Final Project: Airline Passenger Satisfaction Dataset

- Daqian Dang
- Introduction to Data Mining
- DATS 6103
- Dr. Amir Jafari
- June 21, 2022

Five Main Phases

- Basic understanding about the dataset
- Exploratory Data Analysis and data visualization
- Data preprocessing
- Data modeling
- Modeling evaluation

Basic Understanding about the Dataset

- Reason for choosing the dataset
- Kaggle website
- Train and test data
- Classification dataset
- Target variable - Passenger ’s satisfaction level
  - Neutral or dissatisfied
  - Satisfied

SMART Question

What factors could influence an airline passenger satisfaction ?

Data Variable Analysis

- Merge two datasets (train and test)

- Numeric variables

  '*id', 'Age', 'Flight Distance', 'Departure Delay in Minutes', 'Arrival Delay in Minutes'*

- Categorical variables

  9 features with five-point rating scales:

  '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'

- Categorical variable
  - Object
    - *‘Gender’ – Male and Female*
    - *‘Customer Type’ – Loyal Customer and Disloyal Customer*
    - *'Type of Travel’ – Personal Travel and Business Travel*
    - *'Class’ – Business, Eco, Eco Plus*
    - *‘satisfaction’* (target variable) *Neutral or dissatisfied and Satisfied*

Exploratory Data Analysis

- Shape
  - Dataset number of rows: 129,880
  - Dataset number of columns: 25
- Info
  - float64(1), int64(19), object(5)
- Columns

```
Index(['Unnamed: 0', 'id', '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'],
      dtype='object')
```

```
 #   Column                            Non-Null Count   Dtype
---  ------                            --------------   -----
 0   Unnamed: 0                        129880 non-null  int64
 1   id                                129880 non-null  int64
 2   Gender                            129880 non-null  object
 3   Customer Type                     129880 non-null  object
 4   Age                               129880 non-null  int64
 5   Type of Travel                    129880 non-null  object
 6   Class                             129880 non-null  object
 7   Flight Distance                   129880 non-null  int64
 8   Inflight wifi service             129880 non-null  int64
 9   Departure/Arrival time convenient  129880 non-null  int64
 10  Ease of Online booking            129880 non-null  int64
 11  Gate location                     129880 non-null  int64
 12  Food and drink                    129880 non-null  int64
 13  Online boarding                   129880 non-null  int64
 14  Seat comfort                      129880 non-null  int64
 15  Inflight entertainment            129880 non-null  int64
 16  On-board service                  129880 non-null  int64
 17  Leg room service                  129880 non-null  int64
 18  Baggage handling                  129880 non-null  int64
 19  Checkin service                   129880 non-null  int64
 20  Inflight service                  129880 non-null  int64
 21  Cleanliness                       129880 non-null  int64
 22  Departure Delay in Minutes        129880 non-null  int64
 23  Arrival Delay in Minutes          129487 non-null  float64
 24  satisfaction                      129880 non-null  object
dtypes: float64(1), int64(19), object(5)
memory usage: 25.8+ MB
```

Exploratory Data Analysis

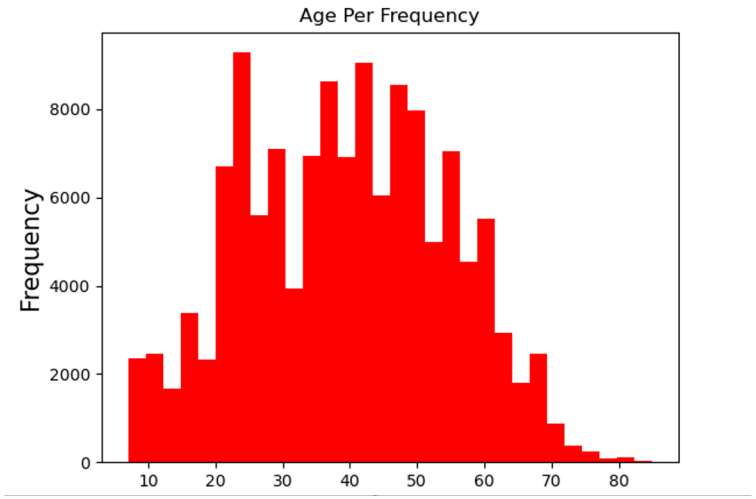
```
Gender
Female    65899
Male      63981
Name: Gender, dtype: int64
```

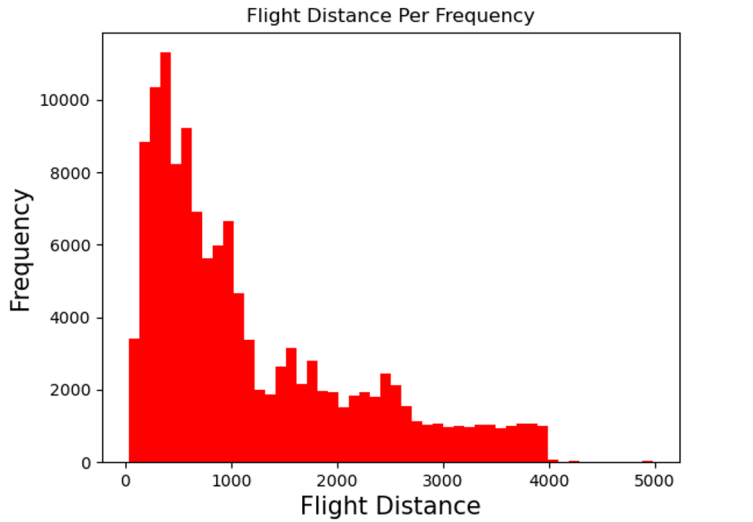
```
Gender  Customer Type
Female  Loyal Customer       53056
        disloyal Customer    12843
Male    Loyal Customer       53044
        disloyal Customer    10937
Name: Gender, dtype: int64
```

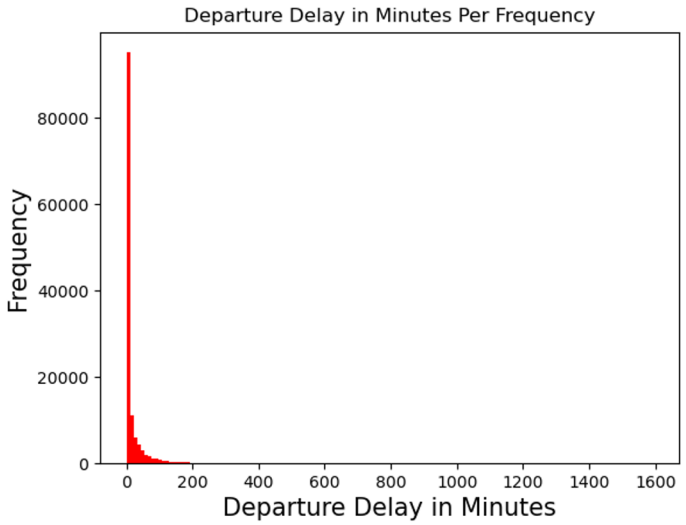
```
Gender  Class
Female  Business    31263
        Eco         29670
        Eco Plus     4966
Male    Business    30897
        Eco         28639
        Eco Plus     4445
Name: Gender, dtype: int64
```

