Comparative Analysis of Logistic Regression, KNN, and Decision Tree on heart and kidney dataset: Much focus on the use of trees

OLAMIDE OSENI[1]

AFRICAN INSTITUTE FOR MATHEMATICAL SCIENCES

January 23, 2024

[1] Supervised by Professor Ernest Fokoue

1. **Introduction**
2. Objectives
3. Methodology
4. Famous script D$_n$
5. Some variables used
6. Analysis
7. Conclusion

INTRODUCTION

The advancement of patient data processing and treatment has been significantly made easier by machine learning (ML). Thanks to machine learning, doctors can now gather and organize patient data, spot trends in healthcare, and suggest treatment.

1. The term “heart disease” refers to several types of heart conditions. The most common type of heart disease in the United States is coronary artery disease (CAD), which affects the blood flow to the heart.
2. CKD is a condition in which the kidneys are damaged and cannot filter blood as well as they should. Because of this, excess fluid and waste from blood remain in the body.

1. Introduction
2. Objectives
3. Methodology
4. Famous script D$_n$
5. Some variables used
6. Analysis
7. Conclusion

OBJECTIVES

1. To do a quick explanatory data analysis.

2. To perform kidney and heart disease detection by implementing three machine learning models as well as to test out how this different models perform on the task.

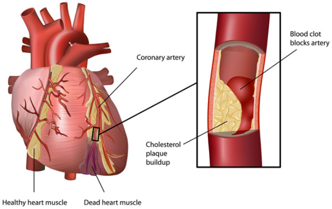


Figure 1: Source: Photo from **Wikipedia**

"A single death is a tragedy; a million deaths is a statistic." — Joseph Stalin

1. Introduction
2. Objectives
3. **Methodology**
4. Famous script D_n
5. Some variables used
6. Analysis
7. Conclusion

METHODOLOGY

$$\mathcal{D}_{\mathrm{n}} \longrightarrow \mathcal{H} \longrightarrow \mathcal{P} \longrightarrow \mathcal{A} \longrightarrow \mathcal{R} \longrightarrow \mathcal{Y} \longrightarrow \mathcal{C}$$

Above is the 7 Wheels of Machine Learning and we have 3 Machine learning models to use on each of the data.

They are:

- **KNN:** $D = \{(\mathbf{x}_i, y_i)\}$, where $\mathbf{x}_i$ is a feature vector and y_i is the class label.

$$\hat{y}(\mathbf{x}) = \mathrm{argmax}_y \sum_{i=1}^{k} I(y_i = y)$$

- **Tree:** Decision at a node:

$$\text{Decision: } f_j(\mathbf{x}) \leq \theta_j$$

Prediction at a leaf:

$$\hat{y}(\mathbf{x}) = c$$

Cont. METHODOLOGY

- **Logisitic regerssion**

  The logistic regression model is represented as:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i \tag{1}$$

Here:

- p_i represents the probability of the binary outcome for the i^{th} observation.
- X_i is the predictor variable associated with the i^{th} observation.
- α is the intercept term.
- β is the coefficient associated with the predictor variable X_i.

1. Introduction
2. Objectives
3. Methodology
4. **Famous script D_n**
5. Some variables used
6. Analysis
7. Conclusion

FAMOUS SCRIPT D$_n$

Table 1: Summary of the datasets.

SN	Dataset	n	k	P + k	k = n/p
1	CKD	400	2	26	16.7
3	Heart	303	1	14	23.3

The benchmark datasets used in this study are rich in terms of varieties and dimensions. We have **multi-class** datasets and they all represent the measure of the information (data) richness because they have a Kappa ≥ 5.

Mathematically represented as:

$$D_n = \{(x_i, y_i) \text{ iid} \sim p_{xy}(x, y), \quad x_i \in \mathbb{R}^2, \quad y_i \in \{-1, +1\}, \quad i = 1, \ldots, n\}$$

1. Introduction
2. Objectives
3. Methodology
4. Famous script D$_n$
5. **Some variables used**
6. Analysis
7. Conclusion

DESCRIPTION OF THE VARIABLES IN THE DATASETS

Variables in kidney dataset

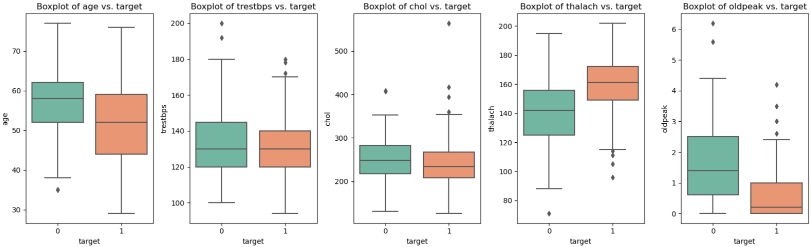
The dataset encompasses various features such as age, blood pressure (bp), sugar level (su), haemoglobin level (hemo), and more. it includes the classification indicating whether the individual has Chronic Kidney Disease.

