Harnessing Data and Statistical Insights to Drive Innovation and Sustainability in Irish Agriculture

**Author: Andrew Mc Guinn**

**Email: sba24458@student.cct.ie**

**Student ID: sba24458**

Word count: 4300

*Abstract – This study evaluates the performance of various machine learning models applied to classification and regression tasks, with a focus on hyperparameter optimisation and model evaluation metrics. Using data with imbalanced class distribution was addressed by using SMOTE, achieving balanced target classes.*

*Grid search was employed to identify the best parameters for models used, Logistic Regression, Decision tree, Random Forest, KNN and others. Performance metrics such as accuracy, precision, recall, Cohens Kappa and r2 were calculated to assess model efficiency. KNN and Random Forest outperformed all other models in regression tasks with 0.88 and 0.94 r2respectively, residuals were then plotted. While Multinomial NB and Logistic Regression demonstrated competitive accuracy in Classification with 0.59 and 0.58 CV accuracy respectively . Hypothesis tests were conducted for the output yield of major European countries against Irelands yield, these tests were evaluated with appropriate metrics such as Cohens D, and the power of a test.*

# Introduction

## Background

Agriculture, one of humanity’s old industries, is undergoing a profound transformation fuelled by advances in big data analytics and machine learning. These innovations are revolutionising farming practices by enabling precision agriculture, optimising resource allocation and enhancing productivity. Ireland, a young country with a rich agricultural heritage, these transformations hold significant potential. However, adoption must be critically evaluated within the frameworks of efficiency, sustainability and economic development and viability.

In Ireland, agriculture remains a cornerstone of the economy providing livelihoods to over 173,000 people and contributing to 7.7% of gross national income. As Irelands largest indigenous sector, agriculture supports rural development, ensures food security. The common agriculture policy has been instrumental in shaping Irelands agricultural landscape, offering income support, infrastructure investment and environmental incentives. (Anon., 2024)

Since its inception in 1962, CAP has evolved significantly to address moder challenges. Its primary goals include increase productivity, stabilizing markets, and ensuring fair living standards for farmers. Successive reforms, including the CAP post-2020 proposals, have emphasized sustainability and environmental stewardship. Notably, they outline new planning requirements, an increased focus on climate action, and significant changes to direct payment structures. (Gov, 2024)

## Scope and Objectives

This study evaluates Europe’s and specifically Irelands agricultural performance and yields using advanced modelling techniques such as regression, classification and analysing public sentiment on farming practices and food through natural language processing (NLP) for sentiment analysis. Evidenced based strategies will be developed to improve decision making, while the socioeconomic impact of CAP reforms will be explored with an emphasis on income support and long-term sustainability.

Statistical tests including hypothesis tests such as Z and T test, ANOVA will validate assumptions, identify relationships between variables and assess differences in agricultural performance across regions.

# Materials and Methods

## Data Acquisition

The datasets were sourced from multiple online repositories, including Reddit for sentiment analysis, Our world in data for precipitation and FAOSTAT for all other data such as temperature change, employment and pesticide usage. These sources provided a diverse range of data critical to understanding agricultural trends and practices in Ireland and Europe. Licensing information was carefully reviewed to ensure compliance with usage rights as seen below.

Data downloaded from FAOSTAT, which specifies that its data is available for public use under the FAO data sharing policy (FAO, n.d.). According to this policy, users are permitted to access, use and share the data for research and non-commercial purposes, provided proper attribution is given to FAO as the source.

In this report, five datasets were sourced from FAOSTAT- temperature, employment, pesticide usage, population, agricultural yield and have been used in accordance with these terms. Proper citation is provided in the references, these datasets were used exclusively for academic and analytical purposes, with no modifications that would misrepresent their content or intent.

Precipitation data, sourced from our world in data, is licenced under a Creative commons attribution 4.0 international (CC BY 4.0) (Ourworldindata, 2024).

Data was extracted for NLP using Reddit API. According to Reddits API terms of service, the use of public data for non-commercial purposes is permitted, provided it complies with reddit usage policies. The extracted data is used solely for research purposes and no personally identifiable information is collected (reddit, 2024).