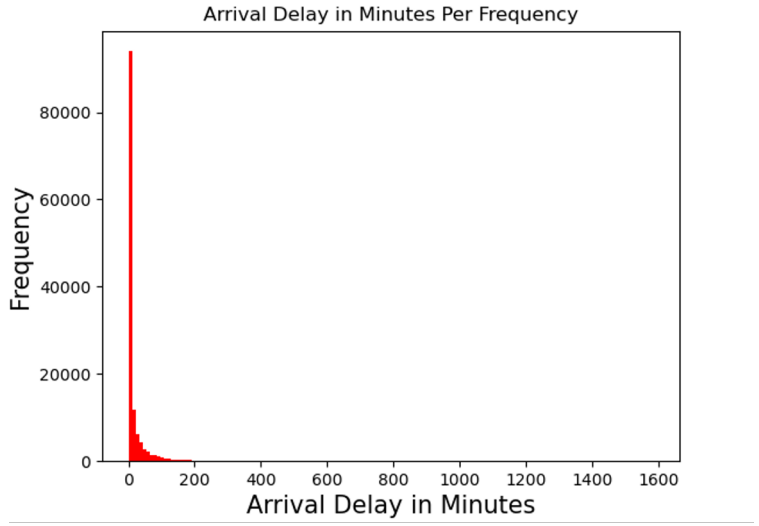
```
Gender  satisfaction
Female  neutral or dissatisfied    37630
        satisfied                  28269
Male    neutral or dissatisfied    35822
        satisfied                  28159
Name: Gender, dtype: int64
```

Data Visualization - Numeric Variables





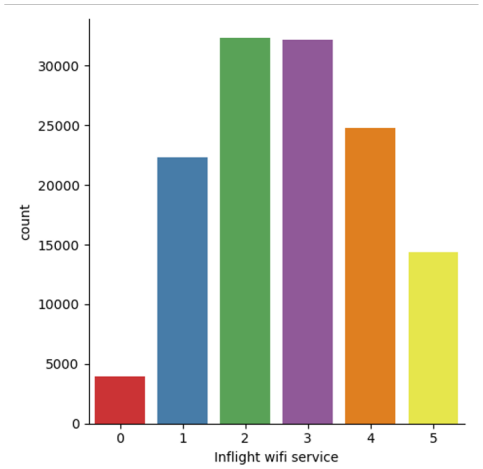


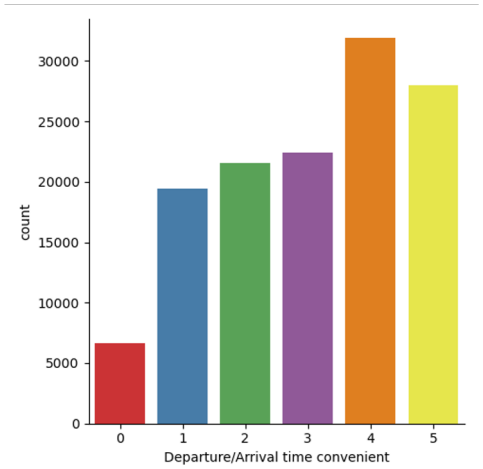


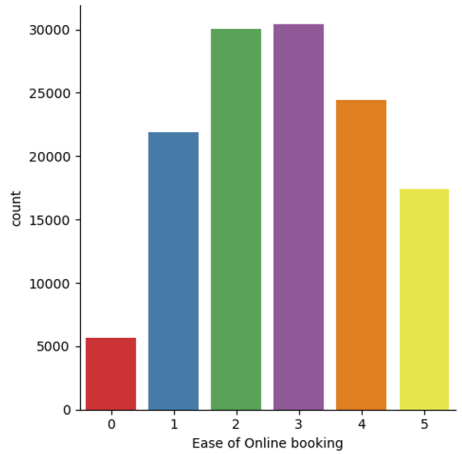
- Age
- Flight Distance
- Departure Delay in Minutes
- Arrival Delay in Minutes

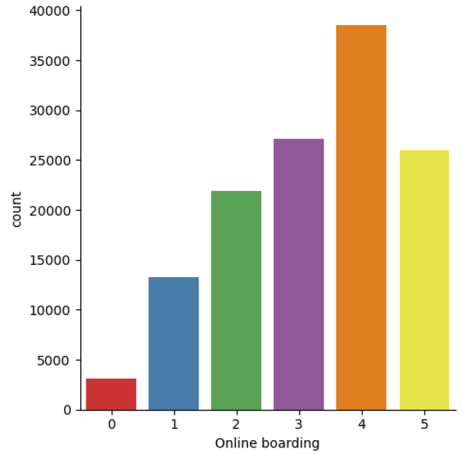
Data Visualization – Categorical Variables

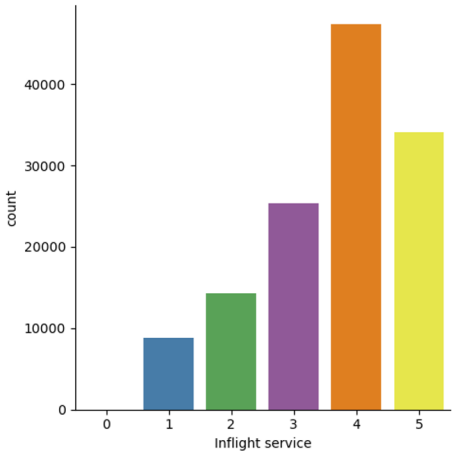
9 features with five-point rating scales

