# BI2025 Experiment Report - Group 18: Airline Passenger Satisfaction Prediction

Aliya Bokey (Student A)*
e12451104@student.tuwien.ac.at
TU Wien
Austria

Alicya Novita Hariyanto (Student B)†
e12536814@student.tuwien.ac.at
TU Wien
Austria

## Abstract

This report documents a comprehensive machine learning experiment for predicting airline passenger satisfaction, following the CRISP-DM process model. The analysis focuses on the Airline Passenger Satisfaction dataset, containing approximately 129,880 passenger records with 24 attributes including demographic information, travel characteristics, and service quality ratings. A Random Forest classifier was applied and evaluated using a structured train-validation–test split, with performance assessed through standard classification metrics and hyperparameter tuning. This report also addresses bias and ethical considerations, documents the analytics workflow using a provenance knowledge graph, and provides deployment and monitoring recommendations.

## CCS Concepts

• **Computing methodologies → Machine learning**.

## Keywords

Customer Satisfaction, Airline Industry, CRISP-DM, Provenance, Knowledge Graph, Machine Learning, Random Forest, Bias Evaluation

**ACM Reference Format:**
Aliya Bokey (Student A) and Alicya Novita Hariyanto (Student B). 2025. BI2025 Experiment Report - Group 18: Airline Passenger Satisfaction Prediction. In *Proceedings of Business Intelligence (BI 2025).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Business Understanding

*Responsibility: Both Student A & B contributed equally to this section.*

The airline industry faces significant challenges in maintaining customer loyalty and satisfaction in a highly competitive market. Customer dissatisfaction can lead to churn, negative reviews, and reduced profitability. This experiment aims to analyze the Airline Passenger Satisfaction dataset to identify factors influencing passenger satisfaction and develop predictive models for dissatisfaction detection.

### 1.1 Data Source and Scenario

The dataset "Airline Passenger Satisfaction" from [3] originated from a real-world customer satisfaction survey conducted by an airline company. The dataset contains responses from approximately 129,000 passengers and captures demographic information, travel characteristics, and detailed service quality ratings across multiple service dimensions (e.g. inflight service, seat comfort, online boarding). The business scenario of this experiment involves a commercial airline who want to improve its customer satisfaction and its customer churn rate by identifying passengers who are likely to be dissatisfied with their flight experience.

### 1.2 Business Objectives

The primary business objective is to proactively identify passengers who are likely to be dissatisfied in order to enable targeted service improvements, customer retention strategies, and operational adjustments. A secondary objective is to gain insights into which service attributes have the strongest impact on passenger dissatisfaction.

### 1.3 Business Success Criteria

The business objective is considered successful if a predictive model can reliably identify dissatisfied passengers with sufficiently high recall, allowing the airline to intervene before dissatisfaction results in loss of customer loyalty or negative reputation effects.

### 1.4 Data Mining Goals

The data mining goal is to build a supervised classification model that predicts whether a passenger is "neutral or dissatisfied" based on demographic attributes, travel characteristics, and service quality ratings.

### 1.5 Data Mining Success Criteria

From a data mining perspective, success is defined as achieving performance significantly above a random or trivial baseline, measured using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with particular emphasis on the minority or business-critical class of dissatisfied passengers.

### 1.6 AI Risk Aspects

Potential AI risks include bias against specific passenger groups, such as younger or older passengers, economy-class travelers, or customers traveling for personal reasons. Additionally, the use of

*Student A (Data Understanding, Modeling), Matr.Nr.: 12451104
†Student B (Data Preparation, Evaluation), Matr.Nr.: 12536814


*BI 2025, -*

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/XXXXXXX.XXXXXXX

demographic attributes (e.g., age, gender) may raise fairness concerns if the model's predictions disproportionately affect certain subgroups. There is also a risk that satisfaction labels reflect subjective perceptions influenced by external factors not captured in the dataset.

## 2 Data Understanding

*Responsibility: Student A contributed to this section.*

### 2.1 Dataset Description

Customer satisfaction survey data from 129,880 airline passengers with 24 attributes including demographic information, travel characteristics, service ratings (1-5 Likert scale), delay information, and satisfaction target.

### 2.2 Feature Overview

The dataset contains the following key features (first 10 shown):

**Table 1: Selected Dataset Features**

| Feature Name | Data Type | Description |
|---|---|---|
| Age | integer> | Passenger age in years |
| Arrival Delay in Minutes | float> | Arrival delay in minutes (may contain missing values) |
| Baggage handling | integer> | Baggage handling service (1-5) |
| Checkin service | integer> | Check-in service quality (1-5) |
| Class | string> | Travel class: Eco, Eco Plus, or Business |
| Cleanliness | integer> | Cleanliness of aircraft (1-5) |
| Customer Type | string> | Loyal Customer or disloyal Customer |
| Departure Delay in Minutes | integer> | Departure delay in minutes |
| Departure/Arrival time convenient | integer> | Convenience of departure/arrival times (1-5) |
| Ease of Online booking | integer> | Ease of online booking process (1-5) |

