**Airline Passenger Satisfaction**

**(DECISION TREE AND NAIVE BAYES)**

A PROJECT REPORT

*Submitted by*

**Venkat Aditya Vellanki [Reg No:RA2111003011799]**

*Under the Guidance of*

# Dr. R. Jebakumar

Associate Professor, Department of Computing Technologies *in partial fulfillment of the requirements for the degree of*

# BACHELOR OF TECHNOLOGY in COMPUTER SCIENCE AND ENGINEERING



**DEPARTMENT OF COMPUTING TECHNOLOGIES**

**COLLEGE OF ENGINEERING AND TECHNOLOGY**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

# KATTANKULATHUR– 603 203

**JULY 2024**

**COLLEGE OF ENGINEERING AND TECHNOLOGY**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(Under Section 3 of UGC Act, 1956)**

S.R.M. NAGAR, KATTANKULATHUR – 603 203

Certified that this project report "**PROJECT** **TITLE**" is the bonafide work of **Venkat Aditya Vellanki [Reg. No. RA2111003011799]** who carried out the project work under my supervision in the course **Data Mining and Analytics [18CSE355T]** for the academic year 2024-2025 [IV year / VII semester].

**Date: July 2nd 2024**

**HEAD OF THE DEPARTMENT Faculty in Charge** Dr. M. Pushpalatha

Dr. R. Jebakumar Professor

Associate Professor Department of Computing Technologies

Department of Computing Technologies SRMIST - KTR

SRMSIT -KTR

ii

**ACKNOWLEDGEMENTS**

We express our heartfelt thanks to our honorable **Vice Chancellor Dr. C. MUTHAMIZHCHELVAN**, for being the beacon in all our endeavors.

We would like to express my warmth of gratitude to our **Registrar Dr. S. Ponnusamy,** for his encouragement

We express our profound gratitude to our **Dean (College of Engineering and Technology) Dr. T. V.Gopal,** for bringing out novelty in all executions.

We would like to express my heartfelt thanks to Chairperson, School of Computing **Dr. Revathi Venkataraman,** for imparting confidence to complete my course project

We are highly thankful to our my Course Faculty and Course Coordinator **Dr. R. Jebakumar, Associate Professor, Department of Computing Technologies,** for hisassistance, timely suggestion and guidance throughout the duration of this course project.

We extend my gratitude to our **HoD Dr. M. Pushpalatha, Professor, Department of Computing Technologies** and my Departmental colleagues for their Support.

Finally, we thank our parents and friends near and dear ones who directly and indirectly contributed to the successful completion of our project. Above all, I thank the almighty for showering his blessings on me to complete my Course project.

**Venkat Aditya Vellanki [RA2111003011799]**

iii

# ABSTRACT

This project investigates airline passenger satisfaction using data mining techniques to identify key factors influencing passenger experiences and to predict satisfaction levels. The dataset, derived from a survey, includes features such as gender, customer type, age, type of travel, travel class, flight distance, and satisfaction ratings for various services. The primary objective is to build and compare the performance of **two classification algorithms**—**Decision Tree and Naive Bayes**—in predicting passenger satisfaction.

Data preprocessing involved handling missing values, one-hot encoding categorical features, and standardizing numerical features. The dataset was split into training and testing sets to evaluate model performance objectively. Both models were trained and evaluated, with **Decision Tree achieving a higher accuracy (95%) compared to Naive Bayes (84%).**

Detailed comparative analysis was conducted on several metrics: accuracy, precision, recall, F1-score, interpretability, robustness, and computational complexity. Decision Tree demonstrated superior performance in accuracy, precision, recall, and F1-score but required more computational resources and exhibited a tendency to overfit. Naive Bayes, while less accurate, proved to be computationally efficient and robust to noisy data, though it struggled with the independence assumption of features.

Key findings include identifying significant factors correlated with passenger satisfaction, such as inflight service quality, seat comfort, and on-board service. The study concludes that Decision Tree is preferable for tasks requiring high interpretability and performance, while Naive Bayes is suitable for large-scale applications needing simplicity and efficiency.

