Python Programming, Data Analysis, and Machine Learning Tasks

Task 1: Python Programming Basics

Objective:

In this task, you will apply basic Python programming concepts to process and analyze a list of strings. The task will test your understanding of Python basics, including loops, conditionals, and string manipulation.

Resources:

- **Python task:** CodeAi-TechAi-Task1.ipynb
- **Python Tutorial:** Python Full Course for Beginners

Task 2: Exploratory Data Analysis (EDA)

Problem Statement: Exploratory Data Analysis of IPL Data

Objective:

Analyze the provided ball-by-ball IPL deliveries dataset to find meaningful patterns and strategic insights about players and teams.

Core Requirements:

- Clean and prepare the dataset for analysis.
- Perform analysis to identify top performers, key matchups, and performance trends (e.g., in powerplay or death overs).
- Present at least three actionable insights supported by visualizations.

Brownie Points:

- Create **interactive visualizations** using a library like Plotly.
- **Insightful and Interactive Visualization**

Submission:

- Push a single, well-commented Jupyter Notebook to a public GitHub repository and submit the link via the provided Google Form.

Resources:

- **Dataset:** [deliveries_updated_mens_ipl_upto_2024](#)
- **EDA tutorial:**
  - [Learn Exploratory Data Analysis (EDA) from Scratch | EDA in 5 hours | Satyajit Pattnaik](#)
  - [Complete Exploratory Data Analysis And Feature Engineering In 3 Hours| Krish Naik](#)

Task 3: Machine Learning (ML) - Model Building

Problem Statement: Predicting Daily Bike Share Demand

Objective:

Your task is to build a machine learning model that accurately predicts the total number of daily bike rentals. This is a classic regression problem where you will use historical data to forecast future demand.

Dataset Overview: Bike Sharing Daily Data

You will be working with a dataset that contains daily records of a bike-sharing system. The data includes various factors that might influence rental demand. Your goal is to predict the cnt column, which represents the total number of rentals for a given day.

Key Features in the Dataset:

- **dteday**: The date of the record.
- **season**: The season (1: spring, 2: summer, 3: fall, 4: winter).
- **yr**: The year (0: 2011, 1: 2012).
- **mnth**: The month (1 to 12).
- **holiday**: Whether the day is a holiday (1) or not (0).
- **weekday**: The day of the week (0-6).
- **workingday**: Whether the day is a workday (1) or a weekend/holiday (0).
- **weathersit**: The weather situation (1: Clear, 2: Mist/Cloudy, 3: Light Snow/Rain, 4: Heavy Rain/Snow).
- **temp**: Normalized temperature in Celsius.
- **hum**: Normalized humidity.
- **windspeed**: Normalized wind speed.
- **casual & registered**: Counts of casual and registered users.
- **cnt (Target Variable)**: The total count of rental bikes (casual + registered).

Your Tasks:

Data Preprocessing:

- Load and inspect the dataset.
- Perform any necessary data cleaning and feature engineering. Think about how you can best represent categorical features and handle date information.

Model Building:

- Build and train a regression model to predict the cnt variable (e.g., Linear Regression, Decision Tree/Random Forest, Gradient Boosting, etc.).
- Clearly document your choice of models and any hyperparameter tuning you perform.

Model Evaluation:

- Rigorously evaluate the performance of your models using appropriate regression metrics (e.g., Mean Absolute Error, Mean Squared Error, R-squared).
- Use techniques like cross-validation to ensure your evaluation is robust.

Conclusion:

- Compare the performance of your models and provide a clear recommendation for which model should be used and why.
- Briefly explain which features you found to be most important for predicting bike rental demand.

Deliverable & Submission:

- **Jupyter Notebook:** Your complete solution should be contained within a single, well-commented Jupyter Notebook (.ipynb) file.
- **GitHub Repository:** Push your final notebook to a public repository on your personal GitHub account. Clearly document your code.
- **Submission:** Submit the direct link to your GitHub repository via the Google Form provided.

Brownie Points:

- Instead of predicting cnt directly, build two separate models to predict casual and registered users. Your final prediction for cnt will be the sum of the outputs from these two models. This demonstrates a deeper understanding of the data.

Resources:

- **Dataset:** [day](day)
- **Machine Learning tutorial:**
  - [Linear Regression, Clearly Explained!!!](#)
  - [Linear Regression Practical Implementation In Hindi](#)