Data Analysis on Diarrhea

Contents

Introduction 2

Data analysis and Visualisation 2

Selection of Data Mining Algorithm 7

Data Pre-processing 8

Data Mining 8

Data Ethics 11

Conclusion 11

Introduction

Diarrhoea is a condition whereby someone has abnormally loose or watery stools. It most times lasts for a few days and can cause dehydration because of fluid loss from the body. It can increase to low urination, fast heart rate, loss of skin colour and a decline in responsiveness as one condition starts to deteriorate.

This disease is most times caused by an infection from either a parasite called gastroenteritis, a virus, or bacteria. It is gotten from water or food that is contaminated by excrement or from an infected person. It is said by scientists that diarrhoea can be prevented by good and improved sanitation, clean drinking water and regular handwashing. Babies and young ones alike are not exempt from Diarrhoea. In fact, young children get diarrhoea on average three times a year. It is the second most common cause of death in children younger than the age of five. It is estimated that about 1.7 to 5 billion cases of diarrhoea are recorded per year, and it is most common in developing countries. These developing countries could be referred to as countries with a low developed industrial base and a low human development index relative to other countries. Although the total deaths due to diarrhoea has reduced significantly by an estimated 1.32m from 1990 till 2013, it is still important to understand key trends and patterns in order to identify reasons as to why the decrease and what can be further done to reduce and better still, eliminate it. Therefore, I have chosen to analyse data on diarrhoea for this coursework.

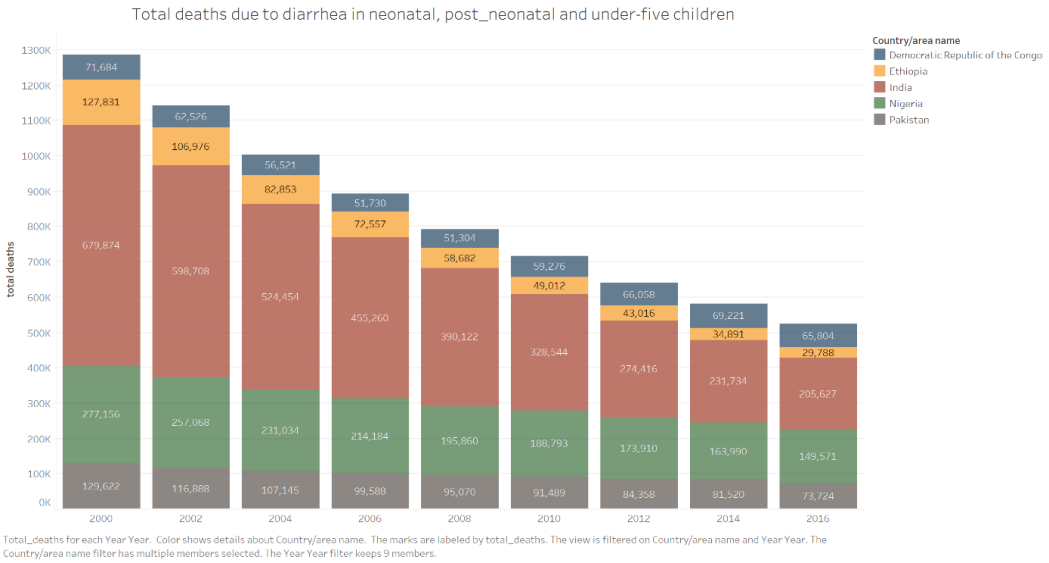
This data set used is collected from the United Nations Inter-Agency Group for Child Mortality Estimation (UN IGME). They produce estimates of child and young adolescent mortality annually, reconciling the differences across data sources and considering the systematic biases associated with the various types of data on child and adolescent mortality. The data set presents the UN IGME's latest estimates through 2018 of neonatal, infant and under-five mortality as well as mortality among children under the age of five (5) years. The variables available in this dataset include:

- Country/Area
- Year
- Neonatal Death Rate from Diarrhoea (Per 1000 live births)
- Post-Neonatal Death Rate from Diarrhoea (Per 1000 live births)
- Under-five Death Rate from Diarrhoea (Per 1000 live births)
- Percentage Of Neonatal Death Due to Diarrhoea
- Percentage of Post-Neonatal Death Due to Diarrhoea
- Percentage of Under-five Death Due to Diarrhoea
- Neonatal Death Due to Diarrhoea
- Post-Neonatal Death Due to Diarrhoea
- Total Neonatal Deaths
- Total Post-Neonatal Deaths
- Under-five Deaths Due to Diarrhoea

