Assignment 1

The goal of this assignment is to learn:

* the basics of understanding the dataset and use respective machine learning category,
* using a machine learning package for a given task
* prepare a short report demonstrating how you have applied it.

Details (with the maximum marks for each aspect of your work):

1. Have a look at the given data, understand the problem based on the dependent variable and select a machine learning category that can solve the task/problem.  Briefly explain why do you think it is the correct ML category for this problem? [2 Marks]
2. Explore and report the data and its distribution among training and testing data. Can we call it imbalanced dataset, explain your answer (yes/no) briefly?  [2 Mark]
3. Research and write down about open-source machine learning package that are freely available and select one that you think will be good and easy for this task. Your report should include a short overview of the main features of the package you have chosen. [2 marks max.]
4. To use the dataset (Plant-dataset) supplied below, you might need to do some work to prepare it for input into the ML package, depending on the ML category requirements. Document any data preparation (e.g., normalisation) steps in your report. [2 marks max.]
5. From the ML package, select two different algorithms from the category you selected and apply to the dataset. In your report, include a clear description of both algorithms. Ensure that you acknowledge all your sources of information.
6. Report the results with and without normalisation of the data. [4 marks max.]
7. Train and test your chosen algorithms using the training set provided in **plant-train.csv**. You should then test your trained models using the test set provided in **plant-test.csv**. Report on the results with appropriate performance metric e.g., accuracy that you consider best for each model on the training set and on the test set. Also include details of the classification models constructed – these may include graphics if appropriate. [ 6 marks max.]
8. Discuss in your report whether the two models give very similar or significantly different results, and why. [2 marks max.]

The goal is to predict **the dependent variable** (which may be one of two classes: **setosa** or **virginica)** based on the other four independent attributes.

Your report should not exceed four A4 pages in total. Therefore, take care to ensure that it is succinct and informative, and not overly superficial. This is an individual assignment. As you are all postgraduate students, I will treat any plagiarism (from another student or other sources) very seriously.

You must submit your report as a single PDF file in Blackboard, on or before the due date. Ensure your name, class and student ID are on it.

N.B. You should submit your assignment well in advance of the deadline. Servers can be very busy close to the deadline, and I will not accept late submissions on the grounds that the server did not allow you to upload a few minutes before the deadline. Multiple uploads are allowed; therefore, I encourage you to submit preliminary versions. I will grade the latest uploaded version only.

Your assignment will be marked out of 20, with the breakdown shown above. If any aspect of your work is plagiarised or is otherwise dishonest, you will receive 0 for the full assignment.

The deadline is **12 October 2022 23:59.**