|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **CCT College Dublin**  **Assessment Cover Page**  *To be provided separately as a word doc for students to include with every submission*   |  |  | | --- | --- | | **Module Title:** | MSc in Data Analytics | | **Assessment Title:** | MSc DA CA1 | | **Lecturer Name:** | David Mc Quaid | | **Student Full Name:** | Stephen Kelly | | **Student Number:** | sba23305 | | **Assessment Due Date:** | 12/11/2023 | | **Date of Submission:** | 17/11/2023 |   **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. | |

Contents

[Abstract 3](#_Toc151146651)

[Programming 3](#_Toc151146652)

[Question 1 3](#_Toc151146653)

[Question 2 4](#_Toc151146654)

[Data Prep 5](#_Toc151146655)

[Question 1 5](#_Toc151146656)

[Question 2 6](#_Toc151146657)

[Question 3 6](#_Toc151146658)

[Question 4 9](#_Toc151146659)

[Statistics 9](#_Toc151146660)

[Question 1 9](#_Toc151146661)

[Question 2 11](#_Toc151146662)

[Question 3 12](#_Toc151146663)

[Question 4 12](#_Toc151146664)

[Machine learning 13](#_Toc151146665)

[Question 1 13](#_Toc151146666)

[Question 2 14](#_Toc151146667)

[Question 3 15](#_Toc151146668)

[Question 4 16](#_Toc151146669)

[Acknowledgements 17](#_Toc151146670)

[References 17](#_Toc151146671)

Table of Figures

[Figure 1: Box plot total population by year 7](#_Toc151147326)

[Figure 2: bar chart total population by gender for each year 8](#_Toc151147327)

[Figure 3: heat map for 2023 values 8](#_Toc151147328)

[Figure 4: scatter chart for each year by citizenship 9](#_Toc151147329)

[Figure 5: histogram displaying population value by frequency 11](#_Toc151147330)

[Figure 6: pie chart displaying total population for all years by gender 11](#_Toc151147331)

[Figure 7: bar chart displaying total population for all years by gender 12](#_Toc151147332)

[Figure 8: machine learning techniques by category 15](#_Toc151147333)

[Figure 9: results from different modelling techniques 16](#_Toc151147334)

[Figure 10: results between different confusion matrixes before and after adjustment 17](#_Toc151147335)

# Abstract

*An analysis of data collected by the Central Statistics Office. By applying the techniques taught at the CCT’s MSc in Data Analytics program and the use of the python programming language the data could be prepared and visualized so, a statistical analysis could be performed giving results in the form of probabilities which could then be used to predict future trends regarding Ireland’s growing population. Where possible machine learning was incorporated to gain an understanding of future possible trends to offer further insight on how Ireland’s demographics may or may not change over the coming years.*

# Programming

## Question 1

Python was used to import data, manipulate the data’s structure, perform calculations and present those results graphically or numerically. The Python programming language was created in 1991 by Guido Van Rossum at CWI Netherlands. Python itself is a high-level programming language that is general-purpose and somewhat easy to learn as it is concise and easy to read. The language does not depend on a specific platform to run and is free to use as well as being open source which allows the community that maintain the language to add improvements as needed. The code is written as a list of commands into scripts with the .py file extension. The code will run from first line/command to the last.

IPython stands for interactive python and provides a terminal with a web-based notebook for computing python. In this assignment IPython was used to import the Image library so images could be displayed within the notebook as well as being the backend for python itself.

“In 2015, the Ipython developers made a significant code reorganization of their project. So, the Notebook is now called the Jupyter Notebook. So, this interface is used with Python and many languages like R and Julia. IPyhton is the name of the Python backend.”[1]

In order to complete the assignment, the following libraries were used:

Pandas is an open-source library and allows for the data structure to be manipulated for quick analysis and easy viewing.

The Numpy library was used to store data in multidimensional arrays and matrices. Numpy provides high-level mathematical functions which allow for calculations to be performed on the stored data.

The Seaborn library was used for visualizing the results of the data in many formats such as heat-maps, histograms, bar charts and boxplots. Seaborn is derived from Matplotlib and is integrated with the Pandas data structures.

The Matplotlib library was used for dimensional plotting, this means results can be viewed interactively as opposed to viewing results statically from a notebook.

The statistics library when imported offered a range of functions for calculating the statistical data. Using the Statistics library with the Scipy library allowed for the importing of the binom, poisson and norm libraries which where needed to complete questions two and three of statistics section of this assignment.

