



Daffodil
International
University

**A Machine Learning Approach to Comparing Different
Regression Models to Predict Bangladesh Life Expectancy Using
Multiple Depend Features**

Submitted By

Fatema Tuj Jannat

ID: 181-35-2440

Department of Software Engineering

Daffodil International University

Supervised by

Khalid Been Md. Badruzzaman Biplob

Lecturer (Senior Scale)

Department of Software Engineering

Daffodil International University

This Project report has been submitted in fulfillment of the requirements for
the Degree of Bachelor of Science in Software Engineering.

APPROVAL

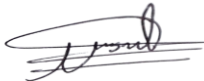
This thesis titled “A Machine Learning Approach to Comparing Different Regression Models to Predict Bangladesh Life Expectancy Using Multiple Depend Features” submitted by **Fatema Tuj Jannat, ID: 181-35-2440** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Software Engineering (SWE) and approved as to its style and contents.

BOARD OF EXAMINERS



Dr. Imran Mahmud
Associate Professor and Head
Department of Software Engineering
Daffodil International University

Chairman



Nusrat Jahan
Assistant Professor
Department of Software Engineering
Daffodil International University

Internal Examiner



Khalid Been Badruzzaman Biplob
Senior Lecturer
Department of Software Engineering
Daffodil International University

Internal Examiner



Professor Dr M Shamim Kaiser,
Professor
Institute of Information Technology
Jahangirnagar University

External Examiner

DECLARATION

I hereby state that I have taken this thesis under the supervision of **Khalid Been Md. Badruzzaman Biplob, Lecturer (Senior Scale), Department of Software Engineering, Daffodil International University**. I also acknowledge that neither this thesis nor any part of this has been submitted elsewhere for the award of any degree previously by others.



Fatema Tuj Jannat

ID: 181-35-2440

Batch 25th

Department of Software Engineering

Faculty of Science & Information Technology

Daffodil International University

Certified by:



Khalid Been Md. Badruzzaman Biplob

Lecturer (Senior Scale)

Department of Software Engineering

Faculty of Science and Information Technology

Daffodil International University

ACKNOWLEDGEMENT

This thesis I am representing was only possible to complete with guidance from some conscientious people. I want to thank each of them. Especially obligated to Daffodil International University for the direction and constant supervision by my honorable teacher **Khalid Been Md. Badruzzaman Biplob, Senior Lecture, Daffodil International University**. I would like to be thankful to my supervisor for his kind support, guidance, and encouragement and I want to express my gratitude towards my parents, teachers, batch mates, and my seniors of DIU for their kind assistance and advice to complete my study.

TABLE OF CONTENTS

Contents

APPROVAL	i
DECLARATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	vi
CHAPTER 1	1
INTRODUCTION	1
1.1 INTRODUCTION	2
1.2 RESEARCH OBJECTIVES	4
1.3 RESEARCH GAPS	4
1.4 ORGANIZATION OF THESIS	4
CHAPTER 2	5
LITERATURE REVIEW	5
2.1 LITERATURE REVIEW	6
CHAPTER 3	12
RESEARCH METHODOLOGY	12
3.1 METHODOLOGY MODEL	13
3.2 DATASET	14
3.3 PRE-PROCESSING	14
3.4 DATA VISUALIZATION.....	14
3.5 MODEL DESCRIPTION	15
CHAPTER 4	18
RESULT AND DISCUSSION	18
4.1 RESULT ANALYSIS.....	19
CHAPTER 5	23
CONCLUSION	23
5.1 FINDINGS AND CONTRIBUTIONS	24
5.2 FUTURE WORK.....	24

REFERENCES	25
PLAGARISM REPORT	28

ABSTRACT

The average longevity of a person is measured by life expectancy. The length of one's life is determined by a number of factors. We utilized GDP, rural population growth, urban population growth, services value, industry value, food production, permanent cropland, cereal production, agriculture, forestry, and fisheries value as indicators of life expectancy. We may examine all of the factors that influence life expectancy, such as the negative link between life expectancy and rural population. We can observe how personality traits are linked to life expectancy and impact how we spend our lives. To evaluate which regression models are the most accurate, we use eight different regression models. The Extreme Gradient Boosting Regressor has the greatest accuracy and the least error of all the models. It was 99 percent correct. K-Neighbors, Random Forest, and Stacking Regressor were all 94 percent accurate. Slightly Stacking was the most accurate of the bunch. We used K-Neighbors, Gradient Boosting, and Random Forest Regressor for the Stacking Regressor, and Random Forest for the meta regressor. Decision Tree has the lowest accuracy of all the models, at 79 percent. The Gradient Boosting Regressor comes in second with 96 percent accuracy. Multiple Linear Regression and Light Gradient Boosting Machine Regressor scored 88 percent and 87 percent, respectively. This study assists a country in enhancing the value of its characteristics in terms of life expectancy

Keywords: Life Expectancy. Regression Algorithms. Machine Learning.

