**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Programming for DA  Statistics for Data Analytics  Machine Learning for Data Analysis  Data Preparation & Visualisation |
| **Assessment Title:** | CA2 – The Rise of Veganism & it’s impact on Agriculture; a comparative analysis. |
| **Lecturer Name:** | Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Rhianne Fahey |
| **Student Number:** | Sbs22034 |
| **Assessment Due Date:** | 22nd May 2022 |
| **Date of Submission:** | 22nd May 2022 |
| **Word Count:** | 3398 excluding references, comments, titles, citations, title page  5374 inclusive of all |
| **Link to Version control Address** |  |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Contents

[Table of Figures 4](#_Toc104140557)

[Abstract 5](#_Toc104140558)

[Intro 5](#_Toc104140559)

[Data Exploration 5](#_Toc104140560)

[Missing values 6](#_Toc104140561)

[Merge Data 6](#_Toc104140562)

[Feature Engineering 7](#_Toc104140563)

[Data Preparation & Comparative Analysis 7](#_Toc104140564)

[Meat Production/ Consumption 7](#_Toc104140565)

[Dairy Production/ Consumption 9](#_Toc104140566)

[Crop Consumption 12](#_Toc104140567)

[Box Plots per food Type 15](#_Toc104140568)

[Inferential Statistics 16](#_Toc104140569)

[Probability Plots 17](#_Toc104140570)

[Shapiro Tests 18](#_Toc104140571)

[Analysis between countries 19](#_Toc104140572)

[Results of Mann-Whitney 19](#_Toc104140573)

[Data correlations 20](#_Toc104140574)

[Modelling 23](#_Toc104140575)

[Sentiment Analysis 23](#_Toc104140576)

[Sentiment Results 24](#_Toc104140577)

[Prediction Model Selection 24](#_Toc104140578)

[Standardise/ Normalise data 25](#_Toc104140579)

[Model Parameters 25](#_Toc104140580)

[Conclusion & Limitations 29](#_Toc104140581)

[Bibliography 31](#_Toc104140582)

# Table of Figures

[Figure 1 Top Ireland Meat Production vs Consumption 8](#_Toc104140583)

[Figure 2 Germany Meat Production vs Consumption 8](#_Toc104140584)

[Figure 3 Top Australia Meat Production vs Consumption 9](#_Toc104140585)

[Figure 4 Fig Ireland Dairy Production vs Consumption 10](#_Toc104140586)

[Figure 5 Other Ireland Dairy Production vs Consumption 10](#_Toc104140587)

[Figure 6 Top Germany Dairy Production vs Consumption 11](#_Toc104140588)

[Figure 7 Other Germany Dairy Production vs Consumption 11](#_Toc104140589)

[Figure 8 Top Australia Dairy Production vs Consumption 12](#_Toc104140590)

[Figure 9 Other Australia Dairy Production vs Consumption 12](#_Toc104140591)

[Figure 10 Ireland Crop Consumption 13](#_Toc104140592)

[Figure 11 Germany Crop Consumption 14](#_Toc104140593)

[Figure 12 Germany Crop Consumption 14](#_Toc104140594)

[Figure 13 Ireland Outlier detection per food item using Box plots 16](#_Toc104140595)

[Figure 14 Descriptive statistics for Ireland Dairy combined 16](#_Toc104140596)

[Figure 15 Descriptive statistics for Ireland Dairy by product 17](#_Toc104140597)

[Figure 16 Probability plots for Ireland dataset 18](#_Toc104140598)

[Figure 17 Ireland dataset for ML correlation between features 21](#_Toc104140599)

[Figure 18 Germany dataset for ML correlation between features 22](#_Toc104140600)

[Figure 19 Australia dataset for ML correlation between features 23](#_Toc104140601)

[Figure 20 Modelling Results 25](#_Toc104140602)

[Figure 21 Feature importance for Decision Tree – Ireland 26](#_Toc104140603)

[Figure 22 Feature importance for Random Forrest – Ireland 26](#_Toc104140604)

