**UNIVERSITY OF WESTMINSTER**

**Msc Big Data Technologies**

Module Code: 7BUIS008W.2

Module Title: Data Mining and Machine Learning

Module Leader: Dr. Panagiotis Chountas

Module Tutor: Dr. Natalia Yerashenia

Assignment: Coursework 1

Submission Date: 23 Mar. 2022

Student ID: W1848650

Student Name: Mohammadamin Shams

Table of Contents

[1 Abbreviations 3](#_Toc98919951)

[2 Covid World Data Analysis 4](#_Toc98919952)

[2.1 Data Analysis 4](#_Toc98919953)

[2.1.1 Understanding The Data and pre-processing and dimensionality reduction 4](#_Toc98919954)

[2.1.2 Clustering 7](#_Toc98919955)

[2.1.3 Analysis of Clustering and Business Values 11](#_Toc98919956)

[3 Radio Listeners Analysis for Retail purposes 12](#_Toc98919957)

[3.1 Data Analysis 12](#_Toc98919958)

[3.1.1 Understanding The Data and Pre-processing 12](#_Toc98919959)

[3.2 Market Basket Analysis 14](#_Toc98919960)

[3.2.1 Change Format of the Data 14](#_Toc98919961)

[3.2.2 Apriori Algorithm 14](#_Toc98919962)

[3.2.3 Business insights 15](#_Toc98919963)

[4 References 16](#_Toc98919964)

[5 Appendix 17](#_Toc98919965)

# Abbreviations

|  |  |
| --- | --- |
| PCA | Principal Component Analysis |
| MBA | Market Basket Analysis |

# Covid World Data Analysis

This part of report presents analysis for the Covid-19 World Data that is captured in the year of 2022. This data is obtained from different countries in the world to monitor and capture the impact of Covid-19 in countries from the start of the date covid was in our lives until now.

In This report, the goal is to divide counties in different groups to understand the way that Covid-19 hit mankind and to see if the measurements against covid for countries paid of or not. To divide countries, it is needed to use clustering algorithms for grouping them and one of the unsupervised ways to do that are using K-means clustering and Hierarchical clustering. The result gives us information to analyse Covid-19 impact on different countries and which measure should be taken to remedy the situation.

## Data Analysis

To cluster the data into different groups, K-means clustering, and hierarchical clustering is used but before clustering, data should be pre-processed and cleaned. So, the outcome would be reliable. Also, in this section data is explained in detail and the python code with explanation and original data can be found in the zip file provided.

### Understanding The Data and pre-processing and dimensionality reduction

#### Description of the Dataset

The raw dataset has 14 variables according to Covid-19 and 227 country or even ships that Covid-19 hit them. The variables recorded in the dataset can be seen in table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Country | Population | Total Cases | New Cases | Total Deaths |
| New Deaths | Total Recovered | New Recovered | Active Cases | Serious, Critical Cases |
| Tot Cases/ 1M pop | Deaths/ 1M pop | Total Tests | Tests/ 1M pop |  |

Table 1- All columns in the dataset

This data from different countries is gathered by world health organization from the data reported in countries for measuring the impacts of covid-19. Each cell of the table provides information about a specific country and one of the variables above from that country.

#### Data Cleansing

First the data need to be cleaned and to do that different task should be done. First understand the data type in each column to see any outliers

Text

Description automatically generated

Figure 1- data Types in the dataset

Population and total deaths have the wrong data type so they should be changed into numeric data type and some of them contains commas too and they should be removed. In attachment in the zip file the code for cleansing the data is provided. Next part is Finding missing values and null values, as can be seen in the figure 2, there is quite a lot of missing values in the dataset, and they need to be changed because the PCA algorithm can’t properly function with missing values.

Text

Description automatically generated

Figure 2- Null values in different columns

Because the new cases and new deaths and new recovered are not important for the whole period of two years that covid-19 has begun we can put zeros for null values. For population the null values should be dropped because if a row doesn’t have population, more than 70 percent of other values is null too. For other rows the mean value of that column is replaced because it can be considered is in the range of others and putting only zeros for those is wrong because covid impact in the best-case scenarios wasn’t good enough to put some values zero and this can lead to mis understanding about covid.

