Megan McHugh

MSc Data Analytics - CA2

27th Jan 2023

Word Count: 2811

**Introduction**

This analysis explores the Producer Price Index (PPI) of Wheat in Ireland vs Croatia. Croatia was chosen due to its relatively similar size to Ireland in terms of land size and population. This analysis further explores the relationship between the price of Wheat and whether a country is in the EU or not, and finally the relationship between the price of Wheat and GDP for a selection of countries over time.

**A Note on Programming, Data Processing and Visualization**

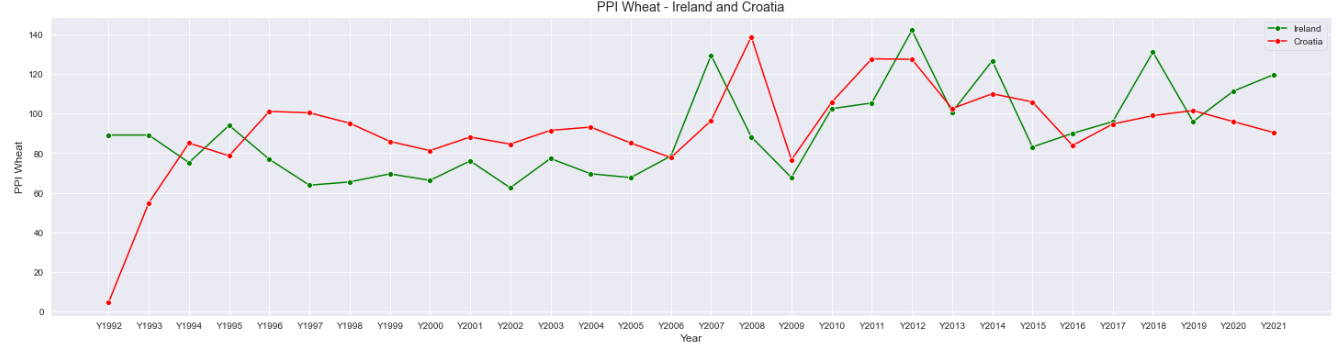
Processing, visualization, statistics and machine learning techniques were carried out in the two Python notebooks attached. The majority of work is found in CA2\_MAIN\_NOTEBOOK, while some secondary processing is in the notebook DATAPROCESSING\_2. I explored a lot of data and used several methods that did not make it into the report due to the word limit. Please see the script for more information. All methodology headings below have a corresponding section in the python notebooks.

**Data Processing**

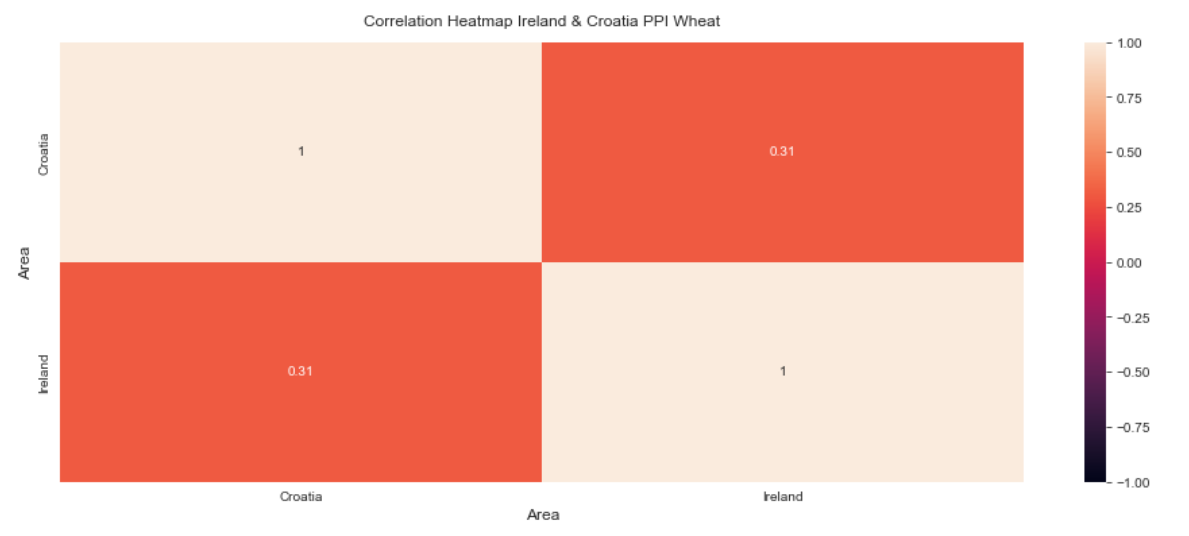
It’s important to have a clean dataset to enable the correct use of data analysis. When the datasets are clean and complete, it makes analysis more straightforward. Data is from FAO.org, which contains a huge amount of data ranging across countries, years, variables and metrics.

Firstly, I cut the massive datasets into the data I require, I did this to make the data more straightforward to explore, understand and cut distractions. In this analysis, we will look at the PPI of wheat for Ireland and Croatia. To make the data easier to work with I created a data frame with just these variables for these countries. I created another data frame with the PPI of Wheat over time by country. I added a column with dummy variables by joining another table to say if the country was in the EU or not (1 or 0). I then proceeded to add the GDP data to the countries and years available for the PPI of Wheat dataset. Unfortunately, the country names were not exactly matching (e.g. Ireland vs IRE) so I had to create a table to join the new name format with the existing names.