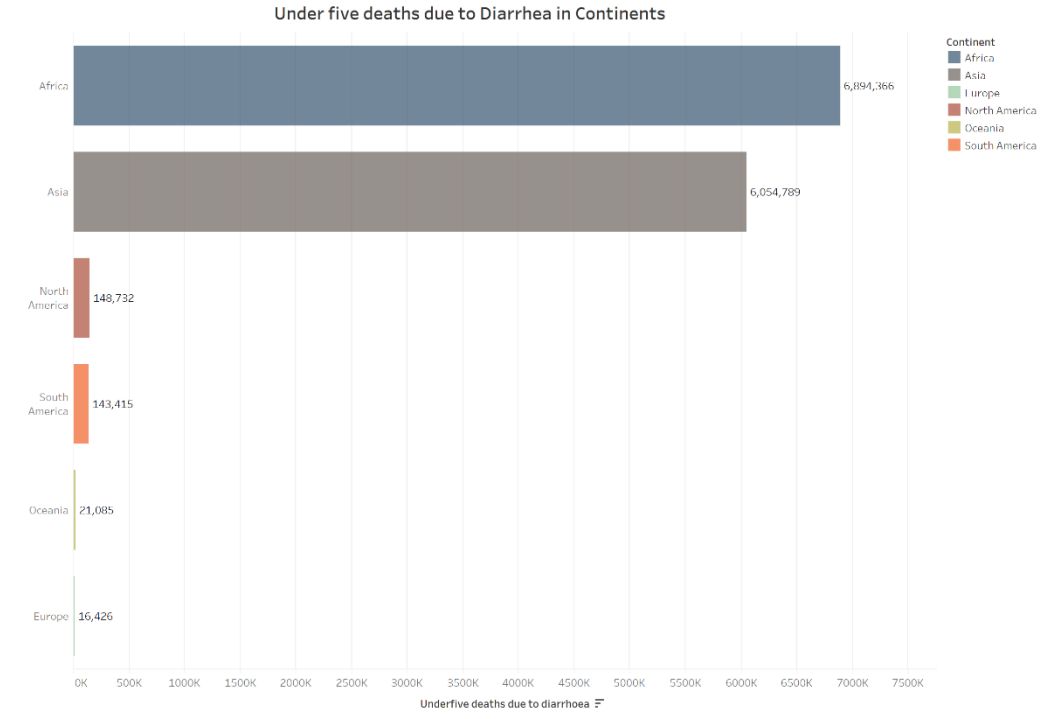
Neonatal refers to newborns during the first month after birth. i.e., 28 days of age. While Post-Neonatal refers to 29 - 364 days of age.

Data analysis and Visualisation

To be able to solve the problem of diarrhoea around the world, it is important to understand the data being gathered in order to understand trends and patterns. Data visualisation serves as the best option as a good visualisation helps tell stories

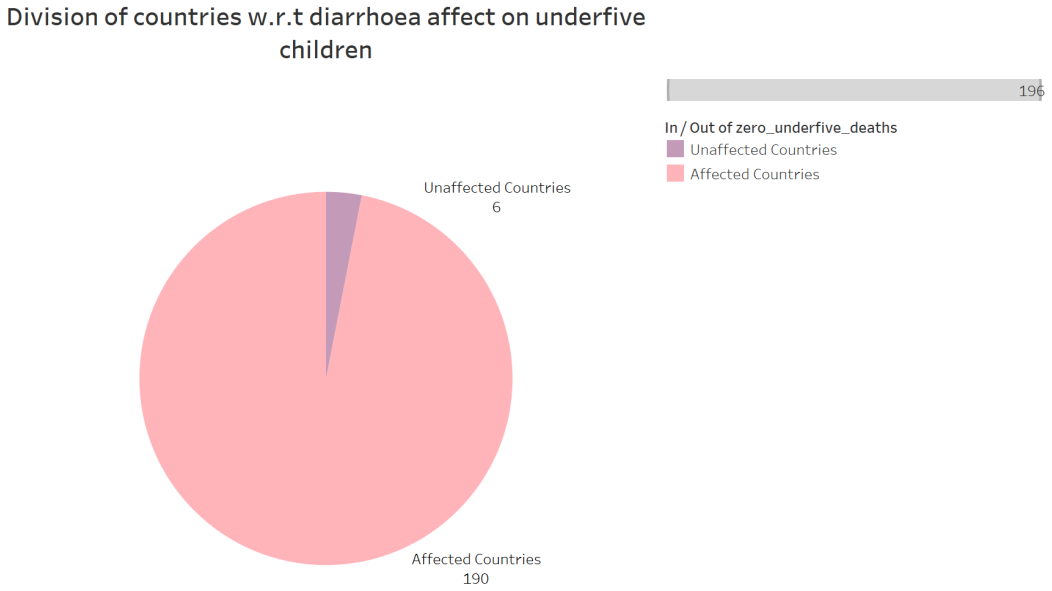


The graph above shows the total deaths due to diarrhoea in neonatal, post-neonatal and under-five children from the year 2000-2016 in the top 5 most affected countries. This visualisation shows that India recorded the highest number of deaths in the world with 679,874 as at the year 2000. It is also worth knowing that of these five countries, India is the most populated. So, it does not necessarily mean they have the highest ratio of death. Of the top five most affected countries, the Democratic Republic of the Congo recorded the lowest number of total deaths with 71,684 as at the year 2000 but fast forward to 2016, Ethiopia has the lowest number of deaths with 29,788. This shows that there has been massive improvement in health care. However, as for The Democratic Republic of Congo, there was a resurgence in total deaths from the year 2010-2014 with death rates climbing until it began regressing. All other countries displayed significant progress through the years with death rates reducing at a rapid rate.



