

The Department of Computer Science

**CIS4515**

**Practical Data Analysis**

Level 7

Coursework 2

2023/2024

### Module Leader: Dr Muhammad Usman

🕿 01695 59 9718

**Email: usmanm@edgehill.ac.uk**

## *Administrators:*

### 🕿 01695 65 7603

|  |
| --- |
| Coursework 2 |

Weighting: **60%**

Draft hand in: **8/4/24 and 15/4/24** (for formative feedback, will not be marked, in class feedback)

Hand in: **24/4/24** (via a Turnitin dropbox on blackboard)

Feedback: **20/5/24**

Learning Outcomes Assessed:

1. Demonstrate proficiency in combining analysis methods and visualisation tools to build complete big data analysis solutions (CW2)
2. Critically interpret and evaluate the results of a big data analysis solution in order to inform the decision-making process (CW2)

|  |
| --- |
| Introduction |

This assignment is the second part of the formal assessment of CIS4515. A case study will present the students with a scenario in which an application requires implementation to meet a set of requirements. Students will be required to consider the requirements of the application and develop a solution which makes use of appropriate advanced data analysis tools and techniques. The final submission will include a critical review of the approach taken and the justification for the decisions made during the design process.

This is an **individual** assignment, and you are required to design, develop and evaluate an appropriate data analysis system for a specific scenario. This will involve the identification of requirements for the projects, critical analysis of suitable data analysis techniques and the creation/presentation of a suitable solution for the scenario.

|  |
| --- |
| Task Description |

You are a member of an investment corporation looking to make a new investment in a third-party Android Application Development (AAD) company. Your team has identified three candidate AAD companies, namely AAD\_1, AAD\_2 and AAD\_3. Each AAD company owns three Android applications which are currently available for download in Amazon. The table below shows the Android application ID codes for the three AAD companies.

Table 1 Android Application codes of the 3 AAD companies

|  |  |
| --- | --- |
| Company Name | Android Application Codes |
| AAD\_1 | B004NWLM8K, B004Q1NH4U, B004LPBTAA |
| AAD\_2 | B004S6NAOU, B004R6HTWU, B004N8KDNY |
| AAD\_3 | B004KA0RBS, B004NPELDA, B004L26XXQ |

Your task is to identify the most successful AAD company (which is likely to return profit to your investment corporation) based on the customer reviews that are posted in Amazon for each Android application. For this, you will design, develop and evaluate an advanced data analysis software which employs sentiment analysis techniques to automatically analyse a real-world dataset consisting of Amazon reviews relevant to Android applications. The developed software will automatically identify whether a given Amazon review expresses a positive, neutral or negative sentiment. You are provided with a training dataset consisting of 20,000 reviews. The training dataset will be used for developing/training the sentiment analysis algorithms. Moreover, you are provided with a test dataset of 20,000 reviews which can be used for evaluation purposes. Note that there is no overlap between the reviews in the training and in the test dataset. Each review, in both training and test datasets, consists of: a) the review text as provided by an Amazon customer, b) a class label with values 1 (negative review), 2 (neutral review) and 3 (positive review) and c) the ID code of the corresponding Android application.

In order to complete the assignment, you are required to complete the following steps:

1. Draw upon the literature to critically analyse and compare existing sentiment analysis systems.
2. Develop at least two sentiment analysis algorithms using an appropriate supervised machine learning approach (e.g., k-Nearest Neighbour or Support Vector Machines).
3. Evaluate the sentiment analysis algorithms on the provided test dataset.
4. Critically analyse the performance of the sentiment analysis algorithms based on evaluation results (i.e., precision, recall, F-Measure) and identify the best algorithm for your task.
5. Apply the best performing sentiment analysis algorithm to all reviews (in the test dataset) of the 9 Android applications (see Table 1 for the corresponding ID codes).
6. Based on your analysis results identify the most successful AAD company.
7. Write a report of approximately 2,000 words detailing all the steps you have undertaken to evaluate the sentiment analysis algorithms. Your report should be structured as follows:
   1. **Introduction**: provide background information; the purpose and objectives of the sentiment analysis task. Do not assume that the reader has specialised knowledge of the area that you are describing and be sure to explain any technical terms that you use.
   2. **Literature review:** discuss and critically analyse existing systems for sentiment analysis. Explain what their advantages and limitations are.
   3. **Methodology**: describe at least two algorithms that you have implemented for this task. Explain how the algorithms learn to discriminate between instances of different classes, their strengths/weaknesses. Provide statistics of the dataset (i.e., features, number of instances in training and test dataset).
   4. **Experiments**: Describe the experiments you have performed to evaluate the sentiment analysis algorithms. Report the performance of the algorithms in terms of different evaluation metrics. Critically analyse the performance of the methods and identify the best sentiment analysis solution.
   5. **Analysis Results:** Automatically analyse the reviews of the 9 Android applications. Based on your analysis results explain which AAD company is likely to be the most successful.
   6. **Conclusions**: provide a summary of the aims of the project and how well you’ve achieved them.

