**Task details**

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. However, if you explore the problem very thoroughly (as you should do in 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 show 'expert' use either KNIME or Python (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), this will attract a better mark.

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

You also need to attend a short oral defence of your classifier of around 5 minutes where you show the classifier (e.g. using the KNIME workflow or Python/R code) and answer some questions about it. Details about the oral defences will be given by email and in class.

**Objectives:**

* This assessment task addresses the following subject learning objectives (SLOs): 3 & 4
* This assessment task contributes to the development of the following Course Intended Learning Outcomes (CILOs): C.1. & D.1

**Using Kaggle**

The Kaggle Competition will be available at a later time.

**Datasets**

Below you will find 3 datasets: a training dataset for training and optimising 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, "ID" and "label". These datasets can also be found on the Kaggle competition page under the "Data" tab.

* [Assignment3-TrainingDataset.csv](https://canvas.uts.edu.au/courses/32639/files/8233355?wrap=1)

 [Download Assignment3-TrainingDataset.csv](https://canvas.uts.edu.au/courses/32639/files/8233355/download?download_frd=1)

 [Assignment3-UnknownDataset.csv](https://canvas.uts.edu.au/courses/32639/files/8233352?wrap=1)

* [Download Assignment3-UnknownDataset.csv](https://canvas.uts.edu.au/courses/32639/files/8233352/download?download_frd=1)
* a sample submission - (available soon)

The attribute description for the dataset is similar to that from assignment 2: [head\_description.csv](https://canvas.uts.edu.au/courses/32639/files/8233358?wrap=1)

[Download head\_description.csv](https://canvas.uts.edu.au/courses/32639/files/8233358/download?download_frd=1)

The Kaggle competition link is here: available soon

**Assessment**

Assessment is real-time. This 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 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.

**Classification task**

Build a classifier that classifies the "label" attribute. The classification goal is to predict the intrusion types. (target attribute: “label“ {(categorical: Mirai-greip\_flood, Recon-OSScan, DictionaryBruteForce)}. You can do various 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 provided training set further into training, validation and/or test sets to accurately set the parameters and evaluate the quality of the classifier.

You can use KNIME to build classifiers, or feel free to use Python scikit-learn or other packages. 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. At the very minimum, you need to produce a classifier for each method we have covered.

**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.

**Submitting via Kaggle**

The predictions on the unknown set should be submitted as a .csv file to the Kaggle competition.

**Submission to Kaggle is not mandatory, but you do need to make predictions on the unknown dataset.**

**Assignment report and oral defence**

**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;
* **Reflection:**  One page reflecting on your learning in Assignment 3. What did you learn about data mining and yourself as a result of doing the assignment? How would you approach the problem differently if you were to do it again? The more incisive and thoughtful your reflection is, the better your mark.

The **report** contributes up to 30 out of the total 50 marks. See the marking criteria below.

**Oral defence**

The **oral defence** contributes up to 20 out of the total 50 marks. At the oral defence, students need to explain how they solved the problem and answer questions about their solutions showing the workflow in KNIME or working code in Python.

You will receive a mark of either 0, 10, 15 or 20 depending on the strength of your response. See the marking criteria via the oral defense portal.

**Note:** Students who fail the oral defence will be permitted to undertake it one more time. If they pass, they will receive a maximum of 10 marks out of 20.

**Submission details**

**Length:**The task requires the submission of a report (approx. 2000-3000 words or 10-12 pages using 11 or 12-point Times or Arial fonts) as well as an oral defence of around 5 minutes. On average you will require between 24 and 36 hours to complete this assignment.

**Weight:** 50%, marked out of 100

**Criteria:** See below Rubric.

**Submit:**

* This assignment is individual work.
* Submit a report in Adobe PDF (preferable) or MS Word Doc with the filename (where xxxxxxxx is your student id):
  + report: "**fda\_a3\_xxxxxxxx.pdf"** or "**fda\_a3\_xxxxxxxx.doc"**

**Feedback:** Marks for the oral defence will be received at the end of the defence. Marks for the written report will be given within 3 weeks of submission. Feedback on the report will be given only to students requesting it. Emails will be sent when marking is complete.

