**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, the European Union, and local authorities. Tools such as statistics and machine learning can be used to acquire insights from the huge amount of data available, which is key for the success of the sector in today’s economy. In this project four datasets about different domains on the agricultural topic were analysed using statistical logic and machine learning. 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? In this project, the framework CRISP-DM was adopted to guide the analytical process from the data understanding step to the evaluation of the modelling results. On the statistical logic part, a confidence interval was used to estimate the organic farming 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](#_Toc123846365)

[2 Materials and Methods 4](#_Toc123846366)

[2.1 Project Management Framework 4](#_Toc123846367)

[2.2 Project Scope and Organization 5](#_Toc123846368)

[2.3 Version Control and Task Management 5](#_Toc123846369)

[3 Data Preparation and Visualization 6](#_Toc123846370)

[3.1 Organic Farming Area 6](#_Toc123846371)

[3.2 Crops and Livestock Products Import/Export 6](#_Toc123846372)

[3.3 Food Price Inflation 6](#_Toc123846373)

[3.4 Tweets Dataset 6](#_Toc123846374)

[3.5 Interactive Dashboard 6](#_Toc123846375)

[4 Statistics Logic 6](#_Toc123846376)

[4.1 Confidence Interval 6](#_Toc123846377)

[4.2 Inferential Statistics Tests 7](#_Toc123846378)

[4.2.1 Crops and Livestock Products Import/Export 7](#_Toc123846379)

[4.2.2 Food Price Inflation Similarities 7](#_Toc123846380)

[4.3 Results 7](#_Toc123846381)

[5 Machine Learning 7](#_Toc123846382)

[5.1 Sentiment Classification 8](#_Toc123846383)

[5.1.1 Naïve Bayes Model 8](#_Toc123846384)

[5.1.2 Logistic Regression 8](#_Toc123846385)

[5.2 Agriculture Import and Export Prediction 9](#_Toc123846386)

[5.3 Results 9](#_Toc123846387)

[6 Conclusions 9](#_Toc123846388)

[7 References 9](#_Toc123846389)

# Introduction

In this project, machine learning (ML) and statistical logic was applied to the Irish agriculture sector data. The sector was analysed under different perspectives and compared with other countries in Europe, such as, Slovakia and Finland. According to Eurostat, these countries have about the same number of habitants, around 5M people (Eurostat, 2022). Thus, only these countries were compared for simplicity and to limit the amount of data to be processed.

This report is organized as follows: Section 2 details the framework, scope and tools used to manage and complete this project. In Section 3, the data preparation tasks are described as well as the interactive dashboard to showcase the information gathered from the datasets. 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, four agricultural datasets were used, that is, Organic Farming Growth, Import/Export of Crops and Livestock, Food Price Inflation and Twitter’s comments about agriculture. Thus, it was very important to use a framework to study and model the data for the different types of analysis proposed. This project also used a version control system and a task management strategy.

Section 2.1 describes the adopted framework, section 2.2 details the scope of the project, while section 2.3 explains the version control system and task management.

## Project Management Framework

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 and definition of the required steps. CRISP-DM suggests the following overall tasks and what they mean in this project:

* **Data understanding:** In this step, the different datasets were studied to understand its formatting and constraints as well as to decide the preparation steps for the analysis undertaken.
* **Data preparation:** In this step, the raw datasets were collected from the source and passed as input to the code that was implemented to prepare the data for the statistical and machine learning analysis.
* **Modelling:** This step was divided into the statistical logic and ML modelling steps, which were applied to the different datasets used in this project. For the statistical part, the data were modelled for the inferential statistics techniques, while in the ML study, the data were prepared for the classification and regression models.
* **Evaluation:** In the statistical part, this step means understanding and stating the conclusions of the inferential tests that were applied to the data. In the ML study, it involves assessing the results and performance of the models that were executed as well as comparing the different strategies results.

## Project Scope and Organization

Based on the framework presented in the previous section, this project was divided into the following parts:

1. **Data preparation and visualization**: In the data preparation and visualization part, the four datasets were pre-processed for the statistical and ML analysis. This step was carried on separately because some datasets were used in both analyses, avoiding 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 found 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.

Thus, the data preparation and visualization were implemented in the **DataPrepVis** accompanying Jupyter notebook. The modelling and evaluation of the results for the statistical logic and ML can be found in the **Statistics** and **ML** Jupyter notebooks, respectively.

## Version Control and Task Management

This project was hosted on a GitHub server, a cloud provider that allows users to have several Git repositories. Users can also keep different organizations under the same account. This way, it was created a public organization called [CCT-MastersDA](https://github.com/CCT-MastersDA) with a repository named [cct-ca2](https://github.com/CCT-MastersDA/cct-ca2), where this project is versioned. It 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

## Organic Farming Area

[Eurostat](https://ec.europa.eu/eurostat/databrowser/view/sdg_02_40/default/table?lang=en) provides a dataset that gathers the percent of utilized agricultural area (UAA) occupied by organic farming (Eurostat, 2022) per country over the years. There are data about 34 countries from 2000 until 2020. The steps performed to retrieve and process this dataset are detailed in Section 3 of the **DataPrepVis** accompanying Jupyter notebook.