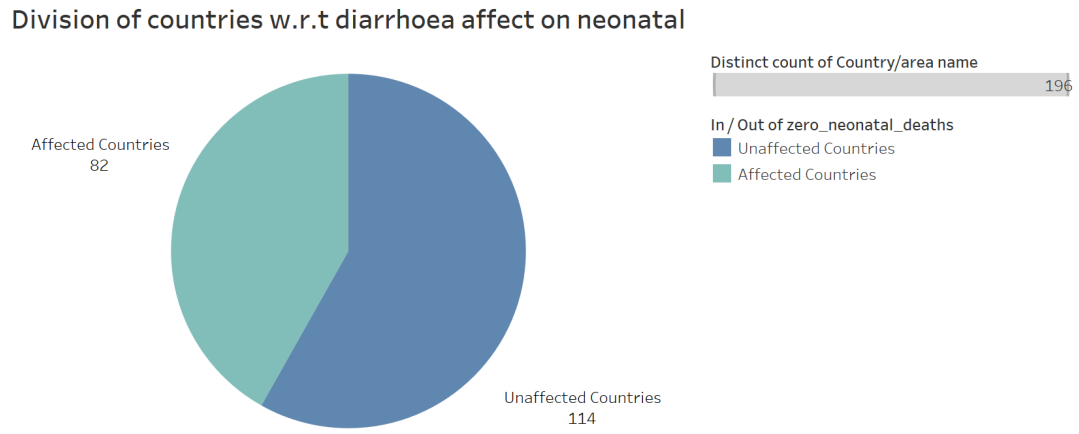
The graph above shows the total number of Under-Five deaths due to Diarrhoea in each continent. Antarctica was an exception as there were no records on it and it is also inhabitable. This visualisation shows that Africa leads as the continent with the Under-Five deaths due to Diarrhoea. They have a whopping total of 6,894,366

with Asia coming close at 6,054,789. It should not be so much of a surprise as these are the two most populated continents in the world and with more countries considered as “third world” countries. However, there is a vast difference with the remaining continents with approximately 5,000,000. Europe has the lowest number of under-five deaths due to diarrhoea.



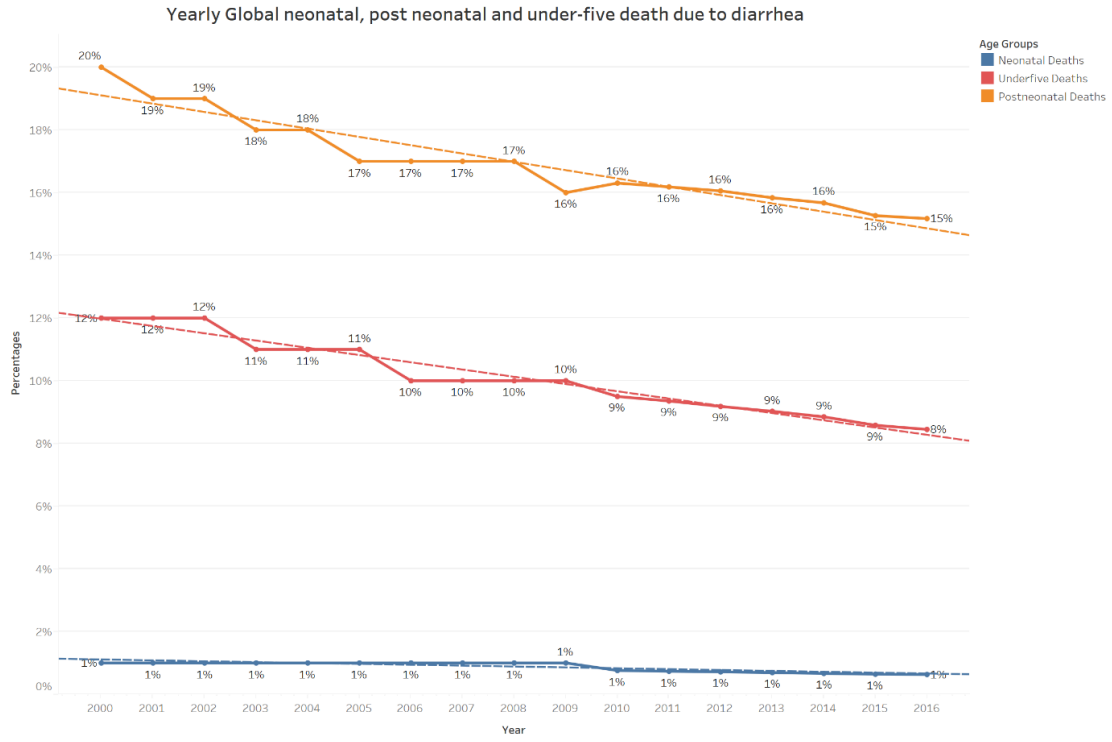
In / Out of zero_underfive_deaths and distinct count of Country/area name. Color shows details about In / Out of zero_underfive_deaths. Size shows distinct count of Country/area name. The marks are labeled by In / Out of zero_underfive_deaths and distinct count of Country/area name.

The pie-chart above shows the number of countries affected by under-five deaths due to diarrhoea. Of course, all continents were affected but not all countries were affected. Out of the 198 countries data was collected from, only six (6) countries were unaffected with the remaining 190 countries having fatalities.

