Comparative Analysis of Regression Models and Exponential Model for Population Growth Prediction

Maisha Haque
School of Data and Sciences
Brac University
Dhaka, Bangladesh
maisha.haque@g.bracu.ac.bd

***Abstract*—Population growth of a country is considered to be an important feature for real-world planning. Predicting the growth rate is commonly calculated with historical data. Without the historical data a proper prediction cannot be done which is why different machine learning algorithms have been used to predict the population growth. As missing feature is a common real-world problem, it is hard to find out which approach works the best in terms of given features. Machine learning models perform differently based on the given features. Also, mathematical models have been used to predict the population but it needs other features This study compares performance four different regression model (Linear Regressor, Polynomial Regressor, Decision Tree Regressor and Random Forest Regressor) and a mathematical model (Exponential Growth Model) when no other features are available except the population. Furthermore, this work tries to predict the population of the next ten years of Bangladesh is predicted with the best performing model.**

***Index Terms*—Population Growth, Machine Learning, Regression, Exponential Model, Prediction**

I. Introduction

Massive population is one of the biggest problems in today's world mostly in developing countries. It affects the nation's overall health, economy, primary needs and environment. For instance, the food limit, economical growth, disease spreading, pollution is highly related with population growth. Bangladesh is one of the most populated countries in the world. The population growth prediction can help a country determining the necessary future steps. As gaining accurate information about the future population size supports planning activities [1], so an accurate information is very much needed.

In this work, we used machine learning algorithms and a mathematical modeling to forecast the population of Bangladesh for the next 10 years using the population data of 1953-2022. The mathematical model used here was the exponential growth model presented by Thomas Mathus [2] which proposesthe assumption that the population grows at a rate proportional to the size of the population. On the other side the machine learning models are used are Linear Regression, Polynomial Regression, Decision Tree Regression and Random Forest Regression. Linear regression is a linear model that assumes a linear relationship between the input variables (x) and the single output variable (y)in this work year and population. The (y) can be calculated from a linear combination of the input variables (x). On the other side, polynomial regression allows for the accurate modeling of non-linear relationships. Decision tree is a model used for supervised learning and can be used to solve regression problem. One of the advantages of this model is that non-linearity doesn't affect the algorithm and lesser data cleaning is required. Although, it has a disadvantage which is the over-fitting problem. The Random Forest Regression model uses multiple decision tree with a technique named bagging. This model has the ability to solve the over-fitting problem of Decision Tree Regressor.

Statistical models can be built for population prediction but the estimation cannot be fully relied. *Python* and Scikit library were used in this work to build the comparison model of all the five models. In Bangladesh, there are too many people living in the country comparing to the area. The population by the years are much more available than other data of a region. We try to find the best model fitted to predict the population only based on the population feature. We also plot graphs and compare the predicated values by the models side by side. The models used in this work which are Linear Regressor, Polynomial Regressor, Decision Tree Regressor and Random Forest Regressor are widely used for forecasting models.

II. Related Works

Wang et al. [3] tries to forecast China's population growth based on two models. On their work the prediction model with net increase data from 2000 to 2003. It only uses data of the past four or five years to predict development status of China in the next twelve or more years. But in this paper, we used more than seventy years data to predict next ten to fifteen years population of Bangladesh. In order to establish one of their models, the used the exponential growth model proposed by Malthus.

Rahima et al. [4] analyses the mathematical models like exponential model, hyperbolic model and logistic growth model with linear regression to predict the future population of the countries of the Indian sub-continent. On the paper the analyses their data using mean absolute percentage error

(MAPE) values of the country and describes that out of all the model the linear regression model's MAPE value is the most accurate in the case of all three countries (Bangladesh, India & Pakistan). Again, for a distant population data the Malthusian and hyperbolic models wasn't a good fit as the MAPE values of hyperbolic model were around 2% whereas in linear regression the value was in the range of 0.04

Zabadi et al. [5] also used mathematical approach to predict the population of Jordan until the year 2020. They applied the simple exponential growth model, logistic growth model and Verhulst logistic growth equation to predict Jordan's population by using the population data of last 60 years to examine the pattern of countries population growth in the long run. Using MiniTab a nonlinear-regression as applied in the work. Here they came to a conclusion their logistic growth model works the best followed by the Verhulst growth equation model and exponential model as logistic model and Verhulst model takes in other parameters which effects the population growth but on the other side exponential models doesn't take in any type of parameters except the population.

