EGCI 425 – Group Project

This project can be done by a group of <=5 students. Max #groups allowed is 8. If there are >8 groups, smallest ones must be dissolved and members must be transferred to bigger groups.

Choose one of the given UCI datasets – at most 2 groups are allowed to work on the same dataset. Check further information from UCI Machine Learning Repository. Do further research if needed. For example, find 2-3 papers that worked on the same or similar dataset as yours.

1. Rename the project folder to "Project_XXX" where XXX = full ID of your group representative. All your materials (data, workflows, report, etc.) must be put in this folder. Build complete workflow(s) for your data analysis, starting from loading original dataset, cleansing, preprocessing, classification, etc. You can use either RapidMiner or Python, or both.

- The workflow files should be named e.g. “step 1 (do xxx)”, “step 2 (do xxx)”, … This is to remind yourself of your own steps and make my grading easier.
- Download the dataset and put the data in your folder. Don't write code or use operator that loads data from URL. All workflows must read data from & write output (if any) to your folder Project_XXX.

2. (Total 15 points) Build workflows for the following tasks:

2.1 Perform necessary data exploration/visualization and data preprocessing.

2.2 Perform at least 3 attribute selection methods.

- If your dataset has only a few attributes, you may use results from these methods only for discussion (see 3.4) but perform classification by using a full set of attributes.
- If your dataset has a lot of attributes, you may use only selected attributes for the classification.

2.3 Run 2 individual classifiers in 10-fold cross-validation mode – one of them should output trees or rules that you can comment/discuss prediction patterns. Also run 1 ensemble classifier in 10-fold cross-validation mode.

- You don't need to use the same types of attributes for all classifiers. For example, you may use nominal attributes for decision tree and transform them to numeric for SVM.
- Due to randomized nature of cross validation, don’t worry about using exactly the same training & testing data for all methods. That is, you can use 3 separate workflows or 3 separate "Cross Validation" operators for the classifiers.
- For each classifier, adjust its parameters (manually or automatically) to get high accuracy. To justify how high is high enough, you may refer to results from other papers.

3. (Total 65 points) Report, with workflow/output captures where appropriate.

Total pages including cover, references, acknowledgment of AI usage, and everything should not exceed 20 (a few exceeding pages are acceptable). The content should include the following:

3.1 (5 points) Dataset overview e.g. data source, what it is about, attributes, target class.

3.2 (10 points) Data exploration and understanding e.g. some descriptive statistics, visualizations.

3.3 (10 points) Data preprocessing. Explain data issues that require each preprocessing or give reasons for each preprocessing.

3.4 (10 points) Classifier selection and their parameter settings. Give reasons for your classification choices and their parameter tuning.

3.5 (25 points) Results and discussion.

- Report classification performance with at least: accuracy, precision & recall (or sensitivity & specificity), F-measure. Compare performance from different classifiers. You may set a positive class, e.g. patient with disease, and compare how each classifier performs w.r.t that class.
- Discuss classification models e.g. interesting patterns from tree/rules, SVM weights.
- From the classification models (e.g. tree nodes, SVM weights) and attribute selection results, discuss which attributes are important or useful for the classification.
- Any limitation.

3.5 (5 points) Others e.g. report tidiness, readability, etc.

In case of AI usage, write your own prompts and perform your own acquisition with the AI. Due to the dynamic nature of generative AI, it is very unlikely that any 2 groups will get identical generated code/text even when using identical prompts.

- Therefore, submitting identical code/text will be counted as cheating.
- Don't use generated content obtained by other groups as your own.
- Don't share generated content you get from the AI with other groups.
- If any suspicious arises, I may ask both groups to show their chat history with the AI. Failure to do so will result in cheating penalty.
- Add acknowledgment of AI usage at the end of the report.

4. **(Total 20 points) Oral presentation** for 10-12 minutes.

4.1 (10 points) Content should cover all key points in 3.1-3.4. Focus on what you did, what you found, and conclusion. No need to add all report details in the slides. A few slides with a few bullet points should be enough.

4.2 (10 points) Others e.g. punctuality, QA, etc. No need for everyone in the group to present.

Submission

1. Put the following files in Project_XXX

- Report in only 1 PDF file. The front page must contain names & IDs of everyone in your group.
- Workflows (rmp or ipynb files) + data files (original Excels, others).
- File readme.txt containing names & IDs of everyone in your group.

2. The group representative zips & submits the whole project to Google Classroom. The other group members submit only readme.txt to Google Classroom.

3. Late submission is not accepted.