IT1244: Artificial Intelligence: Technology and Impact
Semester 1, 2024/2025

Course Project

This course project helps you to apply the AI/ML concepts that you have learnt in IT1244 into real problems. The course project should be done in groups of 4 students. For this project, the members are responsible for dividing up the work equally and making sure that each member contributes to the project.

Revision history of this project document:

- 22th September, 2024 – Initial release
- 26th September, 2024 – Proposal deadline updated

1 Project Submission

Project is to be submitted to Coursemology Project submission folder (one submission per team).

1.1 Due dates

- Project proposal submission[1] – **7th October 2024, 10 pm**
- Final project submission – **30th October 2024, 10 pm**
- Project presentations are scheduled to commence from **4th November 2024**.

1.2 Deliverables

The following are the deliverables for this project:

(a) **Code folder:** Implement your project in Colab or in Jupyter notebook, with proper comments and documentations for your code. This folder should contain your colab/jupyter notebook and any additional source codes that you use for your project.

(b) **Model & Dataset folder:** Place your trained models inside this folder. If you use any external data for your project, include it in this folder as well. For datasets or trained models larger than 10 MB, upload them to your Google Drive and provide a viewable and downloadable link in a text file. Do not include any model or external data exceeding 10 MB in the submission folder.

(c) **Readme.pdf:** Provide step-by-step instructions on how to run your project. The grader will follow these steps to execute your code.

(d) **Project Report (pdf):** Write a clear, succinct 4 page report about your project. For more details, see the subsequent section 2.3.

All the above should be put into a folder and the folder should be archived to a zip file. Submit only the zip file to the Coursemology project submission folder. The following naming format should be used: "IT1244_Team<number>_Project.zip". Example: "IT1244_Team5_Project.zip". Team numbers will be assigned later.

[1] `https://forms.office.com/r/7FGeVbQmfZ`

1.3 Presentation Slides

Project presentation slides should be submitted separately to Coursemology after your presentation with the same naming format. Example, "IT1244_Team5_Presentation.pptx"

2 Grading

The maximum score for this project is 100 marks. The grading rubrics, report formats and presentation details are given below:

2.1 Requirements and Guidelines for Project Submission – (3%)

To earn the 3% mark, it is essential that you strictly follow the requirements below to avoid any delays in project grading:

- Complete the project submission form with all required information in the specified format. Use your email in the following format only: e0123456@u.nus.edu. Friendly NUS email addresses are not acceptable and may cause logistical issues in grading.
- Ensure that your name in all submissions matches exactly the format used in Coursemology.
- Coordinate with your team members to verify the correct information before completing the form.
- Ensure that the project submission zip file follows the format outlined in section 1.2.

Failure to meet any of these requirements will result in receiving 0 out of the 3% project mark.

2.2 Project Main Component – (47%)

2.2.1 Feature Extraction and Visualization

- Are you applying any method for feature selection?
- Have you visualized the features that you are using with respect to the outputs that you need to predict?
- How are you handling rare/missing values?
- What kind of algorithms are you using for effective feature representation?

2.2.2 Related works

- Which articles or research papers did you draw ideas from for your implementation? (We expect you to explore 2-3 relevant sources to provide context for your work). For example, if you're working with an audio dataset, you should reference 2-3 papers related to audio classification.
- What steps did you take to address the limitations identified in the studies you reviewed?

NOTE: Make sure to give the article/research paper references in your submitted code in colab/jupyter notebook, somewhere at the very start for our convenience. Also mention your approach for overcoming their limitation in 2-3 sentences as a comment.

2.2.3 Machine Learning Algorithm Details

- Are the AI/ML solutions adopted appropriate for the problem?
- Are the AI/ML solutions implemented correctly?
- What are the original elements done in the project?
- Code Documentation: Are there sufficient comments provided for the grader to execute and comprehend your code?
- Attributions: Are appropriate references, acknowledgments, or citations given for the code taken from external sources?

2.2.4 Experiments and Results

- Did you implement multiple models (baselines and best)?
- Did you compare with any previously work?
- Did you perform training, cross validation and testing of your model properly?
- Did you use reasonable performance metrics?
- Note that during comparison, you can either improve on some baseline model gradually and show improvement, or you can compare different models that uses completely different perspectives to solve the same problem.

