Vivekanand Education Society's Institute of Technology

Hashu Advani memorial Complex Collector's Colony R C Marg, Chembur, Mumbai 400074

DEPARTMENT OF INFORMATION TECHNOLOGY

MINI PROJECT REPORT ON

"Airline Passenger Satisfaction Prediction"

T.E. (Information Technology)

SUBMITTED BY

Mr. Nishant S Khetal (24)
Mr. Atharv S Nikam (36)
Mr. Pratik M Patil (40)

UNDER THE GUIDANCE OF

Dr Ravita Mishra

(Academic Year: 2024-2025)

Mumbai University

Vivekanand Education Society's Institute Of Technology, Mumbai

DEPARTMENT OF INFORMATION TECHNOLOGY

Certificate

This is to certify that project entitled

"Airline Passenger Satisfaction Prediction"

Mr. Nishant S Khetal (24)
Mr. Atharv S Nikam (36)
Mr. Pratik M Patil (40)

have satisfactorily carried out the project work, under the head - DS Using Python Lab at Semester VI of TE-IT in Information Technology as prescribed by the Mumbai University.

Prof. Guide Name
Dr Ravita Mishra

External Examiner

Dr.(Mrs.) Shalu Chopra
H.O.D

Dr.(Mrs.) J.M.Nair
Principal

Date: / /2025
Place: VESIT, Chembur

LO Mapping

LO1: Understand the concept of Data science process and associated terminologies to solve real-world problems

LO2: Analyze the data using different statistical techniques and visualize the out- come using different types of plots.

LO3: Analyze and apply the supervised machine learning techniques like Classifi- cation, Regression or Support Vector Machine on data for building the models of data and solve the problems.

LO4: Apply the different unsupervised machine learning algorithms like Clustering or Association to solve the problems.

LO5: Design and Build an application that performs exploratory data analysis us- ing Apache Spark.

LO6: Design and develop a data science application that can have data acquisi- tion, processing, visualization and statistical analysis methods with supported machine learning technique to solve the real-world problem

Declaration

I declare that this written submission represents my ideas in my own words and where other's ideas or words have been included, I have adequately cited and ref- erenced the original source. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Mr. Nishant S Khetal (24) - - - - - - - - - - -

Mr. Atharv S Nikam (36) - - - - - - - - - - -

Mr. Pratik M Patil (40) - - - - - - - - - - -

(Signature)

Contents

List of Figures

List of Tables

Abstract

Airlines are constantly striving to enhance passenger satisfaction to retain customers and maintain competitiveness. This project aims to build a machine learning-based predictive system that can determine airline passenger satisfaction using various parameters such as seat comfort, inflight entertainment, food quality, flight delays, and customer service. The system performs data preprocessing, exploratory data analysis, and applies multiple classification algorithms to predict satisfaction. Among the models tested, **Random Forest** achieved the highest accuracy of **96.18%**. The insights derived from the data can help airlines make data-driven decisions to improve overall customer experience.

Keywords – Airline satisfaction, Machine Learning, Logistic Regression, Random Forest, KNN, Naive Bayes, Data Science, Prediction Model

Chapter 1
Introduction

1.1 Introduction

In the competitive aviation industry, customer satisfaction plays a critical role in influencing brand loyalty and revenue. Several variables affect a passenger's travel experience – including check-in service, inflight service, seat comfort, food, and punctuality. A reliable predictive model that accurately gauges satisfaction can help airlines identify pain points and take corrective actions, ultimately improving service quality and customer retention.

1.2 Literature Survey

The main objective of the research paper [1] is to develop a predictive model for airline passenger satisfaction using various machine learning algorithms. This paper is divided into four sections: (i) Data Collection from airline survey responses, (ii) Preprocessing and Feature Selection, (iii) Model Training and Evaluation using classifiers like Random Forest and Logistic Regression, and (iv) Performance Comparison of the models. The results showed that the Random Forest algorithm performed best due to its ability to handle non-linear relationships and feature importance analysis.

The paper [2] proposes a novel hybrid approach that combines Genetic Algorithms and AutoEncoders with machine learning models to improve the accuracy of passenger satisfaction prediction. It discusses the limitations of using individual models and highlights the advantage of ensemble and optimization-based techniques. The study includes a comprehensive analysis of real-world airline satisfaction datasets, emphasizing performance metrics like accuracy and recall. The proposed system showed significant improvement in classifying satisfied and dissatisfied passengers.

1.3 Problem Definition

Airlines often fail to identify the key drivers of customer dissatisfaction due to lack of structured analysis. This project aims to develop a machine learning model that classifies passengers as "satisfied" or "dissatisfied" based on historical feedback data. The goal is to extract actionable insights that can help airlines enhance service quality and reduce churn.

1.4 Objectives

Build a classification model to predict passenger satisfaction.
Perform thorough EDA to understand key satisfaction drivers.
Compare various ML algorithms and identify the best-performing model.
Provide insights into the most important features affecting satisfaction.

1.5 Proposed Solution

We approach this as a **binary classification problem** using supervised learning. The process includes:

1. Preprocessing and cleaning the dataset (handling missing values, encoding, normalization).
2. Performing EDA using plots and statistical summaries.
3. Applying classification algorithms: **Logistic Regression**, **KNN**, **Naive Bayes**, and **Random Forest**.
4. Choosing the model with the highest accuracy – **Random Forest** (96.18%).

1.6 Technology Used

Programming Language: Python
Libraries: pandas, numpy, seaborn, matplotlib, sklearn
Platform: Google Colab
Dataset: Airline Satisfaction Dataset (from Kaggle)

Chapter 2
Pre-processing

2.1 Dataset Description

The dataset contains feedback and demographic information of passengers, including:

- Gender
- Customer Type (Loyal / Disloyal)
- Type of Travel (Business / Personal)
- Inflight Services (Rating scale: 0–5)
- Satisfaction (Target variable: Satisfied / Dissatisfied)

Total Records: ~100,000
Target: satisfaction (Binary classification)

2.2 Handling Missing Data

Checked for null values using df.isnull().sum()
Minimal missing values were handled using:

- **Mode** for categorical features
- **Median** for continuous features

2.3 Handling Outliers

1. Outliers detected using **IQR method**
2. Outliers retained to preserve real-world variance
3. Feature distributions were visualized using boxplots

2.4 Feature Encoding & Scaling

- Categorical variables converted using LabelEncoder or get_dummies

- Scaling applied using StandardScaler for algorithms like KNN and Logistic Regression

