An Analysis of Word Development Indicators and COVID-19 Deaths

DATA COURTESY OF WORLD BANK

Shehzadi Aziz | MA335 Final Project| March 2022| Word Count: 2483

Abstract:

This report will analyze data from the World Bank about Coronavirus casualties in 185 countries around the world, combined with data about each country's socio-economic position (based on the institution's World Development Indicators). We will firstly show that countries like Peru and Hungary counted the highest number of deaths due to coronavirus, and that the continent with the most casualties in this dataset is Europe. The report will then cluster countries in the dataset based on the observations regarding world development indicators (WDI), like life expectancy and GDP per capita – showing that there is small link between how these countries are clustered and what continent they are in. Countries in different continents are still often clustered together based on socio-economic factors. Then, we will examine a logistic regression model to show that after controlling collinearity, world development indicators like 'continent', 'population growth', 'unemployment' and 'Election Access' are the most statistically significant predictors of high COVID casualties ('high' meaning a death toll above the median number of deaths). Three different classification (QDA, LDA, Logistic regression) models can be made for predicting the range of casualties a country may suffer from based on their socio-economic circumstances. The classification models are successfully able to classify most countries in terms of COVID-19 casualties based on selected World Development Indicators from a random forests algorithm. There remain limitations in the data itself due to under-ascertainment and underestimating of casualties in some countries. Hence, the findings still cannot necessarily be applied to casualties of similar pandemics.

Table of Contents

Introduction

The report will begin with a preliminary analysis running some basic descriptive statistics. Clustering algorithms on the countries based on the world development indicators will then be implemented. We will also use a logistic regression model and classification algorithms to predict covid-19 casualties.

PRELIMINARY ANALYSIS

Table 1: A table of summary statistics for each variable (World Development Indicator) for each country.

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Covid.deaths	185	1143.378	1219.881	3	167	1830	6252
Life.expec	177	73.013	7.489	54.239	67.273	78.498	85.078
Elect.access	185	85.731	24.996	6.721	83.5	100	100
Net.nat.income	133	4.281	6.106	-14.379	1.305	6.27	50.172
Net.nat.income.capita	133	2.931	5.879	-17.347	0.53	5.076	47.252
Mortality.rate	172	20.456	19.388	1.6	5.25	31.625	82.4
Primary	115	92.899	12.963	54.729	86.198	101.17	120.447
Pop.growth	184	1.241	1.124	-1.61	0.427	1.988	4.469
Pop.density	185	377.319	1634.906	2.071	35.893	217.008	19223.976
Pop.total	184	41175256.799	149264099.344	33706	2116046.5	30378996.75	1407745000
Health.exp.capita	167	1196.708	1915.711	19.85	73.803	1254.541	10921.013
Health.exp	167	6.432	2.59	1.525	4.398	8.102	16.767
Unemployment	112	7.349	5.238	0.1	3.787	9.952	28.47
GDP.growth	175	2.919	3.108	-7.157	1.283	4.89	19.536
GDP.capita	176	18851.173	28892.493	228.214	2072.459	23347.393	189487.147
Birth.rate	180	19.394	9.888	5.9	10.5	27.135	45.637
Water.services	108	73.6	29.486	5.581	55.386	98.731	100

Table 1 shows us the summary statistics for each numerical variable in the World Development Indicators dataset. From the table, we can see that the mean number of recorded deaths from coronavirus (shown in the first row saying 'Covid.deaths') is 1143 to the nearest significant figure. Meanwhile, the 25th percentile (or lower quartile, denoted by the column Pctl. 25) is 167 deaths and the 75th percentile (or upper quartile, denoted by Pctl. 75) is 1830 deaths. The upper quartile will be useful later when we define the threshold for high covid-19 casualties.

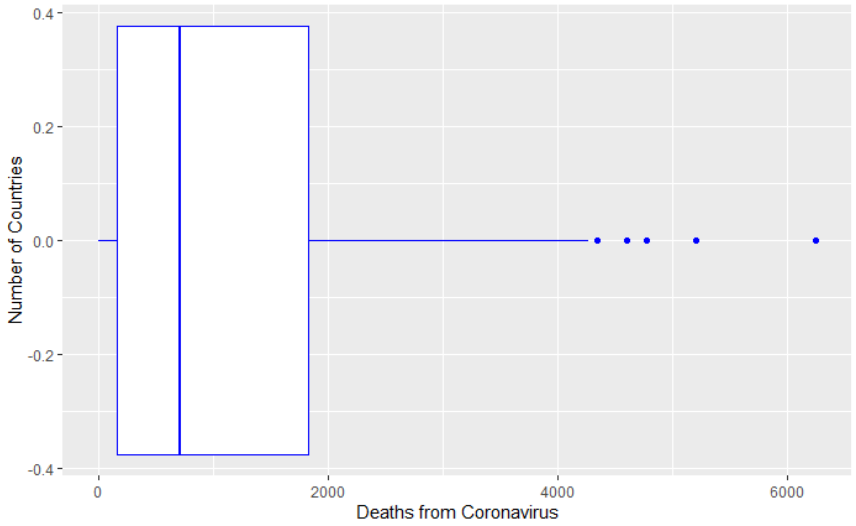


Figure 1: Boxplot showing the distribution of data on COVID-19 deaths.

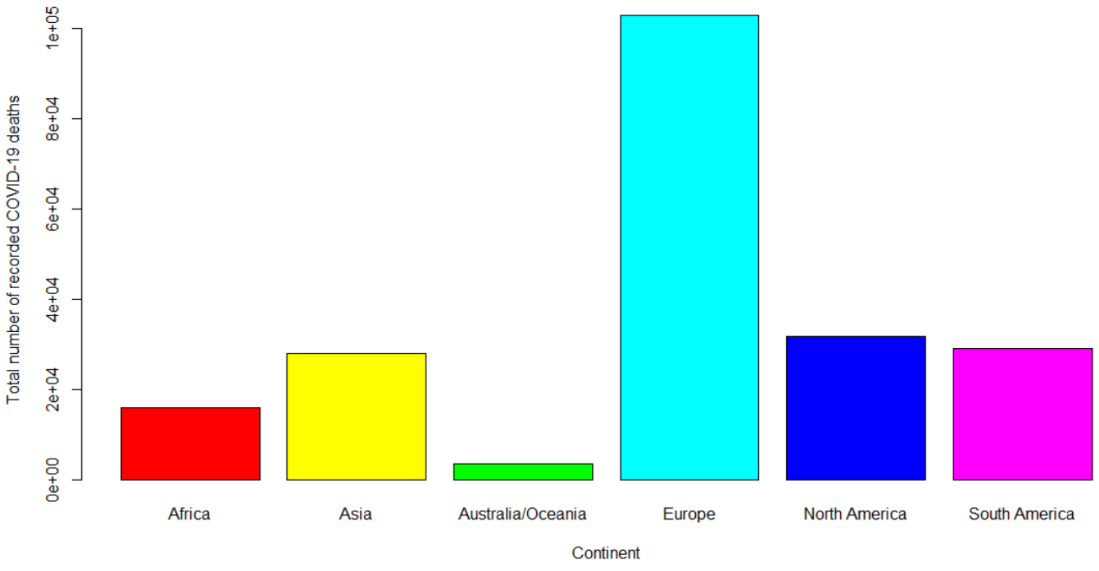
Figure 1 shows the minimum values are close to 0, suggesting that some countries had a very low number of recorded deaths related to coronavirus. The boxplot shows us that there are 5 'outlier' countries with recorded Covid-19 deaths over 4000 – with the most extreme outlier being a country that reported over 6000 deaths.

