AI IN HEALTH TECH

Build A Heart Attack Predictor

- By Shivam Kumar Giri

Heart Attack Predictor

PROBLEM STATEMENT & DATA INFORMATION

Today we will address one burning area in healthcare. An estimated 17 million people die of CVDs (Cardiovascular disease), particularly heart attacks and strokes, in the world every year. Cardiac ailments killed more Indians in 2016 (28%) than any other non-communicable disease, said a new study published in the September 2018 issue of health journal, The Lancet. These are double the numbers reported in 1990 when heart disease caused 15% of deaths in India.

Build a heart attack predictor based on past diagnostic data of patients using a machine learning model such as Logistic Regression or Decision Tree.

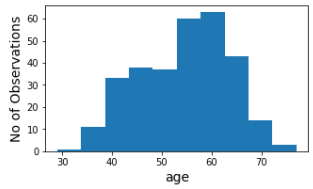
- ✓ **age** : age of the person under observation
- ✓ **sex** : sex of the person under observation
- ✓ **cp** : chest pain.
- ✓ **trestbps** : resting blood pressure in mm Hg
- ✓ **chol** : serum cholesterol in milligram per dl.
- ✓ **fbs** : fasting blood sugar. The value is 1 if yes. 0 otherwise.
- ✓ **restecg** : 1 if this fasting blood sugar is greater than 120 milligram per dl
- ✓ **thalach** : electrocardiographic results, 0 if the result is normal. 1 if ST-T wave abnormality is noticed. 2 if the result shows probable or definite left ventricular hypertrophy by Estes' criteria.
- ✓ **exang** : exercise induced angina. The value is 1 if yes. 0 otherwise.
- ✓ **oldpeak** : ST depression induced by exercise relative to rest'.
- ✓ **slope** : the slope of the peak exercise ST segment. The value is 0 if up-sloping is noticed. 1 if flat. And 2 if down-sloping is noticed
- ✓ **ca** : number of major vessels. The value ranges between 0 and 4
- ✓ **thal** : Thallium Stress Test. The values are 1 if normal, 2 if fixed defect and 3 for reversible defect.
- ➢ **target** : diagnosis of heart disease (angiographic disease status). The value of this variable is 0 if diameter narrowing is less than 50% and 1 if diameter narrowing is greater than 50%

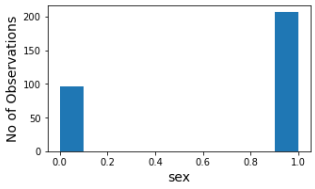
We have **303 entries** for number of people in observation.

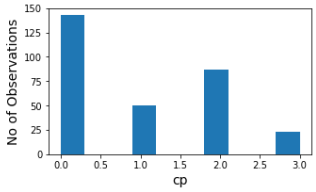
Heart Attack Predictor

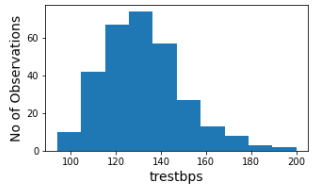
METHODOLOGY

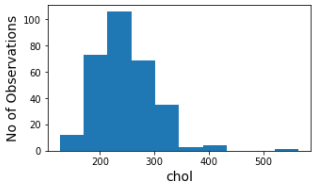
Basic Explanatory Analysis based on various factors of dependence

