FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 8 No. 1, February, 2024, pp 345 - 353
DOI: https://doi.org/10.33003/fjs-2024-0801-2276

COMPARISON OF THREE DISTRIBUTION FREE CLASSIFICATION TECHNIQUES APPLIED TO CRIME DATA OF NIGERIA PRIOR AND POST COVID-19 PANDEMIC

[*1,2]Usman Abubakar, [1]Abdulhamid Ado Osi, [2]Yusuf Ibrahim Muhammad, [1]Ilyasu Abubakar Salisu, [1]Abba Bello Muhammad, [3]Nura Muhammad and [4]Wakili Abubakar

[1]Department of Statistics, Aliko Dangote University of Science and Technology Wudil, Kano State, Nigeria
[2]Department of Statistics, Jigawa State Polytechnic, Dutse, Jigawa State, Nigeria
[3]Department of Social Development, Jigawa State College of Remedial and Advance Studies, Babura, Jigawa State, Nigeria
[4]Department of Computer Science, Jigawa State College of Remedial and Advance Studies, Babura, Jigawa State, Nigeria

*Corresponding authors' email: Usmanabubakar@jigpoly.edu.ng Phone: +2348087962032

ABSTRACT

This paper considers a comparison and evaluation of the performance of the three distribution-free classification methods in classifying states in Nigeria. The methods were CT, KNN, and ANN. The methods classified the state into high and low crime rates using selected variables impacting crime rates. The classification results showed that the CT performed best, correctly classifying 8 as high and misclassifying 3, which yields an apparent error of 27.27%, and also correctly classifying 12 as low and misclassifying 2, which gives an apparent error of 14.29%, 80% accuracy, 80% sensitivity, and 80% specifity for training sample. While for the testing sample, the CT correctly classified 3 as high and misclassified 1, which yields an apparent error of 25%; it correctly classified 6 as low and misclassified 2, which gives an apparent error of 25%, 75% accuracy, 60% specifity, and 75.7% sensitivity, as shown in Table 7, respectively. The KNN method resulted in an apparent error of 66.67%, an accuracy of 41.67%, 42.86% sensitivity, and 40% specifity for testing data. While for training, in Table 3 below, KNN has an apparent error of 66.67%, an accuracy of 88%, 90% specifity, and 86.67% sensitivity, respectively. Lastly, the ANN did not perform well; correctly classified gives an apparent error of 100%, an accuracy of 0%, 0% sensitivity, and 0% specificity in the training sample, while for the testing sample, the method has an accuracy of 50%, 28.57% sensitivity, and 100% specificity. However, it offers many advantages that make it a useful method.

Keywords: k-Nearest Neighbor, Neural Network, Decision Tree, Official Crime Statistics, States of Nigeria

INTRODUCTION

One issue that has plagued humanity throughout its history is crime (Classen & Scarborough, 2012). (Hoffmann & Stuntz, 2021) defined crime as an intentional act committed without defense or excuse in violation of the criminal law (statutory and case law) and punished by the state as a felony or misdemeanor. There are several potential causes of crime, including psychological issues like personality disorders (Abramson, 1944), sociological elements like learning and environment and bio-genetic variables like genetic mutation and heredity (Hooton, 1939). Different viewpoints have been taken when defining crime by academics and social analysts. As such, academics have struggled for years to come up with a general description of the idea. While Farmer (Farmer, 2022) views crime as a category established by law, (Kennedy, 2021) offered a more thorough definition, stating that an act that harms a community, society, or the state in addition to an individual or individuals is a crime or offense. (Chambliss, 2011) describes behaviors that are prohibited under the law, including crimes against the state like murder, theft, resisting arrest, drunk driving, and the possession or sale of illegal narcotics. According to (Albanese, 2014), a crime is an act that transgresses a moral or political law; an act committed by a lone person motivated by personal gain, or perhaps it's an organized crime where gangs of mobsters use violence and murder to further their financial interests at the expense of the public.

The accessibility, authenticity, and trustworthiness of official crime statistics are serious issues in Nigeria (Brody et al., 2022). When available, official crime statistics for Nigeria are not widely obtainable by the general public and are neither recent nor accurate. Here, the first question to consider is whether crime figures in Nigeria as reported by the police are accurate, dependable, and useful. The degree to which the crime figures accurately depict actual criminal activity is referred to as validity. The degree to which crime statistics reflect a consistent measurement of the same occurrence across towns, states, and villages is referred to as reliability (MacDonald, 2002). Utility is the measure of how much the crime statistics add to our understanding. Beyond the more common flaws, police crime statistics are rife with mistakes; the police record is invalid in relation to the actual or true volume of crime in the nation (Logan & Ferguson, 2016).

