PGP-DSE Capstone Project Final Report (Dementia Risk Prediction)

Rajdatt Vernekar Narasimha Bharadwaj C. Rajat Gupta Anukul Dixit P V P Sandeep

Problem understanding

Dementia is a general term for memory loss and decline in thinking and cognitive abilities in a person. We might have seen what is dementia and its effects especially in the older people around us.
The WHO website states that "Worldwide, around 50 million people have dementia, with nearly 60% living in low- and middle-income countries. Every year, there are nearly 10 million new cases."
Future prediction of data suggests that it might reach 82 million by the end of 2030. Here, we would also like to highlight that nearly 60-70 % of the dementia cases are with respect to Alzheimer's Disease.
It should be noted that the Dementia is not only with respect to age, people, even the young people get surrendered to this is known as young-onset dementia (age < 65 years). This can constitute around 9 % of the data.
There by, it might be useful in detecting the mild types of Dementia at earlier stages and proper care and optimal treatment be provided for their better lives.

Current solution to the problem

There is no cure for dementia yet; hence, the early identification of individuals at higher risk of developing dementia becomes critical, as this may provide a window of opportunity to adopt lifestyle changes to reduce dementia risk.
Numerous dementia risk prediction models to identify individuals at higher risk have been developed in the past decade. Three systematic reviews and meta-analyses summarizing dementia risk prediction models were published over the past years. Stephan et al. and Tang et al. mainly focused on the critique of the variables selected for inclusion and the assessment of models' prognostic performance, whereas Hou et al. reviewed published dementia risk models in terms of sensitivity, specificity, and area under the curve from receiving operating characteristic analysis.
Hou et al. recommended four risk prediction models for different populations (midlife, late-life, patients with diabetes, or mild cognitive impairment (MCI)) with acceptable predictive ability, but still concluded that the models showed methodological limitations, such as lack of external validation.

Proposed solution to the problem

We intend to do a small service to hospitals, clinicians, nursing facilities and other healthcare practitioners by modelling a machine learning classification problem with respect to type of Dementia.

It is to be noted that better services and care can be given to the people with dementia.
To detect the early transition of people from non-demented to demented nature, we are looking at different types of prediction models to find out the better model.
It is to be noted that there would be limitations to the proposed solution. They might include the following points but are not limited to:
1. Sample size of the dataset
2. Lack of clinical domain expertise in the project team
3. Scalability of the proposed solution for dementia prediction

Data and its Insights

Data Sources

The data for the Dementia risk analysis is taken from the web page Open Access Series of Imaging Studies (OASIS).

There were no other publicly available data from other authentic resources.

Understanding Data Terminology / Data Dictionary

ID: MRI ID of the subject for each MRI scan.

M/F: Gender of the subject (Male/Female)

Hand: Handedness i.e., whether the person is left-handed or right-handed (L/R)

Age: Age of the subject at time of MRI session (ranging from 18 to 96)

Educ: Education level of the subject (1: less than high school, 2: graduated high school, 3: received college level education, 4: graduated college, 5: received education beyond college.)

SES: Socio-economic status of the subject (1: lower class, 2: lower-middle, 3: middle class, 4: middle upper class, 5: upper class)

MMSE: Mini-Mental State Examination score of the subject (ranging from 0 to 30) (0 - most likely to be demented, 30 - least likely to be demented)

CDR: Clinical Dementia Rating of the subject (Target Variable) (0: non-demented/cognitively normal, 0.5: Very mild dementia, 1: mild dementia, 2: moderate dementia)

eTIV: Estimated total intracranial volume of brain (in mm^3)

nWBV: Normalized whole brain volume (in mg)

ASF: Atlas Scaling Factor i.e., the determinant of an affine transformation matrix of the brain MRI data points

Delay: The interval between the previous and the current MRI session (in days)

Data Information

Number of Attributes	12
Number of Observations	809
Duplicate Observations	Absent
Missing Values	Present
Numerical Attributes	6
Categorical Attributes	6

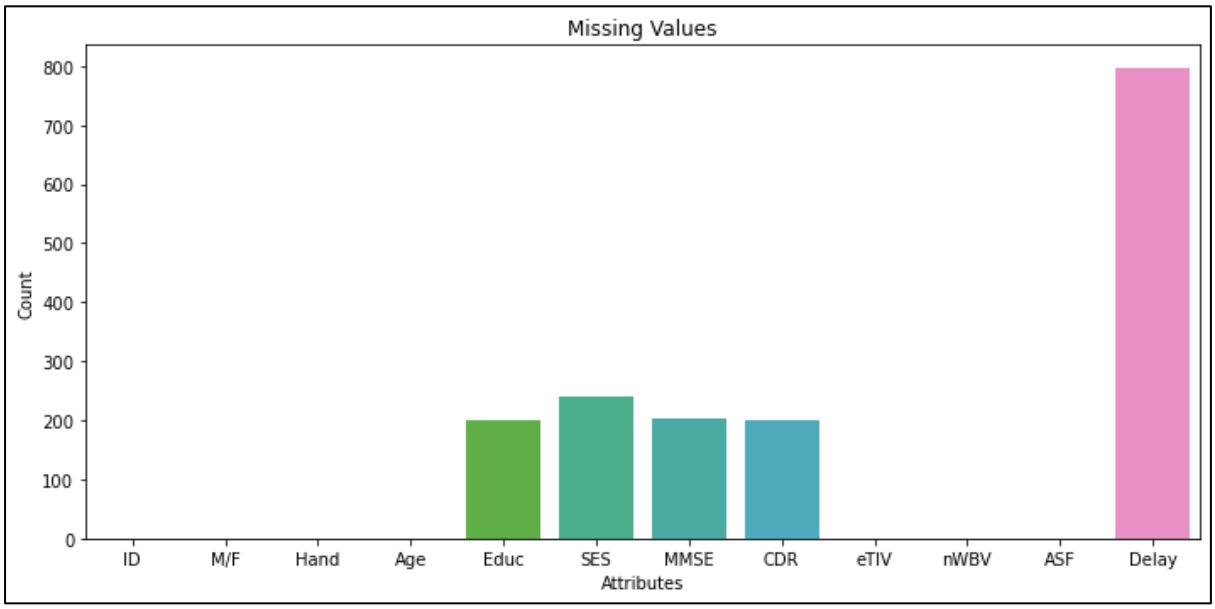
From the initial observation of the data, it was found that the *ID* column and *Hand* column were not significant as the ID column was mere identifier column and the Hand column consisted unique value 'R' as every patient in our data was right-handed.

Sl. No	Variables Name	Type	Remarks
1	M/F	Categorical	
2	Age	Numerical	
3	Educ	Numerical	Modified to categorical data.
4	SES	Numerical	Modified to categorical data.
5	MMSE	Numerical	
6	CDR	Numerical	Modified to categorical data.
7	eTIV	Numerical	
8	nWBV	Numerical	
9	ASF	Numerical	
10	Delay	Numerical	

Pre-processing of the Data

Null Value Treatment

The below plot indicates the null values present in the data.



It can be observed from the graph that more than 95 % of the data is missing for the attribute Delay, which makes the Delay column insignificant for our analysis. So, this column was also not considered for data analysis.

It can be noted that there were quite a few null values present in the column CDR (Clinical Dementia Rating). To avoid misinterpretation of the data, it was considered to drop all the rows where CDR values were missing.

By doing so, we most of the null values were treated except for the columns MMSE and SES which had two missing values and 38 null values respectively.

The SES column was imputed with mode as this was categorical data and the mode for this variable was found to be '2.0'. This was considered as the average standard of living for the data.

