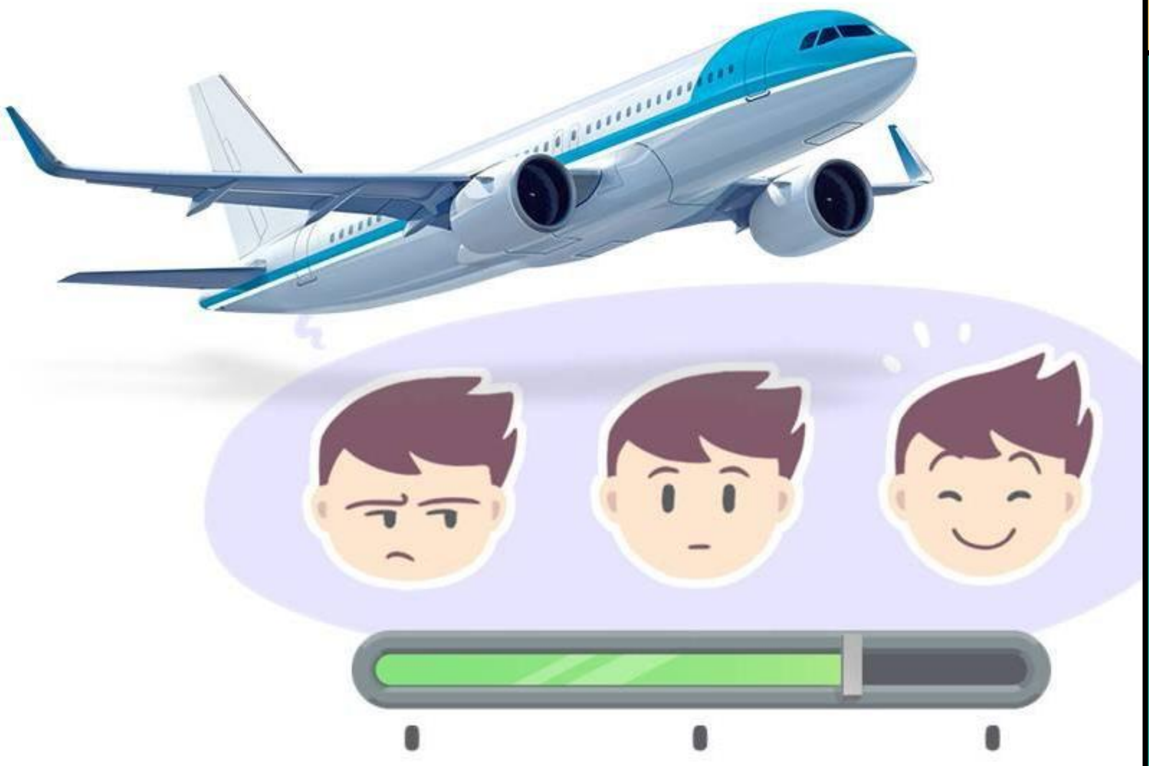
Airline Passenger Satisfaction

- Knowledge Discovery and Data Mining



Team Member

Name : Vanshaj Tyagi

CWID : 20029455

Name : Abhishek Esakkiappan

CWID : 20032119

Name : Havishey Jotaniya

CWID : 20030480

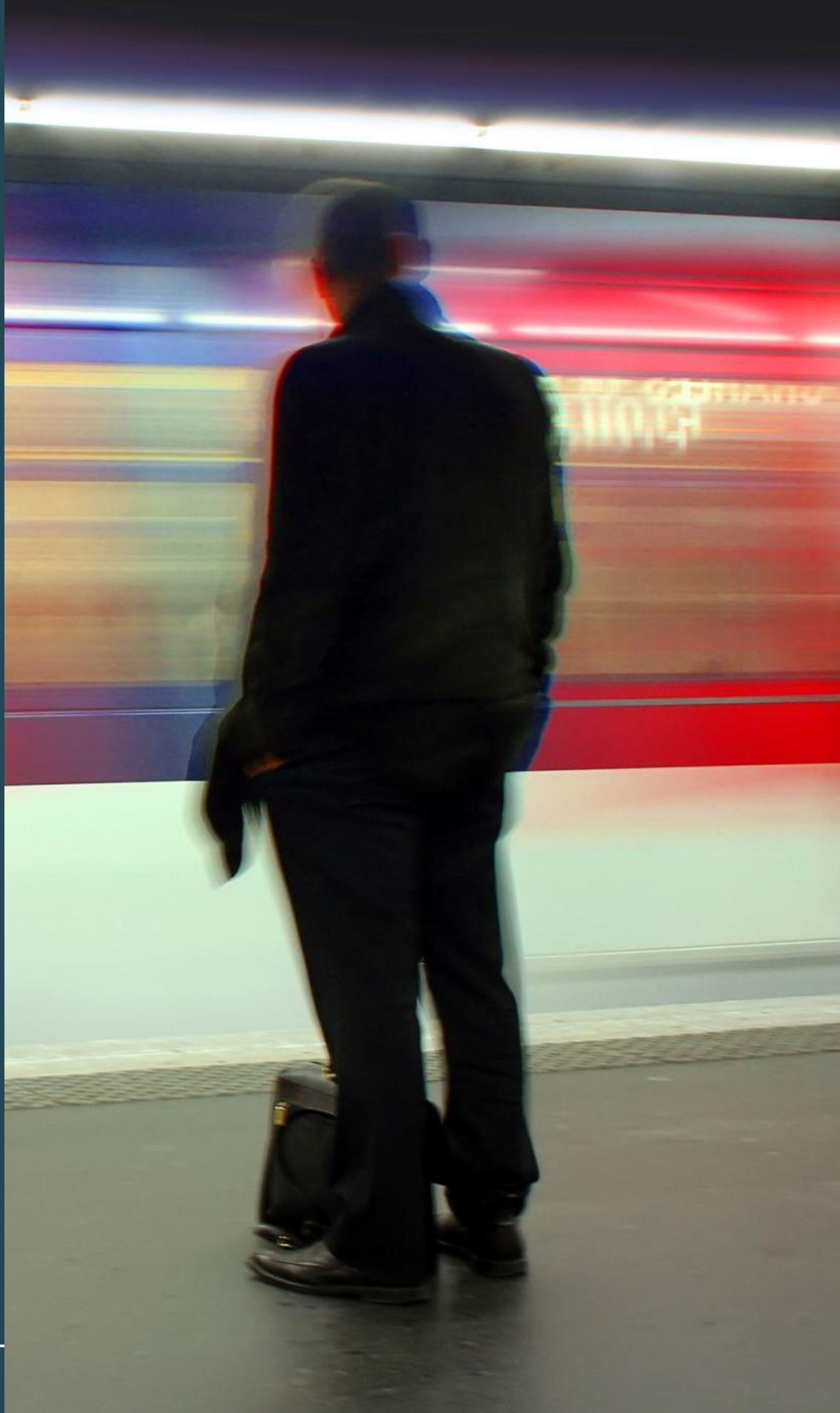
Agenda

- Introduction and Problem Statement
- Data Cleaning and Preprocessing
- Exploratory Data Analysis
- Model Overview
- Comparison and Conclusion

Introduction & Problem Statement

Goal: Analyse airline passenger satisfaction to identify factors influencing customer satisfaction and dissatisfaction.

Problem Statement: Airline companies face challenges in maintaining high customer satisfaction. This analysis aims to help improve service quality by identifying key drivers of satisfaction.



Introduction & Problem Statement

- **Problem Context**: The primary objective is to analyse various factors affecting airline passenger satisfaction and classify passengers as "Satisfied" or "Neutral/Dissatisfied.“
- **Data Utilization:** Utilizing a dataset that includes key demographic, travel, and service-related features (e.g., Type of Travel, Class, Inflight Service Scores, and Delays), the project seeks to process and analyse this data effectively to make accurate classifications.
- **Algorithm Evaluation:** Multiple machine learning algorithms, including Logistic Regression, K-Nearest Neighbours (KNN), Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine (SVM), Artificial Neural Networks (ANN), and Gradient Boosting Machines, are implemented and evaluated. The aim is to determine which algorithm or algorithms are most effective in this classification task.
- **Performance Metrics:** The effectiveness of each algorithm is assessed using key performance metrics such as accuracy, classification report, and confusion matrix. These metrics provide insights into how well each model performs in terms of both overall accuracy and its ability to minimize misclassification.
- **Comparative Analysis:** A comparative analysis of the results from each algorithm is conducted. This involves examining their performance metrics to identify strengths and weaknesses, thereby determining the most suitable model or models for accurately classifying passengers as Satisfied or Neutral/Dissatisfied.

