**University of Strathclyde**

****

**Coronavirus in Scotland**

**CS989 Big Data Fundamentals Coursework**

**SUBMITTED BY**

**Abdulla Sa’ed Hamideh**

**2020677920**

WORD COUNT

**3297**

Table of Contents

[Chapter 1: Introduction 1](#_Toc55339954)

[Chapter 2: Key challenges and problems 2](#_Toc55339955)

[Chapter 3: Data Analysis 3](#_Toc55339956)

[3.1 Source 3](#_Toc55339957)

[3.2 Exploratory Data Analysis 3](#_Toc55339958)

[3.2.1 Confirmed cases in Scotland 3](#_Toc55339959)

[3.2.2 Confirmed cases in the council area 6](#_Toc55339960)

[3.2.3 Confirmed Cases spread between Gender and Age group 9](#_Toc55339961)

[Chapter 4: Unsupervised Learning 11](#_Toc55339962)

[Clustering 11](#_Toc55339963)

[4.1 Agglomerative Clustering 11](#_Toc55339964)

[4.2 K-means 13](#_Toc55339965)

[Chapter 5: Supervised Learning 14](#_Toc55339966)

[5.1 Logistic regression 14](#_Toc55339967)

[5.2 k-nearest neighbors 15](#_Toc55339968)

[Chapter 6: Reflection 16](#_Toc55339969)

[Chapter 7: Conclusion 17](#_Toc55339970)

[Appendix: Environment 18](#_Toc55339971)

[References 19](#_Toc55339972)

**List of Tables and Figures**

**List of Figures**

[Figure 3.1: Covid-19 Monthly cases in Scotland 4](#_Toc55342511)

[Figure 3.2: Monthly Cases Distribution 5](#_Toc55342527)

[Figure 3.3: Highest 10 Scottish Cities with confirmed cases 6](#_Toc55342539)

[Figure 3.4: Covid-19 Monthly cases for the council areas 7](#_Toc55342556)

[Figure 3.5: Crude deaths Rate 8](#_Toc55342566)

[Figure 3.6: Confirmed Cases spread between gender 9](#_Toc55342574)

[Figure 3.7: Confirmed cases spread between the age group 10](#_Toc55342582)

[Figure 4.1: Comparison between the scores 12](#_Toc55342595)

[Figure 4.2: K-means 13](#_Toc55342611)

**List of Tables**

[Table 3.1: The Tables that been used in this study 3](#_Toc55342726)

[Table 5.1: Logistic Regression 14](#_Toc55342738)

[Table 5.2: KNN 15](#_Toc55342744)

# Chapter 1: Introduction

Coronaviruses are single-abandoned RNA infections, around 120 nanometres in breadth. They are susceptible and therefore highly diverse to mutation and recombination(Aronson, 2020). Around 40 different varieties exist, and they mostly infect human and non-human mammals and birds. They live in bats and wild feathered creatures and are capable of spreading to different creatures and consequently to people. The virus that causes COVID-19 is thought to have originated in bats and then spread to snakes and pangolins and then to humans, perhaps by meat contamination from wild animals, as sold in China's meat markets (ibid).

The UK's COVID-19 pandemic is part of the 2019 global coronavirus disease. COVID-19 pandemic caused by extreme acute respiratory syndrome SARS-CoV-2. In late January 2020, the first recorded case in the country. As of 22 October 2020, 810,467 confirmed cases and 44,347 confirmed deaths have been recorded (Coronavirus COVID-19 in the UK, 2020), the world's tenth-highest death rate per 100,000 population. About 90% of those dying had chronic conditions or were over 60 years of age.

However, most of coronavirus analysis reports focus on the UK as the whole picture rather than breaking it into small pictures. As there has been substantial geographical variation in the severity of the outbreak. Breaking it into small pictures will help us to solve more relationships and patterns between the data and will give us a better understanding which states, cities that are suffering the most.

This report will focus on Scotland as the big picture and breaking it into small ones in terms of spreading the virus between the cities, gender, and age. It should be possible to get a proper overview by analysing the data.