2.3 Project Report – (30%)

The project report should be strictly following the AAAI format[2] and should not exceed 4 pages (does not include reference and appendix section). The report should be precise and explain your contributions clearly. The following organization of your report is recommended:

Introduction – Describe the problem that you want to solve using AI/ML, why it is important, and what AI/ML techniques that you plan to use and why. Investigate the works (2-3 recent works) that has been done with respect to the problem, what methods have been used to solve (if any) and what are their drawbacks. You can ignore the abstract section and start the report with introduction.

Dataset – Explain the dataset that you are using, what are the issues with the dataset and what analysis and processing that you did to the dataset (includes visualization and plots), etc. In addition, if you use external data to help with your project, explain the reasoning behind their use and how you used them.

Methods – Explain your technical approach in solving the problem that you stated in the Introduction. You will formulate the problem (mathematically if needed) and explain why you chose some particular AI/ML techniques. You do not need to provide detail description of any model - a high level overview and applicability of this model for your problem will suffice. You should provide figures or flow-charts to explain your method pipeline. If your method solves some limitations of previous works, then mention that as well.

Results & Discussions – You will explain how you have evaluated the solution – how many experiments you have run, what performance metrics you have used to evaluate

[2] https://www.aaai.org/Publications/Templates/AuthorKit23.zip

your model, how did you fine-tune your performance, etc. You will also report the results in tables, charts and figures. You should also list out your findings after running your experiments – explaining with evidence on why a particular model is performing poorly or well. Also, briefly discuss how the performance compares to humans and whether it has to outperform a human to be useful? You should briefly discuss the possible societal impacts of your project considering privacy, fairness, interpretability and impact on jobs, if relevant.

Important Points to Note:

- Avoid describing models and ML techniques in detail; just provide a high level overview even if you are using techniques not covered in lectures.
- Your report appendix should be maximum 2 pages.
- Do not include any codes in your report.
- Ensure all figures are clear, readable, and high resolution.
- Wide figures may span both columns of the report if necessary.

2.4 Presentation – (20%)

You will be given 20 min to present your project and 5 min for Q/A. More details on presentation schedules will be released later. All the members of the team are required to be present during the presentation. Ideally, every member should participate equally in presenting their work. In the final slide of your presentation, you will include the breakdown of the percentage of work completed by each group member.

2.4.1 Quality of presentation

- Is the presentation presented within the allocated time?
- Does the presentation flow well?
- Is the presentation paced appropriately?
- Did the speakers speak in an appropriate tone?
- Do the speakers explain well instead of just reading text in the slides?

2.4.2 Clarity/Understanding of the members

- Do the speakers express the ideas clearly to the audience?
- Do the speakers motivate the importance of the project to the audience?
- Does each member have the technical knowledge of the project?

2.4.3 Presentation Materials

- How are the materials in the slides? Are they neat, colorful or visually creative?
- Have appropriate figures/graphics been used to support the main ideas?

2.4.4 Important Points to Note

- You should not read out the scripts during presentation.
- Avoid showing codes on slides.
- Summarize results in custom made tables instead of taking screenshots from codes.
- Do not describe ML algorithms in detail - instead provide a high level overview and discuss how you have applied the model/technique for your project.
- Keep slide text concise, using bullet points and illustrative figures.
- Each slide in your presentation should be numbered, with a maximum limit of 20 slides.
- Conclude with a slide detailing each team member's contribution percentage out of 100, decided collectively. No complaints will be entertained afterward.

3 Project Marks Distribution Among Team Members

As previously stated, the final slide of your presentation must include the contribution percentage out of 100 for each team member. Individual member's marks will be adjusted accordingly based on this. The intention is not to foster competition within the team, but rather to ensure equitable contributions from all members. Collaborate with each other to achieve equal contributions across all team members. This section serves to penalize the members who are not cooperative or not showing up for regular discussions or meetings.

How should you determine the contribution % of each member?

