I. J. Education and Management Engineering, **2023, 2, 1-9**
Published Online on April 8, 2023 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijeme.2023.02.01

Development of a Prediction Model on Demographic Indicators based on Machine Learning Methods: Azerbaijan Example

Makrufa Sh. Hajirahimova
Institute of Information Technology of the Ministry of Science and Education of the Republic of Azerbaijan, Baku, AZ1141, Azerbaijan
E-mail: hmakrufa@gmail.com
ORCID iD: https://orcid.org/0000-0003-0786-5974

Aybeniz S. Aliyeva*
Institute of Information Technology of the Ministry of Science and Education of the Republic of Azerbaijan, Baku, AZ1141, Azerbaijan
E-mail: aliyeva.a.s@mail.ru
ORCID iD: https://orcid.org/0000-0002-1739-1808

Received: 13 December, 2022; Revised: 02 January, 2023; Accepted: 15 February, 2023; Published: 08 April, 2023

Abstract: The accuracy of population forecasts is one of the most important calculations in demography statistics. However, traditional demographic methods used in population projections are tend to produce biased results. The need for accurate prediction of future behavior in a number of areas require the application of reliable and efficient methods. Recently, machine learning (ML) models have emerged as a serious competitor to classical statistical models in the forecasting community. In this study, the performance and capacity of the four different ML models such as Random forest (RF), Decision tree (DT), Linear regression (LR) and K-nearest neighbors (KNN) to the prediction of population has been examined. The aim of the study is to find the best performing regression model among these machine learning algorithms for forecasting of population. The data were collected from the State Statistical Committee of the Republic of Azerbaijan website were used for the analysis. We used five metrics such as mean absolute percentage error (MAPE), mean absolute error (MAE), root mean squared error (RMSE), mean square error (MSE) and R-squared to compare the predictive ability of the models. As the result of the analysis, it has been known that the all ML models showed high results with correlation coefficient of 0.985 - 0.996. Also the KNN and RF prediction models showed the lowest root mean square deviation, means square error and mean absolute error values compared to other models. By effectively using the advantage of the ML algorithms, the forecast of population growth the near future can be observed objectively, and it can provide an objective reference to the strategic planning in the public and private sectors, particularly in education, health and social areas.

Index Terms: Time series forecasting, population prediction, machine learning, linear regression, decision tree, random forest, k-nearest neighbors.

1. Introduction

The main goal of the state is to provide high-level services to the citizens of the country. For this, it is necessary to know how many people there will be in the future, and in some cases, the future composition of the population, the number by gender and age. Considering the possible changes in the number and age structure of the population plays an important role in determining the steps to be taken for future development. High-level planning can effectively build a healthy national infrastructure and improve the quality of services provided to the population. Without population forecasts, in the field of strengthening the social protection system of the population, reducing poverty, raising the standard of living of the population, etc. It is not possible to develop strategies in these directions, to plan long-term complex socio-economic development. Demographic (population) projections are also widely used for strategic planning in the private sector, by academics and other researchers, particularly in education, health and social sciences.

National and international population forecasts by population size, age and sex are prepared by many reputable organizations of the world. The United Nations [1] is the main organization that prepares regularly updated population

estimates and forecasts. It provided population estimates for 237 countries or territories and projections of numbers by age and sex up to the year 2100 based on an analysis of historical demographic trends beginning in 1950. These forecasts have been published since 1953 in the publication World Population Prospects. Projections are used as input data in modeling global issues such as food security and climate change. Many countries also use these forecast data for national planning [2, 3].

In some countries, these forecasts are prepared by national governments. These forecasts usually involve predictions for the next 40-50 years or so. For example, for major national issues related to infrastructure and workforce planning, the US Census Bureau predicts that the US population will grow over the next 40 years [3]. The Japanese government predicts the future of its population for the next 50 years [4]. The US Social Security Administration projects its budget for the next 75 years, which includes mortality and other population projections for that forecast period [2, 5].

Compiling demographic forecasts was not easy and required complex calculations. The difficulty of preparing these forecasts is that they depend not only on the population size and gender-age structure for the estimated period, but also on trends in demographic processes such as births, deaths and migration. The statistical accuracy of forecasts is particularly important, and the applied methods play a special role here. Population projections are traditionally made using a deterministic mathematical method called the cohort-component method, which has been the dominant method since the 1940s. The long-term use of the cohort-component method is due to its unique advantages. However, the uncertainties in the prediction results of this method are difficult to interpret and lack statistical or probabilistic validity [2].