This study focuses on the application of some machine learning techniques to some selected factors affecting crime management in Nigeria to classify the states as having "high" and "low" crime rates and to evaluate the performance of these techniques. (Osi and Dikko, 2014) have presented good illustrations of the classification using the parametric discriminant analysis method. The data used in the analysis were not initially multivariate normal, so the researchers had to employ some transformation techniques to make the data look more normal. In this case, three machine learning techniques will be used to classify the state into "high" and "low" crime rates (i.e., K-nearest neighbor, classification three, and neural networks). Likewise, each of the techniques will be used to partition the data into training and testing samples, which may be used to evaluate the performance of each technique. These study focuses on the application of some classification techniques to determine the impact of some selected socio-economic and demographic factors on crime management in Nigeria prior and post COVID-19 pandemic, which can lead to classifying the state as having high or low crime rates.

MATERIALS AND METHODS

Source of Data and Data Description

The data on poverty rate for the year 2019 from the online publication of National Bureau of statistics release on May, 2020 (Retrieve for Statista.com), unemployment rate , literacy rate from the online publication of NBS on Labour force statistics (NBS, August 2020), year-on-year internally generated revenue were obtained from an online publication (Budgit.org, retrieved on 2023/3/7), population density from the NPC/NBS website (Population projection of Nigeria by States as of 2022/03/21), drug arrest and seizure arrest index (SAI) from the National Drug Law Enforcement Agency (NDLEA Annual report 2019), literacy rate (UNESCO 2012), number of primary school teachers, and primary school enrolment from Federal ministry of Education as of 2018/2019 academic year (retrieved from statista.com), and police strength (crime in Nigeria, statistics and fact). The dataset has 37 entries; each entry represents the information of a particular state in Nigeria.

Table 1: List of variables

SN	Variables	Description	Type
1	PSE	Primary school enrolment	integer
2	POP.DEN	Population density	Integer
3	YOY.IGR	Year on year internally generated revenue	Integer
4	NPT	Number of primary teachers	Integer
5	DV.ARST	Divisional arrest	Integer
6	PLC.ARST	Police arrest cases	Integer
7	NDP	Number of divisional police	Integer
8	LITRACY	Literacy rate per State	Integer
9	POVERTY	Poverty rate per State	Integer
10	UEMP	Unemployment per State	Integer
11	CRIME	Crime rate	Integer
12	CRC	Crime category	String

K-Nearest neighbor

The K-Nearest Neighbor (KNN) algorithm is a popular machine learning technique used for classification and regression tasks. It relies on the idea that similar data points tend to have similar labels or values (Peterson, 2009). KNN was the earliest non-parametric classification method. It does not assume multivariate normality, but it does assume homogeneity of variance. KNN does not find a function to discriminate groups but classifies an observation based upon the group membership of a number, k, of its nearest neighbor. The KNN classified object based on the group membership, similarities and the distance (Euclidean) between them. The distance between the two observation/object can be calculated as.

$$d(s,t) = \sqrt{\sum_{i=1}^{n}(t_i - s_i)^2} \quad (1)$$

Where s,t are the two observation, and n is the number of spaces.

Classification Trees

Classification trees are non-parametric procedures that classify observations by repeatedly partitioning the data into subsets through a series of decisions (Song & Ying, 2015). The goal is to create a final subset that is homogeneous with respect to the group or class variable. Classification trees start with all of the observations grouped together in one node, referred to as the root node. All of the predictor variables are evaluated to determine which can be used to split the root node into groups that best separate the classes (Lewis, 2000). The best split is determined by evaluating a measure of impurity that quantifies the class makeup of each node (Sandri & Zuccolotto, 2010). The impurity measure is maximized when all the classes are equally mixed in a node and minimized when a node contains only one class. The chosen split is the one that most greatly reduces the impurity. To choose a split rule tree algorithm, evaluate each variable one-by-one. It determines a value for each variable that would provide the best split. Next, it compares the split for each variable to determine which does best and select that variable.

