

Assignment 1: Machine Learning Coding Skills Refresher

Objective: The aim of this assignment is to refresh your machine learning coding skills before diving into more complex topics in the course. This assignment will cover fundamental concepts and techniques commonly used in machine learning projects.

- After completion of this assignment, you should have set up Python development environment on your systems or learn to utilize online development setups such as Google Colab.
- You should also set up the GitHub repository and push this assignment code to it (optional but recommended)

Due date: 5th September 2024, 11.59 PM

Submission: D2L Assignment 1 folder (talk to your peers to understand how to submit if you are working with D2L for the first time)

Instructions:

1. **Setup:** Set up your development environment with the required libraries (e.g., NumPy, Pandas, Scikit-learn).
2. **Data Preparation:** Download any one dataset from D2L (they are free datasets from Kaggle.com, you can choose any of your own dataset as well, just provide its details)
3. **Data Exploration:** Conduct initial data exploration. Display basic statistics, such as mean, median, and standard deviation, and visualizations like histograms or scatter plots on the distribution based on classes and other useful fields/columns (Ask me questions if you have at least 3 days before the due date)
4. **Data Preprocessing:** Perform preprocessing steps, including handling missing values (ignoring incomplete samples, filtering based on string length etc., encoding categorical variables, and splitting the data into training and testing sets).
5. **Model Building:** Choose a basic machine learning algorithm (e.g., Linear Regression, Decision Tree, or K-Nearest Neighbors) and build a model using the training data (you can choose to perform simple supervised ML at this point)
6. **Model Training:** Train the chosen model using the training dataset, evaluate its performance on the testing dataset.
7. **Evaluation:** Calculate and print relevant metrics (e.g., accuracy, mean squared error) to assess the model's performance. Print the confusion matrix and classification report or any other appropriate metric based on the choice of ML method.
8. **Hyperparameter Tuning (optional):** Experiment with different hyperparameters to fine-tune your model's performance. (re-do step 7 above)
9. **Documentation:** Provide clear and concise comments in your code to explain each step and rationale behind your decisions.
10. **Submission:** Submit your Python code file a summary of your findings and a reflection on the assignment in the last cell of your Jupyter Notebook/python code file.

Note: This assignment will refresh your coding skills and understanding of core machine learning concepts. Focus on implementing the basics accurately rather than pursuing complex solutions.

Grading Criteria:

- Correct implementation of data loading, exploration, preprocessing, model building, and evaluation steps.
- Clear documentation of code.
- Effective use of libraries and techniques learned.
- Reasonable and justified decisions during model selection (and hyperparameter tuning if done).

This assignment will serve as a foundation for the more advanced machine learning topics we will cover in this course. Good luck, and don't hesitate to ask for help if needed!