The increasing availability of large amounts of historical data and the need for accurate prediction of future behavior in a number of scientific and applied fields have necessitated the use of reliable and efficient methods. Since the 60s of the twentieth century, linear statistical methods such as autoregressive integrated moving average (ARIMA) models have played an important role in the field of forecasting. Some studies suggest that these approaches provide systematically lower predictive values than simple statistical methods and are valid with extremely low sample sizes. Recently, machine learning models have gained attention and emerged as a serious competitor to classical statistical models in the forecasting community. Although the ARIMA model performed best in many cases, the researchers concluded that machine learning models are better able to predict larger datasets. The widespread use of machine learning algorithms in recent years of research on forecasting time series, as well as demographic time series, proves this once again.

Demographic time series forecasting is one of the most active research topics in Azerbaijan, as in other countries. Many traditional and modern forecasting methods are widely used for forecasting. The cohort-component method [6], the fuzzy time series model [7, 8] and other demographic statistical methods have been used in the studies conducted on population forecasting in the country. Although the cohort-component method, which is traditionally used in population forecasts, has certain advantages, it tends to present a biased result in forecasting. Also the fuzzy-based time series forecasting model faces some problems which come from the difficulty constructing and deconstructing the fuzzy sets, and also from the complexity of the fuzzy logical relationship [9]. The increasing availability of large amounts of demographic time series data and the need for accurate prediction of future behavior in a number of areas require the application of reliable and efficient methods. Last years, ML models have emerged as a serious competitor to classical statistical models in the forecasting community. A machine learning model uses techniques that learn and train using existing data to make connections between input data and output. A successfully trained ML model can analyze the data and make correct predictions even when historical data of the study area is not available and/or some input features are missing. However, while some machine learning algorithms have been used in other studies, they have not been tested in population forecasting in our country. For the first time, four machine learning models (DT, RF, LR and KNN) have been evaluated in the context of population forecasting in Azerbaijan. The purpose of this work is to build the best machine learning model for the time series data of the population and make a correct prediction on the Azerbaijan population size for the next ten years.

The structure of the rest of the paper is as follows: Section 2 reviews some related works. Section 3 provides information on the dataset used in the study and describes the ML methods used for population prediction. In Section 4, the performance evaluation metrics of these methods are presented. In the 5th section, the results of these methods are discussed. Finally, Section 6 concludes with the conclusion.

2. Literature Review

Research on time series forecasting has been widely covered in the literature. This is due to the dynamic nature of the problem and the need for better results. A brief literature review on the application of machine learning techniques in the context of population time series forecasting is presented below.

KNN is one of the widely used methods in classification and regression problems. Despite its simplicity, KNN has been successfully applied to time series forecasting. However, selection of the number of neighbors and feature selection is a difficult aspect of this method. In [10], S. Tajmouati and his colleagues, who used the KNN method for time series prediction, presented two methodologies for selection of the number of neighbors and feature selection: Classical Parameters Tuning in Weighted Nearest Neighbors (CPTO-WNN) and Fast Parameters Tuning in Weighted

Nearest Neighbors (FPTO-WNN). They were compared their proposed methods with some classical approaches to time series forecasting such as SARIMA, Holt-Winters and Exponential smoothing state space model (ETS). The accuracy of the models was evaluated using MAPE metric. The study shown that the proposed approach outperforms the classical approaches in terms of efficiency and accuracy (with lowest MAPE=1.276587) to time series forecasting. However, the large size of the data can raise concerns about the computational complexity.

In [11], four different machine learning methods such as Artificial neural network (ANN), Decision tree (DT), Random forest (RF) and K-nearest neighbors (KNN) were used to predict the population growth rate of an area. The predictive ability of the models is compared according to statistical metrics. The studies have demonstrated that all the machine learning methods used in predicting population growth rates are more than 90% accuracy. However, KNN and RF methods were found to have better prediction accuracy (96.47% and 95.42%) than others methods. The study has shown the relevance of the machine learning models in predicting the population growth rate of an area.

The linear regression method is widely used to make predictions on data and exhibits high predictive ability and statistical accuracy regardless of the size of the data. This model, which has been used in Swedish population growth forecasting [12], Nigerian population forecasting [13], showed high performance and was chosen as the best model for characterizing the population accordingly.

To overcome the problems faced by the census systems in Nigeria, the Nigerian researchers Dr. N. Ashioba and N. N. Daniel developed a Population Forecasting System using a machine learning algorithm [14]. They applied Object Oriented Analysis and Design methodology in developing this system. The obtained results showed that the Linear Regression Model has lower percentage errors than the Average Projection Model and the Natural Fund Growth model.