Artificial Neural Network (ANN)

The ANN is a widely used analytical technique that simulates the functioning of the human brain to help solve complex issues (Bekesiene et al., 2021). NN models the relationship between a set of input signals and an output signal using a model derived from our understanding of how a biological brain responds to stimuli from sensory inputs as presented in Figure 1. Just as a brain uses a network of interconnected cells called neurons to create a massive parallel processor, as the information received as an input pass through the activation function, were information is process and results as output as shown in Figure 2, NN uses a network of artificial neurons, or nodes, to solve learning problems (Abubakar et al, 2023). A neural network calculates classifiers by using $Q = f^{*}(\text{z})$, the input is then process using activation function and bring out the results as the final classification.

Evaluation of Classification Performance

Several parameters can be used for the quality estimation of classification performance, both for fitting and validation purposes (Lucic et al., 2019). Of course, these parameters are related to the presence of error in the results (objects assigned to the wrong classes), even if error can be considered with a difference weigh on the basis of the classification aims. All the classification indices can be derived from the confusion matrix, which is a square matrix with dimension QXQ, where Q is the number of classes. The diagonal element of the confusion matrix represents the correctly classified object, while the up diagonal element represents the object erroneously classified (Forbes, 1995).

Table 2: Confusion Matrix

	Positive (1)	**Negative (0)**
Positive (1)	True Positive (TP)	False Positive (FP)
Negative (0)	False Negative (FN)	True Negative (TN)

By looking at the confusion matrix in Table: (built on fitting or validation outcomes), we can have an idea on how a Classification model is performing; of course, some more informative indices can be derived in order to synthesize this information. First non-error rate (NER) which is also called accuracy can be defined as follows.

$$NER = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

And the error rate (performance error) adopted in this research is classification error, and can be obtain using NER

$$ER = 1 - NER = \frac{FP+FN}{TP+TN+FP+FN} \quad (3)$$

NER and ER can simply describe the performance of a model. There are also indices related to the Classification quality of single class. The sensitivity describes the model ability to correct recognize object belonging to the classes and is defined as;

$$Sensitivity = \frac{TP}{TP+FP} \quad (4)$$

If the object belonging to the classes are correctly assigned, sensitivity is equal to 1. The specifity characterizes the ability of the classes to reject the object of all other classes and is defined as;

$$Specifity = \frac{FP}{FP+TN} \quad (5)$$

RESULTS AND DISCUSSION

The data of the 37 States of Nigeria with 11 variables were analyze using R package (see Appendix for output). Here, one more response variable name crime category were be introduced. This variable is a categorical crime level for each State that is either above average or below average as in Table 5. So as to classify the 37 State in to high crime rate and low crime rate, by using the Classification techniques on the known information. Likewise, the State have been classified as either 'high' or 'low' in crime rate based on the States average, which is 2.702432 per 100 people. Assign '1' to State which are above the average and '0' to those below the average. Thus, we have 19 zero and 18 one for 37 States.

The data contains only 37 entries, which is felt too small data to get the accurate Classification rule due to splitting the data in to training sample and testing sample, there a sample of 25 is taking for training sample while the remaining 12 sample is taking for testing from the base line and test the effectiveness on both 25 and 12 sample of the 37 observation. The data is partition using cross validation, it is divided in two training and testing sample training sample based on proportion, 70% used for training sample while the remaining 30% used for testing sample, so the model learned on train and predicted based on testing sample.

K-Nearest Neighbor Classification

The KNN for k = 13 as illustrated in Figure:1 (ROC Curve) and in Appendix using the Classification rule generated from 37 observations to re-classify them gives the result shown in Table 3 and Table 4 below which summarize the number of misclassification and the apparent error rate of mis-classification as well as the specificity and sensitivity. Table 4 below, on the other hand, shows that the KNN method correctly classified 3 out of 6 states as having low crime, while misclassifying 3 states resulted in an apparent error rate of 50%. Additionally, the method correctly classified 2 out of 6 states as having high crime, and misclassifying 4 states resulted in an apparent error of 66.67%, accuracy of 41.67%, 42.86% sensitivity, and 40% specifity for testing data. In contrast, Table 4 below shows that, during training, KNN correctly classified 9 out of 11 states as having low crime rates, misclassified 2 states, yielding an apparent error rate of 18.18%, and correctly classified 2 out of 6 states as having high crime rates. Additionally, 4 misclassified states resulted in apparent errors of 66.67%, 88%, 90% specifity, and 86.67% sensitivity, respectively. By comparing their accuracy, this result demonstrates that the training sample outperforms the testing sample.

