Project 24-25: (Groups of 2)

Exploratory Data Analysis and Modeling on Web-Scraped Datasets: A Comparative Study of Supervised Learning Techniques

Abstract:

In this project, students will use web scraping techniques (e.g., Selenium, BeautifulSoup) to collect datasets directly from web pages. The project aims to provide a comprehensive understanding of the dataset through exploratory data analysis (EDA), preprocessing, and subsequent application of supervised learning techniques. Datasets obtained through APIs are explicitly excluded to encourage proficiency in direct web scraping methods.

Objectives:

1. Collecting datasets from dynamic or static websites using web scraping tools while adhering to ethical guidelines.
2. Conducting exploratory data analysis to gain insights into the dataset's structure, distributions, and relationships between variables.
3. Preprocessing the dataset to address issues such as missing values, outliers, and feature engineering.
4. Applying supervised learning techniques, including regression and classification, to train predictive models and make informed predictions based on labeled data.

Tasks:

1. **Web Scraping for Data Collection:**
   Each student must identify a website with publicly accessible data that aligns with their project goals.
   - Use tools such as **Selenium** for dynamic web pages or **BeautifulSoup** for static pages to extract relevant data.
   - Ensure compliance with the website’s terms of service and legal restrictions on data collection.
   - Provide documentation on the scraping process, including site selection, scraping tools, and challenges encountered.

2. **Exploratory Data Analysis (EDA):**
   - Conduct descriptive statistics, data visualization, and correlation analysis to understand the dataset's characteristics and identify potential patterns or trends.
   - Highlight any issues in the raw data collected (e.g., duplicate entries, inconsistencies) and propose solutions.

3. **Data Preprocessing:**
    - Handle missing values, outliers, and categorical variables through appropriate techniques such as imputation, scaling, encoding, and feature engineering.
    - Document preprocessing steps, emphasizing adjustments made due to the raw nature of web-scraped data.

4. **Supervised Learning:** (Apply 2 methods)
   a. **Regression:** Apply linear regression or other regression algorithms to predict continuous target variables.
   b. **Classification:** Employ classification algorithms such as logistic regression, ANN, or any other technique that you find interesting such as decision trees, or support vector machines to classify data into predefined categories.

5. **Model Evaluation:**
    - Assess the performance of supervised learning models using evaluation metrics such as mean squared error (MSE) for regression or accuracy, precision for classification.

6. **Comparative Analysis:**
    - Compare the outcomes of the several learning approaches that you used, discussing strengths, weaknesses, and insights gained from each method.
    - Use visual tools such as confusion matrices and boxplots to facilitate comparisons.

7. **Documentation and Presentation:**
    - Document the entire project, including data scraping, exploration, preprocessing steps, model building, evaluation results, and conclusions.
    - Present findings through written reports and presentations.

N.B:

- The trained models should be optimized (accuracy and loss) and saved (e.g., in `.h5` format).
- You are invited t o enhance the project with additional creative elements (e.g., add unsupervised learning techniques and compare them with the supervised ones, real-time data collection, dynamic visualizations, etc.).

Special Note:

- Ensure the scraping process aligns with ethical data usage practices, explicitly avoiding copyrighted or private content.