**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 DA,  Statistics for Data Analytics,  Machine Learning for Data Analysis,  Data Preparation & Visualisation |
| **Assessment Title:** | CA2 |
| **Lecturer Name:** | Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Rita Raher |
| **Student Number:** | sbs22115 |
| **Assessment Due Date:** | 22/05/2022 |
| **Date of Submission:** | 22/05/2022 |

Icon

Description automatically generated with medium confidence

**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.

# Abstract

Some text

# Introduction

Due to the conflict in Ukraine, there has been increased pressure on Irish farmers due to rising input costs (Teagasc, 2022)

Many Irish newspapers have highlighted a bleak outlook for Irish farmers in 2022 due to the many rising input costs particularly when it comes to fertiliser (Blaney, 2022) (Murphy, 2022). There is also rising concern that these soaring costs will have a knock on effect to Irish consumers in the near future (O'Brien, 2022).

The war in Russia has had an effect to the feriliser prices as they produce

Github repo: <https://github.com/RitRa/Msc_CA2>

Link to dashboard: <https://share.streamlit.io/ritra/msc_ca2/notebooks/dashboard.py>

# Methodology

The Cross Industry Standard Process for Data Mining, as known as CRISP-DM, the methodology was developed with DiamlerChrysler, SPSS and NCR in 1996 (Santos, 2008). CRISP-DM, KDD, and SEMMA have all been compared, and CRISP-DM was found to be the most robust process for data mining projects (Qaiser, 2014). It consists of a cycle of 6 steps, as seen in Figure 1 (Chapman, 2000).

Diagram

Description automatically generated

Figure 1: Phases of the CRISP-DM reference model

1. Business understanding: Assess the data project from a business perspective and formulate a plan to achieve those goals.
2. Data understanding: Collect and understand the data to help gain insight and formulate the hypothesis.
3. Data preparation: begin activities for preparing the data for modelling, including pre-processing, transformation and cleaning.
4. Modelling: apply various modelling techniques to the data
5. Evaluation: evaluate and review the modelling techniques applied during the modelling step, and analyse whether the business goals were achieved.
6. Deployment: Create a real-world webpage for the model

# Data Collection

Using the [Central Statistics Office](https://www.cso.ie/en/statistics/), CSO, an Irish dataset on fertiliser prices was sourced with data dating back as far as 1980 to 2022. According to multiple sources, including a leading fertiliser manufacturer in Norway, fertiliser is made up of potassium, nitrogen and phosphate, and during production, it uses a large amount of natural gas (Yara, 2017) . Due to this research, additional features were added to the datasets for potassium, urea (as known as Urea), phosphate and Natural gas. These datasets are readily available on [www.indexmundi.com](http://www.indexmundi.com) and will be used as part of the analysis and models see table .

According to Teagasc, there are multiple factors for the increase in fertiliser prices, they include inflation due to covid recovery, sanctions on Russia due to the war in Ukraine and farmers delaying purchasing fertiliser with hopes that the price will drop (Teagasc, 2022; Liboreiro, 2022). Due to this evidence, feature for inflation is added to the dataset to hopefully try and capture these affects on the price of fertiliser. The dataset is source from <https://db.nomics.world/> and looks at housing, water, electricity, gas and other fuels.

|  |  |
| --- | --- |
| Datasets |  |
| Fertiliser | [source](https://data.cso.ie/table/AJM05) |
| Potassium | [source](https://www.indexmundi.com/commodities/?commodity=potassium-chloride&months=240&currency=eur) |
| Urea (Nitrogen) |  |
| Phosphate | [source](https://www.indexmundi.com/commodities/?commodity=rock-phosphate&months=360&currency=eur) |
| Natural Gas | [source](https://www.indexmundi.com/commodities/?commodity=natural-gas&months=120&currency=eur) |
| Consumer price index ireland | [source](https://db.nomics.world/OECD/MEI/IRL.CP040500.CTGY.M) |

## Twitter

Part of this project is to analysis sentiment related to agriculture. One of the best places to find out what the general population are thinking is Twitter. One way to gain access to tweets, is by using a developer account. This will give you api key, api key secret, access token and access token secret. To get elevated access to twitter you need to explain what you are doing with the tweets. This gives you access to 2M Tweets per month / Project. There is a python library called tweepy which allows for easy access to the twitter api. Unfortunately, Twitter only allows api access up to 5-7 days of tweets or 1,500 tweets in total so the total number of tweets returned was only 9, not enough for this project. Snscrape is a python web scraper for social networking services which allows you to add search criteria and gives back tweets without the restrictions of others like Tweepy. Specific keywords were targeted for focus on agriculture, they included: #farmers, Fertiliser, #farming, #agriculture, farm and each search criteria included “near: Dublin” to try and collected tweets in Ireland only. After appending all the datasets together, it resulted in ~1000 tweets ready for sentiment analysis. [jupyter]

# Data understanding

Descriptive statistics is an effective way of summarising data and identifying patterns.

