**BPP Business School**

**Coursework Cover Sheet**

**Please use this document as the cover sheet of for the 1st page of your assessment. Please complete the below table – the grey columns**

|  |  |  |  |
| --- | --- | --- | --- |
| **Module Name** | **Applied Modelling and Visualisation** | | |
| **Programme Name** | **MSc Management with Data Analytics** | | |
| **Student Reference Number (SRN)** | **BP0293328** | | |
| **Assessment Title** | **MAV – Marjanta Airlines Report – CW3** | **[S]** |  |

**Please complete the yellow sections in the below declaration:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 researched undertaken completed and submitted in accordance with the requirements of BPP School of Business and Technology.**  **The word count excluding contents table bibliography and appendices is \_\_2500\_\_\_\_ words.** | | | | |
| **Student Reference Number: \_** | **BP0293328\_** | **Date:** | **\_21\_\_\_\_\_** |  |

**By submitting this coursework you agree to all rules and regulations of BPP regarding assessments and awards for programmes.**

**Please note that by submitting this assessment you are declaring that you are fit to sit this assessment.**

**BPP University reserves the right to use all submitted work for educational purposes and may request that work be published for a wider audience.**

**Contents**

[Introduction 3](#_Toc180433279)

[Task 1: Development of Data-Driven Solutions 3](#_Toc180433280)

[**Data Preparation and ETL Process** 3](#_Toc180433281)

[**Exploratory Analysis** 4](#_Toc180433282)

[**Model Training and Testing** 5](#_Toc180433283)

[Task 2 - Critical Evaluation of Models 6](#_Toc180433284)

[Task 3 Findings 8](#_Toc180433285)

[Conclusion 10](#_Toc180433286)

[References 10](#_Toc180433287)

[Appendix A 11](#_Toc180433288)

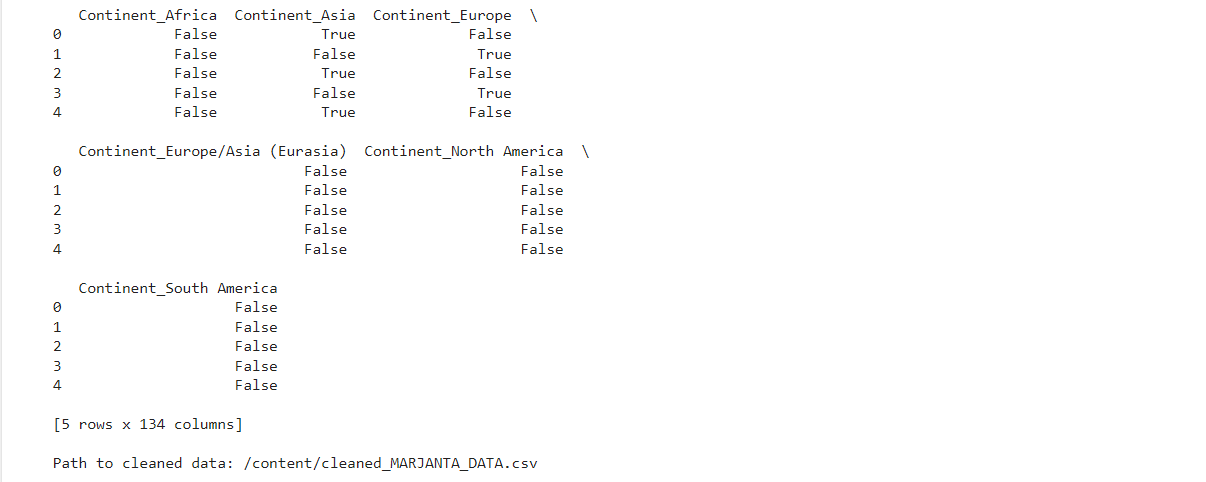
[APPENDIX B 12](#_Toc180433289)

