INFS4203/7203 Project

Semester 2, 2024

Due dates:

16:00 on 13th September 2024 for project proposal (Phase 1, 15%)

16:00 on 25th October 2024 for project report (Phase 2, 20%)

Important Assignment Submission Guidelines:

1. All assignments must be submitted exclusively through the UQ Blackboard. No other forms of submission will be accepted.
2. Failure to submit an assignment appropriately before the due date will result in a penalty, as outlined in the ECP.
3. It is your responsibility to ensure that your assignment is successfully submitted before the designated deadline.
4. Please note that email submissions will not be accepted under any circumstances.

Overview

The assignment aims to assess your ability to apply data mining techniques to solve real-world problems. This is an individual task, and completion should be based on your own design.

For this assignment, you will individually complete a project proposal and implement it to develop data mining methods applicable to test data. You can choose either:

- A data-oriented project, or
- A competition-oriented project.

To complete the project, you need to submit a comprehensive proposal in **Phase 1**, clearly describing the data pre-processing, tuning, model training, and evaluation techniques you plan to apply. Based on this proposal, in **Phase 2**, you will submit a project implementation and a report on the final test results.

Track 1: Data-oriented project

In this data-oriented project, our dataset named *DM_project_24.csv* is designed to closely simulate real-world scenarios, reflecting the inherent complexities found in natural data. In real-world applications, many considerations revolve around deciding how and which data mining techniques to apply to the given data to benefit future predictions in the testing phase. The dataset exhibits a compelling resemblance to naturally occurring data, offering an excellent opportunity to study and develop robust solutions applicable to real-world data analysis.

In this project, you will be provided with a dataset named *DM_project_24.csv* which is the **training data** we will use. Except for the first row, each row in the data file corresponds to one data point. There are 1600 data points in this dataset, formed by micro-array expression data and functional data of 1600 genes of *E. coli*, a bacterium commonly found in the lower intestine of warm-blooded organisms. The first 103 columns ("Num (Col 1)" to "Num (Col 103)") are numerical features describing their expression levels. The following 2 columns ("Nom (Col 104)" and "Nom (Col 105)") are nominal features describing the gene functions. If the gene has a special function, it will be denoted as 1; otherwise, 0. NaN denotes that the feature value is missing at the position. The final column "Target (Col 106)" is the label for the gene indicating whether the gene has the function "Cell communication". In this column, the positive class is denoted as 1, and the negative class is 0.

The main objective, based on the provided labeled data, is to develop a classifier capable of accurately classifying a data point into one of the two classes for unseen data. The classifier's performance will be evaluated by the teaching team **using the test data released in Week 9**, with the ground truth labels accessible only to the teaching team.

Phase 1: project proposal (15 marks)

In the initial phase of the project, you are required to submit a proposal by 16:00 on 13th September 2024. This proposal should outline your comprehensive plan for the project, detailing the learning process and the timeline for Phase 2. It is important to note that this proposal does **not** require the submission of any actual code or the reporting of training, validation, or test results. Furthermore, an abstract is not necessary for this phase. The emphasis should be on providing a clear and detailed outline of your approach and the planned timeline for Phase 2.

The proposal, valued at a total of 15 marks, should be presented as a unified, coherent document that articulates your comprehensive strategy for the project. It is essential that the proposal is not formatted as a series of question-and-answer sections, but rather as a flowing narrative that integrates the following four key aspects seamlessly into the overall discussion. This format is crucial for demonstrating your ability to synthesize information and formulate a strategic plan. Note that you should ONLY use techniques covered in INFS4203/7203 from Week 2 to Week 8, specifically excluding any advanced content taught during these weeks. Techniques beyond those basic ones delivered in the specified weeks, as well as any advanced techniques, are NOT allowed.

1. **Pre-processing Techniques (3 marks):** Based on your analysis of the dataset, discuss whether pre-processing techniques such as outlier detection, normalization, and imputation should be considered. Describe how to determine the appropriate techniques using cross-validation. Then, integrate this

discussion naturally into your overall project strategy by outlining how you will apply these techniques to the current data to optimize the predictive performance of your models.