[Figure 23 Feature importance for Decision Tree – Germany 26](#_Toc104140605)

[Figure 24 Feature importance for Random Forrest – Germany 27](#_Toc104140606)

[Figure 25 Feature importance for Decision Tree – Australia 27](#_Toc104140607)

[Figure 26 Feature importance for Random Forrest – Australia 27](#_Toc104140608)

[Figure 27 Explainer Dashboard Modelling Results Example 28](#_Toc104140609)

[Figure 28 Explainer Dashboard Future Predictions Example 29](#_Toc104140610)

# Abstract

In 2021, the vegan food market was worth over $16 billion and is expected to reach $22 billion by 2025. A few years ago, veganism was not widely known, and sentiment was relatively low. Today, just over 1% of the world are identifying as vegan specifically and it has been noted that if this growth continues it is predicted to increase to one in 10 people within the next 10 years (The Vegan Society, 2021). Veganism or a more plant based diet is known to have multiple health benefits including a decrease in heart disease, diabetes and a reduction in hormone imbalance in females. Specifically in Ireland, according to the National Dairy Council (NDC), “41% of Irish women and 30% of Irish men are now avoiding or limiting their dairy consumption, with one in 10 believing that cow’s milk is unhealthy”, (Sarah Harford 2018). What impact does all this have on agriculture and food production globally? Inspired by the work of Richardson et al in their article on ‘Consumer attitudes to meat Eating 1994, this report details the more modern impact of veganism and plant based diets on Irish production and consumption of meat/ dairy /crops and compares Irish agriculture to other countries where the vegan population is growing at a substantial rate, (Richardson, MacFie and Shepherd, 1994).

# Intro

This analysis will focus on how production and consumption has changed over time in line with consumer diet trends and thus impacted Ireland’s agriculture in comparison to other counties where veganism is also on the rise. Ireland will function as the baseline country for this analysis.

The Irish agriculture findings will then be compared against Germany as both countries benefit from the European Common agricultural Policy. The analysis will be looking for similarities & differences in each countries agriculture industry. Germany has been chosen as a country for analysis as it is one of the countries with the highest % of vegans in Europe, (The Vegan Society, 2021). The analysis will go on to compare Irish agriculture against Australia, one of the largest vegan populations in the world who do not benefit from being a member of the CAP to see if there are any major differences or similarities to between EU and Non-EU countries.

This report will be structured loosely in accordance with the SEMMA Framework (Azevedo and Santos, 2008) and will consist of 4 stages: Data Exploration, Data Preparation & Comparative Analysis, Modelling, Evaluation.

# Data Exploration

The data used in this analysis is predominately taken from 4 separate FAOstat datasets (Europe production, Europe trade, Oceania production, Oceania Trade) containing annual data on all crops and animal products produced, imported and exported in Europe and Oceania from 1961-2020, measured in tonnes (Food and Agriculture Organization of the United Nations (FAO), 2022a). These datasets also contain information on livestock numbers, yields, areas harvested and slaughtering numbers. Trade datasets also contain information on Import and Export values in $1000US, (Food and Agriculture Organization of the United Nations (FAO), 2022b). Population data has also been obtained from Fred economic data and world bank, (World Bank, 2022).

Raw datasets were difficult to read as they had lots of columns for the different years of data. This was first reshaped using the melt() function. Some item names also differ between the datasets. These will be identified and replaced to simplify the datasets. For this analysis, 'meat preparations' (mince) will be categorised alongside meat, cattle / meat, pig accordingly.

## Missing values

Missing values and 0 values looked to reflect that there was no production/ harvest/ yield etc for that period/crop. For early years, it is acceptable to assume 0/ NaN values reflect that a country had not yet began to produce/ trade this item and therefore values remain as NaN. As the removal of these values will not impact the integrity of the data, these values can be removed, (Mostafa S.M, 2020).

However, this does not seem to be the case for all 0 / NaN values in the European production dataset as pending further investigation it looks like quite a lot of items are missing the last 3 years of production data (2018 / 2019 / 2020). Oceania dataset is also missing data for 2020. This data might not have been submitted to official sources and therefore not included in datasets yet due to covid 19 disruptions.

