**Project Report Data Science Fall 2023**

**Airline Customer Satisfaction Prediction**

**M. Abdullah Arshad 23L-2531**

**M. Sohaib 23L-2519**

**M. Awais 23L-2618**

**Department of Computer Science**

**National University of Computer and Emerging Sciences, Lahore**

**November 21, 2024**

**Abstract**

Customer satisfaction prediction in the airline industry is a complex task due to the diverse factors influencing passenger experiences. This project aims to develop a Customer Satisfaction Prediction Model (C.S.P.M) to predict flight customer satisfaction based on historical survey and flight operation data. Utilizing machine learning classification techniques, including Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest, the model analyzes key features such as flight delay status, service quality metrics, seat comfort, in-flight services, and customer feedback scores.

Additional attributes like travel type, age, and customer loyalty status are incorporated to explore patterns and trends in customer satisfaction. The methodology emphasizes data cleaning, feature selection, and model evaluation while addressing challenges such as class imbalance and feature interdependencies. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess model performance. The results highlight significant predictors of customer satisfaction and demonstrate the model’s potential to offer valuable insights for improving service quality in the airline industry. Future work may explore additional datasets and advanced classification techniques to enhance model accuracy and generalizability.

**1 Introduction**

Understanding customer satisfaction in the airline industry is crucial for improving service quality and maintaining customer loyalty. Predicting customer satisfaction is a challenging yet essential task for airlines to optimize operations, enhance passenger experiences, and make data-driven decisions. This project focuses on developing a machine learning-based classification model to predict flight customer satisfaction using historical survey and flight operation data.

The dataset includes key features such as flight delay status, seat comfort, in-flight services, and customer feedback scores, which are essential for understanding passenger sentiment. Additional variables like travel type, customer loyalty status, and demographics are incorporated to explore patterns and trends that influence satisfaction levels. This study contributes to the field of customer analytics by providing a replicable framework for satisfaction prediction and actionable insights for improving service in a competitive industry.

**2 Methodology**

The project methodology is divided into several stages:

**2.1 Data Preprocessing**

* Load and clean the dataset to handle missing values, anomalies, and duplicates.
* Encode categorical features, such as travel type and customer loyalty status, into numerical formats.

**2.2 Feature Extraction**

* Select relevant features such as flight delay status, seat comfort, in-flight services, and customer feedback scores.
* Include additional attributes like travel type, age, and customer loyalty status to enhance prediction accuracy.
* Engineer new features, such as satisfaction categories (e.g., satisfied/neutral/dissatisfied), for better classification.

**2.3 Model Selection**

Implemented following classifiers to predict customer satisfaction.

* Logistic Regression
* K-Nearest Neighbors (KNN)
* Random Forest

**2.4 Model Evaluation**

Model performance was evaluated using metrics such as Accuracy, Precision, Recall, and F1-Score. These helped compare how well each model predicted customer satisfaction. Random Forest performed best in terms of accuracy, while Logistic Regression showed solid results for simpler patterns. K-Nearest Neighbors (KNN) captured local variations well but required careful tuning to avoid overfitting. The evaluation metrics provided clear insights into each model's strengths and guided the final model selection.

**2.5 Visualization**

* **Class Distribution**: Count plot for the satisfaction variable to analyze class balance and calculate the class ratio.
* **Box Plots**: Highlighted outliers in numerical variables like Age and Flight Distance.
* **Histograms with KDE**: Showed distributions and density curves for numerical attributes.
* **Categorical Analysis**: Count plots for categorical variables with enhanced styling.

**3 Experiments**

The experiments involve training and testing the selected classification models on the preprocessed dataset. The dataset is split into training and testing sets (80/20 ratio) to ensure a robust evaluation. Key findings include:

* **Logistic Regression** performed well in identifying linear decision boundaries but may struggle with complex relationships in the data.
* **K-Nearest Neighbors (KNN)** excelled at capturing local patterns but required careful consideration of the number of neighbors to balance bias and variance.
* **Random Forest** effectively handled feature interactions and provided robust predictions, although it required attention to class distributions to avoid bias.

Figures illustrating model performance (e.g., accuracy, precision, recall, F1-score) and confusion matrices are generated to compare results visually. These visualizations provide insights into the strengths and limitations of each model in predicting customer satisfaction.

**4 Results & Discussion**

The results highlight the efficiency of different models in predicting customer satisfaction. Logistic Regression demonstrated consistent performance for linearly separable patterns, while K-Nearest Neighbors (KNN) excelled in capturing local variations but required careful tuning of the number of neighbors. Random Forest emerged as the most robust model, effectively handling feature interactions and delivering high accuracy and F1-scores.

Key trends identified include the strong influence of flight delays, in-flight services, and customer loyalty status on satisfaction levels. The discussion also highlights the challenges faced during the project, particularly in data preprocessing and feature engineering. One significant challenge was dimensionality reduction, where attributes such as flight arrival and departure delays were combined into a single feature. Similarly, multiple in-flight service and on-board service attributes were merged into cohesive categories to streamline the dataset, ensuring relevant information was retained for model training.

Challenges such as class imbalances and overlapping feature effects were addressed, proposing potential improvements like advanced feature engineering and exploring ensemble methods to further enhance prediction accuracy.

**5 Conclusion**

This project successfully developed a machine learning-based model to predict airline customer satisfaction, utilizing key features such as flight delay status, in-flight services, and customer feedback. The results highlight the importance of effective feature engineering and model selection in understanding passenger sentiment. By leveraging machine learning techniques, we were able to uncover valuable insights that can help improve decision-making in the airline industry. The findings suggest that predictive models can be a powerful tool for enhancing customer experience management and optimizing airline operations.

**6 References**

1. Kaggle.