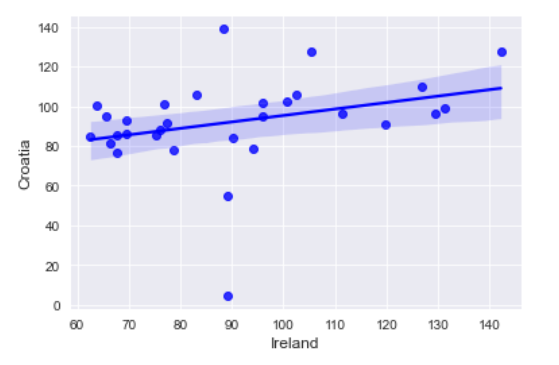
Below is a line graph to illustrate the PPI of Wheat in both countries over time. I did this because I wanted to see if there is any obvious relationship between the PPI of Ireland and Croatia. For this and all subsequent visualisations, I chose green to represent Ireland and red for Croatia where applicable as these are colours on the respective flags, and they contrast nicely to make the graphs each to visualise and understand. For every illustration, I chose the best plot to visualise the data according to the circumstances, and used contrasting colours that are easy to differentiate, while not too abrasive on the eyes. I labelled all axis and titles for the reader’s convenience. Tables are kept in a simple and clean format to avoid distracting the reader or cluttering the page. (forbes.com, 2022) A dashboard of further visualizations is available in the Python Script CA2\_MAIN\_NOTEBOOK attached under the heading ‘Dashboard’. This project is fully documented on GitHub here: [sba22218/MSc\_GitHub\_CA2](https://github.com/sba22218/MSc_GitHub_CA2).



The correlation heatmap below shows us there is a 31% correlation between PPI in Ireland and Croatia, which implies a very weak relationship.



Furthermore, the scatter plot below illustrates there is not a very strong relationship between PPI of Wheat between Ireland and Croatia, the data points do not follow a very obvious pattern, and the trendline is relatively flat.

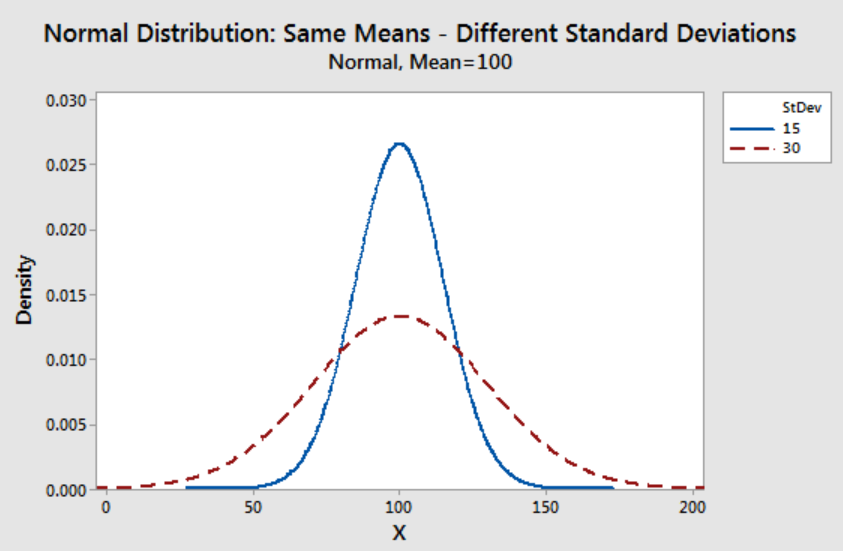


**Statistical Analysis**

1. **Skewness Test for Normality:**

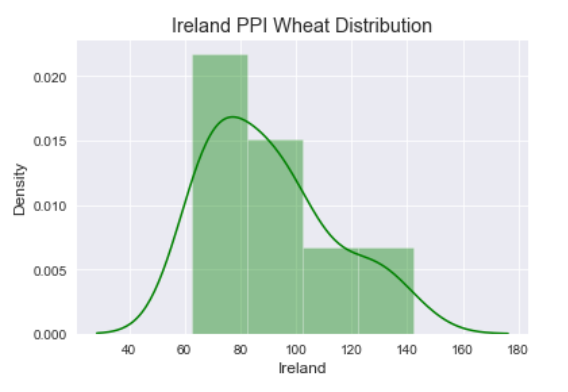
Test of the normality is an important step for deciding the measures of central tendency and statistical methods for data analysis. The normal distribution, also known as the Gaussian distribution, is an important probability distribution in statistics for independent, random variables.