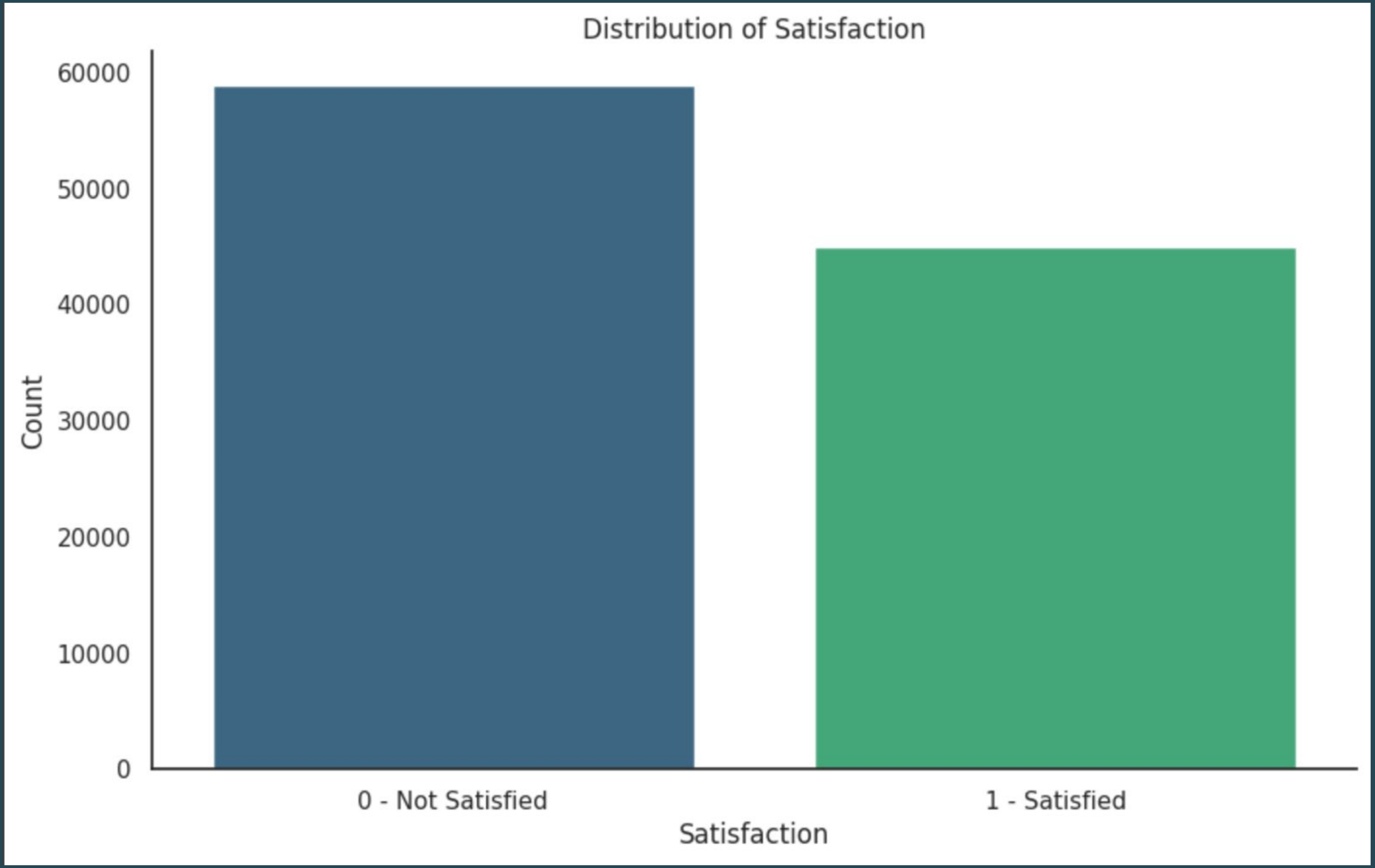
Dataset

Link For Dataset : Kaggle Flight Satisfaction

Data Set Overview

- Description of the Dataset
  - Dataset Source: Contains training data with 103,904 records and testing data with 25,976 records.
  - Features:
    - Demographics: Gender, Age
    - Customer info: Customer Type, Type of Travel, Class
    - Service ratings: Inflight wifi service, Food and drink, Seat comfort, etc.
    - Flight details: Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes
    - Target: satisfaction (Satisfied vs Neutral/Dissatisfied)
  - Challenges: Handling missing data, diverse feature types (categorical, numerical).

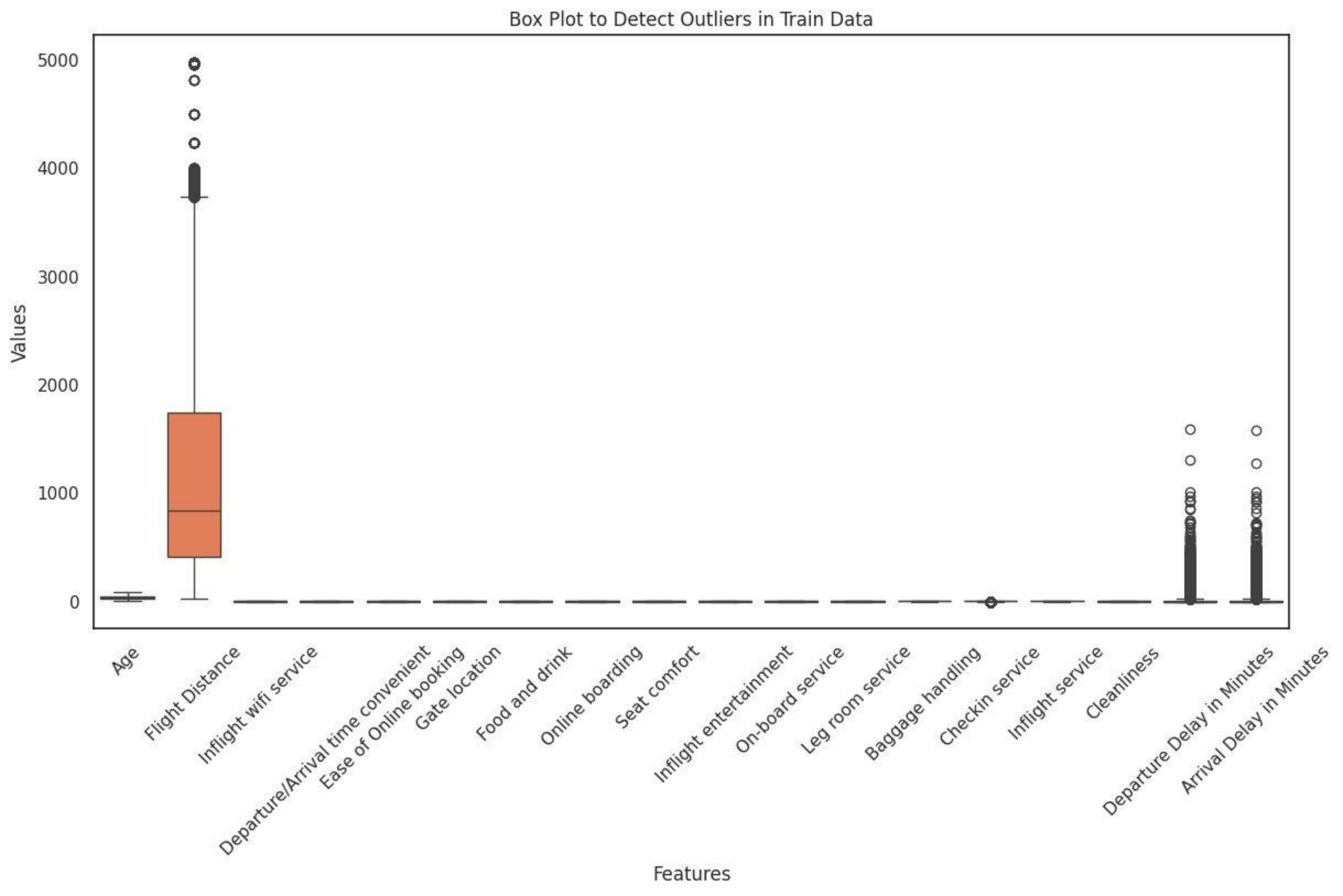
Exploratory Data Analysis 1



- The bar graph shows the distribution of satisfaction levels, where **0 = dissatisfied** and **1 = satisfied**. The **y-axis** represents the count of individuals in each category.

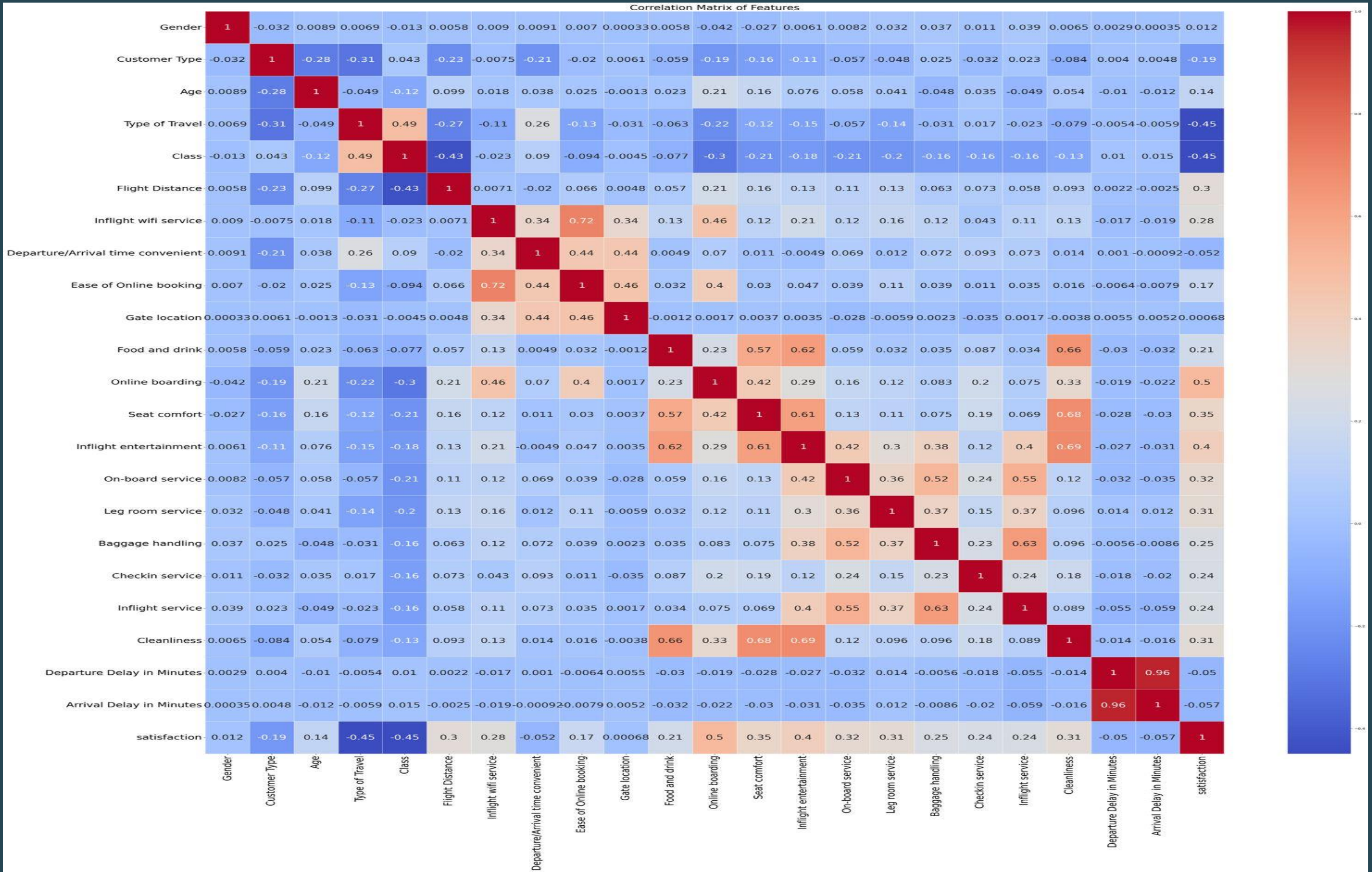
- The taller bar for **0** indicates that more individuals are dissatisfied compared to those satisfied (represented by the shorter bar for **1**).