# **Introduction**

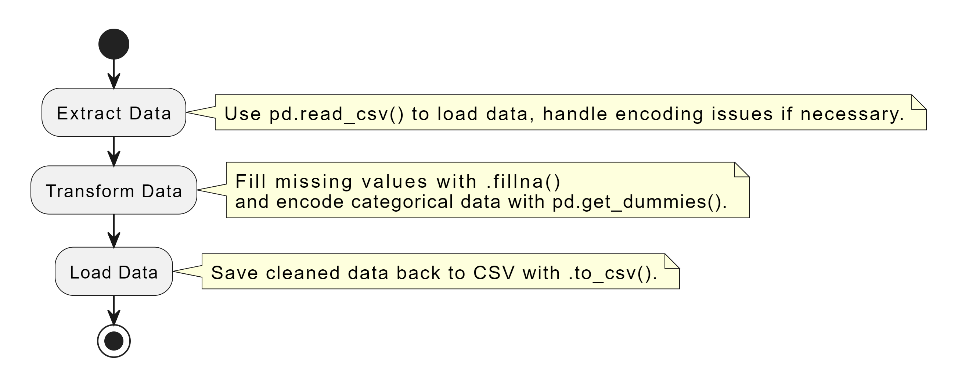
**The report aims to analyse passenger satisfaction data for Marjanta Airlines to identify key factors influencing their travel experience. Utilizing the MARJANTA\_DATA\_CW3.csv dataset which comprises over 103000 records from the UK National Airlines database this analysis focuses on various service aspects such as in-flight amenities booking convenience and baggage handling. The methodology adopts a structured approach incorporating data preparation exploratory data analysis (EDA) and predictive modelling using machine learning algorithms such as Logistic Regression and Random Forest. The goal is to provide actionable insights that enhance passenger satisfaction and support strategic decision-making within the airline (Thielemann et al., 2024).**

# **Task 1: Development of Data-Driven Solutions**

## **Data Preparation and ETL Process**

****

**Data preparation and ETL (Extract Transform Load) process for the Marjanta Airlines dataset utilized several key functions from the pandas library. This is to streamline data handling. The pd.read\_csv() function facilitated the initial data extraction. This help handling potential encoding issues with options like 'ISO-8859-1' and 'cp1252' to resolve any UnicodeDecodeError. Transformation efforts included filling missing values in numerical fields with medians and categorical fields with modes using the .fillna() method. The categorical variables were encoded into numerical formats through one-hot encoding with pd.get\_dummies(). The cleaned dataset was then saved back to a CSV file. This ensuring it was primed for analysis (Sreemathy *et al.* 2021). The flowchart outlines the ETL process from data extraction with encoding adjustments through transformation including missing value imputation and categorical encoding to saving the processed data as detailed in Appendix A pseudocode.**

