Machine Learning Project Proposal

(Using Free Datasets from Kaggle or Other Sources)

Project Requirements:

1. **Team Composition**:
    - Each group must consist of members from the same specialization (e.g., SAD or AI).
    - Teams should have 3-8 members.
2. **Dataset**:
    - The dataset must be **raw** (unprocessed) and **freely available** from sources like Kaggle, UCI.....etc.
    - The dataset should be sufficiently complex to require preprocessing (e.g., handling missing values, encoding categorical variables, feature scaling, etc.).
3. **Model Requirements**:
    - For **classification tasks**, the model must output a **Confusion Matrix**, **Precision**, **Recall**, and **F1-Score**.
    - For **regression tasks**, the model must have at least **4 features** and output evaluation metrics such as **RMSE**, **MAE**, and **R-squared**.
    - **Optional**→ The project include a comparison of at least **two different models** (e.g., Logistic Regression vs. Random Forest, or Linear Regression vs. Decision Tree).
4. **Deliverables**:
    - A **Jupyter Notebook** or Python script containing the code, comments, and explanations (you may use MATLAB).
    - A **report** (PDF) summarizing the problem, dataset, methodology, results, and conclusions.
    - **Visualizations** (e.g., graphs, charts, confusion matrix, etc.).
    - Printed copies of the **Confusion Matrix**, **Precision**, **Recall**, **F1-Score**, or regression metrics.
5. **Evaluation Criteria**:

□ How well the raw data was cleaned and prepared.

□ Justification for the chosen models.

□ Accuracy of the model based on the required metrics.

□ Quality and relevance of the visualizations.

□ Clarity, structure, and depth of the report.

Notes:

- **Deadline**: Projects must be submitted by 13th week
- **Presentation**: Each group will present their project, explaining their methodology, results, and challenges faced.
- **Plagiarism**: Any form of plagiarism will result in disqualification.

Project Ideas (with Free Datasets):

1. Classification: Titanic Survival Prediction

- **Dataset**: Titanic Dataset from Kaggle
- **Task**: Predict whether a passenger survived the Titanic disaster based on features like age, gender, class, etc.
- **Models**: Logistic Regression, Random Forest, or Gradient Boosting.
- **Evaluation**: Confusion Matrix, Precision, Recall, F1-Score.

2. Regression: House Price Prediction

□ **Dataset**: House Prices Dataset from Kaggle

□ **Task**: Predict house prices based on features like square footage, number of bedrooms, location, etc.

□ **Models**: Linear Regression, Decision Tree, or Support Vector Regression.

□ **Evaluation**: RMSE, MAE, R-squared.

3. Classification: Spam Email Detection

□ **Dataset**: Spambase Dataset from UCI

□ **Task**: Classify emails as spam or not spam based on features like word frequency, subject line, etc.

□ **Models**: Naive Bayes, Support Vector Machine, or Neural Networks.

- **Evaluation**: Confusion Matrix, Precision, Recall, F1-Score.

4. Regression: Bike Sharing Demand Prediction

- **Dataset**: Bike Sharing Dataset from UCI
- **Task**: Predict the number of bikes rented per hour based on features like weather, time of day, season, etc.
- **Models**: Linear Regression, Random Forest, or XGBoost.
- **Evaluation**: RMSE, MAE, R-squared.

5. Classification: Credit Card Fraud Detection

- **Dataset**: Credit Card Fraud Detection Dataset from Kaggle
- **Task**: Detect fraudulent transactions based on features like transaction amount, location, time, etc.
- **Models**: Logistic Regression, Random Forest, or Neural Networks.
- **Evaluation**: Confusion Matrix, Precision, Recall, F1-Score.

6. Regression: Student Performance Prediction

- **Dataset**: Student Performance Dataset from UCI
- **Task**: Predict student grades based on features like study time, parental education, attendance, etc.
- **Models**: Linear Regression, Decision Tree, or Support Vector Regression.
- **Evaluation**: RMSE, MAE, R-squared.

7. Classification: Sentiment Analysis on Movie Reviews

- **Dataset**: IMDB Movie Reviews Dataset from Kaggle
- **Task**: Classify movie reviews as positive or negative based on text data.
- **Models**: Naive Bayes, LSTM, or BERT.
- **Evaluation**: Confusion Matrix, Precision, Recall, F1-Score.

8. Regression: Energy Consumption Prediction

- **Dataset**: Appliances Energy Prediction Dataset from UCI
- **Task**: Predict energy consumption based on features like temperature, humidity, time of day, etc.
- **Models**: Linear Regression, Random Forest, or XGBoost.

- **Evaluation**: RMSE, MAE, R-squared.

9. Classification: Heart Disease Prediction

- **Dataset**: Heart Disease Dataset from UCI
- **Task**: Predict the presence of heart disease based on features like age, cholesterol levels, blood pressure, etc.
- **Models**: Logistic Regression, Random Forest, or Support Vector Machine.
- **Evaluation**: Confusion Matrix, Precision, Recall, F1-Score.

10. Regression: Car Price Prediction

- **Dataset**: Car Price Prediction Dataset from Kaggle
- **Task**: Predict car prices based on features like mileage, brand, year, etc.
- **Models**: Linear Regression, Decision Tree, or Random Forest.
- **Evaluation**: RMSE, MAE, R-squared.

Good luck! Dr. Manar