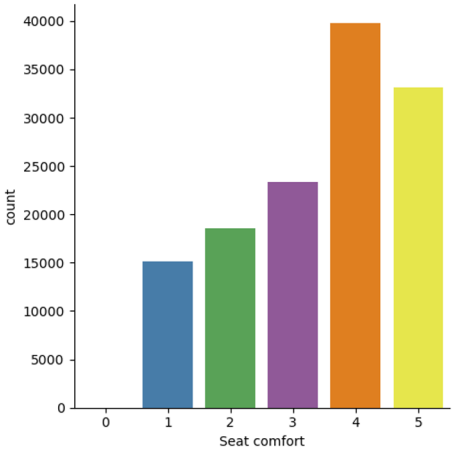


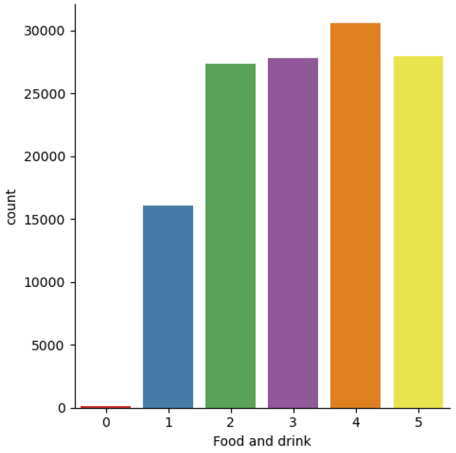


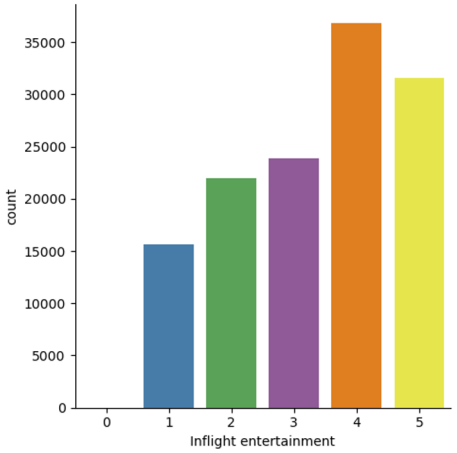


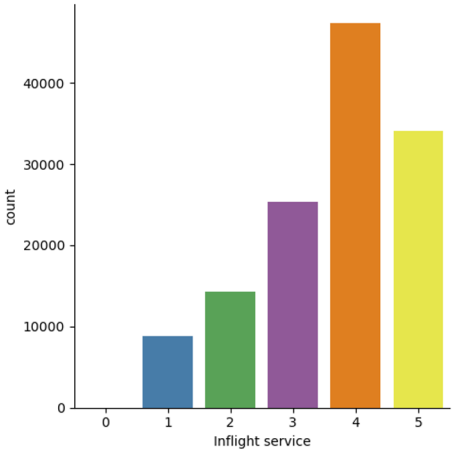




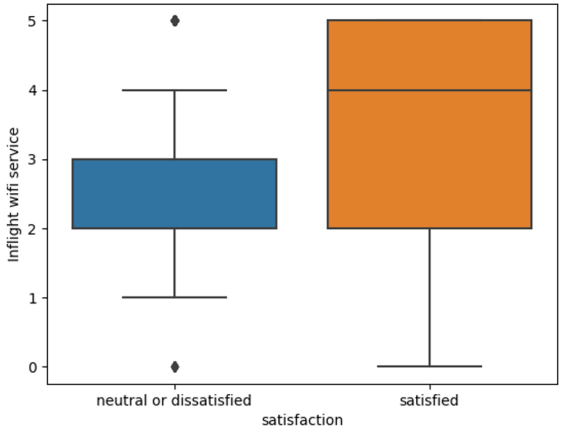


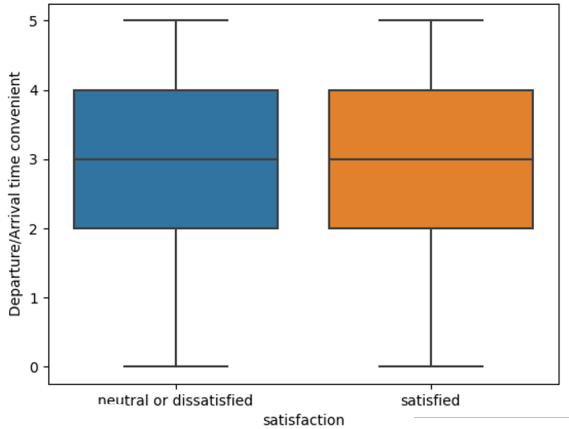


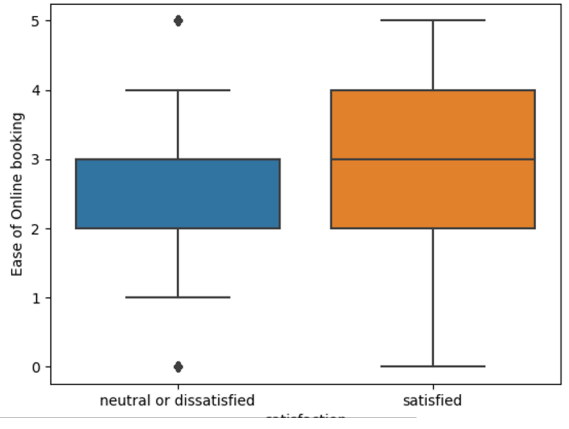


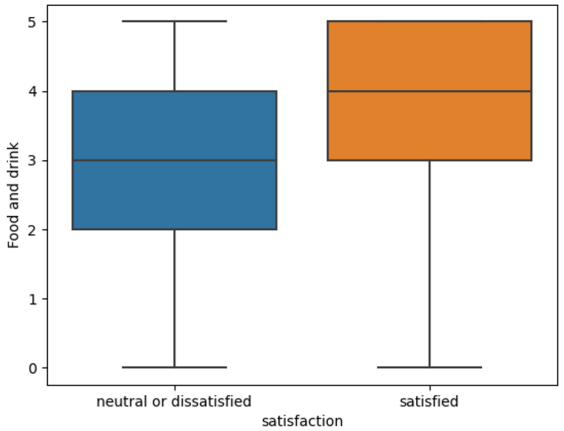


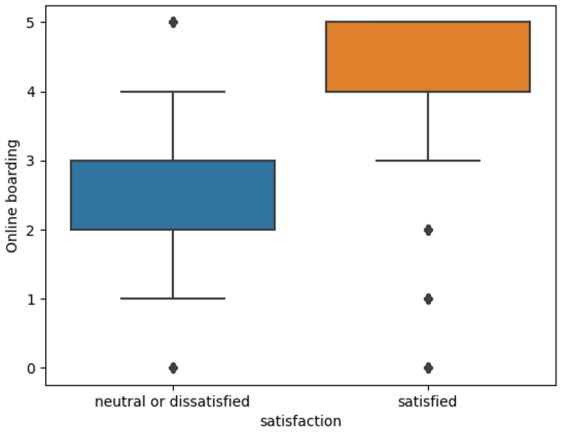
Data Visualization – Significant differences

