A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Title:** | *MSc in Data Analytics* | | |
| **Cohort:** | *MSc in Data Analytics SB+/FT (Sept22 start)* | | |
| **Module Title(s)**: | *Programming for DA*  *Statistics for Data Analytics*  *Machine Learning for Data Analysis*  *Data Preparation & Visualisation* | | |
| **Assignment Type:** | *Individual* | **Weighting(s)**: | *Programming for DA* ***50%***  *Stats for Data Analytics* ***50%***  *ML for Data Analysis* ***50%***  *Data Prep & Vis* ***50%*** |
| **Assignment Title:** | *MSC\_DA\_CA2* | | |
| **Lecturer(s)**: | *Marina Iantorno/John O’Sullivan*  *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid* | | |
| **Issue Date:** | *27th November 2022* | | |
| **Submission Deadline Date:** | *6th January 2023* | | |
| **Late Submission Penalty:** | Late submissions will be accepted up to **5** calendar days after the deadline. All late submissions are subject to a penalty of **10%** of the mark awarded.  Submissions received more than 5 calendar days after the deadline above **will not** be accepted and a mark of 0% will be awarded. | | |
| **Method of Submission:** | **Moodle** | | |
| **Instructions for Submission:** | *Place all files into a* ***standard .zip file*** *and upload.*  *Expected files : Written report (pdf / word), Code files (jupyter notebook), Dashboard (Either python or Jupyter Notebook),* | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | *After the Exam Boards January 2023* | | |

**Ireland Agriculture data - Candy**

Talita Pereira De Souza

SBA22233

[sba22233@studentcct.ie](mailto:sba22233@studentcct.ie)

January 06, 2023

Text

Description automatically generated

# Framework:

# GitHub repository

# <https://github.com/Talitapsouz/CA2_EUAgriculture_Project.gi>

# Collaborators added

# Marina Iantorno - miantorno@cct.ie

# Sam Weiss - sweiss@cct.ie

# Muhammad Iqbal - miqbal@cct.ie

# David McQuaid - dmcquaid@cct.ie

# Introduction

To collect and develop a dataset based on the agriculture topic related to Ireland, you could consider using a combination of primary and secondary sources.

Surveys or questionnaires distributed to farmers, producers, and consumers

Interviews with industry experts or stakeholders

On-site observations or measurements of agricultural practices or conditions.

Secondary sources of data might include:

Agricultural statistics or reports published by government agencies or organizations such as the Central Statistics Office (CSO) in Ireland or the Food and Agriculture Organization (FAO) of the United Nations Online databases or directories of agricultural businesses or organizations in Ireland News articles or other media coverage of agricultural issues in Ireland.

# Data Analysis

Some options for performing data analysis in Python include:

NumPy: a library for working with large, multi-dimensional arrays and matrices of numerical data.

pandas: a library for working with tabular data, such as data stored in a spreadsheet or a CSV file. pandas provide functions for reading, manipulating, and analyzing data.

matplotlib: a library for creating static, animated, and interactive visualizations in Python.

seaborn: a library for creating statistical visualizations in Python, built on top of matplotlib.

scikit-learn: a library for machine learning in Python, including functions for training and evaluating models, and for preprocessing and transforming data.

You can use these tools and libraries in a Jupyter notebook by importing them into your notebook and using the functions and methods they provide. For example, you could use pandas to read in a CSV file and use NumPy to perform calculations on the data, or use matplotlib to create a visualization of the data.

You can find the code load and explored a dataset in a Jupyter notebook attached [Python analysis.ipynb].

It is important to include sound justifications and explanations of your code choices in the project documentation to help others understand the reasoning behind your decisions and to ensure that your code is easy to maintain and modify in the future.

Comment your code: Add comments to your code to explain what each part of the code is doing. This will help others understand the purpose of the code and how it works.

Document your code: Use docstrings to document the purpose and behavior of functions, classes, and modules. This will make it easier for others to understand how to use your code and how it is intended to be used.

Use descriptive variable names: Choose variable names that clearly describe the purpose of the variable. This will make it easier for others to understand the code and will reduce the need for additional comments.

Follow coding standards: Follow established coding standards, such as the Python PEP 8 style guide, to ensure that your code is easy to read and understand. This includes things like using consistent indentation, choosing descriptive function and variable names, and following conventions for naming variables and functions.

Justify your choices: In the documentation, provide explanations for why you made certain choices in your code, such as why you chose a particular algorithm or data structure. This will help others understand the reasoning behind your decisions and will make it easier for them to modify or extend your code in the future.