“Scipy is also used for Data Computation, productivity, high-performance computing, and quality assurance. The various installation packages can be found here. The core Scipy packages are Numpy, SciPy library, Matplotlib, IPython, Sympy, and Pandas.” [2]

## Question 2

No specific programming paradigm was chosen when beginning this assignment. However, after the programming aspects of the assignment were completed an imperative programming paradigm was visible due to each instruction being part of a set instructions to yield the desired result when executed in a sequential order. Writing the code like this solved the problems asked in the assignment as the computer was told exactly what to do and when to do it.

Other options of programming paradigms would include but not limited too are functional, and object oriented. Had these paradigms been used it is possible the code would be cleaner and easier to read as the repetition of repeating cells may have been removed due to the code only having to be wrote once in a secondary file and then imported to the main program file for use.

Three of the defining characteristics of object orientated programming would help to protect the data which would uphold the integrity of the data as it is being worked on. Here are examples how each OOP characteristic could help maintain data integrity. Encapsulation would prevent users or even the system for changing private value the computer terminal depends upon to perform its calculations. Inheritance would remove the possibility of human error taking place when writing multiple lines of repeating code as the code can be wrote once and reused when using inheritance. Polymorphism allows for the inherited code to be changed as needed so the main functionality of the code stays while allow the user to adjust some behaviours of the code to address the problem, they are looking to find a solution for. This save time as a new class for an object does not have to be written from scratch. Similar to inheritance polymorphism also removes the possibility of human error due to writing multiple similar lines of code. Abstraction the fourth characteristic of OOP isn’t directly supported by Python.

“Programming paradigms are different ways or styles in which a given program or programming language can be organized. Each paradigm consists of certain structures, features, and opinions about how common programming problems should be tackled.”[3]

# Data Prep

## Question 1

During the exploratory data analysis phase, the dataset was read into the Jupyter notebook using the Pandas read\_excel function. Next the dataset was split into two variables with each variable being a dataset. The first variable held all of the data imported from the excel spreadsheet while the second variable contained data for 2023 only. The rows containing the 2023 data were first identified by viewing the spreadsheet and then gathered by using the drop function with the range function inserted into its parentheses. The values passed to the range function encompassed all rows containing data belonging to the year 2023.

The head function allowed for the first five rows to be viewed within the Jupyter notebook while the tail function allowed for the last five rows to be viewed. Using these two functions was a quick and easy way to check that no rows of data were lost. A more complete check would use the shape function on the variables, using this command would print the total number of rows and columns to screen. With the total numbers of rows and columns matching that of the spreadsheet it could be said with confidence that no data was lost when reading in the file or dropping rows.

The dtypes function was used to print the data type of each column which were object, int64 and float64. Objects are text/strings and cannot be used for mathematical calculations as they are not numeric values. The int64 data types are integers/whole numbers that are 64bits or 8bytes in memory. The float64 data types are numbers with a decimal point and are also 64bits or 8bytes in memory. Both the int64 and float64 data types allow for mathematical calculations to be performed on them as they are numeric values.

Some columns that held text values could be dropped as they had no use for performing calculations. This time when using the drop function the column names where inserted into the drop function’s parentheses. The drop function was given the parameter of axis=1 so, the computer would know to drop columns. At this point to ensure the correct columns had been dropped the head function was called to check that the correct columns had been dropped.

This next step is not essential but was demonstrated to show it could be done. Renaming the column heading from ‘Sex’ to ‘Gender’ was done using the rename function. Again, it was not essential but it did remove any opportunity for immature jokes to be made.

Next the duplicated function was used to check for duplicate rows. The result of this function when combined with the shape function print to the screen the number of duplicate rows. Before dropping the duplicate rows its good practice to first use the count function which will return the number of rows for each column. The drop\_duplicates function can then be called. The count function was then used again and its result was compared with the previous result from the count function. There were no duplicate values found but the steps described in this paragraph are essential when preparing data for data analytics.

Next step for preparing the data was to remove any null values. Removing the null values ensures analysis on the dataset is based on reliable data and was achieved by first using the isnull function combined with the sum function to get a count of the null values. The dropna function was used to drop the null values and a count was taken and compared with the result of the previous count. Like the duplicate values there were no null values found but the steps described in this paragraph are essential when preparing data for data analytics.