	ï..Country.Name <chr>	Covid.deaths <dbl>
136	Peru	6252
28	Bulgaria	5205
24	Bosnia and Herzegovina	4775
79	Hungary	4604
129	North Macedonia	4344

Table 2: Countries with the highest number of COVID-19 deaths in our data.

Table 2 identifies the outliers in the boxplot, showing us that the country who reported the highest number of deaths related to Covid-19 is Peru, followed by Bulgaria, Bosnia and Herzegovina, Hungry and North Macedonia. Table 2 suggests that Europe may be the continents with the highest number of covid-19 casualties. This is shown to be true in figure 2 (shown below). Figure 2 suggests that overall, the data reports Europe to have had over 100,000 deaths due to coronavirus.

Figure 2: Shows data on COVID-19 casualties by continent.



ANALYSIS

Countries will be clustered based on the data about the World Bank's world development indicators. Differences between observations for each country are displayed in the distance matrix. The dimensions of the data were reduced to include only 50 countries from the data instead of the full sample size of 185 in the figure below. Figure 3 below shows the

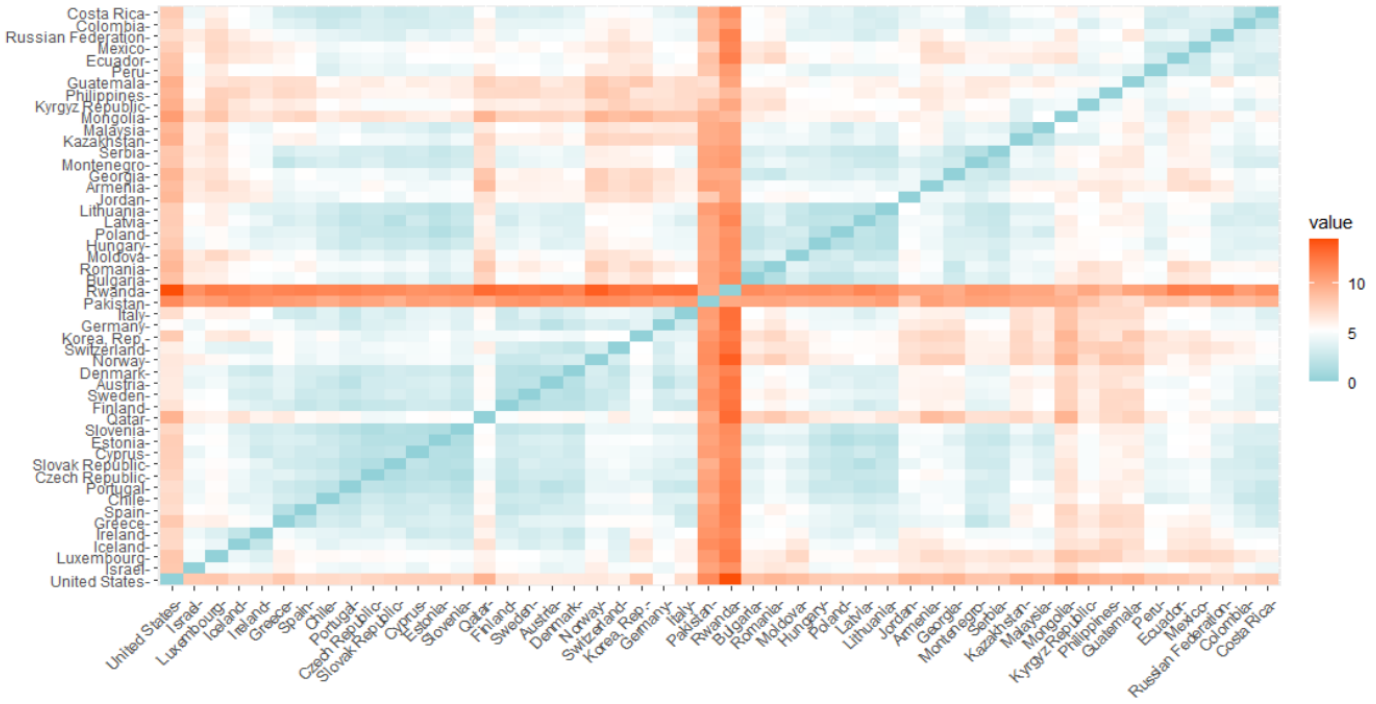


Figure 3: A distance matrix showing the differences between data for each country. The distances run on a spectrum between 0 and 10. Dark orange/red indicates a large distance between the observations for two countries, close to 10, white indicating a medium distance close to 5, and light blue indicating no distance between observations for two countries.

distance matrix indicating which countries are likely to be clustered with or away from each other. Distances are calculated with Euclidean distance metrics ($d = \sqrt{[(x2 - x1)^2 + (y2 - y1)^2]}$). The matrix suggests that Pakistan and Rwanda are going to be in the same cluster when k-means clustering is applied, but not in the same cluster as many, if not all, of the countries in this data sample. On the other hand, countries like Slovenia, Slovak Republic, Cyprus, Spain and Greece – all in Europe - are likely to be clustered with each other as

