CISC7201 DATA SCIENCE PROGRAMMING — Course Project Guidelines

1. Project Purpose

The course project aims to give students hands-on experience with a full **end-to-end data science workflow**, including data collection, data engineering, data analysis, model development, and presentation of findings. Students are expected to demonstrate both technical competency and the ability to apply data science to real-world contexts.

More importantly, this project emphasizes **the use of programming** rather than relying on **off-the-shelf machine learning methods or automated tools**. Students who provide additional effort in data processing, feature engineering, analysis, and visualization will receive corresponding credit in the assessment. In addition, the good use of programming packages will received credit in the assessment.

The focus is on programming proficiency, reproducibility, and the correct implementation of data science techniques, not on achieving the highest model accuracy or using complex models.

---

2. Team Formation

- Each team should consist of **4–5 students**.
  - The project will be assessed using the same standard even if your team has fewer members.
  - The project will be assessed using a higher standard if your team has 6 members.
- Team members are encouraged to bring diverse strengths (programming, analytics, visualization, domain knowledge).
- Every member must contribute meaningfully; workload distribution should be fair and explicitly documented.
- For **contribution assessment**, each member will be asked to evaluate the contributions of their teammates confidentially.

---

3. Project Requirements

A. Motivation & Team Background (Mandatory Section)

Your report must clearly address:

1. **Motivation**
   - Why did you choose this topic?
   - What real problem or question does your project attempt to solve?
   - Why is this problem important or interesting?
2. **Team Background**

- Introduce each team member's strengths or relevant experience.
- Explain how the team's background influenced the topic choice.
- Highlight any domain knowledge that supports the project.

> **A strong motivation section is essential. Projects with unclear motivation will receive lower marks regardless of technical quality.**

B. Data Collection (Benchmark Datasets NOT Allowed)

Your team must **use your own dataset**. Possible sources include:

- Web scraping
- Sensor/IoT data
- Manually recorded logs
- API retrieval (weather, finance, mobility…)
- Surveys or interviews
- Collected images or videos

Please keep your dataset size manageable. We will update you on the maximum dataset size that our repository can accommodate.

Restrictions

- **No benchmarking datasets** (MNIST, CIFAR, UCI, Kaggle defaults).

C. Data Engineering, Cleaning, and Management

You may include any relevant tasks from the following areas (not all are required):

I. Data Cleaning

- Handling missing or inconsistent values
- Removing duplicates
- Normalizing formats and data types

II. Data Engineering

- Feature extraction and creation
- Image or time-series preprocessing
- Building reusable pipelines

III. Data Management

- Clear folder/data organization
- Git/GitHub version control

- Documentation of workflows and metadata
- Ensuring reproducibility of results

D. Data Analysis: Learning, Analytics, Visualization

This is a programming-focused course. The key is to apply programming techniques to analyze data -- **advanced machine learning models are not the main requirement**. You may include any relevant tasks from the following areas (not all are required):

Exploratory Analysis

- Statistical summaries
- Correlation and trend analysis
- Visualizations (plots, charts, dashboards)

Modeling & Evaluation

- Machine learning models (classification, regression, clustering...)
- Performance metrics and interpretation
- Comparison between baseline and improved models

Visualization & Interpretation

- Effective visual storytelling
- Meaningful insights and explanations

E. Conclusion

Your conclusion should include:

- Key findings
- Limitations
- Reflection on what the team learned
- Possible future improvements

---

4. Deliverables

We will create a repository for each project team on Gitea. You only need to **commit your files to the repository** before the deadline.

A. Jupyter Notebook

- Clear structure following the required sections
- Professional English writing

- Proper references
- Codes
- IMPORTANT: list all important programming packages you use and explain why / how you use them

B. Code Structure (Optional)

- Organized GitHub repository
- README with execution instructions
- Structured notebooks/scripts

C. Presentation

- slides in pdf, html, or pptx
- 10–12 minutes in mp4
- Highlight motivation, dataset, engineering pipeline, findings
- Demo or visualization encouraged

D. Readme.md

- Description of your project
- An example can be found in a recent project, https://github.com/showlab/Paper2Video

---

5. Timeline

Stage	Due Date
Team grouping	2025-11-21
Project material submissions	2025-12-14

---

6. Grading Breakdown

Component	Percentage
Motivation & Background	15%
Data Collection, Data Engineering & Cleaning	35%
Data Analysis & Modeling	35%
Conclusion & Reflection	10%
Professionalism	5%

In exceptional cases where the data is collected from laboratory, e.g., biology data, it is acceptable to skip the data collection section. **Please note that the use of benchmark datasets remains strictly prohibited.** However, we expect you to contribute significantly more effort to the data analysis component.

Component	Percentage
Motivation & Background	15%
Data Analysis & Modeling	70%
Conclusion & Reflection	10%
Professionalism	5%