PREDICTION OF POPULATION GROWTH USING MACHINE LEARNING TECHNIQUES

Brintha Rajakumari S[1], Padmanabhan P[2] , Christy S[3], Nandhini M[4]

[1]Associate Professor, Department of Computer Science, Faculty of Arts and Science, BIHER, Chennai.

[2]Assistant Professor, School of Computing Science and Engineering, Galgotias University.

[3]Associate Professor, Department of Information Technology, Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Chennai, Tamil Nadu, INDIA.

[4]Assistant Professor, Department of Information Technology, RRASE College of Engineering, Anna University.

brintha.cse@bharathuniv.ac.in, padmanabhan5312@gmail.com, christymelwyn@gmail.com, nandhini706@gmail.com

Abstract

Population growthprediction shows the future rate of fertility, mortality and migration of people of a country. It is very important for the population and health system. Nowadays, Machine learning concepts are most growing and popular for predicting future values. In order to predict population growth, the machine learning concept applied to build the map between year and population growth. The paper investigates the population growth of Indian government population data using time series forecasting machine learning techniques and analyzed byLinear regression, Support Vector Regression, Multilayer perceptron and Decision tree classifier. The optimum prediction method is based on the technique which gives very less error rate. The increment or degradation of instances in datasets do not affect the performance of the techniques is also analysed. The obtained result shows that the linear regression gives less error than the other classifier to predict population growth of India.

Keywords: Classifier, Machine Learning, Linear Regression, Support Vector Regression, Multi Layer perceptron, Decision tree, Time series forecasting.

1. Introduction:

Population projections are an estimation of birth rate, demise rate and migration of population from one place to another used for future prediction. The prediction and forecasting is very difficult task. With the help of Machine learning prediction and forecasting can be done efficiently. Machine learning consists of different computer algorithms, from where it got the ability to learn.Prediction can be done with two different ways of machine learning techniques such as supervised learning and unsupervised learning. The algorithm is trained on a predefined set of training examples which permit the algorithm to get a prediction from a dataset in supervised learning. In the unsupervised algorithm, the algorithm is given a collection of unlabeled data and it must find pattern relationships and attempt to label the data.

2. Literature Review:

Machine learning approaches to the health of social determinants[1] describes a linear regression of age and gender. The three attributes Prediction, fit, and interpretability were made comparison with different machine learning algorithms. Missing Population in India[2],effectively predict the population in a region. Linear Regression is used to make predictions on the data and it’s results are easy to understand. Gavril Ognjanovski makes a prediction on the Swedish population growth[3]. The data may be of any size it forecast accuracymodel[4] and estimates the statistical relationship.

Other papers also highlights the various problems faced by over population. Variation in the population rate are discussed the current research and importance to provide the necessary sources to solve the problem. To find the accuracy of the regression analysis several alternative functions are used in the regression model.

3. Materials and methods

Forecasting is a technique that predicts the future value based on past value.Usually the information about the products and services is not predictable and most of them are uncertain.Forecasting mainly focus onlong-range planning, budgets and cost controls, future sales, new products in markets, Production of various products and operations. The types of forecasting methods include Qualitative methods and Quantitative methods. The qualitative method is based on subjective opinions got from one or more experts and the quantitative method is based on data and analytical techniques.

The dataset used here has been collected from the Census India from government website link contains the Decadal population census variation which covers the data of all the states of India form the year 1901 to 2011.

Table 1: Population in INDIA from the year 1901 to 2011

Year	Population	Male	Female
1901	238396327	120791301	117358672
1911	252093390	128385368	123708022
1921	251321213	128546225	122774988
1931	278977238	142929689	135788921
1941	318660580	163685302	154690267
1951	361088090	185528462	175559628
1961	439234771	226293201	212941570
1971	548159652	284049276	264110376
1981	683329097	353374460	329954637
1991	846421039	439358440	407062599
2001	1028737436	532223090	496514346
2011	1210854977	623270258	587584719

The dataset load into Weka machine learning tool and then normalize the data to ease out the prediction of growth rate with machine learning. The four columns (Year, Population, Male and Female) from the dataset has been taken and saved as csv file format which will be used for the Weka tool. The data is being processed with different machine learning algorithms like Linear Regression, Support Vector Regression, Multi Layer perceptron and Decision tree classifier techniques and find the Mean error value to find the approximation in the prediction.