Exploratory Data Analysis 2



This box plot reveals variability and outliers in the dataset features. Features like "Flight Distance" and "Departure/Arrival Delay in Minutes" show high variability and numerous outliers, indicating extreme cases that could skew analysis. In contrast, features like "Inflight wifi service" and "Cleanliness" have limited ranges and no significant outliers, suggesting more consistent data. Outliers in certain features may require preprocessing, such as removal or normalization, to enhance model performance. Features with stable distributions might indicate user ratings or low-scale numerical data, while highly variable features could have a greater impact on the outcomes of analysis or modelling tasks.

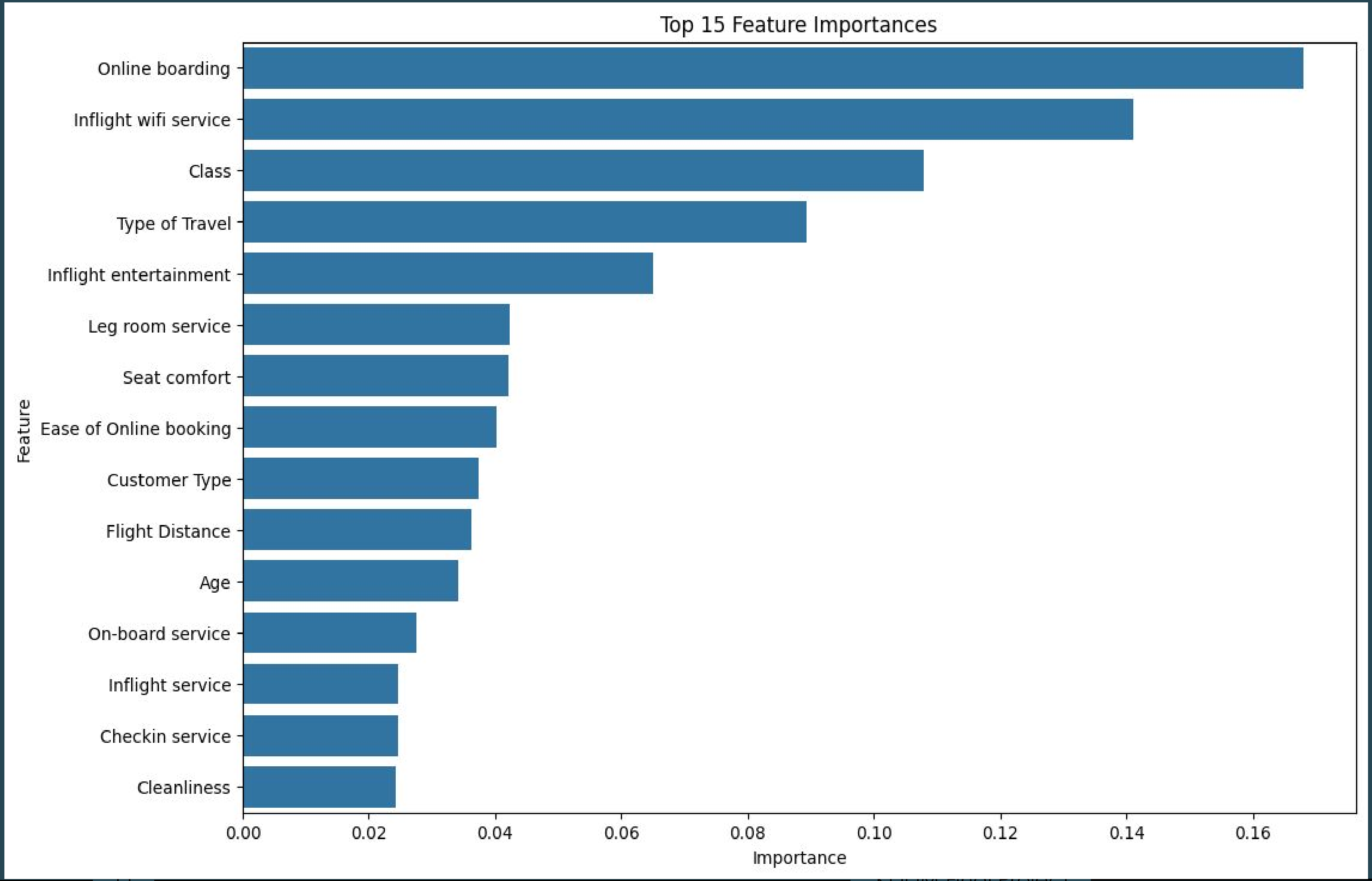
Exploratory Data Analysis 3



	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfaction
Gender	1	-0.032	0.0089	0.0069	-0.013	0.0058	0.009	0.0091	0.007	0.00033	0.0058	-0.042	-0.027	0.0061	0.0082	0.032	0.037	0.011	0.039	0.0065	0.0029	0.00035	0.012
Customer Type	-0.032	1	-0.28	-0.31	0.043	-0.23	-0.0075	-0.21	-0.02	0.0061	-0.059	-0.19	-0.16	-0.11	-0.057	-0.048	0.025	-0.032	0.023	-0.084	0.004	0.0048	-0.19
Age	0.0089	-0.28	1	-0.049	-0.12	0.099	0.018	0.038	0.025	-0.0013	0.023	0.21	0.16	0.076	0.058	0.041	-0.048	0.035	-0.049	0.054	-0.01	-0.012	0.14
Type of Travel	0.0069	-0.31	-0.049	1	0.49	-0.27	-0.11	0.26	-0.13	-0.031	-0.063	-0.22	-0.12	-0.15	-0.057	-0.14	-0.031	0.017	-0.023	-0.079	-0.0054	-0.0059	-0.45
Class	-0.013	0.043	-0.12	0.49	1	-0.43	-0.023	0.09	-0.094	-0.0045	-0.077	-0.3	-0.21	-0.18	-0.21	-0.2	-0.16	-0.16	-0.16	-0.13	0.01	0.015	-0.45
Flight Distance	0.0058	-0.23	0.099	-0.27	-0.43	1	0.0071	-0.02	0.066	0.0048	0.057	0.21	0.16	0.13	0.11	0.13	0.063	0.073	0.058	0.093	0.0022	-0.0025	0.3
Inflight wifi service	0.009	-0.0075	0.018	-0.11	-0.023	0.0071	1	0.34	0.72	0.34	0.13	0.46	0.12	0.21	0.12	0.16	0.12	0.043	0.11	0.13	-0.017	-0.019	0.28
Departure/Arrival time convenient	0.0091	-0.21	0.038	0.26	0.09	-0.02	0.34	1	0.44	0.44	0.0049	0.07	0.011	-0.0049	0.069	0.012	0.072	0.093	0.073	0.014	0.001	-0.00092	-0.052
Ease of Online booking	0.007	-0.02	0.025	-0.13	-0.094	0.066	0.72	0.44	1	0.46	0.032	0.4	0.03	0.047	0.039	0.11	0.039	0.011	0.035	0.016	-0.0064	-0.0079	0.17
Gate location	0.00033	0.0061	-0.0013	-0.031	-0.0045	0.0048	0.34	0.44	0.46	1	-0.0012	0.0017	0.0037	0.0035	-0.028	-0.0059	0.0023	-0.035	0.0017	-0.0038	0.0055	0.0052	0.00068
Food and drink	0.0058	-0.059	0.023	-0.063	-0.077	0.057	0.13	0.0049	0.032	-0.0012	1	0.23	0.57	0.62	0.059	0.032	0.035	0.087	0.034	0.66	-0.03	-0.032	0.21
Online boarding	-0.042	-0.19	0.21	-0.22	-0.3	0.21	0.46	0.07	0.4	0.0017	0.23	1	0.42	0.29	0.16	0.12	0.083	0.2	0.075	0.33	-0.019	-0.022	0.5
Seat comfort	-0.027	-0.16	0.16	-0.12	-0.21	0.16	0.12	0.011	0.03	0.0037	0.57	0.42	1	0.61	0.13	0.11	0.075	0.19	0.069	0.68	-0.028	-0.03	0.35
Inflight entertainment	0.0061	-0.11	0.076	-0.15	-0.18	0.13	0.21	-0.0049	0.047	0.0035	0.62	0.29	0.61	1	0.42	0.3	0.38	0.12	0.4	0.69	-0.027	-0.031	0.4
On-board service	0.0082	-0.057	0.058	-0.057	-0.21	0.11	0.12	0.069	0.039	-0.028	0.059	0.16	0.13	0.42	1	0.36	0.52	0.24	0.55	0.12	-0.032	-0.035	0.32
Leg room service	0.032	-0.048	0.041	-0.14	-0.2	0.13	0.16	0.012	0.11	-0.0059	0.032	0.12	0.11	0.3	0.36	1	0.37	0.15	0.37	0.096	0.014	0.012	0.31
Baggage handling	0.037	0.025	-0.048	-0.031	-0.16	0.063	0.12	0.072	0.039	0.0023	0.035	0.083	0.075	0.38	0.52	0.37	1	0.23	0.63	0.096	-0.0056	-0.0086	0.25
Checkin service	0.011	-0.032	0.035	0.017	-0.16	0.073	0.043	0.093	0.011	-0.035	0.087	0.2	0.19	0.12	0.24	0.15	0.23	1	0.24	0.18	-0.018	-0.02	0.24
Inflight service	0.039	0.023	-0.049	-0.023	-0.16	0.058	0.11	0.073	0.035	0.0017	0.034	0.075	0.069	0.4	0.55	0.37	0.63	0.24	1	0.089	-0.055	-0.059	0.24
Cleanliness	0.0065	-0.084	0.054	-0.079	-0.13	0.093	0.13	0.014	0.016	-0.0038	0.66	0.33	0.68	0.69	0.12	0.096	0.096	0.18	0.089	1	-0.014	-0.016	0.31
Departure Delay in Minutes	0.0029	0.004	-0.01	-0.0054	0.01	0.0022	-0.017	0.001	-0.0064	0.0055	-0.03	-0.019	-0.028	-0.027	-0.032	0.014	-0.0056	-0.018	-0.055	-0.014	1	0.96	-0.05
Arrival Delay in Minutes	0.00035	0.0048	-0.012	-0.0059	0.015	-0.0025	-0.019	-0.00092	-0.0079	0.0052	-0.032	-0.022	-0.03	-0.031	-0.035	0.012	-0.0086	-0.02	-0.059	-0.016	0.96	1	-0.057
satisfaction	0.012	-0.19	0.14	-0.45	-0.45	0.3	0.28	-0.052	0.17	0.00068	0.21	0.5	0.35	0.4	0.32	0.31	0.25	0.24	0.24	0.31	-0.05	-0.057	1

