Project Title	Customer Satisfaction Prediction
Tools	Google Colab
Technologies	Python, SQL, ML, Excel
Domain	Data Science
Project Difficulties level	Advance

Name: Sachin Rathod

UM ID: UMID07112569651

Ph. No: +91 8310619278

Email: sachinmrathod067@gmail.com

Internship Role: Data Science

Contents

INTRODUCTON:

Customer satisfaction plays a crucial role in the success and reputation of any organization, especially in technology-driven industries where customers frequently interact with support services. Efficient handling of customer support tickets, timely responses, and effective issue resolution significantly influence customer experience and satisfaction levels.

This project focuses on analyzing customer support ticket data to understand the factors that impact customer satisfaction. The dataset contains detailed information about customers, products purchased, types of issues raised, support channels used, ticket priority, and resolution timelines. By leveraging data science and machine learning techniques, the project aims to extract meaningful insights from historical ticket data and build predictive models that can estimate customer satisfaction levels.

Such analysis helps organizations improve service quality, optimize support operations, and proactively address issues that may lead to customer dissatisfaction.

OBJECTIVES:

The main objectives of this project are:

• To analyze customer support ticket data and understand common issues and service patterns.

• To study the relationship between ticket characteristics (type, priority, channel) and customer satisfaction levels.

• To evaluate the impact of response time and resolution time on customer satisfaction.

• To build and compare machine learning models for predicting customer satisfaction levels.

• To identify key factors that influence customer satisfaction and provide actionable insights for improving customer support services.

Data Science Application:

1)Customer Support & Issue Analysis

- Analyze support ticket types and subjects to identify the most common customer issues and recurring problems.
- Help organizations prioritize resources by understanding which issues require immediate attention.

2) Customer Behavior & Satisfaction Analysis

- Examine how customer demographics and ticket characteristics influence satisfaction levels.
- Identify patterns that explain why certain customers report higher or lower satisfaction.

3) Service Performance & Trend Analysis

- Evaluate response time and resolution time trends to measure support team efficiency.
- Detect performance fluctuations over time to support data-driven service improvements.

4) Support Channel Effectiveness Analysis

- Compare customer satisfaction across different support channels such as email, chat, phone, and social media.
- Identify the most effective channels for resolving issues with higher customer satisfaction.

5) Product & Service Quality Insights

- Identify products that generate frequent complaints or low satisfaction ratings.
- Provide insights to improve product quality and reduce customer support issues.

Dataset Descriptors:

The dataset used in this project consists of customer support ticket records collected from a technology-based customer service system. It contains detailed information related to customers, products, support tickets, service timelines, and customer satisfaction ratings. The dataset is well-suited for exploratory data analysis, service performance evaluation, and machine learning-based customer satisfaction prediction.

Key Attributes:

• **Ticket ID** – A unique identifier assigned to each customer support ticket.

• **Customer Name** – Name of the customer who raised the support ticket.

• **Customer Email** – Email address of the customer (anonymized for privacy).

• **Customer Age** – Age of the customer.

• **Customer Gender** – Gender of the customer.

• **Product Purchased** – The product associated with the support request.

• **Date of Purchase** – The date on which the product was purchased.

• **Ticket Type** – Category of the ticket such as technical issue, billing inquiry, or refund request.

• **Ticket Subject** – Short description of the issue raised by the customer.

• **Ticket Description** – Detailed explanation of the customer's problem or inquiry.

• **Ticket Status** – Current status of the ticket (open, closed, or pending).

• **Resolution** – Solution provided for closed tickets.

• **Ticket Priority** – Urgency level assigned to the ticket (low, medium, high, critical).

• **Ticket Channel** – Channel through which the ticket was raised (email, chat, phone, or social media).

• **First Response Time** – Time when the customer received the first response from support.

• **Time to Resolution** – Time when the ticket was fully resolved.

• **Customer Satisfaction Rating** – Customer's satisfaction score after ticket resolution on a scale of 1 to 5.

Ethical Considerations and Data Privacy:

• **Customer Data Privacy**
The dataset contains customer-related information such as age, gender, and anonymized email addresses. Personal identifiers were not used for modeling, and sensitive fields were removed during preprocessing to ensure customer privacy and data protection.

• **Responsible Use of Data**
The data was used strictly for analytical and educational purposes. No attempt was made to identify individual customers or misuse the information beyond the project scope.

• **Bias and Fairness**
Care was taken to analyze the model outputs across different demographic groups to reduce potential bias. The model was evaluated to ensure that predictions do not unfairly favor or disadvantage any specific customer group.

• **Transparency and Interpretability**
Machine learning models were selected based on interpretability and performance. Feature importance analysis was conducted to understand the factors influencing customer satisfaction predictions.

• **Ethical Decision Support**
The insights generated from this project are intended to support decision-making and service improvement, not to replace human judgment. All recommendations should be applied with ethical oversight and organizational responsibility.

Customer Satisfaction Prediction

Project Overview:

This project focuses on analyzing customer support ticket data to understand the key factors that influence customer satisfaction. Customer satisfaction is a critical metric for organizations, as it reflects the quality of service provided and directly impacts customer retention and brand reputation. By examining historical support ticket records, the project aims to identify patterns related to customer behavior, service performance, and issue resolution.

