Customer Satisfaction Prediction – Project Documentation

# 1. Project Overview

The Customer Satisfaction Prediction project aims to predict the satisfaction rating of customers based on various factors like ticket details, response times, product information, and resolution speed. This can help companies improve their customer service quality and prioritize issues effectively.

# 2. Objectives

The main objectives of this project were:  
- Analyze historical customer support data.  
- Engineer meaningful features from text, categorical, and numerical data.  
- Apply multiple machine learning models to predict customer satisfaction.  
- Compare models using accuracy, classification reports, and confusion matrices.  
- Optimize the best-performing model using hyperparameter tuning.

# 3. Dataset Description

The dataset contains the following key features:  
- Ticket Description (Text Data)  
- Ticket Type, Priority, Channel, Product Purchased, Status (Categorical Data)  
- First Response Time, Time to Resolution (Datetime Data)  
- Customer Age, Resolution Delay, Response Speed, Resolution Speed (Numerical Data)  
- Customer Satisfaction Rating (Target Variable – 1 to 5 scale)

# 4. Data Preprocessing

Key preprocessing steps included:  
- Handling missing values using mean/median imputation.  
- Converting date/time columns into numeric hour-based features.  
- Encoding categorical variables using one-hot encoding.  
- Text cleaning using TF-IDF for ticket descriptions.  
- Scaling and normalizing numerical features where necessary.

# 5. Feature Engineering

Feature engineering included:  
- Numerical Features: Customer Age, Response/Resolution Times, Delays  
- Categorical Features: Ticket Type, Priority, Channel, Status  
- Text Features: Ticket Description using TF-IDF (Top 500 words)  
- Date Features: Weekday, Month, High-Risk Ticket Flags

# 6. Model Building

Multiple machine learning models were trained including:  
- Random Forest Classifier  
- Gradient Boosting  
- XGBoost Classifier  
  
Random Forest performed best after hyperparameter tuning.

# 7. Hyperparameter Tuning

Hyperparameter tuning was done using GridSearchCV with parameters:  
- n\_estimators: Number of trees  
- max\_depth: Depth of trees  
- min\_samples\_split, min\_samples\_leaf: Node splitting criteria  
- learning\_rate, subsample for boosting models

# 8. Evaluation Metrics & Results

The final tuned Random Forest model achieved:  
- Accuracy: 65.1%  
- Macro F1-Score: ~0.65  
Confusion Matrix showed balanced performance across all satisfaction classes.

# 9. Feature Importance Analysis

Feature importance from Random Forest showed that response time, resolution time, ticket priority, and product type were the most significant predictors of customer satisfaction.

# 10. Tools & Libraries Used

Python, Pandas, NumPy, Scikit-learn, XGBoost, Matplotlib, Seaborn, TF-IDF Vectorizer, WordCloud

# 11. Conclusion & Next Steps

The model provides a baseline system to predict customer satisfaction ratings. Next steps could include:  
- Deploying the model as a web service  
- Integrating real-time prediction in support systems  
- Collecting more data for better accuracy