Final Machine Learning Project (Part 1)

Exploratory Data Analysis and Hypothesis Development

Vision: The ultimate goal of this project is to apply machine learning techniques to address a real-world business problem. Students will integrate machine learning concepts with critical thinking to solve practical challenges organizations face, leveraging data-driven insights to make informed decisions and propose impactful solutions.

Phase 0: Finding Datasets

Objective: Students are encouraged to choose a dataset that aligns with their interests or professional goals.

Tasks:

1. **Choose publicly available datasets from:**
   - Kaggle Datasets https://www.kaggle.com/datasets
   - HuggingFace
   - UCI Machine Learning Repository
   - Google Dataset Search https://datasetsearch.research.google.com/
   - Awesome Public Datasets – A curated list of high-quality datasets. https://github.com/awesomedata/awesome-public-datasets?tab=readme-ov-file
2. **Dataset Requirements:** To keep the project manageable and focused, the dataset must meet the following criteria:
   - Open and Accessible: The dataset must be publicly available without licensing restrictions.
   - Label Column: The dataset must have a clearly defined target variable (label) that can be used for prediction.
   - Number of Features: The dataset must contain at least 10 features (columns), excluding the label column.
   - The features are diverse (e.g., categorical, numerical) to allow for meaningful EDA and preprocessing.
   - Dataset Size: Minimum of 500 rows to ensure there's enough data for meaningful insights.

- Cleanliness: While some missing data or anomalies are expected, avoid datasets that are overly noisy or incomplete (e.g., more than 50% missing data).
- The dataset should relate to a business or real-world scenario that can inspire a hypothesis—examples: Customer churn prediction, student performance analysis, sales forecasting.

*** If you're unable to find a dataset that suits your interests, here are some generic suggestions that are always a good fit:**

- Students Performance in Exams: Analyze factors affecting student success.https://www.kaggle.com/datasets/spscientist/students-performance-in-exams
- Heart Disease UCI: Predict the likelihood of heart disease based on medical attributes. https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data
- Supermarket Sales: Explore sales data from a supermarket chain to uncover patterns. https://www.kaggle.com/datasets/aungpyaeap/supermarket-sales

s

Phase 1: Data Exploration and Cleaning

Objective: Gain a comprehensive understanding of the dataset by analyzing its structure, identifying key patterns, and preparing it for downstream tasks. Use statistical tools and visualizations to uncover trends, anomalies, and correlations.

Tasks:

1. **Understand Your Data:**
    - Inspect Data Types: Distinguish between numerical, categorical, and boolean variables.
    - Identify the roles of variables: target variable vs. predictors (features).
    - Summarize the Dataset: Report the number of rows, columns, missing values, and unique values in categorical features.
    - Evaluate the balance of the dataset
    - Check Correlations: Use correlation heatmaps for numerical variables.
2. **Clean and Prepare:**
    - Normalize numerical features using StandardScaler or MinMaxScaler.
    - Apply label encoding or one-hot encoding for categorical variables.
    - Use box plots to identify outliers and decide whether to remove or transform them (e.g., using winsorization).
3. **Handle Missing Data:**
    - Identify Missing Data Patterns: Categorize missing data as MCAR (Missing Completely at Random), MAR (Missing at Random), or MNAR (Missing Not at Random).
    - Apply imputation techniques such as mean, median, or regression imputation.
    - Justify your chosen imputation method based on the nature of the data.
4. **Visualize and Analyze:**
    - Use box plots, histograms, and scatter plots to identify patterns, outliers, and correlations.
    - Analyze the shape of feature distributions (e.g., skewness, normality) and describe insights.

- Identify clusters or patterns in feature relationships using scatter plots or pair plots.

5. **Feature Engineering:**
    - Combine features to create meaningful interactions (e.g., attendance × participation).
    - Aggregate temporal data or calculate rolling statistics if applicable.
    - Use domain knowledge to craft meaningful derived variables.

Phase 2: Hypothesis Formulation

Objective: Leverage insights gained from the exploratory phase to formulate a business-relevant question → model.

Tasks:

1. **Frame a Business-Oriented Hypothesis:**
    - Frame a Business-Oriented Hypothesis: Use findings from EDA to identify areas of interest, such as:
        - Unique behaviors or patterns in the data.
        - Strong correlations between predictors and the target variable.
        - Clusters or anomalies worth exploring.
    - Define a clear question with actionable value - Explain the significance and relevance of the question using insights from EDA. Examples:
        - What factors drive customer retention in a subscription-based business?
        - How do different product features affect sales volume?
        - Can purchasing patterns predict the likelihood of a customer making a repeat purchase?
2. **Identify the Most Important Trends, Relationships, and Features:**
    - Assess interaction effects between features and the target.
    - Identify critical features predicted to influence the target (feature importance)
    - Explore correlations and clusters to find hidden patterns or relationships.
    - Visualize findings using bar charts, line plots, or other relevant methods.

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.