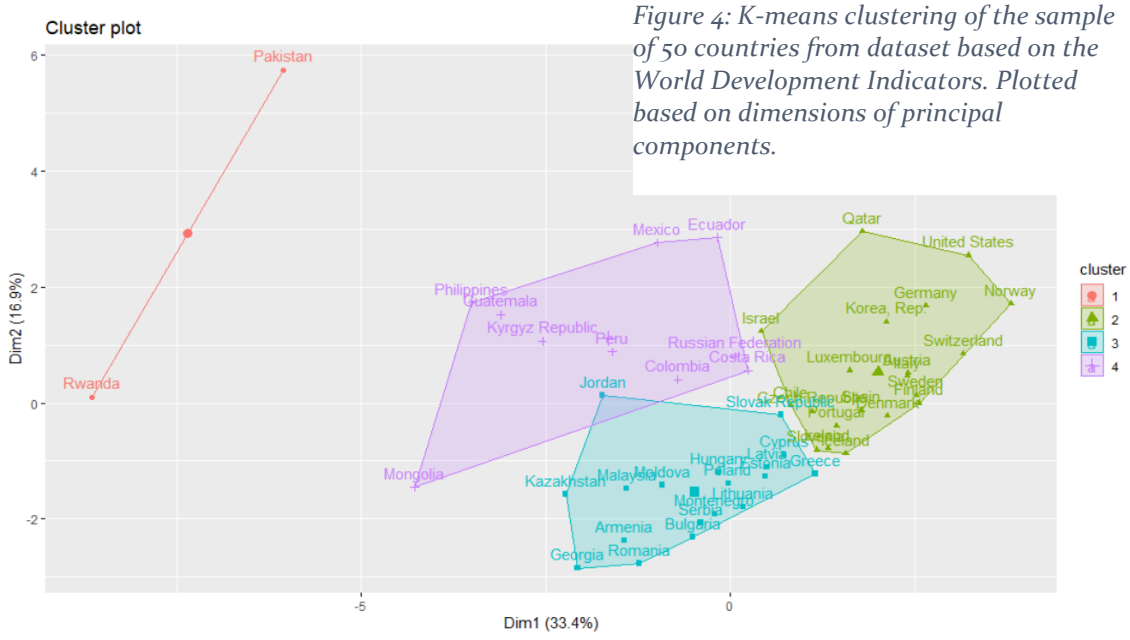
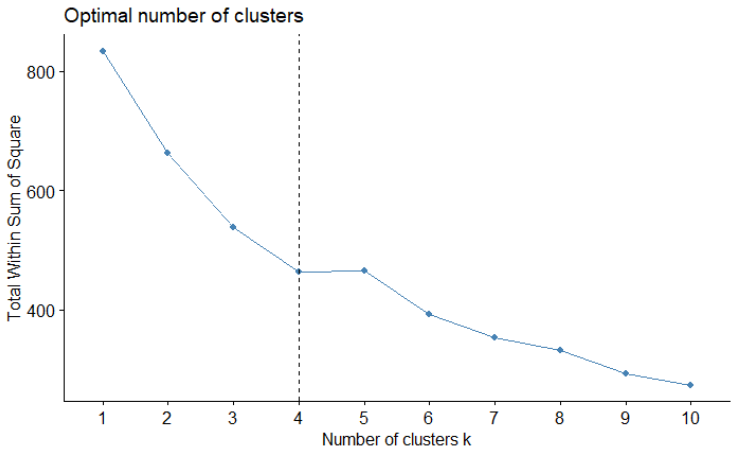


Figure 4: K-means clustering of the sample of 50 countries from dataset based on the World Development Indicators. Plotted based on dimensions of principal components.

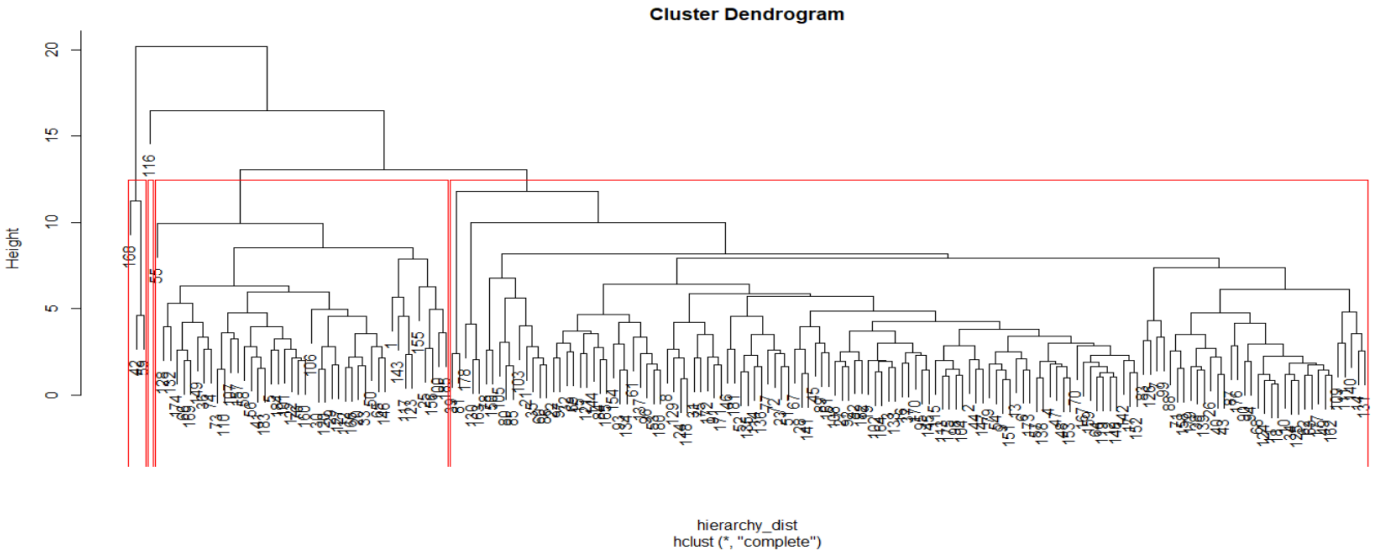
there is little to no Euclidean distance in the observations between them. Applying the k-means clustering algorithm and using the optimal number of 4 k-means clusters (see figure 5), Rwanda and Pakistan are clustered together and far away from the other countries in the dataset. This indicates that countries with similar socio-economic data to Rwanda and Pakistan would be positioned similarly. Cluster number 2 in figure 5 appears to include mainly countries that the World Bank considers 'high income' such as United States, Switzerland, and Germany *(Bank, 2020)*. Cluster 2 is mainly European countries although there are still countries in other continents like North America and Asia (presence of Rep. Korea in cluster 2). Eastern European countries like Hungary, Poland and Latvia are mainly in cluster 3. But cluster 4 consists of both Asian and South American countries. This makes sense when we see that these countries also had relatively small distances between them in the distance matrix in figure 3. Note that although 4 is the optimal number of clusters for categorizing these countries based on the WDI's, there is still a high level of variance between some countries in the clusters since the total sum of squares within clusters is over 400 (see figure 4). Figures 3 and 4 show that there is a small link between what continent a country is in and whether they will be clustered with other countries in the continent based on the WDI's, but often countries in different continents are still placed together as they may share similar WDI data. We should look at the data more broadly. Another way of examining how these countries may be clustered based on the WDI is to look at the dendrogram showing the hierarchical clustering of all the countries. Figure 6 (shown below includes every country in the dataset and once again shows us that countries which are grouped together are often not in the same continent. For instance, on the far left of the dendrogram,

