ARID ZONE JOURNAL OF ENGINEERING, TECHNOLOGY & ENVIRONMENT
AZOJETE June 2023. Vol. 19(2):331-340
Published by the Faculty of Engineering, University of Maiduguri, Maiduguri, Nigeria.
Print ISSN: 1596-2490, Electronic ISSN: 2545-5818
www.azojete.com.ng

ORIGINAL RESEARCH ARTICLE

PREDICTION OF INFANT SIZE AT BIRTH USING PRINCIPAL COMPONENTS BASED ARTIFICIAL NEURAL NETWORK

U. Usman[1], S. Suleiman[2]*, A. B. Muhammad[3] and M. B. Ibrahim[4]

[1]Department of Statistics, Usmanu Danfodiyo University, Sokoto, Nigeria
[2]Department of Statistics, Federal University, Dutsin-Ma, Katsina State, Nigeria
[3]Department of Computer Science, Usmanu Danfodiyo University, Sokoto, Nigeria
[4]Department of Statistics, Usmanu Danfodiyo University, Sokoto, Nigeria
**Corresponding author's email address: macostamkd@gmail.com*

ARTICLE INFORMATION

Submitted 28 February, 2023
Revised 5 April, 2023
Accepted 6 April, 2023

Keywords:
Factor analysis
Artificial Neural Network
Prediction
Infant
Multicollinearity

ABSTRACT

Birth size plays a major role in child's mental development, future physical growth, and survival. The average length of full-term babies at birth is 20 in. (51 cm), although the normal range is 46 cm (18 in.) to 56 cm (22 in.). In the first month, babies typically grow 4 cm (1.5 in.) to 5 cm (2 in.). Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. In neural networks multicollinearity may increase difficulties in data processing; too much feature inputs will also cause the time-consuming training process, prevent the constringency of the training work, and may eventually affect the network performance. This research investigates the effect of some parents' socioeconomic status on the sizes of infants at birth in Zamfara State. Several factors have been identified to have effect on the sizes of infant at birth which is part of major public health challenges in developing countries like Nigeria. Previous studies considered only some maternal factors responsible for the infant sizes at birth. This current study explored the parents' socioeconomic Status and environmental factors that contributes to the status of the infant sizes at birth using factor and neural network analysis. The analysis was carried out using data collected from Federal Medical Centre, Gusau and Ahmad Sani Yariman Bakura Specialist Hospital Gusau, Zamfara State to explore the association between the afore mentioned factors. Multicollinearity results indicated that there is significant correlations among independent variables, factor analysis through principal extraction methods reduced the original 14 correlated variables in to five uncorrelated factors. However, artificial neural network performs better in the presence of multicollinearity as it produces lower mean square error (MSE). It also had a better classification accuracy of 83.5% as compared to 65.9% for Gender of the baby .

1.0 Introduction

Birth size is one of the important predictors of a child's mental development, future physical growth, and survival. It is one of the major risk factors for child morbidity and mortality (Onubogu et al., 2017). As reported by the World Health Organization (WHO), low birth weight (LBW) is defined as an infant birth weight of less than 2,500g (WHO, 2011). Children in this category are considered to have a greater risk of neonatal, post-neonatal death, and morbidity (Dahlui et al., 2016). Children born underweight also tend to have cognitive disabilities and a lower intelligent quotient (IQ), affecting their performance in school and their job opportunities as adults. Previous studies have also linked infant mortality with parents' socioeconomic status (UNICEF, 2004). More than 20 million infants worldwide, representing 15.5 per cent of all births are born with low birth weight, 95.6 percent of them in developing countries (UNICEF, 2004). The level of low birth weight in developing countries (16.5 percent) is more than double the level in developed regions,7 percent (UNICEF, 2004). Birth weight is a strong indicator not only of a birth mother's health and nutritional status but also a newborn's chances for survival, growth, long-term health and psychosocial development (UNICEF, 2008). A low birth weight (less than 2,500 grams) raises grave health risks for children (UNICEF, 2008). Babies who are undernourished in the womb face a greatly increased risk of dying during their early months and years (UNICEF, 2008). The effect of low birth weight on infant mortality is not only additive but also interactive. The magnitude of the contribution of low birth weight to infant mortality is higher in developing countries given that the survival of such infants is dependent on environmental sanitation, effective post-natal nutrition and rehabilitation, and the availability of medical care. Low birth weight is a public health problem all over the world and is connected with a lot of health challenges and even deaths (Isiugo-Abanihe and Oke, 2011). Fetal growth is usually evaluated using birth weight. However, the use of other measurements like the head and length, arm and chest circumferences may be very important in the prediction of long-term health and development results (Neggers et al., 1995).

