A Comparative Model for Predicting Population Census in Nigeria

Orukpe A.
National population commission, Benin City, Nigeria

Imianvan A.
Department of Computer Science, University of Benin, Benin City, Nigeria, Nigeria

Akazue M.A.
Department of Computer Science, Delta State University, Abraka, Delta State, Nigeria

ABSTRACT
Population projections are increasingly employed as tools for understanding and modeling the economic, social and environment futures of a limited area. Population growth projection is the mathematical calculation that depend on the future rate, it is dependent on three main component of population assumptions, which are fertility, mortality and net migration of people of a country. Machine learning has been found useful for projecting future values. For population growth to be projected the machine learning has been applied to construct the map between year and population growth. This is germane for population, planning, budgeting, education, commercial sector and the health system. The study investigates the growth rate of Nigerian population data, employing time series projecting machine methods and analyzed by Linear Regression, k-NN, Support vector machine, and decision tree classifiers. The underlining projection method is hinged on the most accurate machine learning algorithm technique that flag less error rate. The increase and decrease of some example dataset does not impact the behavior of the method is equally analyzed. The outcome of the results reveals that the linear regression gives less error rate than the rest classifiers that was used to project population growth of Nigeria.

Keywords
Classifier, Linear regression, k-NN, Decision tree, and Support vector, Population, growth.

1. INTRODUCTION
Population projections are estimates of the population variables such as fertility. Mortality and net migration,

internal movement of people crossing an administrative boundary to another for various purposes, these components are used for future projection. Prediction and forecasting are projections but projection is not forecasting. Forecasting is based on assumptions that border on the condition that is expected to exist, while projection outlines one or more hypothetical form of actions that is expected to be adopted by the said organization. Predictions and forecasting are somehow difficult. But the advent of the machine learning

projection techniques, has now been simplified, very easy to be employed with the makeup of different computer computational algorithms which is the heart of its study. Machine learning methods are based on three format, supervised, unsupervised and re-enforcement learning. The inbuilt algorithms in these various machine learning are trained on already specified data set for training examples, assist the algorithm to be trained to predict from a dataset based on its previous training from a dataset, such as input and output manner, this method is supervised learning. While in the unsupervised learning the algorithm is handed over a collection of non - prearranged data and is bound to look for domiciled pattern relationship and attempt to annotate the data. We shall use four base learners in this work.

This study will help in population prediction as its changes, it will assist policy makers and private sectors for proper planning and budgeting, (NPC), projected 2006 census data to 2022, using spectrum5 algorithms project the need these data to plan and invest and make decisions. For the purpose of accuracy, it is good we utilize a reliable and efficient tool to project the population total, National Population Commission Nigeria (2020).

2. LITERATURE REVIEW
Employing machine learning techniques to compute the factors that are contributing to the population growth either increasing or decreasing is very important for the circle of government, budgeting, planning, education, commercial sectors etc, the linear regression use for the most sensitive demographic variables such as age and sex gives us an insight apart from the technical details of the isolated three attributes which are prediction, fit and interpretation were used, different machine learning algorithms were employed to analyze and compare them. Therefore, the vacuum years devoid of census in Nigeria was catered for by the utilization of these computational algorithms in machine learning. (Benjamin et al., 2018).

The population of Nigeria was adequately projected using the previous years of census data. Linear regression is applied to project the census data, and their outcome is easy to understand. (Gavril 2018), used the Swedish population growth to ascertain the effective performance of Linear Regression. No matter the size of the population composition we are dealing with, the linear regression can still project the dataset accurately and reveals the hidden estimates of the statistical relationship, (Poongodai, 2019)

Projection is a technique that projects the future data using the base population as a take-off point. Projections usually projects the data for planning, budgets and assisting the commercial sector who also need these data. There are dual techniques used for projecting for future growth rate, this comprises qualitative and quantitative. The qualitative their projection on the expert's opinions which are subjective, and the quantitative methods uses the obtained dataset and analytical method to avoid bias. So the dataset we use in this study is obtained from the Nigerian population database. Therefore, the use of machine learning algorithms on population estimation always make an important contribution to the country.

