**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 agricultural 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 Ireland agricultural 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 over the years? (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 and vectorization. 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](#_Toc123838019)

[2 Materials and Methods 4](#_Toc123838020)

[2.1 Project Management Framework 5](#_Toc123838021)

[2.2 Version Control 5](#_Toc123838022)

[2.3 Tasks Management 6](#_Toc123838023)

[3 Data Preparation and Visualization 6](#_Toc123838024)

[3.1 Organic Farming Area 6](#_Toc123838025)

[3.2 Crops and Livestock Products Import/Export 6](#_Toc123838026)

[3.3 Food Price Inflation 6](#_Toc123838027)

[3.4 Tweets Dataset 6](#_Toc123838028)

[3.5 Interactive Dashboard 6](#_Toc123838029)

[4 Statistics Logic 6](#_Toc123838030)

[4.1 Confidence Interval 6](#_Toc123838031)

[4.2 Inferential Statistics Tests 7](#_Toc123838032)

[4.2.1 Crops and Livestock Products Import/Export 7](#_Toc123838033)

[4.2.2 Food Price Inflation Similarities 7](#_Toc123838034)

[4.3 Results 7](#_Toc123838035)

[5 Machine Learning 7](#_Toc123838036)

[5.1 Sentiment Classification 8](#_Toc123838037)

[5.1.1 Naïve Bayes Model 8](#_Toc123838038)

[5.1.2 Logistic Regression 8](#_Toc123838039)

[5.2 Agriculture Import and Export Prediction 9](#_Toc123838040)

[5.3 Results 9](#_Toc123838041)

[6 Conclusions 9](#_Toc123838042)

[7 References 9](#_Toc123838043)

# Introduction

In this project, four agricultural datasets were analysed under different perspectives using Statistical Logic and Machine Learning (ML), that is, Organic farming growth, import/export of crops and livestock, food price inflation and Twitter’s comments about agriculture.

The main subject of the study was the Irish Agriculture sector, whose data were analysed and compared against other similar countries in Europe, such as, Slovakia and Finland. According to Eurostat (Eurostat, 2022), these countries have about the same number of habitants, around 5M people. Thus, for simplicity and to limit the amount of data to be processed, only these countries data were used in the comparison type of analysis.

This project is divided into three parts: (1) data preparation and visualization, (2) statistical logic and (3) machine learning analysis, which are summarized as follows:

1. 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. 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. 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.

This report is organized as follows: Section 2 details the general framework 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 section it is described the framework adopted to execute this project and to perform the analysis on the different chosen datasets. This way, section 2.1 summarizes the steps of the adopted project management framework, while section 2.2 explains the version control system used. Finally, section 2.3 describes how the requirements of the project were managed.

## Project Management Framework

This project used four datasets covering distinct agricultural domains as outlined in Section 1. Thus, it was very important to use a framework to analyse and model the data for the different types of study proposed in this project.

The main advantage of following 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 a better organization and definition of the steps in a systematic fashion.

This way, the Cross Industry Standard Process (CRISP-DM) (Chapman *et al.*, 2000) was used as a reference framework, which suggests the following steps:

* **Data understanding:** In this step, the different datasets were studied to understand its formatting and constraints as well as to decide which columns or rows could be used in 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 process and 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, this step means assessing the results and performance of the models that were executed as well as comparing the different strategies results.

The application of each step of the framework is described in the next sections in the context of the analysis being discussed.

## Version Control

Version control systems are largely used in industry as they are powerful tools to keep track of the changes in the history of documents, such as source code, providing ways to revert modifications, detect and resolve conflicts while keeping the repository organized. There are several systems in the market, however the main version control system nowadays is Git (Git, 2022).

This project was hosted on GitHub server, which is basically a cloud provider that stores and allows users to manage different Git repositories. Users can also keep different organizations under the same GitHub account where they organize their repositories.

For this project, 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 the entire project is versioned. It basically contains the source code, datasets, and this report. Anyone with a GitHub account can access this project’s repository by executing the following command in any terminal:

git clone https://github.com/CCT-MastersDA/cct-ca2.git

## Tasks Management

Although there are several more efficient ways to manage requirements and tasks during a project, for the sake of simplicity, in this project an Excel file was used for that purpose. Thus, the *CA2-Planning.xlsx* file under the *project-mngmt* folder contains the requirements and tasks that needed to be addressed for a successful completion of this project. This way, it was possible to keep track of the deliverables and mark the items as completed using checkboxes.

# Data Preparation and Visualization

## Organic Farming Area

## Crops and Livestock Products Import/Export

## Food Price Inflation

## Tweets Dataset

## Interactive Dashboard

# 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 (LogR) algorithm was also tested due to its efficiency in predicting classes based on the features relationships. The following sections describe the experiment using NB and LogR 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 LogR 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.

## Agriculture Import and Export Prediction

- Linear Regression

- KNN Regression

The following techniques were also applied K-Fold, GridSearchCV and Cross-Validation.

Each model was executed, and 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).

Polynomial and KNN regressions were applied to all vehicles so the performance of the models can be seen across the board. The dataset with best score results were then tested against Ridge, so the two linear regression types of algorithms could be compared.

The scores were obtained using two approaches: Manual and cross-validation. In both approaches the dataset was split into train (70%) and test (30%) data, the difference is that in the first approach, the scores were obtained manually for each execution of the model with the specific parameters set, while in the second approach a K-Fold cross-validation technique was used. K-Fold basically splits the training set into smaller sets, where the final score is the average of the scores obtained with different train sub-sets given by the K-Fold splitting (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).

Git (2022) ‘Git’. Available at: https://git-scm.com/ (Accessed: 5 January 2023).

Guido, S. and Müller, A.C. (2016) *Introduction to Machine Learning with Python*. Edited by A.C. Müller and S. Guido. O’Reilly Media.

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