Available online at http://scik.org

J. Math. Comput. Sci. 2022, 12:203

https://doi.org/10.28919/jmcs/7451

ISSN: 1927-5307

DATA ANALYSIS OF ORPHANAGE HOMES FUND DISTRIBUTION IN NIGERIA

KAYODE OSHINUBI[1,*], GABRIEL O. ELUMALERO[2], TEMITOPE OMOSEBI[3], OLUWATOSIN OLAOLUWA KOMOLAFE[3], BASIT BOLAJI AFOLABI[4], HAMMED ODIWO[5], GIDEON K. ABEGUNRIN[6]

[1]AGEIS, Laboratory, Université Grenoble Alpes, France.

[2]Department of Crop and Horticultural Science, University of Ibadan, Nigeria

[3]Obafemi Awolowo University, Ile-Ife, Nigeria

[4]Federal University of Technology Akure, Nigeria

[5]Department of Metallurgical and Materials Engineering, Ahmadu Bello University, Zaria, Nigeria

[6]Federal University of Agriculture, Abeokuta, Nigeria

Abstract: Love for humanity is vital in life and with the dire need to cater to the needs of vulnerable children in our society, charity is important. This study analyzed David Adeleke`s donation to orphanages across the states in Nigeria. Data were collated from new media in Nigeria, and it includes states, names of orphanages, the number of children, and funds disbursed in Naira along with the states. The data were subjected to statistical analysis such as Linear regression, Support Vector Machines (SVM), Generalized Additive Model (GAM), and Hierarchical clustering. Results revealed that 8859 children were domiciled in the orphanages chosen and ₦221985548 was distributed. The mean number of children and funds disbursed were 30.4 and 762836.9. From the study, Ebonyi, Adamawa, Kogi, Bauchi, and Kano, Taraba had the lowest numbers of orphanages. Results from the inferential statistics revealed normality test using Jarque-Bera test (p-value < 2.2e-16) while the SVM has (p-value < 2.2e-1) and MAE: 50193 RMSE: 149366. The R-squared for SVM is 0.923 and for Linear regression is 0.9449. GAM and Mixed model have R-squared (adjusted) 0.902, 0.944, and Deviance explained 91.5% and 92.9% respectively. The study concluded that although there was a good spread of the funds, some states were not considered for reasons outside the scope of this study. This study, therefore, recommends intensified action on charity by private individuals to cater to the needs of

*Corresponding author

E-mail address: Kayode.Oshinubi@univ-grenoble-alpes.fr

Received April 25, 2022

the vulnerable in the society, and databanks such as the list of orphanages used in this study can be used to make policy decisions.

Keywords: data analysis; hierarchical clustering; generalized additive model; support vector machines; linear regression; orphanage homes.

2010 AMS Subject Classification: 62P10, 62D05.

1. INTRODUCTION

1.1. Background

There is a common expression that children are the future of nations, and as such, investments in children's welfare and education are paramount for the continued development of every society (Esping-Andersen [1]; Hart [2]). However, the standard of their development is dependent on various factors such as the familial, economic, educational, and social background. WHO [3] discusses that not all children have access to opportunities and proper experience of the factors highlighted earlier to ensure their continued development. However, these factors in positive realization are necessary factors for children. They facilitate environments that protect them from neglect, exploitation, and abuse (Nyathi [4]). In addition to this, each child develops at a different pace, hence there should also be necessary arrangements that will cater to each individual's development (Shah [5]). The summation of this discussion is that to ensure proper growth, children must have access to basic needs and supportive environments.

Many children are deprived of basic needs due to separation from their parents. Many children are separated from their parents at a tender age due to death, diseases, conflicts, or disasters. This limits their access to food, shelter, and even clothing (Edyburn and Meek [6]). Many of these children become orphans experiencing a very poor standard of living. Although it is a global challenge, Africa is considered to have the highest number of such children after Asia (Linn *et al*., [7]). It is important to bear in mind that according to the Wordnet definition, an orphan was defined as a child who has lost both parents (Wordnet [8])

According to Huynh [9], there are about 140 million orphans in the world. It was also revealed in a study by Nar [10], that about 52 million orphans reside in Africa. Although pathetic, this might not be considered unusual. Africa has experienced various challenges which range from extreme hunger and poverty to different epidemics and pandemic-prone diseases such as HIV (human immunodeficiency virus) (Bennell [11]). These have led to many children being separated from their parents.

Orphanages play critical roles in the development of many children, especially in Africa (Pillay, [12]). Orphanages are, usually, the last resort for children who have lost their parents due to circumstances beyond their control (Ahmad and Rashid [13]). Their importance cannot be overemphasized. Orphanages provide homes and companionships to many orphans, and without them, orphans might face hunger, poor shelter, and their education threatened (Kurniawan [14]). In addition, orphans sometimes face stigmatization, and discrimination, from society. Orphanages that are well managed, reduce the effect of such traumatic events on the children (Ntshuntshe and Taukeni [15]). Proper vocational and educational activities will be arranged by the orphanages while providing a guardian in place of their parents (Sagalaeva, Ivahnenko and Landina [16]).

Globally, orphanages have a long history. Its earliest history dates as far back as 400 AD in Rome. Orphanages gradually spread all over Europe and Great Britain (Nathan, [17]). Nowadays, orphanages have extended to every region of the world including Africa. However, orphanages in Africa, are facing many challenges. Some of such challenges include overpopulation, understaffing, and financial difficulties. These have limited their ability to provide adequate care for orphans (Mwoma [18]).

Many orphanages are being run by financial aid. The orphanage sources money through volunteerism; grants; donations; and sponsors (Lyneham and Facchini [19]). However, orphanages in Nigeria have found it difficult to have access to sustainable funding. Although some are supported by the government, it is noteworthy that many orphanages have to source funding/ sponsorship from religious institutions and influential individuals (Muhammad [20]). The most recent of such dominations in Nigeria that got the attention of social media was the 250million Naira donation by Davido, a Nigerian-based artist whose real name is David Adeleke, to orphanages across Nigeria in February 2022. Interestingly, Davido released several images that detailed how the fund was distributed among many orphanage homes across the states in Nigeria on a social media platform called Twitter. It is the numerical information about this distribution as provided by the files that the current research leverage to show the applicability and value of data analysis through several quantitative models which include linear regression, hierarchical clustering, generalized additive model, and support vector machine.

1.2. Review of Recent Literature

Approaches to academic researches and data analysis have often taken either the quantitative or qualitative methodological approach. According to Daniel [21], the qualitative research method

encompasses utilizing data sets that stem from the textual details obtainable from the opinion and experience of individuals as well as the textual information that can be gathered from other sources such as academic studies and newspapers. This implies that studies that utilize the qualitative research method are largely reliant on pieces of textual information that constitute motifs or rather, themes, that are related to the focus of the studies. It is often a subjective assessment of opinions about specific topics.

There are concerns about the validity, reliability, and even generalizability of the studies that utilize this research method (Ahmad, et al [22]; Asper and Corte [23]) This is because the data sets utilized in most qualitative studies, especially those who leverage on the experiences and opinions of individuals, do not have present findings with quantifiable and verifiable values. In addition to this, Galdas [24] discussed that there is the issue of researcher's biases in qualitative studies. This means that qualitative studies tend to collect and report data sets that reflect the biased stances of researchers. The collective implication of all these shortcomings of the qualitative research method is its punctured value when it comes to implementing empirical research.

However, qualitative research has a major advantage which is helping researchers to explain pertinent human and phenomenal factors that influence the quantitative findings that only report figures. Elliot [25] and Ahmad *et al*., [26] unanimously noted that beyond proving numerical insights into phenomena which is what quantitative studies provide, qualitative studies offer exploratory and explanatory insights that enable a better understanding of the rather numerical findings and/or data presentation.

On the other hand, the quantitative research method involves the collection of numerical data sets and the use of empirical data analysis methods to analyze the data sets (Apuke [27]). Qualitative research methods avail researchers the opportunity to implement empirical researches that provide quantifiable and verifiable findings (Queiros, Faria and Almeida [28]) as well as findings that could be used for inference in other studies, that is, studies with generalizability value. Data collection for quantitative studies could be done via several data collection instruments such as questionnaires, interviews, and numerical data from secondary data sources such as financial statements. Numerous empirical data analysis methods exist and are utilized in quantitative studies. These include but are not limited to descriptive statistics, Pearson's correlation analysis, Granger causality test, Chi-square test, regression analysis, support vector machine (SVM), and hierarchical clustering. Given the focus of the current study, this section will focus on carrying out a

methodology review that is focused on the quantitative research method with a specific focus on the methods of analysis used in the current study.