Variables in heart dataset

For heart-related data, the dataset includes features like age, sex, chest pain type (cp), resting blood pressure (trestbps), cholesterol levels (cholestor), and others. The target variable indicates whether the person has a heart condition or not.

1. Introduction
2. Objectives
3. Methodology
4. Famous script D$_n$
5. Some variables used
6. **Analysis**
7. Conclusion

Exploratory Data Analysis Prior to Classification

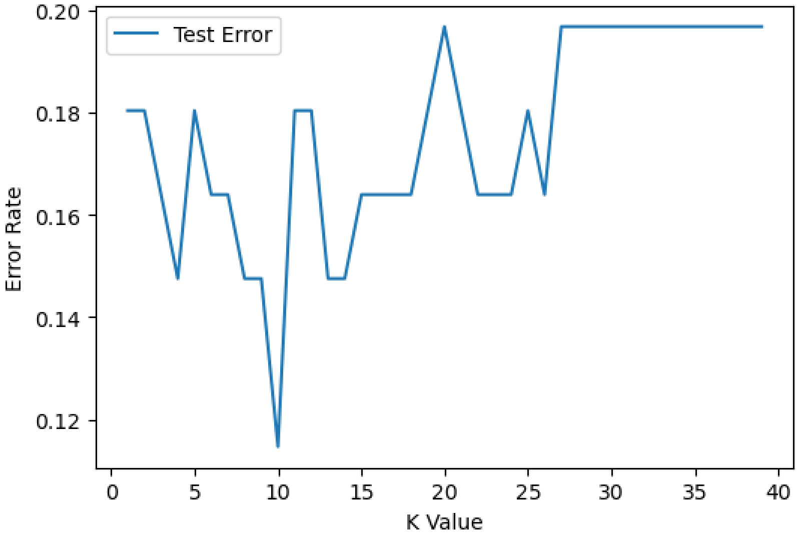


ANALYSIS

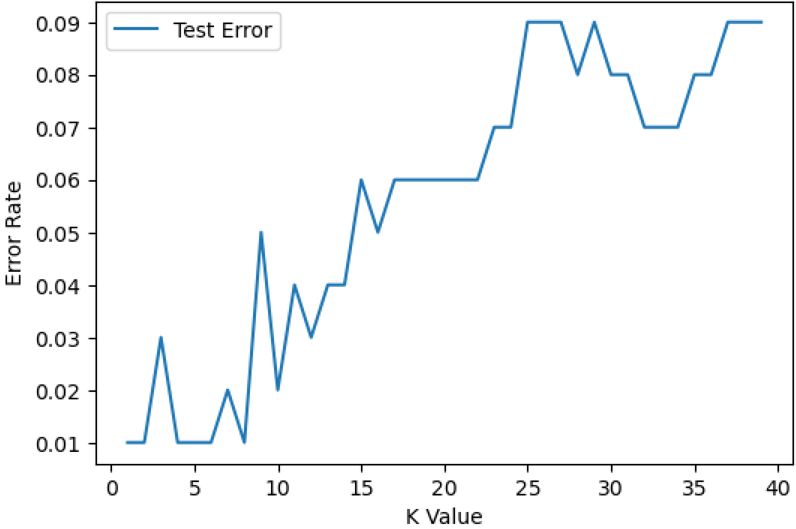
Table 2: Comparison of Models for Heart and Kidney Disease

	Heart Disease			Kidney Disease		
	LR	KNN = 10	Tree	LR	KNN = 5	Tree
X_{train}	242	242	242	300	300	300
X_{test}	61	61	61	100	100	100
Y_{train}	242	242	242	300	300	300
Y_{test}	61	61	61	100	100	100
1	131	131	131	250	250	250
0	111	111	111	150	150	150
Acc	82.20%	88.52%	81.9%	98.00%	89.90%	99.90%
TS %	20%	20%	20%	25%	25%	25%