- Online Boarding: Strong positive correlation (~0.73). Passengers who rate online boarding highly tend to be more satisfied.
- Inflight Entertainment: Positive correlation (~0.56). Entertainment availability improves satisfaction.
- Seat Comfort: Positive correlation (~0.52). Comfortable seating increases satisfaction.
- On-board Service: Positive correlation (~0.50). Quality service during the flight positively impacts satisfaction.
- Leg Room Service: Positive correlation (~0.48). Ample legroom contributes to higher satisfaction.

Exploratory Data Analysis 4



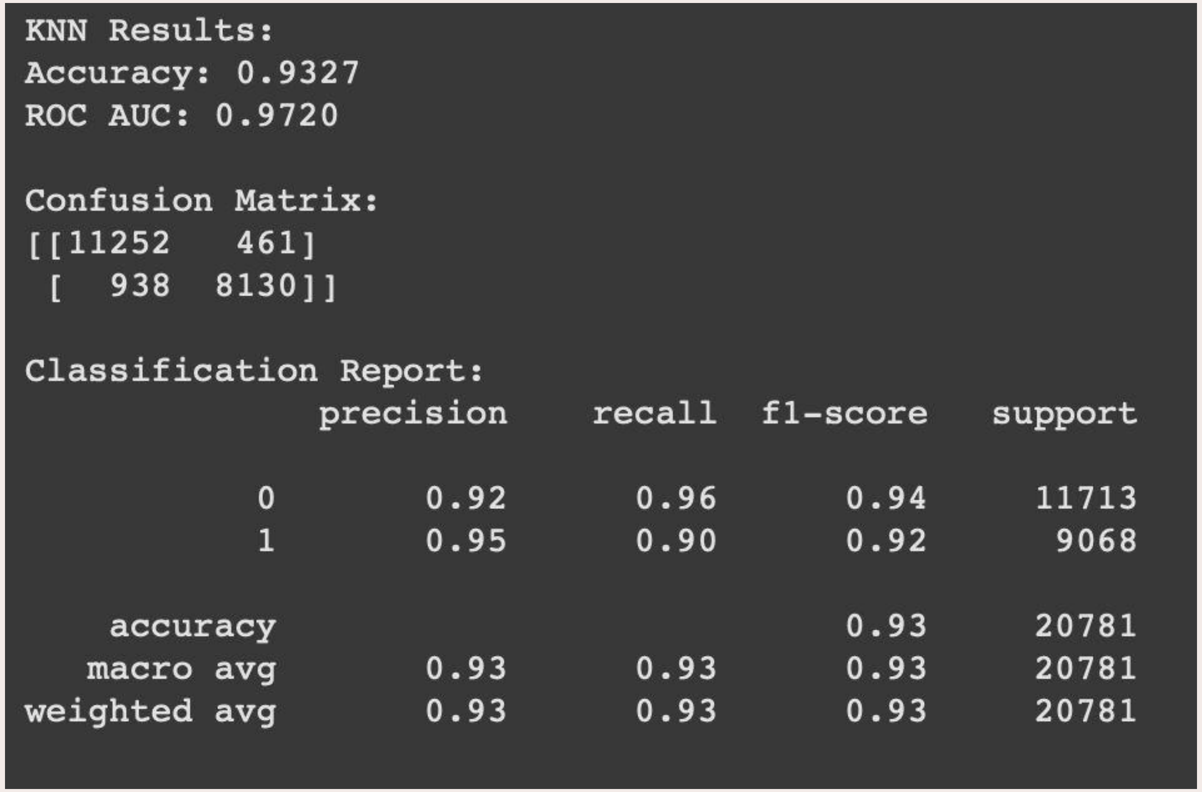
Passenger satisfaction is heavily influenced by both service-related factors (e.g., online boarding, Wi-Fi) and comfort-related features (e.g., class, seat comfort). Features like logistics and minor services also play a role, although to a lesser extent.

Model Overview

- Model Overview
- Models:
  - The project utilizes a Naïve Bayes, KNN, Gradient Boosting Machine, Decision Tree, Random Forest, Logistic Regression, SVM, ANN model for effectiveness in classification tasks.
- Model Training :
  - Training Process: The model was trained using 80% of the dataset, ensuring a substantial amount of data to learn from.
- Model Evaluation
  - Metrics Used: The model's performance was evaluated using a comprehensive set of metrics: Accuracy, Precision, Recall, and F1 Score as Classification Report and Confusion Matrix. These metrics provide a well-rounded view of the model's performance, considering both error types and the balance between precision and recall

kNN

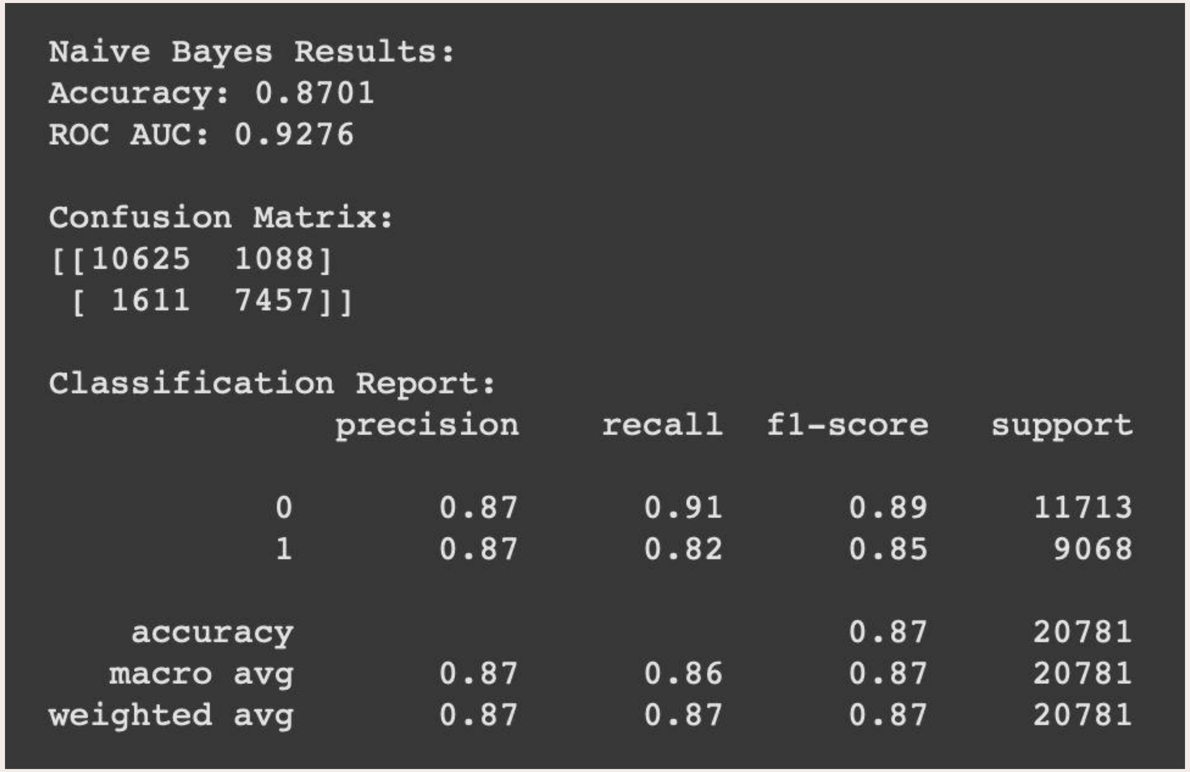
K – Nearest Neighbors



```
KNN Results:
Accuracy: 0.9327
ROC AUC: 0.9720

Confusion Matrix:
[[11252   461]
 [  938  8130]]

Classification Report:
              precision    recall  f1-score   support

           0       0.92      0.96      0.94     11713
           1       0.95      0.90      0.92      9068

    accuracy                           0.93     20781
   macro avg       0.93      0.93      0.93     20781
weighted avg       0.93      0.93      0.93     20781
```

"The kNN compares a new data point to a set of data it was trained on to make predictions. It does this by finding the k-nearest neighbours to the new data point and predicting its class or value based on the classes or values of its neighbours."

