Table of Contents

Table of Figures

List of Abbreviations

Passenger airline satisfaction, machine learning, linear regression, airline travel.

1. Introduction

The International Corporation dataset named as “skywards” provides extensive information related to passengers and their top-tier experiences of flights in terms of satisfaction level. The dataset contains a number of features which includes travel related data, demographic information and service quality metrics. All the factors are contributing to evaluating the passenger’s satisfaction level. So, understanding these factors is crucial as it directly affects satisfaction in order to enhance the quality of service, improve operational efficiency and ensure customer retention for the credibility of passengers (E. Park, 2019).

In this work, we are analyzing the information in the dataset of the passengers who share their information and experience in terms of satisfaction. For that purpose, a dataset (skywards) contains data of over 103,000 passengers for the current calendar year. We are analyzing the sense of data and using analytical models such as linear regression and artificial neural network model for learning the patterns and visualize the data that helps us better understanding of which key factors are playing role on the level of satisfaction for passengers using the airline. Lastly, the accuracy of models is analyzed, and correlation matrix is made to check the effects of features on the output.

2. Key factors that impact on passenger satisfaction

It is crucial to ensure the passengers’ satisfaction of the airline services though, it is influenced by a number of factors. Some of the factors that are considered, and data is collected as given below.

Travel details

Whenever it is asked by the passenger to share their experience, some travel related information mandatory to be asked. These are class, type of travel, destination and flight distance etc.

1. Class: The type of class (such as business, economy and first class) directly influences the satisfaction of passengers as there is a difference in the level of services and comfort provided to the members based on their preferences.

2. Type of travel: the reason for traveling also greatly affects satisfaction as travel can be done for business, leisure, study etc. (Wan, 2015).
3. Destination: The people of different regions have different perceptions towards the services they consider important. So, standards and differences in service expectation might get affected by satisfaction.
4. Distance of flight: Passengers going through airlines have different experiences in different countries. High service standards are expected more in long haul flights as compared to short haul flights (Park, 2019).

Demographics information

Some of factors related to person such as age or age band also impacts the experience in airline satisfaction.

1. Age band: One of the characteristics is age that influence the output so grouping the age into bands for example 7-18, 18-25, 26-35 these can better identify trends and patterns in satisfaction level of services on airline flights.
2. Age: passengers of different ages travel by air that also affects satisfaction. For example, young passengers are interested in services such as food, Wi-Fi and flight entertainment stuff whereas older passengers seek comfort and ease of service (Tsafarakis, 2018).

Quality of Service

1. Arrival time convenience: arrival of flights can help people in judging the discipline of airlines and can influence satisfaction. Nobody likes to travel through airlines with extreme delays in flights.
2. Ease of booking (online): pretravel satisfaction can be enhanced by providing an easy and seamless experience in booking airlines.
3. Wi-Fi facility in the flight: availability and quality of Internet can also play a role in satisfaction especially for those who are traveling for business purposes (Hayadi, 2021).
4. Food and drink: Quality food and snacks, options of the food and drink or beverages also serve to affect satisfaction. People from the home country like to experience their food and they are satisfied if the quality is considered in providing meals.

5. Sitting comfort: The seat space or rest space vary based on the type of travel they are doing. Although, for long flights comfortable seats play an important role in satisfying passengers (Park, 2019).

6. Location of gates: The factor of distance from gate to the flight also creates the thought of satisfaction in people’s minds. Although this factor matters to older passengers as compared to younger ones.

7. On board services: There are additional services provided by cabin crew that can be considered as general quality services.

8. Legroom services: another comfort factor is the adequate area provided to keep their legs so that they can easily sit on their seats for long flights.

9. Boarding process: Online boarding process plays vital role in satisfaction. People like to experience efficient and user-friendly online boarding process.

10. Entertainment facility in flight: Options for entertainment such as movie screens, magazines, etc also comes under the quality of services provided during traveling.

11. Checking services: friendliness and efficient checking services also add up to good traveling experience when traveling on flights (Tahanisaz, 2020).

