A picture containing diagram

Description automatically generated

MSc in Data Analytics

**CA2**

Author: Mara Carcione

e-mail: sba22243@student.cct.ie

Student ID: sba22243

GitHub repository: https://github.com/sba22243/ca2.git

Words count: 3644

## Abstract

In questo documento si esamina e si prevede il sentimento dei tweet sul vaccino covid nell’arco di un anno, da Giugno 2020 a Maggio 2021. Si presentano alcuni dati sui tweer e si stima il sentimento delle persone che usano Twitter per discutere di questo argomento.

## Introduction

Topic: vaccine

this notebook will read the dataset containing only the tweets related to vaccine.  
The dataset rapresent tweets from June 2020 until May 2021. **The dataset was downloaded from the link** [**https://archive.org/details/twitterstream?sort=-publicdate**](https://archive.org/details/twitterstream?sort=-publicdate)

E’ stato scelto l intervallo di tempo che va da Giugno 2020 a Maggio 2021 perche’ e’ in questo periodo che sono stati sviluppati e distribuiti i primi vaccini covid.

## Materials and Methods

The Cross Industry Standard Process for Data Mining (CRISP-DM) represents the most common basic methodology used to standardise data mining processes in all sectors (Hotz, 2018). It includes six steps:

1. Business/Research understanding Phase
2. Data Understanding Phase
3. Data preparation Phase
4. Modelling Phase
5. Evaluation Phase
6. Deployment Phase



Figure - CRISP-DM

1. Business / Research Understanding Phase is the essential phase that focuses on the objectives of the project and, therefore, on the determination of business objectives with a deep understanding of the customer’s needs. At this stage, it is important to determine the availability of resources by making a cost-and-benefit analysis. Finally, it is also very important to define the technical aspect of data mining, producing a project plan that selects the technologies to be used. (Hotz, 2018)
2. Data Understanding Phase is the phase of understanding the data in which the initial data are collected; the data are described by examining their properties, the data are explored by identifying their relationships and finally the quality of the data is verified to examine how dirty or clean this data is. (Hotz, 2018)
3. Data Preparation Phase prepares the final datasets for modelling. In this phase, the data to be used is determined. Then we move on to cleaning the data by correcting or removing incorrect values. If necessary, variables are transformed, and data are reformatted as needed. (Hotz, 2018)
4. Modelling Phase is the shortest phase of the project. The various models are built and evaluated. (Hotz, 2018) It consists of four tasks:
   1. Select modelling techniques - determines which algorithms to try
   2. Generate test design - split data into training, test and validation sets
   3. Build model
   4. Assess model - the data scientist interprets the results of the model applied based on knowledge of the domain
5. Evaluation Phase examines which model best suits the company through:
   1. Evaluate results
   2. Review process
   3. Determine next step
6. Deployment Phase is the final phase and is characterised by four points:
   1. Plan deployment - develop and document a plan for model deployment

The next sections show the phases implemented to create the current report.

## Stage One - Determine Business Objectives and Assess the Situation

Ireland is one of the largest dairy producers in Europe. (EC, 2022) The purpose of this report is to analyse the internal production of milk and its derivatives. Furthermore, a comparison will be made with some European states comparable with Ireland in terms of population.

## Stage Two - Data Understanding

### Collect Data

The data was imported from various institutional sites such as the Irish CSO (CSO, 2022), Agrifood from the European Commission (Agridata, 2022), Population of Europe (European Union, 2021)

To compare Ireland with other European countries, the number of inhabitants of each state within the union was considered.

The data relating to the following topics were imported from the CSO:

* Manufacturing Milk Prices (including VAT) (Euro)
* Value at Current Prices for Output, Input and Income in Agriculture
* Quantity of Agricultural Output
* Intake of Cows Milk by Creameries and Pasteurisers
* Milk Sales (Dairy) for Human Consumption
* Production of Dairy Products

These data refer only to Ireland and will be used for the construction of the dashboard.

The following information was extracted from the Agrifood site, using API calls, offered by the portal itself:

* Raw Milk Price
* Weekly TAXUD imports and exportsDairy Prices
  + Taxud Import weekly data
  + Taxud export weekly data

The extracted information refers only to the countries of interest.

