Airline Passenger Satisfaction Prediction Using Machine Learning Models.

Introduction

Air travel has become one of the most critical industries in global transportation, where passenger satisfaction directly affects airline reputation and competitiveness. Understanding the factors that influence passenger satisfaction allows airlines to improve service quality, customer experience, and operational efficiency.

The provided dataset contains results from an airline passenger satisfaction survey, including demographic details, travel class, flight distance, service ratings, and delay times. This study aims to analyse the dataset and develop Artificial Intelligence (AI) models that can predict whether a passenger is satisfied or dissatisfied based on these factors.

The work involves several stages: loading and cleaning the data, handling missing values, encoding categorical variables, and preparing the data for machine-learning models. Four supervised learning models Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM) were implemented using the training and testing datasets provided. Each model's performance was compared using accuracy, precision, and the confusion matrix metrics to identify the most effective classifier for predicting passenger satisfaction.

Dataset Summary

	Training Dataset	Testing Dataset
Dataset Shape	(103904, 25)	(25976,25)
Dataset features type	id - int64 Gender - object Customer Type - object Age - int64 Type of Travel - object Class - object Flight Distance - int64 Inflight wifi service - int64 Departure/Arrival time -int64 Ease of Online booking -int64 Gate location - int64 Food and drink - int64 Online boarding - int64 Seat comfort - int64 Inflight entertainment - int64 On-board service - int64 Leg room service - int64 Baggage handling - int64 Checkin service - int64 Inflight service - int64 Cleanliness - int64 Departure Delay in Minutes - int64 Arrival Delay in Minutes - float64 satisfaction - object	id - int64 Gender - object Customer Type - object Age - int64 Type of Travel - object Class - object Flight Distance - int64 Inflight wifi service - int64 Departure/Arrival time -int64 Ease of Online booking -int64 Gate location - int64 Food and drink - int64 Online boarding - int64 Seat comfort - int64 Inflight entertainment - int64 On-board service - int64 Leg room service - int64 Baggage handling - int64 Checkin service - int64 Inflight service - int64 Cleanliness - int64 Departure Delay in Minutes - int64 Arrival Delay in Minutes - float64 satisfaction - object
Missing values	Arrival Delay in Minutes 310	Arrival Delay in Minutes 83
Dataset shape after drop missing values	(103594, 25)	(25893,25)
Encoding data columns	Gender - {'Male': 1, 'Female': 0} Customer Type - {'Loyal Customer': 1, 'disloyal Customer': 0} Type of Travel - {'Personal Travel': 1, 'Business travel': 0} Class - {'Eco Plus': 1, 'Business': 0, 'Eco':2 } satisfaction - {'satisfied': 1, 'neutral or dissatisfied': 0}	

Model Summary

Logistic Regression

Training accuracy

Logistic Regression Results

Accuracy : 0.798

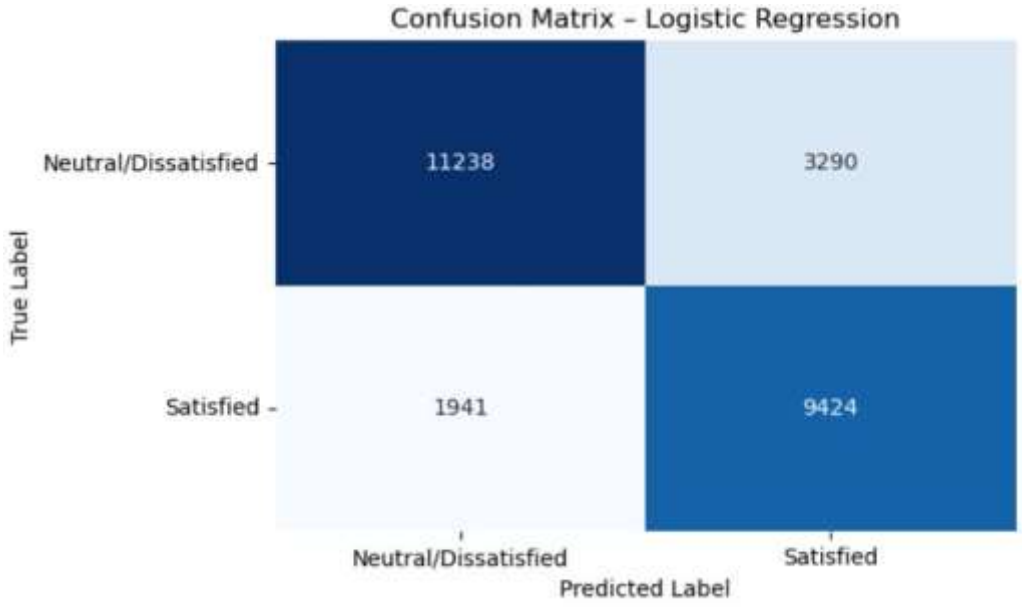
Precision: 0.7412

Testing accuracy and other Parameters

Classification Report:

	precision	recall	f1-score	support
0	0.85	0.77	0.81	14528
1	0.74	0.83	0.78	11365
accuracy			0.80	25893
macro avg	0.80	0.80	0.80	25893
weighted avg	0.80	0.80	0.80	25893

