September 2019

Probabilistic Modeling of Democracy, Corruption, Hemophilia A and Prediabetes Data

A. K. M. Raquibul Bashar
University of South Florida

Probabilistic Modeling of Democracy, Corruption, Hemophilia A and Prediabetes Data

by

A K M Raquibul Bashar

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics and Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Chris P. Tsokos, Ph.D.
Khandethody M. Ramachandran, Ph.D.
Lu Lu, Ph.D.
Getachew A. Dagne, Ph.D.

Date of Approval:
July 19, 2019

Keywords: DIS, CPI, F8, Hemophilia, $K - Means$, Inhibitor, Severity, Diabetes, Prediabetes, Forest, SVM

Dedication

This doctoral dissertation is dedicated to my father, mother, youngest maternal - uncle, my daughter and wife, for their unconditional support in all my endeavors.

Acknowledgements

First and foremost, I would like to thank God Almighty for blessings of being patient and courageous throughout this journey. I am very deeply grateful to my dissertation advisor Dr. Chris P. Tsokos for his extremely valuable advice and directions without which this study would have been impossible.

I am also grateful to the committee members, Dr. Ram, Dr. Dagne, Dr. Lu Lu.

My heartiest appreciation also goes to all the faculties and staff of the department of mathematics who contributed in diverse ways to make my experience a token of memory at University of South Florida. I am fully grateful to all my friends here at USF, especially, Neranga Fernando, NawaRaj Pokhrel, Mahdi Goudarzi, Mohammed Abu Sefa, Tharu Bhikari, Ram C. Kafle and Mariam Habadi for immense emotional support and encouragement I have experienced from them throughout this whole journey.

Most importantly, my heartfelt gratitude goes to my mother Mrs. Rokeya Begum, my father Md. Ali Akbar Khan (Retd. Flight Sergeant, BAF), my uncle Khandaker Ashraf Ali, my wife Monzora Bashar and my cousin Md. Selim Hossain Khan. I wouldn't express my debt of gratitude to them in words. Their incessant love, support, encouragement were the fuel to keep going forward during this endeavor.

Table of Contents

List of Tables

List of Figures

Abstract

Parametric analysis of any real-world data is the most powerful tool to characterize the probabilistic behavior in social, economic, medical, epidemiological, and other areas of study. In the present study, we identify the theoretical Probability Distribution Function(PDF) for Democracy Index Scores (DIS) from the Economist Intelligence Unit (EIU) database and estimate the maximum likelihood estimates of the theoretical PDFS. We also identify the individual PDFs for each of the clusters, Full Democracy, Flawed Democracy, Hybrid Regime, and Authoritarian Regime defined by the Economist Intelligence Unit (EIU).

A statistical model is a convenient instrument to predict the future value of any real phenomenon. In addition to identifying probability distributions, we predict the DIS for 167 countries of the world through a regression model with a high degree of accuracy. Then we do cluster analysis through ($K-means$) clustering algorithm based on the DIS predicted by the corresponding statistical model we have developed.

By extracting Corruption Perception Index (CPI) and World Governance Index (WGI) from Transparency International (TI) and World Bank (WB) databases respectively, we estimate a theoretical PDF of CPI for 175 countries of the world. Moreover, we estimate individual PDFs for each of the clusters - Highly Corrupted, Moderately Corrupted, Fairly Corrupted, and Least Corrupted countries of the world.

We conducted statistical analyses on Hemophilia **A** based on the data retrieved from Centers for Disease Control and Prevention (CDC) CHAMP $F8$ surveillance program to identify the risk factors involved in Severity level of Hemophilia A. We have identified a statistical model for probability prediction of the Severity level of Hemophilia A.

Finally, we have studied some standard machine learning algorithms to compare and identify the best algorithm to classify and predict the correct state of a prediabetes condition in individuals. For this present study, the data was extracted from the National Health and Nutrition Examination Surveys (NHANES), part of the Centers for Disease Control and Prevention (CDC). We compare the identified champion algorithm to the existing machine learning algorithms suggested by some researchers in other countries of the world.

1 Parametric Analysis of Economist Intelligence Units (EIU) Democracy Index Scores (DIS) of 167 Countries in the world

1.1 Introduction

Democracy has been defined in hundreds of ways. However, almost all definitions fit under one of four major types: economic, social, communitarian, or political democracy. Economic, social, and communitarian democracy tend to be defined in terms of outcomes: the equalization of wealth, income, and status, or the creation and maintenance of a feeling of belonging in a community or communities, and the promotion of participation within them. Political democracy is different because it is almost necessarily defined by its procedures and institutions rather than its outcomes. Political democracy does not promise economic equality, social justice, or a feeling of community; whatever outcomes result from political democracy are consistent with this kind of democracy as long as the proper procedures produced them. Procedural political democracy is in itself divided into sub-types [14].

All national states today have a form of representative democracy. Representation is so common that we tend to forget that there is an alternative: direct participatory democracy, in which voters make policy decisions themselves instead of electing representatives to decide for them. Representative democracy, in turn, can vary between a popular sovereignty tendency and more liberal versions. In a popular sovereignty democracy, the majority rules: whatever the people want becomes the law. Liberal democracy limits the power of the majority by guaranteeing some fundamental rights of individuals (and sometimes groups) and by creating constitutional checks on executive, legislative,

and judicial powers. This set of types and sub-types is neither exhaustive nor universally accepted; one could make additional distinctions in the set of liberal representative democracies to distinguish consolidated democracies from transitional ones, parliamentary from 3 presidential democracies, unitary from federal democracies, high-quality from low-quality democracies, and so on. However, this basic typology is useful for describing how political scientists have faced the challenge of measuring democracy.

1.2 EIUs Measure of Democracy

The Economist Intelligence Unit's of democracy [32], on a 0 to 10 scale, is based on the ratings for 60 indicators grouped in five categories mentioned above. Each category has a rating on a 0 to 10 scale, and the overall index of democracy is the simple average of the five category indexes. The category indexes are based on the indicator scores in the category converted to 0 to 10 scale. Adjustments to the category scores are made if countries do not score a 1 in the following critical areas of democracy:

- Whether national elections are free and fair
- The security of the voters
- The influence of foreign powers on government
- The capability of the civil service to implement policies

If the scores for the first three questions are 0 (or 0.5), one point (0.5 points) is deducted from the index in the relevant category (either the electoral process and pluralism or the functioning of government). If the score for 4 is 0, one point is deducted from the functioning of the government category index. The index values are used to place countries within one of four types of regimes:

(1) Full democracies— scores of 8- 10

(2) Flawed democracies— scores of 6 to 7.9

(3) Hybrid regimes— scores of 4 to 5.9

(4) Authoritarian regimes— scores below 4

Threshold points for regime types depend on overall scores that are rounded to one decimal point. Based on the scores defining different kinds of regimes for any country to be fell in the definitions are elaborated in detail. For the scoring system, the EIU has used a combination of a dichotomous and a three-point scoring system for the 60 indicators. According to their claim, a dichotomous 1-0 scoring system (1 for yes and 0 for no) has some drawbacks, but it has several distinct advantages over more refined scoring scales (such as 1-5 or 1-7). Also, they say, for many indicators, the possibility of a 0.5 score is introduced, to capture grey areas where a simple yes or no is problematic with guidelines as to when that should be used. Thus for many indicators, there is a three-point scoring system, which represents a compromise between simple dichotomous scoring and the use of more beautiful scales. They also declare that a crucial, differentiating aspect of their measure is that in addition to experts' assessments they use, where available, public opinion surveys- mainly the World Values Survey (Say, WVS). Indicators based on the surveys predominate heavily in the political participation and political culture categories, and a few are used in the civil liberties and functioning of government categories. In addition to the WVS, other sources that can be leveraged include the Euro-barometer surveys, Gallup polls, Asian Barometer, Latin American Barometer, Afro barometer, and national surveys.

Given below, **Figure 1.1** is the schematic diagram of the complete data set that we have used for pdf estimation.

As part of our preliminary preparation of the dataset, we have checked to see that the data was randomly collected to determine if there is any biasness, and it does not contain any outliers. So, after these tests as mentioned earlier, we proceeded to find the

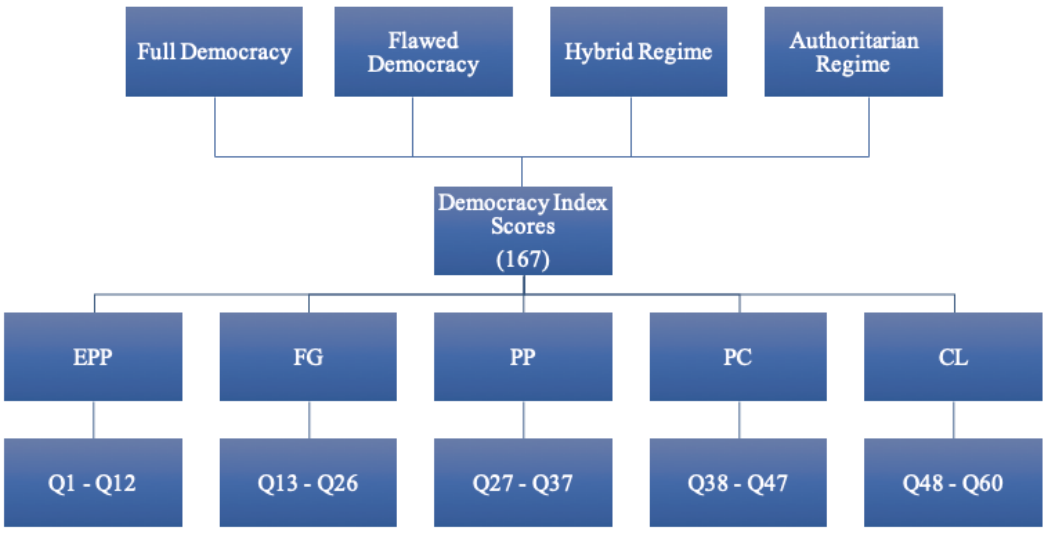


Figure 1.1: Data Diagram of Democracy Index Score (DIS)

best Probability Distribution Functions (PDF) of all the DIS scores and each of the four classifications of Democracy, namely, Full, Flawed, Hybrid, and Authoritarian Regime.

1.2.1 Finding PDF of Democracy Index Scores (DIS)

In the process of finding the best-fitted PDF, we have implemented the methodology of graphing the variable DIS, which will give us an initial idea of what the distribution may look like [55]. Then we shall identify the best candidates for the PDF that characterizes the subject variable. The following, **Table 1.1**, shows the primary statistic of the variable democracy scores (DIS) of 167 countries of the world.

Table 1.1: Descriptive Statistic of DIS of 167 Countries of the World

Descriptive Statistics of DIS Countries				
Mean	Median	Std. Deviation	Skewness	Kurtosis
5.548	5.792	2.177	-0.082	-1.034

From **Table 1.1** above, we see that the average (mean) democracy index score for all the countries of the world is 5.548 and the standard deviation is approximately 2.18. It should be noted that the data is slightly left skewed with skewness value of -0.08153. A

histogram of the scores also supports the same information provided in the descriptive statistics information.

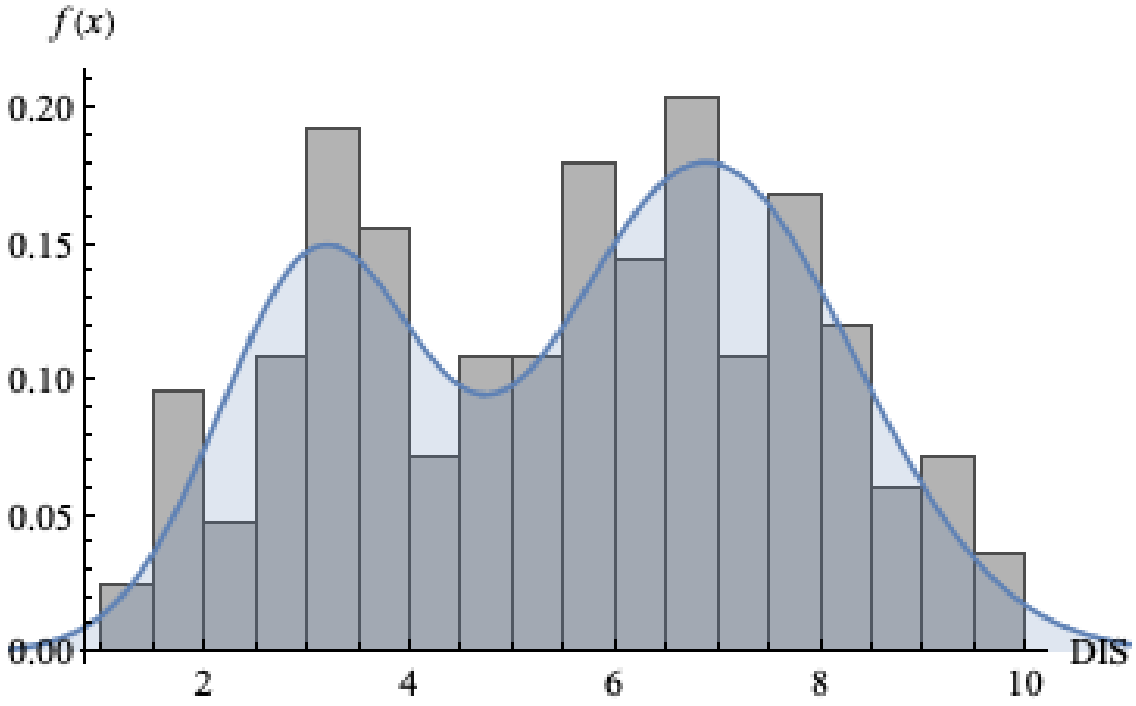


Figure 1.2: Histogram of Democracy Index Scores

1.2.1.1 Goodness-of-Fit tests for All DIS:

We proceeded by testing the goodness-of-fit for a number of well defined PDFs using three statistical tests, namely, **K**olomogrov-**S**mirnov [35], **A**nderson-**D**arling [5] and **Chi-square** [12]. The Kolmogorov-Smirnov test is based on minimum difference estimation. The Anderson-Darling measures whether the data can be transformed into the uniform probability distribution and the Chi-square test for goodness-of-fit is a measure of relative error squared [51]. We have found that the **Mixed Gaussian** PDF best fits all the DIS data as it is supported by the results given in Table 1.2 below.

Table 1.2: Goodness-of-Fit Summary

	α	p-value	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.9993	Do Not Reject
Anderson-Darling	0.05	0.984	Do Not Reject
Chi-Squared	0.05	0.5268	Do Not Reject

Thus, we proceed to discuss and fit the Mixed Gaussian PDF of the DIS of 167 countries of the world.

1.2.1.2 PDF of Democracy Index Score (DIS):

After passing the data through the aforementioned three goodness-of-fit tests [19], the probability distribution that captures the characteristics of DIS the best is the "Mixed Gaussian Probability Density Function". A Gaussian mixture model [17] is parameterized by two types of values, the mixture component weights and the component means and variance/covariance. For a Gaussian mixture model with K components, the K^{th} component has a mean of μ_k and standard deviation of σ_k for the univariate case. In our case $K = 2$. The analytical structure is given by:

$$f(x) = \sum_{i=1}^{k} \phi_i N\left(x|\mu_i, \sigma_i^2\right),$$

$$\text{with,} \tag{1.1}$$

$$N\left(x|\mu_i, \sigma_i\right) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{(x-\mu_i)^2}{\sigma_i^2}\right), \quad -\infty \leq X \leq \infty$$

The mean and the variance is 5.554 and 4.912, respectively, with standard deviation of 2.216. Alternative analytical form of the PDF given in equation 1.1 has the following form of PDF:

$$f(x) = \begin{cases} \frac{\phi_1 e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1(\phi_1+\phi_2)} + \frac{\phi_2 e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2(\phi_1+\phi_2)} \\ 0, \qquad \text{otherwise} \end{cases} \tag{1.2}$$

For our data, the approximate maximum likelihood estimates (MLE) of the parameters (σ_i, μ_i, and ϕ_i) of 1.1 are given in the **Table 1.3** below:

Table 1.3: MLEs of Mixture Distribution fitted to DIS

MLEs of DIS scores					
$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\phi_1}$	$\hat{\phi_2}$
3.107	6.877	0.974	1.437	0.351	0.649

Thus, the estimated analytical form of the subject PDF is given by-

$$f(x) = \begin{cases} 0.144e^{-0.53(x-3.11)^2} + 0.18e^{-0.24(x-6.88)^2}, & 0 \leq X \leq 10 \\ 0, & \text{otherwise} \end{cases} \tag{1.3}$$

The graph of 1.3 is given below by **Figure 1.3**:

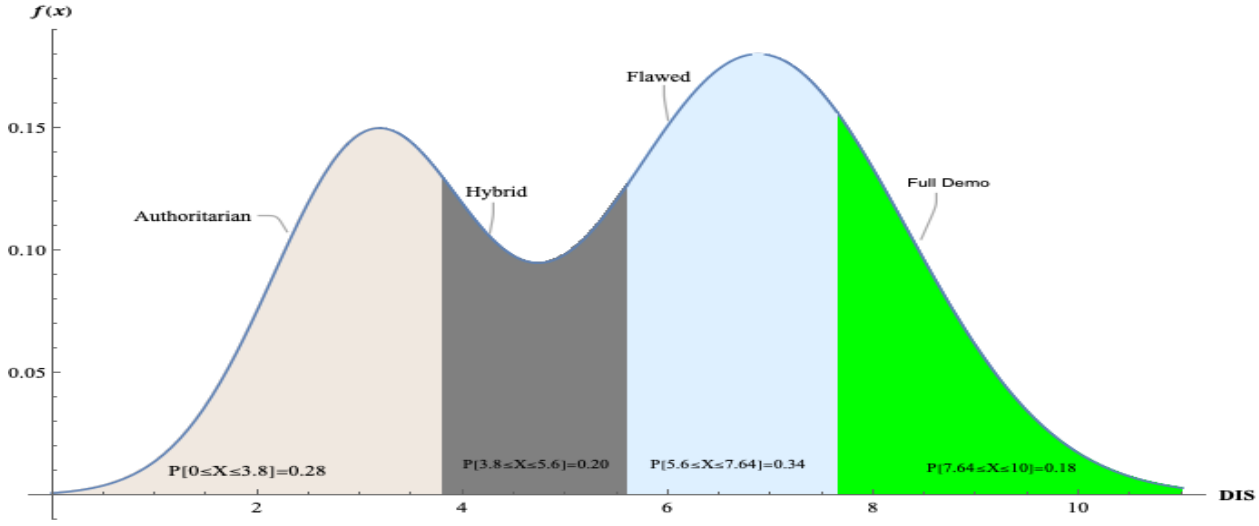


Figure 1.3: PDF plot of DIS (Mixed Gaussian PDF)

Thus, if a country was selected at random from the 167 countries, one can identify the probability of its classification of the four categories of Democracy. By using the plots given in **Figure 1.3**, one can easily identify the areas under the curve for each of the classes of democracy defined by EIU. For example, if anyone calculates the probability of DIS within the range of 7.64 to 10, then the corresponding probability would be the probability of any country falling in the 'Fully Democratic'category and so on. Furthermore, the moment generating function of 1.3 is given by

$$M_X(t) = 0.351e^{(3.11t+0.47t^2)} + 0.65e^{(6.88t+1.033t^2)} \tag{1.4}$$

The moment generating function (MGF) is given in the equation 1.4 can be used to calculate the moments of higher-order and consequently to calculated the mean and variance of the Mixed Gaussian PDF. Thus, if a country is selected at random from the population of 167 countries, we will expect its DIS to be 5.554. Also, we calculate the variance, $V[X] = 4.912$ and standard deviation, $STDV[X] = 2.216$. Note that these estimates are close to the basic statistics given in Table 1.1, which assures the quality of the fit of Mixed Gaussian PDF.

The Cumulative Distribution Function of the DIS is as follows:

$$F(x) = P(X \leq x) = \frac{\phi_1 \text{erfc}\left(\frac{\mu_1 - x}{\sqrt{2}\sigma_1}\right)}{2\left(\phi_1 + \phi_2\right)} + \frac{\phi_2 \text{erfc}\left(\frac{\mu_2 - x}{\sqrt{2}\sigma_2}\right)}{2\left(\phi_1 + \phi_2\right)} \tag{1.5}$$

where, θ_1 and θ_2 are 0.351 & 0.649 respectively.

The graph of cumulative distribution function is by following **Figure 1.4** below,

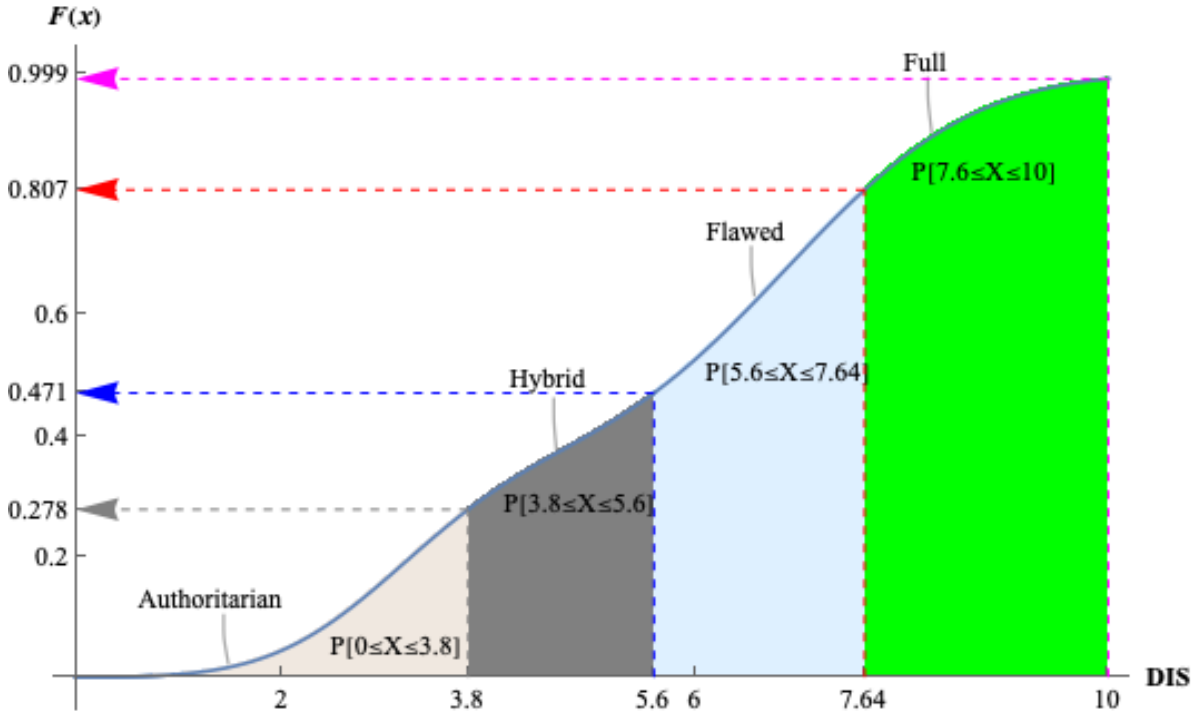


Figure 1.4: CDF plot of Democracy Index Scores

The **Figure 1.4**, is very useful in the cases, for example, if anyone wants to know the probability of any country will have a DIS *less than* 3.8 (i. e. $P[DIS \leq 3.8]$), then from the above figure it is shown that the probability would be 0.278 or approximately 28% of the areas under the cumulative probability distribution curve. Also, if we are curious about the likelihood of any country's DIS less than or equal to 5.6, then from the Figure 1.4, one can easily estimate it and the probability is approximately 0.81 or 81% area under the cumulative curve and so on.

Now we will proceed to find the PDF for each of the four classified categories of Democracy in the following sections.

1.2.2 PDF of "Fully Democratic"Countries

Here we shall proceed to find the probability distribution that characterizes the probabilistic behavior of only the DIS data for **Full Democracy**. To do this, we have implemented the same steps we have used in finding the overall PDF of DIS for all democracy

classifications. For this purpose, we have started with the basic descriptive statistics of **Fully Democratic countries**.

Table 1.4: Descriptive Statistics of DIS of Fully Democratic Countries

Descriptive Statistics of DIS of Fully Democratic Countries				
Mean	Median	Std. Deviation	Skewness	Kurtosis
8.429	8.168	0.633	0.779	-0.469

From the table above, we see that this subset of the overall data is slightly right-skewed with a value of 0.77913, and it has a mean of 8.4292. The histogram of the Full Democratic Countries is given below **Figure 1.5**. From this histogram, the implication is that we need to fit some sort of mixed probability distribution.

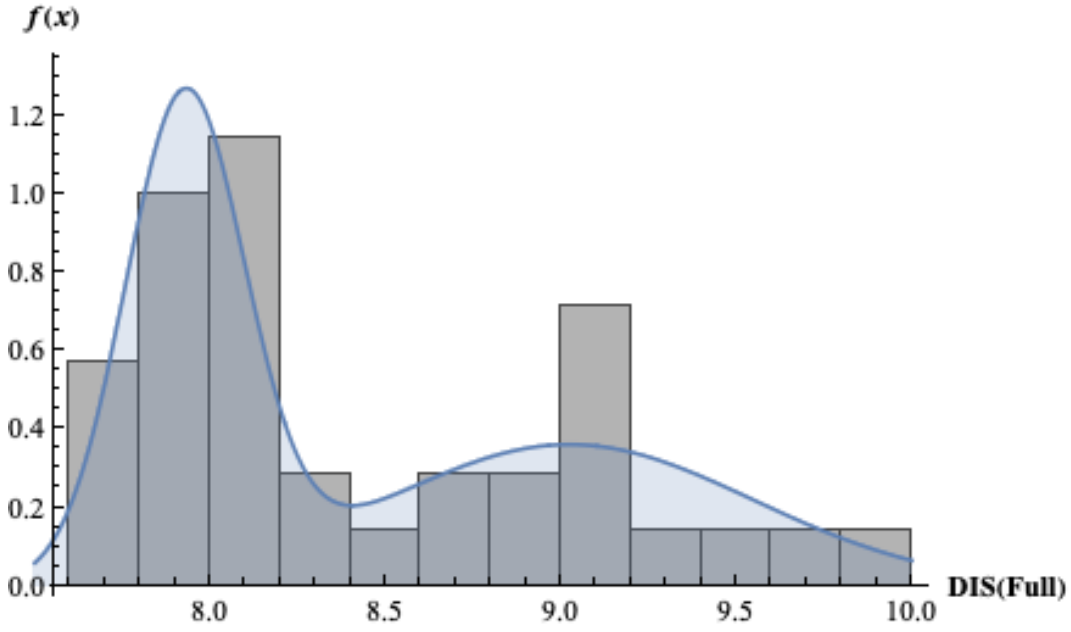


Figure 1.5: Fitted PDF to Histogram of Fully Democratic Countries

Using the three goodness-of-fit tests to the present data of fully democratic countries, we have identified that the data can be characterized probabilistically by the "Mixed distribution of 2– Gaussian PDF". The justification of this selection is confirmed by the three

methods of goodness-of-fit that we used in Table 1.5 given below confirms that the best pdf for the full democratic data is the **Mixed Gaussian PDF**.

Table 1.5: Goodness-of-Fit Summary for Fully Democratic Countries

	α	**p - value**	**Do Not Reject/Reject**
Kolmogorov-Smirnov	0.05	0.916	Do Not Reject
Anderson-Darling	0.05	0.986	Do Not Reject
Chi-Squared	0.05	0.9462	Do Not Reject

Thus, the fitted theoretical PDF of the subject data is given by-

$$f(x) = \sum_{i=1}^{k} \phi_i N\left(x|\mu_i, \sigma_i^2\right), \quad \text{Here,} \quad N\left(x|\mu_i, \sigma_i^2\right) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{(x-\mu_i)^2}{\sigma_i^2}\right) \tag{1.6}$$

The approximate MLEs of the parameters that drive the estimated Mixed Gaussian PDF are given by Table 1.6 below:

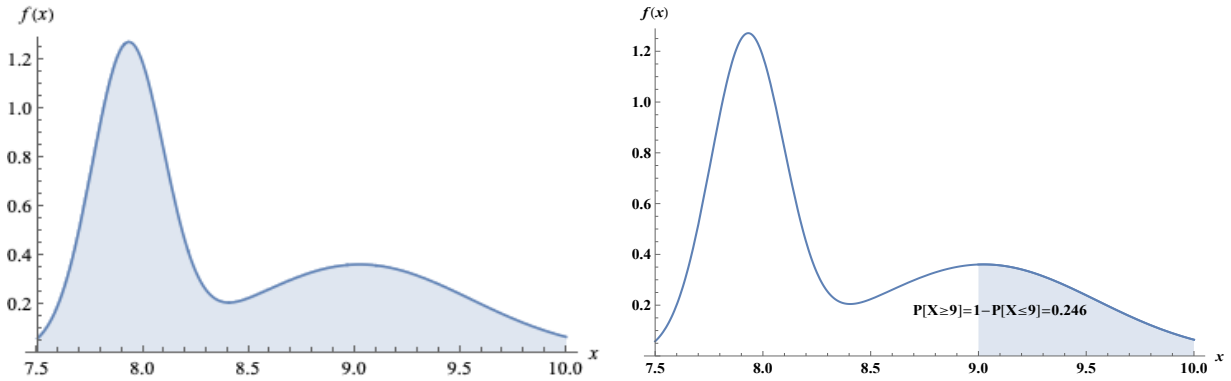
Table 1.6: MLEs of PDF of Fully Democratic Countries

MLEs of Fully Democratic Countries			
$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$
9.024	7.93	0.525	0.1704

also, $\sum_{i=1}^{2} \hat{\phi_i} = \hat{\phi_1} + \hat{\phi_2} = 0.53 + 0.47 = 1.00$, are the weights of 2- Gaussian PDF of the mixed distribution. Thus, the analytical structure of the estimated PDF of Fully Democratic countries of the world is given by

$$f(x) = \begin{cases} 0.36e^{-1.8(x-9.02)^2} + 1.23e^{-17.21(x-7.93)^2}, & 7.6 \leq X \leq 10 \\ 0, & \text{otherwise} \end{cases} \tag{1.7}$$

The graph of the PDF of 1.7 is given in Figure (1.6a) below.



(a) PDF of Fully Democratic Countries **(b)** PDF with shaded area for $P(X \geq 9)$

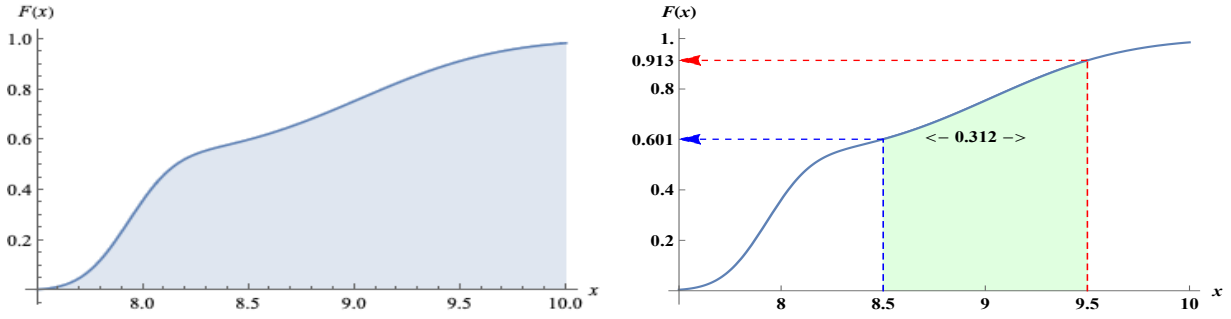
Figure 1.6: Plotting PDF of DIS of Fully Democratic Countries of the World

The expected value and variance of the Fully Democratic data subset is 8.4482 and 0.4453, respectively. That is, if a country is selected at random from this cluster we expect it's DIS will be approximately 8.45. Also, the probability that a country will have a DIS of more than 9 is 0.246 as shown in **Figure (1.6b)**.

