SMARTINTERNZ EXTERNSHIP
APPLIED DATA SCIENCE
PROJECT REPORT

TITLE –

Identifying Airline Passenger Satisfaction Using Machine Learning

MEMBERS –

Ancy Sharmila – 20MIS0211 (VIT Vellore)

Hiya Kulasrestha – 20BCE2828 (VIT Vellore)

N. Madhu Hasitha – 20MIS0203 (VIT Vellore)

Madhumitha S – 20MIS0214 (VIT Vellore)

1. INTRODUCTION

1.1. OVERVIEW

The machine learning project on "Identifying airline passenger satisfaction" aims to develop a predictive model that can accurately determine the satisfaction level of airline passengers. By leveraging machine learning algorithms and techniques, the project analyzes various data points, such as flight details, customer feedback, and demographic information. The model is trained on a labeled dataset containing historical passenger data and their corresponding satisfaction ratings. Through feature engineering and model training, the project seeks to identify patterns, correlations, and key factors that significantly impact passenger satisfaction. The ultimate goal is to create a reliable tool that can assist airlines in enhancing their services and improving customer satisfaction.

1.2. PURPOSE

With the machine learning project on "Identifying Airline Passenger Satisfaction" we aim to leverage advanced analytics techniques to understand and predict customer satisfaction in the airline industry. By applying machine learning algorithms to a comprehensive dataset, the project aims to uncover the factors that significantly impact passenger satisfaction, such as flight delays, in-flight amenities, customer service, and baggage handling. The project's goal is to develop a predictive model that can accurately predict whether a passenger will be satisfied or dissatisfied based on these factors. This knowledge can empower airlines to make data-driven decisions, enhance their services, address pain points, and ultimately improve customer satisfaction and loyalty.

2. LITERATURE SURVEY

- Research by Kumar and Huang (2019) employed sentiment analysis and machine learning techniques to analyze airline customer reviews and predict passenger satisfaction. They used a combination of natural language processing and supervised learning algorithms to classify customer sentiments and determine satisfaction levels.

- In their study, Wijaya et al. (2020) utilized machine learning algorithms, including decision trees and support vector machines, to predict customer satisfaction based on airline service attributes. They found that factors such as in-flight entertainment, staff behavior, and seat comfort significantly influenced passenger satisfaction.

- The work of Zhang et al. (2020) focused on predicting airline customer satisfaction using machine learning models based on a range of features, including flight-related information, customer characteristics, and online reviews. Their results demonstrated the effectiveness of ensemble models, such as random forests and gradient boosting, in accurately predicting passenger satisfaction.

- A study by Joglekar et al. (2021) utilized machine learning techniques, including logistic regression and random forests, to analyze airline customer feedback and identify influential factors affecting satisfaction. Their research highlighted the importance of flight punctuality, baggage handling, and cabin cleanliness in determining passenger satisfaction levels.

- Wang et al. (2021) proposed a hybrid machine learning approach combining feature selection, oversampling, and ensemble learning to predict airline passenger satisfaction. Their study showed that considering both textual reviews and numerical rating data improved the prediction accuracy, providing valuable insights for airlines to enhance customer satisfaction.

2.1. EXISTING PROBLEM

- Lack of Standardized Data Collection: Many studies encounter challenges due to the absence of standardized data collection methods across airlines. Variations in data collection practices, survey formats, and attribute definitions make it difficult to compare results and establish universal benchmarks for passenger satisfaction.

- Subjectivity and Bias in Customer Feedback: Airline customer feedback often contains subjective and biased opinions, making it challenging to extract objective insights. Sentiment analysis techniques may struggle to accurately capture nuanced sentiments, leading to potential misinterpretations of customer satisfaction levels.

- Limited Availability of Comprehensive Datasets: Obtaining large and comprehensive datasets that encompass diverse aspects of the airline passenger experience can be challenging. Access to detailed information about flight operations, customer demographics, and specific service attributes may be limited, which can restrict the model's predictive capabilities.

- Handling Imbalanced Data: Imbalanced datasets, where the number of satisfied and dissatisfied passengers is significantly different, can affect the model's performance. The scarcity of dissatisfied samples may lead to biased predictions, making it challenging to accurately identify and address factors contributing to passenger dissatisfaction.

- Generalizability Across Airlines: Machine learning models trained on data from a specific airline may not generalize well to other airlines due to variations in operational procedures, service quality, and customer preferences. Transfer learning techniques or models trained on aggregated data from multiple airlines can help improve generalizability.

