BPP UNIVERSITY BUSINESS SCHOOL

BPP Coursework Cover Sheet

Programme	
Module name	Applied Modelling and Visualisation
Student Reference Number (SRN)	
Report/Assignment Title	
Date of Submission *(Please attach the confirmation of any extension received)*	

Declaration of Original Work:

- I hereby declare that I have read and understood BPP's regulations on plagiarism and that this is my original work, and that I have researched, undertaken, completed and submitted in accordance with the requirements of BPP School of Business.

- I declare that where I have used any AI tools, it was for the following reasons (highlight as appropriate):
  - □ To review a draft
  - □ To correct language errors
  - □ Other (please describe)

  ____

- I have copied & retained for BPP University's reference, all AI prompts used in the creation of AI content and all AI-generated responses in support of my assignment and attached relevant evidence in the appendices.

- I understand that I may be required to participate in a *viva voce,* where I will be questioned on any aspect of my assignment, including key concepts, theories, examples used, & any sources included.

- The word count, excluding contents table, bibliography and appendices, is ____words.

Student Reference Number: **Date**:

Title: Predictive Modelling of Passenger Satisfaction Using Machine Learning: A CRISP-DM Approach for Empyrean Airlines

Table of Contents

Table Of Figures

1. Introduction

1.1 Background and Scope

In a very result-driven industry like aviation, customer satisfaction is quickly becoming a benchmark of success for the airlines. Understandably, Empyrean Airlines has appreciated the strategic value of understanding what passengers experience and how this relates to loyalty, rebooking, and net promoter score (NPS). What it endeavors to do is to go past the anecdotal feedback and subjective surveys by using data-driven methods. Rather, it aims at exposing practical patterns using extensive passenger data analytics. The current task entails the analysis of a structured data set containing more than 100,000 records describing both quantitative and qualitative passenger attributes. The emphasis is on assessing satisfaction ratings with respect to in-flight experiences, demographics, travel logistics, and delay data (Shimaoka, Ferreira and Goldman, 2024). This report illustrates how, based on the CRISP-DM methodology, a machine learning-based approach can lead to the generation of predictive insight and service strategy as a whole, which can integrate with business performance goals.

1.2 Objectives of the Study

The research seeks to observe and determine the critical factors that affect the passengers' satisfaction in the context of Empyrean Airlines. More specifically, the research goal is to develop predictive models that are capable of classifying whether a passenger is likely to be satisfied or dissatisfied out of a set of possible passenger input features. Key questions addressed include: What services most influence satisfaction? Are some demographic groups especially susceptible to dissatisfaction? What is the effect of travel class on the sentiment? Addressing such questions is the goal of this study, which targets building and testing machine learning models that are not only accurate in their predictions but also explainable (Mohd Rozaini et al., 2024). In that process it promotes data-driven decision-making in areas like service design, route planning, and pricing strategy. Using historical data and predictive analytics, Empyrean Airlines can predict customer sentiment and take action in advance, helping them maintain a competitive advantage.

1.3 Methodological Approach

This study has a CRISP-DM framework underpinning it, a standard approach to approaching data mining that drives systematic progression from business understanding to model deployment. The process first involved the development of relevant business objectives—knowing customer satisfaction drivers and translating them into a predictive classification problem. Data understanding referred to how the variables were distributed and related, while data preparation involved data cleaning, encoding, and scaling. Two techniques of machine learning were actualized at the modelling phase: random forest and decision tree. These were chosen for their ability to handle nonlinear patterns, categorical variables, and feature importance checking. The models were trained and tested on train-test splits and scored on their accuracy, the confusion matrices applied to, and the ROC-AUC scores.

2. Key Factors Influencing Passenger Satisfaction

2.1 Overview of Influential Variables

To understand what key factors affect passenger satisfaction requires a close examination of various data dimensions. Some of the most prominent features in the dataset included not strictly related to service but rather aesthetic factors such as in-flight entertainment, comfort of the seat, quality of the food and drink served, and cleanliness (Mizuno, Ohba and Ito, 2022). These features are subjective measures of in-flight experience and are scored. Air travel logistics are among the other important aspects, i.e., flight distance, arrival and departure delays, and travel class (Palimkar, Shaw and Ghosh, 2021). Moreover, demographic variables—age, gender, and continent of origin—offer a sociocultural perspective on the analysis. The importance of these features was confirmed in the correlation analysis, as well as the feature importance rankings achieved by the Random Forest model. In fact, in-flight entertainment had a 0.62 positive correlation with satisfaction, which made it an area of priority for directed improvement.

2.2 Qualitative and Quantitative Aspects

The dataset had both qualitative and quantitative variables, which played different roles in satisfaction modeling. Quantitative notes such as flight distance, departure delay, and arrival delay testify to objective parameters. These may be harnessed to identify inefficiencies in the flight operation that may be directly perceived in the travel experience. While qualitative, these were converted numerically for machine learning compatibility (Liu, Gu and Wang, 2021). The combination of such types of features highlights the multifactorial character of satisfaction. For instance, the two members of the same flight can have diner satisfaction at different levels depending on the food perceived or the friendliness of the staff.