Artificial neural networks (ANNs) is a machine learning technique widely used for forecasting model in numerous disciplines which include economic, finance, business, engineering, science, health, foreign exchange among others (Suleiman and Sani, 2020). ANNs are powerful tools for modeling. They can learn and identify correlated patterns between input datasets and corresponding target values. Once trained, ANNs can be used to predict the outcome of new independent input data (Suleiman et al.,2019). The ANNs learn these approximate relationships on the basis of actual inputs and outputs. Therefore, they are generally more accurate as compared to the relationships based on assumptions (Suleiman et al., 2016a; Suleiman et al., 2016b). Neural networks are semi-parametric non-linear models, which are able to approximate any reasonable function (Gulumbe et al., 2016). Neural Network is a non-parametric method that can be used in the medical field to classify subjects based on input variables into sick or healthy subjects. Classification and prediction of the patient's condition based on risk factors are an application of artificial neural networks (Gulumbe et al., 2019)

Principal Component Analysis (PCA) is a dimension-reduction tool that can be used to reduce a large set of variables to a small set that still contains most of the information in the large set. It is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal components. PCA is one of the most popular methods used for variable reduction, which can overcome the disturbance of the multicollinearity of the risk factors and has been used in social sciences, health service, and health sciences (Suleiman and Badamasi, 2019). The aim of this paper was to investigate parent's socio-economic status influence on infants' birth sizes (Weight, Height, Head circumference, Gender of the baby and Gestational weeks) in Gusau, Zamfara State.

Taiwo *et al.*, (2018) investigated the association between birth measurements and maternal factors in predicting the birth size of infants in Osogbo, Osun State using canonical correlation analysis (CCA) and multiple linear regressions (MLR). The study shows that, maternal factors could be used predict the eventual birth weight of an infant. In this research, five infants' anthropometric indicators and five parents' socio economic indicators were respectively used as dependent and independent variables. The result revealed that Most of the women were educated but almost half of them fall into the category with low living standard index (47%). Findings in this study suggest CCA to be a better method for assessing the effects of the maternal factors on infant size at birth than usual multivariate techniques like multiple linear regressions.

Affi et al. (2019) used canonical correlation to examine the relationship between neonatal anthropometric indicators and maternal socio-demographic factors. The neonatal anthropometric indicators were considered as dependent (Y) variables whiles maternal socio-demographic factors were considered as independent variables (X). The outcome of the canonical correlation analysis showed that the total shared variance between the two set of variables stood at 69.8%, which indicates a significant relationship between neonatal anthropometric indicators and maternal socio-demographic factors.

1. Methodology

In this study, secondary source of data was employed, pregnant women socio-economic records gathered from Federal Medical Gusau were analyzed using multivariate techniques of Factor analysis based on principal components extraction method and ANN. ANN was built based on the complete data and factor analysis transformed data.