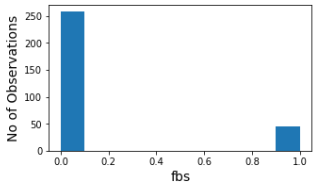


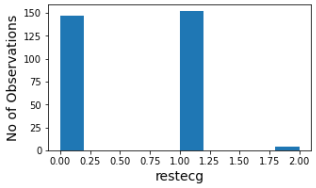


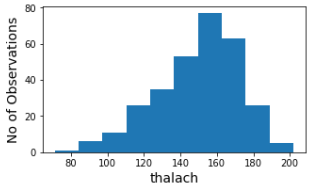


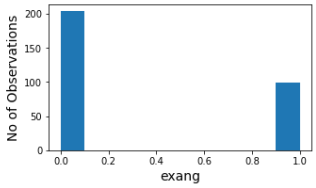


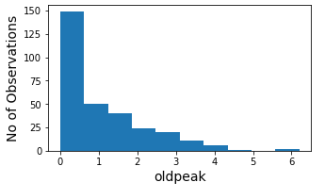


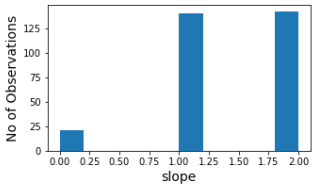


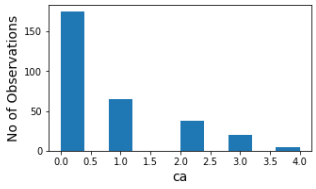


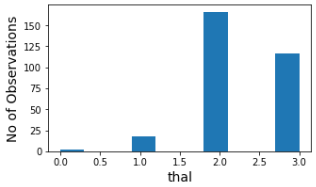




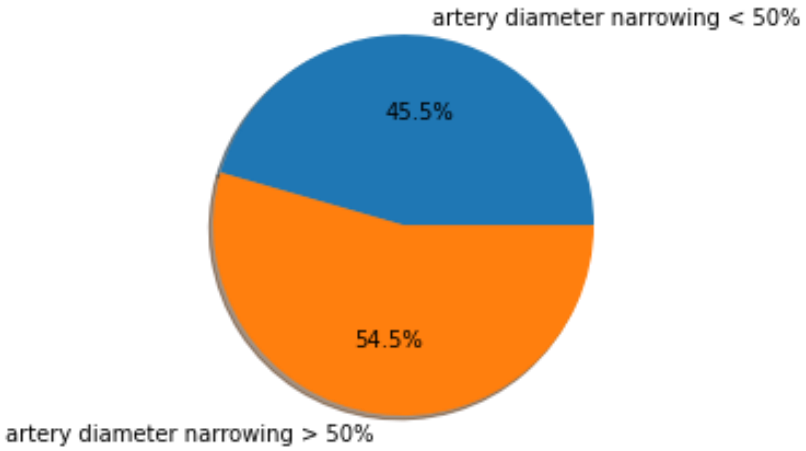








Also the target ratio is given below there are 54.3% people with artery diameter narrowing < 50% amounting to 8022 customers:



Heart Attack Predictor

DECISION TREE WITH VALIDATION OF DATA

I have basically used 3 methods of classification over two the major factors of optimization functions: Gini and entropy:

- o Train-test spilt method
- o K-fold Method
- o Shuffle-split Method

Gini measurement is the probability of a random sample being classified incorrectly if we randomly pick a label according to the distribution in a branch. Entropy is a measurement of information (or rather lack thereof), which calculate the information gain by making a split.

We used both entropy and Gini decision tree and tested their accuracy and the confusion matrix.

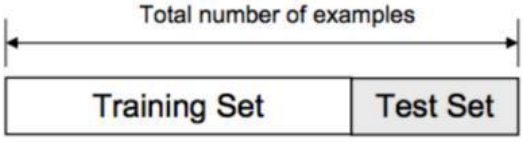
Confusion Matrix

n = 165	Predicted: No	Predicted: Yes	
Actual: No	Tn =50	FP=10	60
Actual: Yes	Fn=5	Tp=100	105
	55	110	

Suppose we have total 165 entries with Actual v/s Predicted Entries. Principle Diagonal Elements represent the data accurately predicted where others represent data misclassified.
Hypothetical Ideal Confusion Matrix has Accuracies 100%:
[[100,0]
[0,65]] //Ideal Confusion Matrix with wrongly predicted value 0;