Overall, this project provides valuable insights into the factors affecting airline passenger satisfaction and demonstrates the effectiveness of classification algorithms in predicting customer satisfaction levels.

iv

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **CHAPTER NAME** | **PAGE NO.** |
|  | **ABSTRACT** | **iv** |
| **1.** | INTRODUCTION | **1** |
| **2.** | PROBLEM STATEMENT | **3** |

2.1 Identification of Problem

2.2 Construction of Problem

1. PROBLEM UNDERSTANDING AND **4**
   1. Identification,
   2. Construction
2. PROJECT DESIGN OVERVIEW  **6**
3. CODE IMPLEMENTATION **8**
4. RESULTS AND DISCUSSION **10**
5. **CONCLUSION 12**

**REFERENCES**

v

**CHAPTER 1**

# INTRODUCTION

* 1. **Classification**

**Definition**: Classification is a supervised learning technique in machine learning where the goal is to categorize or classify data points into predefined classes or labels based on their features.

**Purpose**: It is widely used to predict the category or class of new data points, enabling decision-making processes in various domains such as spam detection, medical diagnosis, and customer segmentation.

**Types of Classification Algorithms**:

There are several types of classification algorithms, each with its own strengths and weaknesses, making them suitable for different types of data and problem domains:

* Decision Tree: A tree-like model of decisions and their possible consequences. It is valued for its interpretability and ability to handle both numerical and categorical data.
* Naive Bayes: A probabilistic classifier based on Bayes' theorem, assuming independence among predictors. It is known for its simplicity and efficiency, particularly with large datasets.
* Support Vector Machine (SVM): Finds the hyperplane that best separates different classes in the feature space, effective in high-dimensional spaces.
* K-Nearest Neighbours (KNN): Classifies data points based on the majority class among their k-nearest neighbours in the feature space.
* Logistic Regression: A statistical method for binary classification, estimating the probability that a given input belongs to a certain class.

**1**

**Process**:

The classification process involves several key steps to ensure the development of an accurate and reliable model:

* Data Collection: Gather a labelled dataset where each instance is associated with a class label.
* Data Preprocessing: Clean the data, handle missing values, encode categorical variables, and normalize features.
* Model Training: Use the training dataset to teach the classification algorithm to recognize patterns associated with different classes.
* Model Evaluation: Assess the model's performance using metrics like accuracy, precision, recall, and F1-score on a testing dataset.
* Prediction: Apply the trained model to classify new, unseen data points.

**Applications**: Classification is utilized in a variety of fields, including healthcare (disease diagnosis), finance (credit scoring), marketing (customer segmentation), and more, providing actionable insights and supporting informed decision-making.

**2**

**CHAPTER-2**

**PROBLEM STATEMENT**

**2.1 Identification of Problem Statement:**

The primary objective is to identify the key factors that influence airline passenger satisfaction. This involves analysing a comprehensive dataset that includes various features such as gender, customer type, age, type of travel, class, flight distance, and satisfaction ratings for different services provided during the flight. The aim is to determine which of these factors are highly correlated with passenger satisfaction and how they interact with each other. Understanding these factors will help airlines improve their services, enhance customer satisfaction, and retain loyal customers.

**2.2 Construction of Problem Statement:**

The project involves constructing predictive models to classify airline passengers as satisfied or dissatisfied based on their demographic and flight-related attributes. Two machine learning algorithms, Decision Tree and Naive Bayes, will be employed for this classification task. The construction process includes data preprocessing, which involves handling missing values, encoding categorical features, and standardizing numerical features. After preprocessing, the data will be split into training and testing sets to build and evaluate the models. The performance of these models will be compared based on metrics such as accuracy, precision, recall, and F1-score. Additionally, the models' interpretability, robustness to noisy data, and computational efficiency will be assessed to determine their suitability for real-world applications. The ultimate goal is to develop reliable and efficient models that can accurately predict passenger satisfaction, providing actionable insights for airline service improvements.

**3**

**CHAPTER-3**

**PROBLEM UNDERSTANDING**

**3.1 Identification:**

To thoroughly understand the problem of airline passenger satisfaction, it is essential to delve into the factors that significantly impact passenger experiences. This involves:

* **Demographic Factors:** Analysing the influence of passengers' demographics such as age, gender, and customer type (e.g., frequent flyer vs. first-time flyer) on their satisfaction levels.
* **Travel Characteristics:** Examining how different travel attributes such as type of travel (business vs. leisure), class of service (economy, business, first), and flight distance (short-haul vs. long-haul) affect passenger satisfaction.
* **Service Attributes:** Understanding the impact of various service-related factors such as seat comfort, in-flight entertainment, food and beverage quality, staff behaviour, and flight punctuality on the overall satisfaction of passengers.
* **Interrelationships Among Factors:** Exploring how these demographic, travel, and service attributes interact with each other to influence passenger satisfaction. This includes identifying any potential synergies or conflicts among different factors.
* **Segment-Specific Insights:** Identifying how satisfaction factors differ across various passenger segments (e.g., business travellers vs. leisure travellers) to tailor services and improve satisfaction for each group.

