**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *Individual* |
| **Lecturer Name:** | *Sam Weiss*  *Marina Iantorno / John O’Sullivan*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Maria Dominguez Alvarenga |
| **Student Number:** | 2019008 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**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. |

Table of Contents

[Construction Materials for 8 countries of the Eurozone. Performance and comparison analysis. 2](#_Toc136028379)

[1. Abstract 3](#_Toc136028380)

[2. Introduction. 3](#_Toc136028381)

[3.– Creative Commons Data. 3](#_Toc136028382)

[4. EDA Analysis 4](#_Toc136028383)

[4.1. Data Sets Overview - Data Frames creation. 4](#_Toc136028384)

[4.1.1. Exports materials: 4](#_Toc136028385)

[4.1.2. Imports materials: 5](#_Toc136028386)

[4.1.3. GDP: 5](#_Toc136028389)

[4.2. Analysis and merging of dataframes. 5](#_Toc136028390)

[4.2.1. NaN values when creating dataframes: 6](#_Toc136028391)

[4.2.2. Outer Merge: 6](#_Toc136028392)

[4.2.3. Dropping NaN values after merging. 6](#_Toc136028393)

[(Table 1) 7](#_Toc136028394)

[4.3. Descriptive Statistics. 7](#_Toc136028395)

[4.4. Detecting categorical data. 8](#_Toc136028396)

[4.5. Inferential Statistics. 8](#_Toc136028397)

[4.5.1. Ireland Imports. 8](#_Toc136028398)

[4.5.2. Ireland Exports. 9](#_Toc136028399)

[4.6. Probability tests. Parametric and non-parametric tests 9](#_Toc136028400)

[4.6.1. One tailed test - Ireland imports variable 9](#_Toc136028401)

[4.6.2. Kolmogorov-Smirnov test – Imports feature. 10](#_Toc136028402)

[4.6.3. ANOVA test (parametric test) – Population: 8 countries vs Exports feature. 10](#_Toc136028403)

[4.6.4. Tukey-Kramer analysis (non-parametric test) 11](#_Toc136028404)

[4.6.6. Two-sample t-tests (Ireland and Denmark exports) 13](#_Toc136028405)

[4.6.7. The Mann Whitney U test - (Ireland and Denmark exports - non-parametric test) 13](#_Toc136028406)

[4.7. Correlation of the Variables. 14](#_Toc136028407)

[4.7.1. Visualizing the correlation between Imports and Gdp features. 14](#_Toc136028408)

[4.7.2. Visualizing the correlation between Imports and Exports features. 15](#_Toc136028409)

[4.8. Positive Skewed to the right data shape 15](#_Toc136028410)

[4.9. Outliers Detection. 16](#_Toc136028411)

[4.9.1. Imports feature outliers. 16](#_Toc136028412)

[4.9.2. Gdp feature outliers 16](#_Toc136028413)

[4.9.3. Winsorisation of Outliers. 17](#_Toc136028414)

[4.9.4. Verifying that shape after applying Winsorization of outliers. 17](#_Toc136028415)

[5. Machine Learning Section. 18](#_Toc136028416)

[5.1. CRISP DM Methodology. 18](#_Toc136028417)

[5.2. Version control link: 18](#_Toc136028418)

[5.3. Supervised Machine Learning models. 18](#_Toc136028419)

[5.4. Detected issue and Strategy of Analysis. 19](#_Toc136028420)

[5.5. Identifying X variables and Y target. 19](#_Toc136028421)

[5.6. Feature Engineering 19](#_Toc136028422)

[5.6.1. Categorical Data Encoding. 19](#_Toc136028423)

[5.6.2. Feature scaling. 20](#_Toc136028424)

[5.7. Implementation of Machine Learning Models 20](#_Toc136028425)

[5.7.1. Splitting Train and test set 20](#_Toc136028426)

[5.7.2. Support Vector Regressor 20](#_Toc136028427)

[5.7.3. Decision Tree and Random Forest Regressors. 20](#_Toc136028428)

[5.7.4. Linear Regression. 20](#_Toc136028429)

[5.7.5. Ridge and Lasso Regression. 21](#_Toc136028430)

[5.8. Comparison ML model performance 21](#_Toc136028431)

[5.9. Conclusion 21](#_Toc136028432)

[6. Sentiment Analysis. 22](#_Toc136028433)

[7. Dashboards (Interactive visualizations) 22](#_Toc136028434)

[8. Other libraries different to pandas for Data Manipulation ( Appendix 4) 23](#_Toc136028437)

[9. Appendices 24](#_Toc136028438)

[Appendix 4 : Other libraries different to pandas for Data Manipulation. 27](#_Toc136028439)

[9. References 28](#_Toc136028440)

