Assignment 1

The objectives of this assignment are as follows:

1. Understand the concept of a classification task.
2. Understand the concept of a regression task.
3. Comprehend the process of training, validation, and testing data split.
4. Learn how to modify model parameters.
5. Gain proficiency in utilizing machine learning frameworks such as NumPy, Pandas, and Scikit-Learn.

Problem Statement 1 (Classification)

Given the MAGIC gamma telescope dataset. This dataset is generated to simulate registration of high energy gamma particles in a ground-based atmospheric Cherenkov gamma telescope using the imaging technique. The dataset consists of two classes: gammas (signal) and hadrons (background). There are 12332 gamma events and 6688 hadron events.

You are required to do the following:

1. the dataset is class imbalanced. To balance the dataset, randomly put aside the extra readings for the gamma "g" class to make both classes equal in size.
2. Split your dataset randomly so that the training set would form 70%, for the validation set 15% and 15% for the testing set (Don't use it while tuning the model parameters).
3. Apply K-NN Classifier to the data Manually (Without Scikit- Learn) once (i.e. you create functions for distance calculation, finding K-nearest neighbors & making predictions based on majority vote or whatever implementation you like).
4. Re-apply K-NN Classifier to the data by using Scikit-Learn.
5. Apply different k values to get the best results in both cases.
6. Add your comments on the results and compare between the models. Plot validation accuracy vs. k values for both implementations, identify optimal k-value and discuss overfitting/underfitting trends
7. Report all of your final trained model's accuracy, precision, recall and f-score as well as confusion matrix for both cases and compare between them.

Problem Statement 2 (Regression)

Given California Houses prices data. This data contains information from the 1990 California census., it does provide an accessible introductory dataset the basics of regression models.

The data pertains to the houses found in each California district and some summary stats about them based on the 1990 census data. The columns are as follows; their names are self-explanatory:

- Median House Value: Median house value for households within a block (measured in US Dollars) [$]
- Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]

Prof. Dr. Marwan Torki

Eng. Youssef El-Ebiary
Eng. Youssef El-Kady
Eng. Ziad Fahmy
Eng. Sarah Mahmoud

- Median Age: Median age of a house within a block; a lower number is a newer building [years]
- Total Rooms: Total number of rooms within a block
- Total Bedrooms: Total number of bedrooms within a block
- Population: Total number of people residing within a block
- Households: Total number of households, a group of people residing within a home unit, for a block
- Latitude: A measure of how far north a house is; a higher value is farther north [°]
- Longitude: A measure of how far west a house is; a higher value is farther west [°]
- Distance to coast: Distance to the nearest coast point [m]
- Distance to Los Angeles: Distance to the center of Los Angeles [m]
- Distance to San Diego: Distance to the center of San Diego [m]
- Distance to San Jose: Distance to the center of San Jose [m]
- Distance to San Francisco: Distance to the center of San Francisco [m]

You are required to do the following:

1. Split your dataset randomly so that the training set would form 70%, for the validation set 15% and 15% for the testing set (Don't use it while tunning the model parameters).
2. Apply linear, lasso and ridge regression to the data to predict the median house value Implement from scratch using matrix operations. Then Calculate weights using normal equation: w = (X^T X)^(-1) X^T y) and apply Gradient descent as an alternative optimization approach and comment on the two results. Apply L2 Regularization (Ridge Regression) and L1 Regularization (Lasso Regression) and try different regularization parameters & plot validation error vs. regularization parameter
3. Re-apply step 3 again using Scikit- Learn and compare both results.
4. Report Mean Square Error and Mean Absolute Errors for all your models.
5. Add your comments on the results and compare between the models.

Grading Scheme

1. Data Splitting & Preprocessing (15%)

    - Proper random splitting implementation
    - Consistent data handling between manual and sklearn approaches

2. Classification Problem (40%)

    - Manual K-NN Implementation (20%):
        - Correct algorithm implementation (10%)
        - Proper evaluation metrics implementation (5%)
        - Parameter tuning and analysis (5%)
    - Scikit-Learn Implementation (10%):
        - Correct usage of sklearn tools (5%)

- Proper evaluation and comparison (5%)
- Analysis & Understanding (10%):
  - Quality of comparison and insights (5%)
  - Demonstration of algorithm understanding (5%)

3. Regression Problem (40%)

- Manual Regression Implementation (20%):
  - Linear regression from scratch (7%)
  - Ridge and Lasso implementation (8%)
  - Evaluation metrics implementation (5%)
- Scikit-Learn Implementation (10%):
  - Correct usage of sklearn regression tools (5%)
  - Proper evaluation and comparison (5%)
- Analysis & Understanding (10%):
  - Quality of regularization analysis (5%)
  - Demonstration of algorithm understanding (5%)

4. Code Quality & Documentation (5%)

- Clean, well-commented code
- Clear markdown explanations
- Professional presentation

Final Notes

1. You should work in groups of three, not finding a team is NOT an excuse.
2. You should deliver a python notebook attached with the comments in markdown cells, you should export it as pdf.
3. We will need both the pdf and the notebook in zipped file.
4. You should deliver with a naming scheme id _assigment.zip.
5. Delivery will be ignored if you didn't follow the naming scheme provided in 4, any one of the team ids can be used.
6. Any form of academic dishonesty, including but not limited to using AI tools (such as ChatGPT or other code generation platforms), copying open-source code without proper attribution, or engaging in 'vibe coding' without genuine understanding, will be considered a serious violation and will be heavily penalized.

Prof. Dr. Marwan Torki

Eng. Youssef El-Ebiary
Eng. Youssef El-Kady
Eng. Ziad Fahmy
Eng. Sarah Mahmoud