2.3 Impact on Business Decisions

Determining satisfaction drivers helps Empyrean Airlines provide targeted business decisions that would support a good customer experience and efficient operations. For instance, if in-flight entertainment is identified as a key indicator of the success of the satisfaction strategy, the firm might spend money on improving digital media or maximizing screen usability. Similarly, delay data may be linked with the airport ground operation or flight crew efficiency, illustrating areas for a process improvement (Disha and Waheed, 2022). The findings that come from the application of the decision tree model can also be explainable to non-technical decision-makers in visual forms such as flowcharts.

3. CRISP-DM-Based Modelling Solution

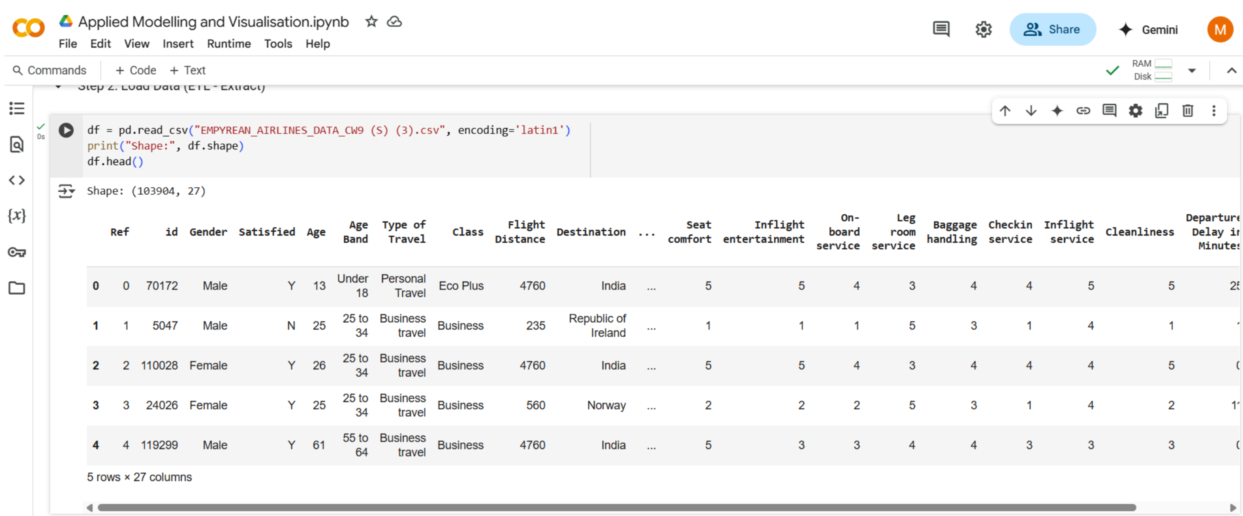
3.1 Business Understanding

At the core of Empyrean Airlines' strategy is the ideology that a data-driven perception of customer sentiment can be beneficial in terms of competitive advantage to the airline.

Leveraging predictive analytics is the strategy chosen by the company to close the gaps that maximize customer satisfaction. Through the modeling of passenger satisfaction as a classification problem, the business will therefore be able to use real-time inputs to predict dissatisfaction and intervene sooner. Business understanding also needs to translate qualitative challenges to quantifiable metrics (Sun et al., 2024). To give a sample, in-flight service quality, which is frequently considered anecdotally, can be metricated using feedback scores. By integrating satisfaction predictions into digital dashboards, Empyrean Airlines attempts to chart operational improvements and streamline staffing and services.

3.2 Data Understanding

The dataset for this case study consisted of 103,904 records and various attributes, ranging from demographic attributes and service ratings to information on logistics (Faiza and Khalil, 2023). A substantial number of variables concerned with satisfaction were analyzed that used a 0 to 5 scale and correlated well with the variable to be predicted. Other variables such as the distance of the flight, nature of travel, and what class one travels also helped contextualize satisfaction in the overall travel experience (Yang, 2023).



Applied Modelling and Visualisation.ipynb
File Edit View Insert Runtime Tools Help
Commands + Code + Text

```
df = pd.read_csv("EMPYREAN_AIRLINES_DATA_CW9 (S) (3).csv", encoding='latin1')
print("Shape:", df.shape)
df.head()
```

```
Shape: (103904, 27)
```

	Ref	id	Gender	Satisfied	Age	Age Band	Type of Travel	Class	Flight Distance	Destination	...	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes
0	0	70172	Male	Y	13	Under 18	Personal Travel	Eco Plus	4760	India	...	5	5	4	3	4	4	5	5	[illegible]
1	1	5047	Male	N	25	25 to 34	Business travel	Business	235	Republic of Ireland	...	1	1	1	5	3	1	4	1	[illegible]
2	2	110028	Female	Y	26	25 to 34	Business travel	Business	4760	India	...	5	5	4	3	4	4	4	5	[illegible]
3	3	24026	Female	Y	25	25 to 34	Business travel	Business	560	Norway	...	2	2	2	5	3	1	4	2	[illegible]
4	4	119299	Male	Y	61	55 to 64	Business travel	Business	4760	India	...	5	3	3	4	4	3	3	3	[illegible]

5 rows × 27 columns

Figure 1 Data Understanding Code section with output

During the early inspection, it was established that though the dataset was reasonably clean, it had 114 duplicated entries, and there were several inconsistencies in string formatting across columns. Gender, class, and destination were categorical data that required encoding.