Similarly, in the paper Uddin et al. [6] predicts the future population growth of Bangladesh using exponential and logistic Models. Their models are used to predict the population from 2000-2050 by using the population data of 2000-2019. The analyzed these models by the score of absolute percentage error for each year and ultimate finding the mean of absolute percentage error (MAPE). The MAPE of the exponential model was 4.7% whare as the logistic growth model has MAPE of 3.7% which is lower than the exponential model.

Tripepi et al. [7] discusses linear regression analysis for the examination of continuous outcome data and logistic regression analysis for the study of categorical outcome data in their paper. Also, they focused on the most important application of multiple linear and logistic regression analyses. For the linear regression analysis that the dependent variable needs to be continuous.The population values of our work shows that it's continuous growing consistently over the years.

On the other side for the human population growth Otom [8] used machine learning approaches to predict future population growth. Here he compared 17 machine learning models and came to a conclusion that random forest outperformed all the other models both in predictive performance and robustness. The models were trained using the Scikit library in python for prediction population growth rate and the accuracy was calculated based on the correctly predicted instances. Some of the models used in the paper are K-nearest neighbors, logistic regression, naive bayes, quadratic discriminant analysis, random forest and more. According to the paper, random forest could be the best model in this case where naive bayes has shown the worst performance among all.Our work of this paper was also done by python using the Scikit library.

Wang et al. [9] used only machine learning strategies to forecast the regional population and analyze of essence in urban and regional planning. Here they used ten years of data to forecast the next seven years population. According to their analysis the MAPE value of Linear Regression model was smaller than Long Short-Term Memory Network model which Determines Liner regression was better.

Agyemang [10]has a regression analysis on the population growth of Ghana based on the population data from 1990-2012. To determine the population growth, he used a linear regression model and performed the data analysis via Minitab. The coefficient of determination was one of the measures of the model which was 72.9% and indicates the accidents were was highly affected the by population.In my paper, the coefficient of determination was used to find the better fitting model.

Khalid [11] has designed a simple linear regression model that predicts 2023 to 2100 model by using the data of the population of Bangladesh from 1953-2022. We have used the same data for our analysis in this paper.

Surada's [12] code in the github repository uses Linear Regression, Ridge, MLP regressor, Random Forest Regressor and more and calculates the co-efficient of determination of the Linear Regressor which achieved the highest value. The dataset used here contains the population of India from 1950-2021. But the code doesn't predict future population.

For the machine learning regression models, we used Dalui's [13] Population prediction model for India using various machine learning algorithms. Here linear regression, polynomial regression, random forest was used to determine the best machine learning algorithm and using that algorithm the future forecasting of the population of India was done. To determine the best algorithm co-efficient of determination and mean squared error was used.

The exponential growth model of Ram's [14] github repository was used in this paper. In the he used a data of the population of countries world wide and using the exponential growth model he compared the population growth of the 7 big countries of the world which are US, Japan, Germany, France, UK, Italy and Canada. In our paper we checked the accuracy of this model for the population growth of Bangladesh. The formula use in this code was the Malthus

$$N(t) = N(0)e^{rt}$$

and the logarithm of the population should be a linear function of time which was $\log N(t) = \log N(0) + rt$.

III. Research Methodology

This work performs a four-stage methodology to evaluate the models in order to forecast the population area of a region. The stages are Dataset Collection, Feature Selection and Processing, Model Training and Model Evaluation.

A. Dataset Collection:

This study uses a population data of a country or a region. The dataset used here is the population dataset of Bangladesh which was collected from a GitHub source [11]. After analyzing the data with the dataset collected from world bank database [15], we can ensure the accuracy of the dataset.

