CART | Random Forest | Artificial Neural Network

Mortality Outcomes for Females Suffering Myocardial Infarction.

This is a Dataset of females having coronary heart disease (CHD. To predict whether the female is dead or alive so as to discover important factors that should be considered crucial in the treatment of the disease.

1. Data Ingestion and EDA:

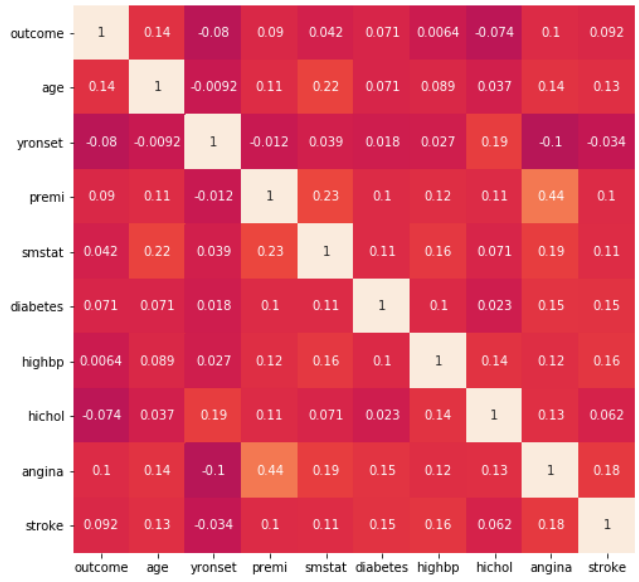
We have 10 features and 1295 records in our dataset. **Age** is a continuous variable ranging from 35 to 69 years and median age being 63; we will be converting age into a categorical variable. **Outcome** is a categorical variable with mortality levels - live and dead. **Yronset** is the year of onset of symptoms with 9 unique values, and it would be treated as a categorical variable.

The remaining **7** variables are categorical variables.

There are no null values. But if we check the unique values, we se e that features- 'diabetes', 'highbp' , 'angina', 'stroke' , 'smstat' , 'p remi' have 'nk' as a unique value, which stands for not known. Ca nnot feed 'Not known' in the data, so will drop the rows where an y of the features have 'nk' value.

New shape of the dataset now is (1124, 10)

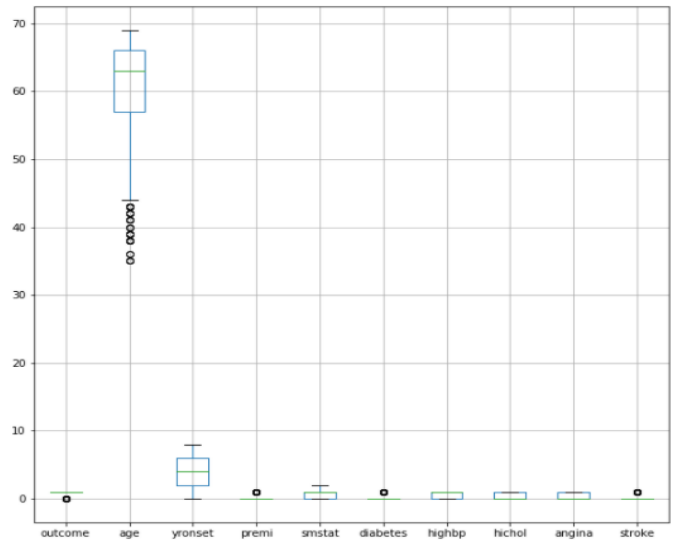
Correlation between variables



If we check correlation of features with 'outcome', yronset seems to have a negative correlation, i.e. the earlier the symptoms set in, the higher chances of chances = dead. Rest all features have positive correlation, with age having the highest correlation of 0.14.

2. Encode the data (having string values) for Modelling. Data Split: Split the data into test and train, build classification model CART, Random Forest, Artificial Neural Network.

Checking for outliers in the dataset, 'age' has multiple outliers.



Age has outliers, but we will be taking age as a categorical variable which has 34 unique values.

