**Continuous Assessment 2: Investigation into Cow’s Milk Production**

*Abstract*

*Agricultural data, particularly diary production contributes to the economies of many countries worldwide. Machine learning can be used to analyse agricultural data in order to classify and predict future output. In this report, under a CRISP-DM framework, I compare the milk output of Ireland to four other European countries with a similar population in 2020. There is variability in milk output. Ireland’s mean milk output is no different to the other countries, however Ireland is the only country with a positive correlation between output and population. In machine learning, both a decision tree and random forest can accurately classify the country based output, year, and population. However it is difficult to predict the future output for all countries. Sentiment analysis shows that overall milk is discussed neutrally on twitter.*

**Introduction**

Cow’s milk is a nutritious and important dietary component for many groups of people, particularly children and pregnant women (Morin et al., 2018). Naturally, the dairy industry contributes to the economies of many countries worldwide (Buleca et al., 2018). Therefore data collected on milk production is an important source of information, which can be used in conjunction with changes in population levels over time, to inform future policies and sustainable development of the dairy agricultural industry.

Machine learning can be used to analyse agricultural data in order to classify and predict future output. There are three categories of machine learning: supervised learning, unsupervised learning, and semi-supervised learning. Supervised learning involves using labelled data (input and output parameters are known) to classify new data. In this report, I will use decision trees and random forests, both of which are supervised learning techniques, to classify countries based on milk output, population and year. Decision trees involve creating a hierarchy of if/else questions that lead to a decision, with the goal of having single classification categories in as few questions (nodes) as possible. Random forests, on the other hand, consist of multiple decision trees created from random subsets of data, and the solution is based on the consensus of predictions from each tree. Random forests tend to be more robust than single decision trees because they use more data to provide evidence. Supervised learning can also be used for time series forecasting, which involves predicting the outcome of variables as closely as possible to the known value.

The CRISP-DM framework was used in this analysis project, see Fig. 1 showing a screenshot of the dynamic Gantt chart used to track progress. More evidence of this framework being used can be found in the layout of the jupyter notebook, as the phases of data preparation, modelling and evaluation are cycled through.

|  |
| --- |
|  |
| Figure 1. Screenshot of Gantt chart taken on 28/12/2022 |

The aim of this report is to analyse the production of cow’s milk in Ireland over time and compare it with four other EU countries. To achieve this, I used two dependant variables (one per data frame); the Gross Production Index (GPI) is a measure of the total output of cow’s milk and Gross per capita Production Index (GPCPI) is a measure of the total value of cow’s milk per person, which is calculated by dividing the Gross Production Index (GPI) by its population across five different European countries. The countries included for comparison were Croatia, Denmark, Finland and Slovakia, which were specifically chosen because their population was similar (within 1 million) to Ireland in 2020. Additionally, Denmark has a similar climate and physical geography to Ireland (Clancy et al., 2009; Ladrech and Little, 2019), this is relevant as cows are predominantly outdoor animals which may be affected my temperature, weather and topography (Krohn et al., 1992). The data consists of the five selected countries, the Gross Production Index Number or Gross per capita Production Index Number (dependant on dataset used), the year of collection and the country’s population for that year. In order to analyse this information, the data was prepared by transforming the raw data into a format useful for analysis, by generating subsets of data for the countries of interest and product of interest (cow’s milk), creating new variables and new data frames. Exploratory data analysis was conducted, assessing the size and features of these data frames, investigating initial patterns and the spread of the data and assessing whether further transformation of variables is required. The data was then assessed using descriptive statistics, such as the mean, median and mode and through graphs exploring the relationships between the different variables in the data set. Five different inferential statistics were conducted including generating confidence intervals, t-tests, ANOVAs, correlation testing and hypothesis testing. Two separate machine learning techniques were used to investigate the data. First, the data was then classified into different countries types using both decision trees and random forests. Hyperparameters were investigated using GridsearhCV. The effectiveness of each method was summarised and the two models were compared. An interactive dashboard was created to dynamically show the change over time of milk production across the five different countries. Second, a time series forecasting machine learning model was used to analyse the milk production index per capita annually and predict future production. The parameters of this model were altered in order to assess the effectiveness of this model. Lastly, a sentiment analysis was conducted using tweets scraped from Twitter to assess public sentiment of milk over a short, medium and long period.