By following these guidelines, you can ensure that your code is well-documented and easy to understand, which will make it easier for others to use and maintain.

## Statistics

## Using statistics and appropriate visualisations in order to summarise the dataset used, and to help justify the chosen models.

## In the pictures 1,2 and 3 you can visualize the feature distribution by plotting the histograms.

## Chart Description automatically generated

Figure 1. This image was extracted from Jupyter notebook [Python analysis.ipynb].

## Chart, scatter chart Description automatically generated

Figure 2. This image was extracted from Jupyter notebook [Python analysis.ipynb].

## Chart, treemap chart Description automatically generated

Figure 3. This image was extracted from Jupyter notebook [Python analysis.ipynb].

# Analyzing the variables using appropriate inferential statistics to gain insights.[¶](http://localhost:8891/notebooks/Documents/CA2/Python%20analysis.ipynb#Analysing-the-variables-using-appropriate-inferential-statistics-to-gain-insights.)

# Mean

## Calculate the sample mean. To calculate the sample mean of a continuous variable, you can use the mean() function from Numpy.

## Text Description automatically generated with low confidence

Sample mean: 0.44

# Median

## Calculate the sample median. To calculate the sample median of a continuous variable, you can use the median() function from Numpy

Diagram

Description automatically generated

Sample median: 0.00

# Mode

Calculate the sample mode. To calculate the sample mode of a categorical variable, using the mode() function from Pandas.

Diagram

Description automatically generated with low confidence

Sample mode: 100 Grand

# Variance

# Calculate the sample variance. To calculate the sample variance of a continuous variable, using the var() function from Numpy.

Text

Description automatically generated

Sample variance: 0.25

# Standard Deviation

# Calculate the sample standard deviation. To calculate the sample standard deviation of a continuous variable, using the std() function from Numpy.

# Text Description automatically generated

Sample standard deviation: 0.50

These steps will allow to calculate a variety of sample statistics for the data in your dataset. These statistics can be used to make inferences about the population and to compare different groups or samples.

# Parametric and non-parametric inferential statistical techniques

# Normality

p-value for equal variance test = 0.006

In this case, the p-value for the normality test for both Ireland and Other Country is 0.000, which means that there is a very low probability (less than 0.001) that the data in these samples followed a normal distribution by chance. This suggests that the data in these samples is not normally distributed.

This could have implications for the statistical analyses that you conduct on these data, as many statistical tests and assumptions are based on the assumption of normality. If the data are not normally distributed, you might need to use different statistical tests or techniques to analyse the data.

# P – VALUE

A p-value is a measure of the probability that a statistical result occurred by chance. When you conduct an equal variance test, you are testing the hypothesis that two samples have equal variances.

In this case, the p-value for the equal variance test is 0.006, which means that there is a small probability (0.006 or 6%) that the variances of the two samples are equal by chance. This suggests that there is some evidence that the variances of the two samples are not equal.

This could have implications for the statistical analyses that you conduct on these data, as many statistical tests and assumptions are based on the assumption of equal variance. If the variances of the two samples are not equal, you might need to use different statistical tests or techniques to analyse the data.

# T – TEST

t-test: t-statistic = 2.850, p-value = 0.006

A t-test is a statistical test used to compare the means of two samples. The t-statistic is a measure of the difference between the means of the two samples, and the p-value is a measure of the probability that this difference occurred by chance.

In this case, the t-statistic is 2.850 and the p-value is 0.006. The t-statistic indicates that there is a significant difference between the means of the two samples.

The p-value indicates that there is a small probability (0.006 or 6%) that this difference occurred by chance. This suggests that the difference between the means of the two samples is statistically significant.

It's important to note that the p-value is dependent on the sample size and the difference between the means. A larger sample size and a larger difference between the means will result in a smaller p-value, indicating a stronger statistical significance.

# ANOVA

ANOVA: F-statistic = 8.125, p-value = 0.006

ANOVA (Analysis of Variance) is a statistical test used to compare the means of three or more samples. The F-statistic is a measure of the difference between the means of the samples, and the p-value is a measure of the probability that this difference occurred by chance.

In this case, the F-statistic is 8.125 and the p-value is 0.006. The F-statistic indicates that there is a significant difference between the means of the samples. The p-value indicates that there is a small probability (0.006 or 6%) that this difference occurred by chance. This suggests that the difference between the means of the samples is statistically significant.

It's important to note that the p-value is dependent on the sample size and the difference between the means. A larger sample size and a larger difference between the means will result in a smaller p-value, indicating a stronger statistical significance.