The process involved API based data extraction and manual file downloads, while this approach enabled access to comprehensive datasets, it presented several challenges. Reddit data extraction required precise query formulation and API integration which posed potential limitations in the granularity of retrieved comments. The downloaded datasets reduced the versatility of integrating the models across platforms as the datasets need to be on the device rather than obtained through a database or API. Another problem was finding relevant and related data for sentiment analysis, when increasing the number of rows for analysis, the corpus is also increased and without more similar data to train the models on, the relevant metrics are quite low.

On the positive side, these datasets enabled detailed cross variable analysis such as integrating weather data with employment data.

## Data cleaning and feature engineering

To ensure data quality and reliability, extensive cleaning and engineering techniques were applied to the dataset using pythons Pandas and Numpy libraries. The initial step involved dropping irrelevant columns that added little or no value, next was reshaping the dataset to have 594 rows, each row representing an EU country and the year from 2000-2021, these years were chosen as the period is a significant amount of time to gather a representative amount of data pertaining to agriculture in Europe spanning different qualities.

The dataset containing yields was reshaped using pivot tables to align data points across “area” and “year”, organising them into interpretable columns, this restructuring facilitated consistency in further analysis as seen (Prince, Jan 23, 2020), (Mckinney, 2024).

To address discrepancies in column naming conventions PEP 8 standards were adhered to throughout this report, snake case naming convention was implemented through a python function. Duplicate columns were removed to prevent redundancy.

For missing values, firstly the data was visualized through a heatmap to get an idea of the scale of the missing values as seen below in figure1.

A purple and yellow background with white text

Description automatically generated

Figure :Missing values

All the columns containing missing values were numerical and missing at random, the next step was to visualise the distribution of these values as seen in figure 2 (Firdose, n.d.).

A close-up of a graph

Description automatically generated

Figure :Missing values distributions

From the above figure, it is visually evident that most of the columns appear to be approximately normally distributed, as for the last column which is ‘agriculture\_value’, the column appears to be quite right skewed, and the data seems to be mostly missing at random. For the missing data, the columns were filled through linear interpolation as this is suitable for normally distributed time series data (Cátia M. Salgado, 2016).

This approach was chosen due to its robustness against outliers compared to other methods, ensuring that the imputed values would preserve the original distribution of the data, as there is a limited amount of data, the priority was to preserve as much data. Below in figure 3 is the figure of the distributions after imputation.

A group of graphs showing different types of data

Description automatically generated with medium confidence

Figure :Columns after median imputation.

The year column was converted to datatype DateTime for visualisation purposes, additionally country codes were mapped the “area” column using a predefined dictionary, enriching the dataset for analysis, this step was crucial for facilitating regional analysis.

For sentiment analysis, the data obtained through the reddit API “reddit\_df” underwent rigorous data cleaning and engineering to transform unstructured reddit data into a usable format for sentiment analysis. Techniques such as tokenisation, lemmatisation and stop word removal enhanced the datasets quality by reducing noise and improving uniformity. Engineering sentiment scores through the library nltk.sentiemnt.vader using polarity scores for the scores, this enriches the data enabling insights and classification (Anurag, 2024). However, the use of rule-based methods like regular expressions for preprocessing may overlook complex cases which could result in losing valuable information.

## Descriptive Statistics and EDA

Descriptive statistics revealed notable trends and variations across the dataset. Most of the datatypes in the dataset are of type float as seen in figure 4, while this is more intensive on memory , as the dataset is relatively small, the key metric is accuracy .

A screenshot of a computer

Description automatically generated

Figure :Data types for df

For instance, “fruit\_primary” had a mean production of 10,984 with a standard deviation of 7,731 reflecting significant variability. However, its maximum value of 44,196 indicates outliers likely influencing the columns skew 1.61. in contrast variables like “milk\_total” showed a much broader range , with values from 1,772 to 100,966 suggesting diverse production levels across regions as seen in figure 5 .

A group of graphs showing different sizes and shapes

Description automatically generated with medium confidence

Figure :Data Distributions

Extreme differences in ranges were also observed for metrics such as “total\_precipitation”, where values spanned from 211 to 1,763. Similarly, “temp\_change” remained relatively stable with a mean of 1.43 but a narrow range of -0.28 to 3.6 (Kumar, Sep 30, 2024).

Outliers significantly influence the datasets analysis, with variables like “Pesticide\_used\_tonnes” , “fruit\_primary” and

”milk\_total” showing extreme values far above their means. These anomalies skew central tendencies and variances, indicating rare but impactful events such as intensive agricultural practices or exceptional climactic conditions. While they may distort results, outliers also highlight critical trends, such as regional disparities or inefficiencies as seen in figure 6.