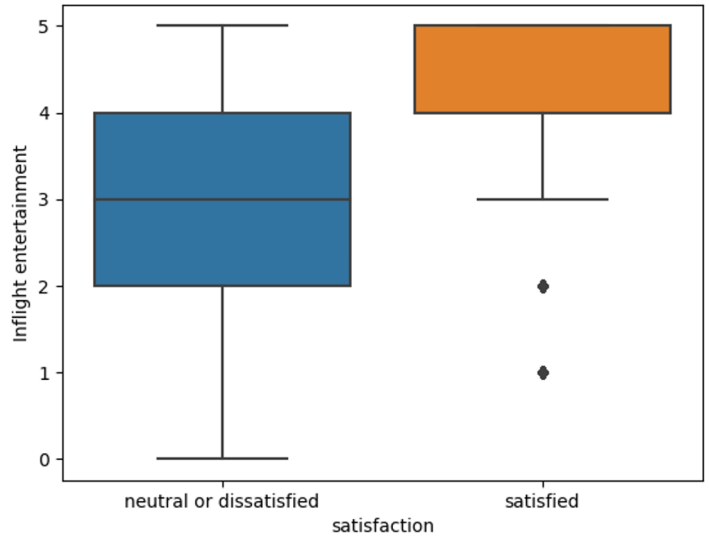


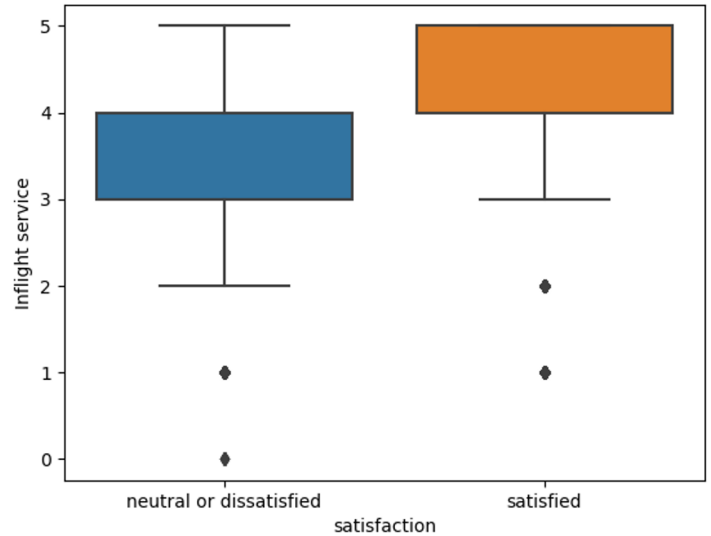


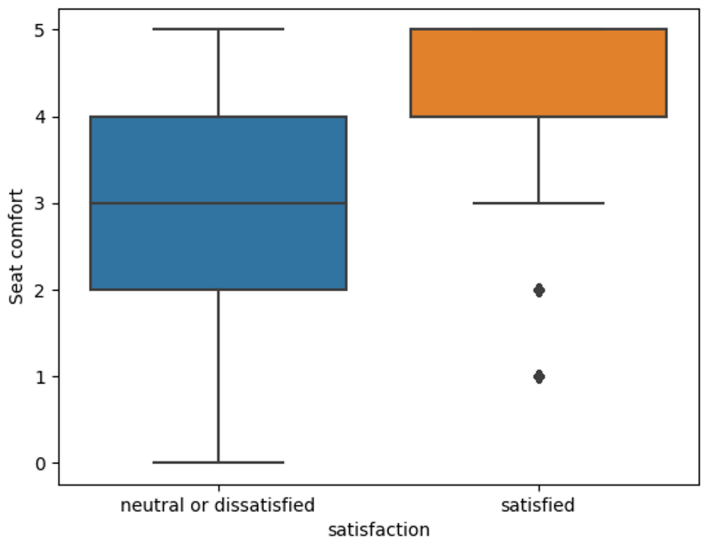


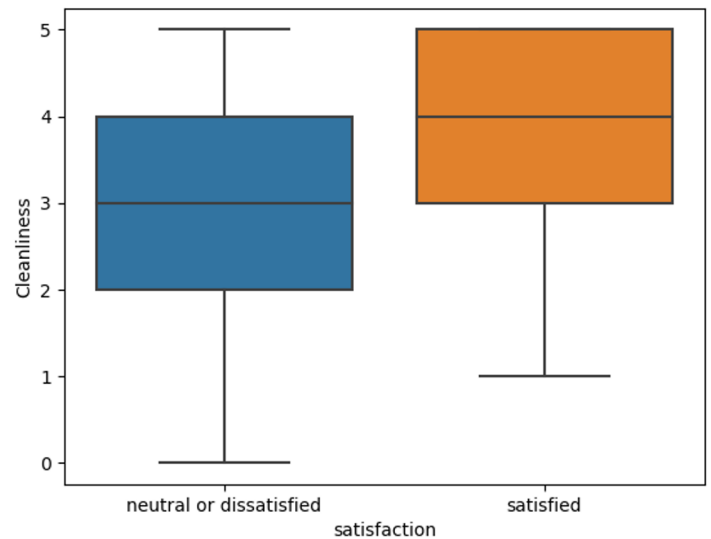





Inflight entertainment
0
1
2
3
4
5
neutral or dissatisfied
satisfied
satisfaction


Inflight service
0
1
2
3
4
5
neutral or dissatisfied
satisfied
satisfaction


Seat comfort
0
1
2
3
4
5
neutral or dissatisfied
satisfied
satisfaction


Cleanliness
0
1
2
3
4
5
neutral or dissatisfied
satisfied
satisfaction

Data Preprocessing

- Missing values
  - Median Imputation for missing values
- Outlier detection
  - Use interquartile range IQR
  - IQR = upper range Q3+1.5*IQR – lower range Q1-1.5*IQR
  - Find and remove outliers
- Data balance
- Drop unnecessary features
  - id and unnames

Missing value handling

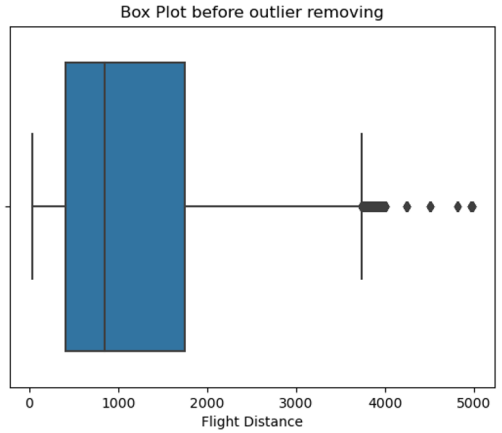
- Missing data - 393 missing data in the feature ‘Arrival Delay in Minutes’