12. Inflight service: Overall services in providing additional stuff such as tissues, snacks, helping sources can become a great deal in the quality of services. Personalized services can be provided based on the traveling class such as business, economy etc. in order to meet the needs of different passengers. For example, leisure travelers focus more on entertainment and food options, whereas business passengers value more to having Internet facilities and workspace comfort zone.

13. Handling baggage: traveling stress is greatly reduced by efficient baggage handling.

14. Cleanliness in flights: Passengers satisfaction and comfort can be affected by cleanliness of air flights and washrooms (Park, 2019).

Metrics related to flights Delay

1. Arrival delay in minutes: Flight delays in arrival can influence pre and post flight experience thus affecting satisfaction.
2. Departure delays in minutes: delays in flights also reduce the satisfaction of travel and thus can frustrate the overall journey of air flights (Wan, 2015).

3. Tasks (with independent details and Independent Research)

The analysis is done in several parts to give thorough insight of dataset and their meaningful derivation. Each task is required to be supported by independent research to ensure the effectiveness of results.

Task 01: Data Preprocessing and exploration

In this step the task involves importing libraries, loading the dataset, checking its description and information, looking for and resolving missing values and performing initial data exploration to understand type of features and patterns and trends of records (Mishra, 2020).

General processing of data involves steps such as handling missing values, normalizing the data and removing outliers in the dataset. Independent research can be done by applying techniques for handling missing values and normalizing the data.

In the phase of output and analysis of that output, initial exploration reveals the structure of dataset, missing values and statistics required for numerical features. The understandability of such aspects is required for analysis of dataset and preparing it for preprocessing phase (İncir, 2024).

Task 2: Visualizing the data

To understand the distribution of features along with their relationships and passenger satisfaction, data needs to be visualized. Visualizations are done in the form of histogram, box plot and heatmaps that can be helpful in understanding the correlations between the features (Srivastava, 2023).

The implementation of these visualization techniques is done in python using seaborn library. These visualizations show the results of satisfaction against the distribution of customers. This plot shows the number of satisfied passengers and unsatisfied passengers by providing a baseline in understanding the overall level of satisfaction. The heatmap, which is used as correlation matrix, also dictates the relationship between overall features with key output that is satisfaction, thus highlighting the key factors that impact the resulting output (Skopal, 2024). The results can be seen in figure 1 and figure 2 as shown below.

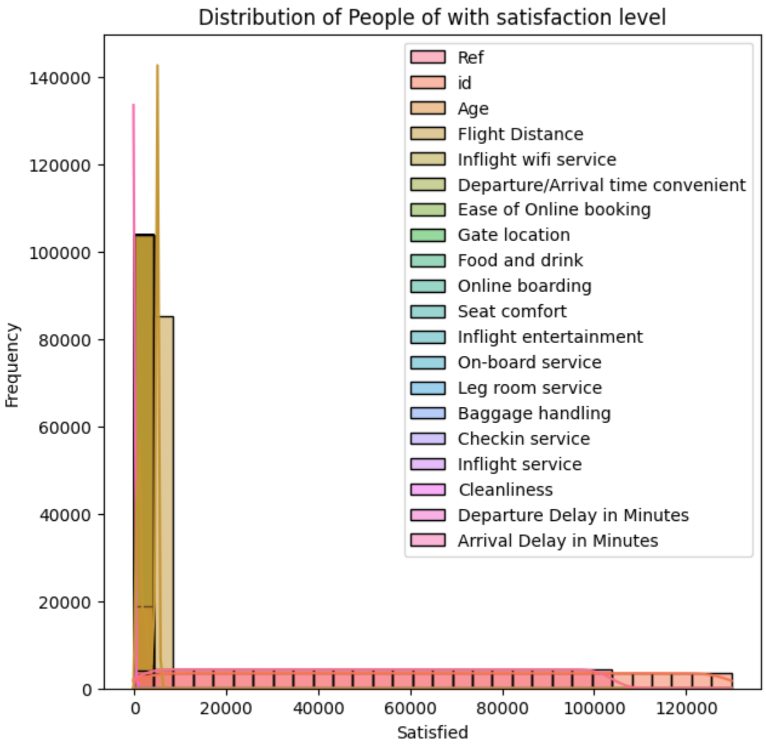

Figure 1: Correlation matrix of features for satisfied passenger (self)

