­CCT College Dublin

|  |  |
| --- | --- |
| Modules Titles: | Data Preparation & Visualisation, Machine Learning for Data Analytics, Programming for Data Analytics, Statistics for Data Analytics |
| Assessment Title: | Ireland Agri-Business |
| Lecturers: | David McQuaid, Dr. Muhammad Iqbal, Sam Weiss, Taufique Ahmed |
| Student: | Ricardo Alves de Souza |
| Student Number: | 2018499 |
| Assessment Due Date: | 18/04/2024 |
| Date of Submission: | 26/05/2024 |

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. |

Title: Ireland Agri-Business

Author: Ricardo Alves de Souza

E-mail: [2018499@student.cct.ie](mailto:2018499@student.cct.ie?subject=Integrated%20CA1)

GitHub Address: https://github.com/ricardoasouz/2024---MSc-in-Data-Analytics---Feb---FT/tree/36c758bf7493b93b5199709afc8300c4e80051d3/%20MSC\_DA\_CA2

Table of Contents

[Scenario 4](#_Toc166598483)

[PROGRAMMING AND DATA SOURCES 5](#_Toc166598484)

[Sentiment Analyses 6](#_Toc166598485)

[- Scraping by “Beautiful Soup” 6](#_Toc166598486)

[- Scraping by “API” 6](#_Toc166598487)

[References 12](#_Toc166598488)

­Table of Figures

[Figure 1: Sentiment Analysis, Polarity and Subjectivity Distributions. 8](#_Toc166751733)

[Figure 2: Heatmap of OBS\_VALUE Over TIME\_PERIOD for Each Country 11](#_Toc166751734)

­­

# 

# Scenario

*“Today, big data is ubiquitous, machine learning applications are thriving, artificial intelligence appears in everyday conversations, and the internet of things is present even in household appliances. Businesses and organizations are increasingly managed through cloud computing and high-performance computing is progressively accessible as a service…More effective operations, reduced uncertainties, and real-time decision support could revolutionize agriculture to a great extent. Food could be produced more efficiently, of higher nutritional quality, in more stable supplies, with less environmental damage, and likely with additional economic, social, and ecological benefits.”*(Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022))

# METHODOLOGY

The methodology is “CRISP-DM”, a robust and well-established framework that outlines a comprehensive process of understanding the business context, data understanding, data preparation, modelling, evaluation, and deployment. It's widely adopted due to its industry-agnostic and flexible nature, facilitating the study and development.

# DATA SOURCES

This study has used the computer programming language “Python” in the format of “Jupyter Notebook” with dedicated libraries for different tasks. It has acquired data to analyse from different sources, including tree files provided together with this assessment:

* Agriculture\_\_crops\_\_livestock\_and\_land\_use\_by\_general\_farm\_type\_\_region\_18042024\_071227.csv
* EuropeanAgriculture\_FarmStructureIndicators\_Eurostat2016.csv
* Irish-agri-food-exports-208-2022\_21032023.csv

https://ec.europa.eu/eurostat/databrowser/view/tag00044/default/table?lang=en&category=t\_agr.t\_apro.t\_apro\_mt

tag00044\_linear.csv

The segment chosen for analysis was:

“Ireland Agri-Business on CAP”

## Scraping by “Beautiful Soup”

A discussion content was obtained from, “Forum4Farming - Britain & Ireland Farming Forum, Agricultural Discussion Boards. IE & UK”. On the following address:

(“https://www.forum4farming.com/forum/index.php?threads/cap-2023-2027.20587/”)

And saved in the format “JSON”, an acronym for, “JavaScript Object Notation” on the file:

(“extracted\_forum4farming.json”)

This format is chosen to make it easy to read by humans and upload to a NoSQL Database without any more conversion.

The library utilised to obtain this data part is popular among Data Analytics by your use facility and good performance and is called “BeatfulSoup”.

