**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**

*Every year Dublin City Council conducts traffic counts at various locations on entry points around Dublin city centre. The locations are chosen to ensure that any person entering the City Centre from outside must pass through one of the locations where the surveys are undertaken. These traffic counts provide a reliable measurement of the modal distribution of persons travelling into, and out of, Dublin city on a year-on-year comparable basis. The data collected is divided into the transport modes allowing a better understanding of the changing usage trends in pedestrian and various vehicle types. In this work, statistics models and machine learning techniques have been applied to extract descriptive measurements and appropriate probability distributions from the data and to make predictions for the future using the historical data. Understanding and predicting the traffic behaviour is beneficial to the city planning as it indicates how the public resources can be applied more efficiently in terms of road infrastructure and public transportation, for example. More specifically, the inbound traffic data of different transportation modes that were collected in Conyngham Road, Dublin have been studied. The machine learning framework CRISP-DM was adopted to guide the analytical process from the data understanding step to the evaluation of the modelling results. Several experiments were performed using different machine learning models and the technique with the best results was KNN Regression algorithm, having a R2 score of 81% using four neighbours as detected by GridSearchCV technique. On the other hand, the models with the lowest performance were Polynomial Regression, with a score of 38% and Ridge Regression, with 35% score, both having their hyperparameters improved by GridSearchCV.*

Contents

[1 Introduction 4](#_Toc122790040)

[2 Materials and Methods 4](#_Toc122790041)

[2.1 Project Management Framework 4](#_Toc122790042)

[2.2 Version Control 5](#_Toc122790043)

[3 Statistics Logic 5](#_Toc122790044)

[3.1 Data Preparation and Visualization 6](#_Toc122790045)

[3.2 Confidence Interval 6](#_Toc122790046)

[3.3 Inferential Statistics Tests 6](#_Toc122790047)

[3.4 Results 6](#_Toc122790048)

[4 Machine Learning 6](#_Toc122790049)

[4.1 Data Preparation and Visualization 6](#_Toc122790050)

[4.2 Experiment Setup 6](#_Toc122790051)

[4.3 Model 1 7](#_Toc122790052)

[4.4 Model 2 7](#_Toc122790053)

[4.5 Results 7](#_Toc122790054)

[5 Interactive Dashboard 7](#_Toc122790055)

[6 Conclusions 7](#_Toc122790056)

[7 References 7](#_Toc122790057)

# Introduction

Several organizations across the world collect data on the agricultural sector, such as the United Nations, the European Union, local governments, etc. 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, the following agricultural domains are analysed under different perspectives using Statistical Logic and Machine Learning (ML): Organic farming, import/export of food and food price inflation. The main subject of the study is the Irish Agriculture sector, whose data are compared against other similar countries in the EU, such as, Slovakia and Finland. The similarity criteria to choose these countries were the size of their population, which is around 5M people according to Eurostat (Eurostat, 2022).

This report is organized as follows: Section 2 explains the scope of the project and how it was managed until its completion. In Section 3, the statistical study is presented with the main insights obtained from the data. Section 4 covers the machine learning techniques that were applied to the agricultural data as well as the sentiment analysis on the same topic. Section 5 presents the interactive dashboard that was implemented to showcase the insights obtained in this project. Finally, in Section 6, the conclusions and challenges faced during this project are summarized.

# Materials and Methods

The main goal of this project is to obtain insights from the selected agricultural data in the context of Ireland and EU. Thus, this projected is divided into two parts (1) statistical and (2) machine learning analysis, which are summarized as follows:

1. In the statistical analysis, it was explored inferential statistics to gain insights about the population based on the sample data provided in the selected dataset as well as the application of inferential statistics tests to compare the agricultural data between Ireland and similar countries. The techniques, steps and results of this analysis are detailed in Section 3.
2. In the ML analysis, it was explored… The techniques, steps and results of this analysis are detailed in Section 4.

The rest of this section is focused on the general approach adopted to execute the project and perform the analysis on the different datasets chosen. This way, section 2.1 summarizes the steps of the project management framework that was adopted, while section 2.2 explains the version control system used. Finally, section 2.3 briefly describes the tasks management used.

## Project Management Framework

This project used three datasets covering distinct agricultural domains as outlined in Section 1. Thus, it was very important to adopt a framework to analyse and model the data for the two different types of study performed in this project. The main reason behind it were for a better organization and definition of the steps required to complete the project 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, as illustrated in the Figure 1.

* **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. The types of processing in this step include handling of null values, removal of unnecessary rows, normalization of data, etc.
* **Modelling:** This step was divided into the statistical logic and ML modelling steps, which were applied to the different datasets used in this project. On the statistical part, the appropriate inferential statistics techniques were applied to the data to gain insights about the population based on the sample and to compare the data between different countries. Regarding the ML study, the supervised regression learning models were applied as well as sentiment analysis.
* **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.

Diagram

Description automatically generated

Figure 1- CRISP-DM Framework.

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

## 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 tools to revert changes, detect and resolve conflicts and keep 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 for a repository, in which every copy should synchronize. Alternatively, every collaborator has its own copy of the entire repository, where the strategy to merge the changes will depend on the organization’s choice.

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 have different organizations under the same account where they can keep different 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 contains basically the source code, i.e., Jupyter files, 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 tasks in a project, it was used an Excel file with all the topics required in the project’s brief for simplicity. This way, it was possible to keep track of the deliverables and mark the items as completed. The file with the full planning and checkboxes is located under the repository’s folder *project-mngmt.*

# Statistics Logic

## Data Preparation and Visualization

## 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.

## Data Preparation and Visualization

## 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.

## Model 1

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.

## Model 2

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

# Interactive Dashboard

# Conclusions

# References