Interpolating or replacing this missing 2020 data is not appropriate as the interpolated data will not be able to account for the disruption caused by Covid on the Agriculture sector for both production & trade. Therefore, this data will be dropped form the analysis using a .loc() function.

2018 / 2019 data should be interpolated, but it will need to be specific to only the last 2 years to ensure data integrity. It is difficult to interpolate data in the format it is currently in as the ‘Crop\_Livestock’ and ‘Year’ columns contain multiple values which would make it harder to specify what exactly to interpolate. To rectify this, Europe data will first be tidied into separate Ireland/ Germany datasets to make it easier to identify what specific data to interpolate using the iloc.replace() function as it allows for specified rows/ values to be amended, thus keeping the earlier years 0 values intact to reflect true circumstances. Unlike the European Dataset, the Oceania dataset does not appear to be missing any 2018/2019 values, meaning all Nan/ 0 values can be removed.

The trade data is not missing large amounts of data as a result of covid disruption. Meaning, any missing data identified here either relates to years there was no imports/exports for a given country for a given crop/livestock and can be removed for European & Oceanic Trade datasets.

Finally, ' Value' dtype will only be changed to int only after missing data has been dealt with in individual ‘Missing Values' sections as this cannot be done when NaN values are present.

## Merge Data

For the sake of this analysis, production is specific on meat, dairy and crops. Livestock data will not be included in this analysis and will be filtered out by specifying the ireland\_prod datafame to only contain 'Production' for element. Livestock is measured in 'head' and '1000 head' so will remove any livestock data.

To calculate consumption of a product; production, export and import data on a particular crop is required. Therefore, the analysis should include data on all 3. An inner merge would be suitable in most circumstances, however there are situations where a country may produce and export an item and have no need to import as supplies are high. Likewise, a country might not produce an item, so imports are high and exports are non-existent. An inner merge would only keep rows where the selected values exist in both the left and right data frames, (Raschka and Vahid Mirjalili, 2019).

Therefore, an outer merge will be used when merging import values. This is most suitable here as to not lose any data where a country would import but not produce. A left merge will be used for export values, so data will only be merged if production values exist."Crop\_Livestock" and "Year" will both used during merge to ensure no duplicate values are added. Missing values are checked once more to ensure no data was lost/skewed during the merge.

## Feature Engineering

In order to analyse whether Irelands consumption has changed over time in line with changing diet trends, a consumption column will need to be feature engineered. To calculate consumption of a product; production, export and import data on a particular crop is required and the following equation will be used:

Consumption = (Production + Imports) - Exports

This calculation is as adaptation of the calculation used to calculate GDP, (Krist, 2016), which is:

GDP = Domestic Consumption (C) + Domestic gross investment (In) + Government spending (G) + [Exports (E)—Imports (I )], or GDP = C + In + G + (E—I)

# Data Preparation & Comparative Analysis

Seaborn & MatPlotlib libraries will be used to visualise Irelands agriculture production from 1961-2019 for meat / dairy/ crops. A line graph is chosen as to show production and consumption over time using a rainbow palette to show distinguish between products. As some lines appear close together, labels have also been added to each lie to make it easier for the viewer to follow, (Kirk, 2019). A loop will be used here to avoid repetition, a function will be created to all this loop for different datasets. These visualisations will then be compared against population values to see if values increase/ decrease with population, (Kirk, 2019).

## Meat Production/ Consumption

In general, Ireland meat production has sowed down since the early 2000’s for most meat types and no longer continues to increase with population at the rate it once did. Irish consumption of Meat products, particularly beef has significantly reduced in terms of population growth and has remained steady since the early 2000’s despite population continuing to increase. This highlights the Irish consumers move away from meat products like beef and other red meats.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure Top Ireland Meat Production vs Consumption

Germany in contrast, has a very different meat market to Ireland. Germany’s highest produced and consumed product is pig meat, more widely known as pork Consumption of pork has declined since the mid 2010’s in line with the population decline however it has only continued to decline despite population levels picking up. Beef production and consumption has been in decline since the early 1990’s with chicken overtaking beef in terms of consumption in recent years. Likewise both production and consumption of Turkey is also increasing as this is considered a leaner, healthier meat.