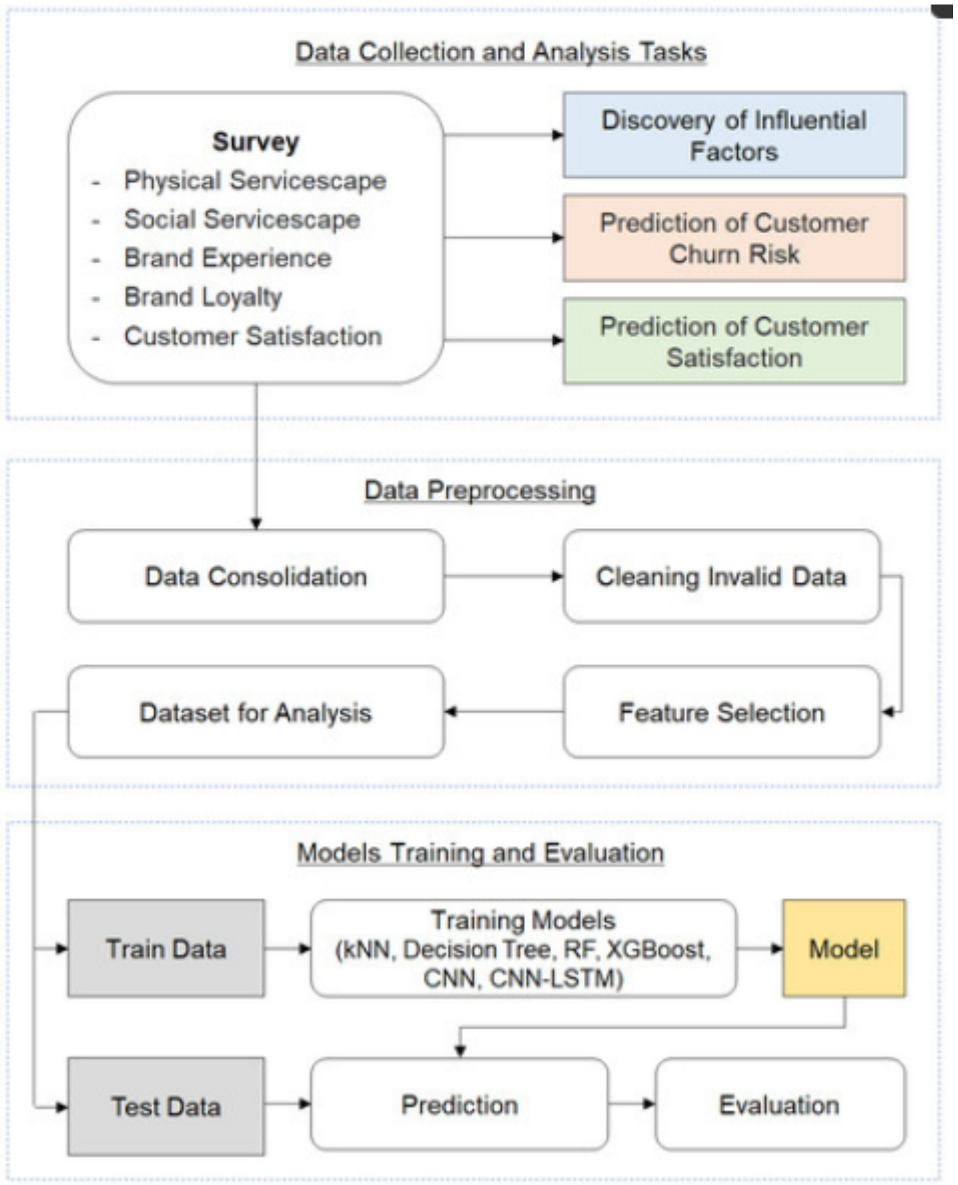
2.2. PROPOSED SOLUTION

The solution we propose for the purpose of "Identifying Airline Passenger Satisfaction" involves collecting a comprehensive dataset comprising passenger feedback, flight details, and service attributes. This dataset is then preprocessed to handle missing values, normalize features, and address imbalances. Various machine learning algorithms, such as logistic regression, decision trees, support vector machines, and ensemble methods, are applied to train predictive models. Feature selection techniques are employed to identify the most influential factors affecting passenger satisfaction. Model performance is evaluated using appropriate evaluation metrics, and techniques for interpretability and explainability are applied to gain insights into the factors driving satisfaction.

The aim is to create an accurate and interpretable predictive model that can assist airlines in improving customer satisfaction by identifying key areas for enhancement.

3. THEORETICAL ANALYSIS

3.1. BLOCK DIAGRAM



3.2. HARDWARE/SOFTWARE DESIGNING

HARDWARE DESIGN

- Computer System: A powerful computer system with sufficient memory and processing capabilities is required to handle the data preprocessing, model training, and evaluation tasks efficiently. This may include a multi-core processor, a suitable amount of RAM, and ample storage capacity to accommodate the dataset.

- Storage: Sufficient storage capacity is needed to store the dataset, preprocessed data, and trained models. Depending on the size of the dataset, a high-capacity hard disk drive (HDD) or solid-state drive (SSD) may be required.

SOFTWARE DESIGN

The software components for the machine learning project would involve various tools and frameworks. Here are some essential elements:

- Programming Language: Python is used for this machine learning projectdue to its extensive libraries and frameworks such as NumPy, Pandas, and Scikit-learn.

- Data Processing and Analysis: Libraries like Pandas and NumPy are employed for data manipulation, preprocessing, and exploratory data analysis.

- Machine Learning Frameworks: Scikit-learn, TensorFlow, andPyTorch are frameworks used for implementing machine learning algorithms, building models, and conducting training and inference.

- Model Evaluation and Validation: Techniques like cross-validation and appropriate evaluation metrics are used to assess the performance of the trained models.

- Visualization: Libraries like Matplotlib or Seaborn can be utilized for visualizing data distributions, feature importance, and model evaluation results.

- Deployment: Once the model is developed, it is integrated into a web application or dashboard using the Flask framework.

4.EXPERIMENTAL INVESTIGATIONS

- Data Collection: Gather a comprehensive dataset containing customer feedback, flight details, service attributes, and satisfaction ratings from multiple airlines. Ensure the dataset covers a diverse range of flights, destinations, and customer demographics to capture a representative sample.

- Data Preprocessing: Clean the dataset by handling missing values, removing duplicates, and addressing outliers. Perform data transformations, such as normalization or feature scaling, to ensure uniformity across variables.

- Feature Engineering: Extract relevant features from the dataset that could impact passenger satisfaction. This may include variables such as flight punctuality, in-flight services, seat comfort, baggage handling, and customer demographics. Consider incorporating additional derived features that could enhance the predictive power of the model.

- Model Selection: Experiment with various machine learning algorithms suitable for classification or regression tasks, such as logistic regression, decision trees, random forests, support vector machines, or neural networks. Explore different algorithms to identify the one that yields the best performance in predicting passenger satisfaction.

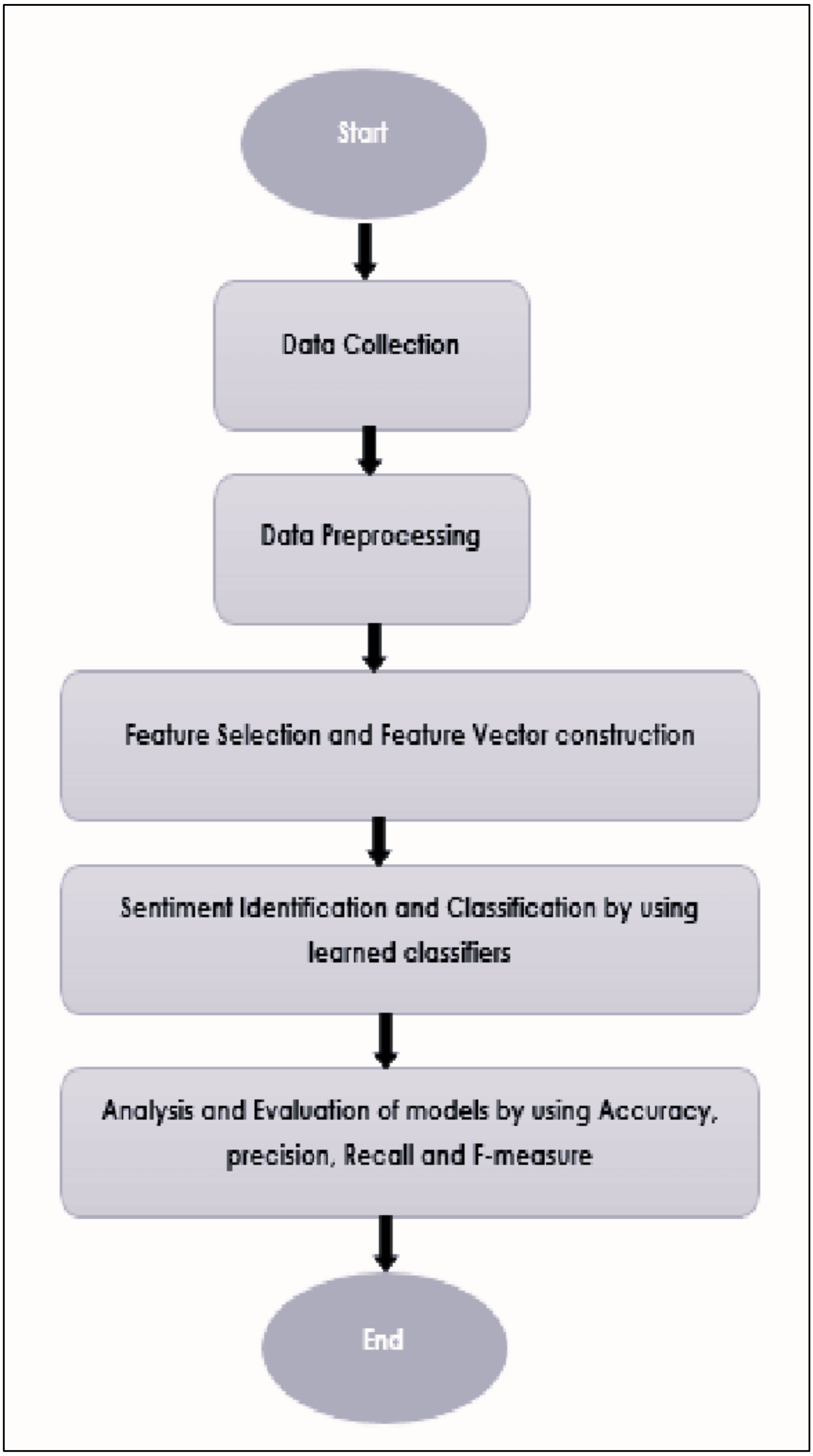
- Training and Evaluation: Split the dataset into training and testing sets, and train the selected models on the training data. Evaluate the models using appropriate evaluation metrics, such as accuracy, precision, recall, and F1 score, to assess their performance in predicting passenger satisfaction. Perform cross-validation to validate the model's robustness.