The dataset includes information such as customer demographics, products purchased, types of support issues, ticket priority, support channels, response time, resolution time, and customer satisfaction ratings. Using data science techniques, the project performs data preprocessing, exploratory data analysis, and feature engineering to prepare the data for modeling.

Machine learning models are then developed to predict customer satisfaction levels based on ticket characteristics and service-related metrics. The results of the analysis provide actionable insights that can help organizations improve support processes, optimize response strategies, and enhance overall customer experience.

Project Steps:

1) Data Preprocessing

• Removed irrelevant and identifier columns such as ticket ID and customer contact details to maintain data privacy.
• Handled missing values, converted date and time fields into proper formats, and prepared the dataset for analysis and modeling.

2) Exploratory Data Analysis (EDA)

• Analyzed distributions of customer satisfaction, ticket types, priorities, and support channels to understand data patterns.
• Identified trends and relationships between service-related variables and customer satisfaction levels using statistical summaries and plots.

3) Feature Engineering

• Created new features such as resolution time in hours by calculating the difference between response and resolution timestamps.
• Transformed the customer satisfaction rating into categorical levels to improve classification performance.

4) Model Building

• Implemented machine learning models including Logistic Regression and Random Forest to predict customer satisfaction levels.
• Split the dataset into training and testing sets and applied feature scaling to ensure consistent model performance.

5) Model Evaluation

• Evaluated model performance using accuracy, precision, recall, and F1-score metrics.
• Compared results across models to identify the most reliable and accurate prediction approach.

6) Visualization

• Used various visualizations such as bar charts, box plots, and heatmaps to represent insights clearly.
• Visualized feature importance and performance metrics to support data-driven conclusions.

Source code and Output:

IN [1]:

#Import Required Libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
```

IN [2]:

```
# Load the dataset
data = pd.read_csv('/content/customer_support_tickets.csv')
data.head()
```

	Ticket ID	Customer Name	Customer Email	Customer Age	Customer Gender	Product Purchased	Date of Purchase	Ticket Type	Ticket Subject	Ticket Description	Ticket Status	Resolution	Ticket Priority	Ticket Channel	Fi Respo T
0	1	Marisa Obrien	carrollallison@example.com	32	Other	GoPro Hero	2021-03-22	Technical issue	Product setup	I'm having an issue with the {product_purchase...	Pending Customer Response	NaN	Critical	Social media	2023- 12:1
1	2	Jessica Rios	clarkeashley@example.com	42	Female	LG Smart TV	2021-05-22	Technical issue	Peripheral compatibility	I'm having an issue with the {product_purchase...	Pending Customer Response	NaN	Critical	Chat	2023- 16:4
2	3	Christopher Robbins	gonzalestracy@example.com	48	Other	Dell XPS	2020-07-14	Technical issue	Network problem	I'm facing a problem with my {product_purchase...	Closed	Case maybe show recently my computer follow.	Low	Social media	2023- 11:1
3	4	Christina Dillon	bradleyolson@example.org	27	Female	Microsoft Office	2020-11-13	Billing inquiry	Account access	I'm having an issue with the {product_purchase...	Closed	Try capital clearly never color toward story.	Low	Social media	2023- 07:2
4	5	Alexander Carroll	bradleymark@example.com	67	Female	Autodesk AutoCAD	2020-02-04	Billing inquiry	Data loss	I'm having an issue with the {product_purchase...	Closed	West decision evidence bit.	Low	Email	2023- 00:1

OUTPUT 1

OUTPUY 2

	Customer Age	Customer Gender	Product Purchased	Date of Purchase	Ticket Type	Ticket Subject	Ticket Description	Ticket Status	Resolution	Ticket Priority	Ticket Channel	Customer Satisfaction Rating	Resolution_Time_Hours	Satisfaction_Level
2	48	2	10	2020-07-14	4	8	53	0	343	2	3	3.0	6.850000	2
3	27	0	25	2020-11-13	0	0	627	0	2549	2	3	3.0	-5.533333	2
4	67	0	5	2020-02-04	0	3	188	0	2657	2	1	1.0	19.683333	1
10	48	1	30	2021-01-19	1	3	1323	0	1368	1	2	1.0	-17.916667	1
11	51	1	27	2021-10-24	2	15	360	0	1366	1	0	1.0	-2.633333	1

data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 8469 entries, 0 to 8468
Data columns (total 17 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Ticket ID                     8469 non-null   int64
 1   Customer Name                 8469 non-null   object
 2   Customer Email                8469 non-null   object
 3   Customer Age                  8469 non-null   int64
 4   Customer Gender               8469 non-null   object
 5   Product Purchased             8469 non-null   object
 6   Date of Purchase              8469 non-null   object
 7   Ticket Type                   8469 non-null   object
 8   Ticket Subject                8469 non-null   object
 9   Ticket Description            8469 non-null   object
 10  Ticket Status                 8469 non-null   object
 11  Resolution                    2769 non-null   object
 12  Ticket Priority               8469 non-null   object
 13  Ticket Channel                8469 non-null   object
 14  First Response Time           5650 non-null   object
 15  Time to Resolution            2769 non-null   object
 16  Customer Satisfaction Rating  2769 non-null   float64
dtypes: float64(1), int64(2), object(14)
memory usage: 1.1+ MB
```

data.describe()

