# **Assessment Task 3: Data mining in action**

**Scenario**

This assignment is a practical data analytics project that follows on from the data exploration you did in Assignment 2.

You will be acting as a data scientist at a consultant company and you need to make a prediction on a dataset. The dataset can be found below.

You need to build classifiers using the techniques covered in the lectures to predict the class attribute. At the very minimum, you need to produce a classifier for each method we have covered (DT, KNN, RF, SVM and NN). However, if you explore the problem very thoroughly (as you should do in the industry), preprocessing the data, looking at different methods, choosing their best parameters settings and identifying the best classifier in a principled and explainable way, then you should be able to get a better mark. If you choose to use KNIME or Python and you show 'expert' use (i.e. exploring multiple classifiers, with different settings, choosing the best in a principled way and being able to explain why you built the model the way you did), optimise and test different models, this will attract a better mark.

You need to write a report describing how you solved the problem and the results you found. See below for requirements.

**Datasets**

Below you will find 3 datasets: a Healthcare-Dataset to build and optimise your model (it contains the target values), an Unknown-Dataset for the final model assessment (it does not have the target values - you need to predict them) and a submission sample which shows you what the file submitted to Kaggle should look like. In particular, you will need to set the column names in your submission file correctly - that is, "row ID" and "Predicted-ExpiredHospital".

[Assignment3-Healthcare-Dataset.csv](https://canvas.uts.edu.au/courses/28396/files/5730840?wrap=1)[Download Assignment3-Healthcare-Dataset.csv](https://canvas.uts.edu.au/courses/28396/files/5730840/download?download_frd=1)

[Assignment3-Unknown-Dataset.csv](https://canvas.uts.edu.au/courses/28396/files/5730841?wrap=1)[Download Assignment3-Unknown-Dataset.csv](https://canvas.uts.edu.au/courses/28396/files/5730841/download?download_frd=1)

[Assignment3-Kaggle-Submission-Random-Sample.csv](https://canvas.uts.edu.au/courses/28396/files/5730842?wrap=1)[Download Assignment3-Kaggle-Submission-Random-Sample.csv](https://canvas.uts.edu.au/courses/28396/files/5730842/download?download_frd=1)

For this dataset, you only have the attribute headings and a brief description of what they mean, which you can find here: [Assignment3-Dataset-Attribute-Description.pdf](https://canvas.uts.edu.au/courses/28396/files/5730906?wrap=1)[Download Assignment3-Dataset-Attribute-Description.pdf](https://canvas.uts.edu.au/courses/28396/files/5730906/download?download_frd=1)

**Classification task**

Build a classifier that classifies the “**ExpiredHospital**” attribute - with 0 if it is not expired and 1 for expired.

You can do different data pre-processing and transformations (e.g. grouping values of attributes, converting them to binary, etc.), providing explanations for why you have chosen to do that. You may need to split the training set into training, validation and test sets to accurately set the parameters and evaluate the quality of the classifier.

You can use KNIME and Python to build classifiers. If you do this, though, please explain more about your classifier - and be sure that you are producing valid results! You don't need to limit yourself to the classifiers we used in class, but if you do use other classifiers you need to describe them in your report and make sure you are producing valid results.

**A hint:** usually it's not a case of having a 'better' classifier that will produce good results. Rather, it's a case of identifying or generating good features that can be used to solve the problem.

**Kaggle Competition**

For this assignment, you will use the Kaggle website (kaggle.com) to submit your assignment solution. Go to this link to sign up for the competition on Kaggle: [https://www.kaggle.com/t/3d3438176c9a4104b3db859e27846abaLinks to an external site.](https://www.kaggle.com/t/3d3438176c9a4104b3db859e27846aba)**.**You need to use the link to access the project because it is a private project for students in 31250 only. Sharing the competition with anyone not relevant to the subject is strictly prohibited. To submit to Kaggle you will need to **make a Kaggle login using your UTS email address**, and set your display name (in My Profile -> Edit Profile -> Display Name) as **UTS\_31250\_xxxx**where xxxx is your student ID. **Submissions will not be considered if they don't meet these criteria.**

**Kaggle Assessment**

Kaggle assessment in real-time means that as soon as you submit the file, Kaggle will assess the performance of your classifier and provide you with the result. You can submit multiple times, but Kaggle has a limit for the number of times you can do this per day.

**Do not use the measure of performance reported by Kaggle as a measure of your test error in the final competition and optimise to it.** This is because Kaggle has two measures: a public measure, which it reports to you, and a private measure, which it keeps hidden. Instead, develop several models and estimate the test error yourself before submitting it to Kaggle. Remember that your estimate of test error is just that: an estimate. The actual private measure will probably be a little bit different.

**Kaggle Submission**

The predictions on the unknown set should be submitted as a .csv file to the Kaggle competition here:  [https://www.kaggle.com/competitions/introduction-to-data-analytics-assignment3-2023sLinks to an external site.](https://www.kaggle.com/competitions/introduction-to-data-analytics-assignment3-2023s). **Submission to Kaggle is mandatory.**

**Assignment report**

Your report should include the following information:

* A description of the **data mining problem;**
* The **data preprocessing and transformations** you did (if any);
* **How you went about solving the problem**;
* **Classification techniques used and summary of the results and parameter settings**;
* The **best classifier that you selected** - the type, its performance, how it solved the problem (if it makes sense for that type of classifier), and reasons for selecting it;
* **Kaggle Submission Score**
* **Appendix:**Screenshot of your KNIME workflow or copy of Python Code (it will not count as the number of pages)

**Assessment Weight**

This assignment is assessed as individual work. The report contributes 35%, marked out of 100 of your final mark.

**Marking Criteria**

Your assignment will be assessed according to the [Assignment3-Marking-Criteria.pdf](https://canvas.uts.edu.au/courses/28396/files/5247247?wrap=1)[Download Assignment3-Marking-Criteria.pdf](https://canvas.uts.edu.au/courses/28396/files/5247247/download?download_frd=1)

**Length**

The task requires a report of about 15 pages with the information and sections described above. Use 11 or 12-point Times or Arial fonts.

**Submission format**

Submit a report in Adobe PDF (preferable) or MS Word Doc with the filename ida\_a3\_xxx.pdf or ida\_a3\_xxx.doc where xxx is your student id. Also, you need to submit your workflow (KNIME) or program file (Python/R/..) together with the report with the filename ida\_a3\_xxx.knwf for KNIME and ida\_a3\_xxx.ipynb for Python.

**Submit to**

UTS Canvas assignment submission button.

**Due Date**

11.59 pm Sunday 05 Nov. 2023.

**Feedback**

Marks for the written report will be given within 2-3 weeks of submission. Feedback on the report will be given only to students requesting it. Emails will be sent when marking is complete.

**Relationship to objectives**

This assessment task addresses subject learning objectives (SLOs) 3, 4, 5 and 6.

This assessment task contributes to the development of Course Intended Learning Outcomes (CILOs) B.1, C.1, E.1 and F.1.

**Academic Standards and Late Penalties**

Please refer to the [subject outline](https://canvas.uts.edu.au/courses/28396/external_tools/1990).