**CCT College Dublin**

MSc in Data Analytics

Integrated CA2 – Individual

Big Data and Agribusiness

**Lecturer Names:**  Marina Iantorno/Vladimir Milosavljevic

Muhammad Iqbal & David McQuaid

**Student:** Claudia Oliveira da Silva

**Student Number:** sbs22026

**Assessment Due Date:** 22nd May2022

**Date of Submission:** 22nd May2022

**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.

1. **Abstract**

In general technological advancement alludes to the disclosure of better methods of producing goods. Changes in innovation lead to an expansion in the efficiency of work, capital, and different variables of production. The use of software algorithms created an avalanche of modernization in the Agriculture business. As the global population increase is decisive to leverage productivity, the profitability of businesses, food production, and the well-being of rural and urban society.

All data were collected on government and open websites, which we understand as a reliable sources. as Agricultural Production is a very specific topic and not linked to being part of our daily routine conversation, this research was not used web scrapping, instead was made a programmatic request for Sentimental analysis.

Performance was compared using evaluation machine learning linear regression prediction. Upon successful completion of experiments, results show that regression-based machine learning models generally showed better results for modeling with lengthier historical data (more than three years). It was used different Datasets for training and test, we found good performance in almost all models (RandomForest Regressor, KNeighbors Regressor, AdaBoost Regressor, Linear Regression, MinMaxScale).

1. **Introduction**

Since the Farmers began to cultivate the land with the support of technology in agriculture, farming has never been the same. Instead, the open data available shows that, from time to time, technological innovations in agriculture provide a paradigm shift in the agricultural production model.

For us the majority of people living in the big cities, it may seem obvious but the wireless and high-speed connection has allowed the introduction of numerous new technologies in agriculture. Now farmers can monitor the farm in real-time, follow machines during work, manage staff, and access data, KPIs, and plantation information on their smartphone screens. All business details can be seen through the cloud, making it much easier to manage the production. The use of technology in the management agriculture industry to change the production system vision, reduce costs, increase productivity and reduce environmental damage.

So, this paper analysed the Price of Land, Type, Ages of Cows, and Price of Fertiliser. The main challenge was to show the possibilities within Data Analysis in Machine Learning and that the most important thing is not the amount of data, but what is done with it.

*Sub RQ: “What Machine Learning approaches are best suited to modelling different types and lengths of data (in terms of model accuracy and processing time), and how to analysis the correlation between apparently different topics.”*

1. **Literature Review**

Many believe that Big Data is a large volume of information that circulates in the digital environment. Big Data, however, also exists in the data analysis part, in extracting relevant facts and applying the data transforming it in information.

Currently, the Big Data is stored in the cloud, the data can be consulted from anywhere, from the farm's seating room or in the middle of the field, and transferred to wherever the business is. In short, Big Data serves the purposes of traceability, prediction, and management of production software systems. With the ability to handle any type of digital record, some subjects are indispensable, and Big Data can record helping Farmers and Agriculture producers, such as geolocation videos, soil characteristics, and the history of water consumption and crop rotation.

Based on 5 principles (of speed, volume, veracity, variety, and value), Big Data uses diverse information collected in real-time to generate expertise. Applied to Machine Learning, data analysis provides information to farmers and machines ensuring the optimization of fertilizers, seeds, and others. It is also possible to identify threats and vulnerabilities and planning preventive actions that avoid losses.

|  |  |
| --- | --- |
| In our process of Data Preparation and Visualization was tried different types of charts and graphics to locate the most common type (compound) of Fertiliser and from there to test the relation between the Price of Land, Productivity, and Fertiliser. In the graphics, you can see there are many input attributes, which it makes difficult to visualize the data. It creates a concern that this dataset with big numbers of attributes will negatively affect the accuracy and training time of the machine learning model.  After research the Principal Component Analysis (PCA) shows to be one way to solve this problem and is used for better data visualization.  *Figure 1* |  |

Here we stipulated two main types of Fertiliser in Ireland and France to be used in the Machine Learning Models and the Statistics and Data Visualization process together with other datasets regarding Cattle Breed, Cattle age, Milk production, and Land prices.

Also, in the figure above we can see the correlation between Fertilisers (compounds) that shows the majority of the producers in the same period using the same Fertilisers. It is not for a feeling or tradition, it is Big Data has been used effectively in improving soil management and consequently increasing agricultural production.