# Chapter 2: Key challenges and problems

One of the main challenges was preparing the data for visualization. As the dataset was recorded for nine months on a daily basis. It was hard to visualize it as daily cases in a graph or a diagram. To simplify these graphs and make them looks better, these daily cases were combined into months as we have monthly basis rather than a daily basis.

The main objective of this study is to identify specific patterns with exploratory data analysis of coronavirus data in Scotland. This will provide an examination of how the spread of the virus evolved, which cities were primarily affected by the virus, and the distribution of different age groups. These statistics and figures should have a good description of what the data are all about and will also provide the first consequences for the further study that follows.

This further research would cover two forms of approaches, one unsupervised and one supervised. Starting with the approach under unsupervised, this study uses two methods under unsupervised learning. Agglomerative clustering and k-means.

The second approach for this study is supervised learning. A regression analysis of the daily cases recorded for each month will be covered. This study demonstrates a trend in how the virus has evolved over this year and whether the governments were effective in controlling the virus. However, with the data provided, the exact explanation for this development cannot be evaluated, but at least it can show a pattern in the development of the virus investigation and whether patterns can be expected in the future.

# Chapter 3: Data Analysis

# Source

All the data used is publicly available on the website of Public Health of Scotland. To build the final dataset, a variety of tables have been combined. All acquired tables contain integers values. For a list of used data, refer to Table 3.1.

This dataset offers details on the number of new reported cases, deaths, negative cases, new hospital admissions of COVID-19 in Scotland, including, where applicable, Council Area levels, NHS Board, Cumulative totals and population rates at Scotland. The dataset also offers details about demographic characteristics, for instance, age, sex.

|  |  |  |
| --- | --- | --- |
| Table Name | Title | Columns |
| daily\_cuml\_scot\_20201022 | Confirmed cases in Scotland | 4 |
| total\_cases\_by\_la\_20201022 | Confirmed cases in the Council Area | 6 |
| trend\_ca\_20201022 | Daily Cases in Council Area | 8 |
| total\_cases\_agesex\_20201022 | Confirmed Cases spread between Gender and Age group | 4 |

Table 3.1: The Tables that been used in this study

# 3.2 Exploratory Data Analysis

Exploratory Data Analysis refers to the essential phase of initial data study to find trends, recognise anomalies, test theories, and verify conclusions using summary statistics and graphical representations.

# 3.2.1 Confirmed cases in Scotland

The overall cases in Scotland dataset contain the cumulative and daily cases and rates for positive coronavirus cases and deaths at Scotland level. The cases are recorded daily however to simplify things and make a better visualization the daily cases have been combined into the basis of a monthly case.

Figure 3.1 shows the change of the virus. it shows two waves. the first wave starts in March and ends in June. The second wave starts in August and still going.

Starting with the monthly cases started with a gradual increase in early March until the end of April. Then cases started going down right after Scotland announced the lockdown measures in April. The monthly cases decreased to 0 in June and remained constant up to until the beginning of August. In Mid of August, the monthly cases had a huge increased in cases and still going to this day as of 22 of October with 23000 cases. The increase in cases due to multiple factors. Firstly, the opening of schools and universities, which caused a lot of international students to come back to Scotland for studying causing an outbreak in the second wave. Secondly opening back businesses and Shopping centres. Lastly eat out help out scheme, which was implemented to help out businesses and causing people to go out more often. All these factors helped the second wave to be stronger than the first wave.

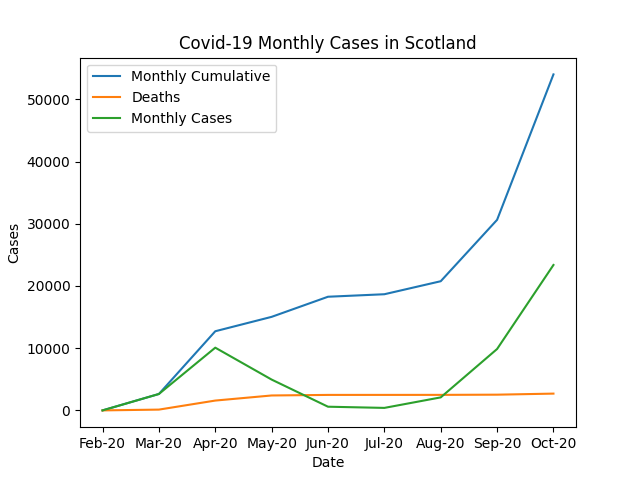
Moving on to the death cases. The cases remained 0 until mid of March. After that, the cases increased gradually at end of March. The figure shows in April a huge increase in death cases with 1400 deaths. After the lockdown, death remained constant until the end of August. However, the deaths did not have a huge increased as the monthly cases. The figure shows that the first wave contributes more than half of the death cases.

