Objective:

Students are tasked with building an ensemble learning model to predict flight delay times based on flight data. The ensemble will consist of models built from scratch, including Support Vector Machine (SVM), Decision Trees, and Logistic Regression, using only pandas and numpy.

Dataset Overview:

You will be provided with a dataset of outbound flights from 2023-2024 for Lahore (LHE), Karachi (KHI), and Islamabad (ISB) airports. The dataset includes the following fields:

- Departure time
- Estimated departure time
- Delay time (in minutes)
- Arrival details (destination, time, etc.)
- Airline and flight information

Task:

1. **Target Variable:**
   - The target variable is **delay_time**, which represents the delay in departure time in minutes.
   - **Binning:** You will bin the delay_time into a total of 8 bins.
2. **Ensemble Model:**
   - You are required to build an ensemble learning model from scratch. ○ Use at least **three models** as learners: Support Vector Machine (SVM), Decision Trees, and Logistic Regression.
   - Implement SVM that supports **multiple kernels** only(e.g., linear, polynomial, RBF). Single-kernel SVMs are not to be used.
   - You are free to choose the number and sequence of models in the ensemble. You can experiment with bagging or boosting techniques. ○ There should be at least 3 different types of models in your sequence. 3. **Model Implementation:**

○ **Build models from scratch only** using pandas and numpy. You are not allowed to use libraries like scikit-learn for the models themselves. ○ Implement the following models from scratch:

■ **Multi-Kernel Support Vector Machine (SVM)**

■ **Decision Tree Classifier**

■ **Logistic Regression**

4. **Data Preprocessing:**

○ **Feature Selection:** You are free to choose the features (columns) to be used for the model. You can mix and match different columns and create new features (e.g., time of day, destination).

○ Handle any missing or incorrect data in the dataset appropriately. ○ Filter out "active" flights only to train your model (flight status == active). The rest of the data cleaning is up to your selection.

5. **Evaluation Metrics:**

○ You will evaluate the ensemble model's performance using accuracy and F1-score.

○ Perform cross-validation to evaluate the robustness of your model.

6. **Deliverables:**

○ Submit a Jupyter Notebook or Python script that:

■ Preprocesses the data

■ Implements each of the models

■ Combines the models into an ensemble

■ Evaluates the performance of the ensemble model

○ Include a report discussing:

■ Feature selection choices

■ The ensemble method used (e.g., stacking, bagging, boosting)

■ Evaluation metrics and model performance

Bonus:

- Implementing more than 3 different types of models in your ensemble.
- Achieving accuracy of more than 93%

Notes:

- Focus on building the models from scratch and explaining your approach. ● Build the Multi-Kernel SVM and experiment with different kernel combinations. ● The students are encouraged to experiment with different combinations of models and ensemble methods.
- Deadline: The deadline to submit the assignment is 14th September 2024. No late submission will be accepted. Correct and timely submission of the assignment is the responsibility of every student; hence no relaxation will be given to anyone.
- Honor Policy: This assignment is a learning opportunity that will be evaluated based on your ability. Plagiarism cases will be dealt with strictly. If found plagiarized, both the involved parties will be awarded zero marks in this assignment, all the remaining assignments, or even an F grade in the course.