B. Feature selection and Processing:

It can be hard to find a lot of features of a region to use in the models. Which is why this work is solely based on the population feature so that population of any area or country can be easily predicted with other related information. There are multiple features in our dataset but we only use the population feature in order to evaluate the model accuracy only on the basis of population of Bangladesh. So, we drop the rest of the columns. After plotting into the scatter graph, we find figure [1] The dataset was divided into train and test where the ratio is 70:30. We created a data frame of the years 2023-2033 to generate future population prediction.

C. Model Training:

Exponential Growth Model: The exponential model was built manually using the logarithmic function used in [cite]. The population was converted into the logarithmic value of the population. Then plot it and plot a fitting line and expand it. Then it was tested with our custom data frame for future prediction.

Linear Regressor: To use the Linear Regression the $Scikit - Learn$ library was used. First, to fit the training data the Linear Regressor was called from the linear model module. Using the test data of the year, the prediction was done. Lastly, using the custom data frame future prediction was made.

Polynomial Regression: Just like the previous model Scikit -Learn library was used. Here, we also had to import the polynomial features from preprocessing module. The polynomial degree was set to 27 and like the linear regression the test data and prediction data is used on it after polynomial fitting.

Decision Tree Regressor: Using the tree module the decision tree regressor used on the data. After fitting the model is used on the test data and predication data frame.

Random Forest Regressor: To use the Random Forest Regressor the ensemble module is used. The random state is assigned 0 and number of estimators is assigned 10 in this work. Like the rest of the models then it was used on the test and prediction data frame.

D. Model Evaluation:

The model was evaluated using coefficient of determination value (R^2), mean squared error (MSE) and the mean absolute percentage error (MAPE). The model which has the whose R^2 value is the closest value to 1 can said to better than the others. In case of mean squared error ad mean absolute percentage error, which model has the closer the value is to 0, is the better model.

IV. Results

A. Graph Analysis

The following graphs give visualizations about the predictions. From the scatter-plot visualization in figure **??** we can see that with the increment of time; the difference between the exponential model's predicated value and the actual value is increasing.

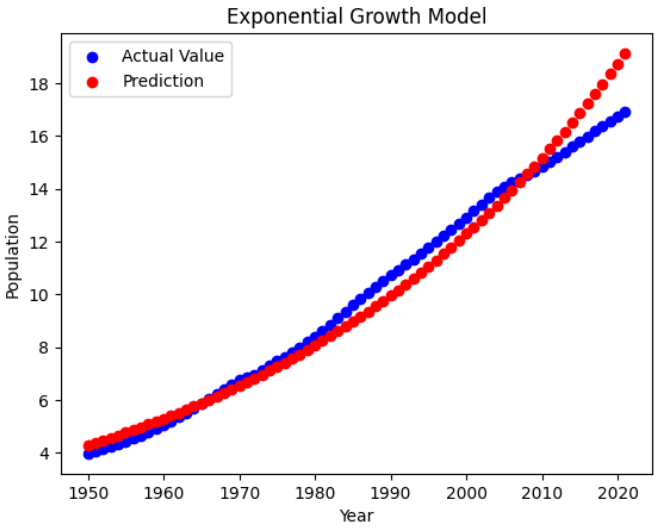

Fig. 1. Scatter plot of the Exponential Growth Model

But the other models in figure 2 don't show this behavior. The values we see in the ML models are very close to the actual value.From the plot we can easily determine that the performance of Exponential Growth Model is poor comparing to the ML models.