1.1 Data Collection

Data on Parents' socio-economic status and infants' anthropometric indicators consisting of 546 pregnant women who delivered at Federal Medical Centre, Gusau and Ahmad Sani Yariman Bakura Specialist Hospital, Gusau Zamfara State between January 2017 and December 2018 were considered in this research. Three hundred and eighty two (382) pregnant women records constituting 70% of cases were randomly selected for training the proposed neural network model and the 30% (164) pregnant women records were utilized for testing the model. Fourteen (14) independent variables used in this research were Maternal Age (MA),Body Mass Index (BMI), Gravida (Gr), Nutrition Habits (NH), mother's education (MsE), father's education (FsE), antenatal care visits (ANC), gestational age (GA), maternal weight gain in pregnancy (MWGP),mother's working status (MsWS),father's income level (FIL),mode of delivery (MD), Birth order (BO), Residence (R) and mother's supplement usage(MsSU). Outcome variables are the birth sizes (weight, height, head circumference and mid-upper arm circumference) of the baby were used as dependent variables. Table 1 describes how the independent variables were measured. It has indicated that all 14 independent variables were used as categorical.

Table 1: Definitions of independent variables.

S/N	Variable	Category 1	Category 2	Category 3	Category 4
1	Maternal age (MA),	16-25yrs=1	26-35yrs=2	35yr+=3	
2	Body Mass Index (BMI),	<18.5=1	18.5-25.0=2	>25.0-30=3	>30=4
3	Gravida (Gr),	1	2	3	4
4	Nutrition habits (NH)	Regular =1	Irregular =2	-	-
5	Mother's education (MsE),	No Education=1	Primary Education=2	Secondary Education=3	High School Education=4
6	Father's education (FsE)	No Education=1	Primary =2	Secondary Education =3	High School Education =4
7	Antenatal care visits (ANCV),	0-2=1	3-5=2	6-8=3	>8=4

8	Maternal weight gain in pregnancy (MWGP)	5-11= 1	>11 = 2	-	-
9	Mother's working status (MsWS),	Working = 1	Not working =2	-	-
10	Father's income level (FIL),	<N 18,000 = 1	>N 18,000 = 2	-	-
11	Mode of delivery (MD)	Normal =1	C-section=2	-	-
12	Mother's supplement usage (MsSU).	Yes = 1	No = 2	-	-
13	Birth order (BO)	1	2-3	4-5	6+
14	Residence	Rural = 1	Urban = 2	-	-

2.2 Variance Inflation Factor

A variance inflation factor is a tool to help identify the degree of multicollinearity. Canonical Multicollinearity exists when there is a linear relationship, or correlation, between one and more of the independent variables or inputs. Multicollinearity creates a problem in the accuracy of traditional model because all the inputs influenced each other. Therefore, they are not actually independent, and it is difficult to test how much the combination of the independent variables affects the dependent variable, or outcome, within the model. While multicollinearity does not reduce a model's overall predictive power, it can produce estimates of the regression coefficients that are not statistically significant (Suleiman & Badamasi, 2019). In a sense, it can be thought of as a kind of double-counting in the model.

Variance inflation factor is calculated using the following formula:

$$VIF_i = \frac{1}{1-R_i^2} \qquad (1)$$

where: R^2_i is the coefficient of determination of the regression equation.

2.3 Mathematical Computation of Factor Analysis

Multiple factors are based on the premise that a large number of questionnaire items could be reduced to only a few dimensions. The multiple factor approach emphasized the goal of extracting a maximum amount of variance from a correlation. The factor analysis describes the covariance relationship among many variables in terms of few underlying, but unobservable, random quantities called factors. Factors analysis believes that the variables can be grouped by their correlations (Suleiman et al., 2019).

Consider the general factor model

$$X_{p\times 1} = \mu_{p\times 1} + L_{p\times m} f_{p\times 1} + \sum\nolimits_{p\times 1} \qquad (2)$$

$X_{p\times 1}$ is the vector of the p correlated variables

$\mu_{p\times 1}$ is the mean vector of the p correlated variables

$\sum_{p\times 1}$ is the variance covariance matrix of the p correlated variables.

The regression coefficient (the partial slopes) for all of these multiple regressions are called i^{th}variable and the j^{th}factor, which could be collected in a matrix as shown below

$$L = \begin{pmatrix} l_{11} & \cdots & l_{15} \\ \vdots & \ddots & \vdots \\ l_{141} & \cdots & l_{14\times 5} \end{pmatrix} \qquad (3)$$

2.4 The Neural Network Model

A neural network is a simplified model of how human brain processes information. It works by simulating a large number of interconnected processing units that resemble abstract versions of neurons (Suleiman and Gulumbe, 2018).

