BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department

Machine Learning and Data Science - ENCS5341

Assignment #1

Submission deadline: 30.11.2023

The attached `cars.csv` file contains a dataset of cars concerning fuel consumption. In this assignment, we will study the dataset and analyze some of its aspects. You have to use Python for the solution.

1- Read the dataset and examine how many features and examples does it have?
(Hint: you can use `Pandas` to load the dataset into a `dataframe`)

2- Are there features with missing values? How many missing values are there in each one?
(Hint: you can use `isnull()` from `Pandas`)

3- Fill the missing values in each feature using a proper imputation method (for example: fill with mean, median, or mode)

4- Which country produces cars with better fuel economy?
(Hint: use box plot that shows the mpg for each country (all countries in one plot))

5- Which of the following features has a distribution that is most similar to a Gaussian: 'acceleration', 'horsepower', or 'mpg'? Answer this part by showing the histogram of each feature.

6- Support your answer for part 5 by using a quantitative measure.

7- Plot a scatter plot that shows the 'horsepower' on the x-axis and 'mpg' on the y-axis. Is there a correlation between them? Positive or negative?

8- Implement the closed form solution of linear regression and use it to learn a linear model to predict the 'mpg' from the 'horsepower'. Plot the learned line on the same scatter plot you got in part 7.
(Hint: This is a simple linear regression problem (one feature). Do not forget to add $x_0=1$ for the intercept. For inverting a matrix use np.linalg.inv from NumPy)

9- Repeat part 8 but now learn a quadratic function of the form

$$f = w_0 + w_1 x + w_2 x^2.$$

10- Repeat part 8 (simple linear regression case) but now by implementing the gradient descent algorithm instead of the closed form solution.

You have to submit both the code (using python) and a short report to summarize the results.
Alternatively, you can use Jupyter Notebook to submit **.ipynb** file that contains both *code cells* and *text cells* that discusses the results. (If you use colab, do not submit links to your notebook. Instead download the .ipynb file and submit it)

Good Luck!