Figure 5: A plot showing the relationship between the sum of squares in terms of distance observations for each country and the number of clusters. This explains why 4 k-means were chosen.



the indices of countries like Thailand (168) and Comoros (42) are closer to each other but very far in terms of the continent where they are located.

Figure 6: A dendrogram showing the hierarchical clustering (k-means = 4) of each country in the dataset based on the distances between the observations for each country. This dendrogram uses complete linkage.



A logistic regression model was built to see which World Development Indicators may act as statistically significant predictors in leading to high COVID-19 casualties. The threshold for what constitutes 'high COVID-19 casualties' was decided based on the upper quartile of the Covid.deaths variable. The upper quartile is 1830 deaths per country. It is important to note the assumptions made by logistic regression models – one being that there is no multicollinearity between the predictor variables in the model (i.e. that continent and health expenditure are not correlated with each other), observations of each country have not come from matched data or repeated measurements and that independent variables are linearly related to the log odds. The generalized linear model was run with 12 out of 19 of the world development indicators the dataset (for which there was no multicollinearity).

Before building the prediction model, positive multicollinearity was found between the continent, life expectancy, net national income, mortality and health expenditure per capita. Negative collinearity was found between GDP.capita and birth rate.[1] Thus, only continent and GDP.capita were included in the logistic regression model. The model detected a statistically significant relationship between only 4 out of the 12 variables and the likelihood of a country having high COVID-19 casualties. These variables were 'Continent', 'Election Access' 'population growth' and 'unemployment'. Continent was found to be the most significant at the <0.1% level, with a p-value of 0.000178. Followed by population growth, which was significant at the <1% level. Many of the fitted probabilities were either very close to zero or one, separation was not detected in the model so this is likely to be due to outliers for some variables. Election access and