Figure 3.1: Covid-19 Monthly cases in Scotland

Figure 3.2 shows the spread for the overall cases in Scotland. the boxplots are a structured way of showing data distribution based on a summary of five numbers (minimum, first quartile, median, third quartile, and maximum).

The mean cases for the monthly cases are 6k if monthly cases are added with the one before, which becomes cumulative cases. It increases the mean cases to 19k. The maximum monthly case was 23379 in October and the minimum cases were recorded in July.

Furthermore, death cases mean is at 1800 cases. The maximum death cases were in April with 1453. The minimum death cases were in July this is because of the lockdown strategy which helps to put the cases down to 0. The government was able to control the death cases after the first wave and kept the death cases either constant or slight increase after August.

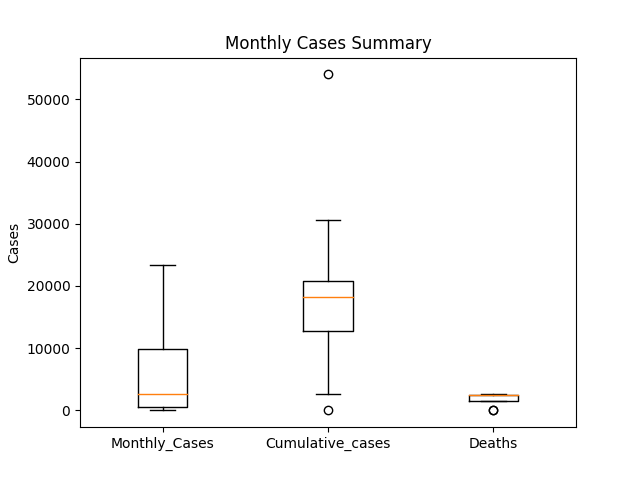


Figure 3.2: Monthly Cases Distribution

# 3.2.2 Confirmed cases in the council area

The Confirmed cases in the council area dataset contain two tables. The first table is the total cases for the 32 council area in Scotland. The table contains the total confirmed cases and deaths for the 32 council area that has been used in this study. The second table is the cumulative and daily cases and rates for positive coronavirus cases and deaths for the council area. The cases are recorded daily however to simplify things and make a better visualization the daily cases have been combined into the basis of a monthly case. The table contains Daily positive cases, deaths, crude rate deaths that have been used in this research.

Figure 3.3 shows the 10 highest council area in Scotland with confirmed cases. The figure shows the East Dunbartonshire is the least cases between the highest cities with 1435 confirmed cases and Glasgow council area leads Scotland with the confirmed cases forming 11123 positive cases. The first 6 cities, the confirmed cases are close to each other which is around the 2000 positive cases, then there’s a huge increase for the next 3 cities making it to 5000 cases, and finally, there is Glasgow area. The only city in Scotland that has above 10000 positive cases.