Chart, line chart

Description automatically generated

Chart, line chart

Description automatically generated

Figure Germany Meat Production vs Consumption

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generatedAustralia beef production took a rather large plunge around 2015but is since recovering. For the most part, major meat production ais increasing steadily, although not as much as you would expect given the rate of population growth. Meat consumption is overall in line with production, apart from beef which as more of less remained steady since the early 2000’s wit some fluctuation but no major increase in line with population.

Figure Top Australia Meat Production vs Consumption

In general, there is a decline in meat consumption across Ireland, Germany and Australia.

## Dairy Production/ Consumption

Dairy production/ consumption is more complex and telling in terms of changing diet trends.

Ireland production and consumption remain unchanged in terms of whole and skimmed milk. However, there is a stark contrast between production and consumption when looking at other dairy products. As butter, cheese and dried products production increases in line with population growth, consumption is far lower in terms of numbers, and most have seen a large reduction in consumption in the last 5 years.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure Fig Ireland Dairy Production vs Consumption

Chart, line chart, histogram

Description automatically generatedChart

Description automatically generated

Figure Other Ireland Dairy Production vs Consumption

Similar to Ireland, Germany’s production and consumption remain consistent for larger dairy products and smaller products like skimmed cheese, butter cream and whole dried milks. Skimmed milk has decreased significantly in both production and consumption since the early 1990’s. Other products like whole cheese, skimmed dried milk and evaporated milks have seen a general decline in consumption in comparison to production.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure Top Germany Dairy Production vs Consumption

Chart

Description automatically generatedChart

Description automatically generated

Figure Other Germany Dairy Production vs Consumption

In contrast to Ireland and Germany, Australia’s major dairy products have been in decline since the early 2000’s. In terms of other dairy products, consumption is far lower for every product. This makes sense as Australia is one of the largest vegan and vegetarian populations in the world and was one of the first countries to adapt these practices into their diets naturally.

Chart, line chart

Description automatically generatedChart, line chart

Description automatically generated

Figure Top Australia Dairy Production vs Consumption

Chart

Description automatically generatedChart, histogram

Description automatically generated

Figure Other Australia Dairy Production vs Consumption

It is clear that dairy consumption is declining across Australia, Germany and Ireland.

## Crop Consumption

Crop consumption is one of the most interesting findings across all 3 countries. Crops such as shelled nuts, rolled oats, avocados and quinoa have been increasing in consumption in all 3 countries since the early 2000’s. These crops happen to be the main ingredients in dairy alternative products.

Irelands top dairy alternatives crops include Soy, Oat coconut and Almond, (Plant Based Ireland, 2022). In Germany, the main crops used for the production of plant-based milk include Oats, Soy, Rice, Coconut and Almond, (Future Grocery Shopping, 2022). In 2018, almond accounted for about 68 percent of the plant-based milk alternatives market across Australia. Other dairy alternatives consisted of soy, coconut and rice, (Hughes, 2022).

This will form an interesting basis for machine learning section.

Chart, histogram

Description automatically generated

Figure Ireland Crop Consumption

A picture containing graphical user interface

Description automatically generated

Figure Germany Crop Consumption

Graphical user interface, chart, histogram

Description automatically generated

Figure Germany Crop Consumption

## Box Plots per food Type

Pivot tables are used to show production per meat/ dairy/ crop item which are used to visualise Outliers using colours that relate to each food group. The FAO datasets are free of outliers for the most part, however, outliers will appear when looking at boxplots per product for all years as this data is presented in Pivot table format and will show 0 values for years where production on an item had not yet started, (Kirk, 2019).

Chart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generated

Chart

Description automatically generated

Figure Ireland Outlier detection per food item using Box plots

## Inferential Statistics

To perform appropriate Inferential analysis between countries, a number of tests will need to be performed on the data to determine what kind of statistical techniques can be used, (Frost, 2019).