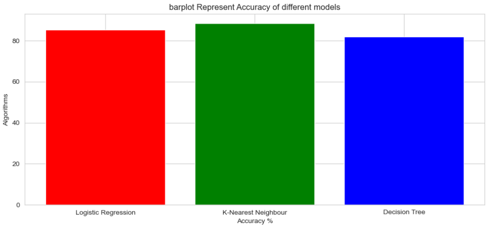
Motivation for picking K in Heart dataset

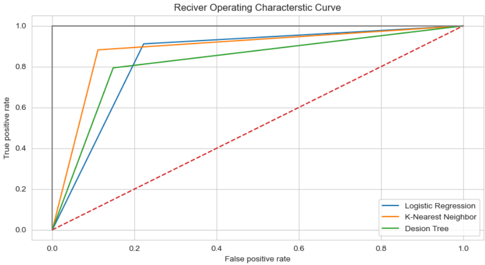


Motivation for picking K in CKD dataset

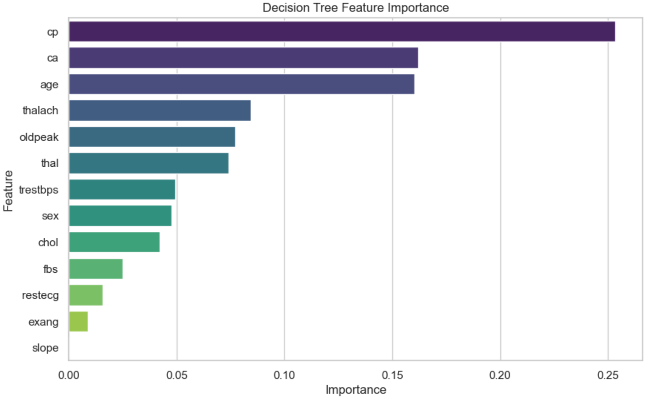


PERFORMANCE OF LOGISTIC REGRESSION, KNN AND TREE MODEL ON HEART DISEASE DATASET


barplot Represent Accuracy of different models

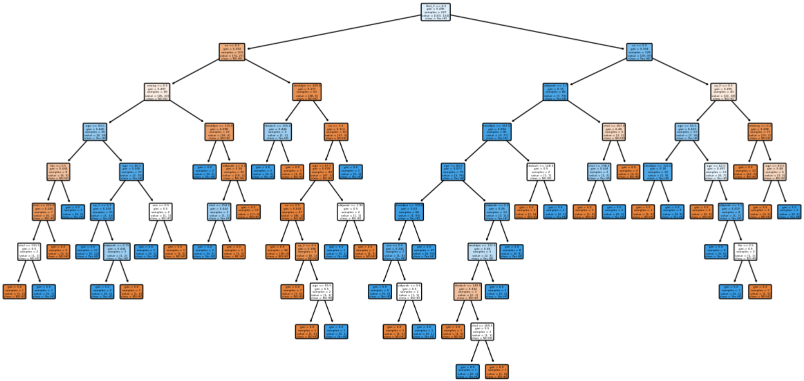

Reciver Operating Characterstic Curve

Decision Tree Feature Importance for Heart disease

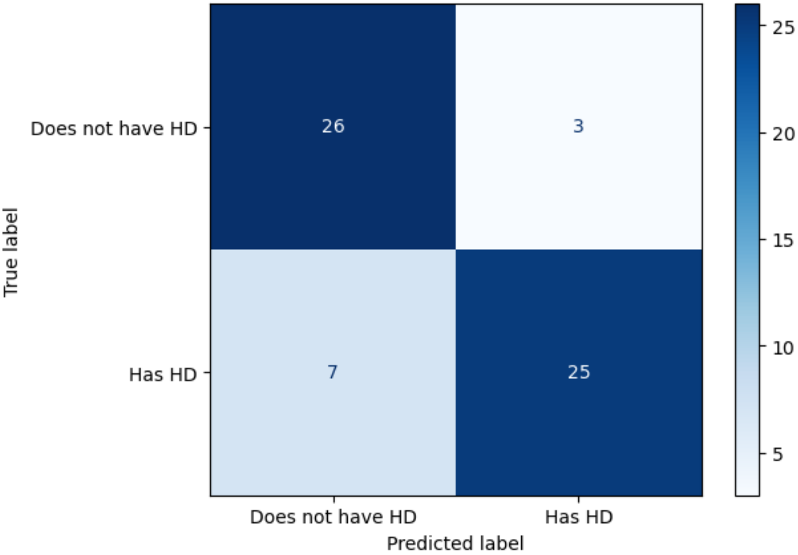


- KNN gives the best Accuracy compared to other models.
- CP, CA and age are major symptoms of heart attack.