TRAIN TEST SPLIT METHOD OF VALIDATION OF DATA

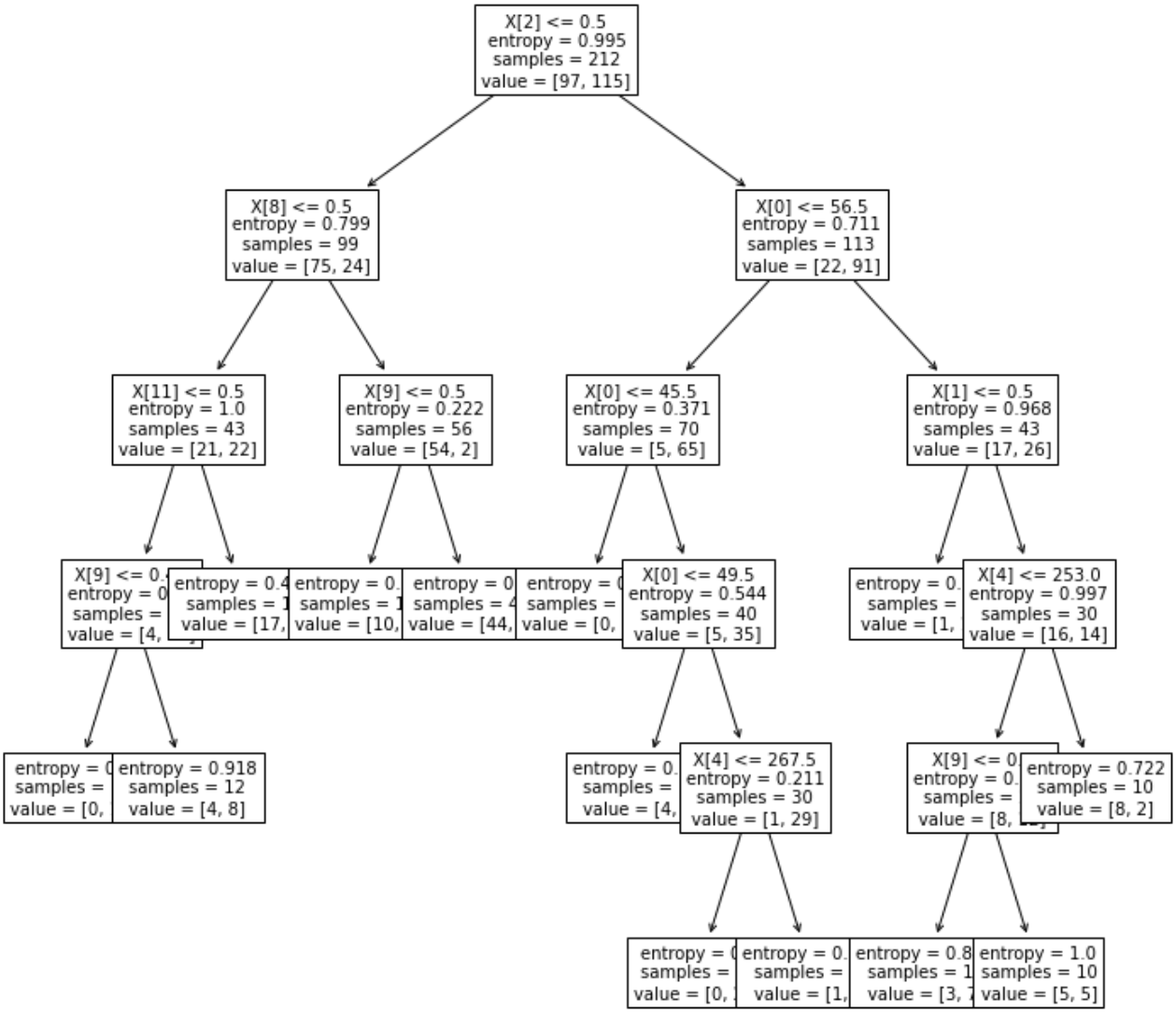
Here Data is splited into set of training data and test data, the training data is trained by decision tree while test data is validated for accuracy.
Here We used 70% of data for training ad rest 30% for testing, which is recommended ratio of splitting.