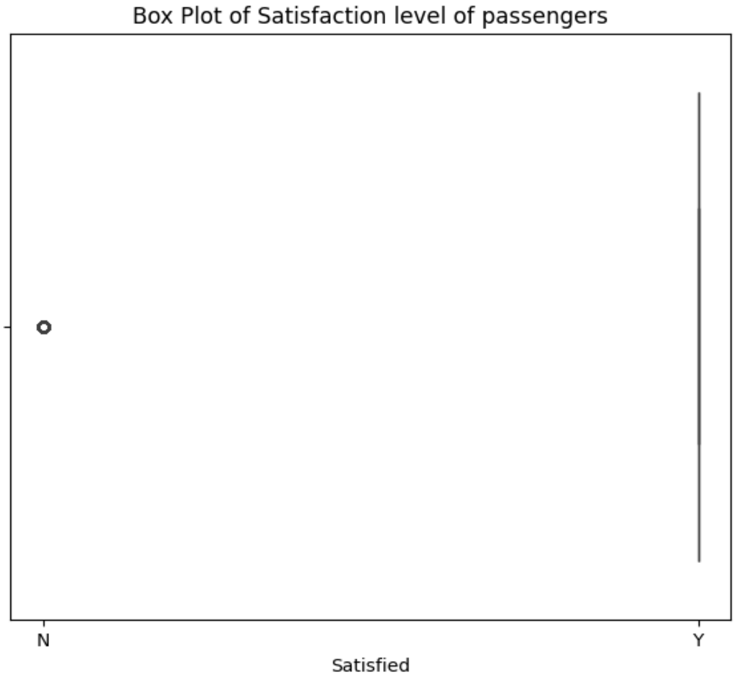

Figure 2: Boxplot of satisfaction level of passengers (self)

Task 3: Feature Engineering and feature extraction

To identify and analyze the key features and remove the features that have zero impact on the target output, a detailed feature engineering is done, which involves feature selection and feature extraction. The performance of models is greatly influenced by performing feature engineering to the dataset to make it more meaningful. Various feature extraction techniques such as correlation analysis and recursive feature elimination can be used to identify more key features (McConnell, 2024).

The output in figure 1 shows the list of ten important features that directly impact satisfaction of passenger as they have highest scores, thus a clear indication of importance in the prediction of passenger satisfaction. Just like other phases of exploratory analysis, feature engineering serves a role in simplifying the model's prediction and focusing on the relevant features that impact the output results.

Task 4: Preparing the model for prediction and evaluation

Passenger satisfaction is predicted on the basis of the data that is passed to the machine learning models for learning and then predict results. There are multiple types of machine learning models and analytical models such as decision tree, linear regression (Hope, 2020), random forest, neural network, bagging methods, boosting methods and complex models for the prediction of passengers' satisfaction. For evaluation of such models there are multiple performance metrics that can be used such as accuracy, recall and precision) for finding the best model. In our case, we use logistic regression model (James, 2023), and its performance is analyzed using mean square root error. Where on the other hand, artificial neural networks (Yang, 2020) are also employed for prediction and accuracy is the metrics used for predicting the results of models.

The results of models are evaluated using accuracy, which is the best measure for machine learning models, thus a better evaluation of results for passengers' satisfaction. The true positives and true negatives along with false positive and false negative are visualized in confusion matrix that can help in better understanding the model's accuracy and misclassification. The linear regression values for mean squared error is 1.05, R^2 score is 0.77, slope (Coefficient) is 2.82 and intercept is 4.39. We run the artificial neural network model up to 50 iterations and at the last iteration the accuracy of the model is 0.84. Figure 3 shows the actual results and predicted results by linear regression model. Figure 4 shows the scatter plot for the model and how the results are predicted by linear regression.

