**Student Declaration of Authorship**

|  |  |
| --- | --- |
| **Course code and name:** | F21DL – Data Mining and Machine Learning |
| **Type of assessment:** | Group Assessment |
| **Coursework Title:** | Airline Survey Analysis |
| **Group Number:** | Dubai\_PG 2 Dubai\_PG |
| **Student Names and IDs:** | AJAY KUMAR PAIDIMARRI (H00467552)  ANUGRAHA KANJOOKARAN ANTOO (H00460688)  FIYIFOLUWA SHEDRACH ONAFESO (H00466745)  HARIHARAKUMAR RATHINAR(H00463082)  SYEDA DANISH MARIAM (H00473375) |

# Declaration of authorship. By signing this form:

* **I declare** that the work I have submitted for individual assessment OR the work I have contributed to a group assessment, is entirely my own. I have NOT taken the ideas, writings or inventions of another person and used these as if they were my own. My submission or my contribution to a group submission is expressed in my own words. Any uses made within this work of the ideas, writings or inventions of others, or of any existing sources of information (books, journals, websites, etc.) are properly acknowledged and listed in the references and/or acknowledgements section.
* I confirm that I have read, understood and followed the University’s Regulations on plagiarism as published on the [University’s website](https://www.hw.ac.uk/uk/students/studies/examinations/plagiarism.htm), and that I am aware of the penalties that I will face should I not adhere to the University Regulations.
* I confirm that I have read, understood and avoided the different types of plagiarism explained in the University guidance on [Academic Integrity and Plagiarism](https://heriotwatt.sharepoint.com/sites/skillshub/SitePages/Academic-Integrity-and-Plagiarism.aspx)

**Student Signature** *(type your names): AJAY KUMAR PAIDIMARRI ANUGRAHA KANJOOKARAN ANTOO  
SYEDA DANISH MARIAM  
HARIHARAKUMAR RATHINAR  
FIYIFOLUWA ONAFESO SHEDRACH*

**Date**: *11/22/2024*

**Data Mining and Machine Learning** (Course Code: **F21DL**)

**Airline Survey Analysis**

**(Course Work – I)**

|  |  |
| --- | --- |
| **Presented By** | |
| Author and Student ID of Author with email IDs | **Ajay Kumar Paidimarri (H00467552)**  [**ap2054@hw.ac.uk**](mailto:ap2054@hw.ac.uk)  **Anugraha Antoo Kanjookaran (H00460688)**  [**ak3034@hw.ac.uk**](mailto:ak3034@hw.ac.uk)  **Fiyifoluwa Shedrach Onafeso (H00466745)**  [**Fso3000@hw.ac.uk**](mailto:Fso3000@hw.ac.uk)  **Hariharakumar Rathinar (H00463082)**  [**hr30000@hw.ac.uk**](mailto:hr30000@hw.ac.uk)  **Syeda Mariam Danish (H00473375)**  [**smd4000@hw.ac.uk**](mailto:smd4000@hw.ac.uk) |
| Version | **v1.0** |
| Working Capacity | **Group** |
| Group Number allocated | **Dubai\_PG 2 Dubai\_PG** |
| Date of submission | **22nd of November 2024** |

**TABLE OF CONTENTS**

[Declaration of authorship. By signing this form: 1](#_Toc182182740)

[1. Overview 4](#_Toc182182741)

[2. Data Analysis and Exploration 4](#_Toc182182742)

[2.1 Data Quality and Missing Values 4](#_Toc182182743)

[2.2 Feature Types and Encoding 4](#_Toc182182744)

[2.3 Feature Engineering 5](#_Toc182182745)

[3. Clustering 6](#_Toc182182746)

[3.1 Rationale for K-Means Selection and Data Preparation 6](#_Toc182182747)

[3.2 Cluster Profiles and Key Insights 6](#_Toc182182748)

[4. Baseline Training and Evaluation Experiments 7](#_Toc182182749)

[4.1 Models Implemented 7](#_Toc182182750)

[4.2 Model Output and Evaluation 7](#_Toc182182751)

[5. Neural Networks 7](#_Toc182182752)

[6. Conclusion 7](#_Toc182182753)

[7. Appendix 7](#_Toc182182754)

# Overview

Airline Survey analysis project is taken for machine learning to answer two question as mentioned below.

