# CI7520 – Assignment 1: Classic Machine Learning

**Overview**

The subject of this assignment is to implement Classic Machine Learning solutions in Python using the Scikit-Learn library and other libraries introduced in the class. Specifically, both clustering and classification methods should be applied to the Breast Cancer Wisconsin. You can work individually or in groups of two (after a discussion and agreement with the tutor).

**Deliverables and Submission**

The coursework must be submitted **by 23:59, on Sunday 23rd February**. Follow the submission guidelines in Canvas. Feedback is due on 15th March. For each submission ensure that you include

* A zip file containing all runnable programs for the first three parts, with code written in Python.
  + For each of the first three parts, dedicate a different folder containing all relevant Python code that implements the given requirements. If the code consists of multiple files, a short README text file should be included in the folder that explains the contents and how the program runs.
* A report (3000-5000 words, excluding references and appendices), in word or pdf format.

**Rules**

* You are encouraged to look in the literature and identify methods that have already been applied to the particular problem. In this case, you must reference the relevant sources (e.g. scientific article, book, webpage).
* Any source code that has not been provided by the Lecturer must be clearly highlighted and referenced. Usage of any third party libraries must be agreed with the Lecturer beforehand.
* In case that the above rules are not obeyed, the submission may be considered for plagiarism and penalised according to the University regulations.

**Project Parts**

**PART I – Application: Load and overview data related to your theme**

The application should be able to load the data and identify its key aspects (number of dimensions, number of classes, datatype of variables).

**PART II – Application: Clustering**

1. You should use at least two clustering methods to partition the dataset into two clusters.
2. Evaluate the clustering methods using appropriate metrics such as the Adjusted Rand index, Homogeneity, Completeness and V-Measure, using the ground truth.
3. Consider and implement any configuration of the parameters of your clustering methods that could further improve the results.

**PART III – Application: Classification: Training and Testing**

1. You should use at least two classification methods to distinguish between the classes. Both the following training/testing protocols should be used:

* Split the data into training (70%) and testing (30%).
* K-fold cross-validation for K=10.

1. For both protocols, evaluate the classification approaches using appropriate metrics such as the Balanced Accuracy, F1-Score, ROC AUC, and drawing ROC curves and a confusion matrices.
2. Consider and implement any configuration of the parameters of your classification methods that could further improve the results.

**PART IV - Report:**

The Project Report should be structured as follows:

* Data: Description of the data, including the information derived in Part I
* Clustering:
  + Outline of the clustering methods used in Part II
  + Comparative analysis of all clustering methods used, including any improvements attempted. Ensure that results are reported in appropriate tables and figures.
* Classification:
  + Outline of the classification methods used in Part III
  + Comparative analysis of all classification methods used, considering both training protocols, including any improvements attempted. Ensure that results are reported in appropriate tables and figures.
* Conclusion and Conclusion:
  + Critical Discussion of any challenges imposed by the specific dataset and the pipelines for clustering and classification.
  + Critical Discussion of results.
* References
* Appendix: Include copies of all the code produced.

**Learning outcomes being assessed**

* Select and specify suitable methods and algorithms relevant for a particular data analysis process;
* Build machine learning and artificial intelligence systems using software packages and/or specialised libraries;
* Articulate and demonstrate the specific problems associated with different phases or tasks of a machine learning or artificial intelligence pipeline;
* Assess and evaluate machine learning methods using datasets and appropriate criteria;

**Marking Scheme (100%)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Criterion | 9-10 | 6-8 | 3-5 | 1-2 | 0 |
| Data handling: 10% | Data is properly loaded. The code and the relevant discussion of the data goes beyond its key aspects, expanding into an insightful analysis. | Data is properly loaded and its key aspects are estimated by the code and properly mentioned in the report. | Data is properly loaded, but its key aspects are not estimated by the code, neither mentioned in the report. | No running code for data. Unclear description of data in the report. | None |
| Technical contribution:  Clustering: 10%  Classification: 10% | A novel implementation of an existing algorithm in the bibliography has been produced and/or a novel sophisticated algorithm is suggested. | Program is mainly based on existing code that has some complexity, as it combines different modules. | Program is mainly based on existing code with no effort to combine different modules. | No running code. | None |
| Variety of methods attempted:  Clustering: 10%  Classification: 10% | A variety of methods from the bibliography have been attempted, including methods that were not presented in the class. | A variety of methods that have been discussed in the class have been attempted. | A single method for clustering/classification that has been discussed in the class has been attempted. | No method has been properly implemented. | None |
| Evaluation:  Clustering: 10%  Classification: 10% | Extensive experimental comparison/evaluation is presented for a variety of methods and/or parameters and presented results outperform the work of other students. | Experimental comparison/evaluation is presented for a few methods and their parameters, using all metrics. Presented results are at or above the average, in comparison to the work of other students. | Results for only one method and only for the default set of parameters. Results presented only for a subset of metrics. | Inadequate results presented. | None |
| Training/Testing protocols used for Classification 10% | Implementation and full results for both protocols were presented. | Implementation for both protocols, but full results only for one protocol were presented. | Implementation and full results for only one protocol were presented. | No training protocol was followed properly. | None |
| Report: 10% | Sharp report that clearly describes both the basic principles and the details of the work, including proper referencing. | Clear description of basic principles of the work and satisfactory presentation of most of the details, including proper referencing. | Adequate description of the work clarifies the basic aspects of work, although many details may have been missed. Adequate Referencing. | Poor description of the work that fails to clarify even basic aspects of the work. Inadequate Referencing. | None |
| Critical Review: 10% | Justified overview of conclusions. Critical view on the theory and the results of selected methods that reflects student’s insights beyond the teaching material. | Honest and justified overview of conclusions. Critical view on the theory and the results of selected methods consistent with the teaching material. | Honest and justified overview of conclusions. Some justification for the selection of the method and some discussion on the results. | Unjustified and dry conclusions. No justification for the selection of the method and no discussion on the results. | None |

**Late submissions:** for details please refer to the Module Guide.

Plagiarism is a serious academic offence and will always be challenged where it is suspected. When submitting work for assessment ensure it is your own and not someone else’s code, ideas, words and images. You must declare that the task submitted for assessment is your own independent work and that all sources used in the submission are referenced.

Please ensure that your name is on all components of the work presented. For any questions or issues, email:  [D.Makris@kingston.ac.uk](mailto:D.Makris@kingston.ac.uk)