**Statistics and Machine Learning Analysis on the Irish Agriculture Sector**

Author: Jefferson William Teixeira

e-mail: [sba22201@student.cct.ie](mailto:sba22201@student.cct.ie)

Student ID: sba22201

**Abstract**

*Several organizations across the world collect data on the agriculture sector, such as the United Nations and local authorities. Statistics and machine learning can be used to acquire insights from the huge amount of data available on that area. In this project, four datasets were analysed using these tools, where the main goal was to compare the Irish agriculture sector with countries worldwide. In this project, the following questions were discussed: (i) how is the organic farming growing in Europe over the years? (ii) how does Ireland compare to other countries in terms of import and export average amounts of crops and livestock? (iii) which machine learning models could predict the import and export average amounts? (iv) how does Ireland compare to other countries in terms of food price inflation? (v) which classification models could be used to perform sentiment analysis on tweets about agriculture? CRISP-DM framework was adopted to guide the analytical process from the data understanding step to the evaluation of the results. On the statistical logic part, a confidence interval was used to estimate the organic farming growth in Europe. The non-parametric statistical tests Shapiro, Levene, Kruskal, Mannwhitneyu, Wilcoxon and one parametric test, ANOVA, were used to discuss similarities between Ireland and other countries. Moreover, several experiments were executed using different machine learning classification and regression models. Naïve Bayes and Logistic Regression models were used to classify tweets based on their sentiment, while Linear Regression and KNN Regression were used to predict the import and export average amounts of agricultural goods in Ireland and similar countries. Different text processing techniques were used in the sentiment analysis such as Lemmatizer, Porter Stemmer, as well as Count and TF-IDF vectorization strategies. The models were evaluated using cross validation and compared using the appropriate scoring system. GridSearchCV was also used to obtain the best hyperparameters for the models.*

Contents

[1 Introduction 4](#_Toc123870647)

[2 Materials and Methods 4](#_Toc123870648)

[3 Data Preparation and Visualization 5](#_Toc123870649)

[4 Statistics Logic 5](#_Toc123870650)

[4.1 Confidence Interval 6](#_Toc123870651)

[4.2 Inferential Statistics Tests 6](#_Toc123870652)

[4.2.1 Crops and Livestock Products Import/Export 6](#_Toc123870653)

[4.2.2 Food Price Inflation Similarities 7](#_Toc123870654)

[4.3 Results 7](#_Toc123870655)

[5 Machine Learning 7](#_Toc123870656)

[5.1 Sentiment Analysis 7](#_Toc123870657)

[5.1.1 Naïve Bayes Model 8](#_Toc123870658)

[5.1.2 Logistic Regression 8](#_Toc123870659)

[5.2 Agriculture Import and Export Prediction 8](#_Toc123870660)

[5.3 Results 9](#_Toc123870661)

[6 Conclusions 9](#_Toc123870662)

[7 References 9](#_Toc123870663)

# Introduction

In this project, machine learning (ML) and statistical logic was applied to the Irish agriculture sector. For this purpose, the following datasets were used: Organic Farming Growth, Import/Export of Crops and Livestock Products, Food Price Inflation and Twitter’s comments about agriculture. The collected data was analysed under different perspectives and compared with other countries in Europe.

This report is organized as follows: Section 2 describes the framework, scope and tools used to complete this project. In Section 3, the datasets and the interactive dashboard are described. In Section 4, the main insights obtained from the statistical analysis are presented, while Section 5 covers the machine learning experiments. Finally, in Section 6, the conclusions and challenges faced during this project are summarized.

# Materials and Methods

In this project, the Cross Industry Standard Process (CRISP-DM) framework (Chapman *et al.*, 2000) was adapted. The advantage of using such a framework is that it can be applied to any domain, so the main tasks are known before the project starts, which contributes to the organization of the required steps. Thus, this project was divided into the following parts:

1. **Data preparation and visualization**: In this part, the exploratory data analysis (EDA) was performed on the datasets, where they were pre-processed for the statistical and ML analysis. This step was carried on separately, so the output data could be consumed by both analyses, avoiding rework and code duplication. In this step, it was also created an interactive dashboard with the main graphs that describe the information collected from the datasets.
2. **Statistical logic**: In the statistical analysis, a confidence interval was calculated to estimate the organic farming growth in Europe based on the countries data. The import/export of crops and livestock as well as the food price inflation datasets were explored using inferential statistics tests to compare Ireland’s indicators with similar countries.
3. **ML analysis**: In the ML part, it was performed sentiment analysis on a curated dataset collected from Twitter’s platform with recent user’s comments about the agriculture topic. More specifically, the sentiment of the tweets was extracted, and classification models were evaluated on these data. Moreover, forecasting analysis was performed on the crops and livestock import/export dataset, where different models were tested.