Encoding data:

To encode the categorical data, we will be using Integer encoding, where each value in the column, will be mapped to an integer. We will be encoding all columns, below is how our dataset looks after integer encoding.

	outcome	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
0	0	27	0	0	2	0	1	1	0	0
1	0	19	0	0	0	0	1	1	0	0
3	0	28	0	0	2	0	1	0	1	0
6	0	27	0	0	1	0	1	0	0	0
7	1	32	0	1	1	0	1	1	1	1

Splitting the data:

Next we have split the data into training and testing dataset, with test size = 30%. Shape of training data = (786, 10) and shape of testing data = (338, 10). We have also split the train and test data into target and independent variables, with 'outcome' being the target variable, marked as **X** and combining the remaining independent variables in a dataset , which we will be calling as **y**

We will be feeding the training data into our model to make the machine learn and recognize the patterns in the data and will be using the testing data to monitor the accuracy of the model prediction.

CART model:

CART is supervised tree –structured graphical representation of getting all possible outcomes to a problem based on conditions. We have features as the internal nodes, branches representing the decision rules and the outcomes are represented by leaf nodes.

We have built a Decision Tree Classifier model with the specified parameters.

criterion = 'gini', max_depth = 30, max_features = 4 , max_leaf_nodes = 50,min_samples_leaf = 3, min_samples_split = 10, random_state = 1

Random Forest Model:

Random forest is an ensemble technique, which uses bootstrap aggregating. RF model builds multiple decision trees and takes a vote of the prediction.

Following parameters are used to build the RF model.

n_estimators = 20 , max_features = 8, max_samples = 50, random_state =1

Artificial Neural Network:

ANN is a computation system which is inspired by and aims to mimics the human brain. It consists of input, output and hidden layers and the output is predicted based on the units/ neurons interconnecting in some pattern to allow communication. We will standardize the data before feeding into our model. **We will be using z score to standardize our data before we feed into the ANN model**

	age	yronset	premi	smstat	diabetes	highbp	hichol	angina	stroke
0	0.997624	-0.729250	-0.554816	0.145332	-0.519413	0.729496	-0.861855	1.300685	-0.373002
1	0.720497	0.459120	1.802401	1.505226	-0.519413	0.729496	-0.861855	1.300685	2.680951
2	-0.803705	-1.521497	-0.554816	0.145332	-0.519413	0.729496	-0.861855	-0.768826	-0.373002
3	0.166241	-0.333127	1.802401	1.505226	-0.519413	0.729496	-0.861855	1.300685	-0.373002
4	0.997624	0.062997	-0.554816	0.145332	-0.519413	0.729496	1.160288	1.300685	-0.373002

Following parameters are used to build the ANN model.

1. **tol = 0.00001**
   This the tolerance or threshold, the lower the tolerance better the accuracy as the model will run till the time the difference of error is at tol
2. **max_iter = 100**
   This is the maximum number of times the model will run to minimize the loss function

3. Performance Metrics- Accuracy, Recall, Confusion Matrix, ROC_AUC Curve

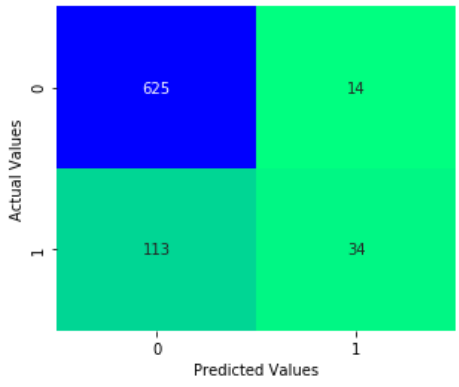
CART

Training data performance metrics

Confusion Matrix is a tabular summary of the number of correct and incorrect predictions made by a model. In this model we are predicting how many patients are dead or alive. In business sense, the model which would be able to identify the dead patients would be successful, as we will be able to understand the impact of features on mortality. Also in future, if we get patients with similar features, we will be able to raise an alarm and monitor and control situation in a better manner.

Hence in our dataset positive i.e. dead =1 and negative, i.e. live= 0

Confusion Matrix

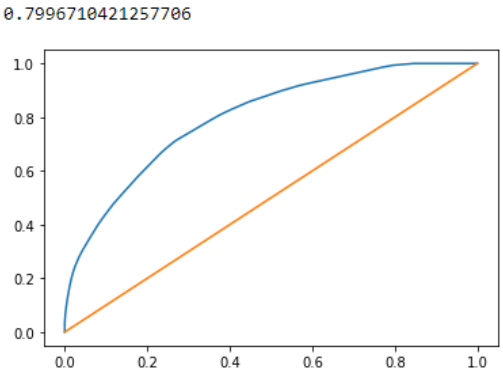


- True Negative, i.e. number of patients model predicted are alive and are actually alive = **625**
- True Positive, i.e. number of patients model predicted are dead and are actually dead = **34**
- False Positive, i.e. number of patients model predicted are alive, but are actually dead = **113**
- False Negative, i.e. number of patients model predicted are dead, but are actually alive = **14**

