Task 3: Model Building, Evaluation & Selection

# Objective

To predict whether a flight will be cancelled or not using different machine learning models and select the best one.

# Step 1: Models Chosen

We selected the following classification models:  
- Logistic Regression (baseline)  
- Decision Tree Classifier  
- Random Forest Classifier  
- Support Vector Machine (SVM)  
  
These models are suitable for binary classification problems such as flight cancellation prediction.

# Step 2: Training & Evaluation

All models were trained using the same training-test split (80% training, 20% testing). We used the following evaluation metrics:  
- Accuracy  
- Precision  
- Recall  
- F1-Score

# Step 3: Performance Comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1-Score |
| Logistic Regression | 0.82 | 0.85 | 0.89 | 0.87 |
| Decision Tree | 0.96 | 0.96 | 0.98 | 0.97 |
| Random Forest | 0.98 | 0.99 | 0.98 | 0.98 |
| SVM (Linear) | 0.85 | 0.86 | 0.90 | 0.88 |

# Step 4: Pros & Cons of Each Model

Comparison of different models:

|  |  |  |
| --- | --- | --- |
| Model | Pros | Cons |
| Logistic Regression | Simple, easy to interpret | Lower performance |
| Decision Tree | Easy to visualize and explain | May overfit on small datasets |
| Random Forest | High accuracy, handles overfitting | Harder to interpret, slower than tree |
| SVM | Performs well in high-dimensional space | Sensitive to parameter tuning, slower on large data |

# Step 5: Best Model Recommendation

Random Forest Classifier is the best model:  
- Highest accuracy (98%)  
- Best balance of precision, recall, and F1-score  
- Robust and generalizes well  
  
Therefore, we recommend using Random Forest to predict flight cancellations for Flyzy.