For the sentiment analysis, the public Twitter tweets related to the topics of interest of this research were extracted.

### Exploratory Data Analysis (EAD)

After having collected the data in the previous phase, we moved on to their processing, in particular, the variables to be analysed for each dataset were selected, the percentages of missing values were calculated, and the columns not appropriate for the purposes of the analysis were deleted.

For all datasets, the analysis of null values was carried out, the statistical description highlighting the count, mean, std, min, max, percentile (25%, 50%, 75%) and Inter Quartile Range (IQR).

Regarding the EU population, we have the following graph.Timeline

Description automatically generated with medium confidence

Figure - EU Population

This shows that Ireland, Slovakia, and Denmark have comparable populations.

Analysing the data collected by the CSO I found that:

* + Manufacturing Milk Prices (including Vat) (Euro), Value at Current Prices for Output and Quantity of Agricultural Output have non-zero values
  + Intake of Cows Milk by Creameries and Pasteurizers has 46% null values, the dataset is therefore not usable because it is not reliable
  + Milk Sales (Dairy) for Human Consumption has 11.98% of null values, the dataset is good
  + Production of Dairy Products has 2.918% null values and is good because the value is below 20%

Analysing the data from Agrifood Europe, the Raw Milk Price, Dairy Prices, Milk Import and Milk Export datasets can be used to make the prediction. In particular, the study is done on the Raw Milk Price of the three countries, Ireland, Slovakia and Denmark.

## Assess Data Quality

### Agrifood Dataset

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Milk Price

In Raw Milk Price the price column shows outliers and in the distribution graph it is noted that the distribution of the Raw Milk Price is not normal.

Chart, bar chart, treemap chart

Description automatically generated

Figure - Correlation matrix for Raw Milk Price

In the correlation matrix price and year are positively correlated (0.56).

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Dairy Price

Chart, bar chart, treemap chart

Description automatically generated

Figure - Correlation Matrix for Dairy Price

Dairy price has outliers in the price column and observing the distribution graph we observe that Dairy Price is not normal.

The correlation matrix shows a positive correlation between Price and Year of 0.4.

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Milk Import

In Milk Import the outliers are shown in euroValue and Kg, while the distribution graph is flat. The outliers are relevant and therefore are not removed.

Chart, treemap chart

Description automatically generated

Figure - Correlation matrix Milk Import

In the Correlation heatmap of Milk Import it is observed that euroValue is slightly correlated with Kg (0.9).

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Milk Export

Milk Export shows outliers in euroValue and in Kg. The distribution graph is flat.

Chart, treemap chart

Description automatically generated

Figure - Correlation matrix for Milk Export

In the Correlation heatmap in Milk Export euroValue and Kg are weakly correlated, in fact the correlation is equal to 0.16.

### CSO Dataset

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Manufacturing Milk Price

Manufacturing Milk Price has outliers in Value and has a non-normal distribution.

Chart

Description automatically generated

Figure - Correlation matrix for Manufactoring Milk Price

The Correlation heatmap between the Date variable and the Value variable has a value of 0.49. The value 0.49 indicates that there is a moderate correlation between the two variables.

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Current Price

In Value at Current Prices for Output, Input and income in Agriculture we note the presence of some outliers in the Value variable.

Chart

Description automatically generated

Figure - Correlation matrix for Current Price

In the Correlation heatmap we note that the correlation coefficient is equal to 0.72, this means that there is a strong positive correlation between the Year variable and the Value variable, consequently the two variables tend to move in the same direction, therefore increasing the value of one variable has a positive effect on the other variable and vice versa.

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Quantity of Agricultural Output

In Quantity of Agricultural Output we note the presence of numerous outliers in the Value variable, the distribution of the graph is not normal.

Chart

Description automatically generated

Figure - Correlation matrix for QUantity of Agricultural Output

The Correlation heatmap shows a strong positive correlation between the Year variable and the Value variable equal to 0.73.

