Random Forest Regression Model For Estimation of Neonatal Levels In Nigeria

C. Managwu[#1], D.Matthias[*2], N. Nwaibu[#3]

[#] *Dept. of Computer Science, Rivers State University, Port Harcourt, Nigeria*

Abstract

Health is considered to be a fundamental necessity of all living beings, especially humans, hence its availability is a critical component of the Human Development Index (HDI) measurement of human health. Neonatal (NMR) and infant (IMR) mortality rates are not topics that require too much time or energy for most of us on our day-to-day thinking, but they are also vital to guaranteeing our standard of life. Nigeria's attempts to minimize under-five mortality have been skewed against neglect of neonates in favour of childhood mortality, and as such the literature lacks adequate knowledge to estimate neonatal rates accurately. Awareness of the rate of neonatal deaths as well as neonatal mortality determinants is important for the design of intervention programs to improve neonatal survival. Hence, this research was conducted to apply Random Forest Regression in predicting neonatal death rates to help better prepare healthcare systems for future occurrences.

This model will be educated using the existing dataset of actual neonatal death rates from 1970 through to 2018.

Keywords — *Neonate, Neonatal Mortality, Big Data, Human Activity Pattern.*

I. INTRODUCTION

Having automated, high-quality results predictions to ensure fair access to knowledge regardless of where they are located, device and current state of health is of great importance. Specifically, reliable forecasts of future performance are a high priority for companies, because they are important for smooth operation, customer loyalty and explain how to manage them, functionality and product safety features. Highly imprecise forecasts can produce very extreme results in the field. Cooperation between undertakings may also suffer from misunderstandings due to bad predictions.

Considering that predictions of machine learning are a very complex problem, the output provided by predictive machine learning models still needs to be continuously retrained and accepted to ensure the required quality. Easy use of software to forecast health outcomes therefore moves the question from growth to evaluation and correction, but does not address it. Consequently, the assessment of the prediction model represents a significant step in minimizing time and costs for companies, and in finding an appropriate way to ensure that the best solution is used as applicable to the application.

Access to computerized systems which make near-correct predictions based on previously trained data is still revolutionary. Regarding this problem, it is necessary to be able to rank the quality of a given prediction-acceptability cannot be assured of a predicted value. Due to their high quantity of in-selling company which further enhances the ability to automatically predict results, the technical documentation places special emphasis on it. Nowadays companies manage the specified prediction problem by outsourcing this task to outside sources. Since the person requesting these predictions does not know the source of the data, it is important to ensure that a person as ordered does the work properly and appropriately and not through an automated predictive system that is poorly trained.

In this context, the objective of this study is to select and implement a machine learning method that generates an algorithm that can predict neonatal deaths on a stable dataset for values on which it has not yet been educated.

II. RELATED WORKS

[1] Used the Random Forest Ensemble technique to predict maternal mortality using Kano State, Nigeria as a case study. This study used different ensemble techniques to determine the predictive performance of different classification trees, predictive relationship of the selected model The best candidate model among the competing models was selected based on information criteria as well as diagnostic checking.

In an attempt to predict the direction of stock market prices, [2] focused on the use of the Random Forest Ensemble technique. The authors propose to minimize forecasting error by treating the forecasting problem as a classification problem, a popular suite of algorithms in Machine learning. The learning model used is an ensemble of multiple decision trees.

[3] Used a data mining approach to predict the daily Internet data traffic of a smart university. The researchers performed data mining analysis was using various learning algorithms such as the Decision Tree, the Tree Ensemble, the Random Forest, and the Naïve Bayes Algorithm on KNIME (Konstanz Information Miner) data mining application and kNN, Neural Network, Random Forest, Naïve Bayes and CN2 Rule Inducer algorithms on the Orange. A minimum accuracy of

55.66% was observed for both the upload and the download IP data on the KNIME platform while minimum accuracies of 57.3% and 51.4% respectively were observed on the Orange platform.

[4] Examined the Nigerian Informal Sector using Data collected between 2014 and 2016 by the National Commission on Salaries, Incomes and Wages (NSIWC) on the informal sector of Nigerian economy. The gathered data was analysed using the algorithm Random Forest Ensemble. This thesis shows the importance or consequences of using data mining methodologies over traditional statistical data analysis.

III. METHODOLOGY

A. *Quantitative Research Methodology*

The method chosen for this analysis is structured according to Quantitative Methodology. Quantitative methods emphasize objective measurements and the statistical, mathematical, or numerical analysis of data collected through polls, questionnaires, and surveys, or by manipulating pre-existing statistical data using computational techniques [5]

Your goal in performing quantitative research studies is to establish the relationship between one aspect [an independent variable] and another within a population [a dependent variable or outcome variable]. Quantitative study designs are either descriptive or experimental [subjects assessed before and after treatment]. A descriptive study establishes intervals only; an experimental study establishes causality.

Quantitative research deals with numbers, logic, and goal position. Quantitative research focuses on numerical and unchanging data, and detailed, convergent reasoning rather than divergent reasoning [i.e. the spontaneous, free-flowing generation of a variety of ideas about a research problem].