**Question 1:** To predict Customer Loyaltyfromprovided data [Appendix](#_Appendix) Dataset 1 & 2.

**Question 2:** To provideCustomer Reaction Analysis based on image-based identification where we will classify customers based on facial reactions.

# Data Analysis and Exploration

Dataset 1 & 2 are excel data. The input CSV file contains 25 columns of customer data with variables capturing demographic details (e.g., Age, Gender), travel specifics (e.g., Flight Distance, Type of Travel), and service experience ratings across multiple aspects such as Online Boarding, Seat Comfort, and Food and Drink aimed to predict customer loyalty i.e segmented into four **Loyalty** classes.

## Data Quality and Missing Values

To ensure data quality, an initial analysis was conducted to assess the completeness of each feature. While most features were complete, a few had minor data gaps. Specifically, **Arrival Delay** showed approximately 0.3% missing values, **Ease of Online Booking** had 4.37% missing values, and **Departure and Arrival Time Convenience** exhibited a 5.1% gap. Additionally, some in-flight service ratings, such as **Leg Room Service**, **In-flight Wifi Service**, and **Food and Drink**, had small data deficiencies.

A screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA screenshot of a graph

Description automatically generated

Figure 1

For continuous features like **Arrival Delay** and **Departure Delay**, missing values were filled using the median, a robust choice against outliers. In cases where certain satisfaction metrics were recorded as zero due to missing feedback, these were replaced with the median of available responses, ensuring consistency in representation and reducing bias introduced by missing feedback. Categorical features, including Gender, Customer Type, Type of Travel, and Class, were label-encoded to convert them into numerical representations that clustering algorithms can process. For example, Gender was transformed into binary values, while Type of Travel (Business or Personal) and Class (Business, Economy, Economy Plus) were similarly encoded to capture their distinct values numerically.

## Feature Engineering

After applying different approaches to select features, we concluded that relying on a single feature selection method (e.g., correlation) along with binning & creating new feature is insufficient for identifying the most impactful features. Instead, combining multiple feature selection techniques provides a more holistic view of feature importance. The binning of age and flight distance did not significantly enhance feature relevance, indicating the need for a more nuanced approach to numerical variables.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Feature / Rank** | **TC** | **Correlation** | **Kmean** | **Chi2Score** | **Lasso** | **RFE** | **Overall rank** |
| Type of Travel | 1 | 7 | 1 | 3 | 2 | 2 | 3 |
| Online Boarding | 2 | 4 | 2 | 7 | 1 | 6 | 4 |
| In-flight Wifi Service | 3 | 17 | 3 | 6 | 14 | 1 | 7 |
| Ease of Online Booking | 4 | 18 | 4 | 8 | 13 | 4 | 9 |
| Age | 7 | 3 | 8 | 2 | 4 | 18 | 7 |
| In-flight Entertainment | 8 | 6 | 9 | 11 | 5 | 7 | 8 |
| Flight Distance | 9 | 2 | 7 | 1 | 3 | 21 | 7 |
| Departure and Arrival Time Convenience | 6 | 16 | 6 | 10 | 12 | 9 | 10 |
| Seat Comfort | 11 | 5 | 10 | 12 | 7 | 15 | 10 |
| Class | 10 | 15 | 5 | 9 | 22 | 3 | 11 |
| Cleanliness | 15 | 8 | 14 | 14 | 17 | 12 | 13 |
| On-board Service | 13 | 11 | 11 | 13 | 9 | 11 | 11 |
| Leg Room Service | 14 | 10 | 13 | 15 | 8 | 14 | 12 |
| In-flight Service | 12 | 19 | 12 | 18 | 16 | 8 | 14 |

The table above combines ranks from various feature selection methods, including Extra Trees Classifier, Correlation, KMeans ranking, Chi-square scores, Lasso, and Recursive Feature Elimination (RFE) from which top 13 features are selected. This approach mitigates the risks of overfitting and underfitting by maintaining a balanced and robust set of features. New critical feature like Type of travel was selected based on this approach along with decision to eliminate features like Departure Delay and Arrival Delay which show showed limited utility across ranking methods.

