Scholarly Journal of Mathematics & Science (https://damaacademia.com/sjms/)
Volume 1, Issue 4, pp.01-07, April, 2019
Published by: Dama Academic Scholarly & Scientific Research Society (www.damaacademia.com)

A Comparison between Logistic Regression and K Nearest Neighbor in Modeling Mortality Amongst Children Under five Years in Ghana

[1]Salifu Nanga | [2]Anani Lotsi, PhD.
Department of Statistics, School of Graduate Studies, University of Ghana,
Sana Research and Consultancy Ghana Limited

Abstract
Child mortality is regarded as one of the most revealing measures of society's ability to meet the needs of its people. The Millennium Development Goal 4 (MDG 4) advocates a reduction of under-five mortality rate by two-thirds between 1990 and 2015. The main objective of this study was to develop a validated set of statistical models and select the most appropriate model between logistic regression and K Nearest Neighbor to predict mortality among children under five and to compare the influence of selected risk factors on the probability of death before the age of 5 years among children in Ghana. The study revealed that the K Nearest Neighbor model was the most efficient in modeling Mortality in Children under five with a CCR of 83%. The Logistic Regression model will also do a good job at predicting mortality in children under five with a CCR of 81%. The highest educational level of mother, Age of mother at birth, Type of toilet facility used by family, alcohol consumption and the wealth index of family were discovered as the most important variables in predicting mortality amongst children under five in Ghana across both models.

***Keywords*:** *Logistic Regression, Neural Networks, Children under five years,*

1.0 INTRODUCTION

One of the most revealing measures of how well a society is performing is meeting the needs of its people is Child mortality (Iram, 2008). Millennium Development Goal 4 (MDG 4) is aimed at reducing under-five mortality rate by two-thirds between 1990 and 2015 (Unicef, 2014). The past two decades has seen an overwhelming increase in under-five deaths; between the period 1990 and 2013, 223 million children died before age five. There has been a 49 percent decline in the under-five mortality rate globally since 1990.This statistic however is still far below the two-thirds reduction targeted to reach the Millennium Development Goal 4(Unicef, 2014). Various factors has been empirically proven in previous research to influence a child's health and survival, including place of residence, breastfeeding, place of delivery, access to postnatal care, maternal age and education (Doctor HV, 2011).

Variations in child mortality rates has been linked to geographic differences in maternal literacy levels and sociocultural practices within countries (Black et al, 2003). Place of delivery, with evidence indicating that women who deliver at health facilities have a lower probability of reporting child death compared to those delivering in home settings has been linked to the likelihood of under-five mortality (Doctor HV, 2011). Postnatal care access has been associated with a drop in under-five mortality, with a study carried out in Bangladesh showing that postnatal home visits within the first 2 days after birth by skilled healthcare workers was significantly associated with a lower likelihood of child death (Baqui, 2009). Maternal education and age has been found to important determinants of child mortality in a study conducted in the developing countries (Deribew et al., 2007).

Evidence also shows that child mortality rates is found to be higher among less educated mothers as compared to mothers who have higher levels of education (You, 2011). The importance of education, mother's education in particular, has been confirmed in many subsequent studies (Murthi et al., 1995; Dre`ze and Murthi, 2001).To the extent that, education is able to improve an individual's ability to undertake these changes, more educated mothers are expected to have healthier babies (Iram et al,2008). Mother's employment status is also considered as a significant factor that affects neonatal, infant and child mortality (Arriaga and Hobbs, 1982). The work status of mother determines the amount of time and care a mother can give to her child, and it may determine the amount of resources (income) available to the mother and thus her access to various goods and services. Data mining applies to advanced data analysis techniques that explore large sets of data to identify useful and unexpected patterns and rules that provide relevant knowledge for predicting future outcomes (Lungu and Bâra, 2012). Nearest Neighbor Analysis is a method that classifies cases based on their similarity to other cases. Machine learning was developed as a way to help recognize patterns of data without requiring an exact match to any stored patterns.