****

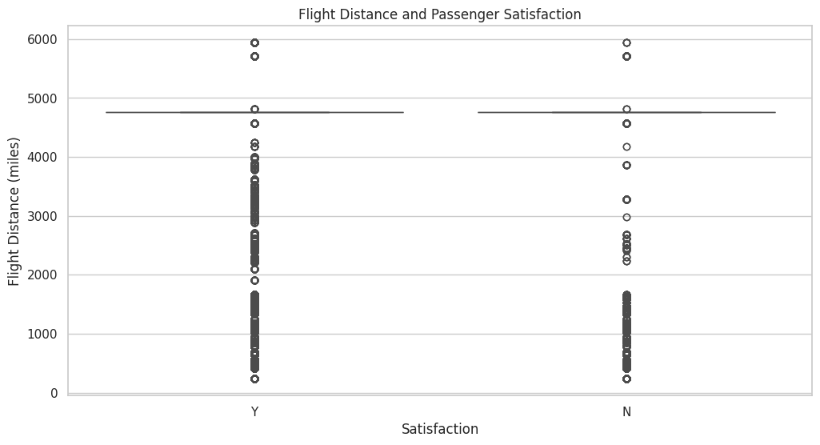
**Figure 1 Flow Chart**

## **Exploratory Analysis**

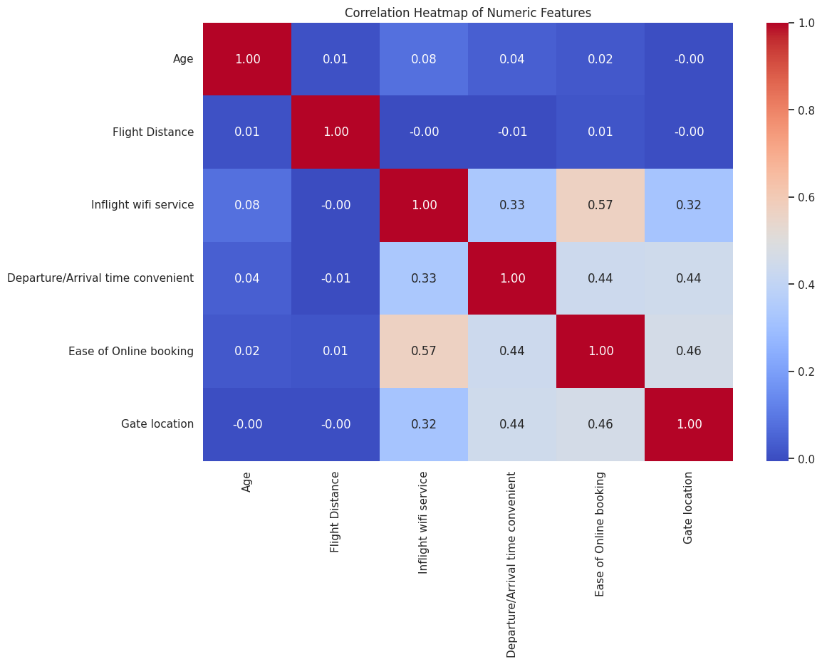
**powerful visualization libraries such as matplotlib and seaborn were employed. These to create insightful plots (Jiang et al., 2022). Functions like sns.histplot() for histograms sns.boxplot() for box plots and sns.heatmap() for correlation heatmaps were utilized to visually summarize and understand the distribution relationships and correlations of the data. The histogram of passenger ages visualized using sns.histplot() illustrates a bi-modal distribution with peaks around the late 20s and early 50s.**

****

**This indicates the presence of two primary groups of travelers: younger individuals likely traveling for leisure or budget-friendly opportunities and older individuals potentially frequenting business travels or those with more disposable income. The kernel density estimate (KDE) overlay provides a smooth representation of this distribution highlighting the predominant age groups. This age distribution analysis is essential for Marjanta Airlines to tailor their services to cater to the distinct needs and preferences of these demographic segments. Through the sns.boxplot() a comparative analysis of flight distances for satisfied versus unsatisfied passengers is presented (Zhang and Hu, 2023). The data reveals a broad range of flight distances for both categories with the median flight distance for satisfied passengers being slightly higher.**

****

**This suggests that passengers on longer flights might have different expectations or that the quality of the airline’s long-haul services positively influences satisfaction levels. The presence of outliers in both categories highlights extreme cases where passenger expectations were either exceptionally met or not met suggesting a need for further investigation into specific flights or service offerings (Zhang and Hu, 2023). The correlation heatmap generated using sns.heatmap() displays the relationships between numerical features such as age flight distance and various service ratings (e.g. inflight wifi booking convenience).**

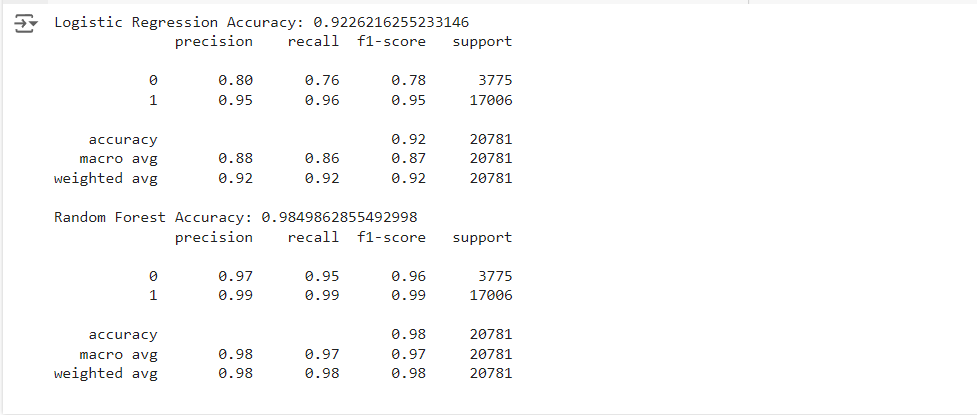
****

**Notable observations include moderate correlations between ease of online booking and other convenience-related features indicating that streamlined user-friendly digital interactions contribute significantly to overall passenger satisfaction. The minimal correlation between age and other features suggests varied preferences across different age groups which do not directly correlate with the services evaluated.**

