Department of Statistics, SPPU

ST-O13 Statistical Learning and Data Mining

ETE Assignment 2023-24

Name: 1) Prajakta Pandurang Madane - 2231

2)Devendra Sanjay Patil - 2234

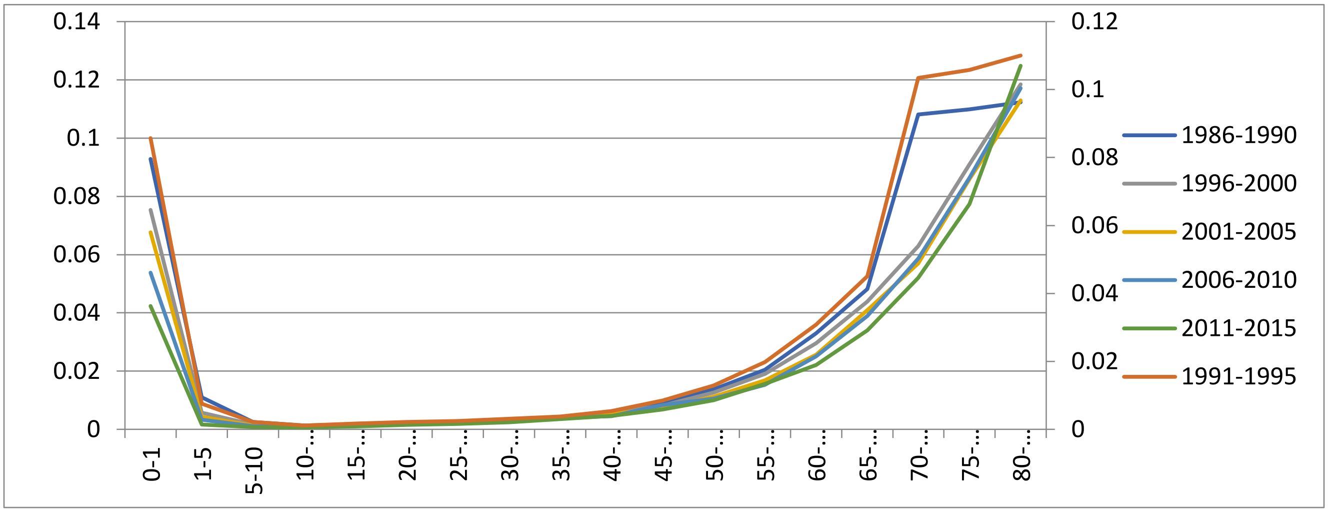
Title: Predictive Modeling and Comparative Analysis of Mortality Rates Across Age Groups and Genders in India

Data Description:

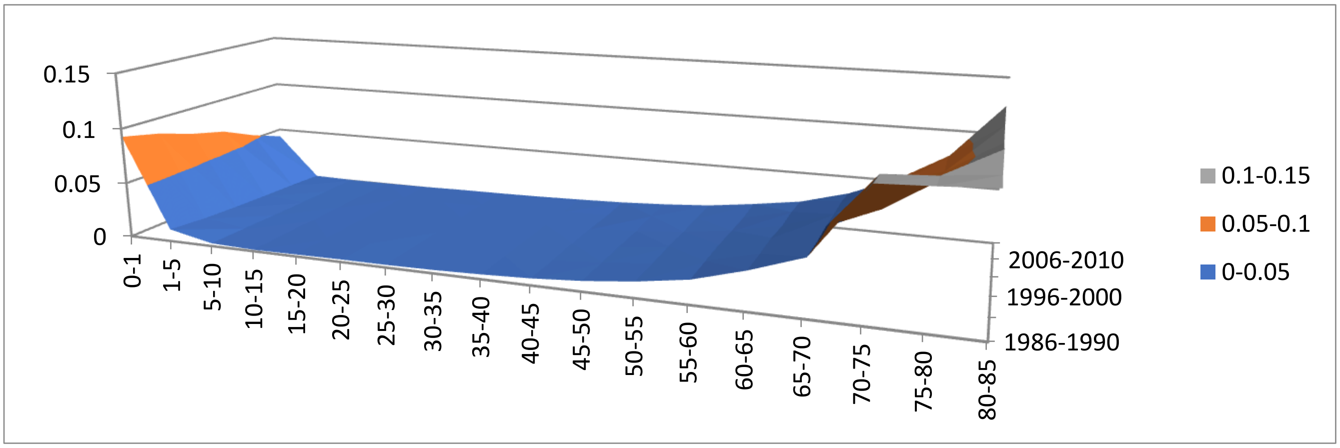
This project revolves around the analysis of a unique dataset titled '**Analyzing Mortality Trends in India**'. This dataset provides a comprehensive view of mortality rates across different age groups for both genders within the Indian population. The data spans over various time points, ranging from the period of 1986-1990 to 2011-2015.

