MINING INTO THE ROOT CAUSES OF VIOLENCE AGAINST WOMEN: A CASE STUDY OF THE MACRO-FACTORS

Dhruvisha Gosai (JC810547) – James Cook University

Abstract

Over the past two hundred years women's rights have seen unparalleled progress in almost all. Despite bridging this gap there is still a significant way to go before women's rights are objectively matched with those of men. A similar trend can be observed for almost all progressive policies, from education, poverty, and unemployment. This correlation is not by coincidence – strong correlations between gender and racial equality in a country are closely tied to the education and economic strength of that country. The Human Development Index (HDI) sourced from United Nations Development Programme that aims to address a wide range of critical performance indicators to measure each country's position relative to each other in terms of poverty and inequality.

This paper aims to extend on the research used to produce this key performance metrics and conduct analysis that helps strengthen the importance of already established 'common day' knowledge against macro-factors, and to find how the metrics that ultimately separates a country with a low HDI – to one with a high HDI score.

These objectives are realized through use of supervised and unsupervised modelling methods and met with partial success in most instances.

Introduction

Violence against women (VAW) and girls is a global issue where on average one in three women are beaten, raped, or otherwise abused in her lifetime with the abuser most likely someone known to her (Moradian, 2009). This crime against women historically wasn't something widely talked about but has become common place in modern day vernacular. In some cultures, it is taboo to even discuss the violent actions endured by women from their intimate or non-intimate partners due to societal and family pressure of being outcast. While more than 60% of incidents go unreported, each year the reported violence cases range between 960,000 and 3,000,000 ("Domestic Violence Statistics", 2018).

As of today, human rights have become more progressive, prior to the mid-1800s the majority of the legal systemd subliminally accepted wife-beating as a legitimate form of authority practice by husbands towards their wives (Daniels & Brooks, 1997).

It wasn't until the end of 1870s that most courts in the US unanimously opposed to the right of domestic violence practiced by husbands (Green, 1876) and by 1920, wife beating was made illegal in all states of the US. It took another 90 years before the establishment of a global organization – UN Women (in 2010) to accelerate progress on meeting needs of women and advocating gender equality.

As of today, there are 49 countries that don't have laws to protect women against domestic violence and a part of the problem can be attributed to the macro-level factors of a country such as gender inequality index, low literacy rates, and socio-economic status (Jahan, 2018).

This creates a foundation for three main objectives for the research:

- Multivariate analysis for dimensionality reduction and data visualization using PCA.
    - This will help identify which macro-societal factors are related to the presence of Low and High HDI

- Clusters of countries grouped together by macro-level factors which have reported VAW using hierarchical clustering.
    - Identifying characteristics that countries share may result in unexpected grouping of countries
- Factors in determining the odds of decade been before 2009 or not.
    - To find if there is a statistically significant relationship between time and progression of ideals in recent decades

Data

As this research involves multiple types of macro data for each country over a time-period, there were 8 datasets sourced from (Human Development Data Center | Human Development Reports, 2021) for 8 macro-level factors reported for each country –

- Demography: *Sex ratio at birth (male to female births)*
- Education: *Literacy rate, adult (% ages 15 and older)*
- Gender: *Violence against women ever experienced (% of female population ages 15 and older)*
- Gender: *Gender Inequality Index (GII)*
- Poverty*: Population living below income poverty line, national poverty line (%)*
- Socio-economic sustainability: *Ratio of education and health expenditure to military expenditure*
- Work, employment, and vulnerability: *Unemployment, total (% of labor force)*
- Human Development Index: *HDI*

Sourced data was a wide table with volumes for individual factors for each year for every country. Data needed to be transposed using the pivot_longer() function from the dplyr package to have countries with volumes for each year – 3 columns for each dataset. Once transposed, year had to converted to numeric to drop trailing characters present prior to the data cleansing stage and to make computations easier. All datasets were merged using inner_join() function by both 'Country' and 'Year' to avoid cartesian join. However, because violence and poverty had records for only as 2019, data had to be joined by just 'Country' for this instance. It was assumed that violence and poverty % remained same for the time-period –between 1999 to 2019.

Once the joins were made and data was pre-processed, the final dataset consisted of 6180 rows with 12 variables (before omitting NAs). List below provides additional information on formats and descriptions of the fields present in the final dataset.

Field Name	Derived	Description	Format
Country		Country	Character
Year		Year	Numeric
HDI		HDI score for each country	Numeric
Violence		% VAW recorded for intimate and non-intimate partners for female population ages 15 and older - result of transpose	Numeric
Poverty		% Population living below income poverty line, national poverty line - result of transpose	Numeric
Expenditure		Ratio of education and health vs. military expenditure - result of transpose	Numeric
Literacy		% Literacy rate, adult ages 15 and older - result of transpose	Numeric
Sex_ratio		Sex ratio of male to female births - result of transpose	Numeric
Inequality		Gender Inequality Index (GII) - result of transpose	Numeric
Unemployment		% Unemployment of labor force - result of transpose	Numeric
Decade_Band	Derived - binary	10-year band created with range 1999 to 2009 and 2010 to 2019 for variable **Year**	Character
HDI_Band	Derived - binary	HDI below or equal to 0.55 are classified as under-developed and over 0.55 are classed as developed countries using **HDI**	Character

Figure 1: List of variables used for analysis.

To conduct the analysis with randomized data, set.seed() function was used and additional data preparation steps were involved- before the data mining techniques.

1. PCA & clustering – data was grouped by the 'Country' and mean value was taken for each numeric variable grouped by year. For both PCA and hierarchical clustering, data was first summarizing and populated with mean values for all numeric variables by country using *dplyr* library, followed by a listwise deletion of the missing variables using na.omit().
2. For logistic regression, NA observations are omitted – leaving the original dataset of ~6K observations down to 2790 which were then standardised. Using set.seed() with 7789 value to replicate the results in

future, data was split into 80%-20% for train and test.

