Transport Mode Prediction & Deal Outcome Analysis

Deepak Arumugam Vivekanandan
deep199907@gmail.com

Contents

DATA DESCRIPTION : 5
Problem Statement : 7
Introduction : 7
Data Summary: 7
Dataset Sample : 8
Descriptive Statistics 9
Univariate analysis : 10
Bivariate analysis : 11
Multivariate analysis : 12
Outlier Treatment: 13
Outlier Treated: 14
Correlation plot : 14
Understanding from Heatmap 15
Label Encoding: 15
a. Logistic Regression Model 17
Training Accuracy (77%): 17
Test Accuracy (75%): 17
Training Data Confusion Matrix: 17
Inferences: 18
Test Data Confusion Matrix: 18
Inferences: 18
AUC and ROC Curve: 19
b. Linear Discriminant Analysis. 20
Test Accuracy (75%): 20
Training Accuracy (76%): 20
Training Data Confusion Matrix: 21
Inferences for Training Data: 21
Test Data Confusion Matrix: 21
Inferences for Test Data: 22
AUC and ROC Curve: 22
AUC Inference (Train and Test): 22
c. KNN Model 23
Train Accuracy (82%): 23

Test Accuracy (73%): 23
Training Data Confusion Matrix: 23
KNN's performance on the training data indicates the following: 24
Test Data Confusion Matrix: 24
Inferences for Test Data: 24
AUC and ROC Curve: 25
Inference (AUC Analysis): 25
d. Naive Bayes Model 25
Train Accuracy (0.76): 26
Test Accuracy (0.80): 26
Training Data Confusion Matrix: 26
Inferences for Training Data: 26
Test Data Confusion Matrix: 27
Inferences for Test Data: 27
AUC and ROC Curve: 27
AUC Inference (Naive Bayes): 28
e. Decision Tree Classifier 28
Training Set Accuracy (1.00): 28
Test Set Accuracy (0.75): 28
Training Data Confusion Matrix: 29
Inferences for Training Data: 29
Test Data Confusion Matrix: 29
AUC and ROC Curve: 30
AUC Inference (Decision Tree): 30
f. Random Forest Model 31
Training Data Accuracy (1.0): 31
Test Data Accuracy (0.78): 31
Training Data Confusion Matrix: 32
Inferences for Training Data: 32
Test Data Confusion Matrix: 32
Inferences for Test Data: 33
AUC and ROC Curve: 33
General AUC Inference (Random Forest): 33
g. Boosting classifier using Gradient boost 34
Training Set Accuracy (0.86): 34
Test Set Accuracy (0.75): 34

Train and Test Data Confusion Matrix: 34
Test Data Confusion Matrix: 35
Inferences for Training Data: 35
Test Data Confusion Matrix: 35
Inferences for Test Data: 35
Additional Considerations: 36
AUC and ROC Curve: 36
General AUC Inference (Boosting Classifier using Gradient Boost): 36
Overall Analysis: 37
High Predictive Accuracy for Public Transport Usage: 37
Business Implications: 38
2. DATASET - Shark Tank Comapnies.csv 38
Part 2: Text Mining 38
Problem Statement : 38
Introduction : 38
Word Cloud: 41
Inference: 42
Word Cloud Analysis: 42
Lack of Conclusive Evidence: 42
Additional Analysis Needed: 43

List of Figures

Fig 1. Bar graph for transport	9
Fig 2. Histplot for Salary and Age	9
Fig 3. Countplot for Transport	10
Fig 4. Countplot for Transport	11
Fig 5. Countplot for Transport	11
Fig 6 . Boxplot for all data	12
Fig 7. Boxplot for engineer column	12
Fig 8. Boxplot for MBA column	12
Fig 9 . Boxplot for all data (outliers treated)	13
Fig 10. Heatmap for all data	13
Fig 11. Confusion matrix (LR-Train)	17
Fig 12: Confusion matrix (LR-Test)	17
Fig 13. AUC-ROC curve (LR-Train)	18
Fig 14. AUC-ROC curve (LR-Test)	18
Fig 15. Confusion matrix (LDA-Train)	20
Fig 16. Confusion matrix (LDA-Test)	20
Fig 17. AUC-ROC (LDA-Train)	21
Fig 18. AUC-ROC (LDA-Test)	21
Fig 19. Confusion matrix (KNN-Train)	22
Fig 20. Confusion matrix (KNN-Test)	23
Fig 21. AUC-ROC (KNN-Train)	24
Fig 22. AUC-ROC(KNN-Test)	24
Fig 23. Confusion matrix (NB-Train)	25
Fig 24. Confusion matrix (NB-Test)	26
Fig 25. AUC-ROC (NB-Test)	26
Fig 26. AUC-ROC (NB-Test)	26
Fig 27. Confusion matrix (DT-Train)	28
Fig 28. Confusion matrix (DT-Test)	28
Fig 29. AUC-ROC(DT-Train)	29
Fig 30. AUC-ROC(DT-Test)	29
Fig 31. Confusion matrix(RF-Train)	30
Fig 32. Confusion matrix(RF-Test)	31
Fig 33. AUC-ROC(RF-Train)	32
Fig 34. AUC-ROC(RF-Test)	32
Fig 35. Confusion Matrix (GB-Train and Test)	33
Fig 36. AUC-ROC (GB-Train)	34
Fig 37. AUC-ROC (GB-Test)	34
Fig 38. 3 Most occurring words	40
Fig 39. Deal customers word cloud	40
Fig 40. No Deal customers word cloud	40

List of Tables

Table 1. Variable Description	5
Table 2. Dataset Sample	7
Table 3. Descriptive stats	8
Table 4. Before label encoding	15
Table 5. Before label encoding	15
Table 6. Train Classification Report- LR	16
Table 7. Train Classification Report- LR	16
Table 8. Train Classification Report- LDA	9
Table 9. Train Classification Report- LDA	9
Table 10. Train Classification Report- KNN	22
Table 11. Train Classification Report- KNN	22
Table 12. Train Classification Report- NB	24
Table 13. Train Classification Report- NB	24
Table 14. Train Classification Report- DT	27
Table 15. Train Classification Report- DT	27
Table 16. Train Classification Report- RF	30
Table 17. Train Classification Report- RF	30
Table 18. Train Classification Report- GB	32
Table 19. Train Classification Report- GB	32
Table 20. Classification Report for all models	35
Table 21. Dataset of Shark Tank Companies	37
Table 22. Deal Customers	38
Table 23. Deal Customers	38
Table 24. No Deal Customers	38
Table 25. Deal Customers character count	39
Table 26. No Deal Customers character count	39

DATA DESCRIPTION :

VARIABLE	DESCRIPTION
Age	Age of the Employee in Years
Gender	Gender of the Employee

Engineer	For Engineer =1 , Non Engineer =0
MBA	For MBA =1 , Non MBA =0
Work Exp	Experience in years
Salary	Salary in Lakhs per Annum
Distance	Distance in Kms from Home to Office
License	If Employee has Driving Licence -1, If not, then 0
Transport	Mode of Transport

Table 1. Variable Description

Problem Statement :

You work for an office transport company and are currently in discussions with ABC Consulting company to provide transportation services for their employees. Your task is to understand the current commuting preferences of ABC Consulting employees between their homes and the office. Using data from the 'Cars.csv' dataset, which includes parameters such as age, salary, work experience, and more, your objective is to predict the preferred mode of transport. This project involves building multiple machine learning models and comparing their performance to finalize the most suitable model for the task.

Introduction :

As representatives of the office transport company, our goal is to assess ABC Consulting employees' commuting preferences for potential transport services. Leveraging the 'Cars.csv' dataset with parameters like age, salary, and work experience, we're building diverse Machine Learning models to predict their preferred mode of transportation. This report encapsulates our extensive analyses, presenting findings and insights from model comparisons. Our recommendations are vital for the success of this significant project, aiding ABC Consulting in providing efficient and preferred transportation options for their workforce, ultimately enhancing employee satisfaction and operational effectiveness.

Data Summary:

The project's primary goal is to analyze the 'Cars.csv' data set and gain insights into the commuting preferences of ABC Consulting employees between their residences and the workplace. Through the examination of key factors like age, salary, and work experience, the aim is to predict their favored mode of transport. To achieve this, multiple Machine Learning models will be constructed and evaluated to identify the most effective model for the task at hand.

Dataset Sample :

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport
...	...	...	...	...	...	...	...	...	...
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

444 rows × 9 columns

Table 2. Dataset Sample

- Based on the information provided, the dataset consists of the following characteristics:
- Number of variables: 9
- Number of unique data: 444
- Categorical variables: 2
- Numerical variables: 7
- Please note that without specific details about the variable names and their attributes, it is difficult to provide further information about the dataset.

Descriptive Statistics

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Age	444.0	NaN	NaN	NaN	27.747748	4.41671	18.0	25.0	27.0	30.0	43.0
Gender	444	2	Male	316	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Engineer	444.0	NaN	NaN	NaN	0.754505	0.430866	0.0	1.0	1.0	1.0	1.0
MBA	444.0	NaN	NaN	NaN	0.252252	0.434795	0.0	0.0	0.0	1.0	1.0
Work Exp	444.0	NaN	NaN	NaN	6.29955	5.112098	0.0	3.0	5.0	8.0	24.0
Salary	444.0	NaN	NaN	NaN	16.238739	10.453851	6.5	9.8	13.6	15.725	57.0
Distance	444.0	NaN	NaN	NaN	11.323198	3.606149	3.2	8.8	11.0	13.425	23.4
license	444.0	NaN	NaN	NaN	0.234234	0.423997	0.0	0.0	0.0	0.0	1.0
Transport	444	2	Public Transport	300	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 3. Descriptive stats

- In Salary column the minimum value is 6.5
- In Salary column the 25% value is 9.8 Out of 57
- In Salary column the 50% value is 13.6 Out of 57
- In Salary column the 75% value is 15.725 Out of 57
- In Salary column the maximum value is 57
- In Yearly_avg_comment_on_travel_page The Maximum value is 815.
- As per Gender ratio Male is high when compared to women
- Due to the abundance of categorical variables, conducting detailed descriptive statistics is impractical. Instead, the forthcoming sections will employ visualization techniques for a comprehensive analysis of the dataset's characteristics and patterns.

