**Juypter notebooks**

**\*Provided with this report is a README file which details the contents of each notebook to highlight the design and how they meet the criteria for CA2 called README.docx**

1.CA2\_Notebook\_1\_GavinDavis\_sba22311- Data preparation, EDA, Visualisation and Machine Learning models

2.CA2\_Notebook\_2\_GavinDavis\_sba22311- Interactive visualisations for European Milk production and Interactive Dashboard with Plotly Dash

3.CA2\_Notebook\_3\_GavinDavis\_sba22311- Statistical comparison of Ireland Germanys Milk production

4.CA2\_Notebook\_4\_GavinDavis\_sba22311- TwitterAPI and Sentiment Analysis

**Contents of report**

1. Introduction

2. Methodology

3. Sentiment Analysis of Dairy research using Twitter data

4. Understanding how Ireland’s Milk production compares to European countries using interactive visualisations and statistics

5. Data exploration, preparation and Machine learning for Irelands Milk production

6. Machine learning model for Milk production in Ireland

7. Conclusion

8. Bibliography

**List of Figures**

Figure 1: A schematic representation of the approach taken for this study using the CRISP-DM methodology

Figure 2: Results from sentiment analysis of tweet data based on research in Dairy.

Figure 3: Classification report for Gaussian naïve bayes classifier model for Twitter sentiment data based on dairy research tweets.

Figure 4: Interactive dashboard displaying Milk production in Europe which takes input of years 1970 to 2021.

Figure 5: Interactive Line graph showing Milk collection from 1968 to 2021 in European Countries

Figure 6: Interactive Line graph showing Milk collection from 1968 to 2021 in European Countries with focus on Ireland, Italy, France and Germany.

Figure 7: Interactive bar plot for milk production in Ireland and Germany by month for 2022

Figure 8: Interactive bar plot for milk production in Ireland and Germany for years 2014 to 2021

Figure 9: Sample of the first 5 rows of the processed synthetic dataset**.**

Figure 10: Results of GridSearchCV for Linear Regression modal for production of Milk Output

Figure 11: Results of GridSearchCV for Decision Tree Regressor modal for production of Milk Output

Figure 12: Results of GridSearchCV for Random Forest Regressor modal for production of Milk Output

**List of Tables**

Table 1: Descriptive statistics of Ireland and Germanys monthly milk production in 2022.

Table 2: Descriptive statistics of Ireland and Germanys yearly milk production in from 2014 to 2021.

Table 3: Descriptive statistics of combined dataset generated to predict milk sales from Ireland

Table 4: Results from Linear Regression, DTRegressor and RFRregressor

**Title :An analysis of dairy production data in Ireland and the EU**

**Abstract**

Agriculture is an important source of income for many countries in Europe and Ireland. Ireland produces large amounts of meat, poultry and dairy each year which generates critical economic revenue. Of the agricultural exports, dairy and in particular milk production and export is of great importance. This study aimed to statistically evaluate and explore datasets related to Irish milk production. Before conducting research on dairy data a sentiment analysis was performed based on dairy research to assess whether there was positive or negative thinking towards research in dairy using tweet data generated from a Twitter API. By exploring datasets related to dairy and milk production an objective of this study was to generate a machine learning model which would be useful to the Irish dairy sector for the prediction of yearly milk production based on a number of carefully chosen influencing factors such as number of dairy cows, feed stuff price, milk manufacturing prices and the production of other dairy products. Data was collected from data.gov and Eurostat and a combined dataset was generated. The necessary data exploration and cleaning was performed and finally machine learning models were employed. Synthetic data was generated to increase the accuracy of Linear regression based machine learning models. The best model, Linear regression, achieved for the data showed excellent accuracy metrics: R squared 0.997 for training and testing, with RMSE 12.87 and MAPE of 0.143%. In exploring Irish dairy datasets statistical and visual comparisons were made using a number of statistical tests and interactive visualisations to understand how Ireland compares against other high milk producing countries such as Germany. The benefit of this work to the dairy sector in Ireland was an enhanced understanding of how Ireland compares to other European countries in terms of milk production, the sentiment around research in dairy and the generation of machine learning models which could predict milk output given a number of influencing agricultural metrics.