A summary of the sampled dataset is as below in Figure 3 –

```
> summary(full_data_x3)
   Country               year          F_HDI          F_violence       F_poverty
 Length:6180        Min.   :1990   Min.   :0.1920   Min.   : 0.00   Min.   : 0.40
 Class :character   1st Qu.:1997   1st Qu.:0.5337   1st Qu.: 0.00   1st Qu.:17.27
 Mode  :character   Median :2004   Median :0.6860   Median :20.00   Median :24.30
                    Mean   :2004   Mean   :0.6590   Mean   :20.82   Mean   :29.46
                    3rd Qu.:2012   3rd Qu.:0.7820   3rd Qu.:33.30   3rd Qu.:41.35
                    Max.   :2019   Max.   :0.9570   Max.   :93.00   Max.   :82.30
                                   NA's   :60                       NA's   :2100
 F_expenditure     F_literacy      F_sex_ratio     F_inequality     F_unemployment
 Min.   : 0.700   Min.   : 22.30   Min.   :1.000   Min.   :0.0000   Min.   : 0.110
 1st Qu.: 4.300   1st Qu.: 71.85   1st Qu.:1.040   1st Qu.:0.2348   1st Qu.: 0.960
 Median : 6.800   Median : 91.45   Median :1.050   Median :0.4139   Median : 1.170
 Mean   : 8.493   Mean   : 82.05   Mean   :1.053   Mean   :0.3882   Mean   : 1.431
 3rd Qu.:10.722   3rd Qu.: 97.83   3rd Qu.:1.060   3rd Qu.:0.5357   3rd Qu.: 1.490
 Max.   :66.500   Max.   :100.00   Max.   :1.170   Max.   :0.8190   Max.   :20.130
 NA's   :1770     NA's   :1680     NA's   :240     NA's   :870      NA's   :390
 Decade_Band          HDI_Band
 Length:6180        Length:6180
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Figure 2: Summary of the variables for final dataset before omitting NAs

Methods

All analysis and statistical procedures were performed in R studio version 1.4.1717 with a list of libraries referenced in Appendix 1.3 (RStudio Team, 2021).

Algorithm 1: PCA

Since PCA is a factor analysis that aims to examine interrelations among a set of variables, a ggpairs plot was created (figure 3) using ggpairs() from *GGally* package to understand the data spread. HDI and literacy were observed to have a correlation of 0.88 and, inequality and literacy with negative correlation of 0.81. With some additive noise, a single variable may suffice to explain the most important aspect of the data – those two variables are linearly correlated.

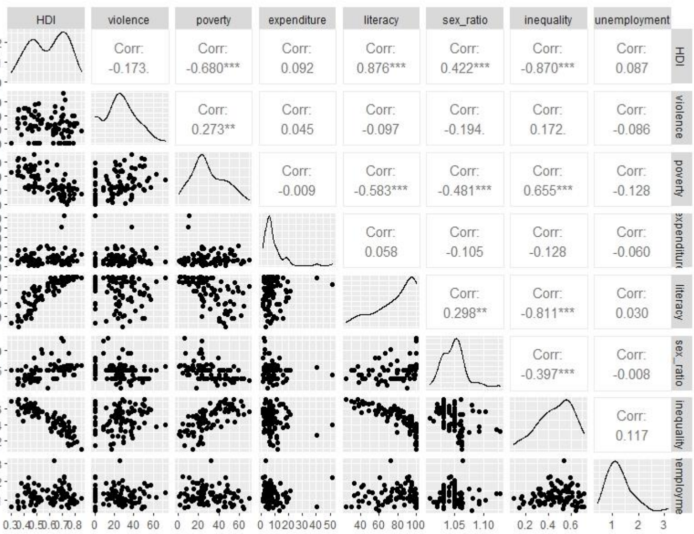


Figure 3: Correlation, density, and linearity check plot of numeric

When performing PCA using the prcomp() function, variables were standardized adding *center = T* and *scale = T* as part of the component analysis. This was done to assign higher weightage to variables with higher variances when later passed through to the PCA function.

A screeplot was produced using the screeplot() function inbuilt in R, and a cumulative variance plot were created to describe variability. Since the output from PCA provides corresponding eigenvalues of each PC, a specified threshold of eigenvalue < 1 was used to discard any variables that could not explain at least one variable's worth of the variability.

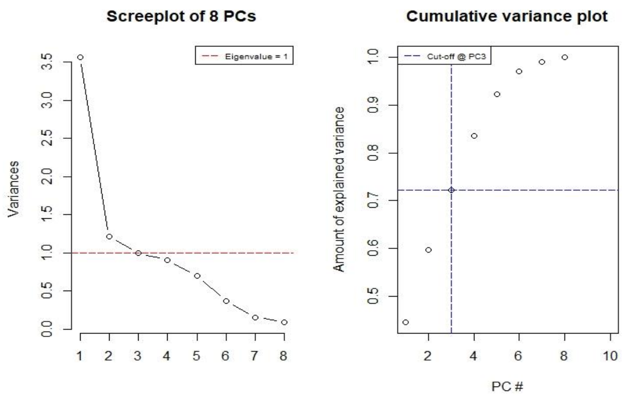


Figure 4: Screeplot and cumulative variance plot of PCA

A biplot of PCA, in addition to HDI band was created to explain the developed and under-developed countries based on only 2 dimensions. An additional biplot was created with 30 contributing attributes that helped determine the 2 dimensions.

Algorithm 2: Hierarchical Clustering (HC)

Prior to performing hierarchical clustering, a distance (dissimilarity) matrix was created using Euclidean distance and visualized using fviz_dist() from the factoextra package as hierarchical clustering requires the distance between each pair of observations.

To assess the agglomerative coefficient for some of the linking methods such as average, single, complete, and ward, a function was created to compute agglomerative coefficient. Of these different methods, the Ward approach resulted in the highest atomic vector (AC) of **0.9273**. For agglomerative hierarchical clustering, the agnes()

function was used with Ward method and Euclidian distance to create hierarchical clusters.

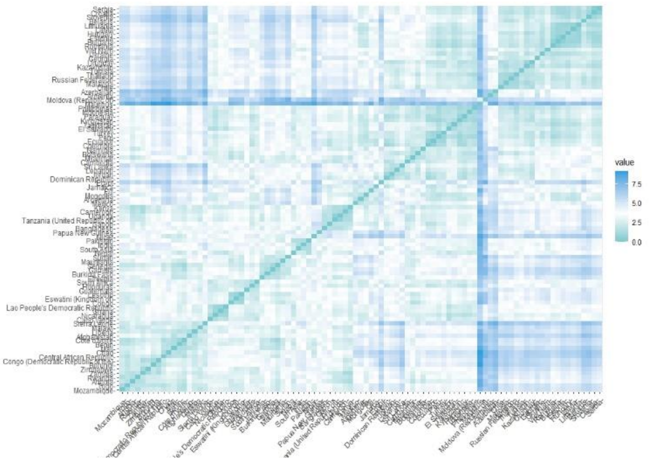


Figure 5: Dissimilarity matrix with Euclidean method

Using fviz_dend(), a dendrogram was created to visualise the clustering before creating the optimal cuts to the cluster. Three different methods were used to determine the optimal clusters – Elbow method, Silhouette method, and the Gap Statistics to ensure the cluster has high <u>intra</u>-class similarity and low <u>inter</u>-class similarity to output a cohesive and distinctive cluster.

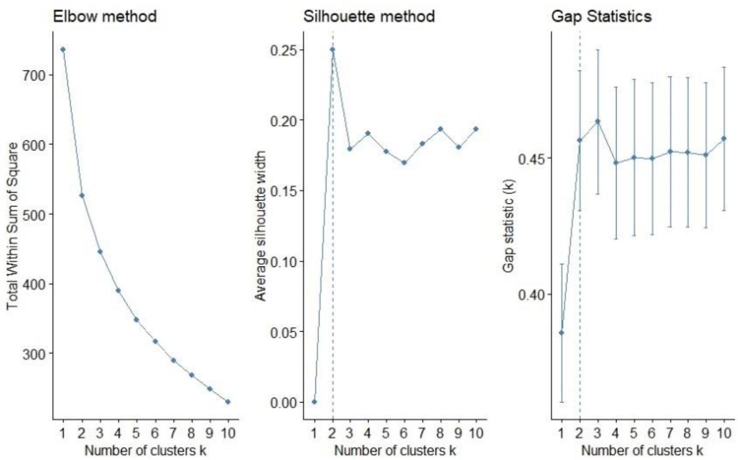


Figure 6: Determining optimal clusters using Elbow, Silhouette, and gap statistics

As particularly evident from the Silhouette method, a suitable K value for this clustering technique is decidedly 2 – characterised by the high peak and sudden drop from 2 onwards.

To cut the cluster into two groups the cutree() function was used with k=2. This subgroup was then mutated to the original dataset to include cluster numbers for each observation.

An additional graph using fviz_cluster() was produced to visualise the cluster results in a scatter plot.

<u>Algorithm 3: Logistic regression classification</u>

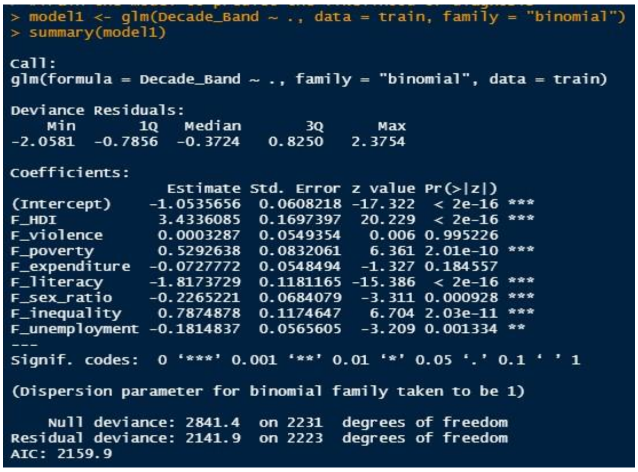
For this classification algorithm, data needed to be split into separated training and test datasets. 80% and 20% split was used for this purpose using the createDataPartition() function and a subsequent proportion table created to identify whether the data was balanced or not.

With no information available about the sampling scheme, it is assumed that the information collected for each year is independent of each other. Exploratory analysis was conducted to further make sure that assumptions were met before implementing the regression.

On plotting of correlation matrix using hierarchical clustering, no variables with correlation greater than 0.9 were highlighted.

Density plot for full data was produced to determine the distribution of the X by Y variable, none of the variables were normally distributed. However, because GLM is a more general class of linear models, it allows the use of linear model even when the dependent variable does not adhere to a normal bell-shape (Phillips, 2021).

The model was created using the decade band as the dependent variable, all other variables were used as the predictors - except HDI banding to avoid introducing bias resulting in a potentially overfitted model. The Glm() function was invoked and the parameter *family* specified as "binomial" used to create the model and predict() functions used to get the log odds, probabilities and confusion matrix for both training and test datasets.