# **Construction Materials for 8 countries of the Eurozone. Regression models Performance and comparison analysis.**

Author: Maria Dominguez, MScs Data Analytics student

Student ID: 2019008

# **1. Abstract**

*The steady increase in imports of construction materials in Ireland, which has an impact on the construction sector is an issue where machine learning regression models can be applied to analyse that variable. This report aims to use supervised regression models which can produce a higher R2 score. 6 regression models have been used and compared to measure performance to predict imports, using data from 8 countries in Europe. Thus, this analysis concluded that performing Support Vector Regressor (SVR), Decision Tree Regressor (DTR), Random Forest Regressor (RFR), Linear Regression (LR), Ridge Regression (RR) and Lasso Regression (LSR) were compared. However, either DTR and RFR performed better, and these two models can be used to make predictions in terms of Imports for Ireland and 8 countries of the eurozone.*

# **2. Introduction.**

A key factor to take into account are construction materials costs when estate market in Ireland is analysed as a result of their impact on the supply of houses, which prices have been steadily increasing over the last twenty-five year (Arigoni et al, 2022). As study conducted by Linesight (2022) about construction materials in Ireland, produced the results that imports of steel, experienced a risen in the price as a result of the global market stressed increased by the Russia-Ukraine conflict. As well as that, Copper prices remain steadily high in 2021 and remain the same through 2022. Therefore, the application of supervised machine learning (ML) algorithm for regression could be useful to make predictions about imports.

# **3.– Creative Commons Data.**

Row data consisted of four datasets, 3 of which were obtained from Knoema site (2023), which were renamed “import materials csv”, “exports materials csv” and “gdp csv”. The original data source was from OECD site, (2023) and it is public domain, which allows me to use it as datasets are not protected by intellectual property laws (Stanford, 2023). I did also verify with the original source about the use of their data, who confirmed their data is public domain as long as it is not used for commercial purposes and the source (OECD) is referenced in the report. An additional dataset “Iso3” was obtained from FAO site (2023), which is open data under the licence “Attibution-NonCommercial-ShareAlike 3.0 IGO”, which allowed me to use in the same terms of stated by OECD. The positive aspects that I gain from this datasets search is that it helps me to feel more confident about chosen datasets in future, which will be our reality in real cases scenarios. However, perhaps a negative aspect, was that many good datasets were only available in paid sites like Statista.com, which memberships were not affordable. Nevertheless, Knoema site is a paid site as well, but inexpensive. The search of the datasets was time consuming, but I took it that the main purpose of this activity is to develop the ability for searching suitable data following the correct criteria for downloading and processing of the same.

# **4. EDA Analysis**

## **4.1. Data Sets Overview - Data Frames creation.**

All the different stages of the EDA analysis have been performed when analyzing and merging the row data, which are going to be explained below and all included in the accompanied Jupiter notebook.

### **4.1.1. Exports materials:**

The original csv file name was “ObservationData\_wnqhzcf.csv”, which was renamed "Exports materials.csv” to avoid any confusion when uploading datasets in the Jupiter notebook.

* The type of data is continuous, and it measures all the units of exports (construction minerals, metals, and non-minerals) for 8 countries of the eurozone (Ireland, Germany, France, Netherlands, Denmark, Poland, Italy, and Spain.
* There is a total of 627 observations and 6 features in this dataset.
* It contains 3 features with numeric, continuous data, 3 columns with categorical data and 1 feature “unit” that only contains “NaN” values.
* Table with description of dataset data and feature type (appendix 1.1.)

### **4.1.2. Imports materials:**

### The original csv file name was “ObservationData\_iwkolud.csv” and it was renamed "Imports materials.csv”.

### The type of data is continuous, and it measures all the units of imports construction materials previously referenced for the same 8 countries listed above.

* There is a total of 627 observations and 6 features.
* It contains 2 features with numeric, continuous data, 3 columns containing categorical data, and 1 feature with only NaN values.
* Table with description of dataset data and feature type (appendix 1.2.)

### **4.1.3. GDP:**

The original csv file name was “ObservationData\_jhykshd.csv” and renamed as "Gdp materials.csv”.

* The type of dataset is continuous, and it measures all the gdp values of imports for construction the same construction materials and countries listed above.
* There is a total of 625 observations and 6 features.
* It contains 2 features containing numeric, continuous data, 2 columns containing categorical data and 1 feature containing only NaN values.
* Table with description of dataset data and feature type (appendix 1.3.)

**Codes (Iso 3 dataset)**

* The type of dataset is nominal values listing the country names and country codes.
* There is a total of 274 observations and 2 features in this dataset. 2 columns of categorical data.
* Table with description of dataset data and feature type (appendix 1.4.)

