APPLICATION OF MACHINE LEARNING TECHNIQUE FOR THE PREDICTION OF NEONATAL MORTALITY USING MULTIPLE RISK FACTORS

1Department of Computer Science, Nigerian Defence Academy, Kaduna State, Nigeria
2Department of Intelligence and Cyber Security, Nigerian Defence Academy, Kaduna State, Nigeria
3Department of Computer Science and Information Technology, Federal University Dutsinma, Katsina State, Nigeria

*Corresponding Author’s Email: pooodion@nda.edu.ng

ABSTRACT
It is a common knowledge that one of the major cause of neonatal mortality is high-risk birth, which can be identified through risk factors. Though, there are many risk factors associated with neonatal mortality (e.g., short birth interval, pre-natal care, etc.), most interventions of government and other agencies target birth based on a single risk factor (i.e., poverty) even though most neonatal deaths are not from the targeted risk factor, thus, failing to curb the prevalence of early-life mortality. Hence, data from Nigerian Demographic and Health Survey was gotten for this study and nine risk factors were used to predict neonatal mortality risk by applying Support Vector Machine algorithm to build a predictive model. KMeansSMOTE was used to solve the problem of class imbalance in the dataset, while model hyper-parameter tuning was applied to the SVM model to get a better prediction. Neonatal mortality risk was estimated as a function of nine risk factors. Risk factors chosen for the study were compared with four (4) risk factors from a previous study. The result gave a sensitivity of 78%, specificity of 44% and area under curve of 60% compared to using only four risk factors which has a sensitivity of 63%, specificity of 37% and area under curve of 50%. The result shows that having more risk factors gives a considerable improvement by predicting more neonatal deaths. This will aid researchers and governments to identify more risk factors causing neonatal deaths especially in Africa.

Keywords: Neonatal mortality, risk factors, high-risk birth, SVM, KMeansSMOTE.

INTRODUCTION
Early-life mortality has become one of the major problems facing the healthcare sector globally. It is estimated that globally, about 2.5 million children died before their first birthday in 2017 (WHO, 2018). A child’s risk of dying is highest during neonatal period (within 28 days of birth) (Darmstadt et al., 2003). According to WHO (2018), 47% of all under-five deaths in 2017 were among newborn infants.

The rate of neonatal mortality has been declining in most developing countries, however, that is not the case in Nigeria (Adebowale, 2017). Nearly 10% of child mortality globally was from Nigeria in 2016 (Owoseye, 2017). The ever high neonatal mortality rate in Nigeria was due to numerous reasons, one of which is high-risk birth (Mogford, 2004). A birth is considered to be of high-risk if there is a potential complication that could affect the baby (Coco et al., 2014). High-risk births in Nigeria remain a public concern because of the salient cultural practices like early marriages and high birth frequency (Adebowale, 2017). Nigeria is a high-fertility country with a total fertility rate of 5.5 and neonatal mortality rate of thirty-two (32) per one thousand (1000) live births (National Population Commission, 2014). High-risk births have been associated with high mortality in Nigeria.

LITERATURE REVIEW
In the health sector, early-life mortality is associated with socio-cultural, socio-economic, maternal, infant factor, proximate and delivery factors (Lamichhane et al., 2017). Maternal age affects the survival of a child during infancy (Issaka, 2016). It has contributed to high-risk pregnancies and births over the last twenty (20) years (Coco et al., 2014). Both adolescents and women of advanced reproductive age having their first child risk poor child health outcomes including death (Finlay et al., 2011). (Centre for Disease Control and Prevention (CDC) 2011; Fagbamigbe, 2017), opined that adequate pre-natal is useful and very essential to ensuring that full-term infants are at healthy weight. Studies suggest that infants born at low birth weight are at increased risk of certain health problems (Hovi et al., 2007). Hence, pre-natal care can reduce the risk of many leading causes to neonatal mortality (CDC, 2012).