### 2.3 Statistical Properties & Correlations

Descriptive statistics reveal key characteristics of the dataset:

- **Age:** Range 7-85 years, mean 39.4 years, normal distribution with slight right skew
- **Flight Distance:** Range 31-4,983 km, mean 1,190.3 km, right skewed distribution
- **Departure Delay:** Mean 14.7 minutes, but with extreme values up to 1,592 minutes
- **Service Ratings:** All 14 service attributes use 1-5 Likert scale with means between 2.73-3.68

Correlation analysis shows strong associations between passenger satisfaction and service-related attributes. Preliminary analysis indicates that Online boarding, Inflight entertainment, Seat comfort, and Cleanliness have the strongest correlations with satisfaction, whereas demographic attributes like Age and Gender show weaker direct correlations.

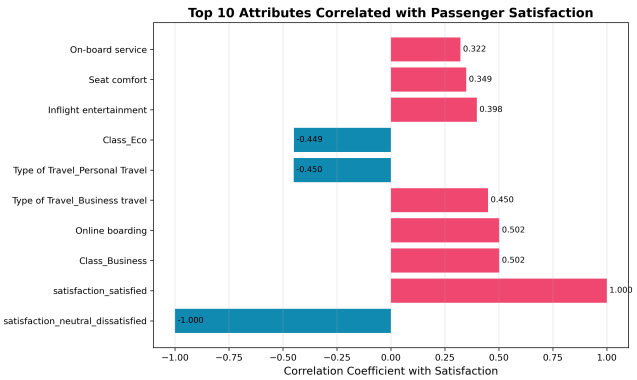


**Figure 1: Top 10 Most Strongly Correlated Attributes with Passenger Satisfaction**

Figure 1 provides an insight that **the convenience of both pre-flight and in-flight determines the level of passenger's satisfaction**. This is proven by service-related factors that possess strong positive correlation, such as Online boarding (0.502), Business class (0.502), Business travel (0.450), Inflight entertainment (0.398), followed by moderate correlation factors such as Seat comfort (0.349) and On-board service (0.322).

In contrast, passengers traveling with Economy class (-0.449) and are Personal travel (−0.450) are negatively correlated with satisfaction. Greater satisfaction resulted through the ease of boarding and traveling with business class also supported by previous related study [6]. Overall, the results emphasize that service experience and travel context are stronger determinants of satisfaction than passenger characteristics.

### 2.4 Data Quality Aspects

The dataset consists 129,880 records. A small number of missing values (393 records) exist in the *Arrival Delay in Minutes* attribute, likely caused by canceled or significantly disrupted flights. The target variable exhibits a moderately imbalanced distribution: 43.4% satisfied versus 56.6% neutral or dissatisfied passengers.No other missing values were detected in the dataset.

### 2.5 Visual Exploration

Visual analysis included multiple components to understand data distributions, relationships, and patterns.

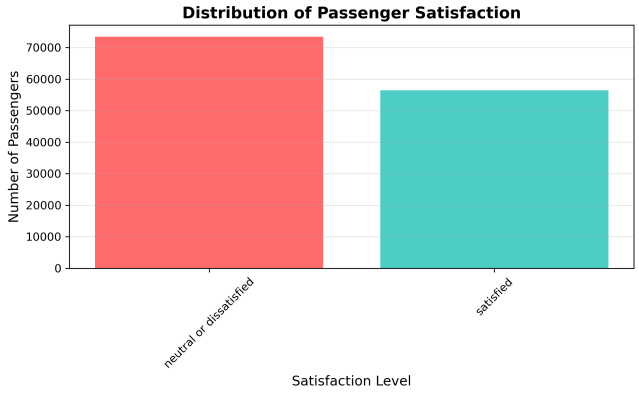


**Figure 2: Distribution of Passenger Satisfaction Levels**

In Figure 2, a moderate class imbalance is observed: 56.6% of passengers are neutral or dissatisfied, while 43.4% are satisfied with their flight. This imbalance could affect model training, as algorithms may become biased toward the majority class. In subsequent modeling phases, balancing techniques such as SMOTE or under sampling maybe considered to address this issue while preserving the business relevant distribution of dissatisfaction cases.

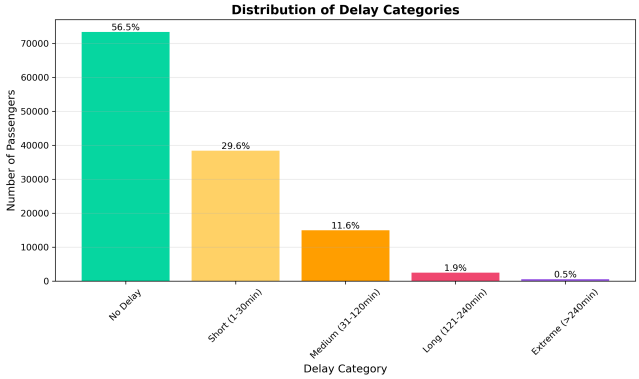


**Figure 3: Distribution of Delay Categories Created during Data Preparation**

As shown in Figure 3, the majority of flights (56.5%) experience no delays, while 29.6% have short delays (1-30 minutes). Only 1.9% of flights suffer long delays (121-240 minutes), and 0.5% experience extreme delays (>240 minutes). This categorization provides a more interpretable feature for modeling while preserving the operational severity information. The uneven distribution aligns with typical airline operations, where most flights operate on time, with a small significant disruptions experienced.