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Milk sales for human consumption

Milk Sales For Human Consumption has no outliers, consequently there are no anomaly values

Chart

Description automatically generated

Figure - Correlation matrix for Milk Sales for human consumption

The Correlation heatmap has a correlation value of -0.089 which indicates that there is a weak negative correlation between the Date and Value variables.

Chart, box and whisker chart

Description automatically generated

Figure - Outliers and Distribution for Production of Dairy Products

In Production of Dairy Products the Value variable has some outliers and the distribution is not normal.

Chart

Description automatically generated

Figure - Correlation matrix for Production of Diary Products

The Correlation heatmap has a correlation value of 0.31 which indicates that there is a moderate correlation between the Date and Value variables.

## Statistics

The prediction is only on Raw Milk price. I’m going to use only raw milk price dataset for statistical analysis.

**Confidence Interval**

The Confidence Interval is used to compare the Raw Milk Price in three countries such as Ireland, Slovakia and Denmark. The calculation is made considering a confidence level of 95%. Below are the values of the calculated limits

Ireland: LL= 33.12268918595459 , UL=35.07768640089987, m=34.10018779342723

Slovakia: LL= 29.43680942753971 , UL=30.747115806105153, m=30.091962616822432

Denmark: LL= 33.99781679894197 , UL=35.705609126983944, m=34.85171296296296

**ANOVA**

Analysis of Variance is a statistical technique used to verify if there are significant differences between the means of two or more groups. You can use Anova to compare milk prices in Ireland, Slovakia and Denmark. However, some conditions must be met to use Anova, such as normality and homogeneity of the variances.

Testing the Raw Milk Price dataset of the three countries in the notebook, we notice that from the plots, the distribution is not normal for each countries and for the whole dataset.

|  |  |
| --- | --- |
| Ireland | Slovakia |
|  |  |
| Denmark | **ALL RAW MILK DATASET** |
|  |  |

Figure - QQ plots

Checking with the Shapiro test, I can conclude that the distribution is not normal, and I cannot apply ANOVA. The test was made for each country and for the whole dataset.

* Ireland: (statistic=0.90357506275177, pvalue=1.684528905254723e-10)
* Slovakia: (statistic=0.9588568210601807, pvalue=2.74332501248864e-06)
* Denmark: (statistic=0.8358032703399658, pvalue=4.377035555156545e-15)
* All raw milk dataset: (statistic=0.8920753002166748, pvalue=1.713981439056487e-21)

As we can see, all p-values are below 5%

**T-Test one population**

Problem Definition:

* H0 : u = average of raw milk price for Ireland is 33 EUR, Slovakia is 31 EUR, and Denmark is 40 Eur
* H1 : u =! 33 (Ireland), 31 (Slovakia), 40 (Denmark)

T-Test one population Results:

* H1 is verified for Ireland, and H0 is rejected
* H1 is verified for Slovakia, and H0 is rejected
* H0 is verified for Denmark

**T-Test two populations**

* We perform the test assuming as H0 that mu1 = mu2 (Ireland = Slovakia)
  + statistic=6.720257056206553
  + pvalue=5.850714791086066e-11
* We perform the test assuming as H0 that mu1 = mu2 (Ireland = Denmark)
  + statistic=-1.1423441276120214
  + pvalue=0.2539511420273216
* We perform the test assuming as H0 that mu1 = mu2 (Slovakia = Denmark)
  + statistic=-8.706361494005543
  + pvalue=6.875662452927534e-17

**Wilcoxon Test**

To compare two paired groups, a Wilcoxon test needs to be used. The test is a nonparametric statistical test. This test calculates the difference between sets of pairs and analyses the differences between them to establish if they are statistically significantly different. (Investopedia, 2022)

The table below shows the result by comparing two countries on a specific product:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Countries | Product | Year | Statistics | pvalue |
| Ireland, Ireland | Raw Milk | 2020, 2021 | 29 | 0.4697265625 |
| Ireland vs Denmark | Raw Milk | 2020 | 18 | 0.10986328125 |
| Slovakia vs Denmark | Raw Milk | 2020 | 0 | 0.00048828125 |
| Ireland vs Denmark | Raw Milk | 2021 | 10 | 0.02099609375 |
| Ireland vs Slovakia | Raw Milk | 2021 | 0 | 0.00048828125 |
| Denmark vs Slovakia | Raw Milk | 2021 | 0 | 0.00048828125 |

**Chi-Squared test**

* H0 : the price of the raw milk is equals between the three countries
* H1 : the price of the raw milk is not equals between the three countries

Text, application

Description automatically generated with medium confidence

Figure - dataset for Chi-squared test

p-val is less then 5% (p-val=4.0344087717482904e-100), the H0 is rejected.