**Keywords** – agriculture, dairy, research, Ireland, milk production, machine leaning

**Abbreviations** – Root mean squared error (RMSE), Mean absolute percentage error (MAPE), application programming interface (API), European union (EU), Confidence interval (CI)

**1.Introduction**

*Project scope:* Agriculture can benefit greatly from the use of machine learning and artificial intelligence. Some of the largest publicly available datasets are from the agricultural sector. One of Irelands and many countries biggest exports are agricultural products, generating large income for the country’s economy, as well as employment. Prominent agricultural products exported by Ireland include meats, poultry and dairy products. Although this sector is critical for Irelands economic performance and revenues, little investment has been put into generating more efficient processes harnessing modern automation and forecasting tools like machine learning and artificial intelligence to benefit the sector. The benefit to the agricultural sector would likely be substantial if time and resources were focused on appropriate data acquisition, generating models for forecasting, prediction, sentiment to name a few. These models, if implemented suitably, would enhance efficiency and sustainability of agriculture in countries where it is a critical part of the economy.

A key area in agriculture for Ireland and many EU countries is dairy production both for the general population and for exports which generate necessary economic revenues. One of Irelands largest agricultural exports is dairy products which generated in excess of 5 billion euros for the Irish economy in 2021, making it one of the largest income generators for the agricultural sector in Ireland. Owing to this, this project has focused on analysing the available data on dairy production in Ireland comparing it to other countries in the EU. Data was collected from multiple sources namely, data.gov, the central statistics office and Eurostat and combined to generate models for exploring the production of milk in Ireland and different countries. The data was explored, prepared and statistically evaluating to understand and make inferences about Irelands milk production performance, and how it compares to other EU countries such as Germany which is the largest producer of milk in the EU.

**2.Methodology**

**2.1.Approach-CRISP-DM**

The CRISP-DM approach is widely used as the industry standard set of criteria for carrying out a data mining project. It is the most widely used and complete methodology to carrying out a data mining project in comparison to SEMMA or KDD, and was thus selected for this study (Schröer, Kruse and Gómez, 2021).

**Diagram

Description automatically generated**

**Figure 1: A schematic representation of the approach taken for this study using the CRISP-DM methodology.** 1-Business understanding describes getting an overview of the business and/or organisation in question, describing the projected project goals and expected outcome. 2-Data understanding describes collecting the data, exploring it and describing it using statistical analysis and visualisation. 3- Data preparation describes preparing the data for modelling by cleaning and feature engineering making the available data suitable to be used in ML models. 4-Modeling describes choosing an appropriate model that that fits the initial question and the gained understanding of the data explaining the choice and the parameters set. 5-Evaluation describes evaluating the results and discussing them in line with the objectives of the projected outcome. 6-Deployment presented as a final report or software component, including the plan for deployment and how to monitor and maintain(Schröer, Kruse and Gómez, 2021).

**2.2.Programming tools**

For statistics, data preparation/visualisation, and machine learning anaconda navigator was used along with Jupyter notebook as a coding interface. The language of choice for the project was python. All code files and outputs are provided alongside this report.

**2.3.Statistics**

As previously stated, statistical analysis was performed in Juypter notebook using python. Python libraries such as scipy.stats and math were used in the analysis of datasets while also harnessing the inbuilt functions which python provides to ascertain statistical measures such as mean, max, min, mode, variance and standard deviation. Confidence intervals, t-tests and ANOVA was also performed with scipy.stats functions.

**2.4.Data preparation and visualisation**

For data preparation and visualisation python libraries pandas, numpy, matplotlib.pyplot, plotly.express, JupyterDash and seaborn were used along with in-built python functions to visualise and prepare data in order to gain an understanding of the data to generate appropriate questions for this study and to prepare visualisations such as bar plots, choropleth maps, interactive visualisations and heatmaps to ascertain any trends within the data that could be seen. Dash and Ploty express were used to generate interactive dashboard.

**2.5.Machine learning**

Given the data provided in datasets was labelled a supervised learning approach was used. Supervised machine learning is used for labelled datasets, given the Trips dataset had labelled features, the supervised ML approach was deemed to be suitable. The models harnessed in this study were chosen according to the available literature on traffic prediction. In light of this KNN, RF, LR, along with DTRegressor, KNRegressor and RFRegressor were used to develop predictive models from the DCC datasets. From a comparison standpoint, Naïve bayes classifier and Support Vector Machine were also applied (Müller and Guido, 2016).