```
> model1 <- glm(Decade_Band ~ ., data = train, family = "binomial")
> summary(model1)

Call:
glm(formula = Decade_Band ~ ., family = "binomial", data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0581  -0.7856  -0.3724   0.8250   2.3754

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.0535656  0.0608218 -17.322  < 2e-16 ***
F_HDI           3.4336085  0.1697397  20.229  < 2e-16 ***
F_violence      0.0003287  0.0549354   0.006 0.995226
F_poverty       0.5292638  0.0832061   6.361 2.01e-10 ***
F_expenditure  -0.0727772  0.0548494  -1.327 0.184557
F_literacy     -1.8173729  0.1181165 -15.386  < 2e-16 ***
F_sex_ratio    -0.2265221  0.0684079  -3.311 0.000928 ***
F_inequality    0.7874878  0.1174647   6.704 2.03e-11 ***
F_unemployment -0.1814837  0.0565605  -3.209 0.001334 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2841.4  on 2231  degrees of freedom
Residual deviance: 2141.9  on 2223  degrees of freedom
AIC: 2159.9
```
Figure 7: Logistic regression model summary

For a final performance assessment, ROC curve and AUC tests were carried out on the model using the prediction() and performance() functions.

Results and Discussion

Algorithm 1: PCA

With the first 3 components of eigenvalue > 1 observed in figure 4, it explains 72.24% of the variance; that is to say, dimensionality could effectively be reduced from 8 to 3 while only losing 27.76% of variance.

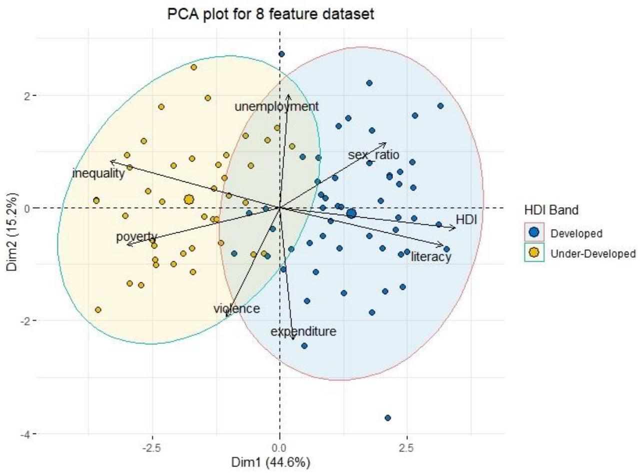


Figure 8: PCA biplot with HDI bands

Visualizing the two components of PCA, data was clearly separated between the two HDI bands – 'Developed' and 'Under-Developed'. The plot describes how each of the 8 variables influence the HDI band with inequality, poverty, and unemployment being highly attributed to the countries that are under-developed. Conversely, countries that are classed as *developed* possess attributes such as literacy, even sex ratio at birth, and expenditure on the education and health.

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.8894 1.1015 0.9979 0.9534 0.83419
Proportion of Variance 0.4462 0.1517 0.1245 0.1136 0.08698
Cumulative Proportion  0.4462 0.5979 0.7224 0.8360 0.92297
                          PC6     PC7    PC8
Standard deviation     0.6099 0.39427 0.2980
Proportion of Variance 0.0465 0.01943 0.0111
Cumulative Proportion  0.9695 0.98890 1.0000
```