Univariate analysis :

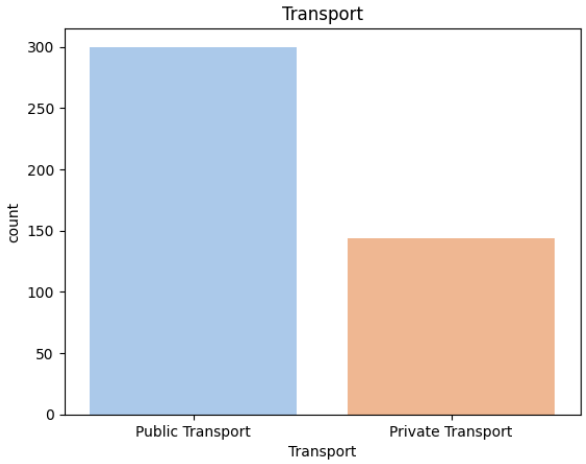


Fig 1. bar graph for transport

```
Public Transport     300
Private Transport    144
Name: Transport, dtype: int64
```

- The analysis indicates that there are 300 users of public transport and 144 users of private transportation, resulting in a ratio of approximately 1:2 between public and private transport users.
- The graph indicates a pronounced preference for public transport over private modes of transportation.
- Notably, when the usage of private transport is increased twofold, it matches the total volume of public transport usage among individuals.
- This underscores a substantial inclination towards public transportation within the surveyed population.

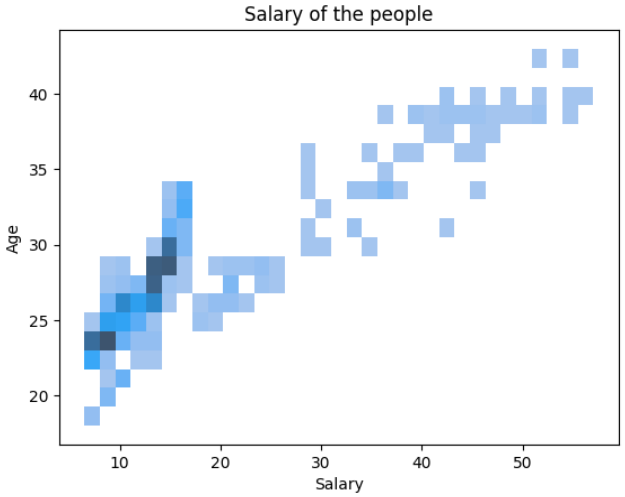


Fig 2. Histplot for Salary and Age

- The plot above reveals a positive correlation between age and salary, indicating that as age increases, so does salary.
- The majority of individuals fall within the age range of 25 to 30 and earn salaries ranging from below 10 to 20, demonstrating a concentration of individuals in this demographic.

Bivariate analysis :

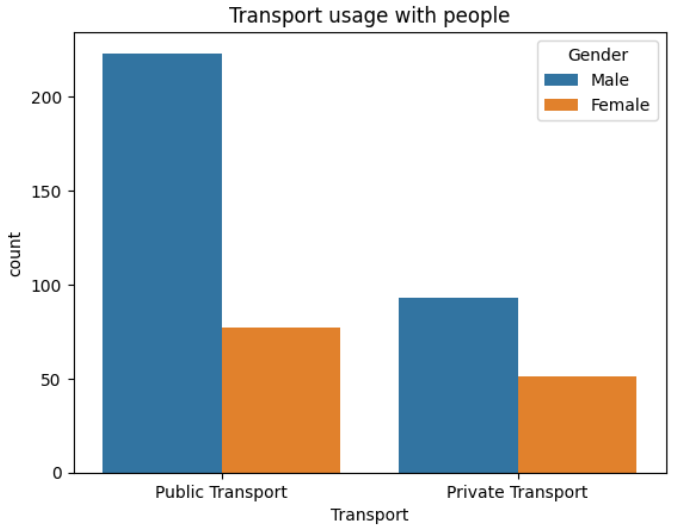


Fig 3. countplot for Transport

- **Gender Disparity:** There exists a notable 2:1 gender disparity in both public and private transport usage, with males predominating.

- **Transportation Factors:** Factors like safety, convenience, affordability, and accessibility likely contribute to this gender disparity.

- **Societal Concerns:** The disparity in transport usage reflects broader societal gender imbalances, requiring attention to promote gender equity in transportation and society as a whole.

Multivariate analysis :

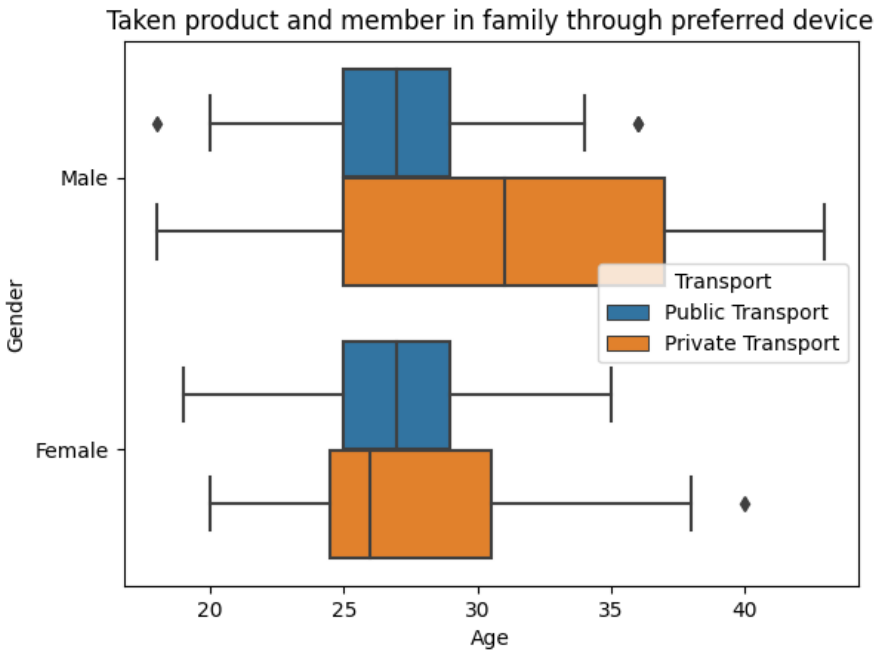


Fig 4. countplot for Transport

- **Age and Private Transportation:** With increasing age, there's a higher inclination to use private transportation.
- **Male Preference for Private Travel:** Males tend to prefer private vehicles for travel, regardless of age.
- **Planning Considerations:** Demographic trends in transportation usage, especially among aging populations and males, should inform transportation planning and policy decisions.

Missing values Treatment:

```
Age          0
Gender       0
Engineer     0
MBA          0
Work Exp     0
Salary       0
Distance     0
license      0
Transport    0
dtype: int64
```

Fig 5.countplot for Transport

There are no missing values in the dataset.

Outlier Treatment:

Fig 6 . boxplot for all data

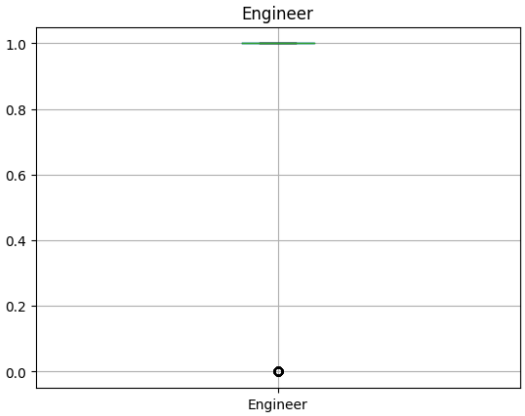


Fig 7. boxplot for engineer column

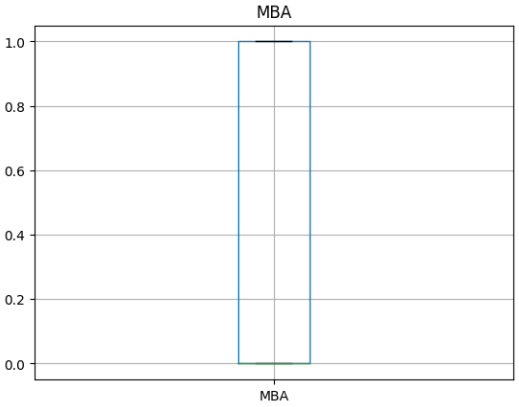


Fig 8. boxplot for MBA column

- Outliers in the dataset were addressed using the Interquartile Range (IQR) method for the whole dataset.
- Outliers were detected in all columns except for the MBA column.
- The Salary column had a notably higher concentration of outliers compared to other columns.

Outlier Treated:

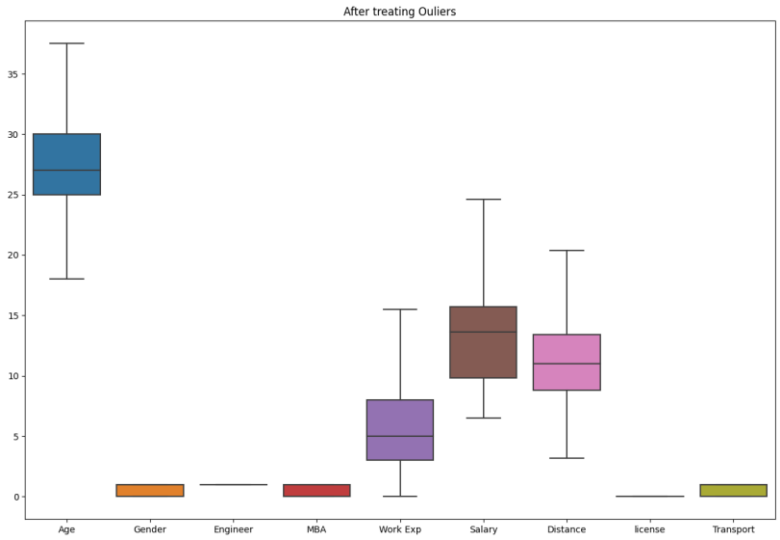


Fig 9 . boxplot for all data (outliers treated)

There are no outliers present in the dataset. Hence this dataset is ready for the regression

and model building process.

