IML'22 Project: Divorce Predictors Dataset

Background. When we talk about relationships between men and women, we usually refer to marriage. However, how can we identify a good relationship? Predicting divorce has been an area of interest for researchers and practitioners for many years, as it has significant social and economic implications. One of the main reasons for predicting divorce is to identify couples who may be at risk of divorce so that interventions can be put in place to help them address any issues and improve their relationship. There are several potential benefits to using machine learning to predict divorce. For example, machine learning algorithms can identify risk factors for divorce that may not be immediately obvious to human analysts.

Dataset. The Divorce Predictors Dataset is a publicly available UCI Machine Learning Repository dataset. A total of 170 couples are included in this dataset, along with the Divorce Predictor Scale variables (DPS), a questionnaire of 54 questions, based on the Gottman couples therapy. Records were collected from face-to-face interviews with couples who were already divorced (49%) or happily married (51%) in various regions of Turkey. A five-point scale was used to measure all responses (0=Never, 1=Seldom, 2=Average, 3=Frequently, 4=Always). The dataset can be downloaded from here.

Requirements.

1. Formulate the machine learning problem by identifying the task (classification or regression), data, and challenges.
2. Perform Exploratory Data Analysis. This includes:
    - Plot relationship between variables
    - Identify correlated features or the ones that highly correlated with the label/outcome, if any
    - Perform data imputation for missing variables
    - Encode your outcome to a one-hot vector, if needed
    - Remove redundant variables by dimensionality reduction techniques
3. Consider splitting the data into Training, Validation, and Testing sets. The suggested split is 70%, 10%, and 20% held-out testing set, respectively. (same split for all tasks)
4. Build and develop the following models (tasks):
    - **Task#1:** Clustering in an unsupervised fashion
    - **Task#2:** Logistic/Linear Regression according to your formulation in step 1
    - **Task#3:** Neural Network for Classification/Regression according to your formulation in step 1

5. For each task, run the model *with* and *without* processing the data, e.g., without normalization or dimensionality reduction, and compare the model's performance after you normalize and/or reduce the dimensionality of the data.
6. For each task, show how you performed the *model selection*. For example, demonstrate the performance of variants of your model with different hyper-parameters, e.g., number of clusters and initialization when it comes to clustering methods.
7. For each task, perform a 5-fold Cross Validation.
8. For each task, run the corresponding evaluation metrics on each fold to demonstrate the performance of your model on the held-out testing set

Submission.

- **Prepare** a well-documented Jupyter notebook demonstrating the following sections:
    - Your name along with your Uni-ID and task assignment within the group
    - The background of the given task
    - Sample of your dataset
    - The rationale behind your ML formulation
    - The exploratory data analysis and data split
    - The developed models (Task #1-#3) including the model selection (hyper-parameters) and their results
    - A brief comparison between different models + concluding remarks
    - Lessons learned from this project
- **Consider** the following naming convention when you name your Jupyter notebook, `DivorcePredictorsDataset_G##`, and consider replacing ## with your group number, e.g, if your group number is 5, then the name of your file should be `DivorcePredictorsDataset_G05.ipynb`
- **Submit** the Jupyter notebook to the following folder https://uni-bonn.sciebo.de/s/tO0R2lXI1RamUWB which allows you to only upload files. You can upload the same file multiple times in case updates/modifications were required.
- **The submission deadline** is 18th Jan. 2023 at 11:59 PM

Grading.

- **Individual-level grading (65%) + Group-level grading (35%)**
    - The problem addressed (clearly stated and understood) – *Group-level*
    - ML problem formulation (rationale) – *Group-level*
    - Exploratory Data Analysis – *Group-level*
    - Task Completion and Findings – *Individual-level*
    - Interpretation and conclusion – *Individual-level*
    - Limitations – *Group-level*
    - Documentation, Submission, and Presentation – *Group-level*