Comparative Analysis of Different Machine Learning Models for Estimating the Population Growth Rate in Data-Limited Area.

Mohammad Mahmood Otoom [1†], Mahdi Jemmali, Yousef Qawqzeh, Khalid Nazim S. A. and Fayez Al Fay ez

Department of Computer Science and Information, College of Science in Zulfi, Majmaah University, Al-Majmaah 11952, Saudi Arabia

Summary
Human population growth plays a key role in any regional planning. However, in many data constraint environment, it is not possible to collect the required demographic data to predict the human population growth rate. In such a context, a tool that could help in predicting human population growth without the need to rely on historical data will be very helpful. This study compares different machine learning (ML) techniques namely k-nearest neighbors (kNN), decision trees, random forest and artificial neural network in their ability to predict the population growth rate ('pgr') of an area. The different demographic variables used to predict population growth rate are human population, population density, life expectancy at birth, female life expectancy at birth, infant mortality rate, under five mortality rate and total fertility rate. The study found that all the ML based models were able to predict the population growth rate with more than 90% accuracy. The top two models are kNN and Random forest with prediction accuracy of 96.47% and 95.42%. The study has demonstrated the relevance of ML models in predicting 'pgr' in data constraint environment.

Key words:
Machine Learning, Population Growth rate, k-nearest neighbors, decision tree, random forest, artificial neural network.

1. Introduction

Human population growth rate estimates are important for various research activities like emergency planning, resource allocation, food security and disease warning system. Commonly, human population growth rate estimates for an area require the historical population data to estimate the human population growth rate. However, it is not always possible to get complete and reliable historical statistics of a region especially in less-developed countries. Consequently, the models to estimate the population growth rate using the current population and demographic details are of considerable significance.

The studies have tried to estimate different demographic details in data constraint environment. One study in Nepal tried to estimate the post-earthquake human migration within Nepal by tracing human movement using their mobile phone data [1]. In another case study, total population in an area is dynamically estimated based on the mobile phone data [2]. Studies on India and Nigeria have tried to use Artificial Neural Network to predict the future population of the country [3,4]. These studies relied on the availability of historical demographic data like population, fertility, mortality, life expectancy, migration for the study area. A study in Peru has focused on estimating the population density of an area based on the area's census data, economic data and satellite based land condition data using different regression and tree based methods (like Random forest and Bayesian Additive Regression Tree) [5]. In a Bangladesh study, infant mortality rate of any area is estimated using iterative proportional fitting method [6]. However, to the best of our knowledge, studies have not focused on estimating population growth rate in data constraint environment.

Various demographic factors have been shown to influence the population growth rate. Total fertility rate (TFR) of the population represents the average number of children that women would have in their lifetime. This impacts the birth rate of population and, consequently, population growth rate. Most countries on higher side of the development have lower fertility rate, which contribute to their lower population growth rate [7]. Mortality and life expectancy of population are other drivers that affect the human population growth rate. Improved population health enables reduction in human mortality and increase in life expectancy of the population which leads to improvement in the population growth rate [8]. Population density is shown to improve the development status of the area [9]. Such improvement in area development status could influence the population growth rate [10].

Human population across the globe tends to follow a similar human population growth curve. In low development scenarios, the high birth and death rates

Manuscript received December 5, 2019
Manuscript revised December 20, 2019

causes low population growth rate. As the area develops, the death rates decrease while the birth rate remains high. Consequently, the country transitions from low to high population growth rate. In the later phases of development, both the birth rate and death rate remains low which lead to low population growth rate [8]. The existence of population growth pattern among the human population irrespective of the region of globe could play a role in estimating the human population growth rate in areas with lack of historical data. This study proposes an approach to estimate the population growth rate based on the data of demographic characteristics and the corresponding growth rates.

Predictive Analytics is an important practice that focuses on predicting the outcomes and trends of future based on current datasets. Machine learning (ML) is a set of techniques that could perform predictive analytics. The algorithms of these techniques try to learn from the existing data to improve their performance. These algorithms do not need any explicit programming for learning and performance improvement [11]. In ML, there are two major types of learning approaches namely unsupervised and supervised learning. Unsupervised learning approach is used for datasets, which are not labeled, i.e., data regarding output variable is not available for any of the data points in the dataset. Supervised ML approach is used for datasets where at least some data points are labeled, i.e., data regarding the output variable is available [12]. This paper explores the supervised ML algorithms to identify the most suitable algorithm for predicting the population growth rate of an area.