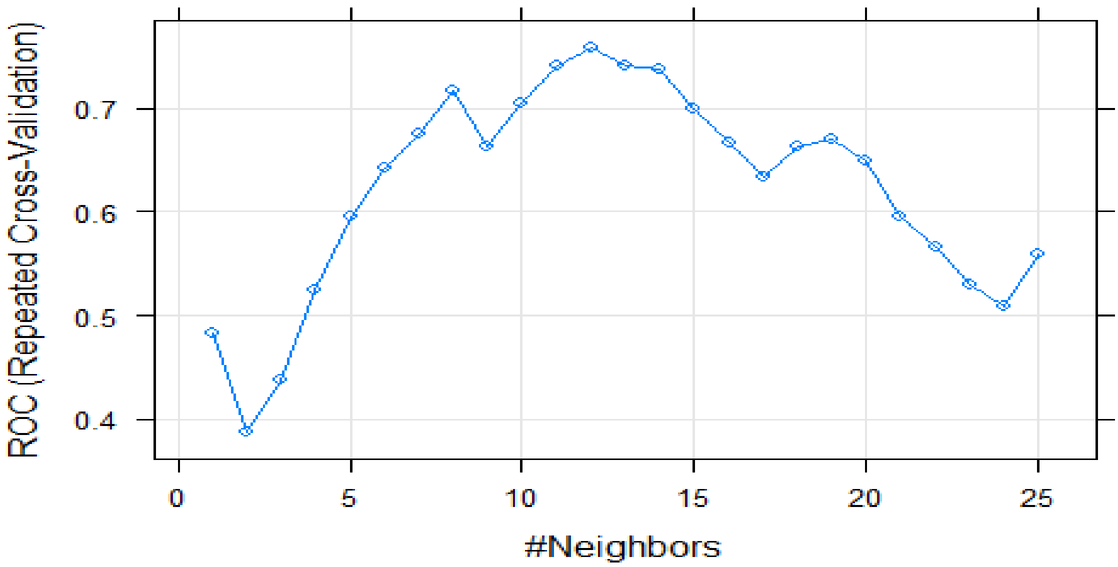

Figure 1: ROC (Repeated Cross Validation) Curve

Table 3: Summary result for testing sample

CATEGORY	Predicted group Membership		Total	Accuracy (%)	Sensitivity (%)	Specifity (%)
	High	Low				
Original count				41.67	42.86	40
High	2	4	6			
Low	3	3	6			
High	33.33	66.67	100			
Low	50	50	100			

Table 4: Summary result for training sample

CATEGORY	Predicted group Membership		Total	Accuracy (%)	Sensitivity (%)	Specifity (%)
	High	Low				
Original count				88	86.67	90
High	9	1	10			
Low	2	13	15			
High	90	10	100			
Low	13.33	85.67	100			

Artificial Neural Network

To identify patterns in the data, a neural network is used. As with the training and testing data sets, the model will classify the State. However, the neural network result presents a challenge because the data was split into training and testing samples. This means that the technique performs poorly if the observation is very small, which results in incorrectly classifying 0 out of 15 as high and misclassifying 15, yielding an apparent error of 100%. It also correctly classifies 0 out of 10 as low and misclassifies 10, yielding an apparent error of 100%, accuracy of 0%, 0% sensitivity, and 0% specificity in the training sample. In contrast, the approach misclassified 5 with an accuracy of 50%, 28.57% sensitivity, and 100% specificity, while correctly classifying 2 as low and 5 out of 10 as high for the testing sample. Consequently, taking this conclusion into account, it is evident that the testing sample outperformed the training sample.

Table 5: Summary result for training sample

CATEGORY	Predicted group Membership		Total	Accuracy (%)	Sensitivity (%)	Specifity (%)
	High	Low				
Original count				0	0	0
High	0	15	15			
Low	10	0	10			
High	0	100	100			
Low	100	0	100			

Table 6: Summary result for testing sample

CATEGORY	Predicted group Membership		Total	Accuracy (%)	Sensitivity (%)	Specifity (%)
	High	Low				
Original count				58.33	28.57	100
High	5	5	10			
Low	0	2	2			
High	50	50	100			
Low	0	100	100			

Classification Tree

However, the result of CT method in the Table 7 below shows that it correctly classifies 8 out of 11 as high and misclassify 3 which yield an apparent error of 27.27% and also correctly classify 12 out of 14 as low and misclassify 2 which give an apparent error of 14.29%, 80% accuracy, 80% sensitivity and 80% specifity for training sample. while for testing sample the CT correctly classify 3 out of 4 as high and misclassify 1 which yield an apparent error of 25%, also correctly classify 6 out of 8 as low and misclassify 2 which give an apparent error of 25%, 75% accuracy, 60% specifity and 75.7% sensitivity as shown in Table 8 respectively. In this case the training sample performed better than the testing sample.

