**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning  Data Preparation & Visualization |
| **Assessment Title:** | MSC\_DA\_CA2 |
| **Lecturer Name:** | Sam Weiss  John O’Sullivan/Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Mustafa Güneş Cetin |
| **Student Number:** | 2022456 |
| **Assessment Due Date:** | Friday, 6 January 2023, 11:59 PM |
| **Date of Submission:** | Friday, 6 January 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**

[1. Introduction 3](#_Toc123910318)

[2. Statistics for Data Analytics 3](#_Toc123910319)

[2.1. Descriptive Statics 3](#_Toc123910321)

[2.1.1. Frequency Distribution 4](#_Toc123910322)

[2.1.2. Central Tendency 4](#_Toc123910323)

[2.1.3. Variability 4](#_Toc123910324)

[2.2. Inferential Statics 5](#_Toc123910325)

[2.2.1. Parametric Inferential Statistical Techniques 5](#_Toc123910326)

[2.2.1.1. T-Test 5](#_Toc123910327)

[2.2.1.2. Analysis of Variance 6](#_Toc123910328)

[2.2.1.3. Pearson Correlation Test 6](#_Toc123910329)

[2.2.2. Non-Parametric Inferential Statistical Techniques 6](#_Toc123910330)

[2.2.2.1. Kruskal Wallis Test 7](#_Toc123910331)

[2.2.2.2. Chi-Square Test 7](#_Toc123910332)

[3. Data Visualization and Preparation 8](#_Toc123910333)

[4. Machine Learning for Data Analytics 11](#_Toc123910334)

[4.1. Sentiment Analysis 11](#_Toc123910339)

[4.2. Supervised Learning 13](#_Toc123910340)

[4.2.1. Random Forest Classification 13](#_Toc123910341)

[4.2.2. Logistic Regression 14](#_Toc123910342)

[4.3. Unsupervised Learning 14](#_Toc123910343)

[4.3.1. K-Means Clustering 15](#_Toc123910344)

[4.4. Forecasting 16](#_Toc123910345)

[5. Conclusion 17](#_Toc123910346)

[6. References 17](#_Toc123910347)

[7. Appendix 18](#_Toc123910348)

[7.1. Github wep URL 18](#_Toc123910349)

[7.2. Word count 18](#_Toc123910350)

1. **Introduction**

In recent years, there has been a significant increase in organic farms, including livestock, in European countries [1]. As a factor of this, activities such as import, export, production, and processing vary in countries that produce organic products. In addition, excess or low production also affects the market sector. With the development of the economy and internet platforms, users comment on agricultural products. Sentiment analysis studies come to the fore to evaluate these comments and produce them according to user needs in the sector [2,3].

* 1. **Project management framework**

Attention has been paid to the fact that the datasets used in the study contain data from each year of the European countries that are the subject of the research. This way, temporal analysis and comparison of all these data are possible. In addition, the selected datasets allow the estimation of data for future years in terms of being a prospective study.

In the study, "the organic operators by status" dataset is suitable for making predictions for the next year according to the values of organic operators in the past years. The organic production data of animal products and those produced by European countries for the last ten years have been analyzed, and statistical tests have been conducted. In addition, this dataset is a suitable dataset for unsupervised learning models as a label-free dataset.

# Statistics for Data Analytics



# Descriptive Statics

Descriptive statistics consist of short information coefficients that summarize a data set. The rows of the data set represent a population, and the columns, the attributes of the population, and descriptive statistics are calculated by taking samples from the population. Measures of central tendency include mean, median, and mode, while measures of variability include standard deviation, variance, minimum, and maximum variables.

# Frequency Distribution

The distribution is a summary of the frequency of individual values or ranges of values for a population. It depicts the frequency or count of the different outcomes in a data set and is presented in a table or a graph. The frequency of the occurrences of values in an interval, range, or specific group accompanies each entry in the table or graph. It allows for a more structured and organized way to present raw data.

# Central Tendency

Measures of central tendency focus on the average or middle values of data sets and describe the center position of a distribution for a data set. The mean, median, and mode are the measures of central tendency. The median refers to the middle score for a data set in ascending order. The mode refers to the score or value that is most frequent in a data set.

# Variability

|  |  |
| --- | --- |
|  | (1) |
|  | (2) |
|  |  |

A measure of variability is a summary statistic that presents the degree of distribution in a sample. The measures of variability determine how far apart the data points are from the center. The range represents the degree of dispersion or an ideal distance between the highest and lowest values within a data set. The standard deviation is used to determine the average variance in a data set and provide insight into the distance between a value in a data set and the mean value of the same data set. The standard deviation formula is shown in equation (1); here, n indicates the number of values and indicates each value of samples, while indicates mean of sample. The variance represents the degree of the spread and is essentially an average of the squared deviations as shown in equation (2).

