# Group 2 Final Report

Anqi Xue (ax2170), Tanisha Aggrawal (ta2709), Vishal Bhardwaj (vb2573)

## Background

Air travel is an essential mode of transportation for many people, but the quality of the experience can vary greatly. Airlines often conduct passenger satisfaction surveys to better understand passenger satisfaction and identify key factors that influence it. The Airline Passenger Satisfaction classification problem seeks to analyze data from one such survey to build a classification model that can accurately predict whether a passenger is satisfied or dissatisfied based on these factors. By analyzing the data and identifying which factors are highly correlated with passenger satisfaction, airlines can make informed decisions about improving the passenger experience and, ultimately, improving overall performance and profitability.

## Objective

Predicting customer satisfaction considering 22 features using various ML techniques.

## Data Preprocessing

We start by analyzing the missing values in the data and remove the information about the passenger (the whole row) if part of their data is missing. Later we observe the distribution of the target variable in the dataset, revealing a slightly imbalanced dataset with 'satisfied': 'neutral or dissatisfied' = 0.43 : 0.57. Next, we drop the columns 'id' and 'satisfaction' for both the development and test dataset and encode the 4 categorical features with String values using Ordinal Encoding. Finally, we scale the data using StandardScaler().

## Exploratory Data Analysis

EDA is performed in two different directions - the features which are categorical and the features that are continuous. The analysis of the categorical features is distributed into the binary categorical features and the features containing more than 2 ordinal categorical features. Following are some insights from the EDA.

- Equitable distribution of the target variable "Customer Satisfaction" with respect to the features "Gender", and "Customer Type - Loyal/Disloyal".
- The significant variation in satisfaction of travelers who traveled by Business and Economy class. 69.5% of the Business class were satisfied whereas only 18.6% of the Economy class were satisfied.
- The contrast in the dataset based on customer loyalty - the dataset contains 4.5 times more loyal customers.
- We observed that there is a high chance that a customer is satisfied if they rate 5 on the features of Inflight Wi-Fi service, Ease of Online Booking, and Online Boarding.
- Highly likely that the customer is unsatisfied if they rate any feature less than 3 except for features such as Gate Location, Departure/Arrival time convenience, and Food and Drink.

- There exists a general trend suggesting that a higher rating implies a higher chance of satisfaction for any feature.
- There are more dissatisfied customers than satisfied ones in the age groups 7-38 years and 61-85 years. There are more satisfied customers than dissatisfied ones in the remaining age group 39-60 years.
- The departure and arrival delay variables are highly correlated as expected but they display no clear trend in terms of overall customer satisfaction.

## Modeling

For modeling customer satisfaction, we tried multiple classification techniques. For all the techniques we are using the processed scaled data and comparing the model performance on the test data using multiple classification metrics. To tune the hyperparameters we are using k-fold cross-validation and using GridSearchCV for performing the analysis.

- Logistic Regression

We start with a simple regression-based model and use the following set of hyperparameters.
*param_grid = {'C': [0.1, 1.0, 10.0, 100.0, 1000], 'penalty': ['l2']}*
Best hyperparameters: {'C': 100.0, 'penalty': 'l2'}
Best cross-validation score: 0.840
Test accuracy of the best model: 0.819
From the above results, we see that logistic regression performs well on the dataset.

- Random Forest

Next, we try a tree-based approach to check whether nonlinear functions fit our data better. We used the following set of hyperparameters.
*param_grid = {'max_features': ['sqrt', 'log2', 0.3, 0.6, 0.9],'n_estimators': [30, 60, 90, 120, 150]}*
Best hyperparameters: {'max_features': 0.3, 'n_estimators': 150}
Best cross-validation score: 0.963
Test accuracy of the best model: 0.964
From the above results, we can see that random forest performs significantly better than logistic regression even on unseen test data.

- XGBoost

We continue with a more complex boosting algorithm to see if the results improve further.
*param_grid = {'learning_rate': [0.001, 0.01, 0.1, 1], 'max_depth': [6, 9, 12, 15], 'n_estimators': [50, 100, 150]}*
Best hyperparameters: {'learning_rate': 0.1, 'max_depth': 9, 'n_estimators': 150}
Best cross-validation score: 0.964
Test accuracy of the best model: 0.964
We infer that the performance of the model is great however, it has not improved significantly as compared to the previous model.

- Support Vector Machines

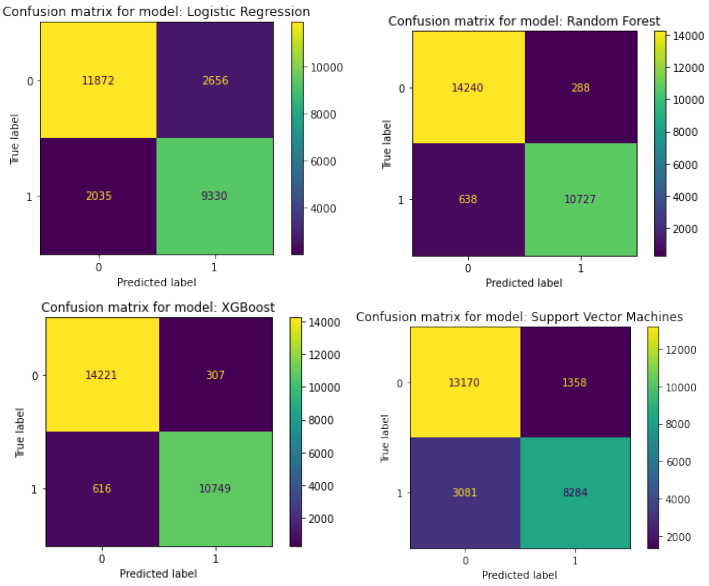
We try a final model, SVMs using SGDClassifier to see if we get better performance.

```
param_grid = {'loss': ['hinge', 'squared_hinge'], 'alpha': [1e-3, 1e-4, 1e-5]}
Best hyperparameters: {'alpha': 0.001, 'loss': 'squared_hinge'}
Best cross-validation score: 0.815
Test accuracy of the best model: 0.829
```

We observe that SVMs perform poorly on our dataset as compared to Random Forest and XGBoost.

## Comparing all the results

From preliminary observation, we can see that Random Forest and XGBoost perform extremely well on our dataset followed by Logistic Regression and SVMs. Following is the confusion matrix for all the models.




From this, we can see that Logistic Regression has very high false positives as compared to the other models. We also observe that SVMs have very high false negatives followed by Logistic Regression. Random Forest and XGBoost have very similar distributions with a slightly higher number of false negatives in Random Forest as compared to XGBoost and XGBoost having a slightly higher number of false positives. Since here we are predicting customer satisfaction, we are more concerned about the customers that are dissatisfied or in other words the false positives. Therefore, on taking that into consideration, for our use case **Random Forest would be a better model**.