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

Introduction to Machine Learning

Project- Spring 2025

1. Project Description

Stroke is a leading cause of death and disability worldwide. Early prediction of stroke risk is crucial for timely medical intervention and reducing mortality rates. This project utilizes the **Stroke Prediction Dataset**, which includes a variety of clinical and lifestyle-related features such as age, gender, hypertension, heart disease, marital status, and more. The goal is to predict whether a patient is at risk of having a stroke and uncover patterns that define high-risk patient groups.

The project will involve exploring the dataset, preparing the data for machine learning, training several classifiers, and evaluating their performance. Additionally, the project will use unsupervised learning techniques like **K-Means** and **Hierarchical Clustering** to find natural clusters of patients and identify groups with higher risks.

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

2. Dataset Attribute Information

- ID: unique identifier
- Gender: "Male", "Female" or "Other"
- Age: age of the patient
- Hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- Heart Disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- Ever Married: "No" or "Yes"

- Work Type: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
- Residence Type: "Rural" or "Urban"
- Average glucose level: average glucose level in blood
- BMI: body mass index
- Smoking Status: "formerly smoked", "never smoked", "smokes" or "Unknown"*
- Stroke: 1 if the patient had a stroke or 0 if not

3. Project Breakdown

3.1. Data Exploration and Visualization

- Perform an initial descriptive analysis for the dataset.
- Visualize correlations using heatmaps and scatter plots.
- Use PCA, LDA and t-SNE to reduce data dimensions and visualize patient distributions in **2D**.

3.2. Data Cleaning and Preprocessing

- Handle missing values (if any) or incorrect values (e.g., zero values for age or BMI).
- Identify any outliers or trends that can affect model performance.
- Deal with categorical variables (e.g., gender, marital status) by one-hot encoding or label encoding.
- Split the data into training, validation, and testing sets

3.3. Training Classifiers

- Train classifiers including Naïve Bayes, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Decision Trees.

- Tune hyperparameters (using **Grid Search** and **Optuna Search**) to improve model performance.
- Plot training and validation accuracy across various hyperparameter configurations.

3.4. Testing Classifiers

- Compute recall, precision, F1-score, and confusion matrix for each classifier.
- Compare the performance of classifiers and identify which one works best for stroke risk prediction.

3.5. Clustering Analysis

- Apply Hierarchical Clustering to identify natural groupings of patients.
- Create dendrograms for hierarchical clustering.
- Apply K-Means Clustering to identify natural groupings of patients.
- Use clustering results to gain insights into the patterns of patients at risk for strokes.
- Compare clusters with the predicted labels from classifiers (e.g., comparing high-risk groups identified by classifiers vs. clustering).

4. 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 24/4/2025

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

5. Milestone 2

This milestone includes training Classifiers (KNN and Decision Tree) and Testing them.

In addition to clustering analysis.

Submission on LMS before 8/5/2025

6. 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

7. Marks Distribution

Data Exploration and Visualization	**2**
Data Cleaning and Preprocessing	**2**
Training Classifiers	**6**
Testing Classifiers	**5**
Documentation	**5**
Presentation	**5**