




GROUP SEPT 2022 - SB+ - 2022 - YR1 - MSc IN DATA ANALYTICS CA2

Author: N. Pereira Linares  
e-mail: sba22223@student.cct.ie  
Student ID: sba22223



---

Module Title:	MSc in Data Analytics - Sept 2022 - SB+ - 2022 - YR1
Assessment Title:	Multivariable forecasting Ireland data Eurostat
Lecturer Name:	
Student Full Name:	Nestor, PEREIRA LINARES
Student Number:	sba22223
Assessment Due Date:	Jan 6 <sup>TH</sup> 2023
Date of Submission:	Jan 6 <sup>th</sup> 2023

---

#### Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced.

I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

## Abstract

*EU farm policy was introduced from the beginning of the creation of the EU and according to serious estimations by 2050 we need to produce double food production to support the development of our society.*

*Now, currently, how to see Ireland with respect to its neighbours and partners in the EU community?*

*In order to respond to this question, it proposes in this project uses the EU Agricultural Price Indexes (API) to compare Ireland with the EU in general, and also with two countries in theory with similar economics to Ireland: Belgium and Netherlands.*

*Based on the data provided by the European Union Eurostat its be comparing the performance of agriculture in Ireland with their neighbours and partners of the European Union based on two indicators (API): The index of price and the Index of expenditure to produce the agricultural products. Also, it will be considered the value of gross domestic product GDP which is one of the principal factors in the index of price.[6]*

*In this project, it will be introduced one sentimental feature which indicates the opinion of the experts about the GDP, whether is positive or negative the economy of the countries. For example, according to the experts if the GDP is higher consequently inflation is increasing, therefore, the index of expenditure (cost to produce the products) and the index of price increase increases, which means that a very higher GDP it is not desirable for the economy.[7][28][24][25]*

*Following the Cross Industry Standard Process **CRISP-DM**, [https://en.wikipedia.org/wiki/Cross-industry\\_standard\\_process\\_for\\_data\\_mining](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining), the phases and plan of the project are available here:*

*<https://github.com/users/sba22223nestorpereira/projects/1>*

*Justification, please see:[28][25]*

*<https://www.investopedia.com/articles/06/gdpinflation.asp>*

*Firstly, it has made a complete statistics analysis to compare the indexes API in Ireland and the EU, and then it has elaborated a machine learning model to predict the indexes API for the new.*

*It has introduced a Neural Network regression model including the historical data and the opinion of the experts about the GDP using sentimental features as input in the model. Also, it applied gradient algorithmics, XGBoost and Light GBM, very popular in the Kaggle competitions consequently of their good results, and compare with classic algorithms such as Random Forest.*

*The final result demonstrates that the index API from Ireland with respect to the EU and also, compare with Belgium and Netherlands, are higher but not higher in the EU, compared with Poland and Romania.*

*The data did not show a strong influence that the GDP variation on the expenditure index or price index that was expected.*

*The project generates several interactive and dynamic graphics to visualize those results following Tufte's 6 principles and produce a Dashboard that can be available via browser and therefore, available from any device PC, tablet or mobile phone.[9] [22][23]*

*Keywords: Machine Learning, Neural Network regression, Random Forest, XGboost Extended Gradient Boosting, **Light GBM**, Inference statistic, Dashboard, Panel, hvplot, Tune ensemble method, Data Wrangling.*

## I. INTRODUCTION

Based on the data provided by the European Union Eurostat it would be comparing the performance of agriculture in Ireland with their neighbours and partners of the European Union based on two indicators: The index of price and the Index of expenditure to produce the products. Also, it will be considered the value of gross domestic

product GDP which is one of the principal factors in the index of price.

In this project, it will be introduced one sentimental feature which indicates the opinion of the experts about the GDP, whether is positive or negative the economy of the countries. For example, according to the experts if the GDP is higher consequently inflation is increasing, therefore, the index of expenditure (cost to produce the products) and the index of price increase increases, which means that a very higher GDP it is not desirable for the economy. [26][27]

Following the Cross Industry Standard Process CRISP-DM, [https://en.wikipedia.org/wiki/Cross-industry\\_standard\\_process\\_for\\_data\\_mining](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining), the phases and plan of the project are available here:

<https://github.com/users/sba22223nestorpereira/projects/1>

Justification, please see:[28][24][25]

<https://www.investopedia.com/articles/06/gdpinflation.asp>

<https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP>

The data in the datasets available in the website Eurostat: [6][7]

<https://ec.europa.eu/eurostat/web/agriculture/data/database>

An Agricultural Price Index shows how agricultural revenue (output) and expenditure (input) are influenced by their price component and is therefore connected with Economic Accounts for Agriculture (EAA).

The agricultural price indices may serve various purposes of economic analysis.

The EU Agricultural Price Indices (API) comprise:

1- the index of purchase prices of the means of agricultural production (input)

Index of variation of the expenditure incurred by farmers in purchasing the means of production (goods and services as well as investment goods), including crop products from other agricultural units for intermediate consumption, over a given period.

2- the index of producer prices of agricultural products (output)

Index of variation of prices reflecting revenue received by the producer for goods and services actually sold to customers over a period.

The strategy to tackle the problem begins to do developing a complete statistical analysis of those indexes in the EU, then extracting and creating a **sentimental feature** to include the opinion of the experts about the GDP, and developing a machine learning model for regression: neural Network, random forest and gradient boosting.

## II. PROBLEM DEFINITION

In order to show how is agriculture production in Ireland it has been compared to the indexes API available on the website Eurostat with the EU in general and with two particular countries: Belgium and Netherlands.

The input price indices cover agricultural inputs including intermediate consumption of goods and services (fertilisers, pesticides, feed, seed, energy and lubricants, maintenance and repairs, etc.) and gross fixed capital formation related to investment goods (machinery and equipment, farms, buildings, etc.) [6]

The output price indices cover agricultural goods and services. They include crops, livestock and livestock products. The producer prices index of agricultural products (output) represents the measure of transaction prices reflecting revenue received by the producer for goods and services actually sold to customers over a period.[6]

It will be introduced a new feature related to the variation of the GDP.[7]

The four components of the gross domestic product are personal consumption, business investment, government spending, and net exports.



All those indexes are impacted by other economic factors but in particular by the GDP - Gross domestic product on output, expenditure and income.[24]

Eurostat publishes annual and quarterly national accounts use and input-output tables: this is an index of GDP and main components (output, expenditure and income).[7]

Data are available **from 2010** in Eurostat.

In order to maintain the consistency and coherence of the data in this project, its development a second part of the analysis from **2010 to 2021**.

Finally, it will be added to the data, characteristics (**Sentimental Categorical features**) based on the opinion of the experts in GDP related when the GDP is negative or positive.[26]

Most economists today agree that a small amount of inflation about 1% to 2% is beneficial, and is essential that the GDP of the countries needs to grow. However, if GDP growth is higher than 2.5% to 3.5% could be dangerous, because causes inflation or even worse hyperinflation.[28]

This economic parameter is essential in the index of producer prices of agricultural products (output) and the index of purchase prices of agricultural production (input) for Ireland and all the countries of the EU.

Therefore, GDP between 0% to 3.5% could be considered "positive", in another way, out of this range, could be considered "negative".[28]

This rule will be applied to this project.