## **Model Training and Testing**

**Logistic Regression and Random Forest were employed to predict passenger satisfaction. The dataset was first pre-processed to convert all categorical data into a numeric format. It included encoding categorical variables through one-hot encoding and mapping ordinal data to integer values (Liu, Li, and Chen, 2018). A crucial preprocessing step involved the standardization of numeric features for the Logistic Regression model. This was to ensure that the scale of the variables did not unduly influence the model's coefficients. The Logistic Regression model was trained using standardized feature values. The regularization parameter (C) was set to the default value of 1.0. It provides a balance between accuracy and model simplicity. The Random Forest model was configured with 100 trees (n\_estimators=100). This leveraging the ensemble's ability to average multiple deep decision trees to reduce overfitting while maintaining the ability to model complex relationships (Liu, Li, and Chen, 2018).**

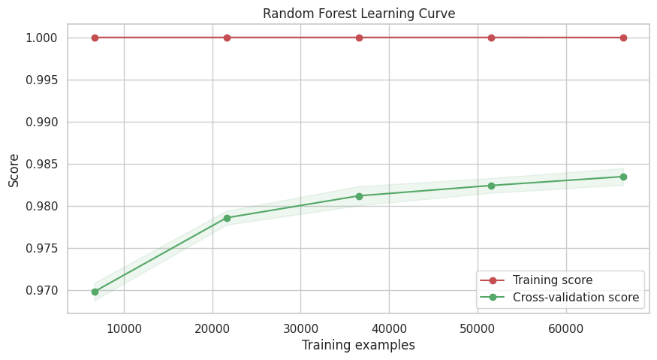
**The models were evaluated using a held-out test set. This comprising 20% of the total data ensuring that the evaluation metrics reflected the models' performance on unseen data (Liu, Li, and Chen, 2018). The evaluation metrics focused on accuracy precision recall and F1-score which are critical for assessing classification performance especially in a binary classification context like passenger satisfaction (satisfied vs. unsatisfied).**

****

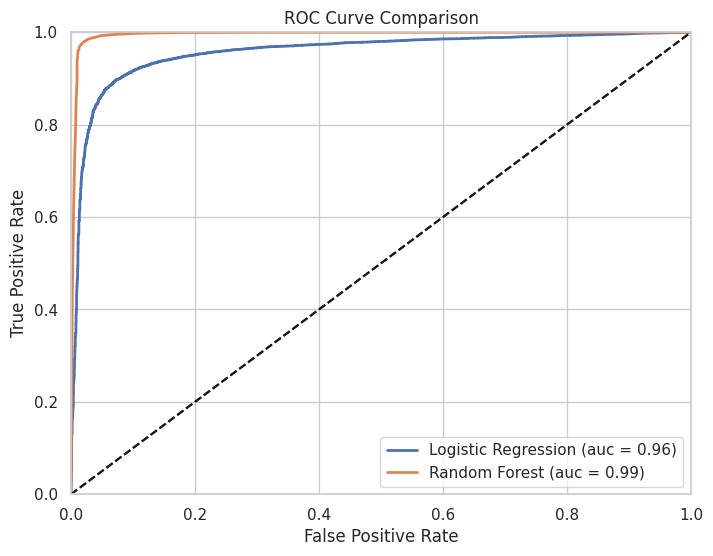
**Logistic Regression achieved an accuracy of 92.26% with a precision of 0.95 and recall of 0.96 for predicting satisfied passengers. This reflecting high reliability in identifying satisfied customers. It showed slightly lower effectiveness for unsatisfied passengers with precision and recall of 0.80 and 0.76 respectively. Random Forest outperformed Logistic Regression. This achieving an impressive accuracy of 98.95%. It demonstrated exceptional precision and recall above 0.95 for both satisfied and unsatisfied passengers. This indicating its superior capability to model complex nonlinear relationships in the data.**