3.3 Data Preparation (ETL)

The ETL process was a series of structured tasks that would guarantee clean, consistent, and modelable data. During the extract phase, the dataset was loaded in using Python's Pandas library with special parameters for encoding to manage special characters.

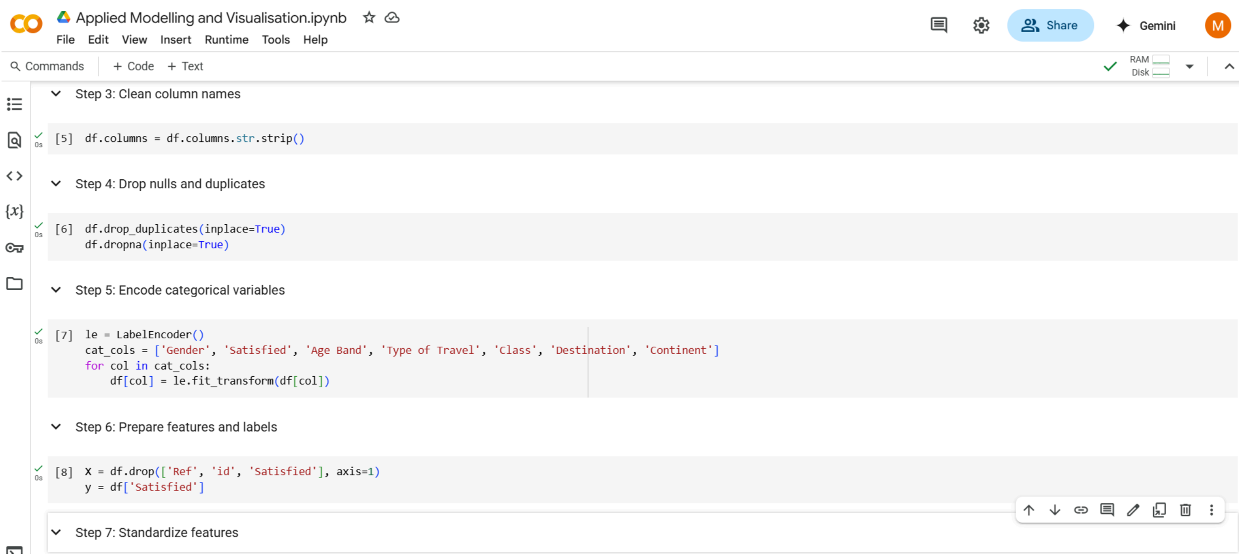


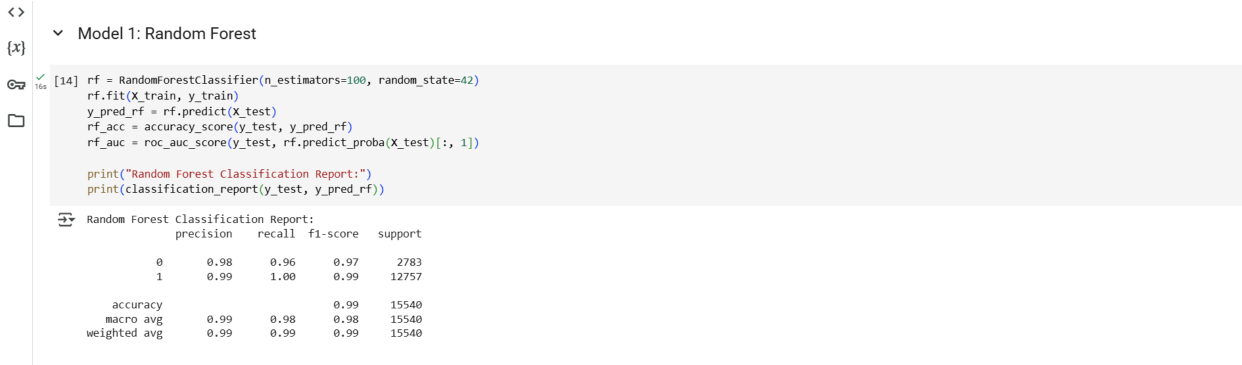
Figure 2 Extract, Transform, and Load Process

Before the transformation, there were 114 duplicate rows removed and the categorical variables encoded by using LabelEncoder. Particular emphasis was laid down on how the variables of gender, class, destination, and continent were to be coded numerically (Indra et al., 2022). Data that were missing were not found, making imputation unnecessary. Numerical variables were standardized using a StandardScaler such that models would not be biased due to changes in scale. Next, the last dataset was split into training, validation, and test sets (70:15:15) so that the original class distribution was retained. The data was rendered reliable and ready to be used to train and test a robust model during the preparation cycle.

Figure 3 ETL : Standarize Features & Split Dataset

3.4 Modelling Techniques Used

Two models were chosen for this classification task. Random Forest and Decision Tree. The three-class random forest classifier was selected because of its capacity to lessen overfitting through ensemble learning. The Random Forest model combines many weak learners to output a powerful prediction and hence enhances generalization over unseen data. The model is resistant to the noise and multicollinearity (Hong et al., 2023). Moreover, it has feature importance rankings to make it useful for finding the degree of influence of each factor on satisfaction. The decision tree classifier was presented with an advantage of transparency and simplicity.