**2.6 Version Control**

In order to maintain work completed and store completed tasks necessary for the project a GitHub repository was set up. GitHub link:

<https://github.com/GavnDavisCCT/CA2_Agriculture_GavinDavis.git>

**3. Sentiment analysis of dairy research using Twitter data**

*\*This section details the use of a Twitter API set up with a Twitter developer account to scrape tweets and perform a sentiment analysis. The code for this section is present in Notebook 4.*

Before exploring data related to milk production it was decided to assess the sentiment surrounding research in dairy by using a Twitter API. To do so, Tweets using the keyword filters “Dairy” and “Research” were used to acquire tweet data to perform a sentiment analysis to understand whether the attitude towards research in the dairy sector was positive or negative. The code for the Twitter API and resultant sentiment analysis is presented in Notebook 4. A Twitter developer account was set up to acquire the necessary access keys such as the access and bearer tokens required to connect to the Twitter API and scrape tweets. After this tweets were scraped based on the keywords Dairy and Research. NLTK was used to process the tweet data by removing frequently used words, also know as stop words, as well as symbols such as hashtag and punctuations, to generate clean text data. The data was then stemmed which combines similar words into a single or abbreviated version of the word. With the resultant cleaned text data the sentiment was assessed using TextBlob, which assesses the sentiment of text data using the measures subjectivity and polarity. The polarity of the tweets was used to derive the sentiment on the basis of how positive or negative the value generated was within the bounds of -1 and 1. Tweets with a polarity greater than 1 were deemed to be positive in sentiment, while tweets less that 0 were negative and finally tweets with a polarity of 0 were deemed neutral. After the tweets were assigned a sentiment, the sentiments were encoded with values of 0, 1 and 2 representing neutral, negative and positive, respectively. As shown in Figure 2, the sentiment around dairy and research is largely positive and neutral with the lowest number of tweets represented by negative sentiment. This indicates that research in Dairy is largely considered positive as well as a large proportion of people being impartial or neutral to the topic. This analysis highlights that generating new ways to research and develop the dairy sector is largely seen as a positive thing giving this project good basis. The basic twitter developer account is limited in its search capabilities to recent tweets, therefore a small dataset was acquired for the analysis. The dataset included 63 tweets. To further utilize this data, a machine learning model which is capable of predicting sentiment from raw tweet data was generated using Count vectorizer which breaks down and vectorizes words which can be identified by machine learning models, which cannot recognize text data. After vectorization, gaussian naïve bayes classifier was used to generate a model which could predict the sentiment based on future text data about dairy research given it was a classification task based on 3 classes, neutral, positive and negative. The results of the model are shown in Figure 3. A accuracy of 0.49 was achieved meaning the model has approximately 50% accuracy. This is a low accuracy for a classification model and means the model only has a 50% chance of being correct. Future work would include increasing the size of the dataset to increase the model accuracy given the dataset was small. This would be done by incorporating the next token field in the twitter API, however when collecting tweet data no next token was generated for the tweet data response to increase the size of the dataset.

Chart, bar chart

Description automatically generated

**Figure 2: Results from sentiment analysis of tweet data based on research in Dairy.**

**Table

Description automatically generated**

**Figure 3: Classification report for Gaussian naïve bayes classifier model for Twitter sentiment data based on dairy research tweets.**

**4.Understaning how Ireland’s Milk production compares to European countries using interactive visualisations and statistics**

*\*This section discusses Sections 1 and 2 of Notebook 3*. *This section details the EDA, data preparation, statistics and visualisations for statistically comparing Milk production in Europe, and in particular, Ireland and Germany. Code for interactive visualisations and interactive dashboard is present in Notebook 2.*