|  |
| --- |
| What you should submit |

1. **Draft version of Report (approximately 2,000 words):** 
   1. This is **optional** but you are strongly encouraged to submit a draft version of your report in order to receive feedback. Draft report cannot be more than a single side of A4 or a structure as per the faculty regulations of draft work. The focus of feedback on the plan will be around the structure and context of the assessment. The focus of the feedback on the draft will be about the academic level attained by the student through demonstration of analysis. Feedback may be verbal and/or written, as appropriate, and may occur via face to face tutorials (groups), emails or on-line tutorials. Generic feedback may be used where appropriate.
   2. **When?** **week 10 (8/4/24)**
   3. **Where? During seminar**
2. **Final submission of report (approximately 2,000 words) (60%)**
   1. **When? 24/4/24 (23:59)**
   2. **Where? TurnItIn on Blackboard**
3. **A copy of your source code and build instructions and any supporting software.**

|  |
| --- |
| What will be assessed |

1. Understanding of the data analysis algorithms, validity and justification of design and implementation choices of the developed solution.
2. Adherence to an academic writing style
3. The developed source code and other assets such as figures to summarise analysis results
4. Documentation of the implemented algorithm and referencing

|  |
| --- |
| Submission Guidelines |

Your documentation and a copy of each completed task should be submitted in *electronic format* through the University VLE, Blackboard. A drop box will be set up near the deadline to allow you to submit them both. Remember that you should leave enough time to upload your work, and if there are lots of other students doing the same thing then you should prepare for this eventuality.

If you have any questions/queries, please contact Muhammad Usman via email: usmanm@edgehill.ac.uk

