VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY

UNIVERSITY OF TECHNOLOGY

FACULTY OF COMPUTER SCIENCE AND ENGINEERING

Course: Machine Learning - 242

Assignment 1

### Class: CC01 - Group LHPD2

Team Members, Team Repository Link and Team Report

Our repository is at this link: https://github.com/pdz1804/ML_LHPD2.

This README that we submitted to you is stored in the repository at ML_LHPD2/notebooks/assignment1/ML_LHPD2_Ass1_README.md.

Also, our team's report which contains detailed information (data analysis, data preprocessing, normalization, models training process and evaluation of those trained models) for this Assignment 1 is stored in the repository at ML_LHPD2/reports/final_project/.

Name	Student ID
Nguyen Quang Phu	2252621
Pham Huynh Bao Dai	2252139
Nguyen Thanh Dat	2252145
Nguyen Tien Hung	2252280
Nguyen Thien Loc	2252460
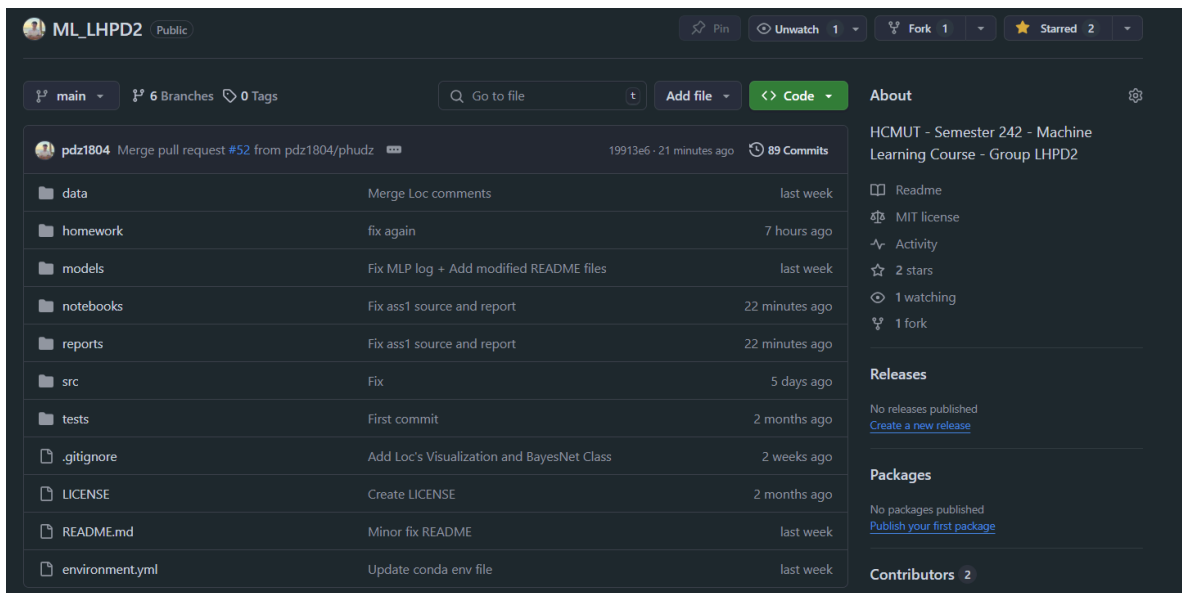
I. Group Project Requirements

Project Organization

Group Formation:

- The group consists of **5 members**.
- **Nguyen Quang Phu** is the **repository owner**.

Collaboration Workflow:

- The **main repository** is managed by the designated team member.
- Other members **fork** the repository and contribute via **pull requests**.
- **Roles and responsibilities** are clearly defined for each team member.

GitHub Repository Structure

We are implementing **ALL machine learning models from the syllabus** using:

- **ONE unified dataset** with one use-case before the midterm (Chapters 2-6) (which we do in this Assignment - Assignment 1).
- **THAT dataset** for that same use-case for the second stage after midterm (Chapters 6-10) (which we will do in our Assignment 2).

Repository Setup:

- The **main repository** is created by the team lead.
- Team members **fork** and contribute via **pull requests**.
- **Branching strategy** is used for different models and features.
- **Comprehensive documentation** is maintained.