(CRISP-DM Phase: Business/ Research Understanding Phase)

### III. METHODOLOGY

Following the Cross Industry Standard Process **CRISP-DM**,

[https://en.wikipedia.org/wiki/Cross-industry\\_standard\\_process\\_for\\_data\\_mining](https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining)

the phases and plan of the project are available here:

<https://github.com/users/sba22223nestorpereira/projects/1>

To tackle the problem and development of the project has been designed and executed a plan based on the phases described in the CRISP-DM framework:

@sba22223nestorpereira's CA2 project perform agriculture	
Title	Phase
1 Design and implementation control version and Planning #2	Business/ Research Understanding Phase
2 Define the problem	Business/ Research Understanding Phase
3 Process of acquiring data (research)	Data Understanding Phase
4 Analysis of raw data	Data Understanding Phase
5 Data wrangling: cleaning, redundancy, duplicate, missing values and outliers	Data Understanding Phase
6 Organize the data by years-countries: analysis of geodata	Data Preparation Phase
7 Create/obtain Sentimental Categorical features	Data Preparation Phase
8 EDA: Descriptive statistics analysis	Data Understanding Phase
9 Inference statistics analysis: Anova one way, Anova two way, Kruskal Wallis	Data Understanding Phase
10 Comparative inference analysis test	Data Understanding Phase
11 EDA: visualization	Data Understanding Phase
12 Verify Tufts principles	Data Understanding Phase
13 Design interactive and dashboard visualization	Data Preparation Phase
14 Implement interactive, dynamic and dashboard	Data Preparation Phase
15 Analysis strategic machine learning approach: Regression Neural Network and ensemble method to regression	Modelling Phase
16 Sentiment analysis: regression model with sentiment feature: justification and implementation	Modelling Phase
17 Analysis regression using Neural Network: justification and implementation	Modelling Phase
18 Analysis regression using Random Forest model: justification and implementation	Modelling Phase
19 Regression analysis using gradient boosting: XGBoost and Light GBM Boost	Modelling Phase
20 Comparative models based on the scoring metrics	Evaluation Phase
21 Tuning the best model: GridSearchCV	Evaluation Phase
22 Saving the models for deployment: final model	Deployment Phase
23 Collect results for final report	Deployment Phase
24 Collect bibliography and papers to support the implementation	Evaluation Phase
25 Generate the final report of the project	Deployment Phase
26 Formal presentation in Moodle	Deployment Phase

Fig. 1. Crisp-DM Phases

The code in the final version and the test version are also available here:

<https://github.com/sba22223nestorpereira>

and the data are available in GitHub.

Fig. 2. Github branches: main, test, data

The Eurostat website available for public access: <https://ec.europa.eu/eurostat/web/main/home> does

not allow reading directly from the website because it's a web application in which needs to choose an option before downloading the excel.

The data was downloaded according the date and store in GitHub:

[https://github.com/sba22223nestorpereira/CCT\\_sba22223nestorpereira/tree/data](https://github.com/sba22223nestorpereira/CCT_sba22223nestorpereira/tree/data)

#### IV. TECHNICAL APPROACH AND EVALUATION STRATEGY

In this section, it is boarded the different algorithms used, assumptions, and the benefits of them.

The strategy has been read the data, melted and joined it according to the period under study from **2010 to 2021**. (CRISP-DM Phase: Data Understanding Phase)

Basically, there are three types of observations from the raw data:

##### - Index of prices (output) by period (*called outa*)

*Price indices of agricultural products, output (2015 = 100) - annual data*

*Price indices of agricultural products, output (2010 = 100) - annual data*

*Price indices of agricultural products, output (2005 = 100) - annual data*

*Price indices of agricultural products, output (2000 = 100) - annual data*

The output price indices cover agricultural goods and services. They include crops, livestock and livestock products. The producer prices index of agricultural products (output) represents the measure of transaction prices reflecting revenue received by the producer for goods and services actually sold to customers over a period.

##### - Index of expenditure (input) by period (*called ina*)

*Price indices of the means of agricultural production, input (2015 = 100) - annual data*

*Price indices of the means of agricultural production, input (2010 = 100) - annual data*

*Price indices of the means of agricultural production, input (2005 = 100) - annual data*

*Price indices of the means of agricultural production, input (2000 = 100) - annual data*

The input price indices cover agricultural inputs including intermediate consumption of goods and services (fertilisers, pesticides, feed, seed, energy and lubricants, maintenance and repairs, etc.) and gross fixed capital formation related to investment goods (machinery and equipment, farms, buildings, etc.)

- GDP - Gross domestic product on output, expenditure and income (*called GDP*)

This is an index of GDP and its main components (output, expenditure and income).

Data are available from 2010 in Eurostat.

<https://ec.europa.eu/eurostat/web/agriculture/data/database>

Finally, it has been generated a **sentimental feature**. (CRISP-DM Phase: Data Preparation Phase)

It has been added as a categorical feature based on the general opinion of the experts in GDP related to when the GDP is negative or positive.

Most economists today agree that a small amount of inflation about 1% to 2% is beneficial, and is essential that the GDP of the countries needs to grow. However, if GDP growth is higher than 2.5% to 3.5% could be dangerous, because causes inflation or even worse hyperinflation.[28][24][26]

This economic parameter is essential in the index of producer prices of agricultural products (output) and the index of purchase prices of agricultural production (input) for Ireland and all the countries of the EU.

Therefore, GDP between 0% to 3.5% could be considered "positive", in another way, out of this range, could be considered "negative".

This rule will be applied to this project.

The strategy to tackle and develop the machine learning model for regression to estimate the Index of price for agriculture products in Ireland, and consequently all EU countries based on the data by Eurostat, would be: (CRISP-DM Phase: Modelling Phase)

1- Applying a Neural Network model for regression over data in 3D included a categorical **sentimental feature as input**.

2- Applying regression models based on the Random Forest model and two types of techniques of gradient boosting framework: XGBoost and Light GBM (light gradient-boosting machine). Also included a categorical sentimental feature as input.

Finally, it uses the GridSearchCV to tune final ensemble method: Light GBM. (CRISP-DM Phase: Evaluation Phase)

Based on the Gridsearchcv technique from Scikit-Learn package it is possible to tuning the Hyper parameter for the model selected: Light GBM.

This facility allows us to find the best hyper parameter combination to obtain the best results.

The develops was made by Python 3.0 under Jupyter, using libraries from PyPi, Scipy, Sklearn, Statsmodels, Panel, hvplot, and dmlc XGBoost, Light GBM.

## V. PRE-PROCESSING AND DATA WRANGLING

Data Wrangling is the process to clean the data and fix many observations invalid, missing or out of the range, in order to have the data ready to apply any exploratory data analysis (EDA) or any algorithms. (CRISP-DM Phase: Data Preparation Phase)

The difference problems treated in this project include:

### A. Join the data of index of expenditure (input).

Groups and join the data frames: df\_ina\_2015, df\_ina\_2010, df\_ina\_2005, and df\_ina\_2000 in order to create a DF from all periods from 2000 to 2021.

Join: it is an inner join in which the "priority" is the newest DF because has the most recent calculation of the index of prices, means from df\_ina\_2015.

### B. Join the data of index of prices (Output)

Groups and join the data frames: df\_outa\_2015, df\_outa\_2010, df\_outa\_2005, and df\_outa\_2000 in order to create a DF from all periods from 2000 to 2021.