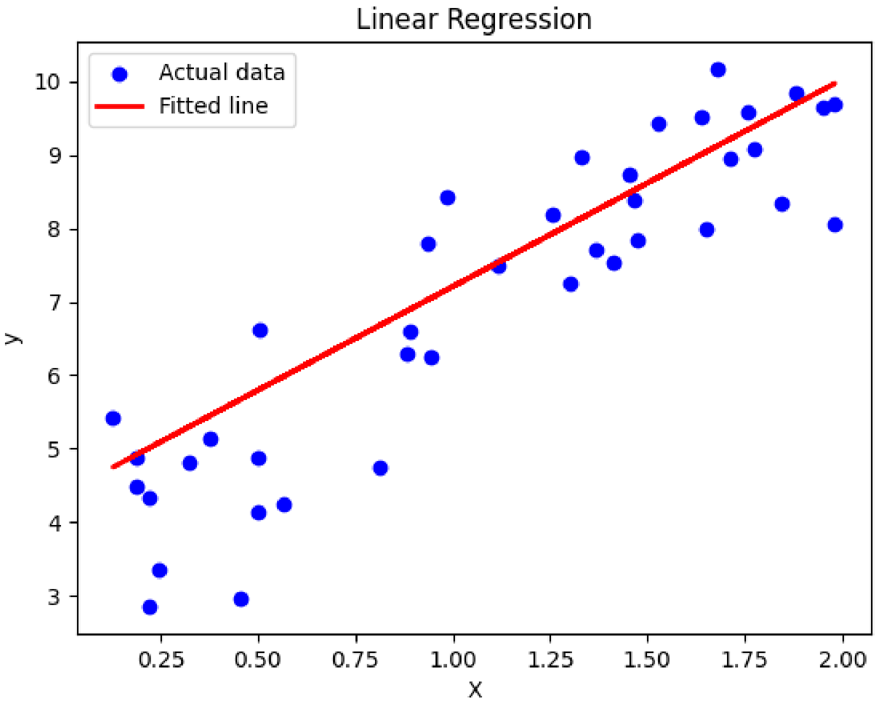

Figure 3: MSE of linear regression model (self)

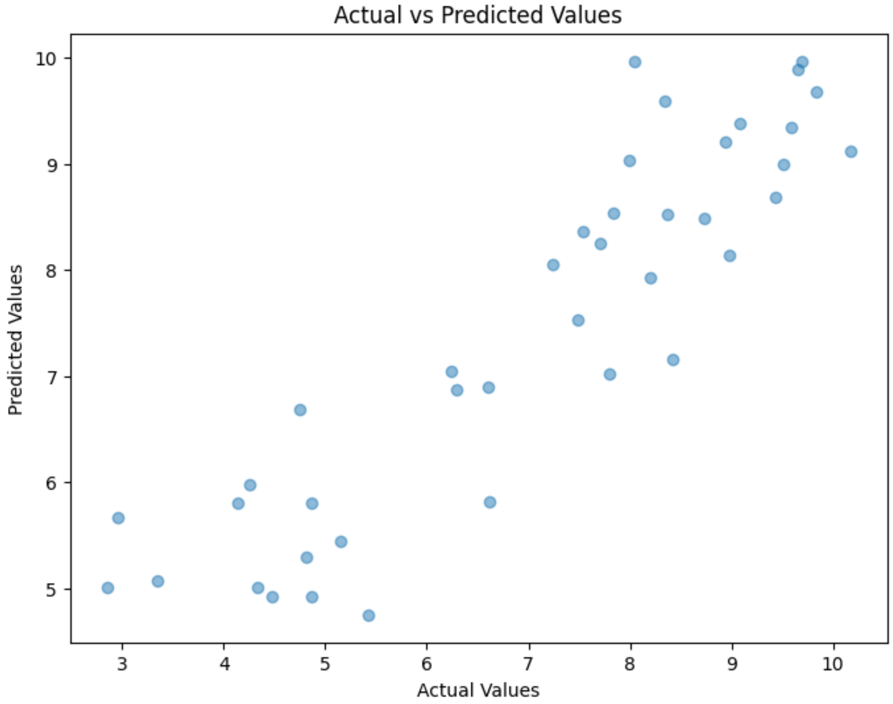

Figure 4: Scatter plot of Linear regression model (self)

Figure 5 shows the accuracy of artificial neural network model which is 0.84 in our case. Figure 6 shows the confusion matrix that shows the results of how many true positives and true negatives are predicted by model.