The processing units are arranged in layers. There are typically three parts in a neural network: an **input layer**, with units representing the input fields; one or more **hidden layers**; and an **output layer**, with a unit or units representing the target field(s). The units are connected with varying connection strengths (or **weights**). Input data are presented to the first layer, and values are propagated from each neuron to every neuron in the next layer. Eventually, a result is delivered from the output layer (Ding et al., 2010).

The network learns by examining individual records, generating a prediction for each record, and making adjustments to the weights whenever it makes an incorrect prediction. This process is repeated many times, and the network continues to improve its predictions until one or more of the stopping criteria have been met (Ding et al., 2010).

Initially, all weights are random, and the answers that come out of the net are probably not closer to the correct known values. The network learns through **training**. , and the answers it gives are compared to the known outcomes. Information from this comparison is passed back through the network, gradually changing the weights. As training progresses, the network becomes increasingly accurate in replicating the known outcomes. Once trained, the network can be applied to future cases where the outcome is unknown (Ding et al., 2010).

Neural networks are simple models of the way the nervous system operates. The basic units are **neurons**, which are typically organized into **layers**, as shown in the following figure 1.

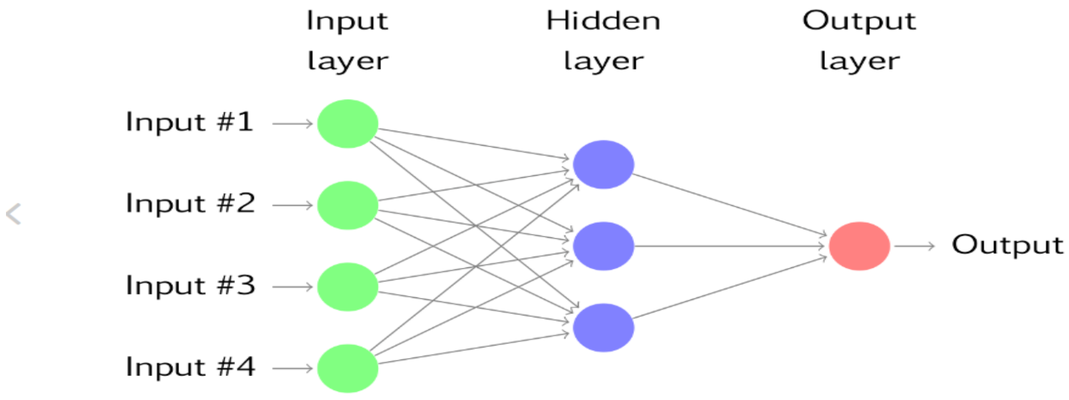

Figure 1: A neural network with four inputs and one hidden layer with three hidden neurons (Ding et al., 2010)

This is known as a *multilayer feed-forward network*, where each layer of nodes receives inputs from the previous layers. The outputs of the nodes in one layer are inputs to the next layer. The inputs to each node are combined using a weighted linear combination. The result is then modified by a nonlinear function before being output (Ding et al., 2010). For example, the inputs into hidden neuron in Figure 1 above are combined linearly to give

$$\boldsymbol{Z_j = b_j + \sum_{i=1}^{4} w_{ij} X_i} \qquad (3)$$

Z_j are the neural network outputs

b_j are the neural network biases or intercepts

w_{ij} are the neural network weights

X_i are the neural network inputs

The parameters b1, b2, b3 and w1, 1, …, w4, 3 are “learned” from the data. The values of the weights are often restricted to prevent them from becoming too large. The parameter that restricts the weights is known as the “decay parameter,” and is often set to be equal to 0.1.