- Comparative Analysis: Compare the performance of different models and identify the most accurate and reliable one for predicting passenger satisfaction. Analyze the strengths and weaknesses of each model to gain insights into their suitability and interpretability.

- Testing on Unseen Data: Evaluate the selected model on unseen data to assess its generalization ability. This step helps ensure that the model can accurately predict passenger satisfaction for new observations outside the training dataset.

- Interpretation and Insights: Analyze the trained model to interpret the importance of different features in predicting passenger satisfaction. Extract insights and actionable recommendations for airlines to improve specific aspects of their services and enhance customer satisfaction.

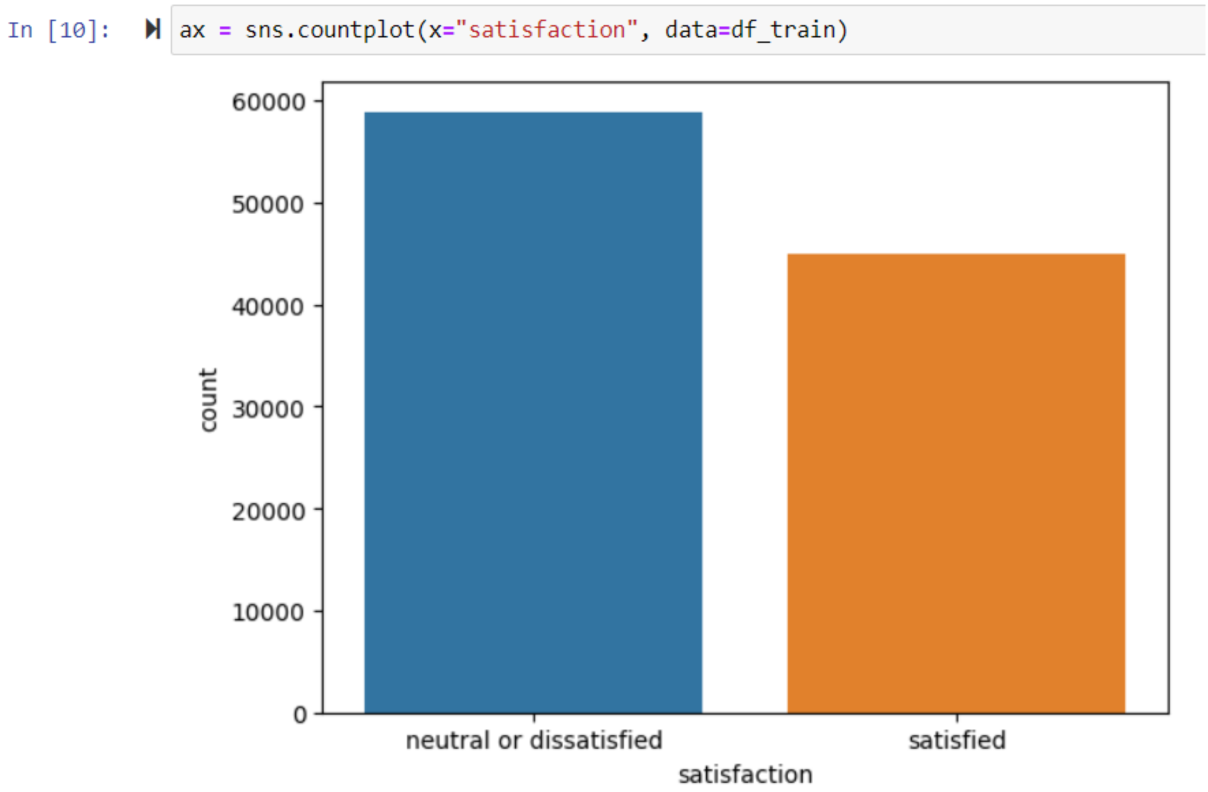
5. FLOWCHART


Start
Data Collection
Data Preprocessing
Feature Selection and Feature Vector construction
Sentiment Identification and Classification by using learned classifiers
Analysis and Evaluation of models by using Accuracy, precision, Recall and F-measure
End

6. RESULT

RESULT OF EXPLORATORY DATA ANALYSIS

- We found out that the majority of the passengers are unsatisfied/neutral with the airline services. Thats a huge concern for the airline and they need to upgrade their values and services.
- The majority of the loyal passengers are unsatisfied/neutral with the airline services, they also need to give extra attention to this category.
- The majority of Business Travel passengers are satisfied and similarlya majority of Personal Travel passengers seem unsatisfied.
- The majority of Eco and Eco Plus Class passengers seem unsatisfied/neutral with the services.
- Most of the columns with passenger input shows positive correlation with the satisfaction level.



```
In [10]:    ax = sns.countplot(x="satisfaction", data=df_train)
```

Figure 1: Satisfied vs Dissatisfied Customers

```
In [17]:  detail_barplot("Customer Type")
```