Given that one of Irelands largest dairy exports is Milk, it was chosen to compare Irelands Milk production against other European countries, to evaluate its performance. To carry out this analysis data was acquired from Eurostat agricultural produce section using monthly and yearly data to create interactive visualisations, dashboards and statistical comparisons to compare Irelands performance against other European countries. According to historical production stocks found at : <https://agriculture.ec.europa.eu/system/files/2021-11/eu-dairy-historical-production-stocks-series_en_0.pdf>., D1110D denotes the total Milk collection for European countries which is a reflection of milk production. This data was used to understand differences between milk production in different European countries as shown in Figure 4, Figure 5 and Figure 6. Figure 4 highlights the variance in milk production in European countries from 1970 to 2021. Figure 4 is an interactive dashboard which would allow stake holders such as farmers to easily query a year of interest and ascertain Irelands milk production performance against the rest of Europe. The choropleth map also has hover capability to see the exact value for a country in a particular year. To prepare this dataset to be used for visualisations and statistics EDA was performed. Null values were removed with dropna along with the removal of unnecessary columns. The dataset also contained other dairy outputs such as cheese, butter, skimmed milk and cream. To create the interactive dashboard with a choropleth map Milk production was isolated from EU agricultural production dataset using the D1110D code using . query and merged with a dataset containing the ISO 3166-1 alpha-3 codes used by plotly to generate choropleth maps, merging the datasets on the basis of alpha 2 code being present in both. A submit button was created using plotly and jupyter dash which allows the user to enter a year and visualise the milk production based on the plotly plasma colour grading scale. Plasma was chosen for colour scheme as it is easy to see differences visually and is widely used on choropleth maps. It can be seen that France and Germany are the highest producers of Milk in Europe from the interactive dashboard created. Focusing in on Ireland, Germany and France in Figure 6 it can be seen that there is a large difference between Ireland’s milk production and that of Germany and France. Plotly was chosen to generate these interactive line plots in Figure 5 and 6 as they allow one to hover over the line for each country and see the exact values at different years. One can also remove countries simply by clicking on the legend to show only countries of interest as shown in Figure 6. This was important for comparing countries of greater interest such as Ireland, Germany and France.

Given Germany is the largest Milk producer in Europe in recent years it was chosen to statistically and visually evaluate differences between Ireland and Germany to confirm the visual differences observed in Figure 4, 5 and 6. To statistically evaluate differences between Ireland and Germany, and test whether this trend is occurring in 2022, monthly data from 2022 and yearly data from 2014 to 2021 was statistically evaluated using 95% confidence intervals, two sample t-tests with equal and unequal variances. Both the monthly and yearly data was checked for normality using a Q-Q plots as can been seen in Notebook 2, Section 1 and 2. Plotly express was used to plot Ireland vs Germany monthly and yearly data by plotting a grouped bar chart with hover capabilities to see the exact value. Blue and dark navy colours were chosen for Ireland and Germany for contrast purposes, to highlight large differences between Ireland and Germany as seen in Figures 7 and 8. Table 1 shows descriptive statistics from Ireland and Germany regarding the 2022 monthly milk production data. It highlights the mean, standard deviation, min, max, 25 %, 50% and 75% quantile values. The mean value of milk production for Germany is much higher that Ireland (almost 2,000,000 tonnes). The units are 1000 tonnes. Despite higher values the standard deviation is higher for Irelands milk production indicating greater variation in Irelands milk production month to month and this is reflected in the visualisation in Figure 7. The greater variation in milk production between Ireland and Germany’s milk production is exemplified by the large variance between the minimum value of milk production for Ireland (188.7) and the maximum value (1200.57). Overall these descriptive statistics, along with 95% confidence intervals, indicate that Germany’s milk production on average is greater per month than Ireland’s with less variation.