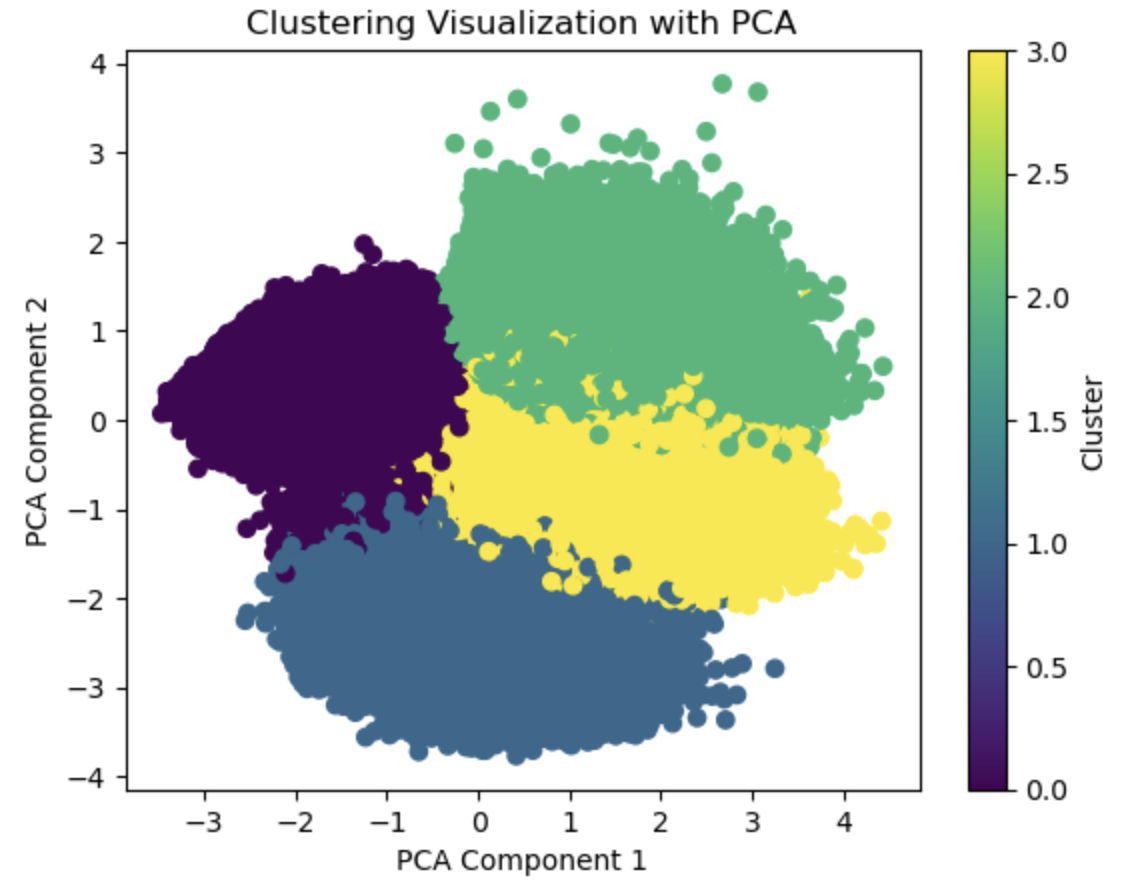
# Clustering

## Rationale for K-Means Selection and Data Preparation

For customer segmentation, K-Means clustering was selected due to its efficiency in handling high-dimensional, continuous data. Our dataset, consisting of metrics on customer service experience (e.g., Online Boarding, Seat Comfort) and demographics, was well-suited to K-Means, which allows for meaningful segmentation based solely on feature similarity without the need for labeled data.