Table 7: Summary result for training sample

CATEGORY	Predicted group Membership		Total	Accuracy (%)	Sensitivity (%)	Specifity (%)
	High	Low				
Original count				80	80	80
High	8	3	11			
Low	2	12	14			
High	72.27	27.73	100			
Low	14.29	85.71	100			

Table 8: Summary result for testing sample

CATEGORY	Predicted group Membership		Total	Accuracy (%)	Sensitivity (%)	Specifity (%)
	High	Low				
Original count	3	1	4	75	75.7	60
High	2	6	8			
Low						
High	75	25	100			
Low	25	75	100			

Comparison of the performance of three techniques

By considering the tables in for the summary of each technique, it is clearly seen that the Classification Tree perform better than the other techniques (i.e. K-nearest neighbor and neural network) when compare their percentage of accuracy and correctness Classification in both training and testing sample.

Table 9: Summary of the techniques performance

Classification Techniques	Accuracy (%)		Sensitivity (%)		Specifity (%)	
	Training	Testing	Training	Testing	Training	Testing
CT	80	75	80	75.5	80	60
NNET	0	58.33	0	28.57	0	100
KNN	41.67	88	42.86	86.67	40	90

CONCLUSION

The research performed here demonstrates a comparative evaluation of three nonparametric statistical classification techniques (i.e., K-nearest neighbor, classification tree, and neural network) in pattern recognition and classification of the Nigerian States in terms of high and low crime rates. The performance evaluation was based on the correct classification rates for each method. Overall, the classification tree performed the best, with the lowest error rate and a larger sensitivity rate. The consideration for everyone in Nigeria is safety, and the state has a high crime rate. Therefore, with the help of machine learning techniques, a method was created that helps in classifying the states as having high and low crime rates. It is shown that certain variables, such as the poverty rate, the number of primary school teachers, and the number of offence arrests by police, among others, are suitable for analysis to classify a state as high or low crime and help in reducing the rate of crime in a state. Thus, this study illustrates the usefulness of machine learning for the classification, prediction of high crime rates, and identification of factors that influence crime rates. The analysis performed has recommended that data subjected to classification and pattern recognition should be large and not partitioned as much as possible in order to access the goodness of their accuracy prior to analysis. It is also recommended that the government double its efforts to reduce the level of poverty, insecurity, unemployment, and literacy through better funding and actualization of small-scale industries and invest in agriculture and education as well. Secondly, the government should employ more police personnel, as this would first provide more hands-on experience in crime management and secondly reduce unemployment. For effective crime prevention and control, reducing high crime, and promoting security through effective law enforcement, intelligence gathering, and analysis, policy criminal statistics need to be useful. In view of this, Nigerian police policymakers should seriously consider the development of non-uniform and non-police personnel who are professionals in criminology and social statistics to handle the data and suggest lines of research and planning. With at least one such qualified professional in each state police command and about five in the Abuja headquarters, the problem of crime statistics should be solved.

REFERENCES

Abubakar, U., Abubakar, A., Sulaiman, A., Ringim, H. I., Salisu, I. A., Osi, A. A., James, I., Sani, A. M., & Haruna, I. S. (2023). APPLICATION OF ARTIFICIAL NEURAL NETWORK FOR PREDICTING HYPERTENSION STATUS AND INDICATORS IN HADEJIA METROPOLITAN. *FUDMA JOURNAL OF SCIENCES*, *7*(1), 284–289. https://doi.org/10.33003/fjs-2023-0701-2052

Albanese, J. (2014). *Organized crime in our times*. Routledge.

Bekesiene, S., Smaliukiene, R., & Vaicaitiene, R. (2021). Using artificial neural networks in predicting the level of stress among military conscripts. *Mathematics*, *9*(6). https://doi.org/10.3390/math9060626

Brody, R. G., Kern, S., & Ogunade, K. (2022). An insider's look at the rise of Nigerian 419 scams. *Journal of Financial Crime*, *29*(1), 202-214.