V. S. Fatih et al. [15] attempted to predict the population of Turkey using machine learning algorithms such as Light Gradient Boosting, Holt-Winters, Exponential, Autoregressive Integrated Moving Average (ARIMA) and Prophet Prediction Model. The models were trained using 1595 different demographics from 262 different countries. The prediction results of the models were compared according to its RMSE and MAPE statistical metrics. The ARIMA model was the most successful among all models with a root mean square error (RMSE) of 136570.57, and a mean absolute percentage error (MAPE) of 0.100%. Although the ARIMA model performed the best result, they concluded that the regression models can better predict a larger dataset.

In [16], LR, Support Vector Regression, Multilayer perceptron and DT machine learning methods were used to predict population growth in India. Prediction results of these machine learning techniques were compared with root mean square error value. The experiments shown that the LR outperform other three techniques with a root mean square error (RMSE=3.5123).

In [17], three tree-based methods such as DT, RF and Gradient Boosted Trees (GBT)) and also ARIMA model were used for forecasting time series data. The results of the research based on real data samples showed that RF model has higher accuracy in data prediction with a root mean square error (RMSE=136570.57) than DT, GBT and ARIMA(0,1,1) models.

M. M. Otoom [18] used 17 different ML methods to predict the population growth rate in cases where the feature data or historical data is not available. The performance of 9 base machine learning algorithms (ANN, DT, KNN, Linear discriminant analysis, Logistic regression, Localized Generalized Matrix Learning Vector Quantization, Naïve Bayes, Quadratic discriminant analysis and Support vector classifier) and 8 ensemble machine learning methods used in the study were compared. Authors evaluated the prediction performance of population growth rate using accurasy as the performance metric for the study. As a result, RF method was found to have better prediction accuracy with high growth rate (0.98%), medium growth rate (0.95%) and negative to low growth rate (0.96%) than others methods. These findings suggest that RF could be the best ML algorithm for performing population growth rate prediction.

In [19], extreme Gradient boosting, CatBoost,linear regression, ridge regression, Holt-Winters, exponential, ARIMA and prophet prediction model were used for forecasting population. The results of experiments showed that the machine learning algorithms performed better than the demographic Cohort component method. The accuracy of the models was measured using MAPE and RMSE metrics. The extreme gradient boosting model was more successful than the other models with a root mean square error (RMSE) of 1723842, and a mean absolute percentage error (MAPE) of 0.0191.

In [20] has been applied three machine learning methods such as Linear Regression model, LSTM model, and XGBoost Regression model for analyzing and forecasting the population growth of large cities in Taiwan. The prediction result of each model is compared based on MAPE. A comparison of the experiments shows that the XGBoost outperform other two approaches with a mean absolute percentage error (MAPE) of 0.02064. The results of the study showed that by effectively using the XGBoost algorithm, it is possible to analyze and objectively predict the population growth of large cities.

ANN models are rarely used in population forecasting, unlike other fields. In [21], predictions generated by ANN models were compared with population predictions from the traditional cohort component method (CMP). Thus, the least percentage accuracy for ANN was 81.02% while that the least for CMP was 64.55%. The highest percentage accuracy of both models was 99.15 and 86.43% respectively. The obtained results showed that a basic ANN model gives more accurate results than the demographic model.

V. Riiman et al. [22] used ANN long-short-term memory models (LSTM) to predict population for Alabama counties. They compared the predictions generated by the LSTM models with population predictions from the traditional cohort-component method in the state of Alabama. Depending on the training data, different results were

obtained for the LSTM model. The results obtained from the cohort-component method were better than the LSTM model developed based on data from all provinces. The LSTM model performed better than the cohort-component method when decennial census data were used.

Many of the statistical analyzing methods described in the literature are not suited for inaccurate, incomplete data. By contrast, ML methods can analyze the data and make accurate predictions even when historical data of the study area is not available or some input features are missing. As can be seen from the studies reviewed above, ML methods (especially KNN, RF, LR, etc.) show better performance for population prediction giving less error than traditional analyzing methods.

3. Materials and Methods

In this section, the dataset and the proposed methods for predicting the yearly number of populations are introduced. First, we are provided a brief introduction to the datasets, and then predictive machine learning models are explained. The effectiveness of the models was experimentally investigated using the Python programming language and its Scikit Learn library.

3.1 Data Collection

The dataset used in our study is based on yearly number of population. The data were collected from the State Statistical Committee of the Republic of Azerbaijan website starting from Dec 1990 to Dec 2021 [23]. This data was plotted on a graph, as shown in Fig. 1. From the graph, it can be seen that positive demographic trends in the country are permanent, and there is a stable growth rate of the population. This is due to the increase in the number of births in the country every year, the life expectancy at birth and the continuation of the positive migration balance. Using these data, the population forecast for the next ten years was made based on machine learning methods.