The code for this extraction can be viewed on file, “StractingForumsBSandAPI.ipynb” and the libraries “request”, “BeautifulSoup”, “re” and “JSON” is utilised to connect to the web address and manage the requisition to realise the download from six pages of the contents mentioned before.

## Scraping by “API”

Another discussion content was obtained from Reddit. On the following address:

(“https://www.reddit.com/r/ireland/comments/1aghrfn/irish\_farmers\_protest\_in\_solidarity\_with\_eu/”)

And also saved in the format “JSON”, on the file:

(“extracted\_redditfarming.json”)

The code for this extraction is also on file, “StractingForumsBSandAPI.ipynb” and the environment variables are hidden on the, “.env” file to not exposure the credentials utilised to access the Reddit API also the, “.env” file is cited inside the file, “.gitignore” making this environment credentials invisible on Open Source platforms like, “GitHub”.

“PRAW”, is the library to access the Reddit API, she can hold the credentials to open access for obtain content submitted from address specified on variable “URL”.

After that, both files have been merged into only one, type, “JSON” and also in format, “CSV” to make it easy to analyse from Machine Learning Models.

(“extracted\_combined.json”)

(“extracted\_combined.csv”).

# SENTIMENT ANALYSIS

With the advance of techniques in machine learning and improvement on code’s dedicatees to analyse what feelings bring each specific word typed by users on the internet, now are possible to detect and or find out what the users want manifest, if this is positive or negative on the question about.

For this case let’s make use of a library called “TextBlob”, with her is possible to process the file previously scrapped and prepared in a formatted dataset, and obtain two variables: polarity and subjectivity.

Polarity measures the positivity or negativity of the text with a range between -1 to 1.

Subjectivity measures the objectivity of the text with a range between 0 to 1.

author statement

0 Franz\_Werfel the famers protests on the continent were abou...

1 lamahorses Over 30% of the European budget is on CAP etc

2 ConnolysMoustache What are they even protesting? The IFA repr...

3 bintags These people are brainwashed.

4 goodforyourself21 Stop bitching at each other and get out and su...

textblob\_sentiment polarity subjectivity

0 (0.0, 0.5) 0.00000 0.500000

1 (0.0, 0.0) 0.00000 0.000000

2 (0.04375, 0.685417) 0.04375 0.685417

3 (0.0, 0.0) 0.00000 0.000000

4 (-0.125, 0.375) -0.12500 0.375000

Now with all phrases analysed by the “TextBlob” library and converted into measured numeric values, is possible to perform descriptive statistics and also plot a graphic to visualize the distribution of these results.