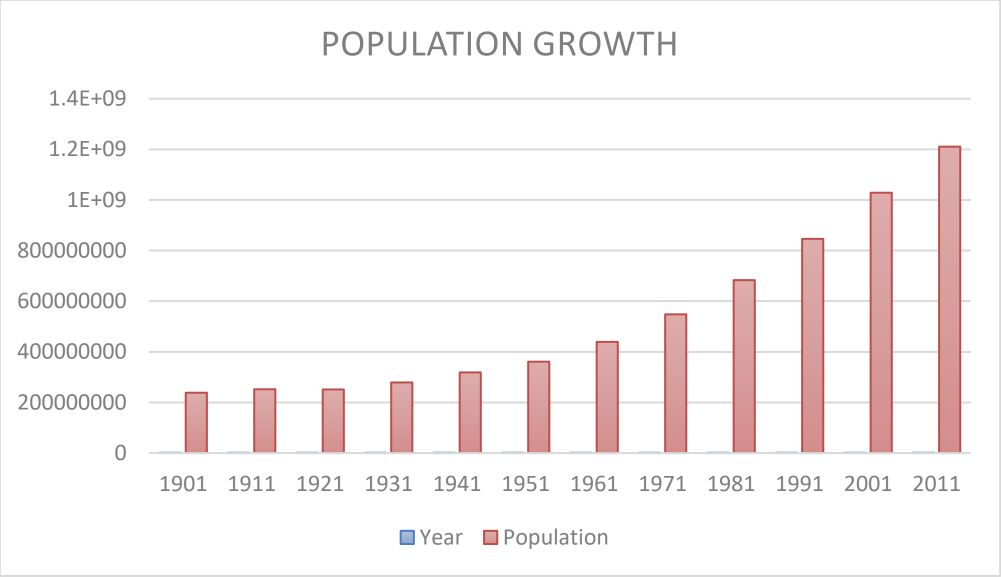

Figure 1: Population Growth in India from 1901 to 2011

4. Prediction by Linear Regression

Linear regression is one among the most basic types of machine learning, in which we train a model thatpredict 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.

For example consider a sales promotion can be evaluated by expecting to increase number of customers than the previous report. For this you have to increase the stalls, and many more staff to serve the customers. 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.

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.

5. Prediction by Support Vector Regression (SVR)

Support Vector Machine(SVM) technique have been made to support multi-class classification and regression problems. The technique of Support VM used for regression function is called Support Vector Regression(SVR). Support vector regression uses a function called SMOregwhich 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 10.

6. Prediction by Multilayer Perceptron

A challenge in the preparation of the data for time series forecasting using multi layer perceptron. In particular, observations of the lags need to be flattened into vectors of the features. The aim of this work is to provide each model’s standalone data which works on each type of time series data set thatdata adapt for specific time series.

6.1 Univariate MLP Models

Multilayer Perceptrons(MLP), is used to model forecasting univariate time series data set. Univariate time series are a dataset consisting of singly serial of temporally ordered observations. This model learn from the series data of past observations and it would predict the next data in the sequence.The data in the set must be prepared before anunivariate series can be modelled. The MLP algorithm will learn a function to map the past observations with the current values.

6.2 MLP Model

A simple MLP model has number of input layers, one hidden node layer, and an output layer which is used to make a forecast. The shape of the input is important in the definition; that is what the model expects data in terms of the number of time series steps to be input for each sample.

For each sample the input dimension is specified on the first hidden layer definition in the input dim argument. Technically, the model is viewed as a single feature each time step, instead of separate time series steps.

7. 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.

8. 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 techniquesLinear regression, Support Vector Regression, Multi layer perceptron and Decision treewere 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, Multi layer perceptron and Decision tree is in the table 2.

Table 2: Prediction of Population in 2021

	Population	**Male**	**Female**
Linear Regression	1488699942	766448362	716829820
SMOreg	1513670634	782050718	733597795
Multilayer Perceptron	1352767306	688999891	663786238
Decision Tree	1459675532	738921463	720754069