Chapter 3
EDA and Visualization

3.1 Measures of Central Tendency

	Gender	Customer Type	Age	Type of Travel	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort
count	25976.000000	25976.000000	2.597600e+04	25976.000000	2.597600e+04	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000	25976.000000
mean	0.492917	0.184747	1.113300e-16	0.305590	8.931021e-17	2.724746	3.046812	2.756775	2.977094	3.215353	3.261665	3.449222
std	0.499959	0.388100	1.000019e+00	0.460666	1.000019e+00	1.335384	1.533371	1.412951	1.282133	1.331506	1.355536	1.320090
min	0.000000	0.000000	-2.155276e+00	0.000000	-1.164343e+00	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	1.000000
25%	0.000000	0.000000	-8.338705e-01	0.000000	-7.808310e-01	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000
50%	0.000000	0.000000	2.504343e-02	0.000000	-3.452494e-01	3.000000	3.000000	3.000000	3.000000	3.000000	4.000000	4.000000
75%	1.000000	0.000000	7.518167e-01	1.000000	5.509472e-01	4.000000	4.000000	4.000000	4.000000	4.000000	4.000000	5.000000
max	1.000000	1.000000	2.998207e+00	1.000000	3.794278e+00	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000	5.000000

3.2 Univariate Analysis

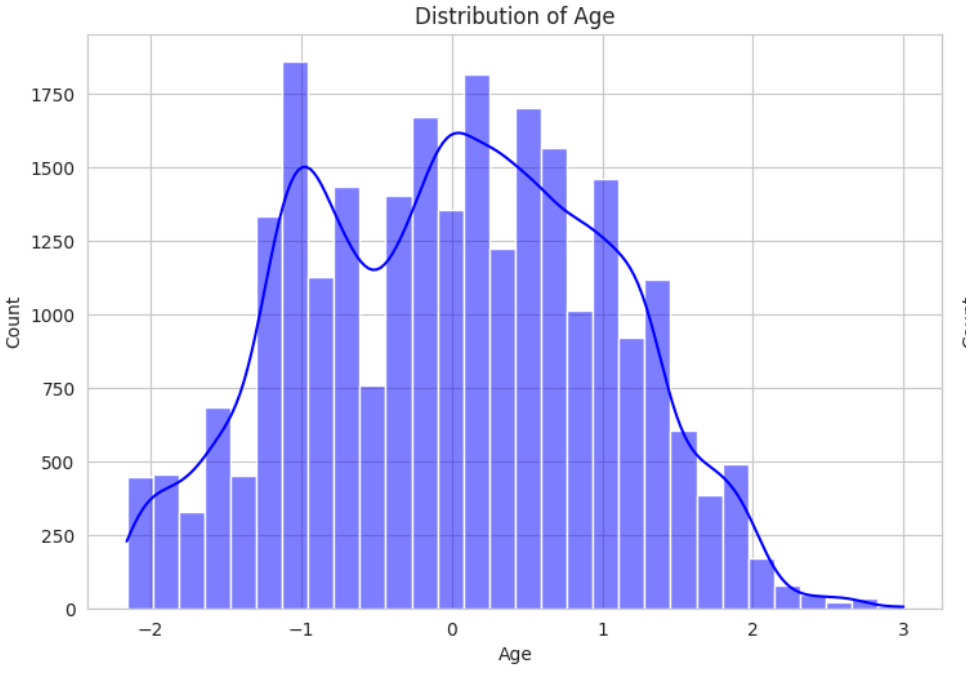


Fig. 3.1 Distribution plot of Flight Distance

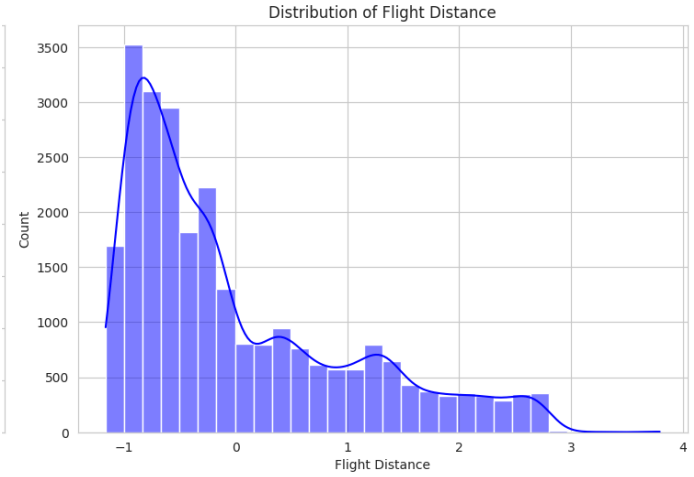


Fig. 3.2 Distribution plot of Age

This shows that the given data is right skewed for Flight Distance, Age

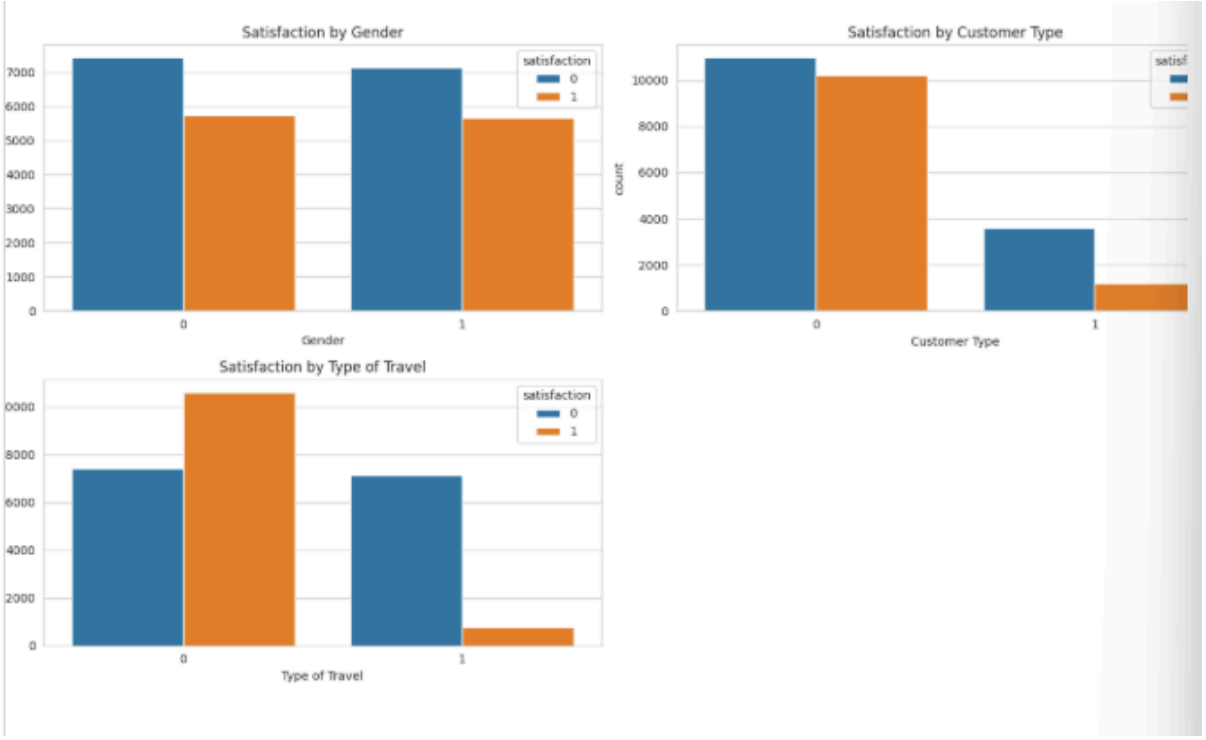


Fig. 3.3 Count plot of all attributes

These graphs show the categorical value distribution of the variables in the dataset

3.3 Distribution plot of Flight Distance

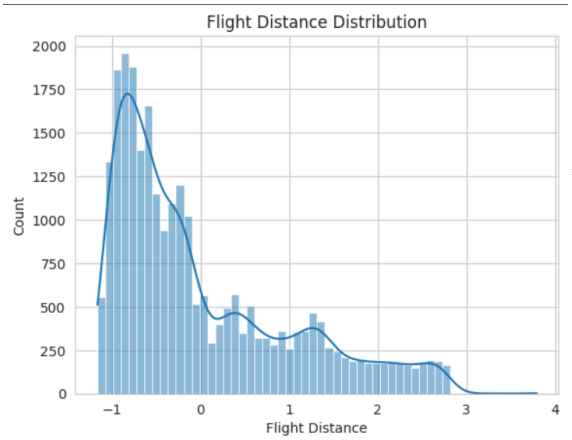
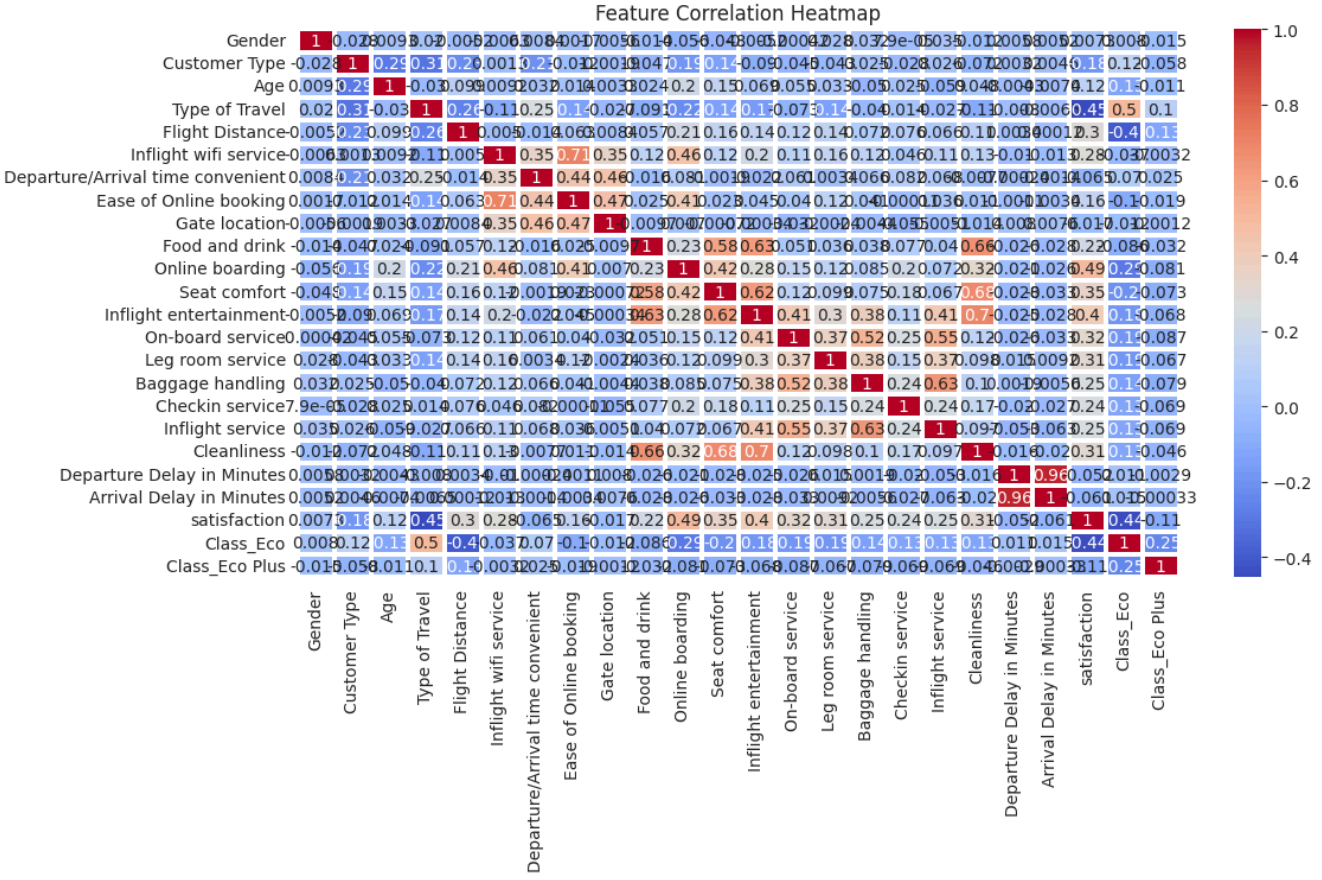


Fig. 3.4 Distribution plot of flight distance

3.4 Correlation Analysis

Correlation Heat Map:

Fig. 3.5 Heat Map showing correlation between all attributes



High positive correlation between:

- satisfaction and OnlineBoarding
- Satisfaction and Inflight Entertainment

Chapter 4
Data Modeling

The problem of predicting whether a passenger is satisfied or not is a classification problem. Therefore, we apply different classification algorithms and compare their performance to determine the best model for our dataset.

Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models a binary outcome; something that can take two values such as true/false, yes/no, and so on.

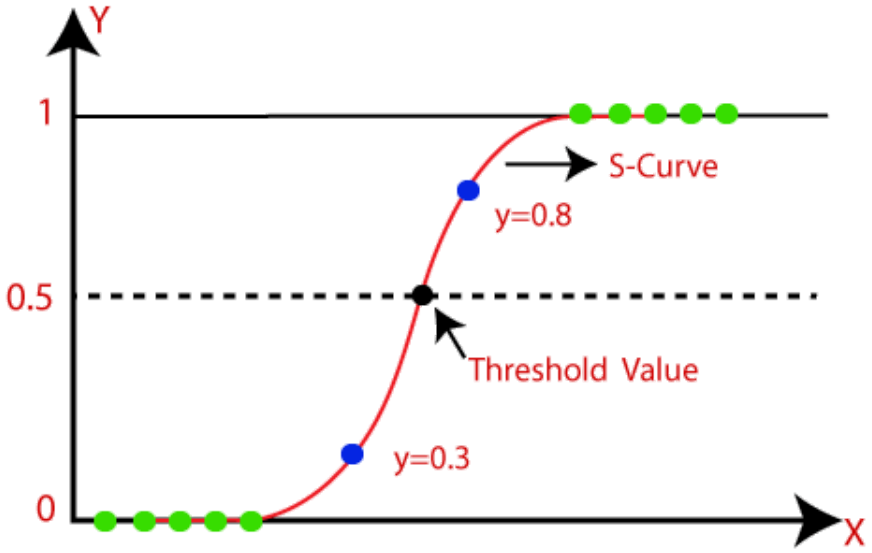


Fig. 4.1 Representation of logistic regression

Code:

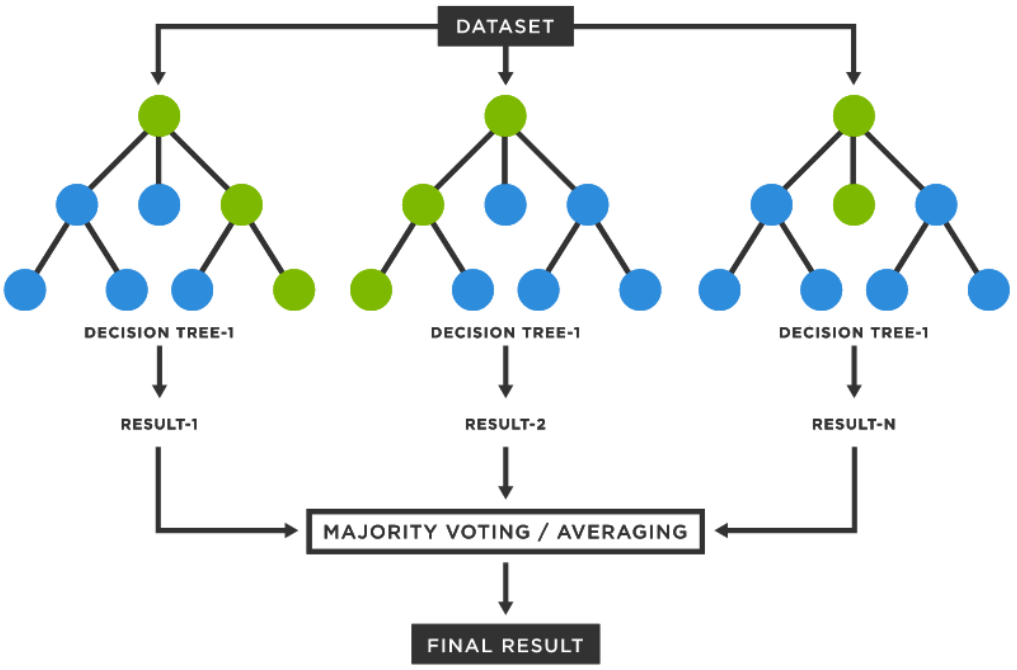
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train, y_train)
y_pred = log_model.predict(X_val)
accuracy = accuracy_score(y_val, y_pred)
print(f"Logistic Regression Accuracy: {accuracy:.4f}")
print("Classification Report:")
print(classification_report(y_val, y_pred))
print("Confusion Matrix:")
print(confusion_matrix(y_val, y_pred))
```

Output:

```
87.64
```

Random Forest

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regressio



n.

Fig. 4.2 Representation of Random Forest Classification

Code:

```
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
y_pred_rf = rf_model.predict(X_val)
accuracy_rf = accuracy_score(y_val, y_pred_rf)
print(f"Random Forest Accuracy: {accuracy_rf:.4f}")
print("Classification Report:")
print(classification_report(y_val, y_pred_rf))
print("Confusion Matrix:")
print(confusion_matrix(y_val, y_pred_rf))
```

Output:

96.18

K-Nearest Neighbors

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. It assumes the similarity between the new case/data and available cases and puts the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K- NN algorithm.

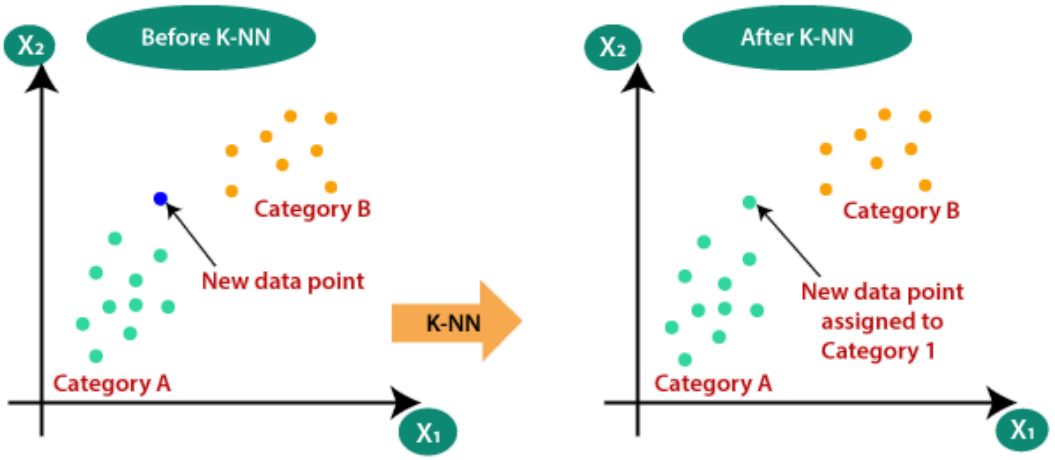


Fig. 4.3 Representation of KNN Classification

Code:

```
from sklearn.neighbors import KNeighborsClassifier
knn_model = KNeighborsClassifier(n_neighbors=5)
knn_model.fit(X_train, y_train)
y_pred_knn = knn_model.predict(X_val)
accuracy_knn = accuracy_score(y_val, y_pred_knn)
print(f"KNN Accuracy: {accuracy_knn:.4f}")
print("Classification Report:")
print(classification_report(y_val, y_pred_knn))
print("Confusion Matrix:")
print(confusion_matrix(y_val, y_pred_knn))
```

Output:

```
92.99
```

Gaussian Naive Bayes Classifier

Gaussian Naive Bayes is an extension of naive Bayes. Other functions can be used to estimate the distribution of the data, but the Gaussian (or Normal distribution) is the easiest to work with because you only need to estimate the mean and the standard deviation from your training data.

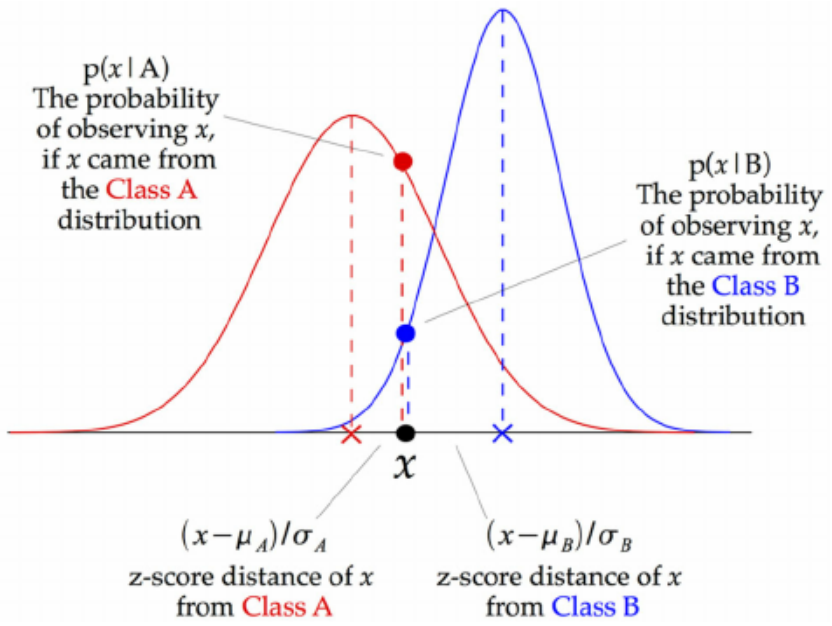


Fig. 4.4 Representation of Naive Bayesian Classification

Code:

```
from sklearn.naive_bayes import GaussianNB
nb_model = GaussianNB()
nb_model.fit(X_train, y_train)
y_pred_nb = nb_model.predict(X_val)
accuracy_nb = accuracy_score(y_val, y_pred_nb)
print(f"Naive Bayes Accuracy: {accuracy_nb:.4f}")
print("Classification Report:")
print(classification_report(y_val, y_pred_nb))
print("Confusion Matrix:")
print(confusion_matrix(y_val, y_pred_nb))
```

Output:

```
86.14
```

SVM Classifier

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

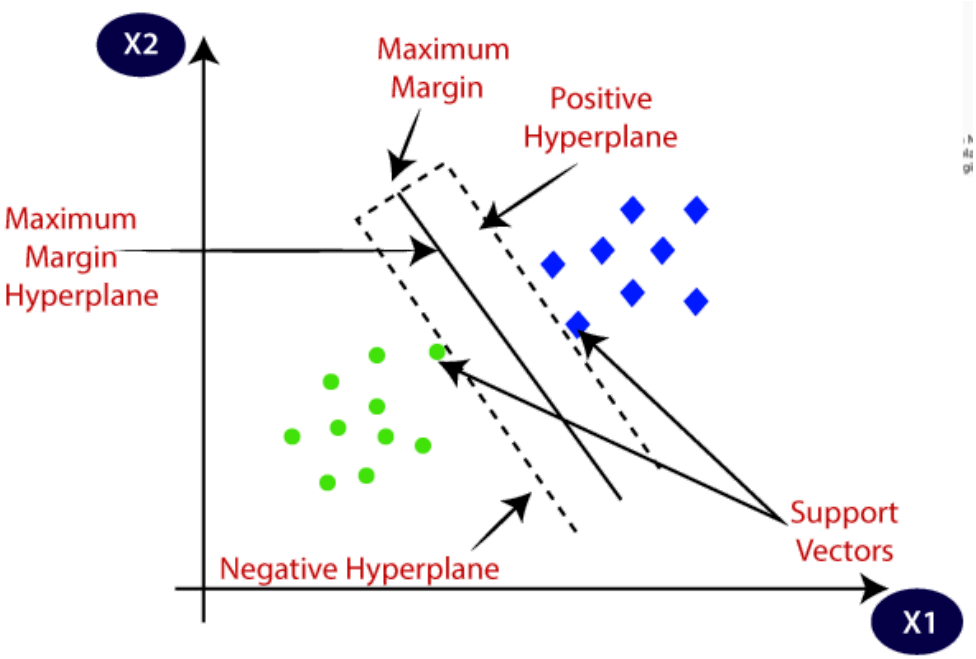


Fig. 4.5 Representation of SVM Classification

Code:

```
from sklearn.svm import SVC
svm_rbf_model = SVC(kernel="rbf", C=1.0, gamma="scale")
svm_rbf_model.fit(X_train, y_train)
y_pred_svm_rbf = svm_rbf_model.predict(X_val)
accuracy_svm_rbf = accuracy_score(y_val, y_pred_svm_rbf)
print(f"SVM (RBF Kernel) Accuracy: {accuracy_svm_rbf:.4f}")
print("Classification Report:")
print(classification_report(y_val, y_pred_svm_rbf))
print("Confusion Matrix:")
print(confusion_matrix(y_val, y_pred_svm_rbf))
```

Output:

```
93.69
```

Gradient Boosting Classifier

Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model. Decision trees are usually used when doing gradient boosting. Gradient boosting models are becoming popular because of their effectiveness at classifying complex datasets

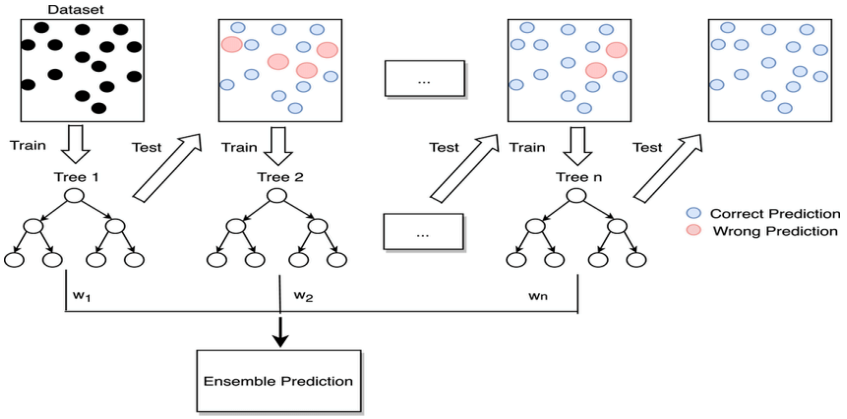


Fig. 4.6 Representation of Gradient Boosting Classification

Code:

```
from sklearn.ensemble import GradientBoostingClassifier
gb_model = GradientBoostingClassifier(n_estimators=100,
learning_rate=0.1, random_state=42)
gb_model.fit(X_train, y_train)
y_pred_gb = gb_model.predict(X_val)
accuracy_gb = accuracy_score(y_val, y_pred_gb)
print(f"Gradient Boosting Accuracy: {accuracy_gb:.4f}")
print("Classification Report:")
print(classification_report(y_val, y_pred_gb))
print("Confusion Matrix:")
print(confusion_matrix(y_val, y_pred_gb))
```

Output:

```
94.13
```

Summary of Data modeling

Model	**Score**
Logistic Regression	87.18
Gaussian Naive Bayes Classifier	85.93
Gradient Boosting Classifier	94.13
Random Forest	95.48
SVC	93.69
K- Nearest Neighbor	91.42

Table 4.1 Summary of Data modeling

The Highest Accuracy among Classifiers is shown by Random Forest => **95.48%**

Chapter 5
Hypothesis Testing

Statistical Hypothesis Testing: Evaluating the Performance Difference Between Random Forest and Logistic Regression

In machine learning, models are typically selected based on their mean performance. However, to ensure the difference in mean performance is statistically significant and not due to chance, we perform a **paired t-test**. Specifically, we use **5×2-fold cross-validation**, which is implemented in the **MLxtend** library by Sebastian Raschka through the paired_ttest_5x2cv() function.

Hypothesis Setup:

- **Null Hypothesis (H0)**: There is no significant difference between the performance of Random Forest and Logistic Regression.
- **Alternate Hypothesis (H1)**: There is a significant difference between the performance of Random Forest and Logistic Regression.

Mathematically:

- H0: d=0
- H1: d!=0

```
from mlxtend.evaluate import paired_ttest_5x2cv
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
import numpy as np
t, p = paired_ttest_5x2cv(
    estimator1=RandomForestClassifier(n_estimators=100, random_state=42),
    estimator2=LogisticRegression(max_iter=1000),
    X=X_train,
    y=y_train,
    scoring="accuracy",
    random_seed=42
)
print(f"P-value: {p:.3f}, t-Statistic: {t:.3f}")
if p <= 0.05:
    print("There is a significant difference between Random Forest and Logistic Regression.")
else:
        print("There is NO significant difference between Random Forest and Logistic Regression.")
```

- **P-value**: 0.000 (less than 0.05)
- **t-Statistic**: 18.03

Since the **p-value is less than 0.05**, we reject the null hypothesis (H0). This means that there is a **significant difference** between the performance of Random Forest and Logistic Regression.

There is a **significant difference** between the performance of **Random Forest** and **Logistic Regression**. Based on the results, we can confidently choose the model with the better performance for further classification tasks.

Chapter 6
Result and Analysis

Performance metrics for the model using Random Forest:

Accuracy : 95.48
Confusion Matrix:

```
[[ 2816   99]
 [136 2145]]
```

Classification Report:

```
Classification Report:
              precision    recall  f1-score   support

           0       0.95      0.97      0.96      2915
           1       0.96      0.94      0.95      2281

    accuracy                           0.95      5196
   macro avg       0.95      0.95      0.95      5196
weighted avg       0.95      0.95      0.95      5196
```

Features important for the classification:

```
importances = pd.DataFrame({'Features': X_train.columns, 'Importance':
np.round(rf_model.feature_importances_, 3)})
importances = importances.sort_values('Importance',
ascending=False).set_index('Features')
importances.plot.bar()
```

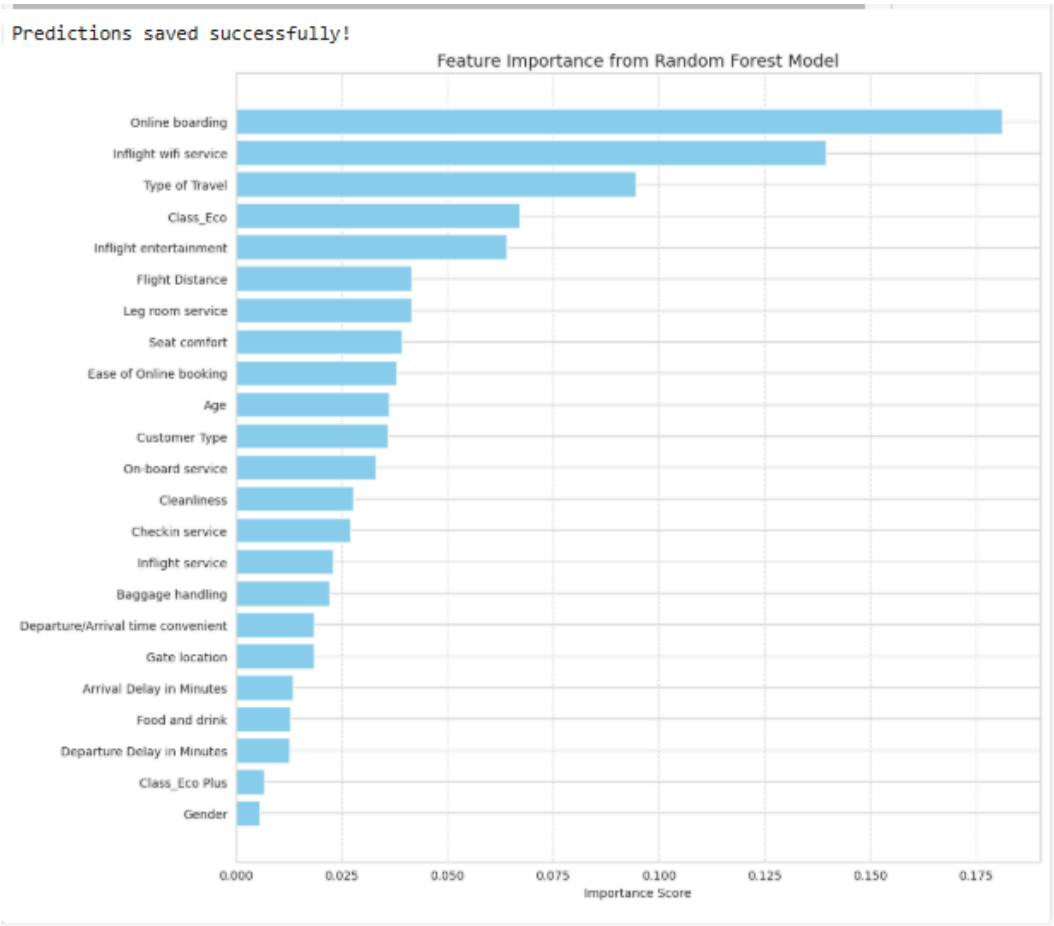


Fig. 6.1 Bar chart showing precedence of importance of features

Online boarding and *Flight Distance* have the maximum importance, while *Gender* and *Gate location* have the least.

Inference:

Passenger satisfaction is most closely related to features such as **Online boarding**, **Flight Distance**, **Class**, **Inflight WiFi service**, and **Inflight entertainment**. These are factors that significantly impact the customer experience during air travel.

The Random Forest model identifies these features as key contributors based on historical data, helping in making accurate predictions about passenger satisfaction for future or unseen instances.

This data-driven analysis helps understand the influence of different flight experience factors on satisfaction, allowing airlines to improve services that matter most to passengers.

Prediction over unseen dataset using Logistic Regression for final submission:

```
X_test = test_df.drop(columns=["satisfaction"], errors='ignore')
test_predictions = rf_model.predict(X_test)
submission = pd.DataFrame({
    "id": test_df.index,
    "satisfaction": test_predictions
})
submission.to_csv("final_predictions.csv", index=False)
```

Chapter 7

Future Scope

The field of predictive analytics for airline passenger satisfaction continues to evolve rapidly with the advancement of artificial intelligence and big data technologies. The current project has demonstrated a robust foundation by achieving a high classification accuracy using the Random Forest model. However, there is vast potential for enhancing the model and expanding its practical implementation.

1. **Integration with Real-Time Feedback**
   Future iterations can focus on integrating real-time feedback during flights through in-flight entertainment systems, mobile apps, or wearable devices. This would enable airlines to monitor satisfaction dynamically and take immediate corrective actions, thereby enhancing customer experience during the journey itself.

2. **Adoption of Deep Learning Models**
   With access to larger datasets, deep learning models such as CNNs (for image data like boarding pass scans) or LSTM/RNNs (for sequential data like real-time logs) can be implemented. These models are capable of capturing complex non-linear patterns and temporal dependencies more effectively than traditional classifiers.

3. **Multimodal Sentiment Analysis**
   In addition to structured feedback, unstructured data like social media posts, online reviews, and call center transcripts can be analyzed using Natural Language Processing (NLP). Sentiment analysis from these sources can enrich the prediction model and offer a 360-degree view of passenger satisfaction.

4. **Cross-Airline Dataset Expansion**
   The current model is trained on data from a single airline. Generalizing the model by incorporating data from multiple domestic and international airlines can significantly improve the reliability and adaptability of the system.

5. **Deployment as a Full-Stack Application**
   A production-grade web application can be developed for airlines. This would include dashboards, satisfaction heatmaps, and predictive alerts for service failure, empowering decision-makers with actionable insights in real time.

Chapter 8

Societal Impact

The deployment of machine learning models to predict airline passenger satisfaction not only advances business strategies but also holds significant societal implications. The ripple effects of such technologies can be seen across various dimensions — from enhancing customer experiences to optimizing airline operations and promoting sustainable development.

1. **Improved Passenger Experience**
   By proactively identifying dissatisfaction drivers such as delayed flights, poor service, or uncomfortable seating, airlines can improve the overall quality of travel. This leads to happier passengers, fewer complaints, and a better travel culture — which is especially critical in post-pandemic recovery phases.

2. **Operational Efficiency**
   Predictive analytics can help airlines allocate resources better — such as adjusting staff levels, improving food quality, or upgrading frequently complained seats. This reduces operational costs while ensuring that services are aligned with what passengers value most.

3. **Data-Driven Decision Making in Aviation**
   Empowering airline management with actionable insights brings a cultural shift from intuition-based to evidence-based decisions. This enhances transparency and trust in airline operations.

4. **Boost to Travel and Tourism Industry**
   Satisfied passengers are more likely to travel again, leave positive reviews, and recommend airlines. This indirectly supports the travel and tourism sector, which contributes significantly to national economies and employment.

5. **Digital Transformation and Skill Development**
   Projects like this one showcase the importance of digital tools and machine learning, encouraging students, professionals, and companies to invest in upskilling. This fosters a future-ready workforce and bridges the gap between academia and industry.

6. **Inclusivity and Accessibility**
   Advanced analytics can uncover patterns that reflect the needs of specific demographics, such as senior citizens or passengers with disabilities. Airlines can then tailor services that foster inclusivity and equality in air travel.

Chapter 9

Implementation

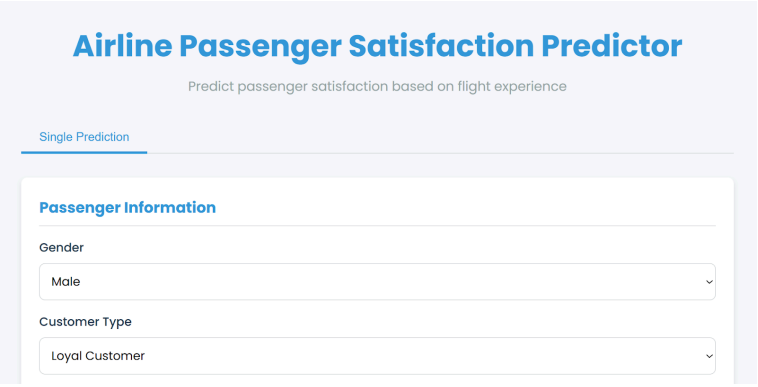


Fig. 9.1 Airline passenger satisfaction predictor

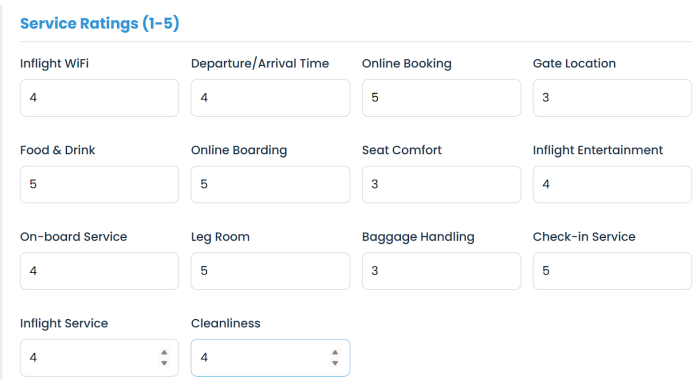


Fig. 9.2 Service based rating in 1-5

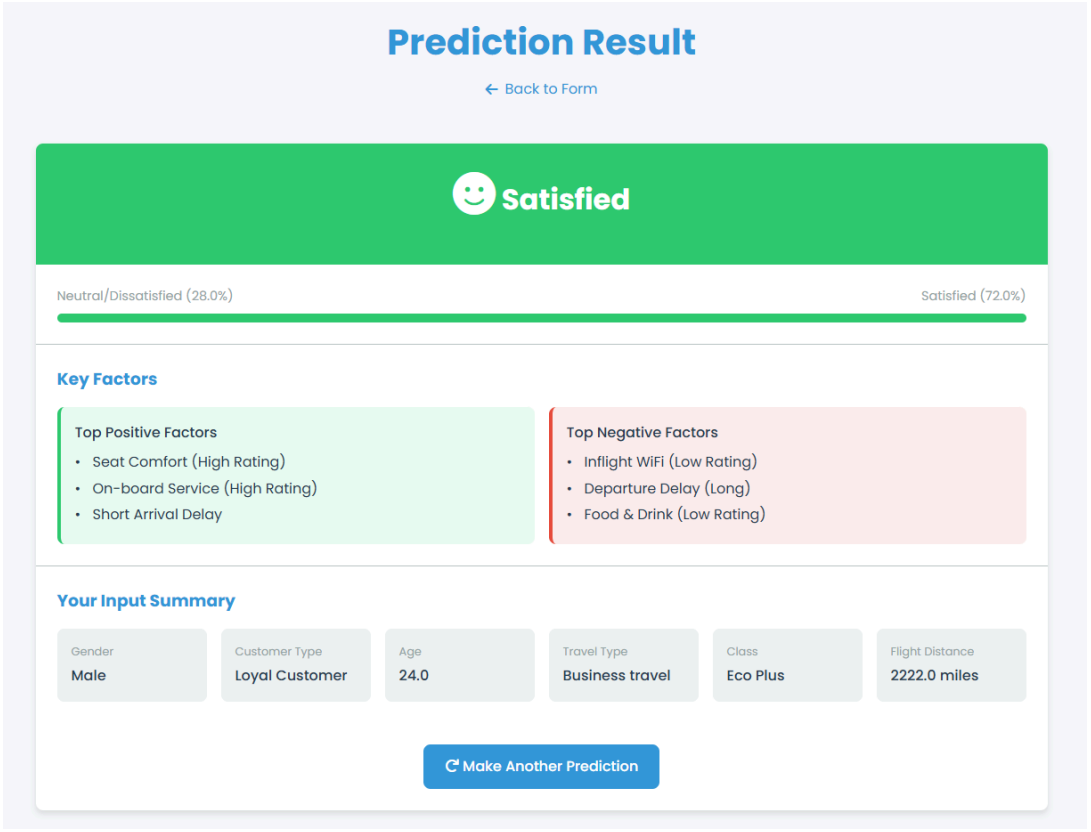


Fig. 9.3 Airline passenger satisfaction predictor Result

Chapter 10
CONCLUSION

In this project, we developed a machine learning classification model aimed at predicting passenger satisfaction based on a variety of features related to their in-flight experience. After exploring and comparing multiple machine learning algorithms, the Random Forest Classifier emerged as the most effective model, achieving a commendable accuracy of **95.48%**. This high level of accuracy underscores the model's robustness and its ability to effectively learn and generalize from complex patterns within the dataset.

The Random Forest algorithm proved particularly suitable for this task due to its ensemble learning approach, which combines the predictions of multiple decision trees to improve performance and reduce overfitting. By leveraging this algorithm, the model was able to accurately distinguish between satisfied and dissatisfied passengers using a diverse set of features.

Key factors influencing passenger satisfaction were identified through feature importance analysis. Among these, **inflight entertainment**, **seat comfort**, and **onboard services** stood out as the most impactful predictors. These insights provide valuable information for airline companies looking to enhance customer satisfaction by focusing on the aspects of service that matter most to their passengers.

Overall, the project successfully demonstrated how machine learning can be applied to real-world scenarios in the airline industry. By building a highly accurate predictive model, we showcased the potential of data-driven decision-making in improving customer experience and operational efficiency.

References

[1] Ashwika, Dishali G. K., Hemalatha Nambisan. "Airline Passenger Satisfaction Prediction Using Machine Learning Algorithms." *Redshine Archive*, Vol. X, Issue Y, pp., 2020.

[2] Lee Ye Hean, Olanrewaju Victor Johnson, Chew XinYing, Teoh Wei Lin, Chong Zhi Lin, Khaw Khai Wah. "An Airline Passenger Satisfaction Prediction by Genetic-Algorithm-Based Hybrid AutoEncoder and Machine Learning Models." *International Journal of Intelligent Systems and Applications in Engineering (IJISAE)*, 2024.

[3] Forecasting Airline Passengers' Satisfaction Based on Sentiments, 2024.

[4] Drivers and Outcomes of Airline Passenger Satisfaction: A Meta-Analysis, 2024.