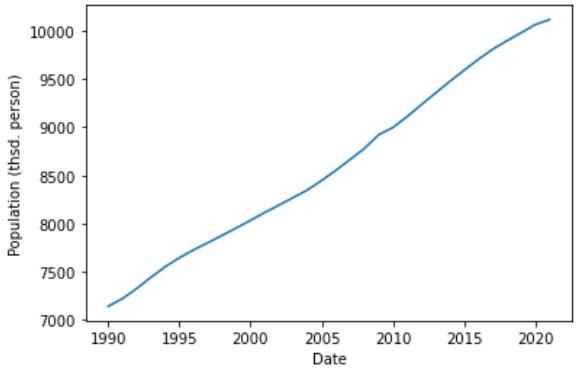


Fig. 1. Time Series Plot of the yearly number of population as on 1990 to 2021.

3.2 Machine learning models

3.2.1. Linear Regression (LR)

In regression modeling, the target class is based on independent features. This method can be used to find the relationship between independent and dependent variables and also for prediction. Linear regression, a type of regression modeling, is the most commonly used statistical method for predictive analysis in machine learning [24]. In linear regression, each observation depends on two quantities (values) - one is the dependent variable and the other is the independent variable. Linear regression determines the linear relationship between these dependent and independent variables. There are two factors (x; y) involved in linear regression analysis. The regression equation showing the dependence of y on x is expressed as follows:

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{1}$$

Here, ε is the error term of the linear regression. The error term is used to account for the variability between x and y, β_0 represents the y-intercept, and β1 represents the slope.

3.2.2. K-Nearest Neighbors (KNN)

K-nearest neighbor algorithm is one of the widely used machine learning techniques in classification and regression. Despite its simplicity, the K-nearest neighbor method has been successfully applied to time series forecasting. KNN is a sample-based non-parametric learning algorithm. KNN is based on the principle of identifying data points with similar values or characteristics that are located near a data point in the database [25]. However, the issue of selecting the number of neighbors and feature selection is a difficult task. In this approach, the value of the unknown output parameter of a data point is determined based on the value of the output parameter of the nearest data

point. A distance metric is used to measure the relative distance between data points. The objective of the distance metric is to minimize the distance between similar data points and maximize the distance between dissimilar data points. Minkowsky, Manhattan, Chebychev and Euclidean metrics are used in the research to measure the relative distance between points [24, 25]. The most commonly used Euclidean distance formula is as follows:

$$d_f(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Here, x and y are the feature vectors, and n is the number of features or parameters. The value of the distance metric ranges from "0" to "1". Here, the minimum distance for all points in the network is "0" and the maximum distance is "1".

3.2.3. Decision Tree (DT)

A decision tree (DT) is a non-parametric machine learning algorithm used for classification and regression problems. It has a tree structure consisting of a root node, branches, internal nodes and leaf nodes. Each tree starts with a single node called the root node, then sub-nodes are created from each branch of the tree, and the process continues to build the decision tree. There are many ways to select the best attribute at each node. Popular metrics for the selection of such attributes for decision tree models are Gini Index and Information Gain [17, 25, 28].

3.2.4. Random Forest (RF)

Random Forest is an ensemble learning approach that generates a large number of decision trees. In this approach, only a subset of the input attributes is used to generate a decision tree. Not all attributes are involved in the construction of an arbitrary tree, but only randomly selected ones. Thus, RF generates several decision trees composed of a random subset of input attributes. To get the final result, he combines the results of different decision trees [28].

4. Model Evaluation Metrics

Accuracy is the most important criterion in statistical model evaluation. In this study, metrics such as mean square error (MSE), mean absolute error (MAE), mean absolute percentage error (MAPE), root mean squared error (RMSE) and R- squared were used to evaluate and compare the performance (accuracy) of learning models. Equations and descriptions of the metrics used in the study are presented in the Table 1.