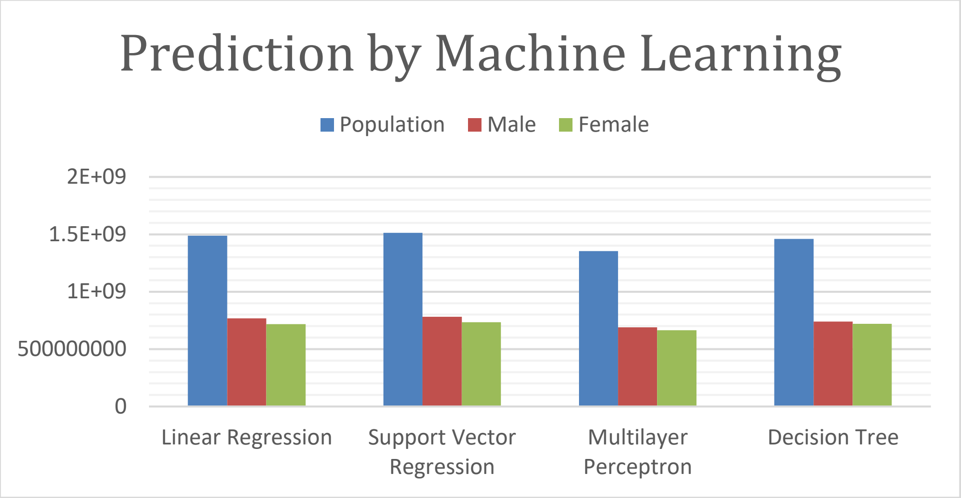

Figure 2: Population Prediction using Machine learning Techniques.

Population prediction results obtained for future decadal variation is shown in figure2. With the help of Machine learning prediction and forecasting techniques the results got different values for the population. The four algorithms, Linear regression, Support Vector Regression, Multi Layer perceptron and Decision tree were used for the prediction techniques and got the above result.

Table 3: MSE value obtained from Prediction Techniques

	Population	**RMSE**
Linear Regression	1488699942	3.5123
Support Vector Regression	1513670634	4.8745
Multilayer Perceptron	1352767306	4.2316
Decision Tree	1459675532	4.3821

The Root Mean Squared Error values are obtained at Cross-validation folds value is 10 while executing the four techniques of Machine Learning. Among these techniques Linear Regression has the minimum mean square error which proves that this is the best way to predict the population growth of India. Since India is suffering by Covid 19, a deadly virus that affects the population growth. This research gives the actual population growth prediction when the country is not affected by any natural disasters or pandemics.

9. Conclusion

Compared to other methods, our current understanding of population growth modelled quantitatively by regression does perform well. Future work, investing 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, Indian Government have to take necessary steps to increase the resources and developments in all the aspects.

References

[1] BenjaminSeligman, ShripadTuljapurkar, DavidRehkopf, Machine learning approaches to the social determinants, *SSM - Population Health*, 2018;4, 95–99.

[2] WenjieHu, JayHarshadbhai Patel, Zoe-Alanah Robert, PaulNovosad, SamuelAsher, ZhongyiTang, MarshallBurke, David Lobell, and Stefano Ermon.. Mapping Missing Population in Rural India: A Deep Learning Approach with Satellite Imagery. In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19),* January 27–28; 2019, Honolulu, HI, USA. ACM, New York, NY, USA, https://doi.org/10.1145/3306618.3314263.

[3] Predict Population Growth Using Linear Regression — Machine Learning Easy and Fun, Gavril Ognjanovski, 4th December 2018, https://medium.com/analytics-vidhya/predict-population-growth-using-linear-regression-machine-learning-d555b1ff8f38.

[4] A Poongodai, R Suhasini, R Muthukumar, "Regression Based on Examining Population Forecast Accuracy, *"International Journal of Recent Technology and Engineering (IJRTE)*", 2019;8, Issue-1S4.

[5] Caleb Robinson, Fred Hohman, BistraDilkina, A Deep Learning Approach for Population Estimation from Satellite Imagery, *GeoHumanities'17*, November 7–10, 2017.

[6] Samir MazidbhaiVohra, "Population Growth – India's Problem", *PARIPEX – Indian Journal of Research, 2015.*

[7] Jeff Tayman, Stanley K. Smith and Stefan Rayer,"Evaluating Population Forecast Accuracy: A Regression Approach Using County Data", *Population Res Policy Rev, Springer*,2011;235–262.

[8] Seung-Joo, Lee Sunghae Jun and Jea-Bok Ryu, “A Divided Regression Analysis for Big Data”, *“International Journal of Software Engineering and Its Applications”*, 2015.

[9] ManjulMayankPandey, RupamTiwari and AnupamaChoubey, “Population Dynamics in India”,International Journal of Scientific & Engineering Research, 2015; 6, Issue-1.

[10] GeetanjaliDiwani, “Nexus between Population, Income, Output and Employment: Econometric Evidence from India”, International Journal of Multidisciplinary and Current Research, 2017.

[11] Divisha S, “Population projection: Meaning, Types and Importance”, http://www.sociologydiscussion.com/demography/population-projections/population-projections-meaning-types-and-importance/3058”.