**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:** |  |
| **Lecturer Name:** | Sam Weiss  Marina Iantorno  Muhammad Iqbal  David McQuaid |
| **Student Full Name:** | Alona Aldushyna |
| **Student Number:** | sba22253 |
| **Assessment Due Date:** | 28 November 2022 |
| **Date of Submission:** | 6 January 2023 |

**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. |

**Irish Agricultural Data Analysis.**

**Comparison of Ireland's agricultural sector with other countries of the world.**

*“Today, big data is ubiquitous, machine learning applications are thriving, artificial intelligence appears in everyday conversations, and the internet of things is present even in household appliances. Businesses and organizations are increasingly managed through cloud computing and high performance computing is progressively accessible as a service…More effective operations, reduced uncertainties, and real time decision-support could revolutionize agriculture to a great extent . Food could be produced more efficiently, of higher nutritional quality, in more stable supplies, with less environmental damage, and likely with additional economic, social, and ecological benefits.”*(Sjoukje A. Osinga, Dilli Paudel, Spiros A. Mouzakitis, Ioannis N. Athanasiadis (2022))

I used github as version control:

<https://github.com/Alona-Aldushyna/CA_2.git>

We looked food and agriculture data. Data taken from the official website:

[https://www.fao.org](https://www.fao.org/)

FAOSTAT provides free access to food and agriculture data for over 245 countries and territories and covers all FAO regional groupings from 1961 to the most recent year available.

Consider the Irish agriculture date, which contains the data of interest to us in columns: Element, Item (Apples, Milk, etc.), Year (1961-2020), Value, Flag Description (Official figure, Estimated value, Imputed value, Unofficial figure).

We will do Exploratory Data Analysis (EDA) is the process of visualizing and analyzing data to extract insights from it. In other words, EDA is the process of summarizing important characteristics of data in order to gain better understanding of the dataset.

Following are the main steps took to process the data.

1. Checked the data for duplicates. There are no duplicates.

2. Checked the data for missing data.

It is not uncommon for samples to miss one or more values for various reasons. Perhaps an error crept into the data collection process, some measurement data turned out to be unacceptable, individual fields could simply remain blank, for example, during a survey. We usually treat missing values as spaces in the data table, or as string placeholders such as NaN (i.e. not a number). Unfortunately, most computational tools are unable to handle such missing values or generate unpredictable results if they are simply ignored. Therefore, before proceeding with further analysis, it was crucial to sort out these missing values.

“dropna” is one of the simplest ways to solve the problem of missing data, which is to simply remove the corresponding features (columns) or patterns (rows) from the data set entirely.

**Using missingno to Identify Missing Data**

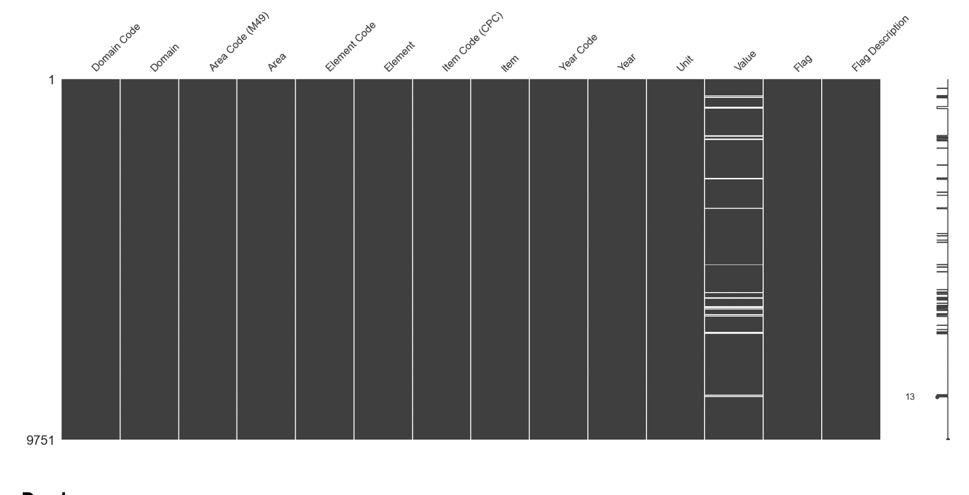
Within the missingno library, there are four types of plots for visualising data completeness: the barplot, the matrix plot, the heatmap, and the dendrogram plot. Each has its own advantages for identifying missing data.

Consider some of them in turn.