Join: it is an inner join in which the "priority" is the newest DF because has the most recent calculation of the index of prices, means from df\_outa\_2015.

### C. Data wrangling: cleaning, missing values and outliers (Tukey fence method).

Basically, the index expenditure illustrates how the expenditure to produce the product has changed since the base period, and the price index illustrates how the price of a product or a basket of products a basket of products has changed.

The **base price** of an index is **100** by agreement (according to Eurostat), meaning that, for instance, an index equal to 110 reflects an increase in the absolute price of 10% and an index equal to 95 a decrease of 5%.

[https://ec.europa.eu/eurostat/cache/metadata/en/apri\\_pi\\_esms.htm](https://ec.europa.eu/eurostat/cache/metadata/en/apri_pi_esms.htm)

This value: **100**, will be considered in order to fix the missing values, meaning that any missing value will be substituted by the base price instead of the mean or median as conventionally used.

### D. GDP and generate a *sentimental feature*.

It will be added categorical feature based on the general opinion of the experts in GDP related when the GDP is negative or positive.

Most economists today agree that a small amount of inflation about 1% to 2% is beneficial, and is essential that the GDP of the countries needs to grow. However, if GDP growth is higher than 2.5% to 3.5% could be dangerous, because causes inflation or even worse hyperinflation. [28]

This economic parameter is essential in the index of producer prices of agricultural products (output) and the index of purchase prices of agricultural production (input) for Ireland and all the countries of the EU.



Therefore, GDP between 0% to 3.5% could be considered "positive", in another way, out of this range, could be considered "negative".[28]

This rule will be applied to this project.

#### E. Preparing the data for annual analysis of GDP by period 2010-2021

All data available, for all countries, is by year, therefore, it is necessary to regroup all data about GDP by year instead of a quarter. In order to do that, it will substitute the values of the four (4) quarters by the mean of the GDP per year for each country.

That means creating a new feature equal to the mean of the 4 quarters, for example, the GDP for 2010-Q1, 2010-Q2, 2010-Q3, and 2010-Q4, will be substituted by only one feature per year, 2010, per country.[7]

It is just for data **from 2010 to 2021**.

## VI. EDA AND STATISTICAL ANALYSIS

Firstly, the next graph shows a general visualization of the distribution and relation between Ireland and the EU (in general) using a scatter plot and histogram. (CRISP-DM Phase: Data Understanding Phase).

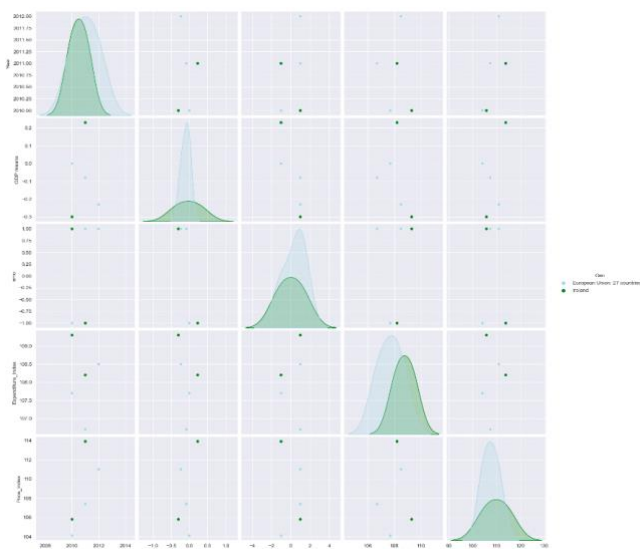


Fig. 3. Full visual comparative Price\_Index between Ireland and EU in general

According to the data, Ireland has a **higher** Price\_Index and also, higher Expenditure\_Index than the EU, see below:

	GDP means	Expenditure_Index	Price_Index
count	12.000000	12.000000	12.000000
mean	1.665833	105.635000	109.423333
std	1.672730	4.907163	8.598597
min	-0.300000	98.500000	95.310000
25%	0.582500	101.632500	103.295000
50%	1.450000	106.840000	108.960000
75%	2.150000	108.925000	114.475000
max	5.780000	113.400000	126.900000

Fig. 4. Descriptive statistics of indexes by Ireland.

	GDP means	Expenditure_Index	Price_Index
count	12.000000	12.000000	12.000000
mean	0.298333	103.084167	105.129167
std	0.498267	4.276559	4.011136
min	-0.650000	97.840000	98.460000
25%	-0.020000	99.227500	102.842500
50%	0.380000	103.750000	104.035000
75%	0.535000	106.950000	107.790000
max	1.250000	108.500000	112.300000

Fig. 5. Descriptive statistics of indexes by the EU.

The Price index and also the Expenditure index of Ireland, by mean, are the ones of the highest values in the EU countries.

However, Ireland has a Price Index, by means, below Poland and Romania despite the fact that the expenditure index is higher than both countries.




Geo	Price_Index			Expenditure_Index		
	min	max	mean	min	max	mean
Poland	100.52	120.21	111.825000	97.68	109.8	103.753333
Romania	101.50	123.11	110.207500	96.19	109.7	103.805000
 Ireland	95.31	126.90	109.423333	98.50	113.4	105.635000
Hungary	96.01	126.60	108.480000	95.20	109.3	102.673333
Sweden	100.20	117.58	107.972500	97.88	108.6	104.205000
France	99.83	115.10	106.685833	96.95	107.6	102.275000
Latvia	91.90	121.45	106.638333	95.29	109.3	101.278333
Bulgaria	92.68	135.10	106.454167	94.95	110.5	102.148333
Germany	98.31	114.00	106.005000	97.91	111.8	104.406667
Cyprus	96.66	113.30	105.821667	93.16	110.3	99.290833
Slovenia	98.34	114.70	105.605833	97.80	110.2	104.190000
Italy	97.10	112.00	105.566667	99.70	111.0	103.979167
Denmark	94.66	121.40	105.317500	99.86	112.4	105.545000
European Union: 27 countries	98.46	112.30	105.129167	97.84	108.5	103.084167
Czechia	88.70	122.90	104.322500	92.46	108.1	99.285833
Lithuania	92.64	118.90	103.945000	84.02	120.3	101.764167
Finland	96.63	119.10	103.668333	97.04	108.5	103.247500
Belgium	88.00	111.70	103.056667	97.09	111.1	102.977500
Luxembourg	94.20	109.40	102.571667	98.18	106.0	101.556667
Slovakia	88.00	116.40	102.400000	90.81	109.4	98.830833
Austria	96.10	108.10	102.317500	96.43	106.9	101.068333
Netherlands	96.39	107.40	102.110833	95.18	107.8	102.172500
Malta	96.40	107.20	102.000833	96.96	110.6	102.890000
Estonia	92.40	113.17	101.671667	93.11	100.0	96.924167
Portugal	94.10	109.83	101.300833	97.04	113.8	104.220000
Croatia	92.10	110.10	100.762500	91.93	111.7	100.129167
Greece	97.50	108.40	100.701667	98.10	107.3	102.999167
Spain	89.50	107.50	99.574167	95.54	110.3	102.422500

Fig. 6. Statistics ordered by Price\_Index by all EU.

In general opinion, Netherlands and Belgium have a similar economy to Ireland so for the purpose of comparison, it would be to obtain statistics values of those countries.

The information below shows that Ireland has indexes higher than their neighbours and partners Belgium and Netherlands.