The CDF of the *Fully Democratic* countries of the world is given by-

$$F(x) = P(X \leq x) = \frac{1}{4}\text{erfc}\left(\frac{\mu_1 - x}{\sqrt{2}\sigma_1}\right) + \frac{1}{4}\text{erfc}\left(\frac{\mu_2 - x}{\sqrt{2}\sigma_2}\right) \tag{1.8}$$

The graph of $F(x)$ in equation 1.8 is given below by Figure (1.7a):



(a) CDF of Fully Democratic Countries **(b)** CDF with shaded area for $P(8.5 \leq X \leq 9.5)$

Figure 1.7: Plotting CDF of DIS of Fully Democratic Countries of the World

The plotting of **Figure 1.7** is very useful in the case if anyone wants to estimate the probability of any country selected at random from this subset of the population and curious about the probability of that country will have a score more than 8.5 but less than 9.5 *i.e.*$P(8.5 \leq X \leq 9.5) = 1 - P(X \leq 8.5) - [1 - P(X \leq 9.5)]$, then that probability is **0.312** as shown in Figure 1.7b.

1.2.3 PDF of "Flawed Democratic"Countries

We shall now proceed to find the probability distribution that characterizes the probabilistic behavior of only the DIS data for **Flawed Democracy**. To do this, we have implemented the same steps we have used in finding the overall PDF of DIS for all democracy classifications. For this purpose, we have started with the basic descriptive statistics of **Flawed Democratic countries**.

Table 1.7: Descriptive Statistics of Flawed Democratic Countries

Descriptive Statistics of Flawed Democratic Countries				
Mean	Median	Std. Deviation	Skewness	Kurtosis
6.665	6.672	0.5592	0.0745	-1.0085

From Table 1.7 above, we see that this subset of the overall data has a mean 6.67. The histogram of the subject dataset is given below Figure 1.8. From this histogram, the

implication is that we need to fit some sort of mixed probability distribution for this data subset as well.

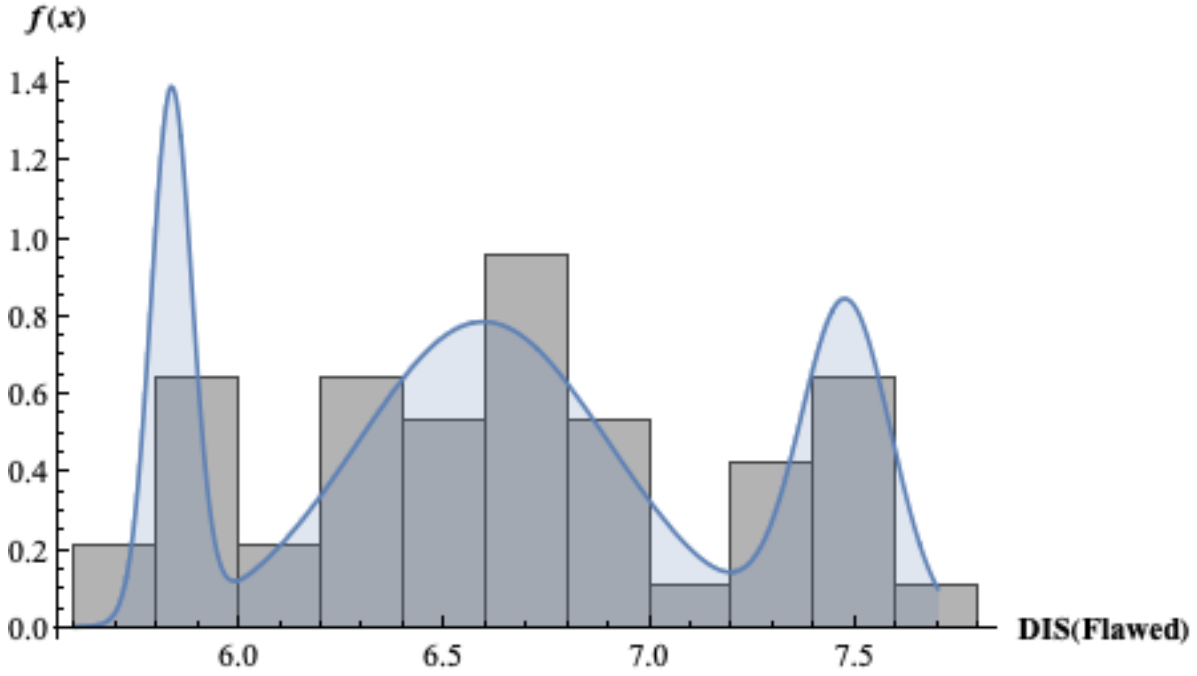


Figure 1.8: Fitted PDF to Histogram of Flawed Democratic Countries

Using the three goodness-of-fit tests to the present data of Flawed democratic countries, we have identified that the data can be characterized probabilistically by the Mixed distribution of 3– Gaussian PDF.

The justification of this selection is confirmed by the three methods of goodness-of-fit that we used in Table 1.8 given below confirms that the best pdf for the Flawed democratic data is **Mixed of 3 - Gaussian PDF**.

Table 1.8: Goodness-of-Fit Summary for Flawed Democratic Countries

	α	**p - value**	**Do Not Reject/Reject**
Kolmogorov-Smirnov	0.05	0.996	Do Not Reject
Anderson-Darling	0.05	0.999	Do Not Reject
Chi-Squared	0.05	0.964	Do not Reject

Thus, the fitted theoretical PDF of the subject data is given by-

$$f(x) = \sum_{i=1}^{k} \phi_i N\left(x|\mu_i, \sigma_i\right)$$

Here,

$$N\left(x|\mu_i, \sigma_i\right) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{(x-\mu_i)^2}{\sigma_i^2}\right), \tag{1.9}$$

Also, $k = 3$ and $\sum_{i=1}^{k} \phi_i = 1$ as well.

The approximate MLEs of the parameters that drive the estimated Mixed Gaussian PDF are given by Table 1.9 below:

Table 1.9: MLEs of PDF of Flawed Democratic Countries

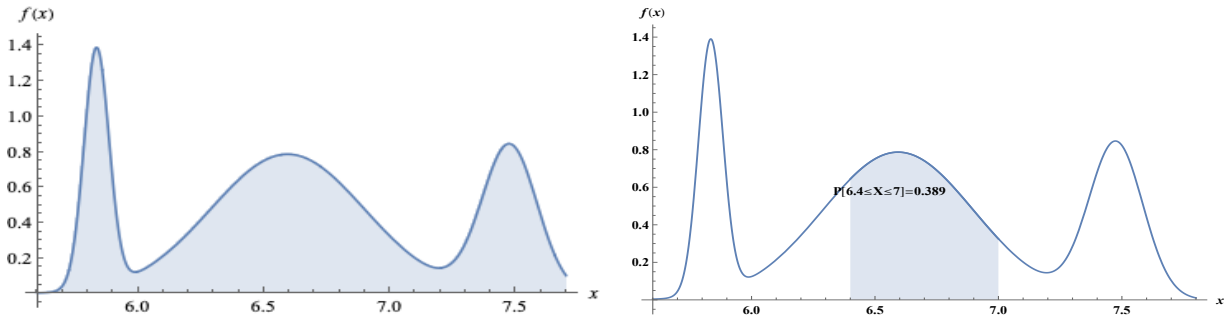
$\hat{\mu_1}$	$\hat{\mu_2}$	$\hat{\mu_3}$	$\hat{\sigma_1}$	$\hat{\sigma_2}$	$\hat{\sigma_3}$
7.475	6.594	5.834	0.109	0.305	0.049

And at the same time the analytical structure of the parameterized probability density function estimated from the data is given in the equation 1.10 as follows:

$$\begin{aligned} f(x) =& 0.83e^{-41.77(x-7.48)^2} + 0.79e^{-5.36(x-6.59)^2} \\ &+ 1.353e^{-202.8(x-5.84)^2}, \quad 5.6 \leq X \leq 7.8 \end{aligned} \tag{1.10}$$

Also, the weights for each of the Gaussian density estimated from the data are $\phi_1 = 0.168518$ $\phi_2 = 0.602757$ and $\phi_3 = 0.228725$ that makes $\sum_{i=1}^{3} \phi_i = 1$.

The graph of the PDF of 1.10 is given in Figure 1.9 below.



(a) PDF of Flawed Democratic Countries **(b)** PDF with shaded area for $P(6.4 \leq X \leq 7)$

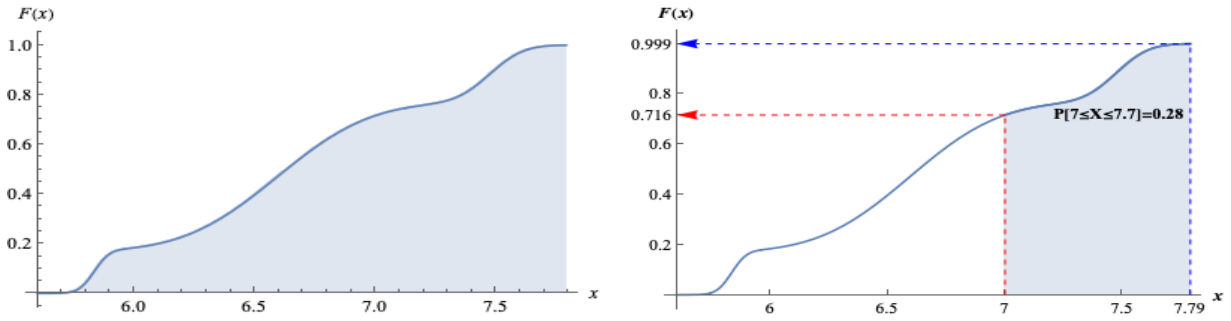
Figure 1.9: Plotting PDF of DIS of Flawed Democratic Countries of the World

The expected value and variance of Flawed Democratic data subset is $E(x) = 6.667$ and $V(X) = 0.328$ respectively and this value closely match with the values given in Table 1.7. Moreover, if a country is selected at random from this cluster, we expect it's DIS will be approximately 6.67. Also, the probability that a country will have a DIS between 6.4 and 7.00 would be approximately 0.4, as shown in Figure (1.9b).

The CDF of the Flawed Democratic DIS is given by-

$$F(x) = P(X \leq x) = \frac{\phi_1 \text{erfc}\left(\frac{\mu_1 - x}{\sqrt{2}\sigma_1}\right)}{2\left(\phi_1 + \phi_2 + \phi_3\right)} + \frac{\phi_2 \text{erfc}\left(\frac{\mu_2 - x}{\sqrt{2}\sigma_2}\right)}{2\left(\phi_1 + \phi_2 + \phi_3\right)} + \frac{\phi_3 \text{erfc}\left(\frac{\mu_3 - x}{\sqrt{2}\sigma_3}\right)}{2\left(\phi_1 + \phi_2 + \phi_3\right)} \tag{1.11}$$

It's graph is given by Figure (1.10a) as follows:



(a) CDF of Flawed Democratic Countries **(b)** CDF with shaded area for $P(X \geq 7)$

Figure 1.10: Plotting CDF of DIS of Flawed Democratic Countries of the World

From the figures given above, we can extract some handy information. Such as the probability of any country's DIS is greater or equal to 7 would be approximately 0.283 as shown in **Figure 1.10b**

1.2.4 PDF of "Hybrid Democratic"Countries

We shall now proceed to find the probability distribution that characterizes the probabilistic behavior of only the DIS data for Hybrid Democracy. To do this, we have implemented the same steps we have used in finding the overall PDF of DIS for all democracy classifications. For this purpose, we have started with the basic descriptive statistics of *Hybrid Regime* countries.

Table 1.10: Descriptive Statistic of Hybrid Democratic Countries

Descriptive Statistics of Hybrid Democratic Countries				
Mean	Median	Std. Deviation	Skewness	Kurtosis
4.96	5.013	0.668	-0.0597	2.113

The table 1.10, describes the basic descriptive statistics of the data subset of the Hybrid democratic countries of the world. The sample mean of this subset is 4.9621. Now the histogram of the Hybrid democratic countries are given in **Figure 1.11**:

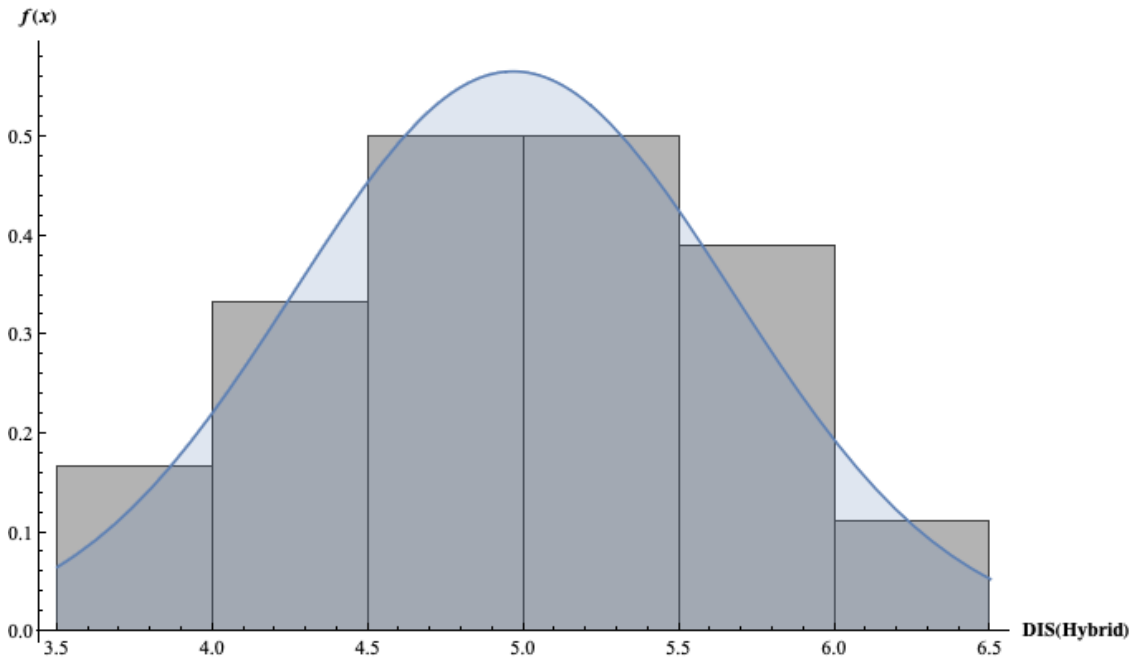


Figure 1.11: Fitted PDF to Histogram of Hybrid Democratic Countries

From Figure 1.11, it can be implied the best-fitted PDF is bell-shaped Normal PDF. The justification for this selection is confirmed by the three methods of goodness-of-fit that we used in Table 1.11. Given below confirms that the best pdf for the full democratic data is Gaussian PDF.

Table 1.11: Goodness-of-Fit Summary for Hybrid Democratic Countries

	α	**p - value**	**Do Not Reject/Reject**
Kolmogorov-Smirnov	0.05	0.977	Do Not Reject
Anderson-Darling	0.05	0.9695	Do Not Reject
Chi-Squared	0.05	0.6472	Do not Reject

The MLEs of this pdf fitted to Hybrid democratic countries data is presented in the following table:

Table 1.12: MLEs of Hybrid Democratic Countries

MLEs of Hybrid Democratic Countries	
$\hat{\mu}$	$\hat{\sigma}$
4.966	0.7044

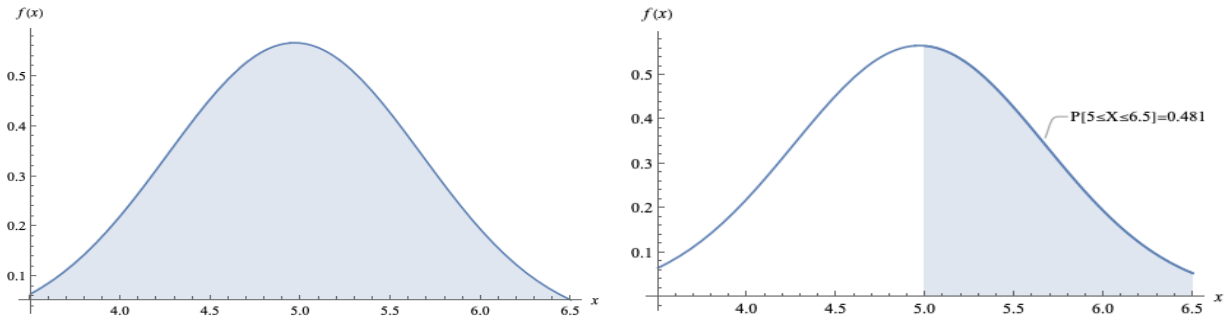
From table 1.12, we see that the estimated value of the population mean $\hat{\mu} = 4.966$, which is also the expected value of the PDF of the hybrid regime is very close to the sample mean **4.9621**. The analytical structure of the PDF of Hybrid democratic countries is given in the equation 1.12

$$f\left(x|\mu,\sigma\right) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{\sigma^2}\right), & -\infty \leq x \leq \infty \\ 0, & \text{otherwise} \end{cases} \tag{1.12}$$

The analytical structure of the PDF of Hybrid regime countries with the estimated parameters is given by:

$$f(x) = \begin{cases} 0.566e^{-1.007(x-4.966)^2}, & 3.5 \leq X \leq 6.5 \\ 0, & \text{otherwise} \end{cases} \tag{1.13}$$

The graph of the pdf given in 1.13 is shown in Figure 1.12a below.



(a) PDF of Hybrid Regime Countries **(b)** PDF with shaded area for $P(X \geq 5)$

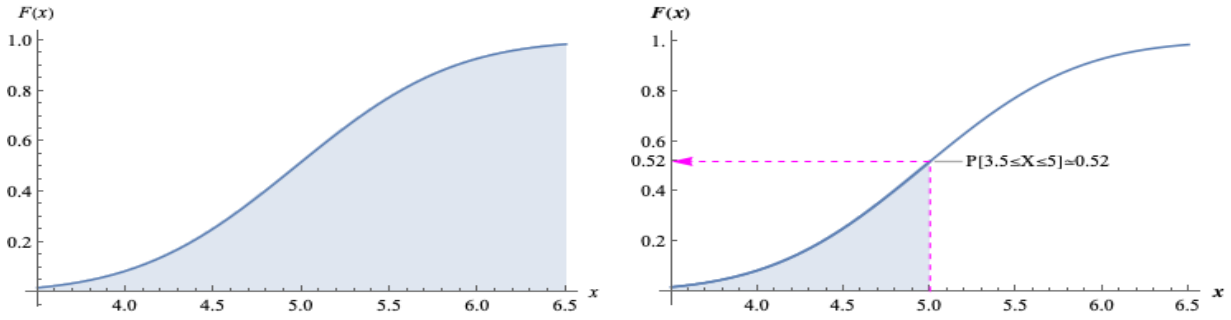
Figure 1.12: Plotting PDF of DIS of Hybrid Regime Countries of the World

From the above, one can calculate the expected DIS score of any country randomly selected from this cluster of the population is $E(X) = 4.966$ and the variance $V(X) = 0.496241$. The estimated expected value is a very close match to the sample mean of 4.9621 given in Table 1.7 and the probability of DIS of any country greater than 5 is 0.481 as per **Figure 1.12b**.

The CDF of Hybrid Regime countries of the world is given by-

$$F(x) = P(X \leq x) = \frac{1}{2}\text{erfc}\left(\frac{\mu - x}{\sqrt{2}\sigma}\right), \quad 3.5 \leq X \leq 6.5 \tag{1.14}$$

The graph of the CDF mentioned in equation 1.14 is postulated as follows:



(a) CDF of Hybrid Regime Countries **(b)** CDF with shaded area for $P(X \leq 5)$

Figure 1.13: Plotting CDF of DIS of Hybrid Regime Countries of the World

From the figures given above, we can extract some beneficial information. Such as, the probability of randomly selected any country's DIS is less than 5 (*i.e.* $P(X \leq 5) = 1 - P(X > 5)$) will be 0.52 as shown in Figure 1.13b.

1.2.5 PDF of "Authoritarian Regime"Countries

We shall now proceed to find the probability distribution that characterizes the probabilistic behavior of only the DIS data for Authoritarian Regime. To do this, we have implemented the same steps we have used in finding the overall PDF of DIS for all democracy classifications. For this purpose, we have started with the basic descriptive statistics of *Authoritarian Regime* countries.

Table 1.13: Descriptive Statistic of Authoritarian Countries

Descriptive Statistics of Authoritarian Regime Countries				
Mean	Median	Std. Deviation	Skewness	Kurtosis
2.85	3.012	0.724	-0.644	2.43

The table 1.13 above shows the basic descriptive statistics of the Authoritarian regime of 49 countries of the world. The sample mean is approximately 2.85. The histogram of the subject data subset of overall DIS data is given below:

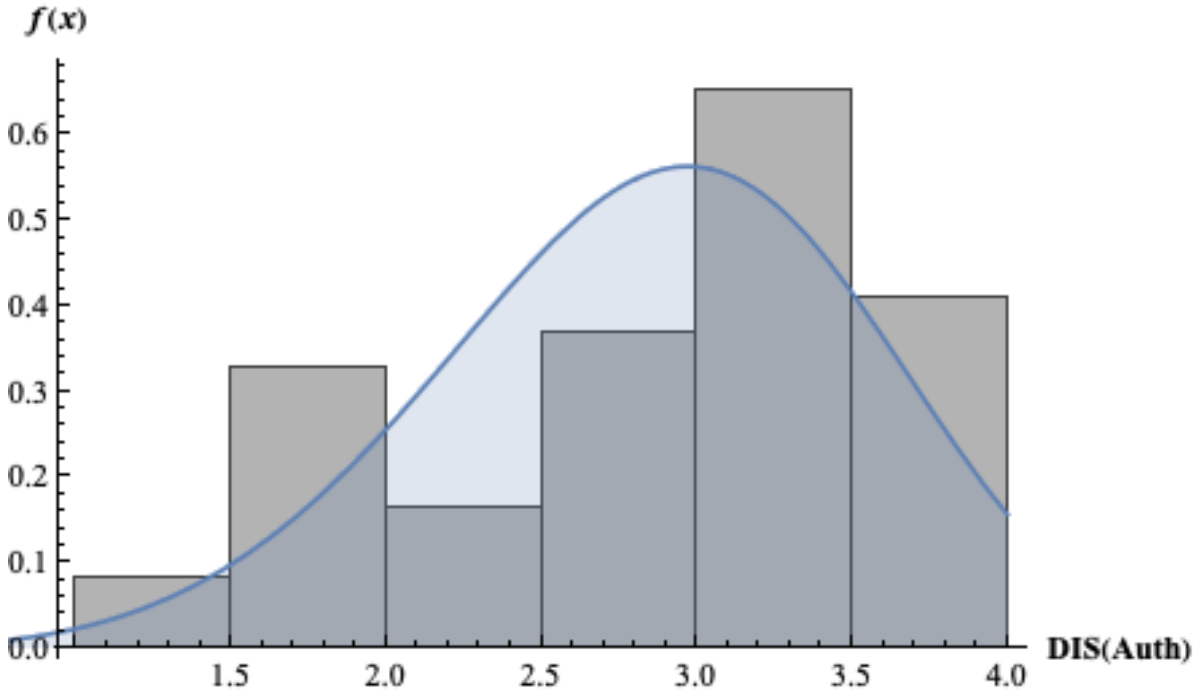


Figure 1.14: Fitted PDF to Histogram of Authoritarian Regime Countries

Figure 1.14, indicates that the population distribution of this data subset might follow some left-skewed probability density functions. This justification is also confirmed by the three methods of Goodness-of-Fit tests. We have found that the population PDF of Authoritarian Regime countries of the world follows the **Weibull Distribution** and this fact is confirmed by the following Table 1.14.

Table 1.14: Goodness-of-Fit Summary for Authoritarian Countries

	α	$p-value$	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.5889	Do Not Reject
Anderson-Darling	0.05	0.4602	Do Not Reject
Chi-Squared	0.05	0.4996	Do not Reject

From **Table 1.14**, it is clear that the population PDF of Authoritarian Regime comes from Weibull Distribution. The approximate MLEs for the given PDF of equation 1.15 given below:

Table 1.15: MLEs of Authoritarian Regime Countries

MLEs of Authoritarian Regime	
$\hat{\alpha}$	$\hat{\beta}$
4.6678	3.126

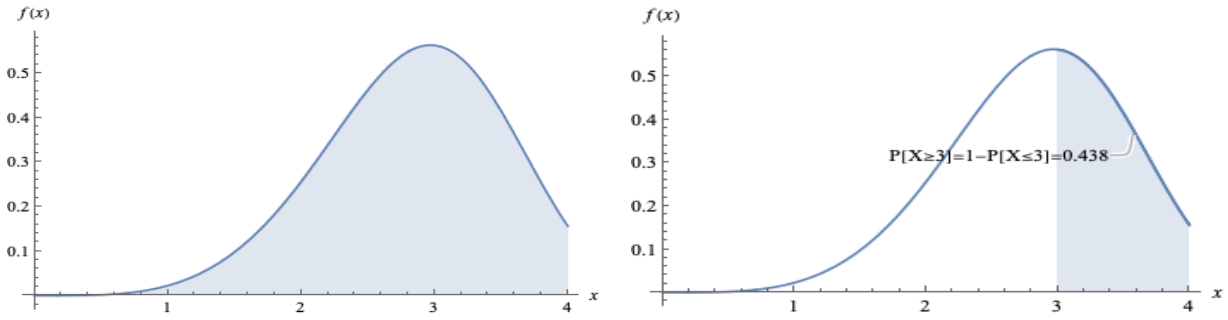
The analytical structure of the PDF of Authoritarian Regime countries of the world is given as follows:

$$f(x) = \begin{cases} \frac{\alpha e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \left(\frac{x}{\beta}\right)^{\alpha-1}}{\beta}, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{1.15}$$

The parameterized PDF of 1.15 is given by

$$f(x) = \begin{cases} 0.0228x^{3.667}e^{-0.005x^{4.667}}, & 0 \leq X \leq 3.89 \\ 0, & \text{otherwise} \end{cases} \tag{1.16}$$

The corresponding graph of 1.16 is given below.



(a) PDF of Authoritarian Regime **(b)** PDF with Shaded area for $P(X \geq 3)$

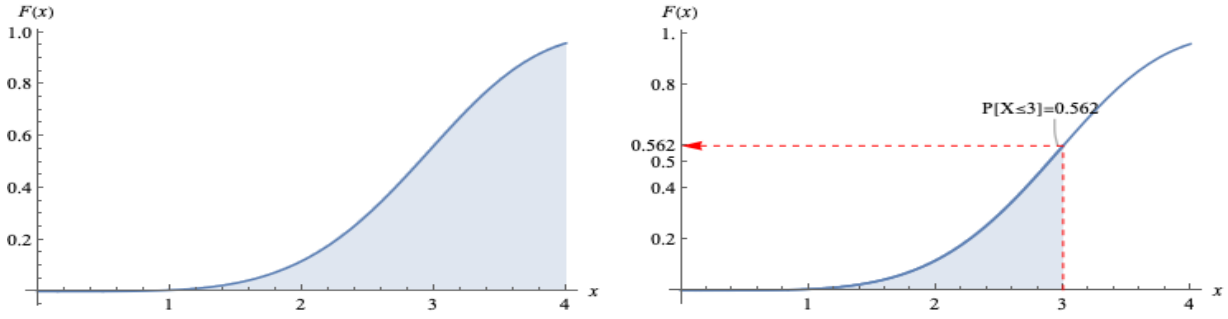
Figure 1.15: Plotting PDF of DIS of Authoritarian Regime Countries of the World

The expected value from this PDF is **2.858**, which is very close to the sample mean $\bar{x} = 2.851$. This indicates that our density estimation process is statistically correct. Also, it tells the fact that, if any country is randomly selected from this population, then the expected democracy index score would be approximately 2.86. Moreover, if a country is randomly selected from this cluster of the population then the probability of that country's DIS is greater 3 will be approximately 0.44, as shown in Figure 1.15b.

The CDF for the Authoritarian Regime countries is given by-

$$F(X) = 1 - P(X \geq x) = 1 - e^{-\left(\frac{x}{\beta}\right)^{\alpha}} \tag{1.17}$$

The plot of the above CDF given in equation 1.17 is as follows:



(a) CDF of Authoritarian Regime **(b)** CDF with Shaded area for $P(X \leq 3)$

Figure 1.16: Plotting CDF of DIS of Authoritarian Regime Countries of the World

Thus, if any country is randomly selected from this sub-population, then the expected democracy index score would be 2.86 and the probability of that country being elected and having scored less than or equal to 3 would be approximately 0.56 as shown in Figure 1.16b.

1.3 Contributions

We have developed the parametric analysis of Democracy Index Scores (DIS) of 167 countries in the world from the data of the Economist Intelligence Unit (EIU), from which we can extract the following useful pieces of information and insights.

1. The probability distribution functions of democracy index scores (DIS) for different subgroups of 167 countries which can be used for other statistical analyses such as statistical inference, clustering on the subject matter.

2. The probability distribution function of overall DIS of 167 countries of the world is Mixer Gaussian Probability, and this can be used to find some useful information like the overall confidence limit of these scores, their expected values, cut off points for determining appropriate cluster/ subgroups.

3. The visual postulation of probability density functions and the respective cumulative distribution functions could be a handy tool to facilitate the decision-makers to identify potential countries to get developmental funds from WB, WHO, IMF, etc.

4. Insights found in this study are essential to the relevant study in Chapter3.

2 Statistical Model that Predicts the Democracy Index Scores of the Countries in the World

2.1 Introduction

In this present study, we have utilized the data extracted from the EIU published DIS and have revisited their methodology and results. This study is conducted to formulate a statistical model that will be used to predict democracy index scores. The methodology used by EIU is descriptive and in other literature used this very database to investigate the progressive characteristics of democracy index[44]. In some literature, the impact of democracy is studied on corruption and other socio-economic attributions[31]. In some articles, the transition to and from the democracy of a country has been studied[56]. In some literature the quality of democracy has been studied[10]. In reference to building a statistical model that predicts the DIS, a very little to no work has been done to the date as per our knowledge.

So, we have used the implicit regression analysis technique to estimated the response variable in order to build a statistical model or in other words regression model that predicts the DIS with a high prediction accuracy. Also, this model has some practical relevance and usefulness by which interested financial institutions, non government organizations and even individuals can have their insights about very complicated and complex social phenomena of democracy.