A group of blue rectangular boxes with text

Description automatically generated with medium confidence

Figure : Data boxplots

Kurtosis values provide insights into the distribution shape of the dataset, helping to understand outliers. A high kurtosis such as

“Pesticide\_used\_tonnes” of 3.8 and “fruit\_primary” of 2.78 indicates leptokurtic tendencies, indicating outliers, lower kurtosis value less than “milk\_total” -1.04 suggest a distribution closer to normal curve, with negative values indicating platykurtic tendencies, as seen in figure 7 (Menon, 2024).

A screenshot of a computer program

Description automatically generated

Figure : Kurtosis values

The skew value of a distribution indicates the asymmetry in the distribution of data, where a positive skew such as “fruit\_primary” of 1.6 and “Pesticide\_used\_tonnes” of 2.2 show positive skew suggesting a concentration of higher values. Negative skew indicates a tail is extending towards lower values like “Beef\_and\_buffalo\_meat\_primary” of -0.58 and “milk\_total” indicating a concentration of lower values as seen in figure 8. Most of the columns do appear to be approximately normally distributed.

A screenshot of a computer

Description automatically generated

Figure :Skew values

A colorful chart with text

Description automatically generated with medium confidence

Figure :correlation coefficients

The correlation matrix in figure 9 provides valuable insights into the relationship between variables, positive correlations such as “barley” and “cereals\_primary” have high correlation of 0.94. Conversely, negative correlations like “eggs\_primary” and “temp\_change” show an inverse relationship. For machine learning models, correlation is important as it helps us understand the relationship between our features and target variable to avoid conditions such as multi collinearity (Gupta, 2024).

A graph with a line going up

Description automatically generated

Figure :Word frequency NLP

EDA for the reddit\_df involves text analysis to understand the content and sentiment of posts, tokenising the “body” column extracts individual words and in figure 10, once stop words were removed, the 30 most frequently used words in the post were graphed. Additionally, a sentiment distribution reveals the counts of positive, neutral and negative sentiments in the data, a class imbalance is evident from the graph.

**A graph of positive and neutral expression

Description automatically generated**

Figure :Class Distribution NLP

Finally, a word cloud in figure 12 visually represents the most frequent words across the posts, providing insights into the sentiment of the data.

**A close-up of words

Description automatically generated**

Figure : Wordcloud most common words

# Visualisations and Statistics

## Visualisations

A graph of a number of people

Description automatically generated with medium confidence

Figure :Average Agriculture Value per country

The graphs in figures 13, and 14 adhere to Tufts principles of effective data visualization by prioritising clarity, accessibility and accuracy, the use of a bar plot effectively communicates average agriculture value and pesticide usage per country allowing for straightforward comparison. The graphs give rise to meaningful trends, however the absence of contextual information such as time periods limit the depth of interpretation, this will be addressed by dynamic visualisations.

A graph of the number of countries/regions

Description automatically generated

Figure : EU Countries Pesticide Usage

A graph with red and blue lines

Description automatically generated

Figure : Agriculture values scaled with temperature change

A graph of different colored bars

Description automatically generated with medium confidence

Figure :Stacked bar plot

Figure 15 explores the relationship between temperature change and agriculture value by normalising both metrics. While it effectively visualizes trends it does lack the knowledge of economic policies or justification for trends. Figure 16 highlights the contributions of agricultural products by country. The plot adheres to Tuft’s principles by maximizing data density while maintaining clarity.

A map of europe with different colored areas

Description automatically generated

Figure :Choropleth of agriculture value

Figure 17, 18 and 20 displays choropleths of Europe in relation to agriculture value, total precipitation and cereal values. The maps effectively visualise the spatial distribution of agricultural metrics across EU countries over time maximising data to ink ratio. Dynamic animations reveal temporal trends, enriching the narrative. The interactive hover features add depth by providing country specific details without cluttering the map.

A map of europe with green squares

Description automatically generated

Figure :Choropleth of total precipitation

A screenshot of a graph

Description automatically generated

Figure :People employed Vs population in terms of agriculture value

Figure 19 provides a compelling representation of the relationship between population ,employment and agricultural value. The animation frame by year as well as labels and hover text effectively provide detailed insights without clutter (Tufte, 2024).