Logistic regression is a mathematical modelling approach that can be used to describe the relationship of several independent variables to a dichotomous dependent variable.Hand (1997) studied the application of k-nearest-neighbor (k-NN) method, as a standard technique in pattern recognition and nonparametric statistics, as a credit scoring techniques for assessing the credit worthiness of consumer loan applicants. Breiman (1996) looked at the instability of different predictors and concluded that neural networks, classification trees and subset selection in linear regression were unstable while the k-th nearest neighbor method was found to be stable. Weinberger and Saul (2007) also

presented a developed algorithm of k-NN. In their proposed model, they applied Mahalanobis distance as the criterion for determination of distance. A developed hierarchical model of k-NN was also introduced by Kubota et al,(2001).Wiginton (1980) gave one of the first published accounts of logistic regression applied to credit scoring in comparison to discriminant analysis and concluded that logistic regression gave a superior result. The purpose of this study was to assess what factors are most important in determining mortality in children under five years in Ghana by comparing Logistic Regression which is a statistical technique and Neural Networks a data mining technique.

2.0 MATERIALS AND METHODS

The study was undertaken using the 2008 Ghana Demographic and Health Survey (GDHS). A sample of 11,888 respondents were considered in the study .A logistic Regression model and K Nearest Neighbor model was applied to the dataset. The variables considered in the study was whether child was alive or not as the dichotomous dependent variable. The explanatory variables were Type of place of residence of family, Highest educational level of mother, Source of drinking water, type of toilet facility used by family, Wealth index of family, alcohol consumption of mother, Whether mother is covered by health insurance, Marital status of mother, Sex of child and age of mother at child birth.

2.1 Logistic Regression

Logistic regression is a mathematical modelling approach that can be used to describe the relationship of several independent variables to a dichotomous dependent variable. It allows the prediction of discrete variables by a mix of continuous and discrete predictors. Letting Y be the binary response variable, it is assumed that $P(Y=1)$ is possibly dependent on $\vec{x}$, a vector of predictor values. The goal is to model

$p(\vec{x}) \equiv P(Y=1 \mid \vec{x})$.

Logistic Regression –With Dichotomous Responses and numeric and/or categorical explanatory variable(s). Model the probability of a particular as a function of the predictor variable(s) Probabilities are bounded between 0 and 1

$$\pi = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}{1+e^{\alpha+\beta_1 x_1+\cdots+\beta_k x_k}}$$

.

The P=primary interest in estimating and testing hypotheses regarding Large-Sample test is the Wald Test:

H_0: $b = 0$ H_A: $b \neq 0$

$$T.S.: X^2_{obs} = \left(\frac{\hat{\beta}}{\sigma_{\hat{\beta}}} \right)^2$$

$$R.R.: X^2_{obs} \geq \chi^2_{\alpha,1}$$

$$P-val: P(\chi^2 \geq X^2_{obs})$$

Odds Ratio

$$\frac{odds(x+1)}{odds(x)} = e^{\beta} \qquad \left(odds(x) = \frac{\pi(x)}{1-\pi(x)} \right)$$

Thus e^b shows the change in the odds of the outcome (multiplicatively) by an increase in x by 1 unit
When $b = 0$, the odds and probability are the same at all x levels (e^b=1)
when $b > 0$, the odds and probability increase as x increases (e^b>1)
When $b < 0$, the odds and probability decrease as x increases (e^b<1)

Testing Regression Coefficients

$$H_0 : \beta_1 = \cdots = \beta_k = 0$$
$$H_A : \text{Not all } \beta_i = 0$$
$$T.S.\ X^2_{obs} = (-2\log(L_0)) - (-2\log(L_1))$$
$$R.R.\ X^2_{obs} \geq \chi^2_{\alpha,k}$$
$$P = P(\chi^2 \geq X^2_{obs})$$

2.2 Nearest Neighbors

a. Number of Nearest Neighbors (k)

The number of nearest neighbors was specified in the model.Values of K will be tentatively chosen in the model. Each value of k in the requested range w tested, and the k, and accompanying feature set, with the lowest error rate was selected. The Euclidean metric and City block metric (Manhattan distance) were used to compute distances between points. Mahalanobis transformation which helps eliminate the correlation between variables and also standardizes the variance of each variable was also computed.