2.2 Non-Response Analysis

In Non-Response Analysis (NRA) developed and utilized by Wooten et. al.[58, 59], one has to follow a process of testing any model's constant. This implies that, if someone has to do the NRA then, he/she has to test the model's constant on all the remaining terms of the model. In this process, there is a alias matrix A, such that, the expected value of the beta coefficients and the constant coefficients are related to each other by the following equation:

$$E\left(\begin{bmatrix}\beta_1\\ \beta_2\\ ....\\ \beta_p\end{bmatrix}\right) = \begin{bmatrix}\beta_1\\ \beta_2\\ ....\\ \beta_p\end{bmatrix} + A\beta_0 \tag{2.1}$$

Here, the alias matrix A, is obtained by the equation:

$$\hat{A} = \left(X_1'X_1\right)^{-1} X_1'X_2 \tag{2.2}$$

where, $X_1 = \begin{bmatrix} x_{11} & x_{21} & x_{31} & ... & x_{p1}\\ x_{12} & x_{22} & x_{32} & ... & x_{p2}\\ x_{13} & x_{23} & x_{33} & ... & x_{p3}\\ ... & ... & ... & ... & ...\\ x_{1n} & x_{2n} & x_{3n} & ... & x_{pn}\end{bmatrix}$ and $X_2 = \begin{bmatrix}1\\ 1\\ 1\\ ...\\ 1\end{bmatrix}$, the bias introduced in the model by constant is equivalent of testing the model given as follows:

$$1 = \beta_1 x X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_p X_p \tag{2.3}$$

Here, $E(\hat{\beta}) = \beta$ and $V(\hat{\beta}) = \sigma^2(X_1'X_1)^1$. The σ, is coming from the fact that, the response variable z, is not measured explicitly in the model, but, assumed to follow $z \sim N(\mu, \sigma^2)$ and $z = h(X_1, X_2, ..., X_p/\Theta)$ is the subject response, and $h(X_1, X_2, ..., X_p/\Theta)$ is measured

in the terms of unknown coefficients, $\Theta = \{\beta_1, \beta_2, \beta_3, ..., \beta_p\}$. In fact, in a good experimental design, varied outcomes in these variables are preferred; otherwise, they are relatively constant and are absorbed by the constant. The measured angle, θ_M, where, θ_M is the angle between SST and SSE given as:

$$SSM = SST + SSE - 2\sqrt{SST \times SSE}\cos(\theta_M)$$

from which we can manipulate the equation for the angle as following:

$$\theta_M = \arccos\left(\frac{SST + SSE - SSM}{2\sqrt{SST \times SSE}}\right) \tag{2.4}$$

2.3 Development of the Statistical Model

2.3.1 Estimating DIS by Non-Response Analysis

After implementing all of the above theories and equations in our data set at hand, we have the following results of the parameter matrix:

$$\beta = \begin{bmatrix} -0.0336288 \\ -0.0565488 \\ 0.052849 \\ 0.1417623 \\ 0.0570979 \end{bmatrix}$$

with the co- efficient of determination as, $R^2 = 0.93472$. After implementing the above theory in our data set, we have ended up with sum of squares presented in the table 2.1. We will use these values for the angel measurement as a term of validating our model to be a representative of the true state of the nature.

Table 2.1: Sum of Squares for the Implicit Regression Model

Sources	Values
SSM	4199.091115
SSE	1147.747441
SST	5346.838556

Using the estimated values from the dataset presented in table 2.1 to the equation 2.4 we have the estimated degree as follows:

$$\theta_M = 62.39889 \approx 62^0$$

A good model should have an angle as close to 90^0 as possible. From our result we can claim that our model is performing quite good enough to be called as a reasonably good model. As a result, we have estimated values of response variable (DIS = Democracy Index Scores) by taking the average of each of the attributes/covariates values estimated from "Rotation Analysis".

2.3.2 Multiple Linear Regression

2.3.2.1 Partitioning Data into Test & Training Data sets

After estimating scores by using Non-response analysis, we have proceeded with the regular process of **Linear Regression**. After, the rotational analysis, we have taken a **random sample** of 80% from the population of **167** data points, and the random sample was stratified random sample from the population because of the fact that data set has a four categories as per the classification of EIU. Therefore, the sample size used in the regression analysis is **135**, that means we will use these 135 data points to analyze and formulate regression model to identify the best model to predict and validate the **Democracy Index Scores** of the countries of the world. Now, we will fit a **Multiple Linear Regression** to

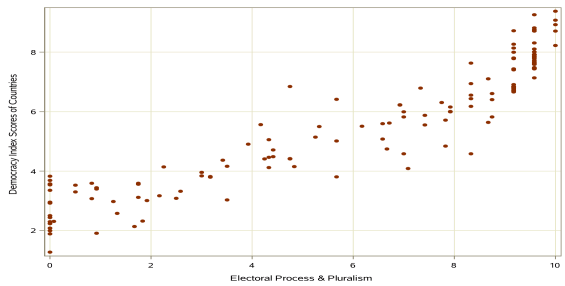
find the best model as per our data set at hand. The prep steps of the sampled data are as follows:

1. Check the relationships between each independent variables and dependent variable using scatter plots and correlations
2. Check the relationships among independent variables using scatter plots and multicollinearities.
3. Use the non- redundant independent variables in the analysis to find the best fitting model
4. Use the best fitted model to make predictions about the dependent variable.

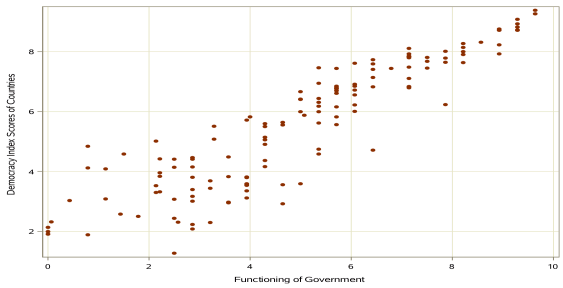
In the following sections, all the process have been enumerated and explained as we go along the analysis process.

2.3.2.2 Checking Co-linearity

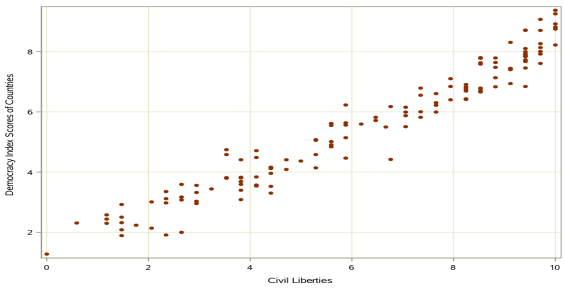
In the following graphs presented in **Figure** 2.1 postulates relationships of dependent variable (DIS) with the Independent variables (such as, EPP, FG, PP, PC, CL). As we can see in **figure 2.1c**, dependent variable has a fairly strong relationship with the independent variable CL so is with the EPP and FG as well. So, these variables might be the effect variables in the prediction model that we are going to formulate. We have calculated the values of the response variable through the non- response analysis briefly mentioned the previous section. Now, considering estimated response variable, we have plotted all the covariates against the dependent variable, democracy index scores of the countries of the world. From the scatter plots give in Figure 2.1,



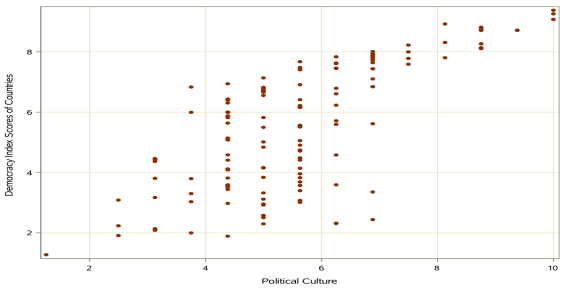
(a) Scatter Plot of DIS vs. EPP



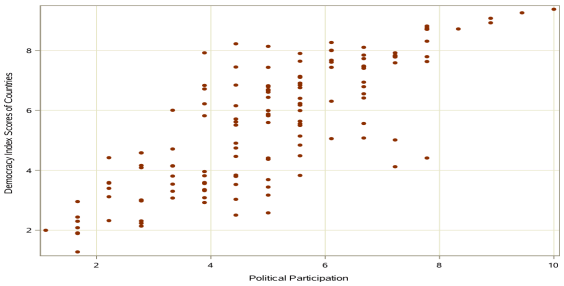
(b) Scatter Plot of DIS vs. FG



(c) Scatter Plot of DIS vs. CL



(d) Scatter Plot of DIS vs. PC



(e) Scatter Plot of DIS vs. PP

Figure 2.1: Scatter plots to determine significant Independent variables

it turns out that the dependent variable DIS (Democracy Index Scores) have moderately correlated to the covariates PP (Political Participation) and PC (Political Culture).

On the other hand, DIS is highly correlated to the covariates named EPP (Electoral Process and Pluralism), FG (Functioning of Government), and CL (Civil Liberties). This fact is also supported by the correlation structure table given in **Table** 2.2.

The summary of scatter plots given in **Figure 2.1a, 2.1b, and 2.1c** shows that DIS has a strong relation with EPP, FG and CL. So, dependent variable (Democracy Index Scores of Countries of the world) and independent variables (EPP, FG, PP, PC, CL) have the relationships as follows:

- CL is highly correlated to DIS
- FG has second highest magnitude of the correlation structure
- EPP is third highest magnitude of correlation to DIS.

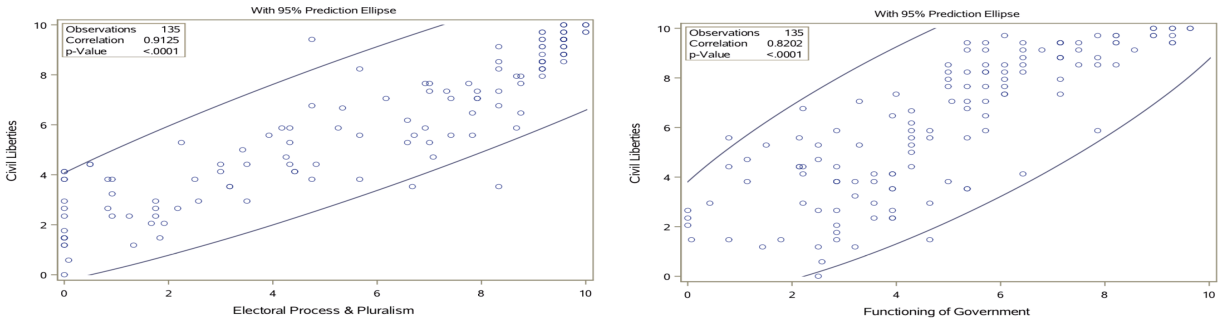
Table 2.2: Correlation Coefficients of covariates under $H_0 : \rho = 0$

	DIS Scores	EPP	FG	PP	PC	PL
DIS Scores	1					
EPP	0.931	1				
FG	0.912	0.777	1			
PP	0.764	0.673	0.652	1		
PC	0.662	0.482	0.662	0.527	1	
PL	0.974	0.913	0.8425	0.686	0.5534	1

Also, one can have an idea from **Table** 2.2 that EPP, FG, and CL is the highly correlated to DIS. So, we will consider these variables in the model. But as a preparation work to formulate a statistical model we have examined the multicollinearity among the Independent variables/ covariates[**?**]. The purpose of testing the multicollinearity is to identify the potential interaction among and between the covariates. In the following section we will examine this fact.

2.3.3 Checking Multicollinearity of IVs

In the process of building this model, we needed to check the multicolinearity among the independent variables (**IVs**). In the following **Figure 2.2**, it is found that variable CL is highly correlated to EPP and FG. In our model we have to include either CL or EPP or FG.



(a) Scatter Plot of EPP vs. CL **(b)** Scatter Plot of FG vs. CL

Figure 2.2: Scatter plots to detect multicollinearity among IVs

But, not all of them together because of the fact that including all the multicolinear variables in the model will effect the contribution of each of these variables which in turn will incorporate over-fitting or under fitting the model. It turns out that, the variable EPP and FG is significantly correlated to CL. In order to formulate a model, we need to use only one of the three highly correlated variables mentioned above.

2.4 Validation of the developed Statistical model

2.4.1 Model validation through R^2, AIC, BIC

After all the preparation and correlation analysis, it turns out the final model consists of three variables (CL = Civil Liberties, FG = Functioning of Government, PC = Political Culture) in the model. The analytical structure of the final model to predict the democracy index scores of countries of the World is in the following equation:

$$\hat{DIS} = \hat{\beta}_0 + \hat{\beta}_1 CL + \hat{\beta}_2 (FG * PC) \tag{2.5}$$

here in this above equation 2.5, we have found that 95% of the variation is explained in the model as per the R^2 of the model and the AIC, and BIC of the aforementioned model given in the table 2.3 indicates that our model is as accurate as 95%.

Table 2.3: Quality of the Model

R^2	Adj. R^2	AIC	BIC
0.9491	0.9483	184.144	195.765

In the table 2.4, it is found that all the co-efficients are statistically significant.

Table 2.4: Estimated values of the co- efficients of model

Co-efficients	Estimated Values	Sig.
$\hat{\beta_0}$	1.114	0.000
$\hat{\beta_1}$	0.702	0.000
$\hat{\beta_2}$	0.028	0.002

and also, CL = Civil Liberties and FG = Functioning of Government and PC = Political Cultures. In the following picture we have postulated the 95% confidence interval ellipsoid of the coefficients. It shows that the ellipsoid is very narrow, that indicates the estimated coefficients are statistically significant enough to be included in the model.

2.4.2 Residual Analysis

As an evaluation process of the model, we have done the residual analysis and it turns out that the distribution of residuals is fairly Normal distribution.

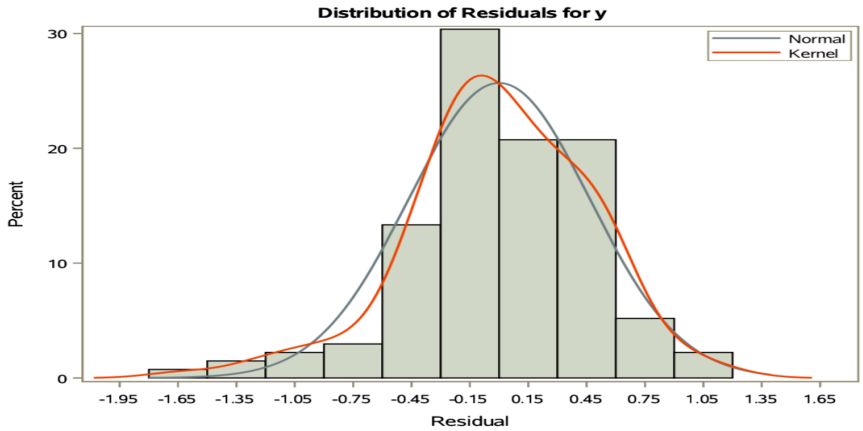


Figure 2.3: Distribution Fit of residuals of the model

and it is confirmed by the **Figure 2.3** given above. From the distribution fit of the residuals of the model, it is obvious that the distribution is Normal with mean 0. As a final plotting of the model, we have postulated the predicted values to the observed values to show the model's prediction accuracy.

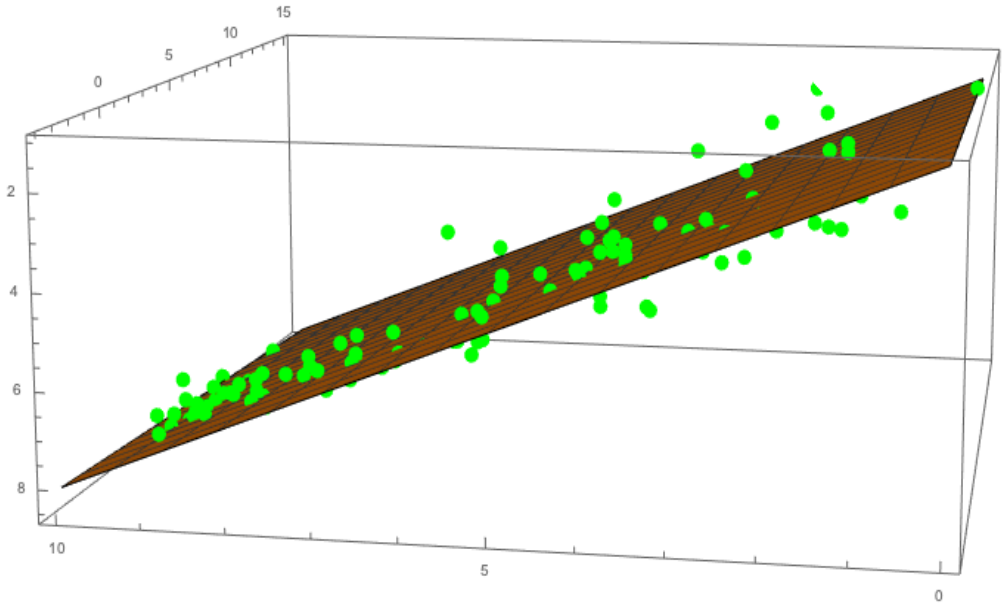


Figure 2.4: Observed *vs* Predicted values of DIS to the fitted regression line

It turns out that, all the observed values of the data set are densely and closely clustered to the fitted/predicted surface of the linear line. It also implicates that, the fitted line is a very accurate fit to the dataset with very small distances around the predicted model.

2.5 Usefulness of the developed model

1. We can use the developed statistical model to predict the democracy score of any countries of the world based on the values of attributable variables contained in the model given in the **Equation 2.5**.

2. We have identified the significant variables that drives the democracy scores.

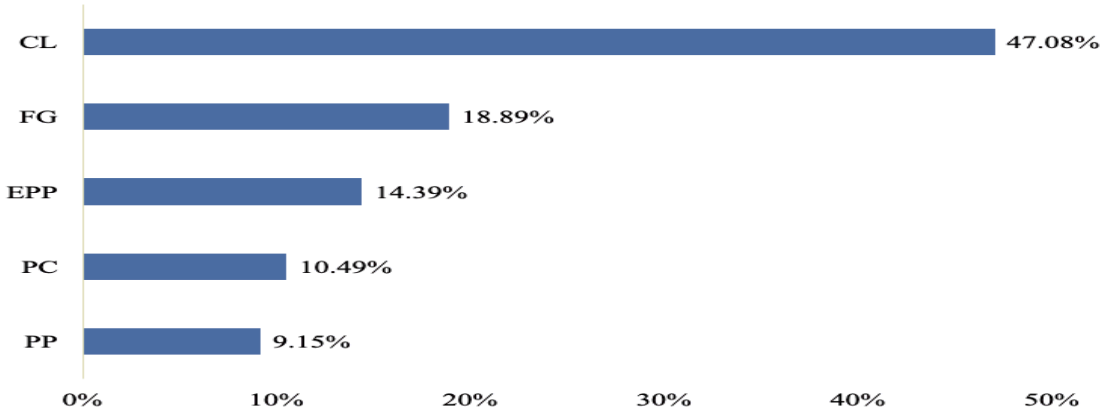


Figure 2.5: Ranking of attributable variables

In the Figure 2.5, we have ranked the descending order of significant variables that drives the democracy scores. It turns out that, the CL (Civil Liberties) is the most contributing or influential variable of the model as per **Figure 2.5**. In this figure, the attribute CL has the highest or in other words almost 47% of contribution to the total variation explained by the estimated model fit with an R^2 of 95% of accuracy.

3. We have identified the significant interaction term that contributed the most to build the prediction model.

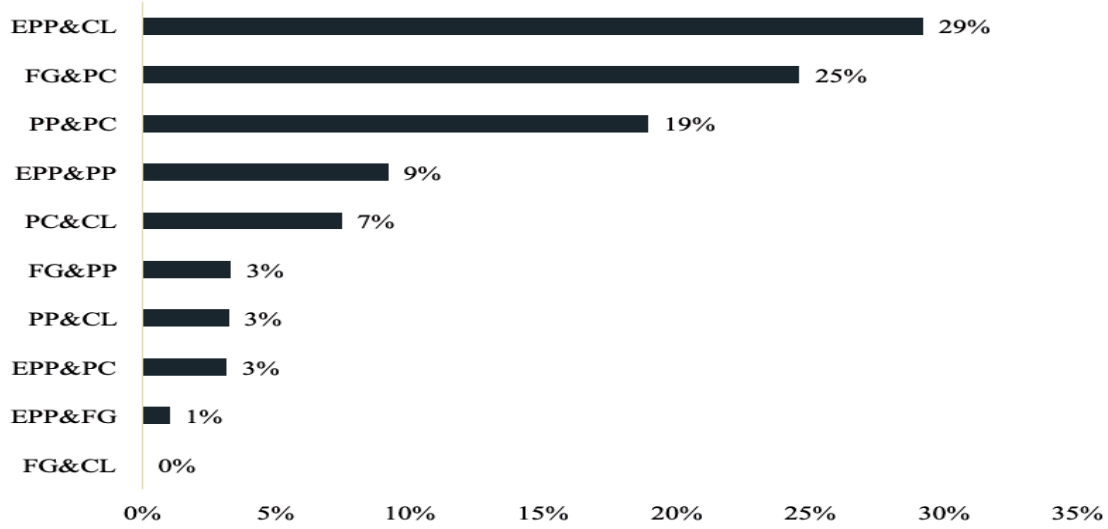


Figure 2.6: Ranking of Interacting terms in the Model

4. Now, we have ranked the covariates included in the final model given in equation 2.5 as per their contribution percentages to the estimated R^2. In the following **Figure 2.7**,

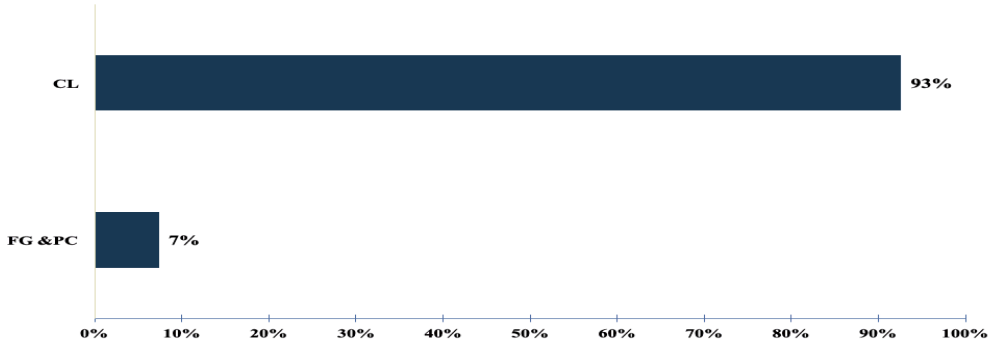


Figure 2.7: Ranking of Attributes according to their contribution in the final model

5. Finally, having this model one can perform a surface response analysis. The objective being what should this attributable variables (risk factors) be so as to maximize democracy scores. (Under Study)

2.6 Discussion of the analysis

The model given in **equation 2.5**, has explained the total variation of the model with approximately 95% accuracy. Also, as per **Table 2.3**, the BIC (Bayesian Information Criterion) and AIC (Akaik Information Criterion) was the lowest among all other models we have tried. Also, the adjusted-R^2 is very good to believe (0.9483) and predict the democracy index score of any countries of the world.

At the same time, the practical relevance of ranking the independent variables are to make statistical survey more economic, time saving and improve the survey methodology and structure. As a process of determining and evaluating a regression model, the most important thing is to determine and rank the significant interaction terms and significant attributable variables. In our analysis, we have satisfied this fact as well. One of the most significant goal of building a statistical model is to make a prediction about the

democracy score of any countries of the world and as per our model's predictability. Here, the prediction quality of the model is depicted with the 95% confidence interval in the following figure :

Figure 2.8: Prediction quality of the model with 95% confidence interval

From the the Figure given above, we have an idea about the prediction quality of the model. In this plot, the observed (Blue Solid line) DIS scores are very closely followed by the predicted scores (Red-Dashed line) by our developed model and all the values are withing the boundary of 95% prediction interval bandwidth.

2.7 Contributions

In the present study we have accomplished the followings:

1. We have very high R^2 & this is consistent with $adjusted - R^2$ because this eliminates the biasness of the interaction introduced in the model due to human interactions in the subject area.

2. Residual analysis, that tell us one or two things, if the residuals are significantly positive it has to be subtracted from the model, and if those are significantly negative they had to be added to the model.

3. Final evaluation of the model is, during the process of developing the model, we left 30% of the observation out of the model building from each strata which are made after the cluster analysis by the $k-means$ and *Multinomial Logistic Regression* analysis chapter.

4. We have used the developed model to predict those countries scores.

5. The practical usefulness of the proposed model would be,

    - to utilize to predict Democracy Index Scores of any new country included in the model with a degree of accuracy.

This statistical model will guide us to avoid misclassification of a country in determining it's democracy scores. Because of misclassification a country might be loosing development funds from monetary organizations like WB (World Bank), IMF, WHO and so on.

3 Parametric Analysis of Corruption Perception Index of Transparency International and World Governance Index of World bank Countries of the World

3.1 Introduction

Corruption is a form of dishonest or unethical conduct by a person entrusted with a position of authority, often to acquire personal benefit. Corruption may include many activities including bribery and embezzlement, though it may also involve practices that are legal in many countries. Government, or 'political', corruption occurs when an office-holder or other governmental employee acts in an official capacity for personal gain. Stephen D. Morris, a professor of politics, writes that [political] corruption is the illegitimate use of public power to benefit a private interest. Economist Ian Senior defines corruption as an action to (a) secretly provide (b) a good or a service to a third party (c) so that he or she can influence certain actions which (d) benefit the corrupt, a third party, or both (e) in which the corrupt agent has authority. Daniel Kaufmann, from the World Bank, extends the concept to include 'legal corruption'in which power is abused within the confines of the law — as those with power often have the ability to make laws for their protection.

3.2 Data Source and Methodology

The Corruption Perception Index (CPI) was established in 1995 as a composite indicator used to measure perceptions of corruption in the public sector in different countries around the world. The method followed by the Transparency International (TI) was in brief as follows: 1) Selection of reliable data collection source; 2) Standardize data sources

3) Aggregate the rescaled data; and 4) Report a measure of uncertainty. By following the aforementioned methods, the TI collected data from various resources from all over the world by segregating the whole world into six different regions and data collected for 175 countries of the world. Also, we have used the data from WB (World Bank) which also considered to be an authentic data sources for corruption index. The Following diagram represent the data structure that has been used in this study:

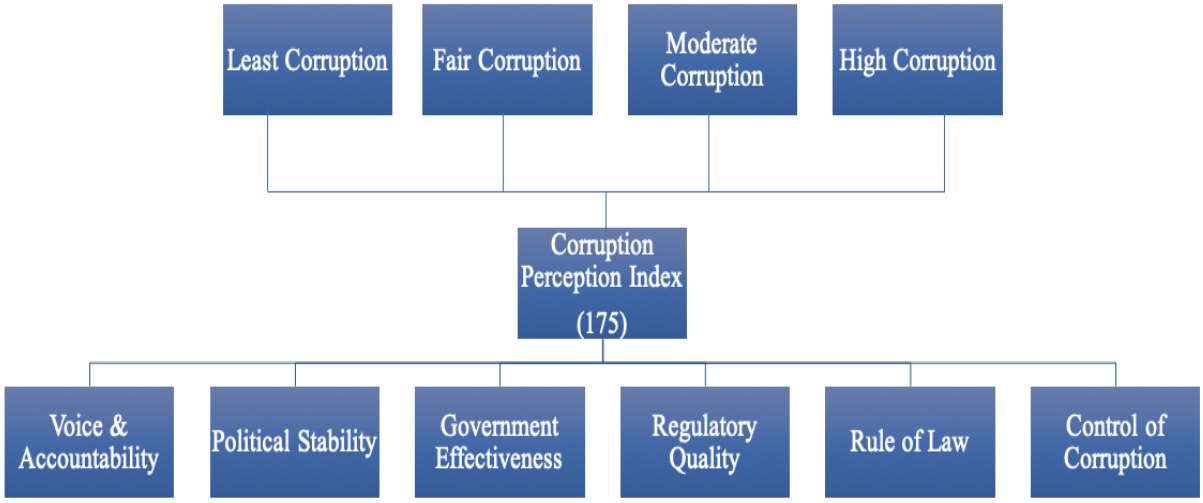


Figure 3.1: Data diagram of Corruption Perception Index (CPI)

For the current study, we will use the data set collected by Transparency International from 175 countries of the world. The diagram of the data used in this study is presented in the Figure 3.1. Our objective is to use the CPI scores calculated by TI and proceed to identify the probability distribution function, PDF, that probabilistically characterize the behavior of the CPI scores of 175 countries.

3.2.1 Finding PDF of Corruption Perception Index

We begin by calculating the basic statistics of the CPI data and the results are given by Table 3.1, below:

Table 3.1: Descriptive Statistics for Corruption Perceptino Index (CPI)

Descriptive Statistics of CPI				
Mean	Median	Standard Deviation	Skewness	Kurtosis
4.739	4.313	2.679	0.249	1.98

From the Table 3.1, we see that the mean of the CPI is 4.73 and median is 4.313 and this data is slightly rigth skewed with a value of 0.249. As a first step of the procedure to determine the probability density function of CPI, we start with the histogram of the data to obtain a visual idea of what type of PDF we should testing goodness-of-fit.

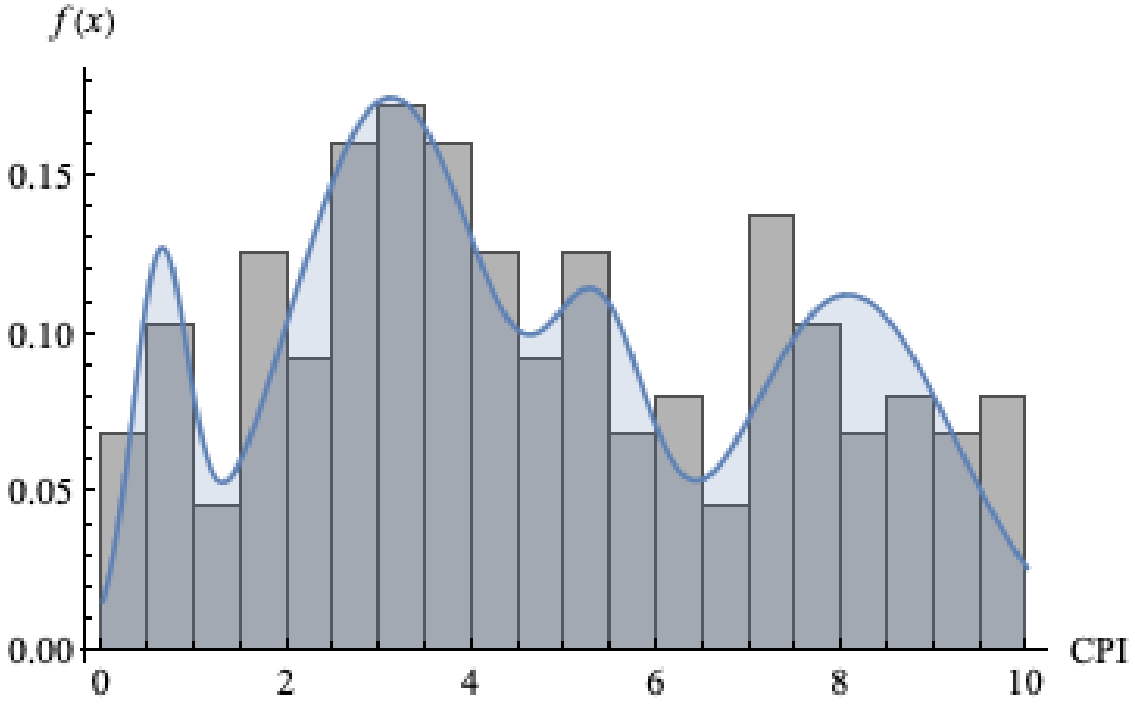


Figure 3.2: Overall Distribution fitting of CPI (Corruption Perception Index)

From Figure 3.2, above, it indicates that the CPI will follow some sort of Mixture distribution. The whole process of estimating the PDF of CPI scores for 175 countries and four categories of CPI scores will be discussed step-by-step in the following sections.