# Inferential Statics

Statistical inference is a set of methods that allows the study of the effect of a behavior of particular population through a statistical sample. It also examines the degree of reliability of the inferences made. Inferences are tested with hypothesis testing methods. The goal is to check if an estimate corresponds to population values. All hypothesis tests have two assumptions the null hypothesis (H0) and the alternative hypothesis (H1). The null hypothesis reflects the idea that a value has a predetermined value. If the null hypothesis is rejected, the alternative hypothesis is accepted.

In the study, countries that are similar in terms of population were discussed and hypothesis tests were applied on the organic agriculture data of these countries. Looking at the population data for the last ten years, the population numbers of Finland, Ireland and Croatia are very close to each other.

Inferential statistical methods are divided into two groups parametric and non-parametric methods. Assumptions should be checked to decide which method to use:

* Normally distributed
* More than 30 samples
* The variance of the data is the same.

If these assumptions are satisfied, parametric statistical methods are used; if at least one of the first two assumptions cannot be satisfied, a non-parametric test is used.

# Parametric Inferential Statistical Techniques

Parametric statistics is a branch that accepts that the data comply with the random distribution principle and makes inferences according to the probability distribution parameters [4]. When parametric tests are used, generalizations can be made about all data according to the test result of a certain sample group.

# T-Test

|  |  |
| --- | --- |
|  | (3) |

The difference between the means of two independent sample groups is tested. To apply this test, the populations must have the same variances. The t-test [5] is the most common application to test whether the means of two populations are different.

In this study, the similarity of "meat of livestock" data in the Finland and Croatia, similar to Ireland in terms of population, was tested. For this, it was checked that all three data groups had normal distribution and the same variance. For the T-Test, the null hypothesis is that the samples have the same mean, but the alternative hypothesis has a different mean as shown in (3). As a conclusion of T-Test results, there is enough evidence to conclude that Finland and Ireland are similar, Finland and Croatia are similar countries. However, there is not enough evidence to conclude that Croatia and Ireland are similar countries.

# Analysis of Variance

Analysis of variance (ANOVA)[6] is used to calculate the significance of the difference between three or more independent means in a normally distributed series. ANOVA cumulatively compares the arithmetic means of three or more groups; If at least one of these comparisons is significant, the ANOVA result is also significant. In this case, the hypotheses are;

H0: There is no difference between their averages.

H1: There is a significant difference between at least two means.

In the study, three groups created with the meat of livestock data in Finland, Ireland, and Croatia were tested with ANOVA. As a result of the test, the was rejected because the p\_value was below 0.05. It means that the mean of at least one of the groups is different.

# Pearson Correlation Test

If the data sets come from a normally distributed population, Pearson's sample correlation coefficient is the best correlation for the correlation found between these two populations estimate variables. The Pearson correlation test is a measure of the covariance of two variables. In the study, the changes in the amount of meat livestock and milk stock in Ireland were analyzed. The result shows that milk and meat production in Ireland has strong relation (%81.5). Since the p\_value is less than 0.05, is rejected. As a result, there is a significant and positive relationship between them.

# Non-Parametric Inferential Statistical Techniques

Non-parametric tests are experiments that do not require the underlying population for assumptions. When non-parametric tests are used, generalizations can not be made about all data according to the test result of a certain sample group.

# Kruskal Wallis Test

It is a technique used to test the significance of the difference between the means of three or more groups in non-normally distributed groups. It has two basic assumptions: the measurement scale is at least ordinal, and the variable is continuous; The groups are mutually independent. In addition, if the result is significant, the Mann-Whitney U test is applied between each pairwise combination of the subgroups to determine the difference between which groups.

The similarity of the import values in the organic operators data for three countries was tested. These data have a normal distribution, but the non-parametric method should be used as the variance homogeneity assumptions are not satisfied. was rejected because the p\_value was less than 0.05 from the applied Kruskal test result. Since the null hypothesis is rejected, at least one means of the Imports for Ireland, Finland, and Croatia is different. Mann-Whitney U test was used to determine the mean difference between data groups, and it was determined that this difference was between Finland and Croatia.

# Chi-Square Test

The chi-square test of independence is used to investigate whether two or more variables are independent. Although this test measures independence, if there is a relationship between the variables, it does not provide sufficient information about this relationship. Also, the chi-square test is sensitive to sample size. As the number of observations increases, the chi-square value and, thus, the significant difference provide observation.