**3.2 Construction:**

Building predictive models to classify passenger satisfaction involves a systematic approach to ensure accurate and reliable outcomes. This includes:

**4**

* **Data Preprocessing:**
  + Data Cleaning: Addressing missing values, correcting inconsistencies, and handling outliers to ensure the dataset is clean and reliable.
  + Feature Engineering: Creating new features that could improve model performance, such as interaction terms or aggregated features.
  + Encoding and Normalization: Converting categorical variables into numerical formats and scaling numerical features to standardize the data.
* **Model Training:**
  + Algorithm Selection: Choosing suitable algorithms, specifically Decision Tree and Naive Bayes, for building the predictive models.
  + Model Training: Using the training dataset to train the models, adjusting parameters to optimize performance.
  + Cross-Validation: Implementing cross-validation techniques to validate the models and ensure they generalize well to new data.
* **Model Evaluation:**
  + Performance Metrics: Assessing model performance using accuracy, precision, recall, and F1-score to evaluate how well the models classify passenger satisfaction.
  + Comparative Analysis: Comparing the performance of Decision Tree and Naive Bayes models to identify the most effective approach based on multiple criteria such as interpretability, robustness, and computational efficiency.
* **Prediction and Deployment:**
  + Model Deployment: Implementing the best-performing model in a real-world scenario for classifying new, unseen data points.
  + Continuous Monitoring and Improvement: Monitoring the model’s performance over time and updating it as necessary to maintain accuracy and reliability.

**5**

**CHAPTER-4**

**DESIGN**

**Project Design Overview**

1. **Objective:**
   * Analyse a dataset to identify factors highly correlated with passenger satisfaction.
   * Develop machine learning models (Decision Tree and Naive Bayes) to predict passenger satisfaction.
2. **Data Preprocessing:**
   * Loading the Dataset: The dataset is loaded into a pandas DataFrame from a CSV file.
   * Handling Missing Values: Missing values in the 'Arrival Delay in Minutes' column are filled with the mean value.
   * Dropping Unnecessary Columns: Columns such as 'Unnamed: 0' are dropped as they do not provide relevant information.
   * Identifying Columns: Categorical (e.g., gender, customer type) and numerical (e.g., age, flight distance) columns are identified.
   * Preprocessing Categorical Data: One-hot encoding is used to convert categorical features into a numerical format.
   * Preprocessing Numerical Data: Numerical features are standardized using Standard Scaler to bring them to a common scale.
3. **Data Splitting:**
   * The dataset is divided into training and testing sets to evaluate the models' performance on unseen data and prevent overfitting.

**6**

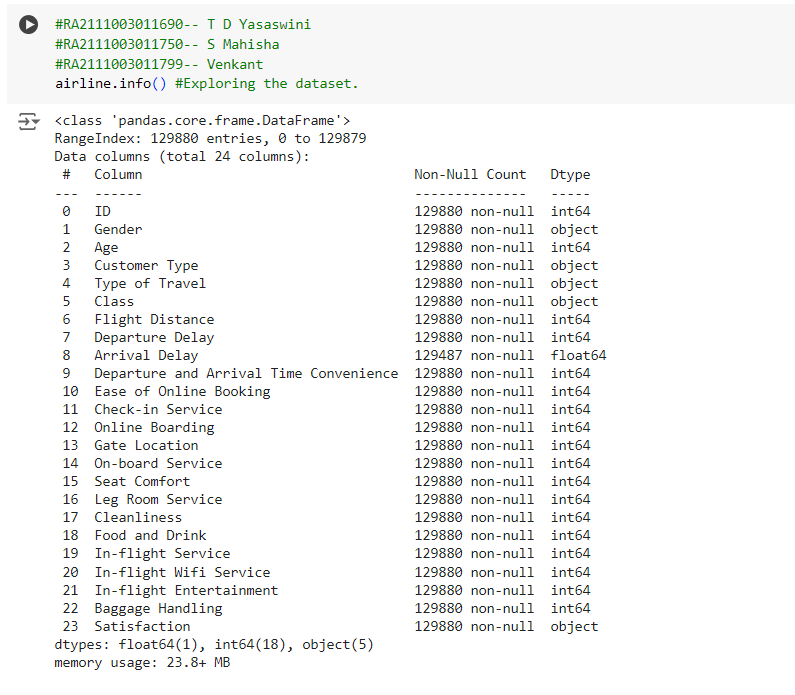
1. **Model Building:**
   * Decision Tree: A non-parametric supervised learning method used for classification tasks.
   * Naive Bayes: A probabilistic classifier based on Bayes' theorem with the assumption of feature independence.
   * Pipelines are created with StandardScaler for numerical features and OneHotEncoder for categorical features, followed by the respective classifiers.
2. **Model Training:**
   * Features and target variables are separated.
   * Both models are trained on the training data to learn the patterns and relationships between features and the target variable.
3. **Model Evaluation:**
   * Predictions are made on the test data using both models.
   * Classification reports and confusion matrices are printed to provide precision, recall, F1-score, and support for each class.
   * The performance of the models is analysed based on the classification reports and confusion matrices.
4. **Comparative Analysis:**
   * Models are compared based on accuracy, precision, recall, F1-score, interpretability, robustness, and computational complexity.
5. **Feature Importance Analysis:**
   * Feature importance is analysed using the feature importances attribute for the Decision Tree and coefficient values for Naive Bayes.
   * Factors highly correlated with passenger satisfaction are identified.