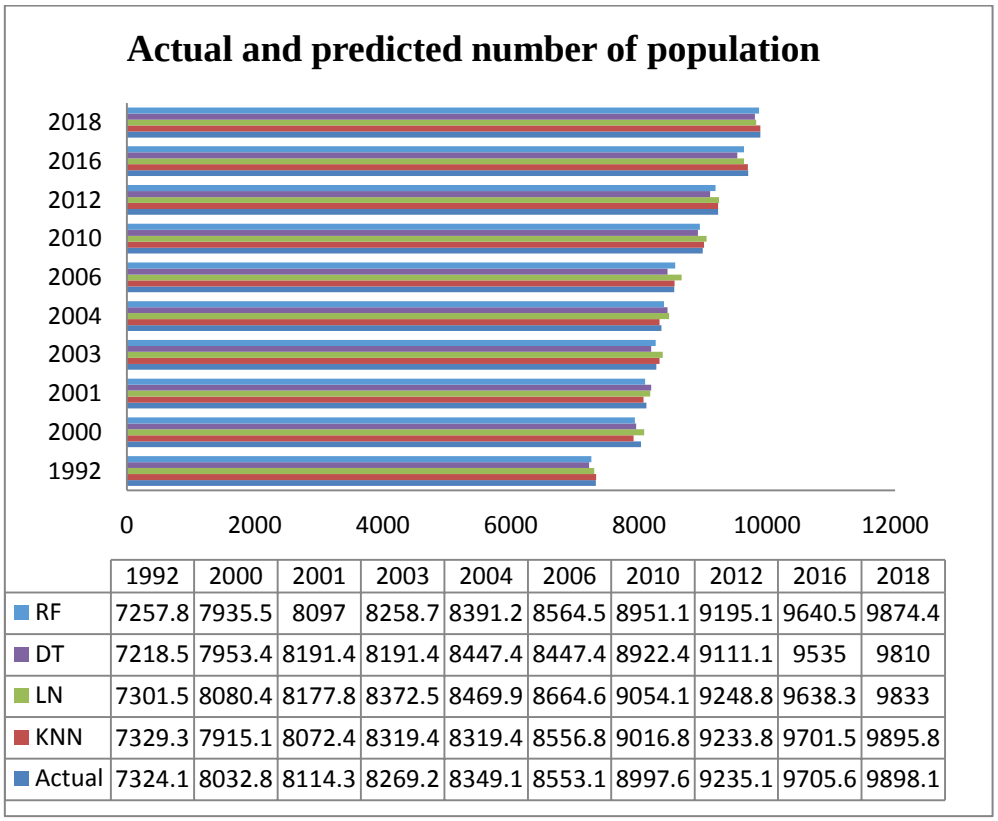
Table 1. Description of the metrics used to evaluate the accuracy of the models.

Accuracy metrics	Equation of metrics	Description of metrics
MAE	$MAE = \frac{1}{N}\sum_{i=1}^{N}(\lvert\hat{y}_i - y_i\rvert)$ (3)	The MAE metric computed as the average absolute difference between the actual values and the predicted values by regression model. The values of the MAE range in the (0, +∞) interval, and the lower the MAE value, the better the model predicts [24, 29, 30].
MAPE	$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert y_i - \hat{y}_i\rvert}{y_i}$ (4)	MAPE is one of the most commonly used metric to measure the accuracy of population forecast. It is the mean of all absolute percentage errors between the predicted values and actual values. The values of the MAPE range from 0 to infinity, and the lower the MAPE value, the more accurate the model predicts. It is advised not to use MAPE when actual values can be at or close to zero [29, 31].
MSE	$MSE = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$ (5)	MSE is another metric used to measure the performance of regression models. It is the average squared differences between the actual and predicted values. The values of the MSE range from 0 to infinity. A lower value of MSE indicates that the performance of the model is high [24, 30, 32].
RMSE	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$ (6)	RMSE is a measure of accuracy, to compare forecasting errors of different models for a given dataset. The RMSE represents the square root of the differences between predicted values and observed values or the quadratic mean of these differences. A lower value of RMSE always indicates a better performance [24, 29, 33].

R-squared	$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}$ (7) where, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ (8)	R-squared or R^2 (the coefficient of determination) is the proportion of the variance in the dependent variable that is predictable from the independent variables. In other words, R-squared determines how well the model fits a dataset. The positive values of the coefficient of determination varies in the [0, 1] interval. The larger the value, the better the performance of the model is considered [34].
Note: In the above equations, $\hat{y}_i$ is the predicted value, y_i is the actual value, $\bar{y}$ is the main of the true values, and N is the number of observations.		

5. Results and Discussions

In this section, the predictive capacity of the ML algorithms (KNN, LR, RF and DT) used in the study were investigated and the results were compared. Using random selection the original training dataset are split into training and testing datasets for machine learning models. This is necessary to train, compare and test the models.. 70% of the data have been used for training the models, and 30% for the test data. Data analysis has been performed using the Python programming language and its Scikit Learn library. The results of the population (in thousands) predictions using the KNN, RF, DT and LR models are presented in Figure 2.