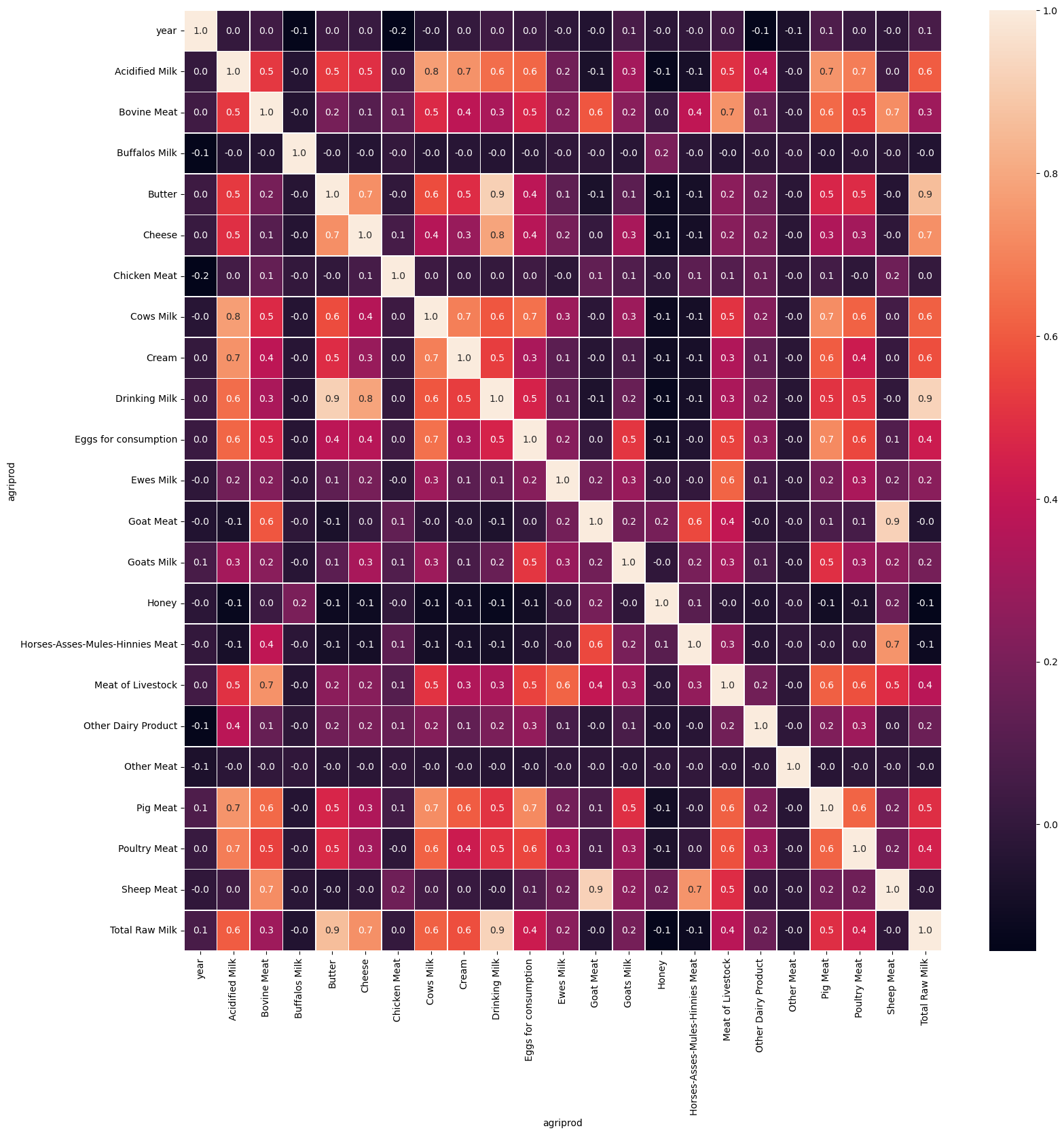
With the chi-square test being successful on categorical data, a test was conducted for Finland and Ireland to compare country and product adherence on milk stock and meat stock categories. Here it can be concluded that they are independent since the p\_value is greater than 0.05. Here, different chi-square tests are performed according to the conditions provided by the expected values, as shown in equation (4). Expected value indicated as exp\_value in the equation. Yates Chi\_Square technique was used because the expected values were between 5 and 25. As a result of the Yates chi-square test, there is no relationship between geography and agricultural product regarding milk and meat stocks for Ireland and Finland since the p\_value is greater than 0.05.

|  |  |
| --- | --- |
|  | (4) |

# Data Visualization and Preparation

The data used in the study are taken from Eurostat and include data for each year. Thanks to the time interval of the data, it is possible to observe the changes in the same year. It is necessary to organize the obtained raw data according to the methods used and remove unnecessary attributes. Data analysis was performed to detect unnecessary features. Exploratory data analysis helps identify patterns, inconsistencies, anomalies, missing data, and other attributes and issues in data sets so that issues can be addressed. In this data set, product categories are structured by converting them into columns.