## **Analysis and merging of dataframes.**

### **NaN values when creating dataframes:**

Null values were detected on the “value” feature when creating 3 data frames (exports, imports, gdp) which were dropped as they will not yield any significant result and imputation of data is not suitable either.

### **Outer Merge:**

This method was chosen to merge the datasets as I wanted to keep information from both datasets. A new dataframe “cons” was created to first merge exports and imports dataframes. Then, a second new data frame “const1” was created to combine “const dataframe” with “gdp dataframe”. Finally, a third new dataframe “construction” was subset to join “const1” with “codes dataframe”. Therefore, whichever information was in one dataframe but it is not in the other, it was filled in the columns as NaN Values.

### **Dropping NaN values after merging.**

* As a result of the outer merge, there were rows that were filled with NaN values, which were dropped before going further in the analysis.
* A screenshot of a computer

  Description automatically generated with low confidenceAfter dropping null values, the “construction” dataframe contains 601 observations and 7 features as below:

(Figure 1)

* New dataframe “construction” contains 4 variables with numeric continuous data, 3 categorical features, The data type has been checked using info() as below:

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Feature** | **Variable type** | **Data type** |
| 1 | Country | Text | String |
| 2 | Group | Text | String |
| 3 | Year | Numeric | Integer |
| 4 | Exports | Numeric | Float64 |
| 5 | Imports | Numeric | Float64 |
| 6 | Gdp | Numeric | Float64 |
| 7 | Code | Text | String |

## 

## (Table 1)

## **4.3. Descriptive Statistics.**

After producing the summary statistics for the numerical features (see Figure 2 ) as below:

* Differences between the mean and median value of the features, which indicates that there is a skew in the data and the data is not spread out evenly.
* The scale of the data has been seen when checking the minimum and maximum values.
* Standard deviation values, it can be seen that feature “exports”, “imports” and “gdp” are less spread out and their standard deviation values are more clustered around the mean.

A screenshot of a data

Description automatically generated with low confidence

(Figure 2)

## **4.4. Detecting categorical data.**

To check the categorical features of “construction dataframe” ( see Figure 3 below ) as follows:

* All categorical features appear 601 times, where “Italy” is the top group category is “metals” with a frequency of 201 times, followed by “Italy” and “ITA” which are displayed 81 times.
* “Country” and “code features have 8 unique values (8 countries chosen)
* “Group” feature has 3 unique values (construction minerals, metals, and non-minerals). The feature engineering method needs to be considered before performing ML models to convert all those string values to numeric values.

A screenshot of a computer code

Description automatically generated with low confidence

(Figure 3)

## **4.5. Inferential Statistics.**

Construction dataframe was sub-set to obtain a confidence interval for population Ireland for “imports” and “exports” populations respectively as follows:

Alpha = 0.90, which produces an output of (3.67981, 4.45099)

### **4.5.1. Ireland Imports.**

|  |  |
| --- | --- |
| **Alpha = 0.90** | **Alpha = 0.99** |
| (3.679813416033894, 4.45099152223771) | (3.45399980097246, 4.676805137299144) |

(See appendix 2.1.)

**Interpretation of results:**

The 90% confidence interval for the true population mean is (3.67-4.45). On the other side, the wider the confidence interval of 99% confidence interval for the true population mean is (3.67-4.67). Notice that this interval is wider than the previous 90% confidence interval (Statology, 2023).

### **4.5.2. Ireland Exports.**

|  |  |
| --- | --- |
| **Alpha = 0.95** | **Alpha = 0.99** |
| (1.5962077918570257, 2.147488504439271) | (1.5063675269201227, 2.2373287693761736) |

(See appendix 2.2.)

**Interpretation of results:**

The 95% confidence interval for the true population exports mean is (1.59-2.14). The larger the confidence level, the bigger the confidence interval will be. The wider the confidence interval. In contrast, with the 99% confidence interval for the true population exports mean is (1.50-2.23). Thus, this interval is wider than the previous 95% confidence interval. (Statology, 2023).

## **4.6. Probability tests. Parametric and non-parametric tests**

### **4.6.1. One tailed test - Ireland imports variable**

**Hypothesis:**

* 𝐻0: 𝜇 = 4 and 𝐻A 𝜇 > 4
* Level of significance is 𝛼 = 0.05.
* **Result:** pvalue=0.38923507692394965 (Appendix 3.1.)

**Conclusion:** The sample mean is 4.06 and pvalue is 0.38, which is greater than 𝛼. Therefore, we fail to reject 𝐻0 as we have not found enough evidence that 𝜇 is equal than 4.

### **4.6.2. Kolmogorov-Smirnov test – Imports feature.**

**Hypothesis**:

* 𝐻0 = Data is consistent with normal distribution.
* 𝐻𝐴 = Data is not consistent with normal distribution.
* Level of significance is 𝛼 = 0.05.
* **Pvalue**=0.11720684769232348 (Appendix 3.2.)