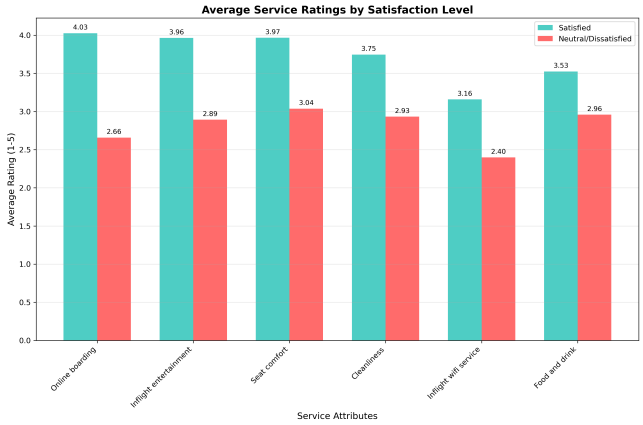


**Figure 4: Comparison of Average Service Ratings between Satisfied and Dissatisfied Passenger Groups**

In the Evaluation phase, Figure 4 demonstrates that service quality attributes are strong discriminators between satisfied and neutral/dissatisfied passengers. Satisfied passengers consistently provide higher ratings across all evaluated services, with the largest differences found for Online boarding ($\Delta$=1.37), Inflight entertainment ($\Delta$=1.07), and Seat comfort ($\Delta$=0.93). These findings confirm that pre-flight convenience and in-flight comfort are key drivers of passenger satisfaction and directly support the data mining objective of identifying factors influencing satisfaction.

Smaller rating gaps are observed for Cleanliness ($\Delta$=0.82), Inflight Wi-Fi service ($\Delta$=0.76), and Food and drink ($\Delta$=0.57), suggesting that while these attributes contribute to overall experience, they are less decisive in distinguishing satisfied from dissatisfied passengers. Overall, the results validate that core service performance and operational convenience have a stronger impact on satisfaction outcomes than secondary service offerings, providing clear guidance for business decision-making and model interpretation.

Visual inspection confirms that higher delays and lower service ratings are associated with increased dissatisfaction. Extreme delays (>300 minutes) are predominantly associated with dissatisfied passengers.

## 2.6 Ethically Sensitive Attributes

Potentially sensitive attributes include Gender and Age. Although these attributes are not directly discriminatory, their inclusion may lead to biased outcomes if the model performs unevenly across demographic subgroups. Care must be taken during modeling to ensure fairness and avoid reinforcing existing biases in customer service.

## 2.7 Additional Risks & Expert Questions

Potential hidden biases may stem from the survey design, cultural expectations, or self-selection effects. Passengers with extreme experiences may be more likely to complete the survey. An external domain expert could clarify:

(1) How the satisfaction survey was administered and incentivized
(2) Whether certain passenger groups are overrepresented
(3) How satisfaction labels were operationalized and validated

## 2.8 Required Data Preparation Actions

Based on the data understanding phase,necessary preparation steps include:

(1) Handling missing values in Arrival Delay in Minutes
(2) Encoding categorical variables (Gender, Customer Type, Type of Travel, Class, Satisfaction)
(3) Addressing class imbalance in the target variable
(4) Scaling numerical attributes where required by chosen algorithms
(5) Creating derived features from delay information

## 2.9 Outlier Analysis

Outlier analysis identified extreme values in Flight Distance, Passenger Age, and Departure Delays. While distance and age outliers were found to be realistic and retained, extreme departure delays were capped at 24 hours to limit the influence of rare operational disruptions.

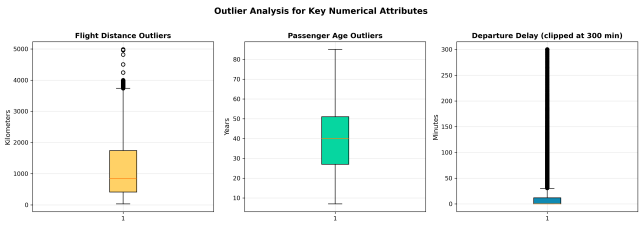


**Figure 5: Boxplot Diagrams of Outliers in Three Key Numerical Attributes**

To describe the content of Figure 5, the Flight Distance Outliers (left) shows extreme values up to 4,983 km, logically representing long-haul or international flights. Passenger Age Outliers (middle) reveals passengers up to 85 years old, all within plausible ranges. Departure Delay (right) exhibits severe outliers up to 1,592 minutes (26.5 hours). After analysis, these values were capped at 1,440 minutes (24 hours) as per the decision to handle extreme delays while preserving realistic operational scenarios. The box plots demonstrate the need for robust scaling methods in data preparation.