b. Feature Selection

Feature selection was based on the approach called wrapper of Cunningham and Delany (2007) and it applies forward selection which begins from Jforced features are entered into the model. Further features are sequentially chosen; the feature chosen at each step is the one that tends to cause the largest decrease in the error rate or sum-of squares error. Let Sj represent the set of J features which currently are chosen to be included, Scj was represented by the set of remaining features and ej represented the error rate or sum-of-squares error associated with the model based on Sj.

c. Stopping Criterion

At each step, the addition of feature whose addition to the model can result in the smallest error (computed as the error rate for a categorical target and sum of squares error for a scale target) was considered for inclusion in the model set. Forward selection was stopped after the specified condition was met.

d. Partition

There was a division of the dataset into training and holdout sets. It specified the partitioning method of dataset into training and holdout samples. The training sample contained data records that was used to train the nearest neighbor model; The holdout sample which is an independent set of data records that is used to assess the final model; the error from the holdout sample helps to give an "honest" estimate of the models ability to predict.

e. Cross-Validation Folds

To determine the best number of neighbors, V-fold cross-validation was used. The "best" number of nearest neighbors is the one that gives the lowest error across folds.

f. The Feature Space Chart

The feature space chart that is interactive in nature. Each axis is represented by a feature in the model, and the location of points in the chart shows what the values of these features represent for cases in the training and holdout partitions**.**

g. Comparison of Models

In order to evaluate the most appropriate technique, the predicted results for each model will be used. The predictive results was determined by the correct classification rate (CCR) test.

3.0 RESULTS AND DISCUSSION

3.1 Logistic Regression

Table 1 displays a summary of the model we see that the -2 Log Likelihood statistic is 10768.032. This statistic measures how poorly the model predicts the decisions, the smaller the statistic the better the model. A more useful measure to assess the utility of a logistic regression model is classification accuracy. From Table 2 the overall classification accuracy of the model was 81%.

Table 1 : Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	10768.032	.020	.034

Table 2 : Classification Table

			Child is alive?		Percentage Correct
			Yes	No	
Step 1	Child is alive?	Yes	8187	1566	83.94
		No	683	1400	67.21
	Overall Percentage				81.00

Table 3 displays the results for the omnibus tests of model coefficients. The presence of a relationship between the dependent variable and combination of independent variables is based on the statistical significance of the model chi-square at step 1 after the independent variables have been added to the analysis. In this analysis, the probability of the model chi-square (244.581) was < 0.001, less than or equal to the level of significance of 0.05. The existence of a relationship between the independent variables and the dependent variable was therefore supported.

Table 3 : Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	244.581	15	.000
	Block	244.581	15	.000
	Model	244.581	15	.000

Table 4 : Variables in Logistic Regression Model

Variables in the Equation						
	β	**S.E.**	**Wald**	**df**	**P-value**	**Odds**
Place_Of_Residence(1)	.050	.073	.475	1	.491	1.051
Educational_Level			25.751	3	.000	
Educational_Level(1)	.515	.007	4.726	1	.030	1.674
Educational_Level(2)	.634	.009	7.183	1	.007	1.501
Educational_Level(3)	.312	.024	1.797	1	.180	1.366
Source_of_drinking_water(1)	.129	.065	4.002	1	.045	1.138
Toilet_facility			57.315	2	.000	
Toilet_facility(1)	.617	.122	25.431	1	.000	1.854
Toilet_facility(2)	.457	.062	54.250	1	.000	1.579
Wealth_index			30.675	2	.000	
Wealth_index(1)	-.492	.089	30.321	1	.000	.612
Wealth_index(2)	-.244	.075	10.666	1	.001	.784
Age_of_mother	.031	.003	82.199	1	.000	1.031
health_insurance(1)	.136	.052	6.952	1	.008	1.146
Marital_status			.949	2	.622	
Marital_status(1)	-.190	.234	.659	1	.417	.827
Marital_status(2)	-.056	.074	.565	1	.452	.946
Sex_of_child(1)	.180	.049	13.471	1	.000	1.197
alchohol_consumption(1)	.195	.060	10.498	1	.001	1.215
Constant	-3.625	.292	153.985	1	.000	.027

The coefficient of any variable is deemed to be significant if P value ≤ 0.05.Hence the significant values in the model are Mothers educational level, Age of mother at birth, Type of toilet facility used, Wealth index of family, whether mother is registered with health insurance, Sex of child and alcohol consumption. A standard error larger than 2.0 indicates numerical problems, such as multicollinearity among the independent variables. From Table 4, it can be deduced that none of the independent variables had a standard error greater than 2.0 hence there is no evidence of multicollinearity.