The data preparation and visualization were implemented in the accompanying Jupyter notebook called **DataPrepVis**. The statistical logic and ML analysis, which consumed the output of the EDA step, can be found in the **Statistics** and **ML** Jupyter notebooks, respectively.

The source code and files used in this project are hosted on GitHub under a public organization called [CCT-MastersDA](https://github.com/CCT-MastersDA). The project’s repository is called [cct-ca2](https://github.com/CCT-MastersDA/cct-ca2), which can be accessed with the following command in any terminal: *git clone* [*https://github.com/CCT-MastersDA/cct-ca2.git*](https://github.com/CCT-MastersDA/cct-ca2.git).

In this project, an Excel file was used to organize the tasks and project requirements due to its simplicity. This way, the *CA2-Planning.xlsx* file under the *project-mngmt* folder was used to keep track of the deliverables and mark the items as completed using checkboxes.

# Data Preparation and Visualization

Table 1 describes the datasets used in this project. Each dataset has a reference to its respective section in the **DataPrepVis** Jupyter notebook, where the EDA tasks were implemented and discussed in more detail.

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset name** | **Description** | **Source** | **EDA Section** |
| Organic Farming | Dataset that gathers the percent of utilized agricultural area (UAA) occupied by organic farming per country over the years. There are data about 34 countries from 2000 until 2020. | [Eurostat](https://ec.europa.eu/eurostat/databrowser/view/sdg_02_40/default/table?lang=en) | DataPrepVis, Section 3 |
| Crops and Livestock Products Import/Export | Dataset on the amount of import and export of crops and livestock products from several countries with data from 1961 until 2020. The main elements tracked are the import and export quantity, in tonnes, and value, in 1000 US$. | [FAO](https://www.fao.org/faostat/en/#data/QCL) | DataPrepVis, Section 4 |
| Food Price Inflation | Dataset on the food price inflation from several countries with data from 2001 until 2022. | [FAO](https://www.fao.org/faostat/en/#data/CP) | DataPrepVis, Section 5 |
| Twitter Dataset | This dataset was collected directly from the Twitter’s platform using their [Developer’s API](https://developer.twitter.com/). | Custom solution. | DataPrepVis, Section 6 |

Table 1- Agriculture Datasets

Due to the large amount of data available, the countries comparisons were limited to Irland, Finland and Slovakia. The reason for that choice was the size of their populations, which, according to Eurostat, is around 5M people (Eurostat, 2022).

In this project, it was created a dashboard called **DataPrepVisDashboard** using Voila tool containing the graphs created in the EDA steps. The HTML version of the dashboard was saved in the project’s repository, which contains all the static and interactive graphs created to visualize the datasets described in this section. The dashboard files are located under the *jupyter/dashboard* folder.

# Statistics Logic

The analysis described in this section was implemented and discussed in more detail in the accompanying Jupyter notebook called **Statistics**.

In section 4.1, it is presented how a confidence interval was calculated to estimate the organic farming growth in Europe. In section 4.2, the import/export of agricultural products and the food price inflation indicators were compared among the selected countries.

## Confidence Interval

The question being solved in this section is: How is the organic farming growing in Europe over the years? This analysis was implemented in the Section 3 of the **Statistics** Jupyter notebook.

Using the Organic Farming dataset, it was calculated the standard deviation and the average of the percent of land under organic farming in Europe based on the available countries and years. This way, it was possible to calculate the confidence interval to estimate the range in which the real percent of land under organic farming lies over the years in Europe with a specified level of confidence.

Table 1 shows some characteristics of the dataset and the parameters used to calculate the confidence interval. The variable being counted is the percent UAA under organic farming because the dataset is provided in terms of percent of land. The dataset only has data ranging from 2000 to 2020. There are 34 countries in the dataset, but not all of them have data for every year, so the maximum degree of freedom in this study was 34, as the null values were removed from the calculation. Since the standard deviation is unknown for this analysis, the distribution used to calculate the interval was T-Student. The alpha parameter means the level of confidence used for the calculation, which was 97%.

