MSc Artificial Intelligence – Machine Learning
Coursework, 2024/2025
Release date: 9 July 2025
Submission date: 6 August 2025 at 11:59pm.
Feedback available on or before 20 August 2025.

Title: Machine Learning Analysis of Stock Price Trends

You are provided with 5 datasets containing historical stock price data for 5 companies, including features such as date, open price, close price, high price, low price, trading volume, and technical indicators. Your task is to investigate how to predict stock price trends using three different machine learning approaches: regression, classification, and clustering.

Dataset Download datasets from Yahoo Finance for the following stocks in US stock markets: 'AXP','V','JPM','GS','HD'.

 Instructions:
1. Regression Approach: Build a regression model to predict the nth-day forward closing price of a specific stock. You need to design and justify regression experiments including data preprocessing (e.g., cleaning, transformation, splitting, preparation), feature engineering (e.g., selection, extraction), model selection (e.g., choice of algorithms, hyperparameter tuning), and evaluation (e.g., choice of metrics, data splitting); conduct experiments and analyse results, making full use of the provided data.
2. Classification Approach: Transform the stock price analysis problem into a classification task by predicting whether the stock price will go up, go down, or remain relatively stable on the next nth-day. Again, you need to design and justify classification experiments including data preprocessing, feature engineering, model selection, and evaluation; conduct experiments and analyse results, making full use of the provided data.
3. Clustering Approach: Now take the stock price analysis problem as a clustering problem – cluster data and then predict by the mean of nth-day prices of a cluster. Again, you need to design and justify clustering experiments including data preprocessing, feature engineering, model selection, and evaluation; conduct experiments and analyse results, making full use of the provided data.
4. After you have completed the above studies, you need to conduct a comparative analysis of the three approaches, using evidence from your experiments, and then make a recommendation to suitable end users.

Submission: Submit a **single zip** file with three files via the ECS8051 Canvas Assignments page – one **Jupyter Notebook**, one **report** (PDF), and one 5-min**video**. The **Jupyter Notebook** should be a python code listing for the project, including explorations to generate evidence to support your decisions and/or arguments. It should be well-structured, and clearly commented to facilitate reading.

The **report** (not exceeding 3000 words) should describe your project in a structured way, demonstrating your critical understanding of regression, classification, and clustering techniques and their applications to the given problem, as well as your awareness of typical problems such as overfitting and imbalanced learning. Your report should include various sections including (1) Introduction (2) Experimental design (and its rationale) (3) Results and analyses (including visualisations) (4) Discussions and (5) Conclusion. Your analysis should highlight the strengths and limitations of each approach and provide insights into their practical applications. Throughout the report, please ensure to provide appropriate references to support your arguments and decisions. Word count excludes Figures, Tables, Titles, References, Code and Cover page. Python code in your Jupyter Notebook must be included in your report as an appendix for plagiarism check.

The **video** should provide an overview of your project, highlighting key processes, experimental results, conclusions and insights, and demonstrating your critical understanding of the subjects.

Assessment Criteria:
- Report: Soundness and completeness and rationale and rigour of the studies, demonstrated in the following parts:
  - Regression approach (20%): data preprocessing (5%), feature engineering (5%), model selection (5%), evaluation and analysis (5%).
  - Classification approach (20%): data preprocessing (5%), feature engineering (5%), model selection (5%), evaluation and analysis (5%).
  - Clustering approach (20%): data preprocessing (5%), feature engineering (5%), model selection (5%), evaluation and analysis (5%).
- Report: Critical comparison of the three approaches (20%): interestingness of recommendation/conclusion (5%); strength of support (5%); use of additional evidence or literature (10%).
- Report: Quality (10%): structure (5%); readability (5%).
- Notebook: Quality (10%): technical execution (5%) -- the extent to which the Jupyter Notebook runs smoothly without errors or issues; comprehension and understanding (5%) – the extent of the student's understanding of the content presented in the Notebook.

Learning outcome mapping:

Learning outcome	Coursework Activity
Demonstrate critical understanding of the theory underpinning core concepts and algorithms in machine learning	Explanation on: 1. Data Preprocessing 2. Feature Engineering 3. Model Selection, Model Evaluation, Cross-Validation and Generalization 4. Hyperparameter Tuning 5. Results Interpretation
Evaluate and compare supervised and unsupervised learning algorithms on problems involving real datasets	1. Data Selection and Preprocessing, Feature Engineering 2. Model Selection, Model Building, Model Evaluation, Cross-Validation and Hyper-parameter Tuning 3. Discussion, and conclusion
Diagnose and rectify common problems that affect the performance of machine learning algorithms	1. Identification of problems (overfitting, underfitting, data imbalance, feature selection issues, the bias-variance trade-off) in the context of their project. 2. Explanation of the problems, how they impact ML in general and specifically in the project. 3. How to rectify the problems in general and specifically in the project.
Design machine learning experiments and justify the procedures employed	1. Problem Formulation (regression, classification, clustering; and why); 2. Experimental Design (what, how, why); 3. Data Selection and Preprocessing, Feature Engineering; 4. Model Selection, Model Building, Model Evaluation, Cross-Validation and Hyper-parameter Tuning; 5. Baseline Models, Ablation Study, Results Interpretation; 6. Discussion, Reflection, Conclusion, and Future Directions.