# Machine Learning Project Proposal

## Project Title:

Analyzing Airline Performance Using Machine Learning to Predict Flight Delays

## Team Members:

• Eman Asif – BSDSF22M003  
• Musqan Yousaf – BSDSF22M004  
• Maham Jamil – BSDSF22M008

## Problem Statement:

Air travel is a critical mode of transportation, yet flight delays cause significant inconvenience and financial losses. Airlines, airports, and passengers benefit greatly from being able to anticipate delays in advance. In this project, we aim to develop a machine learning-based solution that analyzes historical flight data and predicts whether a flight will be delayed or not based on various flight and airline parameters.

## Objectives:

1. To analyze airline performance using key metrics such as delay frequency, on-time performance, and cancellations.  
2. To visualize trends and insights across different airlines and time periods.  
3. To build a classification model that predicts whether a flight will be delayed based on features like airline, origin/destination, time, and distance.  
4. To evaluate the model’s performance and identify the most influential features contributing to delays.

## Proposed Methodology:

1. Data Preprocessing:  
 - Handle missing values and inconsistent formats.  
 - Feature engineering (e.g., extracting hour, weekday from timestamp).  
 - Convert categorical features using encoding.  
  
2. Exploratory Data Analysis:  
 - Use visualizations (matplotlib/seaborn/plotly) to uncover patterns in delays, flight counts, and performance by airline.  
  
3. Classification Modeling:  
 - Logistic Regression  
 - Random Forest  
 - Decision Tree  
 - Support Vector Machine (SVM)  
 - XGBoost (optional)  
  
4. Evaluation Metrics:  
 - Accuracy  
 - Precision/Recall  
 - F1-score  
 - Confusion Matrix

## Dataset Description:

We will use the U.S. Department of Transportation Flight Delay dataset (available on Kaggle and BTS Transtats) which includes:  
- Airline, Flight Number  
- Scheduled and Actual Departure/Arrival Time  
- Delay Duration  
- Origin and Destination Airport  
- Flight Status (On-time, Delayed, Cancelled)  
- Distance

## Expected Outcomes:

- Visual insights on the best and worst-performing airlines.  
- A classification model capable of predicting delays with reasonable accuracy.  
- Feature importance ranking to understand which factors influence delays most.  
- An optional dashboard or report summarizing findings for decision-makers.

## Timeline of Activities:

|  |  |
| --- | --- |
| Days | Task |
| Day 1 | Dataset collection and preprocessing |
| Day 2 | Exploratory Data Analysis and Visualizations |
| Day 3 | Feature engineering and model training |
| Day 4 | Model tuning and evaluation |
| Day 5-6 | Prepare final report and visual dashboard |
| Day 7 | Final testing, results review, and presentation preparation |

## Tools and Libraries:

- Python (Pandas, scikit-learn, Matplotlib, Seaborn)  
- Jupyter Notebook  
- (Optional) Power BI or Plotly Dash for visualizations