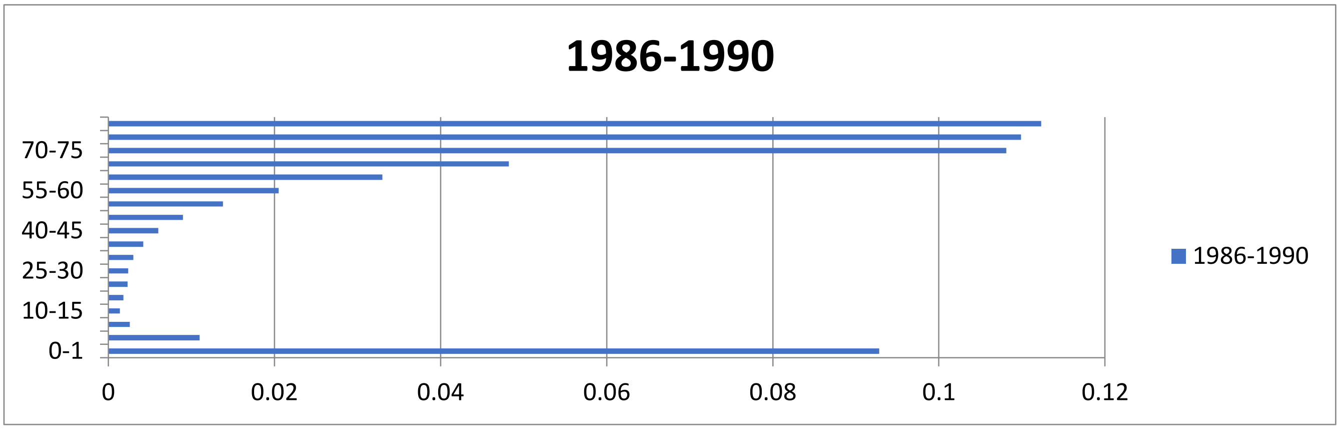
The dataset is structured in a way that the first column represents the age group, while the subsequent columns provide the corresponding mortality rates for the respective time periods. It's important to note that the data for males and females are provided in separate datasets. This segregation allows for a more nuanced understanding of gender-specific mortality trends over time.

Data Visualization:



The line graph, with the y-axis representing mortality rates and the x-axis denoting age groups, reveals a decrease in mortality rates from 1986-1990 to 2011-2015. This downward trend in mortality rates over the years signifies advancements in medical science. 3D Plot of above lines shows below:





The above graph shows mortality rate is high in starting and ending of life.

1.Predictive Modeling of Mortality Rates:

This section will focus on identifying the models that provide the best predictions for mortality rates based on previous data, specifying the regressors and response for each model.

For building a predictive model 1st we have to modified the given data, given data which is shown below that how modified data and original data looks,

1st four values of Original Data

AGE GROUP	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010	2011-2015
0-1	0.0928	0.0857	0.0754	0.0676	0.0538	0.0424
1-5	0.011	0.0076	0.0057	0.0043	0.0033	0.0017
5-10	0.0026	0.0022	0.0017	0.0014	0.0011	0.0008
10-15	0.0014	0.0012	0.0011	0.001	0.0008	0.0007

1st five values of Modified Data

AGE GROUP	Mortality_rate	Gender	Year
0-1	0.0928	M	**1986-1990**
1-5	0.011	M	**1986-1990**
5-10	0.0026	M	**1986-1990**
10-15	0.0014	M	**1986-1990**

In the modified data “Mortality_rate” is **Response** and “AGE GROUP”, “Gender” and “Year” are the **Regressors.**

In predictive modeling, several techniques can be used, including Decision Trees, Bagging, Random Forests, and Boosting **Decision Trees** are flowchart-like structures that help in decision making, where each node represents a test on an attribute, and each leaf node holds a class label **Bagging** is an ensemble method that involves creating multiple subsets of the original data, building a decision tree for each, and averaging the results to get a final prediction. A **Random Forest** is a type of ensemble machine learning algorithm that extends bagging to create a more robust model by introducing randomness into the tree construction. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. **Boosting**, on the other hand, is a sequential process, where each subsequent model attempts to correct the errors of the previous model. The final model (strong learner) is the weighted mean of all the models (weak learners). These methods, when used appropriately, can significantly improve the predictive accuracy and robustness of your model