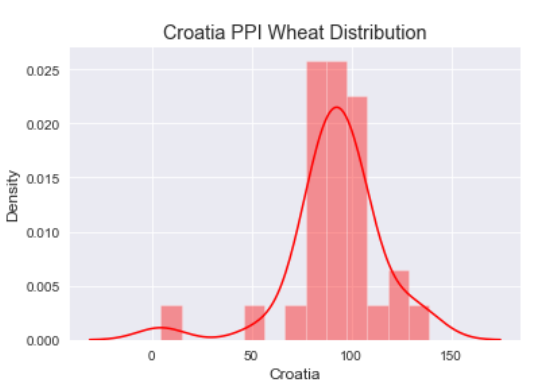
The normal distribution is a continuous probability distribution that is symmetrical around its mean, most of the observations are around the central peak, and the probabilities for values further away from the mean diminish equally in both directions, as illustrated in the figure below. Extreme values in both tails of the distribution are similarly unlikely. (Rohith Gandhi, 2018)



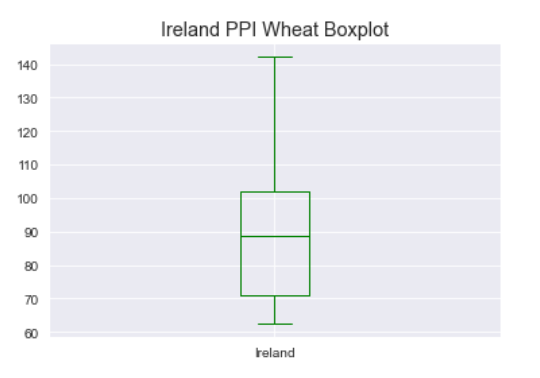
Some statistical tests assume that the data is normally distributed, so it’s important to know if the data we’re using is normally distributed or not.

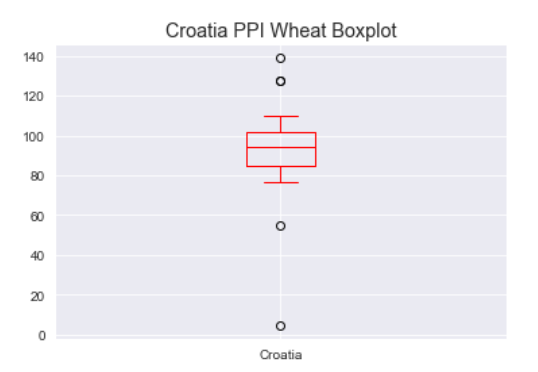
The histograms below illustrate there is a slight skew to the left in the PPI of Wheat in Ireland. Croatia distribution looks more symmetrical, albeit with several outliers.





The boxplot illustrations below show that Croatia has several outliers.

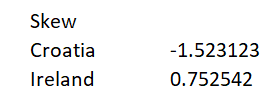




In testing for skewness, the hypothesis are as follows:

* Null hypothesis (H0): The Data the comes from a normal distribution
* Alternative hypothesis (Ha or H1): The Data does not come from a normal distribution

A test for skewness in the Python notebook produces the following results:

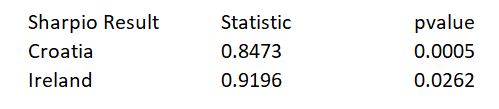


For this data to be normal, the results must be between + / - 0.5. We can therefore reject the null hypothesis, and conclude the distribution of PPI of Wheat in Ireland and Croatia is non-normal.

1. **Sharpio-Wilk Test**

Sharpio-Wilk Test is another test for normality.

* Null hypothesis (H0): The Data the comes from a normal distribution
* Alternative hypothesis (Ha or H1): The Data does not come from a normal distribution



Since the p-values are less than .05, we reject the null hypothesis of the Shapiro-Wilk test. This means there is sufficient evidence to say that the sample data does not come from a normal distribution.

1. **Mann Whitney U Test - For non-normal data:**

The Mann-Whitney U test is used when the data is ordinal or when the assumptions of the t-test are not met, for example, if the data is non-normal like in this case.

* The null hypothesis (H0) is that the two populations (Ireland and Croatia PPI of Wheat) are equal.
* The alternative hypothesis (H1) is that the two populations (Ireland and Croatia PPI of Wheat) are not equal.

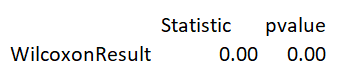
The results of the test show that the pvalue is less than the alpha value of 0.05, therefore we reject Null Hypothesis, and say there is a significant difference between the two samples.



1. **Wilcoxon Test:**

Wilcoxon signed-rank test, also known as Wilcoxon matched pair test is a non-parametric hypothesis test that compares the median of two paired groups and tells if they are identically distributed or not.

* The null hypothesis (H0) is that the median of the population of differences between the paired data is zero.
* The alternative hypothesis (H1) is that it is not.

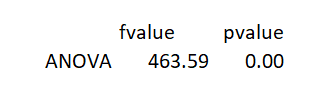


Since the p-value is less than 0.05, we reject the null hypothesis. We have sufficient evidence to say that the true mean is not equal between the two countries.

1. **ANOVA or Analysis of Variance Test**

The one-way ANOVA tests the null hypothesis that two or more groups have the same population mean.

* The null hypothesis (H0) is that there is no difference between the groups and equality between means
* The alternative hypothesis (H1) is that there is a difference between the means and groups



The pvalue is less than 0.05 so we can reject the null hypothesis, and accept the alternative hypothesis that there is a difference between the means of the countries.