Polarity Subjectivity

count 129.000000 129.000000

mean 0.071990 0.409139

std 0.156366 0.207902

min -0.375000 0.000000

25% 0.000000 0.310606

50% 0.052000 0.419643

75% 0.141667 0.512352

Max 0.800000 1.000000

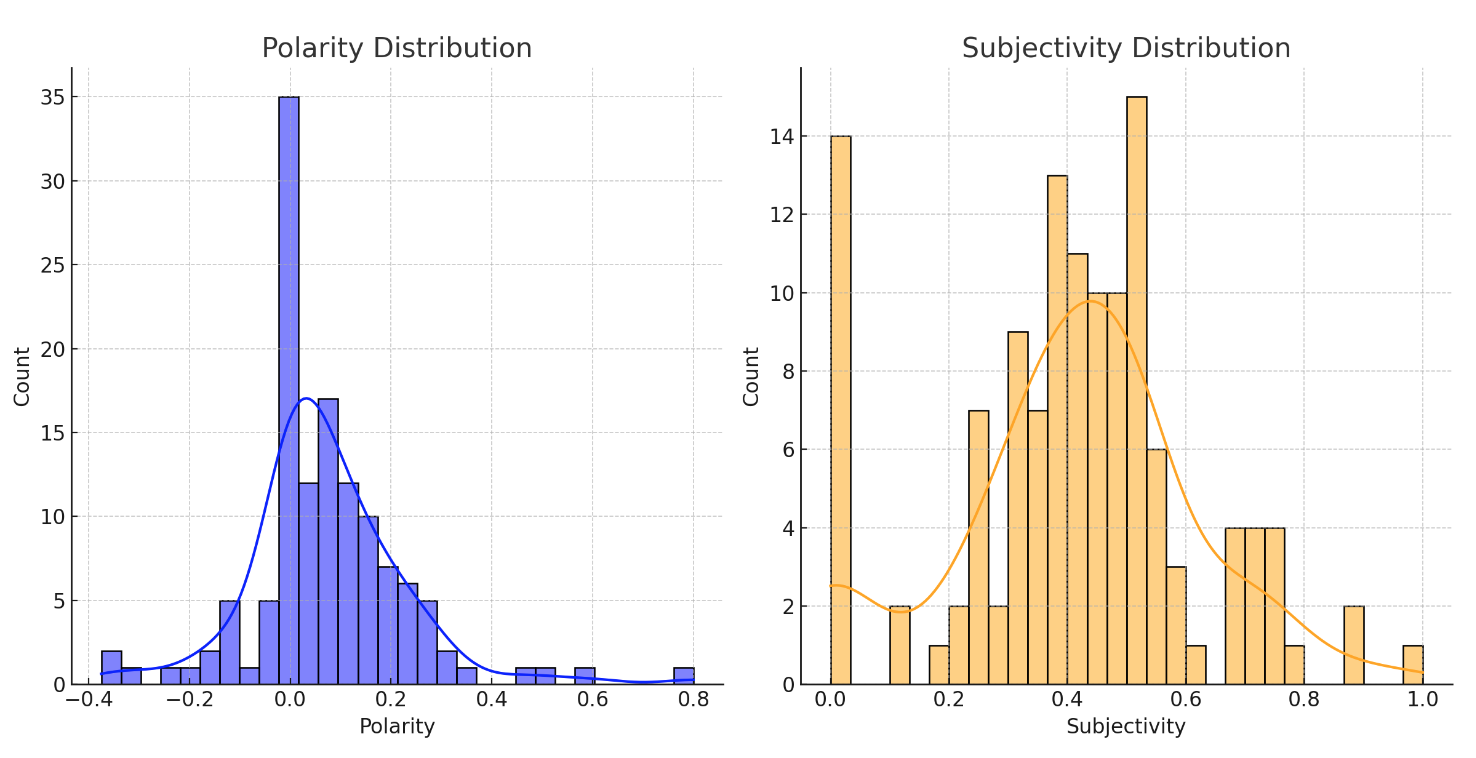


Figure 1: Sentiment Analysis, Polarity and Subjectivity Distributions.

Now with statistics measured and graphically exposed it is easy to see the major part of people manifesting positivity to the subject analysed in this case the influence of CAP upside the farmers in Ireland.

# Top 5 of European Union on PRODUCTION OF MEAT: Cattle

## Obtaining the dataset

For this case study, let us analyse meat production: Cattle between the top five countries in the European Union, from 2012 to 2023. For this, a dataset from the website “Eurostat” can provide us with accurate statistics numbers about countries and the subject cited.

The dataset was obtained from the following address in the format “SDMX-CSV 1.0”,

“https://ec.europa.eu/eurostat/databrowser/view/tag00044/default/table?lang=en&category=t\_agr.t\_apro.t\_apro\_mt”.

After making a selection of the top five countries producers in 2023, “France, Germany, Spain, Italy and Ireland”, to reduce the size of the file and variables, as downloaded and saved with the file name:

“tag00044\_page\_linear.csv”.

*The Dataset page description:*

*Production of meat: cattle*

This indicator covers the carcass weight of bovine animals (calves, bullocks, bulls, heifers and cows) slaughtered in slaughterhouses and on the farm, whose meat is declared fit for human consumption.

