Aston University
BIRMINGHAM UK

# College of Engineering & Physical Sciences
# Assignment Brief

| | |
|---|---|
| CS4730 Machine Learning | Coursework 2 |
| Dr Harry Goldingay<br>h.j.goldingay1@aston.ac.uk<br>Dr Mohammed Hadi<br>m.hadi2@aston.ac.uk | |

Assignment Brief/ Coursework Content:

This assessment consists of three sub-tasks. In the first two sub-tasks, you you will be applying techniques from the unsupervised learning and dimensionality reduction topics covered in this module. The aim of these tasks is to test your ability to apply these machine learning techniques to well-specified tasks and, where applicable, to evaluate their performance. These first two sub-tasks are each worth 10% of the overall module mark.

In the final sub-task, you will be required to solve a more complex task, described in natural language. You will be required to reformulate the problem and design a solution, justifying your design choices in terms of the properties of the problem and of the algorithms you use. This sub-task is worth 60% of the module mark.

Overall, this assessment is worth 80% of the overall module mark.

Follow the instructions below to complete the sub-tasks. The tasks require you to carry out some implementation in Python and to provide a short written justification of your choices. Across the 3 sub-tasks, your justification should be of no more than a total of 1500 words, excluding code. We recommend that you aim to use no more than a total of 500 words of justification for the first 2 sub-tasks, with the remainder used for sub-task 3. Submissions exceeding this word limit will have marks reduced by 10% for every 150 words over the limit (i.e. a reduction of 10% at 1650+ words, 20% at 1800 words, etc).

For each sub-tasks you should produce a Jupyter notebook file integrating your code and written justification. **Do not use a separate document for your written justification – it will not be marked**. The required format for submission is a zip file containing the solution files for all three sub-tasks.

**Sub-task 1: Unsupervised Learning**

Download the file `cluster1.csv` from Blackboard. It contains 600 data points. Each data point has two features.

Using Python, apply k-means to partition the data set into three clusters. Plot a graph of the resulting clusters.

Your (fictional) colleague claims to have done the above. They were visually happy with the resulting clusters. However, they decided to sense-check their results empirically. To do this, they:

- measured the reconstruction error of their clustering (as described above, with k=3)
- measured the reconstruction error of the clusters resulting from applying k-means with k=10.
- compared the two reconstruction errors.

They found that the reconstruction error with k=10 was lower than the error with k=3 and concluded that 10 clusters were more appropriate for the dataset than 3 clusters. Do you agree with your colleague? Justify your answer based on your graph, and on your understanding of the k-means algorithm.

**Sub-task 2: Dimensionality Reduction**

Download the file `kc_house_data.csv` from Blackboard. It contains approximately 21k data points, each of which contains data on houses in King County USA (adapted from this dataset).

Use a dimensionality reduction method of your choice to visualise the data in three dimensions. Evaluate the success of your dimensionality reduction procedure to decide whether the three dimensional projection of your dataset captures the important details of the original dataset. Justify your answer and, if you don’t feel that a three dimensional projection captures these details, design and implement a methodology to determine the appropriate minimum dimensionality to project the data to.

**Sub-task 3**

You have been contracted by an international healthcare provider, who is interested in whether machine learning could be used to help them develop a user-friendly model capable of detecting a stroke in patients. As a starting point, they have asked you to create a machine learning model which, given some data about a patient, can predict whether there is a stroke risk or not (or give the probability of a stroke).

The medical teams provided you with a dataset which contains medical information of more than 600 persons. They also indicated that the dataset does require some pre-processing. In particular, they suggest removing any row in the dataset which contains a missing (NaN) value. The columns are described as shown in the table below.

The team that you are part of includes both medical professionals and data scientists, and your supervisor informed you that the medical team highlighted a few requirements you must take into consideration when preparing the data.

- Concentrate on features which are well known to be contribute to the risk of a stroke. They have provided you with the following resources to help you choose the feature variables for your model.
    - Article#1
- The team recommend that you convert some of the numerical feature variables into categorical ones. To help you in doing so, they have provided you with the following documents:
    - Article#2
    - Article#3

| Column ID | Name | Description |
|---|---|---|
| 1 | RANDID | An identification number for each patient |
| 2 | TOTCHOL | Total serum cholesterol, mg/dL |
| 3 | AGE | Age at exam, years |
| 4 | SYSBP | Systolic blood pressure, mmHg |
| 5 | DIABP | Diastolic blood pressure, mmHg |
| 6 | TIMENI | Time of first angina/spasm |
| 7 | CIGPDAY | Number of cigarettes smoked per day |
| 8 | TIME | Time from baseline to first hypertension |
| 9 | STROKE | Stroke (1=Yes, 2=No) |
| 10 | BMI | Body Mass Index (14–57) |

From your client's perspective, a good solution to this problem would predict (with reasonable accuracy) whether or not a new patient (i.e., one that is not in the dataset) was at high risk of a stroke. It is also important to them that the solution performs well for both well and at-risk persons. As part of your work, they would like to know how well they could expect your solution to perform on both counts.