According to Kumari and Yadav [29], Sir Francis Galton first proposed linear regression in 1894 and this method of analysis is a statistical test utilized for defining a data set and also quantifying the relationship between variables. The authors further itemize and discuss five significances of linear regression. Descriptive, that is linear regression facilitates the analysis of the strength of the association between predictor variables and the outcome; adjustment, means that the method of analysis adjusts the effect of the cofounders or covariates; predictors which help estimate the necessary risk factors that may influence dependent variables; extent of prediction which means that the method analysis helps in analyzing how a change in the independent variable may affect the dependent variables; and lastly, a prediction which refers to the ability of the analysis method to help in quantifying new cases.

Rosenthal [30] provides another discussion with semblance to the discussion of Kumar and Yadav [31] but goes further to discuss that there are three types of linear regression: simple linear regression, multiple linear regression, and hierarchical linear regression. Simple linear regression examines the relationship between an independent and a dependent variable; multiple regression examines the among multiple variables and may also include interaction effects, and hierarchical regression creates theoretically meaningful blocks for independent variables (Rosenthal [32])

But more importantly, are the assumptions that must be met before any form of linear regression is used. These assumptions are five. Firstly, it is assumed that variables that are to be analyzed using linear regression share a linear relationship; secondly, multivariate normality is assumed to be attainable in the data sets to be analyzed; thirdly, in multiple regression analysis, it is assumed that there should be little or no multicollinearity; the fourth assumption is that there is no autocorrelation in linear regression; and lastly, there is an assumption of homoscedasticity in linear regression Meuleman et al [33]. The linear regression method of analysis will enable the current study to check the relationship between the shared fund across the orphanage homes and their populations.

Another analysis model that is adopted in the current study is the support vector machine (SVM). Supervised learning models such as SVM are used to solve classification and regression problems. The models are used to categorize data and understand the relationship between the variables (i.e., both dependent and independent variables) (Ray [34]). However, the SVM algorithm is used to

solve more classification challenges compared to regression analysis (Yue, Li and Hao [35]). This accompanied by the other regression analysis methods enables the study to avail a more robust and comprehensive data analysis. It also facilitates providing a critical comparative discussion on the role of SVM and regression analysis in quantitative research and how both could complement one another to understand empirical phenomena better.

The SVM algorithm developed by Vapnik [36] is based on statistical learning theory and was aimed at creating a hyperplane (N-dimensional) that divides the data into two categories with high precision. Feature selection (a transformed attribute) is done to ensure the most applicable representation is chosen to define the hyperplane that separates the data into clusters. It is important to note that SVMs could capture large feature spaces due to a generalization principle based on the theory of Structural Risk Minimization (SRM) (Eghbalnia [37])

An example of the application of SVM was research carried out by Burbidge *et al*., [38] They revealed that SVM has a high probability to analyze a structure-to-activity relationship. In their study, they compared and contrasted different machine learning algorithms used in solving classification problems, to predict the inhibition of dihydrofolate reductase by pyrimidines. They observed that SVM was significantly better compared to other algorithms except for a manually capacity-controlled neural network which takes a longer time to train.

Clustering analysis is a statistical method of processing and segregating data. This type of analysis is an unsupervised learning algorithm that aims to analyze and cluster unlabeled datasets into similar groups that are different from each other. This led to the discovery of hidden patterns previously undetected in a dataset (Fan *et al*., [39]). Some types of methods used in clustering analysis include the Hierarchical method, Partitioning, Density-based method, Model-based clustering, and Grid-based model (Patel et al, [40]; Rani [41]). (These methods group data points into clusters as described earlier.) This study employed the Hierarchical clustering method as one of its statistical analysis approaches. The Hierarchical clustering approach builds clusters using a top or down approach—similar to a hierarchical system. This clustering method is also subdivided into two methods: Divisive (Top-down approach) and Agglomerate (bottom-up) methods (Bouchefry and Souza [42]).

Although the Hierarchical clustering approach has been proposed and available for a long time (it was first proposed by Ward in 1963[43], Ding and He [44] came up with the merging and splitting process for the approach. They did an extensive analysis of methods (both existing selection and

new methods) that determines the most suitable selection of clusters for split-merge operations was done. It was observed that the best approach for the divisive method was Average Similarity, while the Min-Max Linkage was the best for agglomerative clustering. To effectively assess the quality of clustering, they also introduced an objective function saturation and a clustering target distance (Yogita and Harish, [45]).

In addition, the generalized additive model is also utilized in this study. According to Hastie and Tibshirani [46], real-life situations are not always linear and this necessitates the use of analytical methods that can address nonlinear regressions without prespecifying the type of nonlinear relationship, and this is what the generalized additive model (GAM) is aimed at. Furthermore, Vatter and Chavez-Demoulin [47] discuss that GAMs are a natural extension of linear and generalized models, and the authors also note that GAM is a malleable tool for data analysis in traditionally univariate contexts. It is observable from the discussion of these authors that GAM enables a study to examine the relationship among nonlinear variables and the analysis model is also applicable to various forms of numerical data types or types of nonlinear relationships due to its flexibility. Hence, GAM is adopted in this study to explore to provide a comparative insight as opposed to linear regression on the relationships among the variables in the collected data.

1.3. Organization of the Article

The remainder of this research is divided as follows: In Section 2, we present the descriptive statistics for the data used in the article, in Section 3, we explained the methods used for the analysis, in Section 4, we present results and visualization of our analysis, in Section 5, we discussed the results derived from the analysis and finally, in Section 6, we concluded the article by giving some perspectives and key results of this research.

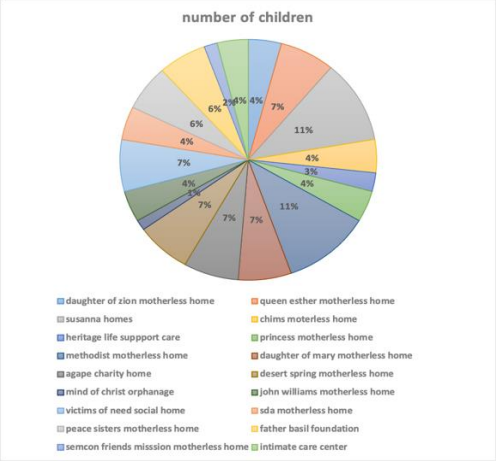
2. DESCRIPTIVE STATISTICS

It was found out from the analysis that the $\bar{x}$ number of children was 30.44 and the $\bar{x}$ of funds disbursed in naira was 762836.9 for all the states. This implies that the average number of children that would benefit from each state would be at least 30.44 and the average number of funds would be around 762836.9. It was also revealed that the standard deviation of the children was 23.07 whilst the standard deviation of the fund was 581983.90. In Figure 1a we present the visualization of the overall descriptive statistics of the data used in this analysis. Also, in Figure 1b we gave a chart of the percentage of children in different orphanage homes in a particular state.

children		*funds*	
Mean	30.44329897	Mean	762836.9347
Standard Error	1.352792946	Standard Error	34116.50096
Median	25	Median	622855
Mode	25	Mode	622855
Standard Deviation	23.07691894	Standard Deviation	581983.9092
Sample Variance	532.5441877	Sample Variance	3.38705E+11
Kurtosis	1.899426069	Kurtosis	1.79597339
Skewness	1.452130937	Skewness	1.437183091
Range	99	Range	2466514
Minimum	1	Minimum	24914
Maximum	100	Maximum	2491428
Sum	8859	Sum	221985548
Count	291	Count	291
Largest(1)	100	Largest(1)	2491428
Smallest(1)	1	Smallest(1)	24914

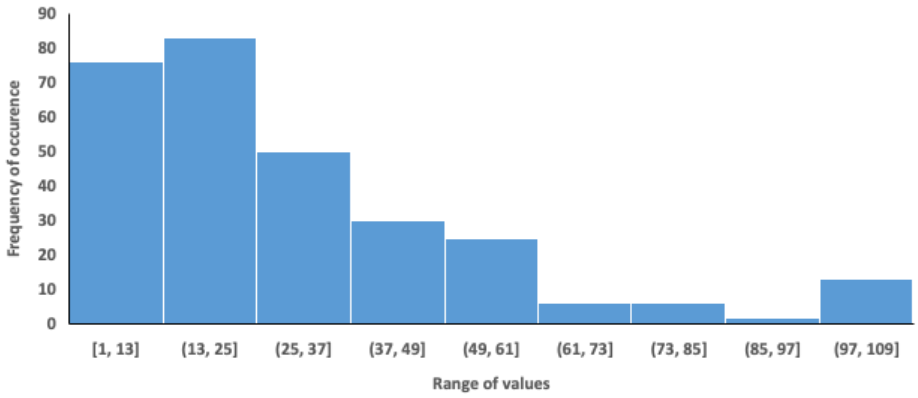
(1a)

Figure 1a: Descriptive statistics



(1b)

Figure 1b: Pie chart of the percentage of children in different orphanage homes in a particular state.