A map of europe with different colors

Description automatically generated

Figure :Choropleth of cereals value

## Inferential Statistics

This test analysed cereal production which is a critical component of agricultural production and contributes significantly to the food security of countries, this test will compare cereal production between Belgium and Ireland on statistically significant differences between mean production values.

A graph of different colored bars

Description automatically generated with medium confidence

Figure :Cereal production

Belgium was chosen as it is the top producer of cereal in the EU, it is evident that there is a difference in production between Belgium and Ireland as seen in figure 21,but is that difference statistically significant, a two-sample independent T-test will be conducted.

A group of graphs with numbers

Description automatically generated with medium confidence

Figure : QQ plots and histograms of distributions

The first test is to check the underlying assumptions for a t test, a test for normality will be conducted, from the visual inspection in figure 22, the distributions appear to be approximately normally distributed. A hypothesis test for normality will be conducted as follows:

**Null Hypothesis (H₀):** The data follows a normal distribution.

**Alternative Hypothesis (H₁):** The data does not follow a normal distribution.

For the hypothesis tests for normality:

A black background with white text

Description automatically generated

Figure : Null and alternate hypothesis

A computer screen with numbers and letters

Description automatically generated

Figure 24:Tests for normality

The distributions are normally distributed as he p values are greater than the alpha set at 0.05, a two samples t-test will be conducted.

A computer screen with white text

Description automatically generated

Figure 25:Results from T-test

The analysis reveals a statistically significant difference between the mean productions. Since the p-value < alpha, the null hypothesis is rejected in favour of the alternative hypothesis (Al-kassab, 2022).

Cohens D was calculated as -1.46 indicating a large effect size indicating that the difference between the two countries is substantial. Using the observed effect size, a test for power was conducted, to achieve aa power of 80%, a sample size of 3.69 is necessary.

A diagram of cereals

Description automatically generated

Figure :boxplots of the two samples

A graph of different colored bars

Description automatically generated with medium confidence

Figure :Barplot of EU BEEF

The analysis examines Ireland beef production by drawing a sample from all EU countries. A Z test will be conducted to determine if the sample is representative of the entire dataset. This allows us to make an inference about Irelands production in relation to the broader EU context.

A black background with white text

Description automatically generated

Figure :Stating hypothesis

A hypothesis test for normality was conducted, Shapiro Wilk and Kolmogorov Smirnov tests were conducted, however Shapiro wilk test is the priority due to its sensitivity to small sample sizes

(Asghar Ghasemi 1, 2012).

A graph and diagram of a graph

Description automatically generated with medium confidence

Figure :QQ plot and Distribution

From figure 27 and 28, it is evident that the data is approximately normally distributed, for certainty appropriate tests will be conducted. The Shapiro Wilk test with a p-value of > alpha 0.05, fail to reject the null hypothesis suggesting the data is normally distributed.

A black screen with white text

Description automatically generated

Figure : Normality tests for beef



Figure : Z-test results

The resulting Z-score -1.41 and p value 0.159 indicates that there is no statistically significant difference between the two means. The Cohens-d effect size of -0.089 indicates a small to medium effect size, this suggests the result may not have a strong practical impact. The required size of 317 is suggested to achieve sufficient statistical power of 80% for detecting this small effect size. This aids in understanding current EU wide policies.

A graph of a number of people

Description automatically generated

Figure : Pesticide usage EU

This analysis examines pesticide usage across various EU countries as seen in figure 30, focusing on comparing pesticide usage relative to the mean. By calculating the mean pesticide usage, a binary “success” column was engineered. Categorising those with above average usage and below average usage.

A black background with white numbers

Description automatically generated

Figure :Results from Binomial

**Null Hypothesis (H₀):** The proportion of countries with pesticide usage exceeding the average of 200.34 tonnes is equal to 0.5 (50%).

**Alternative Hypothesis (H₁):** The proportion of countries with pesticide usage exceeding the average of 200.34 tonnes is not equal to 0.5 (50%).

Figure 31 displays the results of the test showing that approximately 20% of countries exceed the average pesticide usage with 199 of 594 categorised as successes. The binomial test reveals a p-value of 0.2 suggesting no statistically significant difference from the hypothesised 0.5, based on the results, there is insufficient evidence to reject the null hypothesis, therefore we fail to reject the null hypothesis.