Accuracy = **0.8384** (TN+ TP/TN+TP+FN+FP)

ROC_AUC Curve:

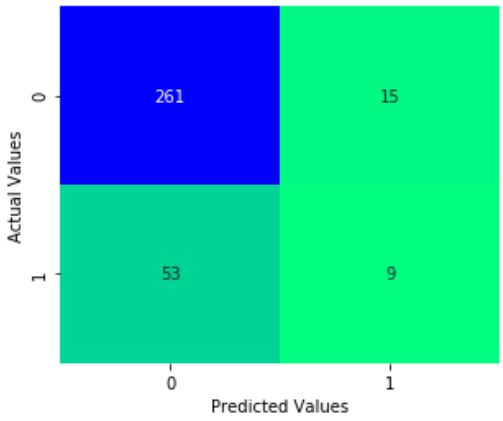
ROC_AUC score describes the ability of the model to differentiate between the two classes. ROC_AUC curve is the graph between true positive rate (tpr) and false positive rate (fpr) , and ROC_AUC score is area under the ROC_AUC curve.



ROC_AUC Score = 0.7996

Testing data performance metrics

Confusion Matrix

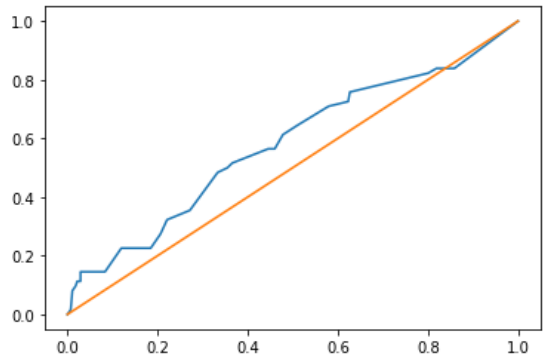


- True Negative, i.e. number of patients model predicted are alive and are actually alive = **261**
- True Positive, i.e. number of patients model predicted are dead and are actually dead = **9**
- False Positive, i.e. number of patients model predicted are alive, but are actually dead = **53**
- False Negative, i.e. number of patients model predicted are dead, but are actually alive = **15**

Accuracy for test data = 0.80
Recall for test data = 0.15

ROC_AUC curve:

0.5791841982234689



ROC_AUC score = 0.5791

Feature Importance for CART model

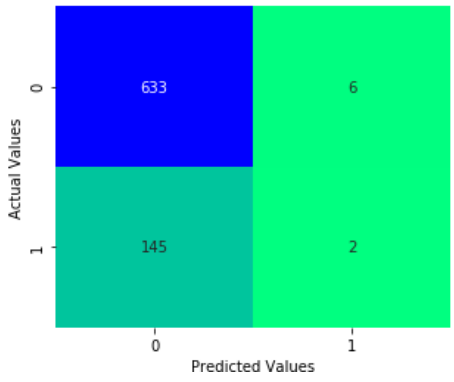
```
age         0.404845
yronset     0.138196
premi       0.026824
smstat      0.157665
diabetes    0.024321
highbp      0.055877
hichol      0.085866
angina      0.038449
stroke      0.067957
Name: 0, dtype: float64
```

The most important feature used for prediction is age, followed by yronset.

Random Forest

Training data performance metrics

Confusion Matrix:

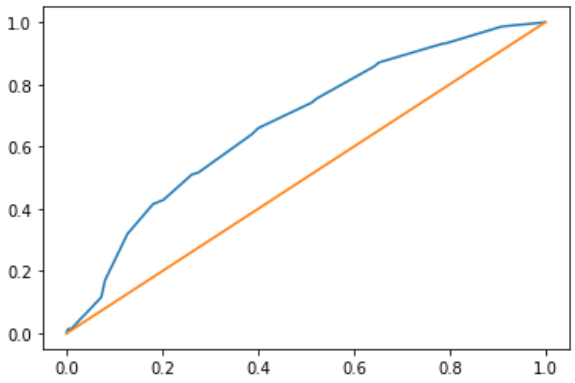


- True Negative, i.e. number of patients model predicted are alive and are actually alive = **623**
- True Positive, i.e. number of patients model predicted are dead and are actually dead = **2**
- False Positive, i.e. number of patients model predicted are alive, but are actually dead = **145**
- False Negative, i.e. number of patients model predicted are dead, but are actually alive = **6**