They are measures that show where the centre of the data line is and are called measures of central tendency, which is the measure of the centre (Weiss, 2017). Central tendency includes: mean, median and mode.

* The mean is the sum of observations and dividing by the total
* The median finds the middle of the data
* The mode finds the most frequently reoccurring value

A boxplot shows the range and spread of data, with the middle of the data being represented by the line in the middle of the box, in this case, the median is 245. Therefore, 50% of the data is represented to the right and left. The first and second quartiles have a smaller spread and are much closer to the median than the third and fourth quartile. This means that the number of observations condensed into the lower quartiles is greater than the upper quartile. The boxplot also identifies a large number of extreme outliers ranging from 620 to 890.

|  |  |
| --- | --- |
|  |  |

In this dataset, they are 24 types of fertiliser [jupyter]. In the 1980s and the 1990s, there were only 11 to 12 types and in the last decade, the number of fertilisers available has exploded to over 20 types [jupyter].

Chart, bar chart

Description automatically generated

## Inferential statistics

Inferential statistics allows you to make inferences or estimations about the population based on the sample data (B. S. Everitt, 2010) . There are a number of ways of sampling methods that can be used, some examples are random sampling and stratified sampling. Random sampling

Stratified sampling is a way of obtaining samples that best represent the population as it divides the data into subgroups and takes a sample from each subgroup (Weiss, 2017).

Using this stratified sampling approach, the dataset was grouped by fertiliser\_type to sample each subgroup of fertiliser type, taking 5 from each [jupyter]. In order to determine whether the sample data are normally distributed, a probability plot was graphed and a Shapiro-Wilk test was performed. The Shapiro-Wilk test revealed that the sample data was not normally distributed as it had a p-value of 0.00001329.

### T-Test

For this project, it will be assumed that it is normally distributed. A t-test is performed on the sample dataset to check whether the sample mean price of fertiliser is equal to 288 yielding a p-value of 0.00097. The null hypothesis is rejected and the alternative hypothesis is accepted. The difference between the mean of the sample data and 288 is statistically significant.

### Anova

After creating sample data [Jupyter: ]the data for selected fertiliser types is checked to see if it is normally distributed using a Shapiro test [Jupyter]. Two out of the five samples were found not to be normally distribution so continuing with three that were normally distribution, a ANOVA test was performed

A sample size of x was selected

To ensure that the sample is

Using a T-test

ANOVA - Analysis of variance

As stated in

## Comparing countries against Ireland

Here are the following tests which were used to compare countries

|  |  |
| --- | --- |
| Parametric | Non-Parametric |
| T-Test | Wilcoxon signed-rank test |
| Anova | Kruskal-Wallis H Test |
|  | Mann-Whitney U Test |

### T-Test

A t-test is a parametric statistical test, used to compare the means of two groups (Brownlee, 2021). The Independent variable is geo, IE and PL (Ireland and Poland) and a t-test can be used to see if there is a difference between the fertiliser consumption volume of each group or if it is the same.

A sample of 15 is randomly selected from each dataset [jupyter] and then a shapiro\_wilk test is performed to check if the data is normal for each one. It is found not to be normal but this project we shall proceed with a t-test. The test output yields a t-statistic of 3.843 and a p-value is 0.001, as the p-value is less than alpha (0.05) and the null hypothesis is rejected. There is a difference between the means which is statistically significant.

### ANOVA

Analysis of variance test can be used to compare groups of more than two.

The Independent variable is geo, IE, PL and DE (Ireland, Poland, Germany).

A random sample of 15 from each country was selected for the test.

The ANOVA test produced an f-statistics of 0.0002, this is less than 0.05 and therefore the null hypothesis is rejected. Th difference between the 3 countries is statistically significant.

### Non-parametric

Non-parametric tests do not assume a normal distribution

### Wilcoxon signed-rank test

Statistical non-parametric tests are used for paired or dependent samples to discover if there is a difference between the samples (O'Loughlin, 2021). It is used when the data does not follow a normal distribution (pythonfordatascienceorg, 2018).

