2023 Sem2

Group Project Stage 3

Due: 11:59pm on Sunday (end of week 12)

Value: 5% of the unit

Note: these instructions are long and somewhat complicated, but the work you need to do is not actually very much. It should be easy to fit into the provided three weeks of your time, if you interact frequently and apply any feedback from the tutors. Don't wait till near the due date to start! If anything in the instructions is unclear or confusing, please ask about it on Edstem, using the category "Group Report", and sub-category "Stage 3".

THE PROJECT WORK FOR THIS STAGE:

Task	Description	Group/ individual
The models created in this stage must all be predicting (in different ways) one common attribute in the one common dataset. You are allowed to use a dataset you already have from Stage 1 or 2, but you are equally free to change dataset and even domain. There are no requirements for particular origin or volume in the dataset for this stage but note that many machine learning techniques do not work well unless the dataset is quite clean. We have made available a dataset (on a topic of our choice) and any group can use that data instead, if they prefer. We recommend that you do some preliminary data analysis to convince yourself that there is some relationship between the other attributes and the one you are going to predict (otherwise predictions will not be very effective). You also need to choose how you will measure the effectiveness of predicting; we recommend that you use one of the measures that is built-in for scikit-learn to calculate, given the test data and the predictions made for those items. For higher levels than pass, you need more than one measure that you will calculate on each model.		
1	Identify an attribute that you will all make predictions about and find a dataset that contains this attribute. The attribute you are predicting may be quantitative or nominal	Group
2	Decide on the measure of success for the predictive models you will be producing. You will need to justify your choice of measure.	Group
3	Divide the dataset into a training set and a test set. We suggest having at least one-tenth of the original dataset in the test dataset.	Group
4	Coordinate in choosing the methods you will use, to each produce a predictive	Group

<table>
<tr><th>Task</th><th>Description</th><th>Group/ individual</th></tr>
<tr><td></td><td>model for this attribute, using the training dataset (the coordination is needed to avoid duplication between members, and to enable a good conclusion for your report).</td><td></td></tr>
<tr><td colspan="3">Each member needs to produce one predictive model, that will predict the chosen attribute from the values of some or all the other attributes. Details are in the marking scheme below. It is required that all the members have different ways to produce their predictive model. So you need to coordinate among the members, in case two members want to do the same approach, one at least will need to change (a bit – maybe you can each use the same general training technique, but scale the data attributes differently, or use a different subset of the input attributes, etc.)!
Each member then needs to work with the training set and the test set, to produce the material for their section in Part A of the report. This will involve writing Python code (we recommend using scikit-learn) to produce a predictive model based on the training set, and then running the model on the test set and calculating the agreed metric for how good the predictions are. Part A needs to include the code you each write, higher levels of mark require additional discussion and explanation (as indicated in the marking scheme)</td></tr>
<tr><td>5</td><td>Use Python (for example, the scikit-learn library) to produce a predictive model for the chosen attribute, from the training dataset, using the kind of model and the training method, which was allocated to you by the group. If your method for training has hyper-parameters, you should adjust them as well as possible, but only using parts of the training dataset in doing so [You must not use any of the test dataset for this.]</td><td>Individual</td></tr>
<tr><td>6</td><td>Evaluate the quality of the predictive model you produced, in terms of the measure of success that the group chose.</td><td>Individual</td></tr>
<tr><td>7</td><td>Write your section in Part A of the report, in which you present the work you have done individually.</td><td>Individual</td></tr>
<tr><td colspan="3">Working together as a group, you need to write up a presentation of what you have found about machine learning approaches. This needs to be written to communicate with readers whose focus is data science, in particular, they want to learn more about machine learning and when different approaches work well or not. We realise that your models are likely to be limited, and indeed it may be that none of the models you produced give good predictions– that's ok, just be honest in saying what you found.
The structure of the report is described in the submission section below. The structure will serve as the basics for grading for this project. From the combined document, you need to produce a PDF. As well, there needs to be a file which compresses a folder, within which are subfolders for each member, the subfolders contain the dataset the member worked with, and the code or spreadsheet for producing their analysis (both summaries and charts). One person submits both PDF and zipped</td></tr>
</table>

Task	Description	Group/ individual
folder, to the submission links on Canvas, on behalf of the whole group. Every member of the group will get the marks earned by the combined submission.		
8	Write Part B of the report, that discusses the different models and their strengths and weaknesses. This should be written for a reader who is interested in machine learning.	Group
9	Produce a PDF of the whole report, with all individual sections and the jointly written Part B and produce the compressed folder with all the data and code from each member. Submit it all.	Group

SUBMISSION FOR STAGE 3

1. There are **two deliverables** in Stage 3 of the Project to be submitted to Canvas site.
2. All two deliverables should be submitted by one person, on behalf of the whole group.
3. The overall mark from this stage will appear under report submission in Canvas gradebook.

