Integrating Recursive Feature Elimination Technique To A Balanced Clustered Mental Health Data

Abstract. The state of mental health has shown to be of significance to an individual's quality of life. There are a couple of factors that may lead to psychological disorders. These may include: biological, social, environmental and many more.
Aim: This study aimed at filtering out the factors by selecting the variables based on their impact to an individual emotional wellbeing. Understanding of the least contributing factors will play a significant role in research studies, health sector, governments bodies and so on. This is through helping them minimize their area of focus while dealing with mental health awareness and making it easier for them to give better mental health care solutions.
Sample: In this study a sample of 10,000 observations from a generated data, comprising of 12 variables was used for the analysis. These variables included: Gender, Age, Marital status, Family members, Residence, Occupation, Medical test, Diagnosis, Cause, Treatment and payment.
Methodology: Random Undersampling balancing technique was first applied to the mental health data. This was in order to deal with the imbalanced nature of the data and thus reduce model selection biasness. After balancing the data K Means clustering technique was then used so as to group the observations into distinct sub-groups. Clustering of the data helped in improving the accuracy of the output of the study. Finally, Recursive Feature Elimination Technique (RFE) was then integrated to the balanced clustered data. Application of RFE helped in selecting the variable that least affects an individuals' mental well-being.
Results: Marital status variable indicated lower values for both the root mean squared error standard deviation (RMSESD) and Mean Absolute Error Standard Deviation (MAESD), with values 0.003605 and 0.0003382 respectively.
Conclusion: The findings given above, have shown that marital status variable was selected as the least contributing factor that leads to psychological conditions.

Keywords. Random Undersampling Technique, K-Means Clustering technique, Recursive Feature Elimination Technique.

1 Introduction

Mental Health is a condition that affects an individuals' way of thinking, emotional behavior and social relations. According to World Health Organization (WHO) (15), there are different types of disorders namely: Anxiety, Depression, Bipolar, Post-traumatic stress disorder, Schizophrenia and many more. Factors that contribute to these disorders according Mental Health Foundation are: Biological, Psychological and social.

The different types of disorders are oftenly treated differently. According to (8), there are three forms of treatment namely: psychotherapy, medication and lifestyle. Due to the variety of forms of treatment individuals have room to choose the one they prefer or as directed by the physician. This choice diversity thus leads to mental health data imbalance. Working with an imbalanced dataset leads to model biasness, inaccuracies and also incorrect predictions. This study aims at dealing with the data imbalance by balancing the dataset based on treatment variable using random undersampling technique. Balancing of the data helps in reducing over-fitting of the majority class,improve model performance thus resulting in accurate predictions.

Studies have shown that there is a great difference in the number of individuals suffering from mental health disorders based on their area of residence. According to (13), the rate of individuals suffering from mental health conditions in the urban area to those in rural areas was 80.6% to 48.9% respectively. This study applies K-means clustering technique as a clustering algorithm. Data clustering is applied to the random undersampled data based on the residence variable. Clustering

of the data will help in reducing the scale of the data and thus simplifying its interpretation.

Finally, Recursive feature elimination (RFE)technique is then integrated to the balanced clustered data. The aim of applying Recursive feature elimination technique is so as to identify the variable that is the least contributor to mental health conditions. This will help improve model efficiency and enable refining of the focus on key factors.

2 Literature Review

Mental health data is often considered as imbalanced. (4),applied three data sampling technique to a mental health dataset with the aim of dealing with data imbalance. The three data sampling techniques were namely: undersampling, oversampling and hybridization. Specific methods namely: Random undersampling, Tomek-Link, Edited Nearest Neighbor, Random oversampling, SMOTE, SMOTE-Tomek and SMOTE-ENN were applied. Random undersampling method turned up to be the best data balancing technique with an ideal value of 1.

In dealing with data imbalance, (6) applied random undersampling and oversampling and SMOTE-Nominal and continuous balancing techniques. This balancing techniques were applied on a data extracted from a database generated from the electronic health records of the mental health service of the Ferrara province, Italy. aplication of the balancing techniques was done to different training sets with different percentages of class imbalance. The three training sets were divided as: 50%-50%, 60%-40% and 70%-30% respectively. Results showed that random oversampling at 50%-50% gave a better model performance.