	1992	2000	2001	2003	2004	2006	2010	2012	2016	2018
RF	7257.8	7935.5	8097	8258.7	8391.2	8564.5	8951.1	9195.1	9640.5	9874.4
DT	7218.5	7953.4	8191.4	8191.4	8447.4	8447.4	8922.4	9111.1	9535	9810
LN	7301.5	8080.4	8177.8	8372.5	8469.9	8664.6	9054.1	9248.8	9638.3	9833
KNN	7329.3	7915.1	8072.4	8319.4	8319.4	8556.8	9016.8	9233.8	9701.5	9895.8
Actual	7324.1	8032.8	8114.3	8269.2	8349.1	8553.1	8997.6	9235.1	9705.6	9898.1

Fig. 2. Actual and predicted number of the population.

In order to make a more detailed comparison to each other algorithms, R-squared, MAPE, MSE, RMSE and MAE error analysis metrics were used for the performance evaluation of the study. Performance results of R-squared, MAPE, MSE, RMSE and MAE error analysis metrics values according to ML algorithms used in the study were presented in Table 2.

Table 2. Performance parameters of the models

Model	Performance factor				
	MSE	RMSE	MAE	MAPE (%)	R^2 (%)
LR	5669,19	75,29	67,21	0,7	99,0
DT	9175,97	95,79	94,38	1,0	98,4
RF	2472,89	49,72	42,02	0,4	**99,5**
KNN	1945,33	44,11	27,52	0,3	**99,6**

The visual representation of the performance indicators of the models (except MSE) is as in Figure 3. As can be seen from Figure 3, each of the machine learning methods used in this study showed high results with correlation coefficient. For the models, the R-squared values range between 0.985 - 0.996 (or 98.5-99.6%). However, KNN and RF models showed superior performance compared to other models.

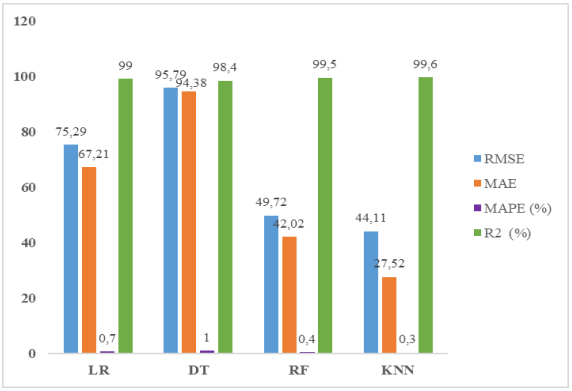


Fig. 3. Performance results of ML algorithms

According to Figure 3, at all results the KNN and RF models achieved the best performance with respect to MAE, RMSE and MAPE values. The performance results of KNN and RF indicate relatively lower values in terms of MAE (27.52 and 42.02), RMSE (44.11 and 49.72) and MAPE (0.3 and 0.4). The error rates between test data and prediction values for LR were 67.21 for MAE, 75.29 for RMSE and 0.7 for MAPE. Also, the results of the DT model were calculated as MAE, RMSE, MAPE as 94.38, 95.79 and 1.0, respectively.

These results show that the KNN and RF models perform better for this problem than other ML algorithms. Using KNN and RF models, we predicted the population for the next 10 years (2022 - 2031). The results of the forecast are given in Table 3.

Table 3. Forecast of the total population (in thousands) for the next 10 years (2022 - 2031)

Years	Models	
	KNN	RF
2022	10198.485	10180.647
2023	10264.144	10242.194
2024	10337.417	10303.741
2025	10405.154	10365.288
2026	10516.117	10426.836
2027	10578.328	10488.383
2028	10649.630	10549.930
2029	10711.271	10611.477
2030	10783.325	10673.025
2031	10884.724	10734.572

Thus, it can be concluded that the machine learning models are effective for forecasting the population of Azerbaijan. At the same time, testing other machine learning models is necessary to choose the model that performs better.

6. Conclusion

Researches carried out over show that machine learning is successfully applied in demographic time series forecasting and show high performance. However, traditional demographic methods used in population projections are tend to produce biased results. The aim of the study is to find the best performing regressor among four different machine learning algorithms in total population prediction. For this purpose, the performance and capacity of the ML regression approaches (Random Forest, DT, LR and KNN) to the prediction of population has been examined. The data were collected from the State Statistical Committee of the Republic of Azerbaijan website were used for the analysis. We used five metrics such as MAPE, MAE and RMSE, MSE and R-squared to compare the predictive ability of the models. As the result of the analysis, it has been known that the all ML models showed high results with correlation coefficient of 0.985 - 0.996. At the same time, the KNN and RF prediction models showed the lowest root mean square deviation (44.11 and 49.72), means absolute error (27.2 and 42.02) and mean absolute percentage error (0.3 and 0.4) values compared to other models, respectively. Thus, by effectively using the advantage of the ML algorithms, the forecast of population growth the near future can be observed objectively, and it can help the state to eliminate errors

in the planning of socio-economic activities in education, employment, health and other areas and to build more reliable and sustainable plans. We intend to carry out the forecasting of other demographic indicators using machine learning algorithms in our future research.