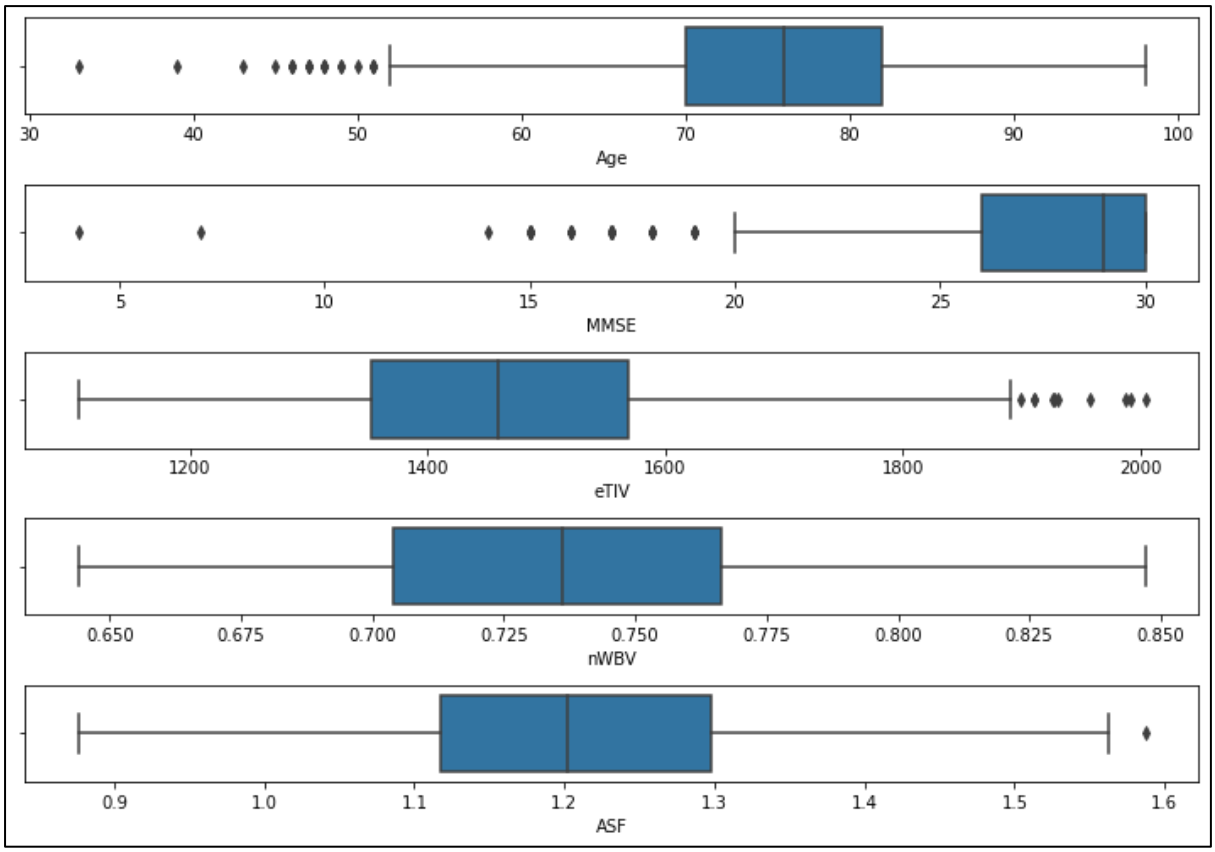
The two MMSE values which were missing was found to be of those whose CDR value was 1.0. It was decided to impute these null values with median MMSE score of CDR category 1.0. This was found to be equal to 21.

Now all the missing data is rightly treated or imputed and the complete data is sent for outlier treatment.

Uni-variate Analysis

Outliers Analysis and Treatment for numerical variables

The below plots suggest the possible outliers present in the numerical attributes of our data.



Inferences:

1. Age and MMSE are data which are left skewed which tells that more older people are present and are scoring around 29/30 in MMSE.
2. The eTIV is right skewed data.
3. ASF contains only one outlier.
4. The outliers in the MMSE column are very important as the MMSE scores tell us whether the person is more likely to be affected with dementia or not and also lower the MMSE scores, higher is the chances of the person getting affected with the disease.
5. It is to be noted that nWBV do not require any outlier treatment.

The data showed us that 70% of the patients above the upper whisker for eTIV feature suffered from the disease. Hence the outliers were clamped to the upper whisker of the boxplot.

It was also observed that 100 % of the data below 52 years of age were healthy records. Hence, the age was clamped to lower whisker value of 52.

ASF was clamped to upper whisker of boxplot.

Data Statistics

	Age	MMSE	eTIV	nWBV	ASF
count	608.000000	608.000000	608.000000	608.00000	608.000000
mean	75.379934	27.213816	1477.062500	0.73713	1.203597
std	9.371628	3.699193	170.653795	0.04267	0.135091
min	52.000000	4.000000	1106.000000	0.64400	0.876000
25%	70.000000	26.000000	1352.500000	0.70400	1.118000
50%	76.000000	29.000000	1460.000000	0.73600	1.202000
75%	82.000000	30.000000	1569.000000	0.76625	1.297500
max	98.000000	30.000000	2004.000000	0.84700	1.587000

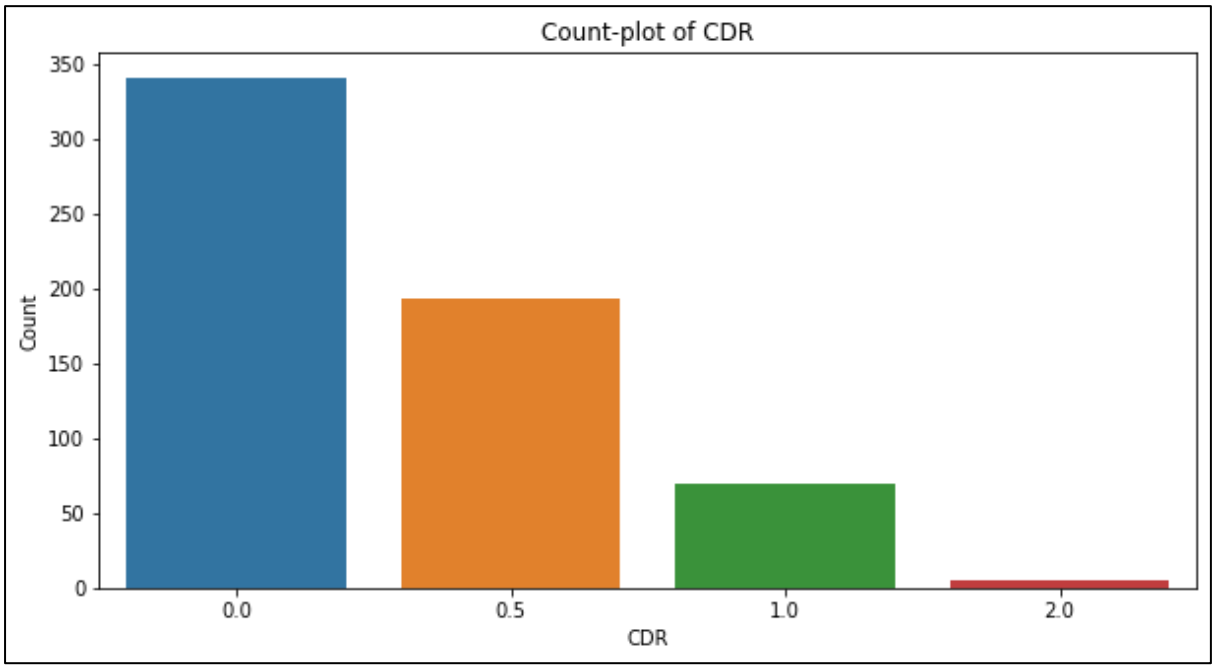
	M/F	Educ	SES	CDR
count	608	608.0	608.0	608.0
unique	2	5.0	5.0	4.0
top	F	1.0	2.0	0.0
freq	369	176.0	206.0	341.0

Inferences:

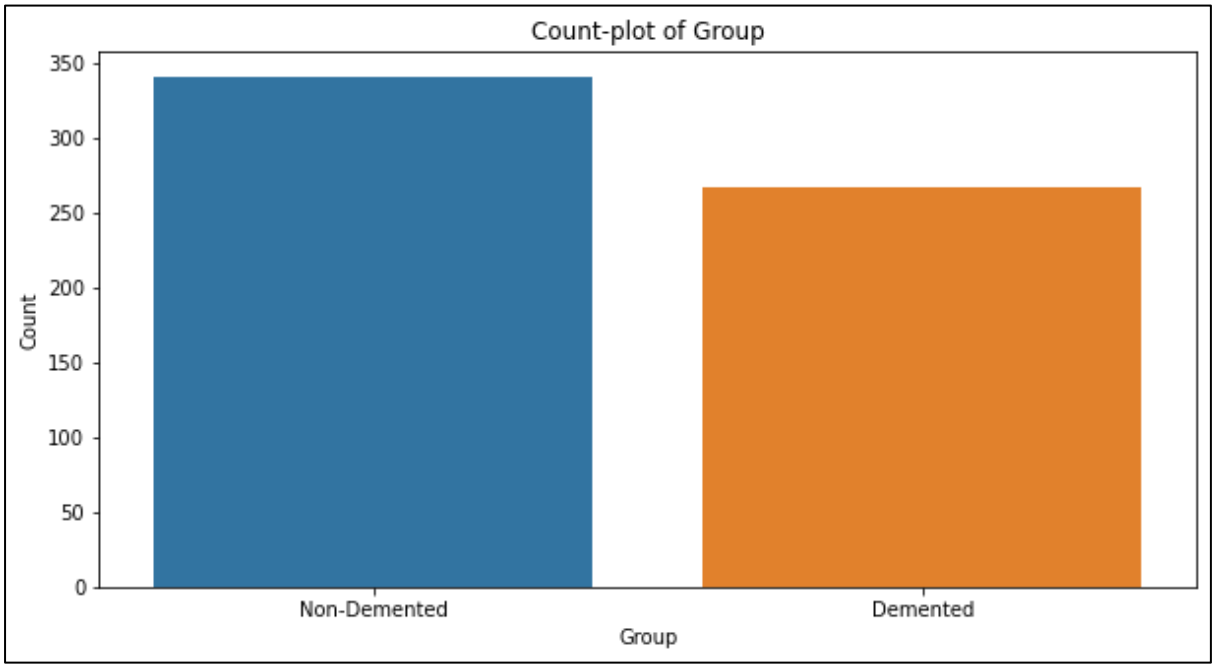
1. There are 608 unique records.
2. There are more females who have taken MRI scan than male.
3. We can see that out of 608 people, 341 people are healthy.
4. Age is ranging from 52 to 98 with an average of 76.
5. The data for MMSE is concentrated highly between 26 to 30 with median being 29.
6. There are other features which can be noted from the above tables.

Understanding the Categorical Variables

- CDR

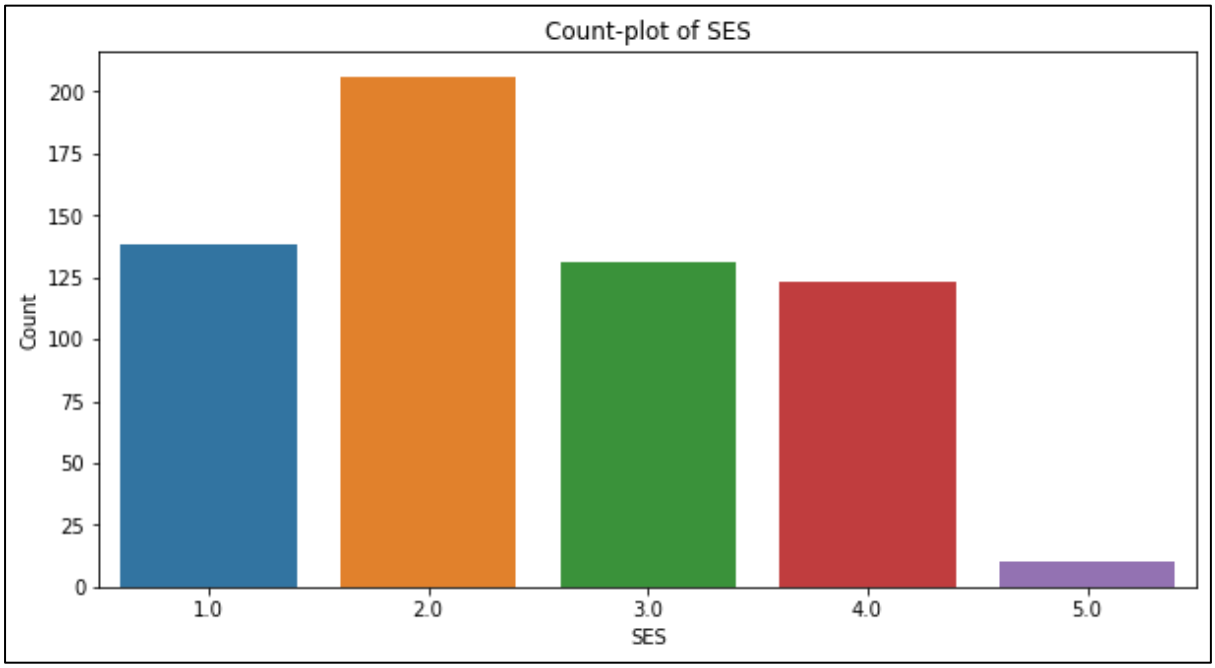


We can see that for category 1.0 and 2.0, the observations are very less and it can cause data imbalance problem. Because of which the problem was converted to binary classification problem. The categories other than zero was considered to be healthy, in our case non-demented. Now data looks balanced.



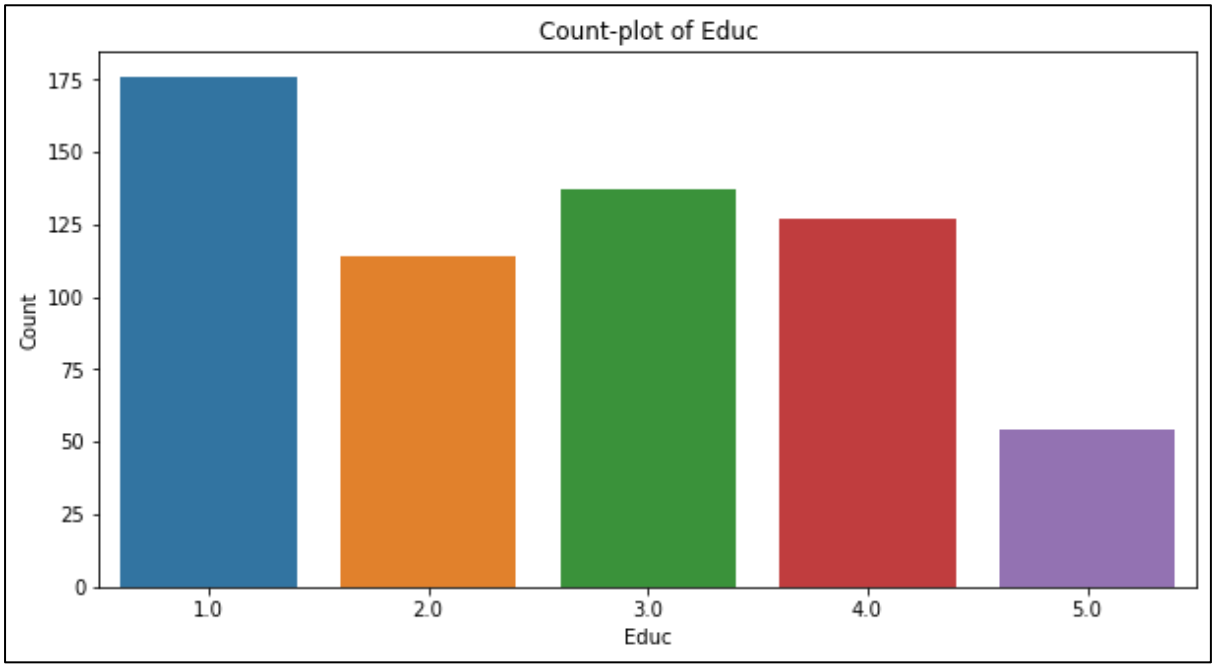
Again, as seen above in data stats section, it can be confirmed that 341 / 608 people are healthy and rest are affected with the disease.