Figure 9: PCA summary

Based on the near Venn-diagram portrayed in figure 8, an additional HDI group of overlapping countries could be categorized as developing.

Algorithm 2: Hierarchical Clustering (HC)

Produced from the Hierarchical Clustering dendrogram as shown in Figure 10, each leaf corresponds to an individual observation, combined into different branches fused at a higher level. The higher the height between fused clusters, indicates the higher amount of dissimilarity between grouped clusters. Consequently, observations that are separated by a vast distance in the fused lines, are separated by a proportional distance in the dataset.

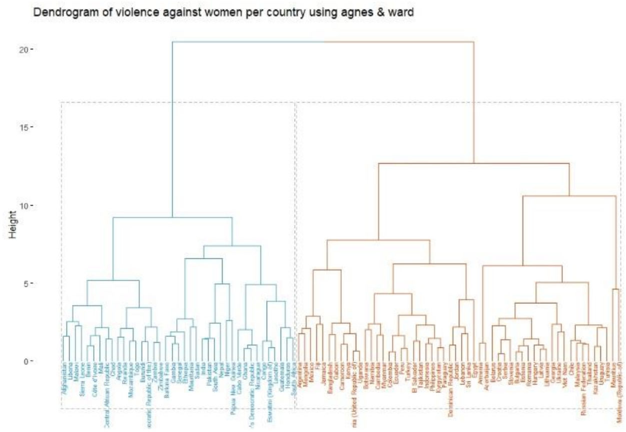


Figure 10: Hierarchical clustering dendrogram using agnes & ward

Using the gap statistics and Silhouette method, a K value of 2 was determined to be the optimum cluster cut for clusters. The majority of cluster 1 consisted of the under-developed countries and cluster 2 with developed countries. However, average violence against women in both the clusters wasn't significantly too different. Cluster 1 with 24.9% VAW compared to 25.54% for cluster 2.

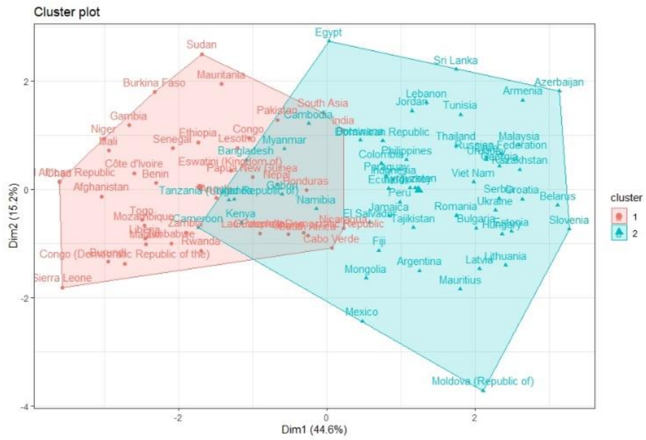


Figure 11: Scatter plot of clusters from hierarchical clustering

Algorithm 3: Logistic regression classification

Given the split of decade band being disproportionate with 66.67% observations falling under older decade '1990-2009' band compared to only 33.33% for the newer decade of '2010-2019' band, there is a possibility of bias towards the older decade. As expected, logistic regression classified older decade more accurately for test data than newer decade as witnessed in figure 11 where only 18.63% observations were correctly classified for newer decade ‘2010-2019’.

```
> confusionMatrix(as.factor(preds_test_1), train$Decade_Band, positive = "2010 - 2019")
Confusion Matrix and Statistics

             Reference
Prediction    1990 - 2009 2010 - 2019
  1990 - 2009        1291         328
  2010 - 2019         197         416

               Accuracy : 0.7648
                 95% CI : (0.7466, 0.7822)
    No Information Rate : 0.6667
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4464

 Mcnemar's Test P-Value : 1.398e-08

            Sensitivity : 0.5591
            Specificity : 0.8676
         Pos Pred Value : 0.6786
         Neg Pred Value : 0.7974
             Prevalence : 0.3333
         Detection Rate : 0.1864
   Detection Prevalence : 0.2746
      Balanced Accuracy : 0.7134

       'Positive' Class : 2010 - 2019
```

From the model summary, other than violence and expenditure on education and heath weren’t significant variables at all and could have been discarded to improve model performance. HDI, poverty and gender inequality index had positive odds compared to unemployment, sex ratio, and literacy with negative odds in determining the decade of the year. Confusion matrix output above shows model accuracy of 0.7648 which is much less than anticipated and this model shouldn’t be reliably used to determine the decade band with sensitivity of 0.5591. This means that the model can correctly classify for newer decade only 55.9% of the time.

With AUC value of 0.8245 and ROC curve observed below in figure 12, model seems to be workable but not perfectly accurate at classifying what decade when provided the odds of other significant variables.