**Machine Learning**

**1. Decision Tree**

Decision Trees are a common form of analysis used to classify two-dimensional data. A decision asks a series of True/False questions to eventually land on a categorical result. Decision Trees can be used to solve both classification and regression problems. Decision trees apply a top-down approach to the dataset that is fed during training. (Rohith Gandhi, 2018)

For this part of the analysis, the hypothesis we can predict if a country is a member of the EU based on its wheat PPI. The null hypothesis, on the other hand, states that there is no change, no difference, no effect, and otherwise no relationship between the independent variable (EU Membership) and dependent variable (PPI of Wheat).

Every decision tree includes the following:

* A root node: This is the note that starts the graph. If it is a normal decision tree, it’s the variable that best splits the data.
* Branches: Each branch tree on a decision presents a decision option, its cost and the probability that it is likely to occur.
* Leaf Nodes: The leaves at the end of the branches show the possible payoffs or outcomes.

We want to establish the sequence of if/else questions that gets us

to the true answer most quickly. In the machine learning settings, these questions are

called tests*.* (Rohith Gandhi, 2018)

Advantages of Decision Trees:

* Decision trees are relatively easy to understand, interpret and visualize.
* Decision trees require relatively little data processing.
* Decision trees are about to handle both numerical and categorical data.
* Are able to handle multi-output problems.
* Uses a white box model. If a given situation is observable in a model, the explanation for the condition is easily explained by Boolean logic.
* It’s possible to validate a decision tree model using statistical tests.

Understanding the Gini Index and Information Gain in Decision Trees:

Entropy is a measurement of randomness or uncertainty. Usually classified between 0 and 1, the higher the figure, the higher the level of disorder.

Information Gain is applied to quantify which variable provides maximal information about the classification. For example, what’s the best way to divide the data to get the most amount of information?

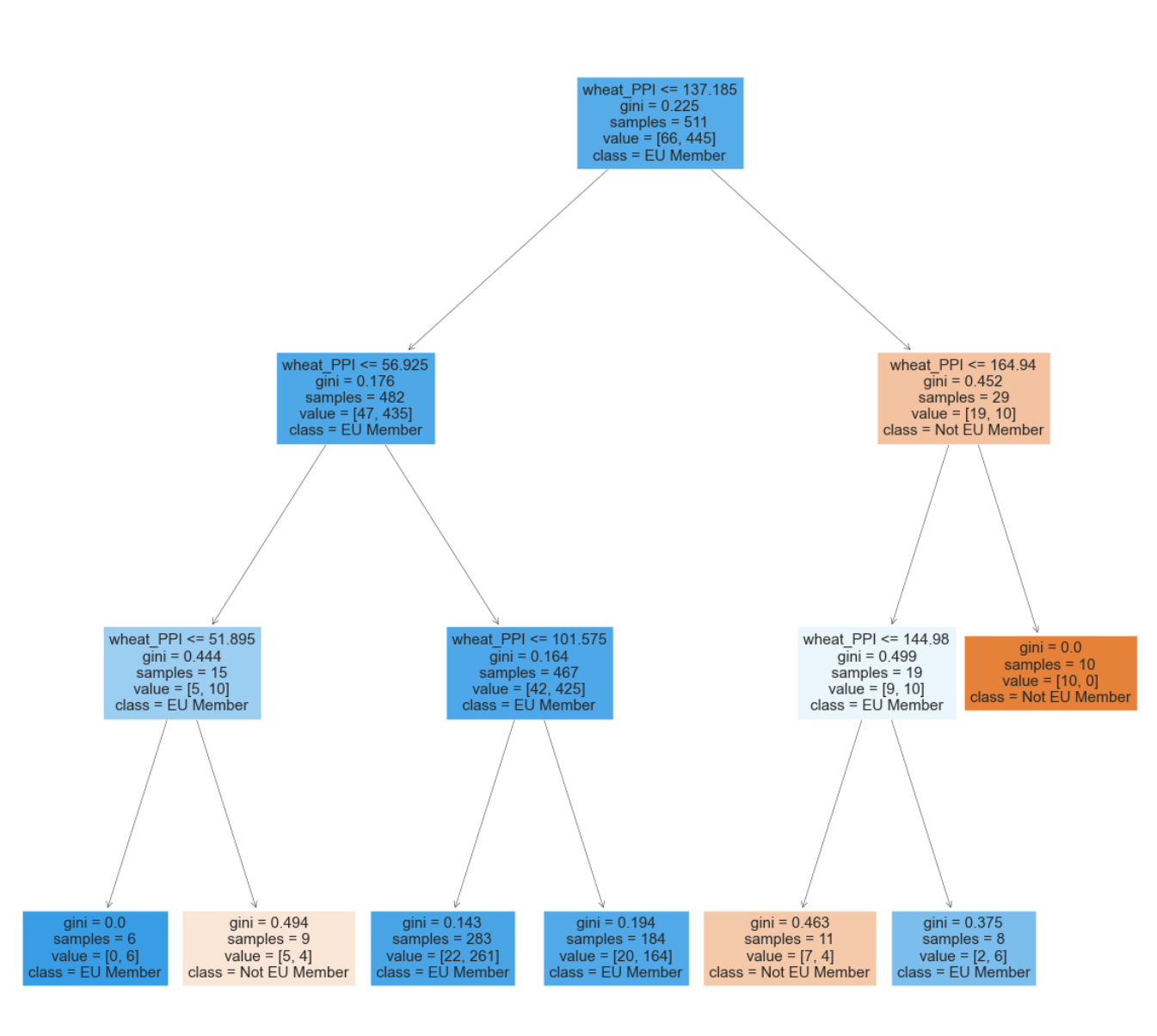
Gini Index (or Gini impurity) calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. The Gini Index also ranges from 0-1. A Gini Index of 0 means all the elements belong to a specified class A Gini distribution of 0.5 indicates an equal distribution of elements across classes. 1 indicates the random distribution of elements across various classes. When designing the decision tree, the features possessing the lowest value of the Gini Index are preferable.  (Rohith Gandhi, 2018)