CHAPTER 1
INTRODUCTION

1.1 INTRODUCTION

The term "life expectancy" refers to the average amount of time a person can anticipate to live. Life expectancy is a measure of a person's projected average lifespan. Life expectancy is calculated using a variety of factors such as the year of birth, current age, and demographic sex. A person's life expectancy is determined by his surroundings. Surrounding refers to the entire social system, not just society. We believe that a person's average life expectancy is determined by gender and age. To put it another way, a person's average life expectancy is influenced by a variety of factors. A person's average life expectancy differs from country to country. The environment is the explanation for this. We will discuss the average life expectancy in Bangladesh here. The average life expectancy depends on lifestyle, economic status (GDP), healthcare, diet, primary education and population. It is true that the death rate in the present is lower than in the past. The main reason is the environment. Lifestyle and Primary Education is one of the many environmental surroundings. Lifestyle depends on primary education. If a person does not receive primary education, he will not be able to be health-conscious in any way. This can lead to premature death from the damage to the health of the person. So that it has an effect on the average life expectancy of the whole country. It is true that the medical system was not good before, so it is said that both the baby and the mother would have died during childbirth. Many people have died because they did not know what medicine to take, or how much to take, because they did not have the right knowledge and primary education. It is through this elementary education that economic status (GDP) and population developed. The average lifespan varies from generation to generation. We are all aware that our life expectancy is increasing year after year. Since its independence in 1971, Bangladesh, a poor nation in South Asia, has achieved significant progress in terms of health outcomes. The expansion of the economic sector there were a lot of good things in the late twentieth century. ramifications all around the globe.

In this paper, we used some features for measure of life expectancy such as GDP, Rural Population Growth (%), Urban Population Growth (%), Services Value, Industry Value Food Production, Permanent Cropland (%), Cereal production (metric tons), Agriculture, forestry and fishing value (%). We will measure the impact of these depend feature for predict life expectancy. Use various regression model to find the most accurate model in

search of find life expectancy of Bangladesh with these depend feature. It will assist us in determining which feature aids in increasing life expectancy. This research aids a country in increasing the value of its features for life expectancy also find which regression model performs best for prediction life expectancy.

1.2 RESEARCH OBJECTIVES

- Predict Bangladesh Life Expectancy Using Machine Learning Models.
- Predict Bangladesh Life Expectancy with 11 Important Features.
- Analysis Features Importance for Bangladesh Life Expectancy.
- Comparison Between Machine Learning Models.
- Select Best Model for Predict Bangladesh Life Expectancy.

1.3 RESEARCH GAPS

- A small number of machine learning models still apply for predict life expectancy.
- Predict life expectancy without enough features that have relationship with life expectancy.
- No one analysis features importance for life expectancy.
- No one compare machine learning regression models for predict Bangladesh life expectancy.

1.4 ORGANIZATION OF THESIS

- Chapter 1: Chapter one produces the introduction of the thesis. Here also describe the research objectives and the research question.
- Chapter 2: This chapter describes the background, literature review and demonstrates previous work related to this study.
- Chapter 3: This chapter depicts the whole proposed model and architecture.
- Chapter 4: This chapter presents the experiment and result and evaluation of the studies.
- Chapter 5: This chapter concludes with future scope and limitations.

CHAPTER 2

LITERATURE REVIEW

2.1 LITERATURE REVIEW

A number of publications, studies, and research articles on life expectancy have previously been produced by a number of different writers. To coincide with our work, we've included some work-related evaluations below.

Beeksmā et al. [1] get their data from seven health-care facilities in Nijmegen, the Netherlands. There are about 33,509 EMRs in the dataset. The keyword model's predictions were accurate to the tune of 29%. While clinicians overestimated life expectancy in 63 percent of erroneous prognoses, causing delays in receiving adequate end-of-life care, the keyword model only overestimated life expectancy in 31% of inaccurate prognoses.

Andrea Nigri et al. [2] based on recurrent neural networks with a long short-term memory, a new technique for projecting life expectancy and lifespan discrepancy was devised. Their projections appear to be consistent with past patterns and physiologically sound, offering a more realistic picture of future life expectancy and disparities. The LSTM model, ARIMA model, DG model, Lee-Carter model, CoDa model, and VAR model are all examples of applied recurrent neural networks. Shown both separate and simultaneous projections of life expectancy and lifespan disparity give fresh insights for a thorough examination of the mortality forecasts, constituting a valuable technique to identify irregular death trajectories. The development of the age-at-death distribution assumes more compressed tails with time, indicating a decrease in longevity difference across industrialized nations.