	Ticket ID	Customer Age	Customer Satisfaction Rating
count	8469.000000	8469.000000	2769.000000
mean	4235.000000	44.026804	2.991333
std	2444.934048	15.296112	1.407016
min	1.000000	18.000000	1.000000
25%	2118.000000	31.000000	2.000000
50%	4235.000000	44.000000	3.000000
75%	6352.000000	57.000000	4.000000
max	8469.000000	70.000000	5.000000

IN [3]:

```
# Drop unnecessary identifier columns

data.drop(['Ticket ID', 'Customer Name', 'Customer Email'], axis=1, inplace=True)

# Convert date columns
data['Date of Purchase'] = pd.to_datetime(data['Date of Purchase'], errors='coerce')
data['First Response Time'] = pd.to_datetime(data['First Response Time'], errors='coerce')
data['Time to Resolution'] = pd.to_datetime(data['Time to Resolution'], errors='coerce')
```

IN [4]:

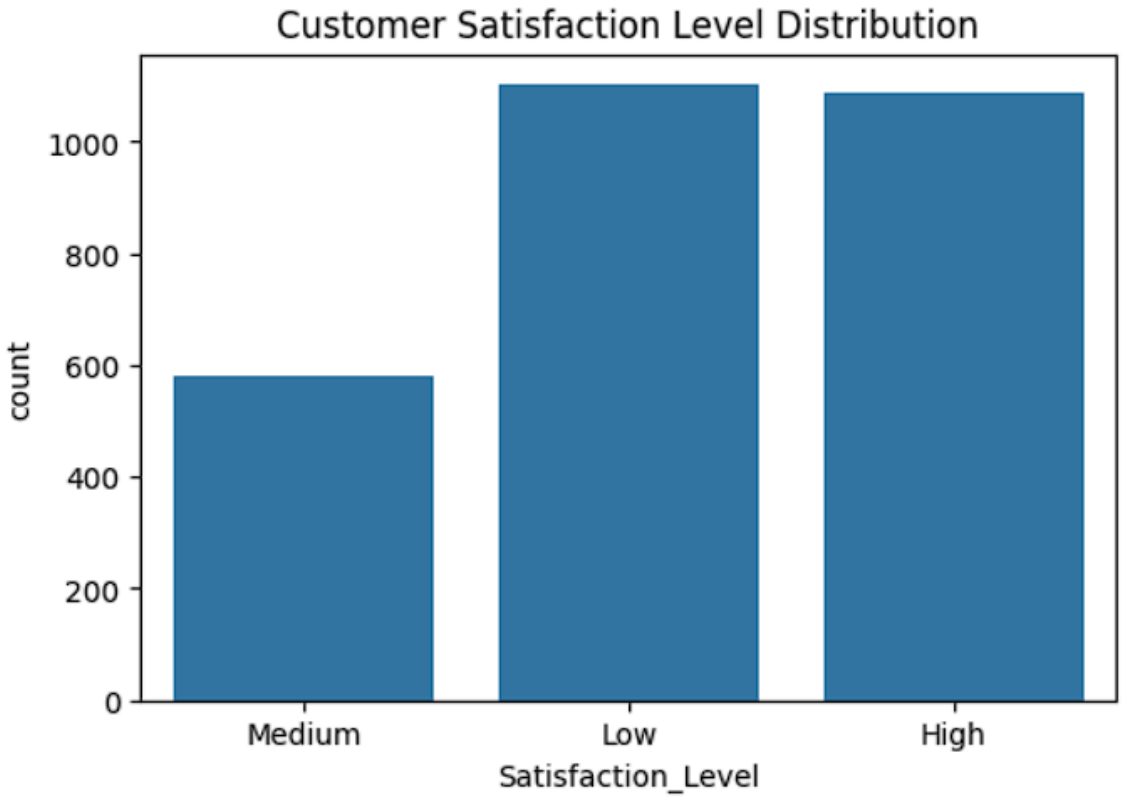
```
# Create resolution time in hours

data['Resolution_Time_Hours'] = (
    data['Time to Resolution'] - data['First Response Time']
).dt.total_seconds() / 3600

# Drop original time columns
data.drop(['First Response Time', 'Time to Resolution'], axis=1, inplace=True)
```

IN [5]:

```
# Remove rows without satisfaction rating

data = data.dropna(subset=['Customer Satisfaction Rating'])

# Convert satisfaction into categories
def satisfaction_level(score):
    if score <= 2:
        return 'Low'
    elif score == 3:
        return 'Medium'
    else:
        return 'High'

data['Satisfaction_Level'] = data['Customer Satisfaction Rating'].apply(satisfaction_level)
```

IN [6]:

```
# Exploratory Analysis

plt.figure(figsize=(6,4))
sns.countplot(x='Satisfaction_Level', data=data)
plt.title('Customer Satisfaction Level Distribution')
plt.show()
```