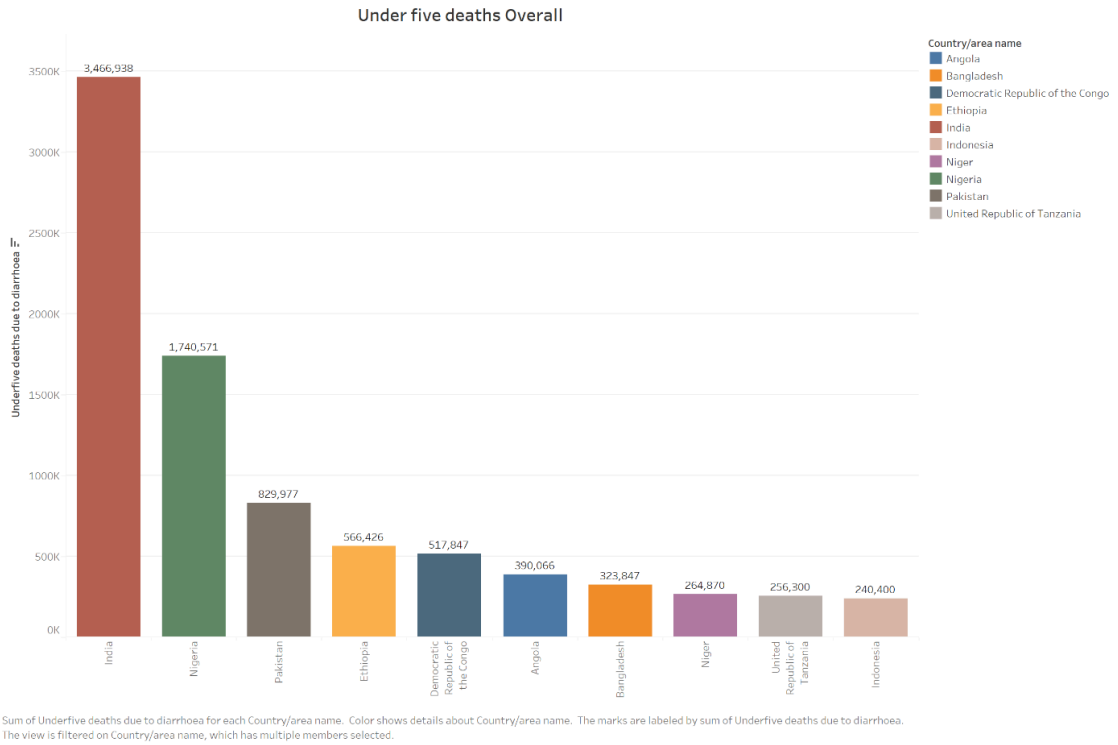


In / Out of zero_neonatal_deaths and distinct count of Country/area name. Color shows details about In / Out of zero_neonatal_deaths. Size shows distinct count of Country/area name. The marks are labeled by In / Out of zero_neonatal_deaths and distinct count of Country/area name.

Unlike under-five deaths due to diarrhoea where six (6) countries were unaffected, this pie chart above shows that up to 114 countries had no fatalities. Therefore, less countries had fatalities 82 being the exact number of countries affected.



This visualisation above shows the yearly global neonatal, post-neonatal and under-five deaths due to diarrhoea. A trend line is placed to see the moving trends and there seems to be a pattern. Since the year 2000 up till 2016, there has been a steady decline in the under-five deaths and post-neonatal deaths. We can also see that there is a very low fatality rate in the neonatal deaths around the world.



This graph shows the countries with the most Under-Five deaths due to diarrhoea. India has the highest number of fatalities with Nigeria coming close second. Again we see a pattern with countries with a very high population having the highest recorded numbers.