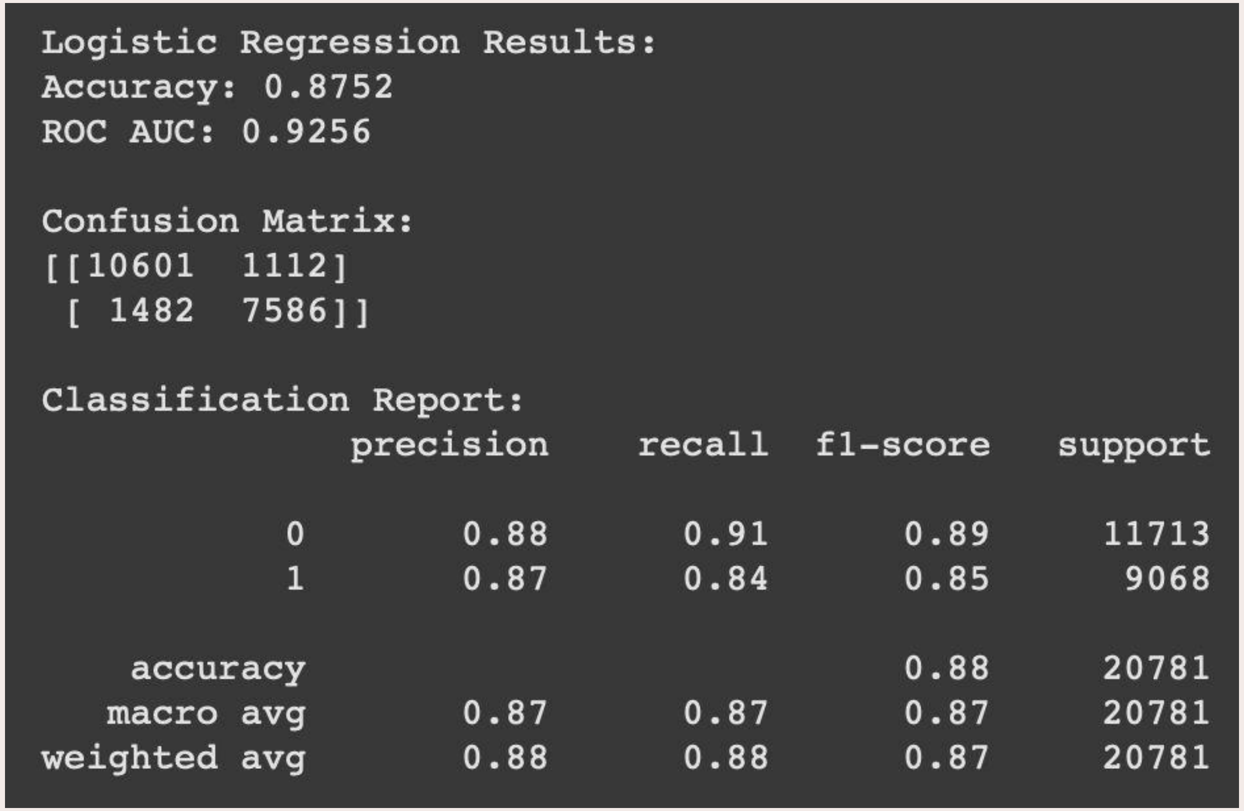
Naïve Bayes



```
Naive Bayes Results:
Accuracy: 0.8701
ROC AUC: 0.9276

Confusion Matrix:
[[10625  1088]
 [ 1611  7457]]

Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.91      0.89     11713
           1       0.87      0.82      0.85      9068

    accuracy                           0.87     20781
   macro avg       0.87      0.86      0.87     20781
weighted avg       0.87      0.87      0.87     20781
```

“Naive Bayes assumes that the effect of each feature in a class is independent of other features. This assumption is called class conditional independence. The algorithm then calculates posterior probabilities for each class and predicts the class with the highest probability.”

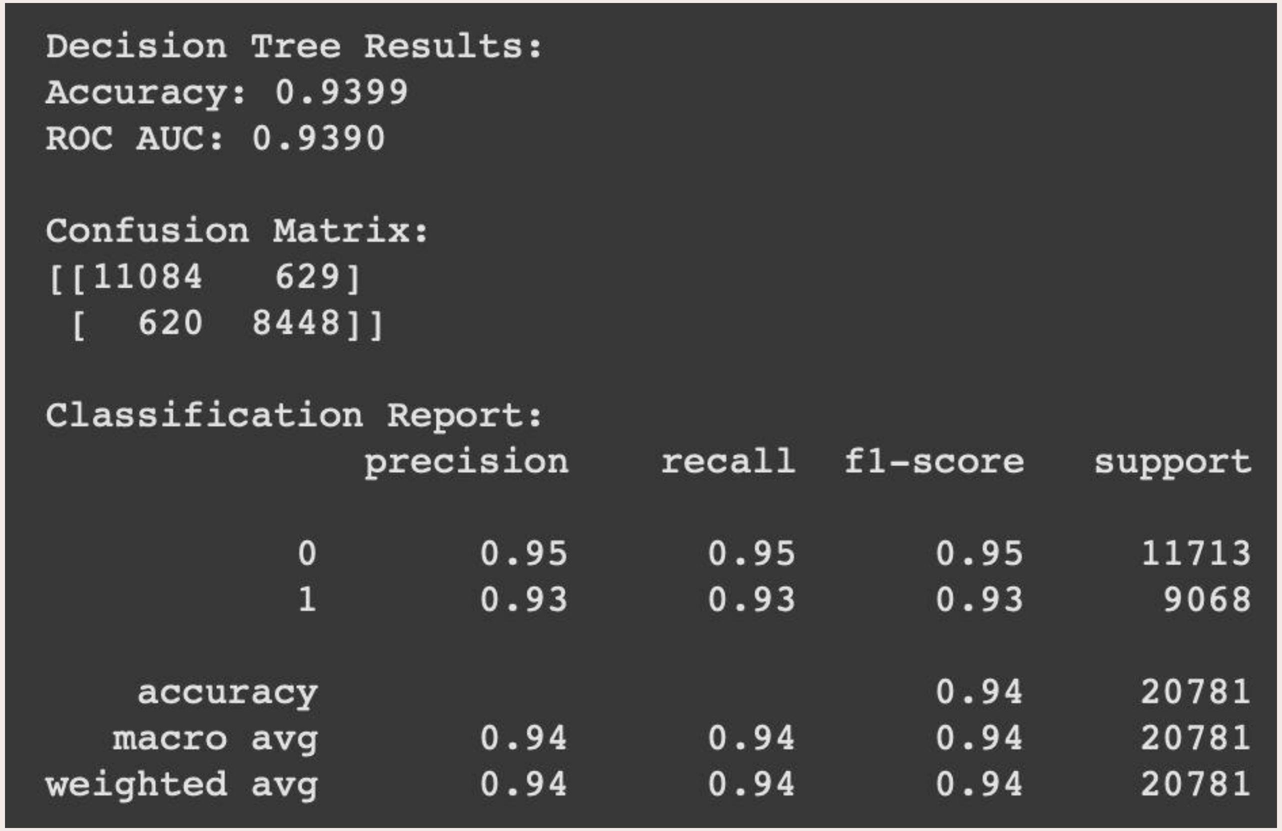
Logistic Regression



```
Logistic Regression Results:
Accuracy: 0.8752
ROC AUC: 0.9256

Confusion Matrix:
[[10601  1112]
 [ 1482  7586]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.91      0.89     11713
           1       0.87      0.84      0.85      9068

    accuracy                           0.88     20781
   macro avg       0.87      0.87      0.87     20781
weighted avg       0.88      0.88      0.87     20781
```

“Logistic regression is a data analysis technique that uses mathematics to find the relationships between two data factors. It then uses this relationship to predict the value of one of those factors based on the other. The prediction usually has a finite number of outcomes, like yes or no.”