- SES



There are 206 people in our data present in lower-middle class. There are only 10 people present in upper class or rich class. Rest of the data looks nicely balanced.

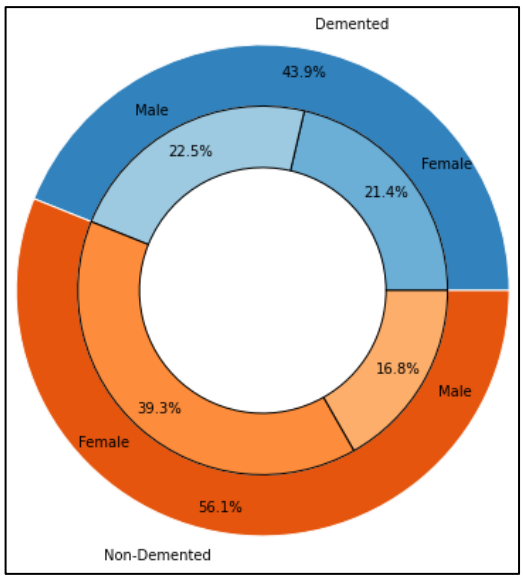
- Educ



Most of the patients in our data are poorly educated. And it can be seen from the graph that only 54 of them are very highly educated. But there is consider amount of balance between all the sub-categories.

Bi-variate Analysis

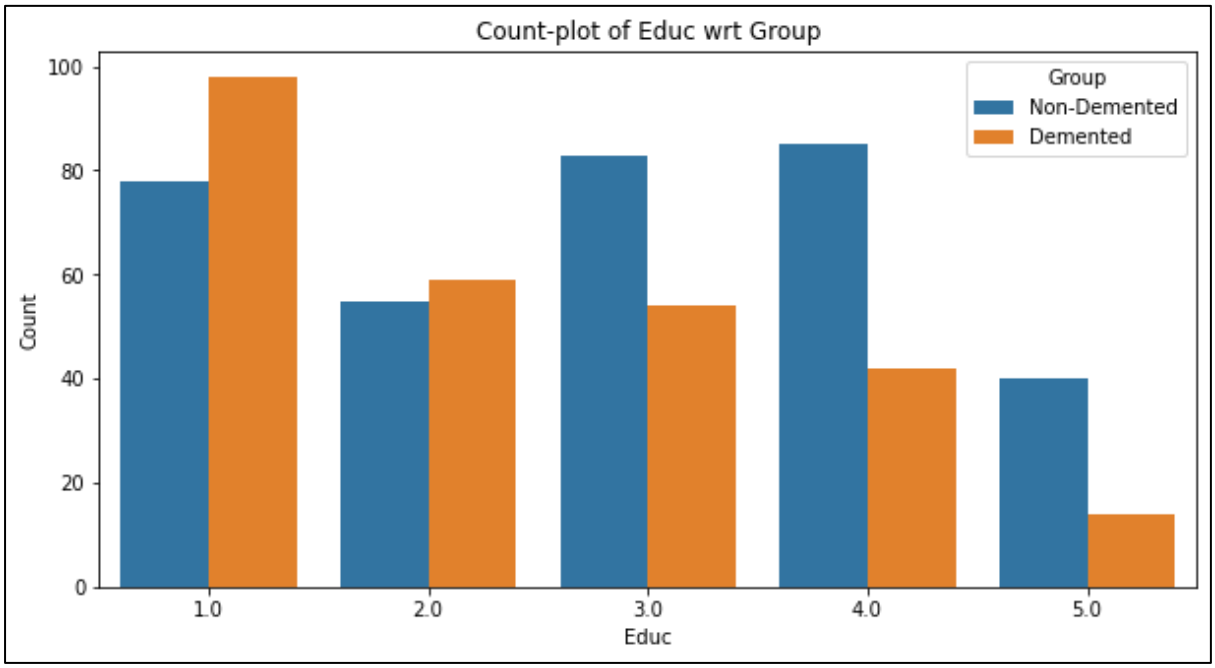
Comparison between Gender and Group



Inferences:

1. The ratio of non-demented to demented ratio for females is very much higher when compared to males.
2. This is very significant insight as it tells us that men are more susceptible to this disease when compared to women.

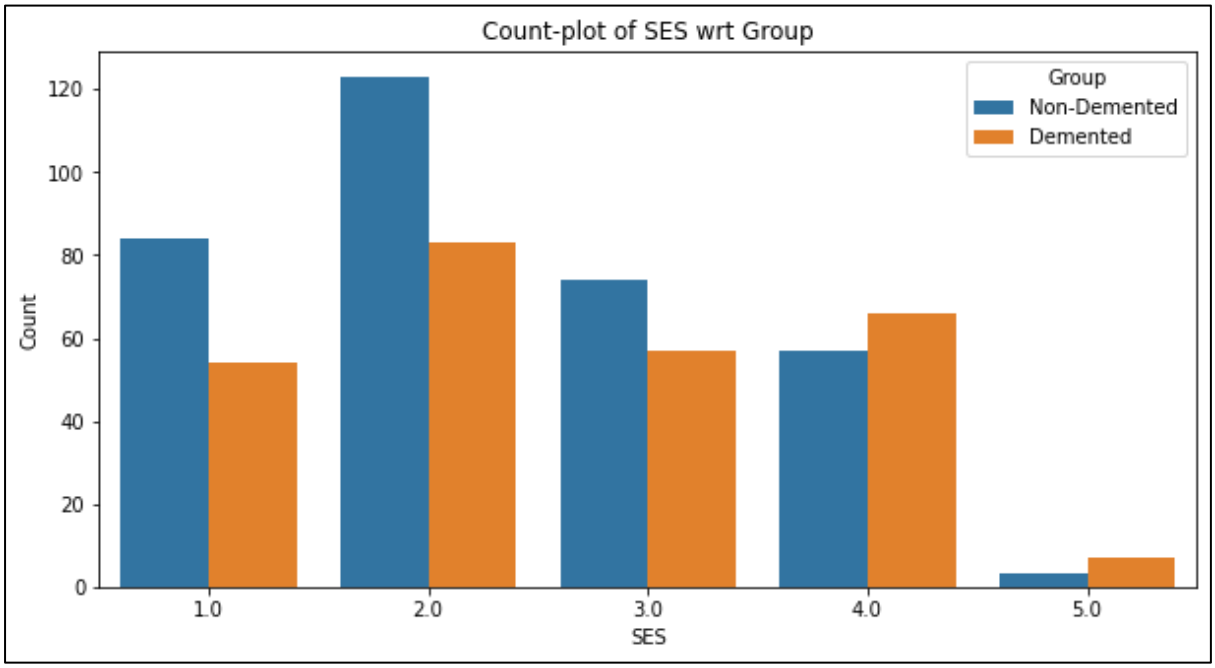
Comparison between Education and Group



This graph also tells us a small story. It can be seen that the patients with very poor education background are more likely to suffer from the Dementia when compared with high education people.