Tareque, M. I. et al. [3] looked at gender disparities in the prevalence of disability and Disability-free Life Expectancy (DFLE) among Bangladeshi senior citizens. They utilized data from a nationally representative survey that included 4,189 senior people aged 60 and above, and they employed the Sullivan technique. Collect data from Bangladesh's Household Income and Expenditure Survey (HIES)-2010, a large nationally representative sample survey conducted by the BBS. The data collecting took a year to complete. There were a total of 12,240 households chosen, with 7,840 from rural regions and 4,400 from urban areas. For a total of 55,580 people, all members of chosen homes were questioned. Males made up 49.54 percent of the total, while females made up the

remainder. They discovered that at the age of 70, both men and women can expect to spend more than half of their lives disabled. Have significant consequences for the likelihood of disability, as well as the requirement for the usage of long-term care services and limitations, including to begin with, the study's data is self-reported. Proxy interviews are not mentioned in the study (HIES-2010). Individual weights are not included in the data to accommodate for the complicated sample methodology. Due to a lack of statistics, the institutionalized population was not taken into consideration. The number of senior individuals living in institutions is tiny, and they have the same health problems and impairments as the elderly in the general population.

Tareque, M. et al [4] explored the link between life expectancy and Disability-free Life Expectancy (DFLE) in the Rajshahi District of Bangladesh by investigating the connections between the Active Aging Index (AAI) and DFLE. Data was obtained during April 2009 from the Socio Demo- graphic Status of the Aged Population and Elderly Abuse study project. They discovered that urban, educated, older men are more engaged in all parts of life and have a longer DFLE. In rural regions, 93 percent of older respondents lived with family members, although 45.9% of nuclear families and 54.1 percent of joint families were noted. In urban regions, however, 23.4 percent were nuclear families and 76.6 percent were joint families, and they face restrictions in terms of several key indicators, such as the types and duration of physical activity. For a post-childhood-life table, Preston and Bennett (1983) estimate technique was used. Because related data was not available, the institutionalized population was not examined.

Tareque, M.I. et al. [5] multiple linear regression models, as well as the Sullivan technique, were utilized. They based their findings on the World Values Survey, which was performed between 1996 and 2002 among people aged 15 and above. They discovered that between 1996 and 2002, people's perceptions of their health improved. Males predicted fewer life years spent in excellent SRH in 2002 than females, but a higher proportion of their expected lives spent in good SRH. The study has certain limitations, such as the sample size is small, and the institutionalized population was not included in the HLE calculation. The subjective character of SRH, as opposed to health assessments based on medical diagnoses, may have resulted in gender bias in the results. In 2002, the response category 'very poor' was missing from the SRH survey. In 2002, there's a chance that healthy persons were overrepresented.

TAREQUE et al. [6] investigated how many years' older individuals expect to remain in excellent health, as well as the factors that influence self-reported health (SRH). By integrating SRH, they proposed a link between LE and HLE. The project's brief said that it was a socioeconomic and demographic research of Rajshahi district's elderly population (60 years and over). They employed Sullivan's approach to solve the problem. For their work, SRH was utilized to estimate HLE. They discovered that as people became older, LE and anticipated life in both poor and good health declined. Individuals in their 60s anticipated to be in excellent health for approximately 40% of their remaining lives, but those in their 80s projected just 21% of their remaining lives to be in good health, and their restrictions were more severe. The sample size is small, and it comes from only one district, Rajshahi; it is not indicative of the entire country. As a result, generalizing the findings of this study to the entire country of Bangladesh should be approached with caution. The institutionalized population was not factored into the HLE calculation.

Zaman SB et al [7] investigate and discover the relationship between healthcare spending and life expectancy and GDP in poor nations, particularly Bangladesh. They utilized STATA and multivariable logistic regression to investigate the relationship between total health spending, GDP, and life expectancy, using data from Bangladesh's "Health Bulletin 2011" and "Sample Vital Registration System 2010" from 1996 to 2006. In bi-variable analysis, they discover a direct link between total health spending and life expectancy.

Also, there is a clear link between GDP and overall health spending. They found no statistical significance between life expectancy and total health spending in multivariable analysis, and total health expenditure is more responsive to gross domestic product than life expectancy. They also have drawbacks, such as a lack of current time series data at the time of study. However, the availability of 10 fiscal year data from 1996 to 2006 provided a unique opportunity to investigate the factors that influence health spending in Bangladesh.