An improved balanced Random Survival Forest model developed by (14), involved the application of four balancing techniques. These balancing techniques were namely: Random undersampling, Random Oversampling ,Hybridization of Random oversampling and undersampling and SMOTE. Findings showed that random undersampling method gave a better model performance.

Area of residence has shown to be a contributing factor to an individual state of mind. (5), applied two clustering techniques namely; K-means and Divisive hierarchical clustering to a balanced mental health data. The aim of the study was to determine which of the two clustering technique is more effective. Findings showed that K-means clustering technique turned out to give a better performance.

In the quest of investigating optimum features for clustering the healthy aging data set, Kouser et al(2024) applied the natural evolution process inherent in genetic algorithms (GA's). This was done combining a number of clustering algorithms. These clustering algorithms were: partitional, density based and Agglomerative. The application of the fore-mentioned aplgorithms was so as to verify the results. The particular clustering techniques used were K-means, DBSCAN, BIRCH and agglomerative clustering. findings indicated that K-means clustering gave a better performance.

An study was done by (11), aiming at identifying specific student profiles based on their similarities in their mental health profiles, demographic attributes and academic performance.The study involved the use of K-means algorithms to a dataset comprising of 500 college students. Results indicated the importance of implementing support services and resources to address the different needs of students and promote positive mental health outcomes.

A novel approach for dementia prediction was introduced by (2). It involved the use of logistic regression model that was improved by integrating recursive feature elimination technique. The model was applied to a dataset composed of 1000 patients. Findings showed that integrating recursive feature elimination to logistic regression model improved the accuracy and efficiencyof the model for the prediction.

In order to improve the investigation of structural abnormalities in schizophrenia (SZ) patients, (10) applied a machine learning method. This method involved the combination of support vector machine (SVM) with recursive feature elimination (RFE) so as to discriminate SZ patients from normal controls using their structural MRI data. Findings showed that an SVM-RFE classifier using the significant structural abnormalities identified by the Voxel-based morphomety analysis, gave the best performance with its accuracy,sensitivity and specificity values of 88.4%, 91.9% and 84.4% respectively.

A study done by (12), involved performing multiclass classification using a hierarchical extreme learnig machine (H-ELM) classifier. 159 structural MRI images of children from the publicly available ADHD-200 MRI images was used. A comparison of the perfomance of the classifier with that of support vector machine and basic extreme learning machine was done. Feature selection was carried out using standard SVM-based recursive feature elimination. Findings showed that RFE-SVM feature selection approach combined with H-ELM effectively enalbled the acquisition of high multiclass clasification accuracy rates for structural neuroimaging data.

Based on the various mental health analysis done in using recursive feature elimination, this study aims at using recursive feature selection to a balanced clustered mental health data. This is with the aim of identifying the variable that is the least

contributor to mental health conditions.

3 Methods

In this study quantitative design was applied in order to achieve its objective. Data comprising of 10,000 observations and 12 variables was used for the analysis .The twelve variables were Gender, Age, Marital Status, Family members, Residence , Occupation, Medical test, Diagnosis , Cause ,Treatment and Payment.

3.1 Experimental Setup.

1. **Random Undersampling Technique**
   In this study data imbalance was dealt with using random undersampling technique based on the treatment variable. The dataset was first divided in two sets that is the training set (80%) and test set (20%) respectively before applying the balancing technique.
   Random under sampling involves balancing of the dataset by randomly extracting observations from the majority class to match that of the minority class.Figure [1]represents an illustration of the process. (3) and (4)

2. **K-Means clustering technique**
   Data clustering was applied to the balanced dataset based on the residence variable.This study Applied K-means clustering technique.Below are the steps that are followed while implementing the fore-mentioned technique: (1)

   - Select K points as initial centroids.
   - Repeat the first step.
   - Form K clusters by assigning each point its closest centroid.
   - Re-evaluate the centroid of each cluster.
   - Repeat until convergence criterion is met.

3. **Recursive Feature Elimination**
   This is a variable elimination algorithm functions by inserting the model and extracting the weakest variables.The process is repeated till the necessary number of variables is attained. The two significant alternative dispositions while applying RFE are:choosing the number of variables to extract and deciding on the technique integrated for singling out the variables.
   The main algorithms used are linear regression (mainly used for regression context) and logistic regression (mainly used for classification problems).The formal procedure works as follows.(7)

   - Insert the model to the dataset.
   - Remove the feature with the smallest coefficient.
   - Repeat steps 1-2 until you reach the number of variables you want. Normalize or standardize the variables to have the same scale.

