**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. Knowing the data and acquiring insights from them using tools such as statistics and modern technology are essential for the success of the sector in today’s economy. In this project four datasets that contain facts about some domain on the agricultural topic were analysed using Statistical Logic and Machine Learning techniques. The goal was to compare Ireland agricultural sector with countries worldwide. Examples of questions discussed in this project are: (i) how is the organic farming growing in Europe over the years? (ii) is it possible to compare the Irish import and export amounts of crops and livestock with similar population countries such as Finland and Slovakia? (iii) which Machine Learning models could predict the import and export evolution over the years in those countries? (iv) how is the food price inflation on these countries, is there enough evidence to say they are similar or not? (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, it was applied confidence interval to estimate the organic farming in Europe, and the following non-parametric statistical tests: Shapiro, Levene, Kruskal, Mannwhitneyu, Wilcoxon and one parametric test, ANOVA, 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 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 different 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](#_Toc123781217)

[2 Materials and Methods 4](#_Toc123781218)

[2.1 Project Management Framework 5](#_Toc123781219)

[2.2 Version Control 5](#_Toc123781220)

[2.3 Tasks Management 6](#_Toc123781221)

[3 Data Preparation and Visualization 6](#_Toc123781222)

[3.1 Organic Farming Area 6](#_Toc123781223)

[3.2 Crops and Livestock Products 6](#_Toc123781224)

[3.3 Food Price Inflation 6](#_Toc123781225)

[3.4 Tweets Dataset 6](#_Toc123781226)

[3.5 Interactive Dashboard 6](#_Toc123781227)

[4 Statistics Logic 6](#_Toc123781228)

[4.1 Confidence Interval 6](#_Toc123781229)

[4.2 Inferential Statistics Tests 6](#_Toc123781230)

[4.3 Results 6](#_Toc123781231)

[5 Machine Learning 6](#_Toc123781232)

[5.1 Experiment Setup 7](#_Toc123781233)

[5.2 Sentiment Classification 7](#_Toc123781234)

[5.3 Agriculture Import and Export Prediction 8](#_Toc123781235)

[5.4 Results 8](#_Toc123781236)

[6 Conclusions 8](#_Toc123781237)

[7 References 8](#_Toc123781238)

# 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 compared against other similar countries in the EU, such as, Slovakia and Finland. The similarity criteria to choose these countries was the size of their population, which is around 5M people according to Eurostat (Eurostat, 2022).

The main goal of this project is to acquire insights from the selected datasets comparing the Irish agricultural sector with other countries. Thus, this project was 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 calculated, and classification models were evaluated with 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 to be used by 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 changes in the history of documents and source code. It is used mainly in collaborative environments, where different people work on the same source, providing ways to revert changes, detect and resolve conflicts while keeping the repository organized. There are several systems with different technologies in the market, however the main version control system nowadays is Git.

Git is a decentralized version control system in which there is not the concept of a single source of truth of a repository. Alternatively, every collaborator has its own copy of the entire repository, where the strategy to merge the changes depends on the organization.

This project was hosted on GitHub server, which is basically a cloud provider that hosts and allows users to manage different Git repositories. GitHub also supports users to keep different organizations under the same account where they manage one or more Git 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

## Food Price Inflation

## Tweets Dataset

## Interactive Dashboard

# Statistics Logic

## Confidence Interval

## Inferential Statistics Tests

## Results

# Machine Learning

Machine learning is about extracting knowledge from the data (Guido and Müller, 2016). The main goal is to enable computers to identify patters in the data in other to extract relevant information or insights that can be used in many applications.

The main algorithms in machine learning fall into two categories: Supervised and Unsupervised (Guido and Müller, 2016). In the first one, the computer is trained with labelled data, while in the second one, the data has no label and the technique applied must identify the relationships in the data. There are two types of supervised problems: Classification and Regression (Guido and Müller, 2016). While the first one is concerned with predicting labels, the second is commonly used to predict continuous numbers. The right technique depends on the problem being solved and the characteristics of the dataset.

For this purpose, the ML models were validated based on the K-Fold cross-validation technique. The ML models were also improved using a hyperparameter tunning technique called GridSearchCV.

## Experiment Setup

The following algorithms were tested against the traffic dataset: Polynomial, KNN and Ridge regression. The steps and the code executed for this part of the project can be seen in the *Machine Learning Models* section of the Jupyter notebook.

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).

Section 3.7.2 discusses the linear regression models based on the results of manual and cross-validation scores. Section 3.7.3 presents the results of KNN model using the manual validation approach. Finally, Section 4 presents the overall results obtained using a technique called GridSearchCV.

## Sentiment Classification

Linear models are the simplest and most classic models for regression problems (Guido and Müller, 2016). They are widely adopted to make value prediction using a linear function of the input variables or features of the dataset.

Classic linear regression models find the function parameters that minimize the mean squared error between predictions and the true regression targets defined on the training set. The mean squared error is the sum of the squared differences between the predictions and the true values, divided by the number of samples. Polynomial regression is a generalization of this technique that tries to fit a N degree polynomial function to the dataset.

Ridge regression uses the same formula as Linear Regression to make predictions. However, this model tries to reduce the influence of the features in the outcome to avoid overfitting, a side-effect that occurs when the model becomes too complex with several parameters. The Ridge model makes a trade-off between the simplicity of the model (near-zero coefficients) and its performance on the training set (Guido and Müller, 2016). This mechanism is controlled by an alpha parameter.

## Agriculture Import and Export Prediction

KNN regression is an adapted version of the classic KNN algorithm commonly used for classification problems (Guido and Müller, 2016). In the regression version, it uses the true target values points to predict the value of the new entries.

The parameter of this model is the number of neighbours (K) it uses to make the predictions, which is calculated based on the average of the neighbours’ target values.

### 

## 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).

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) *Metrics and scoring: quantifying the quality of predictions*, [*https://scikit-learn.org*](https://scikit-learn.org).