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Coursework Marking Criteria** | | | | | | | |
| **Learning Outcome**  **Assessed** | **Mark  (0 – 29)** | **Mark  (30 – 39)** | **Mark  (40 – 49)** | **Mark  (50 – 59)** | **Mark  (60 – 69)** | **Mark  (70 -84)** | **Mark  (85 – 100)** |
| **Grade:**  **Fail** | **Grade:**  **Narrow fail** | **Grade:**  **Pass** | **Grade:**  **Good** | **Grade:  Very Good** | **Grade:  Excellent** | **Grade:  Outstanding** |
| **Knowledge &**  **Understanding**  Combining analysis methods and visualisation tools to design complete big data analysis solutions  **(15%) (LO3)** | Work has no structure or direction, very poor subject knowledge, serious and fundamental factual errors.  Mostly irrelevant material and/or very superficial. | The work shows some effort and minimal understanding of the subject topic(s).  Report shows some understanding of essential concepts of data analysis.  Report shows minimal understanding of the relevant areas. | Evidence of some research carried.  The solution design demonstrates understanding of key principles of implementing data analysis systems but does not identify advanced concepts. | Discusses key components of data analysis implementation such efficiency, ease of implementation, suitability etc.  Alternative data analysis algorithms are identified but not discussed. | As Good but work shows evidence of a good understanding of the material, compiled in a way that is mostly clearly argued and relevant to the set task.  Good research carried out demonstrated using a range of appropriate literature sources with little or no superfluous information. | As Very Good but excellent in-depth research conducted demonstrated by the use of a very wide range of appropriate literature sources with little or no superfluous information.  Identifying a wide range of potential data analysis solutions techniques. | As Excellent but outstanding in-depth research demonstrated by the use of an extremely wide range of appropriate literature sources with no superfluous information.  As excellent and discusses & demonstrates areas of solution that have implemented initiative and work correctly. |
| **Cognitive Skills**  Critically interpret and evaluate the results of a big data analysis solution in order to inform the decision-making.  **(20%) (LO4)** | Nothing is submitted or there is no evidence of a satisfactory attempt. | Partial awareness of possible solutions.  Much of what is presented is irrelevant or incoherent, there are some misunderstandings in the work and no attempt has been made to construct a logical discussion. | Work shows some identification and evaluation of current data analysis systems, including strengths and weaknesses.  No alternative options have been considered, only one data analysis algorithm has been evaluated. | Satisfactory identification and evaluation of current technologies demonstrated by relevant literature.  Satisfactory analysis, evaluation and critique of technologies. | Good identification and critical evaluation of current technical factors demonstrated by the wide range of research.  Alternative data analysis algorithms identified, evaluated and discussed, but with little comparative analysis. | Excellent identification and critical evaluation of current technologies demonstrated by the analysis, evaluation and critique of said technologies supported by inclusion of a range of appropriate situation observations.  Multiple data analysis algorithms have been created and evaluated.  Well-structured and supported work.  Alternative data analysis algorithms discussed and compared thoroughly.  Some evidence of independent thought but not necessarily originality. | Outstanding identification and critical evaluation of current technologies, demonstrated by the extremely broad range of technologies contrasted, and supported by outstanding analysis, evaluation and critique.  Evidence of independent thought and strong, well-organised arguments, relevant analytical and interpretative skills, with some original research. |
| **Computer-related practical skills**  Ability to build complete big data analysis solutions  **(20%) (LO3)** | No working system  Learning outcomes have not been met.  Code contains no comments.  Code design follows none of the concepts from lectures and worksheets. | No working system is provided  Writing style is poor with many errors and not in correct format.  No alternative options have been considered.  Code contains a comment at the start of each program to define its functionality.  Code design follows some concepts from lectures and worksheets but contains significant issues | Data analysis system is functional in part but is not complete.  Code should compile, but may not deploy to, or execute correctly.  Code contains a small number of comments to define its functionality.  Code design generally follows key concepts from lectures and worksheets but may contain major issues. | Developed data analysis system is functional and complete.  Code should compile, deploy and run without errors.  Code contains a number of comments to define its functionality (e.g., most methods and constructs are commented).  There is evidence that the approach to the build is systematic. | Developed data analysis system is fully functional and complete.  Code is mostly well commented to define its functionality (e.g., all methods and constructs are commented).  A mostly systematic approach to each build has been adopted (e.g., following key data analysis libraries such as numpy of scikit-learn). | Developed data analysis system is fully functional and efficient as demonstrated by evaluation analysis.  Code is well commented to define its functionality in detail (e.g., all methods and constructs are commented to include definition of inputs, outputs and processing).  A systematic approach to each build has been adopted (e.g., following key data analysis libraries such as numpy of scikit-learn). | Developed data analysis system has been excellently developed and evaluated, different algorithms are implemented allowing user choices and flexibility. |
| **Transferable Skills**  Ability to communicate information/academic writing and language **(5%) (LO3, LO4)** | Frequent significant errors in spelling, punctuation, and grammar severely affecting the meaning / comprehension of scientific arguments and reasoning.  Little, no or very repetitive use of relevant technical terminology.  Very poor sentence and paragraph structure which seriously affects the clarity of the discussion. | Some significant errors in spelling, punctuation, and grammar affecting the meaning / comprehension of scientific arguments and reasoning.  Limited or repetitive use of relevant technical terminology.  Poor sentence and paragraph structure which affects the clarity of the discussion. | Reoccurring errors in spelling, punctuation, and grammar that may affect the meaning / comprehension of scientific arguments and reasoning.  Some appropriate vocabulary incorporating some technical terminology is present.  Sentence and paragraph structures are partially correct and contain appropriate syntax aiding the clarity of the discussion. | Errors in spelling, punctuation and grammar that do not affect meaning / comprehension of scientific arguments and reasoning and are not recurring.  A variety of appropriate vocabulary incorporating some relevant technical terminology is present and generally effective.  Sentence and paragraph structures are partially correct and contain appropriate syntax and relevant vocabulary, aiding the understanding of the discussion. | Only very minor and not recurring errors in spelling, punctuation that do not affect meaning / comprehension of scientific arguments and reasoning and are not recurring.  Effective and accurate use of a variety of appropriate vocabulary, incorporating adequate and accurately used technical terminology.  Correct sentence and paragraph structures that contain appropriate syntax and relevant vocabulary, aiding the understanding of the discussion. | No or negligible errors in spelling, punctuation, and grammar.  Effective and accurate use of a variety of appropriate vocabulary, incorporating adequate and accurately used technical terminology.  Correct sentence and paragraph structures that contain appropriate syntax and relevant vocabulary, aiding the understanding of the discussion.  Adopts a professional and academic writing style and conventions, with each paragraph following the [SEED](http://eshare.edgehill.ac.uk/7165/1/Paragraph%20Structure%20%28SEED%29.pdf) structure. | No errors in spelling, punctuation, and grammar.  Highly effective and accurate use of a variety of appropriate vocabulary, incorporating adequate and accurately used technical terminology.  Consistent use of correct sentence and paragraph structures that contain appropriate syntax and relevant vocabulary, aiding the understanding of the discussion.  Adopts a professional and academic writing style and conventions, with each paragraph following the [SEED](http://eshare.edgehill.ac.uk/7165/1/Paragraph%20Structure%20%28SEED%29.pdf) structure. |