“An outlier is a datapoint or set of datapoints that are vastly different from other data points in your dataset. Sometimes they can be very high or very low.”[4]

In order to detect and remove outliers, the interquartile range scoring techniques was used. These outliers can be visually seen in boxplots and removing them increases the accuracy of the data. For example, if a dataset has 100 values that range from one to ninety-nine and the last datapoint has a value of one million this single datapoint would skew the total dataset and the resulting graphical displays would also illustrate the skew. The code for removing outliers is in the Jupyter notebook but was not needed.

## Question 2

The data was prepared using the steps outlined in question one. Further preparation of the data was implemented by rescaling the variables and splitting them into either a test or train category. The data was split by putting one column into a variable named ‘X’, this is the independent column and the remaining data was put into a variable called ‘y’ which is the dependent column. The original variables are rescaled to give equal range or variance. The scale function performed on variable ‘X’ standardises the values within the column by using the following mathematical formula: X = X-mean(X)/std(X).

“Scaling your data in machine learning (ML)is important because many algorithms use the Euclidean distance between two data points in their computations/derivations, which is sensitive to the scale of the variables.”[5]

Now the X\_train, X\_test, y\_train, y\_test can be initialized by passing the X and y variables to the train\_test\_split function. The function will also take two other parameter which make four in total. The two other parameters are test\_size and random\_state. The test\_size parameter when set to 0.3 tells the computer that 30% of the dataset goes to the test set and the remaining 70% will be sent to the training set. The random\_state parameter specifies how the data is to be shuffled as the data is passed to the variables. If set to ‘0’ the same results are obtained every time. In our case the random\_state was set to ‘4’. To finish check that the y\_train and y\_test variables have a similar distribution by comparing their mean value.

## Question 3

The boxplot was created using the Seaborn library and is a basic visualization of the distributions. When comparing data between different groups the boxplot is useful because the viewer is not overload and the information about the different groups is accurate. The lines extending from the rectangles are called whiskers. The vertical axis displays the total population values while the horizontal axis displays the year those values represent. The lower whisker for each box shows the minimum value for that year and the upper whisker shows the maximum value for that year. Then the box itself is split between the first quartile and the third quartile with the median being displayed as a line separating the two quartiles. Each box represents eighteen rows of data.

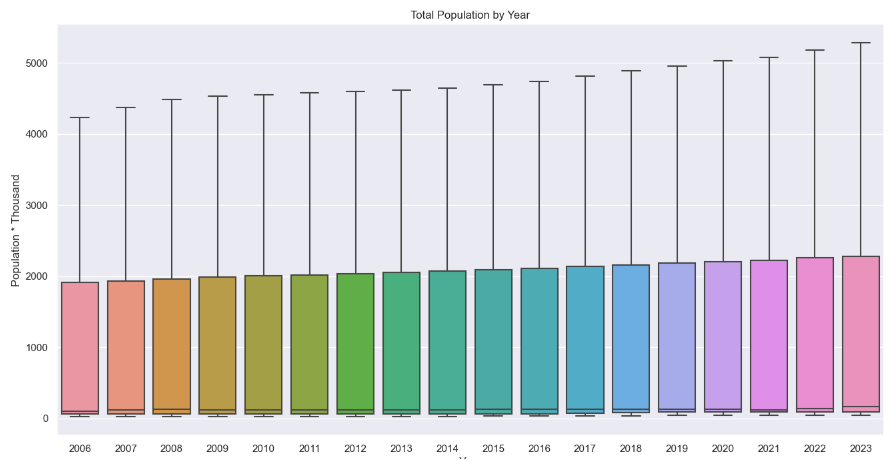


Figure : Box plot total population by year

The bar chart displays the total populations for each gender by year. At a glance it can be seen there are more females in Ireland and this trend is growing year on year. The vertical axis displays the total population value which is multiplies by one thousand to give the actual figure while the horizontal axis displays the year those values represent.

