# MSC DA CA1

|  |  |
| --- | --- |
| Created By: | Stephen Kelly |
| Created On: | 24/10/2023 |
| Updated By: | Stephen Kelly |
| Last Update: | 06/11/2023 |

Contents

[MSC DA CA1 1](#_Toc150203827)

[Assessment Task 2](#_Toc150203828)

[Statistics: (Graded out of 100 & 1200 words) 3](#_Toc150203829)

[Data preparation and Visualization : (Graded out of 100 & 1200 words) 6](#_Toc150203830)

[Machine learning for Data Analytics:(Graded out of 100 & 1200 words) 7](#_Toc150203831)

[Programming**:** (Graded out of 100 & 600 words) 8](#_Toc150203832)

# Assessment Task

Students are advised to review and adhere to the submission requirements documented after the assessment task.

**Scenario:** Population in Ireland

A large amount of data has been collected by The Central Statistics Office in Ireland in relation to the population of Ireland, This data is available at:

<https://data.cso.ie/product/pme>

Chosen File: Population Estimates (Persons in April)

<https://data.cso.ie/>

# array storing values from the Value column

You are required to choose a particular area of interest and formulate the appropriate questions for modelling and analysis. For Example (but not limited to):

* Annual Population Change
* Immigration and Migration
* Population Forecasting
* etc…

You are required to collect, process, analyse and interpret the data in order to identify possible issues/problems at present and make predictions/classifications in regard to the future. This analysis will rely on the available data from CSO and **any additional data you deem necessary** (with supporting evidence) to support your hypothesis for this scenario.

This will require you to employ critical analysis of not only the domain of choice but also for the regression and or classification that you undertake.

**Note: This is an academic exercise and not a hypothetical report to the CSO.**

## Statistics: (Graded out of 100 & 1200 words)

1. Summarise your dataset clearly, using relevant descriptive statistics and appropriate plots. These should be carefully motivated and justified, and clearly presented. You should critically analyse your findings, in addition to including the necessary Python code, output and plots in the report. You are required to plot at least three graphs. [0-35]
2. Use two discrete distributions (Binomial and/or Poisson) in order to explain/identify some information about your dataset. You must explain your reasoning and the techniques you have used. Visualise your data and explain what happens with the large samples in these cases. You must work with Python and your mathematical reasoning must be documented in your report. [0-30]

<https://www.investopedia.com/terms/d/discrete-distribution.asp#:~:text=Common%20examples%20of%20discrete%20distribution,forecasting%20market%20shocks%20or%20recessions>.

A discrete probability distribution counts occurrences that have countable or finite outcomes.

Discrete distributions contrast with continuous distributions, where outcomes can fall anywhere on a continuum.

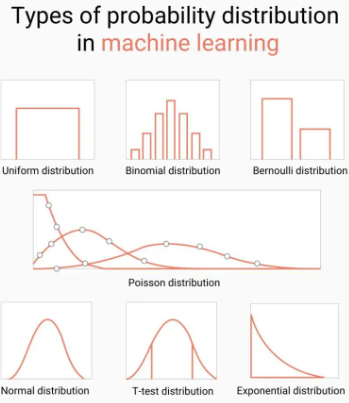
Common examples of discrete distribution include the binomial, Poisson, and Bernoulli distributions.

These distributions often involve statistical analyses of "counts" or "how many times" an event occurs.

<https://www.vedantu.com/question-answer/difference-between-binomial-distribution-class-12-maths-cbse-608a9d6f5ead5f7cec24f577>

|  |  |
| --- | --- |
| Binomial Distribution | Poisson Distribution |
| It is biparametric, i.e. it has 2 parameters n and p. | It is uniparametric, i.e. it has only 1 parameter m. |
| The number of attempts are fixed. | The number of attempts are unlimited. |
| The probability of success is constant. | The probability of success is extremely small. |
| There are only two possible outcomes-Success or failure. | There are unlimited possible outcomes. |
| Mean>Variance | Mean=Variance |

<https://datasciencedojo.com/blog/statistical-distributions/>



1. Use Normal distribution to explain or identify some information about your dataset. [0-20]
2. Explain the importance of the distributions used in point 3 and 4 in your analysis. Justify the choice of the variables and explain if the variables used for the discrete distributions could be used as normal distribution in this case. [0-15]

## Data preparation and Visualization : (Graded out of 100 & 1200 words)

1. You must perform appropriate **Exploratory Data Analysis (EDA)** on your dataset, rationalizing and detailing why you chose the specific methods and what insight you gained. [0-20]
2. You must also rationalise justify and detail all the methods used to prepare the data for ML. [0-30]
3. Appropriate visualizations must be used to engender insight into the dataset and to illustrate your final insights gained in your analysis. [0-20]
4. All design and implementation of your visualizations must be justified and detailed in full., making reference to Tufts Principles [0-30]

From The Visual Display of Q Info.pdf – Lecture4 / data\_visualization folder

Six principles of graphical integrity

1. The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities measured. (page 56)
2. Clear, detailed, and thorough labelling should be used to defeat graphical distortion and ambiguity. Write out explanations of the data on the graphic itself. Label important events in the data. (page 56)
3. Show data variation, not design variation (page 61)
4. In time-series displays of money, deflated and standardized units of monetary measurement are nearly always better than nominal units. (page 68)
5. The number of information-carrying (variable) dimensions depicted should not exceed the number of dimensions in the data. (page 71)
   1. In general, it implies that you should not use 3d to represent something that only needs to be 2d, etc.
6. Graphics must not quote data out of context (page 74)

To summarize Tufte’s principles of how to achieve graphical excellence (page 105)

1. Above all else show the data.
2. Maximize the data-ink ratio.
3. Erase non-data ink.
4. Erase redundant data ink.
5. Revise and edit.

