DATA ANALYSIS AND DATA SCIENCE WITH PYTHON

Task 5: Classification Tasks Overview

Task 1: Student Pass/Fail Prediction

Objective

Predict whether a student will pass or fail based on their study hours and attendance.

Project Steps

1. **Dataset Selection**:
    1. Create or select a dataset with columns:
        - `Study Hours`: Number of hours a student studies per week.
        - `Attendance`: Percentage of classes attended.
        - `Pass`: Binary column indicating pass (`1`) or fail (`0`).
2. **Tasks to Perform**:
    1. **Data Exploration**:
        - Check for missing values or outliers.
        - Plot the relationship between `Study Hours`, `Attendance`, and `Pass` to visualize trends.
    2. **Model Training**:
        - Train a **Logistic Regression** model using `Study Hours` and `Attendance` as features and `Pass` as the target variable.
    3. **Model Evaluation**:
        - Evaluate the model using:
            - **Accuracy**: Proportion of correctly classified instances.
            - **Confusion Matrix**: Breakdown of True Positives, True Negatives, False Positives, and False Negatives.

Deliverables

1. **Classification Model**:
    - Logistic Regression model trained on the dataset.
2. **Evaluation Metrics**:
    - Accuracy and confusion matrix results.
3. **Insights**:
    - Key predictors of student performance.

Task 2: Sentiment Analysis with Natural Language Processing

Objective

Analyze customer reviews to classify sentiments as positive or negative.

Project Steps

1. **Dataset Selection**:
    1. Use a dataset like `reviews.csv` with columns:
        - `Review Text`: Customer reviews.
        - `Sentiment`: Sentiment label (`positive` or `negative`).
2. **Tasks to Perform**:
    1. **Load and Preprocess the Dataset**:
        - Preprocessing steps:
            - Remove stopwords, punctuation, and special characters.
            - Convert text to lowercase.
            - Tokenize and lemmatize the text.
    2. **Text Vectorization**:
        - Convert text to numerical format using **TF-IDF (Term Frequency-Inverse Document Frequency)**.
    3. **Model Training**:
        - Train a **Logistic Regression** model to classify sentiments based on the vectorized text.
    4. **Model Evaluation**:
        - Evaluate using:
            - **Accuracy**: Overall performance of the model.
            - **Precision**: How many predicted positives are actually positive.
            - **Recall**: How many actual positives were predicted correctly.
            - **F1-Score**: Harmonic mean of precision and recall.

Deliverables

1. **Preprocessed Dataset**:
    - Cleaned and tokenized text data.
2. **Sentiment Classification Model**:
    - Logistic Regression model trained to classify sentiments.
3. **Evaluation Report**:
    - Accuracy, precision, recall, and F1-score results.
4. **Insights**:
    - Examples of reviews classified correctly and incorrectly.
    - Common features of positive and negative reviews.

General Guidelines

- **Tools and Libraries**:
  - For Data Analysis: `Pandas`, `NumPy`, `Matplotlib`, `Seaborn`
  - For Preprocessing: `NLTK`, `spaCy`, or `scikit-learn` preprocessing utilities
  - For Modeling: `scikit-learn`
- **Testing**:
  - Use a train-test split (e.g., 80-20) for both tasks.
  - For the sentiment analysis task, consider using cross-validation for robust evaluation.
- **Documentation**:
  - Provide clear steps for preprocessing, model training, and evaluation.
  - Include visualizations (e.g., confusion matrix, feature importance) for better understanding.

Would you like me to assist with specific steps, such as preprocessing guidelines, evaluation metrics, or feature engineering ideas?

Deadline Compliance

- **Restriction**: **Submit the project within 7 days** from the start date.
- **Reason**: Meeting deadlines is crucial in the real-world software development environment. This restriction helps students practice **time management** and **task prioritization**. In professional settings, tight deadlines are often the norm, and learning to meet them without compromising quality is an essential skill.
- **Learning Outcome**: Students will learn to manage their time effectively, complete projects under pressure, and **deliver results on time**, which are all important skills in the workplace.