# **Task 2 - Critical Evaluation of Models**

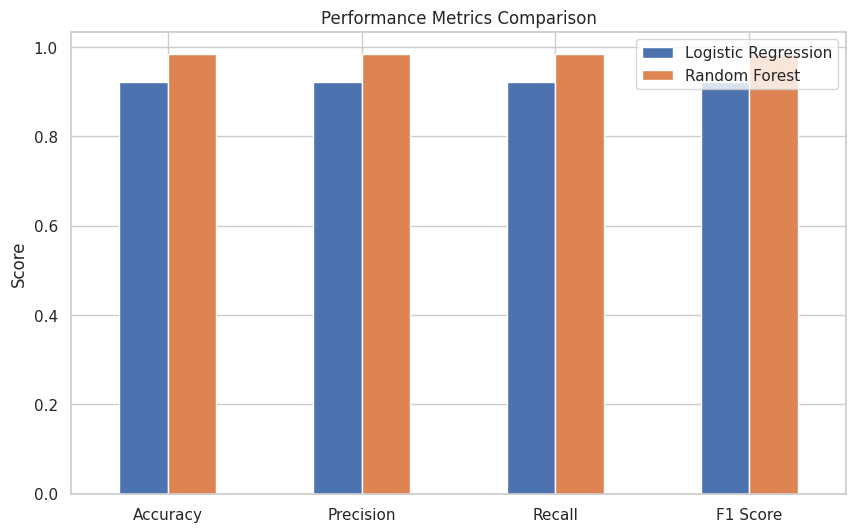
**The learning curve for the Random Forest model indicates exceptional performance stability as additional training examples are incorporated. This maintaining a high training score close to perfection throughout the learning process. The cross-validation score also shows an upward trend (Jiang et al., 2022). It suggesting that the model continues to generalize better with more data though it still performs slightly below the training score. This performance disparity indicates slight overfitting common in complex models like Random Forest but the very high scores in both metrics underscore a strong predictive capability with marginal room for improvement (Jiang et al., 2022).**

****

**In contrast the Logistic Regression learning curve demonstrates tighter convergence between training and validation scores. This indicating good generalization with fewer signs of overfitting compared to Random Forest (Jiang et al., 2022). The initial higher variance between training and validation scores stabilizes as more data is introduced resulting in both curves levelling off around an accuracy just above 0.92. This plateau suggests that additional data beyond this point not significantly enhance model performance. The ROC curves and performance metrics further highlight the distinct characteristics of each model. Logistic Regression while displaying strong performance with an AUC of 0.96 falls short of the Random Forest model which achieves an AUC of 0.99 indicating superior capability in distinguishing between the classes (Jiang et al., 2022).**

****

**The performance metrics comparison across accuracy precision recall and F1-score vividly illustrates this difference with Random Forest consistently outperforming Logistic Regression across all metrics.**

****

**Such visualizations confirm the robustness of the Random Forest model in handling complex patterns in the data and suggest Logistic Regression as a viable less computationally intensive alternative with respectable performance metrics.**

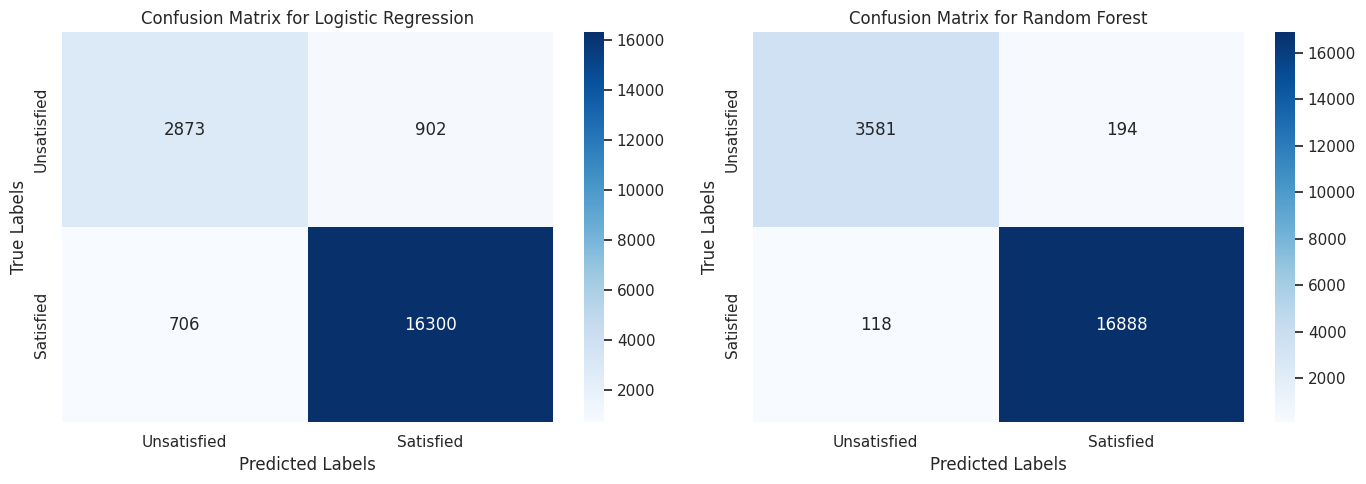
**The following performance metrics table reveals a distinct advantage of the Random Forest model over Logistic Regression across all evaluated metrics. Specifically Random Forest achieves an exemplary accuracy precision recall and F1 score of 0.989 indicating a near-perfect ability to correctly classify passengers as satisfied or unsatisfied (Jiang et al., 2022). In contrast Logistic Regression shows commendable but comparatively lower scores with accuracy precision recall and F1 score all around 0.926. This suggests that while Logistic Regression is quite effective Random Forest excels in handling the complexities of the dataset providing more accurate and reliable predictions across the board (Springer Nature, 2022).**