**Academic Standards and Late Penalties:** Please refer to the subject outline.

**Rubric**

32130 Rubric: Assessment 3 (1)

|  |  |  |
| --- | --- | --- |
| 32130 Rubric: Assessment 3 (1) | | |
| **Criteria** | **Ratings** | **Pts** |
| This criterion is linked to a learning outcome Description of data mining problem:  Concisely and clear description uses appropriate terminology relevant to the field of Data Analytics | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 9 Pts  Full [9]  comprehensive description of the problem you are trying to solve showing an advanced level of relevant terminology | 8.28 Pts  High Distinction [7.56-9)  Clear, succinct description of the problem you are trying to solve using the correct terminology | 7.11 Pts  Distinction [6.66-7.56)  Clear description of the problem you are trying to solve, sometimes but not always using correct terminology | 6.21 Pts  Credit [5.76,6.66)  The description of the problem you are trying to solve is understandable but misuses key terms. | 5.13 Pts  Poor [4.49,5.76)  A vague description of the problem you are trying to solve, misusing key terminology. | 2.25 Pts  Fail [0-4.49)  No attempt / Attempt did not express a data mining problem. | | 9 pts |
| This criterion is linked to a learning outcome Data pre-processing and transformation:  The Steps of data pre-processing were clearly outlined, making appropriate justifications for processes chosen. | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 9 Pts  Full [9]  Comprehensive outline of data preprocessing and transformations made together with insightful justification. | 8.28 Pts  High Distinction [7.56-9)  Clear, succinct outline of data preprocessing and transformations made together with justification. | 7.11 Pts  Distinction [6.66,7.56)  Superior outline of data preprocessing and transformations made together with justification. | 6.21 Pts  Credit [5.76,6.66)  Outlined some of the data preprocessing and transformations made, possibly with justification missing. | 5.13 Pts  Poor [4.49,5.76)  Surface level list of data preprocessing and transformations, with no justification. | 2.25 Pts  Fail [0,4.49)  No attempt / Attempt did not convey the Data pro-processing or transformation processed undertaken. /The Justification offered was not relevant. | | 9 pts |
| This criterion is linked to a learning outcome How to approach the problem:  The decisions undertaken by the student were clearly explained in a sequence of reasonable steps in a logical, creative and thoughtful approach to the problem | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 3 Pts  Full [3]  Comprehensive and insightfull description of how you approached the problem. | 2.76 Pts  High Distinction [2.52,3)  Incisive, creative and thoughtful description of how you approached the problem. | 2.37 Pts  Distinction [2.22,2.52)  Thoughtful and logical description of how you approached the problem. | 2.07 Pts  Credit [1.92,2.22)  Some description of how you approached the problem. | 1.71 Pts  Poor [1.49,1.92)  Surface level description of how you approached the problem. | 0.75 Pts  Fail [0,1.49)  No attempt / Attempt did not convey how the problem was adressed | | 3 pts |
| This criterion is linked to a learning outcome Classification techniques used and summary of the results and the parameter settings:  The description of classification techniques applied to the data to obtain results was clearly stated within the summary, alongside relevant optimised parameter settings. The Results were expressed within the summary, including any insights. | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 18 Pts  Full [18]  A concise yet convincing description of the techniques employed and the results achieved. The summary effectively encapsulates the range of methodologies used, including careful parameter optimization, and communicates the key findings and insights drawn from the results. | 16.56 Pts  High Distinction [15.12,18)  Succinct, believable description of techniques used and obtained results. The summary conveys the (probably many) techniques used, and thoughtful parameter optimisation. The summary conveys the results and insights (if any). | 14.22 Pts  Distinction [13.32,15.12)  Believable description of techniques used and obtained results. Wide range of techniques used and/or careful parameter optimisation. The summary conveys the results and insights (if any). | 12.42 Pts  Credit [11.52,13.32)  List of techniques used and obtained results. Only a few techniques used but some thoughtful parameter optimisation. The results are outlined. | 10.26 Pts  Poor [8.99,11.52)  Very few, or one, technique used and obtained results and not much in the way of parameter optimisation. Some results are outlined. | 4.5 Pts  No Marks [0,8.99)  No attempt / Attempt did not convey which classification techniques / parameter settings were used in obtaining the limited presented results. | | 18 pts |