There are different Supervised ML techniques that are based on different types underlying working principle [13]. The techniques used in the study are decision tree, random forest, k-nearest neighbors (kNN) and Artificial Neural Network (ANN) are some of the commonly used machine learning techniques. A brief review of each of these techniques is provided in next section. This study aims to estimate the population growth rate using these supervised ML techniques.

2. Supervised ML Algorithms

2.1 Decision Tree and Random Forest

Decision tree based algorithm is logic-based algorithm. It is a multi-step process. In this algorithm, nodes are created at which separation of data points in dataset are created based on a feature/input parameter values. At each node, two or more branches could be created. Each branch created at the node represents a value that node could take. Each tree start with single node called root node followed with creation of sub-node at each branch and process continues to create a decision tree. The input parameter selected for a node is the one that could help in best partition of the data points in the data set for output parameter [14]. Some of the common metrics adopted for best-input parameter selection are Gini index [15] and Information gain [16]. This algorithm suffers from the risk of data over fitting. The common approach adopted is to stop the tree formation before the model perfectly fits all the data points [13].

Random Forest is an ensemble learning approach, which creates multiple decision trees. In this approach, a subset of input parameters is only used to create a single decision tree. This way multiple decision trees are created using different subsets of input parameters. Finally, it combines the results from different decision tree to give the final result [17].

2.2 k-nearest Neighbors (kNN)

kNN is an instance-based learning algorithm. kNN relies on the principle that a data point in the dataset would be present near the data point with similar values or properties [18]. Thus, in this approach, the unknown output parameter value of a data point is determined based on the nearest data point output parameter value. The location of data point is positioned in the n-dimensional space, where ‘n’ represents the number of features or parameters. The absolute position is not as important as relative distance between the data points. The distance metric is used to measure this relative distance. Commonly, the objective of a distance metric is to minimize the distance between similar data points and maximize the distance between dissimilar data points. Some of the commonly used distance metrics are Minkowsky, Manhattan, Chebychev and Euclidean [19].

2.3 Artificial Neural Network (ANN)

Artificial Neural Network is a perceptron-based algorithm. ANN consists of multiple units known as neurons. These units are interconnected to each other in pattern. These units are classified into three types namely input units, output units and hidden units. Input units take the input parameter values, which are processed at hidden units and output units provide the results of processing. At the hidden layer, weights are assigned to the input parameters and finally weighted sum for parameter values is computed [13].

Many times, the relationship between input and output parameters is not linear. Thus, in order for ANN to deal

with non-linearity, an activation function is applied to the input parameter value to enable ANN in creating non-linear mapping of input parameters on output values. The common activation functions used are Sigmoid or Logistic function, Hyberbolic tangent function (TanH) and Piecewise linear activation units (PLU) [20,21]. In ANN model, weights assigned to each input parameter are changed to improve the performance of ANN in estimating the output variable value. Feed forward backward propagation is a common approach used to alter the weights of input parameter at hidden layer [13].

3. Methodology

The methodology adopted for the current study uses five stages to ascertain the performance of different ML models used in the study. The different ML algorithms are implemented in Python. These stages are described in detail below:

3.1 Step 1: Dataset Collection Stage

This stage involves the collection of the dataset that will be used for predicting the human population growth rate. A data set consisting of eight attributes (seven input attributes and one output attribute) namely human population, population density, life expectancy at birth, female life expectancy at birth, infant mortality rate, under-five mortality rate, total fertility rate and population growth rate (Table 1). These attributes are chosen, as they are commonly known to have association with population growth rate. The data comprising of 205 countries from 1950 to 2015 was obtained from United Nations database [22] leading to the total dataset of 13530 data points.

Table 1: Attributes used in the study

Parameter	Description/ Values
Population (in thousands)	Range 13.77 to 1397028.55
Population Density (per sq.km)	Range 0.05 to 20098.4
Life expectancy at birth (years)	Range 18.907 to 83.801
Female life expectancy at birth (years)	Range 22.394 to 86.792
Infant mortality rate (infant deaths per 1,000 live births)	Range 1.473 to 290.442
Under-five mortality (deaths under age 5 per 1,000 live	Range 2.058 to 435.155
Total fertility rate (live births per woman)	Range 0.827 to 8.866
Class/output attribute [Population growth rate	Range -7.076 to 17.693

3.2 Step 2: Pre-Processing and Transformation Stage

This stage involves cleaning and transforming the data into the format namely CSV, which is suitable, the ML techniques used in the study. Initially, the data is cleaned, scaled and formatted. The output parameter 'population growth rate (pgr)' is transformed from continuous variable to categorical variable. The 'pgr' value is divided into three categories namely 'high growth rate' (HGR), 'medium growth rate' (MGR) and 'negative to low growth rate' (NLGR). The categorization of the 'pgr' values is performed using K-mean clustering. This approach allowed grouping of closely related values of 'pgr' in one single cluster. Accordingly, three such clusters are created which represent the three categories of 'pgr'.