2. **Application of Classification Techniques (5 marks):** With the pre-processed data, outline the process for applying the four specified classification techniques learned in lectures—decision tree, random forest, k-nearest neighbor, and naïve bayes. Describe the necessary model selection and hyperparameter tuning via cross-validation. Integrate into your discussion the need to establish and provide reasonable ranges for hyperparameter searches, detailing how these ranges are determined and their relevance to the overall strategy. Additionally, consider an ensemble of the classification results from different models to enhance prediction accuracy.
3. **Model Evaluation (3 marks):** Detail your approach for evaluating the models using cross-validation, specifically addressing how the chosen evaluation metric is suited to the data distribution of your project. This section should reflect a thoughtful analysis of why certain metrics are preferred, based on the characteristics of your dataset.
4. **Project Timeline (1 mark):** Provide a detailed timeline for the implementation of your project during Phase 2. This timeline should not merely list milestones but must also include a justified explanation of the time allocated to each activity. Clearly articulate how each segment contributes to the overall project goal, ensuring that the timeline demonstrates a well-planned approach. It should reflect a succinct and feasible plan, showing how the implementation and testing phases can be completed on schedule, with all deliverables ready for submission by the due date of Phase 2.

The **remaining 3 marks** will be awarded based on the presentation of the proposal. The proposal should be well-structured to facilitate understanding and presented in a neat and professional manner, including a correctly formatted bibliography as per the examples in the provided template. Marks will be deducted for formatting, spelling, grammar, bibliography, referencing, or punctuation errors that hinder the comprehension of the proposal.

Hints:

1. **Focus on Application:** In the proposal, avoid detailed explanations of the mechanics of each technique or how to calculate each metric, unless necessary. Instead, emphasize practical considerations such as the criteria for deciding which techniques to apply to your data.
2. **Conjugation of Techniques:** The selection of pre-processing techniques should be directly tied to their impact on classification outcomes. It is crucial to use cross-validation to identify the most effective combination of pre-processing and classification techniques. Some combinations may yield better performance than others due to their synergistic effects.
3. **Evaluating Results:** When assessing the effectiveness of different methods using cross-validation, consider both the mean and standard deviation to determine which result is superior.
4. **Ensemble Techniques:** Consider the potential benefits of using an ensemble of ensembles, such as combining the outputs of a random forest and k-NN through majority voting.
5. **Consistency in Testing:** Ensure that the same pre-processing techniques applied during the training phase are also used during the testing phase.

Using of Generative AI

Artificial Intelligence (AI), such as ChatGPT, provides emerging tools that may support students in completing this assessment task. Students may appropriately use AI in a revision of existing authentic

assessments only. This may involve correcting grammar errors, improving sentence structures, enhancing clarity of expression, or making other relevant revisions. **Students must clearly reference any use of AI in each instance to complete the assessment task.** A failure to reference AI use or any other way of using AI beyond the revision of existing authentic assessment may constitute student misconduct under the Student Code of Conduct.

Specifically, if you use generative AI tools to help revise your proposal, you should

1. Acknowledgment Section: Include a dedicated section in your work where you acknowledge the use of AI tools and mention which parts of the content have been revised by AI.

2. Documentation (not under the four page limit): Keep track of the AI interactions and any generated outputs during the revision process. This documentation can serve as evidence for the use of AI if needed.

Students may be required to demonstrate detailed comprehension of their written submission independent of AI tools through an in-person interview.

Format

The proposal should follow the style of ***Proposal_Template.doc***. The submission should be **within four pages, including all references and illustrations (if needed)**. References should be properly provided once necessary, even if you use contents from lecture slides. Non-peer reviewed web sources could be used and should also be properly cited.

Submission

The proposal should be submitted

- in PDF or Doc (Docx) format, other formats are not acceptable, and
- through the "Proposal submission" Turnitin link provided at Blackboard -> Assessment -> Project -> Proposal submission before the deadline.

You are allowed to submit the proposal multiple times before the due date. Only the last submitted version will be marked. A penalty of 10% of the maximum possible mark will be deducted per 24 hours from the time submission is due, for up to 7 days. After 7 days, you will receive a mark of 0 (see ECP for details).

Do not submit codes in Phase 1, even if you have analyzed data by programming.

Phase 2: Project report (20 marks)

In this phase, you will implement the strategies outlined in your proposal to classify the test data, which will be available in Week 9. It is important to note that you are permitted to use only the techniques covered in INFS4203/7203 from Week 2 to Week 8, specifically excluding any advanced content introduced during these weeks. Additional details about the format, marking standards, and submission guidelines will be released in Week 9.

Regarding the programming tools, you have the flexibility to choose any language that best meets the needs of your project for Phase 2.

Track 2: Competition-oriented project

In this project, you are required to participate in an online data mining competition that aligns with the learning objectives of this course. The competition must **have a leaderboard**, **offer monetary rewards** and **conclude no later than October 24th, 2024**. Additionally, it should have **a minimum of ten competitors** to ensure a competitive environment.

