Alexandria University
Faculty of Computers and Data Science
Data Computation Spring 2025

Final Project

Project Overview:

In this project, teams of **3-5** members will collaborate to explore, analyze, and model a dataset using SVM to either predict an outcome (classification) or estimate a value (regression).

Tasks:

1. Dataset:

Find a dataset containing more than **10,000** records and at least **20** features.

2. Exploratory Data Analysis (EDA):

Conduct a thorough EDA to uncover patterns, anomalies, trends, and relationships within the data. Visualizations should be used to help understand the distribution of data and the relationships between features.

3. Data Cleaning:

This should cover issues like missing values, outliers, and inaccurate data entries.

4. Dimensionality Reduction:

Implement dimensionality reduction technique(s) covered in this course to reduce the number of features while retaining helpful information.

5. SVM Model Development:

Build an SVM model, focusing on either classification or regression. The model should be robust, and its parameters should be fine-tuned to get optimal performance. Evaluate the model using appropriate metrics.

Important Guidelines:

- Make sure to split the data into training and testing datasets BEFORE anything else avoid data leakage and ensure model generalization.
- Avoid dropping records unless it’s extremely necessary and this should be well documented and justified.
- You’re required to provide the complete model pipeline, from data preprocessing to final evaluation.

Bonus:

Deploy the trained SVM model using a framework such as Flask, FastAPI, or Streamlit.

Create a simple web-based user interface (UI) that allows users to:

- Upload or input data
- Receive model predictions

Discussion:

ALL team members should be present in the discussion.

- Early discussions: May 3rd – May 8th (optional if your team is ready early)
- Official period: May 10th – May 15th (available time slots to be announced)

You should have a notebook prepared that includes:

i. An overview of the dataset, explaining the types and nature of features.
ii. Insights and visualizations from the EDA.
iii. Dimensionality reduction technique(s) used.
iv. Modeling pipeline.
v. Hyperparameter tuning and model evaluation.