# Paired t-test

Paired t-test: t-statistic = 4.593, p-value = 0.000

A paired t-test is a statistical test used to compare the means of two related samples. The t-statistic is a measure of the difference between the means of the two samples, and the p-value is a measure of the probability that this difference occurred by chance.

In this case, the t-statistic is 4.593 and the p-value is 0.000. The t-statistic indicates that there is a significant difference between the means of the two samples.

The p-value indicates that there is a very low probability (less than 0.001) that this difference occurred by chance. This suggests that the difference between the means of the two samples is statistically significant.

It's important to note that the p-value is dependent on the sample size and the difference between the means. A larger sample size and a larger difference between the means will result in a smaller p-value, indicating a stronger statistical significance.

On the drawback of using random forest is that although it can be used for both classification and regression tasks, it is not more suitable for Regression tasks.

# Chi - squared test

Chi-squared test: chi2 = 80.981, p-value = 0.000

A chi-squared test is a statistical test used to compare observed and expected frequencies in a categorical data set. The chi-squared statistic is a measure of the difference between the observed and expected frequencies, and the p-value is a measure of the probability that this difference occurred by chance.

In this case, the chi-squared statistic is 80.981 and the p-value is 0.000. The chi-squared statistic indicates that there is a significant difference between the observed and expected frequencies. The p-value indicates that there is a very low probability (less than 0.001) that this difference occurred by chance. This suggests that the difference between the observed and expected frequencies is statistically significant.

## Machine Learning

## 2.1 Logistic Regression Classifier in Python - Basic Introduction[¶](http://localhost:8891/notebooks/Documents/CA2/Python%20analysis.ipynb#Logistic-Regression-Classifier-in-Python---Basic-Introduction)

# [¶](http://localhost:8891/notebooks/Documents/CA2/Python%20analysis.ipynb#Analysing-the-variables-using-appropriate-inferential-statistics-to-gain-insights.)

In logistic regression, you are performing linear regression but applying a sigmoid function for the outcome.

Sigmoid / Logistic Function p=1/1+e−y

Properties of Logistic Regression The dependent variable follows a Bernoulli Distribution Estimation is maximum likelihood estimation (MLE)

Advantages: Straight forward, easy to implement, doesn't require high compute

power, easy to interpret, used widely. Doesn't require feature scaling and provides a probability score for observations.

Disadvantages: Not able to handle a large number of category features/variables. Vulnerable to overfitting. Data Is the candy chocolate? Let's find out because... yum

(This intro was built on a candy-data.csv dataset in DataScience folder also found at <https://github.com/fivethirtyeight/data/blob/master/candy-power-ranking/candy-data.csv> )

Chart, bar chart

Description automatically generated

Figure 4. This image was extracted from Jupyter notebook [Python analysis.ipynb].

## 2.2 Model Development and Prediction [¶](http://localhost:8891/notebooks/Documents/CA2/Python%20analysis.ipynb#Logistic-Regression-Classifier-in-Python---Basic-Introduction)

Import the Scikit Learn Logistic Regression module Fit model on the train set using fit() then perform prediction on test set using prediction

## 2.3 Evaluate model using confusion matrix [¶](http://localhost:8891/notebooks/Documents/CA2/Python%20analysis.ipynb#Logistic-Regression-Classifier-in-Python---Basic-Introduction)

This is basically looking at how well the model did on predictions

## 2.4 Visualize CFM (confusion matrix) using a heatmap

## Chart Description automatically generated

Figure 5. This image was extracted from Jupyter notebook [Python analysis.ipynb].

## 2.5 ROC (Receiver Operation Characteristic) Curve

Plotting true positive rate against false positie rate. Shows tradeoff between sensitivity and specificity.

What is AUC - ROC Curve? (credit <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>) AUC - ROC curve Is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with disease and no disease.

The ROC curve is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis.

Chart

Description automatically generated

Figure 6. This image was extracted from Jupyter notebook [Python analysis.ipynb].

## 2.6 Hierarchical clustering:

Chart, line chart

Description automatically generated

Figure 7. This image was extracted from Jupyter notebook [Python analysis.ipynb].

## 2.7 K-Nearest Neighbor Classifier

Chart, line chart

Description automatically generated

Figure 8. This image was extracted from Jupyter notebook [Python analysis.ipynb].

It is clear from the above graph when n\_neighbors = 3, both the model performs the best. We now use n\_neighbors=3 and re-run the training once again.

## 2.8 Discussion

# 

# Classification Model

It looks like we are working with a classification model and have calculated the accuracy, precision, and recall of the model.

Accuracy is the fraction of correct predictions made by the model, and is calculated as the number of correct predictions divided by the total number of predictions. A model with high accuracy is able to correctly predict the class of a large number of data points.

