Data Mining: Project Proposal

Authors: Leona Hasani / Leona Hoxha / Nanmanat Disayakamonpan

Project Ideas:

Project Title: ***"Comparative Analysis of Imputation Techniques in Australian Rainfall Data"***

- In real life, there is no such thing as a "perfect data set." Instead, there are many problematic ones, like those with lots of missing values. Therefore, our main focus is to explore and compare various imputation methods, including statistical techniques, deletion of rows or columns, and machine learning-based approaches for this kind of data set.
- As our data set, we chose 'Rain in Australia', which is a large weather dataset from Kaggle with lots of missing values in different columns and rows. This data set has about 10 years of daily weather observations from many locations across Australia, where each row represents a particular day with weather information such as wind speed, MaxTemp/MinTemp, humidity, and the target variable being "RainTomorrow", filled by 1 (if tomorrow is rainy) and 0 (if tomorrow is not rainy, then 0).

Implementation Plans:

- **Literature Review:** According to the scientific papers that we have found, they discuss and compare different approaches to imputing missing values. As mentioned, we would like to explore statistical methods, deletion strategies, and machine-learning techniques applied in similar contexts. Each of the methods that will be used from the different scientific papers will be cited and added as a reference in the final project.
- **Data Preprocessing:** We will identify missing values, implement various imputation techniques (mean/mode, KNN, EM, MICE, Hot Deck), and prepare them for machine learning modeling.
- **Imputation Methods:** We will implement various imputation methods, including statistical approaches, deletion of rows/columns, and machine learning models. We utilize tools such as Python and relevant libraries (e.g., scikit-learn, pandas) for implementation.
- **Classification Model Application:** We will employ various classification models such as logistic regression, decision tree, random forest, SVM, ANN, KNN, etc. Another goal of this project is to see how different imputation techniques will perform in the various supervised machine-learning algorithms.
- **Evaluation:** We will also assess the model's performance (accuracy, precision, recall, F1-score, AUC score) for each of the imputation methods in order to compare the effectiveness of different techniques. After that, we visualize and interpret the results.

Purpose of the Project:

- The significance of this project lies in addressing the challenge of missing values in weather data, a common issue in real-world datasets.
- By leveraging scientific literature and experimenting with diverse imputation methods, we aim to enhance our understanding of handling missing data effectively by critically evaluating and considering other factors such as accuracy, time consumption, and computational costs on the performance of each imputation approach. The project contributes to the broader goal of improving data quality and reliability in predictive modeling.

Learning Outcomes:

- We expect to understand various imputation techniques, statistical methods, and machine learning models.
- We expect to gain an overview of how different imputation approaches perform, which helps us evaluate and interpret the performance of different imputation strategies through accuracy scores and visualizations.
- We expect to gain proficiency in data analysis and modeling through Python and have hands-on experience from what we learn in the data preprocessing steps, which is essential for data science skills.
- We expect this project to go beyond academic purposes, but we also expect to offer insights for real-world applications such as agriculture, disaster management, and daily life planning.