Suggested Repository Structure:

```
Project-Root
 |── data/          # Raw & Processed Datasets
 |── notebooks/     # Jupyter Notebooks for EDA & Preprocessing
 |── models/        # ML Model Implementations
 |── other/         # Some other related things here
```

```
|── reports/       # Contain figures and Reports
|── src/           # Contain src code of how to process data, build features and tr.
|── tests/         # Evaluation Results & Comparisons
|── README.md      # Project Documentation
|── requirements.txt # Dependencies
```

Key Requirements

- Well-documented problem statements & model variations
- Consistent structure for model implementations
- Comprehensive testing of all components
- Comparative analysis of models
- Code reviews and pull requests
- Version control best practices

Each Team Member Should:

- **Actively contribute to the repository**
- **Document their work thoroughly**
- **Review and improve others' code**
- **Participate in group discussions**
- **Help maintain code quality**

II. Completed Tasks

Data Collection & Preprocessing

We have **collected and preprocessed** the dataset, and it has been uploaded to **Kaggle**.

Kaggle Dataset: Tweets Clean PosNeg v1

Dataset Preview:

Tasks Completed:

- Collected **raw tweets dataset**
- Applied **data cleaning & preprocessing**
- Uploaded the final dataset to **Kaggle**

Models that we have done in Assignment 1

- **Train and Evaluate Models**
  - Decision Tree
  - Random Forest
  - XGBoost
  - Logistic Regression
  - Naive Bayes
  - GA, HMM, Bayesian Network
  - MLPClassifier and simple Perceptron (ANN)
  - CNN / LSTM for deep learning
- **Compare Model Performance**
- **Create Visualizations for Analysis**
- **Write Final Report & Documentation**

IV. Contributions and Task Distribution

Team Member	Task
Nguyen Quang Phu	Team leader; Repository management; Participate and Ensure everything stays on schedule and verify all work done by other members.
Pham Huynh Bao Dai	Data preparation; Data preprocessing; feature engineering; Visualization; Document Data Collecting; Preprocessing and Merging.
Nguyen Thanh Dat	Consistent model training and evaluating; Model implementation (Decision Tree, Random Forest, XGBoost, Perceptron - ANN, MLP); Document Model Implementation.
Nguyen Tien Hung	Consistent model training and evaluating; Model implementation (GA, HMM Bayesian Network, Logistic Regression, LSTM); Document Model Implementation.
Nguyen Thien Loc	Model evaluation, hyperparameter tuning, Model Comparison, Document Performance analysis.

V. How to Run This Project

Clone the Repository

```
git clone https://github.com/pdz1804/ML_LHPD2
cd ML_LHPD2
```

Setting Up the Environment

If you want to train the model in local (on your computer, in VSC...), to install the necessary dependencies, refer to the environment file (environment.yml) provided in the repository. Ensure you have Conda installed to create and activate the required environment.

```
conda env create -f environment.yml
conda activate ml_env  # Replace ml_env with your environment name
```

Explore the Project

- First, you should revisit the Project structure above to see what is the use of those folders and subfolders in this project.
- In details:
  - To know how we collect all the data, visit the data folder.
  - To see the final preprocessing dataset, visiting this link: Tweets Clean PosNeg v1
  - To know how we preprocessing, make all the datasets consistent and how do we merge them together, visit the location src/data/process.ipynb and its utility file src/data/preprocess.py
  - To know how and what we use to make the features for all the texts/documents in the dataset, visit the location src/features/build_features_utils.py
  - To know how we perform hyperparameter tuning, do k-fold cross-validation and train our models, visit src/models/models_utils.py and src/models/train_model.ipynb
  - We have also created some visualization functions inside the location src/visualization/. Although now we have not used it anywhere in code because all the visualization or plotting now we just use the **in-line-function-plotting**, but we intend to use it later in Assignment 2.
  - The environment.yml contains the information about the libraries or dependencies that we use when performing and executing our project.

VI. License

This project is licensed under the **MIT License**.

VII. Contact

For any questions or contributions, please contact:

- Email: phu.nguyenquang2004@hcmut.edu.vn
- GitHub: https://github.com/pdz1804/

Contributors

We fairly contribute to this repository with dedication and teamwork!