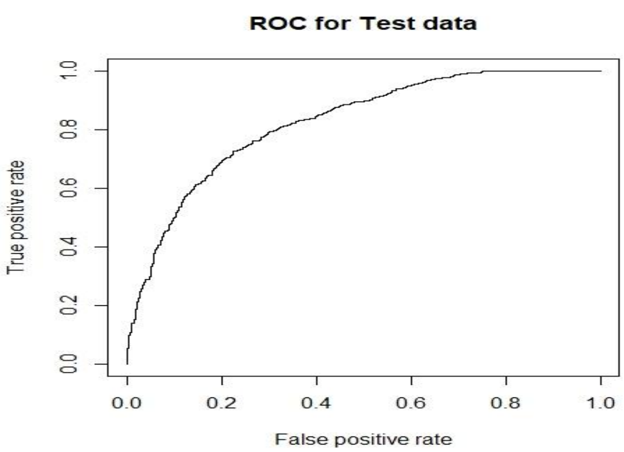


Figure 12: ROC curve for logistic model

Concluding Remarks

The research conducted in this paper extends on a long line of well-established data exploration into determining what metrics impact a country’s view on progressive matters but intends to offer fresh perspective from an unsupervised clustering technique to categorize countries in an atypical way. Additionally, through use of PCA, understanding the common trends between countries that are lower on the HDI spectrum highlights the importance of ensuring that education funding is prioritized and efforts are made to reduce the variance between the births of girls and boys (such as that found in countries where there are 1-2 child policies – or where the birth of a male child is highly regarded as males are generally seen as the breadwinners in the community).

Future research may benefit from having access to a more ‘raw’ data recorded over the same period of time instead of violence against women just recorded once in last 10 years. As one of the limitations in this research was the dependency on using data that was for the most part already summarized to a high level, using a low-level, unrefined data may open more avenues for variable creation and subsequently new hypotheses, and new relationships to be explored in the dataset.

References

"Domestic Violence Statistics". (2018). *The Gateway Center For Domestic Violence Services.* Oregon: City of Portland. Retrieved from The Gateway Center For Domestic Violence Services.

Daniels, C., & Brooks, R. (1997). Feminists Negotiate the State. In *Feminists Negotiate the State: The Politics of Domestic Violence* (pp. 5–10). Lanham: University Press of America.

Green, N. (1876). In *Criminal law reports being reports of cases determined in the federal and state courts of the United States, and in the courts of England, Ireland, Canada, etc.* New York: Hurd and Houghton.

Human Development Data Center | Human Development Reports. (2021). Retrieved from Hdr.undp.org: http://hdr.undp.org/en/data

Jahan, S. (2018, November 19). *Violence against women, a cause and consequence of inequality*. Retrieved from hdr.undp.org: http://hdr.undp.org/en/content/violence-against-women-cause-and-consequence-inequality

Moradian, A. (2009, September). Domestic Violence against Single and Married Women in Iranian Society. *Tolerancy International*.

Phillips, N. (2021). *YaRrr! The Pirate's Guide to R. [online]*. Retrieved from Bookdown.org: https://bookdown.org/ndphillips/YaRrr/regression-on-non-normal-data-with-glm.html

RStudio Team. (2021). RStudio: Integrated Development Environment for R. *RStudio, PBC*. Boston, MA: URL http://www.rstudio.com/. Retrieved from RStudio, PBC: http://www.rstudio.com/

UN Women: The United Nations Entity for Gender Equality and the Empowerment of Women. (2021). Retrieved from UN Women : https://www.un.org/youthenvoy/2013/07/un-women-the-united-nations-entity-for-gender-equality-and-the-empowerment-of-women/