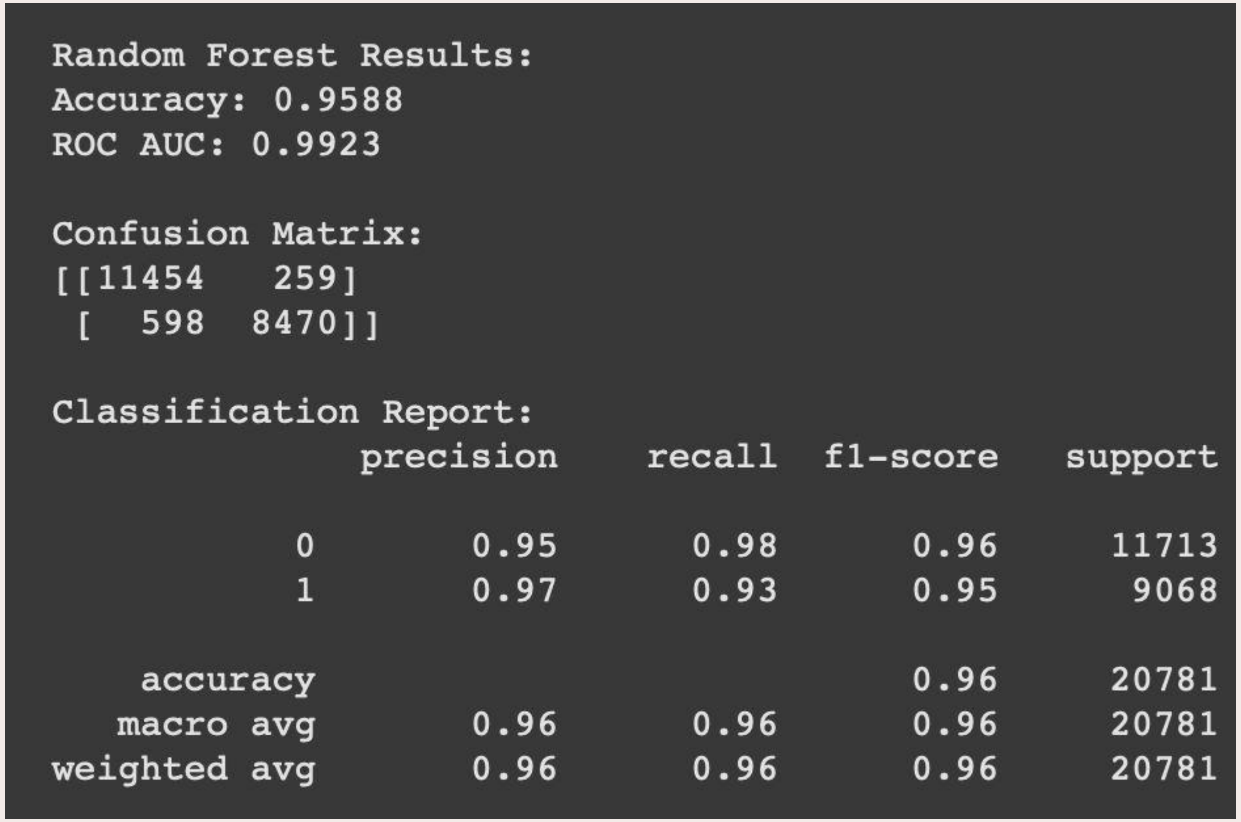
Decision Tree



```
Decision Tree Results:
Accuracy: 0.9399
ROC AUC: 0.9390

Confusion Matrix:
[[11084   629]
 [  620  8448]]

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.95      0.95     11713
           1       0.93      0.93      0.93      9068

    accuracy                           0.94     20781
   macro avg       0.94      0.94      0.94     20781
weighted avg       0.94      0.94      0.94     20781
```

“A decision tree works by dividing data into smaller subsets based on specific attributes at each node, progressively refining the classification until it reaches a final decision at the leaf node, essentially making predictions by following a series of "yes/no" questions based on the data's features, with the goal of identifying the best attribute to split on at each stage to maximize the accuracy of the prediction.”

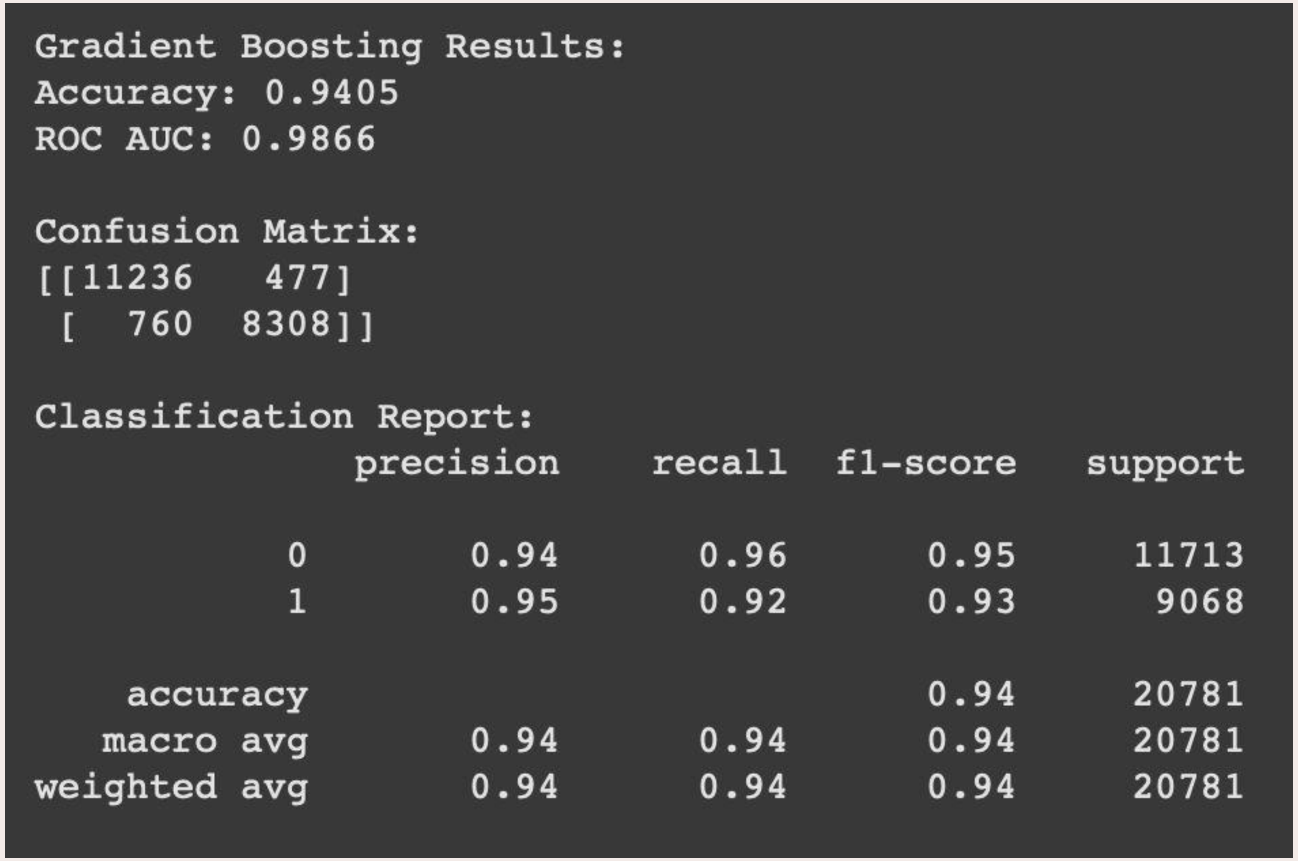
Random Forest



```
Random Forest Results:
Accuracy: 0.9588
ROC AUC: 0.9923

Confusion Matrix:
[[11454   259]
 [  598  8470]]

Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.98      0.96     11713
           1       0.97      0.93      0.95      9068

    accuracy                           0.96     20781
   macro avg       0.96      0.96      0.96     20781
weighted avg       0.96      0.96      0.96     20781
```

“A random forest is an ensemble learning method that combines the output of multiple decision trees to reach a single result. Each decision tree in the forest is trained using a random subset of the training data.”

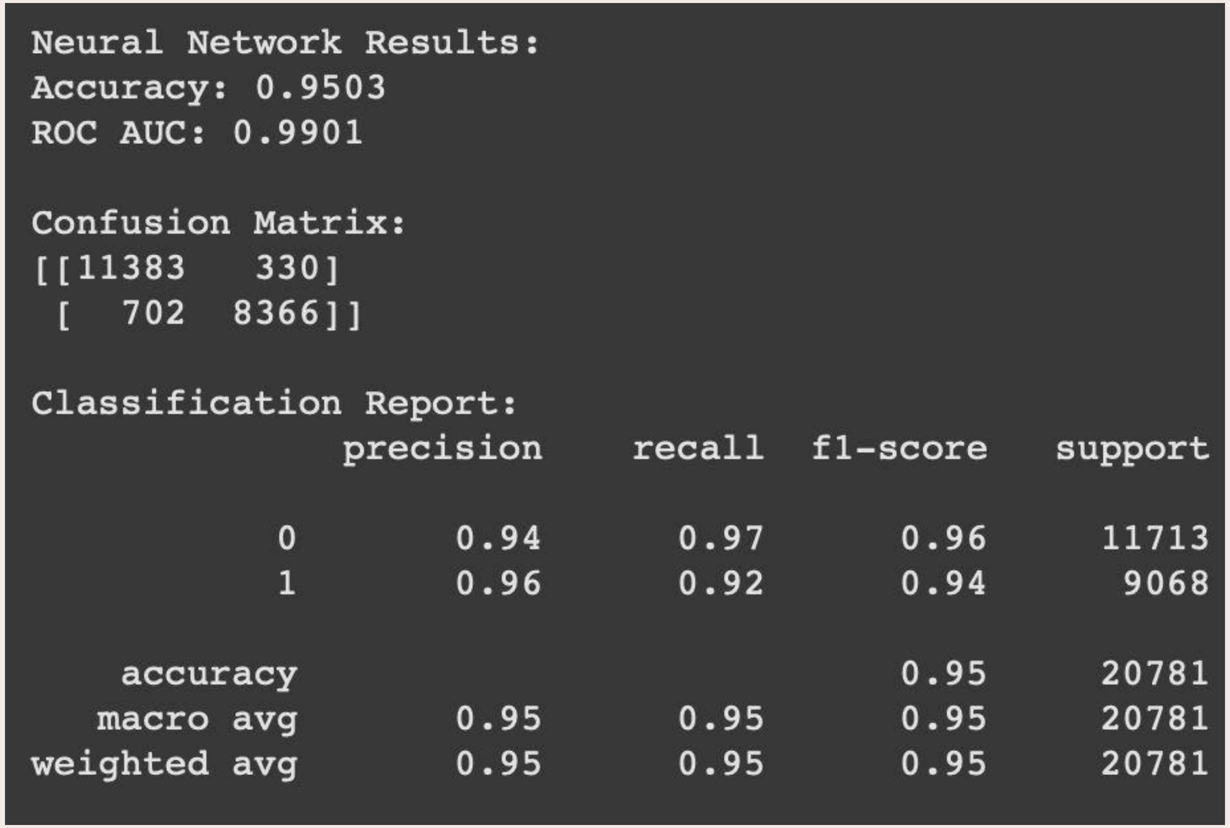
Gradient Boosting



```
Gradient Boosting Results:
Accuracy: 0.9405
ROC AUC: 0.9866

Confusion Matrix:
[[11236   477]
 [  760  8308]]

Classification Report:
             precision    recall  f1-score   support

           0       0.94      0.96      0.95     11713
           1       0.95      0.92      0.93      9068

    accuracy                           0.94     20781
   macro avg       0.94      0.94      0.94     20781
weighted avg       0.94      0.94      0.94     20781
```