- Median imputation
  - The feature ‘Arrival Delay in Minutes’ is extremely skewed
  - It is a good way to replace the missing values

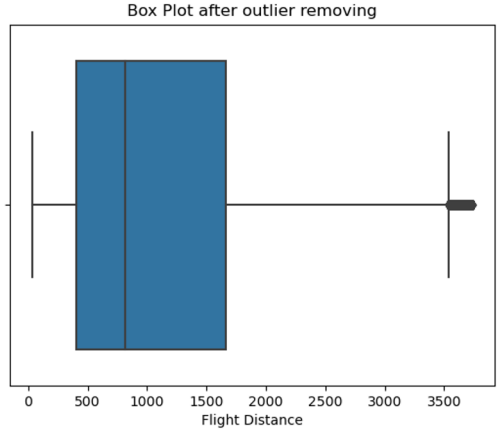
Outlier Detection and Removal

Flight Distance

Before outlier removing

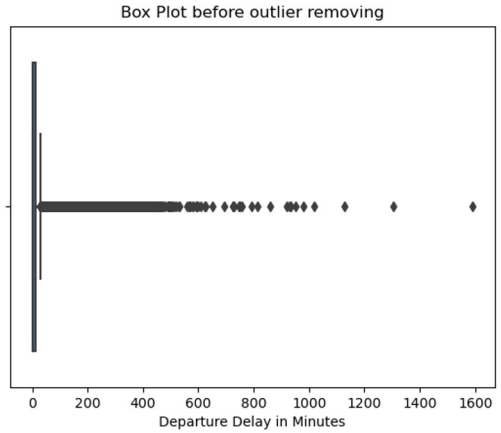


After outlier removing

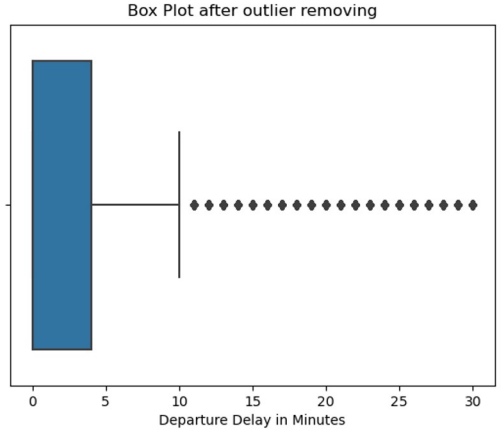


Departure Delay in Minutes

Before outlier removing

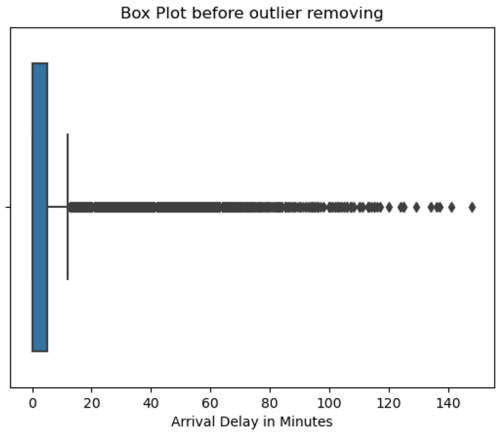


After outlier removing

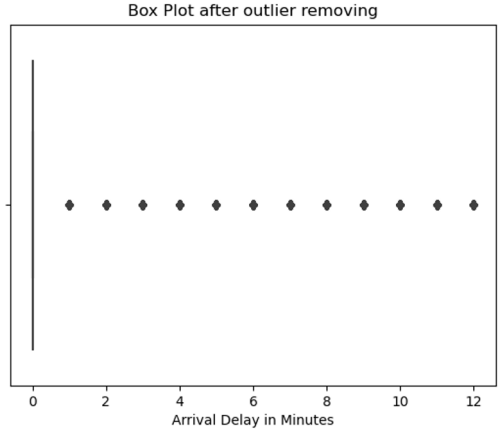


Arrival Delay in Minutes

Before outlier removing



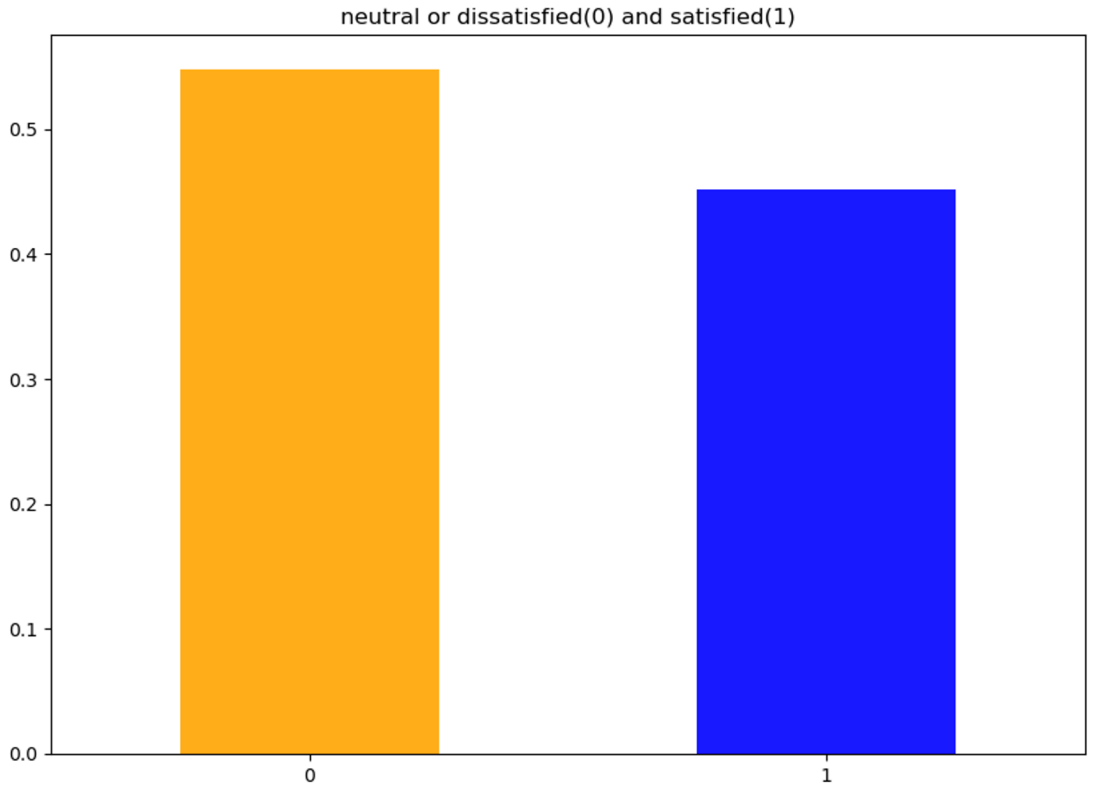
After outlier removing



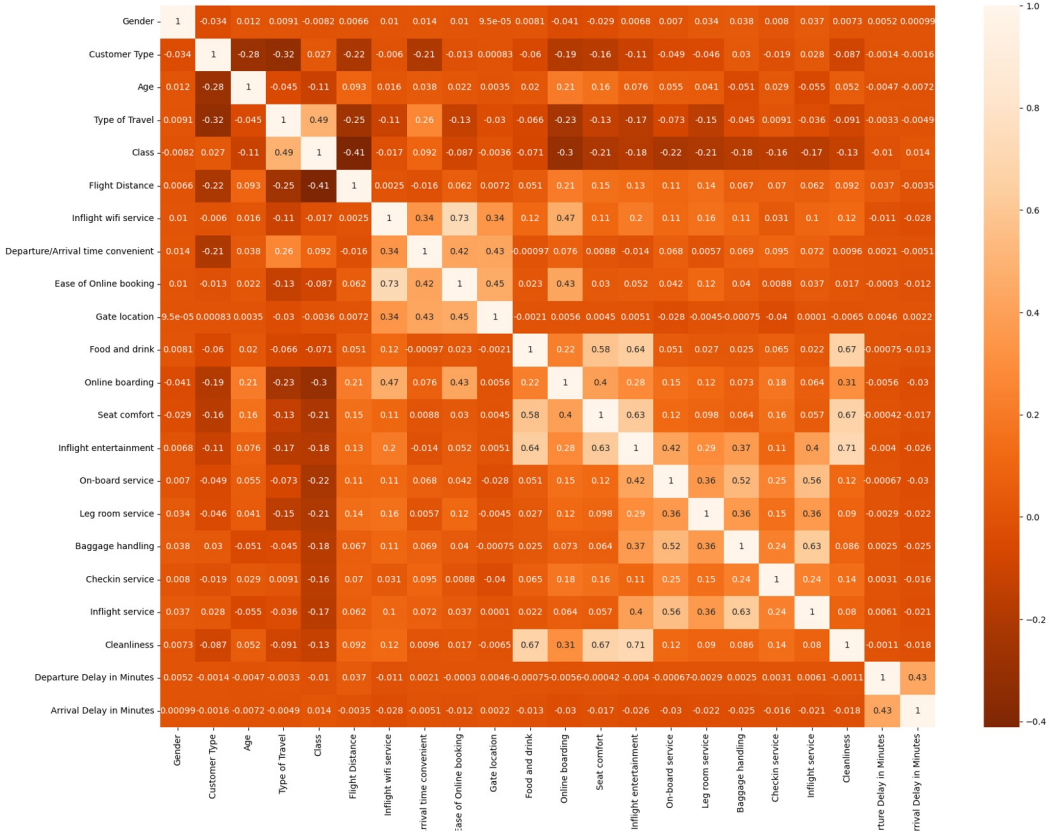
Data Balance

Target variable = satisfaction

Neutral or dissatisfied (0): More than 50%