This help to facilitate the planning of national needs about the country and pave the way for more consistent social, economic and environmental decisions. In this study machine learning algorithms made more accurate population estimation than cohort components method. One of the most commonly used methods to produce population projection is the cohort

components method. Cohort components method is used to estimate the population by using variables such as the total population of the country, birth, death and migration rates, life expectancy at birth and sex ratio at birth. However, different variables can also affect the total population of the country. Machine learning algorithms analyses all variable and produce a more accurate population estimation, Gavril, (2018). With the new wave of Covid-19, projection method for population growth will be better, because the use of questionnaires to bring in interfacing with respondents will be abolished. Factors impacting population growth are many Machine learning algorithms analyses all variable and produce a more accurate population estimation. The variables that impacts population growth are thus stated:

Economic development. Countries who are in the early stages of economic development tend to have higher rates of population growth. In agriculturally based societies, children are seen as potential income earners. From an early age, they can help with household tasks and collecting the harvest. Also, in societies without state pensions, parents often want more children to act as an insurance for their old age. It is expected children will look after parents in old age. Because child mortality rates are often higher, therefore there is a need to have more children to ensure the parents have sufficient children to look after them in old age.

Education. In developed countries, education is usually compulsory until the age of 16. As education becomes compulsory, children are no longer economic assets – but economic costs. In the US, it is estimated a child can cost approx. $230,000 by the time they leave college. Therefore, the cost of bringing up children provides an incentive to reduce family size.

Quality of children. Gary Becker produced a paper in 1973 with H.Gregg Lewis which stated that parents choose the number of children based on a marginal cost and marginal benefit analysis. In developed countries with high rates of return from education, parents have an incentive to have a lower number of children and spend more on their education – to give their children not just standard education but a relatively better education than others. To be able to give children the best start in life, it necessitates smaller families. Becker noted rising real GDP per capita was generally consistent with smaller families.

Welfare payments/State pensions. A generous state pension scheme means couples don't need to have children to provide an effective retirement support when they are old. Family sizes in developing countries are higher because children are viewed as 'insurance' to look after them in old age. In modern societies, this is not necessary and birth rates fall as a result, Wenjie et al., (2019).

Social and cultural factors. India and China (before one family policy) had strong social attachments to having large families. In the developed world, smaller families are the norm, Samir, (2015)

Availability of family planning. Increased availability of contraception can enable women to limit family size closer to the desired level. In the developing world, the availability of contraception is more limited, and this can lead to unplanned pregnancies and more rapid population growth. In Africa in 2015, it was estimated that only 33% of women had access to contraception. Increasing rates would play a role in limiting population growth.

Female labour market participation. In developing economies, female education and social mobility are often lower. In societies where women gain a better education, there is a greater desire to put work over starting a family. In the developed world, women have often chosen to get married later and delay having children (or not at all) because they prefer to work and concentrate on their career.

Death rates – Level of medical provision: often death rates are reduced before a slowdown in birth rates, causing a boom in the population size at a certain point in a country's economic development. In the nineteenth and early twentieth century, there was a rapid improvement in medical treatments which helped to deal with many fatal diseases. Death rates fell and life expectancy increased, Blanpain, (2020).

3. RESEARCH DESIGN AND METHODOLOGY

We used different machine learning algorithms such as Linear regression model, Decision tree model, K -nearest neighbor model, and Support vector machine model to perform projection on the historical data from different years to ascertain the optimal performance on population total projection.

- We use Python database connector to link NPC database.
- How did we establish the connection with the database in python?
- Databases are sourced from where the data is stored.
- Import MySQL. connector
- Establish connection using connect ( )
- MySQL .connector. connect.(host = host name)
- User = user name
- Password = password
- Database = db-name
- Create cursor object
- Obj. name = conn-object. Cursor ( )
- Execute = show database
- Conn obj –name. execute( )
- Once the link to the NPC data base is fetch, then we proceeded to processing our data of reference.

