# COMP20008 Project 2

V1.0: October 20, 2020

## Part 1 - Data Linkage (12 marks)

### Marking guide for Task 1-A:<br>Naïve data linkage without blocking (4 marks)

| | |
|---|---|
| 0 mark | Not reproducible - code does not pass the test script given: this could be due to the source file is not named **task1a.py** and the output file **task1a.csv** is not produced.<br>The output is hardcoded or the output is completely incorrect. |
| 1 mark | The output is in the format as specified. At least one of Precision or Recall are below 0.6. |
| 2 marks | The output is in the format as specified. Both Precision and Recall are above 0.6. |
| 3 marks | The output is in the format as specified. Both Precision and Recall are above 0.8. |
| 4 marks | The output is in the format as specified. Both Precision and Recall are above 0.9. |

## Marking guide for Task 1-B
## Blocking for efficient data linkage (4 marks)

| | |
|---|---|
| 0 mark | Not reproducible - code does not pass the test script given: this could be due to the source file is not named **task1b.py** and the output files **amazon_blocks.csv** and **google_blocks.csv** are not produced; or<br>The output is hardcoded or the output is completely incorrect; or<br>The blocking method is not linear in time-complexity. |
| 1 mark | The output is in the format as specified. PC is below 0.6 and/or RR is below 0.7. |
| 2 marks | The output is in the format as specified. PC is above 0.6 and RR is above 0.7. |
| 3 marks | The output is in the format as specified. PC is above 0.85 and RR is above 0.9. |
| 4 marks | The output is in the format as specified. PC is above 0.95 and RR is above 0.9. |

## Marking guide for Task 1-C:<br>Report on the Data Linkage Project (4 marks)

### Scoring method: How the product comparison works

| | |
|---|---|
| 1 mark | The comparison functions, final scoring function and threshold are appropriate and are clearly and succinctly described, including any preprocessing applied. |
| 0.5 mark | Description of the product comparison is understandable but somewhat unclear. It contains insufficient or excessive unnecessary or incorrect information. Missing some key information on the methods applied. |
| 0 mark | The description misses key information about the product comparison method and/or is incomprehensible. |

### Evaluation of the overall performance of the product comparison

| | |
|---|---|
| 1 mark | Evaluation of the performance is clear and succinct; the potential improvements are sensible and not overly simplistic. |
| 0.5 marks | Evaluation of the performance lacks sufficient details or contains excessive information to give clarity and/or the potential improvements are not convincing or are overly simplistic. |
| 0 mark | Evaluation of the performance is substantially incorrect and/or the potential improvements are incorrect or missing. |

### Blocking method: How the blocking method works

| | |
|---|---|
| 1 mark | The blocking method is appropriate and is clearly and succinctly described; includes any preprocessing applied. |
| 0.5 mark | Description of the blocking method is understandable but somewhat unclear. It contains insufficient or excessive unnecessary information or some incorrect information. |
| 0 mark | The description misses key information about the blocking method and/or is incomprehensible. |

### Evaluation of the blocking method

| | |
|---|---|
| 1 mark | Evaluation of the blocking performance is clear and succinct and the overall performance relates well to the measures used.<br>The potential improvements are sensible and convincing.<br>The time-complexity of the blocking method is correct and clearly indicated or can be clearly inferred. |
| 0.5 mark | Evaluation of the blocking performance is somewhat unclear. It lacks sufficient details or contains excessive information to give clarity; or<br>Overall performance is not well related to the measures used; or<br>The potential improvements are not convincing; or<br>The time-complexity of the blocking method and its correctness are unclear. |
| 0 mark | The evaluation of the blocking performance is substantially incorrect, or the potential improvements are incorrect or missing. |

## Marking guide for Task 2-A:
## Comparing Classification Algorithms (3 marks)

**Please note that your solution program should behave the same way as intended if a similar but different input dataset in an identical format is used. Hard-coding any part of the solution will attract a heavy mark-deduction.**
**Failure to pass the test script will result in 0 mark for the task.**

| | |
|---|---|
| 0 mark | Not reproducible - code does not pass the test script given: this could be due to the source file is not named **task2a.py** and the output file **task2a.csv** is not produced.<br>The classification accuracy of the three classifiers are not printed to standard output.<br>The output is hardcoded or the output is completely incorrect. |
| 1 mark | The output file **task2a.csv** is correct but the classification accuracies are substantially incorrect. |
| 2 marks | The classification accuracies for some of the classifiers are slightly incorrect. |
| 3 marks | The output file **task2a.csv** is produced and has correct number of rows/columns and has correct values.<br>The classification accuracies are correct for all three classifiers. |

## Marking guide for Task 2-B:
## Feature Engineering and Selection(6 marks)

**Please note that your solution program should behave the same way as intended if a similar but different input dataset in an identical format is used. Hard-coding any part of the solution will attract a heavy mark-deduction.**
**Failure to pass the test script will result in 0 mark for the task.**

### Overall

0 mark — Not reproducible - code does not pass the test script given: this could be due to the source file not being named **task2b.py**.