3.2.1.1 Goodness-of-fit tests for CPI:

We proceeded by testing the goodness-of-fit for a number of well-defined PDFs using three statistical tests, namely, Kolomogrov-Smirnov [35], Anderson-Darling [5] and Chi-square [12]. The Kolmogorov-Smirnov test is based on minimum difference estimation. The Anderson-Darling measures whether the data can be transformed into the uniform probability distribution and the Chi-square test for goodness-of-fit is a measure of relative error squared [51].

Table 3.2: Goodness-of-Fit Summary for CPI Scores

	α	$p-value$	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.999	Do Not Reject
Anderson-Darling	0.05	0.994	Do Not Reject
Chi-Squared	0.05	0.977	Do Not Reject

We have found that, the best fitted distribution is the **"Mixture of 4- Gaussian PDF"** best fits all the CPI score data as it is supported by the results given in Table 3.2 above. Thus, we proceed to discuss and fit the Mixture of 4- Gaussian PDF of the CPI scores of 175 countries of the world.

3.2.1.2 PDF of Corruption Perception Index:

After passing the data through the aforementioned three goodness-of-fit tests [19], the probability distribution that captures the characterstics of CPI scores the best is the **4-Mixed Gaussian** PDF. A Gaussian mixture model [17] is parameterized by two types of values, the mixture component weights and the component means and variances/covariances. For a Gaussian mixture model with K components, the K^{th} component has a mean of μ_k and standard deviation of σ_k for the univariate case. In the case of CPI scores, K = 4. The analytical structure is given by:

$$f(x) = \sum_{i=1}^{k} \phi_i N\left(x|\mu_i, \sigma_i^2\right),$$

with, (3.1)

$$N\left(x|\mu_i, \sigma_i\right) = \frac{1}{\sigma_i\sqrt{2\pi}} exp\left(-\frac{(x-\mu_i)^2}{\sigma_i^2}\right), \quad -\infty \leq X \leq \infty$$

Here, $\sum_{i=1}^{k} \phi_i = 1$. The mean and the variance is 4.75 and 7.34, respectively, with standard deviation of 2.71. For our data, the approximate maximum likelihood estimates (MLEs) of the parameters (μ_i, σ_i, and ϕ_i, here, $i = 1, ..., 4$) of 3.1 are given in the following table 3.3:

Table 3.3: MLEs of CPI scores of 175 countries of the world

MLEs of CPI scores							
$\hat{\mu_1}$	$\hat{\sigma_1}$	$\hat{\mu_2}$	$\hat{\sigma_2}$	$\hat{\mu_3}$	$\hat{\sigma_3}$	$\hat{\mu_4}$	$\hat{\sigma_4}$
9.18	0.454	5.6	0.67	2.93	1.46	7.54	0.442

Thus, the estimated analytical form of the subject PDF is given by-

$$f(x) = \begin{cases} 0.113e^{-0.385(x-8.063)^2} + 0.085e^{-1.80(x-5.382)^2} \\ +0.175e^{-0.412(x-3.118)^2} + 0.114e^{-5.634(x-0.622)^2}, & 0 \leq X \leq 10 \\ 0, & \text{otherwise} \end{cases} \quad (3.2)$$

The graph of 3.2 is given below by Figure 3.3:

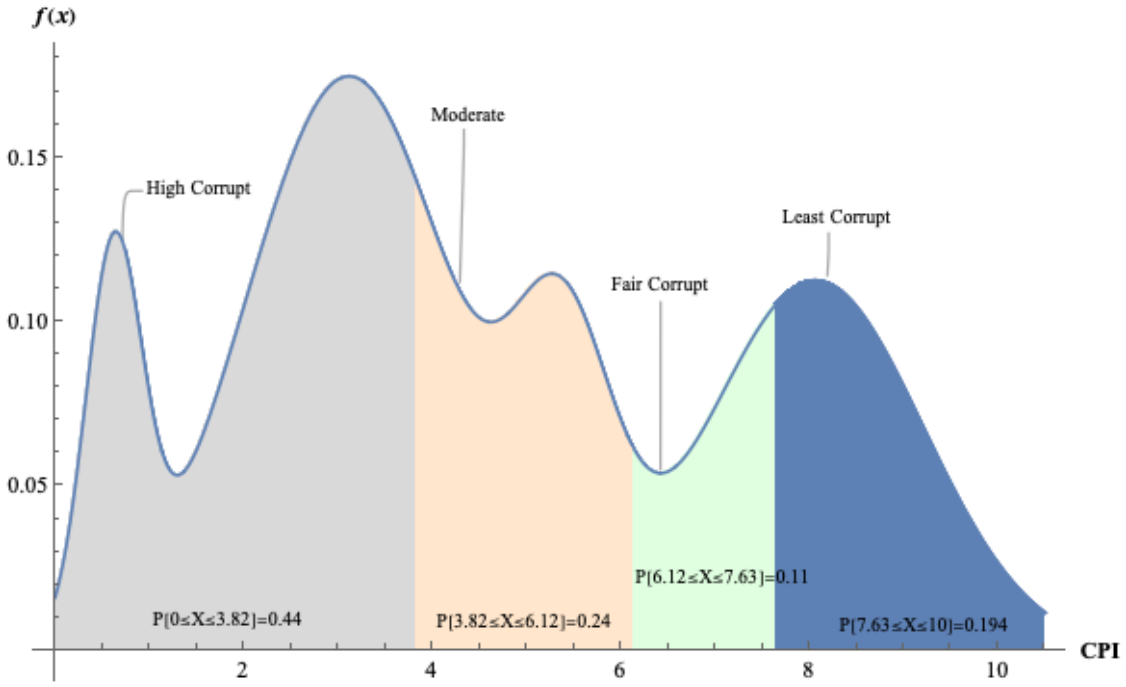


Figure 3.3: PDF of CPI Scores of 175 Countries of the World

Now, suppose that a country was selected at random from the 175 countries, one can identify the probability of its classification of the four categories of Corruption. By using the plots given in Figure 3.3, one can easily identify the areas under the curve for each of the classes of corruption. For example, if anyone caculates the probability of CPI within the range of 7.63 to 10, then the corresponding probability would be the probability of any country falling in the 'Least Corrupted' category and so on. Furthermore, the moment generating function of 3.1 is given by

$$M_X t = \frac{w_1 e^{\frac{1}{2}\sigma_1^2 t^2 + \mu_1 t} + w_2 e^{\frac{1}{2}\sigma_2^2 t^2 + \mu_2 t} + w_3 e^{\frac{1}{2}\sigma_3^2 t^2 + \mu_3 t} + w_4 e^{\frac{1}{2}\sigma_4^2 t^2 + \mu_4 t}}{w_1 + w_2 + w_3 + w_4}$$

or,

$$\begin{aligned} M_X t =& 0.085 e^{0.044t^2 + 0.622t} + 0.112 e^{0.138t^2 + 5.38t} \\ &+ 0.482 e^{0.61t^2 + 3.12t} + 0.32 e^{0.65t^2 + 8.063t} \end{aligned} \tag{3.3}$$

The moment generating function (MGF) is given in the equation 3.3 can be used to calculate the moments of higher order and consequently to calculated the mean and variance of the Mixed Gaussian PDF. Thus, if a country is selected at random from the population of 175 countries we will expect its CPI score to be 4.86. Also, we caculate the variance, V [X] = 7.34 and standard deviation, STDV [X] = 2.71. Note that these estimates are close to the basic statistics given in Table 3.1, which assures to the quality of the fit of **Mixed of 4-Gaussian PDF**.

The cumulative distribution function of the CPI scores is given in the equation below:

$$F(x) = P(X \leq x) = \begin{cases} \frac{w_1 \text{erfc}\left(\frac{\mu_1 - x}{\sqrt{2}\sigma_1}\right)}{2(w_1+w_2+w_3+w_4)} + \frac{w_2 \text{erfc}\left(\frac{\mu_2 - x}{\sqrt{2}\sigma_2}\right)}{2(w_1+w_2+w_3+w_4)} \\ + \frac{w_3 \text{erfc}\left(\frac{\mu_3 - x}{\sqrt{2}\sigma_3}\right)}{2(w_1+w_2+w_3+w_4)} + \frac{w_4 \text{erfc}\left(\frac{\mu_4 - x}{\sqrt{2}\sigma_4}\right)}{2(w_1+w_2+w_3+w_4)} \end{cases} \tag{3.4}$$

The graph of the CDF (Cumulative Distribution Function) of CPI scores is given as follows:

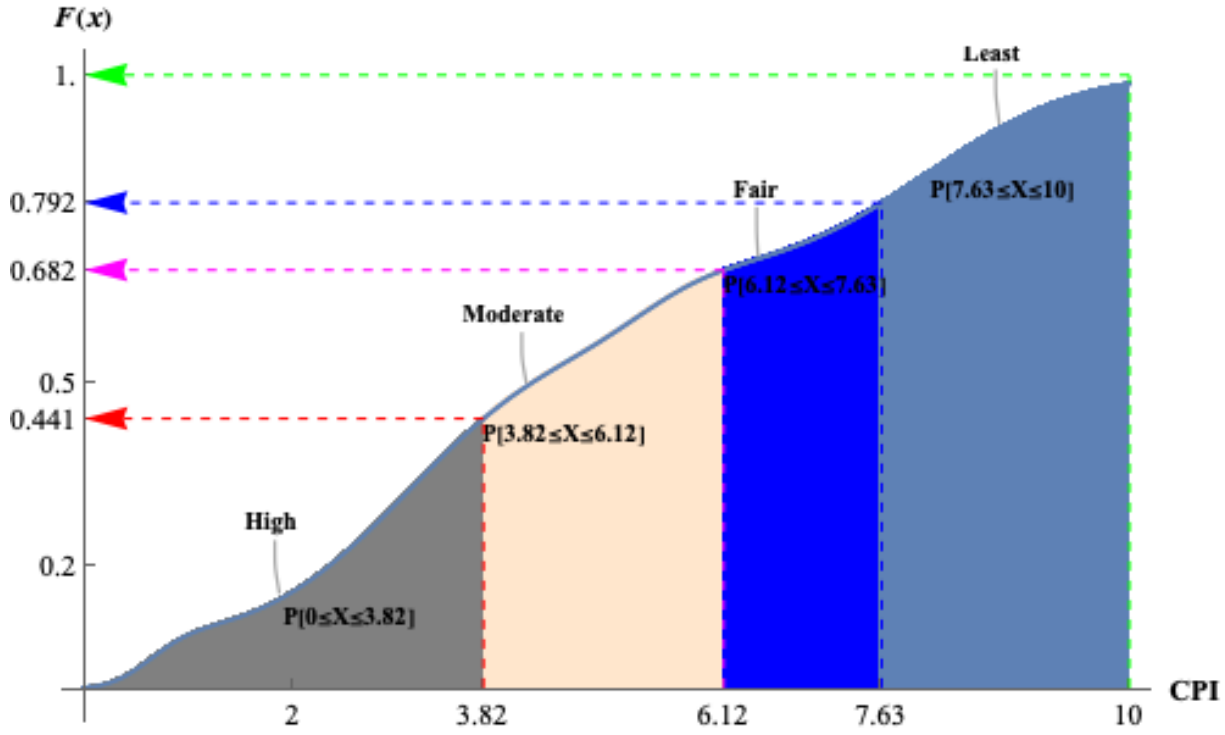


Figure 3.4: CDF of CPI Scores

The Figure 3.4, is very useful in the cases, for example, if anyone wants to know the probability of any country will have a CPI score less than 3.82 (i. e. $P\left[CPI \leq 3.8\right]$), then from the above figure it is shown that the probability would be 0.44 or approximately 44% of the areas under the cumulative probability distribution curve. Also, if we are curious about the probability of any countrys less than or equal to 7.63, then from the Figure 3.4, one can easily estimate it and the probability is approximately 0.792 or 79% area under the cumulative curve and so on.

In next few sections, we will discuss the process of determining PDF of each of the four categories of CPI scores of 175 countries in the world.

3.2.2 PDF of "Least Corrupted"Countries

We shall now proceed to find the probability distribution that characterize the probabilistic behavior of only the CPI data for **Least Corrupted** countries. To do this we have implemented the same steps we have used in finding the overall PDF of CPI scores for all classifications. For this purpose, we have started with the basic descriptive statistics of **Least Corrupted** countries.

Table 3.4: Descriptive Statistics of Least Corrupted Countries of the World

The Least Corrupted Countries CPI scores				
Mean	Median	Standard Deviation	Skewness	Kurtosis
8.77	8.83	0.717	-0.11416	1.685

From the table above, we see that, this subset of the overall data is slightly left skewed with a value of -0.468 and it has a mean 7.714. The histogram of Least Corrupted Countries is given below **Figure 3.5**. From this histogram, the implication is that we need to fit some sort of Mixed probability density function.

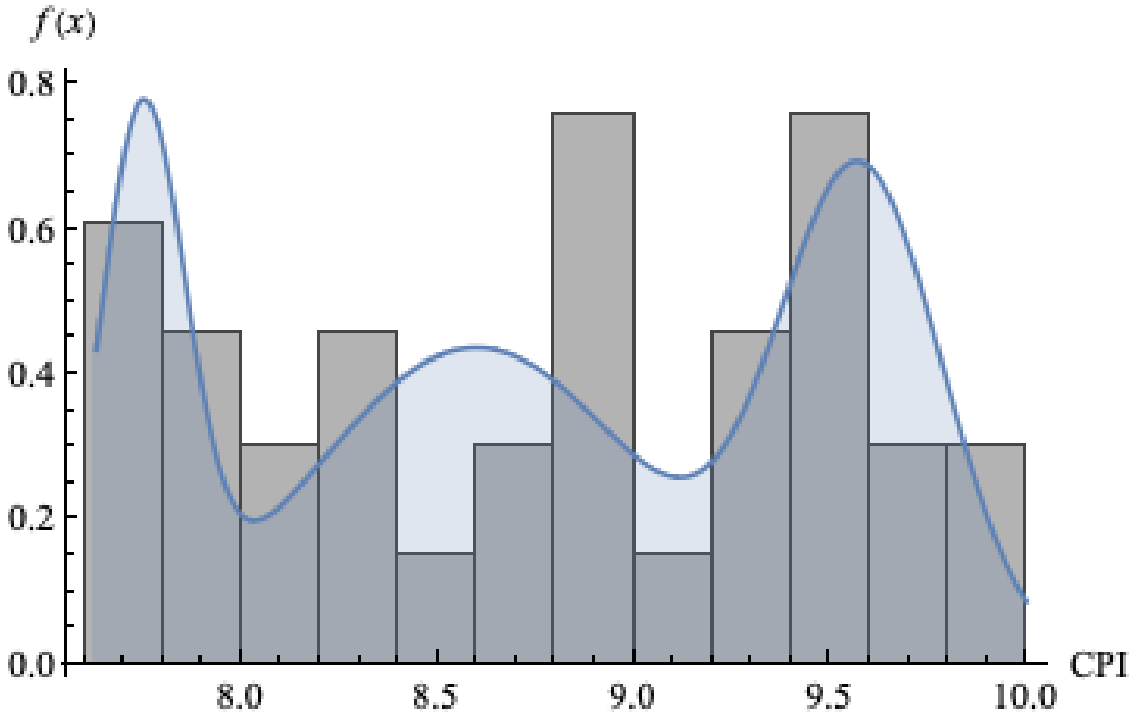


Figure 3.5: PDF fitted to Histogram of Least Corrupted Countries

Using the three goodness-of-fit tests to the present data of Least corrupted countires we have identified that the data can be characterized probabilistically by the "Mixed of 3-Gaussian PDF". The justification of this selection is confirmed by the three methods of goodness-of-fit that we used in Table 3.5 given below confirms that the best fitted pdf for the Least Corrupted data is the **Mixed of 3-Gaussian PDF**.

Table 3.5: Goodness-of-Fit Summary for Least Corrupted Scores

	α	$p-value$	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.9492	Do Not Reject
Anderson-Darling	0.05	0.9720	Do Not Reject
Chi-Squared	0.05	0.7121	Do not Reject

Thus, the fitted theoretical PDF of the subject data is given by-

$$f(x) = \begin{cases} \frac{w_1 e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{\sqrt{2\pi}\sigma_1(w_1+w_2+w_3)} + \frac{w_2 e^{-\frac{(x-\mu_2)^2}{2\sigma_2^2}}}{\sqrt{2\pi}\sigma_2(w_1+w_2+w_3)} + \frac{w_3 e^{-\frac{(x-\mu_3)^2}{2\sigma_3^2}}}{\sqrt{2\pi}\sigma_3(w_1+w_2+w_3)}, & x > 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

The approximate MLEs of the parameters that drive the estimated Mixed Gaussian PDF are given by Table 3.6 below:

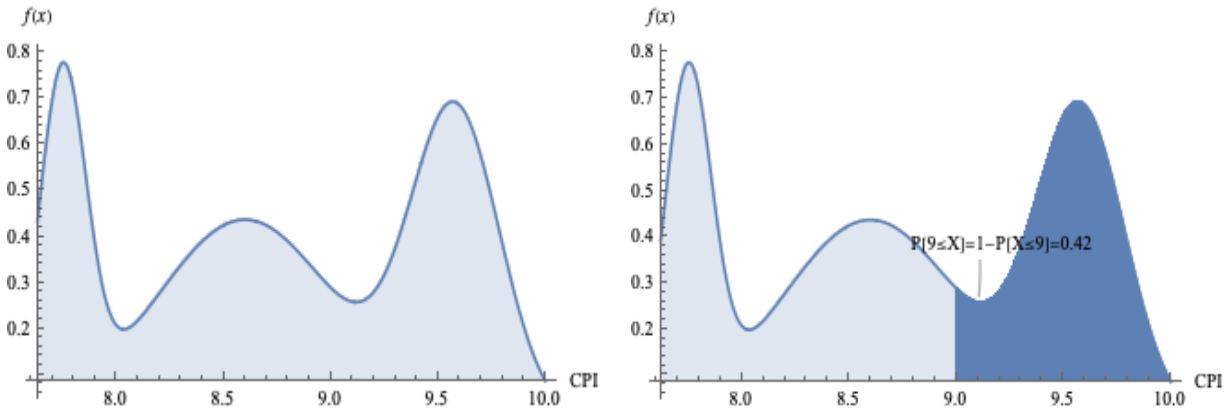
Table 3.6: MLEs of Least Corrupted Countries of the World

MLEs of Least Corrupted Countries					
$\hat{\mu_1}$	$\hat{\sigma_1}$	$\hat{\mu_2}$	$\hat{\sigma_2}$	$\hat{\mu_3}$	$\hat{\sigma_3}$
8.1	0.32	8.84	0.033	9.51	0.211

With the weights $\hat{w_1}$, $\hat{w_2}$, and $\hat{w_3}$ having values 0.43, 0.18, and 0.39 respectively where $\sum_{k=1}^{3} w_i = 1$. So, the analytical structure of the estimated PDF of Fully democratic countries of the world is given by-

$$f(x) = \begin{cases} 0.66e^{-11.63(-9.58+x)^2} + 0.44e^{-2.87(-8.59+x)^2} \\ +0.72e^{-42.05(-7.75+x)^2}, & 7.63 \leq X \leq 10 \\ 0, & \text{otherwise} \end{cases} \tag{3.6}$$

The graph of the PDF of equation 3.6 is given by Figure 3.6 below:



(a) PDF of Least Corrupt Country **(b)** PDF of Least Corrupt for $P(X \geq 9)$

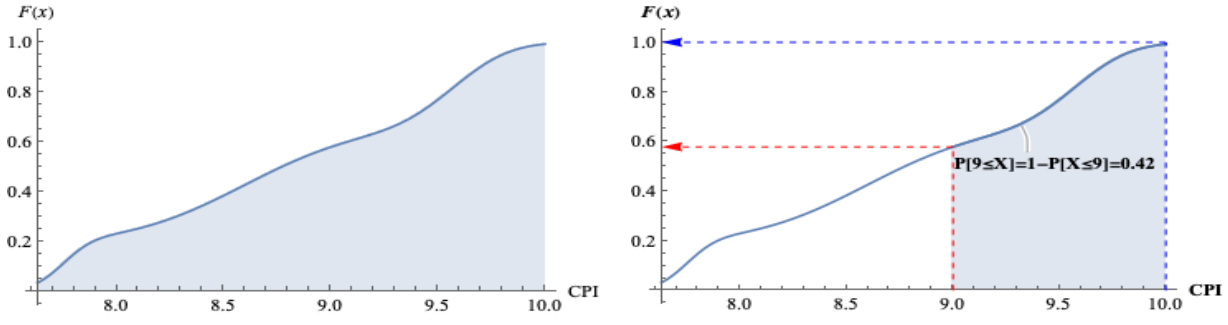
Figure 3.6: Plotting PDF of Least Corrupted Countries of the World

The expected value and variance of Least corrupted data subset is 8.769 and 0.543. That is, if a country is selected at random from this cluster we expect it's CPI score will be approximately 8.769. Also, the probability that a country will have a CPI score of more than 9 is 0.42 as shown in **Figure (3.6b)**.

The CDF of the CPI scores of Least Corrupted Countries of the world is given by-

$$F(x) = P(X \leq x) = \frac{w_1 \text{erfc}\left(\frac{\mu_1 - x}{\sqrt{2}\sigma_1}\right)}{2\left(w_1 + w_2 + w_3\right)} + \frac{w_2 \text{erfc}\left(\frac{\mu_2 - x}{\sqrt{2}\sigma_2}\right)}{2\left(w_1 + w_2 + w_3\right)} + \frac{w_3 \text{erfc}\left(\frac{\mu_3 - x}{\sqrt{2}\sigma_3}\right)}{2\left(w_1 + w_2 + w_3\right)} \tag{3.7}$$

It's graph is given by Figure (3.7) as follows:



(a) CDF of Least Corrupt Country **(b)** CDF of Least Corrupt for $P(X \geq 9)$

Figure 3.7: CDF of Least Corrupt Countries of the World

The plotting of **Figure 3.7b** is very useful in the case if anyone wants to calculate the probability of any country selected at random from this subset of population and curious about the probability of that country will have a score more than 9 (i.e. $P(X \geq 9) = 1 - P(X \leq 9)$), then that probability is **0.42** as shown in Figure 3.7b.

3.2.3 PDF of "Fairly Corrupted"Countries

We shall now proceed to find the probability distribution that characterize the probabilistic behavior of only the CPI data for Fairly Corrupted countries. To do this we have implemented the same steps we have used in finding the overall PDF of CPI scores for all classifications. For this purpose, we have started with the basic descriptive statistics of Fairly Corrupted countries.

Table 3.7: Descriptive Statistics of Fairly Corrupted Countries of the World

Fairly Corrupted Countries CPI scores				
Mean	Median	Standard Deviation	Skewness	Kurtosis
7.038	7.19	0.45	0.785	2.37

From the table above, we see that, this subset of the overall data is slightly left skewed with a value of 0.785 and it has a mean 7.038. The histogram of Fairly Corrupted Countries is given below Figure 3.8. From this histogram, the implication is that we need to fit some sort of Left skewed probability density function.

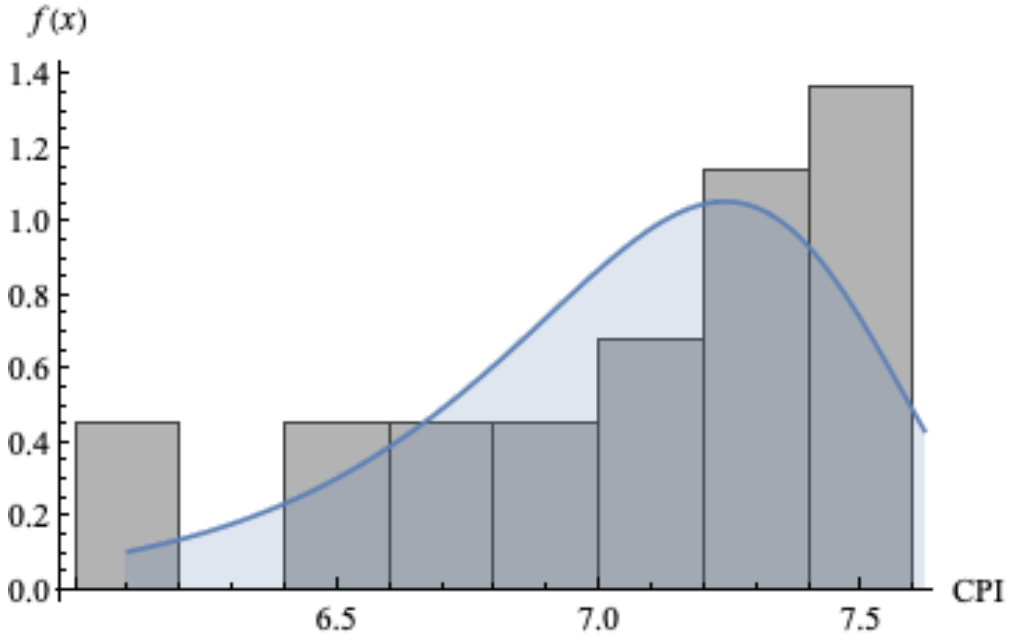


Figure 3.8: PDF fitted to Histogram to the Fairly Corrupted Countries

Using the three goodness-of-fit tests to the present data of Fairly corrupted countires we have identified that the data can be characterized probabilistically by the "Gumbel PDF". The justification of this selection is confirmed by the three methods of goodness-of-fit that we used in Table 3.8 given below confirms that the best fitted pdf for the Fairly Corrupted country data is the **Gumbel probability density function**.

Table 3.8: Goodness-of-Fit Summary for Fairly Corrupted Scores

	α	$p-value$	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.877	Do Not Reject
Anderson-Darling	0.05	0.814	Do Not Reject
Chi-Squared	0.05	0.956	Do not Reject

Thus, the fitted theoretical PDF of the subject data is given by-

$$f(x) = \begin{cases} \frac{e^{\frac{x-\alpha}{\beta}} - e^{\frac{x-\alpha}{\beta}}}{\beta}, & -\infty \leq x \leq \infty \\ 0, & \text{otherwise} \end{cases} \tag{3.8}$$

The following table shows the estimated MLEs of the parameters.

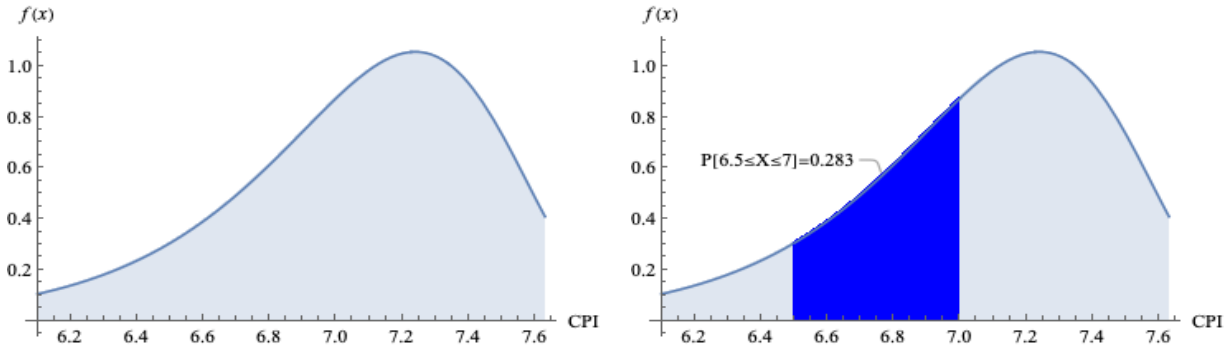
Table 3.9: MLEs of Fairly Corrupted Countries

MLEs of Fairly Corrupted Countries CPI scores	
$\hat{\alpha}$	$\hat{\beta}$
7.237	0.319

So, the analytical form of the pdf of Fairly corrupted countries of the world is given as follows:

$$f(x) = \begin{cases} 2.87e^{2.87(x-7.239)-e^{2.87(x-7.239)}}, & 6.1 \leq X \leq 7.62 \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

The graph of the PDF of equation 3.9 is given by Figure 3.9a below:



(a) PDF of Fair Corrupt Country **(b)** PDF of Fair Corruption for $P(6.5 \leq X \leq 7)$

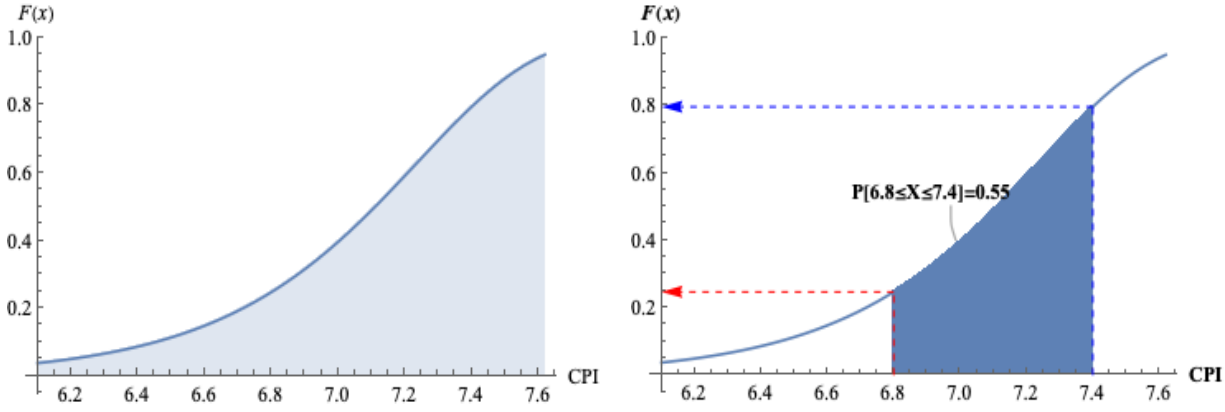
Figure 3.9: Plotting PDF of Fairly Corrupted Countries of the World

The expected value and variance of the Least corrupted data subset is 7.038 and 0.199. That is, if a country is selected at random from this cluster we expect it's CPI score will be approximately 7.038. Also, the probability that a country will have a CPI score between 6.5 and 7 is 0.283 as shown in Figure 3.9b.

The cdf of the CPI of Fairly corrupt countries of the world is given by-

$$F(x) = P(X \leq x) = 1 - e^{-e^{\frac{x-\alpha}{\beta}}}, \tag{3.10}$$

Its graph is given in the Figure 3.10 below:



(a) CDF of Fair Corrupt Contry **(b)** CDF of Fair Corrupt for $P(6.8 \leq X \leq 7.4)$

Figure 3.10: CDF of Fairly Corrupt Countries of the World

The plotting of Figure 3.10b is very useful in the case if anyone wants to calculate the probability of any country selected at random from this subset of population and estimate the probability of that country will have a score between 6.8 and 7.4 (i.e. $P(6.8 \leq X \leq 7.4) = P(X \leq 7.4) - P(X \leq 6.8)$), then that probability is 0.55.