Table

Description automatically generated

**Table 1: Descriptive statistics of Ireland and Germanys monthly milk production in 2022.**

The 95% confidence intervals were 578.09, 1107.16 for Ireland and 2599.27, 2763.42 for Germany. From these monthly milk collection confidence intervals it could be concluded that overall Germany’s average milk production was greater than Ireland’s in 2022 coinciding with the visual representations. Ireland milk production in 2022 was particularly low compared to othe other months. A t test was performed to test whether the value from January was normal and fell within the average at a 5% significance level. The null hypothesis was that the mean was equal to 188.70 (Januarys value) and the alternate hypothesis was that the mean was not equal. The results of the test were that this value was significantly different from the average monthly value Ttest\_1sampResult (statistic=19.336260209629756, pvalue=2.466612597299856e-07). Therefore, the null was rejected. To further assess this, a hypothesis test was used to statistically evaluate whether the mean milk production for Ireland was different to that of Germany’s. The null hypothesis was that the means for Ireland and Germnay were equal, while the alternative hypothesis was that the mean values are not equal. The data was tested for normality using a QQ plot although t tests assume normality. To test whether the null hypothesis was true a two sample t test was used testing with both equal and unequal variances. As expected from the visualisations and confidence intervals for the monthly Milk production for Ireland and Germany the P-value was well below 0.05. Equal variance = Ttest\_indResult(statistic=-15.308579200040041, pvalue=5.6279355912462256e-11) Unequal variance = Ttest\_indResult(statistic=-15.308579200040041, pvalue=5.019350160456182e-08). These tests indicate that the null hypothesis should be rejected and the alternate hypothesis accepted, thus concluding that Ireland and Germany’s monthly milk production is statistically different on average for 2022.

To further analyse differences between Ireland and Germany’s milk production yearly data was acquired from Eurostat on Ireland and Germany’s milk production from 2014 to 2021. The data was cleaned as before, unnecessary columns were removed, null values were removed columns were renamed as the year, and melt/pivot functions were used to organise the data in a way which could be analysed effectively. As before yearly, milk production data was assessed for normality using Q-Q plots for Ireland and Germany from 2014 to 2021. The yearly milk production data for Ireland vs Germany is shown as a bar plot in Figure 8. As with the monthly data blue and dark navy were used as the colours of choice to highlight stark differences between Ireland and Germany’s yearly milk production. Each year Germany out performs Irelands milk production by approximately 6 -fold, however in recent years Ireland’s milk production has increased almost 2-fold making the difference between Ireland and Germany’s milk production smaller.

Table

Description automatically generated

**Table 2: Descriptive statistics of Ireland and Germanys yearly milk production in from 2014 to 2021.**

Descriptive statistics were employed to gain an understanding of the yearly Milk production data in Ireland and Germany as with the monthly data for 2022. As shown in Table 2, the mean for Germanys Milk production between the years of 2014 and 2021 is more than 3 fold higher than Irelands between the same years indicating Germany produces on average more milk yearly than Ireland. Again despite having far greater Milk production than Ireland the variance in Irelands Milk production is greater than Germanys. This greater variance is likely due to the dramatic increase in Irelands milk production between the years 2014 and 2021, as shown by the Irelands minimum production (min =5816.23) and max production (max=9018.37). This large difference between Ireland and Germany is further highlighted in the interactive visualisations shown in Figure 4, 6, and 8.

As with the monthly data from 2022, 95% confidence intervals were generated for Ireland and Germany’s milk production between 2014 and 2021. Ireland 95% CI = (6641.66, 8439.81), Germany 95% CI = (31741.02, 32406.12). These confidence intervals for Ireland and Germany highlight that Germany’s Milk production has remained relatively stable over the years 2014 to 2021 while Ireland CI has larger range indicative of the large change over the years 2014 to 2021. As with the monthly data and in line with visual representations in Figures 4 to 8, two sample t tests with both equal and unequal variance were performed to test whether the difference between Ireland and Germany’s yearly milk production was statistically significant. The null hypothesis was the means are equal while the alternate hypothesis was the means for yearly milk production in Ireland and Germany were unequal. Similar to the monthly data the null hypothesis was rejected meaning the means for yearly milk production data for Ireland and Germany were significantly different. Equal variance = Ttest\_indResult(statistic=-60.51580756208014, pvalue=2.437727370551143e-18), Unequal variance = Ttest\_indResult(statistic=-60.51580756208014, pvalue=6.296630294304647e-13). Taken together these results indicate that Ireland and Germany have vastly different milk production both in 2022 and across the years of 2014 to 2021.

Map

Description automatically generated

**Figure 4: Interactive dashboard displaying Milk production in Europe which takes input of years 1970 to 2021.** Dashboard enables visualisation of Milk production in different European countries based on year. Code and functioning plot present in Notebook 2.