Online data code:

tag00044

Source of data:

Eurostat

Last data update:

02/04/2024 22:00 (1 month ago)

Last structure update:

02/04/2024 22:00

Overall data coverage:

2012 - 2023

Number of values:

428

Data navigation tree location

Agriculture, forestry and fisheries > Agriculture > Agricultural production > Livestock and meat

The source dataset for tag00044 is:

apro\_mt\_pann

https://ec.europa.eu/eurostat/databrowser/product/view/apro\_mt\_pann

tag00044

Time-frequency: Annual

Unit of measure: Thousand tonnes

Item of meat: Slaughtering

Meat product: Bovine meat

## Preparing the dataset for analyses.

Opening the file in a “Pandas Data frame”, and assigning it to a “Pandas data frame” with the variable name “df”, is possible to observe the presence of columns with unique values representing redundant information that can be deleted, and the structure of data can be rearranged and organised to eliminate presence of categorical data, “object” to then make possible and easy to perform statistics analyses and run machine learning models.

Following the original structure of the data frame:

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 60 entries, 0 to 59

Data columns (total 10 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 DATAFLOW 60 non-null object

1 LAST UPDATE 60 non-null object

2 freq 60 non-null object

3 unit 60 non-null object

4 meatitem 60 non-null object

5 meat 60 non-null object

6 geo 60 non-null object

7 TIME\_PERIOD 60 non-null int64

8 OBS\_VALUE 60 non-null float64

9 OBS\_FLAG 12 non-null object

dtypes: float64(1), int64(1), object(8)

memory usage: 4.8+ KB

To clean and organise this data frame, let’s delete the columns ( DATAFLOW, LAST UPDATE, freq, unit, meatitem, meat, OBS\_FLAG ), keeping only the columns, “geo, TIME\_PERIOD, OBS\_VALUE”.

The column “geo” represents the country where the value is observed and comes with only two letters, let’s convert these codes to a full country name with the help of the library, “pycountry”.

Now let’s pivot this table making the column, “TIME\_PERIOD” the index of the table and countries columns with your respective values observed.

The result is a well-structured Data Frame, with only numeric values “float64”, with only twelve rows and five columns ( the countries ), indexed by year, occupying only 576 bytes on memory, how is this possible see performing the “info()” method upside the “df”, data frame:

df.info()

<class 'pandas.core.frame.DataFrame'>

Index: 12 entries, 2012 to 2023

Data columns (total 5 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 France 12 non-null float64

1 Germany 12 non-null float64

2 Ireland 12 non-null float64

3 Italy 12 non-null float64

4 Spain 12 non-null float64

dtypes: float64(5)

memory usage: 576.0 bytes

Let’s plot a heatmap to better visualise this data:

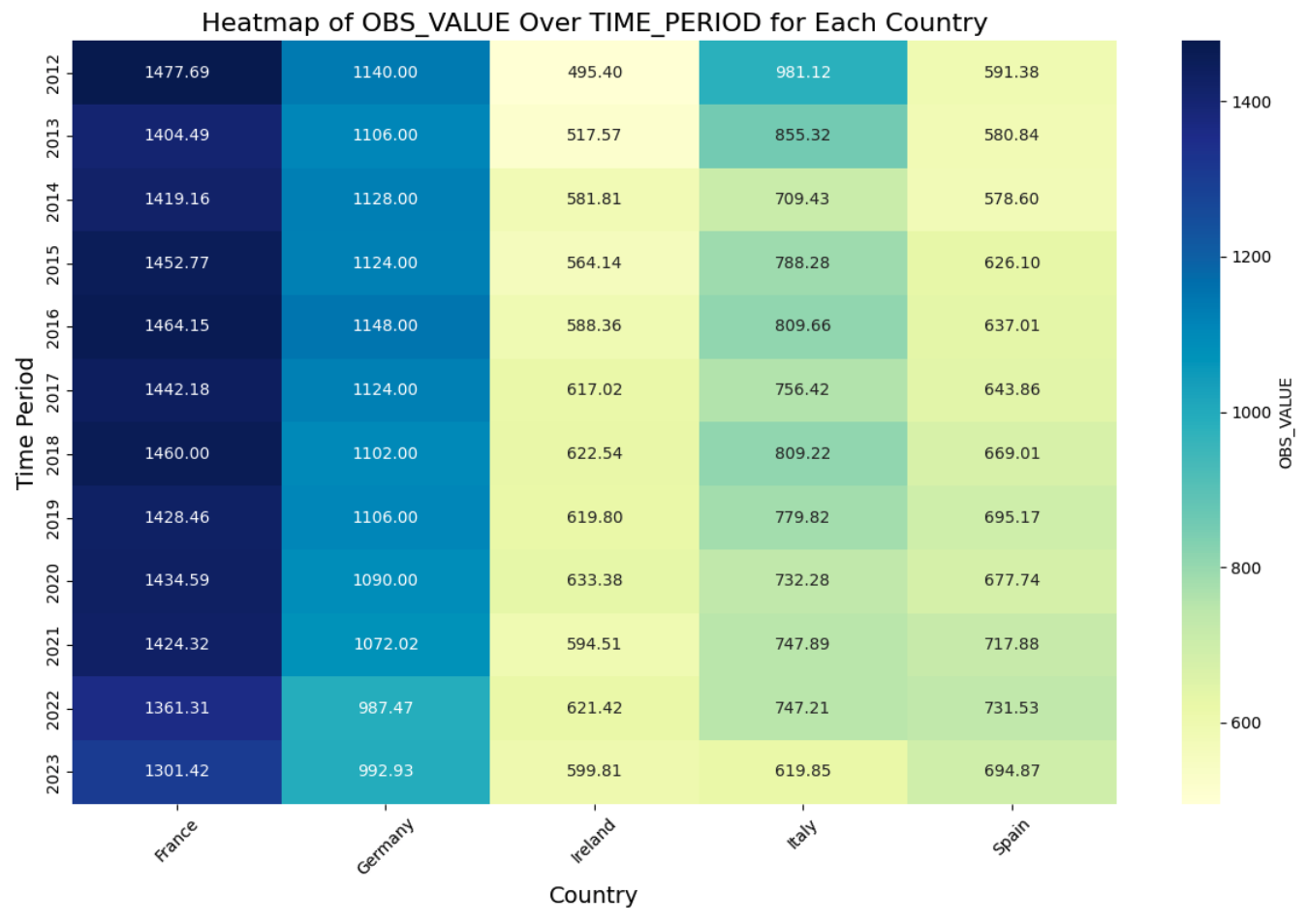


Figure 2: Heatmap of OBS\_VALUE Over TIME\_PERIOD for Each Country

You have been tasked with analysing

Ireland's Agricultural data

and comparing the Irish Agri sector with other countries worldwide.

This analysis should also include

forecasting,

sentiment analysis and

evidence-based recommendations for the sector

as well as a complete rationale of the entire process used to discover your findings.

Your Research could include

export,

import,

trade imbalance,

arable production,

animal stock,

medicinal input,

organic,

gm products etc.

(or any other relevant topic EXCEPT Climate change) with Ireland as your base line.

**Note:**

* **While topical, Agricultural impact on Climate Change SHOULD NOT be chosen as an area of research for this assessment.**
* **Members of the European Union implement the Common Agricultural Policy, and this should be researched as it has a significant statistical impact.**
* **The United Kingdom is NOT part of the European Union**

You must source appropriate data sets from any available repository to inform your research (all datasets MUST be referenced, and the relevant licence/permissions detailed).

**Several Data Sets have been Supplied which you may use as you wish (You do not HAVE to use them)**

**LICENSES for supplied datasets**

*“FAO encourages you to use FAO databases for research, statistical, and scientific purposes. You may access, download, create copies and re-disseminate datasets subject to these Dataset terms.*

*Unless specifically stated otherwise, all datasets disseminated through the databases below are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 3.0 IGO (CC BY-NC-SA 3.0 IGO) “*

***Criteria of Analysis***

Discuss the choice of project management framework you have deemed suitable for this project.