## Crops and Livestock Products Import/Export

[Food and Agriculture Organization of the United Nations](https://www.fao.org/faostat/en/#data/QCL) (FAO) provides data on the amount of import and export of crops and livestock products from several countries with data ranging from 1961 up to 2020. The main elements tracked are the import and export quantity, in tonnes, and value, in 1000 US$. The steps performed to retrieve and process this dataset are detailed in Section 4 of the **DataPrepVis** accompanying Jupyter notebook.

## Food Price Inflation

[FAO](https://www.fao.org/faostat/en/#data/CP) also provides data on the food price inflation from several countries with data from 2001 up to 2022. The steps performed to retrieve and process this dataset are detailed in Section 5 of the **DataPrepVis** accompanying Jupyter notebook.

## Tweets Dataset

This dataset was collected directly from the Twitter’s platform using their [Developer’s API](https://developer.twitter.com/). The steps performed to retrieve and process this dataset are detailed in Section 6 of the **DataPrepVis** accompanying Jupyter notebook.

## Interactive Dashboard

The dashboard solution used was [Voila](https://voila.readthedocs.io/en/stable/index.html), which is a tool that allows users to convert Jupyter notebooks into interactive dashboards. There are several tools for this purpose available. This one was picked for its simplicity and easy installation process and use.

Voila is installed via any supported package manager, such as pip or Conda. There are two ways to generate the dashboard, by using the Voila Jupyter extension, which requires one click to run, and by running the following command on a terminal: *voila jupyter\_file*.

Every Jupyter notebook accompanying this project can be converted into a dashboard using Voila. The outcome of this tool is a web page that is generated with the dashboard, which can run locally or be deployed to a web server for broader audience. For demonstration purposes, a dashboard was generated from the DataPrepVis Jupyter, called **DataPrepVisDashboard**, which contains the static and interactive graphs created to visualize all the datasets.

# Statistics Logic

In section 4.1, the Organic Farming Area dataset was used to find a confidence interval for the average area under organic farming in Europe. In section 4.2, the Crops and Livestock Products Import/Export and the Food Price Inflation datasets were checked against some inferential statistics tests to find similarities between the selected countries.

## Confidence Interval

The dataset used for this analysis was the Organic Farming Area, which is described in section 3.1. Thus, the main question being solved in this section is:

*How is the organic farming growing in Europe over the years?*

For each year, it was found the standard deviation and average of the percent of land under organic farming in all European countries available in the dataset. This way, it was possible to calculate the confidence interval to estimate the range in which the average percent of land under organic farming in the whole Europe lies over the years with a specified level of confidence.

Table 1 shows some characteristics of the dataset and the parameters used to calculate the confidence interval for the organic farming. The variable being calculated 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 1- Confidence Interval Parameters and Dataset Details

## Inferential Statistics Tests

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

### Crops and Livestock Products Import/Export

The dataset used for this analysis is described in section 3.2. Thus, the main 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?*

### Food Price Inflation Similarities

The dataset used for this analysis is described in section 3.3. Thus, the main question being solved in this section is:

*How does Ireland compare to other countries in terms of food price inflation?*

## Results

# Machine Learning

In section 5.1, it was performed a sentiment analysis on the tweet’s dataset about agriculture and food prices. In section 5.2, ML models were used to predict the import and export of agriculture products in Ireland and other countries using the Crops and Livestock Products Import/Export dataset.

## Sentiment Classification

The dataset used for this analysis is described in section 3.4. Thus, the study shown in this section focuses on the classification of tweets based on the user’s comments on the agriculture topic into negative, neutral, or positive categories.

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 is 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 section 3.4, 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, this way the performance of the classifiers on the dataset that uses these techniques were also evaluated.

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. 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. The average accuracy was calculated from each test execution so the results could be evaluated.

### 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 best results in the previous test. Thus, the main parameters tested were related to the vectorizer strategy, that is, keep or not the stop words and which tokenizer provides best results when applied to one of the three versions of the tweets (1), (3) and (4). 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.

## Agriculture Import and Export Prediction

The dataset used for this analysis is described in section 3.2. Thus, this part focuses on applying ML models to make predictions about the average import and export amount in Ireland and Finland.

To apply the ML prediction models to this dataset it had to be split and converted into a timeseries. Thus, a dummy time index column was created to replace the years for each of the following amounts per country (Ireland and Finland): 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. Moreover, for each experiment, the dataset was split into train and test dataset, but due to the small size of the datasets, it was used 40% of the data for test and 60% for train. For each model it was also executed GridSearchCV with K Fold to retrieve the best parameters to be compared with the results given by the manual executions. 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*.