Chart, line chart, histogram

Description automatically generated

**Figure 5: Interactive Line graph showing Milk collection from 1968 to 2021 in European Countries.**

Chart, line chart

Description automatically generated

**Figure 6: Interactive Line graph showing Milk collection from 1968 to 2021 in European Countries with focus on Ireland, France and Germany.**

**Chart, bar chart

Description automatically generated**

**Figure 7: Interactive bar plot for milk production in Ireland and Germany by month for 2022.**

**Chart, bar chart

Description automatically generated**

**Figure 8: Interactive bar plot for milk production in Ireland and Germany for years 2014 to 2021.**

**5. Data exploration, preparation and Machine learning for Irelands Milk production**

*\*This section discusses Sections 1, 2 and 3 of Notebook 1*. *This section details the EDA, statistics and data preparation for the generation and comparison of machine learning models which could predict milk production in Ireland.*

The initial part of the analysis according to the CRISP-DM framework was to understand the aim of the business. In this case the business is the Irish Dairy Sector and the potential desired outcome of exploring, presenting and generating models from Dairy production data. Given there are a vast array of agricultural exports from milk, cheese, butter, skimmed milk powder, and that milk is one of the largest exports, Milk was chosen as the focus of this project. Data was collected from Data.gov for Ireland specific milk production data, and from Eurostat for European milk production data including Ireland. Citing the CRISP-DM framework again, the aim of the project was to explore, compare, statistically evaluate and prepare milk production data from Ireland with the aim of generating a machine learning models which could predict future trends in milk production.

The second point of the CRISP-DM framework is data understanding, after concluding that Milk production was the focus of the project a number of datasets were collected and combined which were directly related to dairy and milk production. These datasets included: Feed stuff price, Number of Dairy cows, Milk consumption, Milk manufacturing prices and production of Dairy products such as skimmed milk, cheese and butter. These datasets were collected imported into Jupyter notebook and combined with the aim of generating a combined dataset incorporating a number of factors which could influence the sale/production of Milk in Ireland. The data preparation required to clean and combine the different datasets addresses the third caveat of the CRISP-DM framework, data preparation. This resulted in a dataset that consisted of 7 rows representing the years 2014 to 2021 with 27 columns representing a number of features which were chosen to influence Milk Sales, the chosen target variable. Citing CRISP-DM once again, the next step in the project was data preparation. A considerable amount of preparation was required to generate this dataset making use of the pivot, melt and merge functions in the Pandas library to align the independent features such as Feed stuff price, Milk manufacturing price, Feed stuff price, Milk consumption, Dairy production and number of dairy cows. The dataset was cleaned of null values which would affect the usefulness of the dataset and effect any machine learning model performance and statistics. Semi-skimmed milk and skimmed-milk sales were remove from the dataset due to the presence of null values in 2020 and 2021 which would affect the performance of potential ML models. To understand the distributions of the accumulated features in the combined dataset histograms were utilized using the .hist function in python which can be seen in Section 2, Notebook 1. The variables showed distinct distributions as expected given the differences in units of measure and large variances in values for each feature.

Table

Description automatically generated

**Table 3: Descriptive statistics of combined dataset generated to predict milk sales from Ireland**

To further understand and assess the differences in values of the features the descriptive statistics was performed on the dataset which provides statistical measures describing each of the features such as count, mean, standard deviation, minimum value, maximum value, as well as quantiles. The mean can give an indication of large variances in data in a dataframe, shown in Table 3. As seen, the mean values of the different features ranges from approximately 0.3 to 300. When considering implementation into machine learning algorithms this is too large a variance to generate an accurate model and therefore at this point in the analysis it was noted that scaling would likely be needed for use of the dataset in machine learning. Another important statistical measure is correlation between features and this can have important implications for the choice of machine learning algorithms which can be effectively used. For this reason a correlation heatmap was generated to assess whether there were correlated features. As shown in Section 2 of Notebook 1, the correlation heatmap highlighted the degree of correlation between features of the combined dataset. Accepting that this heatmap is highly complex given the number of features in the dataset, the colouring of the heatmap was particularly important in highlighting features that were heavily correlated and features that were not. Given many of the features included in the dataset seemed to be correlated it was decided that ML models based on linear regression would be most suitable for the generation of a model which could predict milk sales.