Correlation plot :

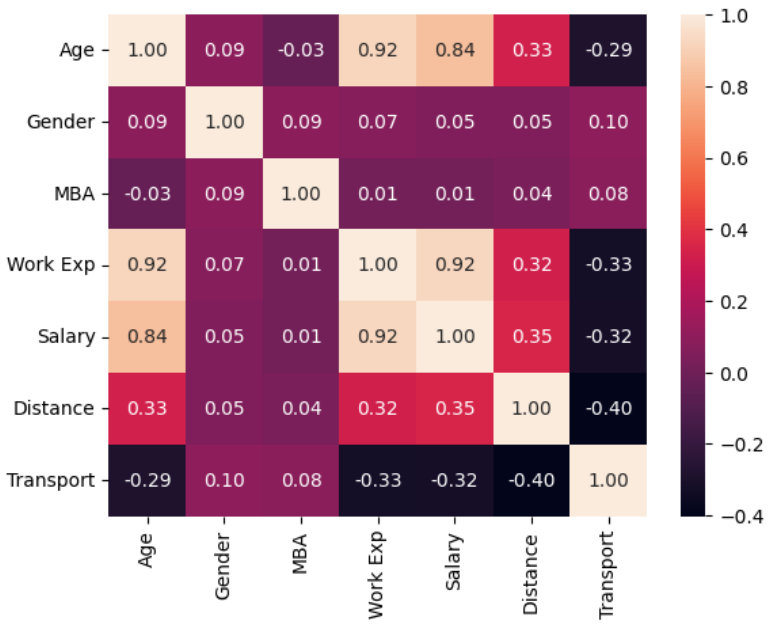


Fig 10. Heatmap for all data

Understanding from Heatmap

- Minimal correlations exist among the variables.
- The most substantial positive correlation, at 92%, is observed between "Work Experience" and "Salary."
- Conversely, the most pronounced negative correlation is found between "Distance" and "Transport," which stands at 40%. A similar correlation of 29% is noted between "Age" and "Transport."
- The likelihood of multicollinearity is low or negligible.

Label Encoding:

Categorical Variables: The dataset contains two categorical variables: "Gender" and "Transport."

Conversion to Numerical:

- To utilize these categorical variables in modeling, a common approach is to convert them into numerical format.

Label Encoding for "Gender":

- For the "Gender" variable, label encoding assigns unique numerical labels to each category. For example, "Male" could be represented as 0, and "Female" as 1.

Label Encoding for "Transport":

- Similarly, label encoding is applied to the "Transport" variable, converting categories like "Public" to 0 and "Private" to 1.

Considerations: While label encoding is straightforward, it may imply ordinal relationships in the data. Ensure that such relationships are justified or consider alternative encoding methods like one-hot encoding if ordinality doesn't exist.

Model Readiness: After encoding, the dataset is prepared for various machine learning models, allowing categorical information to be represented numerically, ready for analysis and prediction.

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28	Male	0	0	4	14.3	3.2	0	Public Transport
1	23	Female	1	0	4	8.3	3.3	0	Public Transport
2	29	Male	1	0	7	13.4	4.1	0	Public Transport
3	28	Female	1	1	5	13.4	4.5	0	Public Transport
4	27	Male	1	0	4	13.4	4.6	0	Public Transport
...	...	...	...	...	...	...	...	...	...
439	40	Male	1	0	20	57.0	21.4	1	Private Transport
440	38	Male	1	0	19	44.0	21.5	1	Private Transport
441	37	Male	1	0	19	45.0	21.5	1	Private Transport
442	37	Male	0	0	19	47.0	22.8	1	Private Transport
443	39	Male	1	1	21	50.0	23.4	1	Private Transport

444 rows × 9 columns

Table 4. Before Label Encoding

	Age	Gender	Engineer	MBA	Work Exp	Salary	Distance	license	Transport
0	28.0	1	1.0	0.0	4.0	14.3000	3.2000	0.0	1
1	23.0	0	1.0	0.0	4.0	8.3000	3.3000	0.0	1
2	29.0	1	1.0	0.0	7.0	13.4000	4.1000	0.0	1
3	28.0	0	1.0	1.0	5.0	13.4000	4.5000	0.0	1
4	27.0	1	1.0	0.0	4.0	13.4000	4.6000	0.0	1
...	...	...	...	...	...	...	...	...	...
439	37.5	1	1.0	0.0	15.5	24.6125	20.3625	0.0	0
440	37.5	1	1.0	0.0	15.5	24.6125	20.3625	0.0	0
441	37.0	1	1.0	0.0	15.5	24.6125	20.3625	0.0	0
442	37.0	1	1.0	0.0	15.5	24.6125	20.3625	0.0	0
443	37.5	1	1.0	1.0	15.5	24.6125	20.3625	0.0	0

444 rows × 9 columns

Table 5. Before Label Encoding

- **Dataset Split:** The dataset is divided into a training set comprising 70% of the data and a test set consisting of 30% of the data. This partitioning allows for model training and evaluation.

- **Imbalanced Data:** The dataset is characterized by imbalanced class distribution, which means that one class may have significantly more instances than the other(s).

- **Scaling Requirement:** Scaling of the dataset is deemed necessary to ensure that all features are on a consistent scale. This is crucial for various machine learning models and analysis techniques.

- **Standard Scaling (Z-score Scaling):** The chosen scaling method is standard scalar (Z-score scaling). It involves transforming features to have a mean of 0 and a standard deviation of 1. This method is effective in addressing the data's varying scales and aids in obtaining accurate and reliable values for analysis and modeling.

- Scaling has been performed using the standard scaler (Z-score scaling) method on the dataset. This scaling method ensures that the features have a mean of 0 and a standard deviation of 1, making the data suitable for various machine learning models and analysis techniques.

a. Logistic Regression Model

Logistic Regression is a binary classification model that predicts the probability of an input belonging to one of two classes. It uses the logistic function to map a linear combination of input features to a probability between 0 and 1, making it a widely used tool for various classification tasks.

Classification Report - Train Set:

	precision	recall	f1-score	support
0	0.72	0.48	0.58	102
1	0.78	0.91	0.84	208
accuracy			0.77	310
macro avg	0.75	0.69	0.71	310
weighted avg	0.76	0.77	0.75	310

Table 6. Train Classification Report - LR

Classification Report - Test Set:

	precision	recall	f1-score	support
0	0.64	0.43	0.51	42
1	0.77	0.89	0.83	92
accuracy			0.75	134
macro avg	0.71	0.66	0.67	134
weighted avg	0.73	0.75	0.73	134

Table 7. Train Classification Report - LR

Training Accuracy (77%):

- The model's accuracy on the training data is 77%, which means it correctly classifies 77% of the training samples.
- A training accuracy of 77% suggests that the model has some level of capability to fit the training data.

Test Accuracy (75%):

- The model's accuracy on the test data is 75%, indicating that it correctly classifies 75% of the test samples.
- A test accuracy of 75% suggests that the model's performance on unseen data is somewhat similar to its performance on the training data.

Training Data Confusion Matrix:

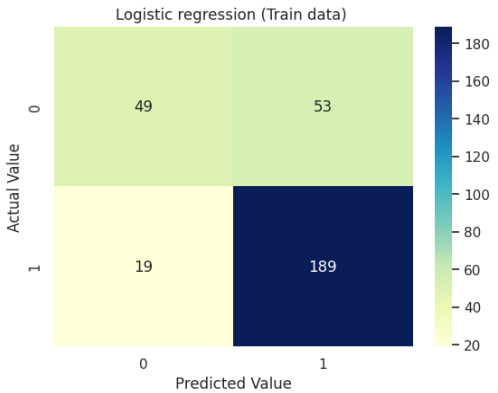


Fig 11. Confusion matrix (LR-Train)

- True Negatives (TN): 49
- False Positives (FP): 53
- False Negatives (FN): 19
- True Positives (TP): 189

Inferences:

- The model correctly predicted 49 instances as negative and 189 instances as positive in the training data.
- However, it also made 53 false positive predictions (instances that were actually negative but predicted as positive) and 19 false negative predictions (instances that were actually positive but predicted as negative).
- The training data's confusion matrix suggests that the model has some ability to identify true positives but also makes a noticeable number of false positive and false negative errors.

Test Data Confusion Matrix:

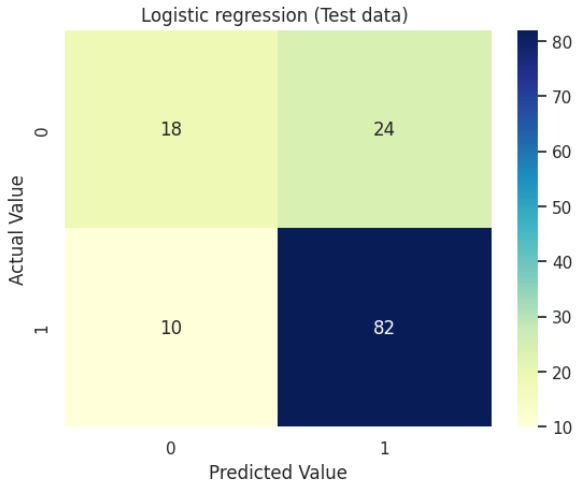


Fig 12. Confusion matrix (LR-Test)

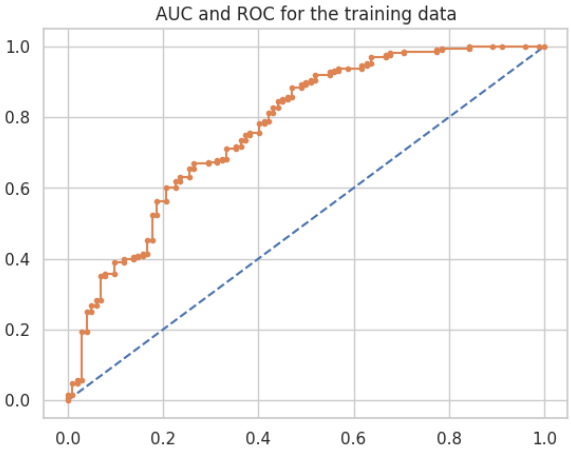
- True Negatives (TN): 18
- False Positives (FP): 24
- False Negatives (FN): 10
- True Positives (TP): 82

Inferences:

- The model correctly predicted 18 instances as negative and 82 instances as positive in the test data.

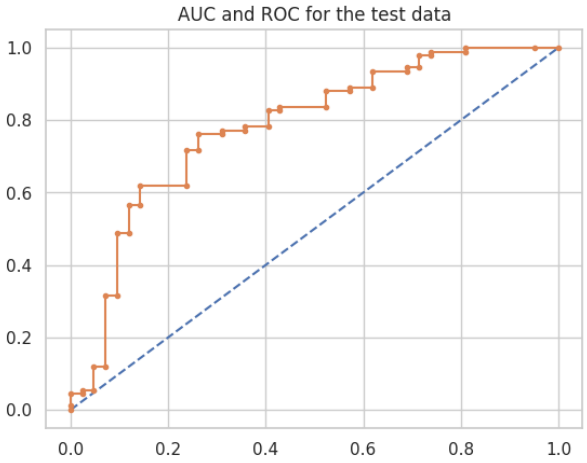
- However, it also made 24 false positive predictions and 10 false negative predictions in the test data.
- The test data's confusion matrix shows a similar pattern to the training data, indicating that the model's performance is consistent across both datasets.