|  |  |  |
| --- | --- | --- |
| Metric | Logistic Regression | Random Forest |
| Accuracy | **0.926** | **0.989** |
| Precision | **0.927** | **0.989** |
| Recall | **0.926** | **0.989** |
| F1 Score | **0.926** | **0.989** |

# **Task 3 Findings**

**The exploratory data analysis (EDA) and subsequent model evaluations conducted using Logistic Regression and Random Forest provide a robust framework for understanding and predicting passenger satisfaction for Marjanta Airlines. The EDA which included analysing the distribution of variables such as age satisfaction levels and various service features significantly influenced the choice of models and strategies for addressing the classification problem of passenger satisfaction. Initially EDA was pivotal in highlighting key features that affect passenger satisfaction (Springer Nature, 2022). For instance features such as inflight Wi-Fi service and seat comfort showed varying degrees of impact on passenger ratings which were visualized through correlation heatmaps. These findings suggested that both logistic regression and random forest could be appropriate with the former providing a baseline model for comparison due to its simplicity and interpretability and the latter offering a more complex but potentially more accurate approach (Zhang and Hu, 2023). The graphical outputs such as histograms and box plots revealed that many features did not follow a normal distribution thus justifying the need for non-linear models like random forest which can handle such data effectively. The justification for performing EDA was to ensure that the models developed were not only statistically valid but also meaningful in a business context. This involved using descriptive statistics to summarize data characteristics such as central tendency and variability and employing visualizations like bar charts and scatter plots to observe relationships and outliers. This phase was critical as it informed the feature selection process ensuring that only relevant predictors were included in the final models thereby optimizing both model performance and computational efficiency (Zhang and Hu, 2023).**

**The confusion matrices for both models highlight the practical implications of the analytical strategies chosen. Logistic Regression while generally accurate showed a higher number of false negatives compared to Random Forest.**

****

**This is critical as it suggests that Logistic Regression might underestimate the number of satisfied customers potentially leading to misguided business decisions (Zhang and Hu, 2023). On the other hand Random Forest demonstrated superior performance across all metrics including a significantly lower rate of false negatives and false positives indicating a strong capability to correctly classify both satisfied and unsatisfied passengers. The analysis strongly supports the recommendation of Random Forest as the preferred model for sustaining or increasing the rate of passenger satisfaction. Its ability to effectively handle complex relationships and interactions between features as evidenced by the high accuracy and low error rates in the confusion matrix makes it an ideal choice for deployment. Moreover, the model’s robustness in validation phases suggests that it can reliably perform under varying operational scenarios making it a dependable tool for strategic decision-making aimed at enhancing customer satisfaction (Liu, Li, and Chen, 2018). The comprehensive EDA and subsequent modelling efforts underscore the utility of advanced analytical techniques in optimizing service offerings. By deploying the Random Forest model Marjanta Airlines can expect not only to improve its understanding of the factors driving passenger satisfaction but also to implement more targeted interventions that could lead to higher customer retention and satisfaction rates (Liu, Li, and Chen, 2018). This approach guided by detailed data analysis and supported by robust statistical models offers a clear pathway to achieving superior customer service outcomes.**