First, the mean/ max/ min/ standard deviation for each variable will be determined using the .describe() function. Some products appear to have negative values included in their consumption figure. As consumption is calculated as '(production + imports) - exports' , a negative consumption figure should not be possible and can only reflect items produced in one year and exported in another year.



Table

Description automatically generated



Figure Descriptive statistics for Ireland Dairy combined

Using pivot tables, items with a negative consumption value can be identified. The majority of products with a negative consumption number tend to be products that could potentially have longer shelf life such as dried milks, hard cheese, butters, truffles etc. Therefore it is acceptable to assume that any negative consumption vales relate to items that have a longer shelf life that could have been produced/ imported and exported in separate years.

Graphical user interface, application

Description automatically generated with medium confidence



Figure Descriptive statistics for Ireland Dairy by product

## Probability Plots

Using probability plots and Shapiro test to determine normality. If the data is not normal, Nonparametric statistical methods will need to be used as these tests do not assume a normal distribution.

As shown by the probability plots, most data seems to follow a right skewed or positively skewed distribution and is not normal. Meaning, the distribution of values of this data are clustered around the left tail of the distribution while the right tail of the distribution is longer.

Chart, line chart

Description automatically generated

Figure Probability plots for Ireland dataset

## Shapiro Tests

There is already a high level of confidence that the data is not normal following the probability plot, but to be absolutely certain, a Shapiro-Wilk Test can be used. Shapiro Test is a hypothesis test that defines:

H0 = data comes from normal distribution

H1 = data does not come from a normal distribution

If p value is less than alpha (0.05), the null hypothesis is rejected, meaning data is not normal and non-parametric statistical tests will need to be used going forward. As proven by the analysis, the data is not normal and the null hypotheses is rejected, (Frost, 2020).

## Analysis between countries

In order to compare like for like between countries, per capita datasets must be prepared. This will be done by dividing each of the relevant columns by the country population to get per capita amounts for production/ imports/ exports and consumption for analysis.

Non-Parametric tests are required as the data in not normal. To compare Ireland against other countries a test that determines whether the values of 2 samples are equal or not is required, (Williams, 2022).

The Mann Whitney test is a rank-based test that can be used to compare values for two groups. If we get a significant result it suggests that the values for the two groups are different.

Focusing on the impact of the Common agricultural policy, and possible diet trends the following null hypothesis's will be analysed using inferential statistics, (Frost, 2020).

On average:

Text

Description automatically generated

## Results of Mann-Whitney

The following observations were obtained:

Text

Description automatically generated

This is an interesting finding that Ireland and Germany show no similarities in meat production/ import/ export/ consumption. However, Germany's meat habits were very different when analysed so it is not unfitting to assume that this is due to cultural differences between the countries.

Looking at Ireland/ Australia, despite production values being similar. From the data it is evident that on average Ireland imports and exports a lot more meat products that Australia, leading to the assumption that Australia relies less heavily on trade and is more self-sufficient, which is known to be the case.

Text

Description automatically generated

These results make sense as there is a general decline in the consumption of dairy in countries with a growing vegan population as all 3 countries have. It is also evident form the data that Ireland produces/ imports/ exports the highest amount of dairy between these 3 countries so it is not surprising that their averages vary. Ireland produces almost than 3 times the amount of dairy per capita as the other 2 countries

Chart

Description automatically generated with low confidence

This is also unsurprising as Ireland and Germany appear to produce, export, and consume similar amounts of crops per capita. These similarities could be due to diet trends such as veganism or do to benefits received from the CAP & being members of the EU.

Australia both produce and consume almost 3 times more crops than Ireland or Germany. This highlights the difference in diet cultures ad shows how much more veg is consumed because of Australia’s high vegan and vegetarian populations.

## Data correlations

Datasets for machine learning were refined down to only include consumption and production values for larger dairy products and crops related to milk alternatives such as shelled nuts and rolled oats. Milk production, imports, exports, and country population will also be included in these datasets.