#### Data Standardization

Since K-means clustering is used in this analysis, the difference in magnitude in the dataset can be problematic. To address this issue data should be normalised and scaled so K-means algorithm run properly. One of the ways to do this is using normalizer function which will find the max of a column and put that as one and min of column as zero and then scale the values in between based on mix and min of values and put them in interval between 0 and 1. This way of normalisation is chosen because of the big difference in population of countries and in this way, we have the best outcome for PCA and clustering.

#### Feature Selection and PCA

One of the problems for performing clustering algorithms like K-means is the high number of dimensions to reduce dimensions there is two possible ways, which is implemented on this data set, first selection of the important features and after that performing principal component analysis. Feature selection is implied by importance of contribution of each feature for analysis of covid-19 impact. To do this objective because the difference between countries wants to be measured, we should choose the ones that are comparative on one million population and for those features that we don’t have these values we should create them in comparison in one million population and daily or weekly items provided as new deaths and new recovered and new cases should be dropped too. The reason is they have a lot of zeros in them because in the previous section we put zeros for nulls and also to understand the covid we do not need the new cases just for a week or a day we need to see what happened till now from the beginning. Another thing is Total recovered and total cases and active cases are created by subtraction or add of each other so we should only choose two of them. In conclusion in the table below the feature selected for PCA analysis can be seen.

|  |  |  |
| --- | --- | --- |
| Population | Tot Cases/ 1M pop | Deaths/ 1M pop |
| Tests/ 1M pop | Active Cases/ 1M Pop | Serious, Critical Cases/ 1M Pop |

Table 2- Columns selected for PCA analysis

Second part for reducing the dimensions is using principal component analysis. PCA is performed to find the optimal number of components that capture most of the variance in dataset. The PCA shows 99 percent of data of the cumulative variance data is explained with only 2 variables. Finally, our data is captured and compacted in 2 dimension which is easy for K-means to analyse and it’s easy to visualize.

Chart, histogram

Description automatically generated

Figure 3- PCA analysis outcome

### Clustering

The aim of cluster analysis is to partition a set of N object into C clusters such that objects within cluster should be like each other and objects in different clusters should be dissimilar with each other. (Srivastava, 2013)

#### K-Means Clustering

One of the unsupervised ways to cluster and divide dataset into groups is K-means clustering. For clustering the data and finding the right number of clusters several ways have been created K-means is used commonly because of its simplicity. The main idea for K-means is assign k centroids for each cluster. Next step is to assure all the point to the nearest centroids, if all the points were assigned to a centroid this part is done. Next step is finding and changing the position of the centroids to be at the centre of cluster, after that we do previous part and assign points to the nearest centroid and this iterative process happens until the number of centroids and point assign to them is found. This unsupervised way of learning for machine is K-means clustering. (Alibuhtto, 2020, p.51).

##### Number of Clusters

Before running K-means on our prepared data we need to understand the right number of clusters. This happens with using 2 methods. First elbow method that shows the right number of clusters by doing an iterative analysis on data and creating different number of clusters to find the optimal number of clusters. In the figure below we can see the outcome of elbow method and optimal number for clusters is 4.

Chart, line chart

Description automatically generated

Figure 4- Elbow method outcome

Another way to find the number of clusters is using silhouette score based on the score for each cluster we can choose the one that all of clusters has at least a minimum accepted silhouette score.

Chart

Description automatically generated

Figure 5- Silhouette score

As can be seen with four clusters we have minimum value for silhouette score of 0.7 and the clusters sizes are acceptable too.

##### Result of K-Means Clustering

The result of K-means clustering assign countries to different groups based on the Covid-19 impact on health f the population. The result for each country in detail and the code is provided in zip file.

Chart, line chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
| Clusters | Number of points | Colour |
| Cluster 1 | 120 | Red |
| Cluster 2 | 50 | Blue |
| Cluster 3 | 34 | Magenta |
| Cluster 4 | 19 | Green |

Figure 6– K-means results

From the figure above we can see the impact of different features on the outcome and importance of them on the K-means algorithm. In the Analysis section, the outcome will be discussed in detail.

