Project | ICS 474 – Big Data Analytics

Dr. Muzammil Behzad, Assistant Professor

Department of Information and Computer Science, King Fahd University of Petroleum and Minerals.
Email: muzammil.behzad@kfupm.edu.sa

Instructions

1. Make teams, and give your team a very cool name.
2. Choose one of the datasets from below links.
3. Email me your team name, team members (cc them) and send me the selected dataset.
    a. This should be done as soon as possible so that you can start the project.
    b. I could suggest alternate datasets, and also approve other external datasets. But please ask!
4. Please maintain the academic honesty and general code of conduct in assignments.
5. The deadline is: **November 15, 11:59 PM**. We will have brief project presentations in the following week.

Project Datasets

1. https://www.kaggle.com/c/ashrae-energy-prediction/data
2. https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london
3. https://www.kaggle.com/datasets/bhanupratapbiswas/uber-data-analysis
4. https://www.kaggle.com/datasets/priyamchoksi/credit-card-transactions-dataset
5. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
6. https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
7. https://www.kaggle.com/datasets/hugomathien/soccer
8. https://www.kaggle.com/code/faressayah/stock-market-analysis-prediction-using-lstm/notebook

Deliverables

1. **Report**
    - **Format**: A well-structured document (PDF) with team names, members and answering the following questions.
    - **Content**:
        - A template of content is provided below.
        - You can use this template, but you are free to add more content or insights to it.
2. **Code**
    - **Submission**: All scripts or notebooks (e.g., Jupyter Notebook) used in your analysis.
    - **Requirements**:
        - Proper documentation and comments.
        - Instructions on how to run the code.
        - List of dependencies and libraries used, etc.
3. **Visualizations**
    - Include all relevant charts, graphs, and plots that support your analysis.
    - Ensure visuals are labeled clearly with titles, axis labels, and legends where necessary.

Part 1: Data Understanding and Exploration

1. **Dataset Overview**
    - What is the source and context of your chosen dataset?
        - *Provide a brief description of the dataset, including its origin and the problem domain it addresses.*
2. **Feature Description**
    - **Question**: What are the features (variables) present in the dataset? Is there a target variable?
        - *List all the features, their data types (e.g., numerical, categorical), and describe their significance.*
3. **Dataset Structure**
    - **Question**: What is the size and structure of the dataset?
        - *Mention the number of rows and columns, and any hierarchical structure if applicable.*
4. **Missing Values and Duplicates**
    - **Question**: Are there missing values or duplicates in the dataset?
        - *Identify any missing or duplicate entries and discuss how they might affect your analysis.*
5. **Statistical Summary**
    - **Question**: What is the statistical summary of the dataset?
        - *Compute summary statistics like mean, median, standard deviation, and provide initial insights.*
6. **Data Distribution**
    - **Question**: How are the features distributed?
        - *Use visualizations like histograms or box plots to show the distribution of key features.*
7. **Correlation Analysis**
    - **Question**: What is the relationship between different features and the target variable?
        - *Calculate correlation coefficients and visualize relationships using scatter plots or heatmaps.*
8. **Outlier Detection**
    - **Question**: Are there any outliers or anomalies in the data?
        - *Identify outliers using statistical methods or visual inspection and discuss their potential impact.*

Part 2: Data Preprocessing

9. **Handling Missing Data**
    - **Question**: How will you handle missing or anomalous data?
        - *Explain your strategy for dealing with missing values (e.g., imputation, deletion) and justify your choice.*
10. **Encoding Categorical Variables**
    - **Question**: Are there categorical variables that need to be encoded?
        - *Describe the encoding techniques you will use (e.g., one-hot encoding, label encoding).*
11. **Feature Scaling**
    - **Question**: Should the data be scaled or normalized?
        - *Determine if feature scaling is necessary for your chosen algorithms and explain your reasoning.*
12. **Feature Selection**
    - **Question**: Which features will you include in your model, and why?
        - *Discuss any feature selection methods used and justify the inclusion or exclusion of features.*

Part 3: Modeling

13. **Algorithm Selection**
    - **Question**: Which machine learning algorithms are appropriate for your task, and why?
        - *Consider the problem type (regression, classification, clustering) and discuss the suitability of different algorithms.*
14. **Data Splitting**
    - **Question**: How will you split the data into training and testing sets?
        - *Explain your method for dividing the data (e.g., hold-out method, cross-validation) and the rationale behind it.*
15. **Model Training**
    - **Question**: How will you train your model?
        - *Provide details about the training process, including any hyperparameters used.*
16. **Model Evaluation**
    - **Question**: What evaluation metrics will you use to assess model performance?
        - *Choose appropriate metrics (e.g., accuracy, precision, recall, RMSE) and explain why they are suitable.*

17. **Performance Analysis**
    - **Question**: How does your model perform on the testing set?
        - *Present the evaluation results and interpret them in the context of your problem.*
18. **Model Improvement**
    - **Question**: Can you improve the model's performance? If so, how?
        - *Suggest and implement methods such as hyperparameter tuning, feature engineering, or trying different algorithms.*
19. **Validation**
    - **Question**: How do you validate your model to ensure it generalizes well?
        - *Discuss techniques like cross-validation or using a validation set.*
20. **Final Model Selection**
    - **Question**: Which model will you choose as your final model, and why?
        - *Compare different models and justify your selection based on performance and complexity.*

Part 4: Visualization

21. **Data Distribution**
    - **Question:** How is the data distributed across different features?
        - Visualize the distribution of numerical features (e.g., histograms, boxplots) and assess any patterns, outliers, or anomalies. For categorical features, use bar plots or count plots.
22. **Feature Importance**
    - **Question:** What are the most important features in your model?
        - After training your model, visualize feature importance using bar charts (e.g., for tree-based models) or coefficients (e.g., for linear models).
23. **Model Performance Across Features**
    - **Question:** How does the model perform across different subsets of features or data?
        - Use visualizations plots to show how different features impact model predictions.
