# **CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a Word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* |
| **Assessment Title:** | *CA2* |
| **Lecturer Name:** | *Marina Lantorno*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | *Marcelo Alves Queiroz* |
| **Student Number:** | *2022010* |
| **Assessment Due Date:** | *22th May 2022* |
| **Date of Submission:** | *20th May 2022* |

## 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.

Summary – Import and Export Production

[CCT College Dublin 1](#_Toc104019601)

[1](#_Toc104019602)

[Declaration 1](#_Toc104019603)

[1. ABSTRACT 3](#_Toc104019604)

[1.1. METRICS, PROBLEM-SOLVING, AND BUSINESS UNDERSTANDING 3](#_Toc104019605)

[1.2. THE CAP IN PRACTICE 4](#_Toc104019606)

[1.3. PROJECT MANAGEMENT 4](#_Toc104019607)

[2. ANALYSIS 5](#_Toc104019608)

[2.1. GATHERING DATA & CLEANING PROCESS - EDA 5](#_Toc104019609)

[2.2. FEATURE ENGINEERING 5](#_Toc104019610)

[2.3. FEATURE EXTRACTION 5](#_Toc104019611)

[2.4. MISSING VALUES 5](#_Toc104019612)

[2.5. NORMALIZING AND STANDARDIZING CONTINUOUS FEATURES 6](#_Toc104019613)

[2.6. FEATURE SCALING 6](#_Toc104019614)

[2.7. OUTLIERS WITH SCATTERPLOT 6](#_Toc104019615)

[2.7.1. ARE THERE ANY OUTLIERS IN OUR DATA? 6](#_Toc104019616)

[2.8. CORRELATION AND HEATMAP 7](#_Toc104019617)

[2.8.1. HOW MANY INDEPENDENT AND DEPENDENT-ON FEATURES DO WE HAVE? 7](#_Toc104019618)

[3. MACHINE LEARNING MODELS 8](#_Toc104019619)

[3.1. KNN – NEIGHBOURS CLASSIFIER 8](#_Toc104019620)

[3.2. SVM – SUPPORT VECTOR MACHINE 9](#_Toc104019621)

[3.2.1. GRID SEARCH CV 9](#_Toc104019622)

[3.3. RIDGE, LINEAR, AND LASSO REGRESSION 10](#_Toc104019623)

[3.4. RANDOM FOREST 10](#_Toc104019624)

[4. RESULTS 10](#_Toc104019625)

[4.1. OPEN DASHBOARD FOR FURTHER VISUALIZATIONS! 11](#_Toc104019626)

[5. WHAT PEOPLE FROM IRELAND ARE THINKING ABOUT BIG DATA AGRICULTURE? 11](#_Toc104019627)

Import and Export Prediction of Arable Production in Ireland Among Countries Around the World

A Data Analysis Study using Machine Learning Algorithms to Predict Import and Export Quantity and Prices for Next Season

# **ABSTRACT**

# **METRICS, PROBLEM-SOLVING, AND BUSINESS UNDERSTANDING**

If it’s not the one, agriculture trade is one of the most vital sectors for economic health for the countries. Ireland is well engaged in trade, and the Irish status is a major food-exporting nation, according to the department of agriculture, food, and the marine.

This project approach and implement machine learning algorithms and data mining methods to analyze arable production in Ireland and then, compare the Irish Agriculture Production among other countries with a similar population, which are going be: Denmark, Finland, and Norway.

All types of crops will be analyzed and compared over the last 20 years, data from FAOSTAT has been gathered for this study case including import and export quantity, types of crops, unit, and value per year, temperature, countries, and type of crops.