#### Price Index comparison

European Union: 27 countries has mean : 105.13  
European Union: 27 countries has std : 4.01

Ireland has mean : 109.42  
Ireland has std : 8.60

Belgium has mean : 103.06  
Belgium has std : 7.61

Netherlands has mean : 102.11  
Netherlands has std : 3.17

Fig. 7. Price Index comparison.

#### Expenditure Index comparison

European Union: 27 countries has mean : 105.13  
European Union: 27 countries has std : 4.01

Ireland has mean : 109.42  
Ireland has std : 8.60

Belgium has mean : 103.06  
Belgium has std : 7.61

Netherlands has mean : 102.11  
Netherlands has std : 3.17

Fig. 8. Expenditure Index comparison.

Despite the fact that in the EU there are common rules and policies about agriculture production, the graph below shows that the indexes are variable in the EU countries, it is because their many other factors to impact those indexes.

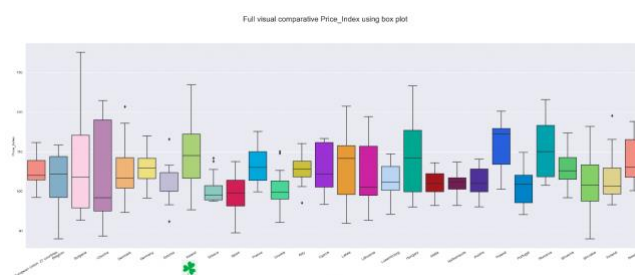


Fig. 9. EU countries comparison.

The index of variation of prices reflects revenue received by the producer for goods and services actually sold to customers over the period. This is a comparison between Ireland and their neighbour Belgium and Netherlands, and also the EU, in general. It demonstrates visually the higher price index of Ireland.

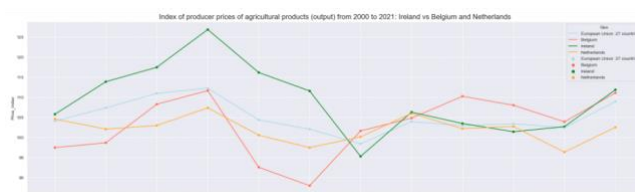


Fig. 10. Index of prices of agriculture products (output) from 2000 to 2021.

The index of variation of expenditure reflects the expenditure incurred by the farmers in purchasing the means of production (goods and services as well as investment goods), including crop products from other agricultural units for intermediate consumption, over a given period. This is a comparison between Ireland and their neighbour

Belgium and Netherlands, and also the EU, in general. It demonstrates visually also the higher expenditure index of Ireland which could explain why has a higher Price Index.

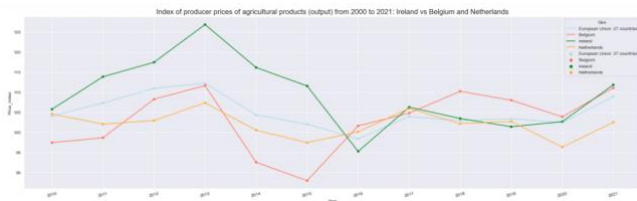


Fig. 11. Index of expenditure of agriculture products (input) from 2000 to 2021.

The analysis of the correlation below shows, surprisingly, a few impacts of the variation of the GDP into the indexes according to this data.

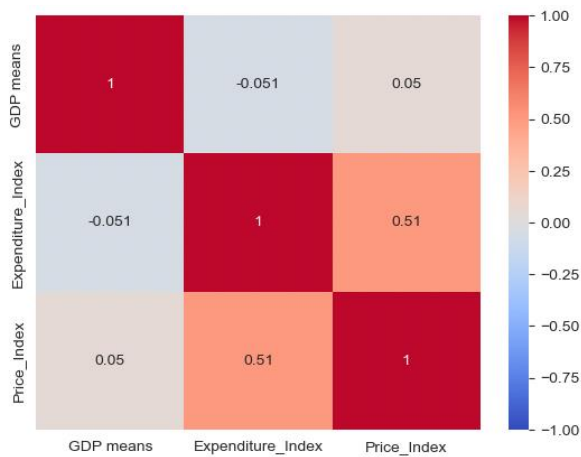


Fig. 12. Correlation between indexes and GDP variation.

#### A. Inference statistics Ireland vs EU: t-student's, Shapiro-Wilk test.

All data comes from countries in the EU therefore, we assume that all those countries follow common rules about the production and sell agricultural products. Therefore, it is a reasonable belief that there are relations (dependencies) between the countries in the EU related to those indexes of the Agricultural Price Index (API) under the evidence that exist Agricultural common policies.

It is can be considered that all these features are correlated and have dependencies between the countries so it can be used "paired dependence test".[21]

The variances in the populations are unknown.

Firstly, applies the **Shapiro-Wilk** test proves that the distribution of the Price Index from Ireland and the EU is almost Normal.[21]

Our null hypothesis  $H_0$  is that the distribution is Normal.

In both cases, since the p-value of the test is greater than  $\alpha = .05$ , the test statistic is 0.975 and 0.934, respectively.

We fail to reject the null hypothesis of the Shapiro-Wilk test.

Therefore, the data is assumed to be normally distributed.

Test for normality for data from Ireland:  
ShapiroResult(statistic=0.9759319424629211, pvalue=0.9620502591133118)

Test for normality for data from EU:  
ShapiroResult(statistic=0.9341965913772583, pvalue=0.4267212450504303)

Fig. 13. Shapiro-Wilk test.

Now, under the assumption that the distributions are normal, it applied the t-Student's test to compare Ireland vs the EU, the hypothesis are:

$H_0$ : Price index mean is equal in Ireland and EU

$H_1$ : Price index mean is non-equal in Ireland and EU

Since the p-value (0.021) is less than 0.05, and the test statistic is 2.687, we reject the null hypothesis.

We have sufficient evidence to say that the price index is different between Ireland and the EU.

#### B. Inference statistics Ireland vs Belgium vs Netherlands: Analysis of variance ANOVA.

According to mentioned before, all countries from the EU follow common rules about the production and sell agricultural products. Therefore, it is a reasonable belief that there are relations (dependencies) between the countries in the EU related to those indexes of the Agricultural Price Index (API) under the evidence that exist Agricultural common policies. means that the populations are not independent.

Hypothesis null,  $H_0$ : Price index means are equal for Ireland vs Belgium and Netherlands (significant difference between the means).

An alternative hypothesis, H1: Price index means are non-equal for Ireland vs Belgium and Netherlands.

ANOVA conditions:

- the distribution of the population is Normal
- the variances of the population are equal
- the population independent

Previous to ANOVA, its necessary to verify the assumptions, so it uses the Shapiro-Wilk test to verify the assumption of Normality.[21]

Our null hypothesis Ho is that the distribution is Normal.

In all cases, since the p-value of the test is greater than  $\alpha = .05$ , the test statistics is 0.975, 0.926 and 0.966, respectively. We fail to reject the null hypothesis of the Shapiro-Wilk test.

Therefore, the data is assumed to be normally distributed.

```
Test for normality for data from Ireland:
ShapiroResult(statistic=0.9759319424629211, pvalue=0.962050259113311)
```

```
Test for normality for data from Belgium:
ShapiroResult(statistic=0.9263584613800049, pvalue=0.343106269836425)
```

```
Test for normality for data from Netherlands:
ShapiroResult(statistic=0.9661213159561157, pvalue=0.866245627403259)
```

Fig. 14. ANOVA: test assumption of normality using Shapiro-Wilk test.