**Conclusion:** We fail to reject 𝐻0. We have not proved that our data comes from normal distribution. It is reasonable to assume that our data is consistent with normal distribution.

### **4.6.3. ANOVA test (parametric test) – Population: 8 countries vs Exports feature.**

It was used the population of the 8 countries and the Exports variable values.

**Hypothesis**

* 𝐻0: Ireland = Germany = Netherlands = Poland = Spain = Denmark = France = Italy (exports are the same)
* 𝐻𝐴: At least one is different from the others.
* Level of significance is 𝛼 = 0.05
* **Pvalue**=3.3689928916359845e-181 (Appendix 3.3.)

**Conclusion:** We have rejected 𝐻0. We have strong evidence to reject 𝐻0. At least one of the populations is different.

### **4.6.4. Tukey-Kramer analysis (non-parametric test)**

A screenshot of a computer

Description automatically generated with low confidence

(Appendix 3.4.)

**Conclusion:**

* **Denmark and Ireland**: Since the confidence interval covers zero and goes from a negative value to a positive value, I do not have evidence of a difference between Ireland and Denmark.
* **Germany and Ireland**: Since the confidence interval is only in negative numbers, I have evidence that there is a difference between these two countries in terms of exports.
* **Ireland and Netherlands**: The confidence interval is positive the two (lower and upper). There is a strong difference between those two countries.

#### **4.6.5.1. Visualization of test - ANOVA test**

Through the visualization of a box plot by group, it is possible to display that there is a difference among countries population (exports). I did use this visualization as I needed to picture the 8 countries export distributions at once (McQuaid, 2023). As well as that, the clarity of the visualization was maintained to inspire to the eye to make a comparison of many pieces of the data (Tufte, 2001)

A picture containing diagram, screenshot, text, rectangle

Description automatically generated

Box Plot of Exports by Group – (Appendix 3.5.)

### **4.6.6. Two-sample t-tests (Ireland and Denmark exports)**

**Hypothesis:**

* 𝐻0: There is no difference in exports between Ireland and Denmark.
* 𝐻𝐴: There is a difference between the means of those two countries.
* Level of significance is 𝛼 = 0.05. (Appendix 3.6.)

**Pvalue**=1.0916435319052207e-22

**Conclusion:** We reject 𝐻0 and we conclude that the means between the two countries are different.

### **4.6.7. The Mann Whitney U test - (Ireland and Denmark exports - non-parametric test)**

* 𝐻0: There is no difference in exports between Ireland and Denmark.
* 𝐻𝐴: There is a difference between those two countries medians.
* Level of significance is 𝛼 = 0.05.
* Pvalue=2.1127054483208824e-19. (Appendix 3.7)

**Conclusion**: We reject 𝐻0 as there is difference between the medians of Ireland and Denmark.

**4.6.7.1. Visualization:** It was chosen a histogram to present the best understanding of a distribution (McQuaid, 2023).

A picture containing text, screenshot, diagram, plot

Description automatically generated

Histogram Exports Ireland - Denmark – (Appendix 3.8.)

## **4.7. Correlation of the Variables.**

A heat map was used to show the correlation between the features through the mapping of the different variables (McQuaid, 2023). There is a strong correlation of 0.83 between “imports” and “exports”. Likewise, there is a moderate correlation of 0.44 between “imports” and “gdp”. Also, there is a weak correlation of 0.24 between “exports” and “gdp”.

**A screenshot of a computer

Description automatically generated with low confidence**

Heat map for correlation of variables (Figure 4)

### **4.7.1. Visualizing the correlation between Imports and Gdp features.**

Scatterplots were used to provide an excellent visualization of the relationship of these two features (McQuaid, 2023). The default colour blue was used to plot these, histogram and box plot diagrams.

**A picture containing text, screenshot, font, plot

Description automatically generated**

Scatter plot to check a moderate correlation between two variables (Figure 5)

### **4.7.2. Visualizing the correlation between Imports and Exports features.**

**A picture containing text, screenshot, map, font

Description automatically generated**

Scatter plot to check a strong correlation between two features (Figure 6)

## **4.8. Positive Skewed to the right data shape**

Through the Histogram visualization of variables, it was possible to get a better insight into the distribution of the data (McQuaid, 2023), which is positive skewed to the right.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Histogram Imports data (Figure 7)

## **4.9. Outliers Detection.**

### **4.9.1. Imports feature outliers.**

Outliers were identified in three of the variables which are going to be used in the ML models. The visualization of those extreme values, was obtained through box plots (McQuaid, 2023)

A picture containing screenshot, diagram, rectangle, line

Description automatically generated

Box Plot Imports (Figure 8)

Median value is 17.57, the extreme values of 132.72, which are outside of the whiskers and minimum value is “0”. The standard deviation is more clustered around the mean.