[1] See Appendix A for collinearity matrix used.

```
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.80922  -0.32337  -0.00005   0.00000   1.55684

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.078e+02  5.843e+01  -1.845 0.065104 .
Continent              9.970e-01  2.660e-01   3.749 0.000178 ***
Elect.access           1.025e+00  5.825e-01   1.760 0.078443 .
Net.nat.income.capita  8.593e-02  1.070e-01   0.803 0.422055
Primary                3.032e-02  5.066e-02   0.599 0.549445
Pop.growth            -1.311e+00  4.292e-01  -3.054 0.002256 **
Pop.density           -2.151e-05  1.869e-04  -0.115 0.908391
Pop.total              1.563e-10  1.882e-09   0.083 0.933813
Health.exp             2.069e-01  1.359e-01   1.522 0.127933
Unemployment           1.459e-01  8.209e-02   1.777 0.075576 .
GDP.growth             1.879e-01  1.739e-01   1.081 0.279771
GDP.capita            -5.426e-07  9.268e-06  -0.059 0.953310
Water.services        -3.049e-02  2.261e-02  -1.348 0.177564
Comp.education        -8.208e-02  1.403e-01  -0.585 0.558547
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3: Coefficients for logistic regression model predicting covid deaths over 1830.

unemployment were significant at the <10% level (see coefficients in table 3 below). The coefficient estimates show that there is a negative correlation between population growth and high COVID casualties, but a positive correlation between election access and unemployment on the response variables. The fact that continent is the most statistically significant variable goes hand in hand with the findings in our preliminary analysis – where Europe had the highest number of recorded COVID-19 casualties. This may be used to suggest that the location of a country, unemployment levels, population and the extent of democracy are the most significant factors in determining whether they are likely to have high coronavirus casualties. Though we tried to control multicollinearity in this model, it is difficult to fully eliminate. Variables, like Election Access, which we found to be significant predictors may not have a truly causal relationship with COVID-19 deaths and may instead be related to the significance of ‘continent’ in terms of predicting COVID casualties. Instead of using a correlation matrix it may therefore be more useful to employ feature selection methods to see what variables are appropriate for a prediction model.

Another way to predict high, as well as low, COVID-19 casualties is to implement classification algorithms. The classes are defined based on the summary statistics about COVID-19 deaths. Casualty numbers between 0-167 would be classed as ‘low’, 167-1143 as ‘medium’, 1143-18630 as ‘high’ and 1830-6260 as ‘very high’. Before applying the multinomial logistic regression to predict what classes a country may be assigned to, feature selection was applied to the world development indicators, and there is no need to remove collinear variables since the random forests algorithm accounts for it. The aim is to see which of these indicators would be the best variables to include in the model. The results are shown in

figure 7 below.

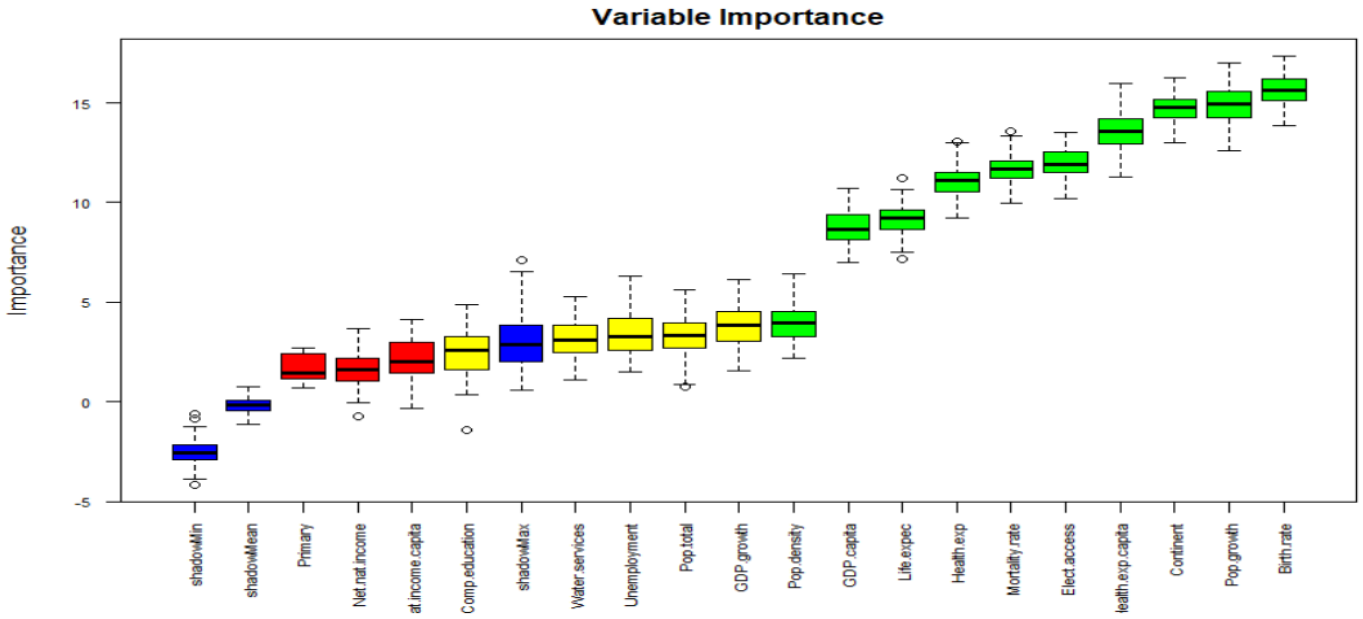


Figure 7: A plot of World Development Indicators based on their importance. Created using the Boruta package in R. The variable with green boxplots represents variables that the feature selection algorithm recommends we include in our models.

Due to the results of the feature selection, and because of the structure of its random forests algorithm. When building the multinomial logistic regression model, only the variables accepted by the feature selection were included. To cross-validate the multinomial logistic regression model, the model was run twice with two sets of testing data. In the first test, 20% of the data was used as test data and the rest as training data. Whereas in the second test, 30% of the data was used as test data and 70% as training data. The second test had more prediction accuracy – with a score of 77.36 % average correct predictions, and 26 countries being correctly classed in 'low', 'high', 'medium' or 'very high'. Whereas the first test scored 72.2% for average number of correct predictions – and 41 countries being placed in the right classes (due to the larger number of test data.[2] A LDA classification model was also developed. This model also only included variables recommended by the feature selection. To cross-validate this model, instead of using a basic data split method as was performed for the logistic regression model, k-fold cross validation was employed. The accuracy of the LDA model itself, after 5 repeats for which the k-fold model took 5 resamples and calculated the mean squared error each time, the overall accuracy was found to be 62%. When the model was used run against the original dataset, the average percentage of correct predictions was 67.03%. 124 countries were correctly placed in the right class. Lastly, a classification model using QDA was created. It is valuable to compare this model to the LDA one because we cannot be sure if every development indicator used in the model has observations that are normally distributed and have homogeneous variance

Table 4: Confusion Matrix for the Logistic, LDA and QDA models. Showing correct predictions against original dataset.

Predicted	Low	Medium	High	Very High
Low	36	8	1	0
Medium	9	42	11	3
High	0	5	6	3
Very High	2	11	8	40

[1] 0.6702703

qda.pred	Low	Medium	High	Very High
Low	44	28	4	1
Medium	0	16	0	1
High	0	5	11	2
Very High	3	17	11	42

[1] 0.6108108

[2] See Appendix B for confusion matrices for the multinomial logistic regression model.

around the mean. QDA does not hold the same requirements for each variable. The overall accuracy calculated by the k-fold cross validation method for QDA was 52.4%. When running this model against the dataset, average number of correct predictions was 61.1%. 113 countries were correctly placed in the right class. Thus, the QDA model scored the lowest for prediction accuracy of COVID-19 casualties. However, it should still be considered as a model to use for predictions since it does not make the same assumptions as the linear discriminant analysis (LDA).

Model	Highest Accuracy after Cross-Validation/recall
Multinomial Logistic Regression	77.36%
Linear Discriminant Analysis	67.03%
Quadratic Discriminant Analysis	61.1%

The multinomial logistic regression model performed the best in terms of prediction accuracy on unseen data. This may be because it used less countries in its testing data. All models are useful for classifying countries as most of their predictions were correct when run against the original dataset.

Discussion

Europe appears to have recorded the most COVID-19 casualties, which is why certain world development indicators were more statistically significant than others. Another finding is that using a feature selection algorithm and building a classification model can predict COVID-19 casualties based on the development indicators – with the three models accurately classifying most countries it was asked to. The findings of this report are remain significantly limited in terms of accuracy, due to the problem of missing data. The under-reporting of disease and virus mortality has been a problem with many other epidemics including Malaria (Whittaker, 2021). Not everyone who contracts the disease seeks testing services or healthcare. Thus, under-ascertainment refers not only to the under-reporting of COVID-19 cases but also the underestimation of COVID-19 deaths. Pandemic researchers argue that excess mortality figures are often a better indicator of pandemic casualties than officially reported deaths from COVID-19 (Whittaker, 2021). The World Bank data suggests that Europe suffered the most COVID-19 casualties, which affected the findings of the prediction models employed in this report. However, this is because European countries counted a higher *proportion* of COVID-19 deaths compared to non-European countries. Vestgaard et al estimate that at least 75% of COVID-19 casualties in Europe have been counted in official statistics, but this number disintegrates when we look at figures for countries and continents outside Europe (Vestergaard LS, 2020). Peru's death toll is the highest in our dataset, only because of the recounting of Peruvian casualties - which tripled official COVID-19 death rates in the country (Dyer, 2021). In addition, many sources estimate that only 10% of casualties in India have been officially reported (AAJS, 2021). This issue also means that our findings about the relationship between WDIs and COVID-19 deaths may not be correct if some countries in the dataset gave lower casualty numbers. In conclusion, the data collected by the World Bank is likely to be inaccurate and thus even our classification models for grouping countries and predicting COVID-19 casualties based on the role of world development indicators may not be fully applicable to predicting casualties of future pandemics. Nevertheless, that is not to say the findings of this data analysis cannot be used in combination with other research (e.g., research that considers excess mortality) to see what factors truly contribute to pandemic casualties.

References

AAJS, S. (2021). hree new estimates of India's all-cause excess mortality during the covid-19 pandemic. *Center for Global Development*. Retrieved from https://cgdev.org/publication/three-new-estimates-indias-all-cause-excess-mortality-during-covid-19-pandemic

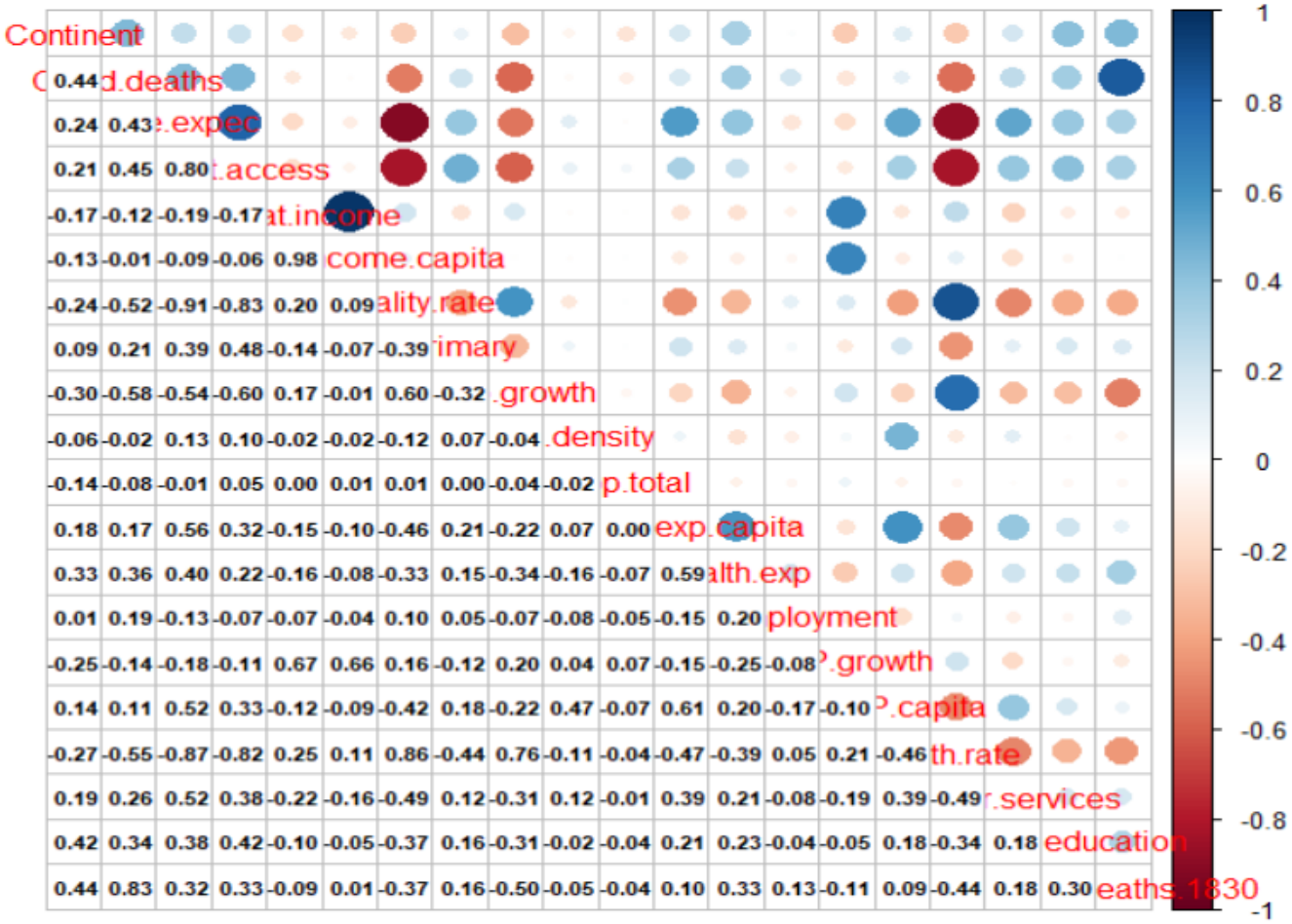
Bank, W. (2020, June). *List of Economies - World Bank DataBank.* Retrieved from databank.worldbank.org: https://databank.worldbank.org/data/download/site-content/CLASS.xls

Dyer, O. (2021). Peru's official death toll triples to become world's highest. *BMJ*, 373. doi:10.1136/bmj.n1442

Vestergaard LS, N. J. (2020). Excess all-cause mortality during the covid-19 pandemic in Europe - preliminary pooled estimates from the EuroMOMO network. *Euro Surveill*. doi:10.2807/1560-7917.ES.2020.25.26.2001214

Whittaker, C. A. (2021). Under-reporting of deaths limits our understanding of true burden of covid-19. *BMJ*, 375. Retrieved from https://www.bmj.com/content/375/bmj.n2239

Appendix A – Collinearity Matrix



Appendix B – Confusion Matrices for multinomial logistic regression model

