RUTGERS
THE STATE UNIVERSITY
OF NEW JERSEY

FINAL PROJECT DESCRIPTION

Automation Lab (540:383)

Fall 2024

I. Task description

Each group will be assigned one of two datasets: Housing Data (Regression Task) or Heart Failure Data (Classification Task). The task involves performing data analysis and modeling following the outline provided in Section 3. Deliverables include a Python notebook with the complete analysis, visualizations, and model implementation, along with a 10-slide PowerPoint presentation summarizing key insights and model performance.

Deadline: Sunday 12/08/2024

II. Goals

By completing this project, you will develop the ability to build and implement essential machine learning models, focusing on the two most common types in supervised learning: Regression and Classification. Additionally, you will enhance your data analysis skills by working with real-world datasets and further refine your Python programming expertise, particularly in the context of Data Science.

III. Project guidelines

1. Problem Definition

- Define the objective clearly. Understand the business problem or the research question you're trying to answer. This ensures the analysis is aligned with the goal.
- Define the type of task you want to perform: Regression/ Classification/Clustering and why?

2. Exploratory Data Analysis (EDA)

- Investigate the dataset through summary statistics, data visualization, and identifying relationships, patterns, or trends.
- Feature Classification
- Detect missing values, outliers, and distributions.
- Formulate hypotheses for further analysis.

3. Data Preprocessing

- Data Cleaning: Handle missing values, duplicates, outliers, and erroneous data.
- Data Transformation: Normalize, scale, or encode categorical variables.
- Feature Engineering: Create new features or variables that may improve the model.

4. Feature Selection

- Reduce the dimensionality by selecting the most relevant features based on domain knowledge or statistical techniques (e.g., correlation analysis, feature importance).
- Choose data for modeling (Input and Output)

5. Modeling

- Data Splitting: Split data into training, validation, and testing sets

- Choose appropriate machine learning models or algorithms based on the problem (e.g., regression, classification, clustering).
- Train the model using the training data.
- Obtain and show models, interpret the results

6. Model Evaluation

- Assess model performance using appropriate metrics (e.g., accuracy, precision, recall, F1-score for classification; RMSE, MAE for regression).
- Compare models based on evaluation metrics and select the best-performing model.
- Cross-validation to ensure the model generalizes well.

7. Model Interpretation and Insights

- Interpret the model's results and make sure they make sense in the context of the problem.
- Explain key drivers or features influencing the model.

IV. Grading rubric

- 7-Step Python Notebook & 10-Slide Presentation: 50%
- Error-Free Execution: 10%
- Consistency and Cohesiveness: 15%
- Accurate Interpretation and Conclusion: 15%
- Valuable Insights Extracted from Data: 10%

Good Luck!