AUC and ROC Curve:



AUC: 0.772

Fig 13. AUC-ROC curve (LR-Train)



AUC: 0.782

Fig 14. AUC-ROC curve (LR-Test)

AUC Curve Analysis:

- AUC is a valuable metric for evaluating the performance of a classification model, especially when dealing with imbalanced datasets or when assessing the overall ability of the model to discriminate between positive and negative classes.
- In this case, the AUC value for the training dataset is 0.772, and for the test dataset, it is 0.782.
- An AUC value close to 1.0 indicates that the model has a strong ability to distinguish between positive and negative instances, while an AUC value close to 0.5 suggests that the model's predictive power is equivalent to random guessing.

Overall Inference:

- When evaluating this logistic regression model, it's essential to consider a combination of metrics, including accuracy, precision, recall, F1-score, ROC AUC, and the confusion matrix.
- The ROC AUC values above 0.5 indicate that the model has some predictive power, and the difference in AUC between training and test data is not substantial.
- However, the model does make a noticeable number of false positives and false negatives, so further optimization or model comparison with other algorithms may be warranted, depending on the specific requirements and goals of the problem

b. Linear Discriminant Analysis.

Linear Discriminant Analysis (LDA) is a dimensionality reduction and classification technique. It maximizes the distance between class means while minimizing the spread within each class. It transforms input features into a lower-dimensional space to better separate classes. LDA assumes Gaussian distribution and equal covariance for all classes, making it useful for supervised classification tasks.

Classification report - train set

	precision	recall	f1-score	support
0	0.71	0.48	0.57	102
1	0.78	0.90	0.84	208
accuracy			0.76	310
macro avg	0.75	0.69	0.71	310
weighted avg	0.76	0.76	0.75	310

Table 8. Train Classification Report - LDA

Classification report - test data

	precision	recall	f1-score	support
0	0.64	0.43	0.51	42
1	0.77	0.89	0.83	92
accuracy			0.75	134
macro avg	0.71	0.66	0.67	134
weighted avg	0.73	0.75	0.73	134

Table 9. Train Classification Report - LDA

Test Accuracy (75%):

- The LDA model achieves a test accuracy of 75%, signifying that it accurately classifies 75% of the test dataset.
- This demonstrates its capability to perform reasonably well on unseen data.
- However, it does not, by itself, indicate the presence or absence of overfitting or underfitting.

Training Accuracy (76%):

- The model's training accuracy stands at 76%, suggesting that it correctly classifies 76% of the training data.
- This training accuracy implies that the model learns from the training set but may not generalize perfectly to new data.
- If the training accuracy were significantly higher than the test accuracy, it would be a potential sign of overfitting, indicating that the model memorizes the training data rather than generalizing from it.

Training Data Confusion Matrix:

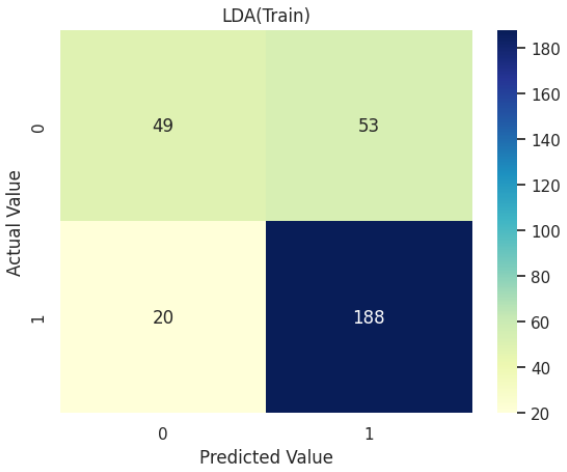


Fig 15. Confusion matrix (LDA-Train)

- True Negatives (TN): 49
- False Positives (FP): 53
- False Negatives (FN): 20
- True Positives (TP): 188

Inferences for Training Data:

- LDA's performance on the training data indicates the following:
- It correctly predicted 49 instances as negative and 188 instances as positive.
- However, it also made 53 false positive predictions (instances that were actually negative but predicted as positive) and 20 false negative predictions (instances that were actually positive but predicted as negative).
- While the true positive count is relatively high, the presence of false positives and false negatives suggests room for improvement in precision and recall.

Test Data Confusion Matrix:

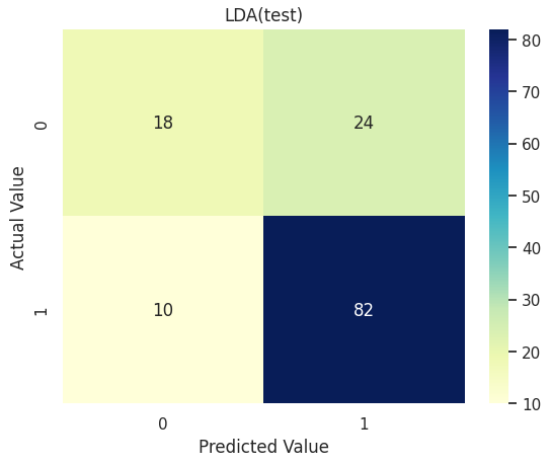


Fig 16. Confusion matrix (LDA-Test)

- True Negatives (TN): 18
- False Positives (FP): 24
- False Negatives (FN): 10
- True Positives (TP): 82

Inferences for Test Data:

- LDA's performance on the test data indicates the following:
- It correctly predicted 18 instances as negative and 82 instances as positive.
- However, it also made 24 false positive predictions and 10 false negative predictions.
- Similar to the training data, the test data results highlight the trade-off between precision and recall, with room for improvement.

AUC and ROC Curve:

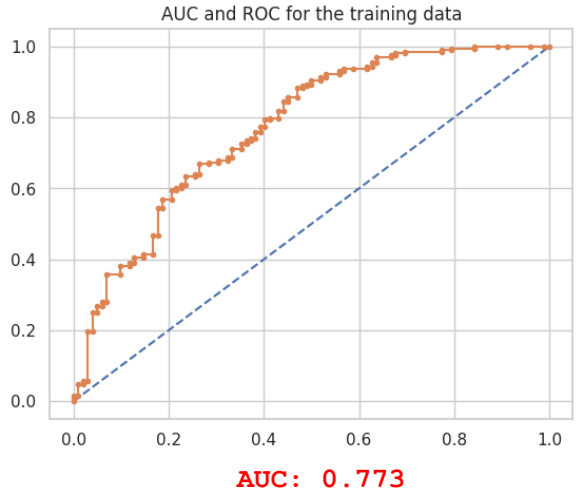


Fig 17. AUC-ROC (LDA-Train)

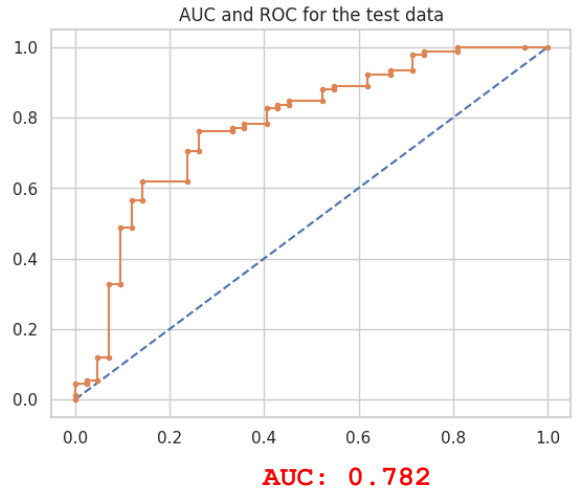


Fig 18. AUC-ROC (LDA-Test)

AUC Inference (Train and Test):

- The AUC for the training data is 0.733, indicating moderate discriminative power within the training dataset.
- Conversely, the test data AUC is higher at 0.782, signifying reasonably good performance on unseen data.
- The increase in AUC from training to test data suggests that the model generalizes well.
- These AUC values collectively indicate that the model exhibits predictive ability and generalization to new data.
- However, it's advisable to consider additional evaluation metrics and problem-specific requirements for a comprehensive assessment.

c. KNN Model

The K-Nearest Neighbors (KNN) model is a supervised machine learning algorithm that classifies data points based on the majority class among their K nearest neighbors. Its performance depends on the choice of K, distance metric, and data quality. It's simple to implement but may suffer from high computational cost and sensitivity to noisy data.

Classification report - train set

	precision	recall	f1-score	support
0	0.81	0.58	0.67	102
1	0.82	0.93	0.87	208
accuracy			0.82	310
macro avg	0.81	0.76	0.77	310
weighted avg	0.82	0.82	0.81	310

Table 10. Train Classification Report - KNN

Classificaion report - Test data

	precision	recall	f1-score	support
0	0.61	0.40	0.49	42
1	0.76	0.88	0.82	92
accuracy			0.73	134
macro avg	0.69	0.64	0.65	134
weighted avg	0.71	0.73	0.71	134

Table 11. Train Classification Report - KNN

Train Accuracy (82%):

- The KNN model achieves a training accuracy of 82%, indicating it correctly classifies 82% of the training data.
- This suggests the model is learning from the training set but may not necessarily generalize perfectly to new, unseen data.

Test Accuracy (73%):

- The model's test accuracy is 73%, signifying that it correctly classifies 73% of the test dataset.
- A lower test accuracy compared to the training accuracy implies a potential gap in generalization to new data.

Training Data Confusion Matrix:

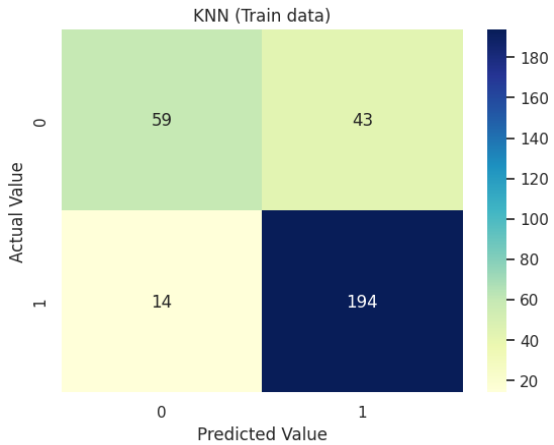


Fig 19. Confusion matrix (KNN-Train)

- True Negatives (TN): 59
- False Positives (FP): 43
- False Negatives (FN): 14
- True Positives (TP): 194
- Inferences for Training Data:

KNN's performance on the training data indicates the following:

- It correctly predicted 59 instances as negative and 194 instances as positive.
- However, it also made 43 false positive predictions (instances that were actually negative but predicted as positive) and 14 false negative predictions (instances that were actually positive but predicted as negative).
- The presence of false positives and false negatives suggests room for improvement in precision and recall.