One of the organic products, 'Rabbit Meat,' was removed because all data were zero or null. In addition, while all products were recorded in tons, egg data was recorded in thousand. In order to delete the Unit column and to provide balance in the dataset, the number of eggs was calculated in tons and updated.



,

Figure 1. Correlation map of agricultural products.

Considering the relationship between the amount of each organic animal product in the last ten years;

* Since milk is used in producing products such as cheese and butter, it can be concluded that milk and dairy products have a strong positive correlation within themselves.
* There is almost no relationship between "Buffalos Milk", "Chicken Meat", "Other Meat" products, and other products.
* Since the correlation is zero, it can be concluded that there is no linear relationship between these two products.

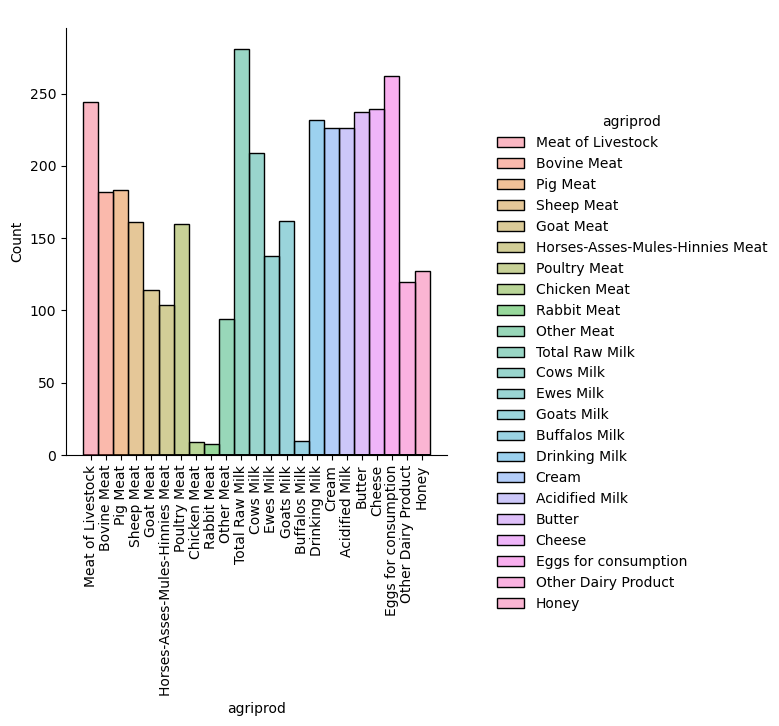


Figure 2. Frequency distribution ofagricultural product for each country

Figure (2) shows the frequency distribution of total agricultural products in all European countries. This distribution shows how low rabbit meat, chicken meat, and buffalo milk production is in European countries. Since the data set consists of nested categories, the data set is divided into cluster countries according to meat production, milk production, and egg production.

For organic operators by status of the registration process, another dataset used in the study, unnecessary columns and rows with empty values were dropped. Categories and countries not included in the study were also excluded from the data set. As seen in the correlation map given in Figure (3), there is a very strong correlation between exporters and importers. Likewise, there is a strong correlation between the processors and the producers.