Its main features are:

1) The data is usually collected using structured research tools.

2) The results are based on larger population-representative sample sizes.

3) *Due to its high reliability the analysis study may normally be replicated or repeated.*

4) Researcher has a clearly identified investigative issue to which reasonable answers are obtained.

5) All aspects of the analysis shall be carefully planned before collecting the data

6) Data is in the form of numbers and statistics, often shown in tables, maps, figures or other non-textual forms

7) Projects may be used to generalize ideas, predict future results or examine causal relationships.

8) Researchers use tools to collect numerical data, such as questionnaires or computer software

A quantitative research study's ultimate aim is to classify characteristics, count them, and construct statistical models in an effort to clarify what's being observed.

The data used for this research work was obtained from UNICEF Neonatal mortality rate from 1970 - 2018.

IV. ANALYSIS OF THE SURVEY DATA AND INTERPRETATION OF RESULTS

A. *Preliminary Analysis*

It is recommended that a lengthy time series data is required for univariate time series forecasting. [6] Recommended that at least 50 observations should be used for such a univariate time series forecasting. If few observations are used this could be problematic. Furthermore, when using data from long time series, it may be common for the series to include a structural break that may involve either the review of a sub-section of the entire data series or the alternative use of interference analysis or dummy variables.

B. *Results and Comparisons*

This section shows the predicted results gotten from the random forest regressor using a total of 100 trees.

TABLE I
Actual and Proposed Model Comparison

Year	Actual Value	Proposed Random Forest Regression Model	Proposed Model vs. Actual value error
1970	67.08845	62.279	4.809649
1971	65.83819	62.279	3.55939
1972	64.52375	62.279	2.244942
1973	63.13051	62.279	0.851705
1974	61.7353	61.707	0.02853
1975	60.30371	60.64	-0.33628
1976	58.95394	60.571	-1.61665
1977	57.53959	57.222	0.317253
1978	56.19048	55.841	0.349719
1979	54.91593	55.005	-0.08898
1980	53.71838	53.892	-0.17405
1981	52.64495	52.834	-0.18862
1982	51.74515	52.003	-0.25782
1983	51.01215	51.132	-0.12022
1984	50.41907	50.573	-0.15407
1985	50.04392	50.136	-0.0919
1986	49.8471	49.911	-0.06356
1987	49.77168	49.806	-0.03446
1988	49.78312	49.793	-0.00988
1989	49.85524	49.853	0.002713

1990	50.03407	50.013	0.02092
1991	50.29586	50.225	0.071303
1992	50.57308	50.39	0.182961
1993	50.73111	50.696	0.035243
1994	50.80556	50.755	0.050595
1995	50.74352	50.721	0.022393
1996	50.49098	50.537	-0.04616
1997	50.05557	50.064	-0.00803
1998	49.35778	49.442	-0.08421
1999	48.50953	49.12	-0.6105
2000	47.54629	47.12	0.4264
2001	46.48953	45.865	0.624646
2002	45.28632	45.42	-0.13337
2003	44.06004	44.273	-0.21323
2004	42.84464	43.11	-0.26581
2005	41.75377	42.018	-0.26459
2006	40.80642	40.923	-0.11688
2007	39.99493	40.173	-0.17825
2008	39.31453	39.462	-0.14751
2009	38.75754	38.896	-0.13874
2010	38.26544	38.355	-0.0899
2011	37.87577	38.085	-0.20903
2012	37.57116	37.74	-0.16888
2013	37.30283	37.234	0.068349
2014	37.03226	37.084	-0.05163
2015	36.82568	36.877	-0.05093
2016	36.61278	36.66	-0.04708
2017	36.40367	36.409	-0.00527
2018	36.0093	36.121	-0.11122