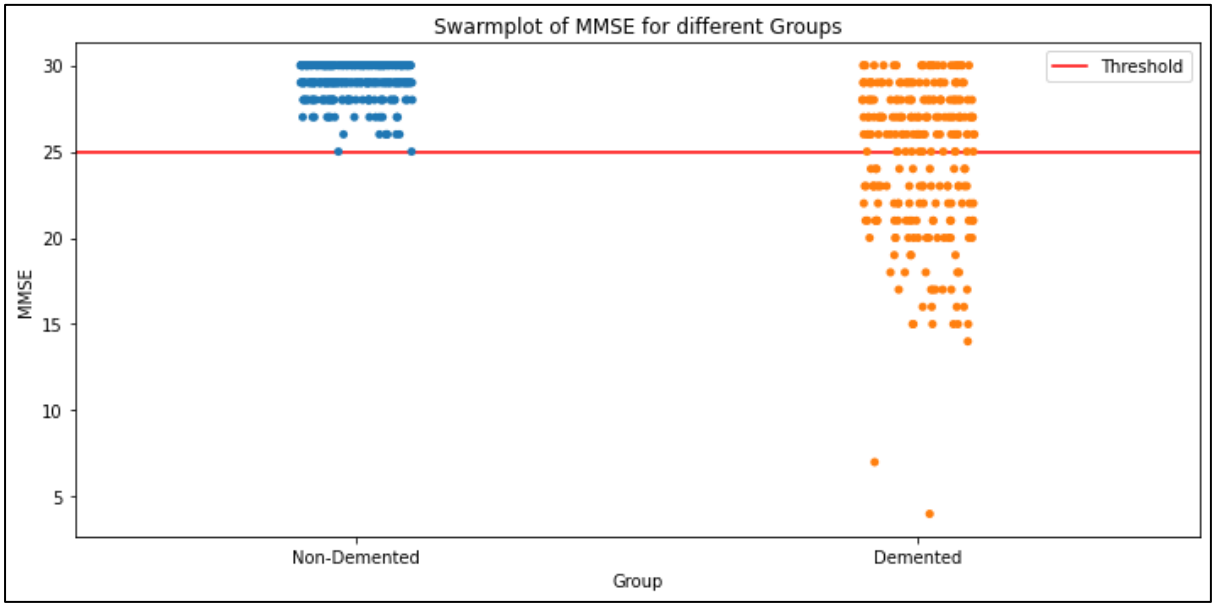
This tells us that keeping our brain more active helps us to avoid getting affected with this disease.

Comparison between SES and Group



From this graph we can observe that people who are rich are more likely to get affected with the disease. From further analysis it was found that in category 5.0, all the members who were affected were female with low education background.

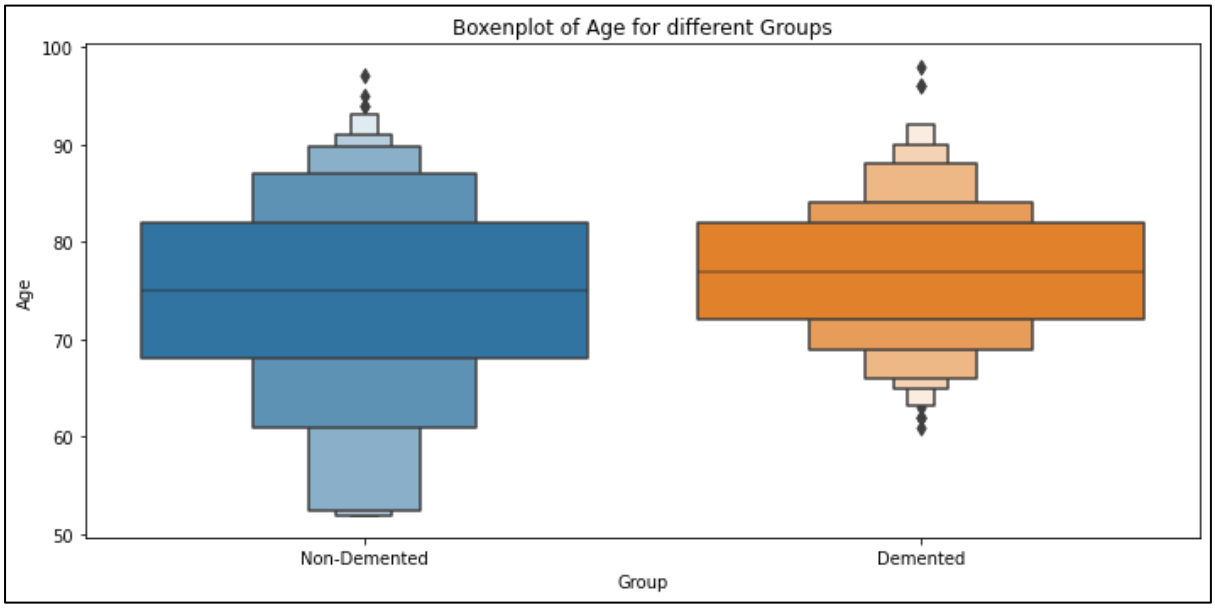
Comparison between MMSE and Group



The plot indicates that all the people who scores less than 25/30 are affected with the disease. It can be also verified from the Data Dictionary as people scoring near 30 are healthier and has the marks decreases the chances of getting affected also increases.

Note: In order to solve our outlier problem, we have considered to treat this MMSE column as categorical column with two categories viz., '> 25' and '<= 25'.

Comparison between Age and Group

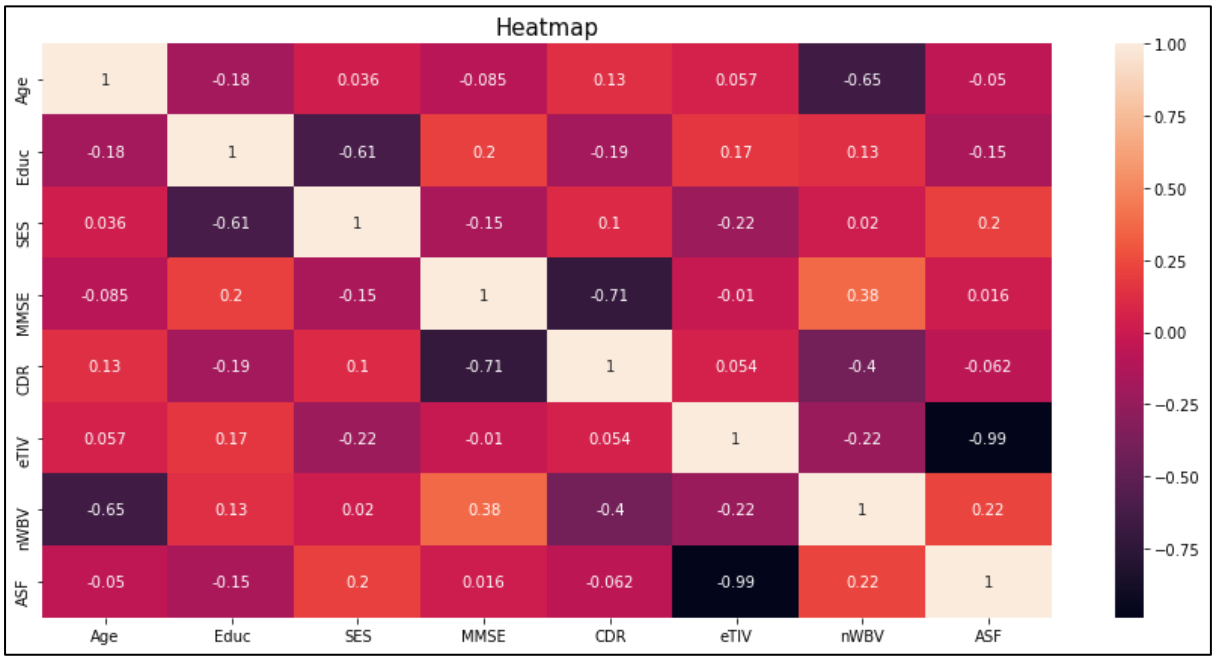


Here, we can compare that the average for demented category is slightly high when compared with non-demented category.

This is regarding the important insights obtained with respect bi-variate analysis. Other features are tested for its significant in statistical hypothesis testing.