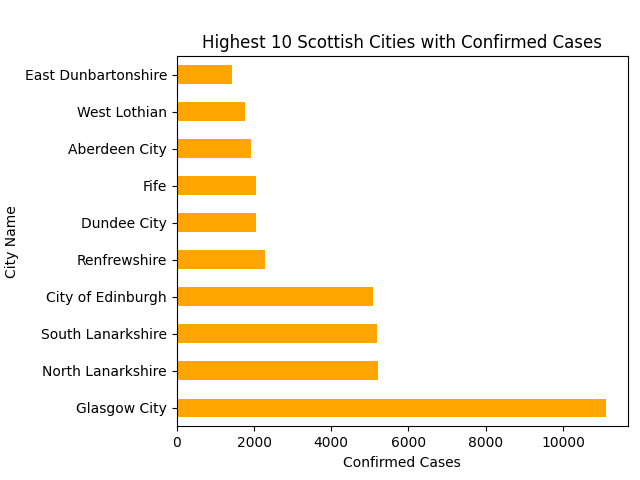


Figure 3.3: Highest 10 Scottish Cities with confirmed cases

Figure 3.4 shows the change of the virus in the 10 highest council areas. All the cities started from 0 in February and some of the cities started with a gradual increase in early of March i.e. Glasgow city however the other cities like East Dunbartonshire the gradual increase started in mid of March, that means the virus started spreading around Scotland from mid of March onwards. The increase of cases kept increasing until the end of April. After the lockdown, all cities started to record fewer cases than march and kept decreasing until all of them are at 0 at the end of June causing the end of the first wave. During July all the cities except Glasgow recorded 0 cases. Glasgow started to increase slowly at end of July, and then a huge increase from august up to until this day.

Moreover, after July all cities started to record more cases early August causing the second wave to start. There is a change in ranks for the cities between the first and second waves. For instance, the City of Edinburgh moved from second place to fourth place in the second wave. The figure points out that the second wave recorded more cases than the first wave and the spread of the virus seems transmitted faster than the first wave due to ease in the restrictions.

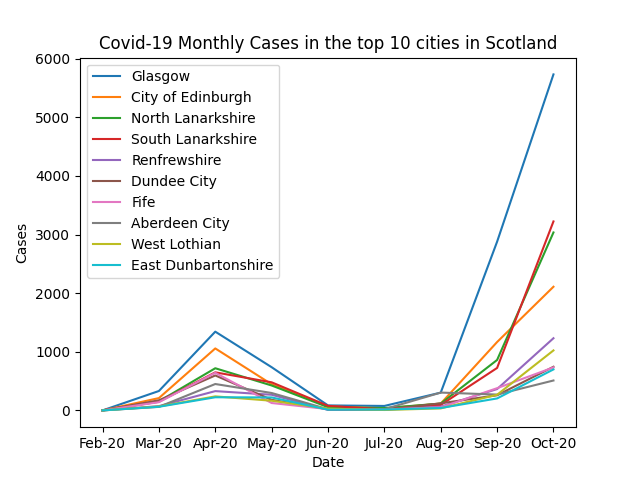


Figure 3.4: Covid-19 Monthly cases for the council areas

Figure 3.5 shows the crude rate of deaths for the 32 council areas in Scotland. the crude death rate was provided in table however it is important to know the calculation of crude rate death. The crude death rate is the number of deaths among the population of a given geographical area within a given year, per 1,000 total mid-year population of that geographical region within the same year.

The pie chart shows only 25 council areas. Seven areas were excluded due to the small crude rate of death they have. The percentage around the pie indicates the per cent of the council takes place in the pie. Figure 3.3 shows Glasgow city leads Scotland with the confirmed cases, however, figure 3.5 shows in terms of crude death rate Midlothian is taking the first place 6.9% from the pie chart with a rate of 95 and Glasgow crude death rate was 66 taking only 4.8% from the pie chart. Midlothian has recorded only 896 cases and has about 100 deaths making the severity of deaths very high.

Moreover, Cities with a big population such as Edinburgh, Glasgow, and Aberdeen. The Crude rate of deaths was small comparing it with small cities. This means cities with big population have a better-hospitalized system or they have more capabilities of holding more people in the ICU than the small cities.

