**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*  *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. 3](#_Toc135955128)

[1. Abstract 3](#_Toc135955129)

[2. Introduction. 3](#_Toc135955130)

[3. Data Base Overview – Creative Commons Data. 3](#_Toc135955131)

[4. EDA Analysis 4](#_Toc135955132)

[4.1. Data Frames creation. 4](#_Toc135955133)

[4.1.1. Exports materials: 4](#_Toc135955134)

[4.1.2. Imports materials: 5](#_Toc135955135)

[4.1.3. GDP: 5](#_Toc135955137)

[4.2. Analysis and merging of dataframes. 6](#_Toc135955138)

[4.2.1. NaN values: 7](#_Toc135955139)

[4.2.2. Outer Merge: 7](#_Toc135955140)

[4.2.3. Dropping NaN values after merging: 7](#_Toc135955141)

[4.3. Descriptive Statistics. 8](#_Toc135955142)

[4.4. Detecting categorical data. 8](#_Toc135955143)

[4.5. Inferential Statistics. 9](#_Toc135955144)

[4.5.1. Ireland Imports. 9](#_Toc135955145)

[4.5.2. Ireland Exports. 9](#_Toc135955146)

[4.6.3. ANOVA test (parametric test) – Population: 8 countries vs Exports feature. 10](#_Toc135955147)

[4.6.4. Tukey-Kramer analysis (non-parametric test) 11](#_Toc135955148)

[4.6.6. Two-sample t-tests (Ireland and Denmark exports) 12](#_Toc135955149)

[4.6.7. The Mann Whitney U test - (Ireland and Denmark exports - non-parametric test) 13](#_Toc135955150)

[4.7. Correlation of the Variables. 14](#_Toc135955151)

[4.7.1. Visualizing the correlation between Imports and Gdp features. 14](#_Toc135955152)

[4.8. Positive Skewed to the right data shape 15](#_Toc135955153)

[4.9. Outliers Detection. 16](#_Toc135955154)

[4.9.1. Imports feature outliers 16](#_Toc135955155)

[4.9.2. Exports feature outliers 16](#_Toc135955156)

[4.9.3. Exports feature outliers 17](#_Toc135955157)

[4.9.3. Winsorisation of Outliers. 17](#_Toc135955158)

[4.9.4. Verifying that shape after applying Winsorization of outliers. 18](#_Toc135955159)

[5. Machine Learning. 18](#_Toc135955160)

[5.1. CRISP DM Methodology. 18](#_Toc135955161)

[5.3. Supervised Machine Learning models. 19](#_Toc135955162)

[5.4. Detected issue and Strategy of Analysis. 19](#_Toc135955163)

[5.5. Identifying X variables and Y target. 19](#_Toc135955164)

[5.6. Feature Engineering 20](#_Toc135955165)

[5.6.1. Categorical Data Encoding. 20](#_Toc135955166)

[5.6.2. Feature scaling. 20](#_Toc135955167)

[5.7. Implementation of Machine Learning Models 21](#_Toc135955168)

[5.7.1. Splitting Training and test set 21](#_Toc135955169)

[5.7.2. Support Vector Regressor 21](#_Toc135955170)

[5.7.3. Decision Tree and Random Forest Regressors. 21](#_Toc135955171)

[5.7.4. Linear Regression. 21](#_Toc135955172)

[5.7.5. Ridge and Lasso Regression. 21](#_Toc135955173)

[5.8. Comparison ML model performance 21](#_Toc135955174)

[5.9. Conclusion 22](#_Toc135955175)

[7. Appendices 24](#_Toc135955176)

[8. References 24](#_Toc135955177)

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

Author: Maria Dominguez, MScs Data Analytics student

Student ID: 2019008

# **1. Abstract**

# **2. Introduction.**

# **3. Data Base Overview – Creative Commons Data.**

Four datasets, which were my row data, were chosen to produce this academic report. 3 of them (import materials,csv”, “exports materials csv” and “gdp csv” which have been taken from the datasets Knoema site (Knoema.com, 2023), all of these are about construction materials (construction minerals, metals a and non-minerals materials). The original data source was taken from OECD site, (OECD.Stats, 2023), the tree datasets downloaded are public domain. I did also contact the original source who confirmed that those data are public and can be used in this study as long as it is not used for commercial pourpuses and the original source (OECD) is well reference in the report. The dataset “Iso3” was downloaded from FAO site (FAO, 2023), which is open data under the licence “Attibution-NonCommercial-ShareAlike 3.0 IGO”, which allowed me to use the data as long as it is not used for commercial purposed and also be properly referenced in this academic paper. The positive aspects that I gain from this search is that it help me to feel more confident about chosen datasets in future. However, perhaps a negative aspect of this research, was the fact that many good datasets were only available in paid sites, which memberships were not affordable. Nevertheless, Knoema site is a paid site, but inexpensive. Additionally, the search of the dataset was time consuming, from my view, the data from the government sites were just concerning to Ireland and not about the others countries in Ireland. Nevertheless, I took it that the ultimate purpose of this activity is to develop the ability of searching for suitable data, which is a reality for any future projects.