A graph showing a number of successful

Description automatically generated

Figure : Binomial Distribution

A graph of a number of colored bars

Description automatically generated with medium confidence

Figure :Milk production distribution

Ireland had a relatively lower milk production value (49685.09), reflecting a smaller output compared to other EU countries. Czechia had a moderate milk production value (73670.95), providing a balanced midpoint. Denmark showed a significantly higher milk production value (87294.32), indicating a much larger output. This diversity allows for a comprehensive understanding of how milk production varies among countries with different production scales, contributing to a more meaningful analysis.

**Null Hypothesis (H₀):**

There is no significant difference in milk production between Ireland, Czechia, and Denmark.

**Alternative Hypothesis (H₁):**

There is a significant difference in milk production between at least one of the countries.



Figure :ANOVA results

In this analysis, the one-way ANOVA was conducted, we reject the null hypothesis, there is a statistically significant difference between at least one of the means.

The next test conducted was a chi squared test to examine the relationship between area and temperature change. These results indicate a statistically significant relationship between these variables.

Figure 35 displays the results; the chi-square stat exceeds the critical value and the p value < alpha of 0.05. we reject the null hypothesis.

A computer screen with white text

Description automatically generated

Figure :Chi squared results

For Ireland this finding is relevant in understanding how localised temperature variations impact agricultural practices and policy changes.

For the last test, a confidence interval for the agricultural value provides a range of estimates within the true mean is likely to fall. Based on a 95% confidence interval, the mean value is approximately 7012, with a lower bound of 6175 and an upper bound of 7849. This suggest that, given the data, one can be 95% confident that the true mean lies between these bounds.

# Machine Learning

## Classification models

The data was loaded and pre-processed using pandas as outlined, it underwent rigorous preprocessing and engineering as outlined in section 2. TF-IDF was employed due to its ability to weigh terms based on their importance across the entire dataset. For sentiment analysis, this is critical since rare words with high importance are less likely to be missed (Anurag, 2024).,

Train\_test\_split was utilised with the standard test\_size of 20% (Brownlee, 2020), as the datset is relatively small, appropriate models were chosen to maximize appropriate metrics. Multiple models such as KNN was chosen as it is straightforward and it treats all features with equal importance, one limitation of this model is its inability to handle imbalanced data. Naïve Bayes was chosen as it is incredibly fast and simple and it serves as a great baseline for text classification as well as others (Wohlwend, n.d.).

On the Initial iteration in figure 36, the models performed reasonably well with Decision tree and Random Forest performing the best with accuracy of 61% and 56%. However as displayed in section 2, the dataset is imbalanced with higher positive comment counts as shown in figure 11.

A screenshot of a graph

Description automatically generated

Figure :Results table

Figure 37 shows, ROC\_AUC scores for the models, this is the best evaluation metric for imbalanced datasets (John T. Hancock, 2023).

A screen shot of a computer

Description automatically generated

Figure :ROC-AUC scores

A screenshot of a computer program

Description automatically generated

Figure :CV validation

The scores for accuracy in figure 38, showing 5-fold cross validation with a mean accuracy of 56.9%. This indicates that it ,ay struggle on the test set.

A screenshot of a computer program

Description automatically generated

Figure :Hyperparameter tuning

Random Forest benefited from hyperparameter tuning, with the best configuration achieving a CV accuracy of 55% , Logistic regression showed a modest increase.

A diagram of different types of confusion matrix

Description automatically generated

Figure : Confusion matrix

A diagram of a logistic regression confusion matrix

Description automatically generated

Figure : Confusion matrix

A graph of different colors

Description automatically generated with medium confidence

Figure :Classification report

Figures 42 and 43 display the results from the best performing models. Clear class imbalance for the negative class is visible.

A graph of a logistic regression

Description automatically generated

Figure :Classification report

To combat the class imbalance SMOTE was implemented as seen in figure 44. However this failed to improve the metrics, one future idea could be to over sample the negative class so there is more data.

A screenshot of a computer

Description automatically generated

Figure :results from SMOTE

## Regression models

A black screen with white numbers

Description automatically generated

Figure :Results

Various models were evaluated for their ability to handle dense small datasets, with a focus on R2 metric. Initial evaluation revealed significant differences in performance, with decision tree and KNN achieving the best scores with 95%.

A group of graphs showing different types of forest predictions

Description automatically generated

Figure :Predicions

The residuals were plotted for visual purposes to gain an understanding of the models performance in figure 46.