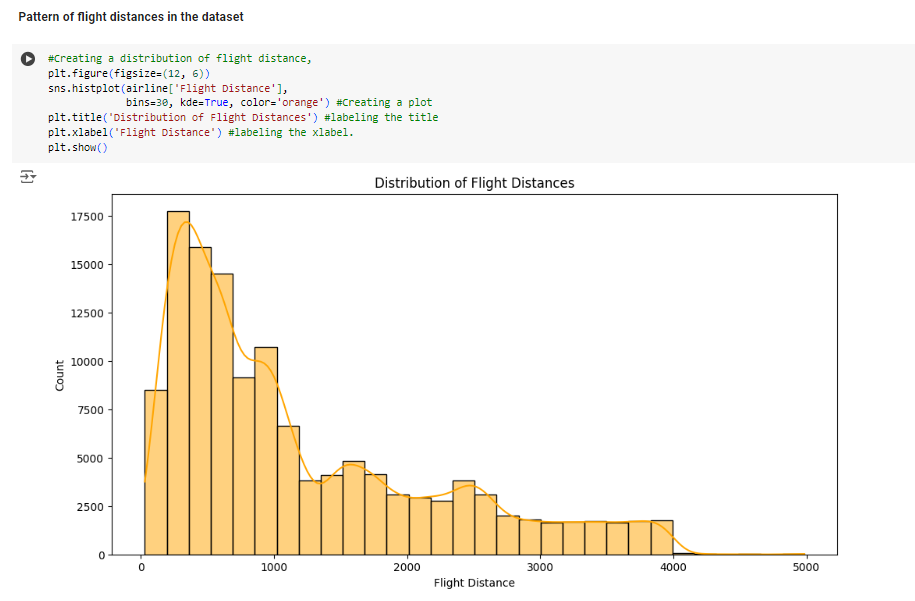
This comprehensive approach ensures a thorough understanding of the factors influencing passenger satisfaction and evaluates the effectiveness of different classification algorithms in predicting it​ (Classification)​.