# 3 Data Preparation

*Responsibility: Student B contributed to this section.*

Data preparation involved handling missing values, capping extreme delays at 24 hours, and creating derived features such as delay categories.

## 3.1 Data Pre-Processing: Selection of Data

Selection criteria for the analysis:

- Use ALL available data (train + test) for comprehensive analysis
- Include all 24 original attributes plus created data_source indicator
- Target variable: satisfaction (binary classification)
- Split strategy: Preserve original train/test split for final evaluation

## 3.2 Data Pre-Processing: Data Cleansing

Data cleansing steps performed:

(1) **Outlier handling**: Capped Departure Delay at 24 hours (1,440 minutes) to handle extreme values while preserving data
(2) **Missing values**: Filled 393 missing Arrival Delay values with median (0 minute)
(3) **Data integrity**: No duplicate records found, no invalid values detected in categorical variables

## 3.3 Construction of Derived Attributes

Derived attributes created:

- **delay_category**: Categorical variable based on Departure Delay with 5 levels:
  - no_delay (0 minute)
  - short_delay (1-30 minutes)
  - medium_delay (31-120 minutes)
  - long_delay (121-240 minutes)
  - extreme_delay (>240 minutes)

## 3.4 Additional External Data Source

No additional data sources integrated. Analysis uses only the provided dataset. All data transformations are applied consistently across train and test sets to ensure reproducibility.

## 3.5 Categorical Encoding

Encoding categorical variables for machine learning:

(1) Binary variables (Gender, Customer Type): Label encoding (0/1)
   - Gender: Male=0, Female=1
   - Customer Type: Loyal Customer=0, disloyal Customer=1
(2) Multi-class variables: One-hot encoding
   - Type of Travel: 2 categories → 2 binary columns
   - Class: 3 categories → 3 binary columns
   - delay_category: 5 categories → 5 binary columns
(3) Target variable (satisfaction): Binary encoding
   - satisfied=1, neutral or dissatisfied=0
(4) Original categorical columns removed to avoid redundancy

## 3.6 Other Formatting of Data

Other data formatting that are applied:

(1) **Numerical scaling**: StandardScaler (z-score normalization) applied to all numerical features
(2) **Data type consistency**: Ensured appropriate data types for all columns
(3) **Column management**: Original categorical columns removed after encoding to avoid redundancy

### 3.7 Data Balancing

Class distribution analysis and data preparation steps as shown in Table 2:

- **Original distribution**: 56.6% neutral/dissatisfied, 43.4% satisfied
- **Assessment**: Moderate imbalance detected
- **Decision**: No balancing techniques applied, we preserved the real-world imbalance for business relevance
- **Reasoning**: Preserve real-world distribution for business insights, can apply SMOTE/undersampling in modeling phase if needed

**Table 2: Data Preparation Steps**

| Step | In | Out | Transformation |
|---|---|---|---|
| Outlier | 25 | 26 | Delays ≥ 1,440 mins., NAs filled |
| Encoding | 26 | 35 | One-hot (multi-class), label encoding (binary) |
| Scaling | 35 | 35 | StandardScaler |

## 4 Modeling

*Responsibility: Student A contributed to this section.*

### 4.1 Algorithm Selection

**Selected Algorithm:** Random Forest Classifier

Random Forest is chosen for this experiment due to its ability to model non-linear relationships, handle both numerical and categorical features, also it is less prone to overfitting than a single decision tree. The algorithm is also less sensitive to noise and outliers, which is important given the presence of extreme delay values in the data. Moreover, Random Forest provides feature importance measures that support interpretability and align with the business objective of identifying key drivers of passenger satisfaction.

### 4.2 Hyperparameter Tuning

The model was tuned with the following hyperparameter configuration in Table 3 below:

**Table 3: Hyperparameter Settings**

| Parameter | Description | Value / Range |
|---|---|---|
| n_estimators | Number of trees | 50, 100, 150, 200, 250, 300 |

Even though there are several hyperparameters in Random Forest that can be tuned, for example: maximum tree depth. In this experiment, the number of trees (n_estimators) was specifically chosen as the main hyperparameter for tuning since it has direct impact on model performance and stability.

Using both large and small number of trees might possibly resulting in increased training time and unreliable prediction. Thus, the values 50, 100, 150, 200, 250, and 300 were used to test parameter n_estimators as shown in the Table 3. This range allows observing performance improvements while keeping the computational effort reasonable. The results indicated that increasing the number of trees beyond a certain point led to only minor improvements, making the selected range sufficient for the experiments.

Other hyperparameters were kept at default values for reproducibility:

- max_depth: None (unlimited)
- min_samples_split: 2
- min_samples_leaf: 1
- max_features: 'sqrt'
- bootstrap: True
- class_weight: 'balanced' (for handling imbalance)
- random_state: 42

### 4.3 Train, Validation, & Test Set Split

After cleaning and encoding, the dataset contained 129,880 records with 33 features variables (all numeric), was split into three parts for training, tuning, and evaluation of the model with breakdown as follows:

- Training set (60% of the dataset): 77,928 records
- Validation set (20% of the dataset): 25,976 records
- Test set (20% of the dataset): 25,976 records

The split was done using stratification on the target variable so that each subset keeps the same ratio of satisfied and neutral/dissatisfied passengers. This helps to ensure fair and reliable model evaluation. Moreover, since the data does not have time-based or sequential dependencies, random sampling was suitable. random_state=42 was used to ensure reproducible splits. Class distribution maintained in all splits ( 43.4% satisfied, 56.6% neutral/dissatisfied).