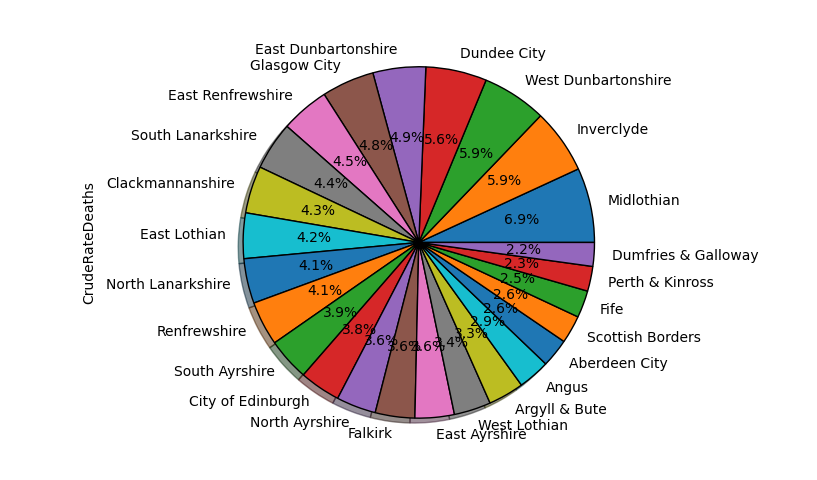


Figure 3.5: Crude deaths Rate

# 3.2.3 Confirmed Cases spread between Gender and Age group

The confirmed cases spread between gender and age group dataset contain the group age and gender counts and rates for positive coronavirus cases and deaths. The gender category contains male and female. The age group category contains from 9 categories. The age group goes from 0 to 85 plus.

Figure 3.6 shows the total positive and total deaths for both genders. It shows that female with positive cases were more than male cases. Female with 29000 cases and males with 24000 cases, however in terms of deaths female and male scored roughly the same.

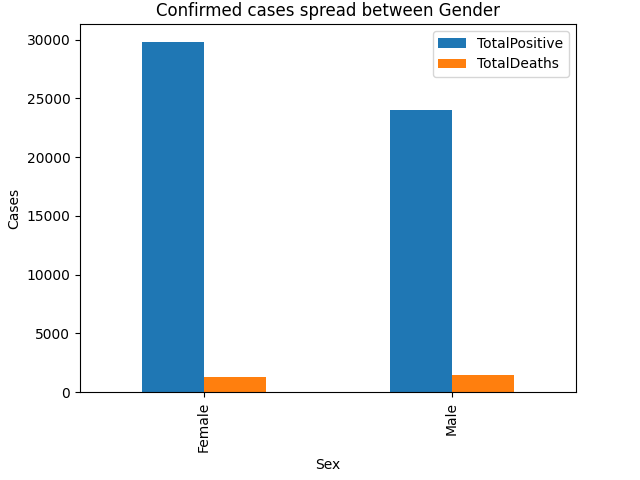


Figure 3.6: Confirmed Cases spread between gender

Figure 3.7 shows the spread of total positive and total deaths between age groups. It shows as the age increases the number of cases increases until it reaches above 65. The most age category with cases is between 45 to 65. The deaths started from 25 to 44 and started increasing as the age increases. However, from 25 to 44 only 20 death case from 15000 cases. People with age 85+ are more likely to die than any other group.

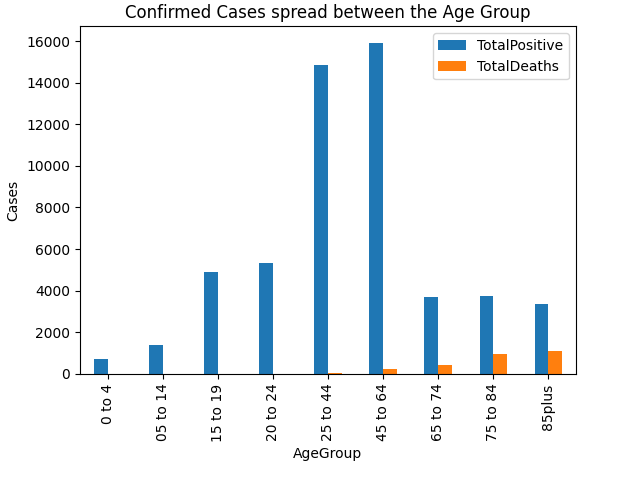


Figure 3.7: Confirmed cases spread between the age group

# Chapter 4: Unsupervised Learning

# Clustering

Clustering aims to evaluate in a collection of unlabelled data the internal classification. The objective here is to find the cities that have a similarly high level of severity of the outbreak. The data will be using is the daily cases for the council area, which is around 7,617 pieces of data. Two methods will be used to cluster the data, hierarchical clustering and K-Means clustering.