**Kruskal Wallis**

E’ un test statistico non parametrico che viene utilizzato per determinare se esiste una differenza significative tra I ranghi medi dei campioni confrontati. Puo’ essere usato in alternativa ad ANOVA che pressupone la normalita’.

* Ireland-Slovakia: The difference, when taking the median into account, is significant as the p-value is less than 5%.
* Slovakia-Denmark: The difference, when taking the median into account, is significant as the p-value is less than 5%.
* Ireland-Denmark: The difference, when taking the median into account, is insignificant as the p-value is much greater than 5%.

# Stage Three - Data Preparation

In this step, I’m going to prepare the data for the raw milk prediction.

Graphical user interface, application

Description automatically generated

Figure - raw milk price prediction dateset - initial

Drop not relevant columns and calculate the percentage of each product in the whole dataset.

* % of organic raw product: 0.061313868613138686
* % or raw product : 0.9386861313868613

The percentage of Organic is only 6% compared with the not organic product. I’m going to remove the organic rows because are not relevant compared with the whole dataset.

Table

Description automatically generated

Figure - dataset without organic milk values

Raw Milk Prices prepared dummies:

Graphical user interface, application

Description automatically generated

Figure - dummies dataset

Text, table

Description automatically generated with medium confidence

Figure - dataset info

There aren’t NaN/Null values. The dataset is ready for the modelling phase. No imputer activity is needed.

# Stage Four - Modeling Phase

The choice of an algorithm depends on many factors, including the size, quality and nature of the data, etc. If the data is labelled, you have a supervised learning problem; if you have unlabeled data, you have an unsupervised learning problem. If the solution involves interacting with the environment and obtaining feedback, there is a learning problem by reinforcement. Furthermore, depending on the data output, there may be a classification or regression problem. If the output is numerical, there is a regression problem; if instead, it is categorical, there is a classification problem.

The table below shows a possible algorithm for model selection.

|  |  |  |
| --- | --- | --- |
| **Algorithms** | | |
| Unsupervised | Clustering | Hierarchical |
| K-Means |
| KNN |
| DBScan |
| LDA |
| PCA |
| SVD |
| Supervised | Classification | Naïve Bayes |
| Random Forest |
| Logistic Regression |
| Linear SVM |
| Decision Tree |
| Gradient Boosting |
| Regression | Linear Regression |
| Random Forest Regressor |
| Poisson Regression |
| Lasso Regression |
| Decision Tree Regressor |
| Multiple Regression |

**Split dataset in training and test**

Before applying any kind of algorithm, the dataset must be split into two or three parts, called train, validation and test set. In the Train Set, the model learns the relationships between the input variables, the X, and the output variables, which are represented by the Y. In this way, the model compares the result of its prediction with the real one and consequently updates the various parameters to minimise the error compared to the previous time.

During this phase, the overfitting phenomenon may occur; that is, the data used are perfectly predicted in the training phase but cannot generalise on new data. In the Validation Set phase, to avoid overfitting the model, data is given that it does not know and based on this data, it makes a forecast. These data are exactly those of train. We will then have a predicted y and a real y that are compared to see how well our model can predict the output variable with a good approximation. The Validation Set phase deals with validating the results obtained in the training set.

The test dataset quota is 30% of the full dataset, hence the train dataset contains 70% of the samples.

**Linear Regression**

Finally, we proceed to test the model, where the functioning of the model is evaluated and displayed.

Linear regression is a data analysis technique that mathematically models the dependent variable and the independent variable as a linear equation.

