Data Mining COMP5009 Assignment

Due Date

The assignment is due Friday 7th of October at 5pm.

Late marks will be applied at a rate of 10 marks per day or part thereof, unless you have an extension. Extensions will only be given **prior** to the deadline so please contact me as early as possible to discuss your situation and organize a new deadline. If you have a CAP that allows for an extension to assignments, you still need to indicate that you need the extra time (but no additional justification is needed).

Weight

The assignment has a total mark of 100, which will count 40% to the total grade for this course. As per the unit outline, you need to demonstrate a reasonable attempt of this assignment. A *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 you perform in the final assessment and the average score.

Changes

If changes need to be made to any part of this assignment, they will be noted on blackboard and a revised version of this document may be created.

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 many 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.

Problem Description

In this assignment, you will perform predictive analytics. You are given an sqlite3 database file (`Assignment2022.sqlite`) which contains a total of 5500 samples across two tables. Table "train" contains 5000 samples have already been categorized into three classes. You are asked to predict the class labels of the 500 samples in the "test" table. You are given the following information:

- The attribute Class indicates the class label. For the "train" table the class label is either 0, 1 or 2. For the "test" table the class label is missing (NULL).
- 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 must describe what strategy was taken to address them, and why you employ that strategy based on the observation of the data. Importantly, the report must also include your prediction of the missing labels.

Tasks

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, choose and perform necessary actions to address it. Note that you will need to apply these actions to both the training and test data at the same time. At the end of this phase, you will have two data sets: one for training and one for the final testing task. 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/standardization) if needed.
- Perform other data preparation operations (This is optional, bonus marks will be awarded for novel ideas).

For each of the above issues your report should:

- Describe the relevant issue in your own words and explain why it is important to address it. Your explanation must consider the classification task that you will undertake subsequently.
- Demonstrate clearly that such an issue exists in the data with 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.

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 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 three 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 200 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 test samples. Clearly explain in detail how you arrive at the prediction.
    - Produce an sqlite3 database file with the name `Answers.sqlite` that contains your prediction in the format: the first column is the index corresponding to the 'test' table, the second and third columns are the predicted class labels. All columns should be integers. This file must be submitted electronically with the electronic copy of the report via Blackboard. An example of such a file is given below:

      ```
      index,Predict1,Predict2
      5000,1,1
      5001,1,0
      5002,0,0
      ...
      5499,0,1
      ```

    - 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).

Reporting

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.
- It conforms to the **page limit of 20 pages**. Anything beyond this limit will not be graded.

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 have 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.

Source Code

In addition to the main report which details your analysis of the assignment tasks, you will also need to submit a Jupyter Notebook that will reproduce your prediction. The notebook will:

- Run without error on **Google Colaboratory** when you select "restart and run all". You should assume that the data file for this assignment is in your root directory.
- Contain a combination of Text and Code cells that describe the tasks that you are attempting. See the practicals for examples of how to do this.
- Produce the `Answers.sqlite` file without need for further modification.
- Be named notebook_surname_SID.ipynb (example notebook_hancock_12345678.ipynb)

Once you have completed your notebook please go to "Edit -> Clear all outputs", and then "File -> Download -> Download .ipynb".

Any notebooks that fail to execute completely, or not reproduce the submitted prediction file, will lose marks.

Mark Allocation

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

- **Satisfactory submission**: 16 marks. This is based on
  - All required 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/standardization.
- **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 and estimated prediction accuracy, the allocation is as follows:

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

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_hancock_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 hancock_12345678.zip):
    - PDF copy of the signed declaration form.
    - Correctly formatted and named prediction file Answers.sqlite.
    - Source code.
    - 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.

Plagiarism

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 always understand this, across all your university studies, with or without warnings like this.