Khan, Hasinur Rahaman et al. [8] utilized the most recent social development measure, Literate Life Expectancy (LLE). They get their information from the Bangladesh Bureau of Statistics' (BBS) 1981 "Statistical Year Book." They were able to gather social disparities in four big demographic groups by measuring the LLE at the national level:

urban men and women, rural men and women, and young men and women. They discovered that there are significant residential and sex differences between urban and rural men and women, confirming Lutz's theory. At birth, urban males had 31.47 years of LLE, but rural men had just 18 years. During the elderly age groups, urban and rural women had fairly similar LLE levels, but from an early age, urban women began to differ significantly. The LLE method has shown to be a cutting-edge system analysis tool for assessing social progress. The use of this novel empirical approach in Bangladesh in 1981 revealed substantial social disparities depending on age group, sex, and home location.

J. Sidey-Gibbons et al. [9] utilize machine learning approaches to create three prediction models for cancer detection based on descriptions of nuclei extracted from breast masses. They developed three prediction models for cancer diagnosis using descriptions of nuclei collected from breast masses, utilizing machine learning techniques. Single-layer Artificial Neural Networks, Support Vector Machines (SVMs) with a radial basis function kernel, and General Linear Model regression (GLMs) were used. The Breast Cancer Wisconsin Diagnostic Data Set was used in this study. The trained algorithms were able to categorize cell nuclei with excellent accuracy (.94-.96), sensitivity (.97-.99), and specificity (.97-.99) according to the University of California Irvine (UCI) Machine Learning Repository (.85 - .94). The SVM technique yielded the highest accuracy (.96) and area under the curve (.97).

Ho J Y et al [10] examine whether decreases in life expectancy happened across high-income countries from 2014 to 2016 with 18 nations. They conducted a demographic study based on aggregated data and data from the World Health Organization's mortality database, which was augmented with data from Statistics Canada and Statistics Portugal, and their contribution to changes in life expectancy between 2014 and 2015. Arriaga's decomposition approach was used. They discovered that in the years 2014-15, life expectancy fell across the board in high-income nations. Women's life expectancy fell in 12 of the 18 nations studied, while men's life expectancy fell in 11 of them. They also have certain flaws, such as the underreporting of influenza and pneumonia on death certificates, the issue of linked causes of death, often known as the competing hazards dilemma, and the comparability of cause of death coding between nations.

S. S. Meshram [11] for the comparison of life expectancy between developed and developing nations, Linear Regression, Decision Tree, and Random Forest Regressor were applied. The Random Forest Regressor was chosen for the construction of the life expectancy prediction model because it had R² scores of 0.99 and 0.95 on training and testing data, respectively, as well as Mean Squared Error and Mean Absolute Error of 4.43 and 1.58. The analysis is based on HIV/AIDS, Adult Mortality, and Healthcare Expenditure, as these are the key aspects indicated by the model. Suggest that India has a higher adult mortality rate than other affluent countries due to its low healthcare spending.

Matsuo K et al [12] investigate survival predictions using clinic laboratory data in women with recurrent cervical cancer, as well as the efficacy of a new analytic technique based on deep-learning neural networks. Their retrospective study, which was authorized by their review board, looked at 157 women who got recurrent cervical cancer among 431 women diagnosed with cervical cancer between January 2008 and December 2014.

Olshansky SJ et al [13] update estimates of the impact of race and education on past and present life expectancy, examine trends in disparities from 1990 to 2008, and situate observed disparities in the context of a rapidly aging society emerging at a time of optimism about the next longevity revolution. They discovered that in 2008, adult men and women in the United States had life expectancies similar to those of all adults in the 1950s and 1960s, and that women in the United States lived longer than males at all ages, a conclusion that is consistent with prior studies. Blacks and Hispanics with sixteen or more years of education lived 7.5 and 13.6 years longer than whites with less than twelve years of education, respectively, and their disparities have widened over time, resulting in at least two “Americas” in terms of life expectancy, delineated by level of education and racial group membership. They used the Multiple Cause of Death public use data file to calculate the number of fatalities in the United States in 2008. They divided the twenty-four categories in the American Community Survey's educational attainment variable into four separate groups. As a consequence, the differences in life expectancy at birth between the most and least educated were 13.4 years for men and 7.7 years for females in 1990, compared to 14.2 years for males and 10.3 years for females in 2008, indicating that the gaps widened during the eighteen-year period.