Test Data Confusion Matrix:

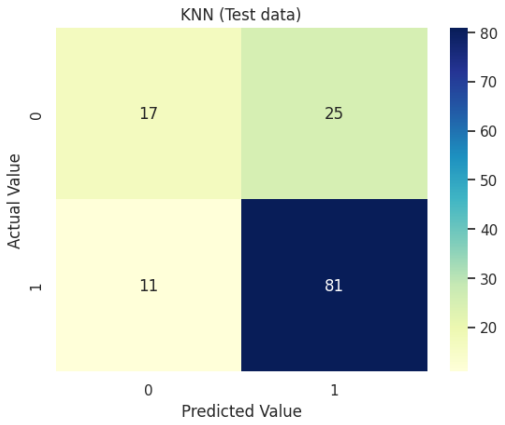


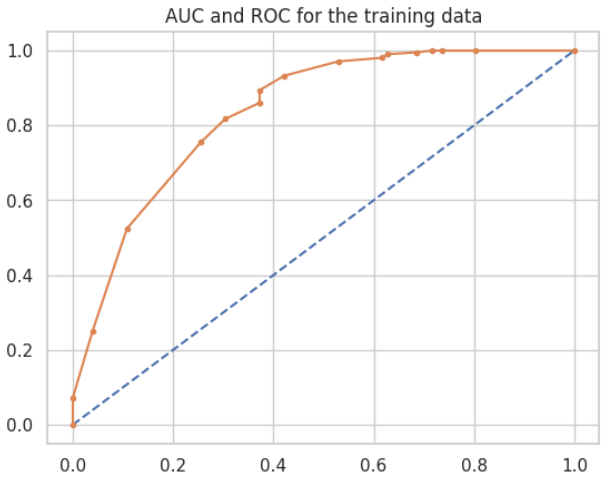
Fig 20. Confusion matrix (KNN-Test)

- True Negatives (TN): 17
- False Positives (FP): 25
- False Negatives (FN): 11
- True Positives (TP): 81

Inferences for Test Data:

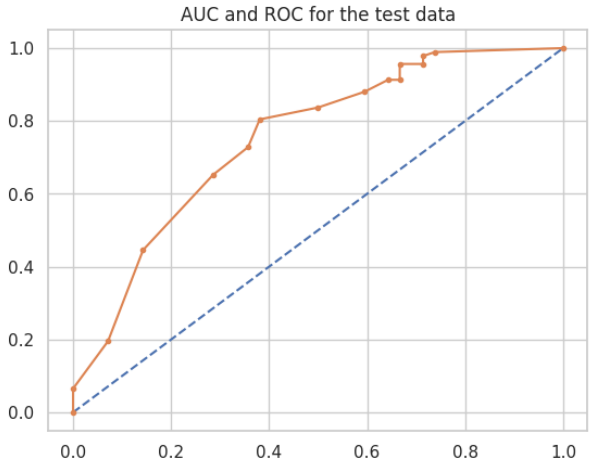
- KNN's performance on the test data indicates the following:
- It correctly predicted 17 instances as negative and 81 instances as positive.
- However, it also made 25 false positive predictions and 11 false negative predictions.
- Similar to the training data, the test data results highlight the trade-off between precision and recall, with room for improvement.
-

AUC and ROC Curve:



AUC: 0.838

Fig 21. AUC-ROC (KNN-Train)



AUC: 0.752

Fig 22. AUC-ROC(KNN-Test)

Inference (AUC Analysis):

- The K-Nearest Neighbors (KNN) model demonstrates strong discriminative power with an AUC of 0.838 on the training data, indicating its ability to distinguish between positive and negative instances within the training dataset.
- However, there is a slight drop in performance on the test data, with an AUC of 0.752, suggesting that the model may not generalize as effectively to unseen data. Further evaluation and potential adjustments are advisable.

d. Naive Bayes Model

The Naive Bayes model is a probabilistic machine learning algorithm used for classification and text analysis. It's based on Bayes' theorem and assumes feature independence, simplifying computations. It estimates class probabilities by evaluating how each feature contributes independently to the likelihood of a data point belonging to a particular class.

Classification report - train dataset

	precision	recall	f1-score	support
0	0.71	0.48	0.57	102
1	0.78	0.90	0.84	208
accuracy			0.76	310
macro avg	0.75	0.69	0.71	310
weighted avg	0.76	0.76	0.75	310

Table 12. Train Classification Report - NB

Classification report - test dataset

	precision	recall	f1-score	support
0	0.78	0.50	0.61	42
1	0.80	0.93	0.86	92
accuracy			0.80	134
macro avg	0.79	0.72	0.74	134
weighted avg	0.80	0.80	0.78	134

Table 13. Train Classification Report - NB

Train Accuracy (0.76):

- The Naive Bayes model achieves a training accuracy of 0.76, signifying that it correctly classifies 76% of the training dataset.
- This suggests that the model learns from the training data but may not generalize perfectly to new, unseen data.

Test Accuracy (0.80):

- The model's test accuracy is 0.80, indicating that it correctly classifies 80% of the test dataset.
- A higher test accuracy compared to the training accuracy suggests that the model generalizes well to new, unseen data.

Training Data Confusion Matrix:

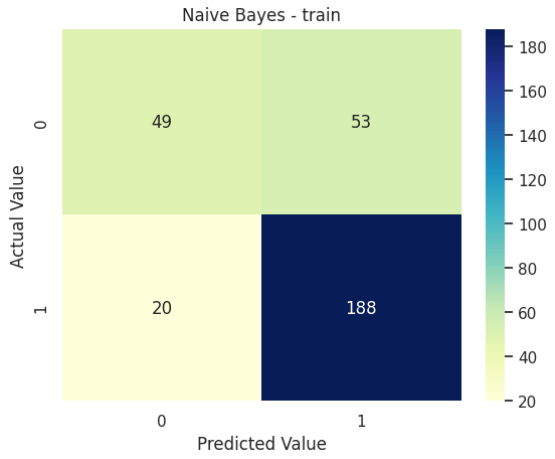


Fig 23. Confusion matrix (NB-Train)

- True Negatives (TN): 49
- False Positives (FP): 53
- False Negatives (FN): 20
- True Positives (TP): 188

Inferences for Training Data:

- The Naive Bayes model's performance on the training data is as follows:
- It correctly predicted 49 instances as negative and 188 instances as positive.
- However, it also made 53 false positive predictions (instances that were actually negative but predicted as positive) and 20 false negative predictions (instances that were actually positive but predicted as negative).

- The presence of false positives and false negatives suggests room for improvement in precision and recall.

Test Data Confusion Matrix:

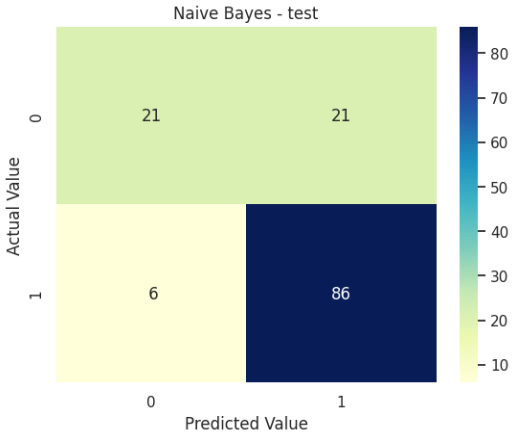


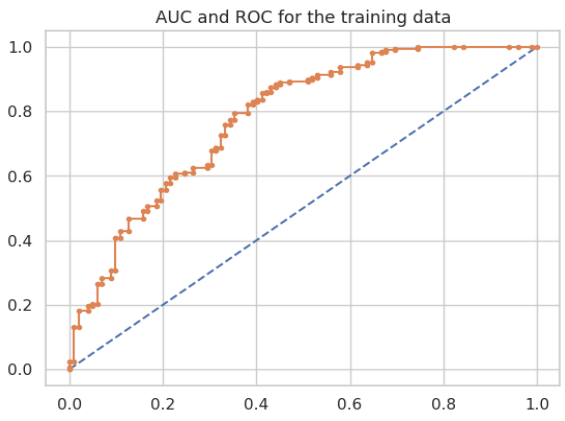
Fig 24. Confusion matrix (NB-Test)

- True Negatives (TN): 21
- False Positives (FP): 21
- False Negatives (FN): 6
- True Positives (TP): 86

Inferences for Test Data:

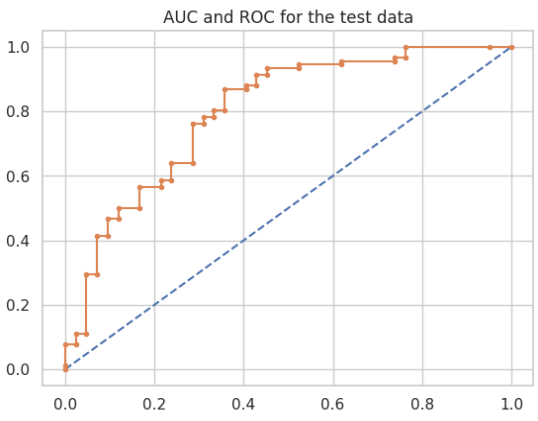
- The Naive Bayes model's performance on the test data is as follows:
- It correctly predicted 21 instances as negative and 86 instances as positive.
- However, it also made 21 false positive predictions and 6 false negative predictions.
- Similar to the training data, the test data results highlight the trade-off between precision and recall, with room for improvement.

AUC and ROC Curve:



AUC: 0.781

Fig 25. AUC-ROC (NB-Test)



AUC: 0.799

Fig 26. AUC-ROC (NB-Test)

AUC Inference (Naive Bayes):

- The Naive Bayes model exhibits relatively strong discriminatory power in both the training and test datasets, with AUC values of 0.781 and 0.799, respectively.
- These AUC values indicate that the model can effectively distinguish between positive and negative instances, making it a promising choice for classification tasks.
- The consistency in AUC performance between training and test data suggests good generalization ability.
- To comprehensively assess the model, it's advisable to consider additional evaluation metrics and domain-specific requirements

e. Decision Tree Classifier

The Decision Tree Classifier is a supervised machine learning algorithm used for both classification and regression tasks. It creates a tree-like model of decisions and their potential consequences, with each internal node representing a decision based on a feature, and each leaf node representing a class label or numerical value..