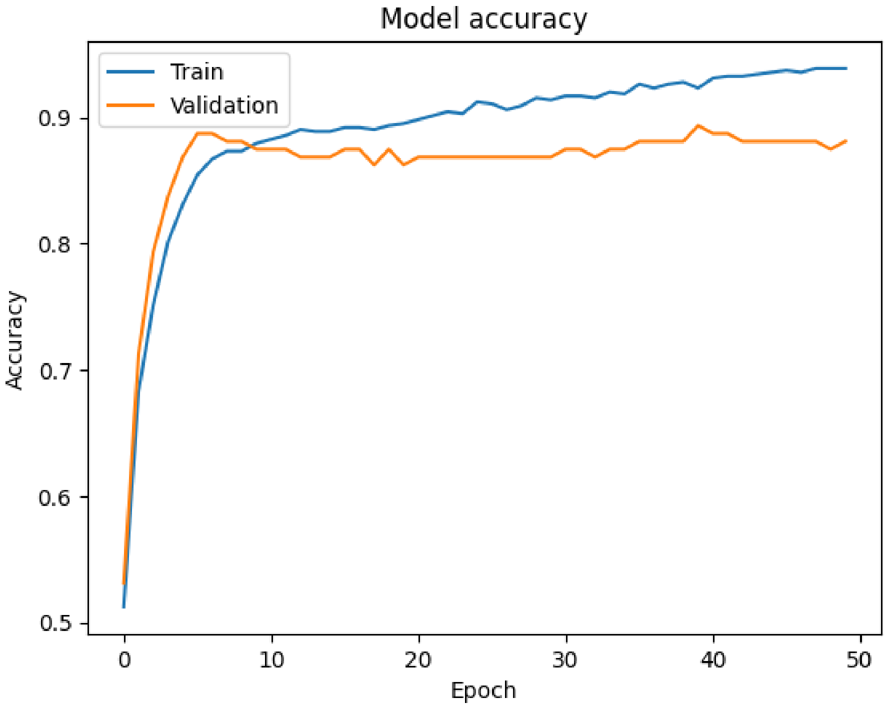

Figure 5: Accuracy of model in terms of passenger satisfaction prediction (self)