Alam et al [14] analyzes the impact of financial development on the rapid growth of life expectancy in Bangladesh using yearly data from 1972 to 2013. They use a structural break unit root test to look at the variables' unit root characteristics. Find some research on the impact of trade openness and foreign direct investment on life expectancy in their literature review. Furthermore, the empirical findings support the occurrence of counteraction in long-run associations. Income disparity appears to reduce life expectancy in the long run, according to the long run elasticities. Finally, their results provide policymakers with fresh information that is critical to improving Bangladesh's life expectancy.

Husain, Abhar Rukh. [15] Conducts a multivariate cross-national study of national life expectancy factors. The linear and log-linear regression models are the first regression models. The data on explanatory factors comes from UNDP, World Bank, and Rudolff's yearly statistics releases (1981). His findings show that if adequate attention is paid to fertility reduction and boosting calorie intakes, life expectancies in poor nations may be considerably enhanced.

M. A. Rubi et al [16] used two independent variables Bangladesh's Gross Domestic Product (GDP) & Population and studied the interaction between GDP and Population with Life Expectancy (LE). They used a long period of data from World Data Open Data (WBOD) and Trends Economics from 1960 to 2020. They find a strong correlation between population sizes with life expectancy after applying the Multiple Linear Regression (MLR) Model and several Artificial Neural Network (ANN) and found 98% accuracy in MLR Model. Their study shows that the population size of a country grows as its population health improves and its socioeconomic status improves and the most fascinating finding is in one word that population size has an impact on life expectancy. In this study, they suggest that future study research be expanded with more data and machine learning algorithms.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 METHODOLOGY MODEL

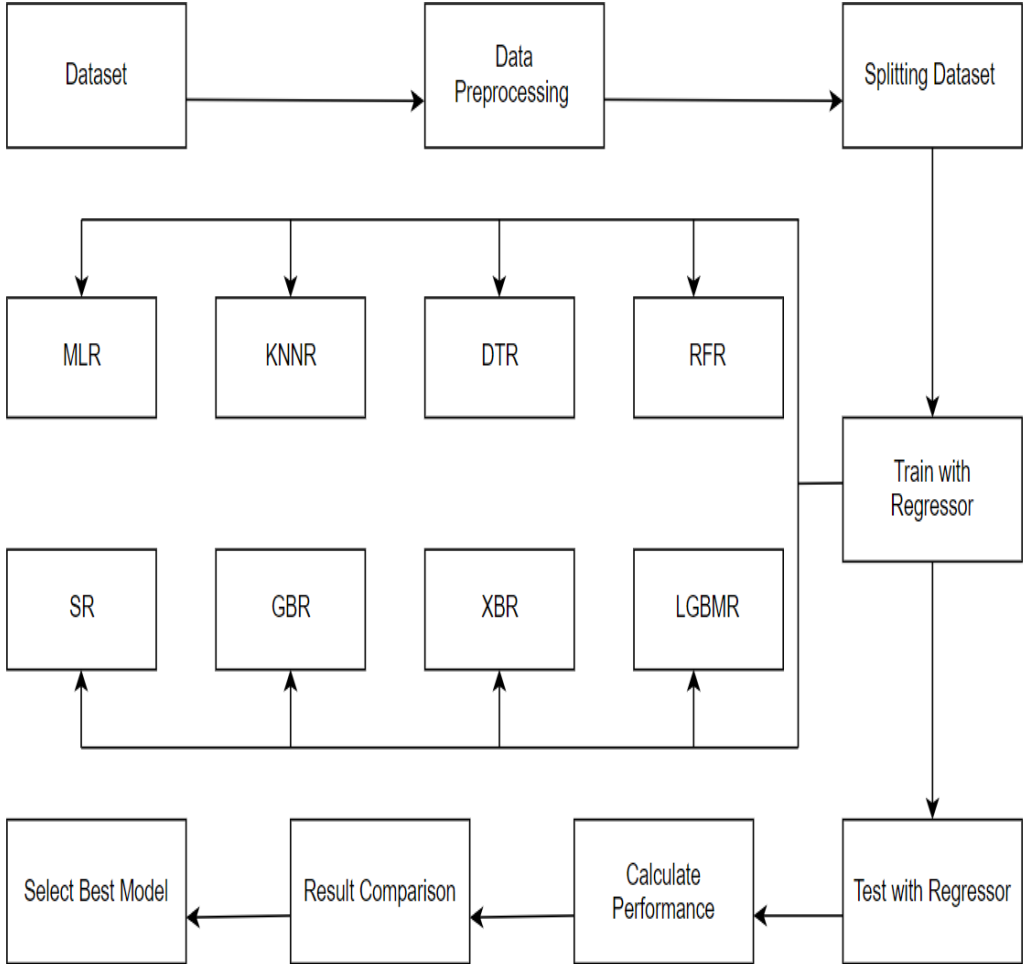


Figure: 3.1.1 Methodology Model

3.2 DATASET

The information was gathered from the Trends Economics. Data set contain data from 1960 to 2020. Combine all of the characteristics that are linked to Bangladesh's Life Expectancy.