3.3 Step 3: Feature Selection Stage

This stage is performed to reduce the number of data dimensions for avoiding dimensionality curse [23]. In this regard, the literature suggests two approaches namely feature combination or selection such that information loss is not significant. Selection approach is chosen for study. In this approach, the promising features are selected from the original features [24]. Pearson correlation coefficient technique is used in this study for selecting the input parameters. Accordingly, five out seven original input attributes namely, human population, population density, life expectancy at birth, under-five mortality rate and total fertility rate. The descriptive statistics of final dataset is shown in Table 2.

Table 2: Final attribute format used for the study

Attributes	***Unit***	***Population Growth Rate (Category)***		
		High	***Medium***	***Negative to Low***
	Number of Data points (%)	346 (2.56)	7205 (44.19)	5979 (53.25)
Population Growth Rate	Mean (Range)	7.39 (17.69 to 5.03)	2.66 (5.02 to 1.67)	0.68 (1.67 to -7.08)
Population	Mean (Range)	6.2 (9.69 to 2.62)	8.16 (13.96 to 2.72)	8.36 (14.15 to 3.38)
Population Density	Mean (Range)	2.79 (7.53 to -3)	3.4 (9.91 to -1.35)	4.27 (9.03 to -1.35)
Life expectancy at birth	Mean (Range)	4.11 (4.37 to 3.55)	4.02 (4.43 to 3.41)	4.21 (4.43 to 2.94)
Under-five mortality rate	Mean (Range)	4.15 (5.8 to 2.02)	4.64 (6.05 to 0.94)	3.34 (6.08 to 0.72)

Total fertility rate	Mean (Range)	1.65 (2.12 to 0.5)	1.66 (2.18 to -0.19)	0.91 (2.09 to -0.05)

3.4 Step 4: Performance Evaluation Stage

In this stage, the classification of 'pgr' data set in terms of corrected classified instances is analyzed from different ML techniques. The various classification techniques used for the study are kNN, decision tree, random forest and ANN. The dataset is divided into 80% training set and 20% test set using stratified sampling. The analysis of performance of different ML techniques in this stage is performed using training dataset.

The accuracy is used as the metric to measure the performance of models obtained from different techniques. Accuracy is determines the number of correctly classified instances. It is calculated as follows:

Accuracy= 100*(number of correctly calculated instances/total number of instances)

The accuracy metric is used to measure the overall model performance as well as model performance in predicting individual category of the 'pgr'. The baseline performance of the dataset is estimated using Zero Rule (ZeroR) classifier. This classifier ignores the predictors, i.e., input parameters and only focus on the class variable, i.e., output parameter. It only predicts the most frequent value or category in the output parameter.

The hyper-parameters used in each of the techniques are tuned to obtain the optimal results (Supplementary Table 1). The study used 3-fold cross-validation for the hyper parameter tuning based on training dataset.

3.5 Step 5: Population Growth Rate Prediction Precision

In this step, the trained models obtained from different ML techniques are used to predict the population growth rate of the test data.

4. Results and Discussions

The study performed classification of the demographic data using different ML techniques in Python. The results on the training dataset is computed and shown in Table 3. The results showed that the models based on different ML techniques are able to predict the 'pgr' with 93.15 to 100% overall accuracy in training data set as compared to only 53.25% overall accuracy from ZeroR model. The results showed that overall accuracy of correct classification of 'pgr' growth rate is highest for kNN and Random Forest technique and lowest for Artificial Neural Network (ANN) technique.

Table 3: Performance of different machine learning models in terms of overall accuracy, high growth rate (HGR) prediction accuracy, medium growth rate (MGR) prediction accuracy and negative to low growth rate (NLGR) prediction accuracy in training data set.

Machine Learning Model	Model Performance in terms of prediction accuracy			
	Overall	HGR	MGR	NLGR
ZeroR	53.25%	0%	100%	0%
kNN	100%	100%	100%	100%
Decision Tree	99.86%	98.84%	99.85%	99.92%
Random Forest	100%	100%	100%	100%
Artificial Neural Network (ANN)	93.15%	50.97%	93.28%	95.07%

In terms of predictability of models towards the individual 'pgr' category for training dataset is found variable across ML techniques. While, kNN and Random forest predicted the HGR, MGR and NLGR category of 'pgr' with 100 percent accuracy in training dataset. Other two models namely decision tree and ANN provided different performance towards different 'pgr' category. NLGR is the most accurate category with prediction accuracy ranging from 95.07% to 99.92%, while HGR is the least accurate category with prediction accuracy ranging from 50.97% to 98.84%. MGR prediction accuracy ranged from 93.28% to 99.85% across decision tree and ANN respectively. This indicates that while; overall performance of different ML techniques could be comparable. The performance at each category level may vary considerably across the techniques.

