A picture containing icon

Description automatically generated

**Table of Contents**

Abstract

Data Preparation/EDA/Visualisation

Statistics

Machine Learning

Programming

Conclusions

References

Appendix

**Abstract**

*This report examined two main areas: the livestock population in the EU and global agricultural food prices. Relevant data was collected from open sources, pre-processed, and then used in statistical analysis or machine learning algorithms. The data was also visualised to deepen the understanding. Livestock in Ireland was compared with Latvia in parametric and non-parametric inferential tests. A sentiment analysis was performed on the farming subreddit with good predictive results when logistic regression and Naïve Bayes models were trained. Food prices were clustered and used to train a support vector machine, logistic regression, and multiple linear regression models. These models all performed well. Hyperparameter tuning and cross-validation were successfully applied.*

110

**Data Preparation/EDA/Visualisation**

Common Agricultural Policy

The Common Agricultural Policy (CAP) is a collection of instruments used by the European Union to achieve its goals in the agricultural sector. It can take many forms such as subsidies, market tools and rural development. Since the beginning, its impact on the production and structure of farms in Europe has been profound, so the influence on the statistics collected from members must be considered. For example, Quiroga et al. found that current policies are reducing production rates as more funding is funnelled towards research (Quiroga et al., 2024).

To aid CAP’s effective implementation Eurostat collects census data. This data is critical as it indicates if current policies are working and highlights areas where future policies should be aimed. Eurostat works with other member states to produce a full census every 10 years. These censuses cover 98% of the EU farming sector (Selenius et al., 2021).

145

Data Acquisition and Licencing

Data was obtained from multiple online open resources. The EU countries' livestock dataset was found on the Eurostat website. This data is collected by Eurostat and published for free use without licence or written permission. The relevant metadata, including the units and the dataset-specific vocabulary, accompany the datasets. The website details information on the quality, methodology and evaluation of the data. It is open source and doesn’t require a written licence for academic use or otherwise. The only stipulation is that the data is properly referenced and any changes to it are made clear (“Copyright notice and free re-use of data - Eurostat,” n.d.).

A dataset containing food prices was taken from the FAO. It is a UN organisation which is a source of global agricultural data. The FAO food price index was taken and used from this source in the ML section. The FAO has an open access policy which allows the data on its website to be used freely without a licence (“FAO Food Price Index | Food and Agriculture Organization of the United Nations,” n.d.).

For the sentiment analysis portion of this report, the Python Reddit API Wrapper was employed. This is a library specifically designed to allow access to Reddit posts through Python. It makes use of HTTP requests but has a lot of built-in functionality making it easier to use. A Python app was set up specifically for this report, the details of which were kept in a .env file for security purposes. Reddit’s policies were studied and adhered to (“Data API Terms - Reddit,” n.d.). Obtaining data for the sentiment analysis was the most difficult. Datasets with sentiment scores already assigned relating to agriculture could not be found. The open-source datasets on Kaggle, for instance, did not relate to agriculture. Therefore, the models were trained using the VADER sentiment classifier.

305

Exploratory Data Analysis

Each of the datasets was explored to detect any issues in them. The livestock datasets were read in as TSV files. The country category contained unnecessary data and so was renamed and reformatted. A column with the livestock type was added to each data frame in preparation for the concatenating and melting in subsequent steps. The extract numbers function was applied to each data frame in line 11. This was an important step as it allowed the numbers to be given the type float and numeric operations to be performed on them. Some of the row values were entered with another letter which provided additional information but needed to be removed. A colon was used in place of missing values, this was changed to the numpy NaN on line 13.

A graph with colored dots

Description automatically generatedThere was a high percentage of missing data, 9%. Linear interpolation was applied to these missing values along the rows on line 14. This used the previous and next year’s values for a given country. This method was decided on as there was not a huge variation from year to year in the number of livestock within a country.189

Figure 1: A selection of the missing data from the EU sheep data frame

A graph with colored dots

Description automatically generated

Figure 2: The results of the linear interpolation

The suitability of this method can be seen in Figures 1 and 2. The missing data is replaced with a string approximation which made sense within this context. This was checked for all categories of livestock (Zheng and Casari, n.d.).

The 4 livestock categories were consolidated into a single data frame using concatenation and melting operations, lines 18 and 20.

The sentiment analysis CSV file was read in and melted to create a single column with the text. The data was cleaned in line 55. ML algorithms cannot deal with strings, the strings must be converted to numeric form. The tokenizer removed punctuation and split each string into individual words and all letters were changed to lowercase. The stemming and lemmatisation were applied using Python libraries. These reduce the complexity of the word and aid ML algorithms. Stemming can be counterproductive but when the algorithms were run without stemming the accuracy was reduced (Zheng and Casari, n.d.).

