Course code: CSE381
Course Name: Introduction to Machine Learning
Instructor: Dr. Mahmoud Khalil
TA: Eng. Engy Ahmed Hassan

Introduction to Machine Learning

Project- Fall 2024

1. Project Description

Heart failure is a leading cause of mortality worldwide, making early detection and risk assessment crucial for improving patient outcomes. This project utilizes the Heart Failure Prediction Dataset, which contains clinical and demographic information about patients, including features such as age, cholesterol levels, resting blood pressure, and maximum heart rate achieved, along with a target variable indicating the presence of heart disease.

The project aims to explore the dataset, uncover patterns through visualization, and develop machine learning models to predict heart failure risk. The analysis begins with data cleaning and preprocessing to prepare the dataset, followed by training and evaluating classifiers, including Naïve Bayes, SVM, KNN, and Decision Trees. Additionally, a dendrogram will be constructed to visualize hierarchical clustering relationships, providing insights into how patient groups are naturally formed based on their attributes.

This is a group project where each group should consist of 3 students

2. Project Breakdown

1. **Data Exploration and visualization:**

   Doing some descriptive analysis for the dataset and visualizing data by reducing the dimensionality of the samples to be 2D using PCA and comment on the results

2. **Data cleaning and processing:**

Appropriate steps to prepare data and get it ready for the rest of the pipeline.

- Handling missing values if exists
- Dealing with Outliers
- Dealing with duplicates
- Splitting data into training, validation and testing data

3. Training Classifiers:

- Train classifiers including **Naïve Bayes**, **Support Vector Machines (SVM)**, **K-Nearest Neighbors (KNN)**, and **Decision Trees**.
- Experiment with different configurations of hyperparamaters.
- Use grid search, random search, or manual tuning methods as needed.
- Plot **training accuracy** and **validation accuracy** across various hyperparameter configurations.
- Comment on the observed performance trends.

4. Testing

- Compute **recall**, **precision**, **F1-score**, and **confusion matrix** for each classifier.
- Compare the performance of different classifiers.

5. Dendrogram Analysis

- Create a dendrogram to illustrate the hierarchical clustering of data points.
- Use proximity measures (e.g., dissimilarity or similarity) to determine when clusters merge.
- Comment on the dendrogram structure, such as the number of clusters and their proximity.
- Discuss how the hierarchical clustering insights align with the PCA and classifier results.

3. Milestone 1

This milestone includes Data Exploration and Visualization, Data Cleaning & Processing, Training Classifiers (Naïve Bayes and SVM) and Testing them.

Submission on LMS before 7/12/2024

4. Milestone 2

This milestone includes training Classifiers (KNN and Decision Tree) and Testing them. In addition to dendrogram analysis.

Submission on LMS before 21/12/2024

5. Project Deliverables

For each milestone you should submit

- **Zipped Folder:**

Include a zipped folder containing a Jupyter notebook with the code. Ensure the notebook is thoroughly documented with markdown cells and code comments. It should execute without errors or faulty outputs when run sequentially. The output for each code cell must be present in the submitted notebook.

- **Report:**

Detailed report containing steps, screenshots of the code, screenshots of the output, visualization of the accuracy change, the outputs and show the reason of using the final values of the hyperparameters.

Note that: In milestone 2 you should submit zipped folder for the **whole** project and report for the **whole** project

6. Marks Distribution

Data Exploration and visualization	**3**
Data cleaning and processing	**2**
Training Classifiers	**8**
Testing Classifiers	**4**
Dendrogram Analysis	**3**
Documentation	**5**