The performance of different ML models on test dataset is computed and shown in Table 4. The results showed that the models based on different ML techniques are able to predict the 'pgr' with 93.01% to 96.47% accuracy in test data set as compared to only 53.25% overall accuracy from ZeroR model. This suggest that all the ML models are able to predict the 'pgr' of any region based on region's demographic details namely total population, population density, life expectancy at birth, under-five mortality rate and total fertility rate. The kNN based ML model has shown highest predictive performance of 96.47% for 'pgr', while decision tree based ML model has shown least predictive performance of 93.01% for 'pgr'. This means that kNN based model is able to identify the 'pgr' category correctly in nearly 96 out of 100 instances, while decision tree based model is able to identify the 'pgr' category correctly in nearly 93 out of 100 instances.

Table 4: Performance of different machine learning models in terms of overall accuracy, high growth rate (HGR) prediction accuracy, medium growth rate (MGR) prediction accuracy and negative to low growth rate (NLGR) prediction accuracy in test data set.

Machine Learning Model	Model Performance in terms of prediction accuracy			
	Overall	HGR	MGR	NLGR
ZeroR	53.25%	0%	100%	0%
kNN	96.47%	82.03%	97.37%	96.42%
Decision Tree	93.01%	63.19%	93.88%	93.71%
Random Forest	95.42%	71.30%	95.82%	96.24%
Artificial Neural Network (ANN)	93.13%	49.50%	93.66%	94.80%

In terms of predictability of models towards the individual 'pgr' category for test dataset is found variable across ML techniques. In case of HGR category, the prediction accuracy of different models ranged from 49.50% to 82.03% as compared to 0.00% accuracy from ZeroR model. The kNN based ML model has shown highest predictive performance of 82.03% for HGR, while ANN based ML model has shown least predictive performance of 49.50% for HGR. The prediction performance in HGR category is low as compared to overall prediction performance parameters of model. Such low performance could be attributed to smaller percentage of data points related to HGR category in training set. This could lead to the model having insufficient number of samples of HGR category to train [13].

In case of MGR category, the prediction accuracy of different models ranged from 93.66% to 97.37% as compared to 100% accuracy from ZeroR model. The ZeroR model showed higher accuracy than ML models for MGR category as it assigns all the datapoints to the MGR category irrespective of their original category. The kNN based ML model has shown highest predictive performance of 97.37% for MGR, while ANN based ML model has shown least predictive performance of 93.66% for MGR.

In case of NLGR category, the prediction accuracy of different models ranged from 93.71% to 96.42% as compared to 0.00% accuracy from ZeroR model. The kNN based ML model has shown highest predictive performance of 96.42% for NLGR, while decision tree based ML model has shown least predictive performance of 93.71% for NLGR.

It is found that among the four models, kNN based model performed consistently better than other three models followed by Random forest. Decision tree is better than ANN in predicting HGR, while comparable with ANN in all other performance categories. Additionally, it is observed that maximum variation in the performance of models has been found in HGR category where coefficient of variation is 20.62% as compared to other categories where coefficient of variation ranged from 1.34 to 1.84%. Further, it is found that kNN and decision tree predicted MGR category with highest accuracy, random forest and ANN predicted NLGR category with highest accuracy.

The results of test dataset are similar to training dataset in terms of performance of different ML techniques. This suggests that model is not over fitting the data. The overall performance is comparable, but the performance at each category level may vary significantly across the techniques. Overall, it is found that kNN is the best classifier for the given type of dataset.

5. Conclusion

This study investigates the different ML techniques for predicting the population growth rate in the data constraint environment. The paper explores the ability of the demographic details of different areas across the globe in enabling the prediction of the population growth rate in any other area of the world. The paper explores and analyses the performance of models based on different ML techniques namely kNN, decision trees, random forest and Artificial Neural Network (ANN). The performance is evaluated using the accuracy of predicting the correct category of population growth rate. It is found that ML techniques could be used to predict the population growth of any area based on its demographic parameter namely population, population density, life expectancy at birth, under-five child mortality rate and total fertility rate.