```
pred.classes Low Medium High Very High
   Low         8      1    0         0
   Medium      1     10    3         2
   High        0      1    1         0
   Very High   0      1    1         7

pred.classes2 Low Medium High Very High
    Low        12      3    0         0
    Medium      2     16    5         1
    High        0      0    1         0
    Very High   0      0    1        12
[1] 0.7222222
[1] 0.7735849
```

Appendix C – Source R Code for Question 1

```
setwd("~/") # set working directory
WDI <- read.csv("project_data1.csv") # read CSV file with project data
View(WDI)
str(WDI)

# convert Covid.deaths variable from character to numeric
WDI$Covid.deaths <- gsub(",", "", WDI$Covid.deaths)
WDI$Covid.deaths <- as.numeric(WDI$Covid.deaths)

library(vtable) # load vtable library
st(WDI) # make table of summary statistics for each numerical variable in the dataset

# make table of covid deaths and sort in descending order
ord.coviddeath <- covid.death[order(-covid.death$Covid.deaths),]
ord.coviddeath

# make boxplot summarising country data for COVID deaths
library(ggplot2)
g = ggplot(data=WDI, aes(x=Covid.deaths))+geom_boxplot(color="blue")+xlab("Deaths from
Coronavirus") +ylab("Number of Countries")
g
# make barchart showing covid deaths per continent

continent.deaths <- aggregate(WDI$Covid.deaths, by=list(Continent=WDI$Continent),
FUN=sum)
y = data.frame(Continent=c('Africa','Asia','Australia/Oceania','Europe','North America',
'South America'),Number=c(16061,27918,3471,103021,31734,29165))
barplot(y$Number, names.arg=y$Continent, xlab="Continent", ylab="Total number of recorded
COVID-19 deaths", col = rainbow(6))
```

Appendix D – Source R code for Question 2