### Feature engineering – interaction terms

1 mark — Evidence of correct implementation of interaction term pairs. Output that support the correct implementation produced and the outputs are correct.

0.5 mark — Evidence of partially correct implementation of interaction term pairs. Output that support the implementation produced and the outputs are partially correct.

0 mark — No output is produced as supporting evidence of implementation of this sub-task **(even if the sub-task is implemented in task2b.py)**

### Feature engineering – clustering

2 marks — Evidence of correct implementation of feature engineered from clustering. Outputs that support the correct implementation produced and the outputs are correct.

Evidence of implementing a good choice of k for k-means clustering. Outputs that support a good selection of k produced.

1 mark — Implementation of feature engineering from clustering and choice of k for k-means clustering are largely correct but contain some minor issues.
Appropriate outputs as evidence of the implementation are produced but choice of k is hard-coded or inappropriate.

0 mark — No output is produced as supporting evidence of implementation of this sub-task **(even if the sub-task is implemented in task2b.py)**

### Feature selection from 211 features

1 mark Appropriate method used for feature selection. Evidence of correct implementation of feature selection. Appropriate outputs that support a correct implementation are produced and the outputs are correct.

0.5 mark Implementation of feature selection is largely correct but contains minor issues. Outputs produced are insufficient to support a correct implementation and a correct method used or outputs contains some errors.

0 mark No output is produced as supporting evidence for implementation of this sub-task **(even if the sub-task is implemented in task2b.py)** or outputs are largely incorrect.

---

### PCA and naïve feature selection of first 4 original features

1 mark Correct implementation of PCA and naïve feature selection. Appropriate outputs that support a correct implementation are produced and the outputs are correct.

0.5 mark Implementations of PCA and naïve feature selection are largely correct but contains minor issues. Outputs produced are insufficient to support a correct implementation and a correct method used or outputs contains some errors.

0 mark No output is produced as supporting evidence for correct implementation of this sub-task **(even if the sub-task is implemented in task2b.py)** or outputs are largely incorrect.

---

### Performing 3-NN with 3 difference feature-sets

1 mark The classification accuracies of the test sets are correctly produced for all three feature-sets. Outputs are correct.

0.5 mark The classification accuracies of the test sets are correctly produced for all three feature-sets. Outputs are partially correct.

0 mark No output is produced as supporting evidence for implementation of this sub-task **(even if the sub-task is implemented in task2b.py)** or outputs are largely incorrect.

---

## Marking guide for Task-2C:<br>Report for Part 2 - Classification(4 marks)

### Reporting of Task-2A

1 mark Explanations of experiments in Task-2A, including pre-processing, are clear and succinct. Correct interpretation and clear reporting of the results.

0.5 mark Explanations of experiments in Task-2A and reporting of the results are reasonably clear. Interpretation of the results are largely correct. Some information lacks clarity or contain minor errors.

0 mark Explanations of experiments in Task-2A and reporting of the results are unclear and difficult to understand. Interpretation of the results are incorrect. Missing key information and contains many errors.

### Reporting of Task-2B

1.5 marks Explanations of experiments in Task-2B, including pre-processing, are clear. Descriptions of cluster-label feature generation are clear and succinct. The method used is correct. Descriptions of feature selection method (from the 211 features) are clear and the method used is correct.

1 mark Explanations of experiments in Task-2B, including pre-processing, are reasonably clear. Descriptions of cluster-label feature generation are reasonably clear The method used is largely correct. Descriptions of feature selection method (from the 211 features) are reasonably clear and the method used is largely correct.
Some information lacks clarity or are incorrect or there are some minor errors in the experiments, feature generation, or feature selection step.

0.5 mark One aspect in the reporting of Task-2B is seriously flawed or largely missing, such as explanations of the experiments, feature generation, and feature selection. *or*
More than one aspect in the reporting of Task-2B contains substantial issues or misses significant amounts of information.

0 marks More than one aspect in the reporting of Task-2B are seriously flawed or largely missing, such as explanations of the experiments, feature generation, and feature selection.

### Conclusions for Task-2B

1.5 mark Correct reporting, interpretations, and justification of the results for Task-2B. Arguments on the reliability of Task-2B are sound and suggestions for improvements are clear, practical and convincing.

1 mark Some errors in the reporting, interpretations, and justification of the results for Task-2B, or
Both of the following aspects have minor issues (incorrect or insufficient): (1) arguments on the reliability of Task-2B and (2) suggestions for improvements.

0.5 mark Some errors in the reporting, interpretations, and justification of the results for Task-2B, and
Both of the following aspects have minor issues (incorrect or insufficient): (1) arguments on the reliability of Task-2B and (2) suggestions for improvements.

0.5 mark Major errors in the reporting, interpretations, and justification of the results for Task-2B, and
Arguments on the reliability of Task-2B are sound and suggestions for improvements are clear, practical and convincing.

0 mark Major errors in the reporting, interpretations, and justification of the results for Task-2B, and
Both of the following aspects have minor issues (incorrect or insufficient): (1) arguments on the reliability of Task-2B and (2) suggestions for improvements.