**Programming for DA Tasks**

**YOU MUST ATTEMPT ALL 6 PARTS**

1. **Programming**: 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. The project documentation must include sound justifications and explanations of your code choices. Code quality standards should also be applied. [30 marks]
2. **Data From Diverse Sources**: In a dedicated section of your report, compare, contrast, and select relevant libraries/techniques to process data from diverse sources [10 marks]
3. **Data manipulation**: In the same section as Part 2, critically appraise aggregation methods (eg combining data) to process and manipulate data from multiple data structures [0-20]
4. **Data structures**: You are required to gather and process data that has been stored in at least two distinct formats. For example, this can be data in a CSV file, from a MySQL database or from a web API in JSON format. [20 marks]
5. **Testing**: In a dedicated section of your report, you are required to document and evaluate a “testing” strategy for your analysis. As part of this, you may want to plan and document how you ensured your code is doing what it is meant to. Note any trade-offs that you've made in these areas. [10 marks]
6. **Optimisation**: In a dedicated section of your report, you are required to document and evaluate an optimisation strategy for your analysis. As part of this, you may want to plan and document how you ensured that the code is making good use of your system’s resources (eg CPU, RAM, time etc). Note any trade-offs that you've made in these areas. [10 marks]

**Total Mark = 30+10+20+20+10+10=100:(100\*0.5=50%)**

**Statistics for Data Analytics Tasks**

* Use descriptive statistics and appropriate visualisations in order to summarise the dataset(s) used, and to help justify the chosen models. **[0-20]**
* Analyse the variables in your dataset(s) and use appropriate inferential statistics to gain insights on possible population values (e.g., if you were working with international commerce, you could find a confidence interval for the population proportion of yearly dairy exports out of all agricultural exports). **[0-20]**
* Undertake research to find similarities between some country(s) against Ireland and apply parametric and non-parametric inferential statistical techniques to compare them (e.g., t-test, analysis of variance, Wilcoxon test, chi-squared test, among others). You must justify your choices and verify the applicability of the tests. Hypotheses and conclusions must be clearly stated. You are expected to use at least 5 different inferential statistics tests. **[0-40]**
* Use the outcome of your analysis to deepen your research. Indicate the challenges you faced in the process. **[0-20]**

*Note: All your calculations and reasoning behind your models must be documented in the report and/or the appendix.*

**Total Mark = 20+20+40+20=100:(100\*0.5=50%)**

**Machine Learning Tasks**

Use of multiple models (at least two) to compare and contrast results and insights gained.

* Describe the rationale and justification for the choice of machine learning models for the above-mentioned scenario. Machine Learning models can be used for Prediction, Classification, Clustering, sentiment analysis, recommendation systems and Time series analysis. You should plan on trying multiple approaches (at least two) with proper selection of hyperparameters using GridSearchCV method. You can choose appropriate features from the datasets and a target feature to answer the question asked in the scenario in the case of supervised learning.

**[0 - 30]**

* Collect and develop a dataset based on the agriculture topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate agricultural topic (e.g., product price, feed quality etc…) for producers and consumers point of view in Ireland.

**[0 - 25]**

* You should train and test for Supervised Learning and other appropriate metrics for unsupervised/ semi-supervised machine learning models that you have chosen. Use cross validation to provide authenticity of the modelling outcomes. You can apply dimensionality reduction methods to prepare the dataset based on your machine learning modelling requirements.

**[0 - 30]**

* A Table or graphics should be provided to illustrate the similarities and contrast of the Machine Learning modelling outcomes based on the scoring metric used for the analysis of the above-mentioned scenario. Discuss and elaborate your understanding clearly.

