CSE 587 Data Intensive Computing
Homework 1

Spring 2025

Due: Feb. 14, 2025 by 11:59 PM

Topics: Big-data Processing and Machine Learning

This assignment is designed to introduce you to fundamental big-data processing skills and basic machine-learning models using a real-world dataset. For this homework, you will use the historical stock prices of Nvidia from July 1, 2022, to July 31, 2024. The dataset contains the following features: Date; Open price; Intraday high and low prices; Close price; Adjusted close price; Trading volume. You will apply data cleaning, preprocessing, visualization, and machine learning techniques such as clustering using KMeans.

Learning Objectives:

- Obtaining, Understanding and Exploring the datasets
- Understanding features and Visualization
- Applying ML models on data
- Coding with Python and libraries
- Independent work and Report writing

General Requirements

- **Work Environment**: This homework must be written **in Python in Jupyter Notebook**. You should explore and read the documentation for using the related libraies.
- **Submission Format**: You will submit a single zip file containing:
  (1) Your `.ipynb` code file
  (2) A PDF file exported from your notebook file with complete results, and
  (3) A comprehensive report (as a PDF) detailing your work, code, explanation, and results

  Your zip file should be named as cse587_hw1_UBIT_PersonNumber
  (for example: cse587_hw1_abcd_12345678).
- **Grading Criteria**: Correctness and completeness of tasks; Clarity of explanations and visualizations; Code readability and documentation.
- **Academic Integrity**: You will get an automatic F for the course if you violate the academic integrity policy. This homework is an individual assignment. You are not permitted to work with anyone else on this assignment. All work submitted must be yours and yours alone.

Notes:

- Late submissions are NOT allowed, therefore, you are encouraged to start working on Homework early on and submit well in time.
- If we are not able to run your code then you will get zero marks. Therefore, you must make sure that the code is executable not only on your device but by anyone having your code.
- AI violation policy will be strictly implemented and code/reports will be checked.

Part 1: Big Data Processing (50 points)

Task 1: Data Cleaning and Exploration (20 points)

1. Load the dataset into a Pandas DataFrame and print the meta data of column information.
2. Check for missing values and handle them appropriately. (any proper processing is acceptable)
3. Convert the date column to a different datetime format (any different proper format is acceptable) and print it.
4. Compute basic statistics (min, max, mean, median, standard deviation) for each numerical feature.
5. Plot the closing price over time using Matplotlib.

Task 2: Feature Engineering (15 points)

1. Create a new column for daily returns based on the adjusted closing price (0 for the first day) and and print the top 10 dates with the highest daily return.
2. Create a new column for the 7-day moving average of the closing price and plot it with Matplotlib. (any proper filling way of the first 6 days is acceptable)
3. Normalize Trading volume column using Min-Max Scaling and print the top 10 dates with the highest volume.

Task 3: Data Visualization (15 points)

1. Create a histogram of daily returns.
2. Generate a boxplot of the trading volume.
3. Display a correlation heatmap of all numerical features. (Don't consider newly added columns)

Part 2: Machine Learning (50 points)

Task 1: Clustering with KMeans (20 points)

1. Select relevant features for clustering (e.g., normalized daily return, trading volume, and adjusted close price).
2. Determine the optimal number of clusters using the elbow method.
3. Apply KMeans clustering and visualize the resulting clusters using a scatter plot.
4. Interpret the clusters and describe potential insights.

Task 2: Other machine learning methods (30 points)

- Adopt two more machine learning methods on this stock dataset to complete certain tasks. For example,
  - **Stock Price Prediction:** Train a regression model (e.g., Linear Regression, Random Forest Regressor, or LSTM) to predict the closing price of the stock based on historical data.
  - **Trend Classification:** Implement a classification model (e.g., Logistic Regression, Support Vector Machine, or Neural Networks) to predict whether the stock price will go up or down the next day.
- The training and test sets should be divided in a ratio of 4:1.
- You can also design other tasks as long as you describe them clearly.
- If at least one of the ML methods you adopt is not mentioned in this course, you will get 5 points bonus for this homework.

Possible References

- Jupyter Notebook: https://github.com/jupyterlab/jupyterlab-desktop.
- JupyterLab: https://github.com/jupyterlab/jupyterlab-desktop.
- VSCode: https://code.visualstudio.com/docs/datascience/jupyter-notebooks.
- pandas: https://pandas.pydata.org/docs/.
- numpy: https://numpy.org/doc/stable/.
- matplotlib: https://matplotlib.org/stable/index.html.