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| Variable | X = Percent of UAA under organic farming |
| Distribution | T-Student |
| Alpha | 0.03 |
| 1 - alpha | 0.97 |
| Years range | [2000, 2020] |
| Degree of freedom (n) | <=34 |

Table 2- Confidence Interval Parameters and Dataset Details

Figure 1 shows the results of the confidence interval calculated for the growth of organic farming in Europe as compared to the actual data from Ireland over the years.

Based on these results, it is possible to say that, with 97% confidence, the percent of UAA under organic farming in Europe in 2020 was between 6.14% and 11.18%.

According to the [Agridata](https://agridata.ec.europa.eu/extensions/CountryFactsheets/CountryFactsheets.html?memberstate=Ireland) website, in 2020, the percent of land under organic farming in Europe was about 9%, which shows that the confidence interval obtained by this method was accurate.

Chart, line chart

Description automatically generated

Figure 1 - Organic Farming Growth in Europe and Ireland

According to the [Agridata](https://agridata.ec.europa.eu/extensions/CountryFactsheets/CountryFactsheets.html?memberstate=Ireland) website, in 2020, the percent of land under organic farming in Europe was about 9%, which shows that the confidence interval obtained by this method was accurate.

## Inferential Statistics Tests

In this section, inferential statistics tests were used to compare the countries indicators on agriculture with a certain level of confidence. Thus, in section 4.2.1, Ireland, Finland and Slovakia are compared in terms of their average import/export amount, whereas in section 4.2.2, the same countries are compared on their food price inflation data.

### Crops and Livestock Products Import/Export

The question being solved in this section is: How does Ireland compare to other countries in terms of import and export average amounts of crops and livestock products? This analysis was implemented in the Section 4.1 of the **Statistics** Jupyter notebook.

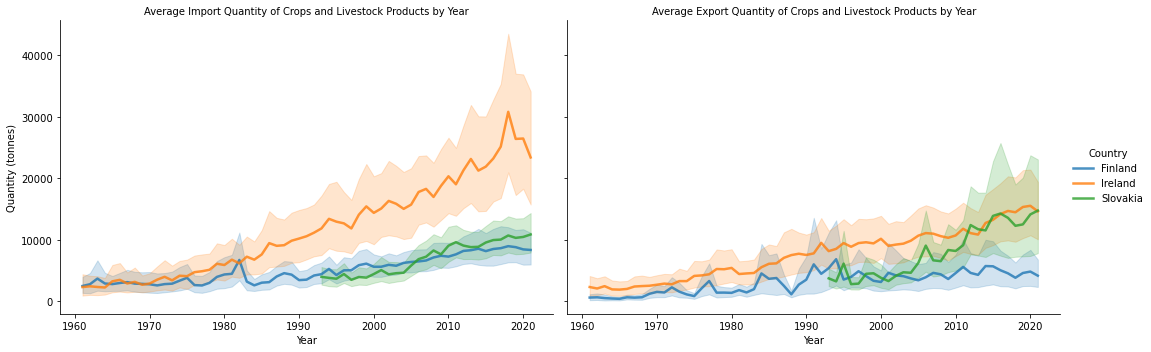


Figure 2 - Average Import/Export Quantity by Year

A picture containing chart

Description automatically generated

Figure 3 - Average Import/Export Value by Year

Figure 2 and Figure 3 give an idea of how the average import and export amounts per year behave for each country. Based on the actual data illustrated by these graphs Ireland's import and export amounts are higher than the other countries, mainly in terms of export values in 1000 US$ (Graph 4). Regarding Slovakia and Finland, they import about the same quantity and values, having slightly different behaviour with regards to the export numbers, where Slovakia shows more exports than Finland.

### Food Price Inflation Similarities

The question being solved in this section is: How does Ireland compare to other countries in terms of food price inflation? This analysis was implemented in the Section 4.2 of the **Statistics** Jupyter notebook.

## Results

# Machine Learning

The analysis described in this section was implemented and discussed in more detail in the accompanying Jupyter notebook called **ML**.

In section 5.1, it is discussed the sentiment analysis performed on the tweet’s dataset about agriculture and food prices. In section 5.2, the ML models used to predict the import and export of agriculture products in Ireland and other countries are presented.

## Sentiment Analysis

The sentiment analysis described in this section focuses on the classification of tweets based on the user’s comments about any topic related to agriculture into negative, neutral, or positive categories. This analysis was implemented in the Section 3 of the **ML** Jupyter notebook.