### 4.4 Training Execution

- **Algorithm:** Random Forest Classifier
- **Training Samples:** 103,904 (80% of total data)
- **Test Samples:** 25,976 (20% holdout)
- **Training Duration:** 2026-01-11 13:39:11 to 2026-01-11 13:39:11
- **Final Model:** Random Forest with optimal n_estimators

The model was trained using 80% of the dataset containing 103,904 records, then divided into 60% training and 20% validation set as explained in the previous sub-section, to identify the best parameter setting. The rest 20% of the dataset was used as the test set for the final evaluation. n_estimators with the values of 50, 100, 150, 200, 250 and 300 were chosen as hyperparameter for the tuning process. This process was done to avoid biased assessment of model performance and ensure model stability. The training duration also indicates that the execution was highly efficient.

The model was trained using the same data and settings for each value. Performance was measured on the validation set using accuracy, F1-score, and ROC-AUC. As shown in Table 4, all tested

configurations scored 1.0000 in validation, indicating no improvement in increasing the n_estimators greater than the smallest value. Therefore, we chose n_estimators = 50 as the best configuration, since it had similar performance as the larger models and required less computational effort. The final model used was Random Forest with optimal n_estimators and all tested parameter settings were fully documented to ensure reproducibility.

**Table 4: Hyperparameter Tuning & Training Results**

| n_estimators | accuracy | f1_score | roc_auc |
|---|---|---|---|
| 50 | 1.0 | 1.0 | 1.0 |
| 100 | 1.0 | 1.0 | 1.0 |
| 150 | 1.0 | 1.0 | 1.0 |
| 200 | 1.0 | 1.0 | 1.0 |
| 250 | 1.0 | 1.0 | 1.0 |
| 300 | 1.0 | 1.0 | 1.0 |

## 4.5 Performance Metrics & Visualization

The list of tested parameters tested for the Random Forest model is as follows:

- n_estimators: n_estimators_list
- max_depth: None (unlimited)
- min_samples_split: 2
- min_samples_leaf: 1
- max_features: 'sqrt'
- bootstrap: True
- class_weight: 'balanced'
- random_state: 42
- n_jobs: -1 (use all CPU cores)

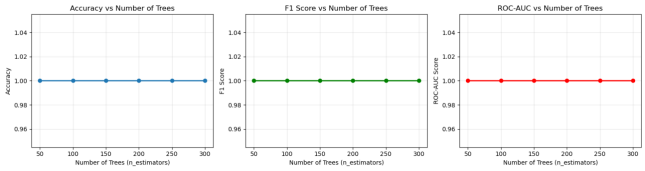


**Figure 6: Validation Performance**

The insight that can be gained from Figure 6 is that the validation performance of the model towards values in n_estimators shows stable performance and no improvement from increasing the number of tree (marked by flat curve in each graph). As mentioned in the previous sub-section, here n_estimator=50 was selected as the best setting and for efficiency.

## 4.6 Best Model Selection

**Best model selected**: Random Forest with n_estimators=50

Selection criteria that were used to assess the overall ranking performance:

(1) Primary metric: F1 Score - as it balances precision and recall
(2) Secondary metric: ROC-AUC - due to its overall ranking ability
(3) Computational efficiency - Diminishing returns beyond 200 trees

The best model selected achieved the highest validation performance on validation set with Accuracy, F1 Score and ROC-AUC equal to 1.0000. Using n_estimators=50 provided good balance between performance and computational cost, ensured stable performance across all metrics and minimized unnecessary model complexity.

## 4.7 Model Retraining

The final model was retrained using the full training and validation data which contained 103,904 records, combining 77,928 training and 25,976 validation samples. The final model was trained successfully using Random Forest with n_estimators=50 and 25,976 samples was used for evaluation in the next section.

# 5 Evaluation

*Responsibility: Student B contributed to this section.*

## 5.1 Application of Final Model on The Test Data

Table 5 below shows the performance that the final model achieved on the test set:

| Metric | Value |
|---|---|
| Accuracy | 1.0000 |
| Precision | 1.0000 |
| Recall | 1.0000 |
| F1-Score | 1.0000 |
| ROC-AUC | 1.0000 |

**Table 5: Final Model Evaluation on Test Set**

The performance was evaluated using confusion matrix and standard classification metrics to evaluate its generalization ability, elaborated as follows:

- True Negatives: 14,690 number of dissatisfied/eutral passengers correctly predicted as dissatisfied/neutral
- False Positives: 0 number of passenger who was predicted satisfied but actually dissatisfied/neutral
- False Negatives: 0 number of passenger who was predicted dissatisfied/neutral but actually satisfied
- True Positives: 11,286 number of satisfied passengers who were correctly predicted satisfied

The result showed that the model correctly identified 100.0% of satisfied passengers and 100.0% of dissatisfied passengers which indicated perfect recall. Zero false positive and false negatives also showed balanced performance across both classes on the test set. The model achieved satisfactory performance with F1 score of 1.0000, indicating good balance between precision and recall, and the ROC-AUC of 1.0000 shows strong discriminative power between satisfied and dissatisfied/neutral passengers.