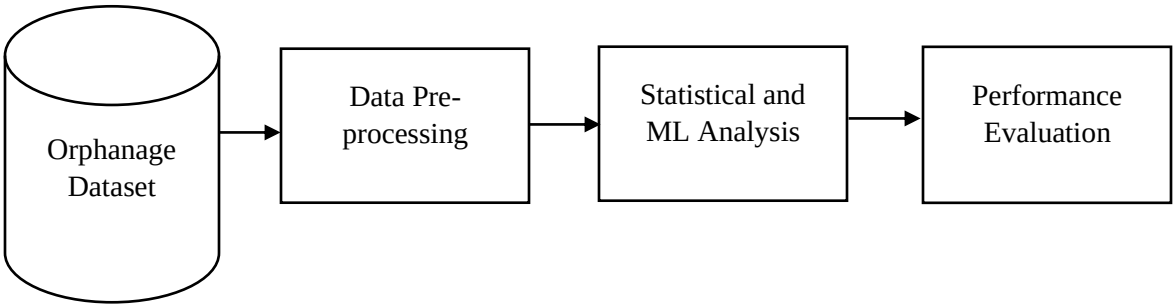
(1c)

Figure 1c: Frequency of the number of children covered as captured in each home.

3. MATERIALS AND METHODS

3.1 Research Methodology

This section covers different phases involved in the orphanage dataset gathering process, pre-processing, statistical and Clustering techniques used, and performance evaluation for this study. The block diagram presented in Figure 2 for the study used statistical and Clustering techniques to assess and identify significant trends in the *Davido* orphanage donation database. This section involves different methodologies used for this study.



3.2 Dataset Gathering

The data used for this study was collected from the 250 million Naira (₦250 million) donated to different Orphanage homes by the famous Nigerian musician David Adeleke, popularly known as Davido, and proceeds donated by the league of friends. The study data includes states, the number of children, name of orphanage homes, date, address of the orphanage homes, and the amounts disbursed. The data were elicited from his timeline (https://twitter.com/davido). The total number of children/orphans that benefitted from this largess stood at 8,859; a total sum of 221,985,548 million Naira was disbursed to these children/orphans in thirty-one states of the Federal Republic of Nigeria. Figure 3 presents the graphical population of the disbursement and the states covered. Lagos state has the highest number of registered orphans with a total number of 1,333 orphans and received the highest fund amounting to 32,089,779 million Naira, while Adamawa state has the lowest number of registered orphans, which stood at 27 children, they also received the lowest fund amounting to 672,683 Naira.

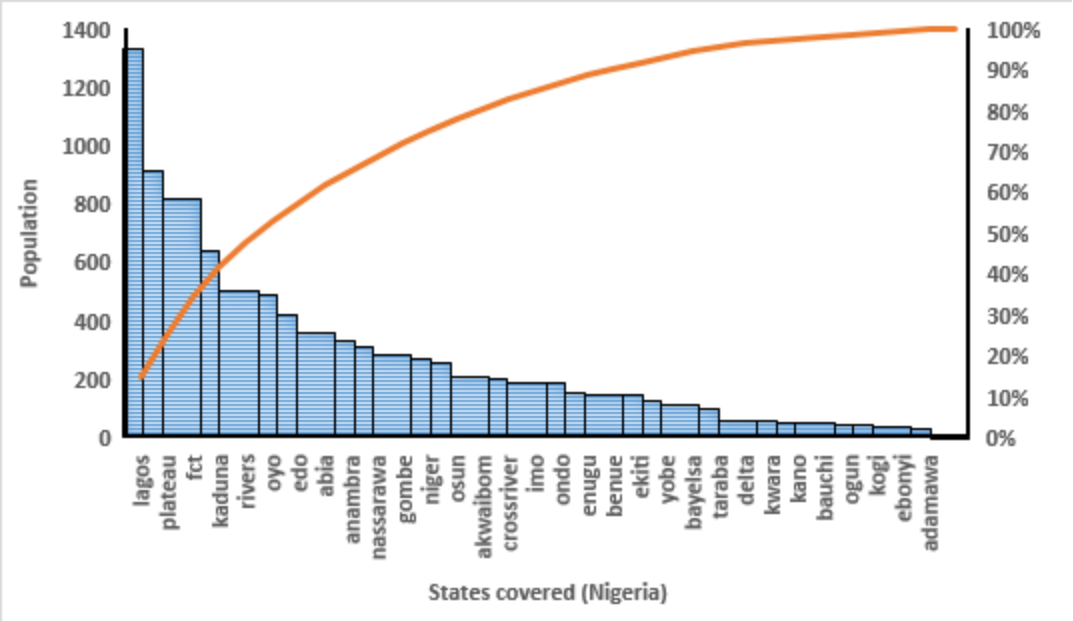


Figure 3. Population and Number of States Covered (Nigeria).

3.3 Dataset Preparation and Cleaning

The Orphanage Dataset raw file was inundated with a large quantity of information collected during the disbursement process. The raw file processing ensured the right attributes to be mined for this study. The raw orphanage dataset was pre-processed and cleaned on Microsoft excel. The data cleaning processes ensure the removal of unwanted attributes and quality dataset creation for an excellent analysis. The data were normalized to generate a cleaned orphanage dataset. Figure 4 presents the steps involved in data preparation and cleaning before applying the statistical methods and the machine learning technique.

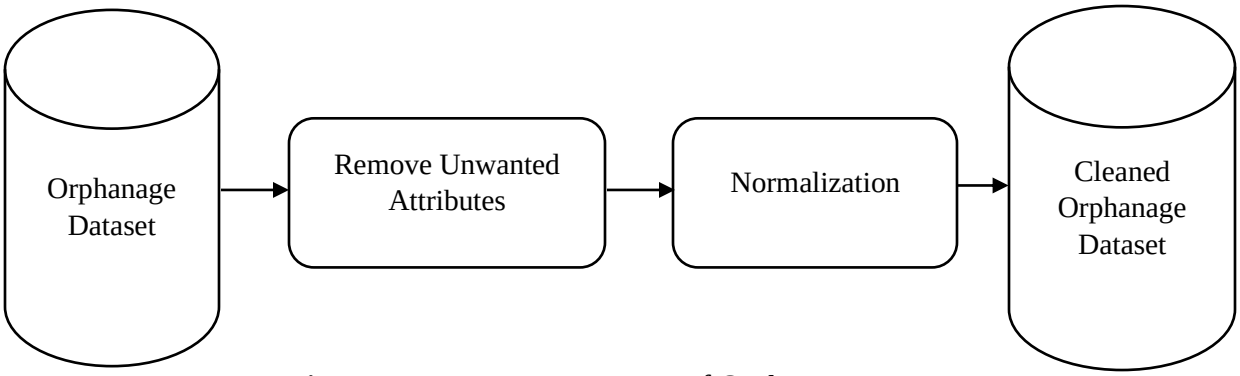


Figure 4. Data Pre-processing of Orphanage Dataset.

For the tool, the data in text format was converted to CSV (Comma-Separated Values) format. The state covered (Nigeria), the number of children/orphans, and the amount disbursed was extracted from the generated data and used to construct a new dataset in CSV (Comma-Separated Values) format.

3.4. Hierarchical Clustering Technique

Clustering is a multivariate data analysis that groups objects into different clusters by measuring distances and identifying individual clusters (Lee *et al.,* [48], Oshinubi,Rachdi and Demongeot *[49])*. Hierarchical Clustering is a commonly used clustering technique to estimate patterns in multi-dimensional datasets. Assessing the groups of data having a similar pattern can lead to a better understanding of the functions and state of orphanage donation. Figure 4 presents the flowchart of the hierarchical clustering algorithm, and the steps involved are:

Step 1: Compute the distance of each data point

Step 2: Consider all the data points as individual clusters

Step 3: Identify the number of cluster groups required

Step 4: Calculate the proximity of the new clusters and merge the two closest clusters to form new clusters.