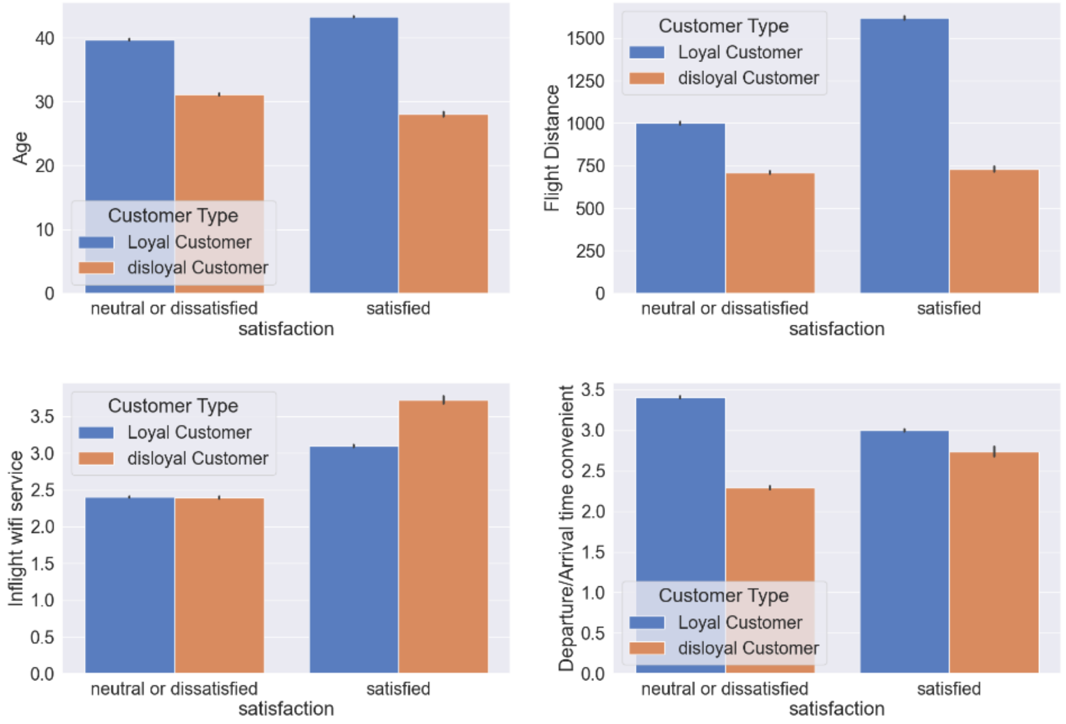


Figure 2: Customer Satifaction Based on Customer Type

```
In [ ]:  detail_barplot("Type of Travel")
```

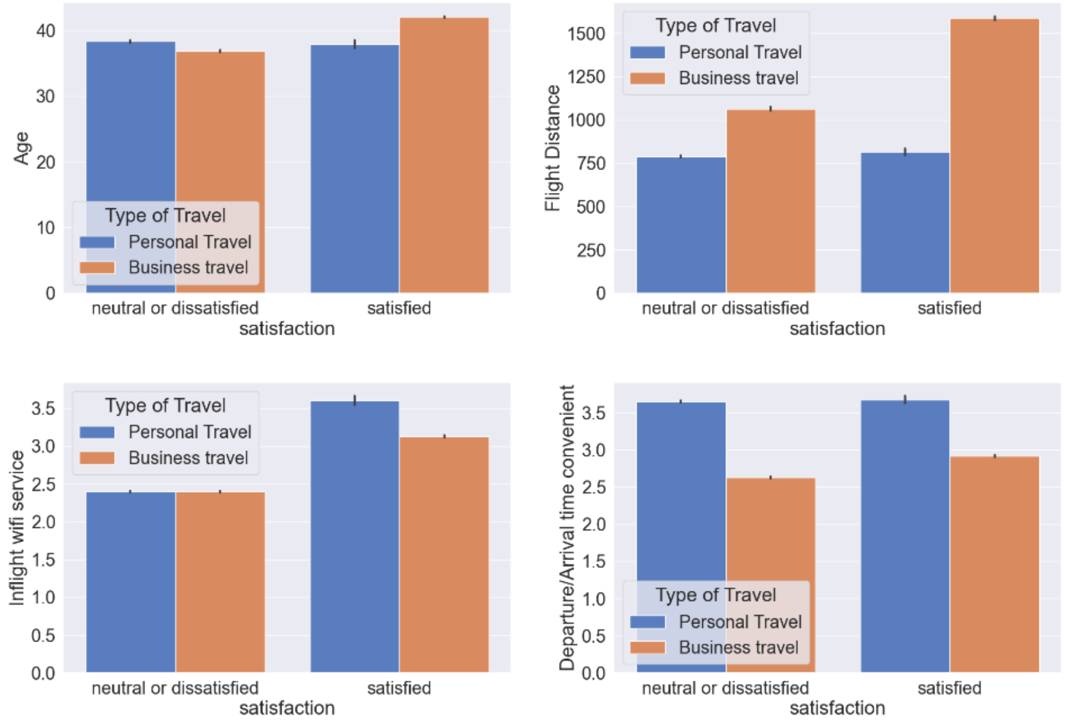


Figure 3: Customer Satisfaction Based on Type of Travel

```
In [ ]:  detail_barplot("Class")
```