“Gradient boosting is an iterative process that sequentially adds new models to improve the accuracy of the previous ones. The final prediction is the sum of all the individual predictions.”

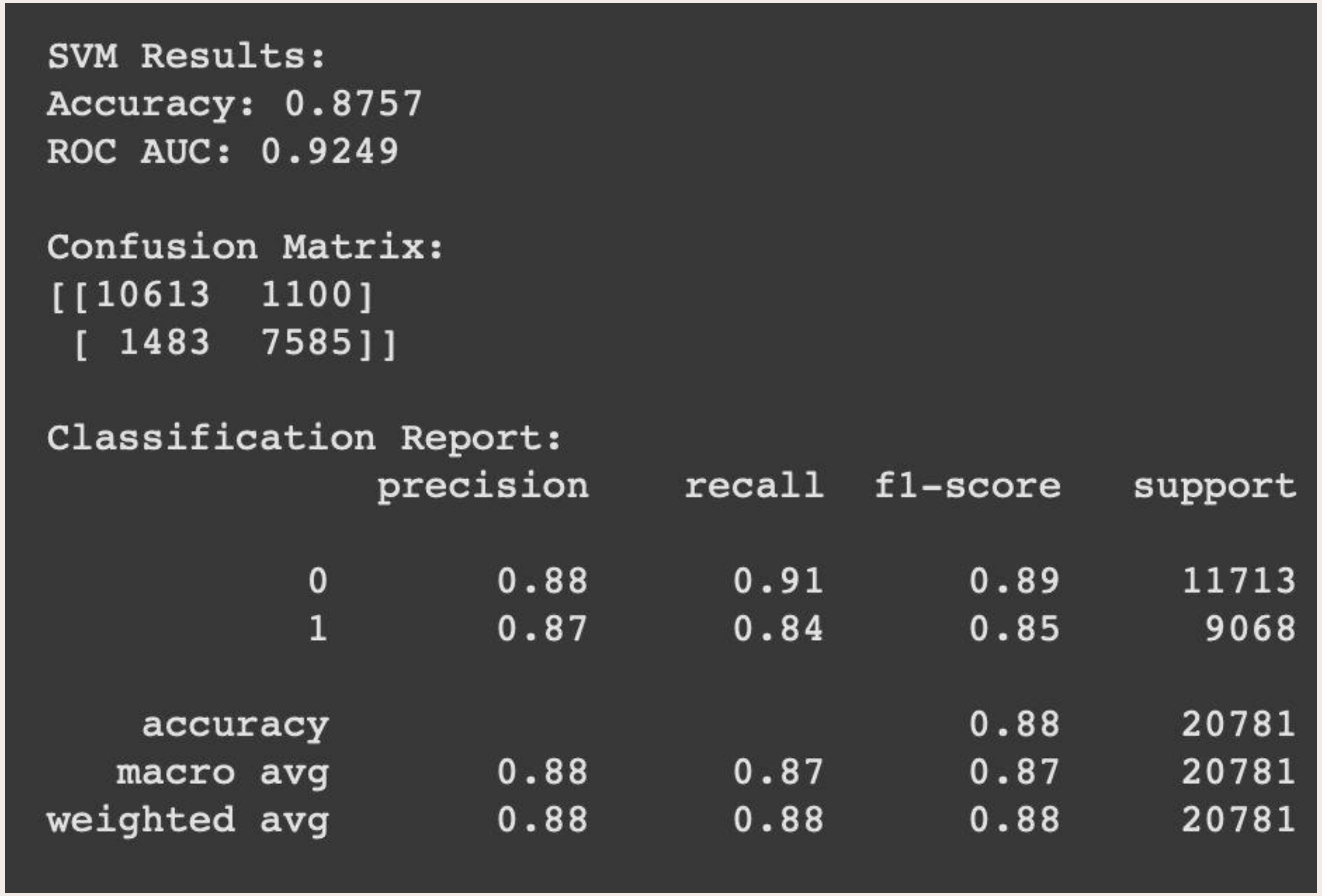
Artificial Neural Networks



```
Neural Network Results:
Accuracy: 0.9503
ROC AUC: 0.9901

Confusion Matrix:
[[11383   330]
 [  702  8366]]

Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.97      0.96     11713
           1       0.96      0.92      0.94      9068

    accuracy                           0.95     20781
   macro avg       0.95      0.95      0.95     20781
weighted avg       0.95      0.95      0.95     20781
```

An Artificial Neural Network (ANN) processes data through layers of interconnected neurons. Each neuron applies a weighted sum, activation function, and passes outputs forward. Through backpropagation, the network adjusts weights to learn patterns and optimize predictions, mimicking the human brain's signal processing in interconnected pathways to handle complex relationships.