Step 5: Finally, merge all the available clusters to form a new cluster.

Ward's method was employed in the hierarchical clustering method to determine the similarity between two clusters. This method was selected for its excellent execution in separating clusters even if there is noise between clusters. The ward's linkage method was used to calculate the distance matrix between the closest features considered in the hierarchical clustering technique. Kaufman and Rousseeuw [50] expressed Ward's method mathematically using Equation 1.

$$D(c_1, c_2) = \delta^2(c_1, c_2) = \frac{|c_1||c_2|}{|c_1|+|c_2|} \|c_1 - c_2\|^2 \quad (1)$$

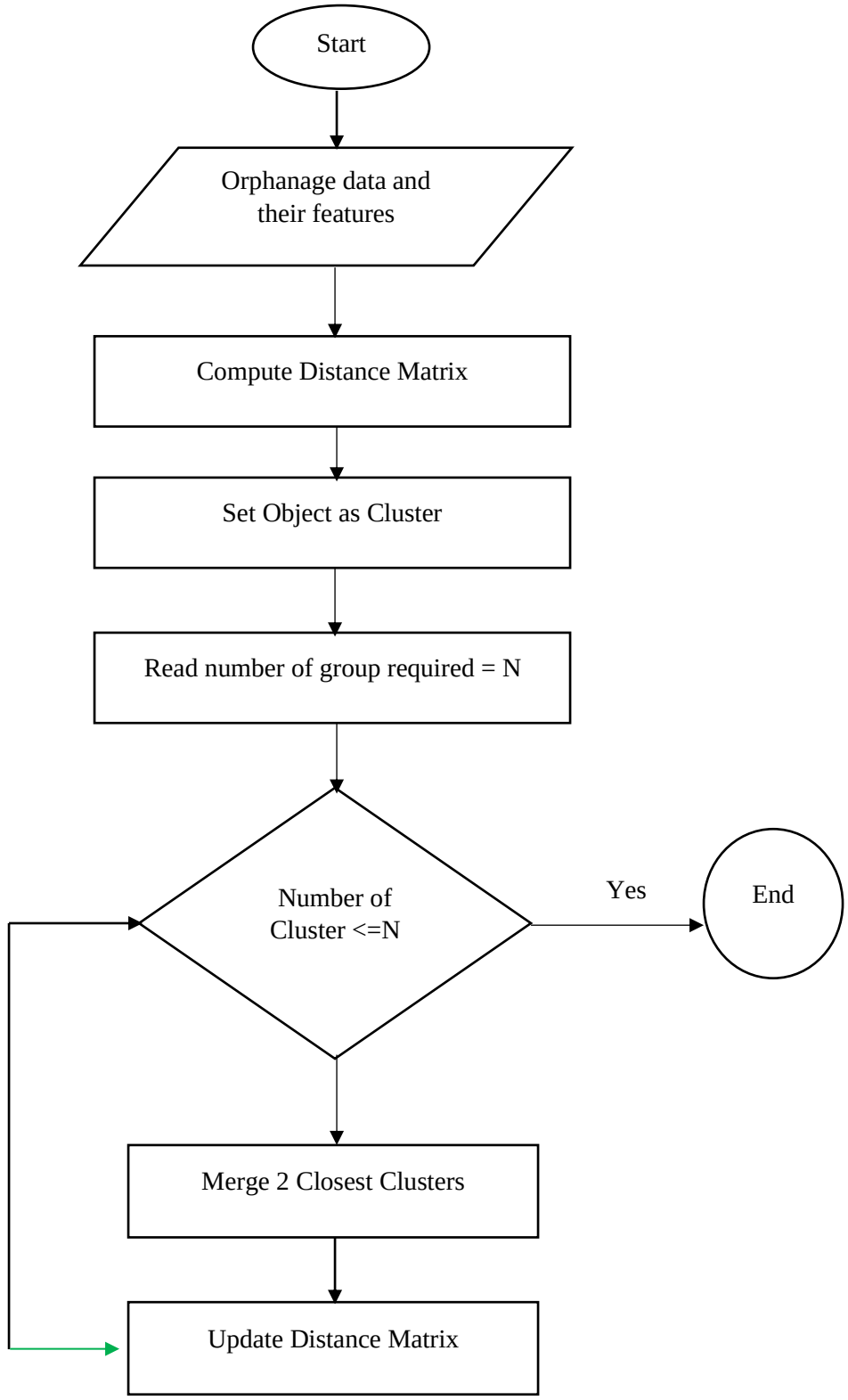


Figure 5. Flowchart of the Hierarchical Clustering Technique.

where c_1 and c_2 represents the two closest cluster.

The hierarchical clustering method was visualized using a Dendrogram. A dendrogram represents the results obtained from cluster analysis; it reveals all the steps taken in the hierarchical algorithm, which includes the distance at which clusters merged. Dendrogram offers a comfortable technique for exploring different possibilities of combining Data (Ackerman and Ben-David [51]).

The parameter employed in the training of hierarchical algorithm includes the number of clusters, affinity, and linkage, as presented in Table 1. All these parameters are set to an optimal value to enhance model optimization. The selected number of clusters to form and the number of centroids to generate was 3. The affinity used for this technique was Euclidean distance, which calls the distance between all points in the selected data. The ward linkage method estimates the proximity between two clusters.

Table 1. Hierarchical Algorithm Parameter

Hyperparameter	**Value**
Number of Clusters	3
Affinity	Euclidean
Linkage	Ward

3.5. Generalized Additive Model (GAM)

Hastie and Tibshirani developed Generalized Additive Models (GAMs) in 1990; GAM overcame the linearity assumption of Generalized Linear Models (GLMs) (Oshinubi, Rachdi and Demongeot [52]). GAMs are effective models capable of creating a relationship between predictor variables and the response. GAMs have the capability of showing the nonlinear association between response and set of predictor variables and also evaluate the output based on non-parametric functions (West *et al.,* 2014, Hastie and Tibshirani, [53]). In this study, the Gaussian family indicating numeric response was employed. Logarithmic GAMs were used to establish the relationship between the large quantity of the response variables and the predictor variables; it can be expressed using Equation 2 (Hastie and Tibshirani, [53]).

$$g(\mu) = \log(\mu) = \sum_{j=1}^{p} f_j(x_j) \tag{2}$$

where f_j represents the regression coefficients, x_j represents measured values for the predictor variables, and μ represents the mean of the response variable.

The GAM model used for this study was evaluated using deviance (D) and coefficient of determination in the regression. Deviance measures the fitted logistic model to a perfect model;

deviance mathematically expressed in Equation 3 is the difference of likelihoods between the saturated model, which equals one, and the fitted model.

$$D = -2loglik(\hat{\beta}) \tag{3}$$

The null deviance D_0 is a benchmark used to evaluate the magnitude of deviance; it compares how much the model has improved by adding the predictors. Null deviance can be expressed mathematically using Equation 4.

$$D_0 = -2loglik(\hat{\beta}_0) \tag{4}$$

R^2 is the coefficient of generalization in multiple regression, and it can be expressed mathematically using Equation 5.

$$R^2 = 1 - \frac{D}{D_0} \tag{5}$$

where D and D_0 represents deviance and null deviance, respectively.

3.6. Support Vector Machine (SVM) Technique

Support vector machines are conventional machine learning techniques used to solve classification problems that involve vast amounts of data; various field applications employ SVM in a big data environment. It is a known fact that SVM is computationally expensive and theoretically complex (Xu *et al.*, [54], Oshinubi *et al.*, [55]). SVM technique uses patterns to learn how to label or tag objects. In addition, support vector machines can be trained to learn features by studying many sample data. Presented in Figure 6 is the flowchart of the Support Vector Machine (SVM) technique for predictive analysis of the orphanage data. This section describes the steps taken in achieving the SVM technique. The technique takes input data comprising funds disbursed and the number of orphans that benefitted; Then, the model carries out a regression analysis process by comparing the features of the input data with the learned features of the data it has been trained.

The parameters for SVM are a set of values defined to enhance the SVM model and overall performance. The technique parameters identified in this study include the SVM-Type, SVM-Kernel, cost, gamma, epsilon, and p-value. These values are set at optimal values to improve the analysis. The best values found are presented in Table 2.