Next in the CRISP-DM framework is modelling. To test whether the dataset could be used to predict Milk sales in Ireland before making any more adjustments to the dataset, a linear regression model was tested. At this point it was noted that skimmed milk sales and semi-skimmed milk sales contained null values in 2020 and 2021, which also likely caused their low correlation with the target variable Milk Sales, so they were removed. All Milk Sales was set as y the target variable and the scaled dataset as X for splitting using train test split from sklearn. This allows the splitting of the data into training and testing sets which can be used to train and test the model and assess its accuracy. This model resulted in a R squared for training accuracy of 1.0 and an R squared for testing accuracy of -0.21. The RMSE and MAPE was 21.54 and 3.33% respectively. This model was likely overfit giving the large difference between training and testing accuracy, with a very low test accuracy. The RMSE and MAPE were acceptable but the model accuracy was unacceptable. To compare another regression model with for this dataset to attempt to achieve higher accuracy, decision tree regressor was used. The decision tree regressor model generated had a training R squared of 1.0 and testing R squared of -0.79, with a RMSE and MAPE of 26.22 and 4.33%, respectively. Although the testing R squared showed better accuracy the training R squared was 1.0, still suggesting overfitting. Finally, Random Forest regressor was implemented generating a training R squared of 0.74 and testing R squared of -0.83, with a RMSE of 26.5 and a MAPE of 3.77%. Random forest regressor results indicated that the model was underfitted given the training R squared was lower than the testing R squared. One anomalous factor is the R squared for training was positive and negative for testing in all models. Given most of the features showed positive correlations from the correlation heatmap this doesn’t make sense. At this point it was decided to carry out further data preparation and processing while changing the target variable from Milk sales to Milk production in tonnes after adding another output from a newly acquired dataset from Eurostat, as shown in Section 3.3, Notebook 1.

In section 3.4 of Notebook 1, linear regression was performed using the target variable all milk sales to assess whether the addition of Milk production from the newly acquired dataset firstly as a independent feature added more accuracy to the model. A training R squared of 1.0 and testing R squared of 0.9 was achieved with a RMSE of 6.1 and a MAPE of 0.97%. Although these metrics were promising, when attempting to cross validate nan values were obtained indicating that the model was unstable for this target variable. Principal component analysis was attempted to try and reduce model complexity further but did not change the accuracy of the model or the outcome of cross validation. After these results, it was decided to test the models for their ability to predict another target variable, Milk output of Ireland in tonnes. If this variable could be predicted Milk sales could be inferred knowing the price of milk per litre for a given year. To increase the chances of achieving a higher model accuracy the dataset set was further processed to reduce complexity by removing the large number of feed stuff variables basing the decision on variables that were least correlated with the target variable.

After changing the target variable to Milk output in tonnes the linear regression ML model was implemented again : Training R squared was 1.0, testing R squared was 0.94 with RMSE of 254.31 and MAPE of 2.69%. Again a training accuracy of 1.0 suggests the model is overfitted and when cross validation was attempted nan values were obtained indicating instability across iterations. At this point given no other data was available for the chosen features in the dataset from Data.gov or Eurostat it was chosen to generate synthetic data using synthetic data vault (sdv) to increase the size of the dataset with the aim of generating a model with enhanced stability. With year chosen as the primary key sdv was employed to generate 60 new rows of data which hypothetically represented 60 years worth of data for the chosen features. After performing dimensionality reduction by feature selection based on poor correlation the dataset now consisted of 60 rows and 8 columns

Table

Description automatically generated

**Figure 9: Sample of the first 5 rows of the processed synthetic dataset.**

As previously, the data was processed by scaling using standard scaler to remove large differences in independent variables. The independent features were defined as X and the dependent or target variable Milk Output was defined as y. The data was split into training and testing sets using train test split using 30% for testing. The accuracy metrics were as follows: R squared training 0.997 and R squared testing 0. 997 with an RMSE of 12.87 and a MAPE of 0.14%. The second last phase of CRISP-DM is validation, to do this accuracy metrics described above were used, along with cross validation and GridSearchCV. Unlike the other models the linear regression model showed good stability across 5 split cross validation with very consistent R squared values ([0.9992493 , 0.99963002, 0.99925024, 0.99983882, 0.9996828 ]). Gridsearch CV was performed and the results are displayed in Figure 10. In order to compare the accuracy of different regressor models to assess whether they would be better suited to the data, Decision Tree regressor (DTRegressor) and Random Forest Regressor (RFRegressor) were implemented and compared to the previously generated linear regression model. GridSearchCV was also performed on these models to assess stability. The results are displayed in Figures 11 and 12. Table 4 shows the R squared for training and testing along with the RMSE and MAPE for each model.