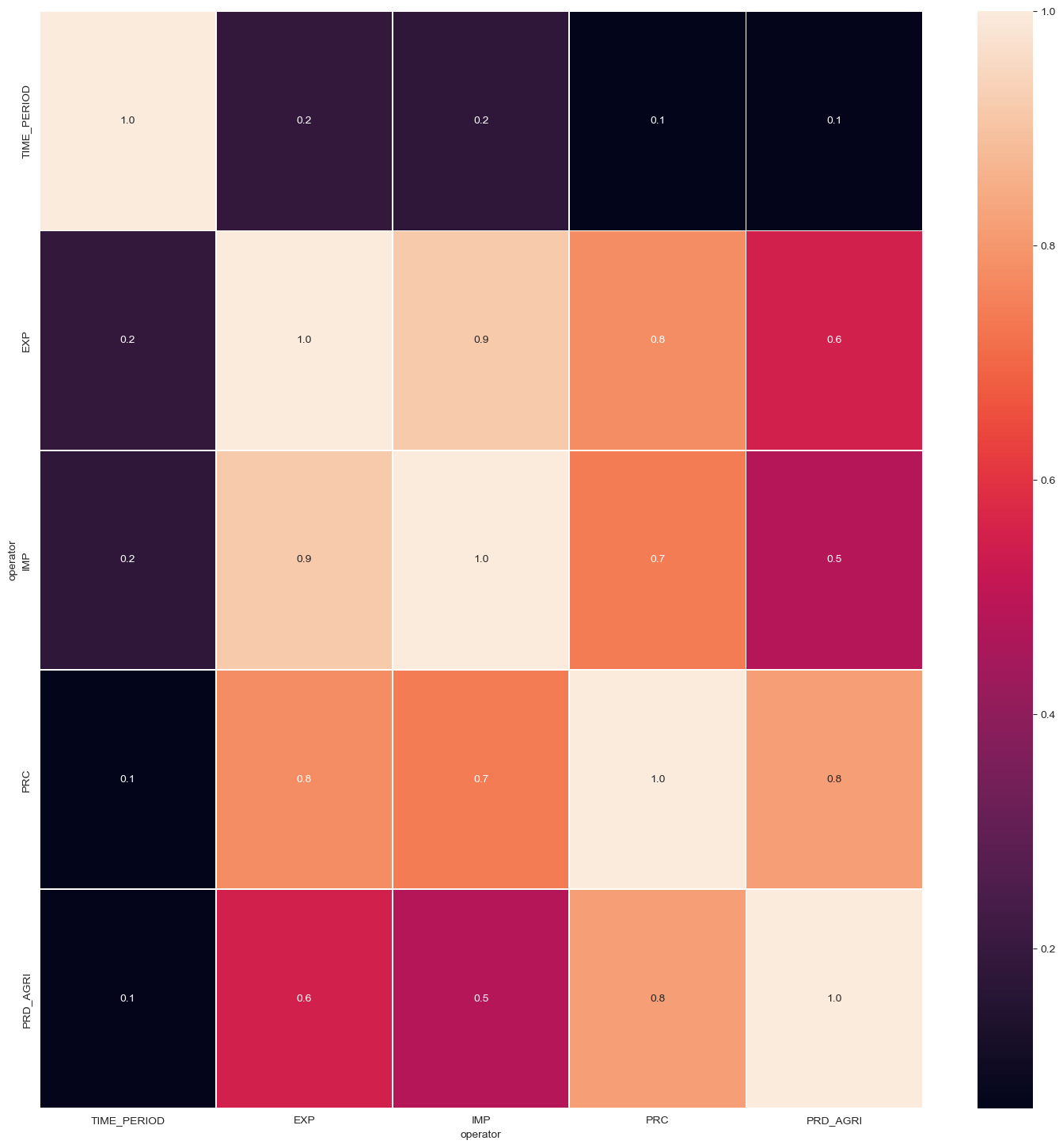


Figure 3. . Correlation map of organic operators.

The exporters and the importers, which are strongly correlated, are visualized in figure 4 with a donut-like pie chart. Here, it seems that each country's importer and exporter rates are close to each other.

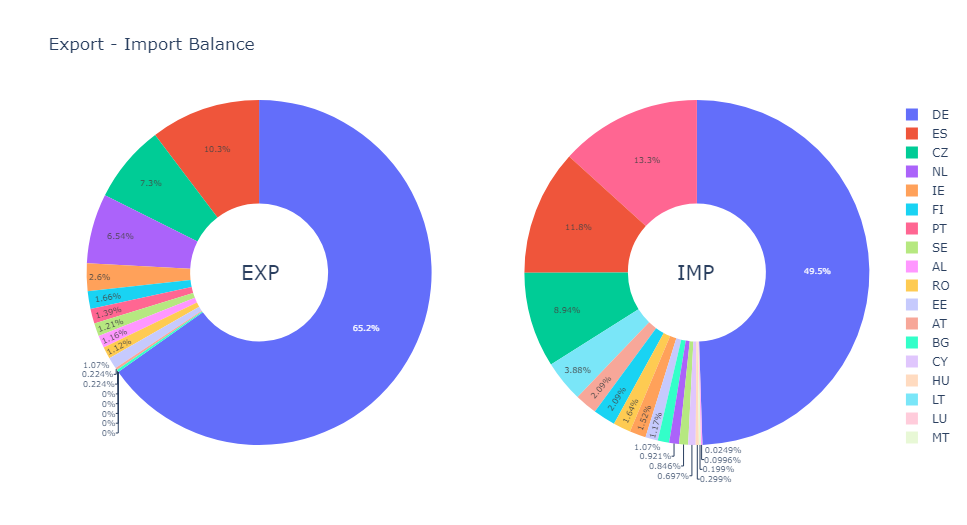


Figure 4. Visualization of exporters and importers rate

# Machine Learning for Data Analytics



# Sentiment Analysis

Sentiment analysis is a text analysis that aims to determine the class (positive, negative, and neutral) that the given text wants to express emotionally. Sentiment analysis is often used to understand customer satisfaction with products. In this context, tweets containing 'organic' and 'price' on Twitter were obtained using snscrape, and a data set was created. However, since tweets contain many unnecessary characters, they require detailed pre-processing steps. Deleting usernames from tweets with mentions; Data containing hashtags, links, emojis, and symbols should be removed. In addition, users use certain abbreviations when writing, which needs to be changed accordingly. After these pre-processing steps were completed, the subjectivity and polarity values of the texts were calculated. Polarity, also known as orientation, is the emotion expressed in the sentence. It is the value that represents whether it is positive, negative, or neutral. Subjectivity is a value that indicates that the text is an explanatory text that needs to be examined in context.