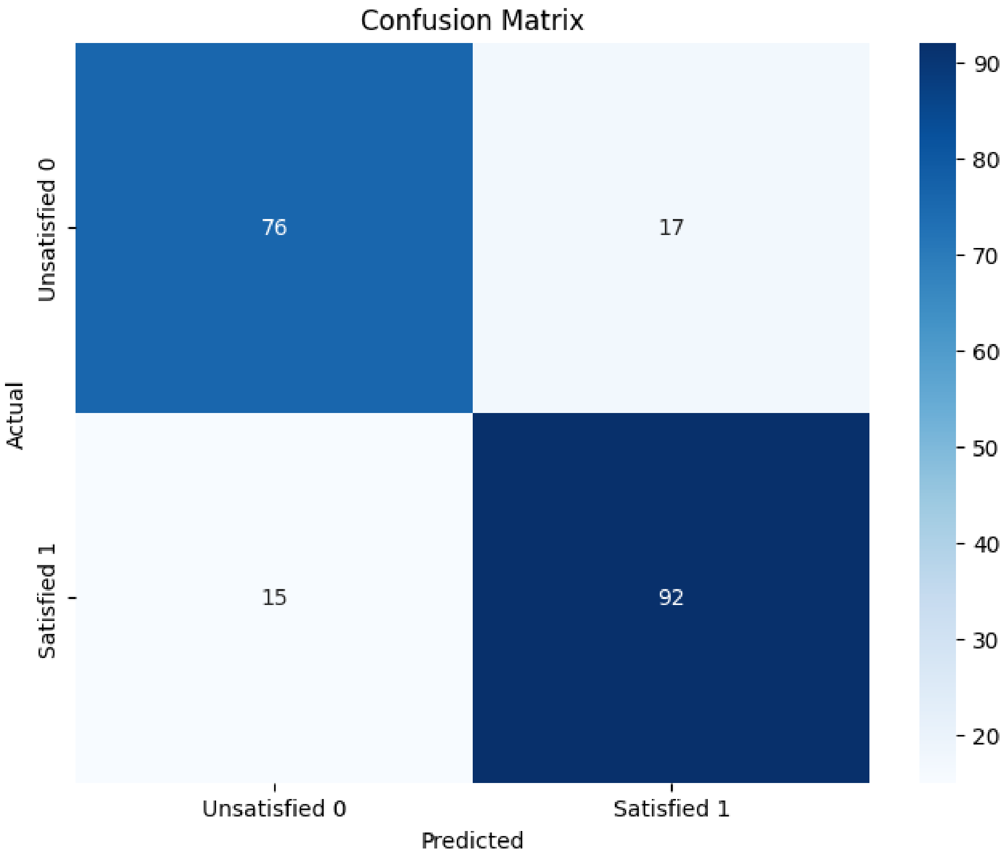

Figure 6: Confusion matrix of model results (self)

4. Recommendations

To enhance customer satisfaction, the following recommendations can be made on the basis of analysis. These recommendations can be incorporated in terms of improving customer satisfaction, optimizing flight operations, improving in-flight services and tailor services given for different classes, online experience, baggage handling, and improved cleanliness.

Improving Inflight Services

The quality of in-flight services can be improved by enhancing the availability of in-flight Wi-Fi and overall comfort to all types of travel classes. A number of factors have been recognized that serve as significant contributors to the satisfaction of passengers (Lucini, 2020).

Flight operations optimization

Flights operations can be improved by reducing the delays of arrival and departure and overall travel experience. This can be solved by implementing efficient real time delay management systems along with scheduling mechanisms to achieve optimized results of airline services (Park, 2019).

Services provided to passengers of different class level

Personalized services can be provided based on the traveling class such as business, economy etc. in order to meet the needs of different passengers. For example, leisure travelers focus more on entertainment and food options, whereas business passengers value more to having Internet facilities and workspace comfort zone.

Online experience enhancement

By improving the user interface and providing streamline of online booking process, it can be helpful for a seamless experience for booking on board. This can significantly enhance satisfaction of customers and reduce the potential stress of passengers (E. Park, 2019).

Cleanliness monitoring and improvement

Cleanliness should be high to maintain high standards by regularly assessing the airlines. Cleanliness is a key factor that directly affects the satisfaction of passengers and therefore is a critical factor to be improved. The cleanliness of washrooms and seating place must be considered with full attention.

Baggage handling in efficient manner

Better handling of baggage is required for ensuring the safe and timely delivery of luggage. Efficient handling of baggage may help in reducing travel stress and thus enhances overall passenger experience. Luggage of different types of users may vary and can contain sensitive stuff so it needs to be handled with care (Park, 2019).

5. Next steps

These recommendations can be ensured by implementing the following steps as given below:

1. Loading the actual dataset: It is to be ensured the dataset is available in a suitable format such as csv file so that it can be loaded and analyzed.

2. Execution of code: The implementation of code is done in the compatible environment so that analysis can be performed.
3. Detailed analysis: To improve customer satisfaction, further investigation is required to identify the factors that directly affect the output results.
4. Implementation of recommendations: use insight gained from the detailed analysis so that further improvements can be made to enhance passenger satisfaction (Park, 2019).

In this work, we are analyzing the information in the dataset of the passengers who share their information and experience in terms of satisfaction. For that purpose, a dataset (skywards) contains data of over 103,000 passengers for the current calendar year. We are analyzing the sense of data and using analytical models such as linear regression and artificial neural network model for learning the patterns and visualize the data that helps us better understanding of which key factors are playing role on the level of satisfaction for passengers using the airline. Lastly, the accuracy of models is analyzed, and correlation matrix is made to check the effects of features on the output.

5. References

E. Park, Y. J. (2019). Determinants of customer satisfaction with airline services: An analysis of customer feedback big data. *Journal of Retailing and Consumer Services*, 186-190.

Hayadi, B. H. (2021). Predicting Airline Passenger Satisfaction with Classification Algorithms. *International Journal of Informatics and Information Systems*, 82-94.

Hope, T. M. (2020). Linear regression. In Machine Learning. *Academic Press.*, 67-81.

İncir, R. &. (2024). A study on effective data preprocessing and augmentation method in diabetic retinopathy classification using pre-trained deep learning approaches. *Multimedia Tools and Applications*, 12185-12208.