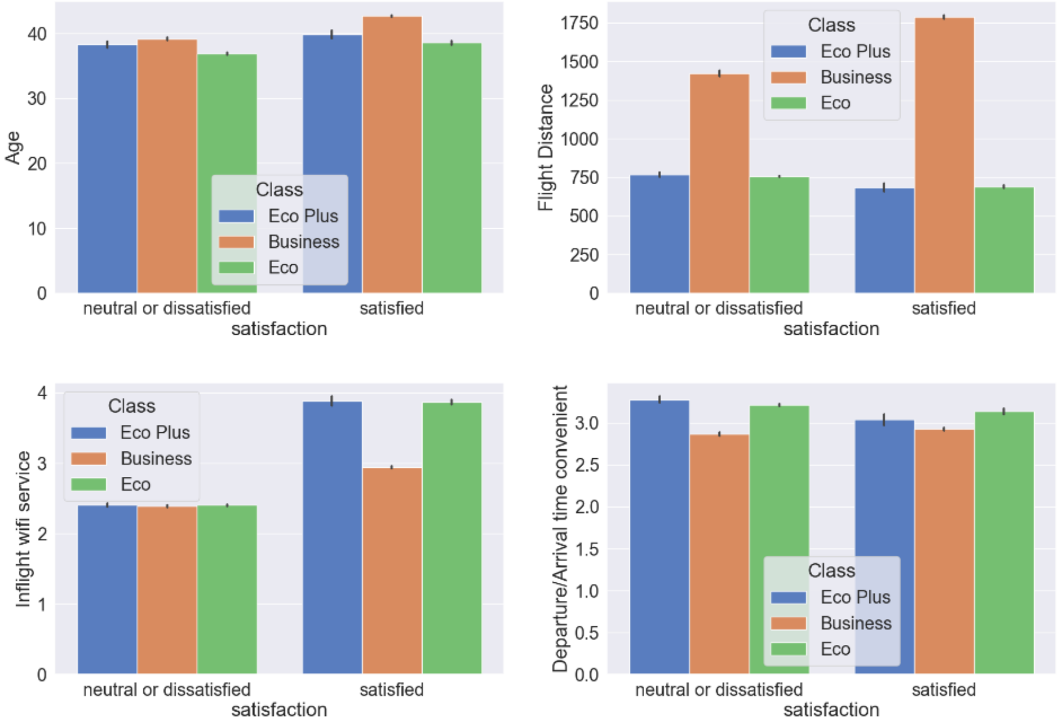


Figure 4: Customer Satisfaction Based on Class

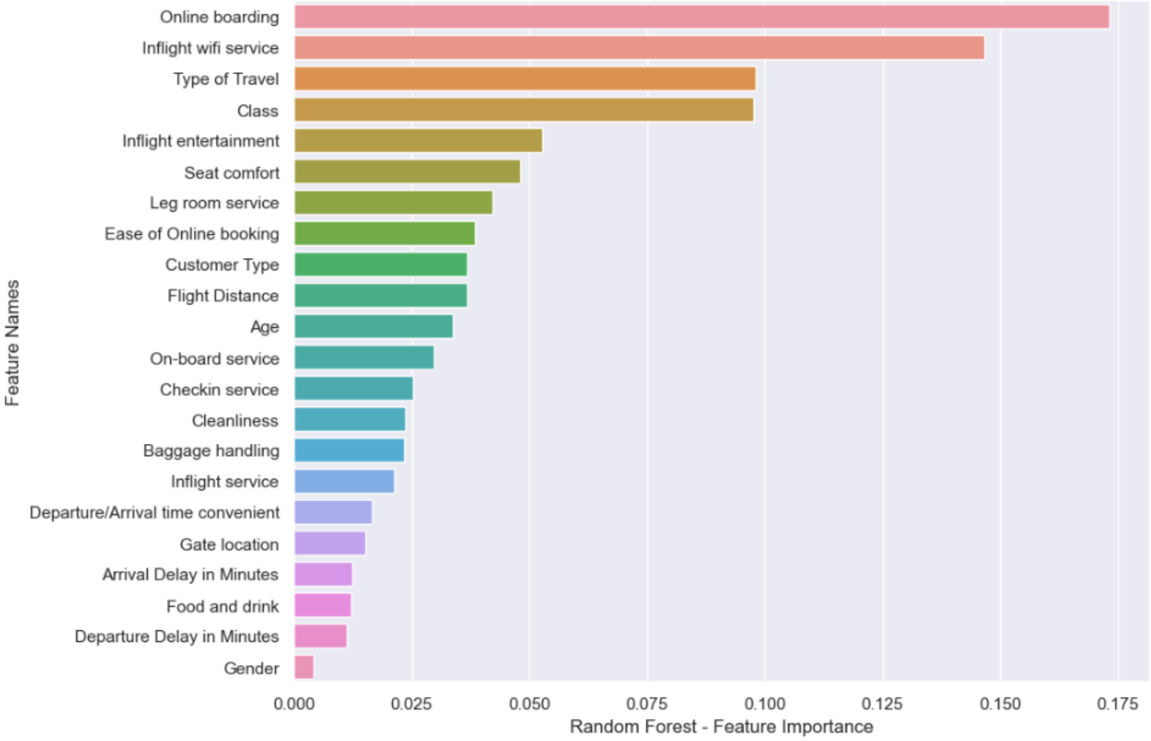
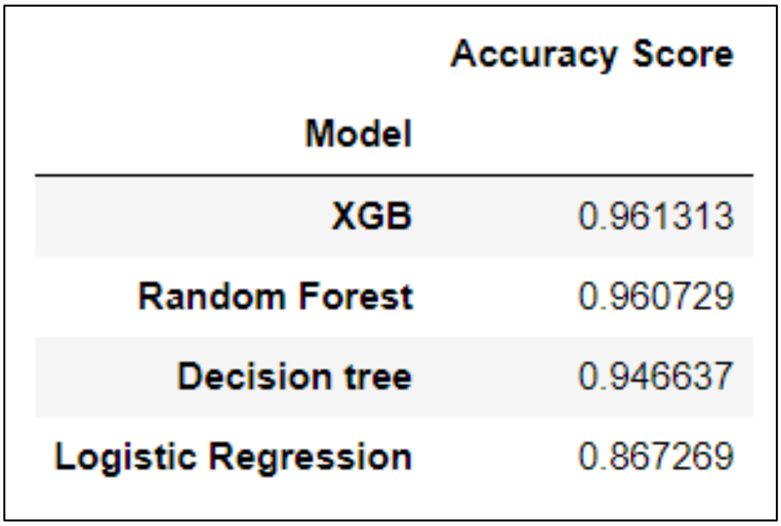


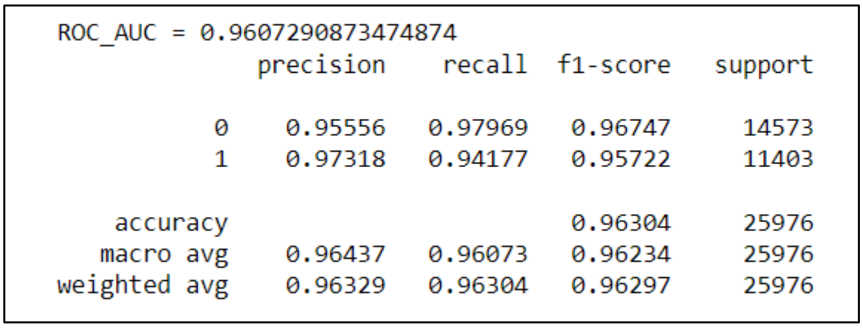
Figure 5: Feature Importance to Analyse Customer Satisfaction

RESULT OF MODEL CREATION

We found that Random Forest had the best accuracy among the models available with an accuracy of 96%.



