Statistical Learning for Data Science- 2025 Project 2

Submit your project by email to `11912714@mail.sustech.edu.cn`. Projects 1 and 2 are due together at the end of the semester, but it is recommended to start early!

Your submission should include:

1. **Technical report** (`TeamID_project1_technical_report.pdf`):
   Report of Tasks 1 and 2. Write your report following the guidelines provided in the tasks' description. **Indicate team members and their contributions in the report.**
2. **Code implementation** (a zip file):
   Each finding should be supported by Python code (Jupyter Notebook is recommended).

Task1. Life Expectancy (40 points)

In this task, we are working with a dataset that includes 12 different features of 211 countries with their corresponding life expectancy. The dataset is named `life_indicator_2008-2018.xlsx`. Our goal is to predict `Life Expectancy at Birth` using these features. The main task involves using data from 2008 to train a model and then predicting life expectancy for 2018 based on this trained model.

This task will include the following steps:

a. Data Understanding (30%)

- Understand each of the 12 features and make a guess which ones are likely to have a significant impact on life expectancy.
- Visualize the relationships between the features using a heatmap to see how they correlate with each other.
- Look at the distribution of life expectancy at birth to understand its range and variability.
- Deal with any missing data by using different methods and comparing their effectiveness.

b. Modeling (30%)

- Try different models for the prediction task, considering factors like complexity and interpretability.
- Evaluate the performance of each model and compared them using metrics like MSE and R^2.
- Identify which features are most important for predicting life expectancy and explain how this determination was made.

c. Analysis of Predictions (20%)

- Visualize the differences between the predicted life expectancy values and the actual values for 2018. Look for any outliers and try to explain why they occurred.
- Examine the distribution of prediction errors to see if they follow any patterns or if there are any unexpected trends.

d. Model Improvement (20%)

- Try advanced techniques like stepwise forward selection to improve the model's performance.
- Try to create new features that could enhance performance, such as in the health status prediction task, Body Mass Index (BMI) derived from weight and height may be a good high-level indicator.

e. Bonus (10 points)

- Is it possible to predict life expectancy for 2025, given the trained model from step d and features (exclude `Life Expectancy at Birth`) ranged from 2008 to 2018?

Task2. Douban Movie Comment Analysis (60 points)

In this project, we aim to predicte whether a film will be loved by the audience from Douban based on textual reviews. The dataset is named `douban_movie.csv`. The model should take a text input (movie review) and output a specific attitude.

Part 1: Machine Learning Approach (40%)

a. Text Preprocessing (30%)

- Filter stopwords, low-frequency words, and special symbols.
- Normalize text (e.g., lowercase, stemming/lemmatization).

b. Text Vectorization (30%)

- Convert text to numerical features using:
    - TF-IDF
    - Word2Vec
    - BERT embeddings

  *(**You only need to select one way to do text vectorization. If you compare the impact of different vectorization methods, you will get extra bonus.**)*

c. Model Training & Evaluation (40%)

- Train and cross-validate (two both need):
    - Logistic Regression
    - Naive Bayes
- Evaluate accuracy, precision, recall, and F1-score and some reasoning analysis.

Part 2: Large Language Model (LLM) Approach (60%)

a. Prompt Design & In-Context Learning (35%)

- Design effective prompts for rating prediction.
- Experiment with few-shot learning (provide examples in the prompt).

b. LLM API Testing (35%)

- Use **at least two open-source LLM APIs** (e.g., Chatgpt-3.5, deepseek) for prediction.
- Compare their accuracy and robustness.

c. Discussion (30%)

- Compare ML and LLM results: strengths, limitations, and insights.

Bonus Tasks (20%)

1. **Advanced Data Analysis (5%)**
    - Generate visualizations (e.g., word clouds, sentiment distribution).
    - Explore correlations between review length/lexicon and ratings.
2. **LLM Prompt Optimization (5%)**
    - Test multi-prompt strategies (e.g., chain-of-thought, role-playing).
    - Analyze how prompt phrasing affects prediction quality.
3. **Fine-Tuning LLMs (10%)**
    - Fine-tune an open LLM (e.g., LLaMA-2) with a small subset of data.
    - Compare performance before/after fine-tuning.