3.2 Performance Metrics

Statistical test metrics was done on both the imbalanced data and the random undersampled data so as to determine the quality and perfomance of the model. (3) and (4). Below are the specific test metrics.

1. **Confusion Matrix**
   This measure evaluates the functionality accuracy of a given model .It is utilized in both binary and multi-class stratification hurdles for the calculation of real and forecasted values
   The output values comprises of actual negative denoted as TN^-, actual positive denoted as TP^+, erroneous negative denoted as FN^- and erroneous positive denoted as FP^+. TN^-, TP^+, FP^+ and FN^- indicates that the prediction were,correctly positive,correctly negative,incorrectly positive and incorrectly negative respectively. The matrix below gives an illustration of the confusion matrix.[1]

$$\begin{bmatrix} & Reference & \\ Prediction & Negative & Positive \\ Negative & TN^- & FP^+ \\ Positive & FN^- & TP^+ \end{bmatrix} \tag{1}$$

2. **Accuracy** Accuracy is a performance metrics that corresponds to the sum of the diagonal elements in the confusion matrix divided by the total number of instances. It is given by the formular below;

$$Accuracy = \frac{(TN^- + TP^+)}{(TN^- + FP^+ + FN^- + TP^+)} \tag{2}$$

3. **Precision** Precision computes the fraction of actual positive values among the values categorized as positive. It is given by the formular below;

$$Precision = \frac{TP^+}{TP^+ + FP^+} \tag{3}$$

4. **Recall** Recall is the fraction of total positive instances accurately categorized as positive. It is given by the formular below;

$$Recall = \frac{TP^+}{TP^+ + FN^-} \tag{4}$$

5. **F Measure** It is a measure that focuses on the analysis of positive class and uses a weighed harmonic mean between precision (ρ) and recall (λ). It is given by the formular below;

$$FS = 2 \times \left(\frac{\rho \times \lambda}{\rho + \lambda}\right) \tag{5}$$

4 Results and Discussion

4.1 Results

1. **Random Undersampling Technique**
Below are the outputs before and after applying data balancing technique.

   - Treatment Variable Training and Test Table
   The figures [2] and [3] below represent the distribution of observations in both the training and test sets respectively.

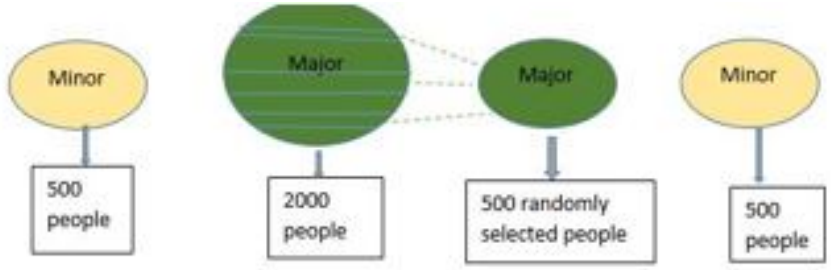

Figure 1: Random Undersampling Demo

TREATMENT	PSYCHOTHERAPY	MEDICATION
NUMBER OF OBSERVATIONS	4,790	3,210

table 1: Train Sampled Treatment Dataset

   - Confusion Matrix
   The matrices given below [6] and [7] are the confusion matrices for both the imbalanced and random undersampled data respectively.

$$\begin{bmatrix} & Reference & \\ Prediction & 0 & 1 \\ 0 & 646 & 377 \\ 1 & 552 & 425 \end{bmatrix} \tag{6}$$

$$\begin{bmatrix} & Reference & \\ Prediction & 0 & 1 \\ 0 & 1198 & 0 \\ 1 & 0 & 802 \end{bmatrix} \tag{7}$$

- Performance Metrics Results.
  Figure [3] indicates the performance output before and after data balancing.

TREATMENT	PSYCHOTHERAPY	MEDICATION
NUMBER OF OBSERVATIONS	1,198	802

table 2-: Test Sampled Treatment Dataset

2. **K-Means Clustering Technique**
   The figures below are the outputs for K-means clustering technique applied to the random undersampled data.

