Final Machine Learning Project (Part 2)

Model Implementation and Evaluation

Vision: Extend insights gained from Exploratory Data Analysis (EDA) in Part 1 by developing, training, and evaluating predictive and clustering models to address specific business questions or challenges.

Phase 1: Supervised Learning

Objective: Build and evaluate predictive models to answer a business-oriented hypothesis.

Tasks:

1. **Model Selection:**
    - Choose the models you will test for the task (e.g., Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVMs)).
    - Explain why you selected these models and how they align with your dataset and problem.
    - Before testing, predict which model you expect to perform best and justify your reasoning based on the data and task characteristics.
2. **Training Models:**
    - Split your data into training and testing subsets.
    - Train your models using the selected algorithms.
    - Perform hyperparameter tuning (e.g., GridSearchCV) to optimize the performance of each model.
3. **Evaluation:**
    - Evaluate each model using metrics relevant to your problem (e.g., accuracy, precision, recall, F1 score, ROC-AUC).
    - Justify why these metrics are important for your specific use case.
    - Visualize the results using tools such as confusion matrices, ROC curves, and performance comparison graphs.
    - Select one final model based on your evaluation. Report its results and discuss any shortcomings.
    - If the results are not satisfactory, explain what could be improved or what might be missing to achieve better performance.

Phase 2: Unsupervised Learning

Objective: Discover hidden patterns and relationships in the data. unsupervised learning can reveal underlying structures, relationships, or clusters within the data that may provide new insights into the features or help improve the predictive power of supervised models.

Tasks:

1. **Clustering Analysis:**
    - Predict which clustering model you expect to perform best and justify your reasoning based on the data and task characteristics.
    - Apply clustering algorithms such as K-Means, Agglomerative Clustering, and DBSCAN.
    - Experiment with different parameters (e.g., number of clusters for K-Means or eps and min_samples for DBSCAN).
    - Describe the features used for clustering and why they are suitable.
2. **Clustering Analysis:**
    - Visualize the clusters and interpret their meaning.
    - Use evaluation metrics such as Silhouette Score and Davies-Bouldin Index to assess clustering performance.
    - Compare the discovered clusters with the original labels (if available): Identify whether clusters align with the labels or if they reveal unexpected subgroups.
    - Use visualizations to explain the relationships between features and clusters.
3. **Business Context:**
    - Relate the discovered clusters to the business problem.
    - Propose actionable insights based on the clustering results (e.g., strategies for targeting specific subgroups, redesigning features, or understanding anomalies).
    - Discuss how these insights can enhance supervised tasks or provide additional value to the business.

Expected Deliverables:

1. Code Submission:
    - A structured notebook (e.g., Jupyter or Colab) that includes all the code used in the project.
    - The notebook should clearly demonstrate the process of dataset exploration, cleaning, hypothesis formulation, and any conclusions drawn.
2. Summary Report:
    - A written document summarizing all the steps taken, the insights derived, and conclusions reached.
    - Include visualizations to support and illustrate key findings.
    - The report should be detailed enough for an external reader to understand the process and outcomes fully.

*** The final project submission (including the structured report and code) is due by midnight, May 8th.**