Table 1: Population of male and female from 2006 to 2022 in Nigeria

YEAR	Total	MALE	FEMALE	
2006	140,431, 790	71,345,488	69,086,302	
2007	144,636,162	73,362,818	71,273,344	
2008	148,987,688	75,458,773	73,528,915	
2009	153,408,431	77,591,819	75,816,612	
2010	157,898,421	79,761,798	78,136,623	
2011	162,450,998	81,964,681	80,486,317	
2012	167,054,454	84,193,850	85,256,694	
2013	171,704,412	86,447,718	85,256,694	
2014	176,438,990	88,742,943	87,696,047	
2015	181,248,792	91,076,923	90,171,869	
2016	186,121,277	93,443,245	92,678,032	
2017	191,053,912	95,839,254	95,214,658	
2018	196,042,933	98,263,945	97,778,988	
2019	201,135,262	100,739,423	100,395,839	

2020	206,283,338	103,242,979	103,040,359	
2021	211,493,324	105,776,862	105,716,462	
2022	216,783,381	108,350,410	108,432,971	

(SOURCE: NPC, 2020)

The four algorithms, Linear regression, support vector machine, k-NN, and Decision tree were used for the projection and obtain the above result. The Nigerian population pyramid for 1963 census to 2006. Showing the population growth over the past 54 years.

4. PREDICTION BY LINEAR REGRESSION.

Linear regression is one among the most basic types of machine learning, in which we train a model that predict the data's actions which on certain variables. The two variables on the x axis and y axis should be linearly associated in linear regression. This estimation of future value is based on the historical data. Assume that 'x' and 'y' are two variables on the regression line. The value will be linearly upward, that is whenever 'x' increases 'y' will also increases, or if 'x' decreases the value of 'y' is also decreases, Poongodai, et al., (2019

Mathematically, a linear regression equation can be expressed as: y = a + bx Where a is y intercept of the line, b is the slope of the line, 'x' is independent variable and 'y' is dependent variable.

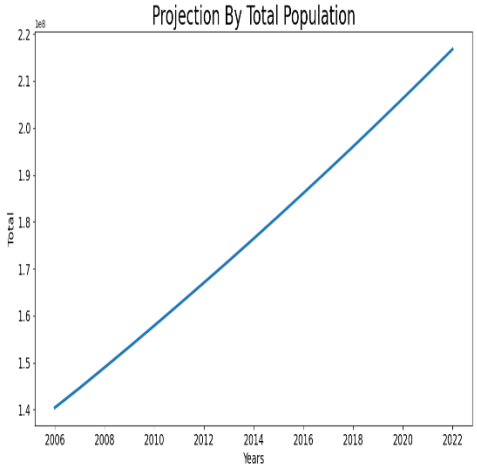

Figure 2: Prediction by Support Vector Regression(SVR)

Support Vector Regression(SVR) technique have been made to support multi-class classification and regression problems. The technique of Support Vector Machine used for regression function is called Support Vector Regression (SVR).

Support vector regression uses a function called SMO reg, which is used to predict the values based on the trained data set. When we run the tool with the function the results will be obtained. The Root Mean Squared Error value is also recorded with the Cross-validation folds' value is 12.

4.1 Prediction by Decision Tree Algorithm

Decision Tree algorithm is one of the techniques for supervised learning. Using decision tree algorithm, regression and classification problems can be solved. Decision tree is used to build a training model which is used to predict the target value. In decision tree the prediction starts from the root node and follow the branch node which has the corresponding value, and move towards the next node. This way prediction is done and the results are obtained, Caleb et al., (2017).

4.2 The K- Nearest Neighbor Algorithm

The K-Nearest Neighbors algorithm uses the whole data set as the training set, rather than splitting the data set into a training set and test. When an outcome is required for a new data instance, the k-NN algorithm goes through the entire data set to find the k-nearest instances to the new instance, or the k number of instances most similar to the new record, and then outputs the mean of the outcomes (for a regression problem) or the mode (most frequent class) for a classification problem. The value of k is user-specified. The similarity between instances is calculated using measures such as Euclidean distance and Hamming distance. The KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm. KNN is very easy to implement. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.) Chomboon, et al., (2015).

