Capstone Project Guidelines

A capstone project involves applying your knowledge to analyze a given dataset. You will conduct extensive research, use critical thinking, and apply practical skills to derive meaningful insights and solutions. This project will demonstrate your expertise in data analysis and your ability to tackle real-world problems.

1. Data Exploration and Cleaning

- Task 1: Load the dataset into a pandas DataFrame and display the first few rows.
- Task 2: Summarize the dataset by providing basic statistics (mean, median, mode, standard deviation, etc.).
- Task 3: Identify and handle missing values. Explain the chosen method for handling them.
- Task 4: Identify and handle duplicate rows if any.
- Task 5: Convert categorical variables to numerical values using appropriate encoding techniques (e.g., one-hot encoding, label encoding).

2. Data Visualization

- Task 6: Create visualizations to understand the distribution of numerical features (e.g., histograms, box plots).
- Task 7: Create visualizations for categorical features (e.g., bar charts, pie charts).
- Task 8: Generate correlation heatmaps to identify relationships between numerical features.
- Task 9: Use pair plots to visualize relationships between features.

3. Feature Engineering

- Task 10: Create new features that might be useful for the analysis (e.g., date-related features from timestamps, interaction terms).
- Task 11: Standardize or normalize numerical features if needed.

4. Model Building

- Task 12: Split the dataset into training and testing sets.
- Task 13: Train a simple linear regression model (if the task is regression) or a logistic regression model (if the task is classification).
- Task 14: Evaluate the model performance using appropriate metrics (e.g., RMSE for regression, accuracy/F1-score for classification).

- Task 15: Experiment with at least two other algorithms (e.g., decision tree, random forest, k-nearest neighbors) and compare their performance.

5. Model Tuning

- Task 16: Perform hyperparameter tuning using GridSearchCV or RandomizedSearchCV.
- Task 17: Evaluate and compare the tuned models' performance.

6. Data Visualization with Power BI

- Task 18: Import the cleaned and preprocessed dataset into Power BI.
- Task 19: Create interactive dashboards to visualize key insights from the data, such as:
  - Distribution of numerical features
  - Comparison of categorical features
  - Correlation heatmap
  - Key model metrics and performance indicators
- Task 20: Use Power BI features to allow dynamic exploration of the data (e.g., slicers, filters).

7. Reporting

- Task 21: Summarize the findings and results in a Jupyter Notebook **(.ipynb file)**, including visualizations and explanations.
- Task 22: Create a final report or presentation summarizing the entire process and key insights, integrating Power BI visualizations.

8. Optional Advanced Tasks

- Task 23: Deploy the model using a simple web application (e.g., Flask, Streamlit).
- Task 24: Perform additional analysis such as time-series forecasting if the dataset has a time component.

Note: Submit the project in .ipynb format along with the presentation file(report).