**Table 4: Results from Linear Regression, DTRegressor and RFRregressor**

Chart, line chart

Description automatically generated

**Figure 10: Results of GridSearchCV for Linear Regression modal for production of Milk Output.**

**Chart, line chart

Description automatically generated**

**Figure 11: Results of GridSearchCV for Decision Tree Regressor modal for production of Milk Output.**

**Chart, line chart

Description automatically generated**

**Figure 12: Results of GridSearchCV for Random Forest Regressor modal for production of Milk Output.**

Overall the best performing machine learning model for the synthetic data generated to predict milk production for Ireland based on a number of dairy outputs and metric appears to be the linear regression model. The linear regression model appears most stable across increasing number of features and the highest average R squared with the lowest RMSE and MAPE. The fact that synthetic data was required to be generated for an accurate model highlights the need for the dairy sector to collect and publish more historical data based on dairy production for the generation of effective machine learning models. The final phase of the CRISP-DM framework involves deployment and translating the findings of the project to stakeholders. This would take the form of this report and also the interactive dashboard and visualisations prepared detailing the data.

**6.Conclusion**

The aim of this study was to compare Irelands milk production performance against other EU countries, along with the generation of a machine learning model which could predict Irelands milk production from a number of chosen dairy related features. To do this statistics, EDA, data preparation, data visualisation and machine learning techniques were used. The sentiment around research in dairy is seen largely positive or neutral according to the sentiment analysis performed in Section 3. This gave a good basis for the study. Given Irelands milk production is an important revenue generator for the Irish economy it was chosen to compare Ireland’s milk production against the rest of the EU. Interactive visualisations and a dashboard was compared which highlighted differences between Ireland and Europe, and in particular, Germany. This dashboard could be used by stakeholders to visualise and query years, while also being extended to other dairy output such as cheese and butter. To deepen the understanding of differences between Ireland and Germany’s milk production both descriptive and inferential statistics were used to validate any visual differences observed, detailed in Section 4. The statistical comparison of Ireland and Germany’s milk production highlighted that Irelands milk production is significantly lower than Germanys on a monthly and yearly basis. Irelands milk production yearly has however been increasing yearly since 2014 indicating an increase level of production over the years 2014 to 2021. After this a number of datasets related to Irelands milk production and dairy sector metrics were combined to generate a model which could be used to predict Irelands milk production resulting in an accurate Linear regression machine learning model detailed in Section 5. The accuracy metrics of the model were as follows: R squared training 0.999, R squared testing 0.999, with an RMSE of 12.87 and a MAPE of 0.143%, indicating that this is a very accurate model, which was further shown by cross validation. It has been openly surveyed and stated that stakeholders in the agricultural sector and dairy sector would support the use of big data to benefit productivity and profit for the agricultural sector (Osinga *et al.*, 2022). This study is one of many studies which attempted to increase the productivity and sustainability of a part of the agricultural sector. Forecasting models for soya yield and optimization of crop yields are but a few of the developing models used to enhance agricultural productivity. What was made clear within this study is the data quality and cohesion for the generation of such models must be increased by implementation of more rigorous data collection policies, which would allow data analytics and machine learning researchers to more rapidly and efficiently develop and deploy models to benefit the sector.

**7.Bibliography**

Müller, A. C. and Guido, S. (2016) *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O’Reilly Media, Incorporated. Available at: https://books.google.ie/books?id=qjUVogEACAAJ.

Osinga, S. A. *et al.* (2022) ‘Big data in agriculture: Between opportunity and solution’, *Agricultural Systems*, 195. doi: 10.1016/J.AGSY.2021.103298.

Schröer, C., Kruse, F. and Gómez, J. M. (2021) ‘A systematic literature review on applying CRISP-DM process model’, *Procedia Computer Science*, 181, pp. 526–534. doi: 10.1016/J.PROCS.2021.01.199.