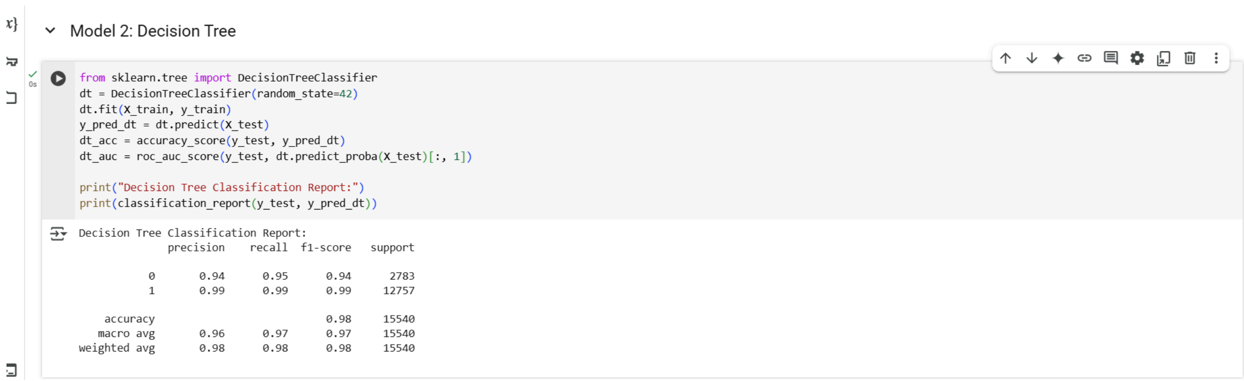


Model 1: Random Forest

```
[14] rf = RandomForestClassifier(n_estimators=100, random_state=42)
     rf.fit(X_train, y_train)
     y_pred_rf = rf.predict(X_test)
     rf_acc = accuracy_score(y_test, y_pred_rf)
     rf_auc = roc_auc_score(y_test, rf.predict_proba(X_test)[:, 1])

     print("Random Forest Classification Report:")
     print(classification_report(y_test, y_pred_rf))
```

```
Random Forest Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.96      0.97      2783
           1       0.99      1.00      0.99     12757

    accuracy                           0.99     15540
   macro avg       0.99      0.98      0.98     15540
weighted avg       0.99      0.99      0.99     15540
```

Figure 4 Random Forest Classification Report: Code+Output



Model 2: Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(random_state=42)
dt.fit(X_train, y_train)
y_pred_dt = dt.predict(X_test)
dt_acc = accuracy_score(y_test, y_pred_dt)
dt_auc = roc_auc_score(y_test, dt.predict_proba(X_test)[:, 1])

print("Decision Tree Classification Report:")
print(classification_report(y_test, y_pred_dt))
```

```
Decision Tree Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.95      0.94      2783
           1       0.99      0.99      0.99     12757

    accuracy                           0.98     15540
   macro avg       0.96      0.97      0.97     15540
weighted avg       0.98      0.98      0.98     15540
```

Figure 5 Decision Tree Classification Report: Code + report

3.5 Evaluation Methods

In order to assess model performance, this study utilized two key metrics. Accuracy and ROC-AUC (Receiver Operating Characteristic—Area Under Curve). Accuracy is the ratio of the correctly guessed observations, and that gives an intuitive feel of how the model is working. However, in the datasets where the class may be imbalanced, accuracy alone is inadequate (Gao and Mavris, 2022). Using the test set, the evaluation worked to mimic actual real-world performance and not overfit the models to the training data. The results showed that Random Forest had more accuracy (99.00%) and ROU-AUC (0.9980) than Decision Tree (97.96% and 0.9779 ROU-AUC), which favors the Random Forest model clearly on both precision and generalizability.

3.6 Deployment and Model Recommendation

Based on superior performance, equipped with the Random Forest model deployment within the business systems of Empyrean Airlines is recommended. It is available with high accuracy and the ability to adjust to new data over time. The model is also capable of being reduced to a customer relationship management dashboard; it is capable of delivering real-time satisfaction predictions according to passenger profiles and the characteristics of flights. In such a role, frontline staff would have the ability to be made aware of possible dissatisfaction risks and provide individual support in advance (Salah-Ud-Din, SS and Al Ali, 2024). Further, the model's feature importance rankings can help in guiding the service development priorities. There should be regular schedules of retraining and performance reviews to uphold the relevancy of a model with data trends changing with time. Decision trees are not the best model but could be held for training and early-phase analysis for their ease of use and interpretation.

4. Model Evaluation and Comparison

```
[20] comparison = pd.DataFrame({
         "Model": ["Random Forest", "Decision Tree"],
         "Accuracy": [rf_acc, dt_acc],
         "AUC": [rf_auc, dt_auc]
     })
     print("Model Comparison:")
     print(comparison)
```

Figure 6 Model Evaluation and Comparison: Code