IN [7]:

```
# Encoding Categorical variables

label_encoders = {}

for col in data.select_dtypes(include='object').columns:
    le = LabelEncoder()
    data[col] = le.fit_transform(data[col])
    label_encoders[col] = le
```

IN [8]:

```
# Train-Test Split
X = data.drop(['Customer Satisfaction Rating', 'Satisfaction_Level'], axis=1)
y = data['Satisfaction_Level']

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.3, random_state=42, stratify=y
)
```

IN [9]:

```
# Feature Scaling
# Convert 'Date of Purchase' to numerical (Unix timestamp) before scaling

X_train['Date of Purchase'] = X_train['Date of Purchase'].apply(lambda x: x.timestamp() if pd.notna(x) else x)
X_test['Date of Purchase'] = X_test['Date of Purchase'].apply(lambda x: x.timestamp() if pd.notna(x) else x)

scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

IN [10]:

```
# Model 1 - Logistic Regression

lr = LogisticRegression(max_iter=1000)
lr.fit(X_train, y_train)

lr_pred = lr.predict(X_test)

print("Logistic Regression Accuracy:", accuracy_score(y_test, lr_pred))
print(classification_report(y_test, lr_pred))
```

```
Logistic Regression Accuracy: 0.4055354993983153
              precision    recall  f1-score   support

           0       0.39      0.48      0.43       326
           1       0.42      0.55      0.47       331
           2       0.00      0.00      0.00       174

    accuracy                           0.41       831
   macro avg       0.27      0.34      0.30       831
weighted avg       0.32      0.41      0.36       831
```

IN [11]:

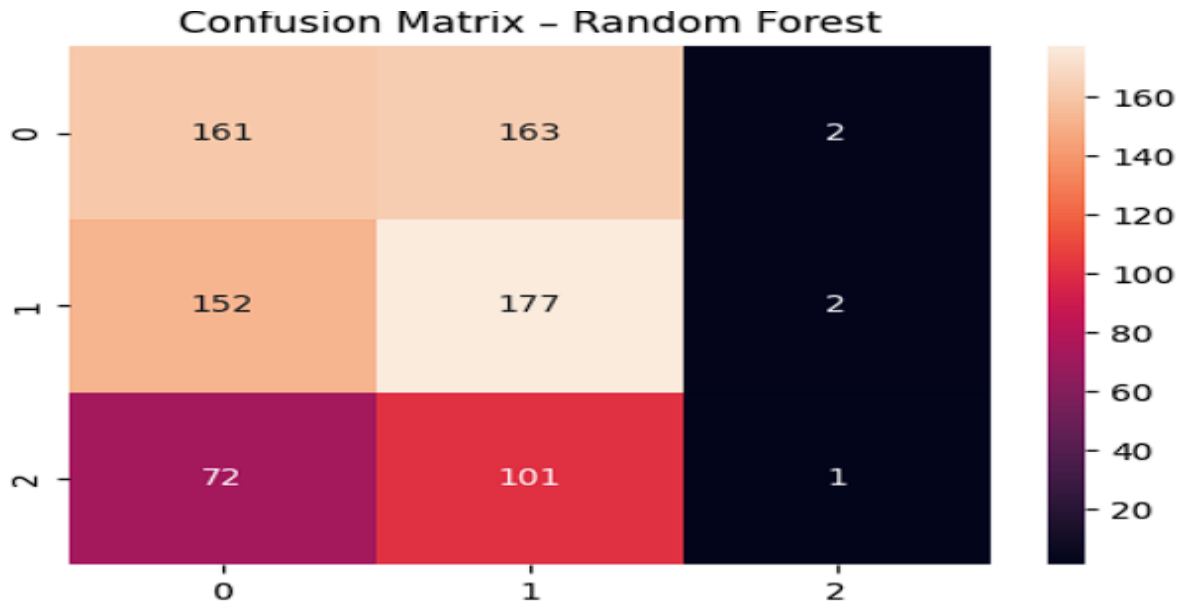
```
# Model 2 - Random Forest

rf = RandomForestClassifier(
   n_estimators=200,
```

```
    max_depth=12,
    random_state=4

rf.fit(X_train, y_train)
rf_pred = rf.predict(X_test)

print("Random Forest Accuracy:", accuracy_score(y_test, rf_pred))
print(classification_report(y_test, rf_pred))
```

```
Random Forest Accuracy: 0.40794223826714804
              precision    recall  f1-score   support