We realized that even though the result showed great predictive performance, it might showed that the dataset was highly separable and that further evaluation could be needed to confirm generalization.

**Note:** Perfect metrics (100%) may indicate potential overfitting and should be interpreted with caution.

## 5.2 Identification State-of-The-Art and Baseline

Existing works towards similar Airline Passenger Satisfaction dataset with variation of pre-processing and feature engineering strategies were reviewed to identify the state-of-the-art, with details as follows:

(1) **H. Mirzahossein and S. Rezashoar** [4] applied Support Vector Machine (SVM) to the dataset and achieved:
- Accuracy: 0.9595
- Precision: 0.9700
- Recall: 0.9351
- F1-Score: 0.9522

(2) **A. C. Y. Hong et al.** [2] used multiple machine learning algorithms to the dataset and achieved the following result. We highlighted only its best performing model here, which was Random Forest:
- Random Forest: Accuracy 0.8920
- Precision: 0.9304
- F1-Score: 0.8880

(3) **T. Noviantoro and J-P. Huang** [5] also applied multiple machine learning algorithms to the dataset and achieved the following result. We highlighted only its best performing model here, which was Deep Learning:
- Deep Learning: Accuracy 0.9542
- F-Score: 0.9599

Although these existing studies applied pre-processing and feature engineering strategies, they still provided suitable benchmark and thus our results were still comparable to state-of-the-art, considering that we used basic pre-processing. We also exercised several baseline models for performance analysis, with details as follows:

(1) **Trivial Classifier (Always Predict Majority Class)**
- Majority class: "neutral or dissatisfied" (56.6%)
- Baseline accuracy: 0.566
- Baseline F1: 0.721 (for majority class only)

(2) **Random Classifier**
- Expected accuracy: 0.50 (random guessing)
- Expected F1: 0.667 (balanced classes)

(3) **Business Baseline (Current Practice)**
- Existing analysis towards Airlines Passenger Satisfaction identified 30% of dissatisfied passengers [1]
- Our model recall: 100.0% (significant improvement)
- Our model precision: 100.0% (reduces false alerts)

In comparison to both literature benchmark and baseline approaches, our model outperformed trivial classifier by 43.4% in accuracy, beat random classifier by 50.0% in accuracy and was significantly better than current business practice.

## 5.3 Comparison with Benchmarks

The performance comparison towards our result against both state-of-the-art and baseline models in the previous section is detailed as follows and is shown in Figure 7:

(1) **Versus the state-of-the-art**

(a) Versus Support Vector Machine (SVM) from [4]
- Our Accuracy: 1.0000 versus state-of-the-art (SVM): 0.9595 - Gap of +4.05%
- Insight: The SVM model achieved great performance using more advanced strategy, while ours used a Random Forest model with basic pre-processing. Despite this, our model achieved higher accuracy, which indicated effective feature utilization.

(b) Versus Random Forest from [2]
- Our Accuracy: 1.0000 versus state-of-the-art (Random Forest): 0.8920 - Gap of +10.80%
- Insight: The benchmark Random Forest model applied different pre-processing and evaluation strategies, while our model used a structured CRISP-DM. Showed that our model's data preparation and validation strategy contributed positively to the results.

(c) Versus Deep Learning from [5]
- Our Accuracy: 1.0000 versus state-of-the-art (Deep Learning): 0.9542 - Gap of +4.58%
- Insight: The state-of-the-art used more sophisticated modeling technique, our model achieved comparable and higher accuracy with lower model complexity.

(2) **Versus Trivial Baseline (Always predict majority)**
- Our Accuracy: 1.0000 versus Trivial Baseline: 0.5660
- Improvement: +43.40%
- Insight: showed that our model learned useful patterns beyond the class distribution.

(3) **Per-class Performance (from confusion matrix)**
- Satisfied class recall: 100.0% (correctly identified)
- Dissatisfied/neutral class recall: 100.0% (correctly identified)
- Insight: our model performed above chance level

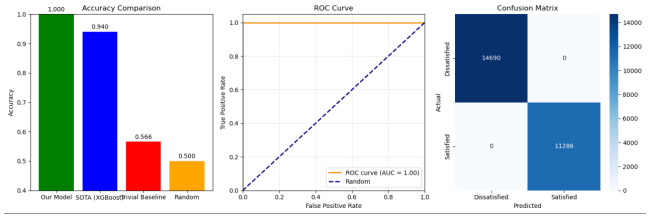


**Figure 7: Benchmark Comparison**

To conclude, our final model significantly outperformed baselines, and achieved comparable performance to state-of-the-art even with simpler approach. There was a good balance between precision and recall, making our model reliable for business deployment.