The second test assumption is the variances of the populations that the samples come from are equal.

Using **Levene's Test** for testing the variances are equal. This test uses the 'mean' which is recommended for symmetric or moderate-tailed distributions.[21]

The hypothesis null, Ho: the variances are equal

An alternative hypothesis, H1: the variances are different.

In the test, the p-value (0.011) is less than .05. The test statistic is 5.122.

This means that we can reject the null hypothesis.

This means we have sufficient evidence to say that the variances in price index between the three countries are significantly different.

```
Levene s test centered at the mean:
LeveneResult(statistic=5.122188959499782, pvalue=0.011551562070365987)
```

Fig. 15. ANOVA: test assumption of variances using Levene's test.

Finally, the assumption the populations are independent.

Unfortunately, there is no formal test to verify that the observations in each group are independent and as was mentioned before, all countries in the EU follow similar rules. Therefore, we can be considered that the samples are not completely independent.

According to the previous results of the tests applied, the best way to continue the analysis is using a non-parametric test.

In this case, it has been chosen to use the **Kruskal-Wallis test**, which is the non-parametric version of the one-way ANOVA.[21]

### C. Inference statistics Ireland vs Belgium vs Netherlands: Kruskal-Wallis test.

A Kruskal-Wallis test is used to determine whether there is a statistically significant difference between the medians of three independent groups.

Kruskal-Wallis test has some assumptions:

- the variable understudied is ordinal or continuous
- the distributions are similar
- the observations in each need to be independent

In this case, the variable price index for each country is not completely independent for the reasons explained before. However, in order to continue the study, it could be right to consider that the **samples** from the three countries are **almost independent**.

The null hypothesis Ho: The median of the price index across the three countries is equal.

The alternative hypothesis H1: At least one of the medians of the price index is different from the others countries.

In this case, the test statistic is 6.14 and the corresponding p-value is 0.0461.

Since this p-value is less than 0.05, we can reject the null hypothesis. We have sufficient evidence to conclude that the median of the **price index** for the three countries are **statistically significant differences**.

#### D. Statistical analysis: *conclusion*.

According to the results, the values of the features from the countries in the EU has some dependency that was expected because of the common rules in the EU.

Ireland has a price index means and expenditure index mean higher than the EU in general and also higher than their neighbours Belgium and Netherlands. However, it is not higher, than Poland or Romania despite the fact that the Ireland expenditure index mean is higher.

The data did not show a strong influence that the GDP variation on the expenditure index or price index that was expected.

### VII. INTERACTIVE AND DYNAMIC GRAPHS

Firstly, it has been organized the data by years-countries using data geodata (geographical data) with the code iso.

Based on Tufte's 6 principles:

- Comparisons: Show data by comparisons (bar charts and the like) to depict differences between an index of price in Ireland vs EU in general, and also EU Countries.

- Causality: Show how the GDP and the index of expenditure impact the index of price.

- Multivariate: simple graphics for easy interpretation from the general audience and the farmer.

- Integration: Incorporate maps with numerical data to show the difference between Ireland and the EU countries.

- Documentation: include attribution, detailed titles, and measurements.

- Context: Show the trend by years from the period 2010 to 2021.

It has created interactive and dynamic graphs of the indexes for all countries with the principal audience the farmers and then, the public in general. Those graphs allow visualisation and comparison between countries in the period of 2010 to 2021.

The colours chosen in the graphics follow the default values of the tools because in general, all countries in the analysis are under the same policy related to the agriculture production of the EU, the colour just represents that it is a different country (do not have another connection).



Fig. 16. Comparison Index of price between EU countries respected Ireland. Source: Eurostat.

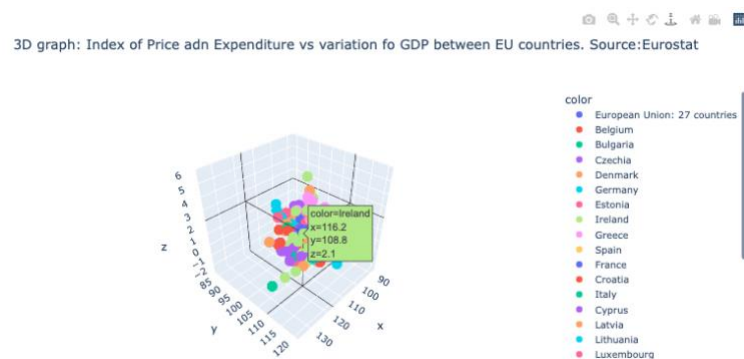


Fig. 17. 3D graph: Index of Price and Expenditure vs variation of GDP between EU countries. Source: Eurostat.



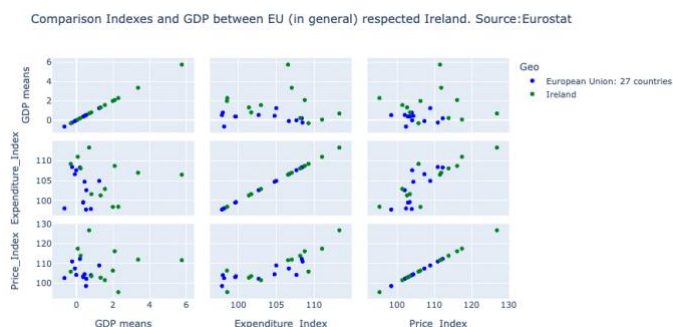


Fig. 18. Comparison Indexes and GDP between EU (in general) and Ireland. Source: Eurostat.

## VIII. DASHBOARD

Firstly, it has been organized the data by years-countries using data geodata (geographical data) with the code iso.

Following Tufte's principles was designed and created a dashboard based on the tools: hvplot, Plotty and Panel.

**hvPlot** is a high-level API for data exploration and visualization

<https://hvplot.holoviz.org/>

It has created a Dashboard available from any browser means that allows access from any device: tablet, PC or Mobile Phone.

Again, the dashboard includes the graphs of the indexes for all countries with the principal audience the farmers and then, the public in general. Those graphs allow visualisation and comparison between countries in the period of 2010 to 2021.

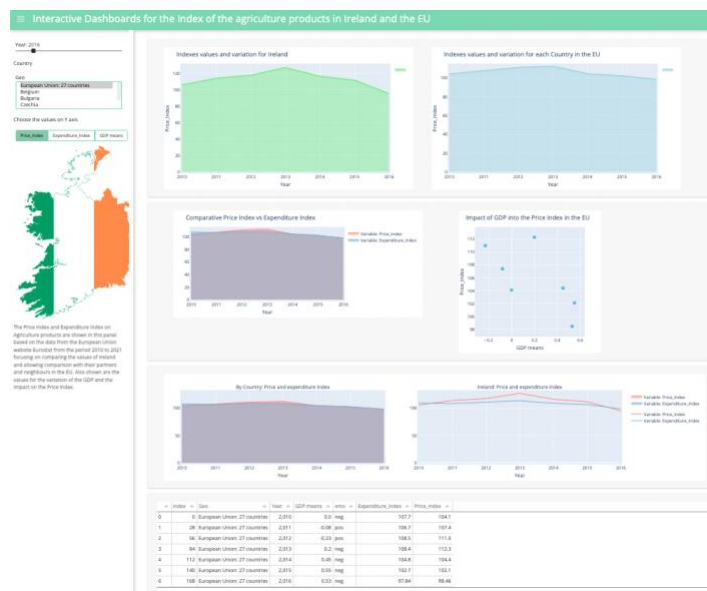


Fig. 19. Dashboard based on hvplot and Panel.