Heart Attack Predictor

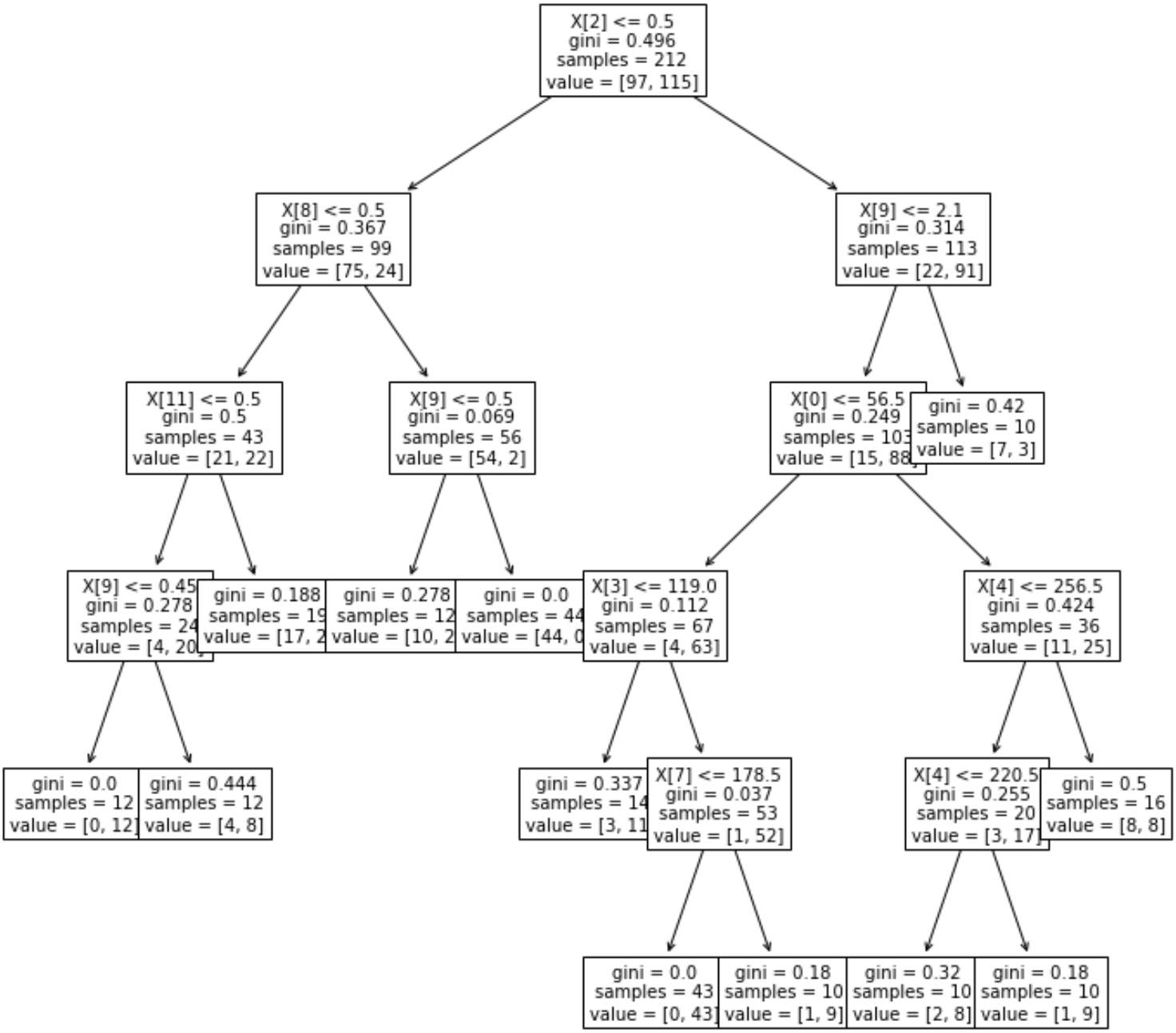
We used both entropy and Gini decision tree and tested their accuracy and the confusion matrix.