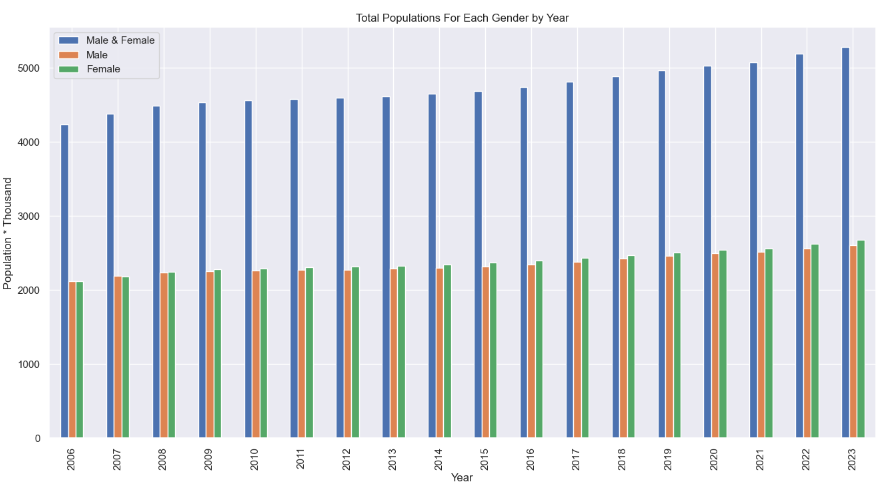


Figure 2: bar chart total population by gender for each year

The heatmap displays all values for the year 2023. Its vertical access displays the six categories of citizenship with their associated values as a row while the horizontal axis displays the gender categories and their associated values as a column. The temperature gauge starts at blue for lower population values and works its way to red to illustrate higher population values.

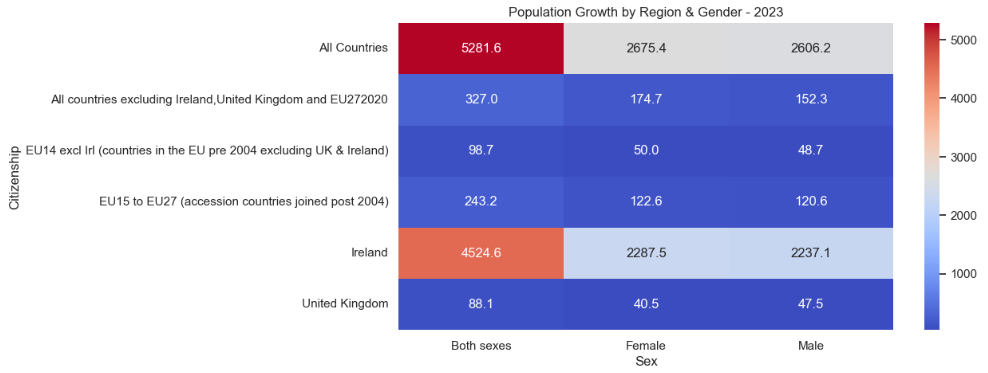


Figure : heat map for 2023 values

The scatter chart shows population totals by citizenship by year. It can be seen that since 2015 there has been a somewhat sharp rise in the total population figures. I was unable to complete the legend, only one label will show and from some research its suggested that this is an issue with pandas.[6]

The block dots represent Ireland, the yellow dots represent the UK, the green dots represent EU14 excluding Ireland, the blue dots represent EU15 to EU27 (accession countries joined post 2004) and the magenta dots represent All countries excluding Ireland, UK and EU27.

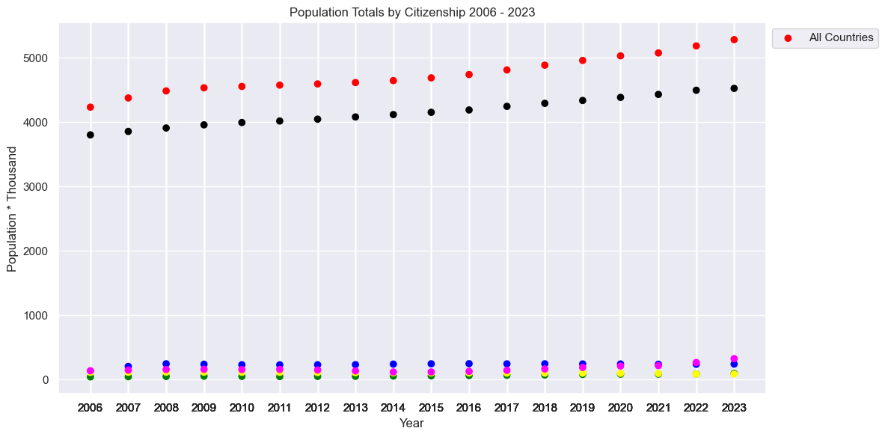


Figure : scatter chart for each year by citizenship

## Question 4

Tufts laid out six principles for graphical integrity. In short, these principles can be summed up as follows:

1. Values presented should be directly proportional to the values being measured
2. To remove any ambiguity towards the values being presented use clear labels
3. Graphics show how the data varies. Don’t distract the viewer with design variation.
4. Show the actual numeric values instead of nominal units
5. Do not add extra variables to the graph that are not present within the data
6. Graphs must be accurate without quoting values out of context

Tufte’s principles were introduced in 1983, at this time graphs were drawn by hand. The following steps focus on how to use actual ink when drawing a graph by hand and these principals also apply to today when creating a graph with the use of a computer. The less is more approach can sum up most of these principles.