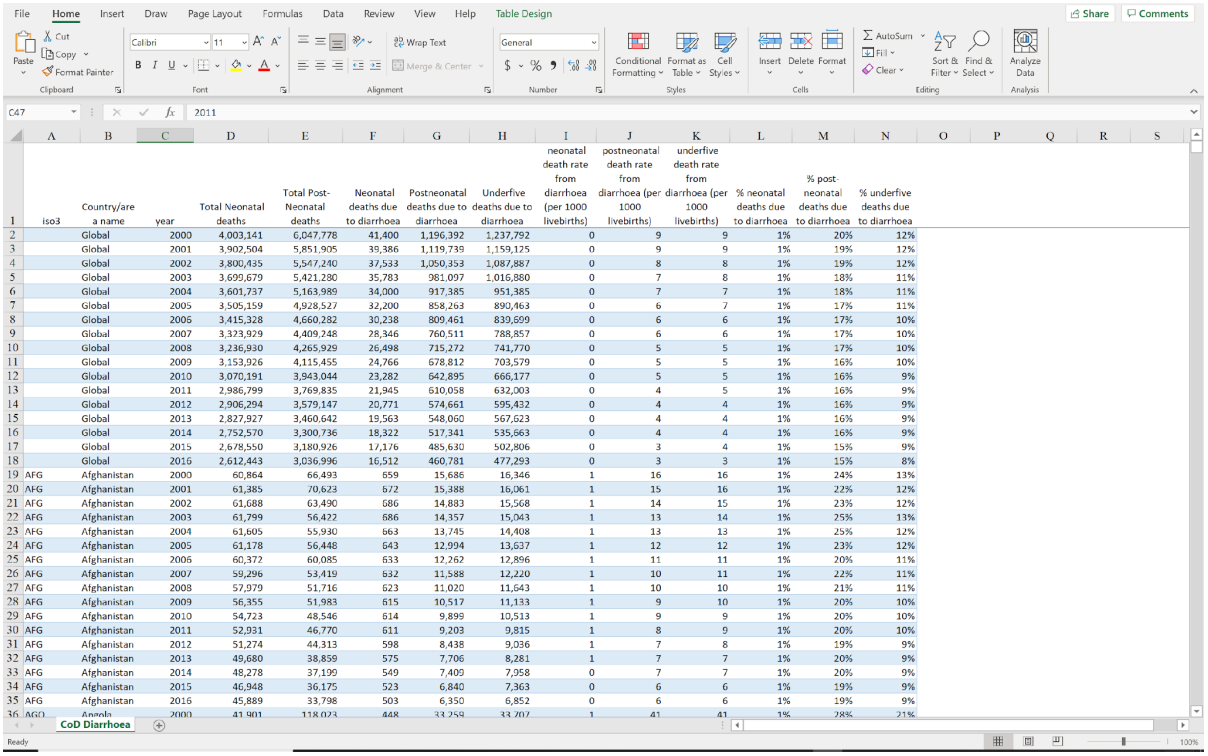
Selection of Data Mining Algorithm

I will use the k-means clustering algorithm for further analysis of the data set. The differences and similarities of a data set appear in this algorithm. K-means algorithm is an iterative algorithm which collects and groups data with multiple attributes. The groups that have data points in a cluster can be defined. The class labels for training data are unknown. This is an unavailable learning algorithm. To use this, you must define the number of clusters (k). For clusters to have centroids, random points from data need to be assigned. The clusters are designated to all points closest to each centroid. The centroids are now re-calculated by the new clusters. The last two steps are then repeated until no other data point is moved.

By using k-means, I can analyse the entire data set much better. The relationship between the different causes of death and where they are located will help me identify the affected areas. This will help to identify what geographical areas need more resources so as to combat the disease.

Data Pre-processing

The data used is Numerical. Therefore, there are no class variables as that is only applicable to a Categorical dataset.



iso3	Country/area name	year	Total Neonatal deaths	Total Post-Neonatal deaths	Neonatal deaths due to diarrhoea	Postneonatal deaths due to diarrhoea	Underfive deaths due to diarrhoea	neonatal death rate from diarrhoea (per 1000 livebirths)	postneonatal death rate from diarrhoea (per 1000 livebirths)	underfive death rate from diarrhoea (per 1000 livebirths)	% neonatal deaths due to diarrhoea	% post-neonatal deaths due to diarrhoea	% underfive deaths due to diarrhoea
	Global	2000	4,003,141	6,047,778	41,400	1,196,392	1,237,792	0	9	9	1%	20%	12%
	Global	2001	3,902,504	5,851,905	39,386	1,119,739	1,159,125	0	9	9	1%	19%	12%
	Global	2002	3,800,435	5,547,240	37,533	1,050,353	1,087,887	0	8	8	1%	19%	12%
	Global	2003	3,699,679	5,421,280	35,783	981,097	1,016,880	0	7	8	1%	18%	11%
	Global	2004	3,601,737	5,163,989	34,000	917,385	951,385	0	7	7	1%	18%	11%
	Global	2005	3,505,159	4,928,527	32,200	858,263	890,463	0	6	7	1%	17%	11%
	Global	2006	3,415,328	4,660,282	30,238	809,461	839,699	0	6	6	1%	17%	10%
	Global	2007	3,323,929	4,409,248	28,346	760,511	788,857	0	6	6	1%	17%	10%
	Global	2008	3,236,930	4,265,929	26,498	715,272	741,770	0	5	5	1%	17%	10%
	Global	2009	3,153,926	4,115,455	24,766	678,812	703,579	0	5	5	1%	16%	10%
	Global	2010	3,070,191	3,943,044	23,282	642,895	666,177	0	5	5	1%	16%	9%
	Global	2011	2,986,799	3,769,835	21,945	610,058	632,003	0	4	5	1%	16%	9%
	Global	2012	2,906,294	3,579,147	20,771	574,661	595,432	0	4	4	1%	16%	9%
	Global	2013	2,827,927	3,460,642	19,563	548,060	567,623	0	4	4	1%	16%	9%
	Global	2014	2,752,570	3,300,736	18,322	517,341	535,663	0	4	4	1%	16%	9%
	Global	2015	2,678,550	3,180,926	17,176	485,630	502,806	0	3	4	1%	15%	9%
	Global	2016	2,612,443	3,036,996	16,512	460,781	477,293	0	3	3	1%	15%	8%
AFG	Afghanistan	2000	60,864	66,493	659	15,686	16,346	1	16	16	1%	24%	13%
AFG	Afghanistan	2001	61,385	70,623	672	15,388	16,061	1	15	16	1%	22%	12%
AFG	Afghanistan	2002	61,688	63,490	686	14,883	15,568	1	14	15	1%	23%	12%
AFG	Afghanistan	2003	61,799	56,422	686	14,357	15,043	1	13	14	1%	25%	13%
AFG	Afghanistan	2004	61,605	55,930	663	13,745	14,408	1	13	13	1%	25%	12%
AFG	Afghanistan	2005	61,178	56,448	643	12,994	13,637	1	12	12	1%	23%	12%
AFG	Afghanistan	2006	60,372	60,085	633	12,262	12,896	1	11	11	1%	20%	11%
AFG	Afghanistan	2007	59,296	53,419	632	11,588	12,220	1	10	11	1%	22%	11%
AFG	Afghanistan	2008	57,979	51,716	623	11,020	11,643	1	10	10	1%	21%	11%
AFG	Afghanistan	2009	56,355	51,983	615	10,517	11,133	1	9	10	1%	20%	10%
AFG	Afghanistan	2010	54,723	48,546	614	9,899	10,513	1	9	9	1%	20%	10%
AFG	Afghanistan	2011	52,931	46,770	611	9,203	9,815	1	8	9	1%	20%	10%
AFG	Afghanistan	2012	51,274	44,313	598	8,438	9,036	1	7	8	1%	19%	9%
AFG	Afghanistan	2013	49,680	38,859	575	7,706	8,281	1	7	7	1%	20%	9%
AFG	Afghanistan	2014	48,278	37,199	549	7,409	7,958	0	7	7	1%	20%	9%
AFG	Afghanistan	2015	46,948	36,175	523	6,840	7,363	0	6	6	1%	19%	9%
AFG	Afghanistan	2016	45,889	33,798	503	6,350	6,852	0	6	6	1%	19%	9%
AGO	Angola	2000	41,901	118,023	448	33,259	33,707	1	41	41	1%	28%	21%

The image above shows the original dataset which was yet to be cleaned.