Accuracy for training set = 0.8078

ROC_AUC curve and score

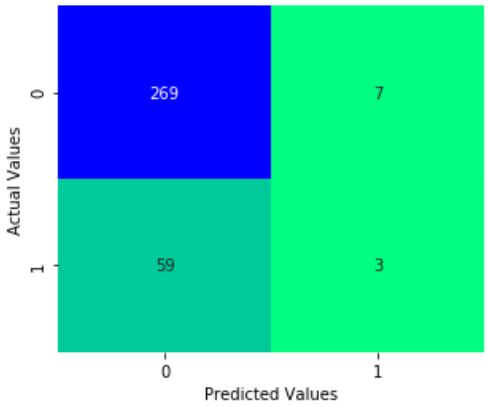
0.6739165149627926



ROC_AUC score = 0.6739

Testing data performance metrics

Confusion Matrix:



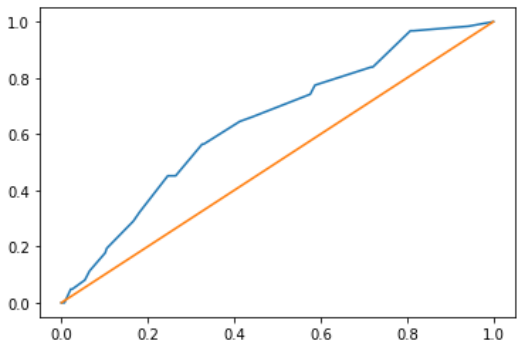
- True Negative, i.e. number of patients model predicted are alive and are actually alive = **269**
- True Positive, i.e. number of patients model predicted are dead and are actually dead = **3**
- False Positive, i.e. number of patients model predicted are alive, but are actually dead = **59**
- False Negative, i.e. number of patients model predicted are dead, but are actually alive = **7**

Accuracy of test data = 0.8047

Recall of test date = 0.05

ROC_AUC curve:

0.6421517064048621



ROC_AUC score for test data = 0.6421

Feature importance for Random Forest

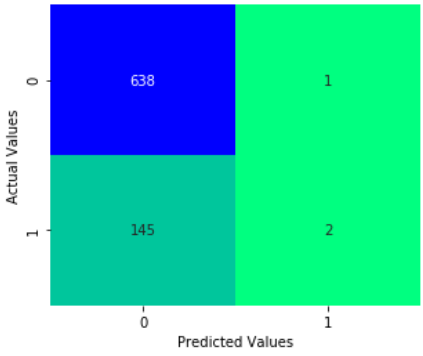
```
age          0.287699
yronset      0.277685
premi        0.051165
smstat       0.088912
diabetes     0.049316
highbp       0.051003
hichol       0.066027
angina       0.060309
stroke       0.067885
Name: 0, dtype: float64
```

In random forest, age is considered as the most importance feature, closely followed by yronset.

ANN

Training data performance metrics

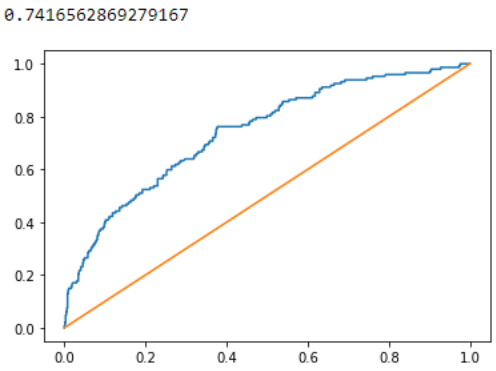
Confusion Matrix



- True Negative, i.e. number of patients model predicted are alive and are actually alive = **638**
- True Positive, i.e. number of patients model predicted are dead and are actually dead = **2**
- False Positive, i.e. number of patients model predicted are alive, but are actually dead = **145**
- False Negative, i.e. number of patients model predicted are dead, but are actually alive = **1**

Accuracy of train data = 0.8142

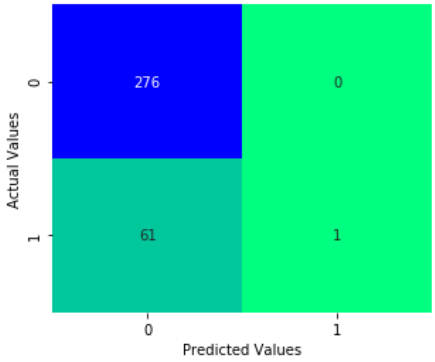
ROC_AUC curve and score



ROC_AUC score = 0.7416

Testing data performance metrics

Confusion Matrix:

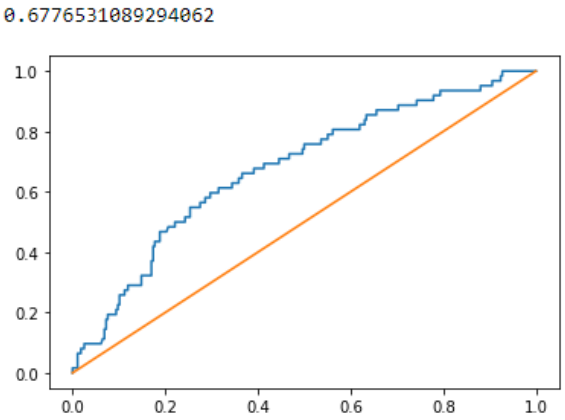


- True Negative, i.e. number of patients model predicted are alive and are actually alive = **276**
- True Positive, i.e. number of patients model predicted are dead and are actually dead = **1**
- False Positive, i.e. number of patients model predicted are alive, but are actually dead = **61**
- False Negative, i.e. number of patients model predicted are dead, but are actually alive = **0**

Accuracy of test data = 0.819

Recall of test data = 0.02

ROC_AUC curve



ROC_AUC score = 0.6776

4 Final Model: Compare all the models and write an inference which model is best/optimized.

	CART	RF	ANN
Accuracy_Train	83.00%	81.00%	81.00%
Accuracy_Test	80.00%	80.00%	81.00%
Recall_Train	23%	1%	1%
Recall_Test	15%	5%	2%

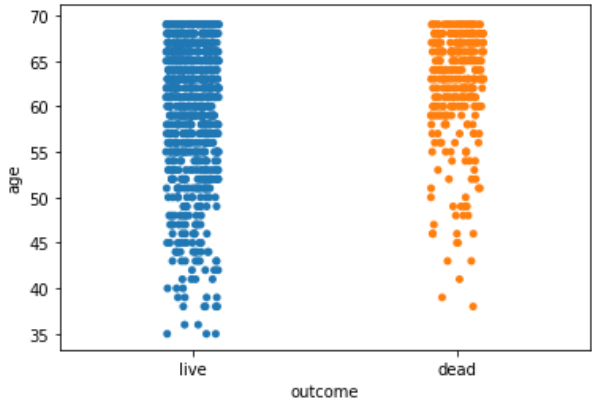
We are more interested in predicting whether the patient is dead or not, than in predicting if the patient is alive. Also our target variable is imbalanced hence accuracy will not be the preferred metrics

The accuracy is high, because it is majorly being able to predict the negative class, i.e. outcome = live.

For this dataset, we need to go with the model which has the highest True positive rate, i.e. we will choose the model with the **highest recall** for outcome= 1, i.e. dead. CART has the highest recall for outcome= dead, so we will go ahead with the CART model.

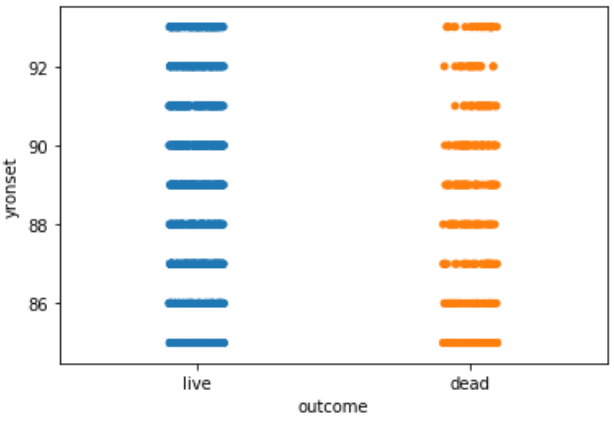
5 Inference: Basis on these predictions, what are the insights and recommendations?

Basis our model, it is clear that age has the highest feature importance. We also saw earlier that age has the highest correlation with outcome.



We can see from the stripplot that outcome= live is spread equally across all ages, for outcome = dead, age has direct correlation, higher the age, higher concentration of dead.

Yronset ranks 3rd on our model feature importance.



Earlier the yronset, higher is the concentration of outcome = dead.

CART gives 15% recall, with importance features rank - age, smstat and yronset