MA335 Final Project

**By**

**SAYALI RAJENDRA KAMBLI**

**Reg no. 2010910**

**Index**

**1. Introduction** 1

**2**. **Analysis** 1

2.1 Question 1 3

2.2 Question 2 4

2.3 Question 3 6

2.4 Question 4 7

2.5 Question 5 9

**3. References** 9

**4.** **Appendix** 9

Word count: 1624

**1. Introduction**

The data set *project\_data.csv* provides a selection of the World Development Indicators (WDI), derived from a primary World Bank database, and the information about the casualties of the COVID-19 pandemic.

The aim of this project is to implement clustering and classification algorithms on the data set and to find relationships between Covid- 19 pandemic impacts and countries with their economic and health profiles.

**2. Analysis**

**2.1: Question 1**

**Analyzing project\_data.csv Dataset**

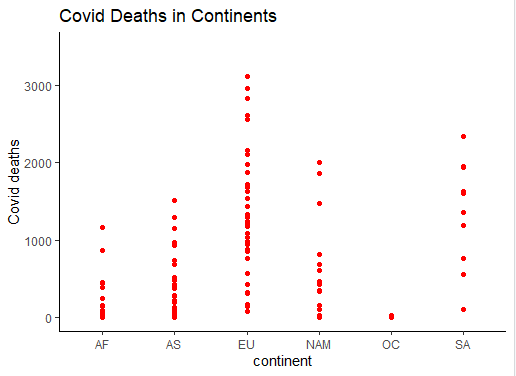
The data set *project data.csv* has observations of 148 countries of 14 attributes about health and general information related to casualties of the COVID-19 pandemic at 2021.

* Data set gives economic profiles with attributes like Income, Import, Export, GDP, Population Density, Inflation rate of countries.
* It also provides information on health attributes like Child mortality rate, Health spending, Life expectancy, Total fertility rate of various countries.
* Countries are from 6 Continents: - Europe, Asia, Africa, North America, South America, Oceania.
* Minimum Covid deaths are 0.30 with average 654.80 and maximum is 5692 which is very high.

Following table shows Minimum, Maximum and Average values of 148 countries can have for some attributes.

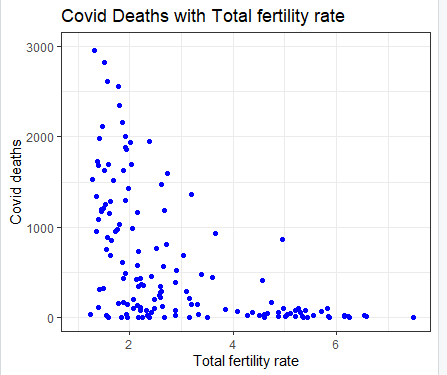
|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes** | **Min** | **Max** | **Average** |
| Child mortality rate | 2.600 | 208 | 39.018 |
| Exports | 0.109 | 175 | 39.848 |
| Health spending | 1.810 | 17.9 | 6.750 |
| Imports | 0.0659 | 154 | 44.7025 |
| Total fertility rate | 1.230 | 7.490 | 2.960 |
| Population density | 2.0 | 1913.0 | 159.1 |
| Income | 609 | 125000 | 17055 |
| GDP | 231 | 105000 | 13259 |

**Table showing summary data of some attributes**

****

**Graph Showing relationship between Covid deaths in continents**

Above graph shows covid deaths variation in 6 continents. It is observed that Europe continent has faced large amount of covid deaths. Whereas in Africa, North America and Asia countries have covid deaths near 1000 per 1M peoples. Oceania continent has very few covid deaths. South America has countries with even spread of covid deaths.

****

**Graph of Covid deaths vs Total fertility rate**

Here, graph shows relationship of total fertility rate with covid deaths in countries.

It is visible that countries having large number of covid deaths i.e. more than 1000 has less total fertility rate (1-3).

As total fertility rate increases covid death count decreases for countries.

