Course Project Description

1. Project Topic

Select a data mining topic that aligns with your research interests or future career goals. Clearly identify the data mining task and the dataset you will use in this project. Then, design and implement a data mining system, pipeline, or solution that addresses the task. Your approach may include components such as data preprocessing, feature extraction and selection, regression, classification, clustering, or other relevant techniques.

	Title
Topic 1	Regression with an Insurance Dataset https://www.kaggle.com/competitions/playground-series-s4e12
Topic 2	Binary Prediction with a Rainfall Dataset https://www.kaggle.com/competitions/playground-series-s5e3
Topic 3	Loan Approval Prediction https://www.kaggle.com/competitions/playground-series-s4e10
Topic 4	Binary Prediction of Poisonous Mushrooms https://www.kaggle.com/competitions/playground-series-s4e8
Topic 5	Regression with a Flood Prediction Dataset https://www.kaggle.com/competitions/playground-series-s4e5
Topic 6	Binary Prediction of Smoker Status using Bio-Signals https://www.kaggle.com/competitions/playground-series-s3e24
Topic #	Kaggle on-going competitions https://www.kaggle.com/competitions

2. Tasks

- Clearly understand and define the data mining task you aim to solve.
- Identify and explore a dataset. Perform necessary preprocessing steps, including data cleaning, feature extraction, and feature selection.
- Design and implement a data mining system, pipeline, or solution that addresses the identified task. You may use techniques such as regression, classification, clustering, or other relevant methods.
- Apply at least three different approaches or models to solve the task, and compare their results and performance

3. Project groups

- Form a team with a size of 1, 2, 3, or 4 students and self-sign up via Canvas – [People] – [Project Groups]

4. Project pitch

1) Your project pitch is a short presentation that outlines your project idea and plan. It should answer the following questions:
    - Background of your project
    - What is the problem that you will study?
    - Why is addressing this problem is important?
    - What are the main challenges (e.g., data challenges, or modeling challenges)?
    - What are the essential tasks in your problem formulation (e.g., can your problem be formulated as a task of clustering, regression, classification, ranking, recommender systems, clustering + classification, clustering + regression, or clustering + ranking?)
    - What is your planned data mining system pipeline?
    - What are your initial data exploration results (check the example of the reading paper: Exploring Millions of Footprints in Location Sharing Services)?
    - Can your data be directly fed to a data mining system?
    - What are your data preprocessing steps?
    - After data collection, preprocessing, are your data quality improved and ready for building predictive models?

   **Failure to meet each expectation will result in point deductions.*

2) *You are encouraged to visit office hours or make an appointment to spend 5 minutes and review your slides with the instructor before your presentation.*
3) You need to submit the following two files. Please submit both the **video recording** and the **PDF file**.
    - A video recording:
      [Start Assignment] – [Media] tab – [Record/Upload Media]
    - A PDF version of your presentation slides:
      [Start Assignment] – [File Upload] tab

4) Each presentation should be **no more than 15 mins**

5. Project progress check point 1

A **project progress presentation** provides an update on your work so far. It should include a brief summary of your project pitch, the key steps and technical details of your proposed approach, current experimental results, and your plans for future improvement.

1) Expectations
   - **Project Pitch Recap**: Begin with a brief summary of your original project pitch, including the problem you are addressing and your initial plan.
   - **Data Exploration & Visualization**: You are expected to explore and visualize your dataset. This helps you observe patterns in the data and gain insights that can guide your predictive modeling.
   - **Preliminary System Development**: Build and present a preliminary data mining pipeline or system. Clearly explain the key steps involved and provide relevant technical details of your approach.
   - **Experimental Results**: Present the results from your current system. Include metrics, visualizations, or comparisons as appropriate.
   - **Analysis & Future Plans**: Analyze potential limitations or issues in your current system. Discuss your plans to improve it for the final version, including any adjustments in methods, features, or model selection.