- Residence variable distribution table

DATA	ACCURACY	RECALL	PRECISION	F SCORE
IMBALANCED	0.5355	0.5392	0.6315	0.5817
RANDOM UNDERSAMPLED	1	1	1	1

table 3: Balanced Data Test Statistics Table

- Optimal Cluster Point Plot
  Figure [5] gives a diagramatic representation of the cluster point plot.

RESIDENCE	URBAN	RURAL
NUMBER OF OBSERVATIONS	6,017	3,983

table 4-: Residence Distribution Table

- Cluster Distribution Plot
  Figure [3] gives a diagrammatic display of the distributions of observations in the various clusters.

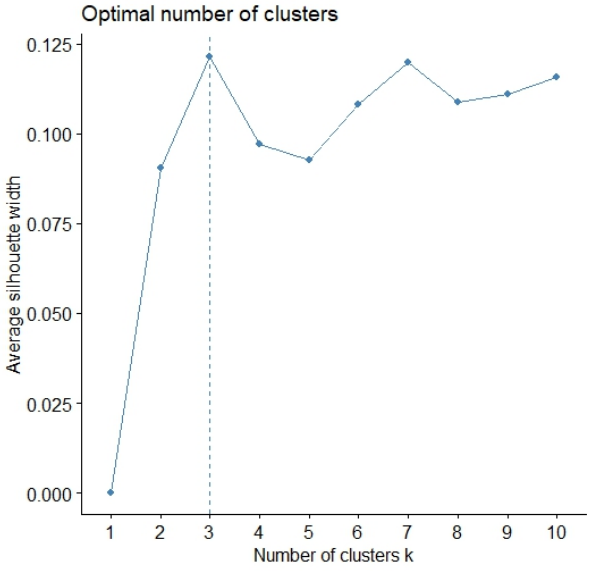

Figure 2: Optimal Balanced Kmeans Cluster Point Plot

- Cluster Distribution table.
  The figure [3] indicates the distribution of observations into the three distinct clusters.

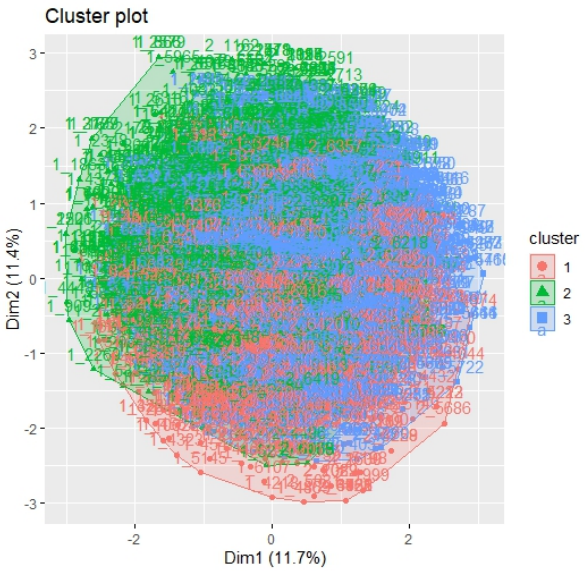

Figure 3: Balanced Kmeans Cluster Plot

3. **Recursive Feature Elimination Technique**
   The figures [4] and [5] given below represents the outputs after applying recursive feature elimination technique.

   - RFE Output Table

K-MEANS CLUSTERS	URBAN (1)	RURAL (2)
1	624	993
2	667	985
3	1256	1895

table 5-: Balanced Kmeans Cluster Table

   - Cross Validation Plot

```
Outer resampling method: Cross-Validated (10 fold)

Resampling performance over subset size:

 Variables  RMSE  Rsquared MAE    RMSESD RsquaredSD     MAESD Selected
         1 1.000 0.0003832   1 0.0003605  0.0004298 0.0003382        *
         2 1.000 0.0006510   1 0.0005736  0.0004420 0.0005245
         3 1.001 0.0006502   1 0.0005470  0.0008578 0.0004941
         4 1.001 0.0006063   1 0.0005755  0.0008603 0.0005010
         5 1.001 0.0007863   1 0.0006722  0.0013209 0.0005969
         6 1.001 0.0008607   1 0.0007287  0.0014323 0.0006498
         7 1.001 0.0009227   1 0.0007429  0.0014875 0.0006634
         8 1.001 0.0009314   1 0.0007491  0.0014676 0.0006695

The top 1 variables (out of 1):
   status
```

Figure 4: RFE Output

4.2 Discussion

Below are the detailed explanations of the results.