Chambliss, W. J. (2011). *Crime and criminal behavior* (Vol. 1). Sage.

Classen, A., & Scarborough, C. (Eds.). (2012). *Crime and punishment in the Middle Ages and Early Modern Age: Mental-historical investigations of basic human problems and social responses* (Vol. 11). Walter de Gruyter.

Dikko A, OSI AA. (2014). Discriminate analysis as an aid to the Classification and prediction of safety across State of Nigeria. *International journal of statistics and application* , 4(3):153-160.

Farmer, L. (2022). The 'market'in criminal law theory. *The Modern Law Review*, *85*(2), 435-460.

Forbes, A. D. (1995). Classification-algorithm evaluation: Five performance measures based onconfusion matrices. *Journal of Clinical Monitoring*, *11*, 189-206.

Hoffmann, J. L., & Stuntz, W. J. (2021). *Defining Crimes*. Aspen Publishing.

Hooton E. (1939). *the American criminals.* cambridge: Harvad university press.

Kennedy, J. (2021). Crimes as public wrongs. *Legal Theory*, *27*(4), 253-284.

Lewis, R. J. (2000, May). An introduction to classification and regression tree (CART) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California* (Vol. 14). San Francisco, CA, USA: Department of Emergency Medicine Harbor-UCLA Medical Center Torrance.

Logan, W. A., & Ferguson, A. G. (2016). Policing criminal justice data. *Minn. L. Rev.*, *101*, 541.

Lucic, B., Batista, J., Bojovic, V., Lovric, M., Krzic, A. S., Beslo, D., & Nadramija, D. (2019). Estimation of Random Accuracy and its Use in Validation of Predictive Quality of Classification Models within Predictive Challenges. *Croatica Chemica Acta*, *92*(3), 1I-1I.

MacDonald, Z. (2002). Official crime statistics: their use and interpretation. *The Economic Journal*, *112*(477), F85-F106.

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia*, *4*(2), 1883.

Sandri, M., & Zuccolotto, P. (2010). Analysis and correction of bias in total decrease in node impurity measures for tree-based algorithms. *Statistics and Computing*, *20*, 393-407.

Song, Y. Y., & Ying, L. U. (2015). Decision tree methods: applications for classification and prediction. *Shanghai archives of psychiatry*, *27*(2), 130.

APPENDIX

Output for R package

CAT OUTPUT FOR TESTING SAMPLE

```
Confusion Matrix and Statistics
          Reference
Prediction 0 1
         0 6 2
         1 1 3
               Accuracy : 0.75
                 95% CI : (0.4281, 0.9451)
    No Information Rate : 0.5833
    P-Value [Acc > NIR] : 0.1916
                  Kappa : 0.4706
 Mcnemar's Test P-Value : 1.0000
            Sensitivity : 0.8571
            Specificity : 0.6000
         Pos Pred Value : 0.7500
         Neg Pred Value : 0.7500
             Prevalence : 0.5833
         Detection Rate : 0.5000
   Detection Prevalence : 0.6667
      Balanced Accuracy : 0.7286

       'Positive' Class : 0
```

CAT OUTPUT FOR TRAINING SAMPLE

```
Confusion Matrix and Statistics
          Reference
Prediction  0  1
         0 12  2
         1  3  8
               Accuracy : 0.8
                 95% CI : (0.593, 0.9317)
    No Information Rate : 0.6
```

```
    P-Value [Acc > NIR] : 0.02936
                  Kappa : 0.5902
Mcnemar's Test P-Value : 1.00000
            Sensitivity : 0.8000
            Specificity : 0.8000
         Pos Pred Value : 0.8571
         Neg Pred Value : 0.7273
             Prevalence : 0.6000
         Detection Rate : 0.4800
   Detection Prevalence : 0.5600
      Balanced Accuracy : 0.8000

       'Positive' Class : 0
```

NUERL NETWORK OUTPUT FOR TRAINING SAMPLE

```
[1] "Nueral Network"
Confusion Matrix and Statistics
          Reference
Prediction  0  1
         0  0 10
         1 15  0
               Accuracy : 0
                 95% CI : (0, 0.1372)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 1.0000
                  Kappa : -0.9231
 Mcnemar's Test P-Value : 0.4237
            Sensitivity : 0.0
            Specificity : 0.0
         Pos Pred Value : 0.0
         Neg Pred Value : 0.0
             Prevalence : 0.6
         Detection Rate : 0.0
   Detection Prevalence : 0.4
      Balanced Accuracy : 0.0

       'Positive' Class : 0
```