2.5 Factor Analysis Artificial neural Network (FA-ANN) algorithm

Factor analysis (FA) was used to reduce the dimensionality of the original data. Then these low-dimensional data was considered as the input of the ANNs, which gives way for the new algorithm based on FA referred to as FA-BP algorithm (Ding et al., 2010). The FA-BP algorithm is summarized in figure 2 below: Records of 546 pregnant women were collected from FMC Gusau, 70% (382) were utilized for training the network and 30 % (164) were used for testing the network.

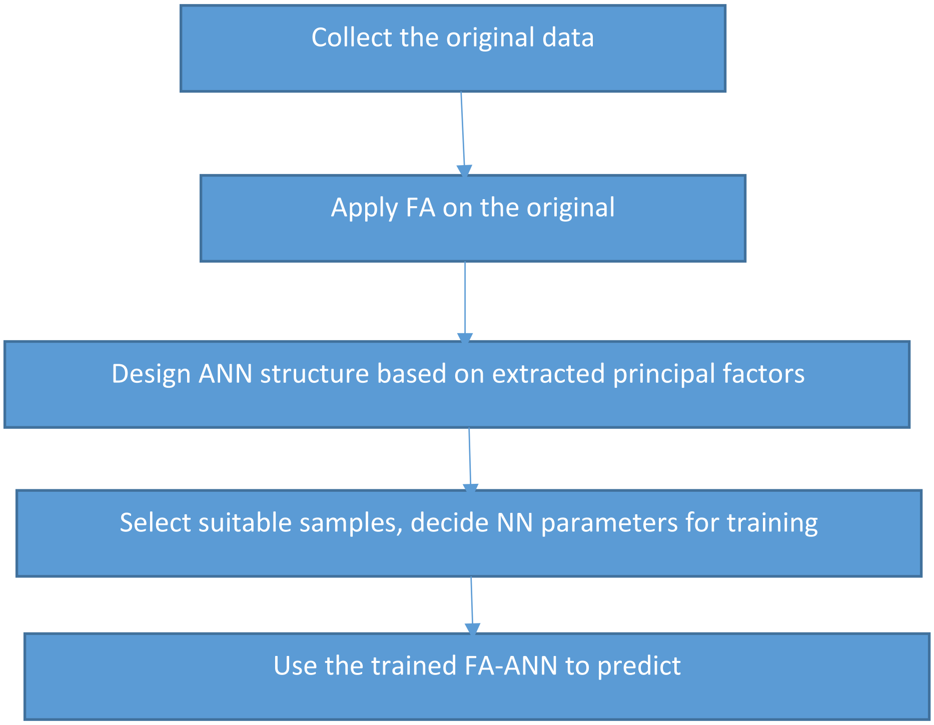
Collect the original data
Apply FA on the original
Design ANN structure based on extracted principal factors
Select suitable samples, decide NN parameters for training
Use the trained FA-ANN to predict

Figure 2: FA-ANN algorithm

3. Results and Discussion

Table 1 indicates that there are moderate correlations among the variables since the variance inflation factor for all the variables ranges from 1 to 5. Multicollinearity existing among the variables was removed using the factor analysis through principal components extraction technique.

Table 1: Collinearity Statistics

Variables	Collinearity statistics	
	Tolerance	Variance inflation factor
Maternal age	0.592	1.689
Mothers Education	0.534	1.874
Fathers Income	0.886	1.129
ANCV	0.826	1.211
MSSU	0.524	1.910
Gravida	0.399	2.505
Birth Order	0.845	1.184
Weight Gain	0.591	1.693
Body Mass Index	0.342	2.923
Fathers Education	0.983	1.018
MSWS	0.972	1.029
Residence	0.846	1.183
MoD	0.918	1.089
Nutritional habit	0.976	1.025

Table 2 presents Kaiser Meyer Olkin (KMO) and Bartlett's test conducted to measure the appropriateness of the data for factor analysis. The result of KMO indicates that sampling is adequate since the coefficient is approximately 0.5. Bartlett's test also indicated that the correlations among the variables is significant. Thus, Factor analysis is suitable for the data reduction.