Precision is a measure of the quality of the model's predictions, and is calculated as the number of true positive predictions divided by the total number of positive predictions made by the model. A model with high precision is able to correctly identify a large number of positive instances, but may have a lower recall (described below).

Recall is a measure of the completeness of the model's predictions, and is calculated as the number of true positive predictions divided by the total number of positive instances in the data. A model with high recall is able to identify a large number of positive instances, but may have a lower precision.

In this case, the accuracy of the model is 88.2%, meaning that the model was able to correctly predict the class of 88.2% of the data points. The precision of the model is 100%, meaning that all of the positive predictions made by the model were correct. The recall of the model is 66.7%, meaning that the model was able to identify 66.7% of the positive instances in the data.

# K-means clustering

It looks like the list provided consists of integers that may represent cluster labels or class labels.

Without more context, it is difficult to interpret the meaning of these labels. In general, cluster labels are used to indicate which cluster a data point belongs to, while class labels are used to indicate which class a data point belongs to.

For example, if the list represents cluster labels, then the integers may represent different groups or categories of data points that have been identified through clustering.

If the list represents class labels, then the integers may represent different classes or categories that the data points belong to, based on some criterion or classification.

Without more information about the data and the context in which the labels were generated, it is not possible to provide a more detailed interpretation of the labels.

It also, looks like we are using the KMeans class from the sklearn.cluster library to perform K-means clustering with 4 clusters.

To use the KMeans class, you will need to provide a dataset as input and fit the model to the data using the fit method. The model will then assign each data point to one of the 4 clusters, and you can use the predict method to predict the cluster labels for new data points.

**Hierarchical clustering**

Hierarchical clustering is a method of clustering data points into groups (also known as clusters) based on their similarity. In hierarchical clustering, the data points

are organized into a tree-like structure, with the most similar data points forming the lowest level of the tree.

To visualize the results of hierarchical clustering, you can use a dendrogram plot, which shows the hierarchical structure of the clusters.

In a dendrogram plot, the data points are represented by horizontal lines, and the clusters are represented by vertical lines that connect the data points. The height of the vertical lines indicates the distance between the clusters, with taller lines indicating a greater distance.

To interpret a dendrogram plot, you can look for patterns in the arrangement of the data points and clusters. For example, you might look for clusters of data points that are closely related to each other, or for data points that are far from the other data points in their cluster

# K-Nearest Neighbor Classifier

It looks like the number provided, 0.9117647058823529, may be a measure of accuracy for a classification model.

Accuracy is a measure of the performance of a classification model, and is defined as the fraction of correct predictions made by the model. A model with high accuracy is able to correctly classify a large number of data points.

A value of 0.9117647058823529 for accuracy would indicate that the model was able to correctly classify 91.2% of the data points. This is generally considered to be a high level of accuracy, and suggests that the model is performing well.

To evaluate the accuracy of a classification model, you can split your dataset into a training set and a test set, and use the model to make predictions on the test set. You can then compare the predicted labels to the true labels of the test set to calculate the accuracy of the model.

## 2.8 Evaluating Model

In the last part was evaluated the model on the testing data using a couple of techniques like confusion\_matrix and classification\_report.

## Data Preparation & Visualization

Here are some factors to consider when acquiring raw data for research:

Positive aspects:

Wide availability: There are many sources of raw data available online, including government agencies, research organizations, and private companies. This makes it relatively easy to find data that is relevant to your research question.

Cost: Many sources of raw data are available for free or at a low cost, which can be a significant benefit for researchers with limited budgets.

Time: It can be time-consuming to collect data from scratch, so using existing data can save time and resources.

Negative aspects:

Quality: Not all sources of raw data are reliable, and it is important to carefully evaluate the quality of the data before using it in your research. This can include checking the sources of the data, the methods used to collect it, and the accuracy of the data.

Relevance: It may be difficult to find data that is directly relevant to your research question, and you may need to adapt your research question to fit the available data.

Licensing/permissions: Some sources of raw data may be protected by copyright or other intellectual property laws, which may require you to obtain permission to use the data. This can be a time-consuming process, and there may be fees associated with obtaining permission.

In summary, acquiring raw data for research can be a positive aspect of the research process, as it allows you to build on the work of others and saves time and resources. However, it is important to carefully evaluate the quality and relevance of the data, and to be aware of any licensing or permission requirements.

## Exploratory Data Analysis (EDA)

Plotting the data:

# Line Chart

Chart

Description automatically generated

Figure 9. This image was extracted from Jupyter notebook [Python analysis.ipynb].