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Figure 5. Visualization of the frequency of use of words belonging to users with wordcloud.

In sentiment analysis, the word cloud is used, which shows the frequency of use of words belonging to users. As seen in the word cloud shown in Figure (\*), there are keywords such as 'buy', 'share', and 'organic growth' as well as 'price' and 'organic' keywords.

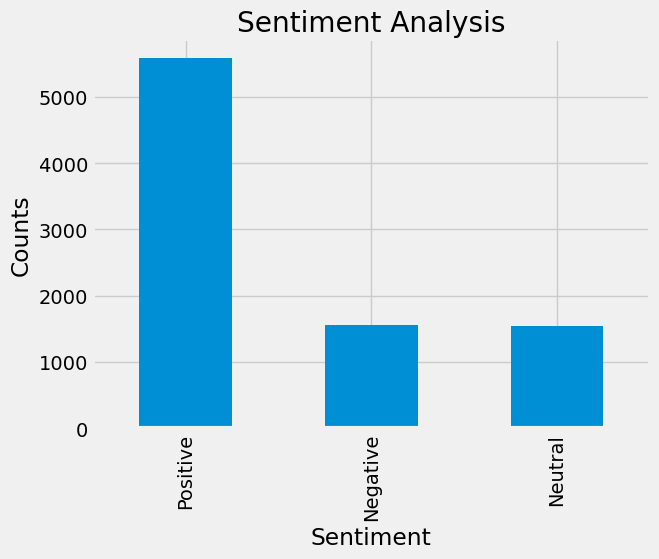


Figure 6. Visualization of groups as a result of sentiment analysis.

Based on the polarity scores, the texts are calculated as negative, neutral, and positive. The distribution of the created groups is visualized with a bar plot and is shown in figure (2).

# Supervised Learning

Supervised learning is learning a function to obtain the desired output set *Y* from a given set of inputs *X* [7]. Supervised learning algorithms aim to learn a function that maps feature vectors to labels based on example input-output pairs [8]. It uses a training set to teach models to produce the desired output.

In machine learning studies, parameters specific to each model are used in model training. These parameters can have multiple values. It is challenging to run all possible parameters for each model and determine the best parameters. For this process, the best parameters can be determined with the GridSearchCV tool provided by the sklearn library. In the study, classification and regression were performed on the data set created for sentiment analysis. For this process, firstly, the texts in the data set were converted into vectors; Then, the data set was divided into train and test at a ratio of 0.8 and 0.2, respectively.

# Random Forest Classification

Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that work by generating many decision trees at training time. The output of the random forest for classification tasks is the class chosen by most trees. Random forest classifiers, as ensemble learning methods, are models that provide high performance. The random forest classifier algorithm was trained with default parameters and the training accuracy was calculated as 64.25%. The GridSearchCV tool was used to select the best parameters for the random forest classifier. For n\_estimators, max\_features, and criterion parameters, it outputs that the best parameters are 50, log2, and entropy, respectively. Then the classifier was run with these parameters, and 71.35% training accuracy was calculated. The random forest classifier created with GridSearchCV has been tested on test data, and the results are given in Table (1).

Table 1. Results of Random Forest Classifier.

|  |
| --- |
| precision recall f1-score support  Positive 0.74 0.98 0.85 1126  Negative 0.91 0.16 0.27 308  Neutral 0.83 0.53 0.65 302  accuracy 0.76 1736  macro avg 0.83 0.56 0.59 1736  weighted avg 0.79 0.76 0.71 1736 |

# Logistic Regression

Logistic regression is a statistical method used to predict the outcome of a dependent variable based on previous observations. Logistic regression measures the relationship between the dependent variable and one or more independent variables. It works by estimating the probabilities with the help of its underlying logistic function. The GridSearchCV tool was used to select the best parameters for the logistic regression. For C and penalty parameters, it outputs that the best parameters are 1000 and l2, respectively. Since the problem is a multiclass problem, and multiclass parameter is selected as multinomial. Then the classifier was run with these parameters, and 79.14% training accuracy was calculated. The logistic regression created with GridSearchCV has been tested on test data, and the results are given in Table (2).