1. Above all else show the data.
2. Maximize the data-ink ratio.
3. Erase non-data ink.
4. Erase redundant data ink.
5. Revise and edit.[7]

# Statistics

## Question 1

Descriptive statistics involves using samples from the dataset so the data can be transformed into information. Having the data organised allows the data analysts to pull variables from the data such as averages which allows for the data to be used as an input for inferential statistics. The right sample of data needs to be pulled to obtain accurate information from the data set. Types of samples that may be used include a random sample, stratified random sample and cluster sample. An example of a random sample would be a weekly lottery draw. The result for that particular week would be random whereas all the balls drawn out over the year would be a more complete sample. The stratified random sample is similar to the previous example except the data has been broken into groups before the random sample is taken. The cluster sample also breaks the data into groups but the difference here is the individual samples are not taken at random instead groups of data are selected so all characteristic attributes of that data can be examined within the cluster.

Using these techniques, it was possible to create graphs and charts to give information about the data at a glance. First a histogram was created which shows the frequency of numerical data using rectangles. There were 324 rows of data within the spreadsheet with one column of numerical values. The histogram’s horizontal axis groups the numeric values by population while the vertical axis displays the frequency of those values. The histogram shows of the 324 rows of data over two hundred of those rows had a value less than one thousand

Note: For the dataset to be more readable, the dataset from the CSO displays its numeric values in units of a thousand. The values displayed need to be multiplied by one thousand to give their actual value. For example, 1000 population \* 1000 per unit = 1,000,000

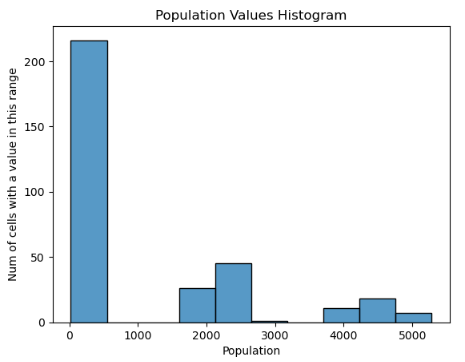


Figure : histogram displaying population value by frequency

The pie chart shows the totals for all the data within the dataset grouped by category. As the data for both sexes are combination of the male and female values it should equate to fifty percent while the values for the total amount of female and male may differ. Here we can see that there are more females in Ireland then there are males. Also, from performing manual calculations this is the expected result.

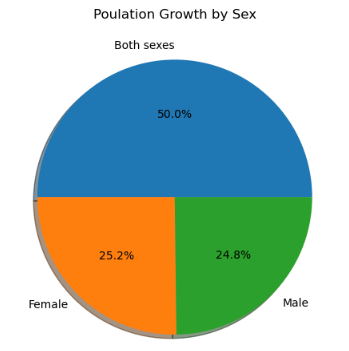


Figure : pie chart displaying total population for all years by gender

The bar chart’s vertical axis shows the total population values over eighteen years. The horizontal axis shows of data grouped by gender. As the numbers for total females and males slightly differ it may be hard to see within the chart.

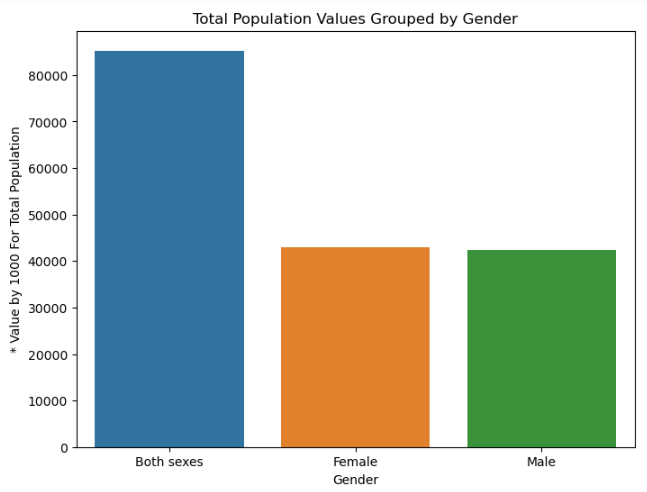


Figure : bar chart displaying total population for all years by gender

## Question 2

The Binomial Distribution is a discrete probability distribution. Distributions like Binomial tend to involve statistical analysis of counts to determine the probability of an event occurring. For this assignment Binomial distribution was used to find the probability of a given number of people being a certain gender within ten people. To get valid results for these probabilities three functions were used. These functions were made available after the binom library was imported from scipy.stats.