Data is visualized using a correlation matrix to determine if features are suitable for machine learning (Kirk, 2020). As people's minds can only interpret so much, it may be helpful to only show the bottom half of our visualization. Similarly, it can make sense to remove the diagonal line of 1s, since this has no real value, (Datagy, 2021).

Parameters such as vmin, vmax are used to anchor the colormap between +1 and -1 and centred at 0 for easier visualisation and understanding of correlations. The cmap parameter is used to distinguish between high/ low correlations using 2 contrasting colours for easy identification.

Chart

Description automatically generated

Figure Ireland dataset for ML correlation between features

There is a moderate/ high corotation between milk consumption and most features bar oat, hazelnut, and butter consumption for Ireland.

Chart

Description automatically generated

Figure Germany dataset for ML correlation between features

There is a moderate/ high corotation between milk consumption and most features bar hazelnut and butter consumption being the lowest but still slight for Germany.

Chart

Description automatically generated

Figure Australia dataset for ML correlation between features

There is a moderate/ high corotation between milk consumption and most features bar soy and butter consumption being the lowest but still slight for Australia.

Overall, there is a high correlation between most features. If the machine learning models perform poorly, the more poorly correlated features can be removed to see if this improves the result.

# Modelling

## Sentiment Analysis

Using sentiment analysis, vegan related tweets will be analysed to see what the sentiment towards veganism is today in order to determine if this could have an impact on the production and consumption of meat/ dairy in future. Due to wordcount restriction, a more detailed account of sentiment analysis approach can be found within the python notebook.

## Sentiment Results

Sentiment for all text can be summarised by adding sentiments together for all tweets and determine whether totals are negative, neutral or negative. In this case, sentiment is positive overall. The average sentiment can also be looked at and here it is on average, positive. For subjectivity as this has no positive or negative scale for vegan related produce, the average is most appropriate to use. On average, tweet subjectivity is closer to objectivity as opposed to public opinion or emotive restated text which highlights that the growing sentiment towards veganism is practical, matter of fact and typically positive.

## Prediction Model Selection

The analysis looks to predict the future milk consumption based on the production/ import/ export data of milk and consumption of other dairy/ dairy alternative produce. As this analysis requires a prediction, regression models would be best suited. Historic data is also present with input/ output variables meaning a supervised learning approach will be taken using a training and testing split. The data also contains very few outliers as original datasets had removed these preciously.

Possible regression models include:

* Linear Regression
* Lasso Regression
* Ridge Regression
* Decision Tree Regression
* Random Forest Regression
* Support Vector Regression
* Gradient Boosting Tree

For this analysis milk consumption will be the dependant variable. Using the pair plot, it was determined that very few features had a strong linear relationship so Linear models will not be used.

Support Vector Regression in general does not work well with large datasets and can take some time to train data. In this instance the SV models took over an hour to run before timing out due to lack of processing power with limited parameters and as such were removed from the analysis.

Decision Tree Regression is easy to interpret and scaling is not required. It also works with linear and non-linear data. However, it can be prone to overfitting.

Random Forest Regression uses multiple decision tree algorithms to improve accuracy. It does not require feature scaling required, it is very stable and handles outliers on its own. It does require slightly longer training period in comparison to other models.

Based on the above, Decision Tree Regression and Random Forrest Regression models will be used.

## Standardise/ Normalise data

The Ireland, Germany, Australia datasets prepared for machine learning have no categorical variables and all units are measured in tonnes, but the features ranges do differ. Therefore, in this circumstance, normalisation would be most appropriate. However, not all machine learning models require feature scaling and it can form a good comparative as to whether the model performs better with or without feature scaling, (Towards AI, 2020). Decision Tree Regression and Random Forest Regression do not require sailing. As models performed quite well without, no sailing will be use in this instance.

## Model Parameters

For each model, parameter selection was performed using GridSearchCV until selected started producing good results, (Beheshti, 2022). The same parameters were used for each of the final 3 country datasets. More details on model parameters can be found in the accompanying python notebook.