2011

D	E	F	G	H
Total Neonatal deaths	Total Post-Neonatal deaths	Neonatal deaths due to diarrhoea	Postneonatal deaths due to diarrhoea	Underfive deaths due to diarrhoea
4,003,141	6,047,778	41,400	1,196,392	1,237,792
3,902,504	5,851,905	39,386	1,119,739	1,159,125
3,800,435	5,547,240	37,533	1,050,353	1,087,887
3,699,679	5,421,280	35,783	981,097	1,016,880
3,601,737	5,163,989	34,000	917,385	951,385
3,505,159	4,928,527	32,200	858,263	890,463
3,415,328	4,660,282	30,238	809,461	839,699
3,323,929	4,409,248	28,346	760,511	788,857
3,236,930	4,265,929	26,498	715,272	741,770
3,153,926	4,115,455	24,766	678,812	703,579
3,070,191	3,943,044	23,282	642,895	666,177
2,986,799	3,769,835	21,945	610,058	632,003
2,906,294	3,579,147	20,771	574,661	595,432
2,827,927	3,460,642	19,563	548,060	567,623
2,752,570	3,300,736	18,322	517,341	535,663
2,678,550	3,180,926	17,176	485,630	502,806
2,612,443	3,036,996	16,512	460,781	477,293
60,864	66,493	659	15,686	16,346
61,385	70,623	672	15,388	16,061
61,688	63,490	686	14,883	15,568
61,799	56,422	686	14,357	15,043
61,605	55,930	663	13,745	14,408
61,178	56,448	643	12,994	13,637
60,372	60,085	633	12,262	12,896
59,296	53,419	632	11,588	12,220
57,979	51,716	623	11,020	11,643
56,355	51,983	615	10,517	11,133
54,723	48,546	614	9,899	10,513
52,931	46,770	611	9,203	9,815
51,274	44,313	598	8,438	9,036
49,680	38,859	575	7,706	8,281
48,278	37,199	549	7,409	7,958
46,948	36,175	523	6,840	7,363
45,889	33,798	503	6,350	6,852
41,901	118,023	448	33,259	33,707

As seen below, the column “iso3” does not have a complete date. Some cells are empty which are meant to be filled with a value.



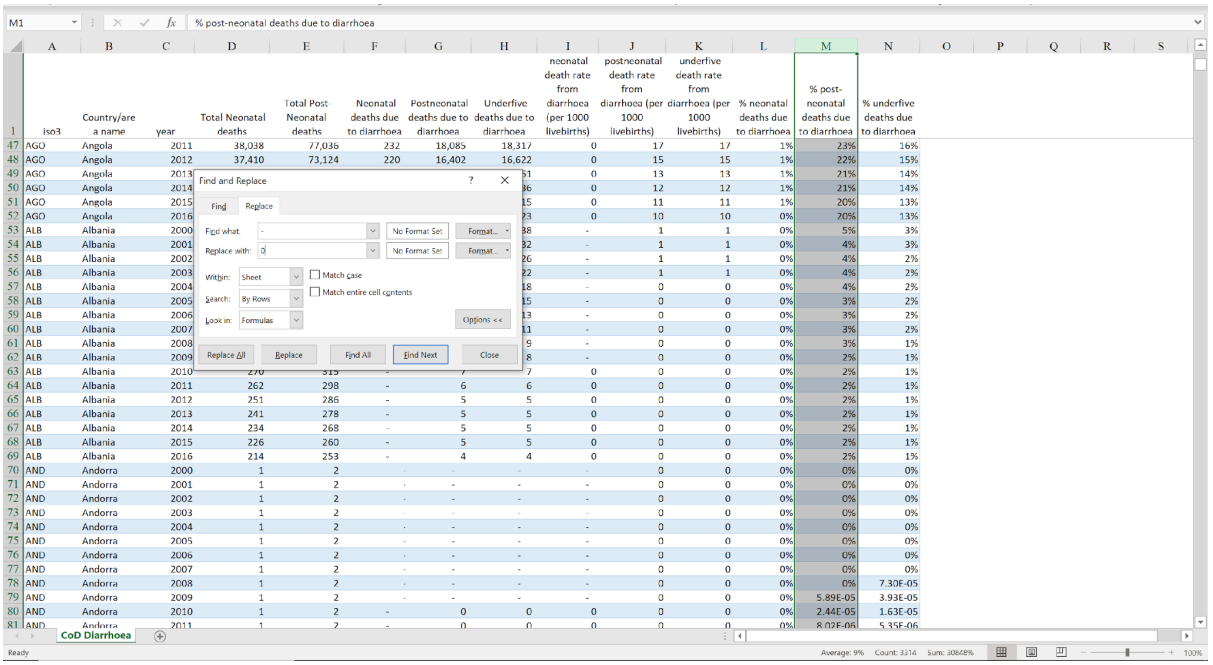
	iso3	Country/area name	year
2		Global	2000
3		Global	2001
4		Global	2002
5		Global	2003
6		Global	2004
7		Global	2005
8		Global	2006
9		Global	2007
10		Global	2008
11		Global	2009
12		Global	2010
13		Global	2011
14		Global	2012
15		Global	2013
16		Global	2014
17		Global	2015
18		Global	2016
19	AFG	Afghanistan	2000
20	AFG	Afghanistan	2001
21	AFG	Afghanistan	2002
22	AFG	Afghanistan	2003
23	AFG	Afghanistan	2004
24	AFG	Afghanistan	2005
25	AFG	Afghanistan	2006
26	AFG	Afghanistan	2007
27	AFG	Afghanistan	2008
28	AFG	Afghanistan	2009
29	AFG	Afghanistan	2010
30	AFG	Afghanistan	2011
31	AFG	Afghanistan	2012
32	AFG	Afghanistan	2013
33	AFG	Afghanistan	2014
34	AFG	Afghanistan	2015
35	AFG	Afghanistan	2016
36	AGO	Angola	2000

