Machine learning challenges!

Agenda- To build a ml model from complex dataset

Task 1: Data Preprocessing & Exploration

- **Objective**: Understand the dataset by cleaning and exploring it.
- **Actions**: Handle missing values, normalize or standardize data, remove outliers, and perform exploratory data analysis (EDA) to gain insights.
- **Deliverables**: Cleaned dataset, EDA visualizations, insights about key features**.**

Task 2: Feature Engineering

- **Objective**: Transform raw data into meaningful features for the model.
- **Actions**: Create new features, perform dimensionality reduction (e.g., using PCA), and handle categorical data with techniques like one-hot encoding.
- **Deliverables**: A refined dataset with engineered features

Task 3: Model Selection & Training

- **Objective**: Choose and train machine learning models.
- **Actions**: Split the dataset into training, validation, and test sets. Train multiple models (like Decision Trees, Random Forests, XGBoost) and tune hyperparameters using cross-validation.
- **Deliverables**: Trained models with performance metrics (accuracy, precision, recall, F1-score, etc.).

Task 4: Model Evaluation & Optimization

- **Objective**: Evaluate the model on the test set and optimize performance.
- **Actions**: Use performance metrics to evaluate the model on unseen data. If necessary, perform model stacking, boosting, or hyperparameter tuning for optimization.
- **Deliverables**: Final model with a performance report and deployment-ready code.

(Here's one dataset each for classification and regression:

1. Classification: Pima Indians Diabetes Database

- **Description**: This dataset contains medical data related to predicting whether a patient has diabetes based on factors like age, glucose level, blood pressure, etc.
- **Task**: Binary classification (predict whether the patient has diabetes or not).
- **Size**: 768 rows and 9 columns.
- **Why it's good**: Simple and clean, ideal for binary classification problems, and involves medical data which makes it practical and relevant.

2. Regression: House Prices - Advanced Regression Techniques

- **Description**: This dataset contains housing data, including features like lot size, number of rooms, and neighborhood characteristics, used to predict house sale prices.
- **Task**: Regression (predict house prices).
- **Size**: 1,460 rows and 80 features.

- **Why it's good**: It's a bit larger with a wide range of features, making it ideal for practicing regression, feature engineering, and model tuning.

Complex dataset:

Market basket analysis.csv

Diabetes dataset.csv) – optional