**2.2: Question 2.**

**Ans:**

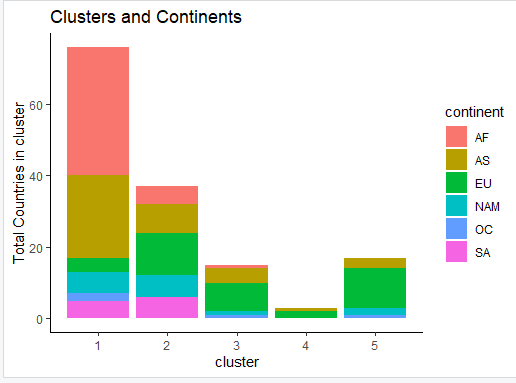
**K-means and Agglomerative Clustering algorithms**

For Clustering, I have selected K=5.

As question says, use 10 variables, I have selected following 10 variables

*child\_mort, exports, health, imports, income, inflation, life\_expec, total\_fer, gdpp and pop\_dens*

**Result after applying k-means algorithm with k=5**



**Graph of countries from various continents vs clusters**

Graph shows clusters having countries count with different continents

From above graph, following observations can be made:

1. Not all countries from a single continent belongs to one cluster, i.e., countries from same continent are divided into 2 or more clusters, which indicates that continent have countries of different economic profiles.
2. So in each continent there will be countries who are performing well but still has covid-19 pandemic impacts.
3. In, Africa and South America continent some countries have similar situation so they belong to clusters 1 and 2.
4. Europe, Asia, North America, Oceania have countries with totally different health and general state, so they are spread into more than 3 clusters.
5. Cluster 1 is largest among all which has 76 countries.
6. Cluster 4 is the smallest cluster with only 3 countries.
7. Cluster 2 is second large cluster with 37 countries.
8. Cluster 3 and cluster 5 have 15 and 17 countries respectively.

I have also performed Agglomerative Clustering with complete linkage method on this dataset.

After comparing both results, accuracy is ***62.83% .***

It is because of clustering algorithms are giving different clusters. So this accuracy can be still considered as good.

**2.3: Question 3.**

**Transform Covid deaths into Binary Variable and Apply Logistic regression model for predicting covid deaths**

**Ans:**

For transforming *covid\_deaths* into binary variable, I have used threshold as 654.798(mean value of *covid*\_*deaths*). So above value of 654.798 covid\_deaths new variable has value as 1 and remaining has 0.

After transforming it 90 rows have value as 0 and 53 rows have values as 1.

As question says use 11 variables, I have used following 11 variable to fit the model

*X, child\_mort, exports, health, imports, income, inflation, life\_expec, total\_fer, gdpp and pop\_dens*

**Result After applying Logistic regression:**

Accuracy is ***85.13 %*** with following confusion matrix:

|  |  |  |
| --- | --- | --- |
| Actual/Predicted values |  |  |
|  | 0 | 1 |
| 0 | 84 | 11 |
| 1 | 11 | 42 |

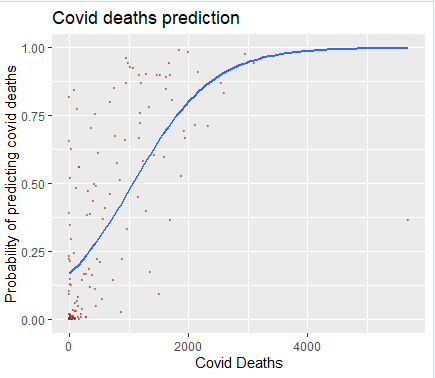
**confusion matrix for Logistic Regression**

Model give*s 88.42 %* of **specificity** and *79.24 %* of **sensitivity** which is very good result.

So, the **Logistic Regression** model has been a very good fit for this data set.

If needed it can be improvised by changing the threshold value while converting covid deaths variable into binary values.

Below graph shows how logistic regression has produced the probabilities of predicting covid deaths.



**Graph showing Logistic Regression for predicting covid deaths using probabilities**

In the graph it can be noticed that as covid deaths increasing probabilities values are getting more accurate. **Logistic Regression** model is good fit for predicting higher covid deaths.