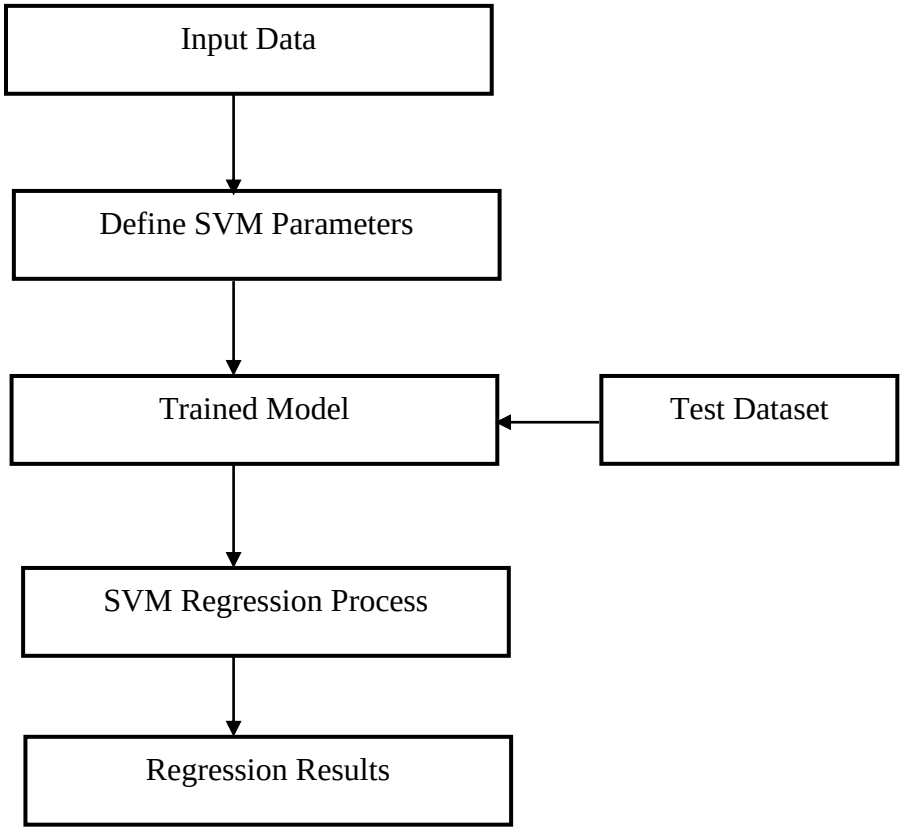


Figure 6. Flowchart of Support Vector Machine Technique

Table 2. Model Parameter for the SVM Model

Hyperparameter	**Value**
Type	Eps-regression
Kernel	Radial
Cost	1
Gamma	1
Epsilon	0.1
p-value	< 2.2e-16

To assess the performance of the support vector machines, three different performance evaluation metrics were employed. The mean absolute error (MAE) and root mean square error (RMSE) was used to assess the performance of the model, RMSE function evaluates possible significant errors. Also, the coefficient of determination (R^2) evaluate how well the developed model is fitted to the study data. Presented in Equations 6, 7, and 8 are the equations for calculating MAE, RMSE, and R^2 respectively.

$$MAE = \frac{1}{N}\sum_i^N (y_{actual,i} - y_{predicted,i}) \tag{6}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_i^N (y_{actual,i} - y_{predicted,i})^2} \tag{7}$$

$$R^2 = 1 - \left(\frac{\sum_i^N (y_{actual,i} - y_{predicted,i})}{\sum_i^N (y_{actual,i} - \hat{y}_{actual})}\right)^2 \tag{8}$$

where $y_{predicted}$ and y_{actual} represents the predicted and actual values of the data.

4. RESULTS

In this section, we present the results of the data analysis using several methods. We trained 80% of the data set and tested 20% to validate our result. We also used Cross-Validation (CV) of five folds to avoid overfitting. Jarque-Bera Normality test was used to check how the residual of the analysis behaved and the Q-Q visualization is shown. Hierarchical clustering was used to cluster different states and different orphanage homes so we can see the relationship across the country. We also used GAM to benchmark the two methods we used for the regression analysis. We used Marginal likelihood to estimate the GAM model coefficients and smoothing parameters and then use Gaussian and Gamma family distribution for the response variable.

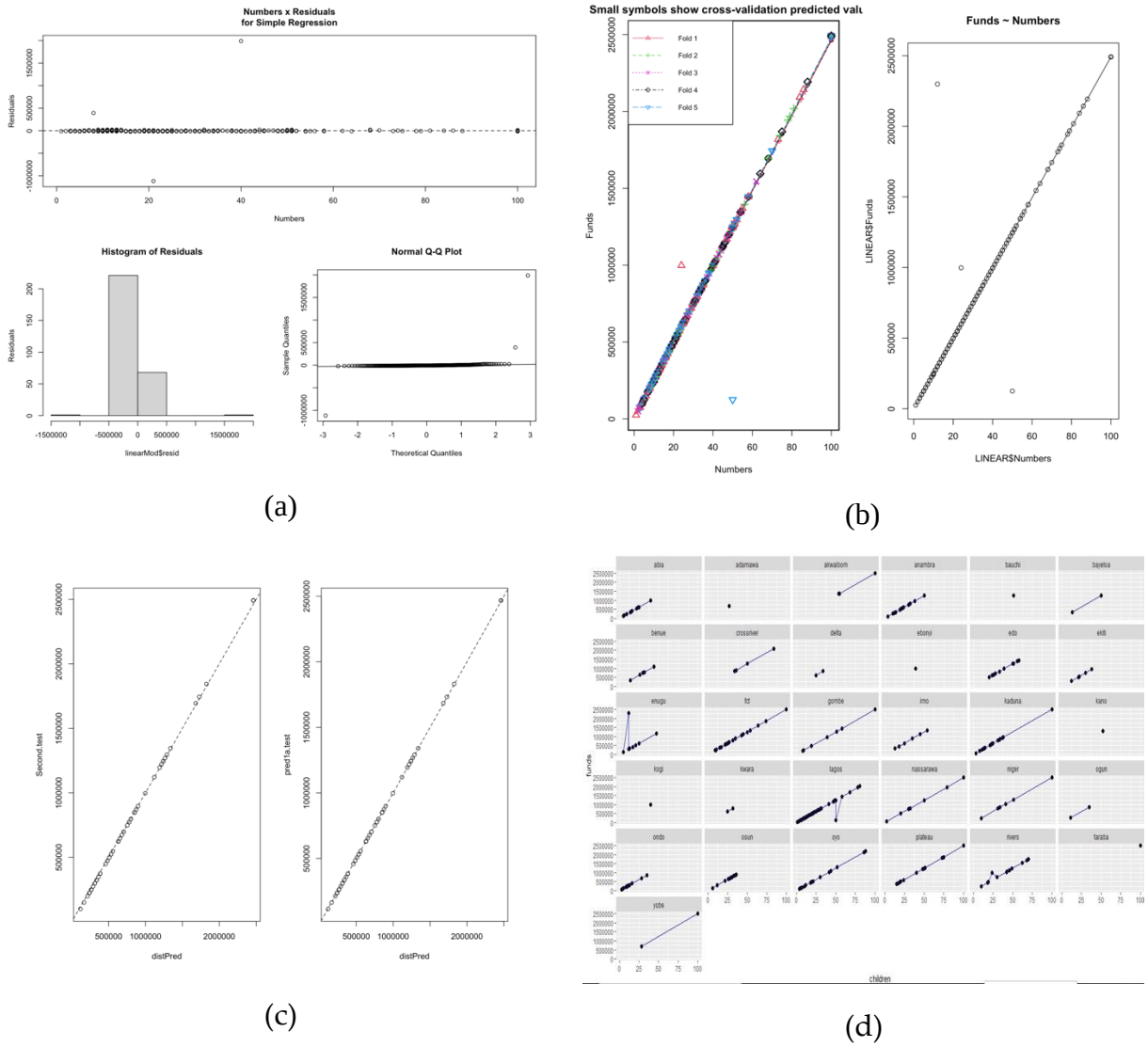
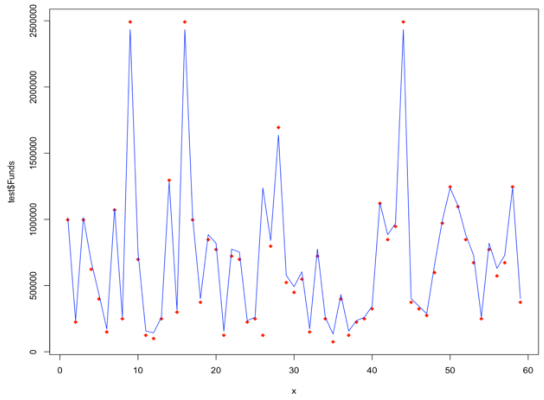
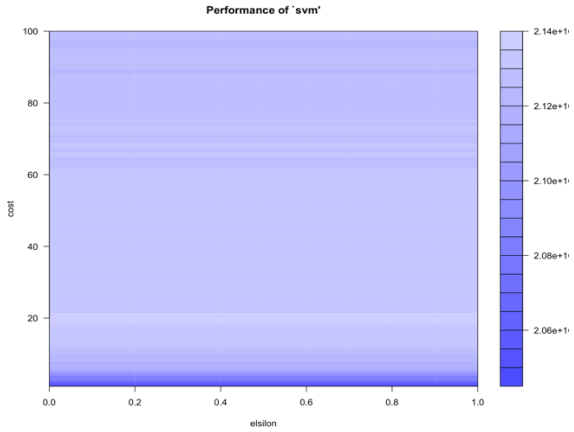