In the investigation of DaVanzo et al., (2008), long birth intervals (at least five (5) years in length) and shorter birth intervals are also predictors of neonatal mortality. Shorter birth intervals were associated with higher neonatal mortality especially if the interval began with a live birth. Other risk factors like first birth order have also been identified in studies to contribute to early child mortality (Dest a, 2011).

Studies have shown high prevalence of early-life mortality in rural settlement than in urban areas (Adebowale, 2017). Urban-rural differences in quality of education and health infrastructures have been identified as some major factors that might be responsible for lower mortality in urban areas compared to rural areas (UNICEF, 2012). Mutunga (2007) also pointed out that neonatal mortality was lower for children who were at birth order two to three (2-3).
In order to mitigate the impact these risk factors have on early-life mortality, the international health organizations have been conducting surveys and collecting data from different countries. These data is studied to identify patterns on major causes of early-life mortality and take preventive action to reduce the rate at which it is occurring (Shamebo et al., 1994). Statisticians have over the years stepped in to apply statistical tools and techniques to assist decision makers in the healthcare sector (Islam et al., 2018). Various scientist have worked on identifying risk-factors associated with high risk birth so as to achieve the then Millenium Development Goal (MDG) of reducing early-life mortality by the year 2015, for which among the 195 countries, 68% failed to achieve it (Victora et al., 2016). This is because most of the researches focus on identifying a single risk factor most especially poverty (Houweling et al., 2010). Ramos et al., (2018) compared using only poverty as a risk factor to using (4) risk factors. The factors include; household income, maternal education, maternal age, and region. The result gotten when applied on a bayesian hierarchical model gave an accuracy of 57% for four risk factors compared with 30% when only one risk factor was used. The supervised classification model used was based on traditional statistics, which can provide ideal results when sample size is tending to infinity (Durgesh & Lekha, 2010) however, only finite samples can be acquired in practice.

RESEARCH METHODOLOGY
In this study, a machine learning based model called Support Vector Machine (SVM) was applied to nine (9) risk factors given from various literature which include age, income level, education, regions, residence, pre-natal, birth order number, birth interval, and religion were used in predicting the neonatal mortality in Nigeria. Other machine learning techniques used include KNN (K Nearest Neighbour) for handling missing values, KMeansSMOTE (K Means Synthetic Minority Oversampling Technique) for tackling the class imbalance and hyper parameter tuning was used to get a better prediction. The study adopted a hybrid methodology process because experts agree that the machine learning cycle is the same across all industries and should follow a defined course, hence the explanation for the hybrid methodology (Petersen, 2018). The methodology begins from data source collection, pre-processing, modelling and evaluation.

Research Area and Data Source
Nigeria was the area where the study was carried out based on the report of Nigerian Demographic report on family health. The source of data used for this study was gotten from Demographic Health Survey (DHS) website available at www.dhsprogram.com after submitting the abstract of the study. The Nigerian Demographic Health Survey (NDHS) 2015 dataset was used because it was the latest survey available as the survey is conducted every 5 years with the support of United States Agency for International Development (USAID).

Data Preparation and Pre-processing
This step involves certain techniques involved in processing raw and unrefined data in order to produce data that is reliable. Attribute selection, data cleaning, data normalization and data re-sampling were used in this phase.

Table 1: Attributes and their values

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>12, 13, 14, ……………., 49 years</td>
</tr>
<tr>
<td>Education Level</td>
<td>No education, Primary, Secondary, higher</td>
</tr>
<tr>
<td>Wealth Index</td>
<td>Poorest, Poorer, Middle, Richer, Richest</td>
</tr>
<tr>
<td>Birth Order Number</td>
<td>1, 2, 3, ……………., 21</td>
</tr>
<tr>
<td>Birth Interval</td>
<td>0, ……………., 68</td>
</tr>
<tr>
<td>Pre-natal</td>
<td>Attended, Not Attended</td>
</tr>
<tr>
<td>Residence</td>
<td>Urban, Rural</td>
</tr>
<tr>
<td>Regions</td>
<td>North West, North East, North Central, South West, South East, South South</td>
</tr>
<tr>
<td>Religion</td>
<td>Muslim, Christian, Traditional, No Religion</td>
</tr>
<tr>
<td>Survive</td>
<td>Yes, No</td>
</tr>
</tbody>
</table>

Data cleaning
Data cleaning is a technique for handling missing data and for smoothing noisy data. In this study we use exploratory statistics to search for missing values. Table 2 presents attributes with missing values.