**2.4: Question 4.**

**Ans:**

**Transform Covid deaths into 4 labels Variable and Apply QDA, LDA and logistic regression for this multiclass classification**

First dataset needs new variable with 4 possible labels of covid deaths.

I have transformed covid\_deaths variable into following for categories

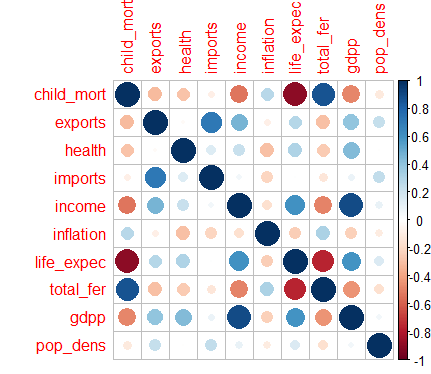
*covid\_deaths > 1900, => "Very High",*

*covid\_deaths between (1000, 1900) => "High",*

*covid\_deaths between (500, 1000) => "Low",*

*covid\_deaths < 500 => "Very Low"*

Figure below shows correlation between attributes from the dataset.

****

After checking correlation between variable, I have removed variable *life\_expec* and *gdpp.*

I have also not taken *country* and *continent* variables.

So final list of attributes used for classifications are: *X , child\_mort , income, exports , health, imports , inflation , total\_fer and pop\_dens.*

**Result:-**

After applying LDA, QDA and Logistic regression for this multiclass covid\_deaths\_label variable, I am getting following accuracy:

**LDA: - 72 %,**

**QDA: - 75.67 %,**

**Logistic regression: - 75%**

From the accuracy values it can be observed that Models are not so bad fit. However, it is not best fit, but it can be considered at good fit.

QDA and Logistic Regression are giving better accuracy than LDA.

Following are the confusion matrices

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| LDA | High | Low | Very High | Very Low |
| High | 17 | 3 | 6 | 7 |
| Low | 1 | 3 | 0 | 2 |
| Very High | 1 | 1 | 6 | 1 |
| Very Low | 7 | 11 | 1 | 81 |
|  |  |  |  |  |
| QDA | High | Low | Very High | Very Low |
| High | 20 | 6 | 0 | 10 |
| Low | 2 | 7 | 0 | 8 |
| Very High | 3 | 1 | 13 | 1 |
| Very Low | 1 | 4 | `0 | 72 |
|  |  |  |  |  |
| QDA | High | Low | Very High | Very Low |
| High | 17 | 3 | 4 | 5 |
| Low | 1 | 4 | 6 | 0 |
| Very High | 2 | 1 | 6 | 0 |
| Very Low | 6 | 10 | `2 | 84 |

For Validation, I have used Cross Validation approach

k-fold cross validation with k=10:

**Results:**

LDA :- 68.89 %.

QDA:- 59.48 %,

Logistic Regression:- 69.01%

After applying validation it is seen that accuracy has decreased that means models was not good fit also there was partitioning of data into labels had some impacts on the low accuracy.

**2.5: Question 5.**

World Development Indicators (WDI) data has very helpful to analyse covid deaths in countries with different economic profiles.

**Is it realistic that the same relationships between the predictor and response variables should be expected in the future?**

Yes. As this is the biggest pandemic happened in world after so long time. It will help to prepare and handle other similar situation in future.

Covid- 19 pandemic affected all countries in all continent. Countries having similar economic profiles faced similar kind of impacts.

One of the helpful attributes which will help is health. As Covid-19 showed how health profile is important for countries, in future countries can improve their health system in order to fight better with pandemic.