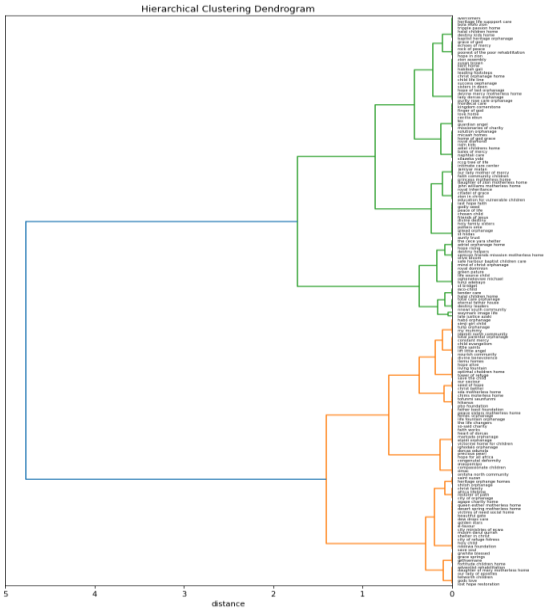
Figure 7. (a) Residual plot for linear regression. (b) Cross-validation for linear regression (c) Test set of the result for linear regression. (d) Analysis of funds distributed per state covered for the donation.



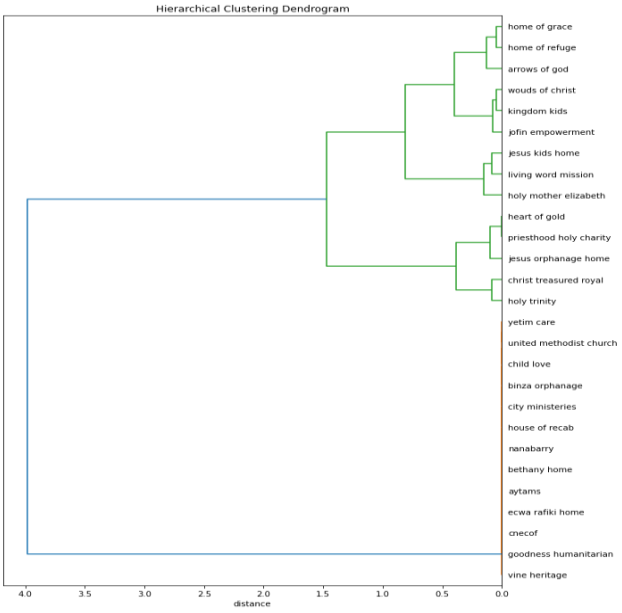
(a)



(b)



(c)



(d)

Figure 8. (a) Fit plot for SVM. (b) Performance of SVM plot (c) Hierarchical clustering for first clusters. (d) Hierarchical clustering for first clusters.

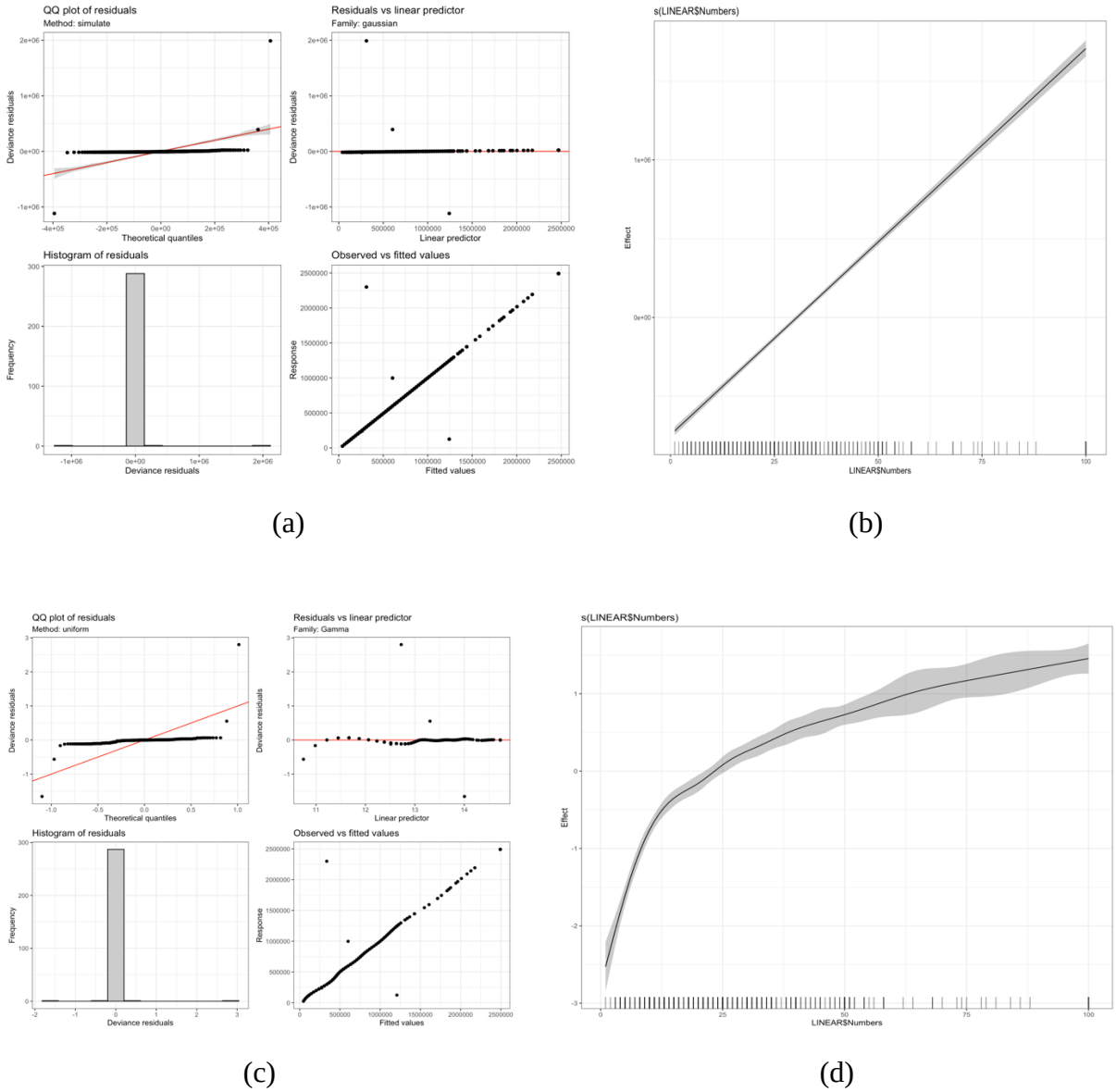


Figure 9. (a) Residual plot for GAM with Gaussian distribution. (b) Prediction and confidence interval for GAM with Gaussian distribution. (c) Residual plot for GAM with Gamma distribution. (d) Prediction and confidence interval for GAM with Gamma distribution.

The result of the normality test using Jarque-Bera test in Figure 7a to 7c is p-value: < 2.2e-16 while for SVM in Figure 8a & 8b is also p-value: < 2.2e-1 and MAE: 50193 RMSE: 149366. The R-squared for SVM is 0.923 and for linear regression, it is 0.9449. GAM and Mixed model in Figure 9 have the result for R-squared (adjusted) = 0.902 with Deviance explained = 91.5% and R-squared (adjusted) = 0.944 with Deviance explained = 92.9% respectively. Based on the result presented

here, reveals that linear regression better predicts our data set even though other methods achieved good results. The hierarchical plot in Figures 8c & 8d clustered orphanage homes based on the geopolitical zone in Nigeria which affirms the real nature of the fund's disbursement.