Minor formatting operations were applied to the raw CSV file for the food prices dataset. The date was converted to a useable format. It was visualised on a scatterplot. Prices were stable before 2005 and the analysis was tried by removing this but the accuracy was worse. The dataset was prepared using scaling. It contained numeric values ranging over large values. The features were individually scaled to maximise shape retention. The normally distributed features used standard scaling and the skewed used min-max scaling. The shape was checked after scaling. Comparing the graphs in lines 70 and 73 it can be seen how closely the data mirrors itself (Zheng and Casari, n.d.). 268

Smote was applied on line 89. This is a useful pre-processing step as it evens out the number of data points which belong to each category in a classification algorithm, as shown in Table 1. 35

|  |  |  |
| --- | --- | --- |
| Category | Pre-SMOTE count | Post-Smote count |
| 2 | 186 | 140 |
| 1 | 118 | 140 |
| 0 | 108 | 140 |

Table 1: number of categories before and after the SMOTE algorithm was applied

SMOTE promotes a more balanced result and reduces bias for a certain category (Fernandez et al., 2018).

Choropleths were experimented with and an interactive app using Streamlit was built. The app showed the variation in the livestock of EU countries from 2012 onwards, indicating trends In CAP countries, see Figure 3. The dashboard allowed the price of various agricultural products to be entered along with the year and based on this the food price index could be calculated. This would give invaluable predictive insights to farmers as this dictates their income. Due to the scaling of the data, there were implementation issues. It was attempted to inverse scale the data but it did not work and could not be resolved due to time constraints.

Tufts principles were taken into account for the visualisations. The area of interest was taken, Europe was focussed on, and the labelling was clear and concise. Only the hovered-over country’s statistics were shown. The map was kept to 2-D to reduce confusion and gridlines were omitted (Tufte, 2001).172

A map of europe with different colored countries/regions

Description automatically generated

Figure 3: Choropleth of the EU Livestock population

**Statistics**

A dataset was pre-processed to make it ready for statistical analysis. The data detailed the population of livestock in the EU. A data frame was put together for each country and stored in a dictionary, line 27. This allowed descriptive statistics to be viewed for the livestock of any country.

As Ireland was the baseline ‘.describe()’ was used. Various countries were plotted on histograms to visualise the spread of the data. Countries with similar land areas to Ireland were focussed on as these provided the fairest comparisons. The chosen countries were Latvia, Czech Republic, Lithuania and Croatia (“Largest country in Europe,” n.d.). Figure 4 shows the distributions. There is a variety of distributions and further analysis was required to gain insight. 121

**A screenshot of a graph

Description automatically generated**

Figure 4: The distribution of the livestock population in similarly sized EU countries

All tests were carried out at 5% significance level (Weiss, 2012). The normality of the country’s data had to be established before parametric testing. Latvia’s overall distribution did not meet the threshold but when divided into categories it was found to have a sufficient p-value in the Shapiro-Wilk test, line 32. The normality of the variance was checked and passed with the Levene test. With these, a QQ-plot was created, and the data was adequately close to the probability plot to be considered normal.

A confidence interval was constructed using this normal dataset. As discussed previously this data was census data and so could be considered population data, therefore a normal interval and student-t tests were performed and compared. A random sample of the population data was taken. There is a 95% chance that the population mean lies between the upper and lower limits. It is advantageous to know the population deviation as it reduces the interval size as shown in Table 2. 176

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lower Limit | Sample Mean | Upper Limit |
| Normal Confidence Interval | 209.5 | 251.5 | 292.9 |
| Student’s t-test | 205.9 | 251.5 | 296.5 |

Table 2: Confidence intervals for Latvian Livestock

Multiple hypotheses were tested. The Irish dataset was tested for normality and found to pass in three of the four categories. The sample mean was tested against the population mean, line 42 and found to be equal with 5% confidence. This was also tested using the larger and smaller parameters in the z-test and H0 was accepted in both cases.

Ireland’s sample was compared to Latvia’s in the previously discussed normally distributed livestock categories using a two-sample t-test, line 47. The two distributions were established to be unequal as the p-value was well below the threshold. Using an ANOVA test, Latvia and Ireland’s livestock categories were established to have differing amounts.

The above analysis was based on the data being approximately normally distributed. Non-parametric are less powerful but can be used on all data distributions (Weiss, 2012). Wilcoxon and Kruskal tests were run. The Wilcoxon test concluded that there was no evidence Ireland had a bigger population than the Czech Republic. The Kruskal test, line 51, determined that of the 4 tested countries, at least one had a different population than the others. 183

The biggest challenge faced in this analysis was acquiring data that was normal. The three tests needed to be passed at a significance of 5%. This was done through trial and error which was time-consuming. Various instruments were used to establish if a particular country had normally distributed data. Histograms, boxplots, Q-Q plots, Shapiro-Wilks and Levene tests were used. Many inference tests depend on this which limits the countries which could be analysed (Rani Das, 2016).76

**Machine Learning**

All train-test splits were 80-20 and the random state set to 0 for consistency.