```
Model Comparison:
           Model  Accuracy       AUC
0  Random Forest  0.990090  0.998019
1  Decision Tree  0.979601  0.967910
```

Figure 7 Model Evaluation and Comparison: Output

4.1 Chosen Loss Function and Metrics

In the case of binary classification tasks, as in this study, loss functions and evaluation metrics reflect the effectiveness of the model. Even though the Scikit-learn defaults were used to train models employed in this study, there is an underlying mechanism of optimization: the minimization of a kind of log loss during the process of probabilistic estimation. Log loss punishes high-confidence wrong guesses more than high-confidence correct guesses (Tian et al., 2021). This loss function is actually implicit in such algorithms as random forests when estimating probabilities. In addition, we used two key metrics for assessment: accuracy and ROC-AUC. The measures of accuracy refer to the percentage of correct predictions and deliver an intuitive understanding of model correctness.

4.2 Comparative Analysis of Models

Having trained both models on the preprocessed dataset, the success of the model was checked on a set of test data that the models had not been trained on to validate the appropriateness of the model for the unseen test set. The Random Forest classifier gave a great test accuracy of 99.00% and a ROC – AUC score of 0.9980, implying high, not just average, performance. Contrary to the same, the accuracy and the ROC-AUC for the decision tree classifier were 97.96% and 0.9679, respectively (Carvalho et al., 2021). These scores remain respectable but play less well to the idea of generalization than the ensemble approach. The difference in the level of this accuracy further accentuates the advantage of using ensemble techniques to form an ensemble of prediction estimates from various learners.

4.3 Strengths, Weaknesses, and Justification

Random Forest is renowned for its robustness and generalizability in problem spaces with different datasets. It reduces overfitting by collaborative prediction of several decision trees with the help of bootstrap aggregation. However, its complexity may render real-time prediction time-consuming and unintelligible (Bunchongchit and Wattanacharoensil, 2021). On the contrary, the situation with the decision tree model is marked by its simplicity and transparency. Each decision node is visually easy to grasp, making it ideal for communicating to non-technical stakeholders.

5. Communicating Insights Through Visualization

5.1 EDA Output Interpretation

Visualization was crucial to understanding the relationships between variables and informed model feature selection. A heatmap was created to visualize correlations with the satisfaction and other features. Strong positive associations with inflight entertainment (0.62), seat comfort (0.55), cleanliness (0.52), and food and drink (0.50) were found. These metrics validated current assumptions about what causes the satisfaction (Gupta et al., 2021). Class-specific count plots indicated that passengers in business class had a higher rate of satisfaction than those in economy class.

Exploratory Data Analysis (EDA)

1s

```
plt.figure(figsize=(12, 8))
sns.heatmap(df.corr(), cmap='coolwarm')
plt.title("Correlation Heatmap")
plt.show()
```

Figure 8 Heatmap EDA Performance: Code section

Heatmap on distance of flight stratified by satisfaction status revealed polarization in satisfaction for longer journeys, implying that the service experience is more important during long flights.

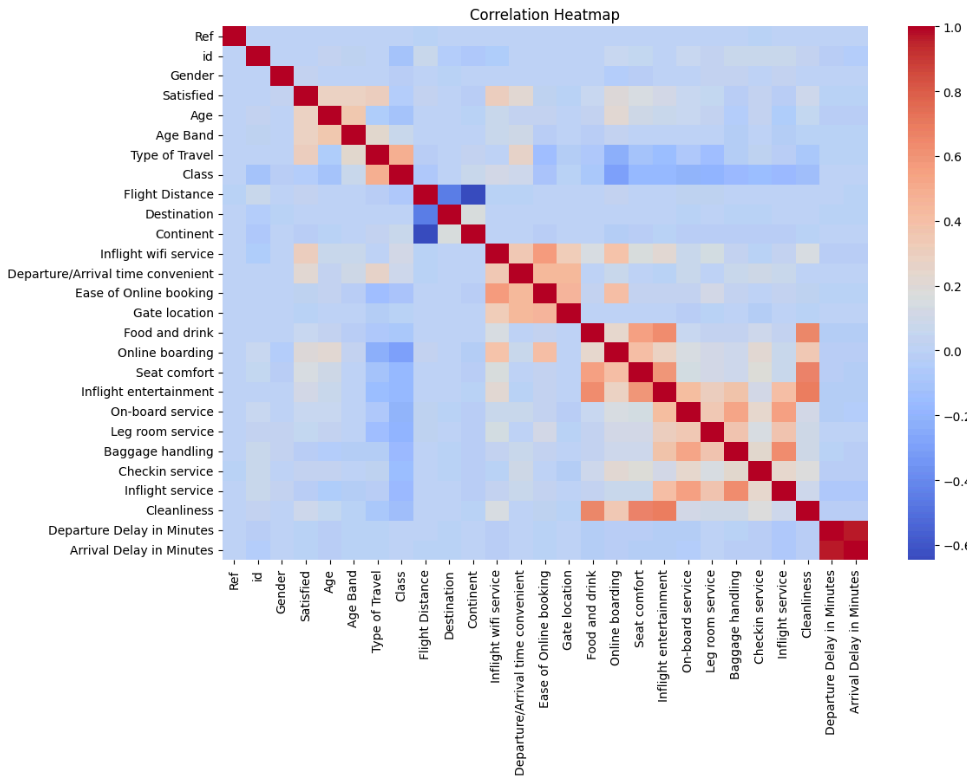


Figure 9 Correlation -Heatmap

5.2 Model Performance Visualization

Confusion matrices were implemented to evaluate classification outcomes of both models. For Random Forest, the matrix showed high true positive and true negative rates with little misclassification. Random Forest indicated a curve near the top left-hand corner of the plot corresponding to high sensitivity and specificity (Guo, Grushka-Cockayne and De Reyck, 2022). The area under this curve was proving excellent separability of the model. These graphical outputs endorsed the numerical assessment and offered intuitive visualizations of the model's performance, which are necessary when presenting findings to decision makers who are unlikely to follow technical measures.

