CIS600 Fundamentals of Data and Knowledge Mining

HW1: Data Preparation and Exploratory Data Analysis

Due: 11:59pm EST, Oct. *26th, 2022*

Homework instructions

- Analyze *employee_attrition.csv* dataset provided. This dataset provides a variety of information about the employees, such as demographics, time on job, etc. and also if they stay with or leave the company (as in binary attribute *"Attrition"* with *No* means stay and *Yes* means leaving).
- Follow CRISP-DM process and conduct Exploratory Data Analysis (EDA)
  - Data preprocessing and cleaning: identify potential data quality issues and properly address those issues as part of data preparation.
  - Data transformation and preparation: discuss all the data transformation techniques you apply to the dataset for both features and records (e.g. discretization/numerization, normalization/standardization) and carry out those transformation operations.
  - Conduct exploratory data analysis (EDA): derive descriptive statistics and apply data visualization to check for interesting data patterns. Particularly look for any correlation between attributes with the target class labels.
- Use Jupyter Notebook to structure your report and submit the html output
  - All the codes and relevant outputs (limit the size of outputs to only include those relevant contents and refrain from printing out excessive amount of irrelevant information or data)
  - Analysis writeup using markdown language (interpretation and discussion of the results with the proper session titles and all the information useful to grade your work)

Grading rubrics

- Include all the key data preparation and preprocessing steps which are neatly structured in the report with all the Python codes and relevant outputs (using proper section titles) (50%)
- In-depth interpretation of the analysis results output using markdown language (30%)
- Effective and proper visualization with multiple techniques to illustrate any noteworthy data patterns (20%)

Submission instructions:

- Submit the html report to the 2U LMS
- Late submission policy: Late submission will incur 10% penalty for every additional 24 hours' delay until all points are deducted