Linear regression models are fairly simple and provide an easy to interpret mathematical formula for making future predictions (AWS, 2022).

In machine learning, algorithms analyse large data sets and work backwards from that data to calculate the linear regression equation. The linear regression analysis must mathematically transform the data values ​​in order to satisfy:

* Linear relationship - there must be a relationship between independent and dependent variables
* Residual independence - the residual represents the difference between the observed data and the predicted value. Residuals are used to measure the accuracy of the forecast
* Normality - The residuals must be distributed normally; if they are not normalised, the data must be tested to identify anomalous values
* Regression can be simple linear or multiple linear.

Linear Regression is used as a model to predict raw milk price. For the estimator, a cross-validation approach it is implemented. K-Fold is initialised with five splits, shuffle as True and a random state of 42. The hyper-parameters contain the number or feature selected as the maximum number of columns in the training dataset. This value is calculated as 6. The GridSearchCV functionality is used in order to find the best fit. The scoring used is r2. The output of this step is “ Fitting five folds for each of 5 candidates, totalling 25 fits”.

The partial view of the cv\_result dataset is shown in the following table.

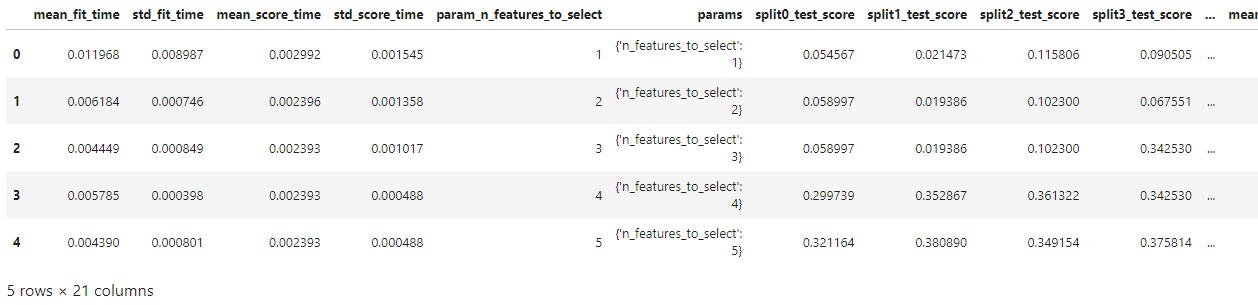


Figure - cv results for Linear Regression

The dataset above contains useful information that enables the researcher to find the best number of features needed to train the model better. Plotting the pair (param\_n\_features\_to\_select, mean\_test\_score) and (param\_n\_features\_to\_select, mean\_train\_score), the diagram is shown in the following picture:

Chart, line chart

Description automatically generated

Figure - optiman number of features for Linear Regression

The diagram shows that the number of features that give the highest r2 is 5, i.e. the maximum number of features present in the dataset.

Calculating the model for n\_features\_to\_select=5, it was obtained as follow

* r2 = , 0.37700637079869503
* MSE= 4.869073758447459
* RMSE= 23.70787926520166

It is possible to visualise the actual values against the predicted ones on a plot. In this way, we can give a better understanding of the result obtained.

Chart, scatter chart

Description automatically generated

Figure - Linear Regression prediction - Actual vs Predicted

**Decision Tree Regression**

Among the supervised methods used in classification and regression, we find Decision Trees or DT. Their goal is to build a predictive model by learning simple decision rules extracted from the data itself.

Advantages of DT:

* Easy to visualise and understand
* A little data preparation is needed, such as dummy variable, normalisation and removal of null value
* the complexity is logarithmic

Disadvantages of DT:

* Possibility to create over-complex trees or overfitting. A pruning method is needed.
* Small variations of data create different trees. Hence they are somehow unstable

For the model, the same approach of the Linear Regression is implemented, i.e. cross-validation and K-Fold initialised with five splits, shuffle as True and a random state of 42. The hyper-parameters contain the number or feature selected as the maximum number of columns in the training dataset. This value is calculated as 6. The GridSearchCV functionality is used in order to find the best fit. The scoring used is r2.

The partial view of the cv\_result dataset is shown in the following table.