```
# remove nas and replace with median / consider changing to for loop
WDI$Life.expec[is.na(WDI$Life.expec)] = median(WDI$Life.expec, na.rm=TRUE)
WDI$Elect.access[is.na(WDI$Elect.access)] = median(WDI$Elect.access, na.rm=TRUE)
WDI$Net.nat.income[is.na(WDI$Net.nat.income)] = median(WDI$Net.nat.income, na.rm=TRUE)
WDI$Net.nat.income.capita[is.na(WDI$Net.nat.income.capita)] =
median(WDI$Net.nat.income.capita, na.rm=TRUE)
WDI$Mortality.rate[is.na(WDI$Mortality.rate)] = median(WDI$Mortality.rate, na.rm=TRUE)
WDI$Primary[is.na(WDI$Primary)] = median(WDI$Primary, na.rm=TRUE)
WDI$Pop.growth[is.na(WDI$Pop.growth)] = median(WDI$Pop.growth, na.rm=TRUE)
WDI$Pop.density[is.na(WDI$Pop.density)] = median(WDI$Pop.density, na.rm=TRUE)
WDI$Pop.total[is.na(WDI$Pop.total)] = median(WDI$Pop.total, na.rm=TRUE)
WDI$Health.exp[is.na(WDI$Health.exp)] = mean(WDI$Health.exp, na.rm=TRUE)
WDI$Unemployment[is.na(WDI$Unemployment)] = median(WDI$Unemployment, na.rm=TRUE)
WDI$GDP.growth[is.na(WDI$GDP.growth)] = median(WDI$GDP.growth, na.rm=TRUE)
WDI$GDP.capita[is.na(WDI$GDP.capita)] = median(WDI$GDP.capita, na.rm=TRUE)
WDI$Birth.rate[is.na(WDI$Birth.rate)] = median(WDI$Birth.rate, na.rm=TRUE)
WDI$Water.services[is.na(WDI$Water.services)] = median(WDI$Water.services, na.rm=TRUE)
WDI$Comp.education[is.na(WDI$Comp.education)] = median(WDI$Comp.education, na.rm=TRUE)
WDI$Health.exp.capita[is.na(WDI$Health.exp.capita)] = mean(WDI$Health.exp.capita,
na.rm=TRUE)


WDI_clean <- WDI[,-1] # makes the country.name column a rowname so it can be treated as
an index
rownames(WDI_clean) <- WDI[,1]
View(WDI_clean)

WDI_clean=subset(WDI_clean, select=-c(1,2,21))
WDI_clean=subset(WDI_clean, select=-c(18))
#WDI_clean <- na.omit(WDI_clean) # remove NAs
#View(WDI_clean)

str(WDI_clean) # check classes of every column/variable in the dataset
WDI_clean$Comp.education <- as.numeric(WDI_clean$Comp.education) # convert factor to
numeric

WDI_clean$Pop.total <- as.numeric(WDI_clean$Pop.total) # convert factor to numeric

library(MASS)
library(ISLR)
library(factoextra)
library(ggplot2)

dim(WDI_clean) # dimensions are 185 by 17

# remove Continent and Covid deaths from subset
cluster = WDI_clean # or WDI_clean[50:100,]
str(cluster)
View(cluster)

# select sample of 50 countries from data
cluster_sample <- cluster[sample(nrow(cluster), 50), ]

# pca
#pca_data <- subset(WDI, select=-c(1,2,3,21))
```

```
#pca1 = prcomp(pca_data, center = TRUE, scale = TRUE)
#summary(pca1)
#cluster_transform = as.data.frame(-pca1$x[,1:6])

# scale data before clustering

scaled <- scale(cluster_sample)

distance_Euclidean <- get_dist(scaled) # get Euclidean distances of data points between
each country

fviz_dist(distance_Euclidean, gradient = list(low = "#00AFBB", mid = "white", high =
"#FC4E07")) # display Euclidean distances in distance matrix

# determine optimal number of clusters
fviz_nbclust(scaled, kmeans, method = "wss")+
geom_vline(xintercept = 4, linetype = 2) # shows us that 4 is the optimal number of
clusters due to Sum of Squares between distances

# implement k-means clustering algorithm with cleaned data
set.seed(123)
kmeans1 <- kmeans(scaled, centers = 4, nstart = 20)
fviz_cluster(kmeans1, data = scaled)

kmeans1

# Hierarchical clustering

WDIx<- WDI[,-(c(1,2,3))]
WDIx<- WDIx[,-(c(18))]

#WDIx$Comp.education <- as.numeric(WDIx$Comp.education)
#WDIx$Pop.total <- as.numeric(WDIx$Pop.total)
WDIx<-scale(WDIx)

# select sample of 50 countries from data
cluster_sample2 <- WDIx[sample(nrow(cluster), 50), ]

#Start my calculating the distance matrix
hierarchy_dist <- dist(WDIx, method = "euclidean")
#Apply hierarchical clustering for differnt linkage methods
fit.complete <- hclust(hierarchy_dist, method="complete")

# complete linkage
plot(fit.complete)
groups.fit.complete <- cutree(fit.complete, k=4)
rect.hclust(fit.complete, k=6, border="red")
# draw dendrogram with red borders around the 4 clusters
rect.hclust(fit.complete, k=6, border="red")

table(groups.fit.complete)
```

Appendix E – Source R code for Question 3

```
# check for multicollinearity
library(corrplot)
fit1<- lm(Covid.deaths~Continent + Life.expec + Elect.access + Net.nat.income
+Net.nat.income.capita + Mortality.rate + Primary + Pop.growth + Pop.density + Pop.total
+ Health.exp.capita + Health.exp + Unemployment + GDP.growth + GDP.capita + Birth.rate +
Water.services + Comp.education, data=WDI_clean)
corr<-cor(log.data)
corr
corrplot.mixed
corrplot.mixed(corr, lower.col = "black", number.cex = .7)