# **4. EDA Analysis**

## **4.1. Data Frames creation.**

I have performed all the different stages of the EDA analysis, which are going to be explained below and all included in the accompanied Jupiter notebook.

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

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

* The type of dataset 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 containing numeric, continuous data, 3 columns with categorical data and 1 feature “unit” will be dropped as only contains “NaN” values and it will not produce any meaning full result in my analysis.

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

A screenshot of a data

Description automatically generated with low confidence

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

### Original csv file name was downloaded as “ObservationData\_iwkolud.csv” and it was renamed as "Imports materials.csv” to avoid any confusion when uploading datasets in the Jupiter notebook.

* The type of dataset is continuous, and it measures all the units of imports (construction minerals, metals and non-metal materials) for the same 8 countries of the eurozone listed above.
* There is a total of 627 observations and 6 features in this dataset.
* It contains 2 features containing numeric, continuous data, 2 columns containing categorical data and 1 feature containing only NaN values, which will be deleted before merging.
* Description of features of dataset ( see appendix 1 )

### **4.1.3. GDP:**

Original csv file name was downloaded as “ObservationData\_jhykshd.csv” and renamed as "Gdp materials.csv” to avoid any confusion when uploading files.

* The type of dataset is continuous, and it measures all the gdp values of imports (construction minerals, metals and non-metal materials) for the same 8 countries of the eurozone listed above.
* There is a total of 625 observations and 6 features in this dataset.
* It contains 2 features containing numeric, continuous data, 2 columns containing categorical data and 1 feature containing only NaN values, will be deleted before merging.

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

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

Description automatically generated**

**Codes (Iso 3 dataset)**

* The type of dataset is continuous, and it measures all the gdp values of imports (construction minerals, metals and non-metal materials) for the same 8 countries of the eurozone listed above.
* There is a total of 274 observations and 2 features in this dataset. 2 columns of categorical data.

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

A screenshot of a computer

Description automatically generated with medium confidence

# **Analysis and merging of dataframes.**

### **NaN values:**

Null values were detected on the “value” feature of 3 dataferames (exports, imports, gdp) which were dropped as it will not yield any valuable information.

### **Outer Merge:**

This method was chosen to merge the datasets as I wanted to keep information from both datasets. It was first merged exports with imports dataframes in a new dataframe “const”. Then, it was created a new data frame “const1! To combine const with “gdp dataframe”. Finally, it was created a new dataframe “construction” to merge “const1 with “codes dataframe”. Whichever information that is 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 which were filled with NaN values, which were dropped.

After dropping null values, the “construction” dataframe will contains 7 features and 601 observations.

A screenshot of a computer

Description automatically generated with low confidence

|  |  |  |  |
| --- | --- | --- | --- |
| **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 |

## Descriptive Statistics.

After producing the summary statistics for the numerical features (see Figure X below), it can be identified the following:

* 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.
* When checking the standard deviation values, it can be seen that the features “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 x below ) depict the top categories and the number of times those categories are display in the dataframe. As it can be seen below, 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. Similarly, for “country” and “code features have 8 unique values ( 8 countries chosen) and “ group” feature has 3 unique values (construction minerals, metals and non-minerals). The feature engineering method needs to be taken into account before performing ML models.

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 1.1.)

**Interpretation of results:**

The 90% confidence interval for the true population mean is (3.67-4.45). The larger the confidence level. 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 ( GeekforGeeks, 2023).

### 4.5.2. Ireland Exports.

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

**Interpretation of results:**

The 95% confidence interval for the true population exports mean is (1.59-2.14).The larger the confidence level, 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.

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

**4.2.1. One tailed test - Ireland imports variable**

**Hypothesis:**

* 𝐻0: M = 4 and 𝐻A > 4
* Level of significance is 𝛼 = 0.05.
* **Result:** pvalue=0.38923507692394965

**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. (See appendix xx

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

**Hypothesis**:

* 𝐻0 = Data is consistent with normal distribution.
* 𝐻𝐴 = Data is not consistent with normal distribution.
* **Pvalue**=0.11720684769232348

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

**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

**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

(Figure xx)

**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.
* **France and Ireland**: Since the confidence interval is only in negative numbers, I have evidence of the statistical difference between these two countries in terms of exports.
* **Germany and Ireland**: Since the confidence interval is only in negative numbers, I have evidence that there is 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 a difference among countries populations (exports). I did use this visualization as I needed to visualize the 8 countries exports distributions at once (McQuaid, D., 2023). As well as that, the clarity of the visualization must be maintained and 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 1.1.)