James, G. W. (2023). Linear regression. In An introduction to statistical learning: With applications in python. *Cham: Springer International Publishing.*, 69-134.

Lucini, F. R. (2020). Text mining approach to explore dimensions of airline customer satisfaction using online customer reviews. . *Journal of Air Transport Management*, 101760.

McConnell, B. V. (2024). On Monitoring Brain Health from the Depths of Sleep: Feature Engineering and Machine Learning Insights for Digital Biomarker Development. *bioRxiv, 2024-02*.

Mishra, P. B. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry*, 116045.

Park, E. (2019). The role of satisfaction on customer reuse to airline services: An application of Big Data approaches. *Journal of Retailing and Consumer Services*, (pp. 370-374).

Skopal, T. P. (2024). Visualizations for universal deep-feature representations: survey and taxonomy. . *Knowledge and Information Systems.*, 811-840.

Srivastava, D. (2023). An Introduction to Data Visualization Tools and Techniques in Various Domains. *Int J Comput Trends Technol*, 125-130.

Tahanisaz, S. (2020). Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry. . *Journal of Air Transport Management*, 101764.

Tsafarakis, S. K. (2018). A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. . *Journal of air transport management*, 61-75.

Wan, Y. a. (2015). An ensemble sentiment classification system of twitter data for airline services analysis. . *IEEE international conference on data mining workshop (ICDMW)* (pp. 1318-1325). IEEE.

Yang, G. R. (2020). Artificial neural networks for neuroscientists: a primer. *Neuron*, 1048-1070.

6. Appendices

1. **Extract, Transform and Load (ETL) the dataset**

```
# Import the libraries for skywards data
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df= pd.read_csv('skywards.csv',encoding='latin-1')

df.head()  #print all the records along with columns



df.info() #display all the information (data types) of columns
df.describe() # more detailed statistics information of numeric fields can be find using 'describe()' method

# check for empty columns
empty_cols = df.columns[df.isnull().all()]
print("Empty Columns: ", empty_cols)   # Display empty columns

#Drop empty columns
df.drop(empty_cols, axis=1, inplace=True)

# Replace np.nan with -200
df.replace(np.nan, -200, inplace=True)

df.head() # print the updated DataFrame
# find missing values in each column
missing_val = df.isnull()

# display the count of missing value in each column, the output shows that our data does not have any missing value
print('Count of missing values in each column')
print(missing_val.sum())
obj = (df.dtypes == 'object') #display the variables of categorical type
object_cols = list(obj[obj].index)
print("Categorical variables:",len(object_cols))

int_ = (df.dtypes == 'int')  #display the variables of integer type
num_cols = list(int_[int_].index)
print("Integer variables:",len(num_cols))
```

```
fl = (df.dtypes == 'float')  #display the variables of float type
fl_cols = list(fl[fl].index)
print("Float variables:",len(fl_cols))
```

2. Exploratory Data Analysis (EDA)

```
# Visualization of results in distribution of people
plt.figure(figsize=(7, 7))
sns.histplot(df, bins=30, kde=True, color='skyblue', edgecolor='black')
plt.title('Distribution of People of with satisfaction level')
plt.xlabel('Satisfied')
plt.ylabel('Frequency')
plt.show()
plt.figure(figsize=(7, 6))
sns.boxplot(x=df['Satisfied'])
plt.title('Box Plot of Satisfaction level of passengers')
plt.show()

labels = np.array(df['Satisfied'])#Target  feature
features= df.drop('Satisfied', axis = 1)
feature_list = list(features.columns)#it store all features except  Reg_capacity price

from sklearn import preprocessing
from sklearn.model_selection import train_test_split
features_org=features;

##############################################################################################
#####################

print('------ Done..')
print('------ Splitting the Data ------- Wait..')
X_train, X_test, y_train, y_test = train_test_split(features, labels, test_size=0.25, random_state=7)
print('Training Features Shape:', X_train.shape)
print('Training Labels Shape:', y_train.shape)
print('Testing Features Shape:', X_test.shape)
print('Testing Labels Shape:', y_test.shape)
print('------ Done..')
print('##############################')
```