**Methods**

*Data Preparation*

The first set of data used in this analysis was sources from the Food and Agricultural Organization of the United Nations. It was published by FAOSTAT, which is a free platform provided access to food and agricultural statistics for over 245 from the 1961 onwards. This data was easily accessible, free and required no licensing or permissions. Additionally it was already processed and cleaned, meaning there were no empty or unexpected values in cells, nor were there any null values. The initially downloaded data frame contained 234 different locations (predominantly individual countries, with some groupings) and 197 different agricultural items along with two outputs for each measure (a Gross Production Index and a Gross per capita Production Index), a value for each measurement and the year. Two dataframes were made of just the cow’s milk data, each containing the five countries of interest, the year and then one data frame of the Gross Production Index and one of the Gross per capita Production Index. Data was checked for outliers and there was no evidence of any point having a disproportionate influence, e.g. high values correlated with high population and vice versa. For this reason outliers were retained in both data frames. A set of population data frames (freely obtained, no permissions required) Eurostat were created and then merged using a loop with both of these data frames to add each country’s population per year. Lastly, a data frame for a sentiment analysis was created by scraping tweets from twitter. Please see Table 1 below listing the libraries imported and used throughout the entire analysis.

|  |  |
| --- | --- |
| Table 1. Libraries imported for report and their function | |
| **Library** | **Function** |
| plotly | high-level data visualization library for quality graphs |
| pandas | data manipulation and analysis |
| seaborn | visualization library |
| matplotlib | 2D plotting library |
| numpy | provides support for large, multi-dimensional arrays and matrices, along with a large collection of mathematical functions to operate on these arrays |
| os | provides functions for interacting with the underlying operating system in a portable way |
| scipy | provides support for scientific and technical computing |
| sklearn | provides a range of tools for machine learning and statistical modeling |
| statsmodels | provides a range of classes and functions for estimating and testing statistical models, |
| warnings | provides a way to issue warnings to the user of your program |
| folium | allows users to create interactive, web-based maps and visualizations of geospatial data |
| ipywidgets | allows users to create interactive, web-based user interface (UI) elements for Jupyter notebooks and web applications |
| nltk | provides tools and resources for working with human language data (text) |
| textblob | provides a simple API for working with text data |
| sys | provides functions and variables that allow you to access and manipulate various aspects of the Python interpreter |
| tweepy | allows users to easily retrieve and post tweets, follow and unfollow users, and perform other actions on Twitter, such as searching for tweets or retrieving user information |
| re | provides functions for working with regular expressions |
| string | functions for formatting and printing strings, for manipulating the case of strings, for searching and replacing substrings, and for generating and manipulating string data in various other ways |
| PIL | provides support for reading and writing a wide range of image file formats |

*Descriptive Statistics*

Standard descriptive statistics were conducted for the GPI and GPCPI data frames. This included looking at the first 20 rows of data to check that the format was as expected. Using the .describe command for object variables, the number of rows and unique data points was assessed. Using the .describe command, the dependant numerical variable, Value and Population were assessed in terms of count, mean, min, max and the quantile ranges. For the dependant variables, the mean (the average) and the median (the middle number, where there are equal number of data points either side) and this was repeated for each country. See Table 2 and 3 for results.

In order to investigate patterns in the data and relationships between variables, it was necessary to graph the data. I first graphed the relationship between the amount of cow’s milk produced (GPI) and the year (see Figure 2) and the I graphed the changed in population over time (see Figure 3).

There were five different types of inferential statistics conducted. First, the 95% confidence intervals for the GPI of cow’s milk was calculated for each country. See Table 4. Second, I wanted to investigate whether there is a difference between the average production of cow's milk between Ireland and Denmark. Data was checked for normality using the Shapiro test. I then performed a U-Mann Whitney test comparing the means of the two sets of data. Third, I compared the mean across all countries. Again, I checked the data for normality using the Shapiro test. I then used a Kruskal-Wallis test to compare the mean across multiple variables. I repeated this test, after removing Denmark, as Denmark had already been compared to Ireland. Next, I compared the means of the GPCPI across all five countries. Again, I checked for normality using the Shapiro test and then performed a Kruskal-Wallis test. Lastly, I wanted to investigate whether the amount of milk produced (GPI) by a country was correlated with the country’s population. I predicted that there would be a correlation. Null: The amount of milk produced by a country is not related to the population. Alternative: The amount of milk produced by a country is related to the population. I used the Spearman’s Correlation test as I have already confirmed that the data is not normally distributed.