**Matrix Plot**

The matrix plot is a great tool if you are working with depth-related data or time-series data. It provides a colour fill for each column. When data is present, the plot is shaded in grey (or your colour of choice), and when it is absent the plot is displayed in white.

Built a matrix plot and saw missing data (fig 1):

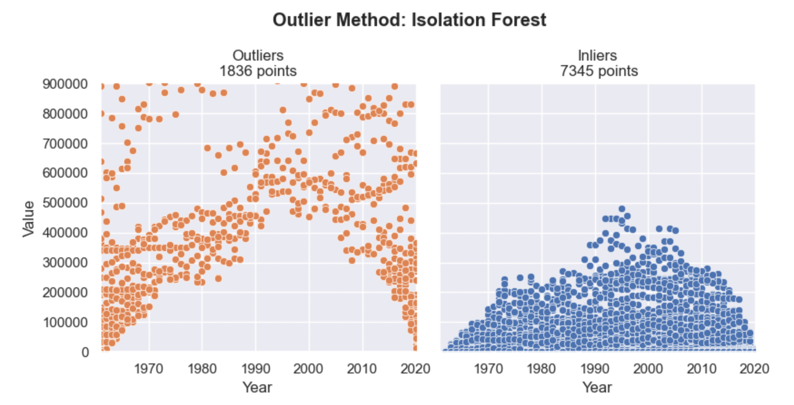


Building the Isolation Forest Model with Scikit-Learn.

From our data frame, we need to select the variables with which we will train our isolated forest model.

In this example, I will only use two variables ('Year', 'Value'). Using two variables allows us to visualize what the algorithm did.

Visualising Anomaly Data using matplotlib:



After removing the outliers, consider and conduct a similar study for the Ukraine dataset.

Next, join the two tables using the merge method.

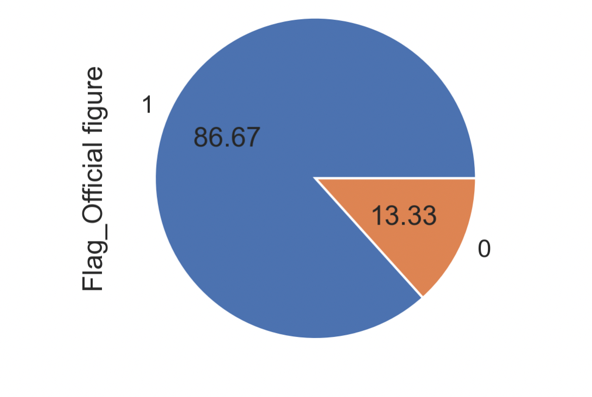
Let’s make a map! Using Geopandas, Pandas and Matplotlib to make a Choropleth map Dealing with an unbalanced data set

Machine learning:

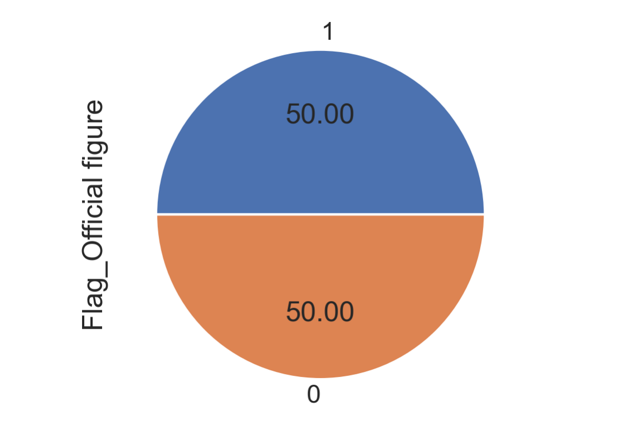
**Working with an unbalanced dataset.**

Error types play an important role when one of two classes is much more common than the other. This is a very common situation in practice. A good example is click-through rate prediction, where each data point represents an "impression" - an item presented to the user. This element can be an announcement, a story, a user of a social network. The goal is to predict whether the user, when shown a given element, will click on it (indicating their interest). Most of what the user sees on the Internet (in particular, advertisements) is not of particular interest to him. You need to show the user 100 ads or articles before they find something interesting enough to click on. This results in a data set in which 99 data points correspond to a “not clicked” situation and 1 data point to a “clicked” situation. In other words, 99% of the examples are in the "no click" class. Datasets in which one class occurs much more frequently than others are often referred to as imbalanced datasets or datasets with imbalanced classes. In reality, imbalanced data is the norm, and it is rare that a class of interest occurs in the data with the same or nearly the same frequency as other classes.

