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 1
Due date is March 12, 2025 at 11:59 PM

The Assignment team consists of **Three members per team**.

In this assignment, you will examine the utility of data preparation and exploration techniques. You need to choose one dataset from the following ones:

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

As part of understanding the dataset, you are required to do the following on the chosen dataset:

Dataset Analysis and Preparation Tasks

a) Display the first and last 12 rows of the dataset.

b) Identify and print the total number of rows and columns present.

c) List all column names along with their corresponding data types.

d) Print the name of the first column.

e) Generate a summary of the dataset, including non-null counts and data types.

f) Choose a categorical attribute and display the distinct values it contains.

g) Identify the most frequently occurring value in the chosen categorical attribute.

h) Calculate and present the mean, median, standard deviation, and percentiles (20

Data Preparation Tasks

a) Apply a filter to select rows based on a specific condition of your choice (e.g., select records where a value exceeds a certain threshold).

b) Identify records where a chosen attribute starts with a specific letter and count how many records match this condition.

c) Determine the total number of duplicate rows and remove them if found.

d) Convert the data type of a numerical column from integer to string.

e) Group the dataset based on two selected categorical features and analyze the results.

f) Check for the existence of missing values within the dataset.

g) If any missing values are found, replace them with the median or mode as appropriate.

h) Divide a chosen numerical column into 5 equal-width bins and count the number of records in each bin.

i) Identify and print the row corresponding to the maximum value of a selected numerical feature.

j) Construct a boxplot for an attribute you consider significant and justify the selection.

k) Generate a histogram for a chosen attribute and provide an explanation for its relevance.

l) Create a scatterplot using two attributes and interpret the relationship observed.

m) Normalize the numerical attributes using StandardScaler to achieve standardized data.

n) Perform PCA (Principal Component Analysis) to reduce dimensionality to two components, and visualize the dataset before and after applying PCA.

o) Analyze the correlation between numerical features using a heatmap.

Practical Analytical Questions

a) Use Python to calculate and display the correlation matrix, and identify potential features relevant for classification.

b) Use Python to find the class distribution of a selected categorical feature and analyze the results.

c) Apply Python techniques to create new features from existing ones (feature engineering) and explain the significance of the new features.

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/cwiVU9S27wJKbbAi6

c) Make sure to comment on every step while coding.

d) Please split the code into cells (Don't write all your code in one cell)

e) Each team should submit only one file with the names and IDs of the other team members.

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