Table 2: Attributes showing valid and missing values

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Age</th>
<th>Education</th>
<th>Wealth</th>
<th>BORO</th>
<th>Birth Interval</th>
<th>Pre-natal</th>
<th>Residence</th>
<th>Regions</th>
<th>Religion</th>
<th>Survive</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6524</td>
<td>6524</td>
<td>6524</td>
<td>6524</td>
<td>4265</td>
<td>6524</td>
<td>6524</td>
<td>6524</td>
<td>6524</td>
<td>6524</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2260</td>
<td>2184</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

As shown in table 2, birth interval and pre-natal contains missing values. We used KNN based imputer to impute the missing values. Each attribute missing values were imputed.
using the mean values of the nearest neighbours found in the training set. It computes the missing value in terms of a standard Euclidean distance. We measure the Euclidean distance between two points in a sample, say, $x_1 = (x_{11}, x_{12}, ..., x_{1k})$ and $x_2 = (x_{21}, x_{22}, ..., x_{2k})$ as the square root of the sum of the squared differences between a new point $(x_1)$ and an existing point $(x_2)$.

$$Euclidean\ dist(x_1, x_2) = \sqrt{\sum_{i=1}^{k}(x_{1i} - x_{2i})^2} \quad (1)$$

**Data Normalization**

Normalization helps to stop attributes with originally wide ranges from outweighing attributes with originally smaller ranges which causes high variance. Age, birth interval have higher ranges of numbers compared to pre-natal which has binary values. Min-max normalization was used and it is computed as

$$\nu^I = \frac{\nu - \min_A}{\max_A - \min_A} \quad (2)$$

Where; $\nu^I$ is the normalized attribute value, $\nu$ is the transformation value, $\max_A$ is the highest value of attribute A, $\min_A$ is the smallest value of attribute A.

**Data Re-sampling**

Re-sampling problems arises when classifiers are to classify categorical variables. The classification algorithm is faced with a problem known as “Class Imbalance”. A dataset is said to be imbalanced if the classification categories are not equally represented as in this case.

<table>
<thead>
<tr>
<th>Name: Survive, dtype: int64</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 6264</td>
</tr>
<tr>
<td>1 260</td>
</tr>
</tbody>
</table>

![Figure 1: Class Distribution of the Dataset](image)

Figure 1 shows how the target distribution of the dataset before re-sampling. To solve this problem of class imbalance, KMeansSMOTE which is an intelligent re-sampling technique that is very good for generalization was used. It uses a KMeans clustering method before applying SMOTE. It cluster group samples together before generating new samples depending on the cluster density while SMOTE create new synthetic samples of the minority class to even the dataset based on spatial Euclidean distance between samples (Nitesh et al., 2002). Of the six thousand five hundred and twenty-four (6524) data, Six thousand two hundred and sixty-four (6264, 96%) are neonates that survived while two hundred and sixty (260). The SVM require the solution of the following optimization problem:

$$\min_{\omega, \xi} = \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{l} \xi_i \quad (3)$$

Here, training vectors $x_i$ (i=1, 2, 3, ...., l) are mapped into a higher-dimensional space by the function $f$. Then the SVM finds a linear separating hyper-plane ($w, b$) with the maximal margin in this higher-dimensional space. $C > 0$ is the cost parameter of the error term. The slack variables, $\xi_i$ (i=1,2,3,.....,l) measure the degree of misclassification of $x_i$. The SVM does not require an estimation of the statistical distributions of classes to carry out the classification task, but it defines the classification model by exploiting the concept of margin maximization.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (4)$$