Now, suppose you are building a classifier that is 99% accurate when solving the click prediction problem. What does it say? 99% accuracy sounds impressive, but it doesn't take class imbalance into account. You can achieve 99% correctness without building a machine learning model, always predicting “no click.” On the other hand, even for imbalanced data, a model with 99% correctness could be quite suitable. However, in this case, the correctness does not allow us to distinguish the model “constantly predicting the absence of a click” from a potentially good model.



After resampling the dataset, we plot it again and it shows an equal number of classes.



Applying ML Models

NOW THE DATA EXPLORATION, ANALYSIS AND VISUALIZATION AND PREPROCESSING HAS BEEN DONE AND NOW WE CAN MOVE TO MODELLING PART.

Based on this data, we need to build a machine learning model that will predict the type of stand for a new set of measurements. But before we apply our model to a new set, we need to make sure that the model actually works and its predictions can be trusted.

Unfortunately, to assess the quality of the model, we cannot use the data that we took to build the model. This is because our model will simply remember the entire training set and therefore it will always predict the correct label for any data point in the training set.

This "remembering" tells us nothing about the generalizing ability of the model (in other words, we don't know if the model will work just as well on new data).

To evaluate the effectiveness of the model, we present it with new labeled data (labeled data that it has not seen before). This is usually done by splitting the collected labeled data into two parts.

One piece of data is used to build our machine learning model and is called training data or training set. The remaining data will be used to evaluate the quality of the model, they are called test data (test data), test set (test set) or control set (hold-out set).

The scikit-learn library has a train\_test\_split function that shuffles the dataset and splits it into two parts. This function selects 80% of the data rows with the corresponding labels in the training set. The remaining 20% of labeled data is declared a test set. The question of how much data to sample in the training and test sets is debatable, but using a test set containing 20% of the data is a good rule of thumb.

We split the dataset into training and test subsets.

The variable "y" was assigned class labels from the column Official figure

variable "X" all other columns. We then used the train\_test\_split function to randomly split x and y into separate training and test datasets.

By setting the test\_size = 0.2 parameter, we assigned 20% of the wine samples to the X\_test and y\_test arrays and the remaining 80% of the samples to the X\_train and y\_train arrays, respectively.

*Models*

We applied several models and algorithms to our dataset, accounting for their merits and demerits based on literature review. These are as follows:

1. Gaussian NB

Naive Bayes is a group of extremely fast and simple classification algorithms, often suitable for very high dimensional datasets. Because of their speed and so few adjustable parameters, they are very useful as a rough reference for classification problems.

Naive Bayes classifiers are based on Bayesian classification methods, which are based on Bayes' theorem, an equation that describes the relationship of conditional probabilities of statistical quantities. In Bayesian classification, we are interested in finding the probability of a label (category) for certain given features that are the results of observations/experiments, denoted

Naive Bayes models share many of the advantages and disadvantages of linear models. They learn and predict very quickly, and the learning process is easy to interpret. The models work very well with high-dimensional sparse data and are relatively robust to parameter changes. Naive Bayes models are great base models and are often used on very large datasets where even a linear model can take too long to train.

1. KNN (K-Nearest Neighbors)

This way we can make a prediction for each stand in the test set and compare it to the actual label (an already known stand). We can evaluate the quality of the model by calculating the accuracy (accuracy) - the percentage of stands for which the model correctly predicted the stand: 0.825

Alternatively, we can use the knn object's score method, which

calculates the correctness of the model for the test set: 0.825

The correctness of this model for the test set is about 0.825, which means that we made a correct prediction for 83% of the stands in the test set. With some mathematical assumptions, this means that we can expect our model to be correct 83% of the time for new testbeds. For us, this high level of correctness means that our model can be reasonably robust in use.

*Linear Models*

Next, we considered Linear Models

Linear models are a class of models that are widely used in practice and have been the subject of detailed study over the past few decades, and their history goes back over a hundred years. Linear models make predictions using a linear function of the input features.

For regression, the general predictive formula of the linear model is as follows:

*y*ˆ = *w* [ 0 ] \* *x* [ 0 ] + *w* [ 1 ] \* *x* [ 1 ] + . . . + *w* [ *p* ] \* *x* [ *p* ] + *b*

For a data set with one feature, this formula looks like:

*y*ˆ = *w* [ 0 ] \* *x* [ 0 ] + *b*

Linear models for regression can be characterized as regression models where the prediction is a straight line for one feature, a flat when two features are used, or a hyperplane for more dimensions (that is, when many features are used).