A screenshot of a computer program

Description automatically generated

Figure :Hyperparameter Tuning

To improve performance, tuning was conducted using GridSearchCV. The Lasso model was chosen to replace the LinearRregression model as it is a more complex choice, r2 improved slightly.

As the dataset was small, appropriate parameters were chosen for all models to give the model a wide base to choose from (ChenDataBytes, 2024).

A group of graphs showing different types of changes

Description automatically generated

Figure :Predictions after tuning

The Decision Tree, Random Forest and KNN achieved the best scores with the decision tree reaching 15 depth and Random Forest reaching 200 trees, N\_neighbors was chosen as 3.

## Project management

Incorporating both agile methodology and the CRISP-DM framework into the project ensured efficient, iterative progress and effective delivery. The Framework provided a structured approach to data acquisition, preparation, modelling and evaluation. Utilising Github classroom for version control and documentation further enhanced the seamless tackling of challenges

## Dashboard

Streamlit was utilised for the dashboard and effectively incorporates Tufte’s principles for data visualization by emphasizing clarity, efficiency and user engagement. The use of Altair and Plotly for interactive visualisations ensures minimising unnecessary elements while maximising the data presented reflected Tufte’s idea of reducing clutter to enhance understanding. The thoughtful design of heatmaps, Choropleths and bar charts provides a clear and comparative view of agriculture across European countries. Additionally, interactive elements and tooltips, enhance user interaction, enabling deeper exploration of the data (Tufte, 2024).

# Programming and optimisation

Multiple libraries were utilised for this project including Pandas for data manipulation, Matplotlib, seaborn and plotly for data visualization, streamlit for dashboard construction and scikit learn for machine learning, these libraries provided a comprehensive suite for the project.

Adherence to PEP8 guidelines for python coding ensured that the code is clean, readable and maintainable. PEP 8 establishes best practices in terms of code style including comments and spacing (Dixit, 2022) . Additionally , modularity and reusability were prioritised to enhance code maintainability. Functions were designed to handle specific tasks independently, promoting a functional programming approach which minimizes redundancy and allows for easier debugging (Dhanush, 15).

Incorporating robust error handling using try-except blocks ensured that the code gracefully handled unexpected errors, this improves the versatility of the code, they were mostly utilised for reading in datasets.

Utilising databases or API’s offers numerous advantages compared to using traditional CSV files. Databases such as SQL provide robust data storage, faster query execution and data integrity. API’s streamline data integration from external sources. Furthermore, these approaches optimise usability of the models deployment and optimise memory usage by only storing necessary data reducing overheads. APS’s and databases also facilitate seamless integration of web application as seen in section 2 .

Data optimisation strategies for memory management included efficient data serialization, selective data fetching and data compression techniques. By reducing unnecessary columns as seen I the data\_collection notebook, as well as applying datatype optimisation, memory consumption was minimised while maintaining the integrity of the project as evidenced in section 2.

Additionally balancing bias-variance trade-off was crucial for model performance. Implementing strategies such as cross validation, feature selection , and regularisation techniques helped manage this trade-off.

To ensure efficient resource utilization and optimise execution time, each code block was carefully timed, this process helped identify performance bottlenecks and allowed for targeted optimisations, timing was performed using the datetime module, most execution times were between 2 and 18 seconds.

# Conclusion and future work

The analysis of agricultural data across the EU countries has revealed significant insights into CAP and the domestic policies of the countries. Through careful examination of distributions and the application of machine learning and sentiment analysis to get a feeling of the attitude towards farming and agriculture. In Ireland, recent sentiment analysis seems positive as well as increasing growth under CAP. However, going forward, more data for both sentiment analysis and regression would be needed, while the models did perform well, more data is needed to refine the classification models.

# Bibliography

Al-kassab, M. M. T., 2022. The Use of One Sample t-Test in the Real Dat. *https://www.researchgate.net/publication/363256658\_The\_Use\_of\_One\_Sample\_t-Test\_in\_the\_Real\_Data,* 21(2347-1921 ), p. 6.

Anon., 2024. *https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-glance\_en.* [Online]   
Available at: https://agriculture.ec.europa.eu/common-agricultural-policy/cap-overview/cap-glance\_en  
[Accessed 15 12 2024].

Anurag, 2024. *medium.* [Online]   
Available at: https://medium.com/@gusainanurag58/tf-idf-vectorizer-explained-373b3f07d23b#:~:text=The%20Term%20Frequency%2DInverse%20Document%20Frequency%20(TF%2DIDF),collection%20of%20documents%20or%20corpus.  
[Accessed 18 12 2024].