References

[1] World population prospects: Summary of Results. New York, United Nations, 2022. Available at: https://www.un.org.development.desa.pd/wpp2022_summary_of_results.pdf

[2] A.E. Raftery and H. Ševčíková, "Probabilistic population forecasting: Short to very long-term," International Journal of Forecasting, 7 October 2021. Available at: https://doi.org/10.1016/j.ijforecast.2021.09.001

[3] J. Vespa, D. M. Armstrong, and L. Medina, "Demographic turning points for the United States: population projections for 2020 to 2060: Current population reports P25-1144," Washington, D.C.: U.S. Census Bureau, 2020.

[4] National Institute of Population and Social Security Research. Population projections for Japan (2016–2065), 2017. Available at: http://www.ipss.go.jp/pp-zenkoku/e/zenkoku_e2017/

[5] Social Security Administration. The 2020 annual report of the board of trustees of the federal old-age and survivors' insurance and federal disability insurance trust funds, 2020. Available at: https://www.ssa.gov/oact/TR/2020/tr2020.pdf.

[6] R. Allahverdiyev and Kh. Nasibov, "Aspects of methodological approach to forecasting of population size," 2017. Available at: http://etsim.az/upload/Image/2017-1/

[7] A. M. Abbasov and M.H. Mamedova, "Application of fuzzy time series to population forecasting," in Proceedings of the 8th International Symposium on ICT and planing and impacts of ICT on Physical Space, Vienna, 2003, pp. 545-552.

[8] Z. G. Jabrayilova, "Development of intelligent demographic forecasting system, ", Eastern-European Journal of Enterprise Technologies, vol.5, no.2(101), pp.18-25, 2019. Available at: DOI:10.15587/1729-4061.2019.178440

[9] P. Singh, "High-order fuzzy-neuro-entropy integration-based expert system for time series forecasting", Neural Comput. Appl., vol. 28, pp. 3851–3868, 2017.

[10] S. Tajmouati, B. Wahbi, A. Bedoui, A. Abarda, and M. Dakkon, "Applying k-nearest neighbors to time series forecasting: two new approaches," arXiv:2103.14200v1 [stat.ME] 26 Mar 2021, pp. 1-20.

[11] M.M. Otoom, M. Jemmali, Y. Qawqzeh, S. A. Khalid Nazim, and F. Al Fayez, "Comparative analysis of different machine learning models for estimating the population growth rate in data-limited area," IJCSNS International Journal of Computer Science and Network Security, vol.19, no.12, pp. 96-101, December 2019.

[12] G. Ognjanovski, "Predict Population Growth Using Linear Regression - Machine Learning Easy and Fun," 4th December 2018. Available at: https://medium,com/analytics-vidhya/

[13] O. D. Odunayo, O. E. Oduntan, and I. R. Olasunkanmi, "Using predictive Machine Learning Regression Model to predict the population of Nigeria," Annals. Computer Science Series, vol. 16, no. 2, pp. 137-142, 2018.

[14] Dr. N. Ashioba and N. N. Daniel, "Population Forecasting System Using Machine Learning Algorithm," International Journal of Computer Trends and Technology, vol. 68, no. 12, pp. 40-43, December 2020.

[15] V. S. Fatih, T.T. Ahmet, and C. Ferhan, "Machine Learning algorithm to forecast the population: Turkey Example," in Proceedings of International Engineering and Technology Management Summit 2019 – ETMS. Available at: www.researchgate.net/publication/33714439

[16] S. B. Rajakumari, P. Padmanabhan, S. Christy, and M. Nandhini, "Prediction of population growth using machine learning techniques," European Journal of Molecular & Clinical Medicine, vol.7, no. 5, pp. 1871-1879, 2020.

[17] E. A. Rady, H. Fawzy, and A.M. Abdel Fattah, "Time Series Forecasting Using Tree Based Methods," Journal of Statistics Applications & Probability, vol.10, no. 1, pp. 229-244, 2021. Available at: http://dx.doi.org/10.18576/jsap/100121

[18] M. M. Otoom, "Comparing the Performance of 17 Machine Learning Models in predicting Human Population Growth of Countries," International Journal of Computer Science and Network Security, vol. 21, no.1, January 2021.