Here, x[0] through x[p] denote features (in this example, the number of features is p+1) for a single data point, w and b are the model parameters estimated during training, and yˆ is the prediction, issued by the model.

When the straight line predictions are compared to the KNeighborsRegressor predictions, using the regression line to get the predictions seems very strict. It seems that all the fine details of the data are not taken into account.

In a sense, this is true. We make a strong (and somewhat unrealistic) assumption that our target variable y is a linear combination of features. However, the analysis of one-dimensional data gives a somewhat distorted picture. For datasets with many features, linear models can be very useful. In particular, if you have more features than the number of data points to train, any target variable y can be nicely modeled (on the training set) as a linear function.

There are different kinds of linear models for regression. The difference between these models lies in the way the model parameters w and b are estimated from the training data and the complexity of the model is controlled. Now we will look at the most popular linear models for regression.

1. Linear Regression

After training and test accuracy of model*:*

training set: 1

test set: 1

Next, we wrote a function to create an adjusted R-Squared.

For our task, we obtained the following values:

R-squared on training set: 1

R-squared on test set: 1

Evaluation of model r2 is more on test data, so no retraining.

One of the most commonly used alternatives to standard linear regression is ridge regression, which we'll look at below.

1. Ridge regression

R-squared on training set: 0.99

R-squared on test set: 0.99

1. Lasso Regression

For our task, we obtained the following values:

R-squared on training set: 0.99

R-squared on test set: 0.999

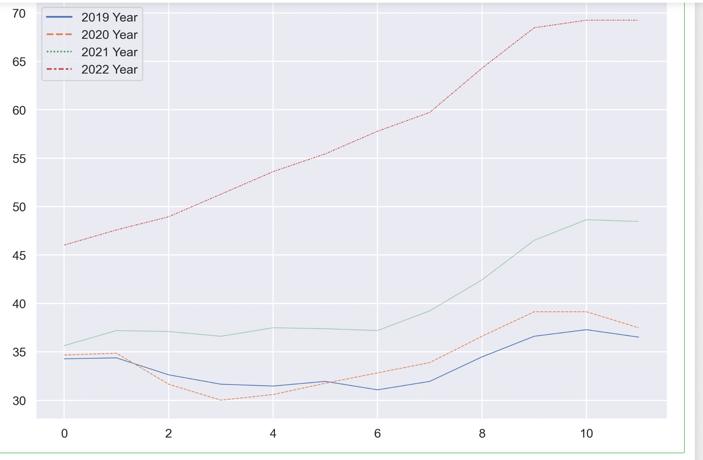
**Sentimental analysis**

Sentiment analysis is contextual mining of text which identifies and extracts subjective information in source material, and helping a business to understand the social sentiment of their brand, product or service while monitoring online conversations. However, analysis of social media streams is usually restricted to just basic sentiment analysis and count based metrics. This is akin to just scratching the surface and missing out on those high value insights that are waiting to be discovered.

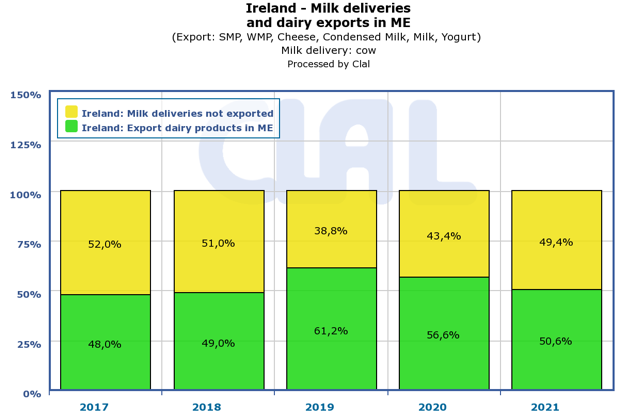
We used: Multinomial Naive Bayes, Bernoulli Naive Bayes

**Milk prices in Ireland**

The graph visually shows how much milk prices have changed in 2022:



In Ireland:

1. accounts for 6,2% of EU milk deliveries (year 2021),
2. exports dairy products, converted into milk equivalent (ME), for the 50,8% of the milk deliveries (year 2021)
3. accounts for 5,7% of EU dairy exports in ME (year 2021) 

On this page, you will find the monthly weighted average prices data for farm-gate raw milk (at real fat and protein content) in Ireland, published by the European Commission (Milk Market Observatory).

