Assignment 1 – Linear & Logistic Regression

A certain car company is planning to manufacture and launch a new car. So, the company's consultants need to study the factors which the pricing of cars depends on. Based on various market surveys, the consultants gathered a dataset of different types of cars across the market. They would like to know which of the car features are significant in predicting the price of a car.

In addition, they would like to predict whether customers will be interested in purchasing the new car. That's why they collected a few records of some of the company's previous customers who either purchased a new car from the company as an upgrade or didn't purchase a new car.

You are required to build a linear regression model and a logistic regression model for this company to predict car prices and purchases based on some features.

1) Data

There are 2 attached datasets *(obtained from Kaggle)*:

- The first dataset **"car_data.csv"** contains 205 records of cars with 25 features per record in addition to 1 target column. These features include the car size and dimensions, the fuel system and fuel type used by the car, the engine size and type, the horsepower, etc. The final column (i.e., the target) is the car price (in some monetary unit).

- The second dataset **"customer_data.csv"** contains 400 records representing some of the company's previous customers. The customer data is composed of the customer's age and salary. The final column (i.e., the target) is a boolean value (0 if the customer didn't purchase a new car and 1 if he/she purchased a new car).

2) Requirements:

Write 2 python programs in which you work on each dataset (each model) separately as follows:

- *In the first program:*

  a- Load the **"car_data.csv"** dataset.

  b- Use **scatter plots** between different features and the target **to select 4 of the numerical features** that are positively/negatively correlated to the car price (i.e., features that have a relationship with the target). These 4 features are the features that will be used in linear regression.

  c- Split the dataset into **training and testing** sets.

  d- Implement **linear regression from scratch** using **gradient descent** (GD) to optimize the parameters of the hypothesis function.

  e- **Calculate the cost (mean squared error)** in every iteration to see how the error of the hypothesis function changes with every iteration of gradient descent.

  f- **Plot the cost** against the number of iterations.

- *In the second program:*

  a- Load the **"customer_data.csv"** dataset.

  b- Split the dataset into **training and testing** sets.

  c- Implement **logistic regression from scratch** using **gradient descent** to optimize the parameters of the hypothesis function. Use the 2 features **(age & salary)** as input and the output to be predicted is **"purchased".**

  d- Use the optimized hypothesis function to **make predictions** on new data.

  e- **Calculate the accuracy** of the final (trained) model on the test set.

3) Notes:

- You will need to **normalize the feature data** before applying regression. You can use minmax normalization where *z* is the normalized value and *z = (x – min) / (max – min).*
- You will need to **shuffle** each dataset's rows.
- You will need to **try different values of the learning rate** and see how this changes the error or accuracy of the model.

4) Submission Remarks:

- The **maximum** number of students in a team is **4** and the minimum is **2**.
- Team members must be from the **same lab** (or at least have the same TA).
- **No late submission** is allowed.
- **Cheaters will take ZERO** and no excuses will be accepted.

5) Grading Criteria:

1st Program [3 marks]	
Scatter plots for feature selection	0.5
Normalizing, shuffling, and splitting the data	0.5
Linear regression (GD)	1.5
MSE (calculation and plot)	0.5
2nd Program [3 marks]	
Normalizing, shuffling, and splitting the data	0.5
Logistic regression (GD)	1.5
"Predict" function	0.5
Accuracy	0.5