# CW02 Introduction to Data Science Assessment ‘Significance Testing & Machine Learning’

The purpose of this assessment is to provide you with practical experience in data science, covering significance testing and basic machine learning models.

It is important that you explain your analysis – see marking scheme available as a separate document on GCULearn.

Write a notebook report, using Google Collaboratory. A single \*.ipynb file should be uploaded using the assessment submission link.

The assessment is divided into two parts as follows;

Part A - significance testing – 25 marks

* Dataset & Significance Test Selection
* Significance Test
* Interpretation

Part B – machine learning – 75 marks

* Linear Regression Analysis
* Logistic Regression Analysis
* Comparison, Analysis & Conclusions

The marking scheme is available in a separate document.

**Part A: Significance Testing – 25 marks**

For part A of the assessment, you will need to find a suitable dataset (not used in the practical examples), conduct data cleaning procedures, determine the appropriate significance test (parametric or non-parametric), implement the correct test and interpret the results accurately.

**Requirements**

**Dataset and Significance Test Selection (10 marks)**

* Choose a dataset (not used in the practical examples) with at least 100 rows, that is relevant to your interests or a problem you wish to investigate.
* Explain where you obtained your dataset, its composition, and what question you intend to answer.
* Based on your dataset selection, determine whether you need a parametric or non-parametric significance test and provide an explanation.

**Significance Test (5 marks)**

* Select and explain the most appropriate significance test based on your question and dataset.
* Consider the following significance tests for your analysis (you need only choose one):
  + Independent t-test: Compare means between two independent groups.
  + Matched pairs t-test: Compare means between two related groups (paired data).
  + Mann-Whitney U-test: Compare medians between two independent groups when data is not normally distributed.
  + Pearson correlation (Pearson r): Assess the linear relationship between two continuous variables.
  + Spearman's rank correlation: Assess the monotonic relationship between two ordinal or continuous variables.

**Interpretation (10 marks)**

* Carefully interpret the results of your significance test referencing probability theory, keeping in mind the context of your question.
* Avoid misinterpretation of statistical results and draw meaningful conclusions.

**Part B: Machine learning models – 75 marks**

For part B of the assessment you are required to analyse the ‘**Apple\_Quality.csv**’ dataset using Linear Regression and Logistic Regression models. The dataset includes features of apples.

**Requirements**

**Linear Regression Analysis: (25 marks)**

* Load and inspect the dataset
* Explain the concept of linear regression
* Split the data into training and testing sets
* Implement linear regression model on appropriate column pairs
* Find the column pair with the highest r\_value and create a scatter plot with a line of regression
* Interpret the findings

**Logistic Regression Analysis: (25 marks)**

* Explain the concept of logistic regression
* Split the data into training and testing sets
* Implement logistic regression model with ‘Quality’ as the target column, highlighting the highest accuracy score
* Create and explain an appropriate confusion matrix
* Interpret the findings

**Comparison, Analysis & Conclusions: (25 marks)**

* Summarise the findings from the linear and logistic regression
* Explain relevant machine learning concepts
* Compare and contrast the results of linear regression and logistic regression analyses

**Learning Outcomes**

The following learning outcomes are addressed in this assessment.

* Understand how probability theory and statistical methods are applied to data
* Details how the output of statistical models are interpreted to form an insight into data
* Understand the basics of Machine Learning including Classification and Regression
* Implement Data Science methods in an appropriate software programming language