3.2.4 PDF of "Moderately Corrupted"countries

In this section of our study, we will proceed to find the probability distribution that characterize the probabilistic behavior of only the CPI data for **Moderately Corrupted** countries. To do this we have implemented the same steps we have used in finding the PDF of CPI scores for previous two categories. For this purpose, we have started with the basic descriptive statistics of Moderately Corrupted countries.

Table 3.10: Descriptive Statistics of Moderately Corrupted Countries

Moderately Corrupted Countries CPI scores				
Mean	Median	Standard Deviation	Skewness	Kurtosis
4.89	4.93	0.68	0.132	1.92

From the table above, we see that, this subset of the overall data has a mean 4.89. The histogram of Moderately Corrupted Countries is given below Figure 3.11. From this histogram, the implication is that we need to fit some sort of right skewed probability density function.

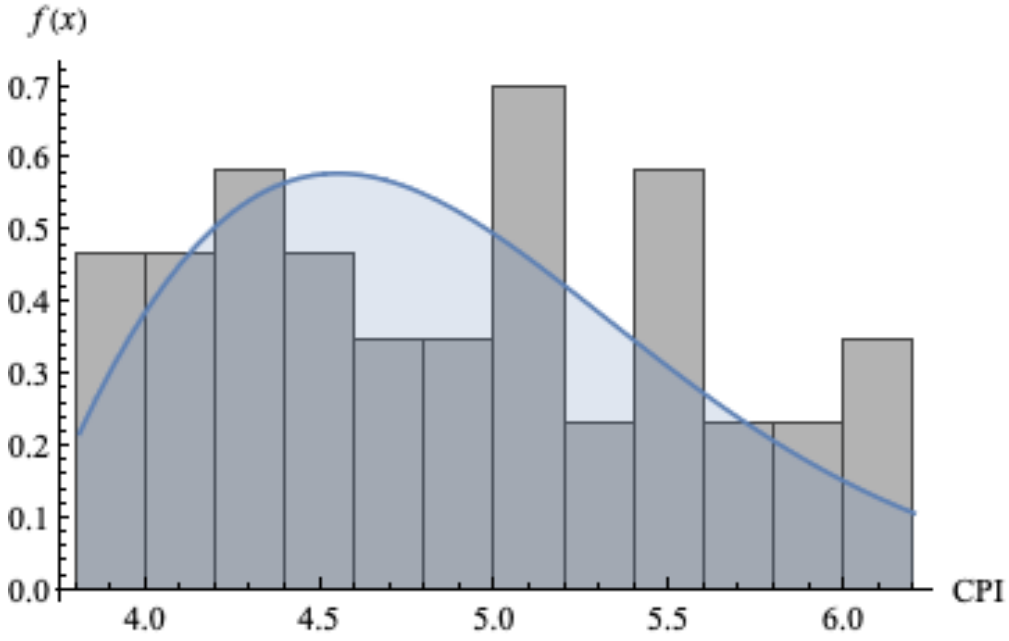


Figure 3.11: Histogram of Moderately Corrupted Countries of the World

Using the three goodness-of-fit tests to the present data of Moderate corrupted countires we have identified that the data can be characterized probabilistically by the *"Weibull PDF"*. The justification of this selection is confirmed by the three methods of goodness-of-fit that we used in Table 3.11 given below confirms that the best fitted pdf for the Fairly Corrupted country data is the **Weibull** $3-P$ **probability density function**.

Table 3.11: G-O-F Summary for Moderately Corrupted Countries

	α	$p-value$	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.881	Do Not Reject
Anderson-Darling	0.05	0.8572	Do Not Reject
Chi-Squared	0.05	0.985	Do not Reject

Thus, the fitted theoretical PDF of the subject data is given by-

$$f(x) = \begin{cases} \frac{\alpha e^{-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} \left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta}, & x > \gamma \\ 0, & \text{otherwise} \end{cases} \tag{3.11}$$

and the MLEs of this PDF of equation 3.11 are given in the following table:

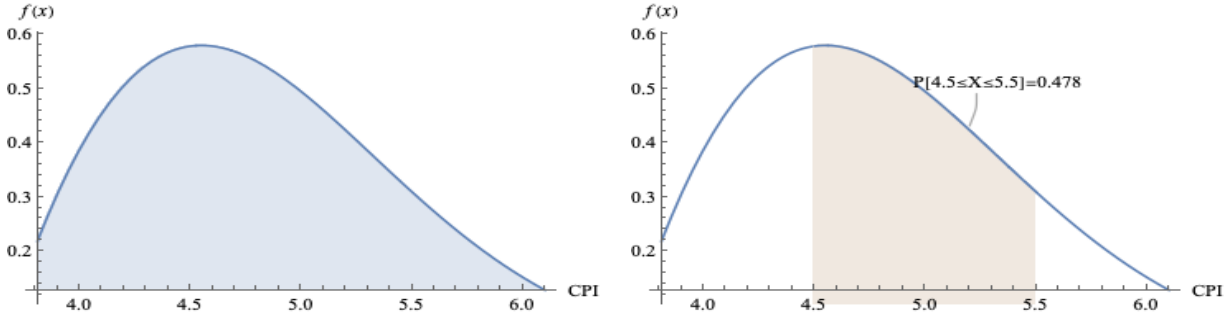
Table 3.12: MLEs of Moderately Corrupted Countries of the World

MLEs of the Moderately Corrupted Countries		
$\hat{\alpha}$	$\hat{\beta}$	$\hat{\gamma}$
1.938	1.41	3.64

The analytical form of this PDF with estimated parametes is given as follows:

$$f(x) = \begin{cases} 0.989e^{-0.544(x-3.64)^{1.82}}(x-3.64)^{0.819}, & x > 3.64 \\ 0, & \text{Otherwise} \end{cases} \tag{3.12}$$

The graph of the PDF of equation 3.12 is given by Figure 3.12a below:



(a) PDF of Moderately Corrupted Country **(b)** PDF for $P(4.5 \leq X \leq 5.5)$

Figure 3.12: Plotting PDF of Moderately Corrupted Countries of the World

The expected value and variance of Moderately corrupted country data subset is 4.89 and 0.499. That is, if a country is selected at random from this cluster we expect it's CPI score will be approximately 4.89. Also, the probability that a country will have a CPI score between 4.5 and 5.5 is 0.478 as shown in Figure **3.12b**.

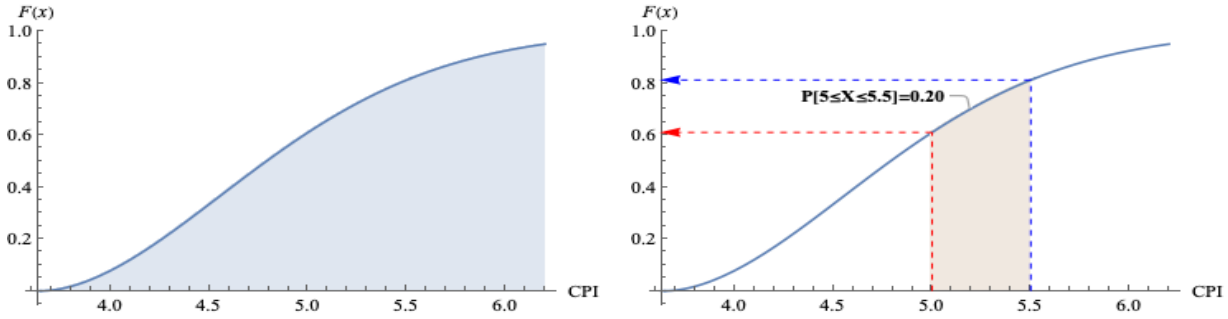
The cumulative distribution function of the Moderately Corrupted Countries of the world is given by-

$$F(x) = P(X \leq x) = 1 - e^{-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}} \tag{3.13}$$

The CDF of the Moderately corrupted country CPI scores with the estimated MLEs of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$ is given below:

$$F(x) = \begin{cases} P(X \leq x) = 1 - e^{-0.544(x-3.65)^{1.82}}, & 3.65 < X \\ 1, & X > 7.62 \end{cases} \tag{3.14}$$

The corresponding plotting of CDF function is as follows:



(a) CDF of Moderate Corruption Country **(b)** CDF of Moderate for $P(5.0 \leq X \leq 5.5)$

Figure 3.13: Plotting CDF of Moderately Corrupted Countries of the World

This particular type of plotting of Figure 3.13b is very useful in the case if anyone wants to calculate the probability of any country selected at random from this subset of population and curious about the probability of that country will have a score more than 5 but less than 5.5 (i.e. $P(5 \leq X \leq 5.5)$), then that probability is 0.20 as per the subject data subset.

3.2.5 PDF of "Highly Corrupted"Countries

In this section of our study of CPI scores, we will proceed to find the probability distribution that characterize the probabilistic behavior of only the CPI data for **Highly Cor-**

rupted countries. To do this we have implemented the same steps we have used in finding the PDF of CPI scores for previous three categories. For this purpose, we have started with the basic descriptive statistics of Highly Corrupted countries.

Table 3.13: Descriptive Statistics of Highly Corrupted Countries

Highly Corrupted Countries CPI scores				
Mean	Median	Standard Deviation	Skewness	Kurtosis
2.269	2.578	1.061	-0.427	1.965

From the table above, we can see that the distribution of the CPI scores of Highly Corrupted countries are some kind of left skewed as it is shown in Figure 3.14.

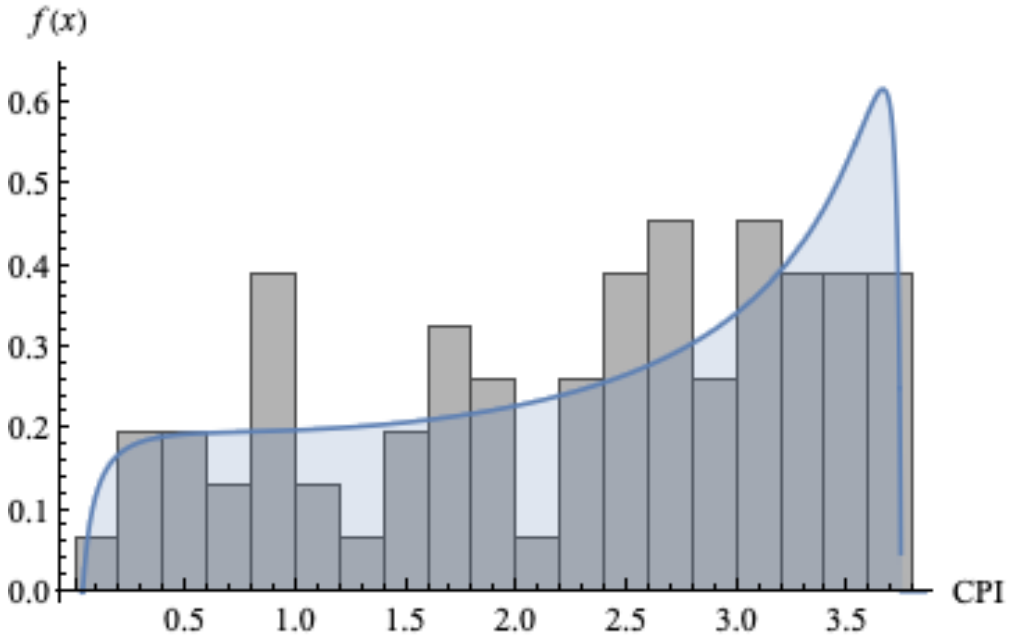


Figure 3.14: Histogram of Highly Corrupted Countries of the World

Using the three goodness-of-fit tests for the best candidate that characterized the Highly corrupted data subset, it turns out that "Johnson SB (4p)"PDF described the probabilistic characteristic of Highly Corrupted data subset and the G-O-F tests confirmed this fact in the following table 3.14 below:

Table 3.14: G-O-F Summary for Highly Corrupted Countries of the World

	α	$p-value$	Do Not Reject/Reject
Kolmogorov-Smirnov	0.05	0.992	Do Not Reject
Anderson-Darling	0.05	0.1342	Do Not Reject
Chi-Squared	0.05	0.747	Do not Reject

The analytical structure of the **Johnson SB (4p)** is given by equation 3.15 below:

$$f(x) = \begin{cases} \frac{\delta\sigma e^{-\frac{1}{2}\left(\gamma+\delta\log\left(\frac{x-\mu}{\mu+\sigma-x}\right)\right)^2}}{\sqrt{2\pi}(x-\mu)(\mu+\sigma-x)}, & \mu < x < \mu+\sigma \\ 0, & \text{otherwise} \end{cases} \tag{3.15}$$

The estimated MLEs for PDF in equation in 3.15 given in the following table 3.15:

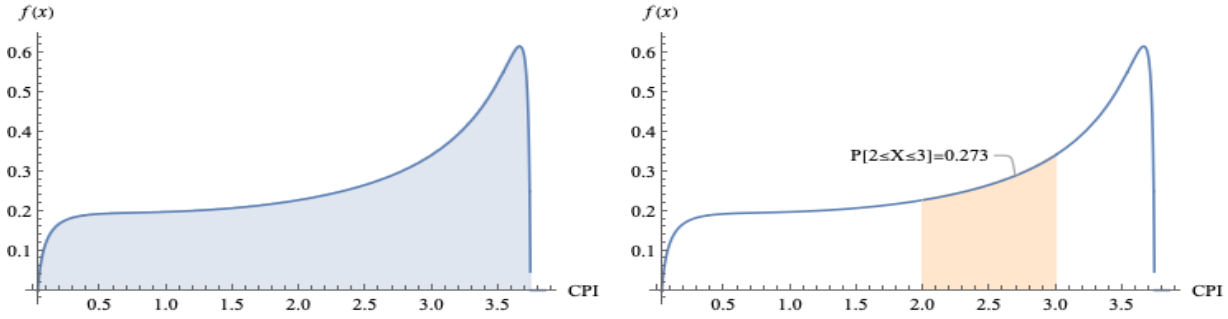
Table 3.15: MLEs of Highly Corrupted Countries PDF

MLEs of Highly Corrupted Countries			
γ	δ	μ	σ
-0.364	0.552	0.0288	3.712

The fitted pdf with estimated parameters is given by equation 3.16 below:

$$f(x) = \begin{cases} \frac{0.82e^{-\frac{1}{2}\left(0.55\log\left(\frac{x-0.029}{3.741-x}\right)-0.36\right)^2}}{(x-0.029)(3.741-x)}, & \text{if } 0 < X < 3.65 \\ 0, & \text{Otherwise} \end{cases} \tag{3.16}$$

The graph of the pdf of equation 3.16 is given in the Figure 3.15a;



(a) PDF of Highly Corrupt Countries **(b)** PDF of Highly Corrupt for $P(2 \leq X \leq 3)$

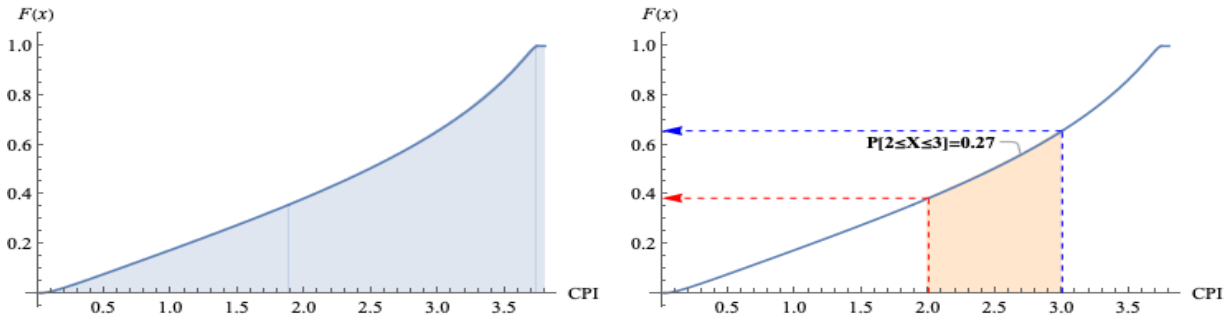
Figure 3.15: Plotting PDF of Highly Corrupted Countries of the World

The expected value and variance of Highly corrupted country data subset is 2.273 and 1.1648. That is, if a country is selected at random from this cluster we expect it's CPI score will be approximately 2.3. Also, the probability that a country will have a CPI score between 2 and 3 is 0.273 as shown in Figure **3.15b**.

The cumulative distribution function is given by-

$$F(x) = P(X \leq x) = \begin{cases} \frac{1}{2}\text{erfc}\left(-\frac{0.552\log\left(\frac{x-0.029}{3.74-x}\right)-0.36}{\sqrt{2}}\right), & 0.03 < x < 1.88 \\ \frac{1}{2}\left(\text{erf}\left(\frac{0.552\log\left(\frac{x-0.029}{3.74-x}\right)-0.36}{\sqrt{2}}\right)+1\right), & 1.88 \leq x < 3.74 \\ 1, & x \geq 3.74 \end{cases} \tag{3.17}$$

The CDF plot of the equation 3.17 is given below:



(a) CDF of Highly Corrupt Countries **(b)** CDF of Highly Corrupt for $P(2 \leq X \leq 3)$

Figure 3.16: Plotting CDF of Highly Corrupted Countries of the World

From this plotting of Figure **3.16b** is very useful in the case if anyone wants to calculate the probability of any country selected at random from this subset of population and curious about the probability of that country will have a score between 2 and 2.5 (i.e. $P(2 \leq X \leq 3)$), then that probability is 0.27.

3.3 Contributions

We have developed parametric analysis of Corruption Perception Index (CPI) of 175 countries in the world from the data extracted from Transparency International and World Bank. The following useful information and insights are very significant in the social sciences and economical paradigm in any countries overall infrastructure perspective.

1. The overall CPI scores for 175 countries to be the Mixture of 4-Gaussian PDF and this PDF can be used to characterize the probabilistic behavior of the CPI scores of any country.

2. We have characterized the probabilistic behavior of CPI scores of each of the four categories of corrupted countries around the world, and we can obtain other useful information such as Expected values for each of the countrys CPI scores as well as the confidence limits of that countrys score.

3. The visual postulation of probability density functions and the respective cumulative distribution functions could be very useful tool to facilitate the decsion makers to identify potential country to be avoided for being fallen into Highly Corrupted country subcategory.

4. These probability density functions can be used to as prior for bayesian inference in future studies.

4 Statistical model for detecting Probability of Severity Level of Hemophilia *A*

4.1 Introduction

As a rare bleeding disorder disease, it is vital to develop and identify the severity level of hemophilia. Typically, this is done by doing several blood tests, also termed as screening tests in the medical science domain. The types of screening tests include - Complete Blood Count (CBC), Activated Partial Thromboplastin Time (APTT) test, Prothrombin Time (PT) Test, Fibrinogen Test, Clotting Factor Tests [30]. In the case of hemophilia study, the last test (CFT - Clotting Factor Tests), is the medical standard to detect and tag the severity level of hemophilia. But, in some study, the severity level based on clotting factor - factor **VIII** or *F8* is one of the predictors in the outcome of Immune tolerance induction [15]. Studies have been done only taking the Inhibitors and F8 into considerations to study these attributes only on African American and Black population [47]. In other literature, the parametric study has been done only on F8 mutation type and Inhibitor development [24]. However, very little has been done on developing a statistical model that can identify and predict the severity level with the concept of probability considering the confidence limit on those probability predictions. For these reasons, the main focus of this study is to develop a statistical prediction model to predict the probability of severity level based on diagnosis reports collected from different HTCs (Hemophilia Treatment Center) in the USA.

4.2 Hemophilia - a Rare Disease

Hemophilia is caused by a mutation or change in one of the genes that provide instructions for making the clotting factor proteins needed to form a blood clot. This change or mutation can prevent the clotting protein from working properly or from being missing altogether. These genes are located on the X chromosome. Males have one X, and one Y chromosome (XY) and females have two X chromosomes (XX). Males inherit the X chromosome from their mothers and the Y chromosome from their fathers. Females inherit one X chromosome from each parent, as shown in the following Figure 4.1:

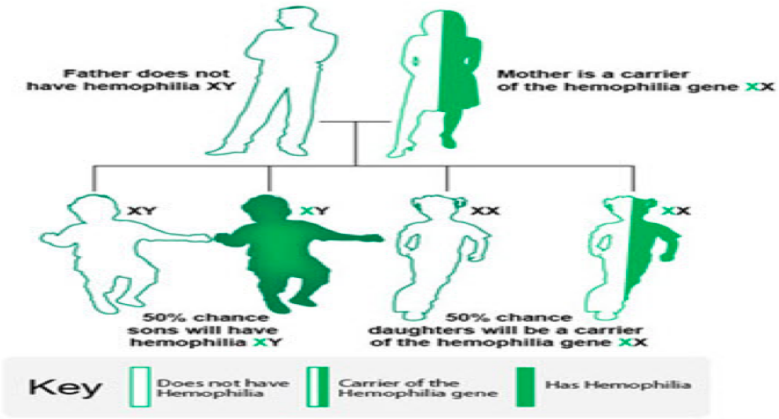


Figure 4.1: Parental relationship to the children (Source:CDC)

The X chromosome contains many genes that are not present on the Y chromosome. This implies that males only have one copy of most of the genes on the X chromosome, whereas females have two copies. Thus, males can have a disease like hemophilia if they inherit an affected X chromosome that has a mutation in either the factor VIII or factor IX gene. Females can also have hemophilia, but this is much rarer. In such cases, both X chromosomes are affected, or one is affected, and the other is missing or inactive. In these females, bleeding symptoms may be similar to males with hemophilia. A female with one affected X chromosome is a "carrier" of hemophilia. Sometimes a female who is a carrier

can have symptoms of hemophilia. Also, she can pass the affected X chromosome with the clotting factor gene mutation on to her children [9].

4.3 Methodology

Because the focus of this study is to identify and develop a statistical model that can predict the severity level of the patient's/individual's Hemophilia, we have collected data from the open and freely accessible database of the CDC called CHAMP 8 [41] database. We will do descriptive the descriptive analysis, and then a model will be identified for classification prediction through some classical model building methods.

4.3.1 Data Description

The CHAMP mutation list was collected and compiled by CDC and made freely accessible and available to public use by downloading at CHAMP Mutation database for statistical reporting and analysis only.

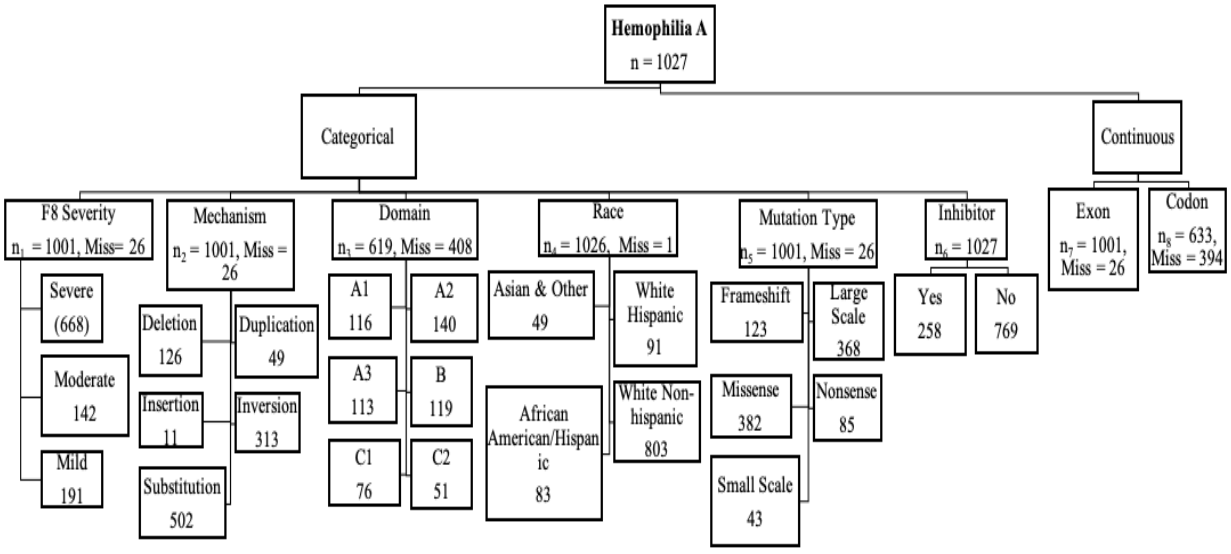


Figure 4.2: Schematic Diagram of CHAMP *F*8 Hemophilia **A** Data

In this present statistical study and analysis, we have considered the Severity level definition based on HGVS-standardized nomenclature [23]. It states that, if the clotting factor - F8 in the blood is between 50% to 100%, then the person/individual is in *Normal* state and does not have hemophilia. On the other hand, if the F8 ranges between $5\% < F8 < 50\%$ then in medical science it is termed as *Mild Level* of Hemophilia A and if any individual's CFT (Clotting Factor Test) shows the results as $1\% \leq F8 \leq 5\%$ or $F8 < 1\%$ then that individual is termed as having *Moderate* or *Severe* level of hemophilia respectively. This categorization/classification has been defined only by HGVS-standardized nomenclature as per medical research and study. So, we have used this as a response variable to the predictor/attributable variables such as Mutation, Mechanism, Subtype, Domain, Race, and Inhibitor information as the categorical attributes and Exon, Codon as the continuous covariates. The schematic diagram of the data set is shown in Figure 4.2.

From the diagram, we see that the data set consists of 6 categorical variables and two continuous variables, and each of the categories for every categorical variable has some missing values. In some cases, so do the continuous variables. Only the categorical variable *Inhibitor* does not contain any missing information, and because of the nature and objectives of this study, we have considered 'F8 Severity Level'the response variable and all other variables were considered as the predictor variables to be.

4.4 Statistical Analysis & Modeling

4.4.1 Uni-variate Analysis

To have a better illustration of the data in hand, we have started to examine each of the variables in the data set and delve through the descriptive statistic to have an idea of the different categorical variables. To see the proportions of the severity level, which is our target variable for this study, the following Figure 4.3 postulates those corresponding proportions.

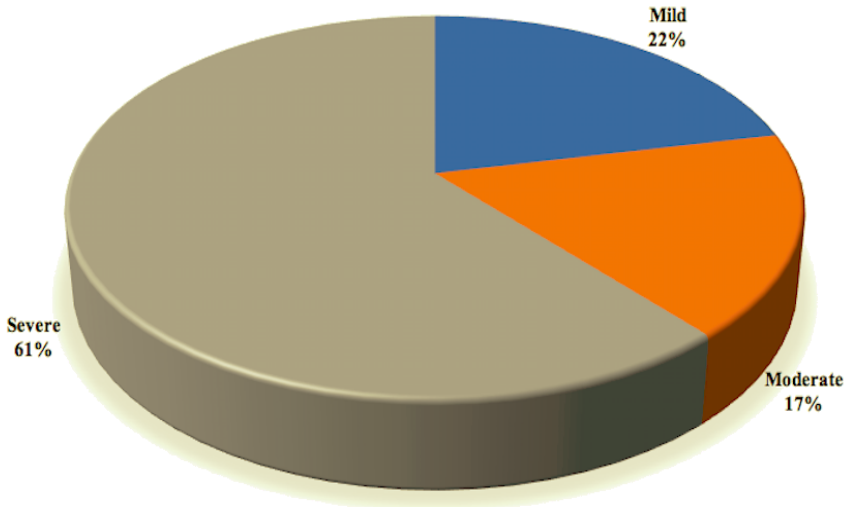


Figure 4.3: Pie Chart for Severity Level

From the same data set, if we take a look at the overall proportion of the individuals for the different races, then we have the following Figure 4.4.

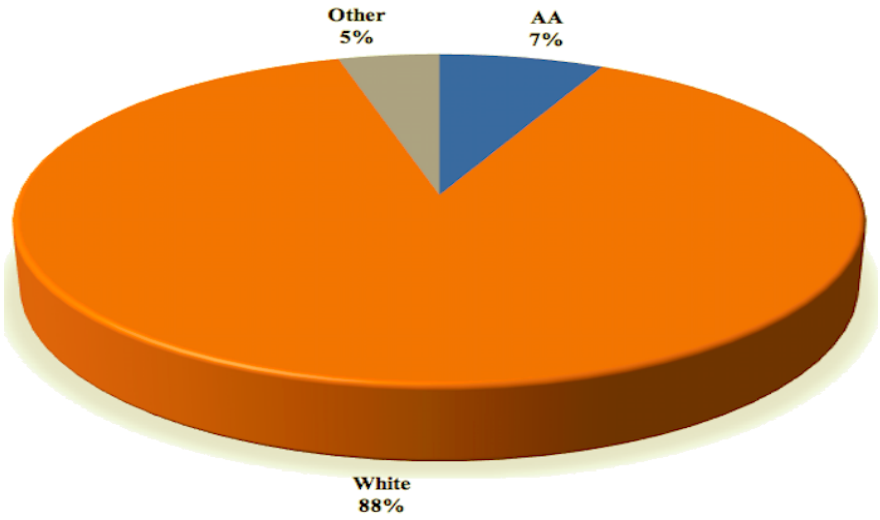


Figure 4.4: Pie Chart for Races

It indicates that about 88% of the individuals are white according to the data we have.

Similarly, we have visualized the other variable, which is History of Inhibitors [3] presence in the body for individuals that plays a very significant role in hemophilia. From the clinical research, it was found that approximately 15%-20% of people with hemophilia will develop an antibody—called an inhibitor—to the product used to treat or prevent

bleeding episodes. Developing an inhibitor is one of the most serious and costly complications of hemophilia [50].

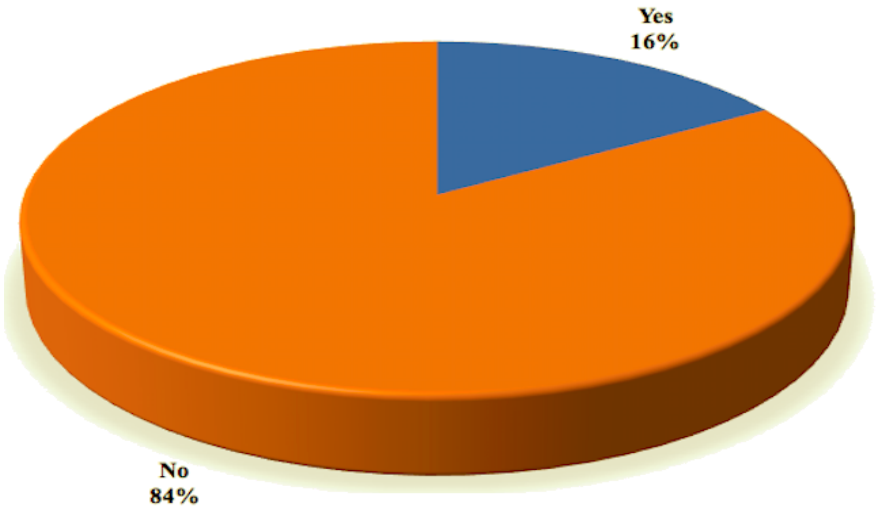


Figure 4.5: Pie Chart for Inhibitor History

84% of the individuals reported for hemophilia for **F8** that have a history of inhibitors. This inhibitor was present in the individuals' blood either due to inheritance from the parents or from the family. So, from this univariate analysis of some categorical variable will be used as a gateway to build a predictive statistical model to classify and estimate the probability of the severity level.