### **4.9.2. Gdp feature outliers**

A picture containing screenshot, rectangle, diagram, line

Description automatically generated

Box Plot Gdp (Figure 9)

Median value is 7.28, the extreme values of 66.44, which are outside of the whiskers and minimum value is “2”. The standard deviation 14.64 is very close to the mean 14.97.

### **4.9.3. Winsorisation of Outliers.**

This technique was used in ML class, tutorial 3 (a) solution (Iqbal, 2023). I decided to apply this technique, so I did not eliminate all my outliers.

**A picture containing rectangle, screenshot, line, design

Description automatically generated**

Box Plot outliers after they have been winsorised (Figure 10)

### **4.9.4. Verifying that shape after applying Winsorization of outliers.**

After applying this technique, histograms were plotted to make sure that the shape of the data has not changed as these diagrams provided a better understanding of the shape distribution of the data (McQuaid, 2023), and the shape remain the same.

**A picture containing text, diagram, screenshot, plot

Description automatically generatedA picture containing text, screenshot, plot, diagram

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated**

**A picture containing text, screenshot, plot, diagram

Description automatically generated**

**A picture containing text, screenshot, diagram, plot

Description automatically generatedA picture containing text, screenshot, diagram, plot

Description automatically generated**

Histograms of features - winsorisation of outliers (Figure 11)

# **Machine Learning Section.**

## **5.1. CRISP DM Methodology.**

[A diagram of data processing

Description automatically generated with low confidence](https://app.diagrams.net/)

**CRISP DM Phases – Construction Materials ( Figure 12)**

The project prepared about Construction Materials for 8 countries, is applicable to the CRISP- DM methodology as this is a flexible methodology, which allowed me to revisit previous stages of the process to correct steps in the analysis. I.e: checking feature scaling in data preparation.

## **5.2. Version control link:**

<https://github.com/mariadominguez2023/assignment2.git>

## **5.3. Supervised Machine Learning models.**

Supervise ML models have been used as my datasets have independent(X) and dependent(y) variables, both are known as labelled data (Moez, 2022) and I did identify the relation between the chosen X and Y (Hack, 2019). On this project, since I have continuous data, I used regression models, which are one of the types of Supervised learning algorithms as the target variable “Imports” is continuous.

## **5.4. Detected issue and Strategy of Analysis.**

Construction materials in terms of imports have increased steadily in Ireland in previous years (Linesight, 2022). and that is why I decided to check whether ML regression models can be performed to predict this variable when analysing independent variables “exports” “year” and “gdp”.

## **5.5. Identifying X variables and Y target.**

Before ML section, it is needed to select the X and y target variable as below:

|  |  |
| --- | --- |
| **X Variables**  **Columns [ 0,1,2,3 & 5]** | **Y (Target)**  **Column “4”** |
| Country | Imports |
| Year |  |
| Group |  |
| Exports |  |
| Gdp |  |

Independent variables are not highly correlated among them. In contrast, the target variable (“imports”) is strongly correlated with one independent variables “exports”. “Codes” feature has been excluded from the analysis as it did not produce any correlation.

## **5.6. Feature Engineering**

### **5.6.1. Categorical Data Encoding.**

Categorical data for “country” and “group” variables need to be transformed into an array of numbers and created dummy variables as regression models read better numeric values. One Hot encoder was used to encode my categorical data, and then concatenate the variables (Iqbal, 2023). Another technique that could have been used to transform categorical data is Label encoder (Iqbal, 2023).

### **5.6.2. Feature scaling.**

It was tested 3 techniques to scale the data as min-max, log transformation, and scaling of the data. It was also compared that the shape of the data has not changed after those techniques were applied (McQuaid, 2023). Log transformation was ruled out as that changed the shape of the data. It was decided to use scale as this function standardises each column as follows:

x = x-mean(x) / std(x)

(Ibqal, 2023)

## **5.7. Implementation of Machine Learning Models**

### **5.7.1. Splitting Train and test set**

For all the ML regression models performed, test data was 0.20% and 0.80% for training data.

### **5.7.2. Support Vector Regressor**

SVR kernel chosen was RBF, but tuning the hyperparameters gamma is not that easy (Saini, 2023). R2 score 0.74 was produced and 0.77 after GridSearchCV was implemented. However, RMSE is 9.50.

### **5.7.3. Decision Tree and Random Forest Regressors.**

Both models are not sensitive to outliers neither feature scaling selection. Both performed well, DTR, with a max\_depth = 5 produced an R2 of 0.91. Similarly, RFR yields an R2 of 0.97%. Both models produced a better RMSE, DTR 5.74 and RFR with 3.025 respectively.