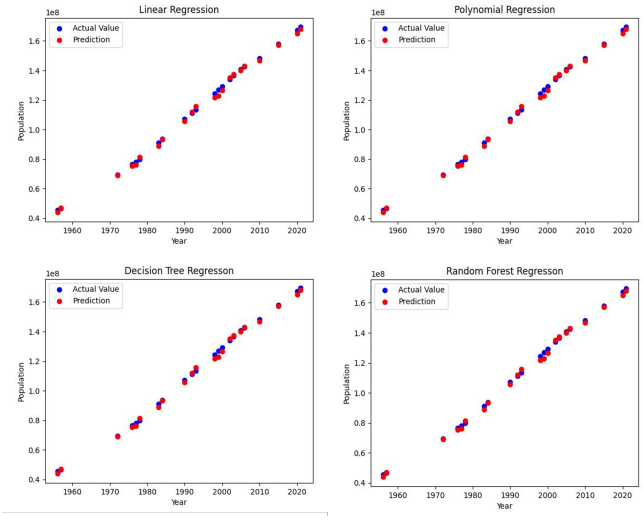

Fig. 2. Scatter plot of the Machine Learning Models

B. Data Analysis

Since the prediction value is so close for all the ML models, the best model can only be determined by data analysis. In table 1 we can see, the actual population (in millions) and the predicted population (in millions) of the years. We can see all the predicted population are really close to the year actual population. So, we try to find the coefficient of determination, mean squared error and mean absolute percentage error for each of the models.

In fig 3 we can see the that the coefficient of determination (r squared) values are really close in this model. As we want

the best model in the terms of r squared, from the table z, we can find that polynomial regression has the highest coefficient of determination.

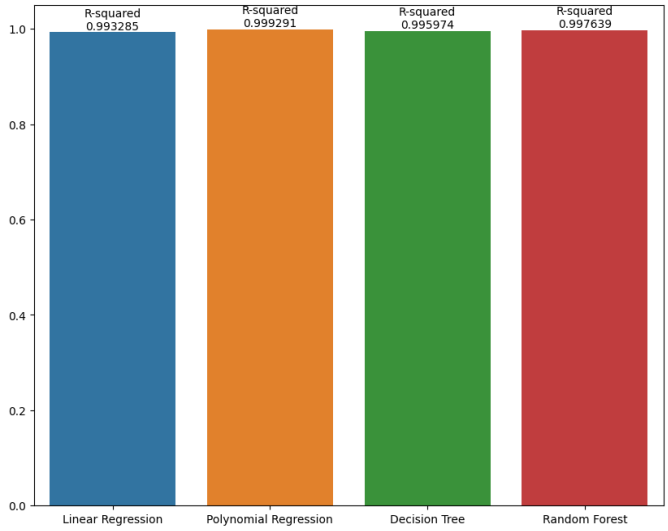

Fig. 3. Coefficient of determination (R^2) of Linear Regression, Polynomial Regression, Decision Tree and Random Forest

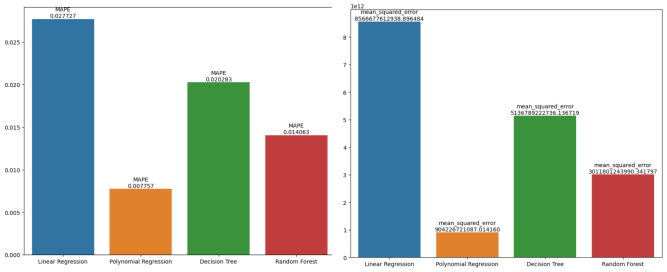

Fig. 4. Mean Squared Error (MSE) and Mean Absolute Percentage (MAPE) Error of Linear Regression, Polynomial Regression, Decision Tree and Random Forest

Again, in fig 4 we can see the bar chart of the MSE and the MAPE values. In both of the graphs Liner Regression has the highest value which are respectively $8.5e+12$ and 0.02772. On the other side, polynomial regression is giving a much less MSE and MAPE than the other models which are respectively $9.40e+11$ and 0.007757. The other two model has the error values between these two models. If we see the table I again, we can choose polynomial regression model in terms of MSE and MAPE as we want the model with the least value here. Furthermore, from table II, it can be identified that the polynomial regression model is giving the closet result to actual value over the years.

C. Future Population Prediction

On the table III a we can see the predicted population of Bangladesh for upcoming 10 years. Since, we find Polynomial Regression Model works the best in predicting the values we can say there is good chance that in the population of Bangladesh will increase $7million$ in the next 10 years and will become $177.01million$.