           0       0.42      0.49      0.45       326
           1       0.40      0.53      0.46       331
           2       0.20      0.01      0.01       174

    accuracy                           0.41       831
   macro avg       0.34      0.34      0.31       831
weighted avg       0.37      0.41      0.36       831
```

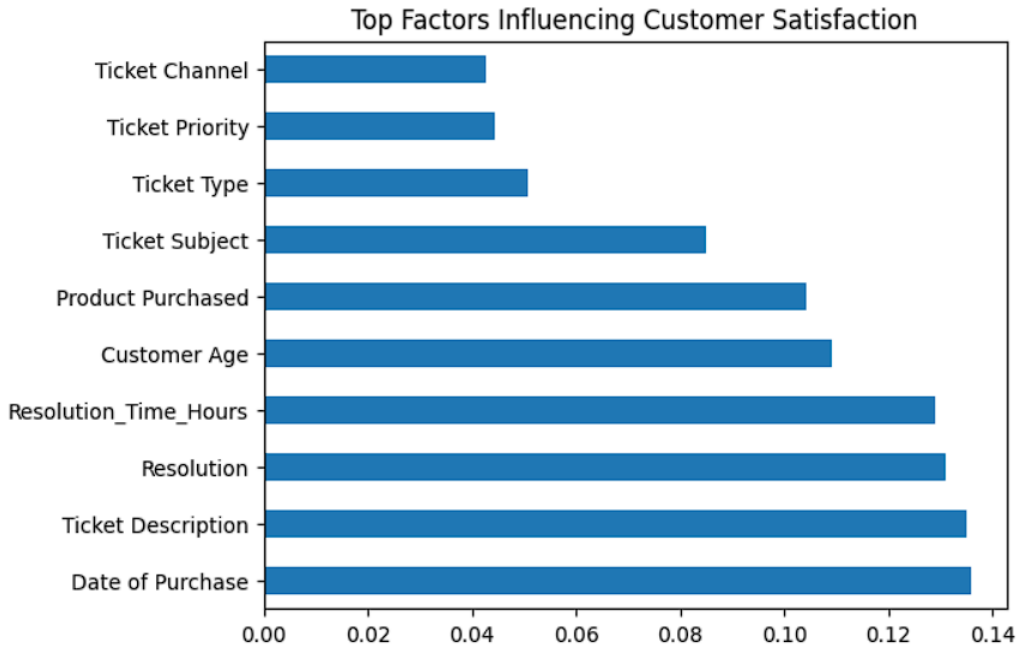
IN [12]:

```
# Comfusion Matrix

plt.figure(figsize=(6,4))
sns.heatmap(confusion_matrix(y_test, rf_pred), annot=True, fmt='d')
plt.title("Confusion Matrix – Random Forest")
plt.show()
```



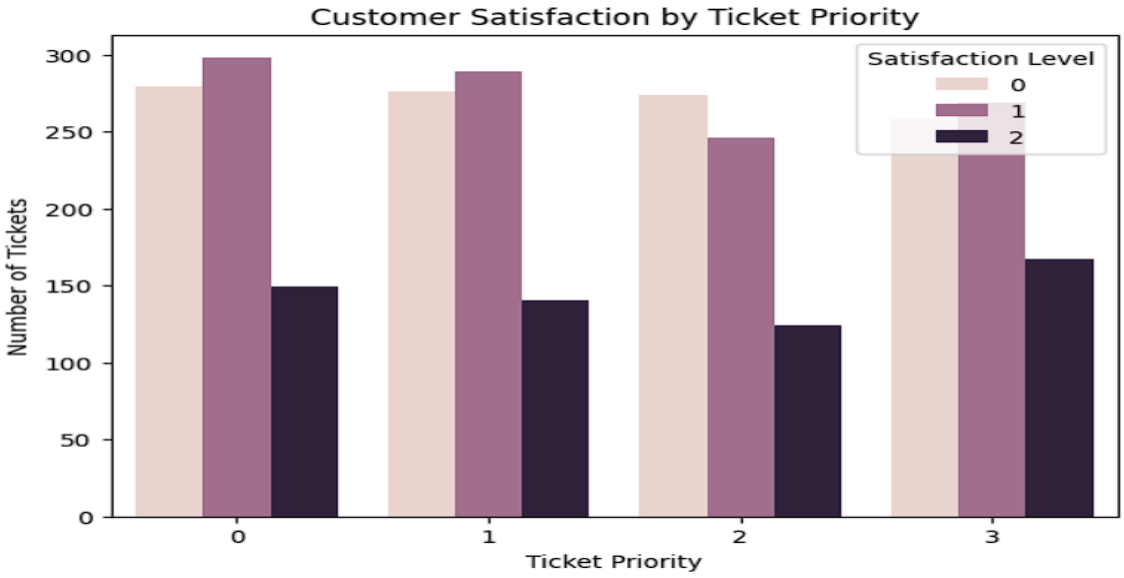
IN [13]:

```
# Feature Importance

importance = pd.Series(rf.feature_importances_, index=X.columns)
importance.sort_values(ascending=False).head(10).plot(kind='barh')
plt.title("Top Factors Influencing Customer Satisfaction")
plt.show()
```



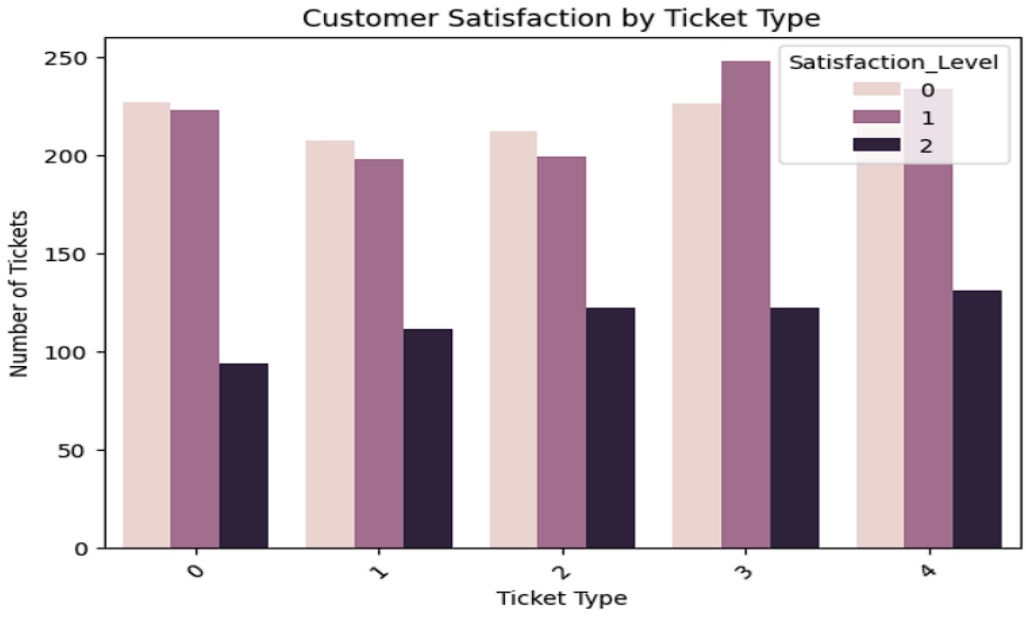
IN [14]:

```
# Satisfaction Level by Ticket Priority

plt.figure(figsize=(7,5))
sns.countplot(
    x='Ticket Priority',
    hue='Satisfaction_Level',
    data=data
)
plt.title('Customer Satisfaction by Ticket Priority')
plt.xlabel('Ticket Priority')
plt.ylabel('Number of Tickets')
plt.legend(title='Satisfaction Level')
plt.show()
```



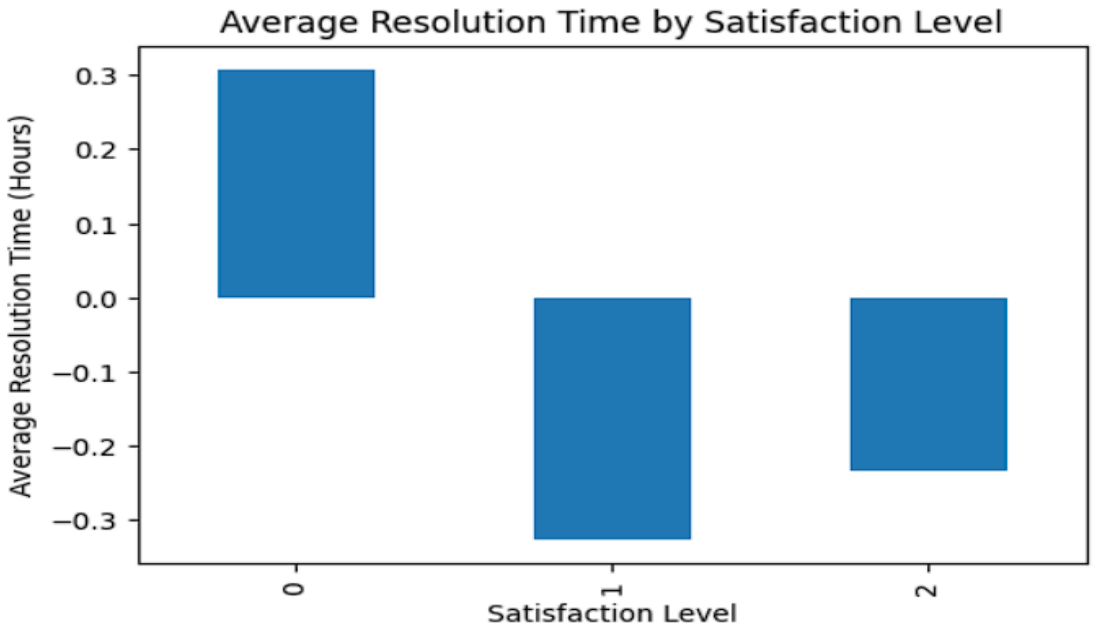
IN [15]:

```
# Satisfaction Level by Ticket Type

plt.figure(figsize=(7,5))
sns.countplot(
    x='Ticket Type',
    hue='Satisfaction_Level',
    data=data
)
plt.title('Customer Satisfaction by Ticket Type')
plt.xlabel('Ticket Type')
plt.ylabel('Number of Tickets')
plt.xticks(rotation=45)
plt.show()
```



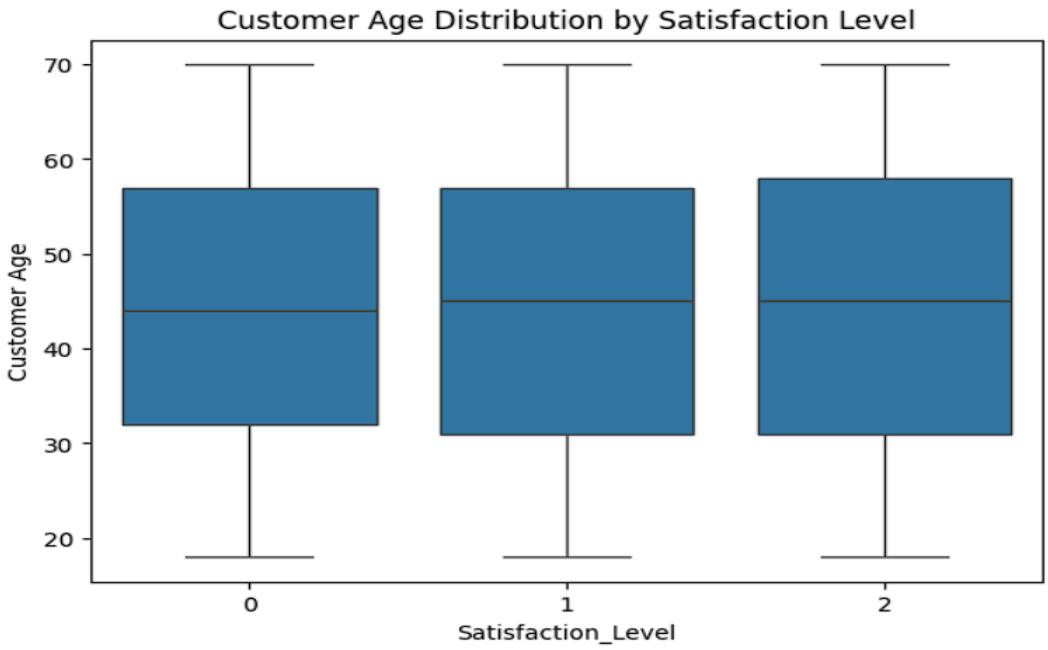
IN [16]:

```
# Average Resolution Time vs Satisfaction Level

avg_resolution = data.groupby('Satisfaction_Level')['Resolution_Time_Hours'].mean()

plt.figure(figsize=(6,4))
avg_resolution.plot(kind='bar')
plt.title('Average Resolution Time by Satisfaction Level')
plt.xlabel('Satisfaction Level')
plt.ylabel('Average Resolution Time (Hours)')
plt.show()
```



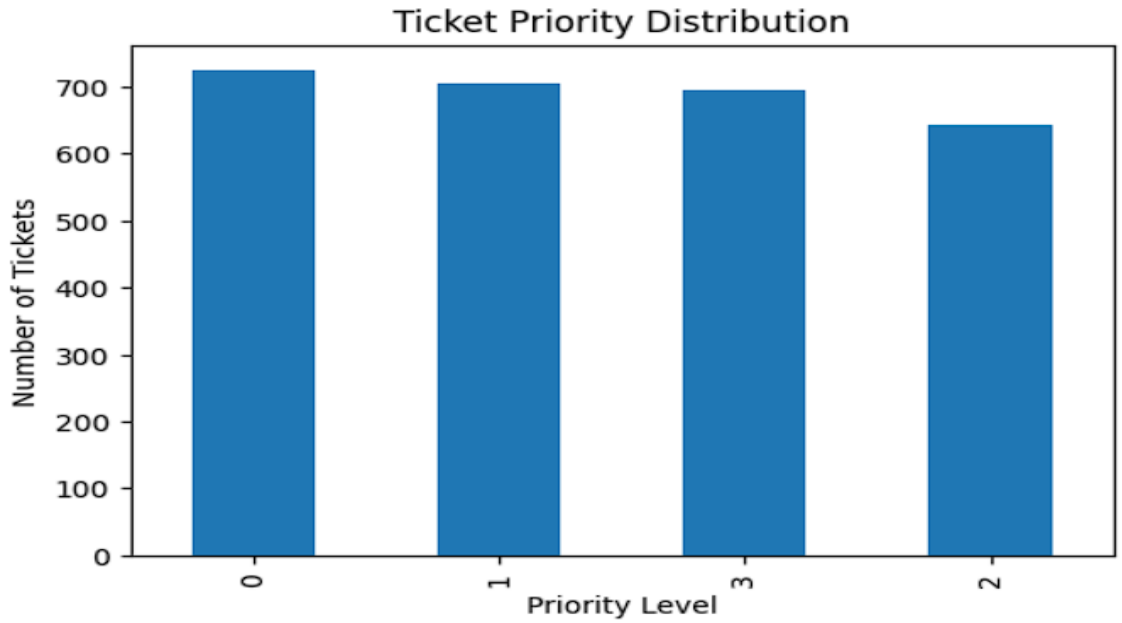
IN [17]:

```
# Customer Age Distribution by Satisfaction Level

plt.figure(figsize=(7,5))
sns.boxplot(
    x='Satisfaction_Level',
    y='Customer Age',
    data=data
)
plt.title('Customer Age Distribution by Satisfaction Level')
plt.show()
```



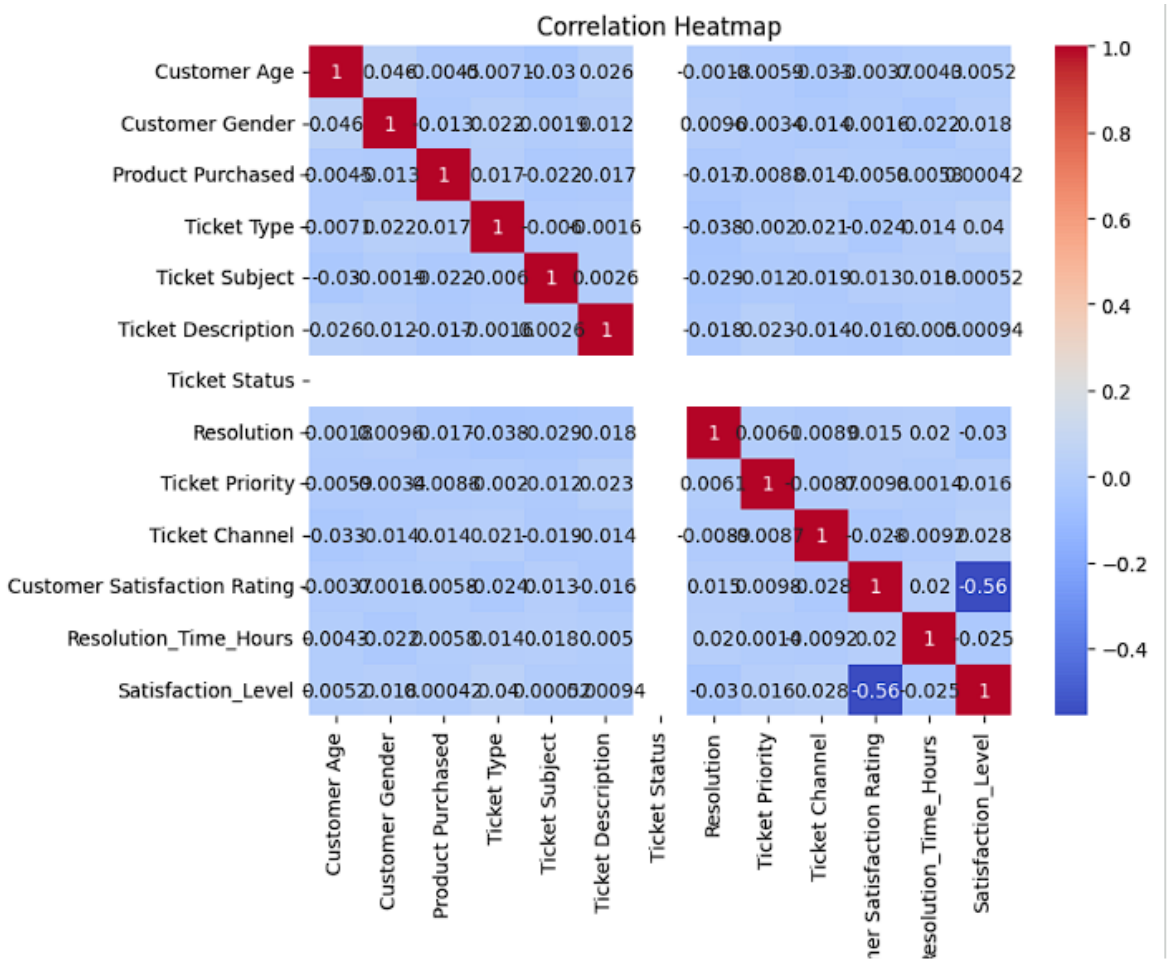
IN [18]:

```
# Ticket Priority Distribution

plt.figure(figsize=(6,4))
data['Ticket Priority'].value_counts().plot(kind='bar')
plt.title('Ticket Priority Distribution')
plt.xlabel('Priority Level')
plt.ylabel('Number of Tickets')
plt.show()
```



IN [19]:

```
# Heatmap – Correlation of Numerical Features

plt.figure(figsize=(8,6))
sns.heatmap(
    data.select_dtypes(include=['int64','float64']).corr(),
    annot=True,
    cmap='coolwarm'
)
plt.title('Correlation Heatmap')
plt.show()
```



Conclusion:

This project successfully analyzed customer support ticket data to understand the factors influencing customer satisfaction. Through data preprocessing, exploratory analysis, and feature engineering, meaningful patterns related to ticket characteristics, service performance, and customer behavior were identified. The analysis highlighted the importance of response time, resolution time, ticket priority, and support channels in determining customer satisfaction levels.

Machine learning models were developed to predict customer satisfaction, and the Random Forest model demonstrated better performance compared to baseline models. Feature importance analysis further revealed that service-related metrics had a stronger impact on satisfaction than customer demographic factors.

Overall, the insights obtained from this project can help organizations improve customer support processes, optimize resource allocation, and enhance customer experience. The predictive model and analytical findings provide a data-driven foundation for making informed decisions aimed at increasing customer satisfaction and service efficiency.

LINK

Customer_Satisfaction_Prediction