*Machine Learning: Decision Tree*

The aim of the decision tree was to investigate whether the country could be identified based on the cow’s milk output (GPI), the population and the year. All required libraries were imported (see Table1.). The format of the data frame did not need to be changed as it was already in an appropriate format for a decision tree. The independent variables included in the decision tree were the cow’s milk output (GPI), the population and the year. The dependant variable was Country (written as “Area” in the jupyter notebook). For this decision tree, 70% of the data was used to train the model, and the remaining 30% was used to test the model. The Gini index was used as a measure of impurity of splits of the decision tree. A max depth of 5 nodes was set to prevent overfitting of the training data to the decision tree. Post hoc analysis was conducted using the mean and standard deviation to identify how accurately this tree could predict the country based on the explanatory variables. A plot of the tree was generated representing the different splits of the five levels of node.

*Machine Learning: Random Forest*

The same data frame in the same format was used for the random forest, as well as the same variables included for both the independent variables and the dependant variable. The test size was set to 0.33 meaning that approx. 66% of the data would be assigned to a training set while 33% would be used for the test model. Two random forests were run. In order to test whether the model would run successfully the first forest contained just 10 trees. The second forest contained 100 trees, because the great the number of trees in the random forest, the more robust the results should be. GridSearchCV was used to optimise the hypermaramters (e.g. finding the best parameter values for an estimate).

*Machine Learning: Traffic Forecasting*

For this machine learning model, a data frame with the Gross per Capita Production Index (GPCPI) was used. Data was checked for null values. Data was then subset into five different data frames for each country. The following steps were repeated for each data frame to get results for each country. The dependant variable “Value” was checked to make sure it was an integer data type. The “Year” column was converted to Datetime type. Irrelevant columns were dropped from the data frame. From this reduced data frame, I created two data frames – train and test – for data before a cutoff year (70% of the data) and after cutoff year (30% of the data). This allowed for using a data frame to train a prediction model and a data frame for comparing prediction results with the actual values for that time period.

The ARIMA (ARIMA) model was used. ARIMA stands for Autoregressive Integrated Moving Average and has three parameters (p,d and q) which are coefficients for the autoregressive, integrated and moving average parts of the model. Data was plot to see if there was a trend. The d value was calculated by how many time the data was differentiated. Importing from the pandas.plotting library, an autocorrelation plot was generated and examined to see which boundary the curve was pitching down after giving a p value. A partial autocorrelation plot (using the statsmodels.graphics.tsaplots library) was generated in order to find the value of the q coefficient, which was calculated by counting the number of outlying data points.

*Machine Learning: Sentiment Analysis*

The aim of this model was to investigate public opinion on cow’s milk over time. I created a developer twitter account in order to access publicly available tweets. I created three data frames using the mostly recently tweeted tweets with the words “cow’s milk”; a ten tweet, a one hundred tweet and a one thousand tweet data frame. The sentiment of the words in the tweets were categorised as neutral, positive and negative and each category was summed to give an overall sentiment score.

*Data Visualisation: Interactive Dashboard*

Using the library “ipywidgets” an interactive dashboard was created, showing the difference in GPI of cow’s milk in different years and between different countries.

**Results**

*Descriptive Statistics*

See Figure 2 below showing the Cow’s Milk output (GPI) by Year and Country and Figure 3 for Population change over time. See Table 2 below for the mean, median and mode results of the GPI data frame

|  |
| --- |
|  |
| Figure 2. Gross Production Index of Cow’s Milk per country over time |

|  |
| --- |
|  |
| Figure 3 Population change per Country over time. |

|  |  |  |  |
| --- | --- | --- | --- |
| Table 2. Mean, Median and Mode Gross Production Index | | | |
| Country | Mean | Median | Mode |
| Ireland | 76.2398 | 81.685 | NA |
| Croatia | 99.9689 | 96.89 | NA |
| Denmark | 92.8716 | 91.77 | NA |
| Finland | 118.0743 | 109.815 | NA |
| Slovakia | 110.7475 | 113.115 | NA |
| Overall | 98.0216 | 96.89 | 81.41 |

See Table 3 below for the mean, median and mode results of the GPCPI data frame.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 3. Mean, Median and Mode Gross per Capita Production Index | | | |
| Country | Mean | Median | Mode |
| Ireland | 95.452 | 99.2 | NA |
| Croatia | 96.7717 | 94.8 | NA |
| Denmark | 101.4561 | 99.735 | NA |
| Finland | 131.037 | 120.43 | NA |
| Slovakia | 111.3592 | 113.88 | NA |
| Overall | 108.0219 | 102.78 | 96.88 |

*Inferential Statistics*

For the confidence interval results see Table 4 below.