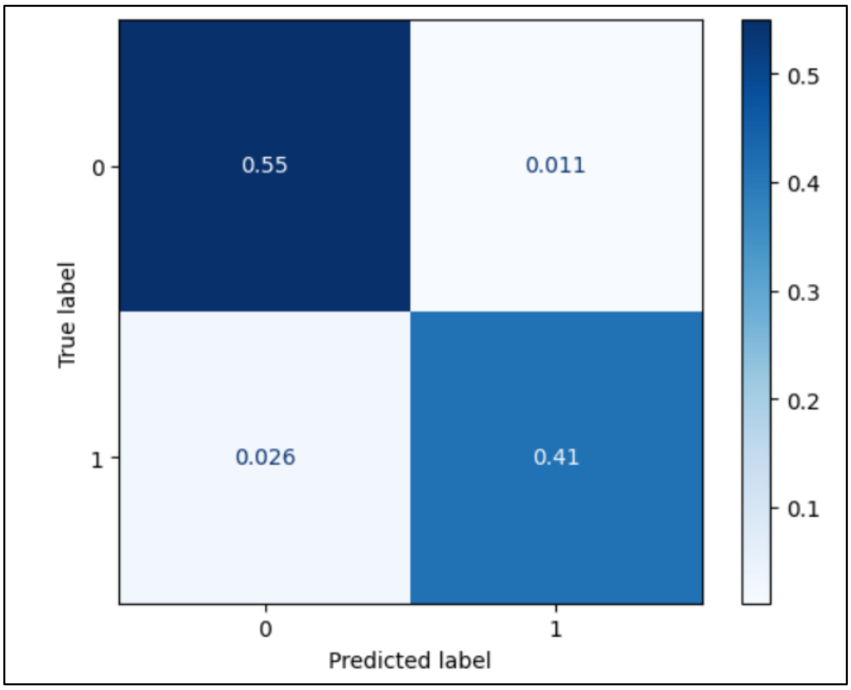
Model	Accuracy Score
XGB	0.961313
Random Forest	0.960729
Decision tree	0.946637
Logistic Regression	0.867269



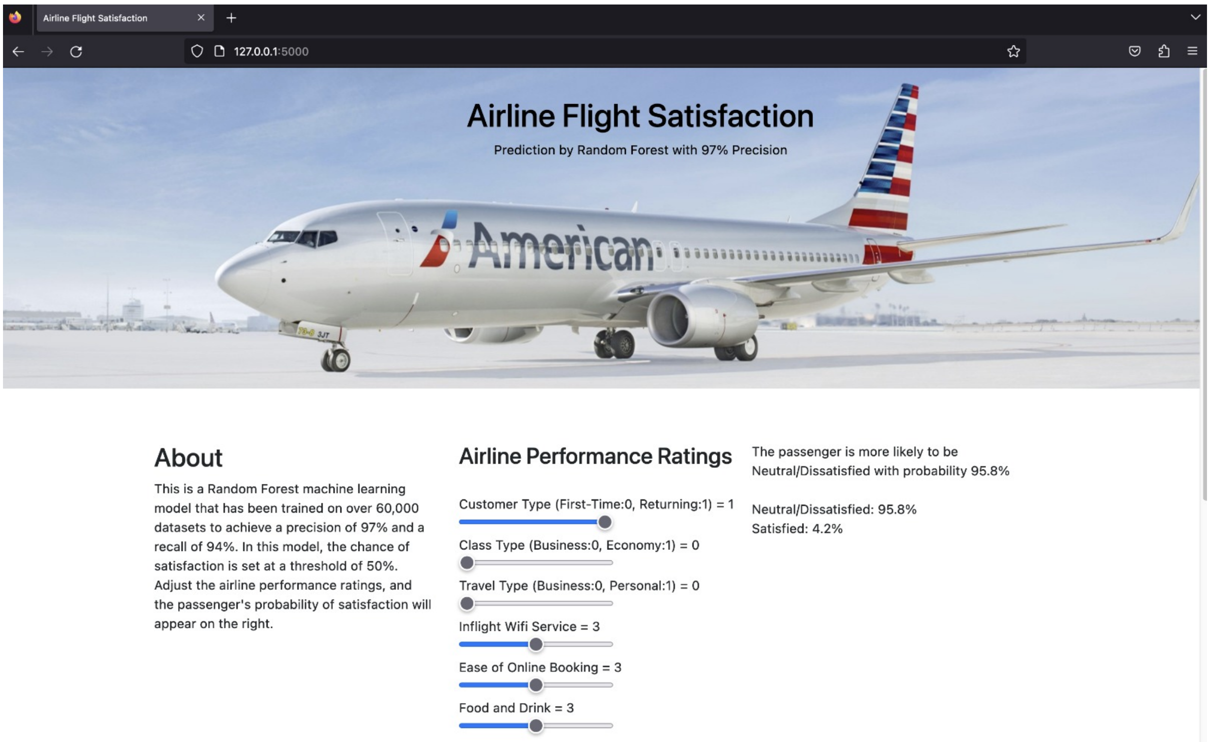
```
ROC_AUC = 0.9607290873474874
              precision    recall  f1-score   support

           0    0.95556   0.97969   0.96747     14573
           1    0.97318   0.94177   0.95722     11403

    accuracy                        0.96304     25976
   macro avg    0.96437   0.96073   0.96234     25976
weighted avg    0.96329   0.96304   0.96297     25976
```



RESULT OF DEPLOYMENT USING FLASK



Airline Flight Satisfaction

Prediction by Random Forest with 97% Precision

About

This is a Random Forest machine learning model that has been trained on over 60,000 datasets to achieve a precision of 97% and a recall of 94%. In this model, the chance of satisfaction is set at a threshold of 50%. Adjust the airline performance ratings, and the passenger's probability of satisfaction will appear on the right.