The tweets dataset cannot be fed directly to the ML algorithms themselves as most of them expect numerical feature vectors with a fixed size rather than the raw text documents with variable length (Scikit-learn, 2022). As a result, The TF-IDF and Count vectorizers techniques were used to extract feature vectors from the tweets, generating the independent variables used for the classification. In this case, the target variable was the sentiment.

The feature vector extraction could be done directly on the raw tweets, however, special characters and multimedia data could interfere in the results, this way, as described in the Section 6 of the **DataPrepVis** Jupyter notebook, four versions of the tweets were generated from the raw text: (1) cleaned with stop words, (2) cleaned without stop words, (3) *lemmatized,* and (4) *stemmerized* tweets. Stop words do not add much information to the text, so their frequency could bias the models, therefore, they are usually removed from the dataset. *Lemmatization* and *Porter Stemmer* are common ways to extract the core meaning from the words (Scikit-learn, 2022), this way these techniques were applied to the datasets to evaluate the performance of the classifiers.

This dataset was collected from a live platform, so it was likely to be unbalanced in terms of the distribution of the sentiment classes, which was confirmed during the experiments as shown in Section 3.1.2 of the **ML** Jupyter notebook. Therefore, it was used a technique called SMOTE to oversample the dataset, turning it into a balanced dataset, because it is known that unbalanced data can led to a poor classification performance.

Every feature in the vectorized tweets can be treated as independent and makes equal contribution to the result, this way the Naïve Bayes (NB) model was used. On the other hand, Logistic Regression (LGR) algorithm was also tested due to its efficiency in predicting classes based on the features relationships. Sections 5.1.1 and 5.1.2 describe the experiments using NB and LGR models respectively.

### Naïve Bayes Model

In this experiment, the four versions of the tweets were tested using TF-IDF and Count vectorizer. The model run with cross validation technique in which 10 folds were generated from the dataset that was split into train and test data. The average accuracy was calculated for each model so the results could be evaluated. This analysis was implemented in the Section 3.2 of the **ML** Jupyter notebook.

### Logistic Regression

In this experiment, GridSearchCV was applied to decide the best parameters to run the LGR model. In this test, TF-IDF option was used, as it provided the best results in the experiment discussed in section 5.1.1.

Thus, the main parameters tested were related to the vectorizer strategy, that is, keep or not the stop words and which tokenizer provides better results when applied to one of the three versions of the tweets (1), (3) and (4), as defined in section 5.1. Version (2) was not tested, because using stop words, or not was part of the test. This way, the result given by GridSearchCV could be evaluated. This analysis was implemented in the Section 3.3 of the **ML** Jupyter notebook.

## Agriculture Import and Export Prediction

This analysis focuses on applying ML models to make predictions about the average import and export amounts in Ireland and Finland. This analysis was implemented in the Section 4 of the **ML** Jupyter notebook.

To run the ML models, the dataset was split and converted into a timeseries. Thus, a dummy time index column was created to replace the years for each of the following data per country: import-quantity, import-value, export-quantity, and export-value. The quantity and value import/export amounts were also scaled in their respective dataset to normalize the data. Therefore, eight datasets were created, i.e., Import-Quantity Ireland, Import-Value Finland, etc. The following graphs were generated to illustrate the datasets after that pre-processing step.

The prediction models were created using supervised regression algorithms, where the independent variable is the time, whereas the target is the import/export average amount. Therefore, based on the nature of the data, which is numerical and continuous, and the problem being solved, which is a prediction problem, the following approaches were used: Polynomial Regression and KNN Regression algorithms.

Each model was executed manually with different predefined parameters, and their performance was calculated using cross validation. The dataset was split into train and test dataset, but due to the small size of the datasets, 40% of the data was used for test against 60% for train. For each model, GridSearchCV with K Fold technique were executed to retrieve the best parameters for the models. The results were compared based on a score function given by the coefficient of determination (R2). This score provides an indication of goodness of fit and therefore a measure of how well unseen samples are likely to be predicted by the model (Scikit-learn, 2022).

## Results

# Conclusions

# References

Chapman, P. *et al.* (2000) *CRISP-DM 1.0 Step-by-step data mining guide*. Available at: http://www.crisp-dm.org/CRISPWP-0800.pdf.

Eurostat (2022) *Facts and figures on life in the European Union*, *Eurostat*. Available at: https://european-union.europa.eu/principles-countries-history/key-facts-and-figures/life-eu\_en (Accessed: 24 December 2022).

Scikit-learn (2022) *Scikit-learn: Machine Learning in Python*, *https://scikit-learn.org*.