|  |  |  |
| --- | --- | --- |
| Table 4. Confidence Intervals for GPI for the five countries of interest | | |
| Country | Mean | Margin of Error |
| Ireland | 70.6163 | 81.86332 |
| Croatia | 95.3592 | 104.5745 |
| Denmark | 91.1957 | 94.5476 |
| Finland | 112.8419 | 123.3066 |
| Slovakia | 106.2021 | 115.2928 |

T-test: The Shapiro test was conducted on the value of cow’s milk output for both Ireland and Denmark to assess the normality of the sample. The Irish test statistic was W = 0.90, and the p-value was p = 0.001 and for Denmark it was W = 0.95 and p = 0.02. Since the p-value is less than 0.05, the null hypothesis that the sample comes from a normally distributed population was rejected, in favour of the alternative hypothesis that the samples do not come from a normally distributed population. There was a significant difference between the Ireland and Denmark milk outputs (Mann-Whitney U = results of the Mann Whitney test were U = 522.00 and p = 2.011e-11), where Denmark had on average a greater milk output compared to Ireland.

ANOVA: GPI

The Shapiro test was conducted on the value of cow’s milk output for the remaining countries (see Table 5 below).

|  |  |  |
| --- | --- | --- |
| Table 5. Shapiro Wilk results for each country | | |
| Country | W | p |
| Ireland | 0.90 | 0.001 |
| Croatia | 0.90 | 0.009 |
| Denmark | 0.95 | 0.020 |
| Finland | 0.86 | 0.001 |
| Slovakia | 0.89 | 0.009 |

The Kruskal-Wallis test indicated that there is a significant difference between the milk outputs of the countries (H = 134.59, p < 0.001). Since Denmark was already confirmed to different from Ireland, the test was run again without that data. Again the Kruskal-Wallis test indicated that the significant difference between the milk outputs of the countries remained (H = 97.12, p < 0.001). See Figure 4 below.

|  |
| --- |
|  |
| Figure 4. Boxplot of milk output per country |

ANOVA: GPCPI

The Kruskal-Wallis test indicated that there is a significant difference between the milk outputs of the countries (H = 63.77, p < 0.001).

Spearman’s Correlation Testing:

There was a positive correlation between milk GPI and population for Ireland, and a negative correlation for Finland and Slovakia. There was no correlation for Croatia and Denmark. See Figure

|  |
| --- |
|  |
| Figure 4 Line graph showing relationship between milk output (GPI) and population |

Please see Table 6 below for results.

|  |  |  |
| --- | --- | --- |
| Table 6. Spearman’s Correlation Results for GPI on all five countries | | |
| Country | Spearmans correlation coefficient | p |
| Ireland | 0.810 | <0.001 |
| Croatia | -0.090 | 0.642 |
| Denmark | -0.070 | 0.595 |
| Finland | -0.948 | <0.001 |
| Slovakia | -0.767 | <0.001 |

*Machine Learning: Decision Tree*

The decision tree was successful in classifying the countries based on the variables supplied. The decision tree had an accuracy of 0.81 at a depth of five nodes. The Gini index was 0 for 8 of the 11 leaf nodes, indicating they were pure classifications. The model accuracy was checked for over and underfitting, it had a high average accuracy (0.66 ± 0.20). Thus the decision tree was deemed successful in classifying country into the correct category. Please see Figure 5 below.

|  |
| --- |
|  |
| Figure 5. Decision Tree of Country classification |

*Machine Learning: Random Forest*

GridsearchCV indicated the best indicators for the random forest classifier (max depth=10, n estimators=50, random state=100). The random forest was successful in classifying the countries based on the variables supplied. It had an accuracy of 0.93 at a depth of five nodes. The model accuracy was checked for over and underfitting, it had a high average accuracy (0.77 ± 0.24). Thus the random forest was deemed successful in classifying country into the correct category.

*Machine Learning: Time Forecasting*

This is was not successful. Please see Figure 6 and Table 7 below for Ireland’s output. Please see appendix one (past the references) for the other countries results.

|  |
| --- |
|  |
| Figure 6. Prediction for Ireland’s Milk output |

|  |
| --- |
| Table 7. ARIMA Output for Ireland |
|  |

*Machine Learning: Sentiment Analysis*

Please see Table 8 for the results of the sentiment analysis.

|  |  |  |  |
| --- | --- | --- | --- |
| Table 8. Sentiment Analysis results based off of 10, 100 and 100 tweets | | | |
| No. of Tweets | Neutral | Positive | Negative |
| 10 | 2 | 1 | 7 |
| 100 | 42 | 34 | 24 |
| 1000 | 478 | 290 | 232 |