Satisfied (1): Between 40% and 50%



Correlation Heatmap



‘*inflight Wi-Fi service’* has positively correlated with ‘*Ease of Online booking’*

‘Seat comfort’ has positively correlated with ‘*Cleanliness*’, ‘*Inflight entertainment*’

‘*Food and drink’* has positively correlated with ‘*Cleanliness*’, ‘*Inflight entertainment*’, and ‘*Seat comfort’*

Preprocessing before building models

- LabelEncoder
  - Categorical variables
  - Target variable - ‘satisfaction’

- Split the dataset into train and test
  - Train – 80%
  - Test – 20%

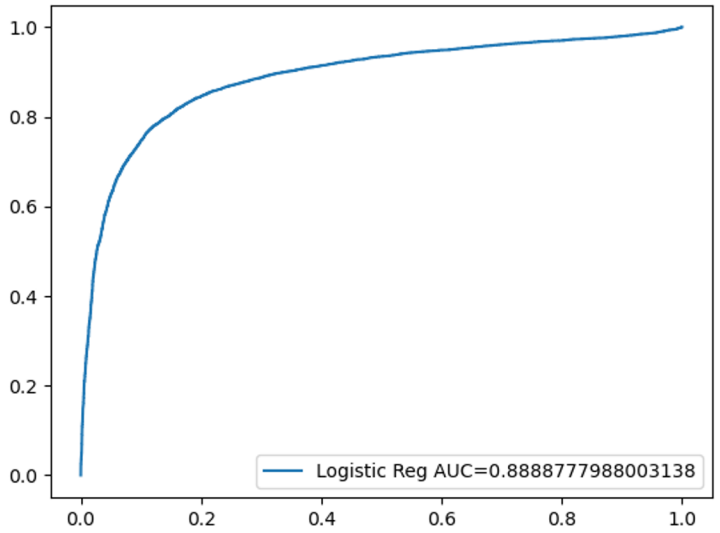
Data Modeling

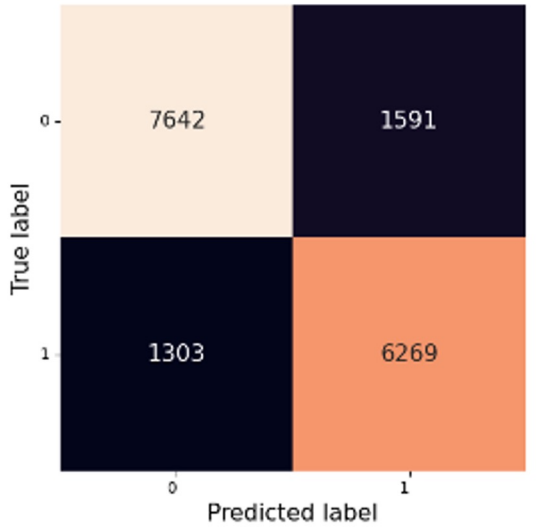
- Logistic Regression
- Naïve Bayes
- Decision Tree (Entropy)
- Random Forest (All features and K features)
- K-Nearest Neighbors