**The Model:**

For this decision tree, I used the column ‘EU\_member’ at the dependent Y variable which we are trying to establish. The Y variable is the PPI of Wheat. 70% of the data was used to train the model, and the remaining 30% was used to test the model. It’s important to give enough data to train the model sufficiently, usually 70-80%. (Severance, 2016)

**The Results:**

The results of the decision tree are illustrated below.

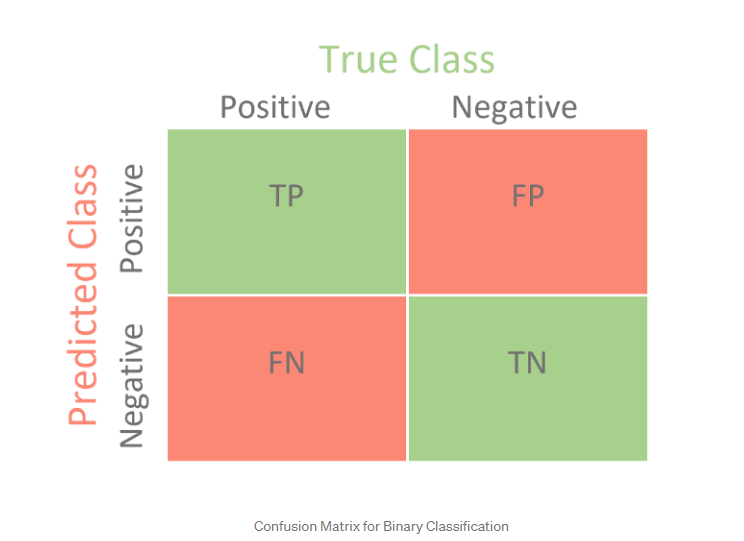


**Confusion Matrix:** Each entry in a confusion matrix denotes the number of predictions made by the model where it classified the classes correctly or incorrectly.

* Positive: A positive result is when the instance is classified as a member of the class that is trying to be identified. In this case, it’s EU member,
* Negative: A negative result is when the instance is classified as not a member of that class that’s trying to be identified, in this case that is Not and EU member.

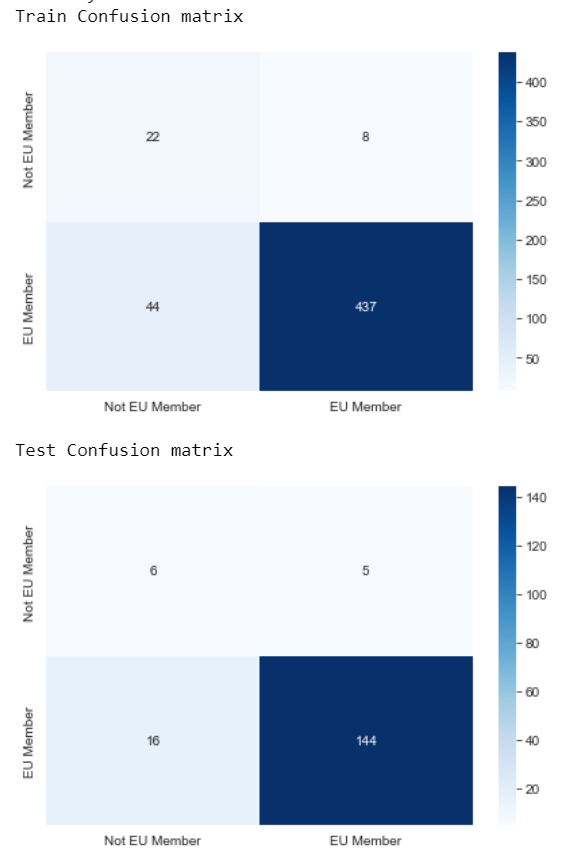
It is possible to get a ‘false’ meaning the model made the wrong prediction. This can happen for both positive and negative classifications, leaving 4 options.

1. True Positive (TP)
2. False Positive (FP)
3. True Negative (TN)
4. False Positive (FP)

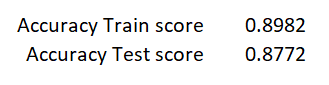


The confusion matrix in this case is as seen in the Confusion Matrix illustration

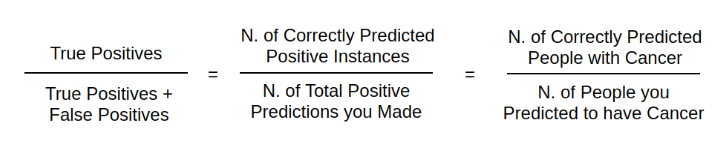
below.