A sentiment analysis was performed on a dataset of Reddit comments about agriculture. Finding pre-classified data was not achieved and so the VADER analyser was used. VADER was chosen as it is commonly used on social media data. It uses a lexicon to classify if a word is positive or negative and based on these scores the overall string. The data was cleaned, see previous sections, and then the sentiment analyser was applied, line 59. Hutto et al. deemed the performance of VADER compared to manually rated strings to be very promising (Hutto and Gilbert, 2014). Before classifying, the strings were transformed using the count vectorizer object. This was chosen over TF-IDF for ease of use. Both performed strongly in the literature (Basarkar, 2017; Suryaningrum, 2023).

A blue squares with white text

Description automatically generatedThe VADER prediction was split into 2 and 3 classifications for comparison. A Naïve Bayes model was trained with positive, negative, and neutral strings. This model performed moderately, with an accuracy of 58%. A logistic regression was performed with the neutral comments removed. This simpler model had a higher accuracy, 70%, see Figure 5. 194

A blue squares with numbers

Description automatically generated

Figure 5: The Naïve Bayes and Logistic Regression Confusion Matrices

A new dataset was read in for the unsupervised portion of the ML. The dataset related to global food prices. Hierarchical clustering was applied to this dataset to group together the years based on the food prices. The dataset was prepared, and various forms of clustering were applied to it, lines 79-84. Multiple linkage methods, cluster sizes and metrics were tried and visualised to determine the optimal one. It was decided that an average linkage method with Euclidean distance and three clusters was chosen as the optimal split, see figure 6.91

A diagram of different colored dots

Description automatically generated

Figure 6: Chosen agglomerative clustering method, distance metric and cluster size

This clustering provided minimal overlap between groups without being overly complex. The average linkage method was found to perform well compared to newer novel approaches (Tokuda et al., 2022).

Next supervised ML algorithms were considered. Based on these groupings each year was low, medium or high in terms of food prices. Models were trained to determine if a year’s category could be predicted based on the food prices.

A red and blue squares

Description automatically generatedSupport Vector Machining was first tried, lines 87-96. Initially, the model did not perform well. The hyperparameters were tuned using GridSearchCV. 89

Figure 7: Heat Map showing the correlation between features and the target variable

Finding the optimal hyperparameters improved the accuracy from 48% to 98%, illustrating the importance of well-chosen parameters.

A logistic regression model was trained to gain insight into the correlation between the various features. This model performed well with the default parameters and so for computational resource reasons a GridSearch was not applied, it had an accuracy of 99%. Figure 7 shows the most strongly correlated values to the type of year were the food price index and cereals.

Lastly, a multiple linear regression was performed, line 103. It was found that a polynomial of degree 2 had the lowest mean square error. A cross-validation was performed on this model. This is a reliable way to measure model accuracy and showed that as the polynomial degree increased the mean square error increased (Rodriguez et al., 2010).135

**Programming**

This report made use of many of the open-source Python libraries associated with data analysis, statistics and machine learning. All of the libraries used were imported in line 1. The functions defined and used were collected and ran in line 2. This prevented a function from being called before it was defined and meant it could be used freely throughout the code. The code was annotated and explained where needed. Markdown cells were made use of. Descriptive variable names were employed to improve clarity. For loops and if statements were used along with functions to minimise copying large blocks of code.

Data was taken from multiple sources. CSV and TSV files from various sources were used and edited using Pandas. The API wrapper for Python and Reddit was utilised. A .env file was deployed with the sensitive information for security purposes. The data scraped from Reddit was saved to a CSV file for analysis. An issue that arose when using PRAW was the 429 error. This occurs when the server detects too much activity and blocks the accessor to prevent server abuse. This limited the number of posts which could be gathered.

There are multiple ways to add and manipulate data. Due to time constraints, this was done in Python, however relational and non-relational database tools could also be used. In a review which compared MongoDB and MySQL, MongoDB was found to have greater flexibility and customisation and to perform better with very large datasets. Non-relational databases have a learning curve and require greater expertise (Győrödi et al., 2015). The file type must also be considered. JSONs store large quantities of data more efficiently but require the use of a nonrelational database due to their unstructured nature (Swami and Sahoo, 2018).