Table 2: Kaiser Meyer Olkin (KMO) and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.467
Bartlett's Test of Sphericity	Approx. Chi-Square	1338.635
	Df	91
	Sig.	.000

Table 3 shows the eigenvalues and total variance explained. Principal component analysis extraction method for factor analysis was employed in this study. Before extraction, the data consists of fourteen linear components. After extraction and rotation, the data was reduced to five distinct linear components with eigenvalue > 1. The five factors are extracted accounted for a combined 56% of the total variance. It is suggested that the proportion of the total variance explained by the retained factors should be at least 50%. The result shows that 56% common variance shared by eleven variables can be accounted by five factors.

Table 3: Eigenvalues and total variance explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.417	17.263	17.263	2.417	17.263	17.263
2	1.831	13.080	30.343	1.831	13.080	30.343
3	1.386	9.903	40.246	1.386	9.903	40.246
4	1.102	7.869	48.116	1.102	7.869	48.116
5	1.078	7.697	55.813	1.078	7.697	55.813
6	.983	7.023	62.836			
7	.950	6.783	69.619			
8	.892	6.372	75.991			
9	.832	5.944	81.935			
10	.782	5.582	87.517			
11	.696	4.971	92.488			
12	.558	3.988	96.476			
13	.326	2.329	98.805			
14	.167	1.195	100.000			

Tables 4 and 5 present results for the original model ANN model and transformed FA-ANN model. ANN model gives smaller prediction errors for all the scale dependent variables. Similarly, ANN model gives higher prediction percent for the categorical variable gender of the baby. These results are consistent with the results of (Suleiman and Badamasi, 2019; Chan et al., 2022; Obite et al., 2020) but contradicts the results of Ding et al., (2010).

Table 4: Neural Network Root Mean Square Error (RMSE) for the continuous dependent variables

Dependent Variables	ANN	FA-ANN
Height	.043	.676
Weight	.247	.576
HCFR	.016	.452
GWeeks	.233	.335

Table 5: Neutral Network Classification performance for Gender of Baby

Model	Observed	Predicted		
		Male	Female	Percent correct
ANN	Male	90	10	90.0%
	Female	17	47	85.8%
	Overall percent	65.2%	34.8%	83.5%
FA-ANN	Male	67	40	62.6%
	Female	16	41	71.9%
	Overall percent	50.6%	49.4%	65.9%

4. Conclusion

This study investigated the effect parent's socio-economic status influence on infants' birth sizes (Weight, Height, Head circumference, Gender of the baby and Gestational weeks). Variance inflation factor was used to establish the presence of multicollinearity among parents' socio-economic status influence. Factor analysis was found to be appropriate approach to reduce the data set in to uncorrelated factors using principal components extraction method. Principal component analysis of parents' socio-economic status affecting birth sizes reveals 56% of the variation account by the first five principal components. The ANN model had a smaller RMSE when the original 14 correlated variables were used. Therefore, this indicated that ANN performance is not affected by the presence of multicollinearity. It is, therefore, recommended that ANN should be used when there is multicollinearity in the data set and in further research variants of back propagation algorithms can be used to train the model.

Reference

Affi, OP., Lartey, NL., Angkyiire, D. and Oppong I. 2019. Canonical Correlation Analysis of Neonatal Anthropometric Indicators and Maternal Socio -Demographic Factors. International Journal of Scientific & Technology Research, 8(7): 93-97.

Chan, JYL., Leow, SMH., Bea, KT., Cheng, WK., Phoong, SW., Hong, ZW. and Chen, YL. 2022. Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. Mathematics, 10, 1283. https://doi.org/10.3390/math10081283.

Dahlui, M., Azahar, N., Oche, OM. and Aziz, NA. 2016. Risk factors for low birth weight in Nigeria: evidence from the 2013 Nigeria Demographic and Health Survey. Global Health Action, 9(1): 34-39. doi:10.3402/gha.v9.28822.