Table 5 : Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	3.229	8	.919

Table 5 shows the Hosmer and Lemeshow Test , since the p – value, 0.919, is greater than the significance level, α = 0.05, we conclude that there is enough evidence that the hypothesized model fits the data set used in predicting mortality among under five children in Ghana.

A. K Nearest Neighbor

The dataset was divided into training and holdout sets. The training sample comprised the data records used to train the nearest neighbor model; 70% of cases in the dataset were assigned to the training sample in order to obtain a model. The holdout samples are independent set of data records that is used to assess the final model; the error from the holdout sample gives an "honest" estimate of the ability of the model to predict. 30% of cases were assigned to the holdout sample.

Table 6 : Error Summary for Nearest Neighbor

Partition	**Percent of Records Incorrectly Classified**
Training	16.8%
Holdout	16.6%

K = 4 was selected because it had the lowest error rate. The error summary in Table 4.15 shows the percentage of incorrect predictions is roughly equal across both the training and holdout samples which is a good sign that the model is a good fit.

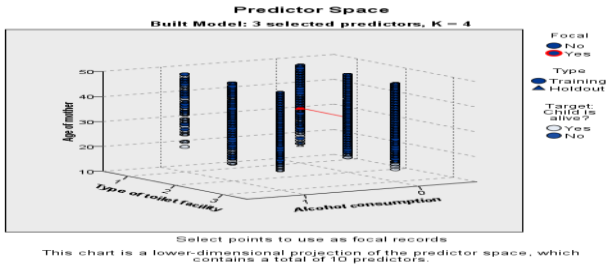

Figure 1 : Predictor Space Chart

In Nearest Neighbor analysis the predictor variables are referred to as feature variables and the dependent variable is referred to as the target variable. Figure 1 displays a feature or predictor space chart with 10 dimensions based on the number of feature variables. In the chart similar cases are near each other and dissimilar cases are distant from each other. The Euclidean metric was used in the computation of distances within the chart. A focal case (point) was randomly selected and marked red in the chart .The analysis will then be based on the focal point randomly selected. From any selected point, the four nearest neighbors to that point will be selected using the Euclidean metric and based on majority vote, nearest neighbors with similar characteristics will be selected as the characteristic of the selected focal point. We can use this concept to make predictions.

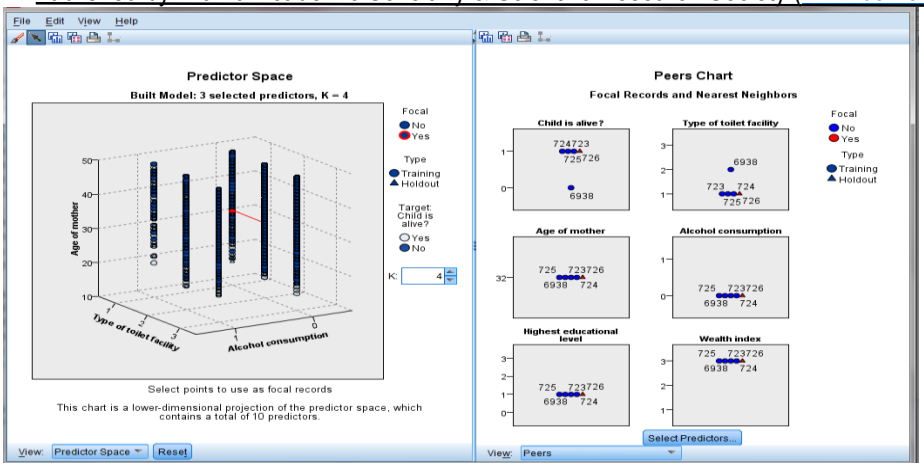

Figure 2 : Peers Chart

Figure 2 is a peer's chart which displays how a randomly selected focal point and its four nearest neighbors are represented in the predictor space taking into consideration the feature variables (independent).