Model Metric Results

Logistic Regression

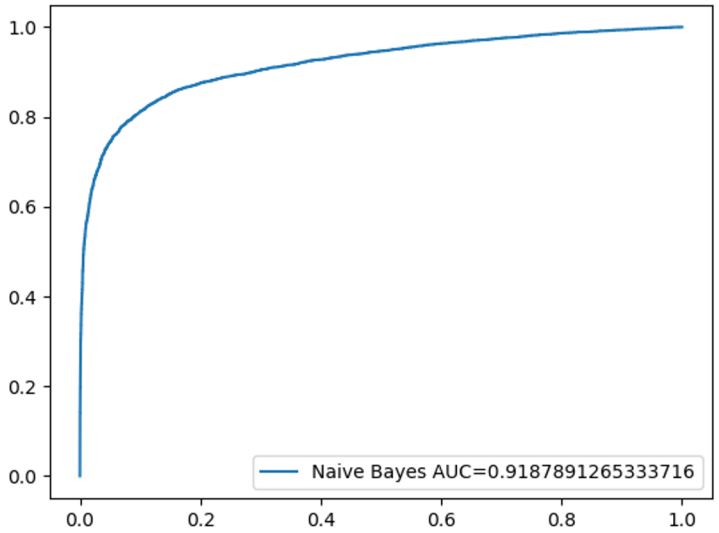
- Accuracy: 82.84
- F1-score: 81.17
- ROC_AUC : 88.89

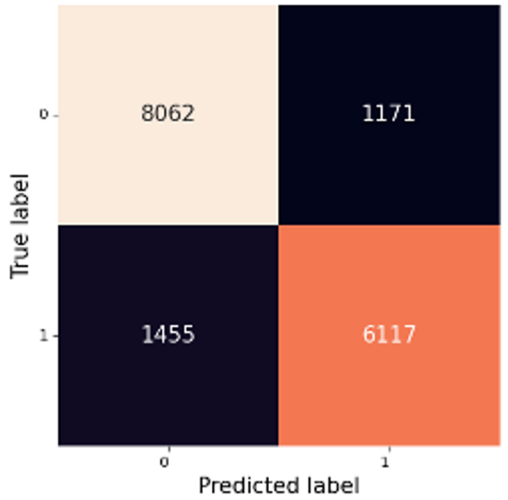




Naïve Bayes

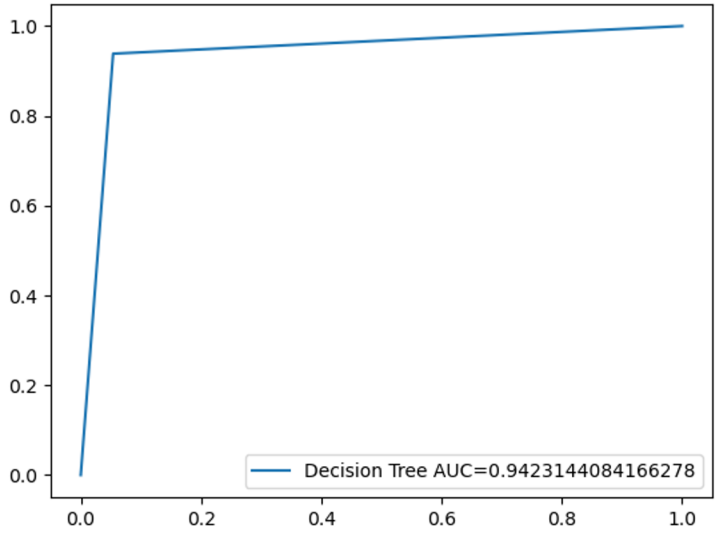
- Accuracy : 85.93
- F1-score : 83.93
- ROC_AUC : 91.88