To interpret the findings of the line chart, you can look at the position of the line to see the values of the two variables, and try to identify any patterns or trends in the data. For example, you might see a positive correlation (where one variable increases as the other variable increases) or a negative correlation (where one variable decreases as the other variable increases).

# Scatter chart

Chart, scatter chart

Description automatically generated

Figure 10. This image was extracted from Jupyter notebook [Python analysis.ipynb].

A scatter chart is a graphical representation of the relationship between two variables. It is a plot where each data point is represented by a marker, and the position of the marker represents the values of the two variables.

To interpret the findings of the scatter chart, you can look at the position of the markers to see the values of the two variables, and try to identify any patterns or trends in the data. For example, you might see a positive correlation (where one variable increases as the other variable increases) or a negative correlation (where one variable decreases as the other variable increases).

In this case, the scatter chart is showing the relationship between the pricepercent column and the winpercent column in the df dataframe. The x-axis is labeled with the text "pricepercent", and the y-axis is labeled with the text "winpercent".

# Histogram

Chart, histogram

Description automatically generated

Figure 11. This image was extracted from Jupyter notebook [Python analysis.ipynb].

A histogram is a graphical representation of the distribution of a dataset. It is a bar graph where the x-axis represents the range of values in the data, and the y-axis represents the number of data points that fall within each range.

To interpret the findings of the histogram, you can look at the x-axis to see the range of values in the data, and the y-axis to see the number of data points that fall within each range. You can use this information to understand the distribution of the data and identify any patterns or trends.

In this case, the histogram is showing the distribution of the pricepercent column in the df dataframe. The x-axis is labeled with the text "pricepercent", and the y-axis is labeled with the text "count".

# Heatmap

A picture containing treemap chart

Description automatically generated

Figure 12. This image was extracted from Jupyter notebook [Python analysis.ipynb].

A heatmap is a graphical representation of data where the values are represented as colours. In this case, the heatmap function is being used to plot the correlations between the columns in the df dataframe. The colour of each cell in the heatmap represents the strength of the correlation between the two columns. Darker colours indicate a stronger positive correlation, while lighter colours indicate a stronger negative correlation.

To interpret the findings of the heatmap, you can look at the colour of each cell to see the strength of the correlation between the two columns. You can also look at the values in the cells to see the exact correlation coefficient.

# Dashboard

# A screenshot of a computer Description automatically generated with low confidence

Figure 13. This image was extracted from Jupyter notebook [Python analysis.ipynb].

**References**

Anggraeni, W. et al. (2018) ‘Agricultural strategic commodity price forecasting using artificial neural network’, in 2018 International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2018. Institute of Electrical and Electronics Engineers Inc., pp. 347– 352. doi: 10.1109/ISRITI.2018.8864442.

Birthal, P., Negi, A. and Joshi, P. K. (2019) ‘Understanding causes of volatility in onion prices in India’, Journal of Agribusiness in Developing and Emerging Economies. Emerald Group Publishing Ltd., 9(3), pp. 255–275. doi: 10.1108/JADEE-06-2018-0068.

Yu, W. et al. (2019) ‘Analysis of Vegetable Price Fluctuation Law and Causes based on Lasso Regression Model’, in Journal of Physics: Conference Series. Institute of Physics Publishing, p. 012002. doi: 10.1088/1742-6596/1284/1/012002

Zong, J. and Zhu, Q. (2012) ‘Apply grey prediction in the agriculture production price’, Proceedings - 2012 4th International Conference on Multimedia and Security, MINES 2012. IEEE, pp. 396–399. doi: 10.1109/MINES.2012.78

Chen, Q. et al. (2019) ‘Price prediction of agricultural products based on wavelet analysis-lstm’, in Proceedings - 2019 IEEE Intl Conf on Parallel and Distributed Processing with Applications, Big Data and Cloud Computing, Sustainable Computing and Communications, Social Computing and Networking, ISPA/BDCloud/SustainCom/SocialCom 2019. Institute of Electrical and Electronics Engineers Inc., pp. 984–990. doi: 10.1109/ISPA-BDCloudSustainCom-SocialCom48970.2019.00142.

Parmezan, A. R. S., Souza, V. M. A. and Batista, G. E. A. P. A. (2019) ‘Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model’, Information Sciences. Elsevier Inc., 484, pp. 302–337. doi: 10.1016/j.ins.2019.01.076

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

G. E. P. Box, W. G. Hunter, J. S. Hunter, Statistics for experimenters: an introduction to design, data analysis, and model building, John Wiley and Sons, 1978.