3.2 Comparison of Models

Table 7 : Classification Table

Prediction		Percentage Correct (Child is Alive)		Overall Percentage
		Yes	No	
Logistic Regression		83.9	67.2	81.0
Kth Nearest Neighbor	Training	85.1	74.3	83.2
	Holdout	85.4	74.3	83.4

Table 8 : CCR Table

		Actual +	Actual -	CCR
Logistic Regression	**Predicted +**	8187	1566	**0.81**
	Predicted -	683	1400	
	Predicted -	726	1358	
Kth Nearest Neighbor	**Predicted +**	8303	1445	**0.83**
	Predicted -	536	1547	

The Correct Classification Rate (CCR) was used to compare the overall classification accuracy of the both models. Table 8 revealed that K Nearest Neighbor was the most efficient in modeling Mortality in Children under five with a CCR of 83% and Logistic Regression with a CCR of 81%.It is evident that the difference in the CCR amongst the models was quite insignificant. It can therefore be concluded that both models will also do a good job in predicting mortality in children under five years in Ghana.

4.0 CONCLUSION

The study sought to compare two different models in the prediction of mortality in children under five years. The models were Logistic Regression and K Nearest Neighbor. The study revealed that both Logistic regression and Neural Network will also do a good job in predicting mortality in children. The study also sought to identify the predictor variables that were most significant or important in the models. Taking a cursory look at the models, five variables cut across as the most significant or important in predicting mortality in amongst children under five in Ghana. Mothers' highest educational level, Age of mother at birth, Type of toilet facility used, alcohol consumption and wealth index were found to be the most important variables in predicting mortality in amongst children under five in Ghana across all models.

References

1. Bâra, A., & Lungu, I. (2012). *Improving Decision Support Systems with Data Mining Techniques*. INTECH Open Access Publisher.
2. Black, R. E., Morris, S. S., & Bryce, J. (2003). Where and why are 10 million children dying every year?. *The lancet*, *361*(9376), 2226-2234.
3. Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.
4. Doctor, H. V., Bairagi, R., Findley, S. E., Helleringer, S., & Dahiru, T. (2011). Northern Nigeria maternal, newborn and child health programme: selected analyses from population-based baseline survey. *The Open Demography Journal*, *4*(11-12), 11.
5. Drèze, J., & Murthi, M. (2001). Fertility, education, and development: evidence from India. *Population and development review*, *27*(1), 33-63.
6. Hand, D. J., & Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *160*(3), 523-541
7. Iram, U., & Butt, M. S. (2008). Socioeconomic determinants of child mortality in Pakistan: Evidence from sequential probit model. *International Journal of Social Economics*, *35*(1/2), 63-76.
8. Johnson, P. D., Hobbs, F., & Arriaga, E. E. (1982). *Techniques for Estimating Infant Mortality* (No. 8). US Department of Commerce, Bureau of the Census
9. Murthi, M., Guio, A. C., & Dreze, J. (1995). Mortality, fertility, and gender bias in India: A district-level analysis. *Population and development review*, 745-782.
10. Thatte, N., Kalter, H. D., Baqui, A. H., Williams, E. M., & Darmstadt, G. L. (2009). Ascertaining causes of neonatal deaths using verbal autopsy: current methods and challenges. *Journal of Perinatology*, *29*(3), 187-194.
11. Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*(Feb), 207-244.
12. World Health Organization, & UNICEF. (2014). Trends in maternal mortality: 1990 to 2013: estimates by WHO, UNICEF, UNFPA, The World Bank and the United Nations Population Division: executive summary.
13. You, D., New, J. R., & Wardlaw, T. (2011). Levels and trends in child mortality. Report 2012. Estimates developed by the UN Inter-agency Group for Child Mortality Estimation