#### Hierarchical Agglomerative Clustering

Another clustering function that can be used is hierarchical clustering. This way of clustering is more complex than K-means and for the large datasets this function is too slow. Agglomerative clustering, in general the merges and splits are greedy in general and hierarchical clustering is usually presented by dendrograms. (Srivastava, 2013)

##### Number of Clusters

To find out the appropriate number of clusters we need to plot the dendrograms and to do that we use ward method for linkage to create the dendrograms. The reason is this method minimises the variance of the merged clusters and the result is really like K-means clustering. After plotting the dendrogram we need to draw a horizontal line through the longest distance possible as can be seen in figure 7. The number of junctions is the number of clusters therefore there are 4 clusters.

Chart

Description automatically generated

Figure 7– Dendrograms

##### Result of Hierarchical Clustering

The result of hierarchical clustering is near to the outcome of K-means. The analysis of the data is provided in the next section. In the next figure we can see the clusters with the name of countries.

A picture containing timeline

Description automatically generated

Figure 8– Clustering with the name of countries

### Analysis of Clustering and Business Values

The clustering analysis in the basic form gives us 4 clusters which means there is a similar trend in Covid-19 impact in countries and could be grouped into 4 different groups. Also, from the figure 5 we can more insight because the displaying of the arrows. The arrows are the projection of each feature on the principal component axis. These  
arrows represent the level of importance of each feature in the multidimensional  
scaling. For example, population and total tests per million and total cases per million contribute more than other features.

One of the things that needs to be considered is the place of Iran near UK and USA. Which in my opinion it is obvious that UK and USA did a better job in comparison to Iran to deal with covid. So, this means the source of the data for countries with untrustworthy governments should not be considered as important.

Other point worth mentioning is the total tests is significant in countries and it was a good measure to control the virus. Also, total tests and population has correlation with each other and as the population grows should tests grow too in order to control covid-19.

To understand Iran situation, we need to know some points. Because of political issues Iran at first didn’t want to declare that covid-19 spread around the country in the first days of starting covid in China and they didn’t report any cases until all countries in the world started to report that their country is impacted by covid. This is one of the reasons that the outcome for Iran is not reliable. Another thing in Iran was dishonesty towards the recovered cases and they reported recovered cases in a way that the total recovered cases in part of the time was even greater than the total number of covid patients. In other countries if someone had a terminal disease and got covid and died, the country will count that as covid death but in Iran they did not report these cases and another reason for having low death rate in Iran was this reason. Iran government did not want to other countries think that they are facing a problematic issue with covid and to show its strength and don’t lose face they taught the best way is reporting not factual data. This led to even greater total rate of covid cases and deaths because people were unaware of these issues at first and they thought covid is not a big deal and that unfortunately led to death of a lot of people. Another point worth mentioning for Iran is these days they want to control people more than before and every time that people about to do a protest they put quarantine rules and accordingly they report more cases than actual one to justify the quantitation rules.

In conclusion its possible to see that covid as a disease hit people and the outcome can be grouped into 4 groups of similar records. This data is valid in most of the cases but in cases such Iran or other countries they may have reported wrong information about the impact of covid and these countries should not be included in the data because they only damage the dataset and do not give insightful ideas.

# Radio Listeners Analysis for Retail purposes

In this part we want to talk about the second task of the coursework 1. In this part we have a data set that contains data about radio listeners from all over the world and the artists that they like to listen.

In this part the goal is to find the frequent artists and groups that people listen to and after finding the frequent item set with the minimum support provided create baskets to offer to listeners for marketing purposes this objective can be done with the help of algorithms like Apriori that perform market basket analysis.

Market basket analysis is a modelling technique that is used by retailers to increase sales by better understanding customer purchasing patterns. It is based on this assumption that if a person buys a certain set of items, they are likely to buy another set of items. In order to this analysis, we need the history of the purchased items (SearchCustomerExperience. 2022).

To perform MBA and understand the pattern of artist that people buy their music Apriori algorithm is used.

In the next section first, we understand the data and pre-process and cleanse it before the Apriori and then find out the frequent items and after that we create rules based on frequent itemset.

## Data Analysis

To perform MBA first we need to understand the data in order to do that pandas profiling is used and other codes to do the first part of pre-processing.