Multi-variate Analysis

Heatmap



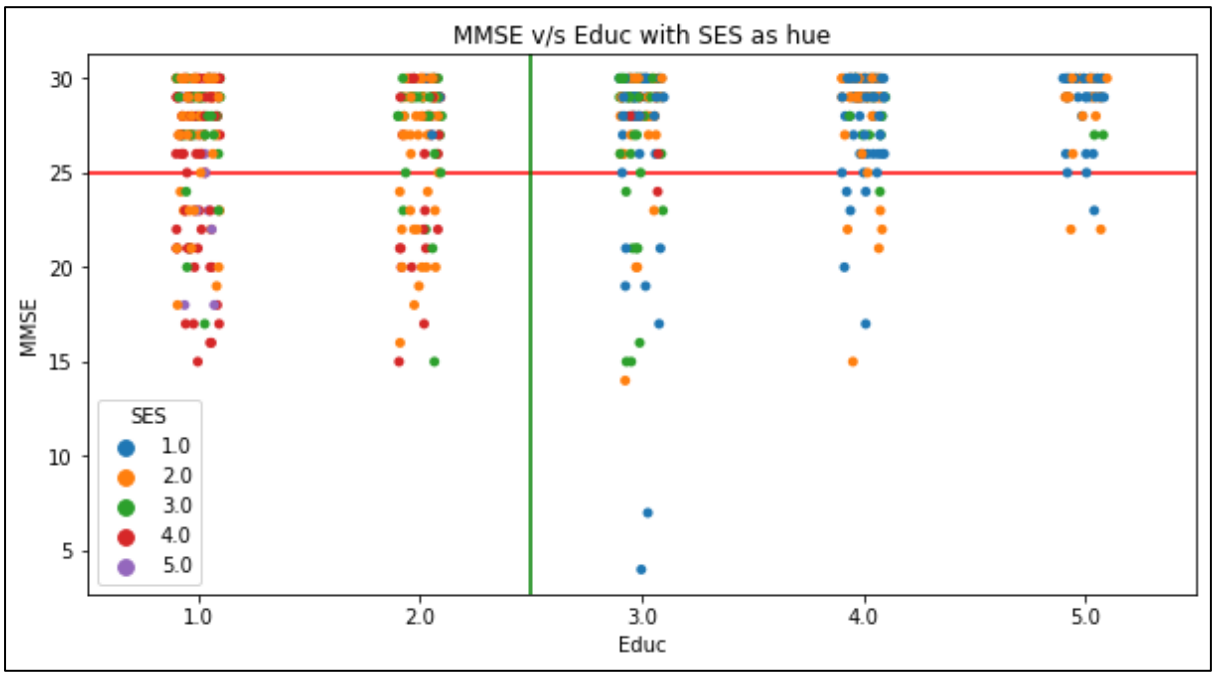
In order to consider target column, we have considered all the variables for heatmap.

Inferences:

1. Strong **negative correlation** between CDR and MMSE i.e., **71%** which implies higher the MMSE greater chances that the person is healthy (CDR = 0.0).
2. **Very strong negative** relationship between eTIV and ASF i.e., **99%**
3. There is **negative relationship** between Age and nWBV i.e.., **65%**
4. There is **mild negative** relationship between CDR and nWBV i.e, **40%**
5. Interestingly we can also observe that the SES and Education have **negative correlation** among them around **60%**

Comparison of SES, Education and MMSE

Here, we are considering MMSE score to study the behaviour of demented people with respect to education and SES.



Consider the region below red line and right of green line. Here, we can see that most of the people who are getting affected comes from low SES background. It tells significant story that people who have pursued higher education and yet settled in lower class get affected by this disease. This can be considered because of the mental stress they put on themselves.

On the other hand, people with low education qualification who are well settled are also suffering from this disease. This can be accounted for less brain activity.

Comparison of Age, MMSE and Group

It was observed that people below the age of 60 years were healthy.

Patients who were above 60 years of age and who scored less than 28/30 in MMSE were significantly suffering from this disease of Dementia.

EDA Conclusion

These were the significant insights obtained from our analysis on the data which would be confirmed by hypothesis testing in the subsequent section.

Statistical Hypothetical Testing

Numerical Variables – Unpaired T-Test

To know the behaviour of the averages of the numerical variables with respect to different groups, we performed two sample independent t-test (unpaired t-test) on them with the Hypothesis as follows:

***H0* - The average for numerical column for demented category and non-demented category are equal.**

***HA* - The average for numerical column for demented category and non-demented category are unequal.**

As a result of this unpaired t-test, we found out that only two numerical variables out of four i.e., *Age* and *nWBV* (normalized whole brain volume) were having p-values less than 5 %.

Hence, we have considered only these two numerical features for our parametric models.

Categorical Variables – Chi-square Test for Independence

Similarly, to analyse the importance of categorical variables with respect to our group category, we performed chi-square test on all categorical features. The Hypothesis considered is as follows:

***H0* - The two categorical variables considered are independent to each other.**

***HA* - The two categorical variables considered are related to each other.**

As a result of the chi-square test for independence of two categorical variables, we found that all the categorical data in our dataset were influencing the group column that is their p-values were less then 5% threshold.

Hence, we considered all the categorical features i.e., *Educ*, *SES*, *MMSE_Level* (Converted variable) and *M/F* for the parametric models.

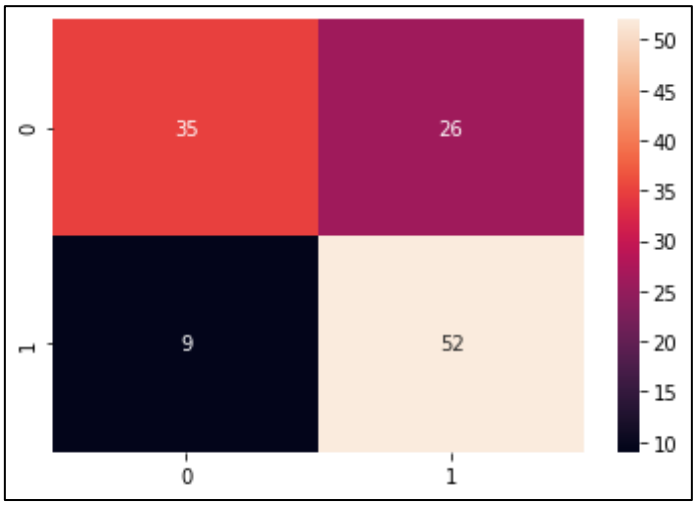
Features for Non-parametric model

For non-parametric model, we shall consider all the features irrespective of their statistical importance.

MODEL BUILDING:

Parametric Models

1. *Logistic Regression*

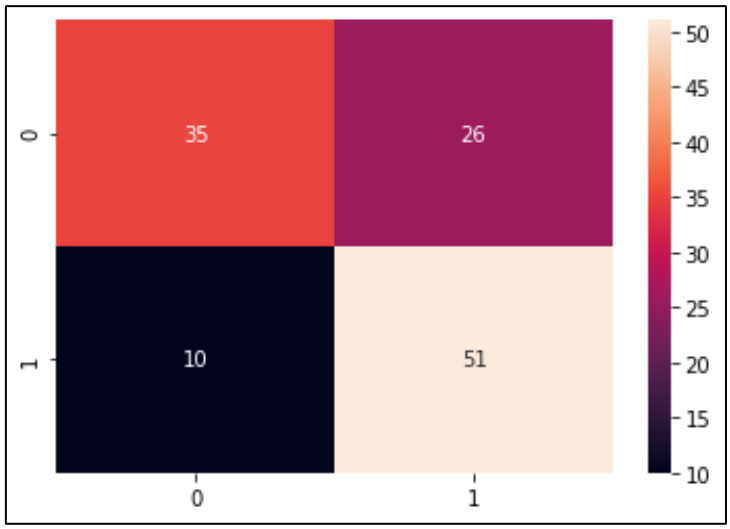


Confusion Matrix

Accuracy Score: 0.6967213114754098
ROC AUC Score: 0.6967213114754097

We can see that the data is underfitting.