# shows collinearity between continent, covid.deaths, Life Expectancy, net.nat.income,
mortality rate, health expenditure per capita. As well as strong negative
multicollinearity birth rate and gdp.capita
     # transform covid.deaths variable from numeric to categorical and then to binary
 # set the threshold for high covid casualties as anything above the upper quartile number
                                    of deaths found in q1

log.data <- subset(WDI, select=-c(1,21))
str(log.data)

summary(log.data$Covid.deaths) # shows us upper quartile for number of covid deaths is
1830
# recode continent as numerical variable instead of categorical
log.data$Continent
log.data$Continent[log.data$Continent == "Asia"] = 1
log.data$Continent[log.data$Continent == "Africa"] = 2
log.data$Continent[log.data$Continent == "Europe"] = 3
log.data$Continent[log.data$Continent == "North America"] = 4
log.data$Continent[log.data$Continent == "Australia/Oceania"] = 5
log.data$Continent[log.data$Continent == "South America"] = 6
log.data$Continent <- as.numeric(log.data$Continent)
log.data$Continent[is.na(log.data$Continent)] = mean(log.data$Continent, na.rm=TRUE)

log.data$Covid.deaths.1830=0
log.data$Covid.deaths.1830[log.data$Covid.deaths>1830]=1
# check dummy variable has been created
table(log.data$Covid.deaths.1830)
highcovid <- table(log.data$Covid.deaths.1830)

# create logistic regression model to predict high covid casualties (use only variables
without multicollinearity)
log_model <- glm(Covid.deaths.1830 ~ Continent + Elect.access +Net.nat.income.capita +
Primary + Pop.growth + Pop.density + Pop.total + Health.exp + Unemployment + GDP.growth +
GDP.capita + Water.services + Comp.education, family=binomial,data=log.data)
library(detectseparation)
# detect separation between predictors and response variable that may be causing overly
high probabilities
glm.det<-glm(log_model,family=binomial("logit"),method="detect_separation")
glm.det # seperation is false meaning probabilities are due to extreme values

summary(log_model)
```

Appendix F – Source R code for Question 4

```
# Logistic regression model with feature selection to keep out noisy variables
```{r}
library(caret)
library(Boruta)

# feature selection for classification algorithms
boruta2<-Boruta(Covid.deaths.categorical~Continent + Life.expec + Elect.access +
Net.nat.income +Net.nat.income.capita + Mortality.rate + Primary + Pop.growth +
Pop.density + Pop.total + Health.exp.capita + Health.exp + Unemployment + GDP.growth +
GDP.capita + Birth.rate + Water.services + Comp.education,data=class_data, doTrace=2)

decision<-boruta2$finalDecision
signif <- decision[boruta2$finalDecision %in% c("Confirmed")]
print(signif)
plot(boruta2, cex.axis=.7, las=2, xlab="", main="Variable Importance")

# logistic multiclass classification
```{r}
class_data <- subset(WDI, select=-c(1,21)) # removes country name
class_data$Covid.deaths.categorical <- cut(class_data$Covid.deaths,
breaks=c(0,167,1143,1830,6260), labels=c("Low", "Medium", "High", "Very High"))

# recode continent as numerical variable instead of categorical
class_data$Continent
class_data$Continent[class_data$Continent == "Asia"] = 1
class_data$Continent[class_data$Continent == "Africa"] = 2
class_data$Continent[class_data$Continent == "Europe"] = 3
class_data$Continent[class_data$Continent == "North America"] = 4
class_data$Continent[class_data$Continent == "Australia/Oceania"] = 5
class_data$Continent[class_data$Continent == "South America"] = 6
class_data$Continent <- as.numeric(class_data$Continent)
class_data$Continent[is.na(class_data$Continent)] = median(class_data$Continent,
na.rm=TRUE)

# load relevant packages
library(MASS)
library(ISLR)
library(boot)
library(nnet)

# start with multinomial logistic regression for mulit-class classification

# Split the data into training and test set

library(caret)
library(tidyverse)
set.seed(123)
training <- class_data$Covid.deaths.categorical %>%
  createDataPartition(p = 0.8, list = FALSE) # 80% training data, 20% testing
train_data  <- class_data[training, ]
test_data <- class_data[-training, ]

set.seed(123)
training2 <- class_data$Covid.deaths.categorical %>%
  createDataPartition(p = 0.7, list = FALSE) # 80% training data, 20% testing
```

```
train_data2  <- class_data[training2, ]
test_data2 <- class_data[-training2, ]


# Fit the multinomial logistic regression model
glm.multi2 <- nnet::multinom(Covid.deaths.categorical ~ Continent + Life.expec +
Elect.access + Mortality.rate + Pop.growth + Pop.density + Health.exp.capita + Health.exp
+ GDP.capita + Birth.rate, data = train_data)
# Summarize the model
summary(glm.multi2)
# Make predictions
pred.classes <- glm.multi2 %>% predict(test_data)
pred.classes2 <- glm.multi2 %>% predict(test_data2) # repeats prediction for second test
set
# check confusion matrix
table(pred.classes, test_data$Covid.deaths.categorical)
table(pred.classes2, test_data2$Covid.deaths.categorical)
# Model accuracy
mean(pred.classes == test_data$Covid.deaths.categorical)
mean(pred.classes2 == test_data2$Covid.deaths.categorical)

# LDA multiclass classification
```{r}
# For 5-fold CV
trControl <- trainControl(method = "repeatedcv", number = 5, repeats=5)
# or trControl <- trainControl(method = "cvRepeat", number = 5, repeats=10)
library(MASS)
lda.fit <- train(Covid.deaths.categorical~Continent + Life.expec + Elect.access +
Mortality.rate + Pop.growth + Pop.density + Health.exp.capita + Health.exp + GDP.capita +
Birth.rate,
                 method = "lda", # could change this to steplda to integrate feature
elimination
                 trControl = trControl,
                 metric = "Accuracy",
                 data = class_data)

Predicted <- predict(lda.fit,class_data)
table(Predicted, class_data$Covid.deaths.categorical)
# Model accuracy
mean(Predicted == class_data$Covid.deaths.categorical)

# model summary
lda.fit

# QDA mutliclass classification
```{r}
# implement QDA classification with cross-validation and feature selection

#qda_model <-qda(Covid.deaths.categorical ~.,data=class_data,subset=1:139)
#qda_model
#qda.predicted <- predict(qda_model,test_data2)$class
#table(qda.predicted, class_data$Covid.deaths.categorical)
#mean(qda.predicted==class_data$Covid.deaths.categorical)

qda.fit <- train(Covid.deaths.categorical~Continent + Life.expec + Elect.access +
Mortality.rate + Pop.growth + Pop.density + Health.exp.capita + Health.exp + GDP.capita +
Birth.rate,
                 method = "qda",
                 trControl = trControl,
                 metric = "Accuracy",
                 data = class_data)
qda.pred <- predict(qda.fit,class_data)
table(qda.pred, class_data$Covid.deaths.categorical)
# Model accuracy
```

```
mean(qda.pred == class_data$Covid.deaths.categorical)

qda.fit
```