### **5.7.4. Linear Regression.**

It was applied this technique taking into consideration that I do not have a large dataset with many variables. However, it was performed to make comparisons with the other regression models, which produced an R2 score of 0.80 and RMSE of 8.84.

### **5.7.5. Ridge and Lasso Regression.**

GridSearchCV was applied with alpha regularization parameters of [0.001, 0.01, 0.1, 1] respectively as tuning these alpha parameters is very important (Iqbal, 2023), but they did not produce better results. However, both scored high in RMSE, 56.78 and 56.11 respectively.

## **5.8. Comparison ML model performance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regression Model** | **Train score** | **Test score** | **R2 score** | **Applying**  **GridSearch CV** | |
| **Train score** | **Test score** |
| Support Vector | 0.68 | 0.74 | 0.7401 | 0.7959 | 0.7757 |
| Decision Tree | 0.9657 | 0.9180 | 0.9180 |  |  |
| Random Forest | 0.9975 | 0.9772 | 0.9772 |  |  |
| Linear Regression. | 0.8255 | 0.8055 | 0.8055 |  |  |
| Ridge Regression. |  |  |  | 0.8254 | 0.8056 |
| Lasso Regression. |  |  |  | 0.8254 | 0.8058 |

**Discussion:** After performing the ML models, DTR R2 score was 0.9180 and RFR with 0.9772. Outliers were not completely removed, and these two models are not sensitive to extreme values. SVR produced an R2 score of 0.7401 and after performing GridSearchCV, the test score was improved to 0.7757. Additionally, LR produced an R2 score of 0.8055, which was not really improved after applying RR GridSearchCV, which produce a test score of 0.8056 and LSR also produced a test score of 0.8058. Overall, the R2 did not improved noticeable after RR and LR models were applied.

## **5.9. Conclusion**

Through the application of the 6 ML models, it was possible to predict the imports target variable. Even though all algorithms yield an R2 score above 0.70%, the ones that would be recommended would be DTR with R2 0.91 and RFR with R2 0.97 to make a more accurate prediction.

# **6. Sentiment Analysis.**

It was performed an analysis from the text scrapped from Reddid of the word “Construction”. Subsequently, it was extracted text, numeric characters, word count, stopwords, average of words using lambda function. All the different actions performed will be explained in detail in the accompanying Jupiter notebook.

Text was cleaned afterwards and Bag of Words for the pre-processed text data using the Count Vectorization and TF-IDF Vectorization approach. A MultinomialNB classifier was trained with 0.30 % to produce the following results:

|  |  |  |
| --- | --- | --- |
| **Classification Model** | **Approach** | |
| **CountVectorizer** | **TF-IDF** |
| Multinomial NB | 0.70 | 0.80 |

Nevertheless, the accuracy yield using CountVectorizer approach would be recommended over TF-IDF as this one produced a precision, recall and f1-score for both categories of sentiment score (0 and 1).

Additionally, the text was analysed to extract the positive, negative, neutral and compound overall sentiment of one piece of text from “Construction” comments.

The previous result shows that the there is no negative information (neg=0). There is some neutral and positive tones (neu=0.682 and pos=0.318). There is a positive tone (0.318). However, the overall sentiment is positive because compound is > 0.05.

# **7. Dashboards (Interactive visualizations)**

# First one using Plotly express library, which allowed me to visualize the different fluctuations in the GDP on the 8 countries chosen,

# Second one was two geo data visualization, which display the data values colouring the countries chosen by region map choropleth, the chosen colour was plasma as this colour shades will be changing according of the year fluctuations in Gdp and Imports (Europe and Earth zone) (McQuaid, 2023).

* Additionally, some pieces of code were saved as “slide” so they can be visualize to display some important insight from the analysis.

# **Other libraries different to pandas for Data Manipulation ( Appendix 4)**

# **9. Appendices**

1. Datasets Data Description

1.1. Exports materials dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Feature** | **Variable type** | **Data type** |
| 1 | Country | Text | String |
| 2 | Variable (Exports) | Text | String |
| 3 | Group | Text | String |
| 4 | Unit | Numeric | Float64 |
| 5 | Date | Numeric | Integer |
| 6 | Value | Numeric | Float64 |

1.2. Imports materials dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Feature** | **Variable type** | **Data type** |
| 1 | Country | Text | String |
| 2 | Variable (Imports) | Text | String |
| 3 | Group | Text | String |
| 4 | Unit | Numeric | Float64 |
| 5 | Date | Numeric | Integer |
| 6 | Value | Numeric | Float64 |