```python
sns.heatmap(confusion_matrix(y_test, y_pred_rf), annot=True, fmt='d', cmap='Blues')
plt.title("Random Forest – Confusion Matrix")
plt.show()
```

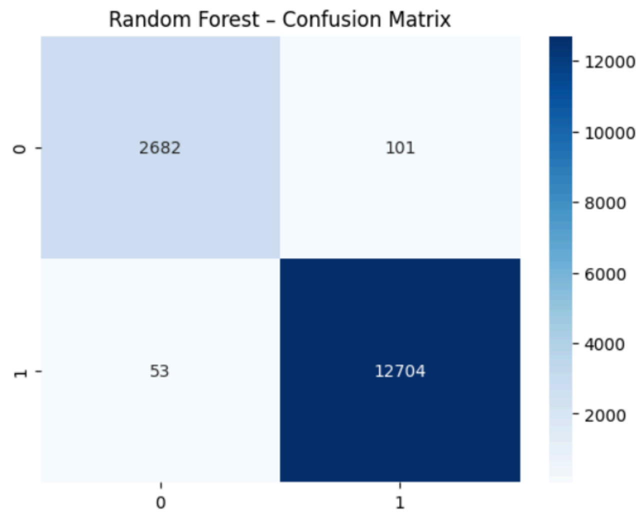


Figure 10 Random Forest – Confusion Matrix : Code+ Output

```python
[18] sns.heatmap(confusion_matrix(y_test, y_pred_dt), annot=True, fmt='d', cmap='Greens')
plt.title("Decision Tree – Confusion Matrix")
plt.show()
```

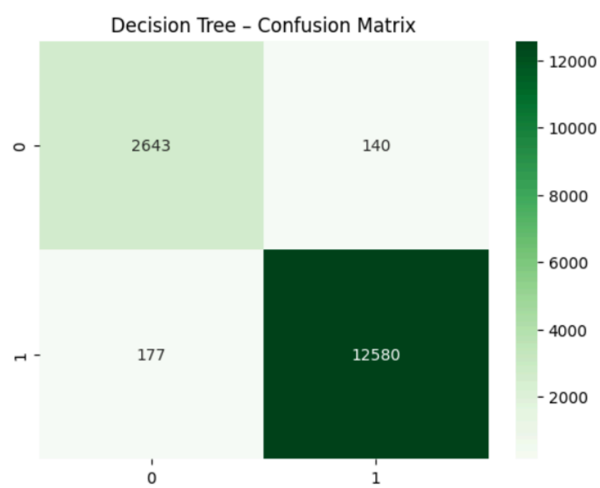


Figure 11 Decision Tree – Confusion Matrix

5.3 EDA Role in Model Selection

The exploratory data analysis conducted initially during the project took a crucial step towards informing model selection and parameter tuning. The recognition of such highly correlated variables as inflight entertainment and cleanliness facilitated a concentration of modeling efforts on the features with outstanding predictive power (Helgo, 2023). Furthermore, plots that indicated skewed feature distributions and outliers influenced preprocessing, such as scaling and encoding (Krishna et al., 2023). With the understanding of these patterns, it was possible to choose the algorithms that are able to manage non-linear relationships and ragged data types such as random forests; assumptions, EDA became a bedrock for data-driven model design and evaluation.

```
[21] print("\nEDA Justification:")
     print("""
     EDA helped identify high-correlation features like inflight entertainment, seat comfort, cleanliness, etc.,
     and guided our model selection. Random Forest was chosen for robustness, and Decision Tree for interpretability.
     Both models used accuracy and AUC for evaluation. Feature scaling was applied to improve generalization.
     """)

     best_model = comparison.loc[comparison["Accuracy"].idxmax()]["Model"]
     print(f"\n Recommended Model: {best_model}")
```

Figure 12 Recommended Model: Code

```
EDA Justification:

EDA helped identify high-correlation features like inflight entertainment, seat comfort, cleanliness, etc.,
and guided our model selection. Random Forest was chosen for robustness, and Decision Tree for interpretability.
Both models used accuracy and AUC for evaluation. Feature scaling was applied to improve generalization.