3. Analytical model (Linear Regression) implementation and evaluation of model

```
# Importing libraries for linear regression
import numpy as np  # Numpy for numerical operations
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split  # For splitting the dataset into training and test sets
from sklearn.linear_model import LinearRegression  # Linear Regression model for prediction
from sklearn.metrics import mean_squared_error, r2_score  #
```

```
# Synthetic data is generated by adding some noise to the dataset
np.random.seed(45)
X = 2 * np.random.rand(100, 1)
y = 3 * X + 4 + np.random.randn(100, 1)

# Here the dataset is splitted into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.4, random_state=42)

# Linear regression model is initialized
model = LinearRegression()

# model is trained on training data
model.fit(X_train, y_train)

#predictions are done after traning
y_pred = model.predict(X_test)

# Mean Squared Error (MSE) and the coefficient of determination (R^2) is calculated
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

# Print all the MSE and R^2 values of the dataset
print(f"Mean Squared Error: {mse:.2f}")
print(f"R^2 Score: {r2:.2f}")

# model parameters are printed (slope and intercept)
print(f"Slope (Coefficient): {model.coef_[0][0]:.2f}")
print(f"Intercept: {model.intercept_[0]:.2f}")

# Visualize the results of linear regression
plt.scatter(X_test, y_test, color='blue', label='Actual data')
plt.plot(X_test, y_pred, color='red', linewidth=2, label='Fitted line')
plt.xlabel('X')
plt.ylabel('y')
plt.title('Linear Regression')
plt.legend()
plt.show()


y_test = y_test.ravel()
y_pred = y_pred.ravel()

import pandas as pd
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
```

```
# Scatter plot to visualize the regression results of LR
plt.figure(figsize=(8, 6))
plt.scatter(y_test, y_pred, alpha=0.5)
plt.xlabel('Actual Values')
plt.ylabel('Predicted Values')
plt.title('Actual vs Predicted Values')
plt.show()
```

3. **Analytical model (Artificial neural network model) implementation and evaluation of model**

```
import numpy as np
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.datasets import make_classification
from sklearn.metrics import accuracy_score

# Generate synthetic data for classification
X, y = make_classification(n_samples=1000, n_features=20, n_classes=2, random_state=42)

# Split the dataset into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Standardize the features (important for neural networks)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Initialize the Neural Network model
model = Sequential()

# Add input layer and first hidden layer with 16 neurons and ReLU activation
model.add(Dense(16, input_dim=X_train.shape[1], activation='relu'))

# Add second hidden layer with 8 neurons and ReLU activation
model.add(Dense(8, activation='relu'))

# Add output layer with 1 neuron and sigmoid activation (for binary classification)
model.add(Dense(1, activation='sigmoid'))

# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

```
# Train the model
history = model.fit(X_train, y_train, epochs=50, batch_size=10, validation_split=0.2, verbose=1)

# Evaluate the model on the test data
y_pred_prob = model.predict(X_test)
y_pred = (y_pred_prob > 0.5).astype("int32")

# Calculate the accuracy
accuracy = accuracy_score(y_test, y_pred)

# Print the accuracy
print(f"Accuracy: {accuracy:.2f}")

# (Optional) Plot training & validation accuracy values
import matplotlib.pyplot as plt

plt.plot(history.history['accuracy'])
plt.plot(history.history['val_accuracy'])
plt.title('Model accuracy')
plt.ylabel('Accuracy')
plt.xlabel('Epoch')
plt.legend(['Train', 'Validation'], loc='upper left')
plt.show()




from sklearn.metrics import accuracy_score, confusion_matrix  # For evaluating the model
import seaborn as sns  # For creating heatmap
import matplotlib.pyplot as plt  # For plotting

cm = confusion_matrix(y_test, y_pred)

# Visualize the confusion matrix using a heatmap
plt.figure(figsize=(8, 6))
sns.heatmap(cm, annot=True, fmt="d", cmap="Blues", xticklabels=["Unsatisfied 0", "Satisfied 1"],
yticklabels=["Unsatisfied 0", "Satisfied 1"])
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```

Link: https://colab.research.google.com/drive/1tbu5sLRI3Ym1jg4ZadlevEUrA-c_mcg9?usp=sharing