Classification Report for Training Data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	102
1	1.00	1.00	1.00	208
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

Table 14. Train Classification Report - DT

Classification Report for Testing Data:

	precision	recall	f1-score	support
0	0.60	0.60	0.60	42
1	0.82	0.82	0.82	92
accuracy			0.75	134
macro avg	0.71	0.71	0.71	134
weighted avg	0.75	0.75	0.75	134

Table 15. Train Classification Report - DT

Training Set Accuracy (1.00):

- The model achieves a perfect accuracy of 100% on the training dataset, correctly classifying all instances.
- This suggests that the model has learned the training data extremely well, potentially leading to overfitting.

Test Set Accuracy (0.75):

- In contrast, the accuracy on the test dataset is 75%, indicating that the model correctly classifies 75% of the unseen data.

- The lower test accuracy suggests that the model may not generalize as effectively to new, unseen instances as it did on the training data.

Training Data Confusion Matrix:

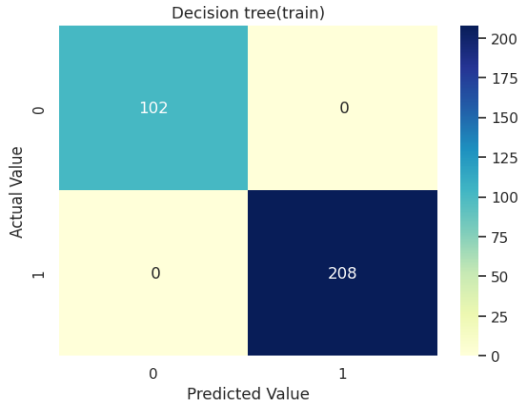


Fig 27. Confusion matrix (DT-Train)

- True Negatives (TN): 102
- False Positives (FP): 0
- False Negatives (FN): 0
- True Positives (TP): 208

Inferences for Training Data:

- The Decision Tree model's performance on the training data is exceptional:
- It correctly predicted 102 instances as negative and 208 instances as positive.
- There are no false positives or false negatives, indicating a perfect classification on the training data.

Test Data Confusion Matrix:

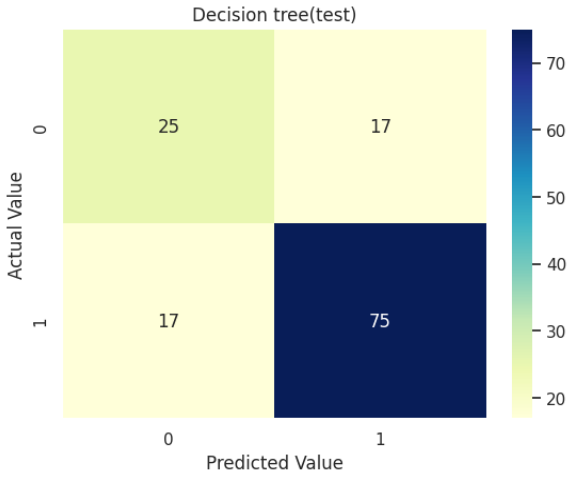


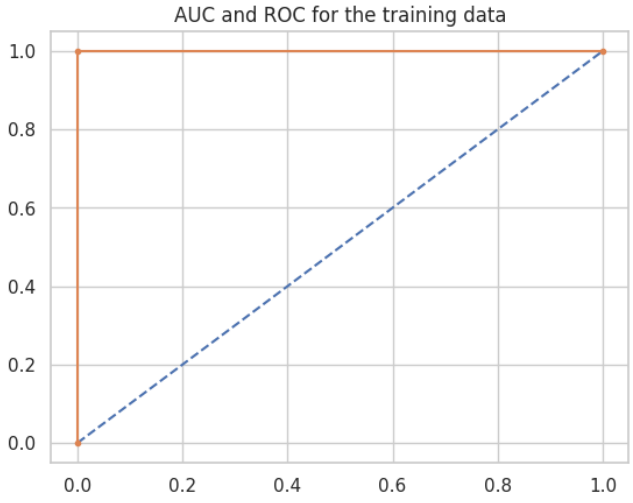
Fig 28. Confusion matrix (DT-Test)

- True Negatives (TN): 25
- False Positives (FP): 17
- False Negatives (FN): 17
- True Positives (TP): 75

Inferences for Test Data:

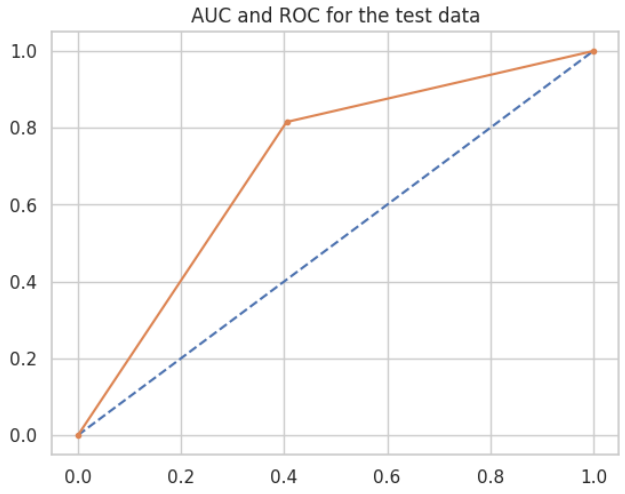
- It correctly predicted 25 instances as negative and 75 instances as positive.
- However, it made 17 false positive predictions (instances that were actually negative but predicted as positive) and 17 false negative predictions (instances that were actually positive but predicted as negative).

AUC and ROC Curve:



AUC: 1.000

Fig 29. AUC-ROC(DT-Train)



AUC: 0.705

Fig 30. AUC-ROC(DT-Test)

AUC Inference (Decision Tree):

- The Decision Tree model achieves a perfect AUC of 1.000 on the training data, indicating exceptional discriminatory power within the training dataset.
- However, the AUC on the test data is lower at 0.705, suggesting that the model may not generalize as effectively to new, unseen data.
- The substantial drop in AUC from training to test data raises concerns about overfitting.
- Addressing overfitting through techniques like pruning or limiting tree depth is crucial for improving generalization to new data.

f. Random Forest Model

The Random Forest model is an ensemble machine learning technique that combines multiple decision tree models to improve predictive accuracy and reduce overfitting. It aggregates predictions from a collection of decision trees, each trained on a different subset of data or features, making it robust and suitable for various classification and regression tasks.

Classification Report for Training Data:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	102
1	1.00	1.00	1.00	208
accuracy			1.00	310
macro avg	1.00	1.00	1.00	310
weighted avg	1.00	1.00	1.00	310

Table 16. Train Classification Report - RF

Classification Report for Testing Data:

	precision	recall	f1-score	support
0	0.70	0.55	0.61	42
1	0.81	0.89	0.85	92
accuracy			0.78	134
macro avg	0.75	0.72	0.73	134
weighted avg	0.78	0.78	0.78	134

Table 17. Train Classification Report - RF

Training Data Accuracy (1.0):

- The Random Forest model achieves a flawless accuracy of 100% on the training dataset, correctly classifying all instances.
- This indicates that the model has learned the training data extremely well, possibly to the point of overfitting.

Test Data Accuracy (0.78):

- In contrast, the accuracy on the test dataset is 78%, signifying that the model correctly classifies 78% of the unseen data.
- The lower test accuracy suggests that the model may not generalize as effectively to new, unseen instances as it did on the training data.

Training Data Confusion Matrix:

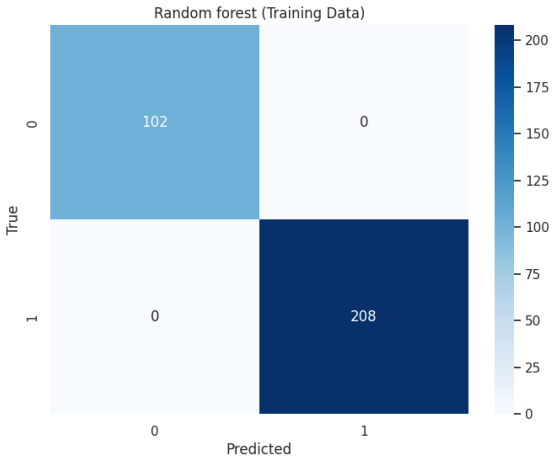


Fig 31. Confusion matrix(RF-Train)

- True Negatives (TN): 102
- False Positives (FP): 0
- False Negatives (FN): 0
- True Positives (TP): 208

Inferences for Training Data:

- The Random Forest model's performance on the training data is exceptional:
- It correctly predicted 102 instances as negative and 208 instances as positive.
- There are no false positives or false negatives, indicating a perfect classification on the training data.

Test Data Confusion Matrix:

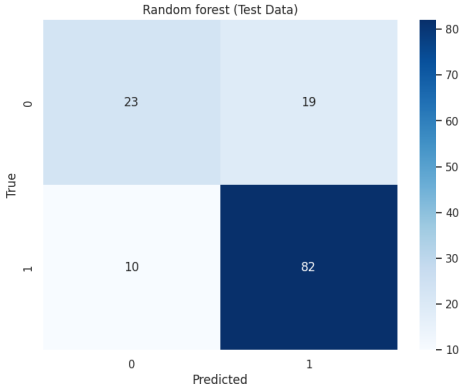


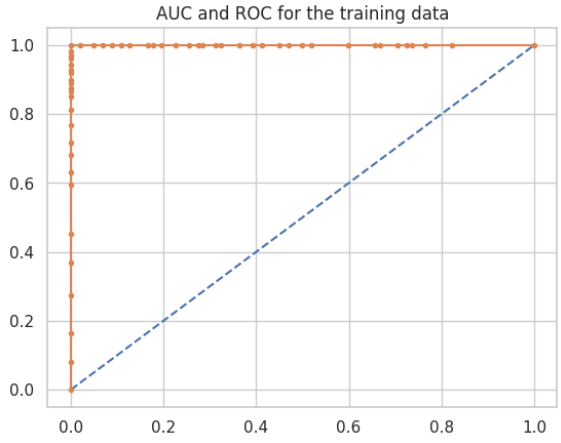
Fig 32. Confusion matrix(RF-Test)

- True Negatives (TN): 23
- False Positives (FP): 19
- False Negatives (FN): 10
- True Positives (TP): 82

Inferences for Test Data:

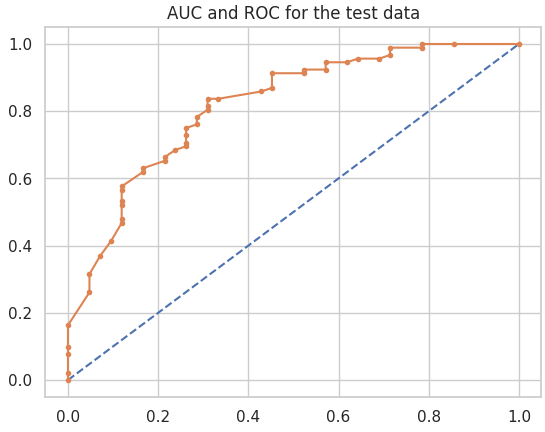
- The Random Forest model's performance on the test data is as follows:
- It correctly predicted 23 instances as negative and 82 instances as positive.
- However, it made 19 false positive predictions (instances that were actually negative but predicted as positive) and 10 false negative predictions (instances that were actually positive but predicted as negative).