Airline Performance Ratings

Customer Type (First-Time:0, Returning:1) = 1

Class Type (Business:0, Economy:1) = 0

Travel Type (Business:0, Personal:1) = 0

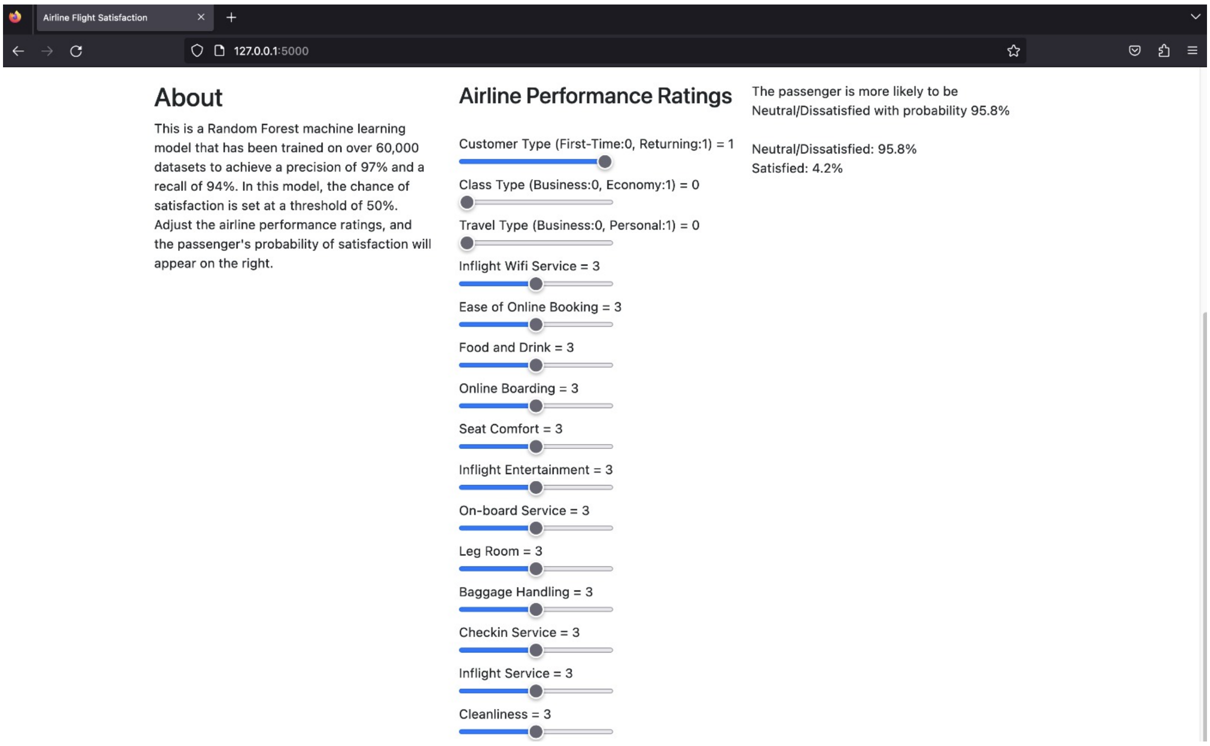
Inflight Wifi Service = 3

Ease of Online Booking = 3

Food and Drink = 3

The passenger is more likely to be Neutral/Dissatisfied with probability 95.8%

Neutral/Dissatisfied: 95.8%
Satisfied: 4.2%



About

This is a Random Forest machine learning model that has been trained on over 60,000 datasets to achieve a precision of 97% and a recall of 94%. In this model, the chance of satisfaction is set at a threshold of 50%. Adjust the airline performance ratings, and the passenger's probability of satisfaction will appear on the right.

Airline Performance Ratings

Customer Type (First-Time:0, Returning:1) = 1

Class Type (Business:0, Economy:1) = 0

Travel Type (Business:0, Personal:1) = 0

Inflight Wifi Service = 3

Ease of Online Booking = 3

Food and Drink = 3

Online Boarding = 3

Seat Comfort = 3

Inflight Entertainment = 3

On-board Service = 3

Leg Room = 3

Baggage Handling = 3

Checkin Service = 3

Inflight Service = 3

Cleanliness = 3

The passenger is more likely to be Neutral/Dissatisfied with probability 95.8%

Neutral/Dissatisfied: 95.8%
Satisfied: 4.2%

7. ADVANTAGES AND DISADVANTAGES

ADVANTAGES

- Accurate Prediction: By employing machine learning algorithms and analyzing a comprehensive dataset, the solution aims to provide accurate predictions of passenger satisfaction.

- Data-Driven Decision Making: The solution enables airlines to make data-driven decisions by leveraging insights derived from the predictive model. By understanding the factors that significantly impact passenger satisfaction, airlines can allocate resources effectively and prioritize initiatives to address specific pain points.

- Enhanced Customer Experience: With a better understanding of passenger satisfaction factors, airlines can take proactive measures to improve their services and enhance the overall customer experience.

- Efficient Resource Allocation: By identifying the specific aspects of the airline experience that influence passenger satisfaction, airlines can focus on areas that have the most significant impact on satisfaction, such as improving in-flight services, enhancing staff behavior, or streamlining baggage handling processes.

- Scalability and Generalizability: The solution can be scaled to accommodate data from multiple airlines, enabling the development of a comprehensive industry-wide passenger satisfaction model.

DISADVANTAGES

- Data Availability and Quality: The success of the solution heavily relies on the availability and quality of the dataset. Obtaining a comprehensive and representative dataset with accurate passenger feedback and relevant attributes from multiple airlines may pose challenges. Incomplete or biased data can lead to inaccurate predictions and hinder the effectiveness of the solution.

- Interpretability of Models: Machine learning models, such as neural networks or ensemble methods, can be complex and lack interpretability. Understanding the underlying factors and features driving the predictions may be challenging, making it difficult to gain actionable insights for improving passenger satisfaction.

- Subjectivity and Bias in Feedback: Airline customer feedback often contains subjective opinions and biases, making it challenging to capture and interpret sentiment accurately. Different interpretations of satisfaction levels and variations in cultural or regional expectations may impact the performance and reliability of the solution.

- Evolving Customer Preferences: Customer preferences and expectations in the airline industry are dynamic and subject to change over time. The predictive models developed using historical data may not capture emerging trends or shifting customer demands, requiring regular updates and retraining to ensure relevance.

8. APPLICATIONS

- Service Improvement: The project's insights can help airlines identify specific areas that need improvement to enhance passenger satisfaction. By understanding the factors that contribute to dissatisfaction, airlines can take targeted actions to enhance services such as in-flight amenities, customer service, on-time performance, and baggage handling.

- Customer Segmentation: The project's predictive model can assist airlines in segmenting their customer base based on satisfaction levels. This segmentation can be used to tailor marketing strategies, personalized offers, and services to different customer segments, thereby improving customer retention and loyalty.

- Operational Efficiency: By analyzing passenger satisfaction factors, airlines can optimize their operations to streamline processes and minimize disruptions. For example, understanding the impact of flight delays or cancellations on passenger satisfaction can guide airlines in implementing strategies to minimize such occurrences and improve overall operations.

- Pricing and Revenue Management: The project's insights can be leveraged in pricing and revenue management strategies. Airlines can adjust pricing based on the level of satisfaction associated with different flight routes, travel classes, or additional services. This enables airlines to optimize revenue generation while considering passenger satisfaction.

- Competitor Analysis: The project's benchmarking capabilities allow airlines to compare their passenger satisfaction levels with competitors. This analysis can help identify areas where an airline is lagging behind or excelling in comparison to industry standards, enabling them to devise strategies to gain a competitive edge.

- Customer Experience Enhancement: By understanding the factors driving passenger satisfaction, airlines can design and deliver a more personalized and enjoyable customer experience. This includes tailoring services to meet specific customer preferences, providing customized recommendations, and delivering personalized interactions throughout the customer journey.

- Feedback Analysis and Response Management: The project's predictive model can be integrated with feedback analysis systems to automatically classify and prioritize customer feedback based on satisfaction levels. This facilitates efficient response management, enabling airlines to address dissatisfied customers and resolve issues promptly.

9. CONCLUSION

In conclusion, our project presents a valuable approach to understanding and improving the airline passenger experience. By leveraging a comprehensive dataset and employing various machine learning algorithms, the project aims to accurately predict passenger satisfaction levels and identify key factors influencing satisfaction.