Confusion Matrix – Logistic Regression



Decision tree

Training accuracy

Decision tree Results

Accuracy : 0.95

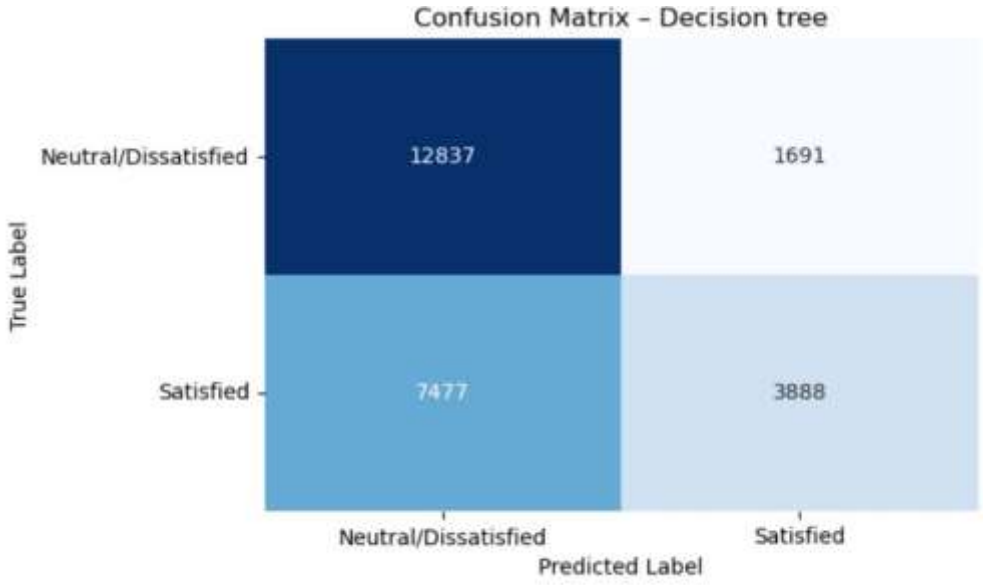
Precision: 0.9449

Testing accuracy and other Parameters

Classification Report:

	precision	recall	f1-score	support
0	0.95	0.96	0.96	14528
1	0.94	0.94	0.94	11365
accuracy			0.95	25893
macro avg	0.95	0.95	0.95	25893
weighted avg	0.95	0.95	0.95	25893