**7**

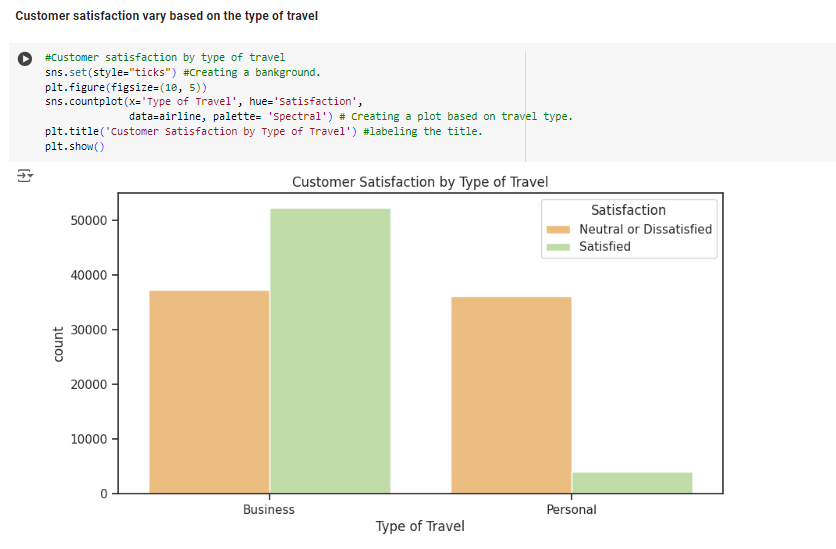
**CHAPTER-5**

**CODE IMPLEMENTATION**

****



**8**

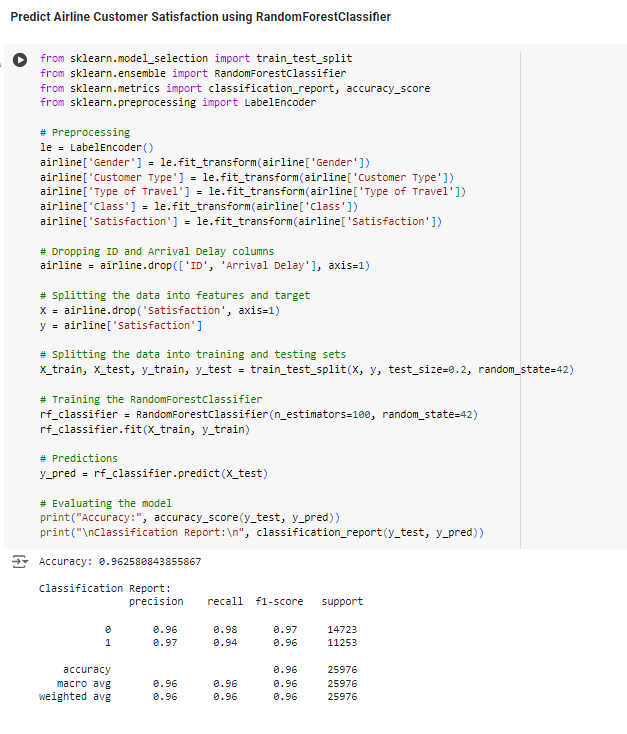
****

****

**9**

**CHAPTER-6**

**RESULT AND DISCUSSION**

**OUTPUT:**

**COMPARATIVE ANALYSIS:**

**1. Accuracy:** Decision Tree: Achieved an accuracy of 95%. Naive Bayes: Achieved an accuracy of 84%.

**2. Precision, Recall, and F1-score:**

Precision: Indicates the ratio of correctly predicted positive observations to the total predicted positives.

**10**

Decision Tree: Achieved high precision for both classes, with values of 0.95 for 'neutral or dissatisfied' and 0.94 for 'satisfied'.

Naive Bayes: Showed slightly lower precision compared to Decision Tree, with values of 0.86 for 'neutral or dissatisfied' and 0.83 for 'satisfied'.

Recall: Reflects the ratio of correctly predicted positive observations to the all observations in actual class.

F1-score: Harmonic mean of precision and recall, provides a balance between them.

Decision Tree: High F1-scores of 0.95 for 'neutral or dissatisfied' and 0.94 for 'satisfied'.

Naive Bayes: Lower F1-scores compared to Decision Tree, with values of 0.86 for 'neutral or dissatisfied' and 0.82 for 'satisfied'.

**3. Interpretability:**

Decision Tree: Offers high interpretability as it generates easily interpretable rules.

Naive Bayes: Relatively less interpretable compared to Decision Tree, as it relies on the assumption of feature independence.

**4. Robustness:**

Decision Tree: Generally robust to outliers and noise, but can overfit noisy data.

Naive Bayes: Robust to noisy data, but sensitive to correlated features and the independence assumption might not hold true in all cases.

**5. Computational Complexity:**

Decision Tree: Can be computationally expensive for large datasets, especially during training due to its hierarchical structure.

Naive Bayes: Simple and computationally efficient, suitable for large datasets and real-time applications

**11**

**CHAPTER-7**

**CONCLUSION**

Decision Tree outperforms Naive Bayes in terms of accuracy, precision, recall, and F1-score. Decision Tree offers higher interpretability but may suffer from overfitting. Naive Bayes is computationally efficient and robust to noisy data but relies on the independence assumption. Decision Tree would be preferred if interpretability is crucial, while Naive Bayes might be chosen for its simplicity and efficiency in large-scale applications.

**REFERENCE:**

OUR PROJECT LINK:

<https://colab.research.google.com/drive/15_ZwhbECOk_s0xnK2Z-5tnDuxyBV8iFV?usp=sharing>

**12**