# 4.1 Agglomerative Clustering

Most common type for hierarchical clustering is agglomerative clustering, where it used to group objects in clusters on the basis of their similarity. The centrality of closeness shows how close a node is to all other network nodes. The average of the shortest path length from the point to every other point in the data.

Since we are taking the distance between points, this can be referred to as a distance-based algorithm. This algorithm takes the minimum distance between data and by using linkage. Scikit-learn documentation has multiple options to minimise the distance.

Complete: Minimize the maximum distance between cluster pairs.

Single: Minimize the distance between clusters to the closest point.

Average: Within pairs of clusters, minimize the mean distance between all points.

Ward: Minimize the variance by reducing the number of square distances between

data within the clusters.

In general, Ward linkage finds clusters of approximately equal size, as the pervasive value of bike commuters is nil, it is doubtful that this would match the information well. Very similarly, Complete and average are alike because the data has several peaks, the better option might be average. In contrast to the others, the single is unlikely to perform well, it needs very few calculations, and as such is less resilient to noisy data. There seems to be no apparent recipe from the study of the dataset, so this may be the least reliable process. Not only must the distance calculation be Euclidean, but the distance also to be travelled in each dimension is summed by Manhattan distance, and the angular distance between points has been measured by Cosine.

Furthermore, there are multiple scores to represent each strategy. Firstly, Silhouette Score used to measure the similarity in the same cluster (range: -1 to 1,). Secondly, Completeness Score The same cluster is assigned to all items of a given class (range 0-1). Finally, Homogeneity Score only items of a single class are included in each cluster (range 0-1).

Figure 3.1 shows the outcome of the scores of different distance measurements for each strategy. All strategies performed well in clustering.

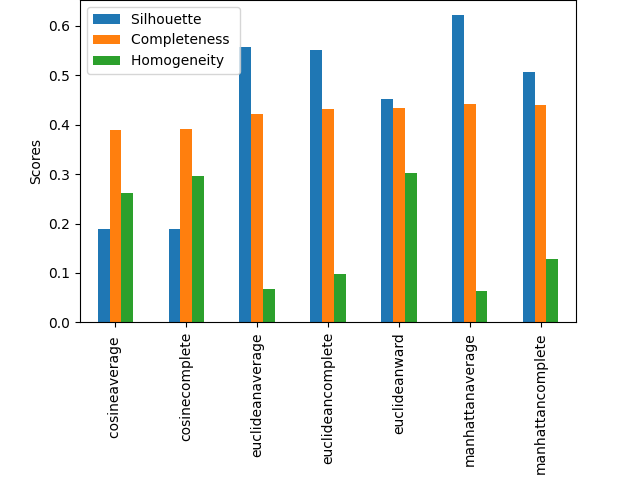


Figure 4.1: Comparison between the scores

# 4.2 K-means

The goal of K-means clustering is to break data into k clusters in such a way that data is split into k clusters. data points in the same cluster are identical and data points are further apart in the various clusters. The algorithm for K-means is not capable of estimating the number of clusters. We need to specify the number of clusters for the algorithm.

Figure 4.2 shows the k-means for a different number of clusters. The number of clusters for this case is from 2 to 50. K-Means performs best where there are four groups. This does not need to be a bad indication, there could be 0, a small amount, a moderate and a high number of cases in four classes. However, as the output metrics have fair performance.

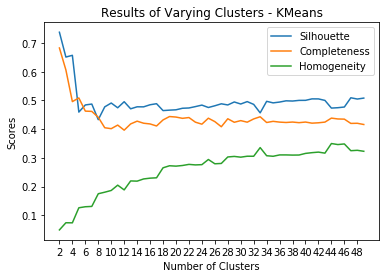


Figure 4.2: K-means

# Chapter 5: Supervised Learning

Supervised learning is an approach to the production of artificial intelligence (AI), where the algorithm is provided with labelled input data and the predicted output results. The AI system is told what to look for, so the model is trained until it is possible to identify the underlying trends and relationships, allowing it to produce good results. In this case, data means the daily recorded cases for the council cities, where the classification is the cities, however, is being replaced with binaries. The objective of this analysis to predict the recorded cases in the future. A series of coefficients that can be evaluated are regression approaches. As such, methods of regression form the bulk of the analysis here. 50 percent of the total is the proportion of data allocated for testing.