The probability value used by the binomial functions was calculated by counting the total population and then dividing that result by the count for the total female population giving a number which represents the probability of the next person you meet in Ireland being a woman. #probability\_of\_female # 0.506551045137837

First the probability mass function was used. To use this function three parameters were passed. The first parameter ‘k’ represents the amount of an occurrence we are looking for. The second parameter ‘n’ represents the total number of occurrences while the third parameter represents the probability of an event happening per iteration. A simple calculation to answer what the probability of three out of ten random people being female was.

The next function used was the cumulative distribution function. Using three parameters the probability of observing less than or equal to x can be obtained, ‘n’ represents the successes in trials, with the probability of success on a single trial being represented as ‘p’. To calculate the probability of having more than ‘k’ out of ‘n’ iterations the survival function was used which can be viewed as 1 – cdf.

The Poisson distribution also uses the same functions except here there are only two parameters passed. The first parameter ‘k’ represents the value we are looking to find the probability of by using the second parameter ‘mu’ which is an average value pre calculated. To use Poisson first the average growth of population over seventeen years was found. This average was set as ‘mu’ and then ‘k’ was given a value of which we want to find the probability of. For example, ‘k’ was set to 100 which represents 100K population ‘mu’ the average growth value was ran against ‘k’ to give the probability of the total population within Ireland growing over one hundred thousand in a year.

## Question 3

Normal distribution also known as Gaussian distribution is a continuous distribution. Graphs created by using normal distribution tend to look like a bell and are commonly called “the Gauss Bell Curve”. The distribution takes into consideration an infinite number of data points where the expected value and variance are finite numbers. The distribution is symmetric so the curve displayed will never be to one side instead the expected average will be in the middle giving us the bell curve previously discussed. By importing the norm library from scipy.stats we gain access to the probability mass function, the cumulative distribution function and the survival function which were previously used for the Binomial and Poisson distributions.

## Question 4

“The binomial distribution model allows us to compute the probability of observing a specified number of "successes" when the process is repeated a specific number of times.”[8]

The variables used for binomial distribution were total population for both sexes in 2023 and the total population of females for 2023. Using these two variables it was possible to find out the probability of a random person in Ireland being female which worked out as 50.65%. This probability value could then be feed to the binomial functions with the ‘k’ and ‘n’ parameters needed to complete each function.

“The Poisson distribution is another probability model that is useful for modelling discrete variables such as the number of events occurring during a given time interval.”[9]

The variables used for Poisson distribution were the sum of differences in population growth from 2006 to 2022. For integrity the following manual check was performed before accepting the result from the computer.

# 2006 total pop + sum of differences = 2023 total pop

# 4,232,900 + 1,048,700 = 5,281,600 -> confirmed true by manual check

The sum of differences was then divided by 17 as we can only calculate the differences 17 times, this is worth highlighting because there are 18 years’ worth of data in the spreadsheet. Using these two variables it was possible to find the average yearly population growth which worked out as 61,690. This average value could then be feed to the Poisson functions as mu with the ‘k’ and ‘n’ parameters needed to complete each function.

“Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.”[10]

Results from the normal distribution’s SF function can be passed to the binomial PMF function. If a Poisson distribution has a high enough mean value it can be used to approximate a normal distribution. With this in mind I would say yes, it is possible for the variables used in discrete distributions to be used in normal distribution.

# Machine learning

## Question 1

All three project management frameworks can be used for data science projects. CRoss Industry Standard Process for Data Mining involves six stages for completing data science projects. The first stage is to identify what the business needs so the stage is called business understanding. The second stage looks at what is data types are there and what is needed for the data to be cleaned, this stage is called data understanding. The third stage looks at organising the collected data and is called the data preparation. With the first three stages complete the data can now be modelled during this the modelling stage but before choosing a concrete model complete step five which is the evaluation stage. Remember the first step what does the business need and with that in mind view the data types again to ensure the correct modelling techniques being used to fulfil the business need. The final stage is deployment here it is decided upon how the stakeholders will access the results when calculations are complete.

CRISP-DM allows for mistakes and the analysist can go back to previous steps to remedy these mistake KDD does not allow for this and is more concreate in its approach. As KDD and SEMMA are almost identical the same can be expected, it either works or it doesn’t and if the model doesn’t work the analyst will have to start over.