Entropy decision tree



Heart Attack Predictor

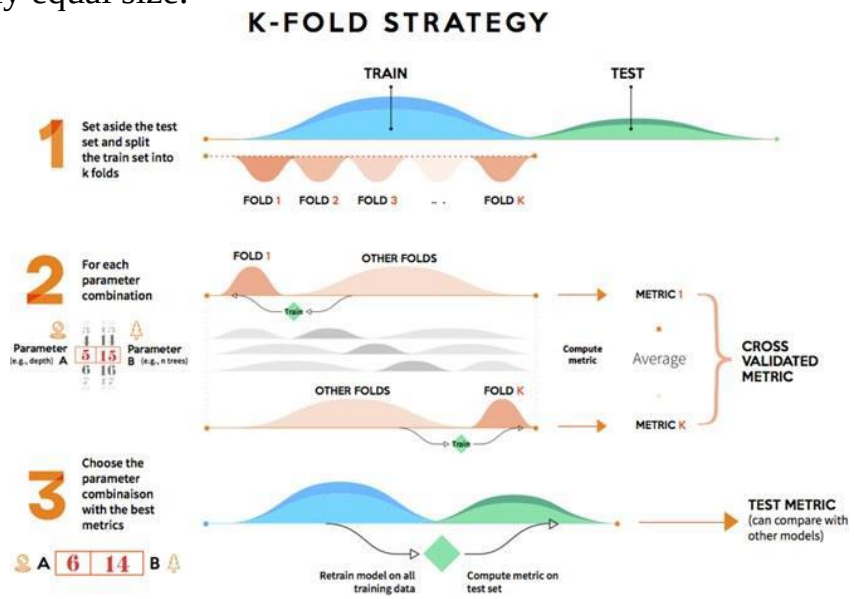
Gini Decision Tree



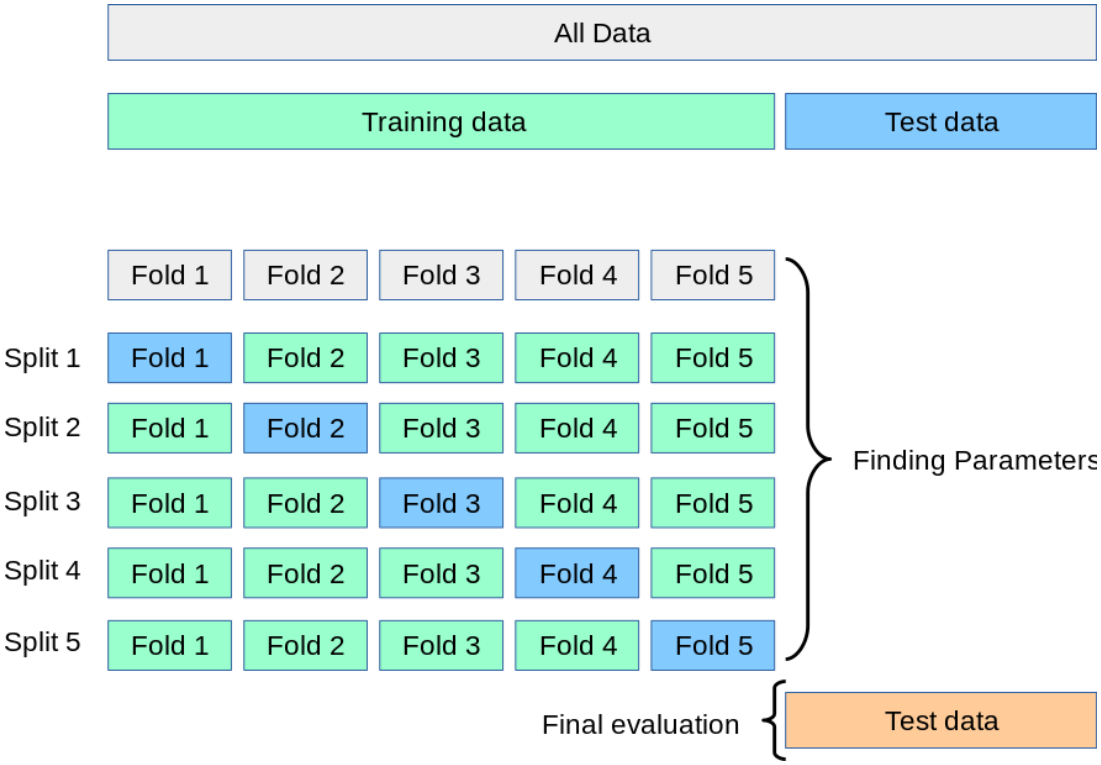
Heart Attack Predictor

K-FOLD METHOD OF VALIDATION OF DATA

This technique involves randomly dividing the dataset into k groups or folds of approximately equal size.