Firstly the data for the test was prepared, focusing on the years 2019 and 2020, and ‘Nitrogen’. Pandas pivot was used to create a dataframe for the analysis with geo, 2019 and 2020 as the columns. A Shapiro test was performed [jupyter] to ensure that the data was suitable for an non-parametric test, it found that the data was not from a normal distribution and a good candidate for a Wilcoxon test. Using the scipy library, wilcoxon test was performed, a p-value of 0.00000210 was returned [jupyter], therefore the null hypnothesis can be rejected. There is a significant difference in fertiliser consumption from 2019 to 2020.

### Kruskal-Wallis H Test

Test of hypothesis to determine whether there is a difference in rank totals between independent groups (O'Loughlin, 2021). This test is an alternative to the ANOVA test; an ANOVA can only be used if the data is normally distributed (Brownlee, 2018).

Using the scipy library, Kruskal test was performed on df\_de, df\_ie and df\_pol [Jupyter notebook]. It returned a p-value of 0.0002879 which means that the ranks of the groups were not the same and rejected the null hypothesis.

### Mann-Whitney Test

A non-parametric test, the Mann-Whitney test compares two sample means that come from the same population (Weiss, 2017). It is an alternative to the T-test and can be used when the data does not follow a Normal distribution. Using the scipy library, mannwhitneyu test was performed using sample data created previously, ie\_sampled and pol\_sampled. The p-value was found to be, 0.0464, which is less than alpha (0.05), therefore the sample means are different and the null hypothesis can be rejected.

# Data preparation

To prepare the data for machine learning algorithms, here are the steps followed:

1. Handling missing values
2. Handle duplicates
3. Remove outliers
4. Encoding values
5. Apply feature scaling

### Handling missing values

After merging new features, the dataset was now missing values from multiple columns, such as consumer price index and milk price. Some fertiliser types were missing pricing data because they were new to the market. From 1980 to 2013 there were only 10 to 11 fertiliser types, however, this grew in 2014 to over 20 types, see Figure 2. After dropping rows earlier than 2015, there were still 14% of data missing from fertiliser types, which is still too large to impute without introducing bias into the dataset. The fertiliser types with the largest missing values were identified and dropped from the dataset [jupyter]. This left 150 missing values, 8.7% which means the rest of the values can be imputated. Using the KKNimputer from sklearn, the missing values are replaced.

Chart, bar chart

Description automatically generated

Figure 2: Fertiliser type growth

### Handle duplicates

Duplicates can create bias in the model, it is important to identify and remove them [jupyter]

### Remove outliers

An outlier is an anomaly outside of the lower and upper quartiles of the data, generally representing either high or low extremes (GRUBBS, 1969). To ensure outliers

### Encoding values

Machine learning models require all input and output variables to be numeric to allow for them to perform mathematical computations and statistical analysis (Casari, 2018). Firstly, categorical data are transformed into numeric values using the category encoders library (Jupyter). To ensure that these values are not given any statistical significance over each other, OneHotEncoding is used to create new columns for each value and the rows are then filled with 1 or 0, representing true or false for if that value exsits [jupyter]. Prior to completing this step of the process, the model had been reporting a precision of 40%, and after it reported 85% precision.

### Apply feature scaling

Machine learning algorithms will give more weight to features with larger numeric values (Roy, 2020)

# Modelling

The goal of this project is to predict the price of fertiliser. Therefore is it a regression problem rather than a classification problem.

There are multiple regression models that can be used to predict the price of fertiliser.

**Chart

Description automatically generated**

The heatmap above uses three distinct colors so it would be easy to distinguish between high, medium and low correlation. The independent variables that have mid to high correlation with fertiliser price are: gas price (0.41), consumer\_price\_index(0.51), phosphate price(0.65) and urea price(0.71). Urea has the strongest correlation with fertiliser price.

**Linear regression**

Linear regression is a supervised learning model which performs a regression task, it can be used on continuous variables. It will predict fertiliser prices based on independent variables passed into the model.

Linear regression can be used when there is a strong relationship between

**Lasso regression**

Some text

Sentiment analysis

Added in stopwords for farm, farms, amp as the topics

Decided to apply the year to

Timeseries analysis

## **ARIMA Model**

<https://siebert-julien.github.io/time-series-analysis-python/overview.html>

Using facebooks prophet for time series analysis, we can predict the future price of fertiliser

With autoregressive model

Time series

# Evaluation

text

# Conclusion

text

Add bib

Ideas for sentiment

positive and negative sentiment

<https://ec.europa.eu/eurostat/databrowser/view/T2020_RN310/bookmark/table?lang=en&bookmarkId=0c6ee6ae-b496-4703-a0dd-66b2b2fc8184>