Testing that code is properly functioning is critical. This was done by using the .head() method, as well as .info|(). These allowed the layout and datatype of the data to be viewed. When merging, melting, and concatenating it is crucial to check the operation has been done as needed. For the ML algorithms classification reporting was used to print the accuracy and determine their performance. The pre-and post-scaling data distribution was plotted using histograms, lines 73 and 76, to ensure the code had maintained the data’s shape.

Where possible previous data frames and variables were written over. This prevented data from being saved multiple times and many redundant copies being created. This can introduce some ambiguity and confusion surrounding the variables. If a variable was not used later in the code, it was overwritten.

**Conclusion**

This report successfully found, manipulated, and analysed a variety of data regarding livestock populations and agricultural food prices. It was concluded that Ireland as the baseline had differing levels of livestock compared to similarly sized EU countries both parametrically and non-parametrically. Farming Reddit posts were analysed, and model was built to determine the sentiment of such a post. Years successfully were grouped based on their food prices. An equation was determined to predict food prices each year and presented on an interactive dashboard.

Word count: 2990

**References**

Basarkar, A., 2017. DOCUMENT CLASSIFICATION USING MACHINE LEARNING (Master of Science). San Jose State University, San Jose, CA, USA. https://doi.org/10.31979/etd.6jmu-9xdt

Copyright notice and free re-use of data - Eurostat [WWW Document], n.d. URL https://ec.europa.eu/eurostat/about-us/policies/copyright (accessed 5.20.24).

Data API Terms - Reddit [WWW Document], n.d. URL https://www.redditinc.com/policies/data-api-terms (accessed 5.23.24).

FAO Food Price Index | Food and Agriculture Organization of the United Nations [WWW Document], n.d. . WorldFoodSituation. URL https://www.fao.org/worldfoodsituation/foodpricesindex/en/ (accessed 5.18.24).

Fernandez, A., Garcia, S., Herrera, F., Chawla, N.V., 2018. SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary. Journal of Artificial Intelligence Research 61, 863–905. https://doi.org/10.1613/jair.1.11192

Győrödi, C., Győrödi, R., Pecherle, G., Olah, A., 2015. A comparative study: MongoDB vs. MySQL, in: 2015 13th International Conference on Engineering of Modern Electric Systems (EMES). pp. 1–6. https://doi.org/10.1109/EMES.2015.7158433

Hutto, C., Gilbert, E., 2014. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. Proceedings of the International AAAI Conference on Web and Social Media 8, 216–225. https://doi.org/10.1609/icwsm.v8i1.14550

Largest country in Europe [WWW Document], n.d. . Statista. URL https://www.statista.com/statistics/1277259/countries-europe-area/ (accessed 5.23.24).

Quiroga, S., Suárez, C., Santos-Arteaga, F.J., Rodrigo, J.M., 2024. Do common agricultural policy subsidies matter for the market-environment trade off? An evaluation of R&D objectives and decisions across farmers. Journal of Agriculture and Food Research 15, 101047. https://doi.org/10.1016/j.jafr.2024.101047

Rani Das, K., 2016. A Brief Review of Tests for Normality. AJTAS 5, 5. https://doi.org/10.11648/j.ajtas.20160501.12

Rodriguez, J.D., Perez, A., Lozano, J.A., 2010. Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 569–575. https://doi.org/10.1109/TPAMI.2009.187

Selenius, J., Wirtz, C., Florescu, D., Lazar, A.C., 2021. Agricultural census 2020 – how to reduce costs and burden? The European Statistical System approach. SJI 37, 327–332. https://doi.org/10.3233/SJI-200656

Suryaningrum, K.M., 2023. Comparison of the TF-IDF Method with the Count Vectorizer to Classify Hate Speech. Engineering, MAthematics and Computer Science Journal (EMACS) 5, 79–83. https://doi.org/10.21512/emacsjournal.v5i2.9978

Swami, D., Sahoo, B., 2018. Storage Size Estimation for Schemaless Big Data Applications: A JSON-Based Overview, in: Hu, Y.-C., Tiwari, S., Mishra, K.K., Trivedi, M.C. (Eds.), Intelligent Communication and Computational Technologies. Springer Singapore, Singapore, pp. 315–323.

Tokuda, E.K., Comin, C.H., Costa, L.D.F., 2022. Revisiting agglomerative clustering. Physica A: Statistical Mechanics and its Applications 585, 126433. https://doi.org/10.1016/j.physa.2021.126433

Tufte, E.R., 2001. The visual display of quantitative information. Graphics Press.

Weiss, N.A., 2012. Introductory Statistics. Addison-Wesley.

Zheng, A., Casari, A., n.d. Feature Engineering for Machine Learning.

**Appendix**

A data project chart with colorful labels

Description automatically generated with medium confidenceTime was taken to properly plan the scope and time frame of this project. A Gantt chart was put together and adhered to where possible. When deviated from it was updated and revised.