So I replaced the empty column with “GLO”.



	iso3	Country/ar	year
2	GLO	Global	2000
3	GLO	Global	2001
4	GLO	Global	2002
5	GLO	Global	2003
6	GLO	Global	2004
7	GLO	Global	2005
8	GLO	Global	2006
9	GLO	Global	2007
10	GLO	Global	2008
11	GLO	Global	2009
12	GLO	Global	2010
13	GLO	Global	2011
14	GLO	Global	2012
15	GLO	Global	2013
16	GLO	Global	2014
17	GLO	Global	2015
18	GLO	Global	2016
19	AFG	Afghanistar	2000
20	AFG	Afghanistar	2001
21	AFG	Afghanistar	2002
22	AFG	Afghanistar	2003
23	AFG	Afghanistar	2004
24	AFG	Afghanistar	2005
25	AFG	Afghanistar	2006
26	AFG	Afghanistar	2007
27	AFG	Afghanistar	2008
28	AFG	Afghanistar	2009
29	AFG	Afghanistar	2010
30	AFG	Afghanistar	2011
31	AFG	Afghanistar	2012
32	AFG	Afghanistar	2013
33	AFG	Afghanistar	2014
34	AFG	Afghanistar	2015
35	AFG	Afghanistar	2016
36	AGO	Angola	2000
37	AGO	Angola	2001
38	AGO	Angola	2002
39	AGO	Angola	2003
40	AGO	Angola	2004

There are some occurrences of noisy data like "-" which were identified and replaced with "0"using the find and replace feature of Microsoft Excel. The image below shows it.



Data Mining

For the first iteration, I declared k to be 3 (3 clusters). The centroids of the clusters returned some interesting information. The first centroid was Nigeria, in cluster 0, revealing that the countries in this cluster are countries with a high death rate for neonatal children. The second cluster has Indonesia as a centroid. The death rates are much lower compared to countries in cluster 0 and 2. Countries in this cluster are the lowest death rates. The countries in the third cluster, cluster 2, had Colombia as its centroid. The death rates are much lower than in the first cluster but are higher than in the second. The difference between the second and third cluster is not that much but the difference between cluster 2 and cluster 0 is big.

		Cluster#		
Attribute	Full Data	0	1	2
	(157.0)	(21.0)	(71.0)	(65.0)
iso	COL	NGA	IDN	COL
country	Colombia	Nigeria	Indonesia	Colombia
year	2008.0127	2008.6667	2011.3521	2004.1538
TotalNeonantalDeath	128819.1911	565837.8095	53720.5775	69659.3538
TotalPostNeonantalDeaths	150050.2038	694326.9048	59016.0845	73644.2308
NeonatalDeathsDueToDiarrhoae	1166.3631	6150.2381	291.0845	512.2615
PostNeonatalDueToDiarrhoae	27093.5924	138223.4286	8074.8451	11964.4308
underfiveDeathsDueToDiarrhae	28259.9745	144373.5714	8365.9718	12476.7231

Cluster centroids for the first iteration.

Because I was unable to point out why some countries were in clusters 1 and 2, I decided to increase the number of clusters to 4. All the cluster centroids stayed the same for the clusters 0, 1 and 3 as in the first iteration but for one cluster, cluster 2. The cluster has Sudan as the centroid which had the lowest death rates. With the additional cluster, I can understand that some countries are able to have lower death rates by investing in infrastructure such as hospitals and schools to train doctors and nurses like Indonesia in this dataset, and other countries have lower death rates due to having low population like Sudan. When there is a

combination of high population and low investment, you can have cases like in Nigeria where the death rates are a lot higher.

		Cluster#			
Attribute	Full Data	0	1	2	3
	(157.0)	(21.0)	(47.0)	(34.0)	(55.0)
iso	COL	NGA	IDN	SDN	COL
country	Colombia	Nigeria	Indonesia	Sudan	Colombia
year	2008.0127	2008.6667	2011.9149	2008.8824	2003.8909
TotalNeonantalDeath	128819.1911	565837.8095	56797.234	53114.4706	70302.8545
TotalPostNeonantalDeaths	150050.2038	694326.9048	61385.0638	60761.8235	73200.3091
NeonatalDeathsDueToDiarrhoae	1166.3631	6150.2381	285.7234	352.2059	519.2727
PostNeonatalDueToDiarrhoae	27093.5924	138223.4286	8028.1489	9592.5	11773.3455
underfiveDeathsDueToDiarrhae	28259.9745	144373.5714	8313.8298	9944.9118	12292.6182

Cluster centroids for the second iteration.