See Figure 7 for the proportion breakdown of 1000 tweets.

|  |
| --- |
|  |
| Figure 7 Sentiment breakdown of 1000 tweets referencing cow’s milk |

**Discussion**

The cow’s milk output (both GPI and GPCPI) varied across countries and years. Ireland had the lowest mean and median of all five countries, while Finland had the highest mean, and Slovakia had the highest median, for both the GPI and GPCPI. For all countries’ confidence interval, the margin of error is greater than the mean, indicating that the confidence interval is relatively wide and there is a higher level of uncertainty about the true value of the milk output mean. Estimates of the population will not be precise using this data. However, we can also see from Figure 2 that there is a lot of variation in the milk output over time, indicating that the estimate may be a true estimate.

Despite having a similar population and climatic conditions Ireland and Denmark do not have similar means. Denmark has a relatively stable cow’s milk output over time, whereas in contrast Ireland has been increasing in cow’s milk output over time. The comparison of means using the GPI and GPCPI also showed a difference in the mean of the countries’ milk output and Ireland. From Figure 2, Ireland seems to be the only country that is (mostly) increasing their output over time, whereas Finland and Slovakia are showing a decline in output and Croatia in more recent years as well. Ireland was the only country with a positive correlation between milk output (GPI) and population, meaning that over time, as the population increased so did the milk output. In contract, both Finland and Slovakia both have a negative correlation between milk output and population. In fact, Slovakia has had a stable population but a steep decline in milk output. There was no correlation evident in Croatia or Denmark.

The first machine learning technique used was the classification method, decision tree and random forest. As expected, the random forest had higher accuracy than the decision tree, although accuracy scores were both high, 0.81 and 0.93 respectively. The random forest is the better fit model, as the random forest is made up of multiple decision trees, thus the generation of multiple decision trees reduced variation and creates a more robust result. The incorporation of the GridsearchCV hyperparameters increased the accuracy of the random forest further.

Time forecasting, I don’t know why this was unsuccessful. Each graph was consistently predicting lower values than the training data. I don’t believe this was due to overfitting with too many parameters, as there were only two parameters per model. For the Ireland data, maybe the issues occurred because the data was trending upwards and the model was predicting the trend based on the training data, but this explain why it didn’t work for the other countries. It could also be a result of seasonal trends that I have failed to capture.

The word sentiment analysis showed differences in sentiment depending on how many tweets were accounted analysed. Analysis of 10 tweets had a predominantly negative sentiment, while the 100 and 1000 tweets analysis has a majority neutral, followed by positive and then negative. The difference in sentiment between 10 and the others may be due to it being a smaller sample size and therefore not giving a true reflection of the sentiment of cow’s milk. It could also be an artifact of the time of year the analysis was conducted. January is also known as “Veganuary” where people are encouraged to go vegan and twitter is a popular forum to push this campaign. Perhaps this sentiment analysis should be conducted again in at a different time to see if the 10 tweet analysis remains negative.

**References**

Buleca, J., Kováč, V., Šubová, N., 2018. Milk Production Related to Price Of Raw Cow’s Milk in Selected European Countries. Potravinarstvo 12, 798–805.

Clancy, D., Kazukauskas, A., Newman, C., Thorne, F., 2009. An investigation of the level of structural change in the agrifood sector of Ireland, Denmark and the Netherlands. Work. Pap., Working Papers.

Krohn, C.C., Munksgaard, L., Jonasen, B., 1992. Behaviour of dairy cows kept in extensive (loose housing/pasture) or intensive (tie stall) environments I. Experimental procedure, facilities, time budgets — diurnal and seasonal conditions. Appl. Anim. Behav. Sci. 34, 37–47. https://doi.org/10.1016/S0168-1591(05)80055-3

Ladrech, R., Little, C., 2019. Drivers of political parties’ climate policy preferences: lessons from Denmark and Ireland. Environ. Polit. 28, 1017–1038. https://doi.org/10.1080/09644016.2019.1625157

Morin, J.-F., Amaral, J.S., Mafra, I., Pissard, A., Baeten, V., 2018. Milk and milk products, in: Lees, M. (Ed.), FoodIntegrity Handbook. Eurofins Analytics France, pp. 1–23. https://doi.org/10.32741/fihb.1.milk

**Appendix 1**

Figures and model outputs from the time forecasting of milk output for Croatia, Denmark, Finland and Slovakia.