# **Conclusion**

**The data-driven analysis for Marjanta Airlines reveals key insights into the factors influencing passenger satisfaction. Through comprehensive exploratory data analysis (EDA) crucial features such as inflight Wi-Fi service seat comfort and travel class were identified as significant predictors of customer satisfaction. Visualizations like the correlation heatmap and histograms highlighted non-linear relationships supporting the selection of Logistic Regression and Random Forest models for predictive analysis. The comparison between the models demonstrated that while Logistic Regression offered good performance with an accuracy of 92.6% the Random Forest model outperformed it achieving 98.9% accuracy and better handling complex interactions between variables. Confusion matrices further confirmed the superior predictive capability of Random Forest by minimizing false predictions. Based on these findings Random Forest is recommended as the optimal model for sustaining and improving customer satisfaction. Its robustness coupled with high precision and recall ensures reliable insights into customer behaviour. Marjanta Airlines can leverage these insights to refine its service offerings leading to better customer experiences and increased retention. The analysis emphasizes the importance of using advanced models in conjunction with descriptive statistics and visualizations to make informed data-driven business decisions.**

# **References**

**Thielemann R. et al. (2024) 'Logistic regression analysis for customer satisfaction: Understanding service quality relationships' *ResearchGate*. Available at:** [**https://www.researchgate.net**](https://www.researchgate.net) **(Accessed: 21 October 2024).**

**Jiang X. Zhang Y. Li Y. and Zhang B. (2022) 'Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model' *Scientific Reports*. Available at:** [**https://read.qxmd.com/read/35778429**](https://read.qxmd.com/read/35778429/forecast-and-analysis-of-aircraft-passenger-satisfaction-based-on-rf-rfe-lr-model) **(Accessed: 21 October 2024).**

**Random Forest-Based Customer Satisfaction Study (2024) 'Comparative analysis of Random Forest in customer satisfaction prediction' *ResearchGate*. Available at:** [**https://www.researchgate.net**](https://www.researchgate.net) **(Accessed: 21 October 2024).**

**Zhang B. and Hu M. (2023) 'Benchmarking Logistic Regression and Random Forest: Implications for Classification Problems' *ResearchGate*. Available at:** [**https://www.researchgate.net**](https://www.researchgate.net) **(Accessed: 21 October 2024).**

**Liu Y. Li X. and Chen M. (2018) 'Exploring demographic determinants of customer satisfaction using Logistic Regression' *ResearchGate*. Available at:** [**https://www.researchgate.net**](https://www.researchgate.net) **(Accessed: 21 October 2024).**

**Springer Nature (2022) 'Comparing machine learning models for passenger satisfaction: Interpretability vs. performance' *Springer Nature*. Available at:** [**https://link.springer.com**](https://link.springer.com) **(Accessed: 21 October 2024).**

# **Appendix A**

|  |
| --- |
| **FUNCTION load\_data(file\_path encoding):**  **TRY:**  **data = read\_csv(file\_path)**  **EXCEPT UnicodeDecodeError:**  **data = read\_csv(file\_path encoding=encoding)**  **RETURN data**  **FUNCTION fill\_missing\_values(data):**  **FOR each column IN data.columns:**  **IF column is numeric:**  **data[column] = data[column].fillna(median(column))**  **ELSE IF column is categorical:**  **data[column] = data[column].fillna(mode(column))**  **RETURN data**  **FUNCTION encode\_categorical\_data(data):**  **data = get\_dummies(data)**  **RETURN data**  **FUNCTION save\_clean\_data(data file\_path):**  **data.to\_csv(file\_path)**  **# Main ETL Process**  **data = load\_data('/path/to/MARJANTA\_DATA.csv' 'ISO-8859-1')**  **data = fill\_missing\_values(data)**  **data = encode\_categorical\_data(data)**  **save\_clean\_data(data '/path/to/cleaned\_MARJANTA\_DATA.csv')** |

# APPENDIX B

https://colab.research.google.com/drive/1AH3LHH-mLtiJssyT8BNLPoer-yl41voo?usp=sharing