**Airlines Customer Satisfaction Report**

**Table of Contents**

1. **Introduction**
   * Overview of the Dataset and Objectives
2. **Part I: Descriptive Analysis**
   * Dataset Overview and Justification
   * Univariate Analysis (Numerical and Categorical Variables)
   * Bivariate Analysis (Correlation, Visualizations, Statistical Tests)
   * Key Takeaways
3. **Part II: Machine Learning Techniques**
   * Overview of Models (Decision Tree, Logistic Regression, Random Forest, KNN)
   * Model Evaluations (Performance Metrics for Each Model)
   * Hyperparameter Optimization
   * Model Comparison
4. **Conclusion**
   * Summary of Findings
   * Model Insights and Limitations
   * Suggestions for Future Work

**Introduction**

Understanding customer satisfaction is a critical component for improving service quality in the airline industry. This report leverages the "Airline Passenger Satisfaction" dataset to explore the key drivers of satisfaction and build predictive models. The dataset, comprising **103,904 instances** and **25 features**, offers a diverse range of variables, including demographic details, service ratings, and operational metrics.

The objectives of this study are twofold:

1. **Descriptive Analysis**: To uncover patterns and relationships in the dataset that influence passenger satisfaction.
2. **Machine Learning**: To build predictive models capable of accurately classifying passengers into Satisfied or Neutral/Dissatisfied categories and identifying key features driving these predictions.

This analysis uses a range of machine learning techniques, including Decision Trees, Logistic Regression, Random Forests, and K-Nearest Neighbors. Hyperparameter optimization was performed to ensure robust performance for the models. The insights gained aim to support data-driven strategies for enhancing customer experiences in the airline industry.

**Part I: Descriptive Analysis**

**1. Dataset Overview and Justification**

The "Airline Passenger Satisfaction" dataset contains **103,904 instances** and **25 variables**. These variables include both numerical (e.g., Age, Flight Distance, Departure Delay in Minutes) and categorical features (e.g., Gender, Class, Type of Travel). This dataset is highly suitable for the objectives of this project, which include:

1. **Classification**: Predicting passenger satisfaction (Satisfied or Neutral/Dissatisfied) using service quality and operational data.
2. **Clustering**: Identifying hidden patterns or groups among passengers based on their ratings and behaviors.

This dataset is appropriate because it combines diverse customer feedback with operational metrics, enabling an exploration of key satisfaction drivers.

**2. Univariate Analysis**

**a) Summary of Numerical Variables**

Key statistics provide insights into passenger demographics and behaviors:

* **Age**: The average passenger is **39 years old** (range: 7 to 85 years).
* **Flight Distance**: Passengers travel an average of **1,189 km**, with some flights covering up to **4,983 km**.
* **Delays**: Both departure and arrival delays show significant skewness. Most flights have no delays, but a few outliers experience delays exceeding **1,500 minutes**.

**b) Summary of Categorical Variables**

The distribution of categorical features highlights passenger characteristics:

* **Gender**: Equally distributed between male and female.
* **Type of Travel**: Business travel dominates, reflecting the dataset's operational focus.
* **Class**: The majority of passengers travel in Business Class, followed by Economy.

**c) Visualizations**

To illustrate these distributions:

* **Histograms**: Continuous variables such as Age, Flight Distance, and log-transformed delays were plotted to highlight their skewness and range.
* **Countplots**: Used for categorical features like Gender, Class, and Type of Travel, showing imbalances favoring business-related attributes.

**3. Bivariate Analysis**

**a) Correlation Analysis**

A correlation matrix reveals key relationships:

* **Positive correlations with satisfaction**:
  + Online boarding (**0.50**), Inflight entertainment (**0.39**), Seat comfort (\*\*0.35`).
* **Negative correlations with satisfaction**:
  + Type of Travel\_Personal Travel (-0.44), Class\_Eco (-0.45).

**b) Conditional Visualizations**

1. **Scatterplot**: Showed a linear relationship between Departure Delay in Minutes and Arrival Delay in Minutes, confirming that delays at departure propagate to arrival.
2. **Boxplots**:
   * Higher ratings for Online boarding and Inflight entertainment were observed for satisfied passengers compared to dissatisfied ones.

**c) Statistical Tests**

1. **T-tests**:
   * Significant differences in Online boarding, Inflight entertainment, and Seat comfort between satisfied and dissatisfied groups (p-values < 0.001).
2. **Chi-squared Tests**:
   * Strong associations between satisfaction and categorical variables like Type of Travel and Class (p-values < 0.001).

**4. Key Takeaways**

1. **Service Quality Matters**:
   * Features like Online boarding, Inflight entertainment, and Seat comfort are the strongest drivers of satisfaction.
2. **Operational Variables Have Limited Impact**:
   * While delays negatively affect satisfaction, their correlation is weaker than that of service ratings.
3. **Demographic and Class Effects**:
   * Business travelers and passengers in Business Class report higher satisfaction, while Economy travelers and personal travel segments report lower satisfaction.

**Conclusion**

The descriptive analysis confirms that service-related variables play a critical role in determining passenger satisfaction. These insights will guide the subsequent machine learning analysis by prioritizing features with strong correlations and significant statistical differences.