Tablo 2. Results of Logistic Regression

|  |
| --- |
| precision recall f1-score support  Positive 0.85 0.94 0.89 1126  Negative 0.75 0.56 0.64 308  Neutral 0.81 0.72 0.76 302  accuracy 0.83 1736  macro avg 0.80 0.74 0.77 1736  weighted avg 0.83 0.83 0.83 1736 |

# Unsupervised Learning

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms locate hidden patterns or data collections without human intervention. Its ability to discover similarities and contrasts in information makes it the ideal solution for exploratory data analysis, cross-selling strategies, and customer segmentation. In this study, the k-means clustering algorithm, an unsupervised learning model, was used on agricultural products and unlabeled data.

# K-Means Clustering

Clustering is a type of data mining approach that groups unlabeled data based on their similarities or contrasts. K-means clustering is a common model of a clustering method where data points are assigned into K groups, where K represents the number of clusters based on the distance from the centroid of each group. The data points nearest to a given centroid will be clustered under the same class.

Columns indicating egg, milk, and meat production were selected as features in the organic product data for the clustering method. The elbow technique was used to determine how many clusters the data should be divided into before clustering. In the graph, which is the output of the Elbow method, the number of clusters is determined according to the point where the line is certainly bent. In the graph shown in Figure (7), the optimal number of clusters was determined to be two.

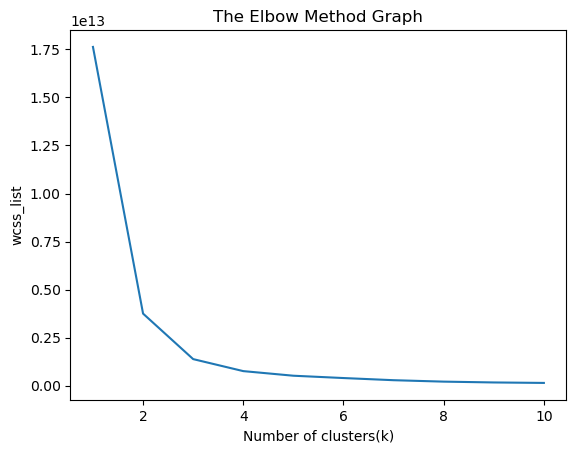


Figure 7. Graph of elbow method.

K-means clustering result data points are divided into two clusters, as in Figure (8). In these data for the last ten years, Austria, Germany, Denmark, France, Italy, and Sweden countries are in a group. Austria, Belgium, Bulgaria, Cyprus, Czechia, Estonia, Greece, Spain, Finland, Croatia, Hungary, Italy, Ireland, Lithuania, Luxembourg, Latvia, Malta, Netherlands, Poland, Romania, Slovenia, and Slovakia are in the other group. As can be seen, Austria and Italy countries are in two groups. This is due to the change in the data of these two countries for the three classes over time. In addition, as seen in the clustering result, Ireland is in the same cluster as the two countries whose similarity was tested before.