5. DISCUSSIONS

Charity is not a new venture in Nigeria due to many factors. Some of these include poor economic conditions, political instability, natural disasters, wars, pandemic and epidemics, poor social capital, unemployment, etc. As a result of these conditions, some people have been constrained physically, economically, and socially and to ensure that the plights of these people are considered, the charity has become a veritable option for helping the vulnerable in the society. Also, to keep up with the ideals of the Sustainable Development Goals (SDGs), there has been intensified action on charity and foundation funding in Nigeria and it was reported by Candid [56] that SDGfunders.org opined that the amount of total foundation funding for the SDGs in Nigeria between 2016 and 2018 was around USD 564.27million.

As found from the study, 8859 children were domiciled in the orphanages chosen and ₦221985548 was distributed. The spread of funds disbursed was evenly distributed throughout all the states as affected by the number of children present in each orphanage/motherless center. This indicates that there was no bias or lopsided distribution in the donation exercise. Furthermore, the frequency of distribution per child (Figure 1c) in the homes covered was more in the range of 13 – 25 per home covered in each state with the lowest occurrence between 85 -97 per home. Ensuring the well-being of the vulnerable and those in the grassroots is not the exclusive preserve of the government and people from the private sector can also take part as exemplified by the donor of these funds (David Adeleke). Countries need all the resources they can amass to ensure development and charity is an avenue for that. The regression analysis (Figure 7a to 7c) explains that 94.4% of every fund disbursed is affected by the number of children and this indicates that there is a significant effect of the funds distributed on the number of children as concerning the project. Hence, the funds were used judiciously as intended for the project as each child was adequately cared for.

The states with the highest fund distribution as presented in Figure 7d are Lagos, Plateau, FCT, Kaduna, and Rivers while states such as Bauchi, Ogun, Kogi, Ebonyi, and Adamawa had the least

amount of funds disbursed. This is affected by the corresponding number of children per orphanage/motherless home represented in each state.

6. CONCLUSION

As discussed above, some humans are constrained by various factors which make them vulnerable. As vulnerable individuals, they are pre-disposable to all forms of vices and health risks, and this necessitates concerted action on their welfare which can only be done through an efficient and effective system that involves all regardless of all divides. The well-being of people should not be the sole preserve of the government and well-to-do individuals can also contribute to this as exemplified by David Adeleke.

Amongst these humans are orphans who are people without parents. They are pre disposable to health risks and this is so because there is a poor social support system in Sub-Saharan Africa which is inadequate to cater to the need of all orphans (UNICEF [57]). Apart from the inadequate support system, they are also constrained socially and might be ostracized from the other members of the society which is expected to have a tremendous impact on their psyche.

The data utilized for this study is the donation by David Adeleke (*Davido*) over the distribution of monies (sum = NGN 250,000,000) to various orphanages in Nigeria on 15th February 2022 (Legit media house, [57]. Out of the ₦250m donated, ₦25m was donated to Paroche Reach out Foundation. The data consist of states, names of orphanages, number of children per orphanage, and amount distributed to the 291 registered orphanages across Nigeria. The data obtained were entered into a Microsoft Excel spreadsheet, treated for missing values, and screened for outliers. The dependent (response) variable in this study was the funds disbursed (NGN) to the orphanages while the independent variables were states, the name of the orphanage, and the number of children. Linear regression analysis was used to analyze the relationship between the funds disbursed and the number of children present in all the orphanages. A graphical representation of the relationship between funds disbursed and the number of children was also studied using a scatterplot. Scatterplots are excellent for spotting a connection between two variables, as well as if the two variables appear to have a functional relationship. The number of children values was plotted on the x-axis, while the funds' disbursed values were plotted on the y-axis.

The Q-Q plot was another tool used to analyze the data to detect whether or not the dataset was normally distributed. It is a unique type of scatterplot that is created using the data's quantiles. The

quantiles of the experimental data are plotted along the x-axis. The quantiles of a predefined distribution with the same mean and standard deviation as the experimental data are plotted along the y-axis. The data were also analyzed using the Generalized Additive Model, a sort of semi-parametric technique based on generalized linear models. The model's smooth functions are intended to reflect non-linear relationships between the independent and dependent variables.

Clustered data to regression models mean that data from known groups ("clusters") are observed. Often these are the result of repeated measurements on the same individuals at different time points. When using clustered data, we have to take into account that observations from the same cluster are correlated. Using a model designed for independent data may lead to biased results or at least significantly reduce the efficiency of the estimates.

Support Vector Machine (SVM) formulation was further used to analyze the data by focusing on the classification boundary, through which it was able to forecast labels based on attributes without having to model or estimate the probabilistic mechanism that creates the labels. The SVM directly minimizes the error rate under the 0-1 loss from a decision-theoretic perspective.

The study concluded that there was a good spread of the funds along the states considered and five states were exempted from the distribution. This study has also revealed the spread of orphanages across the states of Nigeria and the database of orphanages captured can be adopted and built on for several interventions targeted toward the orphans and other vulnerable persons within the society.

DATA AVAILABILITY

Data and codes used for the analysis can be found in the link below:
https://github.com/Honkay/DISTRIBUTION-OF-ORPHANGE-FUNDS-DONATED-BY-DAVIDO.git.

ACKNOWLEDGMENTS

The authors wish to appreciate Mr. David Adeleke (Davido) for his generous donation and for improving the lives of humanity in Nigeria.

AUTHOR CONTRIBUTIONS

Conceptualization, K.O.; methodology, T.O., B.B.A., G.K.A., H.O., G.O.E., and K.O.; software, G.O.E., K.O. and G.K.A.; validation, K.O.; formal analysis, O.O.K., T.O., G.O.E., K.O. and

G.K.A.; investigation, K.O.; resources, K.O.; data curation, G.O.E.; writing—original draft preparation, O.O.K., T.O., B.B.A., G.K.A., H.O., G.O.E., and K.O.; writing—review and editing, O.O.K., T.O., B.B.A., G.K.A., H.O., G.O.E. and K.O.; visualization, B.B.A., G.O.E., K.O. and G.K.A.; supervision, K.O; project administration, K.O.; All authors have read and agreed to the final version of the manuscript.

FUNDING

No specific funding has been received for the research.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] G. Esping‐Andersen, A child‐centred social investment strategy, in: G. Esping-Andersen (Ed.), Why We Need a New Welfare State, 1st ed., Oxford University PressOxford, 2002: pp. 26–67. https://doi.org/10.1093/0199256438.003.0002.

[2] R.A. Hart, Children's participation-The theory and practice of involving young citizens in community development and environmental care, Routledge, 2013. https://doi.org/10.4324/9781315070728.

[3] WHO, Nurturing care for early childhood development: a framework for helping children survive and thrive to transform health and human potential, (2018). https://apps.who.int/iris/handle/10665/272603.

[4] V.S. Nyathi, Equipping orphans and vulnerable children (OVC) with life skills education, in: S.G. Taukeni, J. Mathwasa, Z. Ntshuntshe (Eds.), Advances in Psychology, Mental Health, and Behavioral Studies, IGI Global, 2022: pp. 47–71.

[5] R.K. Shah, Concepts of learner-centred teaching, Shanlax Int. J. Educ. 8 (2020), 45-60.

[6] K.L. Edyburn, S. Meek, Seeking Safety and Humanity in the Harshest Immigration Climate in a Generation: A Review of the Literature on the Effects of Separation and Detention on Migrant and Asylum‐Seeking Children and Families in the United States during the Trump Administration. Soc. Policy Rep. 34 (2021), 1-46.

[7] K. Linn, A. Fay, K. Meddles, et al. HIV-related cognitive impairment of orphans in Myanmar with vertically transmitted HIV taking antiretroviral therapy, Pediatric Neurol. 53 (2015), 485-490.

[8] Wordnet, What is an orphan? (2007). Retrieved April 29, 2012. http://wordnet.princeton.edu/perl/webwn?s=orphan.

[9] H. Huynh, Lessons learned from high-quality residential care centers around the world: A visual story, Int. J. Child Maltreatment: Res. Policy Pract. 2 (2019), 99–116.

