Data Mining
COMP3009

ASSIGNMENT

Due Date: Week 11 - Friday 8-October-2021, 12:00pm Perth time (mid day).
Weight: 40% of the unit mark.

> **Note**: *This document is subject to minor corrections and updates. Announcements will be made promptly on Blackboard and during lectures. Always check for the latest version of the assignment. Failure to do so may result in you not completing the tasks according to the specifications.*

1 Overview

In this assignment, you will solve a real-world data mining problem. This assignment requires you to understand the theory discussed in the workshops, conduct some research into the data mining problem to solve, and use the skills that you should have developed through completing practical exercises to perform various data mining tasks.

> Please note that this is an individual assignment. Whilst you may discuss general data mining topics related to this assignment with other students, you must make sure that your work is not accessible by anyone else. There are a large number of choices to make and therefore it is very unlikely to have identical submissions by chance. Submissions that are very similar will be investigated for academic misconduct.

2 Problem Description

In this assignment, you will perform predictive analytics. You are given a CSV data file (`data2021.student.csv`) which contains a total of 1100 samples. The first 1000 samples have already been categorised into two classes. You are asked to predict the class labels of the last 100 samples associated with IDs from 1001 to 1100. You are given the following information

- The attribute Class indicates the class label. For each of the first 1000 samples, the class label is either 0 or 1. For each of the last 100 samples, the class label is missing. You are asked to predict these missing class labels.
- There are exactly 50 samples from each class in the last 100 samples to be predicted.

- Attributes are either categorical or numeric. Note that some attributes may appear numeric. You will need to decide whether to treat them as numeric or categorical and justify your action.
- The data is known to contain imperfections:
  - There are missing/corrupted entries in the data set.
  - There are duplicates, both instances and attributes.
  - There are irrelevant attributes that do not contain any useful information useful for the classification task.
  - The labelled data is imbalanced: there is a considerable difference between the number of samples from each class.

Note that the attribute names and their values have been obfuscated. Any pre-processing and analytical steps to the data need to be based entirely on the values of the attributes. No domain-specific knowledge is available.

Attempt the following:

- **Data Preparation**: In this phase, you will need to study the data and address the issues present in the data. At the end of this phase, you will need to obtain a processed version of the original data ready for classification, and suitably divide the data into two subsets: a training set and a test set.
- **Data Classification**: In this phase, you will perform analytical processing of the training data, build suitable predictive models, test and validate the models, select the models that you believe the most suitable for the given data, and then predict the missing labels.
- **Report**: You will need to write a complete report documenting the steps taken, from data preparation to classification. In addition, you should also give comments or explain your choice/decision at every step. For example, if an attribute has missing entries, you have to describe what strategy taken to address them, and why you employ that particular strategy based on the observation of the data. Importantly, the report must also include your prediction of the missing labels.

It is expected that you will complete the assignment using a programming language (Python/R) of your choice. It is required that you submit your Python/R program to produce the prediction. If you plan to use any extra tools/packages, you must obtain a written approval from the Unit Coordinator. This is to ensure fairness among students.

3 The Tasks

3.1 Data Preparation

In this first task, you will examine all data attributes and identify issues present in the data. For each of the issues that you have identified, decide and perform necessary actions to address it. Finally, you will need to suitably split the data into two sets: one for training and one for testing, the latter contains 100 samples with missing class labels. Your marks for this task will depend on how well you identify the issues and address them. Below is a list of data preparation issues that you need to address

- Identify and remove irrelevant attributes.
- Detect and handle missing entries.

- Detect and handle duplicates (both instances and attributes).
- Select suitable data types for attributes.
- Perform data transformation (such as scaling/standardisation) if needed.
- Perform other data preparation operations (This is optional, bonus marks will be awarded for novel ideas).

For each issue, you will need to present the following in the report

- Describe the relevant issue in your own words and explain why it is important to address it. You explanation must take into account the classification task that you will undertake subsequently.
- Demonstrate clearly that such an issue exists in the data with a suitable illustration/evidence.
- Clearly state and explain your choice of action to address such an issue.
- Demonstrate convincingly that your action has addressed the issue satisfactorily.
- Where applicable, you should provide references to support your arguments.

3.2 Data Classification

For this task, you will demonstrate **convincingly** how you select, train, and fine tune your predictive models to predict the missing labels. You will must use at least the three (3) classifiers that have been discussed in the workshops, namely k-NN, Naive Bayes, and Decision Trees. You can also select additional classifiers (both base classifiers and meta-classifiers). Attempt and report the following:

- **Class imbalance**: the original labelled data is not equally distributed between the two classes. You need to demonstrate that such an issue exists within the data, explain the importance of this issue, and describe how you address this problem.
- **Model training and tuning**: Every classifier typically has hyperparameters to tune in order. For each classifier, you need to select (at least one) and explain the tuning hyperparameters of your choice. You must select and describe a suitable cross-validation/validation scheme that can measure the performance of your model on labelled data well and can address the class imbalance issue. Then you will need to conduct the actual tuning of your model and report the tuning results in detail. You are expected to look at several classification performance metrics and make comments on the classification performance of each model. Finally, you will need to clearly indicate and justify the selected values of the tuning hyperparameters of each model.
- **Model comparison**: Once you have finished tuning all models, you will need to compare them and explain how you select the best two models for producing the prediction on the 100 test samples.
- **Prediction**:
  - Use the best two (2) models that you have identified in the previous step to predict the missing class labels of the last 100 samples in the original data set. Clearly explain in detail how you arrive at the prediction.
  - Provide your prediction in the report by creating a table, the first column is the sample ID, the second and third columns are the predicted class labels respectively. Observe and comment on the prediction that you have produced.