### Understanding The Data and Pre-processing

The raw data set contains four variables and 289955 observations. The first value gives us the user id and the next one gives artist that they listen and the other one is the listeners sec and the fourth is the country of the listener. We have 1004 distinct artist and 15000 distinct users. About the 70 percent of the listeners are male and 30 percent is female. The most transaction contain four artist and they are Radiohead and the Beatles and Coldplay and red-hot chili peppers. The large amount of people is belonged to United States and Unitec Kingdom and Germany.

Graphical user interface, application

Description automatically generated

Figure 9– Frequent countries

Graphical user interface, application

Description automatically generated

Figure 10– Frequent artists



Figure 11– First rows of the data

The data type of the 4 variables is listed below in fig 12.

Text

Description automatically generated

Figure 12– Data types

#### Missing Values and Duplicates

In the data as can be seen in figure 10 we don’t have any missing values nut on the other hand we have two duplicates, and we need to remove them before analysis.

Graphical user interface, application

Description automatically generated

Figure 13– Duplicates

## Market Basket Analysis

### Change Format of the Data

To do Apriori method on our data we need to change our dataset from only transaction to one hot-encoded table or create a list for each user because Apriori algorithm can work for these inputs. The one hot-encoded table has all the artist as columns and all the users as rows and if a user listens to a artist we put one for that value and for others that he doesn’t listen we put zero and this is done for all the users.

A screenshot of a computer

Description automatically generated with low confidence

Figure 13–Hot-encoded table

### Apriori Algorithm

The next part is performing the Apriori algorithm on the hot-encoded table. Apriori is the most established algorithm for finding frequent items. The basic principle that Apriori uses is that any subset of a frequent itemset must be frequent and based on these frequent itemset we create association rules. In order to find the frequent itemset we need to run the Apriori with minimum support of 0.08 which the outcome is not good enough for creating rules so as Professor suggested we change it to 0.03 in order to find more frequent items and also get not only single frequent items but also has more than one value, in our case we only get frequent items with length of two.

Text

Description automatically generated

Figure 14–Hot-encoded table

Then we need to create our rules for that we put the minimum threshold for confidence at 40 percent level and order our rules by their lift. In this order having these two together and ordering them gives us the three top rules at the above of our association rule.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 15–Top three rules

### Business insights

With the selected parameters for minimum support and confidence level Apriori created 7 rules and we can see top three in the fig 15. We could interpret this rule in the way if listener heard antecedent then consequent is predicated that listener listens to it by the level of support and confidence and lift in the table.

If we change the minimum support and confidence, we can have more rules but with these rules we can suggest that the radio stations should give advertisement based of rules created for example lets only consider top three rules. First rule implies that if someone listens to led zeppelin its probably want to listen to pink Floyd too and we should suggest this item for those who listen to led zeppelin and for the other two rules is the same of the first rule that we explained.

# References

Alibuhtto, M. and Mahat, N., 2020. Distance based k-means clustering algorithm for determining number of clusters for high dimensional data. *Decision Science Letters*, pp.51-58.

Srivastava, K., Shah, R., Valia, D. and Swaminarayan, H., 2013. Data Mining Using Hierarchical Agglomerative Clustering Algorithm in Distributed Cloud Computing Environment. *International Journal of Computer Theory and Engineering*, pp.520-522.

SearchCustomerExperience. 2022. *What is market basket analysis? Definition from WhatIs.com*. [online] Available at: <https://www.techtarget.com/searchcustomerexperience/definition/market-basket-analysis#:~:text=Market%20basket%20analysis%20is%20a,likely%20to%20be%20purchased%20together.> [Accessed 23 March 2022].

# Appendix

1. Original data set for covid-19:



1. Dataset before normalisation and after reduction of Unwanted columns



1. The Whole code for task 1: In the code the detailed explanation of what is happening is created and I didn’t want to bother you with extra details here about different parts of the code that I did.



1. The Data set for radio listeners



1. The Data profiling report



1. The Hot-encoded table



1. The Whole code for second task: In the code the detailed explanation of what is happening is created and I didn’t want to bother you with extra details here about different parts of the code that I did.



1. Frequent Items in the second task



1. Association Rules for second task