We can see the correlation between the use of land that uses more Fertilisers has the last chance to be used as Dairy Farming, which means the last chance to be used in the production of milk.

This research process presented us with the challenge of working data in a very specific market, but at the same time with very detailed data without a common structure for data collection. In this scenario we used different datasets within different Jupyter Notebooks, using PCA technique, functions such as ‘drop’, ‘index’, ‘insa’, ‘arima’, ‘merge’,’ pivot’, more than we anticipated. But allowed us to work with different Machine Learning Models and confirmed that one Machine Learning Model is not always the best for everything.

1. **Methodology**

4.1 Data Collection (Secondary Data)

In this research, we used secondary dataset collection, it means the dataset have already been researched by other organizations. This type of research is also named desk research. The main advantages of secondary data collection are the agility and the broad scope of the research. In this stage of our process, the Agriculture Business presented itself as very broad industrial sector, with very diversity, demanding, and temporal data.

Despite the scenario chosen (Agriculture) had a large number of open databases available, we had great difficulty finding compatibles datasets between Ireland and other Countries, because each Country, organization (governmental or not) or website had a different approach for the same product or technology. There was no need of authorization to collect the data, and all data was collected from governmental organizations which led us to believe in the reliability and security of the source.

We used a programmatic data collection by requesting data in text format (string) for Sentimental analysis, through Tweet, as we were not successful in our web scrapping attempts. Nor do our collection attempts realize that Agriculture Industry is not a topic that the population had the habit of discussing on social media, what was found were data regarding the environment, and health eating, among others data with a poor relationship with our research.

4.2 Data Exploration and Programming

In the process of exploring the chosen datasets we identified the main tasks that were used in all data frames; 1. check for duplicates, and delete when there is two or more variables had the observations and there were two types of duplicates found(see Fig. 2). 1.1. different variables labels but with same observations ., 1.2. different variables labels with different observations names but with the same value for the research(*e.g. name fertilizer code and fertilizer name*(see Fig. 3))., 2. check importance of the variable for our research., 3. find missing values (NAN) and exclude them., 4. find and delete illegal character., 5 identify and check the variable types and change to the correct one, from string to float. (see Fig. 4)., 6. Change the format of the data frame, the value that was variables become observations and vice-versa(see Fig. 5). Another way to prepare the dataset is The Series.mask() Pandas function substitutes non-numeric values for the condition passed True. because we are only interested in numerical values (int).

|  |  |
| --- | --- |
| *Figure 2 duplicities and cleaning* | *Figure 3 duplicities* |
| *Figure 4 type conversions* | *Figure 5 dataframe standardization* |

From this point, was possible to manage and define a single standard for all data frames, and then developed the graphs (plot) for all data frames separately. In an intuitive way and using logical thinking, we analyze the internal correlations of each data frame, to understand the correlation existing in each data frame, and then visualize and better understand the Agricultural context of each product, and as a consequence, we achieve better correlation results between all the data frames to then develop the Machine Learning models. (see Fig. 6, 7 & 8, 9).

|  |  |  |
| --- | --- | --- |
| Figure 6 | | Figure 7 |
| Fig 8 | Fig 9 | | |

In one specific dataset was necessary to metric conversion from hectares to acres to standardize all of the different data frames in the same measurement system. Also, was necessary to merge four datasets and one data frame so that it was possible to add an agricultural product that is sold directly to the consumers. The product chosen was milk and used a seeking to meet one of the criteria of the assessment, the solution found was loop codes because it is possible to directly merge the datasets and create an index to organize data frames with the same layout as the other data frame.

At the end of every Notebook, was generate a new data frame within the standards and with the necessary information to be used in the processes of Machine Learning Models and the statistical analysis.

4.3 Dashboard

In the dashboard we use the Streamlit tool (details in the Appendix) to include our code in python because it needs less code to create the dashboard, there is no need for a callback because the widgets are treated as variables and allow us to deploy and host more than one project with easy to share with recruiters. The only downside is that Streamlit works in .py format, but we have checkbox options like 'multi select' where you can select multiple variables to do the correlation with only the selected variables.