1.3. Gdp dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Feature** | **Variable type** | **Data type** |
| 1 | Country | Text | String |
| 2 | Variable (Gdp) | Text | String |
| 3 | Group | Text | String |
| 4 | Unit (NaN) | Numeric | Float64 |
| 5 | Date | Numeric | Integer |
| 6 | Value | Numeric | Float64 |

1.4. Codes dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **No.** | **Feature** | **Variable type** | **Data type** |
| 1 | Iso3 | Text | Object |
| 2 | Name | Text | Object |

2. Confidence intervals.

2.1. # Create 90% confidence interval – Imports Ireland

st.t.interval(alpha=0.90, df=len(Test\_df\_array)-1,

loc=np.mean(Test\_df\_array),

scale=st.sem(Test\_df\_array))

# Create 99% confidence interval

st.t.interval(alpha=0.99, df=len(Test\_df\_array)-1,

loc=np.mean(Test\_df\_array),

scale=st.sem(Test\_df\_array))

2.2. # Create 95% confidence interval – Exports Ireland

st.t.interval(alpha=0.95, df=len(Test\_df\_array\_exports)-1,

loc=np.mean(Test\_df\_array\_exports),

scale=st.sem(Test\_df\_array\_exports))

# Create 99% confidence interval

st.t.interval(alpha=0.99, df=len(Test\_df\_array\_exports)-1,

loc=np.mean(Test\_df\_array\_exports),

scale=st.sem(Test\_df\_array\_exports))

3. Probabilistic tests.

3.1. One tailed test ( Ireland Imports variable )

from scipy import stats

stats.ttest\_1samp(test\_df['imports'], popmean=4, alternative = "greater")

#test\_df["imports"].mean()

3.2. Kolmogorov-Smirnov test - Imports variable.

xbar = test\_df.imports.mean()

s = test\_df.imports.std()

# mean.imports = pd.mean(test\_df.imports)

stats.kstest(test\_df.imports, 'norm', args = (xbar, s))

3.3. # Perform the one-way ANOVA test - does the mean exports per country differ?

import scipy.stats as stats

stats.f\_oneway(

\*(construction.loc[construction['country']==country, 'exports']

for country in construction['country'].unique())

)

3.4. # perform Tukey's test

import pandas as pd

import numpy as np

from scipy.stats import f\_oneway

from statsmodels.stats.multicomp import pairwise\_tukeyhsd

tukey = pairwise\_tukeyhsd(endog=construction['exports'],

groups=construction['country'],

alpha=0.05)

#display results

print(tukey)

3.5. # Visualizing the results of the Anova test through a Boxplot:

sns.set(rc={"figure.figsize":(8, 10)})

sns.boxplot(data=construction, x="country", y="exports")

plt.title('Box plot of Exports by Country',

fontsize=13);

3.6 # Subsetting:

df1 = Test\_df2[(Test\_df2.country=='Ireland')]

df2 = Test\_df2[(Test\_df2.country=='Denmark')]

# Perform the test:

# Import the library

import scipy.stats as stats

# Perform the two-sample t-test with equal variances:

stats.ttest\_ind(a=df1['exports'], b=df2['exports'])

3.7. stats.mannwhitneyu(x=df1['exports'], y=df2['exports'], method="exact")

3.8 # An appropriate visualization plot to see this:

import seaborn as sns

import matplotlib.pyplot as plt

sns.set(rc={"figure.figsize":(6, 5)})

# Set a grey background (use sns.set\_theme() if seaborn version 0.11.0 or above)

sns.set(style="darkgrid")

sns.histplot(data=df1, x="exports", color="red", label="Ireland", kde=True)

sns.histplot(data=df2, x="exports", color="dodgerblue", label="Denmark", kde=True)

plt.title('Exports Ireland - Denmark (Two-sample t-tests results)',

fontsize=13)

plt.legend()

plt.show()

# **Appendix 4 : Other libraries different to pandas for Data Manipulation.**

Pandas is the most popular library for aggregating, manipulation, and data analysis, which with a clean syntax, is also easy to use (Yildirim, 2022). However, it can perform slow for operations on large datasets, which is its biggest disadvantage (Tan, 2022). In contrast, there are other libraries as the alternative of pandas, which similar syntax, which are able to handle large datasets (Tan, 2022) such as Polars, has been designed to process data faster and can be used an alternative to pandas. Polars can be compared to Pandas as they shared series and DataFrames building blocks (Tan, 2022). However, pandas and polars differs in terms of grouping and aggregation, but on both libraries. agg() and .groupby() methods can be used.

Dask is another library, which is similar to pandas for analysis and data manipulation in Python, In contrast to pandas, Dask can handle larger datasets and pandas is more suitable for smaller data sets (Pedersen, 2023).