## IX. STRATEGIC TO APPROACH THE PROBLEM AND MODELLING THE DATA.

In this project, the approach to tackle the problem to estimate the Index of price for agriculture products in Ireland, and consequently all EU countries based on the data by Eurostat, would be:

(CRISP-DM Phase: Modelling Phase)

1- Applying a Neural Network model for regression over data in 3D included a categorical sentimental feature as input.[1][2][3][17]

2- Applying regression models based on the Random Forest model and two types of techniques of gradient boosting framework: XGBoost and Light GBM (light gradient-boosting machine). Also included a categorical sentimental feature as input.[1][3][11][12][13][29]

### A. ANN Neural Networks for regression.

In this part, it implemented a simple ANN model for regression applying 3-Dimensional data for:

- Country
- Year: 2010 to 2021
- Features available:
  - GDP means (means of GDP)
  - feature "emo" (including **emotional feature** from the opinion of experts)
  - Expenditure\_Index (Index of variation of the expenditure incurred by farmers(input))
  - Price\_Index (feature **target**: index of producer prices of agricultural products (output))

Specifically, this is a problem for multivariable (features) time series forecasting that can be approached using ANN models for

- input shape is 5 features, 28 countries (27 + EU global)
- activation function: rectified linear unit ReLU
- fully connected by using 32 nodes hidden layers

For regression:

One unit with no activation function.

Loss function

for regression: Mean square error.

[1][2][17][18]

The optimizer used was "**Adam**", Adam is an update to the RMSProp optimizer, and finally the metrics was the mean square error MSE.

[https://books.google.ie/books?hl=en&lr=&id=o5qnDwAAQBAJ&oi=fnd&pg=PP1&dq=J.+Brownee,+2020,+Deep+Learning+for+Time+Series+Forecasting&ots=yH63pOsg07&sig=tbBnZAY1Q0G6QiLZG5TKfkkHNW8&redir\\_esc=y#v=onepage&q&f=false](https://books.google.ie/books?hl=en&lr=&id=o5qnDwAAQBAJ&oi=fnd&pg=PP1&dq=J.+Brownee,+2020,+Deep+Learning+for+Time+Series+Forecasting&ots=yH63pOsg07&sig=tbBnZAY1Q0G6QiLZG5TKfkkHNW8&redir_esc=y#v=onepage&q&f=false)

Firstly, it was prepared the data array 3D for the Neural Network Regression algorithm.

(CRISP-DM Phase: Data Preparation Phase)

The model has been created to include the input of a 3D matrix with Countries, Years and features as dimensions, the **sentimental feature** is included as input to the model.[26]

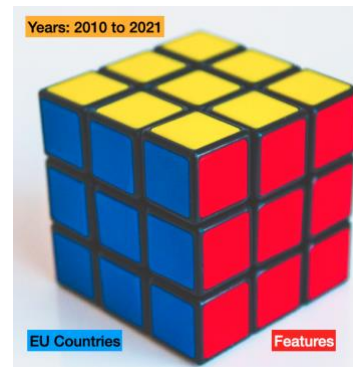


Fig. 20. 3D matrix data to NN model.

The summary of the model design is:

Model: "sequential"		
Layer (type)	Output Shape	Param #
dense (Dense)	(None, 28, 32)	192
dense_1 (Dense)	(None, 28, 32)	1056
dense_2 (Dense)	(None, 28, 1)	33
Total params: 1,281		
Trainable params: 1,281		
Non-trainable params: 0		

Fig. 21. Summary model design 2 levels.

In the beginning, the model produced not very encouraging results, but after applying cross-validation, the results improved substantially.

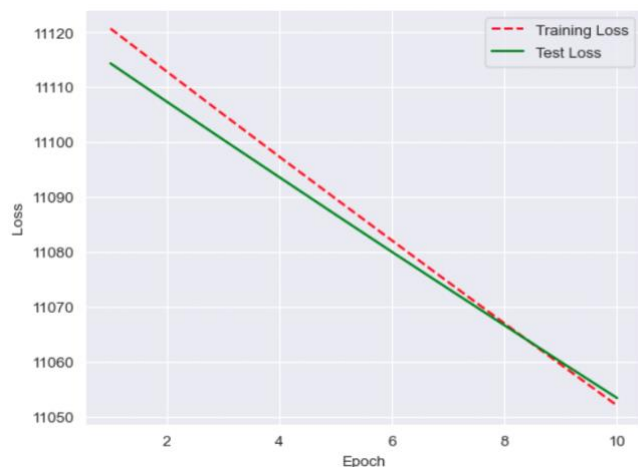


Fig. 22. ANN: training vs test with loss function.

After apply cross-validation, evaluating the neural network model using 10 ten-fold cross-validation, the **MSE 78.68** and standard deviation 37.68.

However, the gradient boosting models obtain better performance that shows the following comparative table:

	KerasRegressor	randomforest	XGBoost	lgmBoost
count	10.000000	10.000000	10.000000	10.000000
mean	78.687719	31.614384	38.478990	32.689700
std	37.684328	11.563304	14.734781	10.333781
min	24.155834	16.703616	20.303456	17.896157
25%	55.177494	23.742906	24.680310	25.304774
50%	77.357929	29.960041	38.294269	32.500643
75%	87.350143	39.685461	49.344442	37.902195
max	150.446655	48.779773	59.491680	50.073500

Fig. 23. Performance MSE comparative table.

B. Using ensemble method to improve performance and accuracy: **Gradient Boosting**. [29]  
(CRISP-DM Phase: Modelling Phase).

In this part, it implemented boosting models for regression:

- 1- Random Forest for regression
- 2- XGBoost or eXtreme Gradient Boosting for regression
- 3- Light GBM or light gradient-boosting machine for regression

All these models below come from the same concept of decision trees with differences in order to obtain performance and avoid overfitting.

- Random Forests (RF is used extensively in the industry because provides good results for many problems)

<https://scikit-learn.org/stable/modules/ensemble.html?highlight=random+forest#forests-of-randomized-trees>

The faster development of algorithms based on the technique of gradient boosting framework has two principal options very popular in the Kaggle competition:

- XGBoost or eXtreme Gradient Boosting from the Distributed (Deep) Machine Learning Community (DMLC) group.

<https://xgboost.readthedocs.io/en/stable/>

- Light GBM or light gradient-boosting machine development by Microsoft.

<https://lightgbm.readthedocs.io/en/v3.3.2/>

This project will be implemented both algorithms applied for a regression problem.

In theory, Light GBM would be better from the point of view of faster training speed and higher efficiency.

<https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/>



The following graph shows a comparative between the models, and also shows the residual analysis for regression:

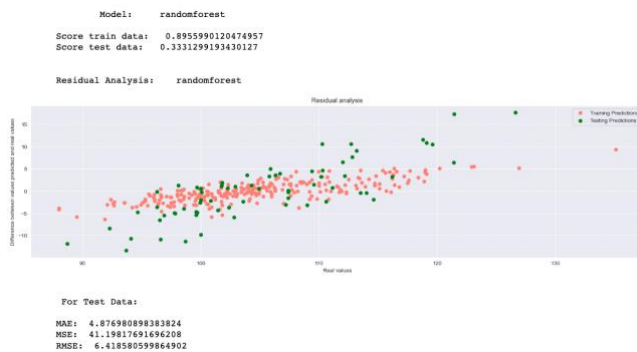


Fig. 24. Random Forest: difference between values predicted and real values.