Further, it is found that kNN is the best performing technique within the given range of predictor values as it showed high accuracy in the predicting the overall number of correct categories as well as individual correct categories. The future study should focus on using more different ML techniques and performance metrics to evaluate the performance of predicting population growth rate.

Acknowledgement

The authors would like to thank the Deanship of Scientific Research at Majmaah University for supporting this work under project number No. R-1441-45.

References

[1] Wilson, R., Erbach-Schoenberg, E. Z., Albert, M., et al., "Rapid and near real-time assessments of population displacement using mobile phone data following disasters: The 2015 Nepal earthquake." PLoS Curr., 8, 1–24 , 2016.

[2] Deville, P., Linard, C., Martin, S., et al., "Dynamic population mapping using mobile phone data." Proc. Natl. Acad. Sci., 111, 15888–15893 , 2014.

[3] Bandyopadhyay, G. and Chattopadhyay, S., "An Artificial Neural Net approach to forecast the population of India." arXiv:nlin/0607058, 1–11, 2006.

[4] Folorunso, O., Akinwale, A. T., Asiribo, O. E., and Adeyemo, T. A., "Population prediction using artificial neural network." African J. Math. Comput. Sci. Res., 3, 155–162, 2010.
[5] Anderson, W., Guikema, S., Zaitchik, B., and Pan, W., "Methods for estimating population density in data-limited areas: Evaluating regression and tree-based models in Peru." PLoS One, 9, e100037 , 2014.
[6] Rose, A. N. and Nagle, N. N., "Validation of patiodemographic estimates produced through data fusion of small area census records and household microdata." Comput. Environ. Urban Syst., 63, 38–49, 2017.
[7] Nargund, G., "Declining birth rate in Developed Countries: A radical policy re-think is required." Facts, views Vis. ObGyn, 1, 191–193, 2009.
[8] Bongaarts, J., "Human population growth and the demographic transition." Philos. Trans. R. Soc. B Biol. Sci., 364, 2985–2990, 2009.
[9] Raouf, B., Peeters, D., and Croix, D. D. Le, "Early literacy achievements, population density, and the transition to modern growth." J. Eur. Econ. Assoc., 5, 183–226, 2007.
[10] Ahlburg, D. and Cassen, R., "Population and development. In International Handbook of Development Economics" (eds. Dutt, A. K., and Ros, J.), Edward Elgar Publishing, Cheltenham, UK, pp. 316–327, 2008.
[11] Langley, P. and Simon, H. A., "Applications of Machine Learning and Rule Induction." Commun. ACM, 38, 55–64, 1995.
[12] Geron, A., "Hands on Machine Learning with Scikit-Learn & TensorFlow First." O'Reilly Media Inc., Sebastopol, USA, 2017.
[13] Kotsiantis, S. B., "Supervised Machine Learning: A Review of Classification Techniques." Informatica, 31, 249–268, 2007.
[14] [Murthy, S. K., "Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey." Data Min. Knowl. Discov., 2, 345–389, 1988.
[15] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., "Classification and Regression Trees." Routledge, New York, USA, 1984.
[16] Hunt, E. B., Marin, J., and Stone, P. J., "Experiments in induction." Academic Press, Oxford, England, 1966.
[17] Breiman, L., "Random Forests." Mach. Learn., 45, 5–32, 2001.
[18] Cover, T. M. and Hart, P. E., "Nearest neighbor pattern classification." IEEE Trans. Inf. Theory, 13, 21–27, 1967.
[19] Chomboon, K., Chujai, P., Teerarassammee, P., Kerdprasop, K., and Kerdprasop, N., "An Empirical Study of Distance Metrics for k-Nearest Neighbor Algorithm." In The Proceedings of the 3rd International Conference on Industrial Application Engineering 2015 Kitakyushu, Japan, pp. 280–285, 2015.
[20] Karlik, B., "Performance analysis of various activation functions in generalized MLP architectures of neural networks." Int. J. Artif. Intell. Expert Syst., 1, 111–122, 2015.
[21] Liao, Z. and Carneiro, G., "On the importance of normalisation layers in deep learning with piecewise linear activation units." In 2016 IEEE Winter Conference on Applications of Computer Vision, WACV 2016 Newyork, USA, pp. 1–9, 2016.
[22] United Nations, "World Population Prospects 2017". 2017.
[23] Bellman, R., "Adaptive Control Processes: A Guided Tour" . Princeton University Press, Princeton, USA, 1961.
[24] Kohavi, R. and John, G. H., "Wrappers for featue subset selection." Artif. Intell., 1, 273–324, 1997.