The dataset provided by the CSO came with labels so a supervised machine learning techniques were used. Had the dataset not come with labels then unsupervised machine learning techniques would have been used. The below image displays the models associated with each technique.[11]

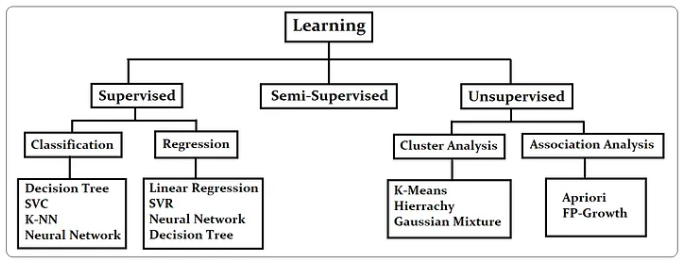


Figure : machine learning techniques by category

## Question 2

I was unable to complete the GridSearchCV or RandomizedSearchCV using the chosen dataset from the CSO office so I will have to reference the example given in class. The data is first loaded into the notebook. The data is then prepared by checking of the correct data types are being used and then removing null values if any are present. The data is split into two variables one being independent with just one column of data and the other being dependent holding the remaining data. The independent column is then scaled as to not skew results with higher valuers. This is similar to removing outliers identified in the boxplots. Once scaled both variables are then pass to the train\_test\_split function and their information is spread across four variables. The spread depends on the test\_size parameter of the train\_test\_split function and the random\_state parameter defines how the data should be shuffled when being split. At this point the mean values can be printed to confirm a similar distribution did take place between the test and train variables. The linear SVM model was then built and had the both train variables belonging to ‘X’ and ‘y’ passed to it. Here a prediction could be made by pass the \_test variable as a parameter to the models predict function with its result being used to create a classification\_report which was an imported fuction from the sklearn.metrics library.

## Question 3

I’m not able to get result with the chosen dataset. This screenshot was taken from the machine learning lecture 4 and demonstrates the differences in results between different modelling techniques. I can understand most the code used to fit the models and create the graphs but I’m unable to prepare the data from the CSO file that it may be modelled for machine learning.

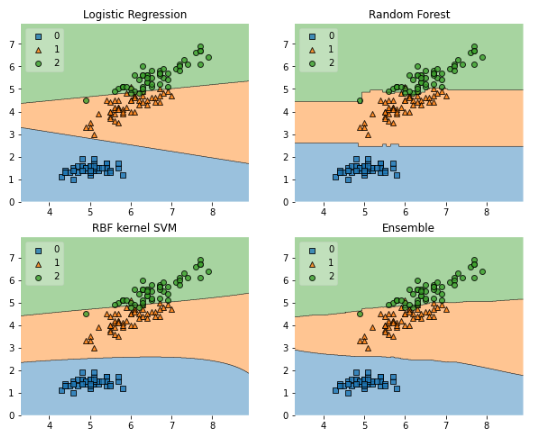


Figure : results from different modelling techniques

## Question 4

The reports are also taken form the machine learning lecture 4

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Non svc\_model no scaling | | | | svc\_model scaled data | | | |
|  | Precision | recall | f1-score | support | Precision | recall | f1-score | support |
| 0 | 0.96 | 0.89 | 0.92 | 55 | 0.98 | 0.96 | 0.97 | 55 |
| 1 | 0.93 | 0.98 | 0.96 | 88 | 0.98 | 0.99 | 0.98 | 88 |
| accuracy |  |  | 0.94 | 143 |  |  | 0.98 | 143 |
| macro avg | 0.95 | 0.93 | 0.94 | 143 | 0.98 | 0.98 | 0.98 | 143 |
| weighted avg | 0.94 | 0.94 | 0.94 | 143 | 0.98 | 0.98 | 0.98 | 143 |

The confusion matrix on the left can be understood as follows. We had 143 women in our test set. Out of 49 women predicted to not have breast cancer, 2 were classified as not having when actually they had (type one error). Out of 86 women predicted to have breast cancer, 6 were classified as having breast cancer whey they did not (type two error).

The confusion matrix on the right can be understood as follows. We had 143 women in our test set. Out of 53 women predicted to not have breast cancer, 1 were classified as not having when actually they had (type one error). Out of 87 women predicted to have breast cancer, 2 were classified as having breast cancer whey they did not (type two error).