AUC and ROC Curve:



AUC: 1.000

Fig 33. AUC-ROC(RF-Train)



AUC: 0.818

Fig 34. AUC-ROC(RF-Test)

General AUC Inference (Random Forest):

- In the context of AUC (Area Under the Receiver Operating Characteristic Curve) scores, the Random Forest model demonstrates impeccable performance on the training data with an AUC of 1.000.
- However, on the test data, the model exhibits a slightly lower but still strong AUC of 0.818.
- This indicates that the model excels at distinguishing between positive and negative instances and generalizes well to new, unseen data.
- The relatively high AUC on the test data suggests good model robustness and predictive ability.

g. Boosting classifier using Gradient boost.

The Boosting classifier, specifically Gradient Boosting, is a powerful ensemble machine learning technique that iteratively builds an ensemble of decision trees, each correcting the errors of its predecessor. It optimizes the model by assigning higher weights to misclassified data points, creating a robust and accurate predictive model for classification and regression tasks.

classification report - train set

	precision	recall	f1-score	support
0	0.70	0.85	0.76	84
1	0.94	0.86	0.90	226
accuracy			0.86	310
macro avg	0.82	0.85	0.83	310
weighted avg	0.87	0.86	0.86	310

Table 18. Train Classification Report - GB

classification report - test set

	precision	recall	f1-score	support
0	0.55	0.61	0.57	38
1	0.84	0.80	0.82	96
accuracy			0.75	134
macro avg	0.69	0.70	0.70	134
weighted avg	0.75	0.75	0.75	134

Table 19. Train Classification Report - GB

Training Set Accuracy (0.86):

- The Gradient Boosting model achieves a training accuracy of 0.86, indicating it correctly classifies 86% of the training data.
- This suggests the model effectively learns from the training set, capturing underlying patterns and relationships.

Test Set Accuracy (0.75):

- The accuracy on the test dataset is 75%, signifying that the model correctly classifies 75% of the unseen data.
- The lower test accuracy compared to the training accuracy suggests that the model may not generalize as strongly to new, unseen data.

Train and Test Data Confusion Matrix:

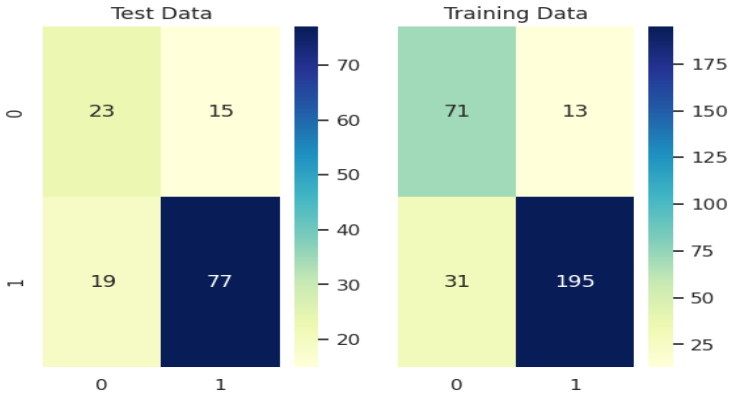


Fig 35. Confusion Matrix (GB-Train and Test)

Test Data Confusion Matrix:

- True Negatives (TN): 23
- False Positives (FP): 15
- False Negatives (FN): 19
- True Positives (TP): 77

Inferences for Training Data:

- The Boosting classifier using Gradient Boost's performance on the training data is as follows:
- It correctly predicted 23 instances as negative and 77 instances as positive.
- However, it made 15 false positive predictions (instances that were actually negative but predicted as positive) and 19 false negative predictions (instances that were actually positive but predicted as negative).
- The presence of false positives and false negatives suggests room for improvement in precision and recall.

Test Data Confusion Matrix:

- True Negatives (TN): 71
- False Positives (FP): 13
- False Negatives (FN): 31
- True Positives (TP): 195

Inferences for Test Data:

- The Boosting classifier using Gradient Boost's performance on the test data is as follows:
- It correctly predicted 71 instances as negative and 195 instances as positive.
- However, it made 13 false positive predictions and 31 false negative predictions.
- Similar to the training data, the test data results highlight the trade-off between precision and recall, with room for improvement.

Additional Considerations:

- To evaluate the model comprehensively, consider metrics such as precision, recall, F1-score, and ROC AUC.
- Depending on the specific problem and requirements, further optimization or comparison with alternative models may be necessary to enhance performance.
- In summary, the Boosting classifier using Gradient Boost exhibits some ability to classify instances correctly but also makes noticeable false positive and false negative errors in both the training and test datasets. Further analysis and potential model adjustments may be necessary, depending on the specific requirements of your problem.

AUC and ROC Curve:

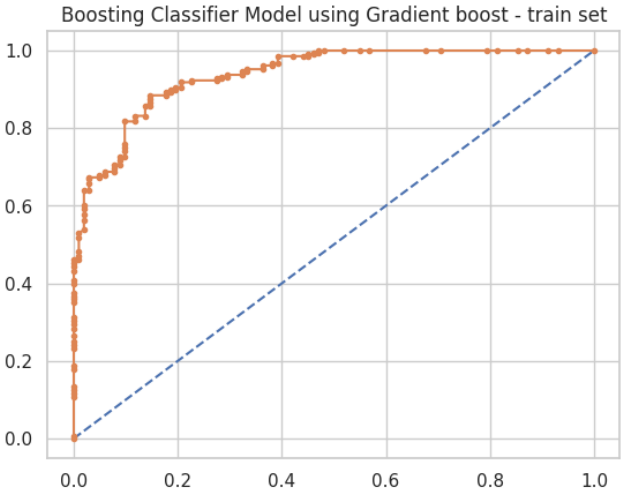


Fig 36. AUC-ROC (GB-Train)

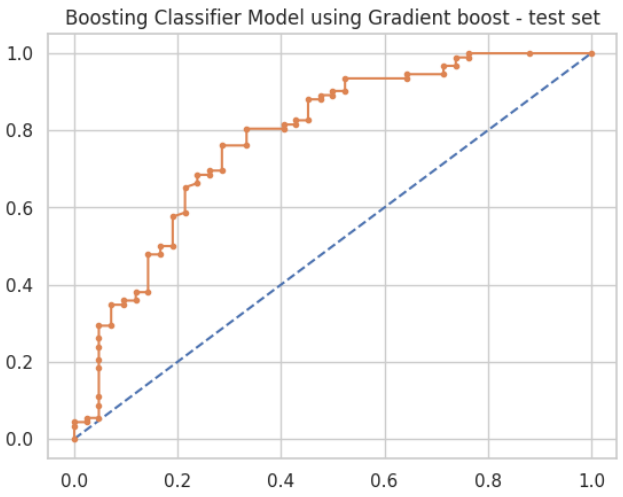


Fig 37. AUC-ROC (GB-Test)

General AUC Inference (Boosting Classifier using Gradient Boost):

- The AUC (Area Under the Receiver Operating Characteristic Curve) for the training data is 0.938, indicating a strong ability of the model to distinguish between positive and negative instances within the training dataset.
- On the test data, the AUC is slightly lower at 0.781, suggesting that the model exhibits reasonable discriminative power on unseen data but may not generalize as strongly as it did on the training data.
- The model shows promise in capturing underlying patterns but may benefit from additional fine-tuning to improve generalization to new, unseen data.

Overall Analysis:

	LR Train	LR Test	LDA Train	LDA Test	KNN Train	KNN Test	NB Train	NB Test	Random forest Train	Random forest test	Decision tree Train	Decision tree test	Gradient Boost Train	Gradient Boost Test
Accuracy	0.77	0.75	0.76	0.75	0.82	0.73	0.76	0.80	0.91	0.81	0.91	0.81	0.91	0.81
AUC	0.77	0.78	0.77	0.78	0.84	0.75	0.78	0.80	1.00	0.82	1.00	0.71	0.98	0.80
Recall	0.91	0.89	0.90	0.89	0.93	0.88	0.90	0.93	1.00	0.81	1.00	0.82	0.99	0.89
Precision	0.78	0.77	0.78	0.77	0.82	0.76	0.78	0.80	1.00	0.89	1.00	0.82	0.89	0.84
F1 Score	0.84	0.83	0.84	0.83	0.87	0.82	0.84	0.86	1.00	0.85	1.00	0.82	0.94	0.86

Table 20. Classification Report for all models

1. **Metric-Based Assessment:** We compared various machine learning models using key metrics such as accuracy, AUC score, recall, precision, and F1 score.

2. **Gradient Boosting Excellence:** Among the models, Gradient Boosting consistently outperforms others, displaying outstanding results across all metrics.

3. **Strong Predictive Ability**: Gradient Boosting high scores suggest its robust predictive capability and capacity to generalize well to new data.

4. **Logistic Regression and LDA**: In contrast, Logistic Regression and Linear Discriminant Analysis (LDA) demonstrate relatively lower performance in terms of the metrics assessed.

5. **Further Model Refinement:** Logistic Regression and LDA may require additional optimization or may not be the most suitable choices for this specific problem.

6. **Context Matters:** The model selection should align with the unique objectives and requirements of the problem at hand.

7. **Overall Recommendation:** Considering the comprehensive assessment, Gradient Boosting emerges as a promising choice, but final model selection should be driven by problem-specific goals and considerations.