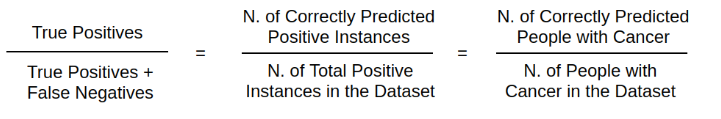
The accuracy scores are as shown below. Any score over 70 is considered good, so we are above the threshold for an acceptable decision tree. (Müller, A.C. and Guido, S. 2017)



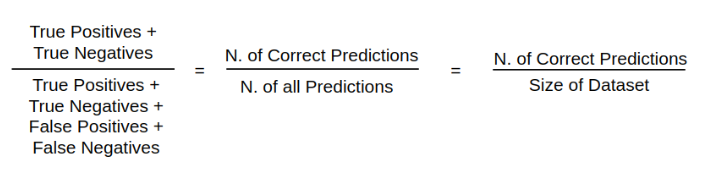
**Precision:** is a measure of how many of the positive predictions made are correct (true positives).



**Recall/Sensitivity:** Is the measure of how many of the positive cases the classifier correctly predicted, over all the positive cases in the data.



**Accuracy:** A key metric used for model evaluation is accuracy, which describe the amount of correct predictions over all predictions and uses the following formula. (Severance, 2016)



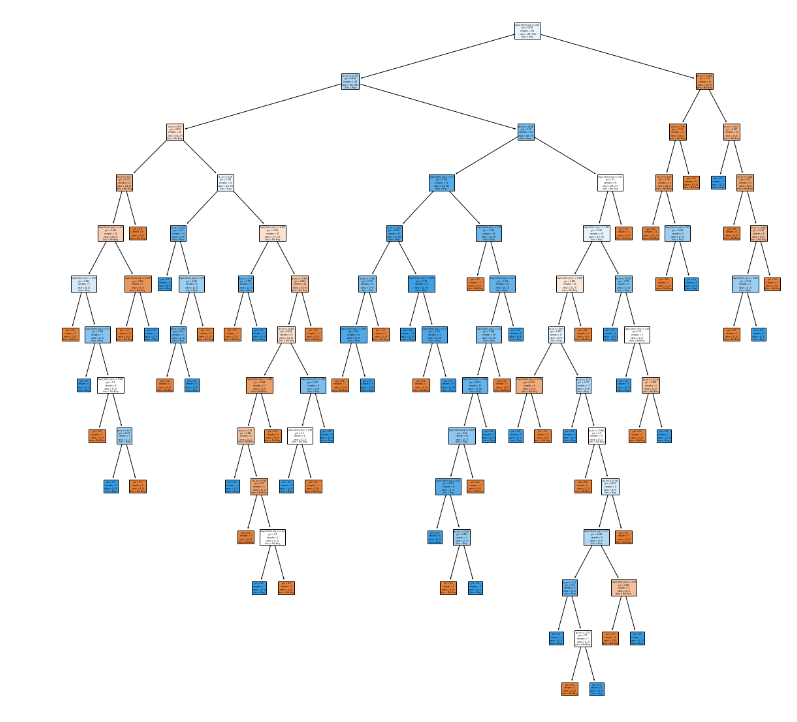
The accuracy of this model is 56%, not much better than you might get flipping a coin to see if Capel Street is busy! In other words, it’s not very accurate.

**Limitations of Decision Trees:**

Though the Decision Tree classifier is one of the most sophisticated classification algorithms, it may have certain limitations, especially in real-world scenarios. (Rohith Gandhi, 2018) Some of its deterrents are as mentioned below:

* Decision Tree Classifiers often tend to overfit the training data.
* Changes in data may lead to unnecessary changes in the result.
* Large trees can sometimes be very difficult to interpret.
* These are biased toward splits on features having a number of levels.

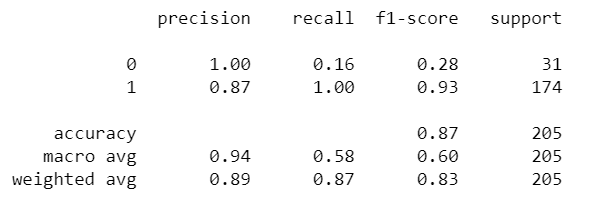
The Decision Tree in this analysis is certainly complicated with many branches which makes it difficult to interoperate as seen in the graph below.



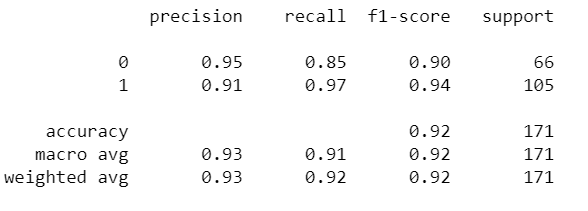
**2. GridSearchCV**

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model. The performance of a model significantly depends on the value of hyperparameters, however, there is no way to know in advance the best values for hyperparameters. Ideally, we would try all possible values to know the optimal values. Doing this manually could take a considerable amount of time and resources. GridSearchCV is a tool used to automate the tuning of hyperparameters.