A screenshot of a computer

Description automatically generated with medium confidence

Figure - cv results for Decision Tree regression

Plotting the pair (param\_n\_features\_to\_select, mean\_test\_score) and (param\_n\_features\_to\_select, mean\_train\_score), the diagram is shown in the following picture:

Chart, line chart

Description automatically generated

Figure - optimal number of features for Decision Tree Regression

The diagram shows that the number of features that give the highest r2 is 5.

Calculating the model for n\_features\_to\_select=5, it was calculated

* r2 = 0.927575602663607
* MSE= 1.6601487229713547
* RMSE= 2.7560937823834197

It is possible to visualise the actual values against the predicted ones on a plot. In this way, we can give a better understanding of the result obtained.

Chart, scatter chart

Description automatically generated

Figure - Decision Tree Regression prediction - Actual vs Predicted

The table below shows the comparison of the two models according to the results obtained.

Graphical user interface, text, application

Description automatically generated

Figure - Comparison of the two models

From the table, we understood that Decision Tree Regression obtained a better result of accuracy than Linear Regression. The Linear Regression model was calculated using all features available, and from this result, we can deduce that more features are needed.

The Decision Tree Regression model needs to be improved in order to increase accuracy.

**Sentiment Analysis**

The method used to classify the text by identifying the various subjects expressed therein is called Sentiment Analysis. A text is classified as neutral, positive or negative, or with a score called polarity, which indicates the strength of the sentiment (Mathworks, 2022).

The dataset used for the sentiment analysis was extracted from Twitter by filtering the tweets with keywords such as milk. Cheese, butter, dairy, Ireland.

In the table below, we have an example of a dataset retrieved from Twitter:

Graphical user interface, text

Description automatically generated

Figure - Datasets containing tweets extracted from the API

The dataset does not contain any label related to sentient analysis for each tweet. Using the polarity calculated by the library TextBlob and classifying according to the following algorithm, the researcher was able to classify every tweet and prepare the dataset for the modelling.

* Polarity>0: Positive
* Polarity=0: Neutral
* Polarity <0: Negative

The final dataset is shown below:

Text

Description automatically generated

Figure - Dataset with calculated sentiment

The dataset was also cleaned using stop words in English. Special chars in the tweets were also removed. In this case, we were able to prepare a clean dataset to be used in the model creation.

The TfidfVectorizer object was used to transform the tweets, and the model was built using the MultinomialNB model.

The classification report for this model is shown below:

Table

Description automatically generated

Figure - Classification Report

## APPENDIX

Below we have the list of relevant jupyter notebooks

* MSC\_DA\_CA2.ipynb - this is the main notebook
* MSC\_DA\_CA2\_DASHBOARD.ipynb - this is the notebook related to the dashboard

## References

* Agridata (2022) *European Commission | Agri-food data portal*. Available at: <https://agridata.ec.europa.eu/extensions/DataPortal/API_Documentation.html> (Accessed: 4 January 2023).
* CSO (2022) *Central Statistics Office*. Available at: <https://data.cso.ie/> (Accessed: 4 January 2023).
* EC (2022) *Milk and dairy products*. Available at: <https://agriculture.ec.europa.eu/farming/animal-products/milk-and-dairy-products_en> (Accessed: 6 January 2023).
* European Union (2021) *Life in the EU*. Available at: <https://european-union.europa.eu/principles-countries-history/key-facts-and-figures/life-eu_en> (Accessed: 4 January 2023).
* Hotz, N. (2018) ‘What is CRISP DM?’, *Data Science Process Alliance*, 10 September. Available at: <https://www.datascience-pm.com/crisp-dm-2/> (Accessed: 9 November 2022).
* Investopedia (2022) *Wilcoxon Test: Definition in Statistics, Types, and Calculation*, *Investopedia*. Available at: <https://www.investopedia.com/terms/w/wilcoxon-test.asp> (Accessed: 6 January 2023).
* Mathworks (2022) *Sentiment Analysis*. Available at: <https://it.mathworks.com/discovery/sentiment-analysis.html> (Accessed: 6 January 2023).