4.4.2 Statistical Model Building

For the study, we have explored the relationship and association (if any) of each of the categorical predictor variables against the response variable 'F8 Severity Level'through Mosaic plot. Then we have established the statistical model through the **Multinomial Logistic Regression**, **Generalized Logistic Regression** and **Cumulative Logistic Regression** (taking the ordered categories of the response variable - *F8 Severity Level* into consideration), and we have compared their results to find the better-fitted model if not the best. To formulate the statistical model that predicts the probability of the severity level of F8,

we have started with the cross-tabulation of predictor variable Mutation Mechanism and response variable F8 Severity Level presented in the table below.

Table 4.1: Cross Tabulation of Severity Level vs. Mutation Mechanism

	DNA change Mechanism					
	Deletion	Duplication	Insertion	Inversion	Substitution	Total
Severe	110	43	9	294	197	653
Moderate	7	5	1	6	116	135
Mild	1	0	0	1	185	187
Total	118	48	10	301	498	975
Missing = 52						

To visualize the above cross tabulation we have used the **Mosaic Plot** [22] to have a better insight of the information presented above given in the Figure 4.6.

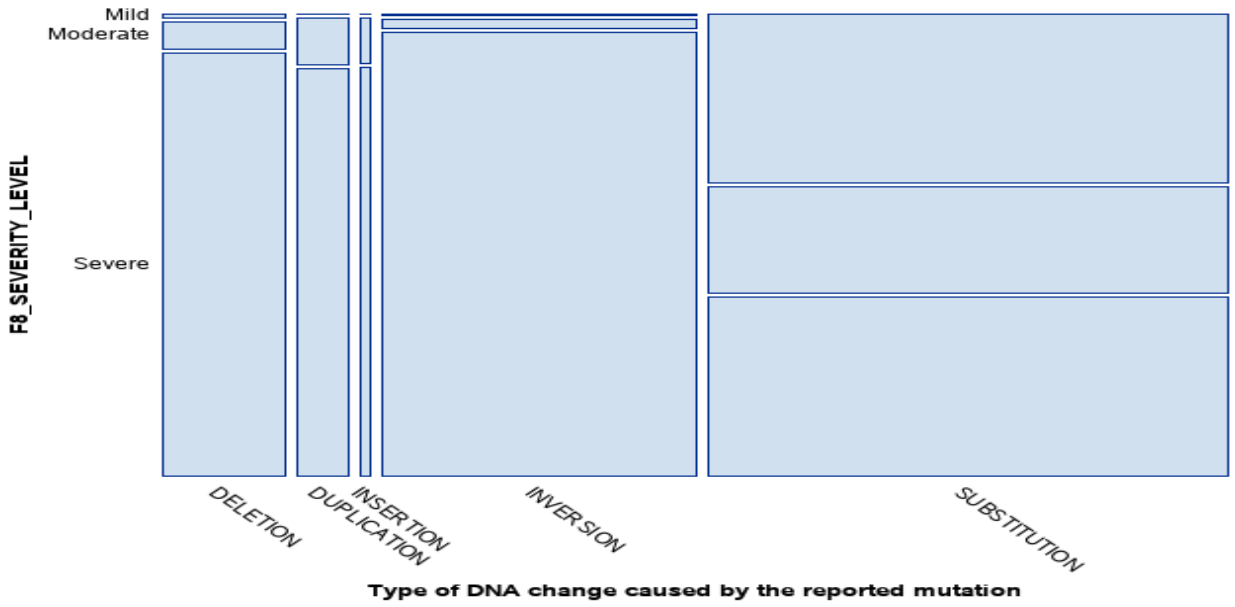


Figure 4.6: Severity Level of *F*8 vs. Mechanism of Mutation

The mosaic plot shows the distribution of mutation mechanism in the DNA categories in the $x - axis$ by dividing that axis into 5 intervals. Its shows that Inversion and Substitution mechanism are greatly associated to the Severity level of F8 in the individuals.

Similarly, we have checked the association between Severity level and Domain of the Mutation given in the cross-tabulation below.

Table 4.2: Cross Tabulation of Severity Level vs. Domain

	Location of Mutation (Domain)							
	A1	A2	A3	B	C1	C2	Other Signals	Total
Severe	50	46	50	106	23	29	364	668
	5%	4.6%	5%	10.59%	2.3%	2.9%	36.36%	66.75%
Moderate	28	23	23	9	26	12	21	142
	2.8%	2.3%	2.3%	0.9%	2.6%	1.2%	2.1%	14.2%
Mild	37	70	36	0	26	10	12	191
	3.7%	6.99%	3.6%	0%	2.6%	1%	1.2%	19.09%
Total	115	139	109	115	75	51	397	1001
	11.5%	13.89%	10.9%	11.49%	7.5%	5.1%	39.66%	100.04%

The mosaic plot of the information given above shows there is association of some of the categories of Mutation Domain with the Severity level of the F8 clotting factor presented in the blood.

In the Figure 4.7 given above, it is obvious that the other signals category of Domain variable has a very large proportion of population having Severe level of F8 clotting factor contained in their blood than all other categories of the Domain location of mutation.

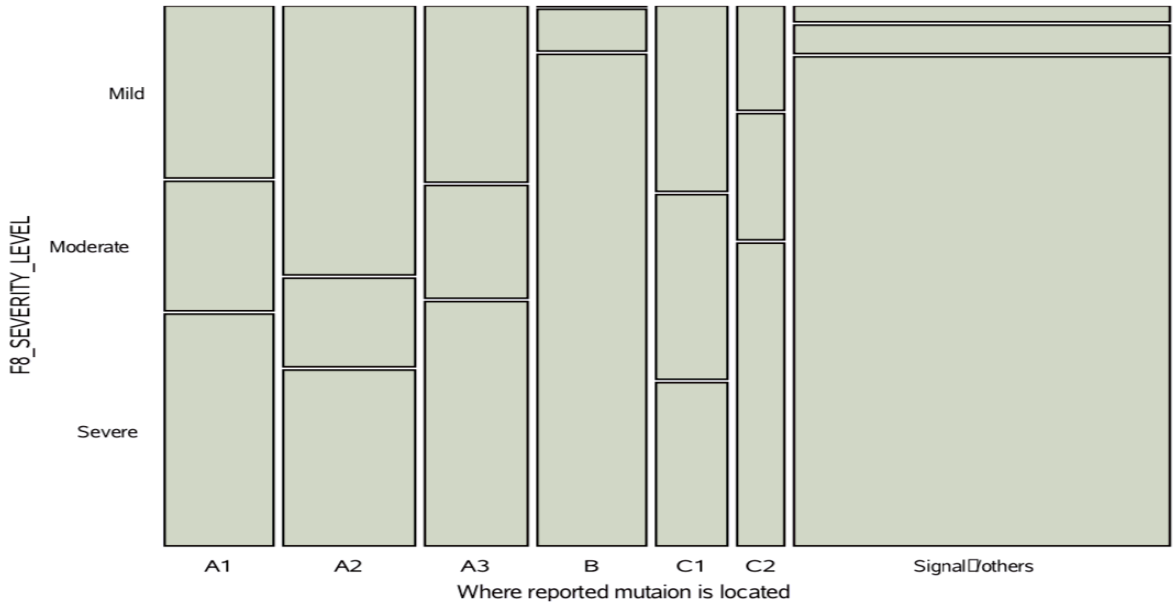


Figure 4.7: Severity Level of *F*8 vs. Domain

Also we have examined the cross-table relationship between Severity level and Race of individuals presented in Table 4.3.

Table 4.3: Cross Tabulation of Severity Level vs. Race

	Race of Individuals				
	Asian & Others	Afro American & Hispanic	White Hispanic	White Non-Hispanic	Total
Severe	33	64	64	506	667
	3.3%	6.4%	6.4%	50.6%	66.7%
Moderate	5	10	12	115	142
	0.5%	1%	1.2%	11.5%	14.2%
Mild	11	4	10	166	191
	1.1%	0.4%	1%	16.6%	19.1%
Total	49	78	86	787	1000
	4.9%	7.8%	8.6%	78.7%	100%
	Missing = 27				

The following figure depicts the relationship presented in the table above.

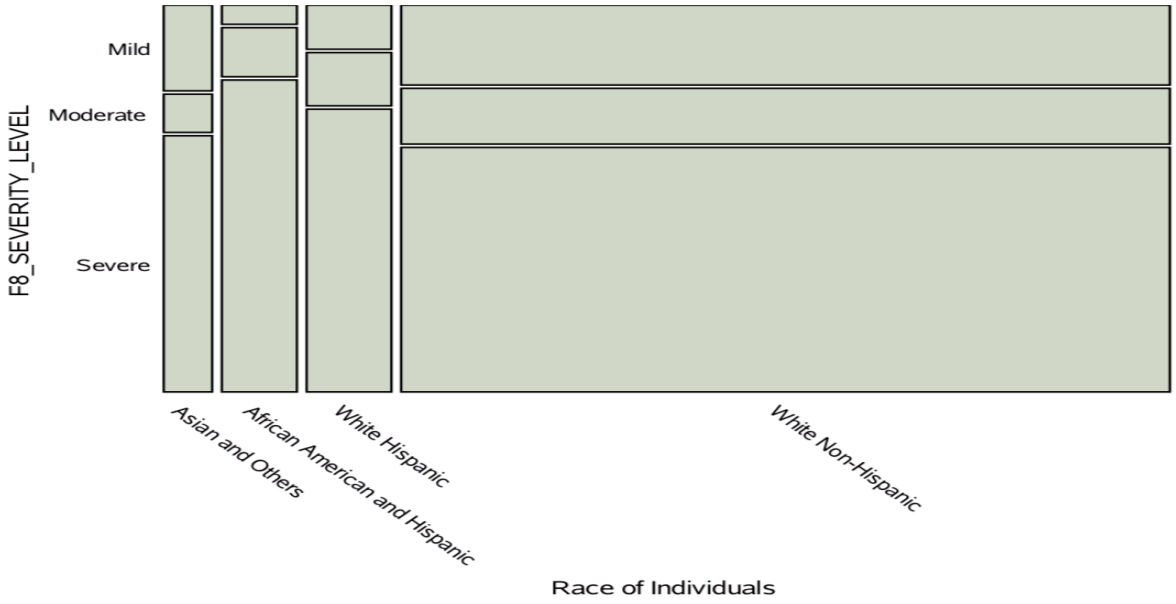


Figure 4.8: Severity Level of *F*8 vs. Race of Hemophilia **A** Data

This figure shows very strong relationship between White Non-Hispanic category of Race variable to the Severe category of Severity level of Hemophilia A. At the same time, we have explored the crosstable relationship between Severity level and Mutation Type variables in Table 4.4.

Table 4.4: Cross Tabulation of Severity Level vs. Mutation Type

	Mutation Type					
	Frameshift	Large Structure	Missense	Nonsense	Small Structure	Total
Severe	108	341	100	79	25	653
	11.08%	34.97%	10.26%	8.1%	2.56%	66.97%
Moderate	9	10	106	2	8	135
	0.92%	1.03%	10.87%	0.21%	0.82%	13.85%
Mild	0	2	175	1	9	187
	0%	0.21%	17.95%	0.1%	0.92%	19.18%
Total	117	353	381	82	42	975
	12%	36.21%	39.08%	8.41%	4.31%	100%
			Frequency Missing = 52			

Also, the mosaic plot of these variables are given in the figure below.

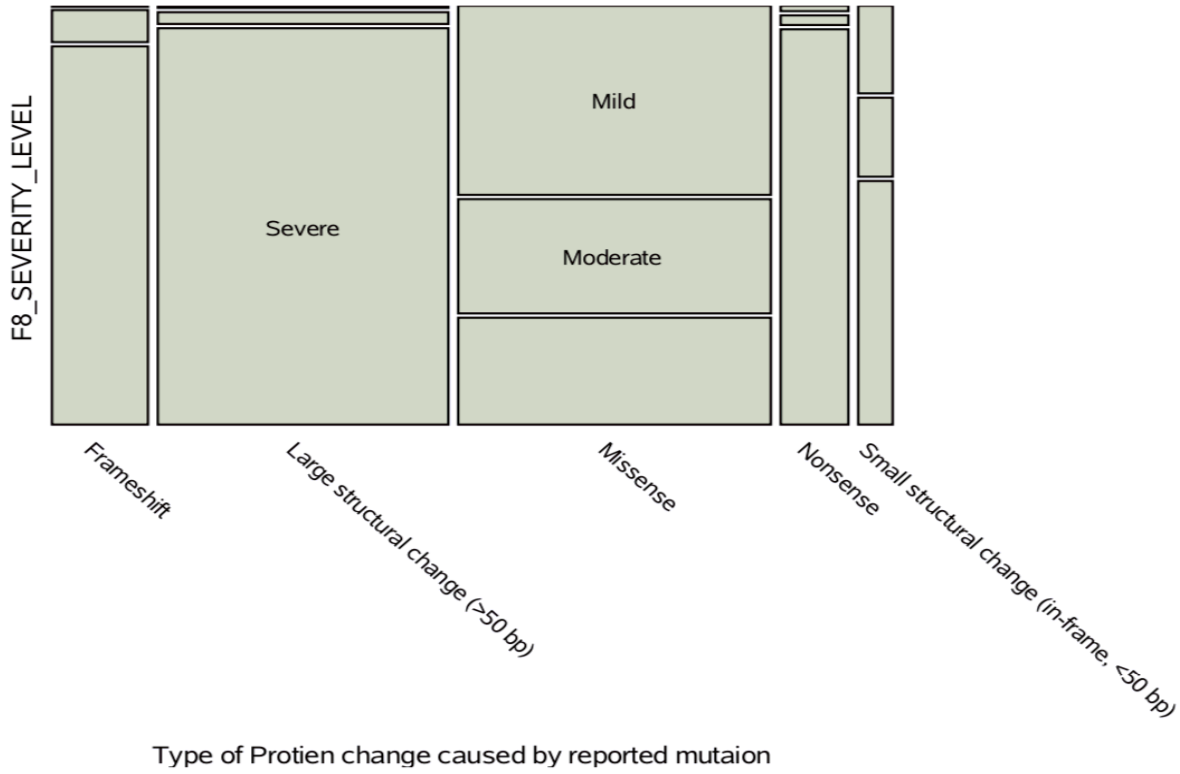


Figure 4.9: Severity Level of *F*8 vs. Mutation Type

Here in the figure above, we see the largest category of mutation type is Large structural change in the protein is strongly associated to the severe level of hemophilia A bleeding disorder. It implicates the fact that large structural change category of Mutation type covariate might have a significant effect on the outcome variable Severity Level.

Table 4.5 represents the association of inhibitor built in protein to help the blood clotting and prohibit the mutation in the DNA causing F8 to reduce is very important variable associated with the Severity level of the Hemophilia A.

Table 4.5: Cross Tabulation of Severity Level vs. Inhibitor History

	Inhibitor		
	No	Yes	Total
Severe	468	200	668
	46.75%	19.98%	66.73%
Moderate	124	18	142
	12.39%	1.8%	14.19%
Mild	177	14	191
	17.68%	1.4%	19.08%
Total	769	232	1001
	76.82%	23.18%	100%
Missing = 26			

This table shows the evidence that approximately 77% of the individuals do not build the inhibitor in their blood that causing almost 47% to have a very severe level of Hemophilia A. Figure 4.10 shows that, the biggest tile in this mosaic is at the intersection of No and Severe category of Inhibitor and severity level variables respectively.

After considering all of the associations presented in the crosstables and mosaic plots, we have built models from several approaches. Since we have a response variable with three categories and those categories are ordered based on the values and nature of the study dataset, the Cumulative Logistic Regression [6] or Ordinal Logistic Regression[3] is the most appropriate modeling approach suggested by some scholars in the medical research[40, 4]. Also, Multinomial Logistic Regression[7] and Generalized Estimating Equations (GEE)[28] are some alternative options considering the response variables are name categories only.

Figure 4.10: Severity Level of *F*8 vs. Mutation Type

So we have built the model considering all the modeling approaches mentioned above, and we have selected the model based on **Log-Likelihood, Akaik Infomation Criterion (AIC) and Bayesian Information Criterion (BIC)**[7]. Considering all the factors, we have started with the "Multinomial Logistic Regression". Then we have modeled the data through the "Generalized Estimating Equation (GEE)"[34, 61, 62] and after this, we have modeled through "Cumulative Logistic Regression"or in other words termed as "Ordinal Logistic Regression". After estimating the parameters for each of the models, those were compared with each other.

There several types of Multinomial Logistic models can be used based on the type of information on response variables in data at hand. If the response variables are in the Nominal scale, then generalized logit models (GLM) and the conditional logit models (CLM) can be used[49]. The GLM consists of estimating parameters of several binary logistic models simultaneously[52]. On the other hand, the CLM is used in biomedical research in order to estimate relative risks in matched case-control studies[8, 36]. Since, we have a Polytomous response variable that is in ordered structure with a set of regressors (attributes), the most appropriate modeling method would be the **Cumulative Logistic Regression**[37]. However, to have the best approximate model of the real-world phenom-

ena, we have estimated all the models above and compared the statistic to make our final decisions, and the following sections will be discussed on our model building methods.

Generalized Logit Model (GLM) - Multinomial

The generalized logistic model focuses on the individual that is considered the unit of analysis, and this GLM model uses individual characteristics as explanatory variables. The explanatory variables that were characteristic of an individual are constant over the choices of the response variable. Considering m nominal choices of the response variable, $\prod_{jk}$ denote the probability that the individual j falls in category k. Also, let X_j represent the characteristic of individual j. The probability of the individual j falling in category k is given by

$$\Pi_{jk} = \frac{exp(\gamma_k' X_j)}{\sum_{l=1}^{m} exp(\gamma_l' X_j)} = \frac{1}{\sum_{l=1}^{m} exp\left[(\gamma_l - \gamma_k)' X_j\right]}$$

Here, $\gamma_1 \cdots \gamma_m$ are m vectors of unknown parameters where each of the estimates are different even though X_j is constant acros other categories or choices. The model to be fit

$$log\left(\frac{\prod_{hjk}}{\prod_{hjl}}\right) = \sum_{l=1}^{m} \gamma_l' X_j \tag{4.1}$$

Cumulative Logistic Model (CLM)

Suppose, Y takes values of $y_1, y_2, \cdots, y_m$ such that $y_1 < y_2, < \cdots, < y_m$, it is assumed that the observed variable is categorized through a continuous latent variable U such that, $Y = y_i \Leftrightarrow \alpha_i - 1 < U \leq \alpha_i$, $i = 1, \cdots, m$ where $-\infty = \alpha_0 < \alpha_1 < \cdots < \alpha_m = \infty$. The assumption on U is that the value will be determined by the attributable variable vector X in the form of a linear function $U = -\lambda' x + \epsilon$ where, λ is vector of regression coefficients and ϵ is a random variable with a distribution function F that assumed to follow $P(Y \leq y_i | x) = F(\alpha_i + \lambda' x) \sim Logistic$ distribution. Moreover, in the CLM concept,

alternatively known as *proportional odds model*, it is assumed that the predictor variable, let X, takes different values for each of the alternative categories and effect of a unit of X is assumed constant acros different alternative categories. Under these assumptions, the probability that an individual j will fall into category k is

$$\Pi_{jk} = \frac{exp(\lambda' X_{jk})}{\sum_{l=1}^{m} exp(\lambda' X_{jl})} = \frac{1}{\sum_{l=1}^{m} exp\left[\lambda'(X_{jl} - X_{jk})\right]}$$

Modeling our Hemophilia A data by *GLM*:

Using the GLM method, the list of significantly useful variables ordered as per the $p - value$ of Wald Chi-square from smallest to the largest in the model are shown in the following Table 4.6.

Table 4.6: Ranking of Covariates in the Generalized Logistic Regression

Covariates (Attributes)	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Mutation (Type of Protein change)	6	42.5087	**<0.0001**
Domain (Location of Mutation Domain)	12	24.3374	**0.0183**
Race of Individuals	6	13.5698	**0.0348**
Inhibitor History	2	3.2079	0.2011
Mechanism (Type of DNA change)	6	5.0811	0.5335
Exon (Exon Number in the Mutation Location)	2	1.1978	0.5494
Codon (Codon Number in the Mutation Location)	2	0.913	0.6335

Taking statistically significant covariates as per the table above into consideration and letting "**Severe**"as the reference category of response variable, the final generalized models are:

$$
\begin{aligned}
log\left(\frac{\pi(Moderate)}{\pi(Severe)}\right) = &-0.8911 + 1.1035(Domain = A1) + 0.9389(Domain = A2) \\
&+ 0.7393(Domain = A3) + 1.9608(Domain = B) \\
&+ 1.6184(Domain = C1) + 0.7799(Domain = C2) \\
&- 0.3863(Race = AAH) - 0.7443(Race = AsO) \\
&+ 0.0641(Race = WH) - 3.3389(Mutation = Frameshift) \\
&- 2.6606(Mutation = LargeScale) \\
&- 4.1569(Mutation = Nonsense) \\
&- 0.5016(Mutation = SmallScale)
\end{aligned}
$$

and the equation for "Mild"category of response variable where "**Severe**"as the base reference category:

$$
\begin{aligned}
log\left(\frac{\pi(Mild)}{\pi(Severe)}\right) = &0.2887 + 0.2897(Domain = A1) + 1.1223(Domain = A2) \\
&+ 0.1228(Domain = A3) - 10.6146(Domain = B) \\
&+ 0.5026(Domain = C1) - 0.4855(Domain = C2) \\
&- 2.1079(Race = AAH) - 0.5492(Race = AsO) \\
&- 0.2666(Race = WH) - 15.6070(Mutation = Frameshift) \\
&- 5.3133(Mutation = LargeScale) \\
&- 4.6073(Mutation = Nonsense) \\
&- 1.4551(Mutation = SmallScale)
\end{aligned}
$$

Modeling our Hemophilia A data by *CLM*

In order to apply CLM method to our data set at hand, let our response variable be **F8 Severity Level** = $Y_j = \{y_{j1}, y_{j2}, y_{j3}\}$ where the ordering is in reverse order of category 1 indicates the most severe case (Severe) and category 3 indicates least severe (Mild). The associated probabilities are $\{\pi_{j1}, \pi_{j2}, \pi_{j3}\}$, and a cumulative probability of a response less than equal to Y_j is: $P(Y_j \leq y_{j3}) = \pi_{j1} + \pi_{j2} + \pi_{j3}$, then the CLM would be:

$$log\left(\frac{P(Y_j \leq y_{j3})}{P(Y_j > y_{j3})}\right) = log\left(\frac{P(Y_j \leq y_{j3})}{1 - P(Y_j \leq y_{j3})}\right) \tag{4.2}$$

The sequence of Cumulative Logit Models will be: $Model(L_1) : log\left(\frac{\pi_{j1}}{\pi_{j2}+\pi_{j3}}\right)$ and$Model(L_2) : log\left(\frac{\pi_{j2}}{\pi_{j3}}\right)$ Or alternatively,

$$L_1 : log\frac{Pr(Mild)}{Pr(\text{Moderate or Severe})} = \alpha_1 + \sum_{j=1}^{7} \beta_j X_{j|\gamma}$$

$$L_2 : log\frac{Pr(\text{Mild or Moderate})}{Pr(\text{Severe})} = \alpha_2 + \sum_{j=1}^{7} \beta_j X_{j|\gamma}$$

Here, γ is indicator for categories of each categorical variable and other notations:

$X_{1|1} = Mechanism(Deletion), X_{1|2} = Mechanism(Duplication),$

$X_{1|3} = Mechanism(Insertion), X_{2|1} = Domain(A1),$

$X_{2|2} = Domain(A2), X_{2|3} = Domain(A3), X_{2|4} = Domain(B), X_{2|5} = Domain(C1),$

$X_{2|6} = Domain(C2), X_{3|1} = Race(AAH), X_{3|2} = Race(Asian\&Others)$

$X_{3|3} = Race(WhiteHispanic), X_{4|1} = Mutation(Frameshift),$

$X_{4|2} = Mutation(Large_{Change}), X_{4|3} = Mutation(Nonsense),$

$X_{4|4} = Mutation(Small_{Change}),$

$X_{5|1} = Inhibitor(No), X_6 = Exon, X_7 = Codon$

Here, we should notice that the intercepts are changing from one model to another but the slopes are equal for all the models. Also, we need to estimate 2 - intercepts and p - slopes. In our case $p = 7$, so we will have 7 slopes for 7 covariates and 2 intercepts should make the full model. The estimated final model is given in equation 4.3.

$$\begin{cases} L_1 & : \textbf{-1.7155} - 3.221X_{1|1} - 3.944X_{1|2} - 3.267X_{1|3} \\ & \quad +1.39X_{2|1} + 2.3739X_{2|2} + 0.988X_{2|3} \\ & \quad +2.023X_{2|4} + 1.42X_{2|5} + 0.828X_{2|6} \\ & \quad -1.39X_{3|1} - 0.44X_{3|2} - 0.2773X_{3|3} \\ & \quad -0.627X_{4|1} - 4.592X_{4|2} - 0.117X_{4|3} \\ & \quad +0.4205X_{5|1} - 0.0727X_6 + 0.000679X_7 \\ L_2 & : \textbf{-0.4102} - 3.221X_{1|1} - 3.944X_{1|2} - 3.267X_{1|3} \\ & \quad +1.39X_{2|1} + 2.3739X_{2|2} + 0.988X_{2|3} \\ & \quad +2.023X_{2|4} + 1.42X_{2|5} + 0.828X_{2|6} \\ & \quad -1.39X_{3|1} - 0.44X_{3|2} - 0.2773X_{3|3} \\ & \quad -0.627X_{4|1} - 4.592X_{4|2} - 0.117X_{4|3} \\ & \quad +0.4205X_{5|1} - 0.0727X_6 + 0.000679X_7 \end{cases} \tag{4.3}$$

The model given in equation 4.3, is formulated considering all the attributable variables in the given data set of hemophilia A disease. But if the covariates are ranked based on their effects on the model, then there are some variables which are not statistically significant. The following Table 4.7 represents the ranking of the statistically significant covariates ranked from most statistically significant to statistically insignificant variables.

In model building, ranking of statistically significant variables are very relevant. From this ranking of siginificant variables, one can construct a model only by including the covariates those are contributing and explaing most of variations in the model. Also, inclu-

sion of the significant variables only in the model will prohibit anyone from overfitting and underfitting.

Table 4.7: Ranking of Covariates in the Cumulative Logistic Regression

Covariates (Attributes)	**DF**	**Wald Chi-square**	**Pr >ChiSq**
Mutation (Type of Protein change)	3	44.4268	**<0.0001**
Race of Individuals	3	13.5435	**0.0036**
Domain (Location of Mutation Domain)	6	16.7199	**0.0104**
Mechanism (Type of DNA change)	3	7.1198	0.0682
Inhibitor History	1	1.9719	0.1602
Exon	1	0.4538	0.5005
Codon	1	0.1836	0.6683

4.4.3 Proposed Model

From the previous section, it is recommended that, in the modeling purpose with the data set at hand, we should include only 3 IVs (Independent Variables) as per Table 4.3, where the statistically significant variables are ranked as per their corresponding $p-values$. Also, the model is compared according to some common model fit statistics given in the table 4.8 below.

Table 4.8: Model Comparison among estimated models

	All variables		**Categorical Variables Only**		**Significant Variables Only**	
	Cumulative	*GLM*	*Cumulative*	*GLM*	*Cumulative*	*GLM*
AIC	989.29	1282.95	1092.51	1107.28	976.89	1108.86
BIC	1077.92	1291.82	1190.14	1292.77	1006.21	1245.54
-2*Log*	-474.65	1278.95	-526.25	1031.28	-536.50	819.385

So, by taking this significance into considerations the final proposed model for the hemophilia dataset we have at our hand is given in the Equation 4.4 below:

$$
\begin{cases}
L_1 \quad : \textbf{-0.6485} \; + 0.4171X_{2|1} + 1.2400X_{2|2} + 0.3267X_{2|3} \\
\qquad +1.1248X_{2|4} + 0.5269X_{2|5} - 0.1468X_{2|6} \\
\qquad -1.4195X_{3|1} - 0.4407X_{3|2} - 0.2060X_{3|3} \\
\qquad -4.0646X_{4|1} - 3.9304X_{4|2} - 4.6403X_{4|3} \\
\qquad -1.1250X_{4|4} \\
L_2 \quad : \textbf{0.6586} + 0.4171X_{2|1} + 1.2400X_{2|2} + 0.3267X_{2|3} \\
\qquad +1.1248X_{2|4} + 0.5269X_{2|5} - 0.1468X_{2|6} \\
\qquad -1.4195X_{3|1} - 0.4407X_{3|2} - 0.2060X_{3|3} \\
\qquad -4.0646X_{4|1} - 3.9304X_{4|2} - 4.6403X_{4|3} \\
\qquad -1.1250X_{4|4}
\end{cases}
\tag{4.4}
$$

Model Evaluation & Assumption Validation

For every classic statistical method, cumulative logistic regression has some assumptions as well. In this section of the study, we have checked all the assumptions of the proposed model as a validation of the final model given in equation 4.4.

⊛ **Assumption #1:** Dependent variable should be measured at the ordinal level;

The severity level of the disease hemophilia is measured in descending order of the magnitude of clotting agent *F8* or Factor 8. So, our target variable (Severity Level) is ordered in categories.

⊛**Assumption #2:** One or more independent variables that are continuous, ordinal or categorical (including dichotomous variables);

This assumption is satisfied because, in our final model, we have used three statistically significant categorical variables (Mutation, Race, Domain).

⊛**Assumption #3:** There is no multicollinearity.

All the independent covariates are categorical, and we have tested the multicollinearity among those covariates using some multicollinearity measurements such as Kendall's $\tau - b$, Kendall's $\tau - c$, and Spearman Correlation given in the following Table 4.9 below.

Table 4.9: Checking Multicollinearity among the Categorical Covariates

	Covariates		
	Race vs. Domain	Domain vs. Mutation	Race vs. Mutation
Kendall's τ - b	-0.024	-0.335	0.05
Kendall's τ - c	-0.017	-0.309	0.034
Spearman	-0.045	-0.383	0.097

It turns out that none of the correlation measurements are statistically significant or alarming in the process of building the final model.

⊛**Assumption #4:** Proportional odds, which is a fundamental assumption of this type of ordinal regression model;

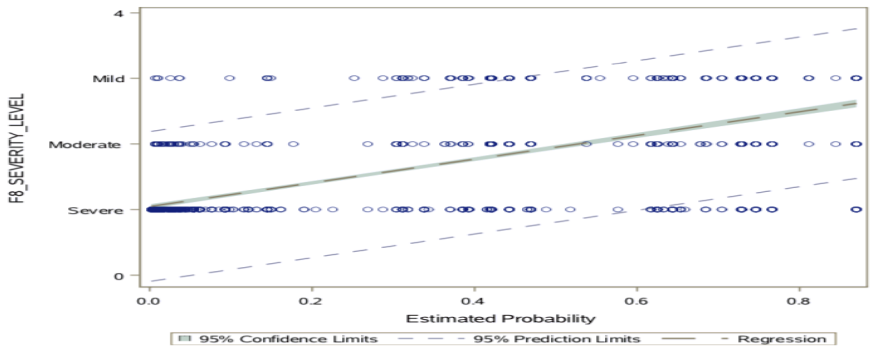


Figure 4.11: Proportional Odds Assumption for Hemophilia A

For the cumulative logistic regression model, proportional odds are the most important assumption to check. As a part of the validation of the model presented in equation

4.4, from Table 4.10, it clearly shows the insignificance of the proportional odds assumption of the Cumulative Logistics Model.