The prices collected over the last four years are listed in the table.

<https://www.clal.it/en/index.php?section=latte_mmo&country=IE>

Descriptive Statistics

The average value of mean and median is 50% of the price in 2021 - 40.33 and 37.44 It is easy to notice that the average price in 2022 increased by 42% compared to 2021. And by 71% compared to 2019.

**Comparing milk prices in Ireland with milk prices in Germany and Italy in 2022**

*Shapiro-Wilk test*:

In the field of frequentist statistics, the Shapiro-Wilk test is a test for normality. This test is generally used for small samples of 50 elements or less.

The null hypothesis for the Shapiro-Wilk test is that the milk price is distributed normally. The alternate hypothesis, therefore, is that the milk price is not distributed normally.

*One-way Analysis of Variance (ANOVA):*

We want to see if the price of one country outperforms another, so our null hypothesis is that the average price in each group is equivalent to the average price in the other groups. For simplicity, we will consider 3 groups (countries "Ireland", "Germany", "Italy") of 12 months each.

*Student's t-test for independent groups:*

The Student’s t-Test is a statistical hypothesis test to compare whether two samples are expected to have been drawn from the same population. It is named for the pseudonym “Student” used by William Gosset, who developed the test.

The test works by checking the means from two samples to see if they are significantly different from each other. It does this by calculating the standard error in the difference between means, which can be interpreted to see how likely the difference is, if the two samples have the same mean (the null hypothesis).

The t-statistic calculated by the test can be interpreted by comparing it to critical values from the t-distribution. The critical value can be calculated using the degrees of freedom and a significance level with the percent point function (PPF).

We can interpret the statistic value in a two-tailed test, meaning that if we reject the null hypothesis, it could be because the first mean is smaller or greater than the second mean.

Non-Parametric Hypothesis t-Test:

*Wilcoxon Signed-Rank*

Tests whether the distributions of two independent samples are equal or not. The same as the One-sample t-test. Assumptions:

\*Observations in each sample are independent & identically distributed(iid)

\*Observations in each sample can be ranked.

*Mann-Whitney U*

tests whether the distributions of two paired samples are equal or not. Sample like Paired t-Test

Assumptions:

Observations in each sample are independent & identically distributed (iid). Observations in each sample can be ranked. Observations across each sample are paired.

*Kruskal-Wallis H*

tests whether the distributions of two or more independent samples are equal or not.

What is the difference between Mann-Whitney and Kruskal-Wallis?

The major difference between the Mann-Whitney U and the Kruskal-Wallis H is simply that the latter can accommodate more than two groups.

The ANOVA (and t-test) is explicitly a test of equality of means of values. The Kruskal-Wallis (and Mann-Whitney) can be seen technically as a comparison of the mean ranks.

Assumptions:

Observations in each sample are independent and identically distributed (iid). Observations in each sample can be ranked.

*Friedman Test*

tests whether the distributions of two or more paired samples are equal or not alternative to the Repeated Measures ANOVA.

It is used to determine whether or not there is a statistically significant difference between the means of three or more groups in which the same subjects show up in each group.

Assumptions: Observations in each sample are independent and identically distributed (iid). Observations in each sample can be ranked. Observations across each sample are paired.

**Creating Choropleth Map Using GeoPandas — Irish cattle Dataset.**

We have created the cartogram shown below. On this map, the colors of the various counties in Ireland are color-coded according to the number of animals.

We breakdown generating the choropleth map into the following steps:

1. Data curation

The number of cattle we took from the website: <https://data.gov.ie/dataset?res_format=CSV&tags=breed>

We need to have access to the map in question. I used the map of the Republic of Ireland, which I downloaded from the following website: <https://www.eea.europa.eu/data-and-maps/data/eea-reference-grids-2/gis-files/ireland-shapefile>

1. Constract to plot

To plot the data in question, we use the MatPlotlib library in Python. I have color-coded each county based on the number of animals. I chose green ( cmap='YlGn' ). One point I would like to make here is that since we are coloring the map based on the set amount of animals, we need a color bar to show how the range of the set amount varies with color, so you need to define a minimum ( vmin ) and maximum ( vmax ) installed power and for the color bar.

Изображение выглядит как карта

Автоматически созданное описание