  - Produce a CSV file with the name `predict.csv` that contain your prediction in a similar format: the first column is the sample ID, the second and third columns are the predicted class labels. This file must be submitted electronically with the electronic copy of the report via Blackboard. An example of such a file is given below

```
ID,Predict1,Predict2
1001,1,1
1002,1,0
1003,0,0
...
1100,0,1
```

  - You must also indicate clearly in the report your estimated **prediction accuracy** for each selected model and explain how you arrive at these estimates.

- **Other inventive steps**: You may also conduct and report other inventive steps not mentioned above (bonus marks will be awarded for novel ideas).

3.3 Report

You will also need to submit a written report. It should serve the following objectives:

- It demonstrates your understanding of the problem, your research skills, and the necessary steps you have attempted to solve the tasks.
- It contains information necessary for marking your work.

Note of the following restriction on the report

> **Page limit:** your report must not exceed 20 pages. Pages beyond 20 will be ignored when marking!

What you should include in the report:

- Structure of the report
  - Cover page: this must show your identity.
  - Summary: briefly list the major findings (data preparation and classification) and the lessons you've learned.
  - Methodology: address the requirements described above for
    - Data preparation
    - Data classification
  - Conclusion: concluding remarks and other comments.
  - References: list any relevant work that you refer to.
  - Appendices: important things not mentioned above.
- Visual illustration to support your analysis which may include: tables, figures, plots, diagrams, and screenshots.

3.4 Source Code

In addition to the main report which details your analysis of the assignment tasks, you will also need to submit fully commented source code that can be used to reproduce your prediction.

You are required to include all source code (Python or R scripts) in your submission. You must provide a `README.txt` file that explains your programs and any known problems. You should also include the original data file `data2021.student.csv` in the same directory as other scripts. Note that your programs must be able to run from the command line. Please make sure you have the following master script:

- Python: `run.py`. Your program must run without error with `python run.py`.
- R: `run.R`. Your program must run without error with `R CMD BATCH run.R`.

Make sure that all your scripts and dependency are placed under the top level of your submission, do not place them under any subfolder. Before submitting your files, properly test your program by unzipping all contents to a directory and execute the above command in a terminal.

4 Mark Allocation

NOTE As per the unit outline, you need to demonstrate a reasonable attempt of this assignment. **Reasonable attempt** has been defined as scoring at least 40 marks out of 100 marks for this assignment. If you do not achieve this basic pass mark you will fail the unit regardless of how well you perform in the final assessment and the average score.

The total mark of this assignment is 100, and it is distributed as follows

- Satisfactory submission: 16 marks. This is based on
  - All requires files are submitted correctly.
  - Declaration correctly executed and submitted.
  - Your source code: your code must run without errors and produce the same prediction that you submitted.
  - Summary,conclusion and references in the report.
  - The overall presentation of the report.
- Data Preparation: 25 marks. This is based on how well you identify and address data preparation issues in the report. This includes: irrelevant attributes, duplicates, missing entries, data types, and scaling/standardisation.
- Data Classification: 29 Marks. This is based on how well you present the class imbalance, training, tuning, validation, comparison of different models, and how you arrive at the prediction as described in the report.
- Prediction: 30 Marks. This is based on two factors: actual prediction accuracy (maximum 24 marks) and your estimate of the prediction accuracy (maximum 6 marks). For the actual prediction accuracy, the allocation is as follows:

Accuracy	Marks
$< 55\%$	0
55%	1
56%	2
57%	3
58%	4
59%	5
60%	6
61%	7
62%	8
63%	9
64%	10
65%	11
66%	12
67%	13
68%	14
69%	15
70%	18
71%-74%	20
$\geq 75\%$	24

For the estimate of the prediction accuracy, the allocation is as follow:

Estimate of Accuracy	Marks
Within ± 2%	6
Within ± 3%	5
Within ± 4%	4
Within ± 5%	3
Within ± 6%	2
Within ± 7%	1
Outside± 7%	0

5 Submission

The assignment is submitted in two parts:

- The main report in PDF format must be submitted through Turnitin. A submission link will be provided on Blackboard. You should name the report file using the following naming convention `report_surname_studentID.pdf`, for example `report_trump_12345678.pdf`
- Other files must be submitted through another assignment submission link. You must put the following files in a single zip file using your surname and student ID as the name of the zip file (for example `trump_12345678.zip`):
  - □ PDF copy of the signed declaration form.
  - □ Correctly formatted and named prediction file `predict.csv`.
  - □ Source code (Python/R).
  - □ Any other files that are relevant, such as notebooks, model files, plots, screenshots that you cannot include in the report and may help explain your approach if needed.

6 Academic Misconduct Plagiarism and Collusion

Please note the following:

Copying material (from other students, websites or other sources) and presenting it as your own work is plagiarism. Even with your own (possibly extensive) modifications,it is still plagiarism.

Exchanging assignment solutions, or parts thereof, with other students is collusion. Engaging in such activities may lead to a grade of ANN (Result Annulled Due to Academic Misconduct) being awarded for the unit, or other penalties. Serious or repeated offences may result in termination or expulsion.

You are expected to understand this at all times, across all your university studies, with or without warnings like this.

END OF ASSIGNMENT