Data standardization was essential, as K-Means relies on Euclidean distance; without standardization, features with larger scales might dominate clustering. We applied StandardScaler to normalize the data, followed by the Elbow Method to determine an optimal cluster count by plotting within-cluster sum of squares (inertia) for clusters ranging from 1 to 10. An “elbow” was observed at 4 clusters, indicating an optimal balance between compactness and simplicity. Silhouette Scores, which measure the separation between clusters, confirmed that 4 clusters provided clear distinctions between customer groups. This was further visualized using PCA for dimensionality reduction, enabling us to view the clusters on a two-dimensional plot for clearer interpretation.

A graph of a number of clusters

Description automatically generated

## Cluster Profiles and Key Insights

The clustering revealed four distinct customer segments with varying service expectations and demographic characteristics.

**Cluster 0** predominantly includes middle-aged customers who prioritize convenience, rating Online Boarding and Leg Room Service highly. This suggests a loyalty driver centered on ease and comfort during onboarding. **Cluster 1** comprises older, mostly business travelers who often take long-haul flights. They rate Food and Drink service well, indicating that in-flight amenities are important to this group’s loyalty. **Cluster 2** represents younger customers on shorter flights, with lower ratings for Online Boarding and On-board Service, suggesting potential gaps in meeting their expectations. These customers may be less loyal, representing an opportunity for targeted improvements. **Cluster 3** includes young leisure travelers who value Leg Room Service, suggesting that targeted offers on comfort and affordability may appeal strongly to this group.

This segmentation allows us to identify loyalty drivers specific to each group. Correlation analysis shows Online Boarding, Total Onboarding Convenience, and Seat Comfort as key factors associated with loyalty, which align with the needs identified within each cluster. By understanding these profiles, we can develop tailored loyalty programs and service improvements that directly address the preferences of each customer segment, enhancing retention and satisfaction across the board.

# Baseline Training and Evaluation Experiments

## Models Implemented

In this analysis, three distinct machine learning models were implemented: Decision Tree, Naive Bayes, and k-Nearest Neighbors (k-NN). Each model was chosen for its unique strengths in handling classification tasks. The Decision Tree Classifier is a powerful model that splits data based on feature criteria, making it adept at capturing non-linear patterns in the dataset. The Naive Bayes model, specifically the Gaussian Naive Bayes, leverages probabilistic methods, assuming conditional independence among features, which often performs well on smaller datasets. Lastly, the k-NN model was used as a non-parametric method that classifies samples based on their proximity to labeled data points, with k set to 5 to balance complexity and performance.

A screenshot of a graph

Description automatically generated

## 4.2 Model Output and Evaluation

Each model’s performance was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 score. The Decision Tree model yielded the highest accuracy of 80%, with corresponding F1, precision, and recall values around 0.81, suggesting it effectively captured the patterns in the data. Naive Bayes produced a lower accuracy of 71%, with an F1 score of 0.73, indicating some limitations in modeling this dataset due to its independence assumption. The k-NN model achieved an accuracy of 77% with an F1 score of 0.77, performing reasonably well though slightly lower than the Decision Tree. Confusion matrices illustrated the models’ performance in detail, with the Decision Tree showing the most balanced results, suggesting that it might be the most reliable model for this dataset.

# Neural Networks

# Conclusion

# Appendix

* GITHUB Repository Link: <https://github.com/HWHospital2024/HW_Hospital_App>
* Dataset Source Details:
  1. Dataset 1: Airline Passenger Satisfaction Dataset
     + Link: <https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction/data>
     + File name: airline\_passenger\_satisfaction.csv
  2. Dataset 2: Airline Passenger Satisfaction
     + Link: <https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction/data>
     + File Name: train.csv & test.scv
  3. Dataset 3: VGGFace2 Dataset
     + Link: <https://www.kaggle.com/datasets/jonathanoheix/face-expression-recognition-dataset>
     + Folder name: "train" & "Validate" folders