NUERL NETWORK OUTPUT FOR TESTING SAMPLE

```
Confusion Matrix and Statistics
          Reference
Prediction 0 1
         0 2 0
         1 5 5
               Accuracy : 0.5833
                 95% CI : (0.2767, 0.8483)
    No Information Rate : 0.5833
    P-Value [Acc > NIR] : 0.62023
                  Kappa : 0.25
 Mcnemar's Test P-Value : 0.07364
            Sensitivity : 0.2857
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.5000
             Prevalence : 0.5833
         Detection Rate : 0.1667
   Detection Prevalence : 0.1667
      Balanced Accuracy : 0.6429

       'Positive' Class : 0


pred2 0 1
    0 2 0
    1 5 5
[1] 0.4166667
```

KNN OUTPUT FOR TESTING SAMPLE

```
[1] "KNN"
k-Nearest Neighbors
25 samples
```

```
10 predictors
 2 classes: 'No', 'Yes'
Pre-processing: centered (10), scaled (10)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 23, 23, 23, 22, 23, 22, ...
Resampling results across tuning parameters:
  k   ROC        Sens       Spec
   1  0.5416667  0.6500000  0.4333333
   2  0.7583333  0.7166667  0.5000000
   3  0.7916667  0.7166667  0.6333333
   4  0.8000000  0.6833333  0.6333333
   5  0.8750000  0.7000000  0.9000000
   6  0.8750000  0.7500000  0.8000000
   7  0.8416667  0.7166667  0.7000000
   8  0.7500000  0.7166667  0.5666667
   9  0.7583333  0.7333333  0.6333333
  10  0.7750000  0.7666667  0.4666667
  11  0.7833333  0.8500000  0.6000000
  12  0.8250000  0.8333333  0.3333333
  13  0.8500000  0.8500000  0.3333333
  14  0.8333333  0.9500000  0.3000000
  15  0.8166667  1.0000000  0.2000000
  16  0.7416667  1.0000000  0.1333333
  17  0.7750000  1.0000000  0.2000000
  18  0.7166667  1.0000000  0.1000000
  19  0.6333333  1.0000000  0.0000000
  20  0.6000000  1.0000000  0.0000000
  21  0.6000000  1.0000000  0.0000000
  22  0.5500000  1.0000000  0.0000000
  23  0.5000000  1.0000000  0.0000000
  24  0.5000000  1.0000000  0.0000000
  25  0.5000000  1.0000000  0.0000000

ROC was used to select the optimal model using the largest value.
The final value used for the model was k = 13.
ROC curve variable importance
      Importance
PARST     100.00
DARST      97.01
PSE        79.10
NDP        71.64
DENS       59.70
UEMP       50.75
LTR        35.82
NPT        35.82
IGR        29.85
PVT         0.00


Confusion Matrix and Statistics
          Reference
Prediction No Yes
       No   3   3
       Yes  4   2
               Accuracy : 0.4167
                 95% CI : (0.1517, 0.7233)
    No Information Rate : 0.5833
    P-Value [Acc > NIR] : 0.9274
                  Kappa : -0.1667
 Mcnemar's Test P-Value : 1.0000
            Sensitivity : 0.4286
            Specificity : 0.4000
         Pos Pred Value : 0.5000
         Neg Pred Value : 0.3333
             Prevalence : 0.5833
         Detection Rate : 0.2500
   Detection Prevalence : 0.5000
      Balanced Accuracy : 0.4143
```

```
       'Positive' Class : No
KNN OUTPUT FOR TRAINING SAMPLE
Confusion Matrix and Statistics
          Reference
Prediction No Yes
       No  13   1
       Yes  2   9
               Accuracy : 0.88
                 95% CI : (0.6878, 0.9745)
    No Information Rate : 0.6
    P-Value [Acc > NIR] : 0.002367
                  Kappa : 0.7541
 Mcnemar's Test P-Value : 1.000000
            Sensitivity : 0.8667
            Specificity : 0.9000
         Pos Pred Value : 0.9286
         Neg Pred Value : 0.8182
             Prevalence : 0.6000
         Detection Rate : 0.5200
   Detection Prevalence : 0.5600
      Balanced Accuracy : 0.8833
       'Positive' Class : No
```