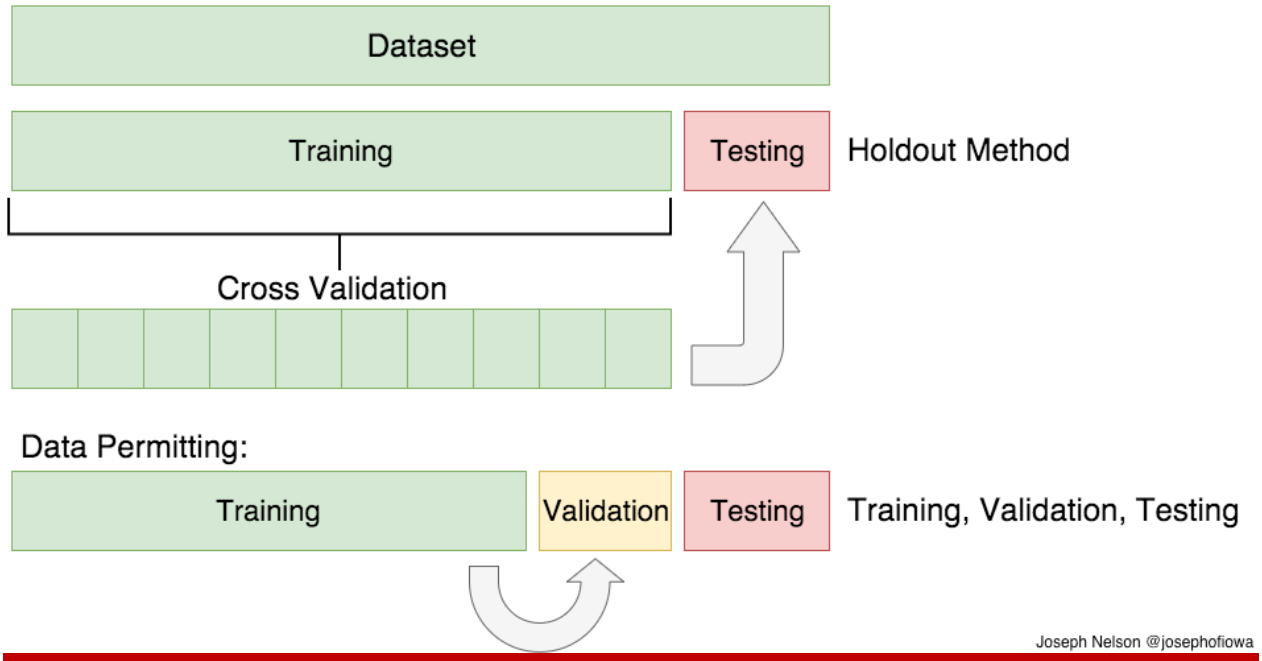
Each fold is then used once as a validation while the k 1 remaining folds form the training set. Here we use k=4 ideal for dataset having 30% testing data as per the formula: K=N/(N*(0.3)) =0.3333, hence k=4, give maximum efficiency.
We used both entropy and Gini decision tree and tested their accuracy and the confusion matrix.



Heart Attack Predictor

SHUFFLE-SPLIT METHOD OF VALIDATION OF DATA

Shuffle-Split is a Random permutation cross-validator. It Yields indices to split data into training and test sets. We used both entropy and Gini decision tree and tested their accuracy and the confusion matrix.