Confusion Matrix – Decision tree



Random Forest

Training accuracy

Random Forest Results

Accuracy : 0.9621

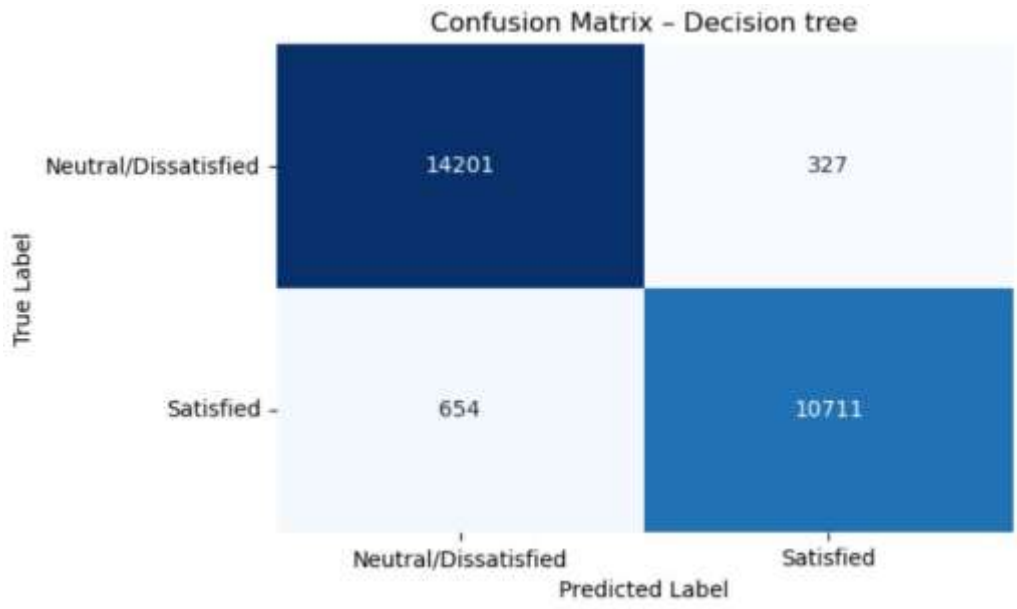
Precision: 0.9704

Testing accuracy and other Parameters

Classification Report:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	14528
1	0.97	0.94	0.96	11365
accuracy			0.96	25893
macro avg	0.96	0.96	0.96	25893
weighted avg	0.96	0.96	0.96	25893

Confusion Matrix – Decision tree



SVM Model

Training accuracy

SVM Results

Accuracy : 0.6459

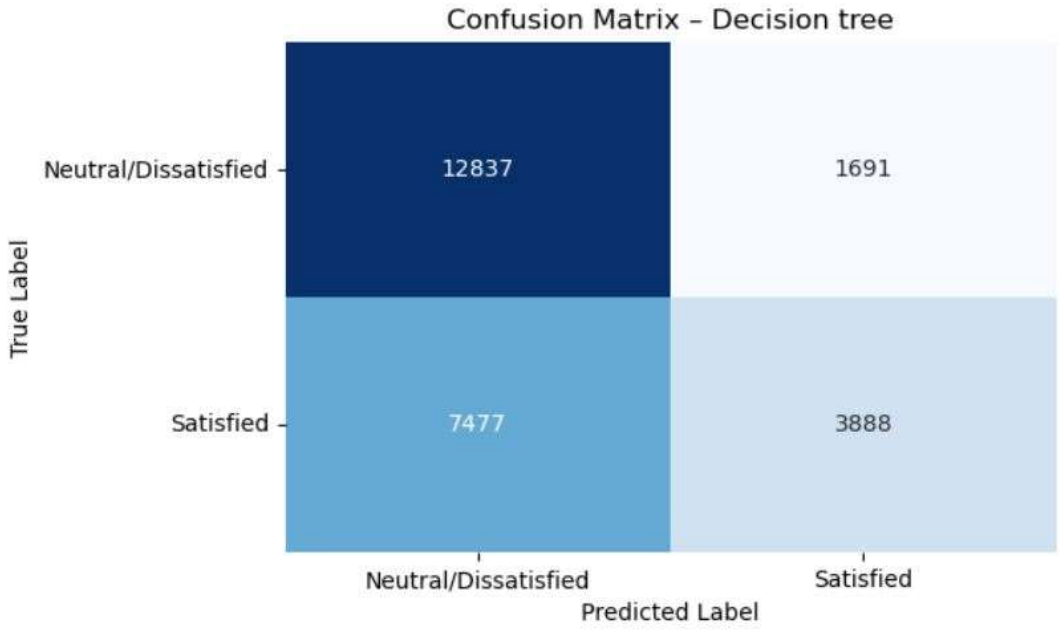
Precision: 0.6969

Testing accuracy and other Parameters

Classification Report:

	precision	recall	f1-score	support
0	0.63	0.88	0.74	14528
1	0.70	0.34	0.46	11365
accuracy			0.65	25893
macro avg	0.66	0.61	0.60	25893
weighted avg	0.66	0.65	0.61	25893

Confusion Matrix – Decision tree



Conclusion

After data preprocessing and model development, the performance of all four AI models was evaluated.

- The Logistic Regression model achieved moderate accuracy (≈ 0.80), serving as a simple linear baseline.
- The Decision Tree improved prediction accuracy to around 0.95 by capturing non-linear relationships.
- The Random Forest model delivered the highest accuracy (0.9621) and precision (0.9704), confirming that ensemble learning provides robust and reliable results by combining multiple decision trees.
- The SVM model produced lower accuracy (≈ 0.65), likely due to scale sensitivity and the dataset's high dimensionality.

Overall, the Random Forest classifier was identified as the most effective model for predicting airline passenger satisfaction. Key influencing factors included online boarding, inflight Wi-Fi service, travel class, and type of travel.
This analysis demonstrates how data driven AI models can provide valuable insights into customer experience, enabling airlines to enhance service quality and maintain higher levels of passenger satisfaction in real-world operations.