# 5.1 Logistic regression

Logistics regression works on the theory of the probability of satisfying a binary condition based on some set of inputs. The cities names replaced with 0s and 1s. the cities that have under 1200 cases were numbered 0 and cities that have over 1200 were numbered 1. The 1 indicates to there is an outbreak and 0 means there is no outbreak.

Table 5.1 shows the regression output. The total scores seem reasonably rational, putting the two binaries in the correct category. The dataset is slightly skewed, with most of the chosen dataset is zero.

Precision: Number of actual positives to all positives.

Recall: The number of all positive items that have been defined as positive.

F1: A weighted accuracy average and recall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Binary | precision | recall | F1-score | Support |
| 0 | 0.79 | 0.91 | 0.85 | 2247 |
| 1 | 0.84 | 0.66 | 0.74 | 1561 |
| Total | 0.81 | 0.79 | 0.80 | 3808 |

Table 5.1: Logistic Regression

# 5.2 k-nearest neighbors

The KNN algorithm is a non-parametric that assumes that in close proximity, similar items happen. Similar objects, in other words, are close to each other. Despite its simplicity, KNN can outperform more efficient classifiers and is used in several applications such as economic forecasting, the KNN algorithm is a simple and robust classifier that is also used as a benchmark. The KNN classifier district has higher cities with the virus than others.

Table 5.2 the KNN model using the distance parameter. The total scores seem reasonably rational, putting the two binaries in the correct category.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Binary | precision | recall | F1-score | Support |
| 0 | 0.80 | 0.86 | 0.83 | 2247 |
| 1 | 0.77 | 0.69 | 0.73 | 1561 |
| Total | 0.79 | 0.79 | 0.79 | 3808 |

Table 5.2: KNN

# Chapter 6: Reflection

I maybe would not have chosen this data if redoing the task, or possibly would have considered an inquiry that was better liable utilizing the procedures learned in class.it is better to stick into one big table and use all the technique that was taught in the class.

Furthermore, the dataset only included a limited number of variables, which limited my analysis. If they included more variables like the symptoms that could allow me to run more analysis like the correlation and linear regression between the variables and to find what kind of relationship I would be expecting.

Finally, the dataset was given a daily basis I had to manipulate the data first and make it a monthly basis, to create appropriate graphs and tables.

# Chapter 7: Conclusion

Overall, the analyses on the key human factors showing the spread of Covid-19 between cities and age groups in Scotland.

The Exploratory data analysis shows how the virus spread in the first and second wave in Scotland level. To a smaller picture the activity of monthly cases for the council areas and how Scotland government measures affected the virus to stop from spreading. Lastly, it makes sense that the most affected age group was 85+ as it is hard for their immune system to fight it.

Unsupervised learning methods did fit the data and they performed fairly. For the supervised learning also did a good job in clustering the data as the precision of the model scored fairly.

# Appendix: Environment

Python version: 3.8.5

PyCharm IDE version: 2020.2.3

Data derived from: <https://www.opendata.nhs.scot/dataset/covid-19-in-scotland>

Packages used:

• pandas

• matplotlib.pyplot

• sklearn

• cluster from sklearn

• metrics from sklearn

• scale from sklearn.preprocessing

• LabelEncoder from klearn.preprocessing

• LogisticRegression from sklearn.linear\_model

• KNeighborsClassifier from sklearn.neighbors

# References

Aronson, J. (2020) Coronaviruses - *A General Introduction - CEBM*. Available at: <[https://www.cebm.net/covid-19/coronaviruses-a-general-introduction/](https://www.cebm.net/covid-19/coronaviruses-a-general-introduction/%20)> (Accessed 1 November 2020).

Coronavirus.data.gov.uk. 2020. *Coronavirus (COVID-19) In The UK*. Available at: <[https://coronavirus.data.gov.uk/](https://coronavirus.data.gov.uk/%20)> (Accessed 1 November 2020).