"Agriculture hasn't always been an industry people associate with big data analytics or data centers, but that's changing. From time immemorial, farmers have been quick to turn to new technologies that allow them to improve their crop yields and work more efficiently. As big data analytics have become more widely available across industries, many farmers and companies in the agricultural sector are using those tools to become more productive and competitive." (McGuirre, June 8, 2018)

In this analysis, machine learning will be implemented to create a model to predict the 10 most import and export crops and values using the dataset from FAOSTAT. A very well-known modeling technique with works out the relationship between a dependent and independent variable is called Regression Analysis, the models used in this project are:

* Random Forest Regression
* Support Vector Machine
* Decision Tree Regression
* KNN Neighbours Classifier
* Linear Regression
* Lasso Regression

Based on recommendations and predictions from data of import and export, we can advise farmers on what to expect for this area following harvest, with import and export predictions we can prepare for the supply and demand of the country, and we could even play with the prices, thanks to the **Common Agricultural Policy**, playing with prices is not allowed, lunched in 1969 the CAP is a partnership between agriculture and society and farmers among European countries.

# **THE CAP IN PRACTICE**

“Farming is unlike most other businesses, as the following special considerations apply:

* despite the importance of food production, farmers’ income is around 40% lower compared to non-agricultural income;
* agriculture depends more on the weather and the climate than many other sectors;
* there is an inevitable time gap between consumer demand and farmers being able to supply – growing more wheat or producing more milk inevitably takes time.” (Europe, 2018)

The CAP is meant to help farmers with national strategic plans and measure the performance and results from farms, also implemented the environmentally friendly farm green.

*OBS: The 4 countries analyzed in this study are part of the European Union*

# **PROJECT MANAGEMENT**

The methodology used to understand and analyze the data is CRISP-DM. The most common methodology used nowadays in data analysis, from Daimler Chrysler (formerly Daimler-Benz), SPSS (then ISL), and NCR introduced the Cross-Industry Standard Process for Data Mining **(CRISP-DM)** in late 1996. This model illustrates the evolution of refinements throughout time. It is divided into six stages or phases, which will be followed as below:

Business

Understanding

Data

Understanding

Data

Preparation

Modeling

Evaluation

Deployment

**Data**

CRISP-DM Diagram. Inspired by [Medium](https://medium.com/analytics-vidhya/understanding-crisp-dm-and-its-importance-in-data-science-projects-91c8742c9f9b).

# **ANALYSIS**

# **GATHERING DATA & CLEANING PROCESS - EDA**

Datasets from FAOSTAT have been collected for this study as .csv a format which is famous and easy to read in Python, as our data is not too big, we should have no issue in regards to data format. The website contains data from all countries around Europe and the world, more than 300 types of crops, and import and export information such as price per tonne, it also offers huge statistics of national and international Agri data.

When applied **.describe()** some insights have been gained such as the years, FAOSTAT has data from 1744 until 2020, as our study will analyze data from the last 20 years, .query has been applied and filtered to show data only from this interval, and .query also on Area feature, to show only our desired countries previously described.

Another dataset is updated as temperature, which contains a median and standard deviation of the temperature per year over the past years, also from FAOSTAT.

When applied .merge() we've created a single dataset with columns from the import and export dataset and temperature and standard deviation from another, after merging these two datasets, we have a dataset of 125.787 rows and 7 columns.

# **FEATURE ENGINEERING**

After having a look at the dataset with **.describe()** and **.head()** we can notice that Import and Export values are all in one column as **Element**.

# **FEATURE EXTRACTION**

Feature extraction is the action of manipulating your dataset as gaining insights from your analysis,

After a quick look at the data, It's noticed that the columns Element has 4 other features that we could use in our Data Analysis, which are: Import Values, Export Values, and Quantity values were extracted from the **Element** feature and changed its names to make it easier to recognize that are data of import, export, quantity, and unit.