1. **Random Undersampling Technique**
   Data balancing was carried out based on the treatment variable where by the imbalanced data set that comprised of 10,000 observations was split into two sets. The split was of 80% to 20% of the training and test set respectively. Figure [2] and [3] indicate the observations of the observations undergoing the distinct mode of treatment. Psychotherapy being the majority class indicates 4790 and 1198 number of observations in the training and test set respectively, while medication which represents the minority class has 3210 and 802 number of observations respectively.

   The confusion matrices [6] and [7]indicate the entries of the respective outputs before and after balancing of the dataset. The rows and columns are labeled as predictions and references respectively, with values 0 and 1 representing the majority and minority classes respectively . For the imbalanced data confusion matrix [6], entries [0,0] and [1,1] indicate the values that were correctly predicted as 646 and 425 for classes 0 and 1 respectively. Entries [0,1] and [1,0] indicate values 377 and 552 that were incorrectly predicted . According to the confusion matrix for the random undersampled dataset [7], it is noted that the model correctly predicted values 1198 and 802 for entries [0,0] and [1,1] respectively and none of the observations were incorrectly predicted as per entries [0,1] and [1,1] respectively.

   Figure [4] shows the results of the various test metrics that were used to evaluate the performance of the model before and after balancing of the data. From the test metrics outputs, it is noted that applying random undersampling technique has helped in the improvement of model performance. This is due a change in the values of the accuracy, recall, precision and F-score from (0.5355, 0.5392, 0.6315 and 0.5817) for the imbalanced data to 1 which is an ideal value for the random undersampled data.

2. **K-Means clustering technique**
   Clustering of the data was done based on the residence variable with the distribution of observations indicated in figure [5] as 6,017 and 3,983 for those residing in the urban and rural areas respectively. Silhoutte method was applied so as to define the optimal number of clusters as show in plot [6]. According to the plot it is noted that the optimal cluster point is 3, hence the data should be divided into three distinct clusters.

   Figure [3] displays a diagrammatic representation of the observations in the three distinct clusters. Due to the overlap viewed in the plot, the distribution of the observations may not be clearly reflected. Hence, figure [4] indicates the number of observations distributed in the three clusters. For the two original clusters urban and rural denoted as cluster 1 and 2 respectively, it is noted that 624 and 985 were correctly placed in their respective clusters. Observations from cluster 1 that were incorrectly placed in distinct k-mean clusters 2 and 3 were 667 and 1256. Cluster 2 observations that were incorrectly placed in distinct k-means clusters 1 and 3 were 993 and 1895. Observations that fell at the borders of the distinct k-means clusters 1, 2 and 3 and were not properly placed are 3470 and 110 that is from original clusters 1 and 2 respectively.

3. **Recursive Feature Elimination Technique**
   According to figure [5] ,the marital status (status) variable indicated lower values of the Root Squared (Rsquared), Root Mean Squared Error Standard Deviation (RMSESD), Root Squared Standard Deviation (RSquaredSD) and Mean Ab- solute Error Standard Deviation (MAESD) with values 0.0003832, 0.003605, 0.0004298 and 0.0003382 respectively. The Root Mean Squared Error Cross Validation (RMSE-CV) Plot [10] shows that there is a slant towards the variable that has a lower RMSE value thus indicating that it is the least important variable. Finally, based on the two figures [4] and [5] it is concluded that marital status is the least important variable.

5 Conclusion

Balancing of the data using random undersampling technique improved the performance of the model. This is due to an increase in the accuracy, recall, precision and F score values to 1 which is regarded as an ideal value. Applying K-means clustering technique to the balanced dataset divided the observations into three distinct clusters. Grouping of this dataset helped in exposing insights that may have not been detected through manual scrutiny.

Integrating recursive feature elimination technique to the balanced clustered data extracted marital status as the least contributing factor that leads to mental health conditions. This findings will be of great significance to health practitioners and even the government bodies in refining the focus on the key contributing factors to mental health disorders.