Use your knowledge of machine learning and of Python, supported by personal research where necessary, to design and implement a solution to the problem described above and to answer the healthcare provider's question about performance. Note that, as described in the marking scheme, you will be marked primarily on your approach to the task and your understanding as evident from your written justification rather than the final performance of your solution (although a high performing solution may be evidence that your approach is a good one). As such, make sure to include more than just your final solution in your write-up. Also include information on solution methodologies that you tried and rejected.

Note also that the content of the module is sufficient for you to complete this task well. However, as you will see from the mark scheme, marks in the higher ranges are characterised by use of citations and of independent research. To get you started, the following references contain overviews of topics which may be useful to you when preparing your solution. They are much too broad for you to implement all of their suggestions in your work, but if you find any part of them interesting or relevant then they, and the references within them, could give you a starting point for your own literature search:

- He, H. and Garcia, E.A., 2009. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9), pp.1263-1284.
- Hutter, F., Kotthoff, L. and Vanschoren, J., 2019. *Automated machine learning: methods, systems, challenges*. Springer Nature.
- Dong, X., Yu, Z., Cao, W., Shi, Y. and Ma, Q., 2020. A survey on ensemble learning. *Frontiers of Computer Science*, 14, pp.241-258.

Descriptive details of Assignment:

- Preferred Format: Jupyter Notebook
- Word Count: 1500 words (code does not count towards word limit). This is a total for all three sub-tasks.
- Preferred reference style: Harvard referencing

Recommended reading/ online sources:

- All units of CS4730
- Articles referenced and linked to in sub-task 3.

Key Dates:

| | |
|---|---|
| 09/12/2024 | Coursework set |
| 16/01/2025 | Submission date |
| 13/02/2025 | Expected feedback return date. |

Submission Details:

- As discussed at the start of the document, for each sub-tasks you should produce a Jupyter notebook file integrating your code and written justification. Submit your file as a zip file containing the solutions to each of the sub-tasks through the link on Blackboard.

Marking Rubric:

Sub-tasks 1-2 (10% of module mark each):

- **0-39%** Brief, irrelevant, confused, incomplete. Does not come close to meeting the required learning outcomes.
- **40-49%** Evidence that some learning outcomes have been achieved or most learning outcomes achieved partially. Although work may include brief signs of comprehension, it contains basic misunderstandings or misinterpretations, demonstrates limited ability to meet the requirements of the assessment.
- **50-59%** The applied element of the task has been completed in a broadly correct fashion but may have some flaws in application or methodology. Written answers and justification have been attempted but are largely descriptive and show some limitations in understanding.
- **60-69%** The applied element of the task has been completed correctly. The written answers and justification show clear understanding of any models used and of the implications of the results of the applied element.
- **70-79%** The approach to the applied element of the task has been carefully designed or chosen to produce the evidence needed for the written element.
- **80%+** As above, but with additional evidence of some or all of: attention to quality throughout the implementation, thorough understanding in experimental design, excellent justification.

Sub-task 3 (50% of module mark):

- **0-39%** Brief, irrelevant, confused, incomplete. Does not come close to meeting the required learning outcomes.
- **40-49%** Evidence that some learning outcomes have been achieved or most learning outcomes achieved partially. Although work may include brief signs of comprehension, it contains basic misunderstandings or misinterpretations, demonstrates limited ability to meet the requirements of the assessment.
- **50-59%** The given problem has been reformulated as a machine learning problem and a solution has been proposed, implemented, and evaluated. Some attempt has been made to pre-process the data appropriately. The solution has some value, but the quality may be compromised by a range of factors including misunderstanding how to best approach the problem, flaws in implementation or in evaluation methodology.
- **60-69%** Appropriate machine learning methods have been used to address the given problem, including ensuring good performance for both fraud and non-fraud cases. The written justification shows a clear understanding of why the methods employed are appropriate given their properties and those of the problem. Experimental work to empirically support this justification has been undertaken.
- **70-79%** The approach for choosing solution methodologies is systematic. An appropriate range of options has been considered and critically analysed for the suitability, both in terms of their properties and, where appropriate, through experimental comparison. This has led to a well-designed solution to the problem. The written justification documents the rationale for all choices with supporting evidence, including relevant references.
- **80%+** The chosen approach and written justification show insight into the problem. The work makes use of information from the student’s independent research and draws on academic work outside of the texts discussed in the module.