MODELLING

SVM algorithm was used in creating the machine learning model. It classifies the target samples into whether the new born survived or not. Given a training set of data points n pairs $x_n, y_n$ with $y_n = 1, -1$, denotes a constant, a class to which $x_n$ belongs and $n$ = number of samples. The SVM does not require an estimation of the statistical distributions of classes to carry out the classification task, but it defines the classification model by exploiting the concept of margin maximization.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (4)$$

Furthermore, equation 4 is called the kernel function, with $x_n, y_n$ with $y_n = 1, -1$, denotes a constant, a class to which $x_n$ belongs and $n$ = number of samples. The SVM does not require an estimation of the statistical distributions of classes to carry out the classification task, but it defines the classification model by exploiting the concept of margin maximization.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j) \quad (4)$$
solved. Although new kernels are being proposed by researchers, the radial basis function (RBF) is a reasonable first choice (Kaur & Singh, 2013). The RBF kernel nonlinearly maps samples into a higher-dimensional space, so it can handle nonlinear classification issues. With certain parameters (C, \( \gamma \)) the RBF kernel has the same performance as the linear kernel or the sigmoid kernel. There are two parameters while using RBF kernels: C and \( \gamma \). It is not known beforehand which C and \( \gamma \) are the best for one problem; consequently, some kind of parameter search must be done. Cross-validation is commonly utilized to identify good (C, \( \gamma \)) so that the classifier can accurately predict unknown (independent) data. A common strategy is to separate the data set into two parts, of which one is considered unknown. The prediction accuracy obtained from the “unknown” set reflects more precisely the performance of the model in classifying an independent data set.

In this study, we used the grid search for tuning the best parameter set and got (C=1000, \( \gamma =0.01 \)). The dataset was split into training set and testing set with 80% for training and 20% for testing. The training set was used to train the SVM model, while the independent testing set was used to test the model’s performance.

**Model Evaluation**

Model Evaluation is the process of assessing how well a machine learning model performed against real data. The evaluation metrics used in this study are sensitivity, specificity and Area Under Receiver Operating Curve (AUROC). These metrics are more suitable for imbalanced health dataset.

**Sensitivity (also known as True Positive Rate)**

When the actual value is positive, how often is the prediction correct? It is the ratio of true positives to the sum of true positives and false negatives computed mathematically as

\[
Sensitivity = \frac{TP}{TP + FN}
\]  
(5)

Sensitivity in this research is the measure that tells what proportion of the infants that had actually died and was predicted by the algorithm as having died.

**Specificity (also known as True Negative Rate)**

Specificity relates to the ability of the classifier to identify negative results. It is the ratio of true negatives to the sum of true negatives and false positives which is calculated as

\[
Specificity = \frac{TN}{TN + FP}
\]  
(6)

In this research, specificity is the proportion of children that survived and were predicted as survived.

**Threshold**

Threshold is used to convert predicted probabilities into class predictions. Threshold can be adjusted to increase specificity or sensitivity though it is one of the last thing to be done in model building. In this study, threshold of the model was moved to 40% for the maximization of the sensitivity while keeping other metrics relatively good.

**Receiver Operating Characteristic (ROC) Curves**

ROC is the graph plotted to choose a threshold that balances true positive rate and false positive rate in which each dimension is a strict columnar ratio so do not depend on a class distribution. Area under the Curve (AUC) or Area under Receiver Operating Curve (AUROC) is a portion of the area of the unit square, and its value will always be between 0 and 1.

**RESULTS AND DISCUSSION**

Result

![Feature Importance](image)

**Legend**

<table>
<thead>
<tr>
<th>Probability</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.218503</td>
<td>22%</td>
</tr>
<tr>
<td>0.205235</td>
<td>21%</td>
</tr>
<tr>
<td>0.19874</td>
<td>20%</td>
</tr>
<tr>
<td>0.105631</td>
<td>11%</td>
</tr>
<tr>
<td>0.084264</td>
<td>8%</td>
</tr>
<tr>
<td>0.073347</td>
<td>7%</td>
</tr>
<tr>
<td>0.053519</td>
<td>5%</td>
</tr>
<tr>
<td>0.033006</td>
<td>3%</td>
</tr>
<tr>
<td>0.027754</td>
<td>3%</td>
</tr>
</tbody>
</table>

Figure 2: Risk Factor importance
Figure 2 shows the importance or significance of the risk factors obtained from the dataset. The legend besides it by the right presents the percentage of each risk factor used for the study.