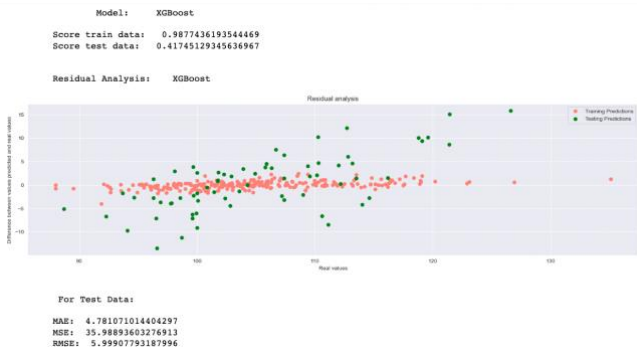


Fig. 25. XGBoost: difference between values predicted and real values.

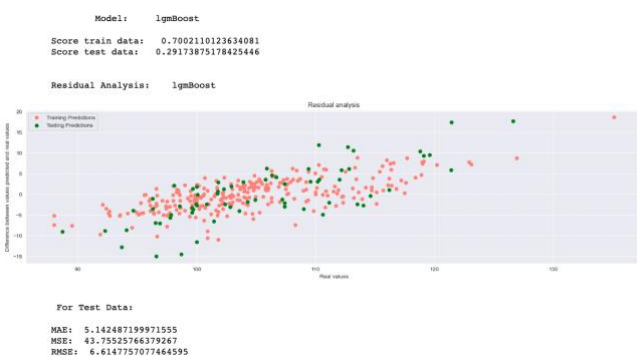


Fig. 26. Light GBM: difference between values predicted and real values.

C. Final comparison between the models based on metrics: MSE.  
(CRISP-DM Phase: Evaluation Phase)

Despite the fact that the mse mean in the Random Forest model is less than the mean on the Light gradient-boosting, **31.61 vs 32.68**, the standard deviation is less in the Light gradient-boosting mean that the values of mse are most stable. Less dispersion, therefore, will be chosen the Light **gradient-boosting** as the best model for this problem.

	KerasRegressor	randomforest	XGBoost	lgbmBoost
count	10.000000	10.000000	10.000000	10.000000
mean	78.687719	31.614384	38.478990	32.689700
std	37.684328	11.563304	14.734781	10.333781
min	24.155834	16.703616	20.303456	17.896157
25%	55.177494	23.742906	24.680310	25.304774
50%	77.357929	29.960041	38.294269	32.500643
75%	87.350143	39.685461	49.344442	37.902195
max	150.446655	48.779773	59.491680	50.073500

Fig. 27. Comparative models based on the scoring metrics MSE.

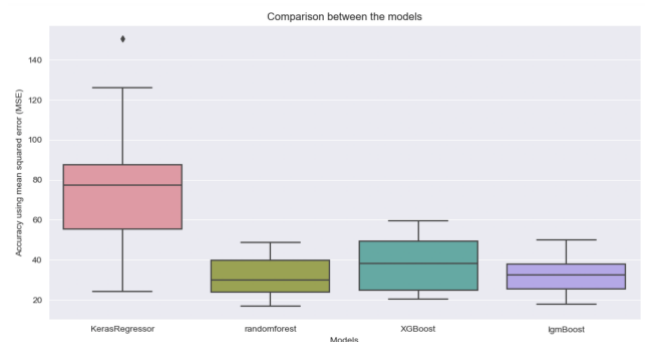


Fig. 28. Comparative models based on the scoring metrics MSE.

## X. TUNNING LIGHT GBM MODEL USING GRIDSEARCHCV (CRISP-DM Phase: Evaluation Phase)

Based on the Gridsearchcv technique from Sci kit-Learn package it is possible to tuning the Hyper parameter for the model selected Light GBM.

This facility allows us to find the best hyper parameter combination to obtain the best results.

This tuning of the parameters:

```
{'max_depth': 6,
'n_estimators': 50, '
num_leaves': 10}
```

allow improve the previous results, MSE: **31.98**.

The following graph shows how was obtained this improvement.

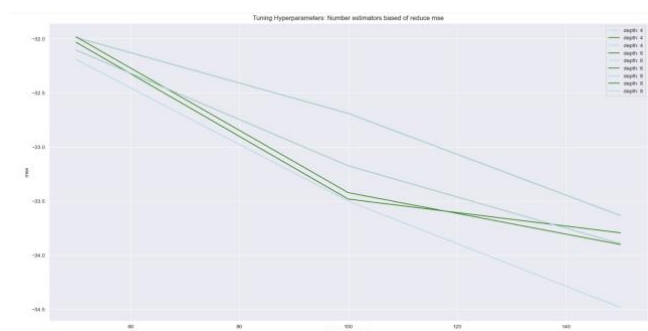


Fig. 29. Tunning Hyperparameters.

Finally, the Price index is shown in the next graph comparison with the real values,

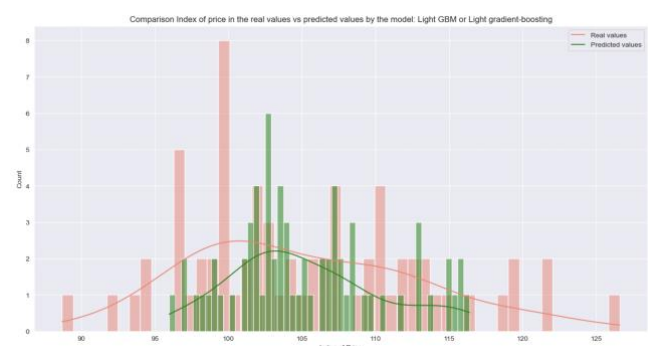
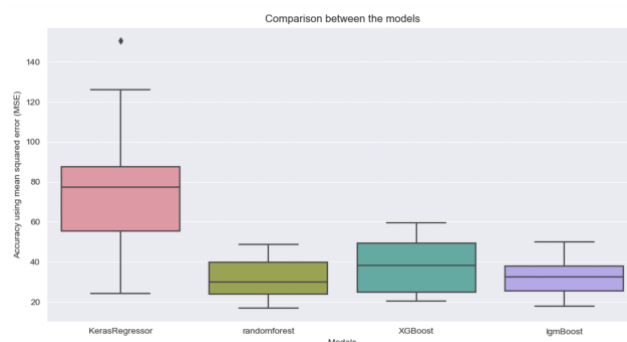


Fig. 30. Comparison Price Index estimated vs real values for the model Light GBM.

## XI. FINAL RESULTS AND THE BEST MODEL TO FIX THE PROBLEM

Using coherent comparative methods, in this project was compared different algorithms for regression to find the best trustily model to fix the problem and prognostic the index API.