More research work can be done using other machine learning balancing and clustering techniques, while integrating the dataset to the recursive feature elimination technique. Based on the Root Mean Squared Error Cross Validation (RMSE-CV)Plot (figure 4.73) the graph seems to slant towards variable one which has a lower RMSE value thus indicating that it is the least important variable.

References

[1] Aggrawal.C and Reddy.C. Data Clustering:Algorithms and Applications.*Taylor and Francis Group,LLC*,ISBN 13:978-1-4665-5822-9,(2014).

[2] Ahmed.R,Fahad.N,Miah.S.U,Hessen.J,Morol.K,Mahmed.M and Rahman.M. A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction .*ELSEVIER*,(2024).

[3] Fernandez,A.,Garcia,S.,Galar,M.,Patri,R.C.,Krawczyk,B. and Herrera,F. Learning from Imbalanced data sets.*Springer Nature Switzerland AG*,ISBN 978-3-319-98074-4,(2018).

[4] Chege.W.L,Waititu.H.W,Nyakundi.C.O. Comparative Analysis of Data Balancing Techniques in Mental Health Data:Application to Treatment Modalities.*International Journal of Statistics and Application*,(2024).

[5] Chege.W.L,Waititu.H.W,Nyakundi.C.O. Comparative Analysis of K-Means and Divisive Clustering Techniques on Balanced Mental Health Data.*Asian Journal of Probability and Statistics*,(2024).

[6] Gentili,E.,Franchini,G.,Zese,R.,Alberti,M.,Domenicano,I. and Grassi,L. Machine Learning from Real Data: A mental health registry case study.*Computer Methods and Programs in Biomedicine*.5,100132(2024).

[7] Guyon,I.,Weston,J. and Barnhill,S.. Gene selection for cancer classification using support vector machine.*Kluwer Academic Publishers*.(2002).

[8] HealthDirect.MentalIllness .*https://www.healthdirect.gov.au/mental-illness*.

[9] Kouser.K,Priyam.A,Gupta.M,Kumar.S and Bhattacharjee.V.Genetic Algorithm-Based Optimization of clustering Algorithms for the Healthy Aging Dataset.*MDPI*,2024.

[10] Lu.X,Yang.Y,Wu.F,Gao.M,Xu.Y,Zhang.Y,Yao.y,Du.X,Li.C,
Wu.L,Zhong.X,Zhou.Y,Fan.N,Zheng.Y,Xiong.D,Peng.H,
Escudero.J,Huang.B,Li.X,Ning.Y and Wu.K.Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images .*Medicine*,2016.

[11] Ouyang.X.Application and Effectiveness Assessment of Big Data analysis algorithm in college students' mental health education.*Journal of Electrical Systems* , 2024 .

[12] Qureshi.M.N.I,Min.B,Jo.H and Lee.B. Multiclass classification for the differential diagnosis on the ADHD subtypes using recursive feature elimination and Hierarchical extreme learning Machine Structural MRI study .*Journal of Electrical Systems* , 2016 .

[13] Reddy . Prevelance of mental and Behavioral Disorders in India. *India Journal Of Psychiatry* , 1998.

[14] Waititu .H.W. Improved Balanced Random Survival Forest for the analysis of right censored data:apllication in determining under five child mortality . *Moi University Open Access Repository*, 2021.

[15] WHO.Mental Health Disorders . *WHO*, 2022.

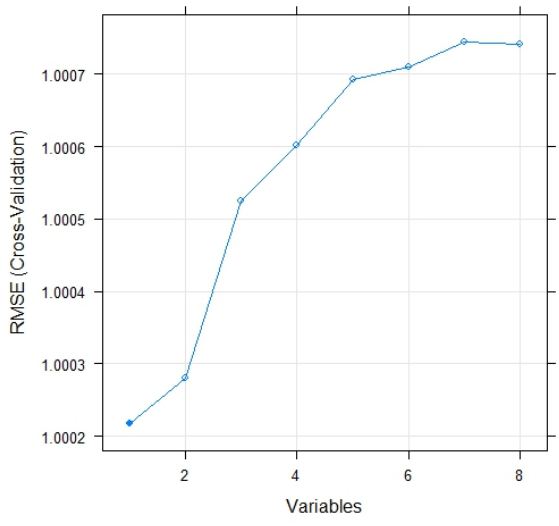

Figure 5: RFE Plot