Following machine learning, the below results were obtained for each model using GridSearchCV to select the best model parameters. Both modes performed extremely well in both training and testing meaning no sailing, feature refinement should be necessary for better results.

Table

Description automatically generated

Figure Modelling Results

Using explainer dashboard, a dashboard was created to display more detailed finding in accordance with this analysis, (Explainer Dashboard Documentation, 2021). It details feature contribution to the modelling outcomes and clearly visualises the performance of each model.

As shown, feature importance varies between country and model. Rando Forrest models seem to show more variance in feature spread whereas Decision Tree models seems to be overly influenced by 1 or 2 features.

Graphical user interface

Description automatically generated with low confidence

Figure Feature importance for Decision Tree – Ireland

A picture containing application

Description automatically generated

Figure Feature importance for Random Forrest – Ireland

Graphical user interface, application, Word

Description automatically generated

Figure Feature importance for Decision Tree – Germany

Graphical user interface, application

Description automatically generated

Figure Feature importance for Random Forrest – Germany

Text, application

Description automatically generated

Figure Feature importance for Decision Tree – Australia

Application

Description automatically generated with low confidence

Figure Feature importance for Random Forrest – Australia

This dashboard also shows model performance. More details on performance of each model can be found in the dashboard.

Chart, line chart

Description automatically generated

Figure Explainer Dashboard Modelling Results Example

It also allows for individual predictions by filling in numbers for each feature allowing farmers to predict milk consumption based on other dairy/ non dairy consumptions rising or falling.

A screenshot of a computer

Description automatically generated

Figure Explainer Dashboard Future Predictions Example

# Conclusion & Limitations

This analysis focused on how production and consumption has changed over time in line with consumer diet trends and thus impacted Ireland’s agriculture in comparison to other counties where veganism is also on the rise with Ireland as the baseline. Irish findings were then compared against Germany and Australia. It is clear form the analysis than on average meat and dairy consumption is decreasing and there is moderate justification that changing diet trends and the over towards veganism has a part to play in this as production and consumption of plant-based alternatives has risen significantly since the early 2000’s and continues to rise which would ideally be addressed in a later revision. This can and will have an increasing influence on agriculture and will impact how farmers deal with the general decline in meat/ dairy consumption globally. In accordance with these findings, farmers should look for more vegan friendly farming solutions and produce to stay current and competitive.

However, there were a number of limitations to this analysis.

* Ideally, data would have been available at a monthly frequency to allow for more detailed analysis and a larger training/ testing set for modelling.
* The availability of data on the growing vegan population is also scarce as this trend has only come to fruition in recent years, as the number of vegans worldwide is only available from 2018/2019 onwards.
* According to google trends, (Google Trends, 2022) veganism hit its peak in 2020. Unfortunately, due to covid interruptions, this analysis had to remove 2020 data. Ideally in future iterations, analysis would include 2020, 2021, 2022 at a more granular detail.
* Due to time constraints this analysis focused on a small number of the larger milk alternatives. Given more time, a larger number of milk and meat alternatives could be included for more substantial findings.
* Sentiment analysis was limited the last 7 days due to twitter licensing restrictions. Twitter would have been the ideal source for sentiment as the growing vegan population are very present on social media. It would have been ideal to include sentiment data for the past 10 years to see if there was an upwards/ downwards trend in sentiment. This should be included in future iterations.

# Bibliography

Azevedo, A. and Santos, M.F. (2008). *KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*. [online] www.iadisportal.org. Available at: http://www.iadisportal.org/digital-library/kdd-semma-and-crisp-dm-a-parallel-overview [Accessed 22 May 2022].

Beheshti, N. (2022). *Cross Validation and Grid Search*. [online] Medium. Available at: https://towardsdatascience.com/cross-validation-and-grid-search-efa64b127c1b.

Datagy (2021). *Calculate and Plot a Correlation Matrix in Python and Pandas • datagy*. [online] datagy. Available at: https://datagy.io/python-correlation-matrix/.