1. **Machine Learning**

In this research, we stayed with the regression model, because it was the one that best suited the chosen dataset, as we obtain in collecting data from different products and subjects to create a challenge in finding correlations between them (Fertilizers, cattle, milk, and land) the Regression models where it is possible to predict (forecast) prices of final products such as milk, and also to investigate correlation is to use these correlations to predict prices of other commodities such as the price of land. Since machine learning models cannot deal with variables in a text (string) format, so the approach chosen was to use 'dummies' function on the variables. The objective was to specify which observations should be used and when to use them (fig. 10). It is a similar way to creating a Matrix to be used as a preparation for our Machine Learning Models process.

|  |  |
| --- | --- |
| Fig10 | Fig 11 |

Another obstacle encountered was a large number of variables with the same meaning. From their concern, we try to find a way to know exactly which one of those variables is the best one to be used in each dataset, that has the best interrelationships between them so that this way decrease the number of variables to make the analysis simpler as one. For this the solution found was the Principal Component Analysis (PCA) dimensionality reduction technique, with this technique we located and reduced the number of variables. that is, changing the dimensions of some data frame. The results were very satisfactory, managing to reduce the number of variables of a single dataset from 8 to 2 variables with a correlation ratio of 0.9776 in the PCA *(eg. Notebooks; CA2\_Fertiliser, CA2\_Cattle )*. This meant that with only 2 variables it was possible to explain/inform the same content of 8 variables of the same dataset.

Regarding Training and Testing Machine Learning Models, we opted for a test size the proportion of 70% for training and 30% for testing, we opted for a test size in the proportion of 70% for training and 30% for testing due to the number of variables in each dataset is not large, but the observation diversity is large(Fig 11). After training each machine learning model, we evaluated and all of the visualized models as metrics from each training model and compare them to the test results of the chosen model.

|  |  |
| --- | --- |
| Fig 12 | Working on the Notebook CA2\_land & CA2\_milk  Despite the thorough preparation of the data frames before the application of the machine learning models, the results were not satisfactory. The model with the best performance was the Ada Boost Regressor with 0.160620, which means a regular results model. But when we check the forecast versus the average calculated over the data frame, the result shrinks even further(fig 12). |

5.1 Times Series (Notebook CA2\_milk)

In the development of our time series, we chose the CA2\_milk Notebook because it is the dataset found with the best time variables (month and year). As with other datasets, it was necessary to transform variables from 'string' format to number (int). The objective was to create a data frame to be used as a time series, because is necessary numbers to work with Time series and then after transforming all observations to ‘int’ format, we sum two variables month and year(fig 13), and transformed them into date format.

|  |  |
| --- | --- |
| Fig 13 |  |

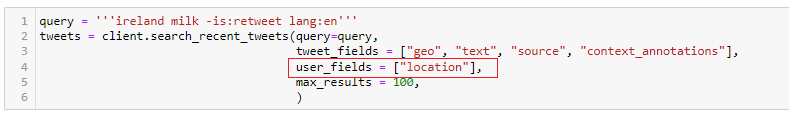
The next step on Time Serie was used a statistic model Adfuller to verified if the quantity of milk sold in the past year, has influence on the price of milk that will be sold in the future. For this analysis imported library Statistic ‘ Adfuller (statsmodels.tsa.stattools.adfuller). Another Time series forecast used ARIMA model, to build a linear regression model to validate the previous forecast result, and the Dashboard as well.

5.2 Sentiment Analysis

To begin our Sentiment Analysis code, imported the libraries: a.) import tweepy b.)import re c.)import string, d.) import punctuation, e.)import nltk, f.)import stopwords, g.)nltk.download('stopwords'), h.)import CountVectorizer, i.)import tensorflowand and f.)import TfidfTransformer.

As mentioned before, because the web scraping procedure did not give the expected return, we used the social media Twitter. We follow the documentation to collect the twitters according with their architecture, however the parameters did show the location requested ( Ireland / France), was not possible to know exactly the reason for this fault, or if this problem is from our access type or this part of Tweepy documentation is not working for all location(fig 14) .

