BIRZEIT UNIVERSITY

Department of Electrical and Computer Engineering
Summer Semester, 2023/2024
Intelligent Systems Lab, ENCS5141
Case Study #1: Data Cleaning and Feature Engineering for the Bike Sharing Dataset

In this case study, you will perform essential data preprocessing steps on the **modified version of the Bike Sharing Dataset**. The bike-sharing rental dataset includes a two-year historical log (2011-2012) from Capital Bikeshare in Washington D.C., aggregated on an hourly basis, with additional weather and seasonal information. The dataset comprises several fields such as the date, season, year, month, hour, whether the day is a holiday or not, day of the week, whether the day is a working day, and count of total rental bikes, among others. A detailed description of the data and the fields can be found in the ReadMe file.

Follow these steps:

1. You may download the dataset using the following link https://github.com/mkjubran/ENCS5141Datasets/tree/main/ENCS5141_BikeSharingDataset_Modified
2. Perform initial data exploration to understand the dataset's structure, features, and any missing values. Summarize the dataset's statistics and gain insights into the data.
3. Address any data quality issues, such as missing values and outliers. Decide on an appropriate strategy for handling missing data, such as imputation or removal of rows/columns.
4. Analyze the relevance of each feature for your machine learning task by using feature selection techniques.
5. If the dataset contains categorical variables, encode them into a numerical format suitable for machine learning models.
6. Split the dataset into training and testing subsets to evaluate the performance of your machine learning models.
7. Scale or normalize the numerical features to ensure consistent scaling across variables.
8. Apply suitable dimensionality reduction techniques to reduce the size of the data while preserving important information.
9. Validate your preprocessing pipeline by training and evaluating a machine learning model, such as the Random Forest model, on the preprocessed data.
10. Compare the results to the model trained on the raw data (before feature filtering, transformation, and reduction) to ensure that preprocessing has improved model performance.

Submissions:

- 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"."

Important notes:

- Make sure to add descriptive comments and headings using markup language, such as Markdown, in your Google Colab notebook or Jupiter notebook.
- Deadline: Firday, 2 August 2024 at 11:59 pm. Please submit your case study solution and report through Ritaj as a reply to this message.