   *Please note that this is a **checkpoint** for your **preliminary** system. It is not expected to be final and should be viewed as a work-in-progress that can and should evolve.

   <u> Failure to meet each expectation will result in point deductions.</u>*

2) *<u>You are encouraged to visit office hours or make an appointment to spend 5 minutes and review your slides with the instructor before your presentation.</u>*
3) You need to submit the following two files. Please submit both the **video recording** and the **PDF file**.
   - A video recording:
     [Start Assignment] – [Media] tab – [Record/Upload Media]
   - A PDF version of your presentation slides:
     [Start Assignment] – [File Upload] tab
4) Each presentation should be **no more than 15 mins**

6. Project progress check point 2

For **Checkpoint 2**, you are expected to build upon the data mining pipeline developed in Checkpoint 1. Specifically, you should:

1) Expectations:
   - **Explore Multiple Predictive Methods:** Apply a variety of predictive techniques (e.g., regression, classification, clustering, etc.) to your problem and compare their performance.
   - **Analyze Results and Identify Issues:** Evaluate the outcomes of your experiments, identify any limitations or challenges (e.g., overfitting, data imbalance, low accuracy, etc.), and analyze the underlying causes.
   - **Propose and Test Improvements:** Based on your analysis, suggest and implement improvements to enhance your model performance (e.g., hyperparameter tuning, feature engineering, model selection, etc.).
   - **Deliver Enhanced Results:** Present updated experimental results that demonstrate improvement over your initial models.
   - **Outline Your Future Plan:** Discuss the next steps for completing your project, including any additional methods or enhancements you plan to explore.

   * The goal of Checkpoint 2 is to demonstrate meaningful progress and refinement in your data mining pipeline and results, setting the stage for your final project submission.

   * *Failure to meet each expectation will result in point deductions.*

2) *You are encouraged to visit office hours or make an appointment to spend 5 minutes and review your slides with the instructor before your presentation.*
3) You need to submit the following two files. Please submit both the **video recording** and the **PDF file**.
   - A video recording:
     [Start Assignment] – [Media] tab – [Record/Upload Media]
   - A PDF version of your presentation slides:
     [Start Assignment] – [File Upload] tab
4) Each presentation should be **no more than 15 mins**

7. Final project report

Each group is required to submit a **comprehensive project report** (minimum of **five pages**, excluding references and appendices).

1) Report Structure should follow the structure below:

2) **Title, authors and Abstract**
    - Clearly state the project title
    - Contain the authors' information
    - Provide a concise abstract summarizing the problem, methodology, and key findings (150-250 words)
3) **Section 1: Introduction**
    - Background: Provide context and motivation for the problem.
    - Problem Description: What specific problem are you addressing?
    - Importance: Explain why solving this problem matters (e.g., societal impact, business relevance, etc.).
    - Related Work: Briefly review relevant literature or prior work.
    - System Overview: Summarize your approach and system architecture.
    - Data Collection: Describe the dataset used (source, type, size, etc.).
    - ML System Components: Outline the key stages of your pipeline (e.g., preprocessing, modeling).
    - Initial Results: Highlight any preliminary experimental findings.
4) **Section 2: Problem Formulation**
    - Key Definitions: Define important concepts, variables, data types, and the prediction target.
    - Formal Problem Statement: Clearly describe the objective, constraints, and assumptions of your task.
5) **Section 3: Overview of Proposed Approach/System**
    - Provide a high-level explanation of your proposed data mining or machine learning system.
    - Include design rationale and how it addresses the problem.
6) **Section 4: Technical Details**
    - Feature Engineering: Explain how features were extracted and selected.
    - Predictive Modeling: Describe models used, parameter tuning, and training process.
7) **Section 5: Experimental Evaluation**
    - Dataset Description: Include key statistics and properties of your dataset.
    - Evaluation Metrics: Define metrics used (e.g., accuracy, precision, recall, RMSE).
    - Baseline Methods:

        - Select appropriate baseline methods based on your task type (e.g., classification, regression, clustering, ranking, recommendation).
        - Example baseline models include Decision Trees, Random Forest, AdaBoost, SVM, Naive Bayes, Logistic Regression, and k-Nearest Neighbors.
        - Clearly distinguish between your proposed method (best model) and baseline methods (for comparison)
    - Results and Analysis:
        - Provide performance comparison across methods.
        - Use tables, graphs, or visualizations to present results.
        - Include ablation studies or component-wise evaluations (e.g., showing the impact of a specific feature or model component).