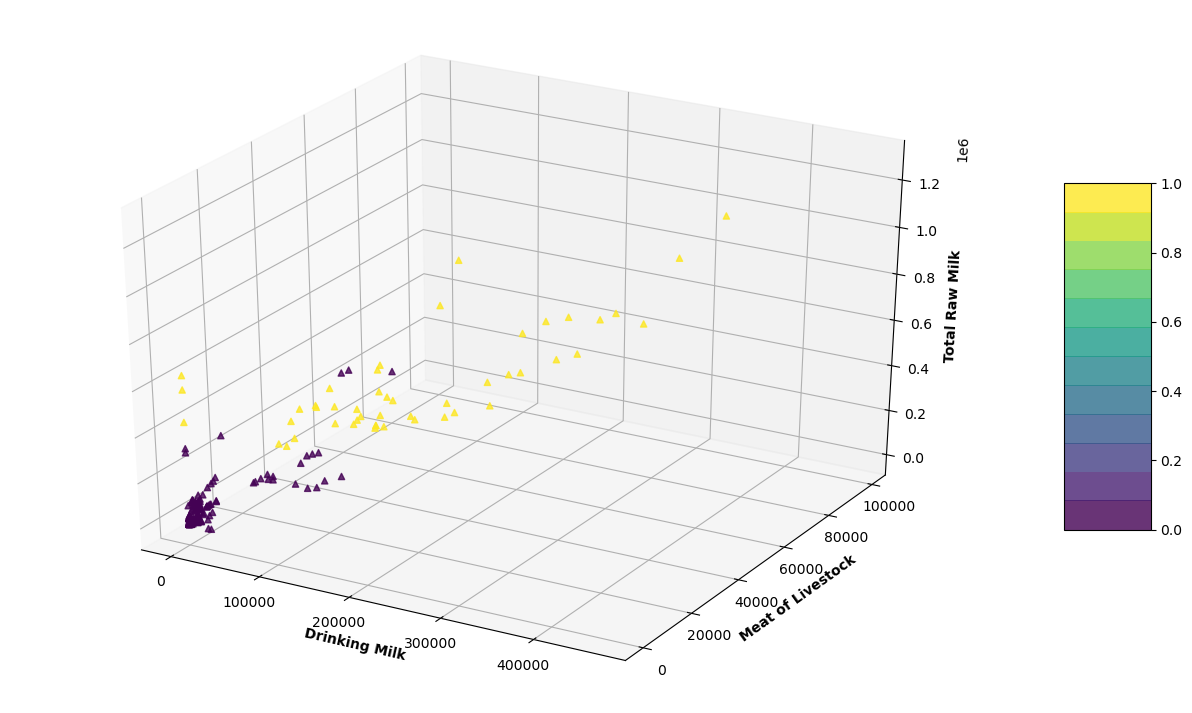


Figure 8. Visualization of clustering result

# Forecasting

Forecasting is a technique that uses historical data as input to make informed predictions that are predictive in determining the direction of future trends. In this study, the data for the next year were estimated by using the data set covering the last ten years of organic transactions. For this process, the machine learning library pycaret tool was used. In this context, one-hot encoding was applied to the data. Results were obtained for all models, parameters were determined for the model providing high performance, and k-fold was applied. For the forecasting process, the data is separated as train and test. In addition to the data set, blank data for the future covering one year were created for each country and category. Random forest regressor was the model that provided the highest performance among all the models applied. The model with the second highest performance was the extra trees regressor. These two models were trained and tested with the test dataset. The train and test accuracy for the random forest regressor are 94.65 and 97.65%, respectively. The test accuracy for the extra trees regressor was calculated as 98.15%, respectively. In Figure 9, the prediction error graph of both algorithms is given. Then, estimates were made for the year 2022, which was not in the dataset, and the dataset was added.

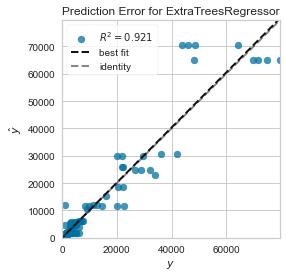
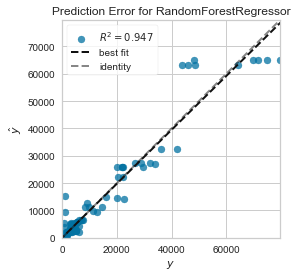


Figure 9. Visualization of prediction error for regressor algorithms

# Conclusion

In the study, data from Ireland and other EU countries were analyzed using data on the organic production of animal products. Based on these data, other countries similar to Ireland were identified, and similarity tests were applied. Analysis was made on the data produced by Finland, Ireland, and Croatia countries, which are similar in population. As a result, it was concluded that while Croatia and Ireland are quite similar in meat production, they differ in milk production. In addition, when clustering was made, these three countries were included in the same cluster. In addition, organic operators made by all countries in 2022 were estimated in the study.

# References

[1] Hermansen, J. E. (2003). “Organic livestock production systems and appropriate development in relation to public expectations”. *Livestock production science*, *80*(1-2), 3-15.

[2] Novák, J., Benda, P., Šilerová, E., Vaněk, J. and Kánská, E. (2021) “Sentiment Analysis in Agriculture", *AGRIS on-line Papers in Economics and Informatics*, Vol. 13, No. 1, pp. 121-130. ISSN 1804-1930.

[3] Cao, Y., Sun, Z., Li, L.,Mo,W. (2022). “A Study of SentimentAnalysis Algorithms for AgriculturalProduct Reviews Based on Improved BERT Model”. Symmetry, 14, 1604.

[4] Geisser S, Johnson WM. (2006). “Modes of Parametric Statistical Inference”, John Wiley & Sons.

[5] Pfanzagl, J., and O. Sheynin. (1996). "Studies in the history of probability and statistics XLIV A forerunner of the t-distribution." Biometrika 891-898.

[6] Stigler, Stephen. (2005) "Fisher in 1921." Statistical Science 32-49.

[7] Alpaydın, E. (2010). Introduction to Machine Learning. Londra: The MIT Press. s. 8. ISBN 978-0-262-01243-0.

[8] Stuart J. Russell, Peter Norvig (2010). “Artificial Intelligence: A Modern Approach”, Third Edition, Prentice Hall ISBN 9780136042594.

# Appendix

# Github wep URL

<https://github.com/MGunesCetin/MSC_DA_CA2.git>

# Word count

3285 (not including code, code comments, titles, references or citations)