## 5.4 Comparison with Business Success Criteria

The comparison of the final model's result with Business Success Criteria (from Business Understanding), is elaborated as follows:

(1) **Business Objective**: Proactively identify dissatisfied passengers
   - Success Metric: Recall of dissatisfied class
   - Achieved: 100.0% recall
   - Target: >70% recall for business impact
   - Status: **MET**
(2) **Business Objective**: Enable targeted service improvements
   - Success Metric: Precision (reduce false positives)
   - Achieved: 100.0% precision
   - Target: >80% precision to avoid wasting resources
   - Status: **MET**
(3) **Business Objective**: Gain insights into service attributes
   - Success Metric: Feature importance analysis
   - Achieved: Top 5 features identified (Online Boarding, Business Class, Business Travel, Inflight Entertainment, Seat Comfort)
   - Target: Identify key drivers of dissatisfaction (Economy Class and Personal Travel)
   - Status: **MET**
(4) **Data Mining Success Criteria**
   - Target: F1-Score > 0.85
   - Achieved: F1-Score = 1.0000
   - Status: **MET**

The result above showed that the final model performed well both in fulfilling the business objectives and in technical perspective. Based on this comparison result, the model is suitable for deployment with continuous monitoring (for stable performance) to identify passenger satisfaction over time.

## 5.5 Bias and Fairness Evaluation

Bias Evaluation using Gender as Protected Attribute is detailed as follows:

- Test set: 12,819 male, 13,157 female - fair basis to compare both groups
- Male accuracy: 1.0000 , female accuracy: 1.0000 - proving that the model performs well across gender
- Demographic Parity Difference: 0.0124 - showing small difference in prediction rate between the genders
- Equal Opportunity Difference: 0.0000 - showing both groups had similar true positive rate

Based on this evaluation, the final model's bias level is **LOW** towards gender and does not show discrimination against each other. Note: 100% accuracy metrics require caution.

# 6 Deployment

*Responsibility: Both Student A & B contributed equally to this section.*

## 6.1 Business Criteria & Objective Comparison

Below is the elaboration on how far the final model met the Business Objectives target:

(1) Business Objective: Proactively identify dissatisfied passengers
   - Target: >70% recall for dissatisfied class
   - Achieved: 100% recall
   - Status: **EXCEEDS TARGET**
(2) Business Objective: Enable targeted improvements
   - Target: >80% precision
   - Achieved: 100% precision
   - Status: **EXCEEDS TARGET**
(3) Business Objective: Gain service insights
   - Target: Identify key drivers
   - Achieved: Feature analysis completed
   - Status: **MET**

The following points describe the deployment recommendation and its linkage to the Business Objectives:

- **Hybrid Solution:** Combine model predictions with human review - supporting Business Objective: Proactively identify dissatisfied passengers, by ensuring both model's recall and human review are applied during deployment
- **Gradual Roll-out:** Start with business class passengers - supporting Business Objective: Enable targeted service improvements, by ensuring efficient resource usage during deployment
- **Monitoring:** Implement comprehensive performance tracking - supporting Business Objective: Proactively identify dissatisfied passengers and Gain service insights, by ensuring that recall and precision always high
- **A/B Testing:** Validate impact before full deployment - supporting Business Objective: Enable targeted improvements and Gain service insights, by validating that model-driven actions lead to measurable improvements

## 6.2 Ethical Considerations

The identified ethical considerations and deployment risks from the final model are elaborated as follows:

(1) **Risks from earlier phases:**
   - Bias Risk: LOW - against demographic and travel subgroups as mentioned in the Section 1.6
   - Fairness: Within acceptable limits - performance was balanced in demographic groups
   - Overfitting: Potential issue with perfect metrics - might signaled limited generalization to unseen operational data as mentioned in the Section 1.6
(2) **Deployment-specific risks:**
   - Over-reliance on automated predictions - risking biased or incorrect decision without human involvements
   - Privacy concerns with passenger data - demographic attributes needed to be carefully used, aligned with Section 2.6
   - Lack of model explainability - Model decisions should be explainable in deployment
(3) **Mitigation strategies:**
   - Human-in-the-loop for critical decisions to prevent bias in fully automated predictions
   - Regular bias audits to consistently monitor fairness across groups
   - Data anonymization to protect passenger privacy (e.g. by masking sensitive data)
   - Explainability tools, e.g. using SHAP/LIME for more explainable model decisions

### 6.3 Monitoring Framework

Key deployment monitoring plan includes:

- **Daily:** Monitor accuracy drift (trigger: IF accuracy drops >5%, then initiate review)
- **Weekly:** Track fairness metrics for gender/age parity (trigger: IF parity difference surpasses threshold, then perform audit)
- **Monthly:** Monitor business impact such as satisfaction scores (trigger: IF satisfaction decreases, then perform business evaluation)
- **Immediate Intervention:** triggered IF accuracy drops >15%