In other words, the coefficients that describe the relationship between Severe level of disease versus Mild and Moderate categories of the response variable (Severity Level) are the same as those coefficients that describe the relationship between Moderate and Mild category of response variable which is, in other words, called parallel regression assumption. From the figure above, it indicates that the proportional odds assumption is not violated and the diagonal parallel lines indicate the parallel regression assumptions as well. This is also confirmed by the chi-square test of the model given in the following Table 4.10.

Table 4.10: Testing for Proportional Odds Assumption

Chi-Square	DF	Pr >ChiSq
17.2278	13	0.1891

4.5 Results & Discussion

The model given in equation 4.4 is the reasonably best model. From this model, we can interpret that 1 unit change of Domain Protein (A2), we expect that there will be approximately 42% increase in the log odds being in the lower level of the severity level from Sever to Mild when all other covariates are held constant. So, there is always a positive increase in the log odds for one unit increase in each of the domain protein category. On the other hand, one unit increase or change from one category of Race to other categories of the Race will have a negative impact on the log odds of being from Mild to Moderate category of Severity level so is the third categorical variable, *Mutation*.

Now, we want to rank the attributable variables concerning their effects on the model. Table 4.6 shows the ranking of the attributable variables for their $p - value$ in the model. The first variable in the case of GLM is Mutation, and the last variable is Codon, and it is

also the same when the modeling is done with CLM also. However, things to be noted in the CLM method is that the ranking of statistically significant variables changes as the modeling technique changes. For example, the third variable in GLM from the top is Race of individuals, but in the CLM method, the third variable from the top is Domain (Location of Mutation Domain) at 5% level of significance.

Here, for instance, it is essential to discuss the proposed model given in equation 4.4 and model in equation 4.3. The following table shows the comparison of some vital information about the full model and the model that is built considering the statistically significant covariates only.

Table 4.11: Comparison of Models Accuracy

	Model	
	CLM (All Covariates)	CLM (Only sig. Covariates)
% Concordant	81.40	85.80
% Discordant	17.80	10.30
% Tied	0.80	3.90
Somers' D	0.63	0.75
C (ROC)	**0.81**	**0.87**

From the table above, we can see that the percentage of concordant pairs of CLM while considering all the covariates in the model is approximately 4% less than that of the CLM built with statistically significant covariates. Also, % of the area covered is measured by C statistic as an alternative for the ROC curve while Cumulative Logistic is used. CLM captures 6% higher areas estimated by the significant covariates only than that of the CLM with all covariates. After comparing the statistic given in Table 4.11, it is statistically conclusive that the model given in equation 4.4 is a better predictive model to predict the probability of the severity level of hemophilia A with *F*8 as per our dataset at hand. The predicted probabilities and actual probabilities are compared in Figure 4.12 below.

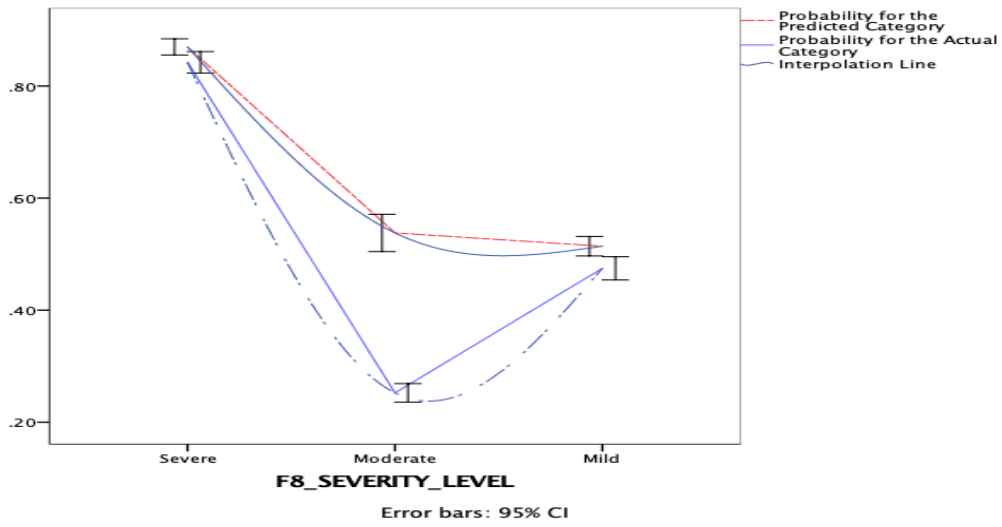


Figure 4.12: Predicted vs. Actual probabilities of CLR

The objectives of this study were to identify statistically significant variables that affect the outcome variable "Severity Level". Also, we wanted to see statistically significant categories of variables interacting with each other affecting the categories of response variable if any. At the same time, we wanted to rank the main effect variables that are contributing to the response and eventually come up with a model that is statistically significant and robust with a high degree of prediction probability and convergence.

So, we have identified statistically significant variables that affect the Severity Level by implementing a various methodology for model building, and we have compared them through some statistics. It turns out that the attributable variables **Race**, **Domain** and **Mutation type** are the most significant variables to predict the probability of the severity level of hemophilia A shown in Table 4.7. Also, we have investigated the interaction terms among the attributable variables, and it turns out that there were no interactions among variates as per out data concerns.

In terms of finding the best model driven by our data at hand, the *Cumulative Logistic Regression* considering the statistically significant covariates only as the independent variables has fast convergence rate and best results based on our data. This model is pre-

dicting the probability for each of the response category with about 87% accuracy under the *ROC* as per ***C*** statistic shown in Table 4.11.

In the context of the disease, the model given in **Equation 4.4**, has not violated the proportional odds assumption of the CLM. In brief, as per the model is given in equation 4.4, it indicates that proteins A1, A2, A3, B, C1 of Domain mutation will affect the probability of severity level of any individual in an increasing manner, i. e., if the doctors and scientists can identify these proteins in the blood then they have to provide some treatments that will locate these proteins and reduce their positive effects to increase patients probability of being in the Mild category (lowest category of Severity Level in Hemophilia A) from very severe category of the disease and it has to be opposite in case of C2 protein to improve or alleviate the severity level of any patient.

As per model in Equation 4.4, Race is affecting the probability of individuals being in one of the three categories of response variable, and there is no real-life treatment of to change Race of individuals we might conclude that being in the different categories of Severity level by race categories are totally in the hands of mother nature. However, it is conclusive that the majority of patients in the severe category comes from White Non-Hispanic rather than other categories of Race. On the other hand, the term $X_{4|3} = Mutation(Nonsense)$ has the smallest coefficient in the model mentioned above indicates that Nonsense type mutation change in the gene of individual will have maximum effect on the response outcome and to change/alleviate the severity level of any individual, it is suggested to attack/reverse this particular type of mutation cause and consequently change other types of mutations such as $X_{4|1} = Mutation(Frameshift), X_{4|2} = Mutation(Large_{Change}), X_{4|4} = Mutation(Small_{Change})$ in this order as per data at hand suggests.

4.6 Contribution

In this study, we have conducted uni-variate and multivariate analysis for some significant variables related to hemophilia A. During this process, we have achieved some insights listed as follows:

1. We found a cumulative logistic model and estimated its parameters.

2. This model can be used to predict the probability of any individuals severity level of hemophilia A with 87% accuracy.

3. Medical doctors and professionals will find useful for classifying any individual as one of the three levels of hemophilia based on information regarding individual's Domain (Location of mutation change), Mutation (Type of Protein change) and Race

4. Based on proper severity level, medical doctors, physician, and other medical organizations will be able to decide proper treatment program for that individual after having the genetic profile analyzed.

5 A Machine Learning Classification Model for Detecting Prediabetes

5.1 Introduction

Insulin is one of the many essential hormones produced by our pancreas that works as an accelerator to break the blood sugar and process those sugars (glucose) in such a way so that micro-cells in our body can absorb those to produce energy and heat in the human body. If the cells in the body do not normally respond to insulin, then this state of sugar insulation is termed as the **Prediabetes** [11]. Approximately 84 million American Adults - more than 1 out of 3 - have prediabetes. Among those with prediabetes, about 90% do not know they have this hormonal condition. If this stage goes untreated, then there is an increased risk of developing type -2 diabetes, heart disease, and stroke as per CDC [43]. In terms of preventing some risks involved with prediabetes condition, it is very significant and vital to detect the prevalence of prediabetes [53]. Also, the risk of cardiovascular disease and mortality is almost two times as high in individuals with a condition of prediabetes [16, 42]. Early detection, diagnosis, and intervention for prediabetes is a highly desired preventive measure that can be taken by anyone to avoid all the complications, prevent the transition of state for individuals from prediabetes to other type of diabetes (type - 2) and the model can be deployed to detect this condition with a very cost-effective way [54, 29].

In recent years, artificial intelligence research has been used to quantify almost all areas of human intervention with disease diagnosis and treatment selection. Machine learning is one of the broad areas of artificial intelligence that uses statistical methods for data classification and clustering. There are a handful of machine learning techniques

have been utilized and applied in the clinical domain to predict any disease condition and have implied higher accuracy for diagnosis rather than classical methods [60].

5.2 Methodology and Materials

5.2.1 Data Source

The National Center for Health Statistics (NCHS), Division of Health and Nutrition Examination Surveys (DHANES), part of the Centers for Disease Control and Prevention (CDC), has conducted a series of health and nutrition surveys since the early 1960s. The National Health and Nutrition Examination Surveys (NHANES) were conducted periodically from 1971 to 1994. In 1999, NHANES became continuous. Every year, approximately 5,000 individuals of all ages are interviewed in their homes and complete the health examination component of the survey [1].

5.2.2 Data Description

In this dataset, Risk for prediabetes is the response variable, and all other covariates are subdivided into different variable clusters based on the attributes of those variables such as Demographic, Diet Behavior, Weight, Height, Physical Activities, and Symptoms as shown in Figure 5.1. The NHANES sample represents the total non-institutionalized civilian US population residing in the 50 states and District of Columbia. As with previous NHANES samples, a four-stage sample design was used in NHANES 2011–2014. The first stage consisted of selecting PSUs from a frame of all US counties.

5.2.3 Risk Factors

At the beginning of variable inclusion in the machine learning algorithm, all the attributes under the sub-cluster of covariates are taken into the model feed, and sequentially variables are ranked in the final machine learning model according to their importance determined by the relative importance calculated using the actual model. In this

modeling, Age, Weight, Height, Poverty Ratio, and Blood pressures, are the continuous variables, and the rest of the attributes are categorical in variable measurements.

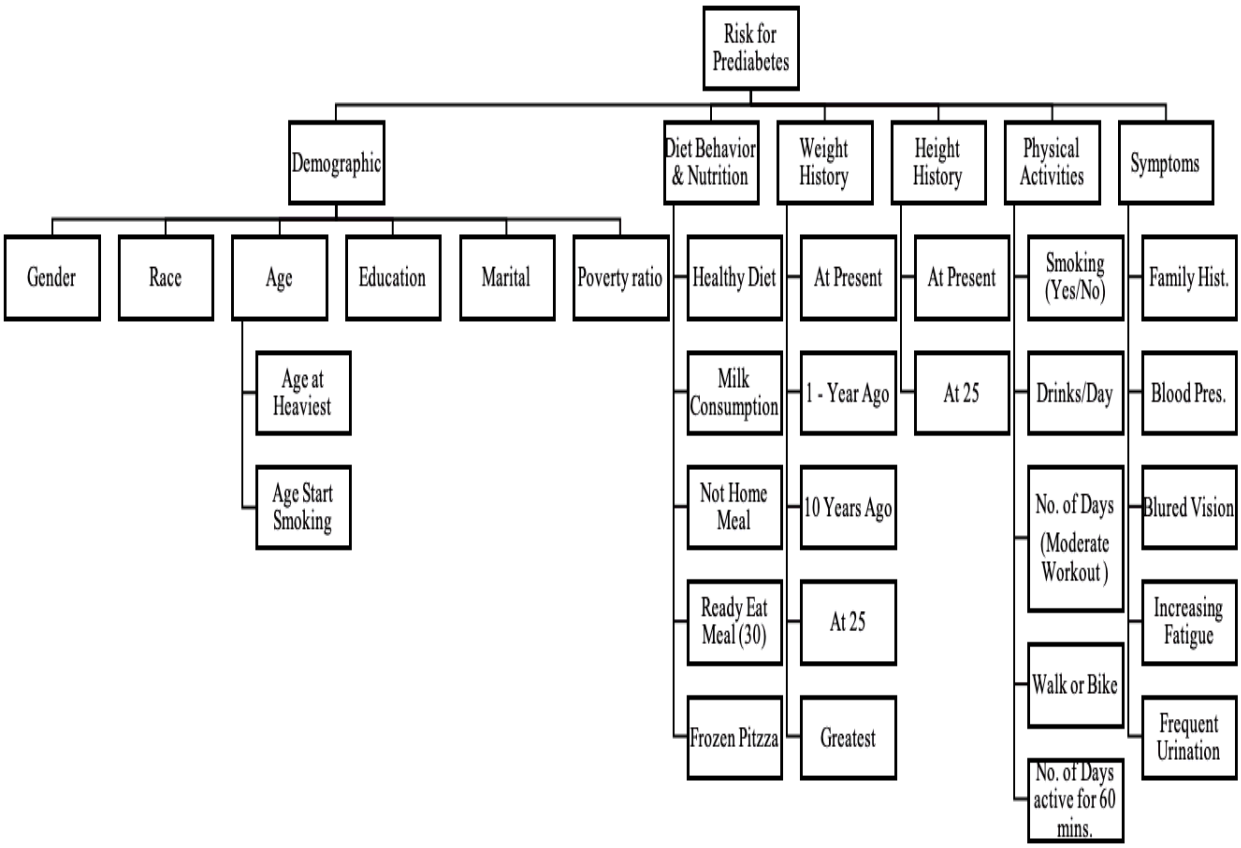


Figure 5.1: Schematic Diagram of Prediabetes Data

5.2.4 Machine Learning Modeling

In this study, we have used a supervised learning algorithm such as Decision Tree, Support Vector Machine (SVM), Gradient Boosting, Random Forest, Logistic Regression, and Neural Network. Also, for all the algorithms, all the observations was subdivided in 65% for **Training Set**, 25% for **Validation** and 10% for **Testing**. Then all the models were compared according to their *Average Squared Error* ***(ASE)***, *Captured Response Percentage* ***(CRP)*** and *Areas Under* ***(ROC)***. The champion model is selected with the lowest value of ASE and the highest value of CRP and areas under ROC. In the following sections, some machine learning algorithms are discussed very briefly.

5.2.4.1 Decision Tree

The decision tree algorithm falls under the category of supervised learning [39]. They can be used to solve both regression and classification problems. The decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label, and attributes are represented on the internal node of the tree. We can represent any Boolean function on discrete attributes using the decision tree. While using the decision tree, there are some basic assumptions are made as follows:

- at the beginning, the whole training dataset is considered as the root
- featured values are preferred to be categorical
- based on attributable values records are distributed recursively
- statistical methods are used for ordering attributes as root or internal node

5.2.4.2 Support Vector Machine (SVM)

More formally, a support-vector machine [46] constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, which can be used for classification, regression, or other tasks like outlier detection. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin), since in general the larger the margin, the lower the generalization error of the classifier.

5.2.4.3 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion as other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. [21]

5.2.4.4 Forest Model

A forest is an ensemble of decision trees [48, 21], each one able to predict its own response to a set of input variables. The results from the individual trees are combined to provide the final prediction. For a categorical target, the forest model's prediction is either the most popular class (as determined by a vote) or the average of the posterior probabilities of the individual trees. For an interval target, the forest model's prediction is the average of the estimates from the individual decision trees. The forest algorithm uses the following process to build each tree:

1. The algorithm selects a sample of cases, with replacement, from the original training data.
2. Then, for each node, the algorithm selects a sample of input variables from all available inputs.
3. From this sample, the input that has the strongest association with the target is used in the splitting rule for that node.

Therefore, the method of selecting the input variable for a splitting rule is different for a forest than it is for the split-search process used to build an individual tree. Each tree is created on a different sample of the cases, and each splitting rule is based on a different sample of the inputs. This process ensures that the individual models in the ensemble are more varied. The process that the forest algorithm uses to build the individual trees and then combine the results of the predictions is a more stable model than a single tree. Training each tree with different data reduces the correlation of the predictions of the trees. This, in turn, is likely to improve the predictions of the forest as compared to the naïve method of using the same data to build all the trees in a forest. The forest algorithm also takes random samples of the inputs. Therefore, the trees in the forest use different combinations of cases and inputs to determine the splits. This additional perturbation leads to greater diversity in the trees and often better predictive accuracy. Each sample of the original training data that is selected to train a specific decision tree is called bagged

data. For each tree in the forest, the data that are withheld from training are called an out-of-bag sample. Model assessment measures (such as misclassification rates and average squared error) and iteration plots are constructed on both the entire training data set and the out-of-bag sample.

5.2.4.5 Artificial Neural Network (ANN)

An artificial neural network (ANN)[27] is a network of simple elements called artificial neurons, which receive input, change their internal state (activation) according to that input, and produce output depending on the input and activation. An artificial neuron mimics the working of a biophysical neuron with inputs and outputs but is not a biological neuron model. The network forms by connecting the output of certain neurons to the input of other neurons forming a directed, weighted graph. The weights, as well as the functions that compute the activation, can be modified by a process called learning, which is governed by a learning rule.

5.2.5 Statistical Analyses

Ranking Variable Importance

In this study, we have taken 20 covariates into the initial consideration but considered 16 variables in total to build the machine learning model determined by the TREE SPLIT procedure [2, 18]. It measures variable importance based on the following metrics:

- Count-based variable importance counts the number of times in the tree that a particular variable is used in a split.
- Surrogate-count-based variable importance tallies the number of times that a variable is used in a surrogate splitting rule.
- RSS-based variable importance measures variable importance based on the change of RSS when a split is found at a node.

Figure 5.2 shows the ranking of relative importance by this process for those variables considered initially, and Age is the most important risk factor for predicting prediabetes found by this process.

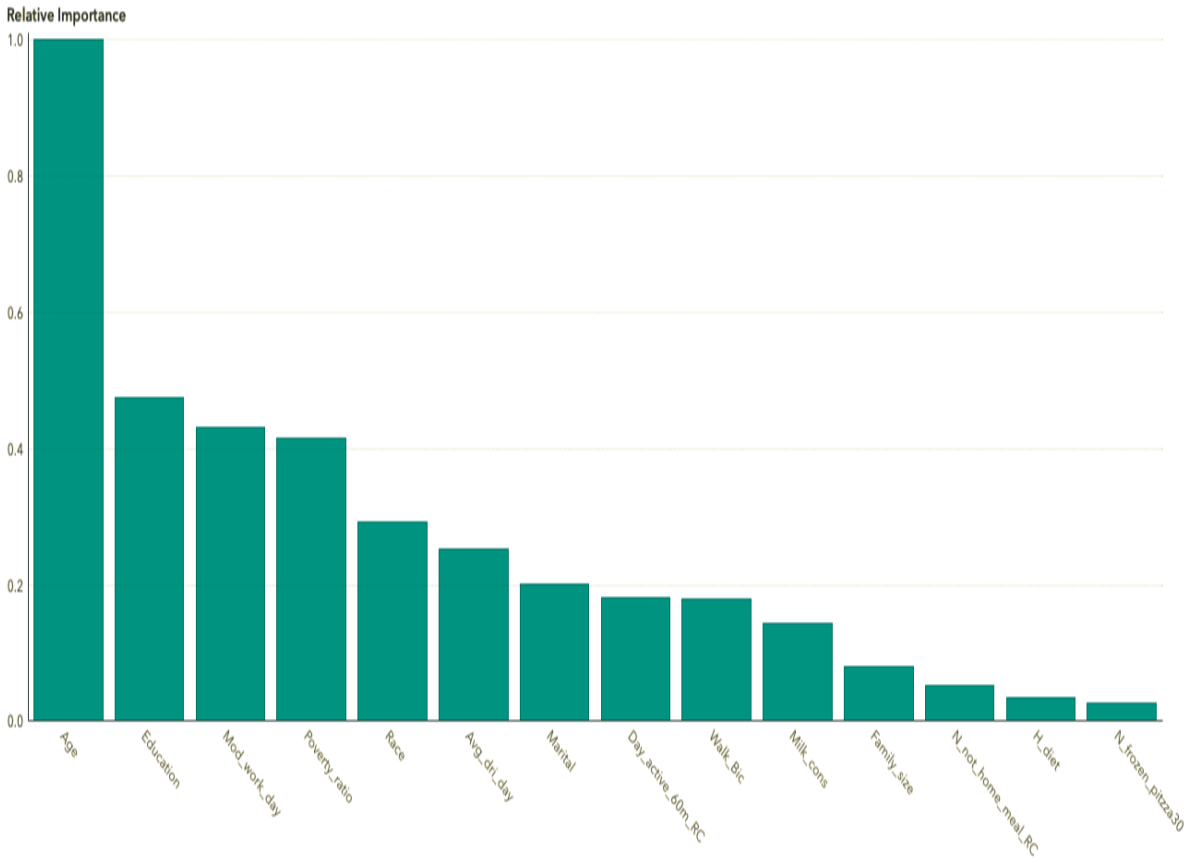


Figure 5.2: Ranking of Important Covariates

Missing Data Imputation

Because the data was collected from the survey and missing information is inherent characteristics for this dataset, we have used multiple imputation [45] method to impute missing information for each of the attributable variables considered in the model building process.

5.2.5.1 Variable Selection

In our analysis, we have considered those variables accepted by all the algorithms because each algorithm has its own selection criterion to be considered in the final analysis [25]. After running the input selection criterion for each of the algorithms we have ended up with selecting **16** inputs as shown in the **Table 5.1**.

Table 5.1: Variables Selected by all Algorithms

1	RISK DIAB	BINARY	TARGET	⊙
2	AGE	INTERVAL	INPUT	
3	AVG DRI DAY	INTERVAL	INPUT	
4	DAY ACTIVE 60M RC	NOMINAL	INPUT	
5	EDUCATION	NOMINAL	INPUT	
6	FAMILY SIZE	NOMINAL	INPUT	
7	GENDER	BINARY	INPUT	
8	GREATEST WEIGHT	INTERVAL	INPUT	
9	HEIGHT	INTERVAL	INPUT	
10	H DIET	NOMINAL	INPUT	
11	MARITAL	NOMINAL	INPUT	
12	MILK CONS	NOMINAL	INPUT	
13	MOD WORK DAY	NOMINAL	INPUT	
14	N FROZEN PIZZA30	INTERVAL	INPUT	
15	RACE	NOMINAL	INPUT	
16	SMOKING	NOMINAL	INPUT	
17	WEIGHT	INTERVAL	INPUT	
18	N NOT HOME MEAL	NOMINAL	REJECTED	Combination Criterion
19	N READY EAT30	NOMINAL	REJECTED	Combination Criterion
20	WALK BIC	NOMINAL	REJECTED	Combination Criterion

It turns out that, **No. of not a home meal, No. of ready to eat meal in the last 30 days, and Walking-biking** variables are rejected by all the algorithms.

After selecting appropriate covariates for the model building, we have tried most of the commonly known machine learning algorithms mentioned in the section **??** and at the end of the analysis, we have compared all the models to determine the champion model based on ASE (Average Squared Error) [57], CRP (Captured Response Percentage), and ROC [26]. Figure 5.3 shows the complete flow chart of the analysis.

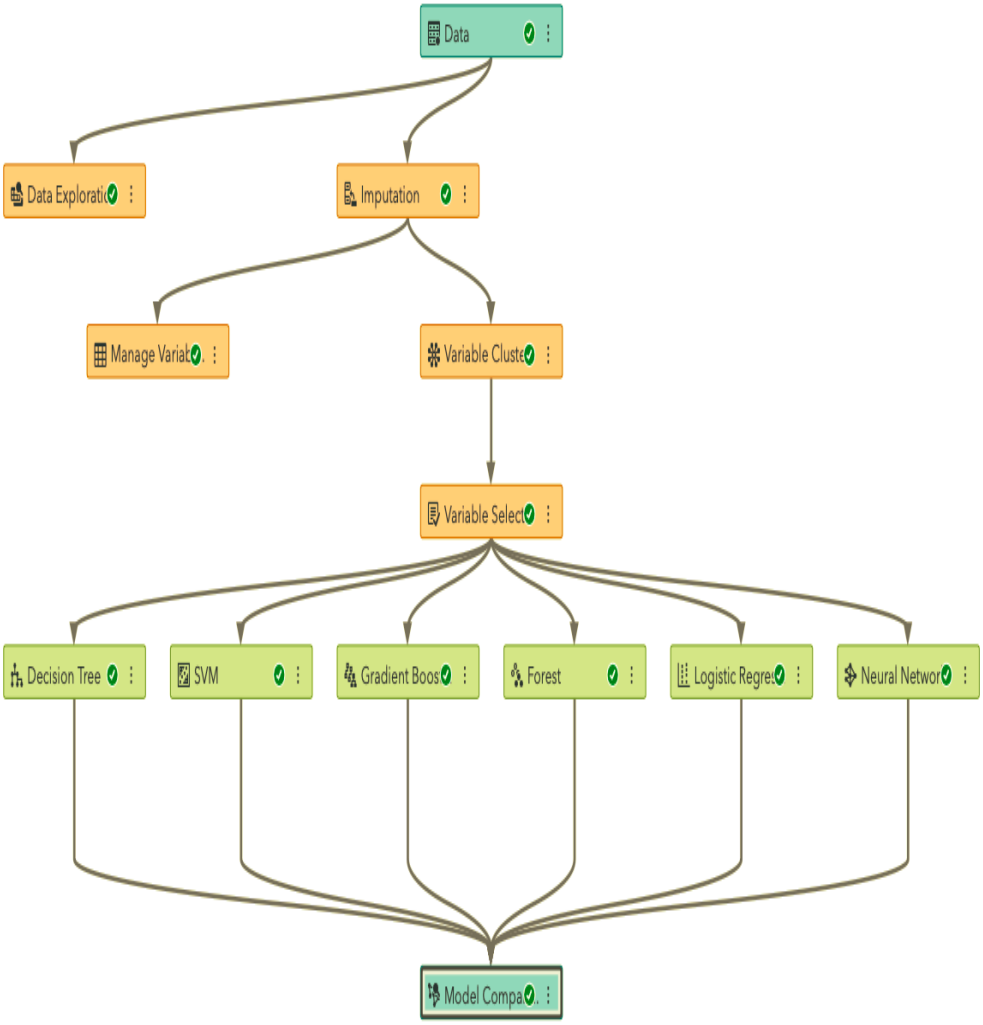


Figure 5.3: Flow chart of the Analysis

In the flow chart figure above, the whole process of data analysis and machine learning model building is postulated with respect to our analysis. In this workflow, we have

considered six most commonly known machine learning algorithms, among those, five are under supervised machine learning algorithms (Decision Tree, SVM, Gradient Boost, Forrest, Logistic Regression) and one is under supervised and unsupervised machine algorithm (Neural Network) both.

5.3 Proposed Champion Model

In the process of building a machine learning model, we have implemented six different types of algorithms and compared their results with each other to determine the best model. After considering the numerical values of ROC, ASE, Captured response percentage, and KS (Youden)[20]and it turns out the **Random Forest** model is the champion machine learning model for classifying the prediabetes patients. In the following Table 5.2, the comparative results are shown.

Table 5.2: Model Comparison For Prediabetes Data

Algorithm Name	ASE	KS (Youden)	ROC Area	CRP	Champion
Forest	0.115	0.1298	0.593	5.226	✠
Neural Network	0.249	0.0000	0.500	5.074	
SVM	0.192	0.0320	0.501	5.175	
Logistic Regression	0.117	0.0720	0.539	5.124	
Decision Tree	0.118	0.0730	0.552	5.256	
Gradient Boosting	0.117	0.0903	0.545	4.819	

From the above table, we see that the greatest KS (Youden) among all the models is for Forest about 0.1298 and areas under the ROC curve is 0.593 and this model has the smallest ASE (Average Squared Error) among all the model algorithm as per our analysis. So, to determine the champion model, we have selected the "Forest"to be the best model among machine learning algorithms.

5.4 Results & Discussion

Since the Forest algorithm is the champion one as a machine learning model, we will discuss the results of this model in detail. Forest model is the ensemble of Decision Tree, and the options we have used to build this model such that, the number of trees used was 50, during the Tree splitting options, the class target criterion used is the Entropy, maximum depth of the Tree was 12, minimum leaf size used was 15, the number of bins for the continuous variable was 100. After setting all the values in the algorithm, we have acquired the Forest model, and as per our champion model, we have ranked the attributable variables according to the relative importance in Figure 5.4 below.

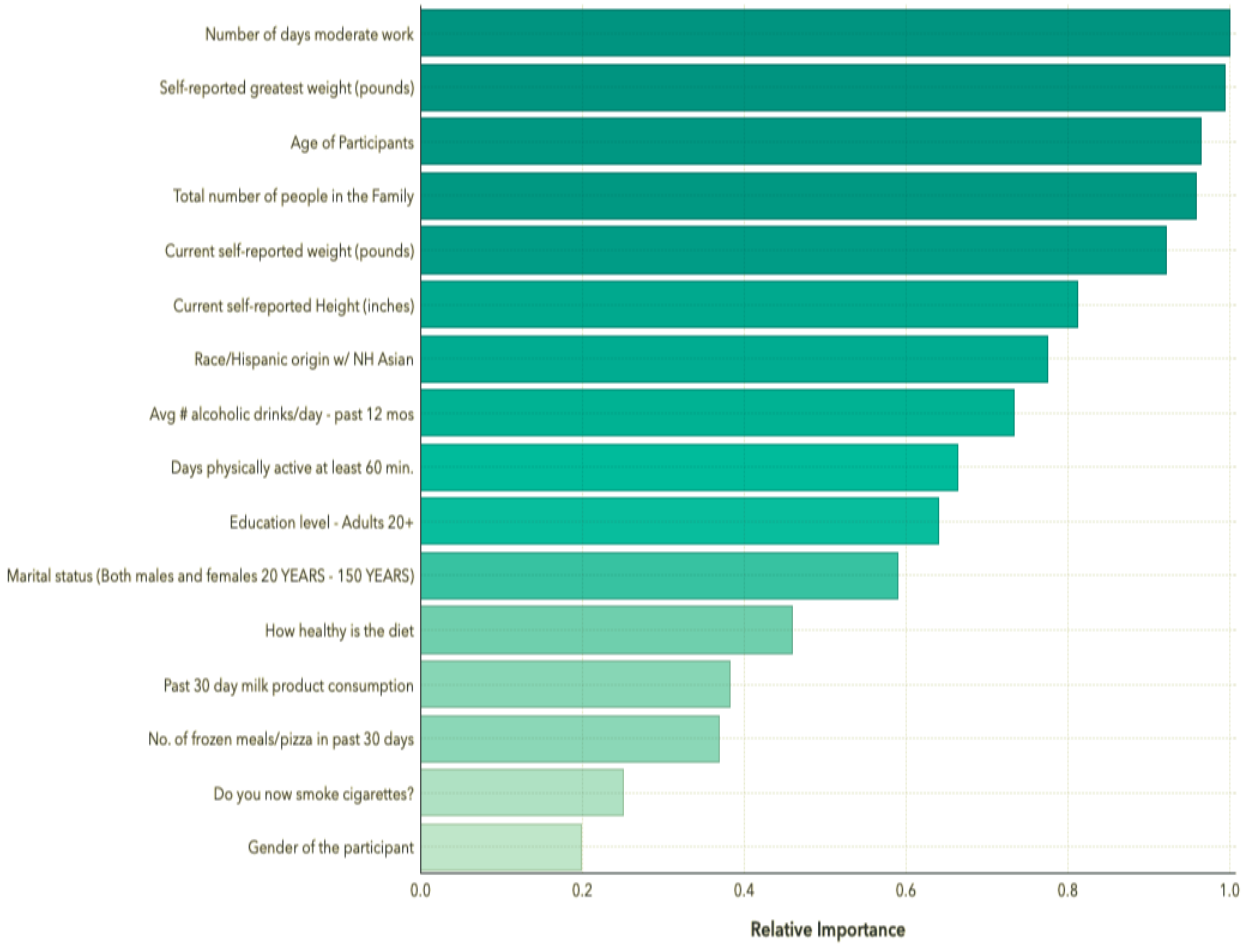


Figure 5.4: Ranking of Important Variables in the Champion Model

The five most important factors are *Number of days moderate work, Self-reported the greatest weight (pounds), Age of Participants, Total number of people in the Family, and Current self-reported weight (pounds).*

The next Figure 5.5 below shows the Average Squared Error for the proposed model. It is very important to see that the ASE decreases as the model trees grow larger, but after 20 trees it becomes flat and remains flat for all the training, validation and test data partitions through 50 trees that were the option used for the number of trees in the algorithm.