**Part II: Machine Learning Techniques**

**1. Overview of Methods**

Several supervised machine learning models were implemented to classify passenger satisfaction (Satisfied or Neutral/Dissatisfied) based on features extracted from the dataset. These include:

1. Decision Tree
2. Logistic Regression
3. Random Forest
4. K-Nearest Neighbors (KNN)

Grid Search was applied to optimize hyperparameters for Decision Tree and Random Forest models, ensuring the best performance.

**2. Model Evaluations**

**a) Decision Tree**

The Decision Tree classifier achieved:

* **Accuracy**: 95%
* **Precision**: 94%
* **Recall**: 94%
* **F1-Score**: 94%
* **AUC (ROC)**: 94%

Key Observations:

* The model performed well across all metrics, indicating strong classification capabilities.
* The confusion matrix revealed a slight bias in predicting the majority class (False), but overall the model balanced precision and recall effectively.

**Visualizations**:

1. **ROC Curve**: AUC of 94% demonstrated strong separation between classes.
2. **Tree Visualization**: Highlights feature splits, showing Online boarding and Inflight entertainment as key decision points.

**b) Logistic Regression**

After scaling the data, Logistic Regression achieved:

* **Accuracy**: 88%
* **Precision**: 88%
* **Recall**: 84%
* **F1-Score**: 86%
* **AUC (ROC)**: 93%

Key Observations:

* The model underperformed compared to Decision Tree and Random Forest but remained robust with a high AUC.
* Recall was slightly lower, indicating potential difficulty in correctly identifying all satisfied passengers.

**Visualizations**:

1. **ROC Curve**: An AUC of 93% suggests good discriminative power despite lower accuracy.
2. **Confusion Matrix**: More false negatives compared to the Decision Tree.

**c) Random Forest**

The Random Forest model, both with default settings and optimized parameters, consistently outperformed other models:

* **Accuracy**: 96%
* **Precision**: 97%
* **Recall**: 94%
* **F1-Score**: 96%
* **AUC (ROC)**: 99%

Key Observations:

* Random Forest showed exceptional performance, particularly in handling complex relationships and reducing overfitting through bagging.
* Optimized parameters slightly improved precision and recall, achieving the best overall results.

**Visualizations**:

1. **ROC Curve**: An AUC of 99% highlights the model's robustness.
2. **Feature Importances**: Key predictors included Online boarding, Inflight entertainment, and Seat comfort.

**d) K-Nearest Neighbors (KNN)**

The KNN model, with default settings (k=5), performed poorly compared to other methods:

* **Accuracy**: 60%
* **Precision**: 63%
* **Recall**: 45%
* **F1-Score**: 50%
* **AUC (ROC)**: 61%

Key Observations:

* KNN struggled to generalize due to its sensitivity to feature scaling and imbalanced data.
* High false positive and false negative rates limited its reliability.

**3. Hyperparameter Optimization**

Using Grid Search, the best parameters were identified for Decision Tree and Random Forest models:

* **Decision Tree**:
  + max\_depth: 30
  + min\_samples\_split: 5
  + min\_samples\_leaf: 2
* **Random Forest**:
  + n\_estimators: 300
  + max\_depth: 20
  + min\_samples\_split: 5
  + min\_samples\_leaf: 1

Optimized models further improved performance metrics, especially for the Random Forest.

**4. Model Comparison**

| **Model** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **AUC (ROC)** |
| --- | --- | --- | --- | --- | --- |
| Decision Tree | 95% | 94% | 94% | 94% | 94% |
| Logistic Regression | 88% | 88% | 84% | 86% | 93% |
| Random Forest | 96% | 97% | 94% | 96% | 99% |
| KNN | 60% | 63% | 45% | 50% | 61% |

**5. Key Insights**

1. **Best Performing Model**:
   * Random Forest demonstrated the best overall performance, excelling in precision, recall, and AUC.
2. **Role of Feature Importance**:
   * Features such as Online boarding, Inflight entertainment, and Seat comfort were consistently influential across models.
3. **Model Limitations**:
   * KNN's poor performance highlights its sensitivity to feature scaling and high-dimensional data.

**Conclusion**

The machine learning analysis confirms that Random Forest is the most effective model for predicting passenger satisfaction. This result aligns with the exploratory analysis, emphasizing the importance of service-related features. Future work could include exploring ensemble methods or deep learning approaches for further improvements.

**Overall Conclusion**

This study highlights the critical role of service-related variables, such as Online boarding, Inflight entertainment, and Seat comfort, in determining passenger satisfaction. The descriptive analysis confirmed these features as the strongest correlates of satisfaction, while operational metrics like delays had a comparatively weaker impact.

Among the machine learning models applied, the Random Forest algorithm emerged as the most effective, achieving an **accuracy of 96%** and an **AUC of 99%**. This model demonstrated superior predictive performance and robustness, reinforcing the importance of advanced ensemble methods for handling complex relationships in the data.

While the findings are promising, there are limitations to consider:

1. **Data Imbalance**: The distribution of certain categories, such as Type of Travel, may influence the model's ability to generalize.
2. **Feature Complexity**: Some variables, like delays, may benefit from additional preprocessing to capture their nuanced effects.

Future work could focus on:

* Implementing ensemble methods like Gradient Boosting or XGBoost for comparison.
* Incorporating external factors, such as economic trends or customer loyalty programs, to enrich the dataset.
* Exploring deep learning models for more sophisticated feature extraction.

Overall, this report provides actionable insights and a robust foundation for using machine learning to improve passenger satisfaction in the airline industry.