**What can be learned?**

As this dataset had various economic attributes like Income, Import, Export, GDP, Population Density, Inflation rate of countries it is better to find out how Covid-19 or any other pandemic impacts differently on countries.

|  |
| --- |
| child\_mort exports health imports income inflation life\_expec total\_fer gdpp pop\_dens |
| 1 63.63 31.00 5.92 44.47 4433.72 10.42 64.96 3.86 1948.09 125.83 |
| 2 17.41 44.84 6.56 43.70 16883.78 8.11 73.89 2.09 9236.22 149.92 |
| 3 12.97 51.39 7.82 48.61 32440.00 4.27 78.07 2.03 26533.33 369.60 |
| 4 5.00 92.33 6.35 64.77 93000.00 5.52 80.60 1.88 87700.00 163.67 |
| 5 5.02 49.12 10.00 40.93 46876.47 2.58 80.49 1.82 47735.29 141.12 |

Above table shows how countries from different continents are clustered based on various attributes used.

* For child mortality rate there are big differences in centres value which indicate some countries really have higher rate where as in some it is very low.
* Same relation is there with GDP. Countries with higher GDPP will have different impact than less GDP countries.
* It is obvious that not all countries will have same income. Here there is very large difference between incomes of countries. Some have very large income rate , while others have not so good income rate.
* Countries having similar count of Exports and imports values belong to same clusters.
* Population density is somewhat same for 4 clusters and high for one cluster.

This type of data is definitely going to help in future because if big pandemic occurs then all countries will be affected as it is hard to control the spread of viruses and all. So similar or improvised analysis can be done in future so that we can get results faster and will help controlling bad impacts. There might be need to introduce new variables information related to deaths and vaccination as going forward is future. As time goes data need to be updated in order to find the appropriate analysis. Also, countries which had less casualties of the COVID-19 pandemic can be observed for next analysis in order to get perfect model of countries.

**3. References:**

1.<https://daviddalpiaz.github.io/r4sl/the-caret-package.html>

2.http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

3. <https://moodle.essex.ac.uk/course/view.php?id=15075&section=3>

**4 Appendix**

#loading library

library(ggplot2)

library(dplyr)

library(MASS)

library(nnet)

library(caret)

library(corrplot)

#importing dataset

setwd('C:\\Users\\hp\\Desktop\\Masters\\Summer Term\\335- Modelling\\Final project\\code')

dataset<- read.csv("project\_data.csv",header = TRUE)

#fixing uppercase error in continent

dataset$continent <- toupper(dataset$continent)

df<- dataset

#Question 1

#dimension of dataset

dim(df)

attach(df)

#finding summaries details of attributes

summary(covid\_deaths)

summary(child\_mort)

summary(exports)

summary(health)

summary(imports)

summary(total\_fer)

summary(pop\_dens)

summary(income)

summary(gdpp)

attach(df)

#first graph to show relation ship between covid deaths and continents

ggplot(df,aes(x=continent,y=covid\_deaths)) + geom\_point(color="red")+

ylab("Covid deaths")+ylim(0,3500)+theme\_classic()+ggtitle("Covid Deaths in Continents")

# graph to show relation ship between covid deaths and Total fertility

ggplot(df,aes(y=covid\_deaths,x=total\_fer)) + geom\_point(color="blue")+xlim(1,7.5)+ylim(0,3000)+

ylab("Covid deaths")+xlab("Total fertility rate")+theme\_bw()+ggtitle("Covid Deaths with Total fertility rate")

#question2

#taking attributes 3 to 12(10)

# K-means clustering with k=5

k=5;

kdata <-df[,c(3:12)]

names(kdata)

set.seed(123)

kd <- kmeans(kdata,centers=K, nstart=20)

kdata$cluster <- kd$cluster

kdata$continent<-as.factor(df$continent)

kdata$covid<-df$covid\_deaths

kc<- as.data.frame(kdata$cluster)

#ploting graph of clusters and continent

ggplot(kdata,aes(x=cluster,fill=continent)) + geom\_bar()+theme\_classic()+ggtitle("Clusters and Continents")+

ylab("Total Countries in cluster")

kd$size

#Agglomerative Clustering

hdata <- dist(df[,c(3:12)])

hd <- hclust(hdata, method="complete")

hc <- as.data.frame(cutree(hd, k=5) )

#binding both cluster values

cbind(hc ,kc)

#calculating accuracy value

mean(hc==kc)