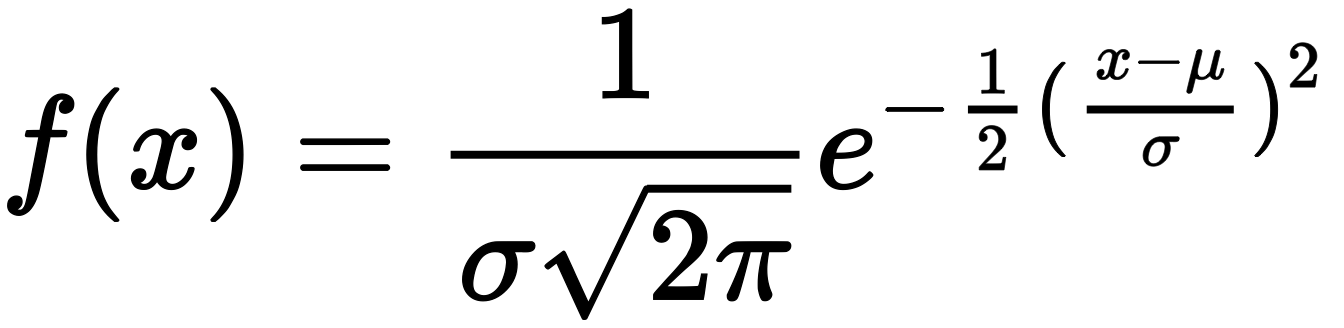
# **MISSING VALUES**

Using **.isnull().**sum we could have a look at missing values from our data set, which represent **2.38%** of da data, which is a very little percentage that will not affect our analysis, all missing data has been removed using **.isnan()** from the features import and export values.

# **NORMALIZING AND STANDARDIZING CONTINUOUS FEATURES**

After dealing with missing values and feature extraction, normalizing data became a standard procedure nowadays when it comes to continuous data analysis. **Normalization** function, native to **python library** has been applied to get all our continuous data at the same range.

Below formula for normalization:



*Formula Feature Normalization*

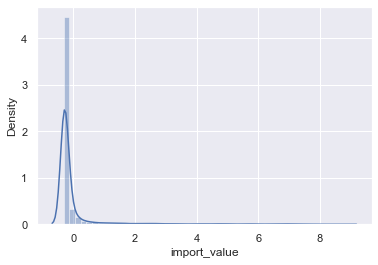
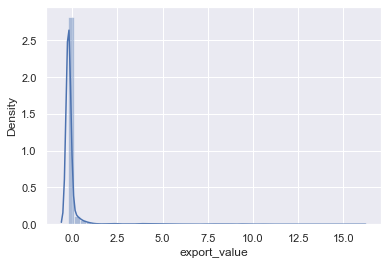
Data from Import and export features are now in the same range.

# **FEATURE SCALING**

After normalized our data, a distribution plot has been suggested to visualize the parametric distribution of our data “A picture is worth a thousand words".

*Data is a positive skew*

*Data is a positive skew*

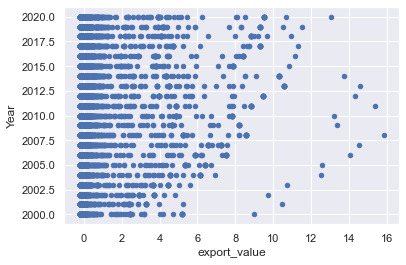


# **OUTLIERS WITH SCATTERPLOT**

# **ARE THERE ANY OUTLIERS IN OUR DATA?**

Let's torture our data a bit more, the definition of outliers is: "An outlier is a value that differs significantly from other values in a population's random sampling." (NIST, National Institute of Standards and Technology)

For the plot below is documented that a few numbers from our features export and import values are out of range:



*Export values* ***Outliers***

*Import values* ***Outliers***

Although we have outliers in our dataset, after careful investigation, we cannot remove them as contains valuable information about the import and exportation values.