Asghar Ghasemi 1, S. Z. 1., 2012. Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *https://pmc.ncbi.nlm.nih.gov/articles/PMC3693611/,* p. 4.

Brownlee, J., 2020. [Online]   
Available at: https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/  
[Accessed 18 12 2024].

Cátia M. Salgado, C. A. H. P. &. S. M. V., 2016. https://link.springer.com/chapter/10.1007/978-3-319-43742-2\_13. In: M. c. data, ed. *Secondary Analysis of Electronic Health Records .* s.l.:Springer Nature Link, pp. 143-162.

ChenDataBytes, 2024. *medium.* [Online]   
Available at: https://medium.com/@chenycy/accelerate-your-machine-learning-skills-with-hyperparameter-tuning-00dfd27e261c  
[Accessed 18 12 2024].

Dhanush, 15. *dev.to.* [Online]   
Available at: https://dev.to/dhanush9952/a-guide-to-functional-programming-18h9  
[Accessed 18 12 2024].

Dixit, C., 2022. *DATACAMP.COM.* [Online]   
Available at: https://www.datacamp.com/tutorial/pep8-tutorial-python-code  
[Accessed 18 12 2024].

FAO, n.d. *FAO.* [Online]   
Available at: https://www.fao.org/faostat/en/#data/QCL  
[Accessed 01 12 2024].

Firdose, T., n.d. *medium.* [Online]   
Available at: https://tahera-firdose.medium.com/filling-missing-values-with-mean-and-median-76635d55c1bc  
[Accessed 18 12 2024].

Gov, I., 2024. [Online]   
Available at: https://www.google.com/url?sa=i&url=https%3A%2F%2Fassets.gov.ie%2F88207%2Fd8e7465f-496b-4d4f-b1b2-c08abd208bfc.pdf&psig=AOvVaw3sr6WSgN-MA623bJigQhe9&ust=1734606226524000&source=images&cd=vfe&opi=89978449&ved=0CAMQjB1qFwoTCPCcvofosYoDFQAAAAAdAAAAABAE  
[Accessed 15 12 2024].

Gupta, A., 2024. *simplilearn.* [Online]   
Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/what-is-correlation-in-statistics  
[Accessed 18 12 2024].

John T. Hancock, T. M. K. &. J. M. J., 2023. Evaluating classifier performance with highly imbalanced Big Data. *Springer open ,* Volume 42.

Kumar, A., Sep 30, 2024. *.simplilearn.* [Online]   
Available at: https://www.simplilearn.com/what-is-descriptive-statistics-article  
[Accessed 18 12 2024].

Mckinney, W., 2024. pandas: a Foundational Python Library for Data Analysis and Statistics. *research gate* , 15 12, p. 9.

Menon, K., 2024. *simplilearn.* [Online]   
Available at: https://www.simplilearn.com/tutorials/statistics-tutorial/skewness-and-kurtosis  
[Accessed 19 12 2024].

Ourworldindata, 2024. [Online]   
Available at: https://ourworldindata.org/grapher/average-precipitation-per-year?tab=chart&time=2000..2021&showSelectionOnlyInTable=1&country=ESP~IRL~AUT~HRV~CYP~CZE~DNK~EST~FIN~FRA~DEU~GRC~HUN~ITA~LVA~LTU~LUX~NLD~POL~PRT~ROU~SVK~SVN~SWE~BEL~BGR  
[Accessed 01 12 2024].

Prince, M., Jan 23, 2020. *medium.* [Online]   
Available at: https://princeshady5.medium.com/simplifying-your-data-with-pivot-tables-using-python-cb358003c406  
[Accessed 15 12 2024].

reddit, 2024. *REDDIT API.* [Online]   
Available at: https://redditinc.com/policies/data-api-terms  
[Accessed 15 12 2024].

Tufte, E. R., 2024. *The Visual Display of Quantitative Information.* [Online]   
Available at: https://kyl.neocities.org/books/%5BTEC%20TUF%5D%20the%20visual%20display%20of%20quantitative%20information.pdf  
[Accessed 18 12 2024].

Wohlwend, B., n.d. *Medium.* [Online]   
Available at: https://medium.com/@brandon93.w/classification-algorithms-knn-naive-bayes-and-logistic-regression-515bdb085047  
[Accessed 18 12 2024].