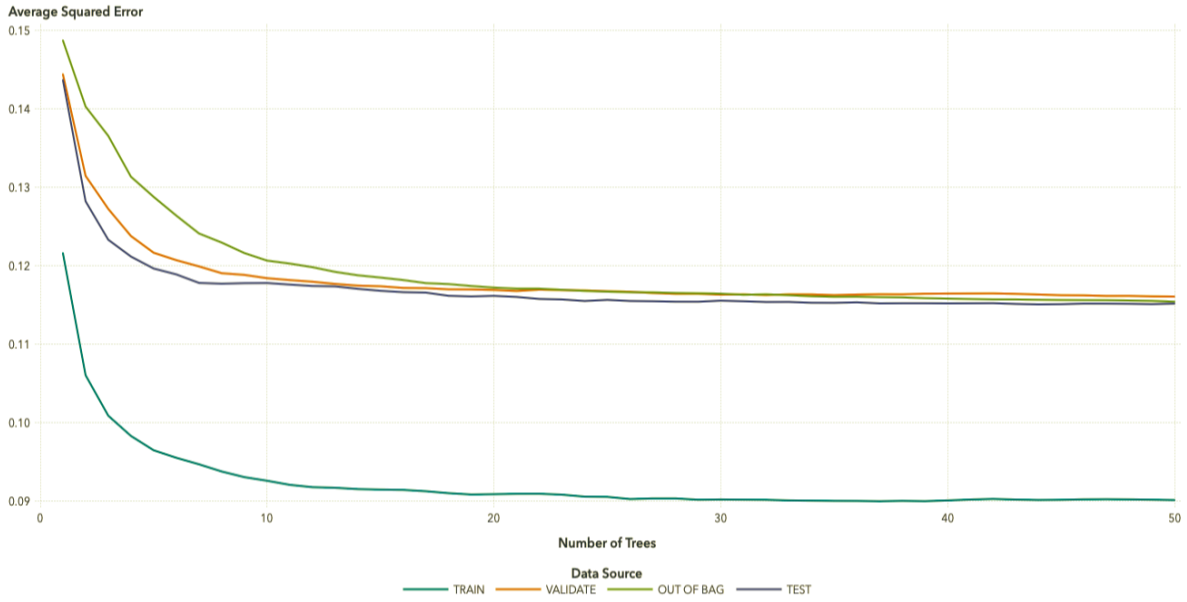


Figure 5.5: Avg. Sq. Error for Proposed model (Forest)

In the case of assessing the model through the ROC curve, it is a plot of sensitivity (the true positive rate) against 1-specificity (the false positive rate), which are both measures of classification based on the confusion matrix. These measures are calculated at various cutoff values. To help identify the best cutoff to use when scoring data, the KS Cutoff reference line is drawn at the value of 1-specificity where the greatest difference between sensitivity and 1-specificity is observed for the VALIDATE partition. The KS Cutoff line is drawn at the cutoff value 0.85, where the 1-specificity value is 0.574, and the sensitivity value is 0.687. Cutoff values range from 0 to 1, inclusive, in increments of 0.05. At each cutoff value, the predicted target classification is determined by whether the Risk of Pre-

diabetes, which is the predicted probability of the event "2"(category - NO) for the target Risk_diab, is greater than or equal to the cutoff value. When P_Risk_diab2 (category - NO) is greater than or equal to the cutoff value, then the predicted classification is the event; otherwise, it is a non-event.

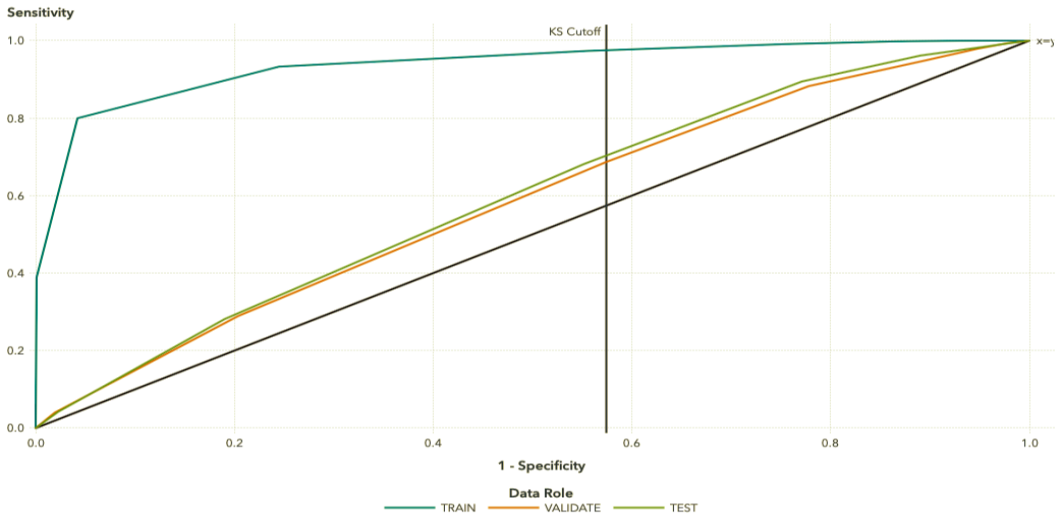


Figure 5.6: ROC for Proposed model (Forest)

The confusion matrix for each cutoff value contains four cells that display the true positives for events that are correctly classified (TP), false positives for non-events that are classified as events (FP), false negatives for events that are classified as non-events (FN), and true negatives for non-events that are classified as non-events (TN). True negatives include non-event classifications that specify a different non-event. Sensitivity is calculated as $TP/(TP + FN)$. Specificity, the true negative rate, is calculated as $TN/(TN + FP)$, so 1-specificity is $FP/(TN + FP)$. The values of sensitivity and 1-specificity are plotted at each cutoff value. A ROC curve that rapidly approaches the upper-left corner of the graph, where the difference between sensitivity and 1-specificity is the greatest, indicates a more accurate model. A diagonal line where sensitivity = 1-specificity indicates a random model.

Captured response percentage is calculated by sorting each partition in descending order by the predicted probability of the target event P_Risk_diab2, which represents the

predicted probability of the event "2"(category - NO) for the target Risk_diab. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed. Captured response percentage is the percentage of the total number of events that are in that quantile. With no model, it is expected that 5% of the events are in each quantile.

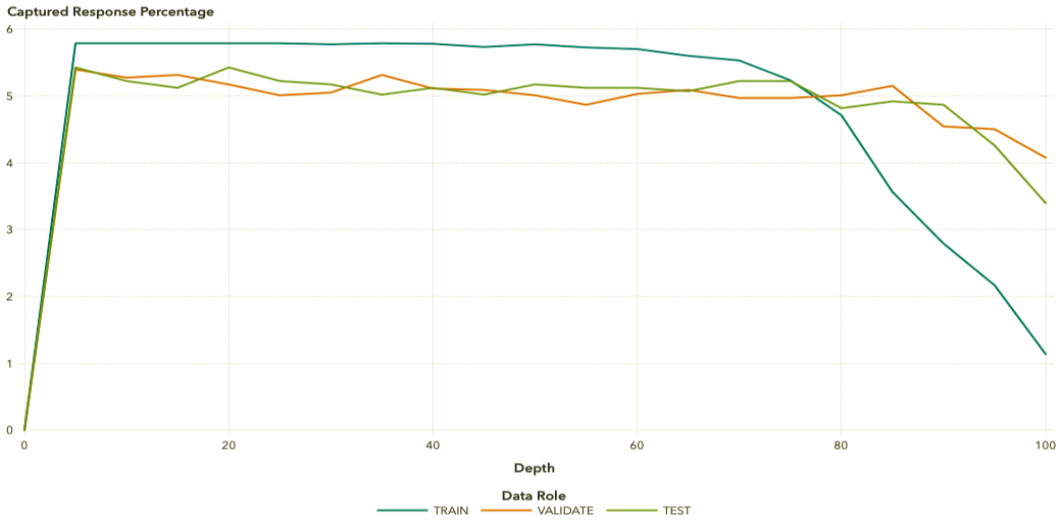


Figure 5.7: Captured Response Percentage for Proposed model (Forest)

At the 5% quantile (depth of 5), the VALIDATE partition has a Captured response percentage of 5.4 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 5.8. At the 5% quantile (depth of 5), the TRAIN partition has a Captured response percentage of 5.8 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 5.79. At the 5% quantile (depth of 5), the TEST partition has a Captured response percentage of 5.4 (compared to the expected value of 5 for no model). The best possible value of Captured response percentage for this partition at depth 5 is 5.83.

As part of model validation, we have used another plot called *Cumulative Lift*. The cumulative lift for a particular quantile is the ratio of the number of events across all quantiles up to and including the current quantile to the number of events that would

be there at random, or equivalently, the ratio of the cumulative response percentage to the baseline response percentage. The cumulative lift at depth 10 includes the top 10 percent of the data, which is the first two quantiles, which would have 10% of the events at random. Thus, cumulative lift measures how much more likely it is to observe an event in the quantiles than by selecting observations at random.

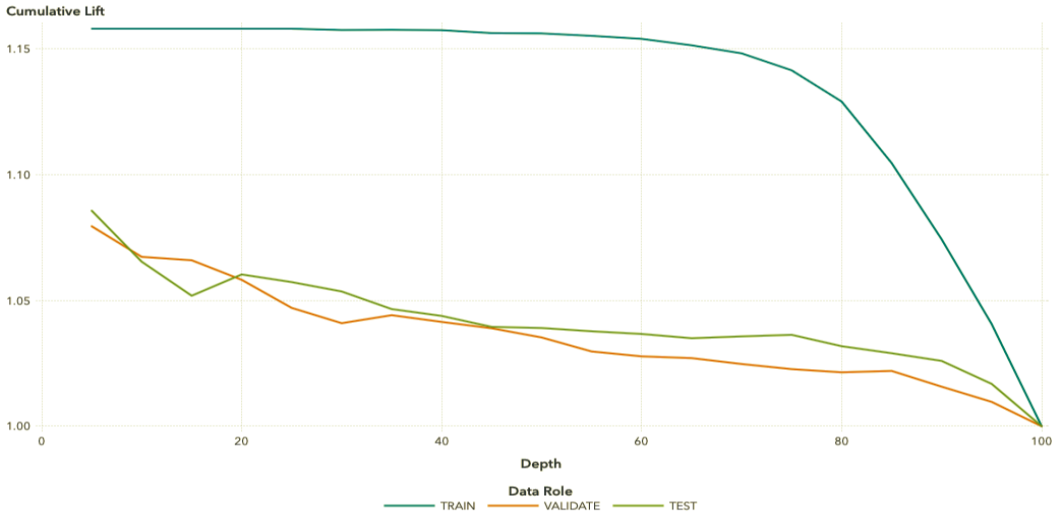


Figure 5.8: Cumulative Lift for Proposed model (Forest)

The VALIDATE partition has a Cumulative Lift of 1.07 in the 10% quantile (depth of 10) meaning there are about one times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition. The TRAIN partition has a Cumulative Lift of 1.16 in the 10% quantile (depth of 10) meaning there are about one times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is better to use your model to identify responders than no model, based on the selected partition. The TEST partition has a Cumulative Lift of 1.07 in the 10% quantile (depth of 10) meaning there are about one times more events in the first two quantiles than expected by random (10% of the total number of events). Because this value is greater than 1, it is

better to use your model to identify responders than no model, based on the selected partition.

Cumulative lift is calculated by sorting each partition in descending order by the predicted probability of the target event P_Risk_diab2, which represents the predicted probability of the event "2"(category - NO) for the target Risk_diab. The data is divided into 20 quantiles (demi-deciles, with 5% of the data in each), and the number of events in each quantile is computed.

In terms of model accuracy, we have plotted the accuracy plot for this model. Accuracy is the proportion of observations that are correctly classified as either an event or non-event, calculated at various cutoff values. Figure 5.9 below shows the accuracy of the proposed model. Cutoff values range from 0 to 1, inclusive, in increments of 0.05.

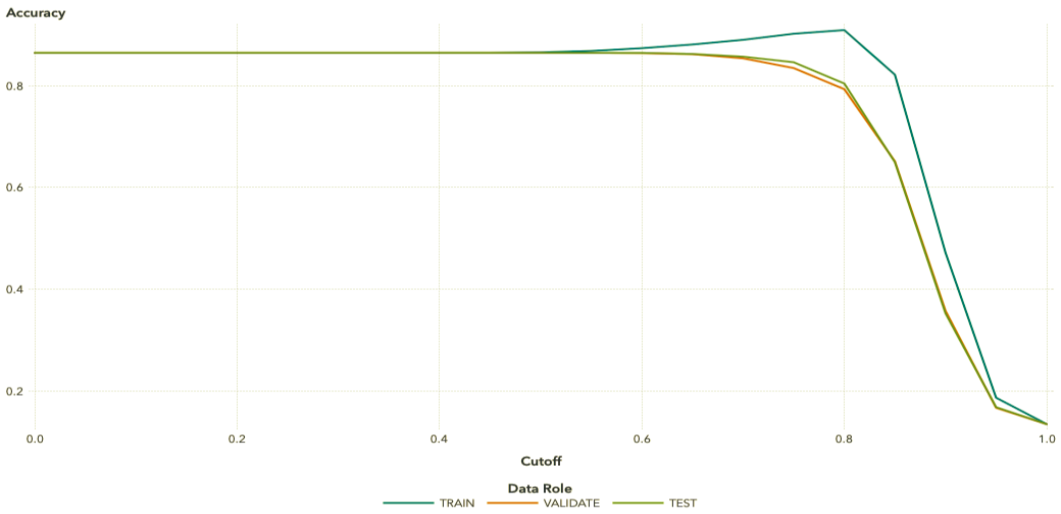


Figure 5.9: Accuracy Plot for Proposed model (Forest)

At each cutoff value, the predicted target classification is determined by whether P_Risk_diab2, which is the predicted probability of the event "2"(category - No) for the target Risk_diab, is greater than or equal to the cutoff value. When P_Risk_diab2 is greater than or equal to the cutoff value, then the predicted classification is the event, otherwise it is a non-event. When the predicted classification and the actual classification are both events (true positives) or both non-events (true negatives), the observation

is correctly classified. If the predicted classification and actual classification disagree, then the observation is incorrectly classified. Accuracy is calculated as $(\text{true positives} + \text{true negatives})/(\text{total observations})$.

From the results of the present study, it indicates that the Forest model that we have developed performed better than any other existing model used for classifying and predict the prevalence of Prediabetes condition among the US population. Although, ANN could be used as the best model for this study purpose which also suggested by some machine learning model for screening individuals for prediabetes developed by Choi et al. [13] where six risk factors were used to build the model from the Korean population on prediabetes, but considering 16 risk factors included in our study produced Forest is the best machine learning algorithm. On the other hand, Meng et al. [38] did comparative study among the performances of logistic regression, ANNs, and decision tree models for predicting diabetes as well as prediabetes in Chine population using common risk factors. Regarding Meng et al. study, the ANNs model was the least suggested model with the most inferior performance in terms of accuracy. These indicate that our model is consistent with their machine learning model. Also, if any clinical study or research wants to use the model to classify the individual prediabetes state with more risk factors involved, then our proposed model will perform the best at a higher accuracy.

Although there are some common risk factors (covariates/attributes) that were included in our model and other machine learning models developed for prediction of prediabetes condition [13, 38, 33], they have considered less number of risk factors (covariates) than our model. In the case of a large number of risk factors included in the model, the Forest model will perform relatively better.

5.5 Contribution

In this present study, we have constructed a reasonably better classification algorithm for prediabetes in the USA population. Also, we have achieved crucial insights that will be very useful in terms of practical relevancy of this study.

1. In case of interested countries, they can implement a similar type of methodologies to classify the individuals with prediabetes condition.

2. This classification algorithm can be used by government agencies, scholars, and researchers to develop region or state-specific machine learning models.

3. The development of such a model can be deployed as a web application with a user-friendly calculator program. This will enable the access of mass people, including the medical scholars and professionals.

4. Early diagnosis or preventive measures with correctly identified prediabetes state will impact the public health issue on this subject matter.

5. This type of classification technique will help to reduce the incidence of other health issues related to prediabetes conditions such as heart disease, stroke, and obesity among early diagnosed and undiagnosed portion of the population.

Conclusion & Contributions

In this dissertation, we have found the probability distribution function, PDF, of the **Democracy Index Scores**, DIS, that have been documented by the Economist Intelligence Unit, EIU. Having identified the PDF of the subject data, we can characterize the probabilistic behavior of different types of Democracy of different countries of the world. The EIU collected information of 167 countries in the world and descriptively classified each country as (1). Full Democracy, (2). Flawed Democracy, (3). Hybrid Regime, and (4). Authoritarian Regime. Thus, we have characterized the probabilistic behavior of all the DIS scores or the DIS for each of the four categories of Democracy around the globe and obtain other useful information.

As a continued study of the DIS scores from EIU, we have formulated a model with a very high R^2 & this is consistent with $adjusted - R^2$ because this eliminates the biasness of the interaction introduced in the model due to human interactions in the subject area. The final evaluation of the model is, during the process of developing the model, we left 30% of the observation out of the model building from each stratum that is made after the cluster analysis by the $k - means$ and *Multinomial Logistic Regression* analysis chapter. We have used the developed model to classify those democracy index scores.

The practical usefulness of the proposed model would be, to utilize to predict Democracy Index Scores of any new country included in the model with a certain degree of assurance. A misclassification of a country in determining its democracy scores could be detrimental to that country. Because, if WB (World Bank) decides not to give out grants to the non-democratic country that is mistakenly classified as "Hybrid Democratic"country, but in reality, it should be classified into the "Flawed Democratic"criterion and so on.

In the present study, we have identified the probability distribution function, PDF, of the **Corruption Perception Index**, CPI, that have been documented by the Transparency International, TI, and WGI of WB. Having identified the PDF of overall CPI, we characterized the probabilistic behavior of Corruption categories for 175 countries of the world. The TI collected information of 175 countries, and we have classified each of the countries into one of **four** different categories as (1) Least corrupted, (2) Fair Corrupted, (3) Moderately Corrupted, and (4) Highly Corrupted Countries. Then, we proceeded to find the PDF and CPDF of each of the four categories.

Furthermore, while we were studying F8 mutation on Hemophilia A, we have found that race and inhibitor history are independent of each other. We also found that the severity level of hemophilia A is statistically related to the races of the individuals of the US population to some degree based on our analysis. However, we have found that the severity level is highly dependent on the History of Inhibitors for the US population. Also, from the local and cumulative odds ratios indicates that the Odds of Whites being in the Mild and Severe level of the disease are significantly higher than those of the African Americans and other race categories.

Finally, the **Forest** model that we have identified while studying prediabetes risk performed better than any other existing model used for classifying and predicting the prevalence of Prediabetes condition among the US population. Also, ANN (Artificial Neural Network) and SVM (Support Vector Machine), our second and third best models, respectively. They could be used as one of the champion models for predicting prediabetes suggested by some researchers who studied the machine learning models for screening individuals on prediabetes condition, developed by Choi et al.[13] where six risk factors were used to build the model from the Korean population on prediabetes. However, considering 16 risk factors included in our study determined that **Forest** is the best machine learning algorithm. On the other hand, Meng et al. [38] did a comparative study among

the performances of logistic regression, ANNs, and decision tree models for predicting diabetes as well as prediabetes in the Chinese population using common risk factors.

Regarding Meng et al. study, the ANNs model was the least suggested model with the weakest performance in terms of accuracy. These indicate that our model is consistent with their machine learning model. Although there are some common risk factors (covariates/attributes) that were included in our model and other machine learning models developed for prediction of prediabetes condition [13, 38, 33], they have considered few risk factors. In the case of a finitely large number of risk factors included in the model, the **Forest** model will perform relatively better.

References

[1] Nhanes 2015-2016 questionnaire data. https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2015. National Health and Nutrition Examination Survey.

[2] The treesplit procedure - variable importance. https://documentation.sas.com/?docsetId=casml&docsetTarget=viyaml_treesplit_details20.htm&docsetVersion=3.0&locale=en. SAS Viya Data Mining and Machine Learning: Procedures Guide.

[3] Robert D Abbott. Logistic regression in survival analysis. *American Journal of Epidemiology*, 121(3):465–471, 1985.

[4] Cande V Ananth and David G Kleinbaum. Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333, 1997.

[5] Theodore W Anderson and Donald A Darling. A test of goodness of fit. *Journal of the American statistical association*, 49(268):765–769, 1954.

[6] Ralf Bender and Ulrich Grouven. Ordinal logistic regression in medical research. *Journal of the Royal College of physicians of London*, 31(5):546–551, 1997.

[7] Dankmar Böhning. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 44(1):197–200, 1992.

[8] Norman E Breslow, Nicholas E Day, et al. *Statistical methods in cancer research. Vol. 1. The analysis of case-control studies.*, volume 1. Distributed for IARC by WHO, Geneva, Switzerland, 1980.

[9] Beverly Britton. Myths & facts...about hemophilia. *Nursing*, 33(12):78, Dec 2003.

[10] Marc Bühlmann, Wolfgang Merkel, and Bernhard Wessels. The quality of democracy: democracy barometer for established democracies. 2008.

[11] Martin Buysschaert and Michael Bergman. Definition of prediabetes. *Medical Clinics*, 95(2):289–297, 2011.

[12] Herman Chernoff and EL Lehmann. The use of maximum likelihood estimates in $\chi 2$ tests for goodness of fit. *The Annals of Mathematical Statistics*, pages 579–586, 1954.

[13] Soo Beom Choi, Won Jae Kim, Tae Keun Yoo, Jee Soo Park, Jai Won Chung, Yong-ho Lee, Eun Seok Kang, and Deok Won Kim. Screening for prediabetes using machine learning models. *Computational and mathematical methods in medicine*, 2014, 2014.

[14] Michael Coppedge, John Gerring, David Altman, Michael Bernhard, Steven Fish, Allen Hicken, Matthew Kroenig, Staffan I Lindberg, Kelly McMann, Pamela Paxton, et al. Conceptualizing and measuring democracy: A new approach. *Perspectives on Politics*, 9(2):247–267, 2011.

[15] A Coppola, M Margaglione, E Santagostino, A Rocino, E Grandone, PM Mannucci, G Di Minno, and AICE PROFIT STUDY GROUP. Factor viii gene (f8) mutations as predictors of outcome in immune tolerance induction of hemophilia a patients with high-responding inhibitors. *Journal of Thrombosis and Haemostasis*, 7(11):1809–1815, 2009.

[16] Mario Coutinho, Hertzel C Gerstein, Yong Wang, and Salim Yusuf. The relationship between glucose and incident cardiovascular events. a metaregression analysis of published data from 20 studies of 95,783 individuals followed for 12.4 years. *Diabetes care*, 22(2):233–240, 1999.

[17] Brian S Everitt. Mixture distributions—i. *Encyclopedia of statistical sciences*, 7, 2004.

[18] George Fernandez. *Statistical data mining using SAS applications*. CRC press, 2010.

[19] Jack M Finkelstein and Ray E Schafer. Improved goodness-of-fit tests. *Biometrika*, 58(3):641–645, 1971.

[20] Ronen Fluss, David Faraggi, and Benjamin Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(4):458–472, 2005.

[21] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[22] Michael Friendly. Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association*, 89(425):190–200, 1994.

[23] AC Goodeve, PH Reitsma, and JH McVey. Nomenclature of genetic variants in hemostasis. *Journal of Thrombosis and Haemostasis*, 9(4):852–855, 2011.

[24] Samantha C Gouw, H Marijke van den Berg, Johannes Oldenburg, Jan Astermark, Philip G de Groot, Maurizio Margaglione, Arthur R Thompson, Waander van Heerde, Jorien Boekhorst, Connie H Miller, et al. F8 gene mutation type and inhibitor development in patients with severe hemophilia a: systematic review and meta-analysis. *Blood*, 119(12):2922–2934, 2012.

[25] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

[26] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[27] Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (10):993–1001, 1990.

[28] James W Hardin and Joseph M Hilbe. *Generalized estimating equations*. Chapman and Hall/CRC, 2002.

[29] William H Herman, Thomas J Hoerger, Michael Brandle, Katherine Hicks, Stephen Sorensen, Ping Zhang, Richard F Hamman, Ronald T Ackermann, Michael M Engelgau, and Robert E Ratner. The cost-effectiveness of lifestyle modification or metformin in preventing type 2 diabetes in adults with impaired glucose tolerance. *Annals of internal medicine*, 142(5):323–332, 2005.

[30] ND Hicks and WR Pitney. A rapid screening test for disorders of thromboplastin generation. *British journal of haematology*, 3(2):227–237, 1957.

[31] Christine Kalenborn and Christian Lessmann. The impact of democracy and press freedom on corruption: Conditionality matters. *Journal of Policy Modeling*, 35(6):857–886, 2013.

[32] Laza Kekic. The economist intelligence unit's index of democracy. *The Economist*, 21:1–11, 2007.

[33] Yong-ho Lee, Heejung Bang, Hyeon Chang Kim, Hee Man Kim, Seok Won Park, and Dae Jung Kim. A simple screening score for diabetes for the korean population: development, validation, and comparison with other scores. *Diabetes care*, 35(8):1723–1730, 2012.

[34] Kung-Yee Liang and Scott L Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

[35] Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

[36] Daniel L McFadden. Econometric analysis of qualitative response models. *Handbook of econometrics*, 2:1395–1457, 1984.

[37] Richard D McKelvey and William Zavoina. A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*, 4(1):103–120, 1975.

[38] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. *The Kaohsiung journal of medical sciences*, 29(2):93–99, 2013.

[39] Tom M Mitchell and Machine Learning. Mcgraw-hill science. *Engineering/Math*, 1:27, 1997.

[40] Ann A O'Connell. *Logistic regression models for ordinal response variables*. Number 146. Sage, 2006.

[41] Amanda B Payne, Connie H Miller, Fiona M Kelly, J Michael Soucie, and W Craig Hooper. The cdc hemophilia a mutation project (champ) mutation list: a new online resource. *Human mutation*, 34(2):E2382–E2392, 2013.

[42] Sidney C Port, Mark O Goodarzi, Noel G Boyle, and Robert I Jennrich. Blood glucose: a strong risk factor for mortality in nondiabetic patients with cardiovascular disease. *American Heart Journal*, 150(2):209–214, 2005.

[43] What Causes Prediabetes? Centers for disease control and prevention. *Retrieved from https://www.cdc.gov/diabetes/basics/prediabetes.html*, July, 2019.

[44] M. S. Rahman. Statistical analysis of democracy index. *Humanomics*, 30:373–384, 11 2014.

[45] Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

[46] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[47] John Schwarz, Jan Astermark, Erika D Menius, Mary Carrington, Sharyne M Donfield, Edward D Gomperts, George W Nelson, Johannes Oldenburg, Anna Pavlova, Amy D Shapiro, et al. F8 haplotype and inhibitor risk: results from the hemophilia inhibitor genetics study (higs) combined cohort. *Haemophilia*, 19(1):113–118, 2013.

[48] Mark R Segal. Machine learning benchmarks and random forest regression. 2004.

[49] Ying So and Warren F Kuhfeld. Multinomial logit models. In *SUGI 20 Conference Proceedings*, pages 1227–1234, 1995.

[50] J M Soucie, J Symons, 4th, B Evatt, D Brettler, H Huszti, J Linden, and Hemophilia Surveillance System Project Investigators. Home-based factor infusion therapy and hospitalization for bleeding complications among males with haemophilia. *Haemophilia*, 7(2):198–206, Mar 2001.

[51] Michael A Stephens. Edf statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, 69(347):730–737, 1974.

[52] Thérèse A Stukel. Generalized logistic models. *Journal of the American Statistical Association*, 83(402):426–431, 1988.

[53] Adam G Tabák, Christian Herder, Wolfgang Rathmann, Eric J Brunner, and Mika Kivimäki. Prediabetes: a high-risk state for diabetes development. *The Lancet*, 379(9833):2279–2290, 2012.

[54] Li Tian, Jun Zhu, Lisheng Liu, Yan Liang, Jiandong Li, and Yanmin Yang. Prediabetes and short-term outcomes in nondiabetic patients after acute st-elevation myocardial infarction. *Cardiology*, 127(1):55–61, 2014.

[55] Chris P Tsokos. Probability distributions: An introduction to probability theory with applications. 1972.

[56] Jay Ulfelder and Michael Lustik. Modelling transitions to and from democracy. *Democratisation*, 14(3):351–387, 2007.

[57] D Wallach and B Goffinet. Mean squared error of prediction as a criterion for evaluating and comparing system models. *Ecological modelling*, 44(3-4):299–306, 1989.

[58] RD Wooten, K Baah, and Joy D'Andrea. Implicit regression: Detecting constants and inverse relationships with bivariate random error. *arXiv preprint arXiv:1512.05307*, 2015.

[59] Rebecca D Wooten. Introduction to implicit regression. *arXiv preprint arXiv:1602.00158*, 2016.

[60] Tae Keun Yoo, Sung Kean Kim, Deok Won Kim, Joon Yul Choi, Wan Hyung Lee, and Eun-Cheol Park. Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. *Yonsei medical journal*, 54(6):1321–1330, 2013.

[61] Scott L Zeger and Kung-Yee Liang. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, pages 121–130, 1986.

[62] Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pages 1049–1060, 1988.