Deliverable	Description
Report	The report should have a front page, that gives the group name, and lists the members involved (giving their SID and unikey, not their name), and then the body of the report has **two parts** as follows:
	Part A
	1. There is an initial section which briefly states the domain and the dataset you are using, and which attribute will be predicted. It also indicates how you split this into training and test data. This section is not marked as such, it is just so the marker can understand the setting for the rest of the report. (max 1 page) 2. There should be one section for each member (the section should state the SID/unikey of the group member who did the work reported in this section, max 2 pages per individual). In this section, there should be some subsections. a. A description of the way you produced the predictive model, including the Python code you wrote that produces the model, and any pre-processing e.g. rescaling some attributes. If possible, you should also give the predictive model itself (e.g. for a linear regression, you would report what coefficients each attribute has in the model; for a decision tree you would state the different decision points) b. The evaluation of how well your predictive model does in predicting; this must include the code you wrote that calculates

	some measure of effectiveness (on the test data), as well as stating the actual value of this measure for your predictive model. For higher marks, textual discussion is also needed.
	Part B
	3. This section is jointly written by the group. It is written for readers who are interested in machine learning In it, you describe the different ways the members produced predictive members, and comment on the evaluations, to draw conclusions about the benefits of the different approaches (see the marking scheme for more guidance on what is expected here). (max 2 pages)
	Write whatever is needed to show the reader that you have earned the marks, and don't say more than that! In most cases, the code to produce a summary or chart will be fairly short (a few dozen lines at most), and the evaluation of a chart should not take more than a half- page.
Data and Code	This should be submitted through the Canvas system, as a single zip or tar.gz file. So you should put have a single folder, with *<u>subfolders for each member</u>*. The subfolder for a member should contain the Python code to calculate a predictive model and calculate some measure of effectiveness of the model (as well, if you have done any further transforms on attributes before training/testing, the code for these should also be part of what is in your folder). Compress the top folder (with all these subfolders and their contents), then submit the single compressed file.

MARKING

The score (out of 5) is the sum of separate scores for each of the five components. A student's overall Stage 2 mark will come from the group (20%) and individual marks (80%). However, if all members agree to share the same mark within the group, this should be made explicitly on the front page of the report.

- M1&2 are individual marks for the person who write the individual part of Section 2 of the report.
- M3 is the group mark, and all members receive that same score.

M1: Producing Predictive Models [2 points] - Individual

This component is assessed based on the corresponding subsections of each separate member section in Part A of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Distinction criteria hold, and *there is a clear explanation of any method that is not presented in the Grok ML module, and an argument for why this is a reasonable approach to consider for the task (this discussion should go well beyond simply reporting that the model predicts well, to argue that one could reasonably hope that it might be good, in several ways)*.
Distinction	The Pass criteria hold, and *at least one of the methods used must go beyond what is covered in the Grok ML module*.
Pass	Every member (except when the situation is reasonably explained in a "Note to Marker") uses Python and the agreed training dataset, and with these correctly produces a predictive model for the agreed attribute; The code that each member wrote to produce their model (including doing any preliminary attribute transformations) must be explicitly shown in the report. The ways in which the various members' models are produced should all be different from one another (this could be different algorithmic training techniques, different choice of hyper-parameters, different scaling, or choice of input attributes, etc).
Flawed	Some predictive model is produced using Python.

M2: Evaluation of Predictive Models [2 points] - Individual

This component is assessed based on the corresponding subsections of each separate member section in Part A of the report; the uploaded data and code may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Distinction criteria hold, and, *for each approach, there is a reasonable discussion relating the outcome of the measurements to the nature of the training approach, characteristics of the dataset and any transformations done*.
Distinction	Each member (except when the situation is reasonably explained in a "Note to Marker") has correctly reported on *more than one* measure of performance of the model on the test dataset; the code that does this measurement must be explicitly shown in the report. Also, *for each approach there is a sensible discussion of the interpretation of the measurements (for example, whether it is indicating overfitting or underfitting)*.