2. *Naïve Bayes Algorithm*

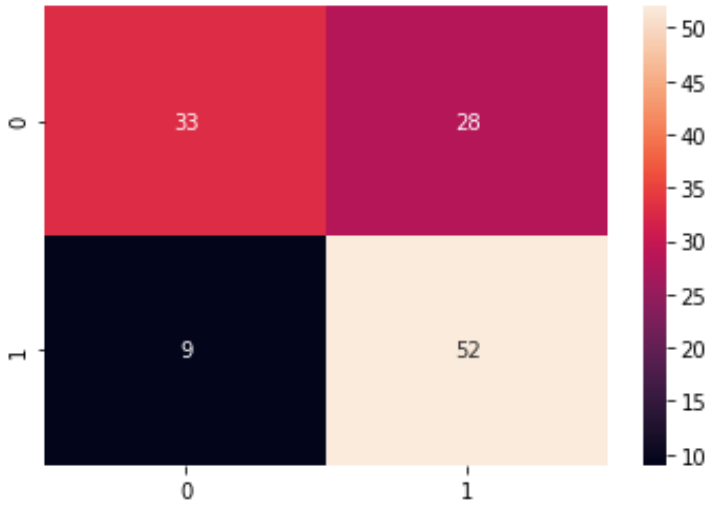


Confusion Matrix

Accuracy Score: 0.7049180327868853
ROC AUC Score: 0.7049180327868853

We can see that the data is underfitting in this model too.

3. *K-Nearest Neighbors Algorithm*



Confusion Matrix

Accuracy Score: 0.6967213114754098
ROC AUC Score: 0.6967213114754097

4. *K-Nearest Neighbors Algorithm – Hyper parameter tuning*

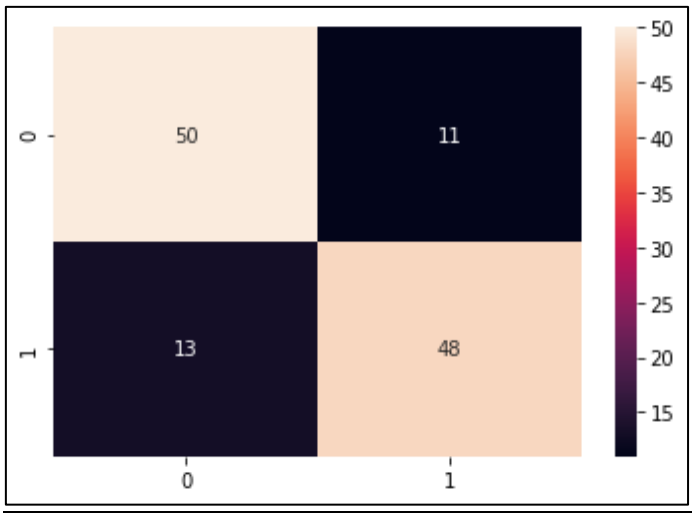
KNN TUNED MODEL: KNeighborsClassifier(
metric='manhattan',
n_neighbors=43,
weights='distance')

KNN TUNED SCORES: [0.79285714 0.88571429 0.6641604]
KNN Tuned Bias Error: 0.21908939014202178
KNN Tuned Variance Error: 0.11125902523276573

We are getting a high bias error.

Non-Parametric Models:

1. *Decision Tree Classifier Algorithm*



Confusion Matrix

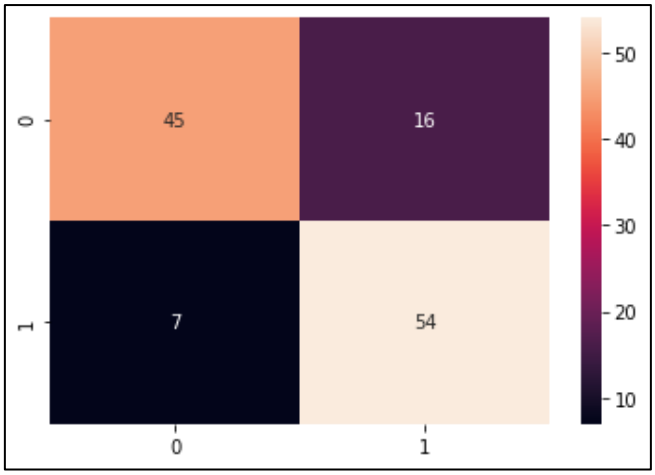
Accuracy Score: 0.8032786885245902
ROC AUC Score: 0.8032786885245902

2. *Decision Tree – Hyperparamerter Tuning*

Decision Tree TUNED MODEL: DecisionTreeClassifier(
criterion='entropy',
max_depth=6,
random_state=10)

Decision Tree TUNED SCORES: [0.81853583 0.84036382 0.86021835]
Decision Tree Tuned Bias Error: 0.1602940047484741
Decision Tree Tuned Variance Error: 0.0208490444307445

3. *Random Forest Classifier Algorithm*



Confusion Matrix

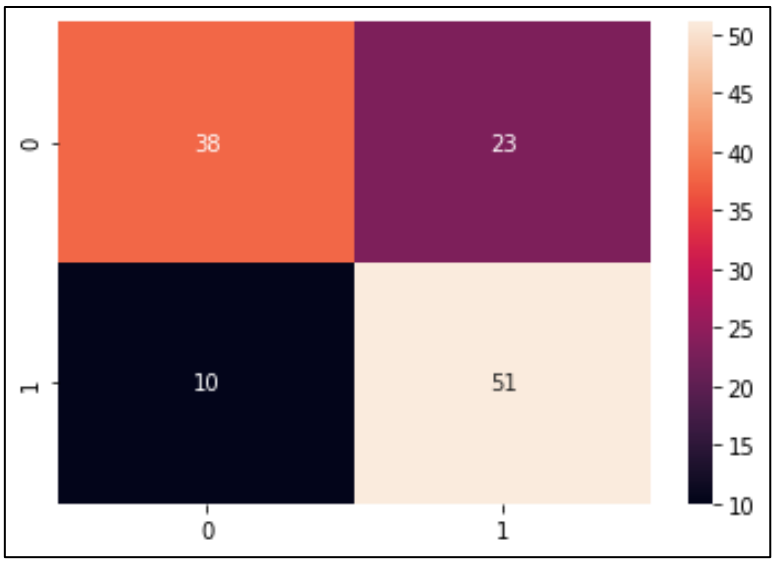
Accuracy Score: 0.8114754098360656
ROC AUC Score: 0.8114754098360655

4. *Random Forest – Hyperparameter Tuning*

Random Forest TUNED MODEL: RandomForestClassifier(
max_depth=7,
n_estimators=38,
random_state=10)

Random Forest TUNED SCORES: [0.84267913 0.88574238 0.90909091]
Random Forest Tuned Bias Error: 0.1208291945451817
Random Forest Tuned Variance Error: 0.033690063306017266

5. *Adaboost Classifier Algorithm*



Confusion Matrix

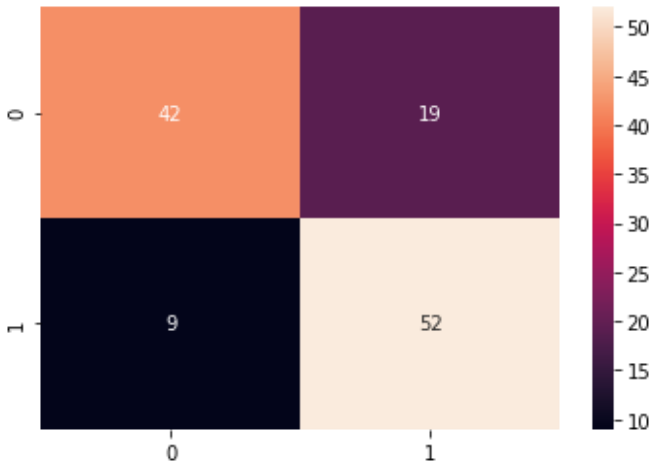
Accuracy Score: 0.7295081967213115
ROC AUC Score: 0.7295081967213115