3.3 PRE-PROCESSING

Pre-processing, which includes data cleansing and standardization, noisy data filtering, and management of missing information, is necessary for machine learning to be done. Any data analysis will succeed if there is enough relevant data. The information was gathered from the Trends Economics. Data set contain data from 1960 to 2020. Combine all of the characteristics that are linked to Bangladesh's Life Expectancy. After you've checked for null values. We find some null values that fill up with the mean values. We examining the relationship between the dependent and independent features. For prediction, we select accurate features. After defining independent and dependent features, the data set is split. Divide 80 percent of the data for training and 20 percent for testing. This study was carried out in jupyter Notebook using the Python programming language.

3.4 DATA VISUALIZATION

A	B	C	D	E	F	G	H	I	J	K
Year	Life Expectancy	GDP	Rural Population Growth(%)	Urban Population Growth(%)	Services Value	Industry Value	Food Production	Permanent Cropland(%)	Cereal production (metric tons)	Agriculture, forestry and fishi
1960	45.82931707	4274893913		5.135	5.32E+11	1430991500				
1961	46.45792683	4817580184	2.620711539	5.278	5.35E+11	1571098000	26.13	2.112621956	14523696	5.208136666
1962	47.08397561	5081413340	2.543085144	5.498	5.39E+11	1942448400	25.01	2.151033264	13408128	5.246177198
1963	47.69246342	5319458351	2.574066034	5.727	5.63E+11	1992778000	28.33	2.112621956	16042664	-3.364688932
1964	48.25543902	5386054619	2.655145112	5.964	5.70E+11	2694671100	28.52	2.151033264	15842004	9.538324061
1965	48.68953659	5906636557	2.749118469	6.211	6.00E+11	3040176600	28.72	1.997388031	15852140	-1.066446185
1966	48.89541463	6439687598	2.876922837	6.467	6.20E+11	3188444700	27.58	1.935929938	14464297	2.5904944
1967	48.83814634	7253575399	2.945648049	6.733	6.57E+11	2821175000	31.18	1.935929938	16876448	-2.693533641
1968	48.53670732	7483685474	2.845475794	7.009	6.55E+11	3327191200	31.95	1.843742798	17160028	10.27067214
1969	48.05153659	8471006101	2.541141117	7.296	7.05E+11	3735268200	33.26	2.035799339	18187398	0.972407135
1970	47.52541463	8992721809	2.126148837	7.593	6.97E+11	3767914400	31.74	1.997388031	16905215	5.422769301
1971	47.13858537	8751842840	1.668529113	7.901	6.50E+11	3227892100	29.57	1.982023508	15088878	-4.508098223
1972	47.03431707	6288245867	1.307691731	8.221	6.13E+11	2287953900	29.11	1.982023508	15323537	-10.70660626
1973	47.2955122	8086725729	1.144859431	8.553	6.66E+11	7620165000	32.09	1.982023508	18021512	0.272871769
1974	47.93012195	12512460520	1.092699617	9.034	6.60E+11	13329167900	31.19	1.982023508	17105414	6.438208427
1975	48.87080488	19448348073	1.022261983	9.836	6.99E+11	20043400600	34.35	1.997388031	19322564	-4.563524652
1976	49.98773171	10117113333	1.274448782	10.701	7.22E+11	21708357000	32.94	1.997388031	17908126	8.387863927
1977	51.10568293	9651149302	1.450569154	11.63	7.75E+11	24977056000	35.08	1.997388031	19772481	-3.658242081
1978	52.08982927	13281761143	1.53116518	12.629	8.02E+11	28736810700	36.31	2.012752554	19984399	7.834220164
1979	52.88729268	15565480322	1.483663549	13.701	7.87E+11	37162248400	36.08	2.035799339	19658591	-0.64807058
1980	53.48804878	18138049096	1.344209046	14.851	8.57E+11	56483000000	36.65	2.043481601	21698327	-1.518205102
1981	53.92802439	20249594002	1.515033479	15.801	8.87E+11	65534000000	36.4	2.089575171	21591389	3.303884672
1982	54.30307317	18525399202	2.123518263	16.212	9.16E+11	74409000000	37.72	2.104939694	22339648	1.023737189
1983	54.68956098	17609048822	2.098112053	16.631	9.52E+11	87325000000	38.66	2.112621956	22990185	3.889743452
1984	55.11443902	18920840000	2.089691663	17.06	9.92E+11	1.01E+11	39	2.112621956	23256274	4.905098749
1985	55.59065854	22278423077	2.093522784	17.496	1.03E+12	1.18E+11	40.41	2.112621956	24134965	0.190430821
1986	56.11519512	21774033333	2.094166609	17.941	1.07E+12	1.33E+11	40.81	2.151033264	24266331	3.325031425
1987	56.66858537	24298032258	2.07542695	18.395	1.11E+12	1.48E+11	40.87	2.151033264	24304321	0.095033045