- How much effort did the team member put? Even if their ideas aren't explicitly mentioned in the project code/report, you should still count that as major contribution.
- How much did he/she participate in implementation and idea generation?
- Did he/she attend the team meetings regularly?
- What is his/her contribution in the report writing and slide preparation?
- How positive was he/she in fostering a good team environment?

If you are not sure on any of the above or if you have further questions on this part, you can discuss with Prof.Prabhu.

Recommendations for Good Team Ethics

- Establish a common communication channel using platforms like Telegram, Microsoft Teams, or WhatsApp to facilitate easy communication among team members.
- Aim to have regular face-to-face meetings or virtual meetings via Zoom at least once a week to discuss project progress and address any issues or concerns.
- Maintain a shared Google Doc where each member can document their individual contributions, work progress, and discussions from weekly meetings.
- Divide the project into distinct components and assign responsibility for each component to different team members to ensure efficient implementation.
- Utilize collaborative tools like Overleaf or Google Docs for writing and editing project reports to enable seamless collaboration among team members.

4 Datasets

Here, we provide some datasets that are curated by IT1244 project coordinators. We divide the datasets into several categories to make it easier for you to browse and select. Each dataset is accompanied by a README file that provides detailed information about the dataset. To help you select a dataset, we've provided an estimated difficulty level for each task, categorized as **Easy**, **Medium**, or **Challenging**:

- **Easy**: Datasets with well arranged tabular features. There is no (or small) additional learning required for you to work on this dataset – it's likely that you only need to apply some knowledge of what you've learned in this course.
- **Medium**: Text, image, or multi-table tabular dataset. To perform well, some preliminary analysis and a literature review may be required to understand potential issues and determine the best approach for utilizing the data.
- **Challenging**: Moderately complex image, text, or audio data. Additional learning is necessary to effectively use this dataset with machine learning models. In some cases, further analysis may also be required. To achieve optimal performance, reviewing a few research papers and applying recent machine learning techniques can be beneficial.

Naturally, the higher the difficulty, the more learning opportunities arise, offering greater room for improvement, and the more we will value your work. If you choose to work with a medium or challenging dataset, you may be awarded a bonus of up to 5%. However, the maximum marks you can receive for this project will be capped to 100%.

4.1 Tabular Datasets

4.1.1 Housing Price [Easy]

Resale transacted prices in Singapore from 1990 to present, managed by the Housing and Development Board (HDB).

4.1.2 Bank Telemarketing Success [Easy]

Predictions of whether the client will subscribe (yes/no) a term deposit (variable y) after advertisement through telemarketing.

4.1.3 Cancer Detection [Medium]

Early stage cancer detection dataset containing details of the DNA fragment length frequency feature for each sample and corresponding labels such as healthy and cancer

4.1.4 Fraud in Electricity and Gas Consumption [Medium]

Fraudulent activities data from an electricity and gas company; containing information about client data and billing history from early 2000.

4.2 Textual Datasets

4.2.1 Code Review Comment [Medium]

Line level code review dataset where the correct type of operation (insertion/deletion/replacement) intended by the reviewer needs to predicted.

4.2.2 Movie Description [Challenging]

A large-scale sentiment analysis dataset based on public movie reviews in the format of raw text (document).

4.2.3 DNA Binding Protein [Medium]

A large protein sequence dataset containing DNA binding and non-DNA binding protein sequences

4.3 Clustering Datasets

4.3.1 Airline Sentiment [Challenging]

Clustering twitter posts about airline service based on their sentiment.

4.4 Time-series Datasets

4.4.1 Stock Market [Challenging]

Multiyear stock market data on a few hundred companies in US.

4.5 Image Datasets

4.5.1 Traffic Sign Recognition [Medium]

Traffic sign recognition from image.

4.5.2 Rice Leaf Disease [Easy]

Dataset containing RGB leaf images of healthy and diseased plants.

4.6 Audio Datasets

4.6.1 Spoken Digit [Medium]

A speech dataset containing noisy spoken digit in English from several speakers in a total of a few thousands recordings.

4.6.2 Urban Sound [Medium]

An audio classification dataset consisting of several thousand samples of 10 types of urban sound.

End of the Document