Methods	MSE_TrainData	MSE_TestData
DecisionTree	7.084686e-05	0.0001211611
RandomForest	4.523294e-04	0.0004629950
Bagging	1.030998e-04	0.0001043220
Boosting	9.150363e-05	0.0001558567

The MSE is a measure of the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual value. The lower the MSE, the better the model's performance.

From the given data, we can observe that:

- The **Decision Tree** method has the lowest MSE for the training data, which is approximately

  7.084686×10−5

- The **Bagging** method has the lowest MSE for the test data, which is approximately

  0.0001043220

This suggests that the Decision Tree model fits the training data best, while the Bagging model generalizes better to unseen data.

It's also mentioned that k-fold cross-validation was used with a fold of 3 (i.e., k=3) for each method. K-fold cross-validation is a resampling procedure used to evaluate machine learning models on a limited data sample. It generally results in a less biased model compared to other methods. Because the algorithm is trained and evaluated multiple times on different data, the performance estimate is more reliable.

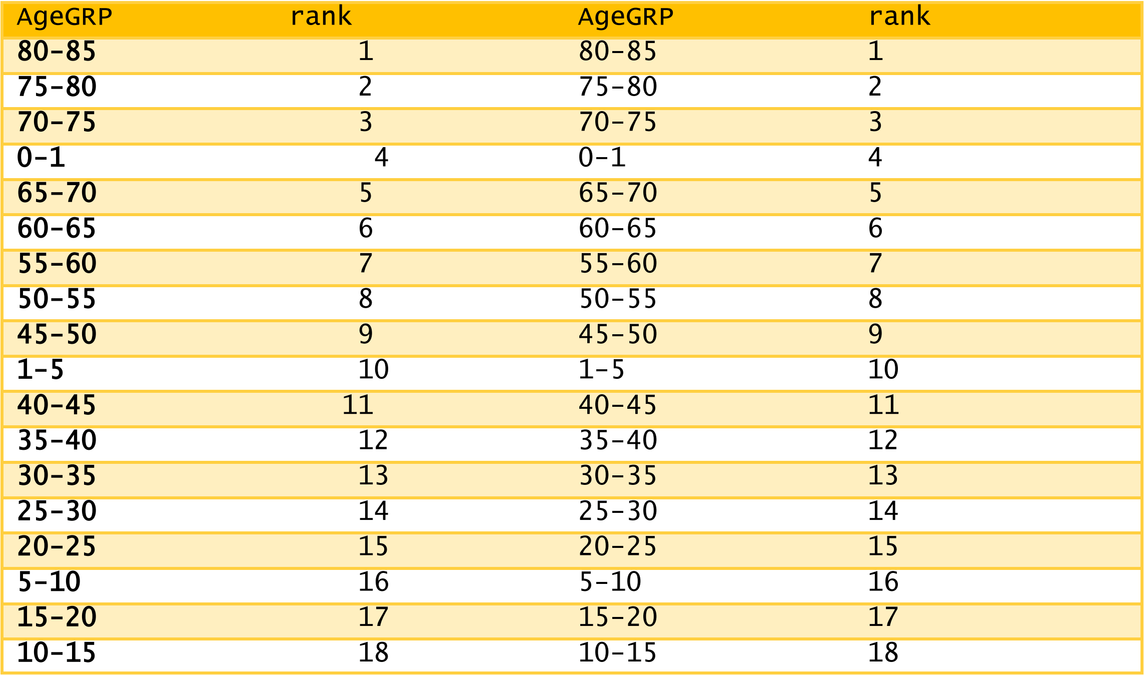
In conclusion, while the Decision Tree model had the best performance on the **training data**, the **Bagging model** had the best performance on the **test data**, suggesting it may be the most robust model for this particular task.

2.Comparative Analysis of Mortality Trends Across Age Groups and Genders:

In this section, we will be employing the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) for our comparative analysis. TOPSIS is a multi-criteria decision-making method that calculates the relative closeness of various alternatives to an ideal solution.