Table 2. Population growth projection using Linea regression, k NN, Decision tree and support vector machine

MACHINE. LEARNING	MALE	FEMALE	TOTAL POPULATION
LINEAR REGRESSION	100,739,423	100,395,839	201,135,262
K- NN	103,242,979	103,040,359	206,283,338
DECISION TREE	105,776,862	105,716,462	211,493,324
SUPPORT VECTOR MACHNE	108,350,410	108,432,971	216,783,381

Table 2: Projection of some states Population in 2023

Population growth projection using Linea regression, k NN, Decision tree and support vector

Table 3: MSE value obtained from projection methods

	Population	RMSE
LINEAR REGRESSION	201,135,262	3.4123
K- NN	206,283,338	4. 1942
DECISION TREE	211,493,324	4. 2351
SUPPORT VECTOR	216,783,381	4.6786

5. RESULTS AND DISCUSSION

In the dataset, year, growth of the population, male and female, is used to predict the future growth in value. Prediction output results of four machine learning techniques, Linear regression,

Support Vector Regression, k- NN and Decision tree were compared with mean square error value. The technique which gives less mean square error will be the best technique. The forecasting of population of male and female using Linear regression, Support Vector Regression, K- NN and Decision tree is in the table 2.

From the table value, the Mean square error that was got at cross-validation folds' value is 12 when the four methods of machine learning were employed. From these methods linear regression has the least mean square error which reveals that this is the most efficient form to project the population of growth of Nigeria.

That Linear projection perform better in regards to population projection of Nigeria with a lesser mean square error margin.

6. CONCLUSION

There has not been census in the past 16 years, there is need to develop a better prediction model, hence we design a system for projection of Nigeria population based on the previous census figure. The use of the four different machine learning algorithm, using different factors as metrics to evaluate the contributing variable to population growth. We realized that for the government of Nigeria to do better in area of the variables contributing to the population growth, they should hold on to good planning and budgeting so that these key variables won't suffer and thereby translating to the citizen wellbeing. Compared to other methods, our current understanding of population growth modelled quantitatively by regression does perform well. Future work, investing in other machine learning approaches, with other set of data as well as exploring the models. Finding new ways of analyzing and understanding population growth should be pursued. Since the populations are going on increasing, Nigeria Government have to take necessary steps to increase the resources and developments in all the aspects.

7. REFERENCES

[1] National Population Commission Nigeria (2020). Nigeria Population Projection and Demographic Indicators-State and National. Abuja, Nigeria.

[2] Benjamin Seligman, Shripad Tuljapurkar, David Rehkopf, Machi (2018). modeled Investing other learning approaches to the social determinants, SSM - Population Health.

[3] Wenjie Hu, Jay Harshadbhai Patel, Zoe-Alanah Robert, Paul Novosad, Samuel Asher, Zhongyi Tang, Marshall Burke, David Lobell, and Stefano Ermon. (2019). Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery. In AAAI/ACM Conference on AI, Ethics, and Society.

[4] Gavril Ognjanovski, (2018). Predict Population Growth Using Linear Regression Machine Learning Easy and Fun, https://medium.com/analyticsvidhya/predict-population-growth-using-linear-regression-machine-learningd555b1ff8f38.

[5] Poongodai, A Suhasini, R Muthukumar, R. (2019) "Regression Based on Examining Population Forecast Accuracy, "International Journal of Recent Technology and Engineering.

[6] Samir Mazidbhai Vohra, (2015) "Population Growth – India's Problem", PARIPEX – Indian Journal of Research,

[7] Biau, G. and Scornet, E. (2016). A random forest guided tour. Test, 25(2):197–227.

[8] Blanpain, N. (2020). Is the Ageing of the French Population Unavoidable? Economie et Statistique / Economics and Statistics, this issue.

[9] Caleb Robinson, Fred Hohman, and Bistra Dilkina. (2017).