Figure: 3.1.2 Dataset

3.5 MODEL DESCRIPTION

- 1) Multiple Linear Regression (MLR): A statistical strategy for predicting the outcome of a variable using the values of two or more variables is known as multiple linear regression. Multiple regression is a type of regression that is an extension of linear regression. The dependent variable is the one we want to predict, while the independent or explanatory factors are used to predict the dependent variable's value. The formula for multiple linear regression is as follows:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \dots + \beta_nX_n + \epsilon \quad (1)$$

- 2) K-Neighbors Regressor (KNNR): It's a non-parametric strategy for logically averaging data in the same neighborhood to approximate the link between independent variables and continuous outcomes. To discover the neighborhood size that minimizes the mean-squared error, the analyst must define the size of the neighborhood.
- 3) Decision Tree Regressor (DTR): A decision tree is a tree structure that looks like a flowchart and is used to model decisions. A supervised learning approach, the decision tree technique is categorized. It may be utilized with both categorical and continuous output variables. The Decision Tree method has become one of the most commonly used machine learning algorithms. The use of a Decision Tree can help with both classification and regression difficulties.
- 4) Random Forest Regressor (RFR): A random forest is a meta estimator that fits a number of classification decision trees on distinct sub-samples of the dataset and utilizes averaging to increase prediction accuracy and control over-fitting. The forest's total amount of trees. A Random Forest is an ensemble method for solving regression and classification problems that use several decision trees with the Bootstrap and Aggregation methodology. Rather of relying on individual decision

trees to decide the ultimate outcome, the fundamental concept is to combine many decision trees. Random Forest employs several decision trees as a foundation learning paradigm.

- 5) **Stacking Regressor (SR):** The phrase "stacking" or "stacked" refers to the process of stacking objects. Each estimator's output is piled, and a regressor is used to calculate the final forecast. By feeding the output of each individual estimate into a final estimator, you may take use of each estimate's strengths. Using a meta-learning technique, it learns how to combine predictions from two or more fundamental machine learning algorithms. On a classification or regression problem, stacking has the benefit of combining the talents of a number of high-performing models to create predictions that surpass any one model in the ensemble.
- 6) **Gradient Boosting Regressor (GBR):** Gradient Boosting Regressor is a forward stage-wise additive model that allows any differentiable loss function to be optimized. At each level, a regression tree is fitted based on the negative gradient of the supplied loss function. It's one of the most efficient ways to build predictive models. It was feasible to build an ensemble model by combining the weak learners or weak predictive models. The gradient boosting approach can help with both regression and classification issues. The Gradient Boosting Regression technique is used to fit the model that predicts the continuous value.
- 7) **Extreme Gradient Boosting Regressor (XGBR):** Extreme Gradient Boosting is an open-source application that executes the gradient boosting approach efficiently and effectively. Extreme Gradient Boosting (EGB) is a machine learning technique for regression, classification, and other problems that builds a prediction model from a set of weak prediction models, most commonly decision trees. The resultant technique is called gradient boosted trees, and it often beats random forest when a decision tree is the weak learner. It uses the same step-by-step approach as previous boosting approaches, but it broadens the scope by allowing optimization of any differentiable loss function.
- 8) **Light Gradient Boosting Machine Regressor (LGBMR):** Light Gradient Boosted Machine is an open-source toolkit that efficiently and effectively implements the

gradient boosting approach. LightGBM enhances the gradient boosting approach by incorporating automated feature selection and focusing on boosting situations with larger gradients. This might result in a considerable boost in training speed as well as improved prediction accuracy. As a result, LightGBM has been the de facto technique for machine learning contests when working with tabular data for regression and classification predictive modeling tasks.