A purple and yellow background with white text

Description automatically generated

Figure :Missing values

A close-up of a graph

Description automatically generated

Figure :Missing values distributions

A group of graphs showing different types of data

Description automatically generated with medium confidence

Figure :Columns after median imputation.

A screenshot of a computer

Description automatically generated

Figure :Data types for df

A group of graphs showing different sizes and shapes

Description automatically generated with medium confidence

Figure :Data Distributions

A group of blue rectangular boxes with text

Description automatically generated with medium confidence

Figure : Data boxplots

A screenshot of a computer program

Description automatically generated

Figure : Kurtosis values

A screenshot of a computer

Description automatically generated

Figure :Skew values

A colorful chart with text

Description automatically generated with medium confidence

Figure :correlation coefficients

A graph with a line going up

Description automatically generated

Figure :Word frequency NLP

**A graph of positive and neutral expression

Description automatically generated**

Figure :Class Distribution NLP

**A close-up of words

Description automatically generated**

Figure : Wordcloud most common words

A graph of a number of people

Description automatically generated with medium confidence

Figure :Average Agriculture Value per country

A graph of the number of countries/regions

Description automatically generated

Figure : EU Countries Pesticide Usage

A graph with red and blue lines

Description automatically generated

Figure : Agriculture values scaled with temperature change

A graph of different colored bars

Description automatically generated with medium confidence

Figure :Stacked bar plot

A map of europe with different colored areas

Description automatically generated

Figure :Choropleth of agriculture value

A map of europe with green squares

Description automatically generated

Figure :Choropleth of total precipitation

Figure :People employed Vs population in terms of agriculture value

Figure :People employed Vs population in terms of agriculture value

Figure :People employed Vs population in terms of agriculture value

A screenshot of a graph

Description automatically generated  
Figure :People employed Vs population in terms of agriculture value

A map of europe with different colors

Description automatically generated

Figure :Choropleth of cereals value

A graph of different colored bars

Description automatically generated with medium confidence

Figure :Cereal production

A group of graphs with numbers

Description automatically generated with medium confidence

Figure : QQ plots and histograms of distributions

A black background with white text

Description automatically generated

Figure : Null and alternate hypothesis

A computer screen with numbers and letters

Description automatically generated

Figure 24:Tests for normality

A computer screen with white text

Description automatically generated

Figure 24:Tests for normality

Figure :boxplots of the two samples

A diagram of cereals

Description automatically generated

Figure :boxplots of the two samples

A graph of different colored bars

Description automatically generated with medium confidence

Figure :Barplot of EU BEEF

A black background with white text

Description automatically generated

Figure :Stating hypothesis

A graph and diagram of a graph

Description automatically generated with medium confidence

Figure :QQ plot and Distribution

A black screen with white text

Description automatically generated

Figure : Normality tests for beef



Figure : Z-test results

A graph of a number of people

Description automatically generated

Figure : Pesticide usage EU

A black background with white numbers

Description automatically generated

Figure :Results from Binomial

A graph showing a number of successful

Description automatically generated

Figure : Binomial Distribution

A graph of a number of colored bars

Description automatically generated with medium confidence

Figure :Milk production distribution



Figure :ANOVA results

A computer screen with white text

Description automatically generated

Figure :Chi squared results

A screenshot of a graph

Description automatically generated

Figure :Results table

A screen shot of a computer

Description automatically generated

Figure :ROC-AUC scores

A screenshot of a computer program

Description automatically generated

Figure :CV validation

A screenshot of a computer program

Description automatically generated

Figure :Hyperparameter tuning

A diagram of different types of confusion matrix

Description automatically generated

Figure : Confusion matrix

A diagram of a logistic regression confusion matrix

Description automatically generated

Figure 41: Confusion matrix

A graph of different colors

Description automatically generated with medium confidence

Figure :Classification report

A graph of a logistic regression

Description automatically generated

Figure :Classification report

A screenshot of a computer

Description automatically generated

Figure :results from SMOTE

A black screen with white numbers

Description automatically generated

Figure :Results

A group of graphs showing different types of forest predictions

Description automatically generated

Figure :Predicions

A screenshot of a computer program

Description automatically generated

Figure :Hyperparameter Tuning

A group of graphs showing different types of changes

Description automatically generated

Figure :Predictions after tuning