Pass	Each member (except when the situation is reasonably explained in a "Note to Marker") has correctly reported on some measure of performance of the model on the test dataset; the code that does this measurement must be explicitly shown in the report. The ways in which the various members' models are produced should all be different from one another (this could be different algorithmic training techniques, different choice of hyper-parameters, different scaling, or choice of input attributes, etc).
Flawed	Some reasonable attempts to evaluate the effectiveness of some of the predictive models.

M3: Conclusion [1 point] - Shared

This component is assessed based on Part B of the report. Material in Part A, or the submitted data and code, may be checked by the marker as supporting evidence for claims made in the report.

Full marks	The Conclusion section has all the Distinction criteria, and also *discusses honestly and with insight, the limitations and uncertainties about the comparisons made* between different machine learning techniques (for example, what are limitations of the measurements which were used). *It draws the reader in and engages their attention with vivid and stylish prose*.
Distinction	The Conclusion section provides some accurate and clear information about the different machine learning methods that were used for this task and *provides useful insight into strengths and weaknesses* of the different machine learning methods for this task. *It also indicates features of the dataset that impact on the outcomes. It clearly links to the readers' background and aims*. The structure needs to be *logical and well-organised*.
Pass	The Conclusion section provides some accurate and clear information about the machine learning techniques that were used for this task, and how the resulting predictive models performed.
Flawed	The Conclusion section describes the machine learning techniques that were used.

GROUP ISSUES

Rules

Membership changes will only be made following the process described below. If there is any group of 4 where all the members are happy to keep the group unchanged, then it will not be forced to change. Note however that a new Canvas group has been created for this stage of the project (so that any changes made now, do not affect the marking of stage 2). Similarly, any group of 3 members from Stage 2, can choose to stay together; however, they may receive an extra person joining the group for Stage 3.

If for any reason any members in a group want to leave, then they should inform the tutor at the start of week 10 lab, by not joining physically / the breakout room for their former group. If someone who wants to leave a group will not be at the week 10 lab, they need to urgently contact their lab tutor (your tutor may the unit coordinator *if necessary*), naming the lab and group they wish to leave. The lab tutor will endeavor to form groups of the proper size, by combining people who have left groups, and/or by adding such people to existing groups with less than 4 members. If several people (from a previous group) all want to leave that group but stay together with one another, then they can let the tutor know; the tutor will try to achieve this, but it is not guaranteed. Similarly, if someone wants to join a specific existing group which less than four members, they should tell the tutor, but again this can't be certain. Note that whenever a move occurs, all members of the former group may continue to make use of any data, code or documents that had been produced in Stage 1 or 2, by the group they were part of during that stage.

Sometimes, people ask to have someone else removed from a group (usually, for non- contributing in Stage 1 or 2). This is not allowed. Instead, the people who are unhappy with someone, can choose to leave the group themselves (as described above), thus leaving the other person in the former group.

Group process

During the project, you need to manage the work among the group members. *We insist that every person do each activity and describe what they did and found in the appropriate section of the report and in the appropriate subfolder of the compressed folder that gets submitted*. We intend for the members to compare regularly and learn from one another (as well as from tutor feedback during lab sessions). Because any member's poor work will reduce everyone's score, make sure to quickly report any difficulty in working together to the unit coordinator as described above.

DISPUTE RESOLUTION

If there is a dispute among group members that you can't resolve, or that will impact your group's capacity to complete the task well,

- You need to inform your lab tutor (your tutor may contact the unit coordinator *if necessary*).
- Make sure that your email names the group, and is explicit about the difficulty

  - also make sure this email is copied to all the members of the group (including anyone you are complaining about).
- If necessary, the coordinator will split a group, and leave anyone who didn't participate effectively, in a group by themselves (they will need to achieve all the outcomes on their own).
  - This option is only available up until Monday of week 11, which is the last day with time to resolve the issue before the due date.
  - For any group issues that arise after Monday of week 11, you will need to try to resolve the problem on your own, and you will continue to be treated as a single group which all get the same mark for this Stage, based on whatever is submitted (though you should still let the coordinator know about them).

LATE WORK

As announced in the unit outline, late work (without approved special consideration or arrangements) suffers a penalty of 5% of the maximum marks, for each calendar day after the due date. That is, we subtract 0.25 marks per day from what you would otherwise get for the work. No late work will be accepted more than 10 calendar days after the due date.