Model	R squared	MSE	MAPE
LR	0.993285	8.57E+12	0.27727
PR	0.999291	9.04E+11	0.007747
DTR	0.995974	5.14E+12	0.020283
RFR	0.997639	3.01E+12	0.14063

TABLE I
COEFFICIENT OF DETERMINATION (R^2), MEAN SQUARED VALUE (MSE) AND MEAN ABSOLUTE PERCENTAGE ERROR (MAPE) OF THE ML MODELS

Year	Actual P	LR	PR	DTR	RFR
1956	45.4077	42.7772	45.7547	44.3155	43.9021
1957	46.5605	44.7037	46.9925	47.7428	47.0802
1972	69.3467	73.6011	70.0143	68.3762	69.1234
1976	76.3801	81.3071	77.4276	74.7003	75.2459
1977	78.1378	83.2336	79.3526	74.7003	76.1419
1978	80.0075	85.1601	81.3045	81.9082	81.3895
1983	91.0455	94.7926	91.4294	88.5553	88.8332
1984	93.5342	96.7191	93.5195	95.9591	93.2289
1990	107.148	108.278	106.405	104.894	105.543
1992	111.272	112.131	110.796	109.243	111.994
1993	113.419	114.058	113.002	115.615	115.833
1998	124.35	123.69	124.061	122.039	121.607
1999	126.755	125.617	126.265	122.039	122.786
2000	129.193	127.543	128.461	131.67	126.639
2002	134.14	131.396	132.82	131.67	135.23
2003	136.503	133.322	134.979	138.79	137.366
2005	140.913	137.175	139.242	138.79	139.859
2006	142.629	139.102	141.341	144.136	143.067
2010	148.391	146.808	149.452	146.707	146.8
2015	157.83	156.44	158.721	155.961	157.108
2020	167.421	166.073	166.622	165.516	165.161
2021	169.356	167.999	167.996	171.186	168.18

TABLE II
POPULATION PREDICTION COMPARISON OF THE MACHINE LEARNING MODELS

D. Discussion

None of the models show absolute accuracy but all of them except the exponential growth models shows a really close prediction. Based on all the evaluation it can be said that polynomial regression works the best when there is only one feature. On the previous works we can see Linear Regression and Random Forest Regression works the best when there are other features included [4], [8], [10] but with only one feature the performance of them is lower. The lowest performing model in this work was Linear Regressor followed by Random Forest Regressor and Decision Tree Regressor.

V. CONCLUSION

Population count is one of the most important studies as it is need in order to build a proper planning of infrastructure, economy and more. Bangladesh is one of the most populated countries in the world with a very high population growth rate. We analyze in the work five different models which

Year	Prediction (PR)	Growth Rate
2023	170.503	
2024	171.6278	1.124836
2025	172.6606	1.032784
2026	173.5966	0.935974
2027	174.4309	0.83428
2028	175.1584	0.727571
2029	175.7742	0.615716
2030	176.2727	0.498577
2031	176.6488	0.376021
2032	176.8967	0.247903
2033	177.0107	0.114085

TABLE III
POPULATION PREDICTION OF THE NEXT 10 YEARS IN BANGLADESH

are exponential growth model, Linear Regressor, Polynomial Regressor, Decision Tree Regressor and Random Forest Regressor. We evaluate to the models in terms of the coefficient of determination, mean squared error and mean absolute percentage error and found out that when no other feature is affecting the population growth the best model to predict the population is Polynomial Regressor. It has the highest coefficient of determination and the lowest MSE and MAPE. On the other side, performance of Linear Regressor was good but bad comparing to the other models as it has the highest MSE and MAPE. Furthermore, using the Polynomial Regressor the population of Bangladesh is predicted when no other factors are affecting the population and according to that prediction in the next ten years the population will be 7 million more than the current population and the growth rate will fall down from 1.12 to 0.11 which is a good sign for the country. Although, it should be mentioned this value isn't fully accurate as a lot of factors like economy, environment, politics, disease and more affects the population of a country directly. Not including other affecting features is a drawback of the work. For further research, the model's efficiency can be analyzed when a certain feature like mortality and migration affects the population of a country. Also, other machine learning models can be implemented to find out the best result Furthermore, this analysis can be done countries with a negative population growth e.g., Japan and find out how the model behaves.