Explainer Dashboard Documentation (2021). *ExplainerHub — explainerdashboard 0.2 documentation*. [online] explainerdashboard.readthedocs.io. Available at: https://explainerdashboard.readthedocs.io/en/latest/hub.html [Accessed 22 May 2022].

Food and Agriculture Organization of the United Nations (FAO) (2022a). *Crops and livestock products Datssets*. [online] www.fao.org. Available at: https://www.fao.org/faostat/en/#data/TCL [Accessed 26 Apr. 2022].

Food and Agriculture Organization of the United Nations (FAO) (2022b). *FAOSTAT*. [online] www.fao.org. Available at: https://www.fao.org/faostat/en/#data/TCL [Accessed 26 Apr. 2022].

Frost, J. (2019). *Introduction to statistics : an Intuitive guide for analyzing data and unlocking discoveries*. State College, Pa: Jim Publishing.

Frost, J. (2020). *Hypothesis testing : an intuitive guide for making data driven decisions*. State College, Pennsylvania: Statistics By Jim Publishing.

Future Grocery Shopping (2022). *The German Plant-based Milk Market*. [online] Future Grocery Shopping. Available at: https://www.futuregroceryshopping.com/blog/the-german-plant-based-milk-market#:~:text=Among%20the%20most%20available%20brands [Accessed 22 Dec. 2022].

Google Trends (2022). *Google Trends*. [online] Google Trends. Available at: https://trends.google.com/trends/explore?date=all&geo=FR&q=vegan [Accessed 22 May 2022].

Hughes, C. (2022). *Australia: non-dairy milk alternatives market share*. [online] Statista. Available at: https://www.statista.com/statistics/1231058/australia-non-dairy-milk-alternatives-market-share/#:~:text=In%202018%2C%20almond%20accounted%20for [Accessed 10 Apr. 2022].

Kirk, A. (2019). *Data Visualisation : A Handbook for Data Driven Design*. Los Angeles: Sage Publications.

Krist, W. (2016). *Chapter 3: Trade Agreements and Economic Theory*. [online] Wilson Center. Available at: https://www.wilsoncenter.org/chapter-3-trade-agreements-and-economic-theory.

Mostafa, S.M. (2020). Missing data imputation by the aid of features similarities. *International Journal of Big Data Management*, 1(1), p.81. doi:10.1504/ijbdm.2020.106883.

Plant Based Ireland (2022). *Dairy Archives*. [online] Plant Based Ireland. Available at: https://plantbased.ie/news\_category/dairy/ [Accessed 17 May 2022].

Raschka, S. and Vahid Mirjalili (2019). *Python machine learning : machine learning and deep learning with Python, scikit-learn, and TensorFlow*. Birmingham (Uk): Packt Publishing.

Richardson, N.J., MacFie, H.J.H. and Shepherd, R. (1994). Consumer attitudes to meat eating. *Meat Science*, 36(1-2), pp.57–65. doi:10.1016/0309-1740(94)90033-7.

Sarah Harford (2018). *The Irish food board wants to know how to ‘win back’ vegetarians and vegans*. [online] TheJournal.ie. Available at: https://www.thejournal.ie/ireland-vegetarians-vegans-2-4058390-Jun2018/ [Accessed 17 Nov. 2019].

The Vegan Society (2021). *Worldwide growth of veganism*. [online] The Vegan Society. Available at: https://www.vegansociety.com/news/media/statistics/worldwide.

Towards AI, T.A. (2020). *How, When, and Why Should You Normalize / Standardize / Rescale Your Data? – Towards AI — The Best of Tech, Science, and Engineering*. [online] Towards AI. Available at: https://towardsai.net/p/data-science/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff.

Williams, L.H. (2022). *Choosing a statistical test: A cheat sheet*. [online] Lee Hulbert-Williams. Available at: https://leehw.com/crib-sheets/statistical-test-cheat-sheet/.

World Bank (2022). *Population, Total for Ireland*. [online] FRED, Federal Reserve Bank of St. Louis. Available at: https://fred.stlouisfed.org/series/POPTOTIEA647NWDB [Accessed 22 May 2022].