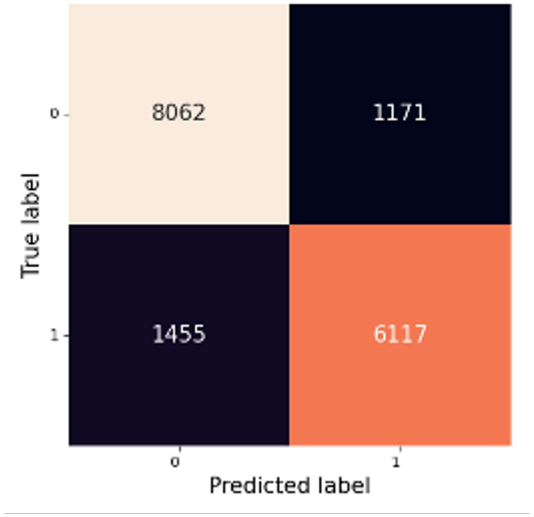




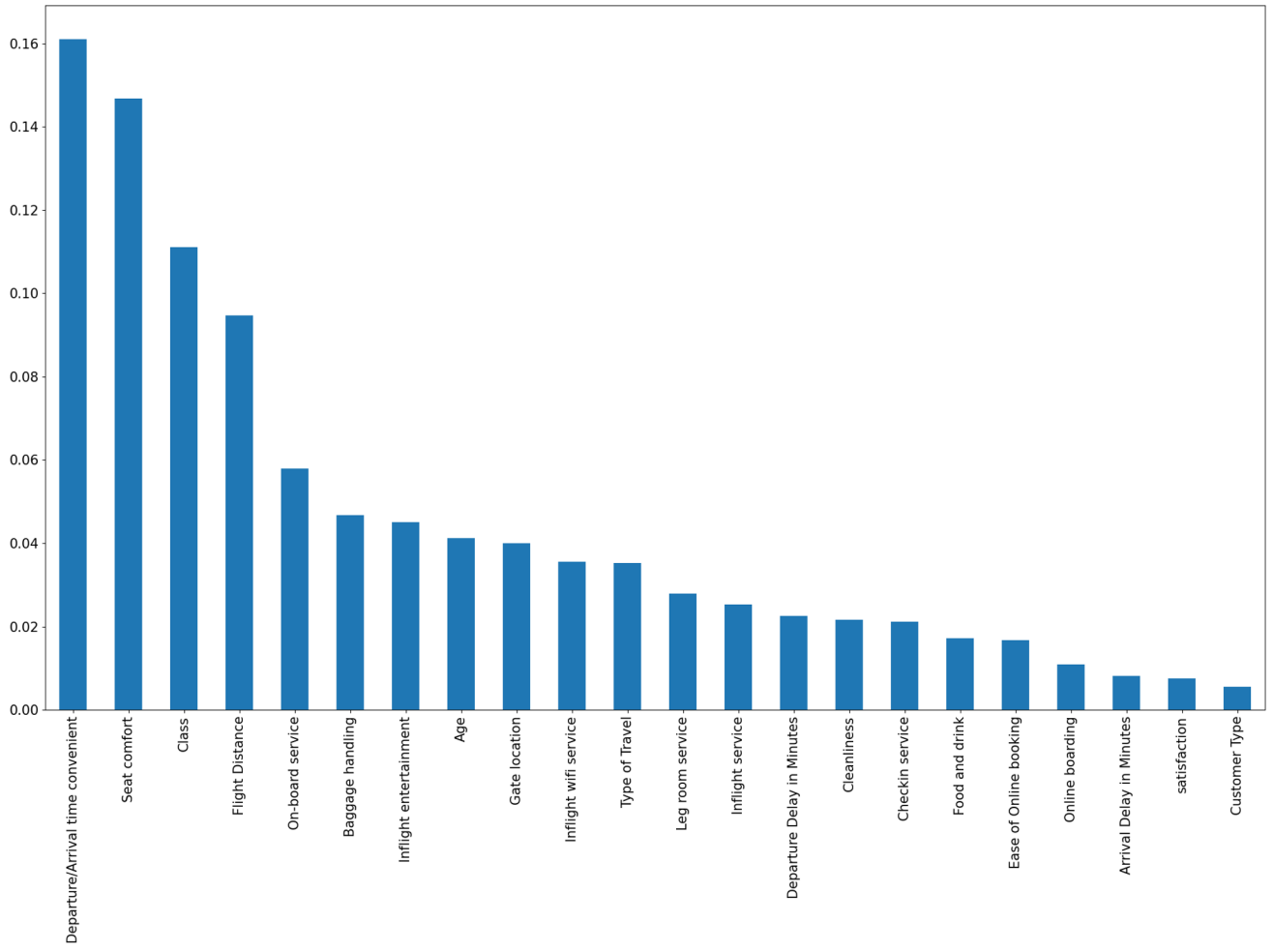
Decision Tree with Entropy

- Accuracy : 94.27
- F1-score : 93.65
- ROC_AUC : 94.23





Random Forest – Feature Importance

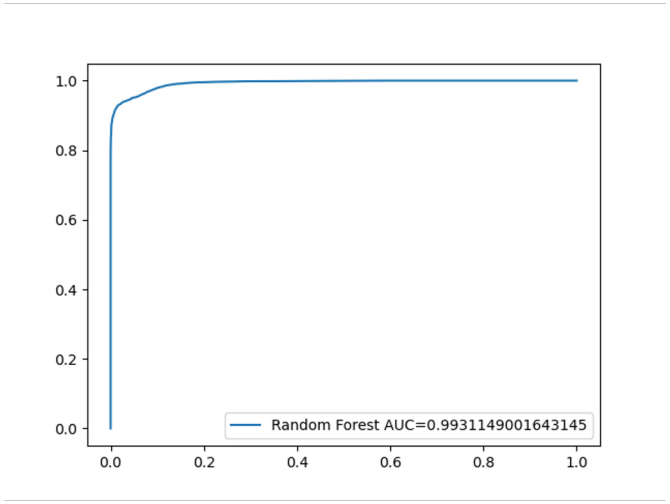


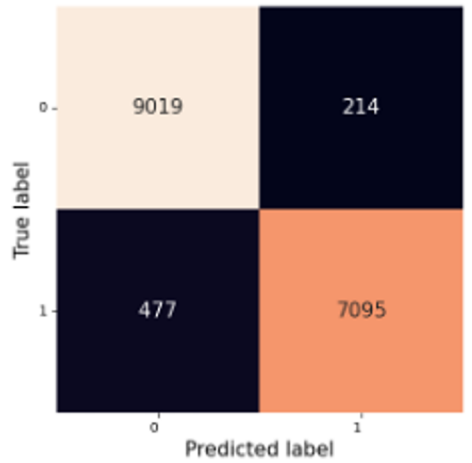
Random Forest with All Features

- Accuracy : 95.75
- F1-score : 95.21
- ROC_AUC : 99.32

Random Forest with K Features

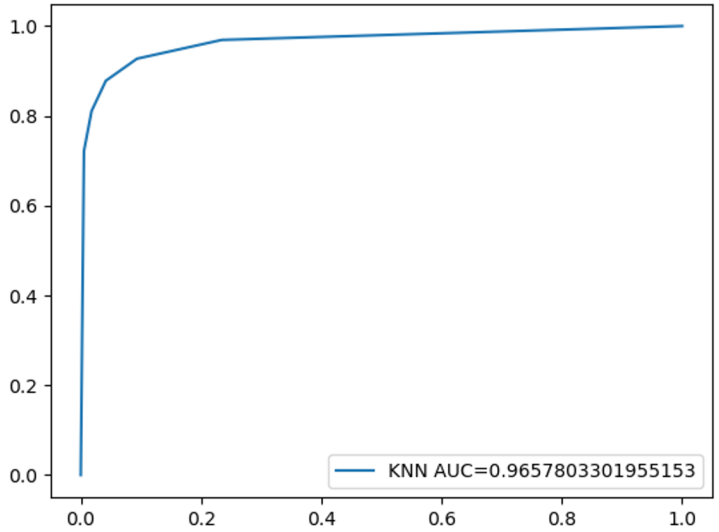
- Accuracy : 95.78
- F1-score : 95.21
- ROC_AUC : 99.34

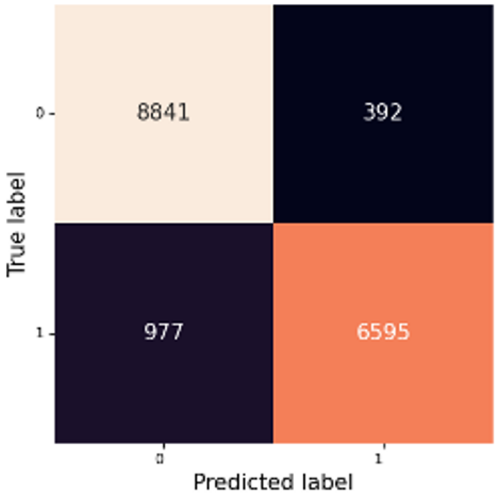




K-Nearest Neighbor

- Accuracy : 92.22
- F1-score : 91.04
- ROC_AUC : 96.58





Model Comparison and Conclusion

Model	Accuracy score	F1-score	ROC_AUC score
Logistic Regression	82.84	81.17	88.89
Naive Bayes	85.93	83.93	91.88
Decision Tree with Entropy	94.27	93.65	94.23
Random Forest	95.78	95.21	99.34
KNN	92.22	91.04	96.58

- Random forest model has the highest accuracy score 96%, F1-score 95%, and ROC_AUC score 99% when compared with other models.

- Random forest model performs the best on classification accuracy on the dataset.

Any Question?