The project offers numerous advantages, including accurate predictions, data-driven decision-making, enhanced customer experience, efficient resource allocation, benchmarking capabilities, scalability, and continuous improvement. These benefits empower airlines to make informed decisions, optimize their services, and prioritize initiatives that enhance customer satisfaction.

However, the satisfaction feature could have been much better if neutral/unsatisfied were separate category. The data could have been divided into 3 groups as satisfied, neutral, and dissatisfied passengers. This can offer meaningful prediction as it's hard to divide between neutral and unsatisfied passengers.

Despite these challenges, the applications of the project are vast, ranging from service improvement and customer segmentation to operational efficiency, pricing and revenue management, competitor analysis, customer experience enhancement, and feedback analysis.

10. FUTURE SCOPE

The future scope of the machine learning project on "Identifying Airline Passenger Satisfaction" is promising. With advancements in data collection techniques, such as leveraging real-time customer feedback through social media and customer service interactions, the project can incorporate more diverse and dynamic data sources for improved predictions.

Additionally, the integration of natural language processing (NLP) techniques can enable the project to extract deeper insights from unstructured textual data, further enhancing the accuracy of passenger satisfaction predictions. Furthermore, the project can explore the integration of other emerging technologies, such as sentiment analysis, recommendation systems, and personalized marketing, to provide a holistic and tailored passenger experience.

As the project evolves, collaborations between airlines, data scientists, and researchers can facilitate the creation of a shared industry-wide platform that fosters continuous learning and benchmarking, ultimately leading to enhanced customer satisfaction across the entire airline industry.

11. BIBLIOGRAPHY

- Chen, K. Y., & Chang, R. D. (2019). Predicting customer satisfaction in the airline industry using machine learning techniques. Journal of Air Transport Management, 81, 101707.

- Goel, R., & Gupta, M. (2019). Airline passenger satisfaction prediction using machine learning algorithms. In 2019 2nd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-6). IEEE.

- Alamri, A. (2020). Predicting airline customer satisfaction using machine learning techniques. Journal of Retailing and Consumer Services, 52, 101930.

- Chen, Z., Chen, H., & Zhong, M. (2020). Airline customer satisfaction analysis using machine learning approaches. Journal of Air Transport Management, 84, 101774.

- Kuo, Y. C., & Chen, L. H. (2020). Improving airline service quality using machine learning techniques: A case study of passenger satisfaction prediction. Applied Sciences, 10(2), 523.

- Ma, Z., Zhang, Y., &Jin, L. (2020). Predicting airline customer satisfaction using machine learning algorithms and social media data. Journal of Air Transport Management, 85, 101822.

APPENDIX

A.SOURCE CODE

```
import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px

import xgboost as xgb

from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.decomposition import PCA
from sklearn.model_selection import train_test_split, GridSearchCV, StratifiedKFold
from sklearn.linear_model import LinearRegression, LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix, roc_auc_score, accuracy_score,
plot_confusion_matrix, classification_report
from scipy import stats

import warnings
from sklearn.exceptions import DataConversionWarning
warnings.filterwarnings(action='ignore', category=DataConversionWarning)

import os

train = pd.read_csv('/Users/ancysharmila/Desktop/Python/train.csv')
test = pd.read_csv('/Users/ancysharmila/Desktop/Python/test.csv')
train.head()
```

```
def transform_gender(x):
    if x == 'Female':
        return 1
elif x == 'Male':
        return 0
    else:
        return -1

def transform_customer_type(x):
    if x == 'Loyal Customer':
        return 1
elif x == 'disloyal Customer':
        return 0
    else:
        return -1

def transform_travel_type(x):
    if x == 'Business travel':
        return 1
elif x == 'Personal Travel':
        return 0
    else:
        return -1

def transform_class(x):
    if x == 'Business':
        return 2
elif x == 'Eco Plus':
        return 1
elif x == 'Eco':
        return 0
    else:
        return -1

def transform_satisfaction(x):
    if x == 'satisfied':
        return 1
elif x == 'neutral or dissatisfied':
        return 0
    else:
        return -1

def process_data(df):
df = df.drop(['Unnamed: 0', 'id'], axis = 1)
df['Gender'] = df['Gender'].apply(transform_gender)
df['Customer Type'] = df['Customer Type'].apply(transform_customer_type)
df['Type of Travel'] = df['Type of Travel'].apply(transform_travel_type)
df['Class'] = df['Class'].apply(transform_class)
df['satisfaction'] = df['satisfaction'].apply(transform_satisfaction)
df['Arrival Delay in Minutes'].fillna(df['Arrival Delay in Minutes'].median(), inplace = True)

    return df

train = process_data(train)
test = process_data(test)

#Define our features and target (this is helpful in case you would like to drop any features that harm model performance)
features = ['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']
target = ['satisfaction']

# Split into test and train
X_train = train[features]
y_train = train[target].to_numpy()
X_test = test[features]
y_test = test[target].to_numpy()

# Normalize Features
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.fit_transform(X_test)

corr = train.corr(method='spearman')
# Generate a mask for the upper triangle
mask = np.zeros_like(corr, dtype=np.bool)
mask[np.triu_indices_from(mask)] = True

# Set up the matplotlib figure
f, ax = plt.subplots(figsize=(20, 18))

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, annot = True, mask=mask, cmap="YlGnBu", center=0,
            square=True, linewidths=.5)

def run_model(model, X_train, y_train, X_test, y_test, verbose=True):
    if verbose == False:
model.fit(X_train,y_train, verbose=0)
    else:
model.fit(X_train,y_train)
y_pred = model.predict(X_test)
```

```
roc_auc = roc_auc_score(y_test, y_pred)
    print("ROC_AUC = {}".format(roc_auc))
    print(classification_report(y_test,y_pred,digits=5))
plot_confusion_matrix(model, X_test, y_test,cmap=plt.cm.Blues, normalize = 'all')

    return model, roc_auc

params_rf = {'max_depth': 25,
         'min_samples_leaf': 1,
         'min_samples_split': 2,
         'n_estimators': 1200,
         'random_state': 42}

model_rf = RandomForestClassifier(**params_rf)
model_rf, roc_auc_rf = run_model(model_rf, X_train, y_train, X_test, y_test)

results=pd.DataFrame({'Model':['Random Forest','XGB','Logistic Regression', 'Decision tree'],
                    'Accuracy Score':[roc_auc_rf,roc_auc_xgb,roc_auc_log,roc_auc_dt]})
result_df=results.sort_values(by='Accuracy Score', ascending=False)
result_df=result_df.set_index('Model')
result_df

model = pickle.load(open('model.pkl','rb'))

print(model.predict([[1.8]]))

pickle.dump(model_xgb, open('model.pkl','wb'))

model = pickle.load(open('model.pkl','rb'))
print(model.predict([[1.8]]))
```