Appendix

Built with R Version:** 4.0.5**

```
# Install packages if not available already
if (!require("car")) install.packages("car")
if (!require("datasets")) install.packages("datasets")
if (!require("ggplot2")) install.packages("ggplot2")
if (!require("dplyr")) install.packages("dplyr")
if (!require("tidyverse")) install.packages("tidyverse")
if (!require("qqplotr")) install.packages("qqplotr")
if (!require("ggfortify")) install.packages("ggfortify")
if (!require("ggthemes")) install.packages("ggthemes")
if (!require("hrbrthemes")) install.packages("hrbrthemes")
if (!require("ISLR")) install.packages("ISLR")
if (!require("caret")) install.packages("caret")
if (!require("GGally")) install.packages("GGally")
if (!require("knitr")) install.packages("knitr")
if (!require("MASS")) install.packages("MASS")
if (!require("ROCR")) install.packages("ROCR")
if (!require("corrplot")) install.packages("corrplot")
if (!require("ggridges")) install.packages("ggridges")
if (!require("klaR")) install.packages("klaR")
if (!require("psych")) install.packages("psych")
if (!require("yaml")) install.packages("yaml")
if (!require("cluster")) install.packages("cluster")
if (!require("factoextra")) install.packages("factoextra")
if (!require("reshape2")) install.packages("reshape2")
if (!require("broom")) install.packages("broom")
if (!require("aod")) install.packages("aod")
if (!require("ggpubr")) install.packages("ggpubr")
if (!require("gridExtra")) install.packages("gridExtra")
if (!require("fpc")) install.packages("fpc")
```

```r
if (!require("readr")) install.packages("readr")
if (!require("dendextend")) install.packages("dendextend")
if (!require("tibble")) install.packages("tibble")
if (!require("ggforce")) install.packages("ggforce")
if (!require("FactoMineR")) install.packages("FactoMineR")

# Loading relevant R packages
library(car, warn.conflicts = F, quietly = T)
library(datasets, warn.conflicts = F, quietly = T)
library(ggplot2, warn.conflicts = F, quietly = T)
library(MASS, warn.conflicts = F, quietly = T)
library(dplyr, warn.conflicts = F, quietly = T)       # for piping
library(tidyverse, warn.conflicts = F, quietly = T)
library(qqplotr, warn.conflicts = F, quietly = T)     # for qq plots
library(ggfortify, warn.conflicts = F, quietly = T)   # for visualisations
library(ggthemes, warn.conflicts = F, quietly = T)    # for ggplot themes
library(hrbrthemes, warn.conflicts = F, quietly = T)  # for ggplot background themes
library(ISLR, warn.conflicts = F, quietly = T)        # for data
library(caret, warn.conflicts = F, quietly = T)       # for splitting the data
library(GGally, warn.conflicts = F, quietly = T)
library(knitr, warn.conflicts = F, quietly = T)       # to add appendix in the end
library(ROCR, warn.conflicts = F, quietly = T)
library(corrplot, warn.conflicts = F, quietly = T)    # Correlation matrix
library(ggridges, warn.conflicts = F, quietly = T)
library(klaR, warn.conflicts = F, quietly = T)
library(psych, warn.conflicts = F, quietly = T)       # Visualise
library(yaml, warn.conflicts = F, quietly = T)
library(cluster, warn.conflicts = F, quietly = T)
library(factoextra, warn.conflicts = F, quietly = T)  # clustering visualization
library(reshape2, warn.conflicts = F, quietly = T)    # reshaping data
library(broom, warn.conflicts = F, quietly = T)
library(aod, warn.conflicts = F, quietly = T)         # for wald test
library(ggpubr, warn.conflicts = F, quietly = T)
library(gridExtra, warn.conflicts = F, quietly = T)
library(fpc, warn.conflicts = F, quietly = T)
library(readr, warn.conflicts = F, quietly = T)
library(dendextend, warn.conflicts = F, quietly = T) # for comparing dendrograms
library(tibble, warn.conflicts = F, quietly = T)
library(ggforce, warn.conflicts = F, quietly = T) # PCA graph
library(FactoMineR, warn.conflicts = F, quietly = T) # PCA

#-------------------------------#
#          IMPORT DATA          #
#-------------------------------#

Below_poverty_line_population <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data M
ining/Assignment 3 - Capstone project/Import_data/Below_poverty_line_population.csv"
                                          ,header = TRUE
                                          ,sep=",")

Education_health_expenditure <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mi
ning/Assignment 3 - Capstone project/Import_data/Education_health_expenditure.csv"
                                         ,header = TRUE
                                         ,sep=",")

Literacy_rate <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mining/Assignment
3 - Capstone project/Import_data/Literacy_rate.csv"
                             ,header = TRUE
                             ,sep=",")

Sex_ratio <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mining/Assignment 3 -
Capstone project/Import_data/Sex_ratio.csv"
                                ,header = TRUE
                                ,sep=",")

Gender_Inequality_Index <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mining/
Assignment 3 - Capstone project/Import_data/Gender_Inequality_Index.csv"
                                    ,header = TRUE
                                    ,sep=",")

HDI <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mining/Assignment 3 - Capst
one project/Import_data/HDI.csv"
                ,header = TRUE
                ,sep=",")

Unemployment_rate <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mining/Assign
ment 3 - Capstone project/Import_data/Unemployment_rate.csv"
                              ,header = TRUE
                              ,sep=",")
```

```
Violence_against_women <- read.csv("D:/Dhru Folder/JCU - Master of Data science/MA5810 - Introduction to Data Mining/A
ssignment 3 - Capstone project/Import_data/Violence_against_women.csv"
                                   ,header = TRUE
                                   ,sep=",")

#---------------------------------------------------------#
#  Data Transformation: Clean and transform as required   #
#---------------------------------------------------------#

#---------------------------
# Transpose the data to join
poverty_transpose     <- Below_poverty_line_population %>%
    pivot_longer(-Country, names_to = "year", values_to = "poverty") %>%
    transform(year_clean = substr(year,2,5))
expenditure_transpose <- Education_health_expenditure %>%
    pivot_longer(-Country, names_to = "year", values_to = "expenditure") %>%
    transform(year_clean = substr(year,2,5))
literacy_transpose  <- Literacy_rate %>%
    pivot_longer(-Country, names_to = "year", values_to = "literacy") %>%
    transform(year_clean = substr(year,2,5))
sex_ratio_transpose     <- Sex_ratio %>%
    pivot_longer(-Country, names_to = "year", values_to = "sex_ratio") %>%
    transform(year_clean = substr(year,2,5))
inequality_transpose  <- Gender_Inequality_Index %>%
    pivot_longer(-Country, names_to = "year", values_to = "inequality") %>%
    transform(year_clean = substr(year,2,5))
HDI_transpose         <- HDI %>%
    pivot_longer(-Country, names_to = "year", values_to = "HDI") %>%
    transform(year_clean = substr(year,2,5))
unemployment_transpose <- Unemployment_rate %>%
