Birzeit University
Department of Electrical & Computer Engineering
First Semester, 2024/2025
ENCS5141 Intelligent Systems Lab
Assignment 1
Due Date November 25, 2024

Assignment Description:

This assignment consists of two parts. In the first part, you will focus on **Data Cleaning and Feature Engineering** using the **attached Dataset**. In the second part, you will perform a **Comparative Analysis of Classification Techniques**, comparing **Random Forest (RF), Support Vector Machine (SVM),** and **Multilayer Perceptron (MLP)** for the data set.

Part 1: Data Cleaning and Feature Engineering for the Bike Sharing Dataset

Objectives:

- To explore and preprocess the dataset by addressing missing values, encoding categorical variables, and scaling numerical features.
- To apply feature selection and dimensionality reduction techniques for effective data preparation.
- To evaluate the impact of preprocessing on model performance compared to using raw data.

Procedure:

- **Data Exploration**: Summarize dataset statistics to understand the structure, features, and any missing values.
- **Data Visualization:** Explore relationships between features by generating various visualizations, including box plots, scatter plots, and histograms, to reveal underlying patterns and insights.
- **Data Cleaning**:
  - Address missing values and outliers by selecting appropriate imputation methods or removing irrelevant data.
  - Apply necessary transformations to ensure data quality. what do we mean exactly?
- **Feature Engineering**:
  - Analyze the relevance of each feature for the machine learning task.
  - Encode categorical variables into numerical formats (e.g., one-hot encoding).
  - Scale or normalize numerical features for consistency.

        - Use dimensionality reduction techniques to reduce data dimensionality while retaining important information. LDA, PCA or IG, Variance and Correlation??
- **Model Evaluation**:
    - Split the dataset into training and testing subsets.
    - Train a Random Forest model on the preprocessed data.
    - Compare the performance of the model trained on preprocessed data vs. raw data (before applying feature selection and scaling).

Experiments:

- Compare the results of a Random Forest model trained on the preprocessed data versus the raw data.
- Analyze the effect of various preprocessing techniques on model performance using metrics such as accuracy, precision, and recall.

Results:

- Summarize the performance of the model on preprocessed vs. raw data.
- Show improvements in model accuracy, consistency, and training speed due to preprocessing.

Part 2: Comparative Analysis of Classification Techniques

Objectives:

- To compare the effectiveness of Random Forest (RF), SVM, and Multilayer Perceptron (MLP) models.
- To study the effect of parameter tuning for each model.

Background:

In this part, we aim to assess the performance of three classification techniques—RF, SVM, and MLP—on the dataset from Part 1. Each model has different strengths in terms of handling data complexity and noise. A thorough comparison of their performance helps in understanding which model is better suited for predicting varying levels of bike demand.

Procedure:

- **Data Preparation**: Use the same preprocessed dataset from Part 1.
- **Model Training**:
    - Train three models: Random Forest (RF), SVM, and Multilayer Perceptron (MLP).
    - Evaluate the models using training and testing datasets.
- **Model Comparison**: Analyze the models in terms of:
    - Accuracy, precision, and recall.
    - Computational efficiency and training time.
- **Effect of Preprocessing**: Investigate how data cleaning, and feature engineering impact model performance.

- **Effect of Model Parameters**: Investigate how parameters affect model performance. Try different combinations of parameters.

Experiments:

- Compare the performance of RF, SVM, and MLP on the same classification problem.
- Measure each model's classification accuracy, precision, recall, and F-Measure.
- Measure how model parameters affect performance.

Results:

- Provide a summary of model performances, discussing the strengths and weaknesses of each model.
- Compare the models based on their classification performance and computational efficiency (exectuion time and memory used).
- Conclude which model provides the best balance between prediction accuracy and computational time for this dataset.

Deliverables

- You need to submit the code in .ipynb format. You can obtain this file in Google Colab by navigating to the File menu and selecting Download > Download .ipynb.
- Additionally, write a report detailing the case study. Ensure adherence to the report preparation guidelines outlined in the "ENCS5141 Case Study Report Guidelines.pdf" document. If you opt to write the report using LaTeX, utilize the provided report template "ENCS5141 Sample Report.tex"."