#this is to get clusters centers

round(kd$centers,2)

#Question 3

#To convert into binary variable finding threshold

mean(covid\_deaths)

#creating new binary variable

df<-df %>% mutate(covid\_deaths\_bin=ifelse(covid\_deaths>654.798,1,0))

attach(df)

summary(as.factor(df$covid\_deaths\_bin))

#fitting logistic regression model with 11 attributes

model=glm(covid\_deaths\_bin ~ X + child\_mort + exports + health + imports + income + inflation + life\_expec + total\_fer + gdpp + pop\_dens ,family = binomial,data=df)

modelpredict <- predict(model, type="response" )

#for converting probabilities values into categorical

modelpredict1 = rep(0 ,148)

modelpredict1[modelpredict>0.5]=1

df$Covid\_deaths<-covid\_deaths

df$Prediction\_probability<-modelpredict

#finding accuracy of model

mean(df$covid\_deaths\_bin==modelpredict1)

#confusion matrix

table(covid\_deaths\_bin,modelpredict1)

#graph

ggplot(df,aes(x=Covid\_deaths,y=Prediction\_probability))+geom\_point(size=0.5,color="brown",alpha=0.5)+stat\_smooth(method="glm", se=FALSE, method.args = list(family=binomial)) +

labs(title = "Covid deaths prediction", x = "Covid Deaths",y = "Probability of predicting covid deaths")

specificity(actual=covid\_deaths\_bin,predicted=modelpredict)

sensitivity(actual=covid\_deaths\_bin,predicted=modelpredict)

#Question 4

data<-dataset

#converting into labels for classification

data1<-data %>% mutate(covid\_deaths\_labels=case\_when(covid\_deaths > 1900 ~ "Very High",

between(covid\_deaths,1000,1900) ~ "High",

between(covid\_deaths,500,1000) ~ "Low",TRUE ~"Very Low"))

summary(as.factor(data1$covid\_deaths\_labels))

#finding correlation

corrplot(cor(data[,3:12]))

#LDA

set.seed(123)

lda=lda(covid\_deaths\_labels ~ X + child\_mort + income+ exports + health + imports + inflation +total\_fer + pop\_dens , data=data1 )

ldapredict <- predict(lda)$class

mean(ldapredict==data1$covid\_deaths\_labels)

#QDA

qda=qda(covid\_deaths\_labels ~ X + child\_mort + income+ exports + health + imports + inflation +total\_fer + pop\_dens , data=data1 )

qdapredict <- predict(qda)$class

#Logistic Regression for multi-classification

test <- multinom(covid\_deaths\_labels ~ X + child\_mort + income + exports + health + imports + inflation + total\_fer + pop\_dens,data=data1)

glmpredict <- predict(test)

#Finding accuracy for models

mean(ldapredict==data1$covid\_deaths\_labels)

mean(qdapredict==data1$covid\_deaths\_labels)

mean(glmpredict==data1$covid\_deaths\_labels)

#confusion matrix

x1<-table(predicted =ldapredict ,actual= data1$covid\_deaths\_labels)

x2<-table(qdapredict , data1$covid\_deaths\_labels)

x3<-table(glmpredict , data1$covid\_deaths\_labels)

x1;x2;x3

#k-fold cv validation with k=10

set.seed(123)

t <- trainControl(method = "cv", number = 5)

#for LDA

ld <- train(covid\_deaths\_labels ~ X+child\_mort + income+ exports + health + imports + inflation + total\_fer + pop\_dens,

data = data1, method = "lda",trControl = t)

ld$results$Accuracy

#for QDA

qd <- train(covid\_deaths\_labels ~ X+child\_mort + income+ exports + health + imports + inflation + total\_fer + pop\_dens,

data = data1, method = "qda",trControl = t)

qd$results$Accuracy

#for Logistic regression

lg <- train(covid\_deaths\_labels ~X+ child\_mort + income+ exports + health + imports + inflation + total\_fer + pop\_dens,

data = data1, method ="multinom",trControl = t)

lg$results$Accuracy