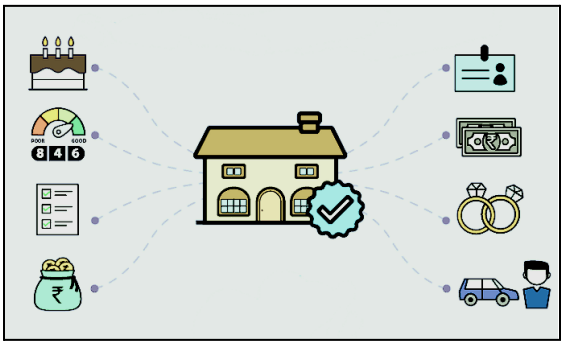
Cairo University
Faculty of Computers and Artificial Intelligence
Machine Learning Course

Assignment 1: Linear and Logistic Regression

A housing finance company offers interest-free home loans to customers. When a customer applies for a home loan, the company validates the customer's eligibility for a loan before making a decision.

Now, the company wants to automate the customers' eligibility validation process based on the customers' details provided while filling the application form. These details include gender, education, income, credit history, and others. The company also wants to have a predictive model for the maximum loan amount that an applicant is authorized to borrow based on his details.



You are required to build a linear regression model and a logistic regression model for this company to predict loan decisions and amounts based on some features.

Datasets:

There are two attached datasets:

- The first dataset **"loan_old.csv"** contains 614 records of applicants' data with 10 feature columns in addition to 2 target columns. The features are: the loan application ID, the applicant's gender, marital status, number of dependents, education and income, the co-applicant's income, the number of months until the loan is due, the applicant's credit history check, and the property area. The targets are the maximum loan amount (in thousands) and the loan acceptance status.
- The second dataset **"loan_new.csv"** contains 367 records of new applicants' data with the 10 feature columns.

Note: These datasets are modified versions of the "Loan Eligibility Dataset". The original datasets were obtained from Kaggle.

Requirements:

Write a Python program in which you do the following:

a) Load the **"loan_old.csv"** dataset.

b) **Perform analysis** on the dataset to:
   i) check whether there are missing values
   ii) check the type of each feature (categorical or numerical)
   iii) check whether numerical features have the same scale
   iv) visualize a pairplot between numercial columns

c) **Preprocess** the data such that:
   i) records containing missing values are removed
   ii) the features and targets are separated
   iii) the data is shuffled and split into training and testing sets
   iv) categorical features are encoded
   v) categorical targets are encoded
   vi) numerical features are standardized

d) **Fit a linear regression model** to the data to predict the loan amount.
   -> Use sklearn's linear regression.

e) **Evaluate** the linear regression model **using sklearn's R^2 score.**

f) **Fit a logistic regression model** to the data to predict the loan status.
   -> Implement logistic regression from scratch using gradient descent.

g) **Write a function** (from scratch) to calculate the **accuracy** of the model.

h) Load the **"loan_new.csv"** dataset.

i) Perform the **same preprocessing** on it (except shuffling and splitting).

j) **Use your models** on this data to predict the loan amounts and status.

Remarks:

- You can use functions from **data analysis and computing libraries** (e.g. Pandas and NumPy) as you please throughout the entire code.

- You can use **machine learning libraries** such as Scikit-learn for preprocessing and metrics **but NOT for "from scratch" requirements.**

- The **train/test split** has to be **performed before** the encoding and standardization steps.

- **The categorical features of the test set (and of the new data) should be transformed (encoded) using the encoder fitted on the train set.**

- **The numerical features of the test set (and of the new data) should be standardized using the mean and standard deviation of the train set.**

- We will use **R^2 score** to evaluate the linear regression model as it **provides a measure of how well observed outcomes are replicated by the model** (based on the proportion of total variation of outcomes explained by the model). **The best possible score is 1**, but the score can be negative as the model can be arbitrarily worse.

Deliverables:

- **You are required to submit ONE zip file containing the following:**
  - Your **code (.py)** file.

    If you have a (.ipynb) file, you have to save/download it as (.py) before submitting.
  - A **report (.pdf)** containing the team members' names and IDs, and the code with screenshots of the output of each part.

    If you have a (.ipynb) file, you can just convert it to pdf.

- **The zip file must follow this naming convention:**
  ID1_ID2_ID3_ID4_ID5_Group

Submission Remarks:

- The **maximum** number of students in a team is **5** and the **minimum is 4**.
- Team members must be from the **same lab** (or have the same TA).
- **No late submission** is allowed.
- A **penalty** will be **imposed for violating** any of the assignment rules.
- **Cheaters will get ZERO** and no excuses will be accepted.

Grading Criteria:

Both the code and the report must include:	
Analysis	**2 marks**
Preprocessing	**6 marks**
Linear regression and R^2 score	**2 marks**
Logistic regression (gradient descent)	**6 marks**
Accuracy	**2 marks**
New predictions	**2 marks**
The total is 20 marks (will be scaled to 5 marks)	