**[0 - 15]**

**Total Mark = 30+25+30+15=100:(100\*0.5=50%)**

**Data Preparation & Visualisation Tasks**

* Discuss in detail the process of acquiring your raw data, detailing the positive and/or negative aspects of your research and acquisition. This should include the relevance and implications of any and all licensing/permissions associated with the data (This will require research outside of class material). **[0-15]**
* Exploratory Data Analysis helps to identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so problems can be addressed. Evaluate your raw data and detail, in depth, the various attributes and issues that you find. Your evaluation should reference evidence to support your chosen methodology and use visualizations to illustrate your findings.**[0-25]**
* Taking into consideration the tasks required in the machine learning section, use appropriate data cleaning, engineering, extraction and/or other techniques to structure and enrich your data. Rationalize your decisions and implementation, including evidence of how your process has addressed the problems identified in the EDA (Exploratory Data Analysis) stage and how your structured data will assist in the analysis stage. This should include visualizations to illustrate your work and evidence to support your methodology.**[0-30**]
* Modern farming has a great dependence on technology and relies upon visualizations to communicate information, this includes web based, mobile based and many other digital transmission formats. Develop an interactive dashboard tailored to modern farmers, using tufts principles, to showcase the information/evidence gathered following your Machine Learning Analysis. Detail the rationale for approach and visualisation choices made during development making reference to Tufts Principles. **Note you may not use Powerbi, RapidMiner, tableau or other such tools to accomplish this (at this stage).[0-30]**

**Total Mark = 15+25+30+30=100:(100\*0.5=50%)**

**Additional notes :**

All:

* Your documentation should present your approach to the project, including elements of project planning ( timelines).
* Ensure that your documentation follows a logical sequence through the planning / research / justification / implementation phases of the project.
* Ensure that your final upload contains a **maximum of 1 Jupyter notebook per module.**
* Please ensure that additional resources are placed and linked to a logical file structure eg, Scripts, Images, Report, Data etc…
* Ensure that you include your raw and structured datasets in your submission
* 3000(+/- 10%) words in report (not including code, code comments, titles, references or citations)
* Your Word count MUST be included.

**(it is expected that research be carried out beyond class material)**

**Submission Requirements** All assessment submissions must meet the minimum requirements listed below. Failure to do so may have implications for the mark awarded.

All assessment submissions must:

* Jupyter Notebook, Word Document, Dashboard
* Be submitted by the deadline date specified or be subject to late submission penalties
* Be submitted via Moodle upload
* Use [Harvard Referencing](http://40.115.124.2/sp/subjects/guide.php?subject=harvardref) when citing third party material
* Be the student’s own work.
* Include the CCT assessment cover page.

**Additional Information**

* Lecturers are not required to review draft assessment submissions. This may be offered at the lecturer’s discretion.
* In accordance with CCT policy, feedback to learners may be provided in written, audio or video format and can be provided as individual learner feedback, small group feedback or whole class feedback.
* Results and feedback will only be issued when assessments have been marked and moderated / reviewed by a second examiner.
* Additional feedback may be requested by *contacting Your Lecturer,* Additional feedback may be provided as individual, small group or whole class feedback. Lecturers are not obliged to respond to email requests for additional feedback where this is not the specified process or to respond to further requests for feedback following the additional feedback.
* Following receipt of feedback, where a student believes there has been an error in the marks or feedback received, they should avail of the recheck and review process and should not attempt to get a revised mark / feedback by directly approaching the lecturer. Lecturers are not authorised to amend published marks outside of the recheck and review process or the Board of Examiners process.
* Students are advised that disagreement with an academic judgement is not grounds for review.
* For additional support with academic writing and referencing students are advised to contact the CCT Library Service or access the [CCT Learning Space](http://learningspace.cct.ie/subjects/index.php).
* For additional support with subject matter content students are advised to contact the [CCT Student Mentoring Academy](https://moodle.cct.ie/mod/forum/view.php?id=55148)
* For additional support with IT subject content, students are advised to access the [CCT Support Hub](https://moodle.cct.ie/course/view.php?id=1861).

# References

Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022): “Big data in agriculture: Between opportunity and solution”

<https://doi.org/10.1016/j.agsy.2021.103298>,(https://www.sciencedirect.com/science/article/pii/S0308521X21002511)