[19] F.V. Şahinarslan, A.T. Tekin, and F. Çebi, "Application of machine learning algorithms for population forecasting," International Journal of Data Science, vol. 6, no. 4, pp. 257–270, 2021.

[20] C.Y. Wang and S.J. Lee, "Regional Population Forecast and Analysis Based on Machine Learning Strategy," Entropy, vol. 23, no. 656, pp. 1-12, 2021. Available at: https://doi.org/10.3390/e23060656

[21] O. Folorunso, A. Akinwale, O. Asiribo, and T. Adeyemo, "Population prediction using artificial neural network," African Journal of Mathematics and Computer Science Research, vol. 3, no. 8, pp. 155–162, 2010.

[22] V. Riiman, A. Wilsony, R. Milewiczz, and P. Pirkelbauerx, "Comparing Articial Neural Network and Cohort-Component Models for Population Forecasts," Population review, vol. 58, no. 2, pp.100-116, 2019.

[23] www.stat.gov.az, State Statistical Committee of the Republic of Azerbaijan, 2022.

[24] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-Won On, W. Aslam, and G. S. Choi, "COVID-19 Future Forecasting Using Supervised Machine Learning Models," IEEE Access, vol.8, pp.101489-101499, 2020.

[25] M. Sh. Hajirahimova, L. R. Yusifova, "Experimental Study of Machine Learning Methods in Anomaly Detection," Problems of Information Technology, vol. 13, no. 1, pp. 9-19, 2022.

[26] What is a Decision Tree? Available at: https://www.ibm.com/topics/decision-trees

[27] M. P. Frías, M. D. Pérez, and A. J. Rivera, "A methodology for applying k-nearest neighbor to time series forecasting," Artificial Intelligence Review, vol.52, no. 3, 2019.

[28] J. Brownlee, "Random Forest for Time Series Forecasting," November 2, 2020, Available at: https://machinelearningmastery.com/random-forest-for-time-series-forecasting/].

[29] M. Sh. Hajirahimova and A. S. Aliyeva, "Forecasting the COVID-19 confirmed cases and deaths in Azerbaijan using Prophet," National Supercomputing Forum (NSCF-2021), Russia, Pereslavl-Zalessky, Program Systems Institute of the RAS, November 30 – December 03, 2021, Available at: https://2021.nscf.ru/TesisAll/05_AI_MachineLearning/250_AliyevaAS.pdf

[30] V. Verma, R. K. Aggarwal, "Accuracy Assessment of Similarity Measures in Collaborative Recommendations Using CF4J Framework", International Journal of Modern Education and Computer Science (IJMECS), vol.11, no.5, pp. 41-53, 2019. DOI: 10.5815/ijmecs.2019.05.05

[31] B. Nazlı, Y. Gültepe, and H. Altural, "Classification of Coronary Artery Disease Using Different Machine Learning Algorithms," International Journal of Education and Management Engineering (IJEME), vol.10, no.4, pp.1-7, 2020. DOI: 10.5815/ijeme.2020.04.01
[32] S. Allwright, "What is a good MAPE score?", 15 Aug 2022, Available at: https://stephenallwright.com/good-mape-score/
[33] D. Chicco, M.J. Warrens, G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," Peer J Computer Science 7:e623, 2021, https://doi.org/10.7717/peerj-cs.623
[34] R. Agrawal, Know the Best Evaluation Metrics for Your Regression Model. Available at: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

Authors' Profiles

Makrufa Sh. Hajirahimova, PhD, is an Associate Professor and a department head at the Institute of Information Technology of the Ministry of Science and Education of the Republic of Azerbaijan. Her research areas are Electron demography, social network security, anomaly detection, Big Data Analytics, intellectual potential and Machine learning. She is the author of more than 125 papers.

Aybeniz S. Aliyeva is currently working as senior researcher of Information Technology of the Ministry of Science and Education of the Republic of Azerbaijan. She has graduated from Applied Mathematics faculty of Baku State University. Her current research interests include Electron demography, intellectual migration, Big Data Analytics and machine learning. She has published more 65 research articles.

How to cite this paper: Makrufa Sh. Hajirahimova, Aybeniz S. Aliyeva, "Development of a Prediction Model on Demographic Indicators based on Machine Learning Methods: Azerbaijan Example", International Journal of Education and Management Engineering (IJEME), Vol.13, No.2, pp. 1-9, 2023. DOI:10.5815/ijeme.2023.02.01