RMIT UNIVERSITY

COSC3126 - Data Mining

Assignment 3: Course project

Due: 17:00, May 23rd, 2025 (week 12)

This assignment is worth 20% of your overall mark.

1. Overview

This assignment covers the topics of Naive Bayes and advanced data mining techniques, with a focus on classification and text mining. You are also asked to carry out a data mining investigation on real-world data sets. You are required to write a report on your findings. Your assignment will be assessed on demonstrated understanding of concepts, algorithms, methodology, analysis of results and conclusions. Please make sure your answers are labelled correctly with the corresponding part and sub-question numbers, to make it easier for the marker to follow.

- This is a group project. It should be carried out in groups of a maximum of 3 students. It is up to you to form a group. Once you have formed your group, you should register your name in your group on Canvas.
- You must register your group by May 14th, 2025, at the latest. Anyone without a group after this day will be assigned to a group by the teaching staff.
- If you want to complete the assignment with less than 3 members, you need to send an email to the lecturer, explaining your reasons. However, bear in mind that the requirements and available marks will be the same as for a group of 3.
- Please submit what percentage each member contributed to the assignment and include this in your report. The contributions of your group should add up to 100%. The ones with too little contribution (e.g. less than 15% contribution) will have their marks reduced.

2. Learning Outcomes

This assessment relates to the following learning outcomes of the course.

• CLO 1: Demonstrate advanced knowledge of data mining concepts and techniques.

• CLO 2: Apply the techniques of clustering, classification, association finding, feature selection and visualisation on real world data.

• CLO 3: Determine whether a real-world problem has a data mining solution.

• CLO 4: Apply data mining software and toolkits in a range of applications.

• CLO 5: Set up a data mining process for an application, including data preparation, modelling, and evaluation.

• CLO 6: Demonstrate knowledge of ethical considerations involved in data mining.

3. Assignment Details

3.1. Task 1: Classification with Bayes method (10 points)

Task 1.1: Naïve Bayes for classification (5 points)

This part involves the following file: **Obesity.csv** given in with this assignment. A description of this data set can be found here:

https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition

The main objective is to accurately predict obesity levels using the Naïve Bayes classifier in Weka.

1. Load the data set, then run OneR (i.e., a very basic classifier) and J48 (a typical decision tree) to establish two baselines, for comparing with Naïve Bayes classifier.

2. Run Naïve Bayes classifier and present the results in terms of accuracy, confusion matrix and ROC area. What do you observe? Provide your explanation.

3. Run 5 different machine learning (ML) models (such as Ibk, Logistic, ensemble classifiers, neural networks) with the best parameters you can find for each model.

- Provide the results (including accuracy, Precision, Recall, F-Measure, time taken to build the model, and time taken to test the model) for comparison in a table.

- Provide your analysis on whether there are differences between Naïve Bayes and the others. Is Naïve Bayes method better than the previously used classifiers? Provide your explanation.

4. From the above experimental runs and result analysis, explain whether (or not) Naïve Bayes classifier should be considered as an effective data mining method.

Task 1.2: Bayesian methods for sentiment analysis (5 points)

This part involves data file **sentiment.zip** given with this assignment. A description of this data set can be found here:

https://archive.ics.uci.edu/dataset/331/sentiment+labelled+sentences

1. Preparing the data: Load the data file into Weka, then use Weka filter "StringToWordVector" to convert the text content into a list of words which will be used as a list of attributes. How many attributes are produced as the result of applying this filter?

As the instances are from three websites and stored in three files (in the zip file), you need to combine them into one data file for the experiments in the following sub-tasks. Give a short description of how you combine these instances from the provided data files.

Note: The target (label) must be the last column to run experiments in Weka.

2. From Weka, run Naïve Bayes classifier and BayesNet classifier then compare the results in terms of accuracy, confusion matrix and ROC area in a table. What do you observe? Provide your explanation.

3. Improve the results by further fine tunning some of the parameters in "StringToWordVector". Run Naïve Bayes classifier and BayesNet classifier again then compare the results in terms of accuracy, Precision, Recall, F-Measure, time taken to build the model, and time taken to test the model.

4. Try different machine learning (ML) models. Which one provides the best performance? Give a brief explanation.

Report Length: Up to five pages.

3.2. Task 2: Data mining using advanced techniques (5 points)

For this task, you will explore advanced data mining techniques. These can be classification, numeric prediction, clustering, and/or association rules finding, depending on your choice. You need to select one of the following data sets and work on it.