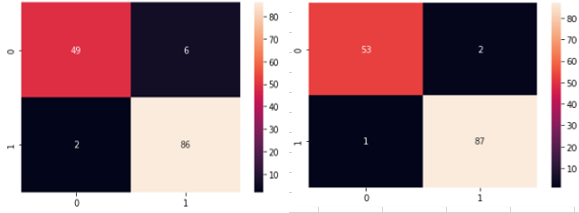


Figure : results between different confusion matrixes before and after adjustment

This one example can show the importance of tuning a machine learning model for better results. The cells in the bottom left of each matrix show people who were told they didn’t have breast cancer when they actually did have breast cancer. It can be thought of as one less law suit but it’s better to think of it as one less person going around thinking they are fine when they actually need medical attention and have time to do something about it.

The cell in the top right has similar impact, instead of telling 6 people they have breast cancer they better trained model can see that the number is 2 and this works out as four families having less stress due to their loved one’s ailments.

# Acknowledgements

Thank you to CCT for putting the effort into making the MSc fun and enjoyable to take part in. Apart from the material referenced all material came from either the lectures training material or my own understanding which I gained by attending class and studying the training material.

# References

[1] IPython display - javatpoint (2021) www.javatpoint.com. Available at: https://www.javatpoint.com/ipython-display#:~:text=It%20is%20an%20interactive%20command,single%20line%20of%20Python%20code. (Accessed: 16 November 2023).

[2] Team, G.L. (2023) *Top 30 python libraries to know in 2024*, *Great Learning Blog: Free Resources what Matters to shape your Career!* Available at: https://www.mygreatlearning.com/blog/open-source-python-libraries/#:~:text=Python%20Libraries%20are%20a%20set,data%20manipulation%20applications%2C%20and%20more. (Accessed: 16 November 2023).

[3] Cocca, G. (2022) *Programming paradigms – paradigm examples for beginners*, *freeCodeCamp.org*. Available at: https://www.freecodecamp.org/news/an-introduction-to-programming-paradigms/#:~:text=What%20is%20a%20Programming%20Paradigm%3F,programming%20problems%20should%20be%20tackled. (Accessed: 16 November 2023).

[4] Quaid, D.M. (2023) ‘Lecture 2 - What is Exploratory Data Analysis?’, *Data Preparation and Visualisation*. *Online Class*, Dublin: CCT, 26 September.

[5] Khoong, W.H. (2023) *Why scaling your data is important*, *Medium*. Available at: https://medium.com/codex/why-scaling-your-data-is-important-1aff95ca97a2#:~:text=Scaling%20your%20data%20in%20machine,the%20scale%20of%20the%20variables. (Accessed: 16 November 2023).

[6] Kassies, R. 2017 *legend only shows one label when plotting with pandas-pandas*, *Hire Developers, Free Coding Resources for the Developer*. Available at: https://www.appsloveworld.com/pandas/100/4/legend-only-shows-one-label-when-plotting-with-pandas?expand\_article=1 (Accessed: 17 November 2023).

[7] Tufte, E.R. (2007) *The visual display of quantitative information*. Second edition. Cheshire, CT, Connecticut: Graphics Press.

[8] LaMorte, W.W. (2016) *The role of probability*, *The Binomial Distribution: A Probability Model for a Discrete Outcome*. Available at: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_probability/bs704\_probability7.html#:~:text=The%20binomial%20distribution%20model%20allows,a%20success%20or%20a%20failure. (Accessed: 17 November 2023).

[9] LaMorte, Wayne W. (2016) *The role of probability*, *Central Limit Theorem*. Available at: https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704\_probability/BS704\_Probability12.html (Accessed: 17 November 2023).

[10] Chen, J. (2023) *Normal distribution: What it is, properties, uses, and Formula*, *Investopedia*. Available at: https://www.investopedia.com/terms/n/normaldistribution.asp#:~:text=Error%20Code%3A%20100013)-,What%20Is%20a%20Normal%20Distribution%3F,as%20a%20%22bell%20curve%22. (Accessed: 17 November 2023).

[11] Hshan.T (2020) *Figure 1: Different learning problems.*, *What is Supervised, Semi-Supervised and Unsupervised Learning?* medium. Available at: https://hshan0103.medium.com/what-is-supervised-semi-supervised-and-unsupervised-learning-12cd8160f99a#:~:text=Supervised%20learning%20is%20a%20problem,of%20supervised%20and%20unsupervised%20learning. (Accessed: 17 November 2023).