![Confusion Matrix](image)

Figure 3: Model Confusion Matrix

Figure 3 present the confusion matrix of the model. The model has a sensitivity of 78% and specificity of 43%.

![ROC of the Model](image)

Figure 4: ROC of the Model

Table 3: Comparative Analysis of the Two Experiments

<table>
<thead>
<tr>
<th></th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM with 4 Risk Factors</td>
<td>63</td>
<td>37</td>
<td>50</td>
</tr>
<tr>
<td>SVM with 9 Risk Factors</td>
<td>78</td>
<td>43</td>
<td>60</td>
</tr>
</tbody>
</table>
DISCUSSION OF RESULT
The result presented in figure 2 shows that all the risk factors should have an impact on the model output. However, birth interval is the most significant to the survival of a neonate, while the second most significant is age, followed by regions, BORD, wealth, prenatal, education, residence and religion respectively. Since none of the factors was less than 3%, all the risk factors were used and fed to the model.

Figure 3 presents the confusion matrix of the model. It predicts 38 out of 49 neonate deaths resulting in having 78% sensitivity. Though, the model only predicted 540 out of all the 1256 survived neonates resulting in 43% specificity. Since, we are more concerned with predicting neonate deaths, specificity is not as important as sensitivity. Figure 4 show that the model has an AUROC of 60%. This shows that the probability in percentage for predicting an instance with the available information, whether each class taken at random is 60% (i.e the probability of predicting the survival or death of a neonate is 0.6).

This approach of using nine (9) risk factors was compared with the approach of Ramos et al. (2018) that used four (4) of the risk factors.

Table 3 shows that the performance of the model with nine (9) risk factors gave a better performance. It also shows that having nine (9) risk factors gave a sensitivity of 78%, specificity of 43% and AUROC of 60% while with four (4) risk factors, it gave a sensitivity of 63%, specificity of 37% and AUROC of 50%. This further confirms that having more risk factors predicts more deaths. The findings from this study showed the model with 9 risk factors predict neonate deaths 78% of the time which is an improvement on the previous study. This shows that the model is more sensitive towards neonates death compared to survived neonate leading to an AUROC of 60%. More important risk factors need to be identified to get better prediction.

SUMMARY AND CONCLUSION
The aim of this study is to apply a Machine learning techniques to predict neonatal mortality by using multiple risk factors to identify high-risk birth in Nigeria. In solving the problem, the study used a hybrid methodology to achieve the project goals. The process involved research area and data source, data preparation and pre-processing, modelling and evaluation. The research used the 9 risk factors. This factors are; wealth index, education, region, residence, age, religion, preceding birth interval, prenatal care, birth order number to predict neonate death which were all gotten from NDHS 2015 dataset. An SVM model was built for the prediction of neonate death. In the process of modelling, it was discovered that the data distribution of the target class were imbalanced and affected the model building. Therefore, intelligent re-sampling technique called KMeansSMOTE to even the distribution and improve the model so as to achieve the project objective.

The 9 risk factors identified were compared with the 4 risk factors used in a previous study. The result with 9 risk factors 78% of neonate deaths compared to 63% of deaths when 4 risk factors identified from previous study were used. The results gotten from this study show that having more risk factors gave a superior performance as elaborated in the results section. This confirmed that more risk factors are needed to predict neonate deaths accurately.

The findings from this research will help policy makers both within and outside Nigeria to develop programs to assist in reducing neonatal mortality by targeting high-risk births. It will also encourage other researchers to work on identifying additional risk factors that could give an improved prediction of neonatal mortality.

REFERENCES
APPLICATION OF MACHINE...


©2020 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.