 Recommended Model: Random Forest
```

Figure 13 Recommended Model : Output

6. Conclusion and Recommendations

6.1 Key Findings

The main purpose of this study was to model and to forecast passenger satisfaction based on the data collected by Empyrean Airlines. After applying the CRISP-DM methodology, Python-based tools were used to cleanse, transform, and analyze the data. Two models, Random Forest and Decision Tree, were tested. Experiments indicated that in both accuracy and AUC, Random Forest achieved superior results to Decision Tree. Specifically, Random Forest reached 99% accuracy and a 0.9980 ROC-AUC score. The exploratory data analysis indicated that such factors as in-flight entertainment, comfort of seats, and cleanliness are the best determinants of satisfaction. The visualizations facilitated the EDA and the model evaluation process.

6.2 Business Recommendations and Next Steps

From analysis, we recommend that Empyrean Airlines should reallocate investments into improving areas of satisfaction to determine the needs of customers most relevant to satisfaction, such as in-flight entertainment, seat comfort, and cleanliness. These could be achieved through digital entertainment platform upgrades, better ergonomics of seats to be provided, and better cleaning protocols. The prediction and decision support should be implemented in the Random Forest model within internal analytics platforms in real-time. It could help customer service units to determine the passengers with high potential to be dissatisfied so they could be catered to on time.

7. References

- Bunchongchit, K. and Wattanacharoensil, W., 2021. Data analytics of Skytrax's airport review and ratings: Views of airport quality by passengers types. *Research in Transportation Business & Management*, *41*, p.100688. https://doi.org/10.1016/j.rtbm.2021.100688

- Carvalho, L., Sternberg, A., Maia Goncalves, L., Beatriz Cruz, A., Soares, J.A., Brandão, D., Carvalho, D. and Ogasawara, E., 2021. On the relevance of data science for flight delay research: a systematic review. *Transport Reviews*, *41*(4), pp.499-528. https://doi.org/10.1080/01441647.2020.1861123
- Disha, R.A. and Waheed, S., 2022. Performance analysis of machine learning models for intrusion detection system using Gini Impurity-based Weighted Random Forest (GIWRF) feature selection technique. *Cybersecurity*, *5*(1), p.1. https://doi.org/10.1186/s42400-021-00103-8
- Faiza and Khalil, K., 2023. Airline flight delays using artificial intelligence in COVID-19 with perspective analytics. *Journal of Intelligent & Fuzzy Systems*, *44*(4), pp.6631-6653. https://doi.org/10.3233/JIFS-222827
- Gao, Z. and Mavris, D.N., 2022. Statistics and machine learning in aviation environmental impact analysis: A survey of recent progress. *Aerospace*, *9*(12), p.750. https://doi.org/10.3390/aerospace9120750
- Guo, X., Grushka-Cockayne, Y. and De Reyck, B., 2022. Forecasting airport transfer passenger flow using real-time data and machine learning. *Manufacturing & Service Operations Management*, *24*(6), pp.3193-3214. https://doi.org/10.1287/msom.2021.0975
- Gupta, M., Kumar, R., Walia, H. and Kaur, G., 2021, October. Airlines based twitter sentiment analysis using deep learning. In *2021 5th International Conference on Information Systems and Computer Networks (ISCON)* (pp. 1-6). IEEE. 10.1109/ISCON52037.2021.9702502
- Helgo, M., 2023. Deep learning and machine learning algorithms for enhanced aircraft maintenance and flight data analysis. *Journal of Robotics Spectrum*, *1*, pp.090-099. Doi: https://doi.org/10.53759/9852/JRS202301009

- Hong, A.C.Y., KHAW, K.W., Chew, X. and Yeong, W.C., 2023. Prediction of US airline passenger satisfaction using machine learning algorithms. *Data Analytics and Applied Mathematics (DAAM)*, pp.7-22. https://doi.org/10.15282/daam.v4i1.9071
- Indra, E., Suwanto, J., Sitompul, D.R.H., Sinurat, S.H., Situmorang, A., Ruben, R. and Ziegel, D.J., 2022. Comparison of classification algorithm in classifying airline passenger satisfaction. *Jurnal Sistem Informasi dan Ilmu Komputer*, *6*(1), pp.92-98. https://doi.org/10.34012/jurnalsisteminformasidanilmukomputer.v6i1.2848
- Krishna, R., Sariki, T.P., Sigamsetty, S.M. and Kumar, G.B., 2023, December. Aviation safe landing prediction utilizing long-short term memory algorithm. In *2023 6th International Conference on Recent Trends in Advance Computing (ICRTAC)* (pp. 562-568). IEEE. 10.1109/ICRTAC59277.2023.10480721
- Liu, C., Gu, Z. and Wang, J., 2021. A hybrid intrusion detection system based on scalable k-means+ random forest and deep learning. *Ieee Access*, *9*, pp.75729-75740. Doi: 10.1109/ACCESS.2021.3082147
- Mizuno, S., Ohba, H. and Ito, K., 2022. Machine learning-based turbulence-risk prediction method for the safe operation of aircrafts. *Journal of Big Data*, *9*(1), p.29. https://doi.org/10.1186/s40537-022-00584-5
- Mohd Rozaini, N.A.A., Abd Talib, N.F.A., Ul–Saufie, A.Z., Ibrahim, N. and Jamil, S.A.M., 2024, August. Analysis of airline customer satisfaction using data mining techniques. In *AIP Conference Proceedings* (Vol. 3123, No. 1). AIP Publishing. https://doi.org/10.1063/5.0223954
- Palimkar, P., Shaw, R.N. and Ghosh, A., 2021. Machine learning technique to prognosis diabetes disease: Random forest classifier approach. In *Advanced computing and intelligent technologies: proceedings of ICACIT 2021* (pp. 219-244). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-2164-2_19
- Salah-Ud-Din, M., SS, B.T.L. and Al Ali, H., 2024, April. Exploratory Data Analysis and Prediction of Passenger Satisfaction with Airline services. In *2024 New Trends in Civil Aviation (NTCA)* (pp. 295-302). IEEE. Doi: 10.23919/NTCA60572.2024.10517814
- Shimaoka, A.M., Ferreira, R.C. and Goldman, A., 2024. The evolution of CRISP-DM for data science: Methods, processes and frameworks. *SBC Reviews on Computer Science*, *4*(1), pp.28-43. **DOI:** https://doi.org/10.5753/reviews.2024.3757

- Sun, Z., Wang, G., Li, P., Wang, H., Zhang, M. and Liang, X., 2024. An improved random forest based on the classification accuracy and correlation measurement of decision trees. *Expert Systems with Applications*, *237*, p.121549. https://doi.org/10.1016/j.eswa.2023.121549
- Tian, H., Presa-Reyes, M., Tao, Y., Wang, T., Pouyanfar, S., Miguel, A., Luis, S., Shyu, M.L., Chen, S.C. and Iyengar, S.S., 2021. Data analytics for air travel data: a survey and new perspectives. *ACM Computing Surveys (CSUR)*, *54*(8), pp.1-35. https://doi.org/10.1145/3469028
- Yang, S., 2023, September. Machine Learning-Based Airline Passenger Satisfaction Evaluation. In *2023 International Conference on Industrial IoT, Big Data and Supply Chain (IIoTBDSC)* (pp. 11-15). IEEE. DOI: 10.1109/IIoTBDSC60298.2023.00012

8. Appendices

Google Colab Link for Code:

https://colab.research.google.com/drive/1Jg5b0O2BY8S0NpIcuE7v1vo88ufyXuRc?usp=sharing