For the third iteration, I decided to leave only post neonatal deaths and under-fives deaths. Removing these attributes will allow me to analyse how good after care is for neonatal children for the different countries. After running the algorithm, there was no change in the results of the clusters and the centroids. This shows that the relationship between neonatal and post neonatal deaths are close. This shows that the higher the child death rates are, the more babies die, and the more neonatal babies die which also means more die due to diarrhoea.

		Cluster#			
Attribute	Full Data	0	1	2	3
	(157.0)	(21.0)	(47.0)	(34.0)	(55.0)
iso	COL	NGA	IDN	SDN	COL
country	Colombia	Nigeria	Indonesia	Sudan	Colombia
year	2008.0127	2008.6667	2011.9149	2008.8824	2003.8909
TotalNeonantalDeath	128819.1911	565837.8095	56797.234	53114.4706	70302.8545
TotalPostNeonantalDeaths	150050.2038	694326.9048	61385.0638	60761.8235	73200.3091
NeonatalDeathsDueToDiarrhoae	1166.3631	6150.2381	285.7234	352.2059	519.2727
PostNeonatalDueToDiarrhoae	27093.5924	138223.4286	8028.1489	9592.5	11773.3455
underfiveDeathsDueToDiarrhae	28259.9745	144373.5714	8313.8298	9944.9118	12292.6182

Cluster centroids for the third iteration.

The best iteration can be the second one because the additional cluster made it easier to understand and the changes made to the data for the third iteration did not change much. The data model fits very well as it puts countries with similar death rates into clusters. One problem is that it is not able to differentiate between countries that have a higher or lower population which affects death rates. This might just be because it was not included in the data. The sum of squared errors was low and tolerable considering the size of the dataset. There is some difference in percentage in each cluster from the training set and the percentage split set but the type of countries in them are very similar.

```
======

Number of iterations: 5
Within cluster sum of squared errors: 233.98126405714555
```

Sum of squared errors for the second iteration.

```
Clustered Instances

0       17 (  7%)
1       97 ( 41%)
2      107 ( 45%)
3       17 (  7%)
```

Clustered instances for the training set.

```
Clustered Instances

0       13 ( 16%)
1       24 ( 30%)
2       19 ( 23%)
3       25 ( 31%)
```

Clustered instances for the second iteration.

		Cluster#			
Attribute	Full Data	0	1	2	3
	(238.0)	(17.0)	(97.0)	(107.0)	(17.0)
iso	NGA	IND	RUS	BRA	NGA
country	Nigeria	India	RussianFederation	Brazil	Nigeria
year	2008	2008	2011.7113	2004.6355	2008
TotalNeonantalDeath	133842.8697	934587.5882	43191.3608	69476.0374	255477.4118
TotalPostNeonantalDeaths	156245.8277	837266.7647	44363.1031	77365.4953	610096.6471
NeonatalDeathsDueToDiarrhoae	1248.5084	10881.5882	230.7835	459.1215	2390.9412
PostNeonatalDueToDiarrhoae	28759.2521	193055.9412	5562	12367.3925	99995.6471
underfiveDeathsDueToDiarrhae	30007.7857	203937.4706	5792.8041	12826.5701	102386.5294

Cluster centroids for the training set.

For example, the first cluster has countries that have very high death rates which is the same on the percentage split set, followed by cluster 3. The difference between cluster 1 and 2 is very little which is also the same as in the percentage split set.

Data Ethics

There are numerous problems with data analysis related to data collection, use and storage. These problems can be classified into ethical, juridical, and professional questions. More companies collect and use data because of the importance of data analytics in companies growing. Most of the data is gathered online and it is important to use only information necessary and secure when users leave a digital footprint while using the internet.

Confidentiality is an ethical issue to consider. Individuals are entitled to retain private information. They are free of unwanted attention and have a right to be anonymous. The concern is that if private information is gathered, Individuals that may not be intended may be identified. Also, the security of private information is a problem when it is released as name, address, and political opinions. When collecting the data, companies and organisations must also clarify for what they use the data. Transparency must be given to what is happening and if something is wrong.

Legally, the rapid development of data technology and technology is problematic. In order to address these legal problems, the GDPR (General Data Protection Report) has been introduced in the EU. The principles covered are accuracy, limitation of storage, the completeness and secrecy of the data. It also covers limiting purposes and minimising data, as well as the legitimate, fair and transparent use of the data.

There are professional issues in data analysis involving professionals. These professionals must maintain the privacy of personal data. They must abide by fundamental principles of data protection and inform users of the risks to privacy. In addition, users should be able to customise confidentiality settings, safety levels and acceptable recording, monitoring, and tracking levels.

Conclusion

From the data set used, there shows that there has been a great improvement over the years with the attention to diarrhoea. There has been an overall steady decline in the death rates due to diarrhoea. Countries such as Nigeria and India have the highest numbers but compared to where they were in the year 2000, there seems to be very significant progress in curbing the disease.

The analysis shows that countries with a larger population have higher death rates. However, countries that have more resources or are considered as developed countries have low death rates. Therefore, resources from those developed countries can be used to reduce mortality in heavily affected areas.