We request only original and Irish twitter, but because we could not apply properly the location, we can not be sure that all twitter are from Ireland, but we can assume that all twitter in this research are regarding Irish Milk.

 fig 14

We collected all twitter from or about Irish Milk and USA Milk to create two data frame and so that in this way it was possible to continue the Machine Learning process. To get the best perform our Sentiment Analysis Model, was use the train scope used in class with the same parameters. But, we did some steps difference, we limited the words counting to 5000 features, because my device has enough power to run more features or unlimited. In the created a Neural Model ( Sequential) with 3 layers and 5000 neurons , function ‘relu’ and ‘droupt 0.5’ in each layers to avoid overfit.

1. **Statistic**

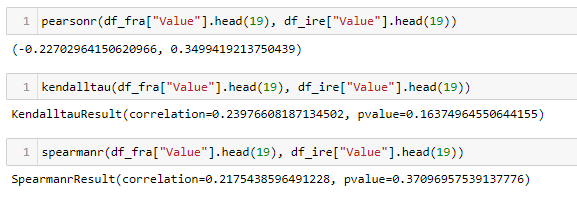
All Jupyter Notebooks have Analysis Statistics, the great difficulty to Statistic was to find a dataset for two countries with the same variables. A solution for this issue was to choose one dataset only for Statistic Analysis. The subject chosen was the Price of the land in Ireland and France.

We used two Notebooks to the first part of the Statistics for Data Analysis was in the Notebook: CA2\_milk\_Statistic, and the second Notebooks is exclusive to work with the similarities between Ireland and France, applying parametric and non-parametric techniques(Figure 15).

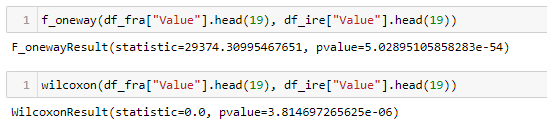
The analysis for the data frame demonstrated there is LINEAR correlation, however all test takes in consideration the size of the data frame (size of the sample). Always in mind that every statistics test come from hypotheses.

|  |  |
| --- | --- |
|  | Fig 15 |

The difference between the values of mean and deviation was considered large and some of the tests. The Statistic test carried out sought to verify whether there is a correlation between the price of land in France and the price of land in Ireland. The results were relatively low, around 0.2 (correlation value). Also, the variable Pvalue results demonstrate that despite assuming that the variables have a normal distribution there is no correlation between the Price of Land In Ireland and France. Remember that the statistical tests of Spermanr and Kendaltau (fig 16) do not need the hypothesis that works with a data frame with normal distribution. The results confirm an equal outcome of the Methodo Personr results. It is important to emphasize the difficulty of finding data that are not only a time series (with normal distribution) to run the Statistics Test.

 figure 16

Concerning the Anova test, which tests whether two groups have the same mean. The first step verifies that the samples are independent and that each sample is from a normally distributed population. However, we did not find any data frame with the population/sample with same deviations in the groups. Also, the Wilcoxon test does not require the hypothesis of normal distribution, but that the samples have to be dependence, which again was not what was reported in our data frame. But we continue the tests to demonstrate that when we cannot reach all the hypotheses, on the two tests and the result was null. And this was the result we got with the PValue well below 0.05(fig 17), which is the measurement standard for this test.

Figure. 17

1. **Discussion and Conclusion**

This paper aimed to understand and develop the process of Machine Learning Models, Statistics Tests, Programming and Visualization (Dashboard) for Data Analysis. Given the huge variety of agricultural products, as defined in our research analysis of a single product cow milk, and then choose others to products work on. The reason for that was to create a challenging learning process. Instead, look for one single product, where we had much more guaranteed correlation if for example, cattle and pasture land price only, or looking for a date on cereal production. We prefer to work with products from a view of a part without immediate correlation for a person with no knowledge in the agricultural sector.

However, the difficulty of finding an adequate database to proceed with the development of the tasks was crucial. It was possible to apply all the tasks requested in the assignment, but the outcome was not always satisfactory. We were able to understand and put into practice all the Machine Learning models, we collected enough information for our Sentiment Analysis and Time series, as well as the five statistical tests, and we efficiently exploited our data. But probably more knowledge and experience in the search and collection of data would change the results we found.