In the context of our project, the ‘alternatives’ are the different age groups and genders, and the ‘criteria’ are the mortality rates. The ‘ideal solution’ would be the age group/gender with the most favorable mortality trend.

By applying TOPSIS, we aim to rank the age groups and genders based on their relative performance in terms of mortality trends. This method will allow us to identify which age groups/genders show better or worse trends compared to others.



AgeGRP	rank	AgeGRP	rank
80-85	1	80-85	1
75-80	2	75-80	2
70-75	3	70-75	3
0-1	4	0-1	4
65-70	5	65-70	5
60-65	6	60-65	6
55-60	7	55-60	7
50-55	8	50-55	8
45-50	9	45-50	9
1-5	10	1-5	10
40-45	11	40-45	11
35-40	12	35-40	12
30-35	13	30-35	13
25-30	14	25-30	14
20-25	15	20-25	15
5-10	16	5-10	16
15-20	17	15-20	17
10-15	18	10-15	18

Male data **Female data**

Based on the results from the TOPSIS analysis, we can indeed observe that some age groups show better or worse trends than others in terms of mortality. Here’s a summary:

1. The age group **80-85** shows the best trend in terms of mortality, with the highest TOPSIS score of **1.000**, ranking **1st**. This suggests that this age group has the most favorable mortality trend among all age groups.

2. The age group **75-80** also shows a relatively favorable trend, with a score of **0.749**, ranking **2nd**.
3. On the other hand, the age group **10-15** shows the least favorable trend with a score of **0.000**, ranking **18th**.

These results indicate that older age groups (80-85 and 75-80) tend to have high mortality rate compared to the younger age group (10-15).

4. Males in the top five age groups, as ranked by the TOPSIS score, exhibit a higher risk of mortality.
5. Conversely, males in the last three age groups have a comparatively lower risk of mortality.
6. Males in the middle age groups present a moderate risk of mortality.

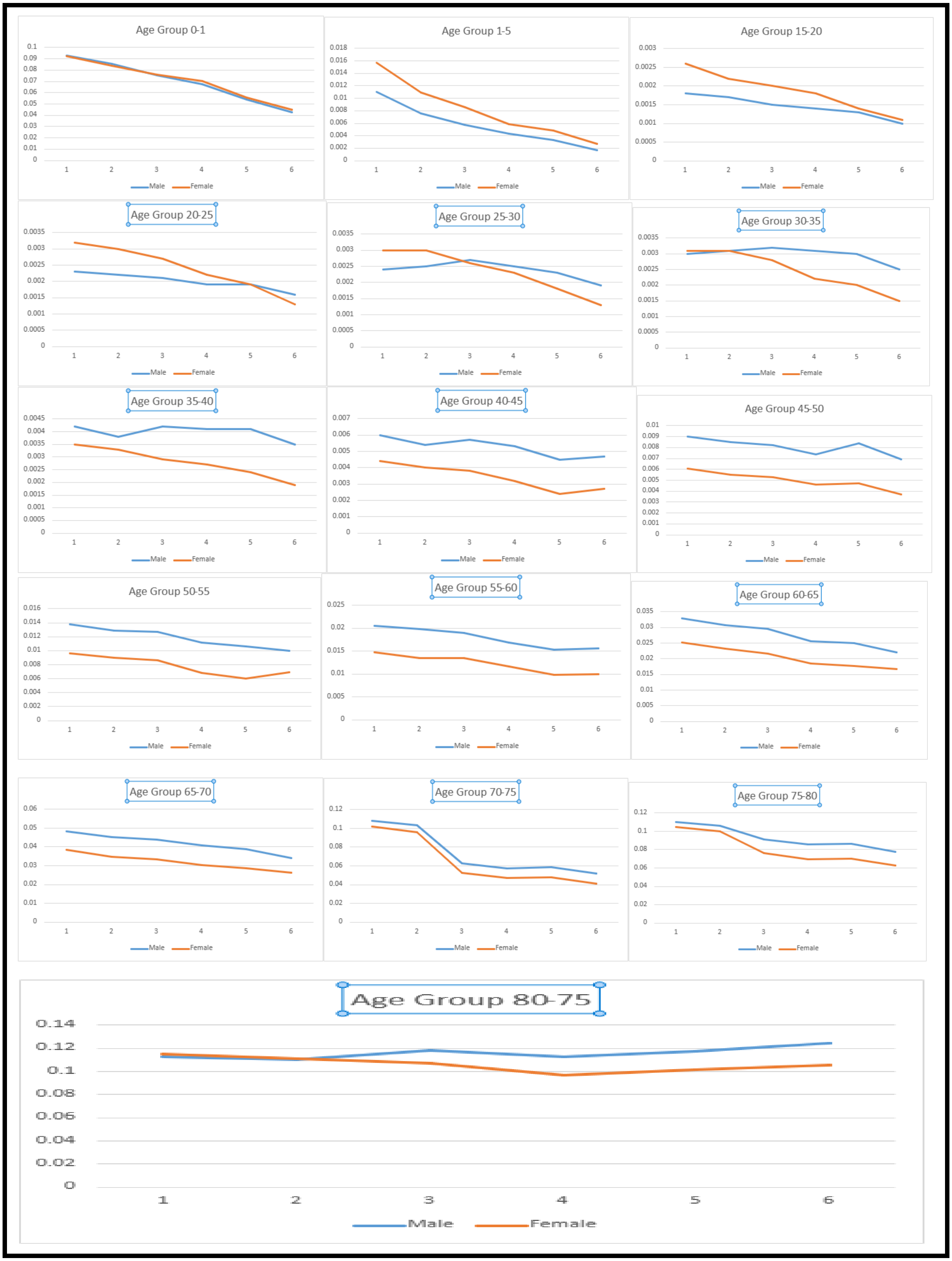
These observations provide valuable insights into the varying risk levels across different age groups within the male population.