| This criterion is linked to a learning outcome Justify the classifier selected:  The justification for the classifier selected demonstrates an understanding of a range of classifier techniques, and a knowledge of pros and cons between at least 2-3 comparative techniques. | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 6 Pts  Full [6]  Insightful and comprehensive justification addressing the best classifier chosen. | 5.52 Pts  High Distinction [5.04,6]  Insightful and creative justification addressing the best classifier chosen, making reference to at least 2 other possible techniques, with reference to why / why not it was appropriate, explaining how it informed the selectionof the correct evaluation techniques. | 4.74 Pts  Distinction [4.44,5.04)  A clear justification of the best classifier making reference to at least 2 other possible techniques, with reference to why / why not it was appropriate, explaining how it informed the selection chosen using mostly correct evaluation techniques. | 4.14 Pts  Credit [3.84,4.44)  Some justification of the best classifier making reference to at least 2 other possible techniques, with reference to why / why not it was appropriate, explaining how it informed the selection provided using some evaluation techniques | 3.42 Pts  Poor [2.99,3.84)  Minor justification of the best classifier making reference to at least 1 other possible techniques, with reference to why / why not it was appropriate, explaining how it informed the selection chosen provided using minimum of evaluation techniques | 1.5 Pts  Fail [0,2.99)  No attempt / Attempt did not justify why the classifier was selected, and makes no/irrelevant/poor reference to other techniques. | | 6 pts |
| This criterion is linked to a learning outcome Reflection:  The reflection was insightful, creative and thoughtful; addressing all of the questions presented in the Task description. | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 12 Pts  Full [12]  Insightful and creative reflection on problem-solving. | 11.04 Pts  High Distinction [10.08-12)  Insightful and creative reflection addressing all of the questions asked. | 9.48 Pts  Distinction [8.88,10.08)  A clear and thoughtful reflection addressing most or all of the questions asked. | 8.28 Pts  Credit [7.68,8.88)  There is a reflection but it wasn’t deep and thoughtful and addressed few of the questions. | 6.84 Pts  Poor [5.99,7.68)  Brief and surface level reflection. | 3 Pts  Fail [0,5.99)  No attempt / Attempt did not make relevant reflections on the assignment / Did not address the questions presented in the task | | 12 pts |
| This criterion is linked to a learning outcome Writing and presentation:  The writing style and presentation quality was of a high quality, without basic grammar or spelling errors, and a clear layout with all graphics or images appropriately labelled. | |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | 3 Pts  Full [3]  The report is meticulously structured and thoughtfully presented. | 2.76 Pts  High Distinction [2.52,3)  Document writing skills are appropriate for University level. Spelling and grammar is error-free and the report is extremely well presented and organised | 2.37 Pts  Distinction [2.22, 2.52)  Document writing skills are appropriate for University level. Spelling and grammar is mostly error-free and the report well presented and organised. | 2.07 Pts  Credit [1.92, 2.22)  Document writing skills are somewhat appropriate for University level. Spelling and grammar is mostly acceptable with some areas of improvement needed with respect to the presentation or organisation. | 1.71 Pts  Poor [1.49,1.92)  Document writing skills can be improved for University level. Spelling and grammar may be vague with many aspects needing improvement. | 0.75 Pts  Fail [0,1.49)  No attempt / Attempt was not readable. | | 3 pts |
| This criterion is linked to a learning outcome Oral Defence  Oral defence criteria Clearly explain how the problem was solved with a demonstration of the workflow in KNIME (or working code in Python, R, or other tools).  Be able to accurately defend the solution by responding to questions about it. | |  |  |  |  | | --- | --- | --- | --- | | 40 Pts  Excellent  Demonstrated an in-depth investigation examining multiple classifiers (using Python/ R/ KNIME) with valid data preprocessing, parameter estimation and model evaluation. | 30 Pts  Good  Undertook an investigation with many classifiers (using Python/ KNIME or other tools), with basic data preprocessing, parameter estimation and model evaluation | 20 Pts  Pass  Used baseline classifiers and are able to satisfactorily answer questions about them. | 0 Pts  Fail  Unable to demonstrate use of classifiers or satisfactorily answers questions about them. Note: Students who fail the oral defence will be permitted to undertake it one more time. If they pass, they will receive a maximum of 10 marks out of 20. | | 40 pts |
| Total points: 100 | | |