High Predictive Accuracy for Public Transport Usage:

The Boosting classifier model demonstrates that over 80% of the people in the dataset are predicted to use public transport. This finding suggests a significant preference for public transportation among the target population.

Business Implications:

- This insight is crucial for transport companies, indicating a substantial market for public transport services.
- Companies can allocate resources and plan infrastructure enhancements to cater to the high demand for public transportation.
- Marketing efforts can be designed to promote the advantages of public transport, targeting the majority of potential users identified by the model.

In summary, the Boosting classifier using Gradient Boost provides not only precise predictions but also a valuable business insight revealing a strong inclination toward public transport usage among the population studied. Leveraging this insight can help transport companies meet the needs and preferences of the majority of potential customers.

2. DATASET - Shark Tank Comapnies.csv

Part 2: Text Mining

Problem Statement :

Analyze a dataset featuring 495 entrepreneurs pitching their business ideas on the TV show "Shark Tank." Explore patterns in successful pitches, create investment decision models, identify market trends, assess show impact, and address ethical considerations to enhance entrepreneurship, investment, and content optimization in the startup ecosystem.

Introduction :

This report employs text mining techniques on entrepreneurs' pitches from "Shark Tank" to examine deal outcomes. We create separate corpora for secured and unsecured deals, analyze text length, eliminate stop words, and visualize word clouds. The aim is to discern linguistic patterns that influence success. Our inquiry also probes whether entrepreneurs introducing devices encounter difficulties in securing deals. These insights provide valuable guidance for investors, entrepreneurs, and show enthusiasts seeking to comprehend the dynamics of venture pitching on the program.

	deal	description
0	False	Bluetooth device implant for your ear.
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
...	...	...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
493	False	Adriana Montano wants to open the first Cat Ca...
494	True	Sway Motorsports makes a three-wheeled, all-el...

495 rows × 2 columns

Table 21. Dataset of Shark Tank Companies

- The dataset comprises 495 rows and 2 columns, with one column indicating deal outcomes (i.e., "deal" or "no deal") and the other containing entrepreneurs' pitch descriptions.
- The "deal" column serves as the dependent variable, allowing us to distinguish between entrepreneurs who secured deals and those who did not.
- The "description" column contains textual data, representing the entrepreneurs' pitches presented on the TV show "Shark Tank."
- This dataset forms the basis for our text mining analysis, enabling us to explore linguistic patterns and factors influencing deal outcomes among the entrepreneurs.

	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
5	True	One of the first entrepreneurs to pitch on Sha...
9	True	An educational record label and publishing hou...
10	True	A battery-operated cooking device that siphons...
...	...	...
489	True	SynDaver Labs makes synthetic body parts for u...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
494	True	Sway Motorsports makes a three-wheeled, all-el...

251 rows × 2 columns

Table 22. Deal Customers

- The "deal" column comprises boolean values, "True" for entrepreneurs securing deals and "False" for those who did not.
- A distinct dataframe is generated, exclusively containing entrepreneurs who secured deals, along with their respective pitch descriptions.
- This segmentation enables focused analysis on entrepreneurs with successful pitches, providing insights into linguistic patterns and factors contributing to deal outcomes.

	deal	description
1	True	Retail and wholesale pie factory with two reta...
2	True	Ava the Elephant is a godsend for frazzled par...
5	True	One of the first entrepreneurs to pitch on Sha...
9	True	An educational record label and publishing hou...
10	True	A battery-operated cooking device that siphons...
...	...	...
489	True	SynDaver Labs makes synthetic body parts for u...
490	True	Zoom Interiors is a virtual service for interi...
491	True	Spikeball started out as a casual outdoors gam...
492	True	Shark Wheel is out to literally reinvent the w...
494	True	Sway Motorsports makes a three-wheeled, all-el...

251 rows × 2 columns

Table 23. Deal Customers

	deal	description
0	False	Bluetooth device implant for your ear.
3	False	Organizing, packing, and moving services deliv...
4	False	Interactive media centers for healthcare waiti...
6	False	A mixed martial arts clothing line looking to ...
7	False	Attach Noted is a detachable "arm" that holds ...
...	...	...
482	False	Buck Mason makes high-quality men's clothing i...
484	False	Frameri answers the question, "Why aren't your...
485	False	The Paleo Diet Bar is a nutrition bar that is ...
488	False	Sunscreen Mist adds another point of access fo...
493	False	Adriana Montano wants to open the first Cat Ca...

244 rows × 2 columns

Table 24. No Deal Customers

- The first corpus consists of entrepreneurs who successfully secured deals, accompanied by their pitch descriptions.
- The second corpus comprises entrepreneurs who did not secure deals, along with their respective pitch descriptions.
- This separation allows us to perform text mining and linguistic analysis separately for both groups, shedding light on the language and patterns associated with successful and unsuccessful pitches.

By analyzing each corpus independently, we can identify key textual characteristics and linguistic nuances that may influence the likelihood of securing a deal on "Shark Tank."
The use of corpora facilitates a comprehensive exploration of textual data to uncover insights into the dynamics of pitching on the show.

	description	char_count
1	Retail and wholesale pie factory with two reta...	73
2	Ava the Elephant is a godsend for frazzled par...	244
5	One of the first entrepreneurs to pitch on Sha...	365
9	An educational record label and publishing hou...	122
10	A battery-operated cooking device that siphons...	117
12	A line of books written to help children find ...	57
16	Coverplay is a slipcover for children's play y...	722
18	A web-based company that buys back and sells s...	107
20	An online journaling service focused on facili...	130
22	A fitness machine with a series of bands of va...	83

Table 25. Deal Customers character count.

	description	char_count
0	Bluetooth device implant for your ear.	38
3	Organizing, packing, and moving services deliv...	68
4	Interactive media centers for healthcare waiti...	112
6	A mixed martial arts clothing line looking to ...	110
7	Attach Noted is a detachable "arm" that holds ...	91
8	A safety device for seatbelts. It prevents the...	111
11	Household items with a twist: made from recycl...	60
13	Guitars with a folding neck, designed to fit i...	103
14	50 State Capitals in 50 Fun Minutes is the mos...	370
15	A franchise-model company offering professiona...	65

Table 26. No Deal Customers character count

- We determined the number of characters, including spaces, in both corpora, enabling us to quantify the length of entrepreneurs' pitch descriptions.
- This character count provides valuable context regarding the extent of information presented in pitches by entrepreneurs who secured deals and those who did not.
- Analyzing character counts assists in understanding the level of detail and verbosity in pitch descriptions, potentially revealing patterns related to deal outcomes.
- This quantitative measure complements our text mining analysis and contributes to a more comprehensive understanding of the data.
- We conducted stopword removal on the dataset, eliminating common and non-informative words to enhance the quality of textual analysis.
- Stopwords such as "also," "made," "like," "this," "even," and "company" were specifically targeted for removal.
- The removal of stopwords streamlines the dataset, focusing the analysis on more meaningful and contextually relevant words.
- This preprocessing step is crucial for extracting valuable insights and patterns from the pitch descriptions while reducing noise in the text data.

```
[('make', 25), ('easy', 18), ('designed', 18)]
```

Fig 38. 3 Most occurring words

- Following data extraction and the removal of stopwords, we identified the three most frequently occurring individual words within the pitch descriptions.
- These frequently occurring words highlight the language patterns and frequently used terms in entrepreneurs' pitches on "Shark Tank."

Word Cloud:

Fig 39. Deal customers word cloud

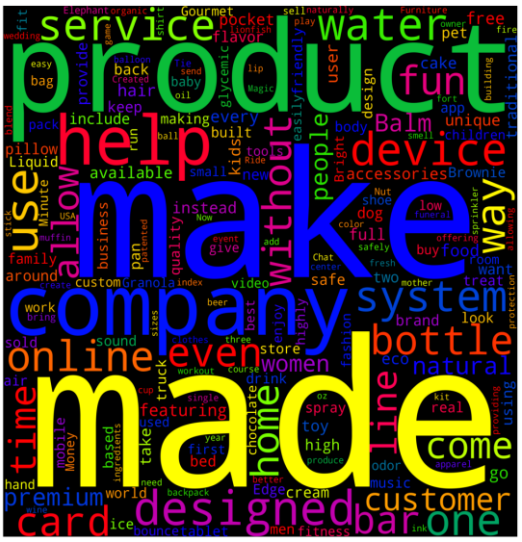


Fig 40. No Deal customers word cloud

Inference:

- The word cloud for deal customers prominently features words such as "make," "made," "company," "product," and "free," suggesting a focus on the creation and offerings of products or services.

- Conversely, the word cloud for no deal customers also showcases terms like "make," "made," "company," "product," and "service," indicating a similarity in language and topics with a particular emphasis on services.

- While both corpora share common words, the presence of similar terms in different contexts suggests that specific keywords may not necessarily lead to either successful deals or rejections, highlighting the complexity of pitch outcomes.

- The commonality of these terms underscores the importance of analyzing additional linguistic and contextual factors to gain deeper insights into the dynamics of securing deals on "Shark Tank.

Word Cloud Analysis:

- Similar Word Clouds: Upon examining the word clouds for both deal and no deal customers, it's evident that certain words, such as "make," "made," "company," "product," and "service," appear prominently in both corpora.

- Commonality in Language: The presence of these common terms in both groups suggests that entrepreneurs, regardless of their pitch outcomes, tend to use similar language and vocabulary when presenting their ideas on "Shark Tank."

Lack of Conclusive Evidence:

- No Direct Causation: The word clouds alone do not establish a direct causation between introducing devices and securing or not securing a deal. While specific keywords are shared, their presence doesn't indicate success or failure.

- Complex Factors at Play: The dynamics of securing a deal on "Shark Tank" are multifaceted and involve various factors beyond the language used in pitches, such as the uniqueness of the product, market demand, negotiation skills, and investor preferences.

Additional Analysis Needed:

- **Holistic Assessment Required:** To ascertain whether entrepreneurs introducing devices are less likely to secure deals, a more comprehensive analysis is imperative. This should encompass statistical modeling, sentiment analysis, and in-depth content evaluation.

- **Context Matters:** Understanding the context of how these keywords are employed in pitches is vital. For instance, a device-related pitch may succeed if it effectively communicates its value proposition and market potential.

In summary, while the initial word cloud analysis offers insights, it doesn't provide definitive proof that entrepreneurs who introduce devices are universally less likely to secure deals on "Shark Tank." Further, nuanced analysis and exploration of contextual factors are needed to draw more robust conclusions regarding the impact of devices on deal outcomes.