REFERENCES

[1] M. Y. Dawed, P. R. Koya, and A. T. Goshu, "Mathematical Modelling of Population Growth: The Case of Logistic and Von Bertalanffy Models," *Open journal of modelling and simulation*, vol. 02, no. 04, pp. 113–126, 9 2014. [Online]. Available: https://doi.org/10.4236/ojmsi.2014.24013

[2] T. R. Malthus, *An Essay on the Principle of Population*, 7 2017. [Online]. Available: https://doi.org/10.4324/9781912281176

[3] G. Wang, J. Li, J. Sun, and X. Huang, "A structure analysis and trend prediction of the population development in china," *2009 International Conference on Business Intelligence and Financial Engineering*, 2009.

[4] N. Rahima, Suravi, S. Jahan, and H. M. S. Ali, "A STUDY ON "COMPARATIVE ANALYSIS TO PREDICT THE FUTURE POPULATION GROWTH IN INDIA, BANGLADESH, AND..." *ResearchGate*, 1 2023. [Online]. Available: https://www.researchgate.net/publication/367408901_A_STUDY_ON_COMPARATIVE_ANALYSIS_TO_PREDICT_THE_FUTURE_POPULATION_GROWTH_IN_INDIA_BANGLADESH_AND_PAKISTAN_USING_MATHEMATICAL_MODELS_EXPONENTIAL_HYPERBOLIC_LOGISTIC_GROWTH_AND_LINEAR_REGRESSION_MODEL

[5] A. M. Zabadi, R. Assaf, and M. Kanan, "A Mathematical and Statistical Approach for Predicting the Population Growth," *ResearchGate*, 8 2017. [Online]. Available: https://www.researchgate.net/publication/319083098_A_Mathematical_and_Statistical_Approach_for_Predicting_the_Population_Growth

[6] M. N. Uddin, "Prediction For Future Population Growth Of Bangladesh By Using Exponential Logistic Model - IRE Journals," 8 2019. [Online]. Available: https://www.irejournals.com/index.php/paper-details/1701470

[7] G. Tripepi, K. J. Jager, F. W. Dekker, and C. Zoccali, "Linear and logistic regression analysis," *Kidney International*, vol. 73, no. 7, pp. 806–810, 4 2008. [Online]. Available: https://doi.org/10.1038/sj.ki.5002787

[8] M. M. Otoom, "Comparing the Performance of 17 Machine Learning Models in Predicting Human Population Growth of Countries," *ResearchGate*, 1 2021. [Online]. Available: https://doi.org/10.22937/IJCSNS.2021.21.1.28

[9] C. Y. Wang and S.-J. Lee, "Regional Population Forecast and Analysis Based on Machine Learning Strategy," *Entropy*, vol. 23, no. 6, p. 656, 5 2021. [Online]. Available: https://doi.org/10.3390/e23060656

[10] B. Agyemang, "Regression Analysis of Road Traffic Accidents and Population Growth in Ghana." *International journal of business and social research*, vol. 3, no. 10, pp. 41–47, 11 2013. [Online]. Available: https://EconPapers.repec.org/RePEc:mir:mirbus:v:3:y:2013:i:10:p:41-47

[11] Kazikhalid, "GitHub - kazikhalid757/Bangladesh-population-prediction-Model: Read README file." [Online]. Available: https://github.com/kazikhalid757/Bangladesh-population-prediction-Model

[12] NaveenSurada, "GitHub - NaveenSurada/Population$_{G}rowth_{P}rediction$: $PopulationGrowthPrediction.''$

[13] Arijit-Dalui, "India-Population-Prediction-Model/Prediction.ipynb at main · arijit-dalui/India-Population-Prediction-Model." [Online]. Available: https://github.com/NaveenSurada/Population_Growth_Prediction/blob/main/Population.ipynb

[14] Yoavram, "SciComPy/population-growth.ipynb at master · yoavram/SciComPy." [Online]. Available: https://github.com/yoavram/SciComPy/blob/master/notebooks/population-growth.ipynb

[15] "World Bank Open Data." [Online]. Available: https://data.worldbank.org/indicator/SP.POP.TOTL?locations=BD&most_recent_value_desc=true