**Results Without GridSearchCV:**



**Results Without GridSearchCV:**



Using GridSearchCV on our model increases our accuracy score from 0.87, to 0.90. In addition to precision, recall, and accuracy which were explained earlier, we also have F1 score. F1 score is a weighted average of precision and recall. F1 score is usually more useful than accuracy, especially if there is an uneven class distribution, which is true in this case. In conclusion, are all well above the threshold for ‘good’ (Severance, 2016) but the model metrics were improved with the use of GridSearchCV.

**3. K-Means Clustering**

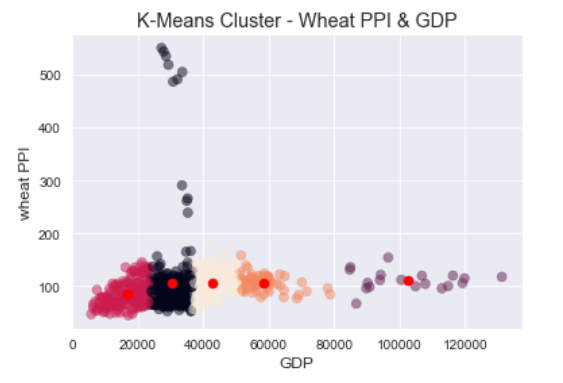
K-means clustering is an unsupervised machine learning algorithm that divides a set of observations into clusters (k). Clustering is a method of assigning comparable data points to groups using data patterns. Clustering algorithms find similar data points and allocate them to the same set.

“K means” refers to:

* The number of clusters you specify (K).
* The process of assigning observations to the cluster with the nearest centre (mean).

K means clustering generates groups comprised of observations with similar characteristics by forming groups that minimize the variances between the data points and the cluster’s centroid.

This cluster analysis explores Wheat PPI and GDP by country from 2000-2021. Below is an illustration of the 5 centroids in red, and the clusters of data points grouped in different colours.



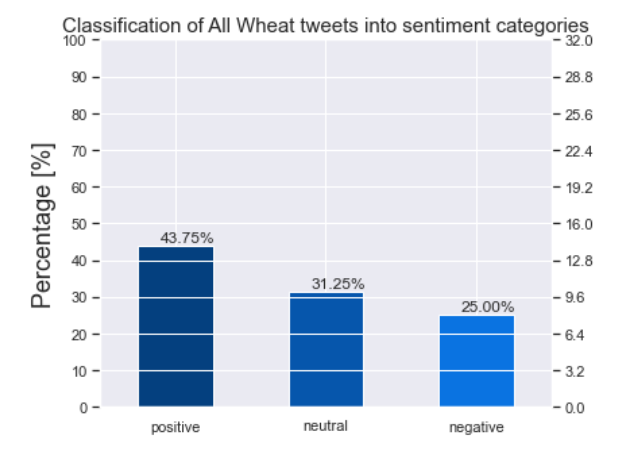
One of the limitations of K-Means Clustering is it does not attempt to predict existing or known group labels (therefore not a supervised learning method), so while the data was clustered successfully, more investigation is required to discover the defining characterizes of each cluster. **Parallel Coordinates Plot**s show individual data points sit across all variables*.* By looking at how the values for each variable compare across clusters, we can get a sense of what each cluster represents. (Müller, A.C. and Guido, S. 2017)

**Sentiment Analysis**

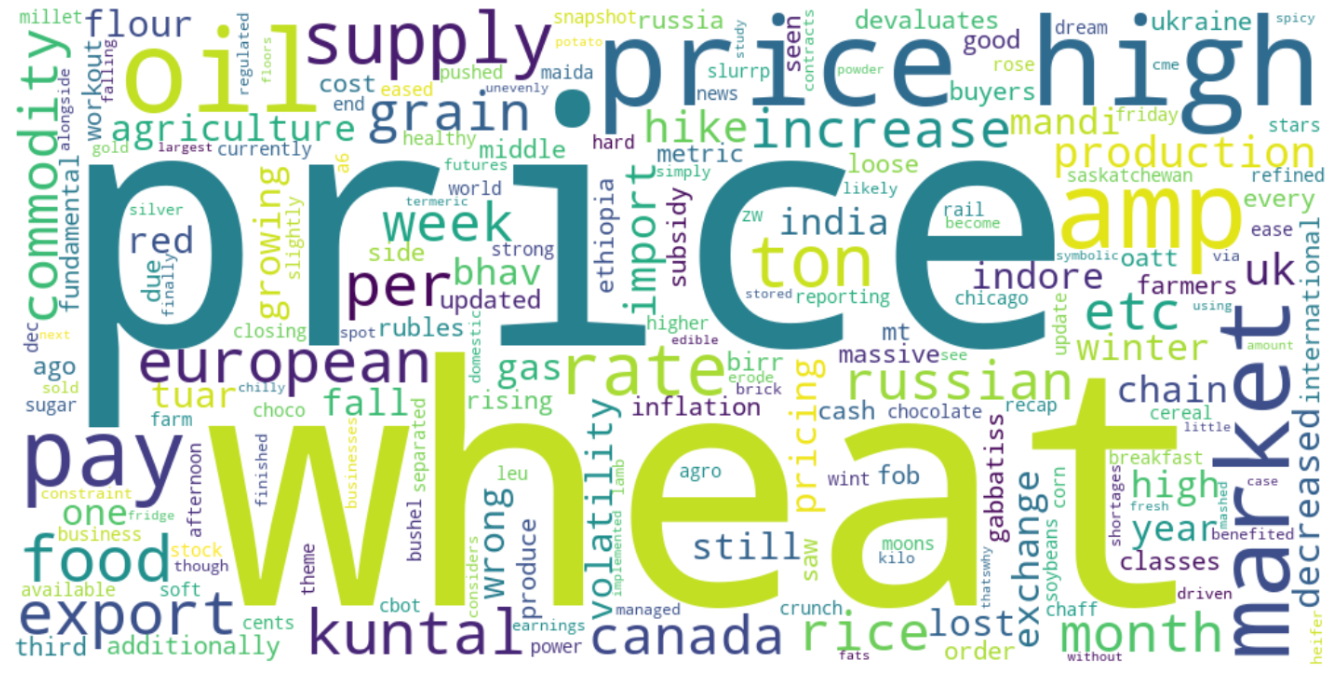
Sentiment Analysis is a case of Natural Language Processing (NPL) which falls under the category text classification. Text is classified into various sentiments, like positive, negative or neutral. The goal is to decipher the underlying mood or sentiment of a text. It’s also known as opinion mining.

Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter’s API. Twint was used in this analysis as  it does not require any authentication developer credentials to connect to twitter which can be timely to acquire.

The results of the sentiment analysis are presented in the bar chart below.



The majority of the sentiment is positive, at 43.75%, while only 25% is negative. The remaining 31.25% is neutral. In order to explore the words associated with ‘wheat’ and ‘price’, the Word Cloud below was created. The number of times a word is mentioned is represented by its size, the larger the word the more times it was mentioned. The Word Cloud below shows *Wheat* and *Price* as the two largest words, and has a variety of economic words- *high price, rate, pay, supply, export, commodity*.



I reran the analysis excluding the word *price* in the search criteriaas this may be causing a bias in the results. The results are in the Word Cloud below. There are a lot fewer economic terms found. In this instance, the words associated with Wheat are *friends, brewing, beer, peach, berries,* and *style*. This suggests that when people are talking about Wheat on Twitter, they’re largely not talking about price.



**Conclusion:**

In conclusion, we found that there are statistical differences between the PPI of Wheat in Ireland vs Europe. We learned that the PPI of Wheat is a relatively good way to classify whether or not a country is in the EU or not, using Decision Tree and then GridSearchCV to improve the results further. We found several clusters evident when exploring Wheat PPI GDP across countries and time, but further analysis is required in order to discover what defines each cluster. When people Tweet about Wheat, they largely do not do so to speak of the price of wheat, but it’s more associated with beer brewing.

**Reference list**

ManeroBastin (2018). *7 Classical Assumptions of Ordinary Least Squares (OLS) Linear Regression - DataScienceCentral.com*. [online] Data Science Central. Available at: <https://www.datasciencecentral.com/7-classical-assumptions-of-ordinary-least-squares-ols-linear/>.

Chun-Houh Chen and Al, E. (2016). *Handbook of data visualization*. Berlin: Springer.

GeeksforGeeks. (2020). *What is Data ?* [online] Available at: <https://www.geeksforgeeks.org/what-is-data/>.

https://www.facebook.com/jason.brownlee.39 (2018). *How To Implement The Decision Tree Algorithm From Scratch In Python*. [online] Machine Learning Mastery. Available at: <https://machinelearningmastery.com/implement-decision-tree-algorithm-scratch-python/>.

Müller, A.C. and Guido, S. (2017). *Introduction to machine learning with Python : a guide for data scientists*. Beijing: O’reilly.

Python, R. (2020). *Python Machine Learning – Real Python*. [online] realpython.com. Available at: <https://realpython.com/tutorials/machine-learning/> [Accessed 11 Nov. 2022].

Rohith Gandhi (2018). *Introduction to Machine Learning Algorithms: Linear Regression*. [online] Medium. Available at: <https://towardsdatascience.com/introduction-to-machine-learning-algorithms-linear-regression-14c4e325882a>.

Severance, C.R. (2016). *Python for everybody : exploring data using Python 3*. Ann Arbor, Mi: Charles Severance.

Stephen Allwright (2022) *What is a good accuracy score? simply explained*, *Stephen Allwright*. Stephen Allwright. Available at: https://stephenallwright.com/good-accuracy-score/ (Accessed: January 18, 2023).

Statistics by Jim (2021) Statistics By Jim. Available at: https://statisticsbyjim.com/ (Accessed: January 18, 2023).

Murray, E. (2019) The importance of color in data visualizations, Forbes. Forbes Magazine. Available at: https://www.forbes.com/sites/evamurray/2019/03/22/the-importance-of-color-in-data-visualizations/ (Accessed: January 18, 2023).