German International University of Applied Sciences
Informatics and Computer Science
Dr. Caroline Sabty
TA. Merna Said
TA. Sandra Samuel
TA. Amany Hussein
TA. May Magdy

Introduction to Data Science, Spring 2025
Assignment 2
Due date is April 24, 2025 at 11:59 PM

The **Teams will remain the same as in Milestone 1**, with three members per team. In this assignment, **you will continue working on the same dataset that was selected** and used in Milestone 1 for data preparation and exploration tasks: **Check your team number from CMS**

a) **Employee Attrition Prediction Dataset**:
Link: https://www.kaggle.com/datasets/ziya07/employee-attrition-prediction-dataset/data

b) **Heart Failure Prediction Dataset**:
Link: https://www.kaggle.com/datasets/endofnight17j03/heart-failure-prediction-dataset

c) **Obesity Prediction Dataset**:
Link: https://www.kaggle.com/datasets/adeniranstephen/obesity-prediction-dataset

d) **Warranty Claims Dataset**:
Link: https://www.kaggle.com/datasets/amanneo/df-cleancsv

e) **Medicine Quality Assessment Dataset**:
Link: https://www.kaggle.com/datasets/chaitanya205/medicine-quality-assessment-dataset

f) **Loan approval Dataset**:
Link: https://www.kaggle.com/datasets/suryadeepthi/loan-approval-dataset

g) **Cirrhosis Patient Survival Prediction Dataset**:
Link: https://archive.ics.uci.edu/dataset/878/cirrhosis+patient+survival+prediction+dataset-1

h) **Differentiated Thyroid Cancer Recurrence Dataset**:
Link: https://archive.ics.uci.edu/dataset/915/differentiated+thyroid+cancer+recurrence

Objective:

Using the preprocessed dataset from Milestone 1, apply four different machine learning algorithms to solve a classification problem and compare their performance.

Chosen Algorithms:

- K-Nearest Neighbors (KNN)
- Naive Bayes
- Any other sutiable two(e.g., Decision Trees, Support Vector Machine, or Random Forest)

Bonus Point:

If possible, apply a deep learning model for additional performance comparison.

Steps:

1. Data Preparation:

a) **Categorical Encoding:** If your dataset contains categorical (non-numeric) features, onvert them into numerical format using techniques such as LabelEncoder or OneHotEncoder.

b) **Split the Data:** Separate the dataset into features (denoted as `X`) and output/target (denoted as `y`).

c) **Training and Testing Sets:** Split the features and target into training and testing sets.

2. Apply Machine Learning Algorithms:

a) **K-Nearest Neighbors (KNN):**

- Train a KNN model using the training set.

b) **Naive Bayes:**

- Train a Naive Bayes classifier using the training set.

c) **Additional Model:**

- Choose and train two additional suitable machine learning models (e.g., Decision Tree, Support Vector Machine, or Random Forest).

d) **Bouns:**

- Apply a deep learning model, if possible, for further comparison.

3. Model evaluation:

For each model, compute the following evaluation metrics in the test set.

- **Accuracy:** Overall percentage of correctly predicted instances.
- **Confusion Matrix:** A table that visualizes true vs. predicted classes.
- **Recall:** The model's ability to capture all relevant cases (i.e., true positives).
- **Precision:** The quality of the positive predictions made by the model.

Comparison: Compare the models based on these metrics and decide which algorithm performs best. Provide a clear reasoning behind your choice, considering factors such as data distribution, model assumptions, and performance metrics.

Deliverables:

a) Your code needs to be submitted on the Google form (make sure the code runs and no errors in it).

b) Google form link: https://forms.gle/ZU7tVitLphDUe9RQ8

c) Make sure the code runs without any errors.

d) Avoid writing all your code in a single cell; organize it logically into multiple cells.

e) The detailed code for data splitting, training, prediction, and metric calculations should be accompanied by inline comments to explain each step.

f) Each team should submit only one file with the names and IDs of the other team members (file format .ipynb).

PLAGIARISM IS NOT TOLERATED AND COPIED WORK WILL BE AWARDED 0 POINTS FOR BOTH PERSONS INVOLVED! **There will be an an an individual evaluation.**