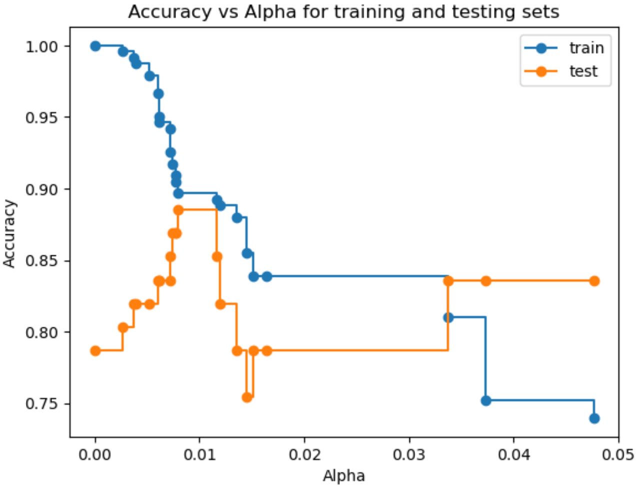
Preliminary Classification Tree



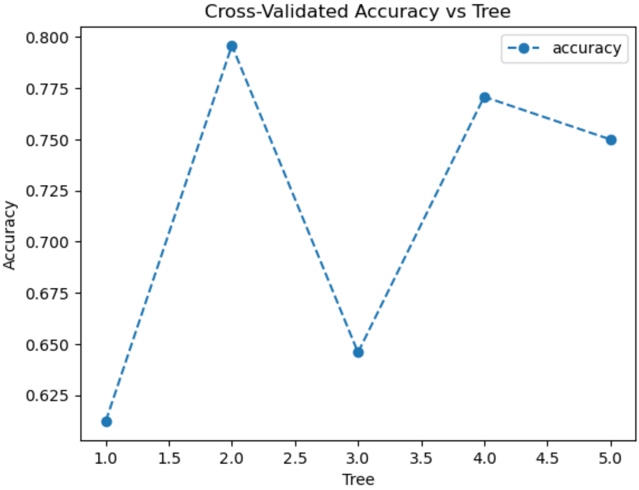
Accuracy of the trees using the Training Dataset and the Testing Dataset as function of Alpha.



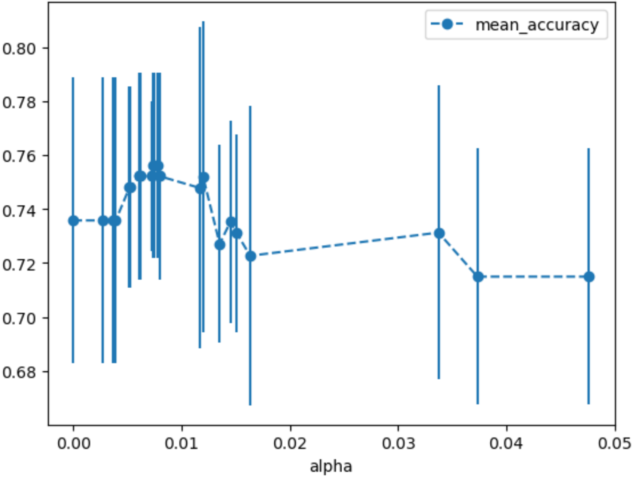
Confusion matrix: We are trying to see how it performs on the Training Dataset by running the training dataset down the tree