# **9. References**

Arigoni, F., Kennedy, G., and Killeen, N. (2022). Central Bank of Ireland. [online] Available at: [No.12 Rising construction costs and the residential real estate market in Ireland (Arigoni, Kennedy and Killeen) (centralbank.ie)](https://www.centralbank.ie/docs/default-source/publications/financial-stability-notes/rising-construction-costs-and-the-residential-real-estate-market-in-ireland.pdf) [Accessed 20 May 2023].

Bruce, P., Bruce, A., and Gedeck, P. (2020) Practical Statistics for Data Scientists 50 + Essential Concepts Using R and Python. California, O’Reilly Media, Inc.

Diagrams.net (2023). [online] Available at: [CRISP DM Flow chart.drawio](https://1drv.ms/u/s!Aq0wa_kD53uikDxNP9YJRIX5AYjF?e=02X96L) [Accessed 24 May 2023].

Food and Agriculture Organization of the United Nations (2023). ISO 3 Code list (Global, Region). [online] Available at: [ISO 3 Code list (Global, Region, Country) - Datasets - "FAO catalog"](https://data.apps.fao.org/catalog/dataset/iso-3-code-list-global-region-country) [Accessed 15 May 2023].

Iqbal, M. (2023). Case Study (Linear SVM). Machine Learning for Data Analytics module. [Accessed 20 May 2023].

Iqbal, M. (2023). Linear Models for Regression & Classification Week 3 lecture. Machine Learning for Data Analytics module. [Accessed 20 May 2023].

Iqbal, M. (2023). Case Study (Linear, Ridge and Lasso Regression). Machine Learning for Data Analytics module. [Accessed 20 May 2023].

Hack, S. (2019) Machine Learning for Beginners: A Math Guide to Mastering Deep Learning And Business Application. Understand How Artificial Intelligence, Data Science, And Neural Networks Work Through Real Examples. First Edition, Italy, Amazon Italia Logistica.

Knoema (2023). [Material resources. [online] Available at: Public Knoema Data Hub](https://public.knoema.com/rzrojv/material-resources) [Accessed 10 May 2023].

# McQuaid, D. (2023). Data Visualization and Communication. What visualization should I used? Data Preparation and Visualization module. (Slides 5,6,10, 12) [accessed 24 May 2023].

McQuaid, D. (2023). Feature Scaling or Normalization. Data Preparation and Visualization module. (p 1) [Accessed 23 May, 2023].

Moez, A. (2022) Supervised Machine Learning. [online] Available at: <https://www.datacamp.com/blog/supervised-machine-learning> [Accessed 20 May 2023].

Linesight (2022). Ireland, Country Insights and Commodity Report, Q1 2022. [online] Available at: [Ireland Q4 2021 (ctfassets.net)](https://assets.ctfassets.net/1lsus2dflm8x/2rcZHm7dUkbLVP6EpQJFys/d773f9c56c4938de39f46f6571a4188b/Linesight_Ireland_Country_Insights_and_Commodity_Report_Q1_2022.pdf) [Accessed 20 May 2023].

Tufte, E. (2001). The Visual Display of Quantitative Information. The Second Edition, United States, Graphics Press LLC.

OECD.Stats (2023). Material resources. [online] Available at: [Material resources (oecd.org)](https://stats.oecd.org/Index.aspx?DataSetCode=MATERIAL_RESOURCES) [Accessed 10 May 2023].

Pedersen, U., T. (2023). Python Pandas vs. Dask DataFrames: A Comparative Analysis. [online] Available at <https://pub.towardsai.net/python-pandas-vs-dask-dataframes-a-comparative-analysis-c0f59dad5eeb> [ Accessed 25 May 2023 ].

Saini, A. (2023). Support Vector Machine(SVM): A Complete guide for beginners. [online] Available at: <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/> [Accessed 22 May 2023].

Stanford Libraries (2023) Welcome to the Public Domain. [Welcome to the Public Domain - Copyright Overview by Rich Stim - Stanford Copyright and Fair Use Center](https://fairuse.stanford.edu/overview/public-domain/welcome/) [Accessed 22 May 2023].

Statology (2023). How to Calculate Confidence Intervals in Python [online] Available at: <https://www.statology.org/confidence-intervals-python/> [ Accessed 20 May 2023]

Tan, E. (2022). 8 Alternatives to Pandas for Processing Large Datasets. [online] Available at: <https://towardsdatascience.com/8-alternatives-to-pandas-for-processing-large-datasets-928fc927b08c> [Accessed 22 May 2023].

Yildirim, S. (2022). 3 Methods for Aggregating Data with Python Pandas [online] Available at: <https://towardsdatascience.com/3-methods-for-aggregating-data-with-python-pandas-14ceb75b6f6e> [Accessed 22 May 2023].