1. Incident management process enriched event log Data Set. More details can be found from the following UCI webpage about this dataset: https://archive.ics.uci.edu/ml/datasets/Incident+management+process+enriched+event+log
2. Online Shoppers Purchasing Intention Dataset Data Set. More details can be found from the following UCI webpage about this dataset: https://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

Alternative choices of datasets: You can propose another data set to work on for **Task 2**. However, the data set must be at least at the level of complexity (in terms of size and data types) with the data sets given above and must be with the same tasks. You need to send an email to the teaching staff with a detailed description of the data set and the tasks that you will work on for this project. You need to get written permission from the teaching staff before working on your proposed dataset.

Task 2.1: Defining data mining tasks (1 point)

You need to define at least three (3) data mining tasks for the selected data set. The tasks must extract the most useful information from the data to identify the trends and patterns for businesses to understand customers and make out important decisions.

Examples of tasks or questions are:

- To predict if a user ends up with shopping or not.
- To determine whether there are significant associations in the data.
- Which is a group of potential customers? (Customer segmentation problem).
- What are “golden nuggets” in the dataset?

Task 2.2: Advanced data mining techniques (3 points)

Your task is to analyse the selected data set with appropriate classification, clustering, association finding, attribute selection and visualization techniques selected from the Weka software menus to solve the defined tasks and identify any “golden nuggets" in the data.

You need to use at least three advanced techniques such as: Ensemble models (Bagging, Boosting, Random Forest), Neural networks, Bayes network for these tasks.

You need to provide a report for this analysis, focusing on the following aspects: Describe the strategies you have adopted, your methods, the runs you performed, any “golden nuggets" you found, your conclusions, and recommendations.

Task 2.3: Innovative techniques (1 point)

In this task you will explore or propose some innovative techniques for data mining. Examples of an innovative technique are as below:

- An attribute selection technique that selects a combination of attributes that gives the best performance (in terms of computational complexity and accuracy) for the tasks.
- A complex ensemble model or a complex combination of multiple algorithms.
- A new approach (method or technique) for the problem.

Give a short explanation about the methods used and the results obtained from the innovative techniques. If you use a model (technique) from any research work, you must provide a reference for the work.

Report Length: Up to four pages.

3.3 Task 3: In-class presentation (5 points)

You need to prepare a maximum of 15 slides for the in-class presentation and demonstration.

Task 3.1. Slide and presentation (2 points)

- Slides should follow the RMIT University template.
- Slides and presentation must clearly present the research question(s), the methods used for solving the problem(s), the results, and recommendations.

Task 3.2. Demo (1 point): The runs are without error and show the exact results as presented in the report.

Task 3.3. Q&A (2 points): Answer the questions by the lecturer and other students clearly and convincingly.

4. What to Submit, When, and How

The assignment is due at **17:00, May 23rd, 2025** (Week 12).
Assignments submitted after this time will be subject to standard late submission penalties.
You need to submit the following files:

- Your **report.pdf** file of max 12 pages (in single column format, including tables and figures, excluding a cover page and references) with a font size 11 or 12 points.
- Presentation slides.

They must be submitted as ONE single zip file, named as your group number (for example, 1.zip if your group ID is 1). The zip file must be submitted in Canvas: Assignments/Assignment 3. Please do NOT submit other unnecessary files.

Late submission after the in-class presentation is not allowed.

Assessment declaration: When you submit work electronically, you agree to the assessment declaration
https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/assessment-declaration

5. Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge, and ideas. You should take extreme care that you have:

• Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed, or mentioned in your assessment through the appropriate referencing methods.

• Provided a reference list of the publication details (sources of materials) so your reader can locate the source if necessary. This includes material taken from Internet sites, AI tools. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate reference, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

• Failure to properly document a source. “Any ideas or outputs generated by AI must be referenced accurately in your academic work, otherwise it is considered plagiarism.” -

RMIT.

- Copyright material from the internet or databases.
- Collusion between students.

For further information on our policies and procedures, please refer to the following: https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity

6. Marking guidelines

Factors contributing to the final mark will include the number of tasks attempted, the amount of exploration and demonstrated understanding of the algorithms, methodology, logical analysis, presentation of results and conclusions (see the marking rubrics in Canvas).

Marking criteria

Task 1.1	Points
1	0.5
2	1
3	2.5*
4	1

Task 1.2	Points
1	1
2	1
3	1.5
4	1.5*

Task 2	Points
1	1
2	3
3	1

Task 3	Points
1	2
2	1
3	2

* Scores will be given based on the performance of your best model relative to all students in the class.
  + Full points (100%): Top 10% of submissions
  + 70% points: Next 25% of submissions
  + 50% points: The rest of submissions
  + 0 points: No attempt or submissions with incomplete work, lack of experimental rigor, or not meeting the basic requirements.