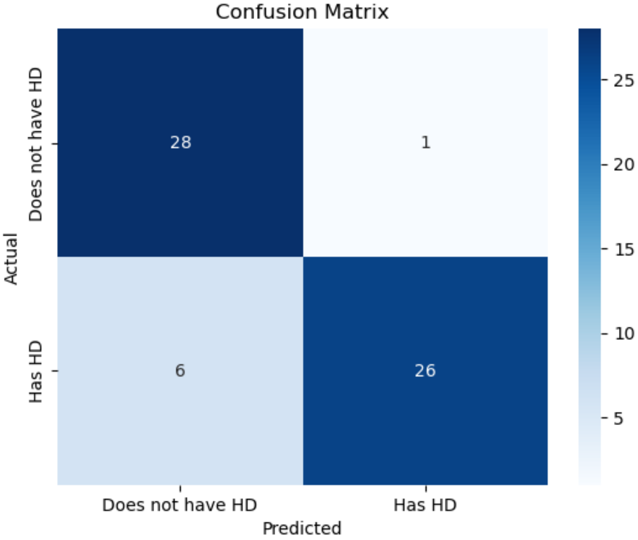
Creating the tree with alpha = 0.035



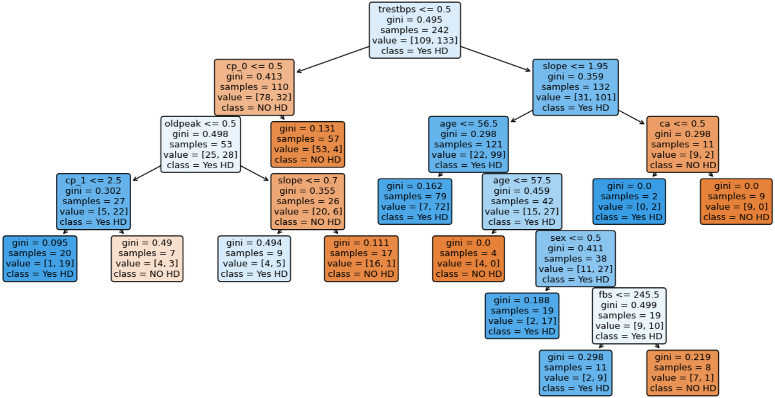
Looping each candidate of alpha to see it’s performance



Plotting the confusion matrix after using the new alpha = 0.03374655647382921



Pruned$_c$*lassificationtree*



trestbps <= 0.5
gini = 0.495
samples = 242
value = [109, 133]
class = Yes HD

cp_0 <= 0.5
gini = 0.413
samples = 110
value = [78, 32]
class = NO HD

slope <= 1.95
gini = 0.359
samples = 132
value = [31, 101]
class = Yes HD

oldpeak <= 0.5
gini = 0.498
samples = 53
value = [25, 28]
class = Yes HD

gini = 0.131
samples = 57
value = [53, 4]
class = NO HD

age <= 56.5
gini = 0.298
samples = 121
value = [22, 99]
class = Yes HD

ca <= 0.5
gini = 0.298
samples = 11
value = [9, 2]
class = NO HD

cp_1 <= 2.5
gini = 0.302
samples = 27
value = [5, 22]
class = Yes HD

slope <= 0.7
gini = 0.355
samples = 26
value = [20, 6]
class = NO HD

gini = 0.162
samples = 79
value = [7, 72]
class = Yes HD

age <= 57.5
gini = 0.459
samples = 42
value = [15, 27]
class = Yes HD

gini = 0.0
samples = 2
value = [0, 2]
class = Yes HD

gini = 0.0
samples = 9
value = [9, 0]
class = NO HD

gini = 0.095
samples = 20
value = [1, 19]
class = Yes HD

gini = 0.49
samples = 7
value = [4, 3]
class = NO HD

gini = 0.494
samples = 9
value = [4, 5]
class = Yes HD

gini = 0.111
samples = 17
value = [16, 1]
class = NO HD

gini = 0.0
samples = 4
value = [4, 0]
class = NO HD

sex <= 0.5
gini = 0.411
samples = 38
value = [11, 27]
class = Yes HD

gini = 0.188
samples = 19
value = [2, 17]
class = Yes HD

fbs <= 245.5
gini = 0.499
samples = 19
value = [9, 10]
class = Yes HD

gini = 0.298
samples = 11
value = [2, 9]
class = Yes HD

gini = 0.219
samples = 8
value = [7, 1]
class = NO HD

1. Introduction
2. Objectives
3. Methodology
4. Famous script D$_n$
5. Some variables used
6. Analysis
7. Conclusion

CONCLUSION

To summarize, within the context of heart disease, factors such as Chest Pain Type (CP), the number of major vessels colored by Fluoroscopy (Ca), and age are identified as significant contributors by the model.

of the $28 + 1 = 29$ people that did not have heart disease, 28 (96%) were correctly classified. This is an improvement. Of the $26 + 6 = 32$ people with heart disease, 26(81%) were correctly classfied.

Tree is one of the best machine learning model for classification.

References

- Adebayo, Ezekiel and Fokoue, Ernest (2019). An Empirical Demonstration of the No Free Lunch Theorem. Mathematics Applications, 10.13164/ma.2019.11, volume 8, pages 173- 188.
- Pei, Eddie and Fokoue, Ernest (2022). On Some Similarities and Differences Between Deep Neural Networks and Kernel Learning Machines. 10.13164/ma.2022.07, volume 11, pages 75 - 106.

Thanks!