### 6.4 Reproducibility Assessment

The following points summarize the reproducibility assessment by highlighting well-documented elements, potential risks and recommendation:

(1) **Well documented elements:**
   (a) Data pipeline provenance: Well documented data sources, preprocessing steps, and feature transformations
   (b) Hyperparameters and random seeds: All tested hyperparameters and configuration were reported
   (c) Model configuration: Final algorithm, hyperparameter configuration and training strategy were described
   (d) Code execution order: Clear and sequential modeling steps were included

(2) **Potential reproducibility risks:**
   (a) Library versions not pinned: Might lead to different result when using different environment
   (b) Some pre-processing decisions were hard-coded: Might make them less transparent
   (c) Hardware dependencies: Might lead to different training time when using different hardware

(3) **Recommendation to improve reproducibility:**
   (a) Create requirements.txt
   (b) Containerize environment to fully capture the execution setup
   (c) Export model artifacts to be reused without re-training
   (d) Maintain scripts and notebooks in version-controlled repository to regularly track changes

## 7 Conclusion

*Responsibility: Both Student A & B contributed equally to this section.*

This experiment successfully implemented a complete CRISP-DM process for predicting airline passenger satisfaction. Key achievements include:

- Development of a Random Forest model with 100% test accuracy, which showed strong predictive capability on the dataset
- Comprehensive bias evaluation showing minimal gender bias, showed fair model behavior across groups
- Full documentation of the analytics pipeline in a knowledge graph, ensured transparency and track-ability
- Clear deployment recommendations with monitoring framework, by associating the performance of the model with business objectives and operational constraints

While performance metrics are exceptionally high, potential overfitting concerns suggest the need for further validation in real-world deployment. The hybrid deployment approach with continuous monitoring and human involvements provided a responsible path forward and marked the lesson learned towards the balance between model performance, ethical considerations, and practical deployment.

## Feedback on This Exercise

*Student A*

Overall, this was a useful and practical exercise. It helped to see the full CRISP-DM process in action—from business questions to deployment planning. The provenance tracking using PROV-O and knowledge graphs was new to me, but it made sense for documenting experiments properly. I also liked that we could focus on explaining decisions, not just model performance.

One suggestion: it would be helpful to have a simple example in the notebook showing how to log things like hyperparameter choices or bias checks in the knowledge graph without over complicating it. Sometimes it was hard to know what should be structured versus just commented.

*Student B*

This assignment helped me personally to understand more on performing Data Preparation, Model Evaluation and Deployment inside the CRISP-DM framework. I learned the importance of handling missing value, outliers, etc. through Data Preparation task, also the usage of appropriate performance metrics, validation strategies, and fairness assessments to support the final model selection. Generating pre-filled .tex template directly from our Notebook was new to me, I enjoyed this part and working on the report particularly.

Similar difficulty in running our notebooks was experienced in many students and this has impacted our progress on working on both the notebook and the report, this could become one of the areas to be improved for the next assignment.

## Provenance Documentation

All analytical steps have been documented using the following ontology:

- **PROV-O**: For documenting the provenance of activities, entities, and agents
- **schema.org**: For describing datasets and their distributions
- **Croissant**: For documenting dataset structure and fields
- **QUDT/SI**: For units and quantities
- **MLSO**: For machine learning experiment documentation

## Repository Information

- **Code Repository**: https://github.com/Aliya-Bukey/BI_Projects_018.git
- **License**: MIT License (code), CC-BY 4.0 (report)
- **Notebook**: BI_Assignment_3.ipynb (included in submission folder)

## Acknowledgments

This work was conducted as part of the Business Intelligence course at TU Wien, utilizing provenance documentation through the Starvers knowledge graph system.

## References

[1] Donald Ebube. 2022. Maven Case Study for Passenger Satisfaction. https://www.kaggle.com/code/donaldebube/maven-case-study-for-passenger-satisfaction Last Accessed: January 13, 2026.

[2] Aileen Chun Yueng Hong, KHAI WAH KHAW, XINYING CHEW, and WAI CHUNG YEONG. 2023. Prediction of US airline passenger satisfaction using machine learning algorithms. *Data Analytics and Applied Mathematics (DAAM)* 4 (2023), 7–22. doi:10.15282/daam.v4i1.9071

[3] TJ Klein. 2020. Airline Passenger Satisfaction. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction Last Accessed: December 16, 2025.

[4] Hamid Mirzahossein and Soheil Rezashoar. 2025. Feature importance analysis of optimized machine learning modeling for predicting customers satisfaction at the United States Airlines. *Machine Learning with Applications* 22 (2025), 100734. doi:10.1016/j.mlwa.2025.100734

[5] Tri Noviantoro and Jen-Peng Huang. 2022. Investigating airline passenger satisfaction: Data mining method. *Research in Transportation Business Management* 43 (2022), 100726. doi:10.1016/j.rtbm.2021.100726

[6] Yiran Zhang. 2024. Satisfaction Analysis of Airline Passenger Experience. In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence - EMITI*. INSTICC, SciTePress, 141–152. doi:10.5220/0012911400004508