## Machine learning for Data Analytics:(Graded out of 100 & 1200 words)

1. Explain which project management framework (CRISP-DM, KDD or SEMMA) is required for a data science project. Discuss and justify with real-life scenarios.

<https://medium.com/international-school-of-ai-data-science/data-science-project-frameworks-for-managers-c577d58b03db#:~:text=The%20CRoss%20Industry%20Standard%20Process,(or%20machine%20learning)%20project>.

<https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb>

Provide an explanation of why you chose a supervised, unsupervised, or semi-supervised machine learning technique for the dataset you used for ML modelling. **[0 - 20]**

<https://www.ibm.com/blog/supervised-vs-unsupervised-learning/>

The main distinction between the two approaches is the use of labelled datasets.

<https://www.datacamp.com/blog/what-is-labeled-data>

1. Machine learning models have a wide range of uses, including prediction, classification, and clustering. It is advised that you assess several approaches (at least two), choose appropriate hyperparameters for the optimal outcomes of Machine Learning models using an approach of hyperparameter tunning, such as GridSearchCV or RandomizedSearchCV. **[0 - 30]**

**Saturday class SVM case study**

1. Show the results of two or more ML modelling comparisons in a table or graph format. Review and critically examine the machine learning models' performance based on the selected metric for supervised, unsupervised, and semi-supervised approaches. **[0 - 30]**
2. Demonstrate the similarities and differences between your Machine Learning modelling results using the tables or visualizations. Provide a report along with an explanation and interpretation of the relevance and effectiveness of your findings. **[0 - 20]**

Discussed 31/10/2023

After Neil (CCT Founder) left the call

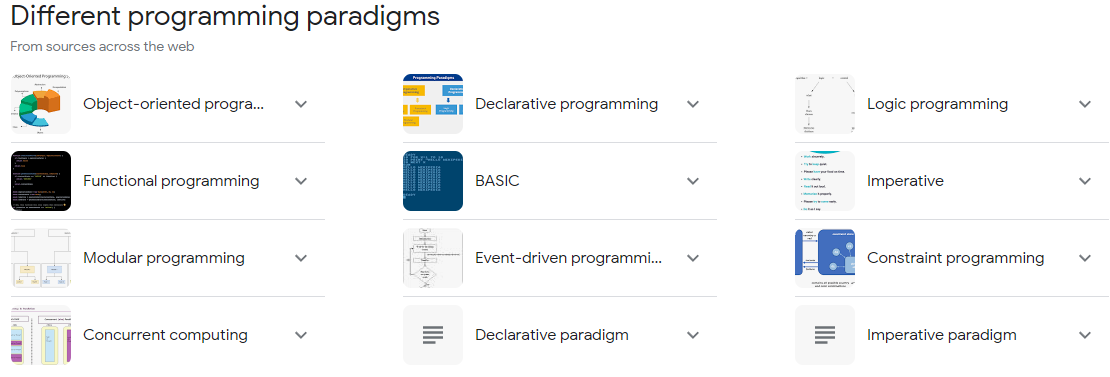
## Programming**:** (Graded out of 100 & 600 words)

1. The project must be explored programmatically, this means that you must implement suitable Python tools (code and/or libraries) to complete the analysis required. All of this is to be implemented in a Jupyter Notebook. Your codebook should be properly annotated. The project documentation must include sound justifications and explanation of your code choices (code quality standards should be applied). **[0-50]**

**Please recall that simply performing the analyses is a requirement to achieve a grade of PASS. Critical analysis and independent research are required for higher marks.**

Finish .ipynb files.

1. Briefly discuss your use of aspects of various programming paradigms in the development of your project. For example, this may include (but is not limited to) how they influenced your design decisions or how they helped you solve problems. **Note: marks may not be awarded if the discussion does not involve your specific project.** **[0-50]**



<https://blog.felgo.com/5-popular-programming-languages-examples>

Object-oriented programming is probably one of the most popular programming paradigms. It is based on organizing application elements around the data by defining the application as a set of objects. Each of them contains parts of necessary data as member variables and methods.

There are four main principles of object-oriented programming:

* Encapsulation – Access to objects variables and methods should be restricted in a way that only necessary data is available for other objects to use. This ensures that data is secure, and the chance of data corruption is reduced.
* Abstraction – Objects are hiding any unnecessary implementation code. The example of following this principle is creating getters and setters.
* Inheritance – Sub-classes containing logic of parent class can be created. It greatly improves code reusability.
* Polymorphism – Objects can be interpreted differently according to the context.

Object-oriented programming is commonly used when creating complex applications.