```
sns.countplot(data=df, x='Class', hue='Satisfied')
plt.title("Satisfaction by Class")
plt.xticks(rotation=45)
plt.show()
```

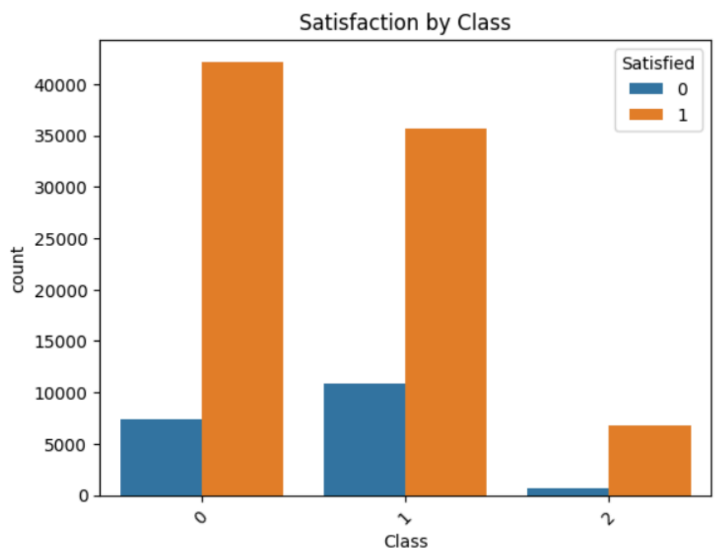


Figure 14 Satisfaction by Class

```
sns.histplot(data=df, x='Flight Distance', hue='Satisfied', bins=30, kde=True)
plt.title("Satisfaction vs Flight Distance")
plt.show()
```

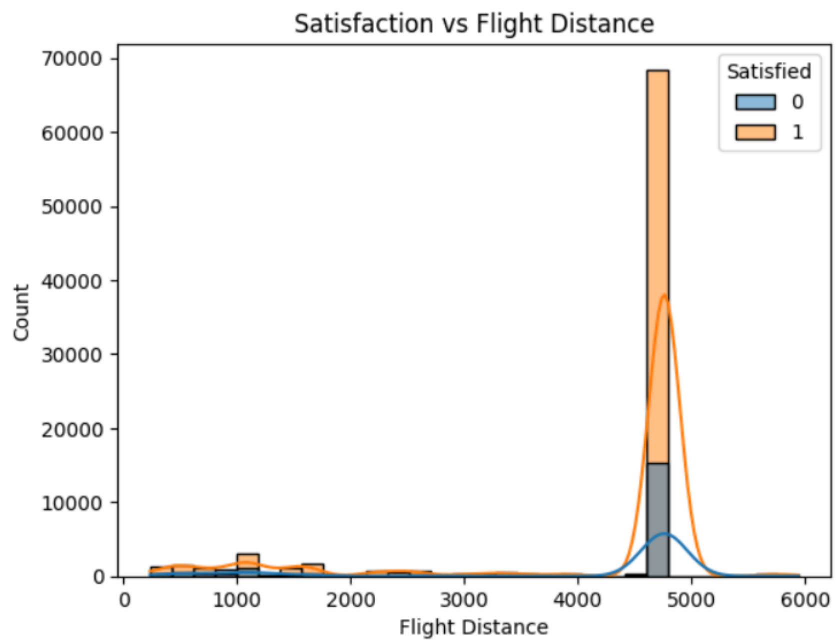


Figure 15 Satisfaction Vs Flight Distance: Code+ Output