	KerasRegressor	randomforest	XGBoost	lgmBoost
count	10.000000	10.000000	10.000000	10.000000
mean	78.687719	31.614384	38.478990	32.689700
std	37.684328	11.563304	14.734781	10.333781
min	24.155834	16.703616	20.303456	17.896157
25%	55.177494	23.742906	24.680310	25.304774
50%	77.357929	29.960041	38.294269	32.500643
75%	87.350143	39.685461	49.344442	37.902195
max	150.446655	48.779773	59.491680	50.073500



Despite the fact that the mse mean in the Random Forest model is less than the mean on the Light gradient-boosting, **31.61 vs 32.68**, the standard deviation is less in the Light gradient-boosting mean that the values of mse are most stable. Less dispersion, therefore, will be chosen the **Light gradient-boosting** as the best model for this problem.

In this approach, it was found that perhaps the best solution is to use more than one method applied in a coherent way to support a trusty solution.

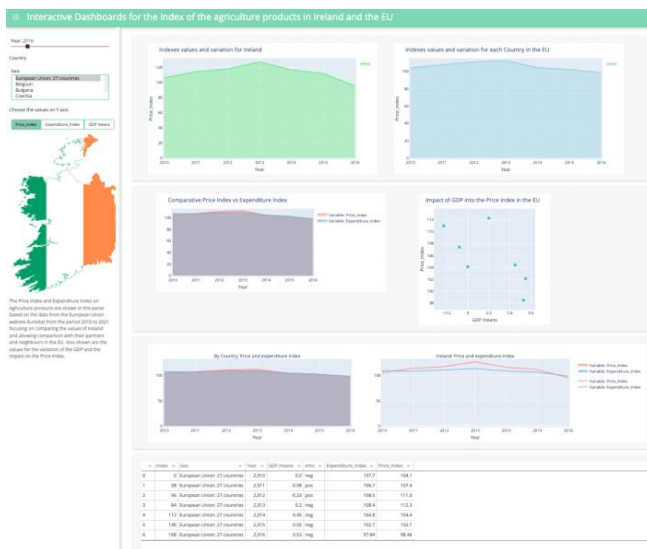
## XII. CONCLUSION AND FUTURE WORK

According to the results, the values of the features from the countries in the EU has some dependency that was expected because of the common rules in the EU.

Ireland has a price index means and expenditure index mean higher than the EU in general and also higher than their neighbours Belgium and Netherlands. However, it is not higher, than Poland or Romania despite the fact that the Ireland expenditure index mean is higher.

The data did not show a strong influence that the GDP variation on the expenditure index or price index that was expected.

Finally, the potential to use interactive and dynamic graphs to provide information to the farmer such as Dashboard allows help for future decisions about the production from the farm industry in Ireland.



It would be interesting to expand this project to prognostic those indexes in more detail showing the different categories of products: crop production, animal production, organic farms, etc.[6][7]

## REFERENCES

- [1] A. Géron, "Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow" (p. 28, 136, 195, 238, 243) 2019. O'Reilly
- [2] A. Albon, "Machine Learning with Python Cookbook: Practical Solutions from Preprocessing to Deep Learning (p. 33, 81, 109, 287)", 2018, O'Reilly, First edition. Kindle Edition.
- [3] J. VanderPlas, "Python Data Science Handbook" (p. 119, 390) 2016. O'Reilly.
- [4] statsmodels statistical models. <https://www.statsmodels.org/stable/index.html>. Access: October 2022.
- [5] A. Kassambara, "Machine Learning Essentials: Practical Guide in R." (P. 44-56). 2017. STHDA.
- [6] Eurostat, Agriculture prices and price indices (April), 2022 <https://ec.europa.eu/eurostat/web/agriculture/data/database>
- [7] Eurostat, GDP and main components (output, expenditure and income), 2022, [https://ec.europa.eu/eurostat/databrowser/view/NAMQ\\_10\\_GDP\\_custom\\_3761889/bookmark/table?lang=en&bookmarkId=4eef75c1-4ab8-4e39-865e-6301e3390d28](https://ec.europa.eu/eurostat/databrowser/view/NAMQ_10_GDP_custom_3761889/bookmark/table?lang=en&bookmarkId=4eef75c1-4ab8-4e39-865e-6301e3390d28)
- [8] W. McKinney. Python for Data Analysis (p. 191, 205, 227, 268, 296). 2018 O'Reilly.
- [9] C. Chen, W. Hårdle, A. Unwin. Handbook of Data Visualization (P. 152). 2008 Springer
- [10] Gradient boosting. Wikipedia. [https://en.wikipedia.org/wiki/Gradient\\_boosting](https://en.wikipedia.org/wiki/Gradient_boosting). Access: October 2022
- [11] Sklearn. Ensemble methods. <https://scikit-learn.org/stable/modules/ensemble.html>. Access: October 2022
- [12] Sklearn. Gradient Tree Boosting methods. <https://scikit-learn.org/stable/modules/ensemble.html>. Access: October 2022
- [13] XGBoost extended gradient boosting. <https://xgboost.ai/about>. Access: October 2022
- [14] Sklearn. Classical linear regressors, Regressors with variable selection. <https://scikit-learn.org/stable/modules/classes.html#classical-linear-regressors>. Access: October 2022.
- [15] T. Hastie, R. Tibshirani, J. Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". (p. 37-40, 219, 226, 597). 2017. Springer.
- [16] J. Brownlee, 2018, "How to Reduce Variance in a Final Machine Learning Model", <https://machinelearningmastery.com/how-to-reduce-model-variance/> Access: July 2019
- [17] J. Brownlee, 2020, Deep Learning for Time Series Forecasting
- [18] J. Brownlee, 2019, "Machine Learning Mastery with Python"
- [19] Scikit-learn developers, 2019, "Ensemble methods", <https://scikit-learn.org/stable/modules/ensemble.html#gradient-tree-boosting>. Access: October 2022.
- [20] Wikipedia. [https://en.wikipedia.org/wiki/Elbow\\_method\\_\(clustering\)](https://en.wikipedia.org/wiki/Elbow_method_(clustering)) Access: October 2022
- [21] N. Weiss Introduction STATISTICS (p 273-280). Pearson. 2017
- [22] hvplot, 2022, <https://hvplot.holoviz.org/>
- [23] plotly, 2022, <https://plotly.com/>
- [24] International Monetary Fund, FD, Gross Domestic Product AN ECONOMY'S ALL, 2022, <https://www.imf.org/en/Publications/fandd/issues/Series/Back-to-Basics/gross-domestic-product-GDP>
- [25] Farm Price Index (FPI), Investopedia, 2022, <https://www.investopedia.com/terms/f/farmprices.asp>
- [26] Sentimental Analysis, Brand 24, 2022, <https://brand24.com/blog/sentiment-analysis/>
- [27] Stock-prediction-using-twitter-sentiment-analysis, Kaggle, 2022, <https://www.kaggle.com/code/kirolosat/stock-prediction-using-twitter-sentiment-analysis#Load-the-dataset>
- [28] The Importance of Inflation And GDP, Investopedia, 2021, <https://www.investopedia.com/articles/06/gdpinflation.asp>
- [29] Analytics Vidhya, Which algorithm takes the crown: Light GBM vs XGBOOST? 2020, <https://www.analyticsvidhya.com/blog/2017/06/which-algorithm-takes-the-crown-light-gbm-vs-xgboost/>

The develops was made by Python 3.0 under Jupyter, using libraries from PyPi, Scipy, Sklearn, Statsmodels, dmlc XGBoost, Light GBM, Panel, Plotly, hvplot.