1. **Appendix**

|  |  |
| --- | --- |
| Dashboard - Streamlit | <https://github.com/ClaudiaOSBrennan/CA_2_MSc.git>  In Anaconda Prompt  streamlit run [sbs22026\_dash.py](https://github.com/ClaudiaOSBrennan/CA_2_MSc/blob/main/Dashboard/sbs22026_dash.py) |
| Git Hub – version control | <https://github.com/ClaudiaOSBrennan/CA_2_MSc.git> |
|  |  |

1. **Reference**
2. Dataset, May 2022., <https://ec.europa.eu/eurostat/databrowser/product/view/ORG_COPTYP?lang=en>
3. Dataset, <https://data.gov.ie/dataset?q=agriculture&theme=Agriculture&sort=score+desc%2C+metadata_created+desc>
4. Dataset, <https://ec.europa.eu/eurostat/web/agriculture/data/database>
5. Dataset, <https://w3.unece.org/PXWeb/en/>
6. [Laura O’Mahony](https://medium.com/@_lauraomahony?source=post_page-----600fbd5b5558--------------------------------), Oct 22, 2021., Getting Started with Data Collection Using Twitter API v2 in Less than an Hour, <https://towardsdatascience.com/getting-started-with-data-collection-using-twitter-api-v2-in-less-than-an-hour-600fbd5b5558>
7. Stats Models for Python, documentation is for the 0.6.1., https://www.statsmodels.org/0.6.1/generated/statsmodels.tsa.stattools.adfuller.html
8. Check for NaN in Pandas DataFrame (examples included), September 10, 2021., https://datatofish.com/check-nan-pandas-dataframe/
9. Bord Bia, 2022, Selected EU Prices Steers Meat Plant Prices(€/kg dw excluding VAT), <https://www.bordbia.ie/farmers-growers/prices-markets/cattle-trade-prices/cattle-selected-eu-prices/>
10. Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis, 12 June 2021., Big data in agriculture: Between opportunity and solution., <https://www.sciencedirect.com/journal/agricultural-systems>
11. UCD Dublin, 2022., Research, Innovation & Impact, <https://www.ucd.ie/agfood/research/>
12. Jia Chen, Shouwu He, Xiaoying Li, Oct 2021., A Study of Big Data Application in Agriculture., <https://iopscience.iop.org/article/10.1088/1742-6596/1757/1/012107/pdf>
13. Margaret Nolan, June2021., Money Mentor, the importance of Data Protection in Farming Families, <https://www.farmersjournal.ie/money-mentor-the-importance-of-data-protection-for-farming-families-622790>
14. Keith H. Coble\*, Ashok K. Mishra, Shannon Ferrell, and Terry Griffin, October 2017., https://www.researchgate.net/publication/323836711\_Big\_Data\_in\_Agriculture\_A\_Challenge\_for\_the\_Future
15. Aditya Sharma**,** January 1, 2020, Learn about PCA and how it can be leveraged to extract information from the data without any supervision using two popular datasets: Breast Cancer and CIFAR-10. <https://www.datacamp.com/tutorial/principal-component-analysis-in-python>
16. Mala Deep, March, 5 2021, Build Your First Data Visualization Web App in Python Using Streamlit., <https://towardsdatascience.com/build-your-first-data-visualization-web-app-in-python-using-streamlit-37e4c83a85db>
17. Ranjit Maity, Jul2021, Min Max Scaler, <https://medium.com/@ranjitmaity95/min-max-scaler-b2411ab3136d>
18. Nidhay Pancholi, Jul 23, 2021, Easily deploy web apps from Jupyter notebook., https://medium.com/analytics-vidhya/deploy-flask-streamlit-apps-from-jupyter-notebook-and-google-colab-5c3729dc3647
19. sklearn.preprocessing.MinMaxScaler, Scikit-learn 1.0.2., 2022, <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>
20. sklearn.ensemble.AdaBoostClassifier, Scikit-learn 1.0.2., 2022,<https://scikit->learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html?highlight=adaboost#sklearn.ensemble.AdaBoostClassifier
21. sklearn.model\_selection.GridSearchCV, 2022, version 1.0.2, <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>
22. scipy.stats.norm, 2008-2022, The SWciPy Community, <https://docs.scipy.org/doc/scipy-1.8.0/html-scipyorg/reference/generated/scipy.stats.norm.html>
23. NumPy, 2022, <https://numpy.org/doc/>
24. sklearn.ensemble.RandomForestClassifier, 2022, vesion 1.0.2 , <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html?highlight=random%20forest%20classifier#sklearn.ensemble.RandomForestClassifier>