification," Mobile Information Systems, vol. 4, pp. 1-11, 2020.

[10] I. K. A Enriko, M. Suryanegara and D. Gunawan, "Heart disease prediction system using k-Nearest neighbor algorithm with simplified patient's health parameters," Journal of Telecommunication, Electronic and Computer Engineering, vol. 8, no. 12, pp. 59-65, 2016.

[11] M. Saw, T. Saxena, S. Kaithwas, R. Yadav and N. Lal, "Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-6, 2020.

[12] fedesoriano, September 2021, “Heart Failure Prediction Dataset,” [online] retrieved on 18 September 2021 from https://www.kaggle.com/fedesoriano/heart-failure-prediction.