- Similar to the male data, the female data also shows varying levels of mortality risk across different age groups.
- Females in the top five age groups, as per the TOPSIS ranking, exhibit a higher risk of mortality.
- Conversely, females in the last three age groups have a comparatively lower risk of mortality.
- Females in the middle age groups present a moderate risk of mortality.

These findings suggest that both males and females show similar patterns in mortality trends across different age groups.

3.Additional Insights on Mortality Rates

The line plot presented below illustrates the comparison of mortality rates between males and females across various age groups. The X-axis of the graph represents the years, spanning from 1986-1990 to 2011-2015, while the Y-axis denotes the mortality rate. This visualization provides a comprehensive view of how mortality rates have evolved over time for different age groups and genders.

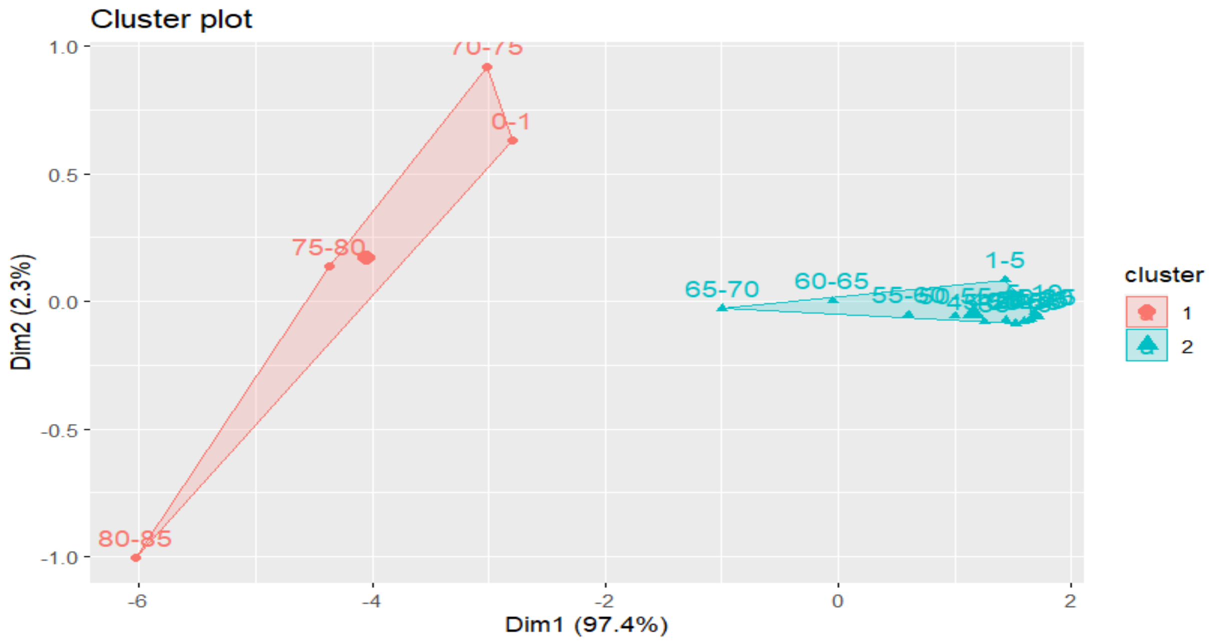

Age Group 0-1
0.1
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
1
2
3
4
5
6
Male
Female
Age Group 1-5
0.018
0.016
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
Age Group 15-20
0.003
0.0025
0.002
0.0015
0.001
0.0005
Age Group 20-25
0.0035
Age Group 25-30
Age Group 30-35
Age Group 35-40
0.0045
0.004
Age Group 40-45
0.007
0.006
0.005
Age Group 45-50
0.009
0.007
Age Group 50-55
Age Group 55-60
0.025
0.02
0.015
Age Group 60-65
0.035
0.03
Age Group 65-70
0.06
Age Group 70-75
0.12
Age Group 75-80
Age Group 80-75
0.14

From the analysis of the plots, it’s evident that at birth, both males and females exhibit similar mortality rates. However, a divergence in trends is observed from the age group 1-5 to 25-30, where males demonstrate a lower mortality rate. Interestingly, a reversal in this trend occurs beyond the 25-30 age group, with females showing a lower mortality rate up to the age group 80-85.

Another significant observation is the consistent downward slope in the graphs from the period 1986-1990 to 2011-2015. This trend indicates a steady improvement in mortality rates over time, suggesting advancements in medical facilities and healthcare services. The decrease in mortality rates over the years is a positive sign of progress in public health.