If you identify a suitable competition, please express your interest by registering through this link: Express of Interest Form. Availability is limited to 10 students and will be determined based on the timeliness of your EOI submission and the relevance of the competition to the course's learning objectives.

Please note that the competition must offer monetary rewards. Therefore, most entry-level Kaggle competitions labeled under the tags “Getting Started,” “Playground,” or “Community” are NOT eligible for this project. You can explore competitions outside these categories to find those that qualify.

Since this project will be evaluated based on your position in the leaderboard, please format your Kaggle username as INFS4203_2024_sxxxxxxxx (where xxxxxxxx is your eight-digit UQ student number). Failure to do so will result in the project not being marked in Phase 2.

Phase 1: project proposal (15 marks)

In the first phase of this project, you are required to submit a proposal, which will contribute a total of 15 marks to your overall score. You can earn up to 12 marks by clearly and comprehensively describing the following aspects and the processes based on them to achieve optimal generalization performance. In this track, you are permitted to use data mining techniques beyond those covered in INFS4203/7203, based on your own exploration of the subject.

1. **(2 marks)** Provide a description of the competition task and outline the basic statistics of the provided dataset.
2. **(3 marks)** Analyze the dataset to determine if the following pre-processing techniques should be considered: outlier detection, normalization, imputation, etc. Explain how to select appropriate techniques through cross-validation and their application to the current data.
3. **(5 marks)** Describe the process of applying four classification techniques—learned in lectures or additional methods such as SVM, logistic regression, neural networks, and boosting—to the pre-processed data. Include details on model selection and hyperparameter tuning through cross-validation. Also, consider integrating an ensemble of results from different classifiers at the end of the learning phase.
4. **(1 mark)** Outline the method for evaluating the model using cross-validation on the current dataset.
5. **(1 mark)** Provide a timeline for the second phase of your project, detailing a justified, comprehensive, and feasible list of milestones.

The remaining **3 marks** will be awarded based on the structure and presentation of your proposal. The proposal should be well-organized and easy to comprehend, neatly formatted, and professionally presented, including a properly formatted bibliography as per the example in the provided template.

Deductions will be made for formatting, spelling, grammar, bibliography, referencing, or punctuation errors that impede the understanding of the proposal.

Using of Generative AI

Artificial Intelligence (AI), such as ChatGPT, provides emerging tools that may support students in completing this assessment task. Students may appropriately use AI in a revision of existing authentic assessments only. This may involve correcting grammar errors, improving sentence structures, enhancing clarity of expression, or making other relevant revisions. **Students must clearly reference any use of AI in each instance to complete the assessment task.** A failure to reference AI use or any other way of using AI beyond the revision of existing authentic assessment may constitute student misconduct under the Student Code of Conduct.

Specifically, if you use generative AI tools to help revise your proposal, you should

1. Acknowledgment Section: Include a dedicated section in your work where you acknowledge the use of AI tools and mention which parts of the content have been revised by AI.

2. Documentation (not under the six page limit): Keep track of the AI interactions and any generated outputs during the revision process. This documentation can serve as evidence for the use of AI if needed.

Students may be required to demonstrate detailed comprehension of their written submission independent of AI tools through an in-person interview.

Format

The proposal should follow the style of ***Proposal_Template.doc***. The submission should be **within six pages, including all references and illustrations (if needed)**. References should be properly provided once necessary, even if you use contents from lecture slides. Non-peer reviewed web sources could be used and should also be properly cited.

Submission

The proposal should be submitted

- in PDF or Doc (Docx) format, other formats are not acceptable, and
- through the "Proposal submission" Turnitin link provided at Blackboard -> Assessment -> Project -> Proposal submission before the deadline.

You are allowed to submit the proposal multiple times before the due date. Only the last submitted version will be marked. A penalty of 10% of the maximum possible mark will be deducted per 24 hours from the time submission is due, for up to 7 days. After 7 days, you will receive a mark of 0 (see ECP for details).

Do not submit codes in Phase 1, even if you have analyzed data by programming.

Phase 2: Project report (20 marks)

In this phase, you will implement the ideas outlined in your proposal and use the developed models to achieve a strong position on the competition's public leaderboard.

There are no restrictions on the type of programming languages you can use in Phase 2; you may select any language with which you are comfortable.

Since this project will be evaluated based on your position in the leaderboard, please format your Kaggle username as INFS4203_2024_sxxxxxxxx (where xxxxxxxx is your eight-digit UQ student number). Failure to do so will result in the project not being marked in Phase 2.

Details regarding the marking standard, format and submission process will be released in Week 9.

---End---