Ding, S., Jia, W., Xu, X. and Zhou, H. 2010. Neural Networks Algorithm Based on Factor Analysis. Conference: Advances in Neural Networks. Proceedings of 7th International Symposium on Neural Networks, June 6-9, Shanghai, China, Part I, 319-324.

Gulumbe, SU., Suleiman, S., Asare, BK. and Abubakar, M. 2016. Forecasting Volatility of Nigerian Crude Price Using Non-Linear Auto-Regressive with Exogenous (NARX) Inputs Model. Imperial Journal of Interdisciplinary Research (IJIR), 2(5): 434-442.

Gulumbe, SU., Suleiman, S., Badamasi S., Yusuf, AT. and Usman, U. 2019. Predicting Diabetes Mellitus Using Artificial Neural Network Through a Simulation Study. Machine Learning Research, 4(2):33-38. doi: 10.11648/j.mlr.20190402.12

Isiugo-Abanihe, U. and Oke, O. 2011. Maternal and environmental factors influencing infant birth weight in Ibadan, Nigeria. African Population Studies, 25(2): 11-19.

Obite, CP., Olewuezi, NP., Ugwuanyim, GU. and Bartholomew, DC. 2020. Multicollinearity Effect in Regression Analysis: A Feed Forward Artificial Neural Network Approach. Asian Journal of Probability and Statistics, 6(1):22-33. DOI: 10.9734/AJPAS/2020/v6i130151

Onubogu, CU., Egbuonu, I., Ugochukwu, EF., Nwabueze, AS. and Ugochukwu, O. 2017. The influence of maternal anthropometric characteristics on the birth size of term singleton South-East Nigerian newborn infants. Nigerian Journal of Clinical Practice, 20: 852.

Suleiman, S. and Badamasi, S. 2019. Effect of Multicollinearity in Predicting Diabetes Mellitus Using Statistical Neural Network. European Journal of Advances in Engineering and Technology, 6(6):30-38.

Suleiman, S. and Gulumbe, SU. 2018. Application of Back Propagation Gradient Based Neural Network Training Algorithms to Volatility Forecasting. Edited Proceedings of 2nd International Conference Professional Statisticians Society of Nigeria, 2: 273-282.

Suleiman S., Gulumbe, SU., Asare, BK. and Abubakar, M. 2016a.Comparative Study of Backpropagation Algorithms in Forecasting Volatility of Crude Oil Price in Nigeria. Science Journal of Applied Mathematics and Statistics, 4(3): 88-96, doi: 10.11648/j.sjams.20160403.11

Suleiman, S., Gulumbe, SU., Asare, BK. and Abubakar, M. 2016b.Training Dynamic Neural Networks for Forecasting Naira/Dollar Exchange Returns Volatility in Nigeria. American Journal of Management Science and Engineering, 1(1): 8-14, doi: 10.11648/j.ajmse.20160101.12

Suleiman, S., Lawal, A., Usman, U., Gulumbe, SU. and Muhammad, AB. 2019. Student's Academic Performance Prediction Using Factor Analysis Based Neural Network. International Journal of Data Science and Analysis, 5(4): 61-66.10.11648/j.ijdsa.20190504.12

Suleiman, S. and Sani, M. 2020. Application of ARIMA and Artificial Neural Networks Models for Daily Cumulative Confirmed Covid-19 Prediction in Nigeria. Equity Journal of Science and Technology**,** 7(2):83 – 90**.**

Taiwo, AO., Bolanle, SO. and Kehinde, AB. 2018. Evaluation of Maternal Factors on Infant's Birth Size Using Canonical Correlation Analysis in Osogbo, Osun State. NOUN Journal of Physical and Life Sciences, 2(1): 67-83.

UNICEF, 2004. The State of World's Children 2004. Girls' Education and the Promise of Education for All. 2004. 111.

UNICEF, 2008. The State of World's Children 2008. The Most Comprehensive Analysis of global Trends affecting Children, 106-113.

WHO, 2011. World health statistics 2011. 20 Avenue Appia, 1211 Geneva 27, Switzerland: WHO Press, World Health Organization