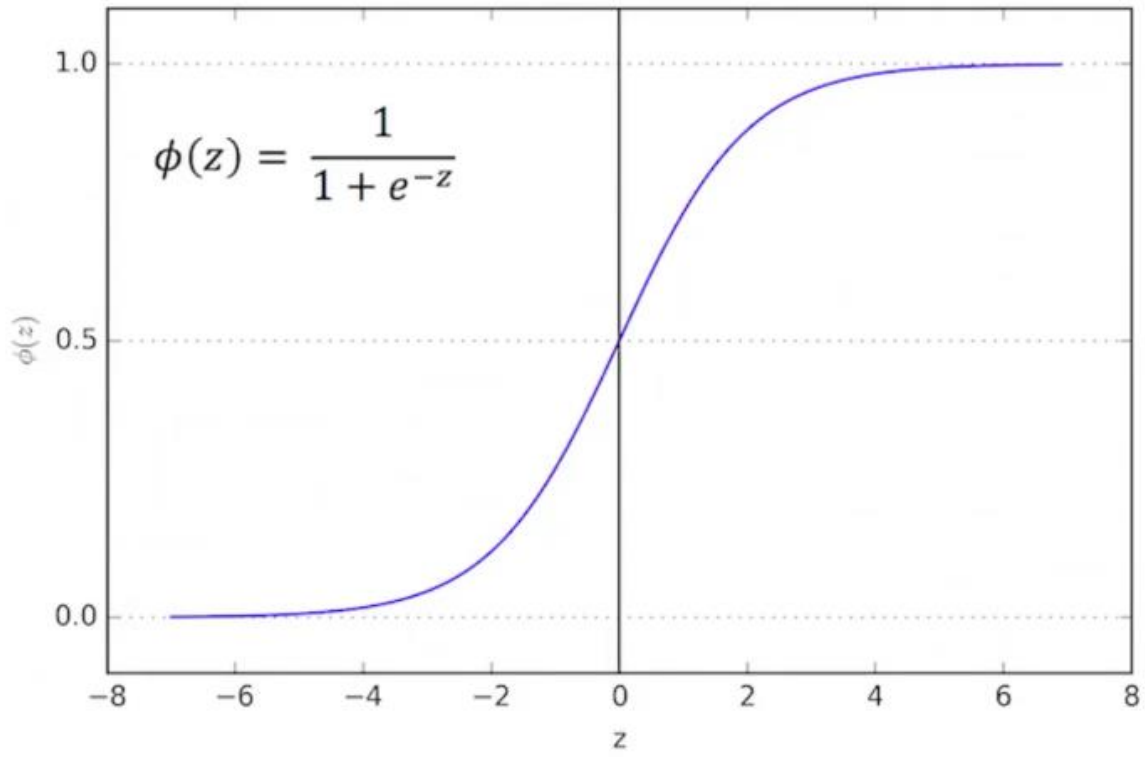
LOGISTIC REGRESSION

Logistic Regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

We had use 2 type of regularization technique:

1. **Ridge Regression:**
   - Performs L2 regularization, i.e. adds penalty equivalent to square of the magnitude of coefficients
   - Minimization objective = LS Obj + α * (sum of square of coefficients)
2. **Lasso Regression:**
   - Performs L1 regularization, i.e. adds penalty equivalent to absolute value of the magnitude of coefficients
   - Minimization objective = LS Obj + α * (sum of absolute value of coefficients)

Heart Attack Predictor



OBSERVATION

Accuracy by various Validation Method of decision tree

Validation Methods	Entropy Method	Entropy Confusion Matrix	Gini Method	Gini confusion Matrix
Train-test split:	73.63	[[28 13] [11 39]]	67.03	[[29 12] [18 32]]
Kfold(4-fold):	78.95 +/- 0.075	[[21 7] [13 34]]	78.95 +/- 0.075	[[21 7] [13 34]]
Shuffle_split:	81.13 +/- 0.048	[[21 20] [13 37]]	81.60 +/- 0.036	[[15 26] [11 39]]

Accuracy by various Penalty Method of Logistic Regression

Penalty Methods	Accuracy	Confusion Matrix
Lasso Penalty :	0.736264	[[29 12] [12 38]]
Ridge Penalty :	0.78022	[[29 12] [8 42]]

Heart Attack Predictor

CONCLUSION

I have successfully predicted the risk of heart attack using decision tree and logistic regression. Logistic regression **with L2 penalty gave highest accuracy of 78.02%.** Decision tree with **Shuffle split under Gini Criteria has got maximum Accuracy 81.60% with standard deviation of 0.036.**

Hence the heart-attack predictor with Decision tree and Logistic Regression is completed and hence the decision tree with Shuffle-Split under Gini optimization function is highest performer.