### 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 those two countries.
* Level of significance is 𝛼 = 0.05.

**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.
* Level of significance is 𝛼 = 0.05.
* Pvalue=2.1127054483208824e-19.

**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, D., 2023).

A picture containing text, screenshot, diagram, plot

Description automatically generated

Histogram – (Appendix 1.1.)

## 4.7. Correlation of the Variables.

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

**A screenshot of a computer

Description automatically generated with low confidence**

Heat map for correlation of variables (Appendix 1.xx.)

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

Scatterplots were used to provide an excellent visualization of the relationship of two features (McQuaid, D., 2023).

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

Description automatically generated**

Scatter plot to check a moderate correlation between two variables (Appendix 1.3.)

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

Description automatically generated**

Scatter plot to check a strong correlation between two features (Appendix 1.3.)

## 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, D., 2023), which is positive skewed to the right.

A picture containing text, screenshot, plot, diagram

Description automatically generated

## 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 section. The visualization of those extreme values, was obtain through box plots (McQuaid, D., 2023)

A picture containing screenshot, diagram, rectangle, line

Description automatically generated

### 4.9.2. Exports feature outliers

A picture containing screenshot, rectangle, diagram, line

Description automatically generated

### 

### 4.9.3. Exports feature outliers

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

Description automatically generated**

### 4.9.3. Winsorisation of Outliers.

This technique was used in ML in tutorial 3 (a) solution (class when performing Linear and other regression models Lasso and Sasso regression. I decided to apply this technique so I did not eliminate all my outliers

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

Description automatically generated**

### 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, D., 2023), which still remain.

**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**

# 5. Machine Learning.

## 5.1. CRISP DM Methodology.

[A diagram of data processing

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

**Figure – CRISP DM Phases – Construction Materials**

The project prepared about Construction Materials for 8 countries of the eurozone, is applicable to the CRISP DM methodology, which requires the understanding of the variables chosen, obtain the dataset from different sources and analyse the data, which involves the next step of data preparation, handling outliers, removing null values. Modelling was implemented with 6 ML for regression, which have been evaluated with metrics for regression. Finally, the deployment stage consists on the preparation of this report, which contains the models that can be recommended.

## **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 is continuous “Imports”. Supervised algorithms SVM, DTR, RFR, LR, RR, and LR algorithms, which are included in this supervised learning models, which will learn from the training data to be able to make a prediction. (Moez, 2022).

## 5.4. Detected issue and Strategy of Analysis.

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

Before ML section, it is needed to select the X and y target variable taking into account the strong correlation between “imports and “exports” features as follows:

|  |  |
| --- | --- |
| **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 strong correlated with one independent variables. “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 that need to be transformed to 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. 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 changed the shape of the data. It was decided to use scale as this function standardises each column.

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

(Ibqal, 2023)

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

6 supervised ML models for regression have been applied, trained as I have continuous data. Outliers were not removed, but winsorized instead as outliers make algorithms in regression models difficult to perform (Bruce et al., 2020) and I would like to see how this ML models perform after this technique was applied.

## 5.7.1. Splitting Training 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

This technique was used knowing that is sensitive to outliers ( Kanamori et al, 2014) and also choosing the kernel (KRF for this model) and but tuning the hyperparameters gamma is not that easy (Saini, 2023). Hence, the results were slightly improved after the GridSearchCV was implemented.

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

Both models are not sensitive to outliers neither feature scaling selection. Both performed well, DTR produced an R2 of 0.91. Similarly, RFR yield an R2 of 0.97%. Both models produced a good root mean squared error, DTR 3.025 and RFR with 5.74 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.

### 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). Even though both scored slightly better in test set in comparison with Linear Regression model. Both scored high in the mean absolute percentage error of prediction, 56.88 and 56.15 respectively.

## 5.8. Comparison ML model performance

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Regression Model** | **Train score** | **Test score** | **R2 score** | **Applying**  **GridSearch CV** | |
| **Train score** | **Test score** |
| Support Vector Regressor |  |  | 0.7401 | 0.7959 | 0.7757 |
| Decision Tree Regressor | 0.9657 | 0.9180 | 0.9180 |  |  |
| Random Forest Regressor | 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, DTR performance 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 improved slightly by applying RR GridSearchCV, which produce a test score of 0.8056 and LR also produced a test score of 0.8056. 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.

1. **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 panda por 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).

# 7. Appendices

# **8. References**

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].

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

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 SVM). Machine Learning for Data Analytics module. [Accessed 20 May 2023]

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

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

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

Kanamori, T., Fujiwara, S., and Takeda, A., (2014) Breakdown Point of Robust Support Vector Machine. Cornell University [online] Available at: [https://arxiv.org/abs/1409.0934#](https://arxiv.org/abs/1409.0934) [ Accessed 07 April 2023].

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]

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].

# ISO 3 Code list (Global, Region, Country)