In addition to these observations, it would be interesting to investigate the factors contributing to these trends. For instance, changes in lifestyle, dietary habits, access to healthcare, and public health initiatives could have influenced these mortality trends. Understanding these factors could provide valuable insights for future health policies and interventions.

Cluster Analysis

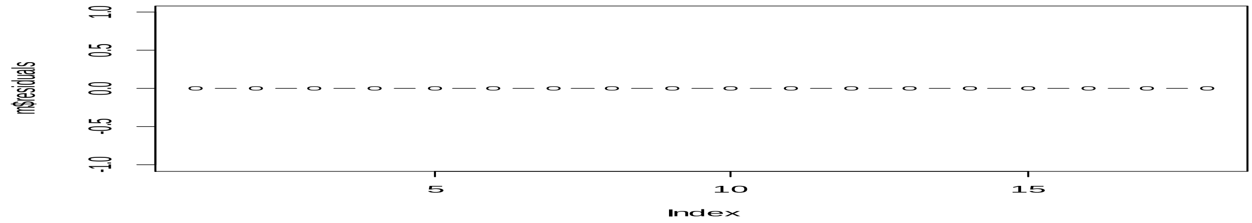


The application of the K-means clustering method on the data has in two divided into parts for both male and female data. Interestingly, the age groups 0-1, 70-75, 75-80, and 80-85 form one cluster, while the remaining age groups form the other cluster. This clustering suggests that the mortality rates for the age groups 0-1, 70-75, 75-80, and 80-85 are similar to each other.

Furthermore, this clustering pattern might indicate underlying similarities in the factors affecting mortality rates within these age groups. For instance, the age group 0-1 is typically vulnerable due to infancy, while the age groups 70-75, 75-80, and 80-85 are more susceptible due to age-related health issues. These shared vulnerabilities could explain the similar mortality rates within this cluster.

On the other hand, the second cluster comprising the remaining age groups might represent a different set of factors influencing mortality rates. Further investigation into these factors could provide additional insights into the patterns observed in the mortality rates across different age groups.

Multiple Linear Regression on Original Data: In the given project, a Multiple Linear Regression (MLR) model was fitted to the data. The response variable was the mortality rate for the years 2011-2015, and the predictors or regressors were the ‘Age Group’ and the ‘Average of all mortality rates from 1986-1990 to 2006-2010’. After fitting the model, an interesting insight was observed: the residuals were zero. This implies that the model’s predictions for the mortality rate for the years 2011-2015 were exactly the same as the actual mortality rates. This can be visualized in the residual plot, where the residuals are parallel to the x-axis at y=0.This result indicates a perfect fit of the model to the data, which is quite rare in real-world scenarios. It suggests that the chosen predictors are highly predictive of the response variable. However, it’s important to be cautious with such results, as they may also indicate overfitting, where the model is too closely fit to the training data and may not generalize well to new, unseen data


m$residuals
Index