C. *Visualizing the output*

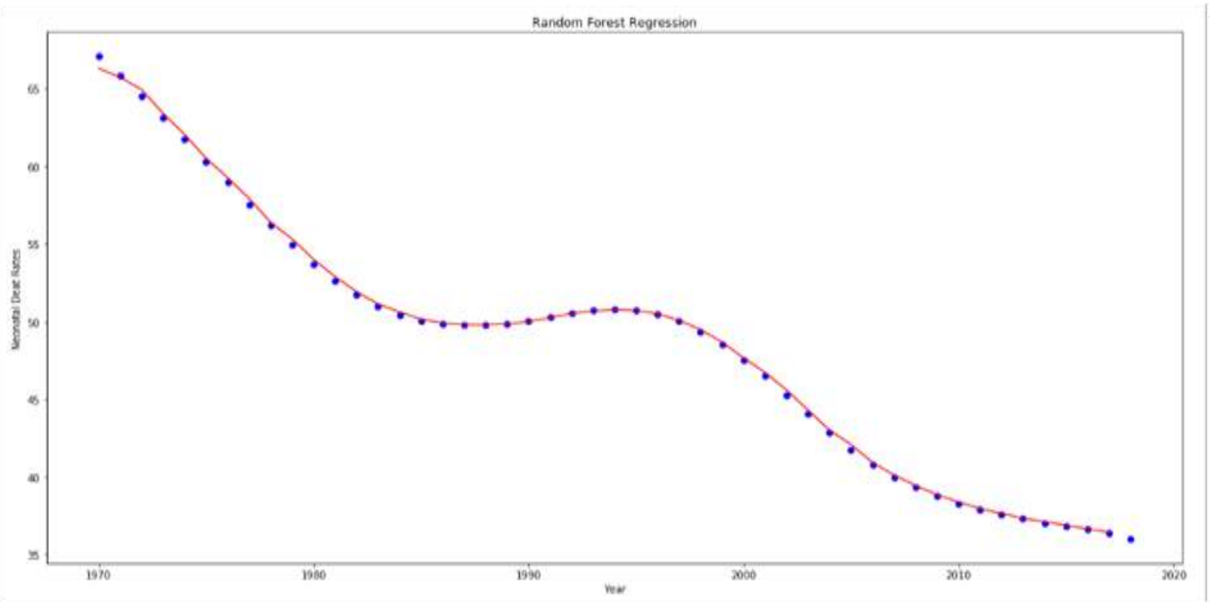

Fig 1: Actual values (blue) vs. Predicted values (red)

TABLE III
Accuracy measures for Random Forest Regression

Property	Value
MAE (mean absolute error)	0.5518035570200055
RSME (root mean square error)	0.8264859357303544
MSE (mean squared error)	0.6830790019600796

V. CONCLUSION

The objective of this research was to develop a supervised machine learning model to predict neonatal death rates Data from 1970 to 2018, on yearly bases, were collated from United Nations Children's Fund (UNICEF). An analysis on the yearly neonatal death rates assumed stable mean except during the years 1990 to 1995, which recorded significantly higher neonatal deaths as a result of extreme deplorable health and childbirth conditions

Random Forest Ensemble technique was used in modeling the data in the Pandas Library. The study identified several Decision tree models which best fitted the data. After the estimation of the parameters of selected models, a series of accuracy tests were performed.

As we increase the number of retained features the accuracy of the model increases as expected. Although not significantly the training time also increases

VI. RECOMMENDATIONS

In light of the knowledge acquired from this research, the tremendous value contribution to academic research and to the Health organizations and even to users, the researcher is recommending that this model be deployed alongside other forecasting machine learning ensemble techniques as tertiary predictive parameter in healthcare.

REFERENCES

[1] Abdulkarim, Kamaluddin & Kajuru, Jibril & Kurfi, Usman & Iliyasu, R. (2017). "*Application of Data Mining Technique To Maternal Mortality Prediction in Katsina State, Nigeria*".

[2] Saha, Snehanshu. (2016). "*Predicting the direction of stock market prices using Ensemble Learning."* 10.13140/RG.2.1.3202.6482.

[3] Adekitan, A.I. & Salau, Odunayo. (2019). "*The impact of engineering students' performance in the first three years on their graduation result using educational data mining*". Heliyon. 5. e01250. 10.1016/j.heliyon.2019.e01250.

[4] Paul, E. ., & Olumide, O. . (2019). "*Data Mining Approaches in the Study of the Nigerian Informal Sector*". I. J. Of Advances in Scientific Research and Engineering-IJASRE (ISSN: 2454 - 8006), 5(10), 237-250. https://doi.org/10.31695/IJASRE.2019.33565.

[5] Babbie, Earl R. "*The Practice of Social Research. 12th ed. Belmont*", CA: Wadsworth Cengage, 2010; Muijs, Daniel. Doing Quantitative Research in Education with SPSS. 2nd edition. London: SAGE Publications, 2010.

[6] Meyler, Aidan & Kenny, Geoff & Quinn, Terry, 1998. "*Forecasting irish inflation using ARIMA models*," MPRA Paper 11359, University Library of Munich, Germany.

[7] Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). "*Time series analysis, forecasting and control (3rd Ed.)"*. New Jersey: Prentice Hall, Englewood Clifs.

[8] Caiado, J. (2009). "*Performance of combined double seasonal univariate time series models for forecasting water consumption*". Munich: Munich University Personal RePEc Archive.

[9] Raschka, S., (2015). "*Python Machine Learning*", Packt Publishing Ltd., Birmingham, UK.

[10] Brownlee, J. (2016). "*Supervised and Unsupervised Machine Learning Algorithm*". Machine Learning Mastery, Volume 16, Issue 3, March 2016.

[11] Caldwell, John. (1979). "*Education as a Factor in Mortality Decline an Examination of Nigerian Data. Population Studies*" 33. 10.2307/2173888.

[12] Cech, T. V. (2005). "*Principles of water resources: History, development, management, and policy (2nd Ed.)"*. Hoboken: John Wiley and Son, Inc,

[13] Cheng, C. (2002). "*Study of the inter-relationship between water use and energy conservation for a building*". Energy and Buildings, 34, 261-266.

[14] Chenoweth, J. (2008). "*Looming water crisis simply a management problem*". Retrieved on November 20, 2012 from http://www.newscientist.com/ article/mg19926700.100-looming-water-crisis-simply-a-management-problem.html

[15] Cui, F. (2011). "*ARIMA models for bank failures: Prediction and comparison*". University of Nevada, Las Vegas.