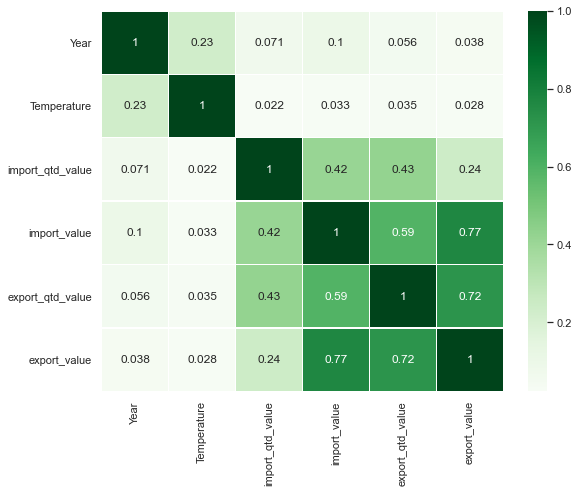
# **CORRELATION AND HEATMAP**

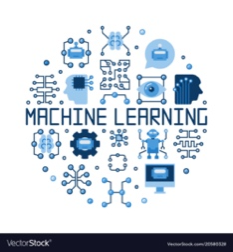
# **HOW MANY INDEPENDENT AND DEPENDENT-ON FEATURES DO WE HAVE?**

Exploring our data deeper with correlation and heatmap we can find the independent and dependent features just with a few lines of codes which will bring us a cool visualization with heatmap and, it is easier to check for correlation by visualizing something instead of looking at numbers.

Below heatmap with the color green as we are talking about agriculture data which is a natural color:

*Heatmap for correlation features*



****

# **MACHINE LEARNING MODELS**

Is there any similarity between import and export?

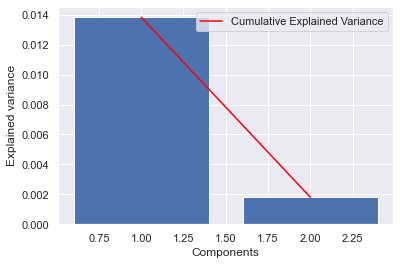
Which country imports more? Which country exports more?

Six different machine learning models will be built to answer those questions and predict how good or bad next season will be in regarding the import and export of grains production.

# **KNN – NEIGHBOURS CLASSIFIER**

K-Nearest Neighbors Classifier is the simplest and easier to apply for classification and regression problems, the KNN will look for similarities between import and export data and find if any similarities can help us to understand in deep and face our problem.

Below we can see the results applied PCA – Principal Component Analysis to reduce the redundancy in our data and dimensionality. Below graphic explaining the variance in the training and testing set:

****

Train score after PCA 0.5354601461387841 %

Test score after PCA 0.38845553822152884 %

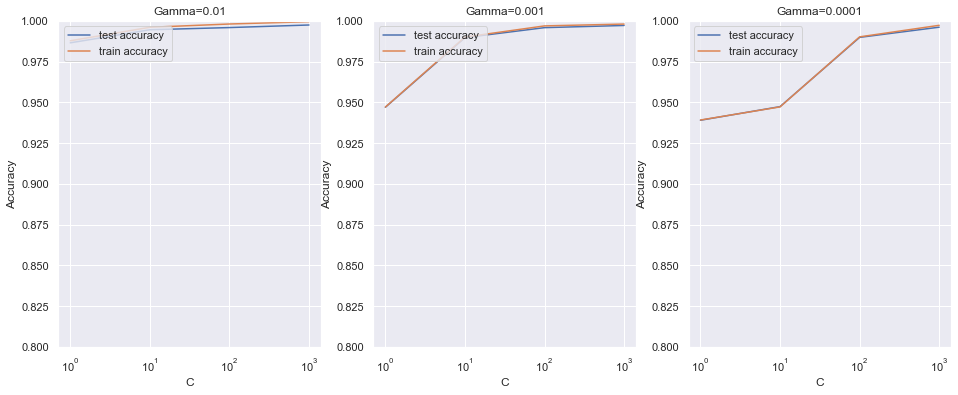
The results before and after applied PCA are the same!

# **SVM – SUPPORT VECTOR MACHINE**

Support Vector Machine is a supervised learning model with associated learning algorithms which can analyze classification and regression data for two groups of classification problems, After applying all types of kernels ('linear', 'poly', 'rbf', 'sigmoid', 'precomputed') we got a better result using **Sigmoid** kernel, the accuracy of 93%

# **GRID SEARCH CV**

Grid Search CV is a library function from sklearn model package, which is pre-defined hyperparameters and estimators, below different results from grid search cv with gamma parameters:

****

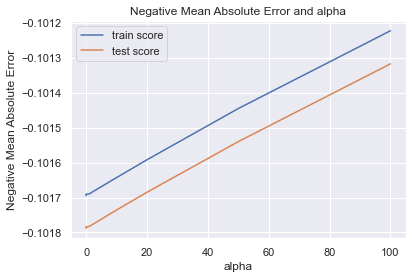
The best test score is 0.9974688106567173 corresponding to hyperparameters {'C': 1000, 'gamma': 0.01}

The higher accuracy as illustrated above with 0.99 corresponding to the gamma 0.01

# **RIDGE, LINEAR, AND LASSO REGRESSION**

Regression models can develop and help predict a continuous outcome, in this case, import and export prices and volume exported and imported, these three models have been implemented tunning **Alpha** parameters to better results, **Cross-Validation**, and **Grid Search CV.**

As the results, we’ve got R-squared of the training set and testing set, Root mean squared error, and mean absolute percentage error of the prediction

****

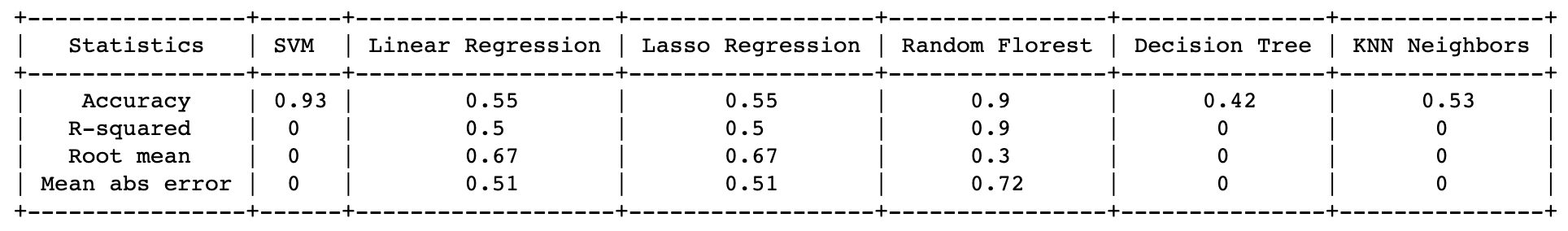
See above illustrations of negative mean absolute error and alpha from Jupyter notebook

# **RANDOM FOREST**

“Random forest is a commonly-used machine learning algorithm trademarked by Leo Bierman and Adele Cutler, which combines the output of multiple decision trees to reach a single result. Its ease of use and flexibility have fuelled its adoption, as it handles both classification and regression problems.” (IBM, no date)

The random forest has been applied in order to check the performance and compare it with the others models. The number of estimators applied was 100 and the random state was 42. Multiple combinations were applied, and the estimators and random state applied were the ones with the best performance

# **RESULTS**

Below table developed in python shows the results of each model with their respective particularities and differences with properly parameters side by side:

This research has thrown up many questions in need of further investigation for the sake of import and export volume. As per the graphics on the **Dashboard**, we can quickly see which country has had the highest volume of imports and export throughout the past 10 years:

# **OPEN DASHBOARD FOR FURTHER VISUALIZATIONS!**

Overall, all models performed well, SVM and Random Forest performed better in this case with an accuracy of nearly 90%, these two models would enhance the predictions.

Further work needs to be done such as crop predictions for future harvest so that farmers can prepare what do they need to get better results in regarding the future

Below we will perform a sentiment analysis about this theme, big data for agriculture.

# **WHAT PEOPLE FROM IRELAND ARE THINKING ABOUT BIG DATA AGRICULTURE?**

Sentiment Analysis is simple using machine learning to teach computers to extract sentiments from texts with word vectors corresponding to each word.

This project will focus on tweets from @Agriculture\_ie, the official Twitter account of Agriculture in Ireland to better understand what are they talking about and, how positive or negative these tweets are.

For data collection, API from Twitter will be used as a tool to gather our data, credentials, and permissions website official has been applied and requested elevated access for study purposes (Twitter API v2) has given access to 500 tweets per month which will be used on this research. Consumer Keys and Authentication Tokens have been properly set to gather the tweets.

Below is a word cloud of words that have appeared in the sentiment analysis from Twitter:



*WordCloud of Sentiment Analysis*