    pivot_longer(-Country, names_to = "year", values_to = "unemployment") %>%
    transform(year_clean = substr(year,2,5))
violence_transpose    <- Violence_against_women %>%
    pivot_longer(-Country, names_to = "year", values_to = "violence") %>%
    transform(year_clean = substr(year,2,5))
    # mutate(violence_flag = ifelse(is.na(violence), "Missing", "Reported"))

# Creating year variable as numeric and dropping year_clean column
poverty_transpose$year      <- as.numeric(poverty_transpose$year_clean,replace = T)
poverty_transpose           <- subset(poverty_transpose, select = -c(year_clean, year)) #Drop column

expenditure_transpose$year  <- as.numeric(expenditure_transpose$year_clean,replace = T)
expenditure_transpose       <- subset(expenditure_transpose, select = -c(year_clean)) #Drop column3

literacy_transpose$year     <- as.numeric(literacy_transpose$year_clean,replace = T)
literacy_transpose          <- subset(literacy_transpose, select = -c(year_clean)) #Drop column

sex_ratio_transpose$year    <- as.numeric(sex_ratio_transpose$year_clean,replace = T)
sex_ratio_transpose         <- subset(sex_ratio_transpose, select = -c(year_clean)) #Drop column

inequality_transpose$year   <- as.numeric(inequality_transpose$year_clean,replace = T)
inequality_transpose        <- subset(inequality_transpose, select = -c(year_clean)) #Drop column

HDI_transpose$year          <- as.numeric(HDI_transpose$year_clean,replace = T)
HDI_transpose               <- subset(HDI_transpose, select = -c(year_clean)) #Drop column

unemployment_transpose$year <- as.numeric(unemployment_transpose$year_clean,replace = T)
unemployment_transpose      <- subset(unemployment_transpose, select = -c(year_clean)) #Drop column

violence_transpose$year     <- as.numeric(violence_transpose$year_clean,replace = T)
violence_transpose_new      <- subset(violence_transpose, select = -c(year_clean, year)) #Drop column

# Creating a single table with data for HDI, deaths and affected -> cleaning the year to remove x
full_data_x1 <-   left_join(HDI_transpose, violence_transpose_new, by=c("Country"="Country")) %>%
    left_join(.,poverty_transpose,              by=c("Country"="Country")) %>%
    left_join(.,expenditure_transpose,          by=c("year" = "year","Country"="Country")) %>%
    left_join(.,literacy_transpose,           by=c("year" = "year","Country"="Country")) %>%
    left_join(.,sex_ratio_transpose,              by=c("year" = "year","Country"="Country")) %>%
    left_join(.,inequality_transpose,           by=c("year" = "year","Country"="Country")) %>%
    left_join(.,unemployment_transpose,         by=c("year" = "year","Country"="Country"))


full_data_x2 <- full_data_x1 %>%
    group_by(Country) %>%
    mutate(New_HDI          = ifelse(mean(HDI, na.rm = T)          < 0,NA,mean(HDI, na.rm = T)) ,
           New_violence     = ifelse(mean(violence, na.rm = T)     < 0,NA,mean(violence, na.rm = T)),
           New_poverty      = ifelse(mean(poverty, na.rm = T)      < 0,NA,mean(poverty, na.rm = T)),
```

```r
           New_expenditure  = ifelse(mean(expenditure, na.rm = T)  < 0,NA,mean(expenditure, na.rm = T)),
           New_literacy     = ifelse(mean(literacy, na.rm = T)     < 0,NA,mean(literacy, na.rm = T)),
           New_sex_ratio    = ifelse(mean(sex_ratio, na.rm = T)    < 0,NA,mean(sex_ratio, na.rm = T)),
           New_inequality   = ifelse(mean(inequality, na.rm = T)   < 0,NA,mean(inequality, na.rm = T)),
           New_unemployment = ifelse(mean(unemployment, na.rm = T) < 0,NA,mean(unemployment, na.rm = T)) )

full_data_x3 <- full_data_x2 %>%
    group_by(Country,year) %>%
    dplyr::mutate(F_HDI          = ifelse(!is.na(HDI),         HDI,         ifelse(!is.na(New_HDI)         ,New_HDI,
NA)),
                  F_violence     = ifelse(!is.na(violence),    violence,    ifelse(!is.na(New_violence)    ,New_violen
ce,     NA)),
                  F_poverty      = ifelse(!is.na(poverty),     poverty,     ifelse(!is.na(New_poverty)     ,New_povert
y,      NA)),
                  F_expenditure  = ifelse(!is.na(expenditure), expenditure, ifelse(!is.na(New_expenditure) ,New_expend
iture,  NA)),
                  F_literacy     = ifelse(!is.na(literacy),    literacy,    ifelse(!is.na(New_literacy)    ,New_litera
cy,     NA)),
                  F_sex_ratio    = ifelse(!is.na(sex_ratio),   sex_ratio,   ifelse(!is.na(New_sex_ratio)   ,New_sex_ra
tio,    NA)),
                  F_inequality   = ifelse(!is.na(inequality),  inequality,  ifelse(!is.na(New_inequality)  ,New_inequa
lity,   NA)),
                  F_unemployment = ifelse(!is.na(unemployment),unemployment,ifelse(!is.na(New_unemployment),New_unempl
oyment, NA)),
                  Decade_Band    = case_when(year >= 1990 & year <= 2009 ~ "1990 - 2009",
                                             year >= 2010 & year <= 2019 ~ "2010 - 2019"),
                  HDI_Band       = case_when(F_HDI <= 0.55  ~ "Under-Developed",
                                             F_HDI  > 0.55  ~ "Developed") ) %>%
    dplyr::select(Country,
                  year,
                  F_HDI,
                  F_violence,
                  F_poverty,
                  F_expenditure,
                  F_literacy,
                  F_sex_ratio,
                  F_inequality,
                  F_unemployment,
                  Decade_Band,
                  HDI_Band)

summary(full_data_x3)

#--------------------------------------#
# PCA % Clustering data - summarised
#--------------------------------------#
cluster_pca_df <- full_data_x3 %>%
    group_by(Country) %>%
    summarise(HDI = mean(F_HDI),
              violence = mean(F_violence),
              poverty = mean(F_poverty),
              expenditure = mean(F_expenditure),
              literacy = mean(F_literacy),
              sex_ratio = mean(F_sex_ratio),
              inequality = mean(F_inequality),
              unemployment = mean(F_unemployment)) %>%
    mutate(HDI_Band       = case_when(HDI <= 0.55  ~ "Under-Developed",
                                      HDI  > 0.55  ~ "Developed")) %>%
    dplyr::select(Country,
                  HDI,
                  violence,
                  poverty,
                  expenditure,
                  literacy,
                  sex_ratio,
                  inequality,
                  unemployment,
                  HDI_Band)

cluster_pca_df <- na.omit(cluster_pca_df)# listwise deletion of missing
cluster_pca_df_nocountry <- data.frame(column_to_rownames(cluster_pca_df, var = "Country")) # made countries to be row
names

#--------------------------------------#
# ALGORITHM 1: PCA  #
#--------------------------------------#

pca_df <- cluster_pca_df_nocountry
```

```r
pca_df$cluster <- as.factor(pca_df$HDI_Band)

# Correlation Matrix to explore existing correlation
M <- round(cor(pca_df[,1:8]), 2) # Create the correlation matrix
corrplot(M,order="hclust", tl.cex = 0.90, method = 'square', type = 'lower', diag = FALSE) # Create corr plot

# plot variables to understand the spread
gg <- GGally::ggpairs(pca_df[,1:8])
                      # , upper = "blank")
gg

pc <- prcomp(pca_df[ ,1:8], center = T, scale = T)
summary(pc) # Get the summary of pca - first 3 components explain 70.55% of the variance, whereas