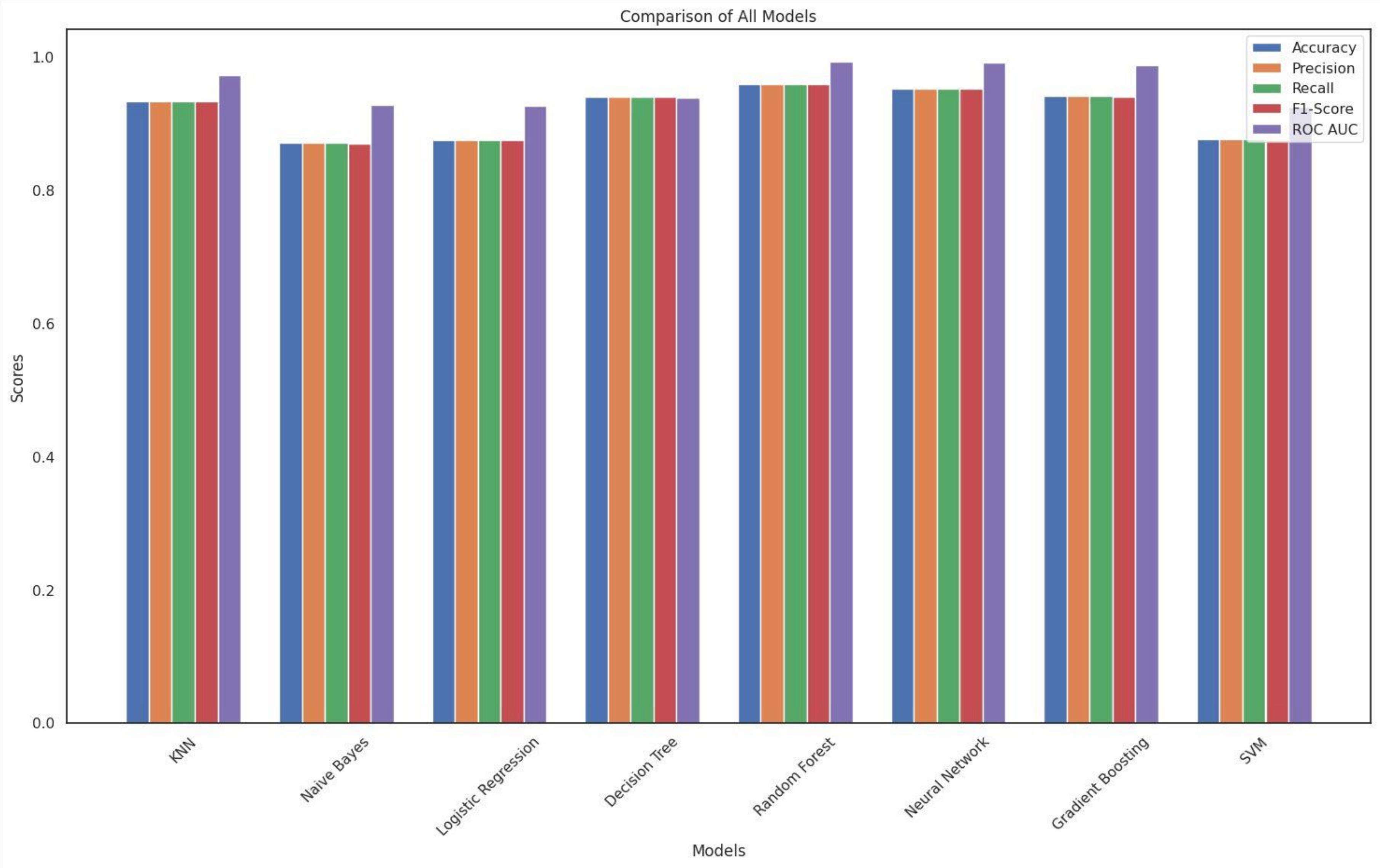
Support Vector Machine

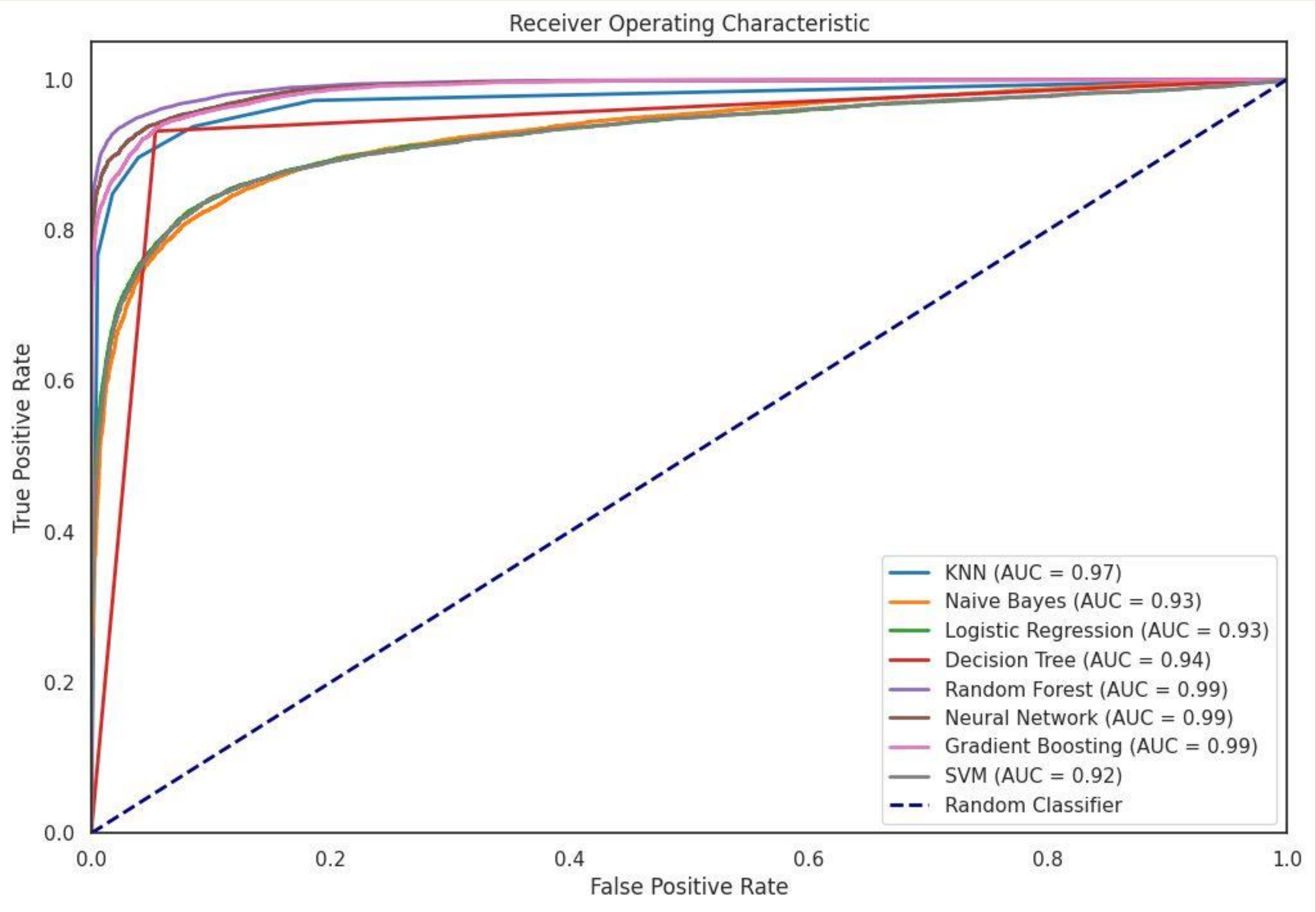


```
SVM Results:
Accuracy: 0.8757
ROC AUC: 0.9249

Confusion Matrix:
[[10613  1100]
 [ 1483  7585]]

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.91      0.89     11713
           1       0.87      0.84      0.85      9068

    accuracy                           0.88     20781
   macro avg       0.88      0.87      0.87     20781
weighted avg       0.88      0.88      0.88     20781
```

“SVMs find a hyperplane that separates data points of one class from those of another class. The hyperplane is chosen to maximize the distance between the two classes, or the margin. The closest points to the hyperplane are called support vectors, and they help identify the boundary line.”

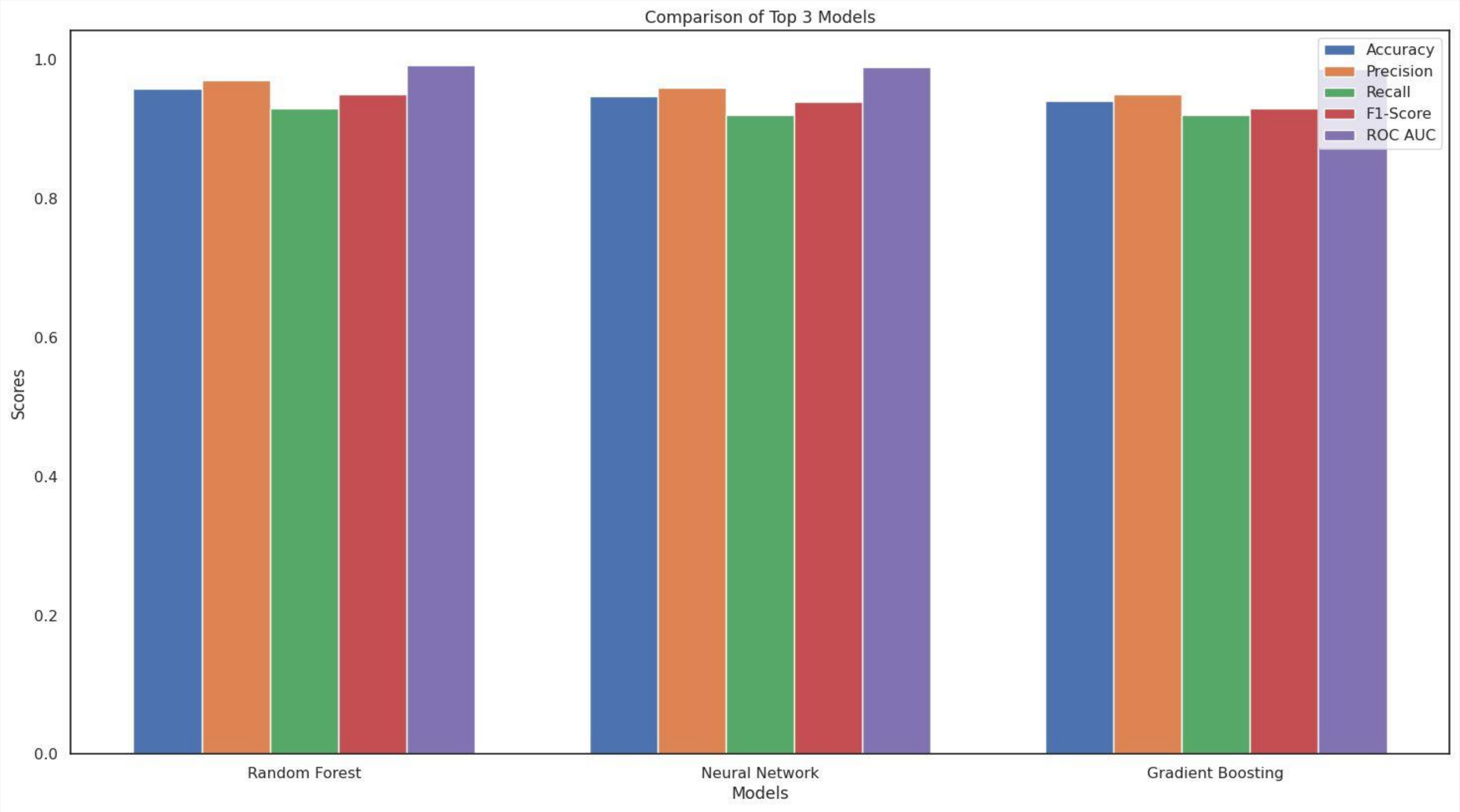
Comparison Of Model




Receiver Operating Characteristic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
KNN (AUC = 0.97)
Naive Bayes (AUC = 0.93)
Logistic Regression (AUC = 0.93)
Decision Tree (AUC = 0.94)
Random Forest (AUC = 0.99)
Neural Network (AUC = 0.99)
Gradient Boosting (AUC = 0.99)
SVM (AUC = 0.92)
Random Classifier

ROC - AUC

Top 3 Models


Comparison of Top 3 Models
Accuracy
Precision
Recall
F1-Score
ROC AUC
Scores
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest
Neural Network
Gradient Boosting
Models

Summary

The study utilizes the "Airline Passenger Satisfaction" dataset, which includes variables such as demographics, flight details, and service quality ratings. Key steps include data preprocessing (handling missing data, encoding categorical variables), feature selection, and applying predictive models like logistic regression and random forests. The analysis highlights service quality, seat comfort, and customer service as major contributors to satisfaction, with model performance evaluated using metrics like accuracy and F1-score.

Thank you