8) **Section 7: Conclusion**
    - Summarize the outcomes, lessons learned, and significance of the project.
    - Suggest directions for future work or improvement.

9) **Team member responsibilities**
    - If your project group contains two or more members, please include a responsibility table that clearly outlines:
        - Each member's assigned tasks (e.g., data preprocessing, model development, evaluation, report writing, etc.) A brief description of what each member actually completed during the project. Do not include tasks that were completed jointly by the entire team.
    - If you work individually, this section is not required

10) **References**
    - Include all sources, datasets, libraries, and prior work cited in your report.

8. Report format

- Use the following template
    - ACM proceeding template –
    - https://www.acm.org/publications/proceedings-template
    - IEEE proceeding template –
    - https://www.ieee.org/conferences/publishing/templates

- DO NOT paste your code or snapshot into the PDF. At the end of your PDF, please include a website (Github, Dropbox, OneDrive, GoogleDrive) address that can allow the TA to access your code.
- Useful links: Writing Technical Articles, Writing a Technical Report, Paper Writing and Presentations

9. Important dates: check the syllabus

10. Final advice and key principles

1) *Start earlier.*
2) *You are encouraged to meet me to discuss the directions and technical issues of your progress. However, it doesn't guarantee your reports will be scored well. It all depends on your workload and your technical depth as shown in your reports. Your score relates to your efforts of implementing different ideas to overcome data and model challenges, improve performances, and develop novel methods.*
3) *Final project report, presentations, and peer evaluations are important for grading.*

11. Evaluation criteria

Your project report will be evaluated based on the following 3 major aspects:

Aspects	Score	Descriptions
Technical novelty and contribution	9-10	develop a novel/new algorithm/model/system/framework that is different from existing methods to solve your problem.
	8-9	try or combine multiple existing methods to solve your problem.
	7-8	apply one existing method to solve your problem.
	0-7	unclear about what is the problem and which method is appropriate.
Comprehensive experiment designs and improved performances	9-10	include data description, baseline methods, evaluation metrics; evaluate the proposed method using different evaluation metrics and from different perspectives; use different visualization/graphs to present experimental results; correctly describe the results; provide in-depth explanations and interpretation; identify insightful/interesting/surprising findings if possible.

	8-9	include data description, baseline methods, evaluation metrics; evaluate the proposed method using different evaluation metrics but only from the single perspective of overall accuracies; use different visualization/graphs to present experimental results; correctly describe the results; provide explanations and interpretation.
	7-8	include data description, baseline methods, evaluation metrics; evaluate the proposed method using different evaluation metrics but only from the perspective of overall performances/accuracies; use different visualization/graphs to plot experimental results; correctly describe the results.
	0-7	missing one or all of the components: data description, baseline methods, evaluation metrics, overall performance comparisons, result description.
Reporting Structure and Writing	9-10	follow the suggested report structure; easy to follow; free of typos; a beautiful paper shape/layout.
	8-9	follow the suggested report structure; satisfactory clarity; some typos; a satisfactory paper shape/layout.
	7-8	follow or do not follow the suggested report structure; acceptable clarify OR lack of clarity but somewhat understandable; an acceptable or OK paper shape/layout; quit a few typos.
	0-7	DONOT follow the suggested report structure; difficult to follow or even unreadable; full of typos; a bad paper shape/layout.