# the second component explains the remaining 29.45%

par(mfrow=c(1,2))
# dimensionality can be reduced from 8 to 6 as 2 components have Eigenvalue > 1
# that explains almost 90% of variance - while only “loosing” about 10% of variance
screeplot(pc, type = "l", npcs = 8, main = "Screeplot of 8 PCs")
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"),
       col=c("red"), lty=5, cex=0.6)

cumpro <- cumsum(pc$sdev^2 / sum(pc$sdev^2))
plot(cumpro[0:10], xlab = "PC #", ylab = "Amount of explained variance", main = "Cumulative variance plot")
abline(v = 3, col="blue", lty=5)
abline(h = 0.7224, col="blue", lty=5)
legend("topleft", legend=c("Cut-off @ PC3"),
       col=c("blue"), lty=5, cex=0.6)

# Principal components + tree
par(mfrow=c(1,1))
fviz_pca_biplot(pc, geom.ind = "point", pointshape = 21,
                pointsize = 2,
                fill.ind = pca_df$HDI_Band,
                col.ind = "black",
                palette = "jco",
                addEllipses = TRUE,
                label = "var",
                col.var = "black",
                repel = TRUE,
                legend.title = "HDI Band") +
    ggtitle("PCA plot for 8 feature dataset") +
    theme(plot.title = element_text(hjust = 0.5))

# Change the color by groups, add ellipses
fviz_pca_biplot(pc, label="var",
                select.ind = list(contrib = 30),
                col.ind = "black",
                palette = "jco")+
    ggtitle("Biplot of variables and 30 contributing observations") +
    theme(plot.title = element_text(hjust = 0.5))


#---------------------------------------#
# ALGORITHM 2: HIERARCHIAL CLUSTERING  #
#---------------------------------------#

cluster_df_scaled <- scale(cluster_pca_df_nocountry[,1:8]) # standardize variables

head(cluster_df_scaled, n=6)

# For reproducibility
set.seed(7789)

# Get distance with default Euclidean (others possible)
dMatrix <- dist(cluster_df_scaled, method="euclidean")
# Visualise distance matrix
fviz_dist(dMatrix, gradient = list(low = "#00AFBB",
                                   mid = "white", high = "#2E9FDF"))

# linking methods to test
measure <- c( "average", "single", "complete", "ward")
names(measure) <- c( "average", "single", "complete", "ward")
# function to compute agglomerative coefficient
ac <- function(x) {
    agnes(cluster_df_scaled, method = x)$ac
}
#map function that transforms the input by applying a function
```

```
#to each element of a list or atomic vector and returning
#an object of the same
map_dbl(measure, ac) # Ward's method identifies the strongest clustering structure of the four methods assessed.

# Create clusters
hc_agnes_ward <- agnes(cluster_df_scaled,metric = "euclidean", method = "ward")
hc_agnes_ward$ac
pltree(hc_agnes_ward, cex = 0.6, hang = -1,
       main = "Dendrogram of Violence against women in countries using agnes & ward")

# Cut in 2 groups and color by groups
fviz_dend(hc_agnes_ward, k = 2, # Cut in four groups
          cex = 0.5, # label size
          #Colour choice
          k_colors = c("#2E9FDF", "#FC4E07"),
          color_labels_by_k = TRUE, # color labels by groups
          rect = TRUE # Add rectangle around groups
          )+
    ggtitle("Dendrogram of violence against women per country using agnes & ward")

# Ward's method
hc_ward <- hclust(dMatrix, method = "ward.D2" )
plot(hc_ward, cex = 0.6, main="Dendrogram of violence against women per country using ward")
rect.hclust(hc_ward, k = 2, border = 2:5)

# Elbow method
p1<-fviz_nbclust(cluster_df_scaled, FUN = hcut, method = "wss")+
    ggtitle("Elbow method")
# Silhouette method
p2<-fviz_nbclust(cluster_df_scaled, FUN = hcut, method = "silhouette")+
    ggtitle("Silhouette method")

p3<- fviz_gap_stat(clusGap(cluster_df_scaled, FUN = hcut, nstart = 25, K.max = 10, B = 50))+
    ggtitle("Gap Statistics")

# Display plots side by side
gridExtra::grid.arrange(p1, p2, p3, nrow = 1)


# Cut tree into 2 groups/clusters
sub_grp <- cutree(hc_ward, k = 2)
cluster_df_new <- cluster_pca_df_nocountry %>%
                    mutate(cluster = sub_grp)

cluster_df_new2 <- tibble::rownames_to_column(cluster_df_new, "Country")

# Number of members in each cluster
table(sub_grp)

# Better interpretation of cluster
fviz_cluster(list(data = cluster_df_scaled, cluster = sub_grp))+
    theme_bw()

as.matrix(table(as.factor(cluster_df_new2$HDI_Band),as.factor(cluster_df_new2$cluster)))

#---------------------------------------#
# ALGORITHM 3: Logistic regression     #
#---------------------------------------#
# Violence against women in developed and underdeveloped countries is equal

# Test for correlation
#***********************
logistic_df <- na.omit(full_data_x3)# listwise deletion of missing# remove diagnosis for correlation matrix
M <- round(cor(logistic_df[,3:10]), 2) # Create the correlation matrix

# Remove highly correlated variables to improve model performance
highlyCor <- colnames(M)[findCorrelation(M, cutoff = 0.9)]
logistic_df <-logistic_df[, which(!colnames(logistic_df) %in% highlyCor)]

corrplot(M,order="hclust", tl.cex = 0.90, method = 'square', type = 'lower', diag =  FALSE) # Create corr plot

logistic_df_scaled <- data.frame(scale(logistic_df[,3:10])) # standardize variables

new_df <- logistic_df %>%
    dplyr::select(Country, year, HDI_Band, Decade_Band)

logistic_df <- cbind(new_df[,3:4], logistic_df_scaled)

logistic_df <- logistic_df %>% dplyr::select(-HDI_Band)
```

```r
# Test for distribution
#************************
#Plot histograms of variables group by diagnosis - Is data normally distributed? (does not affect glm)
# gg <- GGally::ggpairs(logistic_df[,4:12])
# gg

#split into training (80%) and test
set.seed(7789)
split <- createDataPartition(logistic_df$Decade_Band, p = 0.8, list = F)

train <- logistic_df[split, ]
test <- logistic_df[-split, ]

c(nrow(train), nrow(test)) # print number of observations in test vs. train
table(train$Decade_Band) %>% prop.table()*100 # Proportion (in %) by Diagnosis

train$Decade_Band <- as.factor(train$Decade_Band)
#Train the model to predict the likelihood of diagnosis
model1 <- glm(Decade_Band ~ ., data = train, family = "binomial")
summary(model1)

# Make predictions on test data
lodds_1  <- predict(model1, train, type = "link")#log odds
probs_1  <- predict(model1, train, type = "response")#probabilities
preds_1 <- ifelse(lodds_1 > 0,  "2010 - 2019", "1990 - 2009") #using log odds
confusionMatrix(as.factor(preds_1), train$Decade_Band, positive = "2010 - 2019")

## Make predictions on test data
lodds_test_1  <- predict(model1, train, type = "link")#log odds
probs_test_1  <- predict(model1, train, type = "response")#probabilities
preds_test_1 <- ifelse(lodds_test_1 > 0,   "2010 - 2019", "1990 - 2009") #using log odds
confusionMatrix(as.factor(preds_test_1), train$Decade_Band, positive = "2010 - 2019")

# AUC on Test data
print(paste("AUC for Test accuracy using logistic regression is: ",
            prediction(probs_test_1, train$Decade_Band) %>%
                performance(measure = "auc") %>%
                .@y.values
))
# ROC on test data
prediction(probs_test_1, train$Decade_Band) %>%
    performance(measure = "tpr", x.measure = "fpr") %>%
    plot(main = "ROC for Test data")
#----------------------------
```