6. *AdaBoost – Hyperparameter Tuning*

Adaboost TUNED MODEL: AdaBoostClassifier(
base_estimator=RandomForestClassifier(random_state=10), l
earning_rate=0.1,
 n_estimators=1,
random_state=10)

Ada Boost Classifier TUNED SCORES: [0.90172313 0.90821042 0.92944597]
Ada Boost Classifier Tuned Bias Error: 0.08687349048205517
Ada Boost Classifier Tuned Variance Error: 0.014500515579085127

7. *Gradient Boosting Algorithm*



Confusion Matrix

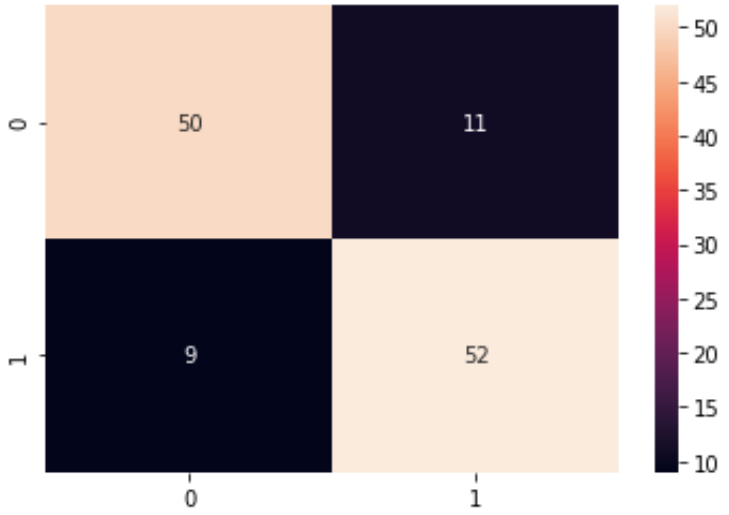
Accuracy Score: 0.7704918032786885
ROC AUC Score: 0.7704918032786885

8. *Gradient Boosting – Hyperparameter Tuning*

Gradient Boosting TUNED MODEL: GradientBoostingClassifier(
learning_rate=0.15,
n_estimators=93,
random_state=10)

Gradient Boosting Classifier TUNED SCORES: [0.87052181 0.88495575 0.92429344]
Gradient Boosting Classifier Tuned Bias Error: 0.1067430004963578 8
Gradient Boosting Classifier Tuned Variance Error: 0.02783038174061353

9. XG boost Algorithm



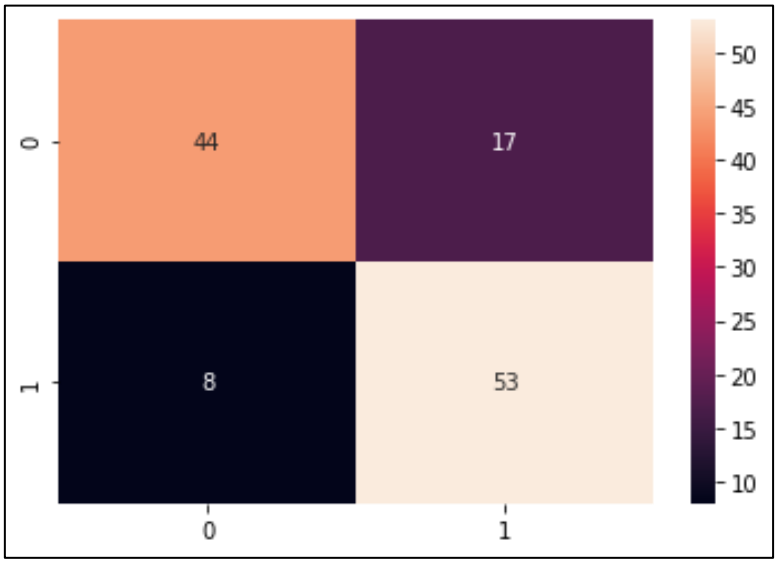
Confusion Matrix

Accuracy Score: 0.8360655737704918
ROC AUC Score: 0.8360655737704918

10. XG boost – Hyperparamter Tuning

XG Boost Classifier TUNED SCORES: [0.8605919 0.89164208 0.92072238]
XG Boost Classifier Tuned Bias Error: 0.1090145466194018
XG Boost Classifier Tuned Variance Error: 0.03007061485901785

11. Light GBM Algorithm



Confusion Matrix

Accuracy Score: 0.7950819672131147

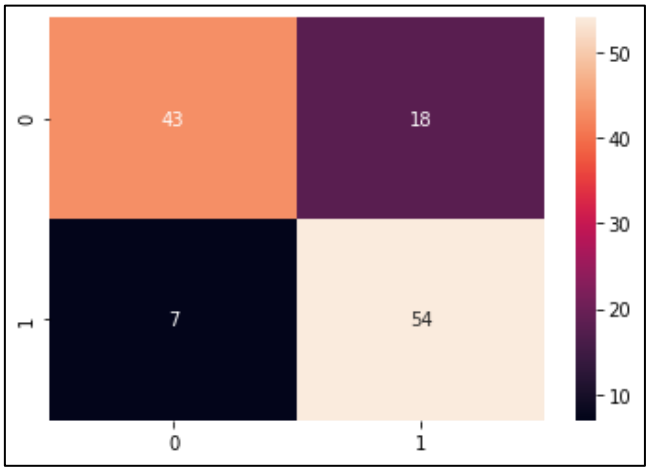
ROC AUC Score: 0.7950819672131147

12. Light GBM – Hyperparameter Tuning

LGBM TUNED MODEL: LGBMClassifier(
learning_rate=0.15,
n_estimators=69,
random_state=10)

LGBM Classifier TUNED SCORES: [0.86360981 0.88466077 0.91255994]
LGBM Boost Classifier Tuned Bias Error: 0.11305649236375415
LGBM Boost Classifier Tuned Variance Error: 0.02455477509736993

13. Catboost Algorithm



Confusion Matrix

Accuracy Score: 0.7950819672131147
ROC AUC Score: 0.7950819672131147

14. Catboost – Hyperparameter Tuning

Catboost Classifier TUNED SCORES: [0.8654595 0.88190757 0.90205081]
Catboost Classifier Tuned Bias Error: 0.09853398055857837
Catboost Classifier Tuned Variance Error: 0.018326724778936647

Applied Models Summary

	ROC AUC	Bias Error	Variance Error
Ada Boost Tuned	0.91	0.0800	0.010
Light GBM Tuned	0.90	0.0700	0.020
Gradient Boost Tuned	0.89	0.0900	0.020
XG Boost Tuned	0.89	0.0900	0.020
Cat Boost Tuned	0.88	0.1000	0.018
Random Forest Tuned	0.87	0.1000	0.030
Decision Tree Tuned	0.83	0.1500	0.020
Random Forest	0.81	0.1300	0.050
Decision Tree	0.80	0.0900	0.100
KNN Tuned	0.79	0.1000	0.110
Light GBM	0.79	0.1300	0.070
Cat Boost	0.79	0.1474	0.050
Gradient Boost	0.77	0.1500	0.070
XG Boost	0.77	0.1600	0.060
Ada Boost	0.72	0.1800	0.080
Logistic	0.71	0.2100	0.073
Naive Bayes	0.70	0.2100	0.081
KNN	0.69	0.2200	0.070

We have found that AdaBoost tuned model with Random Forest has provided the best results.

Since our data was not large, the findings of the model and the insight cannot be levied onto a larger population. This is a limitation.

In the future, better models can be built on top of this model owing to larger availability of data.