COMP3010J 2025 - 2026 Project Notes

Each of you got a project problem and a related dataset. Using the dataset you need to solve the given problem.

There is no one single correct way to do a project even with the same dataset. One can take different decisions and reach the correct analysis. Although there will be a general format of doing things when done correctly.

The main point to be noted is that **I will be looking how well you have understood the machine learning concepts and the processes and if you have applied what you have learned correctly.** The final model score is not very important. For example, if two students are using the same dataset, and one student gets a final F1-score of 0.71 and another student gets a final F1-score of 0.74, then that does not necessarily means that the project with F1-score of 0.74 is an automatically better project. **I will be looking into how you have performed the steps, what steps have you done and how did you explain what you have done.**

Writing just the code is not enough, even if it follows the correct steps. Whatever step you will be doing, **you need to explain in English in the notebook file. Make use of the notebook's Markup cell feature to explain in details what you are doing**. I will be grading based on how well you are explaining what you do. To do this, divide the notebook in sections, use heading using Markdown cells, write what are you going to do next, then write code, run the code for the output, then write about the results which you have got. For example, you can make a section, "Model Selection" in the notebook, and perform model selection, first you can start by writing what is model selection, if you will be comparing multiple algorithms, and for each algorithm what hyperparameter combinations will you be exploring. How do you want to compare the different algorithm-hyperparameter combinations and select the best one, etc. Then you write the code for that. Next, show the result of the code, and then explain what does the outputs indicate in a Markdown cell. You need to explain what method are you using, like if you will do crossvalidation, stratified crossvalidation, or train-validation-test partition method, or a combination of them. Explain why you decided to select one.

General Guidelines

Here are some general outline to give you an idea of what can be done. Think about what you want to do, use your learnings from the course, learn more from internet. **Note that the below points are not strict, but just explains the overall process.** You can change the arrangement of the sections in your notebook, add more or less. The below sections indicate the overall things which I would like to see, and then I will evaluate if they are done in a correct way. Note that, there can be many correct ways.

Introduction

+ Using only Markdown cells, explain what is the problem and what are you going to do, a general overview of the project.

Load and analyse data

+ Load data, convert each variable to correct formats, identify target attribute, etc.
+ Show a data health report, mean, median, min, standard deviation, percent of missing value for an attribute, etc. You need to decide what you want to show in this.

Data Cleaning

Some of the examples are

+ Identify obvious variables to remove, like, ID fields, or similar fields which uniquely identifies a row
+ For example, maybe convert a date attributes in format "dd-mm-yyyy" to three seperat attributes "dd", "mm", "yyyy"
+ If there are ordinal attributes, then convert them to integers. For example if there is an attribute which can have values "low", "medium", "high", then replace 1 with "low", 2 with "medium" and 3 with "high". If you have categorical attributes, you can use one-hot encoding to perform this, or leave it for the algorithm to make the conversion.
+ Identify how much missing data is present. Depending on how much data, take correct action. Explain it what you did.
+ Perform normalisation if you think it is needed.
+ Explain what you did and why in Markdown cells.

Data Visualisation

+ Perform some interesting data visualisations, and plots which gives some insights. You can try several plots but show a few which may show some interesting thing or trends in the data. Explain what you are seeing in the plots in Markdown cells.

Attribute Selection

+ Perform some kind of attribute selection and explain in details.
+ If you decide to even keep all the attributes, that is okay, but explain that why. Show that using experiments.

Model Selection and Experiments

+ Write strategy of your evaluation method and how you will be selecting the final algorithm, what evaluation metric, what evaluation techniques and what hyperparameters will you be using.
+ Write the code
+ Explain the results, show the comparative tables between different algorithms, plots, etc. and justify what model you have selected for what reason. Use one or more appropriate evaluation metrics.
+ Try comparing multiple algorithms, and for each algorithm try to explore multiple hyperparameters. For example, in kNN use different values of k, for Decision Trees, use different objectives (Information Gain, Gini, Gain Ratio, Maximum tree depth, etc.).
+ Perform crossvalidation, stratified crossvalidation, train-validation-test, or a combination of these, whichever you think correct.
+ Finally tell what is the best model as per you and what do you think about it.

Final Model Training

+ Do a final training using the selected best algorithm and hyperparameters and report the appropriate metric or metrics.
+ Get the final score on test partition.
+ Note that in this case you will need to do a combination of crossvalidation or train-validation experiments during the Model Selection and Experiments section, and then in this step, a train-validation-test step or a crossvalidation step.

Further Analysis and Discussion

Try to do something interesting with the final model you got.

+ Try to visualise the model if available easily
+ Given a few datapoints try to explain how those points are being classified. For example, in the case of kNN you can try to find the nearest neighbour points of one test point and show them in a table to analyse manually.

Discussion

+ Write some final discussion and comments based on the results and conclude your project
+ What is the overall message one can take from your project
+ Mention some positive points as well as some drawbacks of your project
+ Mention some future work which you might want to do