**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming for Data Analysis, Machine Learning for Data Analysis, Statistics for Data Analytics, Data Preparation & Visualisation |
| **Assessment Title:** | CA2 50% Integrated Assessment |
| **Lecturer Name:** | *Marina Iantorno*  *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Patrick O’ Connor |
| **Student Number:** | sbs23060 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Github Repository:** repo:sbs23060/POC-CA2

**Word Count: -**

**Project Management Framework**

The CRISP-DM project management framework was chosen for implementing this project due to its focus on the business question at hand, its flexibility and the level of supporting material available. These three factors are considered central considerations when implementing any data science project compared with alternative approaches such as SEMMA or KDD (Dåderman & Rosander 2018).

**Raw Data Gathering**

The datasets selected for this research project are chosen with the below requirements in mind:

|  |  |
| --- | --- |
| **Specific Requirements** | **Module** |
| Gather and process data that has been stored in at least two distinct formats | Programming |
| Document and evaluate a testing and optimisation strategy for your analysis     Plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). | Programming |
| Inferential stats for possible population values | Statistics |
| 5 different inferential statistics tests Ireland + 1 other country to compare  Parametric and non-parametric testing | Statistics |
| At least 2 models to compare and contrast | ML (Machine learning) |
| At least 2 approaches (prediction, classification etc.) with GridSearchCV | ML |
| Collect and develop a dataset based on the construction topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate construction topic (e.g., house price, availability of labour etc…) for producers and consumers point of view in Ireland. | ML |

The above requirements dictate the kind of data and data sources that need to be considered when choosing a dataset for this study. One example of this includes the requirement to gather and process data that has been stored in at least two distinct formats. To fulfil this requirement, data needs to be chosen from two separate formats and collated together in one format in order to be analysed, or a dataset can be split into two with one of the split halves transformed to an alternative format. The timeline of Github commits demonstrates the project planning associated with this research project.

**Bibliography**

Dåderman, A. and Rosander, S. (2018) *Evaluating Frameworks for Implementing Machine Learning in Signal Processing : A Comparative Study of CRISP-DM, SEMMA and KDD*. [Online]. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-235408 (Accessed: 7 April 2023).