[10] C. Nar, Orphan report. Instanbul: Pelikan basam, R. Osei Sarpong, C. Mensahankrah, (2018). Adoption practices fueling child trafficking in Ghana. SSRN Electronic Journal, 1–10, (2020).

[11] P. Bennell, The educational attainment of orphans in high HIV countries in sub-Saharan Africa: An update. Int. J. Educ. Develop. 82 (2021), 102358.

[12] J. Pillay, Early Education of orphans and vulnerable children: A crucial aspect for social justice and African development. Koers: Bulletin for Christian Scholarship, 83 (2018), 1-12.

[13] S. Ahmad, S. Rashid, Conflict and Orphans: An exploration of challenges faced in educational sector by Orphans in Kashmir, The Communications, (2019), 96-102.

[14] B. Kurniawan, N. Neviyarni, S. Solfema, The relationship between self-esteem and resilience of adolescents who living in orphanages, Int. J. Res. Counsel. Educ. 1 (2018), 47-52.

[15] Z. Ntshuntshe, S.G. Taukeni, Psychological and social issues affecting orphans and vulnerable children, In Addressing Multicultural Needs in School Guidance and Counseling (pp. 20-31), IGI Global, (2020).

[16] E.S. Sagalaeva, S.N. Ivahnenko, O.V. Landina, Social and legal norms ensuring the child's right to live and be brought up in a family, J. Adv. Res. L. & Econ. 10 (2019), 674.

[17] R.S. Bagnall, K. Brodersen, C.B. Champion, et al. (Eds.). The encyclopedia of ancient history (Vol. 1), Wiley-Blackwell, (2013).

[18] T. Mwoma, J. Pillay, Psychosocial support for orphans and vulnerable children in public primary schools: Challenges and intervention strategies, South Afr. J. Educ. 35 (2015), 1092.

[19] S. Lyneham, L. Facchini, Benevolent harm: Orphanages, voluntourism and child sexual exploitation in South-East Asia, Trends Issues Crime Criminal Justice, (574) (2019), 1-16.

[20] A.A. Muhammad, Common practices in orphanages: A case study of Bauchi Nigeria, Int. J. Umranic Stud. 3 (2020), 35- 42.

[21] E. Daniel, The usefulness of qualitative and quantitative approaches and methods in researching problem-solving ability in science education curriculum, J. Educ. Practice, 7 (2016), 91-100.

[22] S. Ahmad, et al. Qualitative v/s. quantitative research-A summarized review, J. Evid. Based Med. Healthc. 6 (2019), 2828-2832.

[23] P. Asper, U. Corte, What is qualitative in qualitative research, Qual. Sociol. 42 (2019), 139-160.

[24] P.M. Galdas, Revisiting bias in qualitative research: reflections on its relationship with funding and impact, Int. J. Qual. Meth. 16 (2017), 1-2.

[25] V. Elliot, Thinking about the coding process in qualitative data analysis, Qual. Rep. 23 (2018), 2850-2861.

[26] S. Ahmad, et al. Qualitative v/s. quantitative research-A summarized review, J. Evid. Based Med. Healthc, 6 (2019), 2828-2832.

[27] O.D. Apuke, Quantitative research methods a synopsis approach, Arab. J. Bus. Manage. Rev. 6 (2017), 40-47.

[28] A. Queirós, D. Faria, F. Almeida, Strengths and limitations of qualitative and quantitative research methods, Eur. J. Educ. Stud. 3 (2017), 369-387.

[29] K. Kumari, S. Yadav, Linear regression analysis study, J. Pract. Cardiovasc Sci. 4 (2018), 33-36.

[30] S. Rosenthal, Regression analysis linear, 2017. [Online] Available at: https://www.researchgate.net/publication/320928517_Regression_Analysis_Linear[Accessed 19 February 2022].

[31] K. Kumari, S. Yadav, Linear regression analysis study, J. Pract. Cardiovasc Sci. 4 (2018), 33-36.

[32] S. Rosenthal, Regression analysis linear, 2017. [Online] Available at: https://www.researchgate.net/publication/320928517_Regression_Analysis_Linear[Accessed 19 February 2022].

[33] B. Meuleman, G. Loosveldt, V. Emonds, Regression analysis: Assumptions and diagnostics. In: H. Best & C. Wolfs, eds. The Sage handbook of regression analysis and causal inference. London: Sage, 2015.

[34] S. Ray, A quick review of machine learning algorithms. In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), (2019), pp. 35-39.

[35] S. Yue, P. Li, P. Hao, SVM classification: Its contents and challenges, Appl. Math. 18 (2003), 332-342.

[36] V. Vapnik, The nature of statistical learning theory, N.Y.: Springer, (1995).

[37] H.R. Eghbalnia, A complex-valued overcomplete representation of information for visual search: A learning theoretic approach based on multiscale symmetry, The University of Wisconsin-Madison, (2000).

[38] R. Burbidge, M. Trotter, B. Buxton, et al. Drug design by machine learning: support vector machines for pharmaceutical data analysis, Computers Chem. 26 (2001), 5-14.

[39] M. Fan, X. Luo, J. Liu, et al. Graph embedding based familial analysis of android malware using unsupervised learning, In: 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), (2019), pp. 771-782.

[40] D. Patel, R. Modi, K. Sarvakar, A comparative study of clustering data mining: Techniques and research challenges, Int. J. Latest Technol. Eng. Manage. Appl. Sci. 3 (2014), 67-70.

[41] P. Rani, A survey on STING and CLIQUE grid based clustering methods, Int. J. Adv. Res. Computer Sci. 8 (2017), 1510-1512.

[42] K. Bouchefry, R.S. Souza, Learning in big data: introduction to machine learning. In: P. Skoda & F. Adam, eds. Knowledge discovery in big data from astronomy and earth observation, Amsterdam: Elsevier, 2020, pp. 225-249.

[43] J.H. Ward Jr. Hierarchical grouping to optimize an objective function, J. Amer. Stat. Assoc. 58 (1963), 236–244.

[44] C. Ding, X. He, K-means clustering via principal component analysis, In: Proceedings of the twenty-first international conference on machine learning, (2004), p. 29.

[45] R. Yogita, R. Harish, A study of hierarchical clustering algorithm, Int. J. Inform. Comput. Technol. 3 (2013), 1115-1122.

[46] T. Hastie, R. Tibshirani, Generalized additive models, Stat. Sci. 1 (1986), 297-310.

[47] T. Vatter, V. Chavez-Demoulin, Generalized additive models for conditional dependence structures, J. Multivar. Anal. 141 (2015), 147-167.

[48] S. Lee, J. Kim, J. Hwang, et al. Clustering of time series water quality data using dynamic time warping: A case study from the bukhan river water quality monitoring network, Water, 12 (2020), 2411.

[49] K. Oshinubi, M. Rachdi, J. Demongeot, Modelling of COVID-19 pandemic vis-à-vis some socio-economic factors, Front. Appl. Math. Stat. 7 (2022), 786983.

[50] L. Kaufman, P.J. Rousseeuw, Finding groups in data: An introduction to cluster analysis, John Wiley & Sons, (2009).

[51] M. Ackerman, S. Ben-David, A characterization of linkage-based hierarchical clustering, J. Mach. Learn. Res. 17 (2016), 8182-8198.

[52] K. Oshinubi, M. Rachdi, J. Demongeot, Analysis of daily reproduction rates of COVID-19 using current health expenditure as gross domestic product percentage (CHE/GDP) across countries, Healthcare, 9 (2021), 1247.

[53] T.J. Hastie, R.J. Tibshirani, Generalized additive models, Stat. Sci. 1 (2010), 297–310.

[54] Y. Xu, W. Ahmad, A. Ahmad, et al. Computation of high-performance concrete compressive strength using standalone and ensembled machine learning techniques, Materials, 14 (2021), 7034.

[55] K. Oshinubi, F. Al-Awadhi, M. Rachdi, et al. Data analysis and forecasting of Covid-19 pandemic in Kuwait based on daily observation and basic reproduction number dynamics, Kuwait J. Sci. Special Issue, (2021), 1-30.

[56] United Nations Children's Fund/The Joint United Nations Programme on HIV/AIDS. Children orphaned by AIDS. Frontline responses from eastern and southern Africa. UNICEF/UNAIDS; 1999. Available at: http://www.unicef.org/pub_aids_en.pdf

[57] Legit Media House, Available online: https://legit9ja.com/2022/02/davido-keeps-his-word-donates-n250-million-to-292-orphanages-around-the-country-see-list.html. (2022). (accessed on 17th February 2022).