- 9) Mean Absolute Error (MAE): The MAE is a statistic for evaluating regression models. The mean absolute error of a model with regard to a test set is the average of all individual prediction errors on all occurrences in the test set. The discrepancy between the true and expected value for each occurrence is referred to as a prediction error. The following is the formula:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |A_i - A|$$

- 10) Mean Squared Error (MSE): MSE indicates how near it is to a set of points. It accomplishes this by squaring the distances between the points and the regression line. Squaring is required to eliminate any undesirable signs. Inequalities with greater magnitude are also given more weight. The fact that you're computing the average of a series of errors gives the mean squared error its name. The better the prediction, the smaller the MSE. The following is the formula:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n |[\text{Actual} - \text{Prediction}]|$$

- 1) Root Mean Square Error (RMSE): The residuals' standard deviation is RMSE. Residues measure the distance between data points and the regression line, and the RMSE is a measure of how spread out these residuals are. The following is the formula:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n |A_i - B_i|^2}$$

CHAPTER 4
RESULT AND DISCUSSION

4.1 RESULT ANALYSIS

The life expectancy of a country is determined by a number of variables. Figure: 4.1.1 depicted the pairwise association between life expectancy and a variety of independent characteristics such as GDP, Rural Population Growth (%), Urban Population Growth (%), Services Value, Industry Value Food Production, Permanent Cropland (%), Cereal production (metric tons), Agriculture, forestry and fishing value (%).

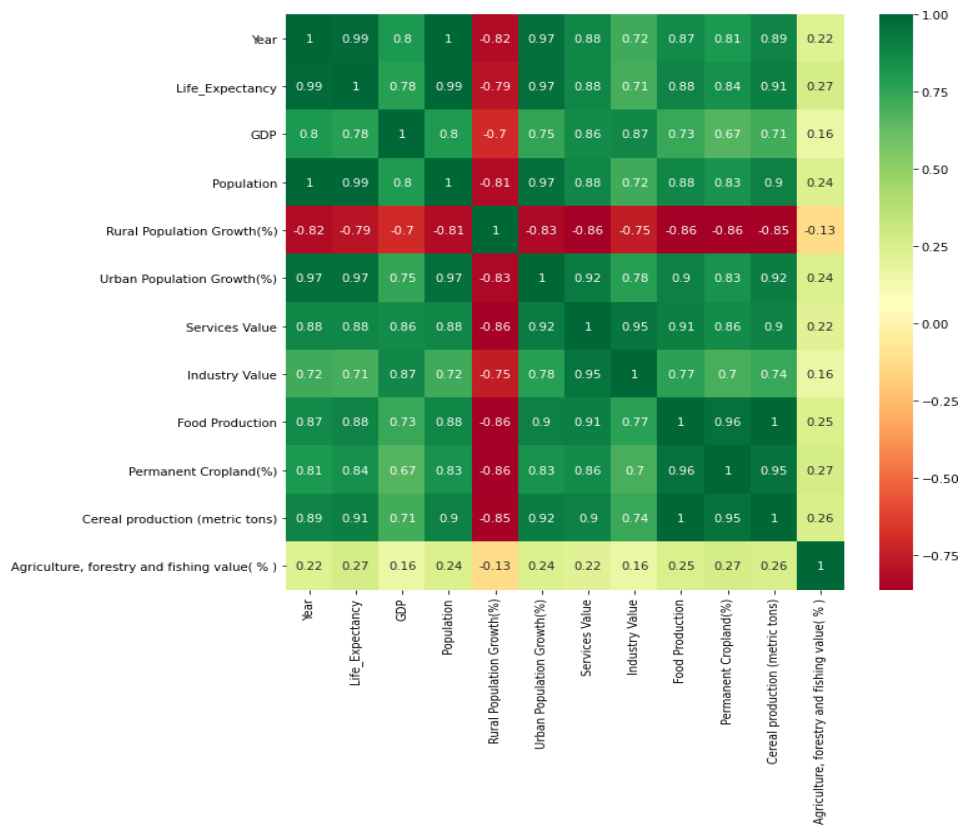


Figure: 4.1.1. Correlation Between Features

Figure: 4.1.2 shows that the data reveals the value of GDP that has risen steadily over time. As a consequence, GDP in 1960 was 4274893913.49536, whereas GDP in 2020 was 353000000000. It was discovered that the value of GDP had risen. The two factors of life expectancy and GDP are inextricably linked. The bigger the GDP, the higher the standard of living will be. As a result, the average life expectancy may rise. Life expectancy is also influenced by service value and industry value. The greater the service and industry values are, the better the quality of life will be. As can be seen, service value and industry value have increased significantly year after year, and according to the most recent update in 2020, service value has increased significantly, and now stands at 5460000000000. And the industry value was 7540000000000, which has a positive impact on daily life. Food production has an influence on life expectancy and quality of life. Our level of living will improve if our food production is good, and this will have a positive influence on life expectancy. From 1990 through 2020, food production ranged between 26.13 and 109.07. Agriculture, forestry and fishing value percent also shortly involved with life expectancy.

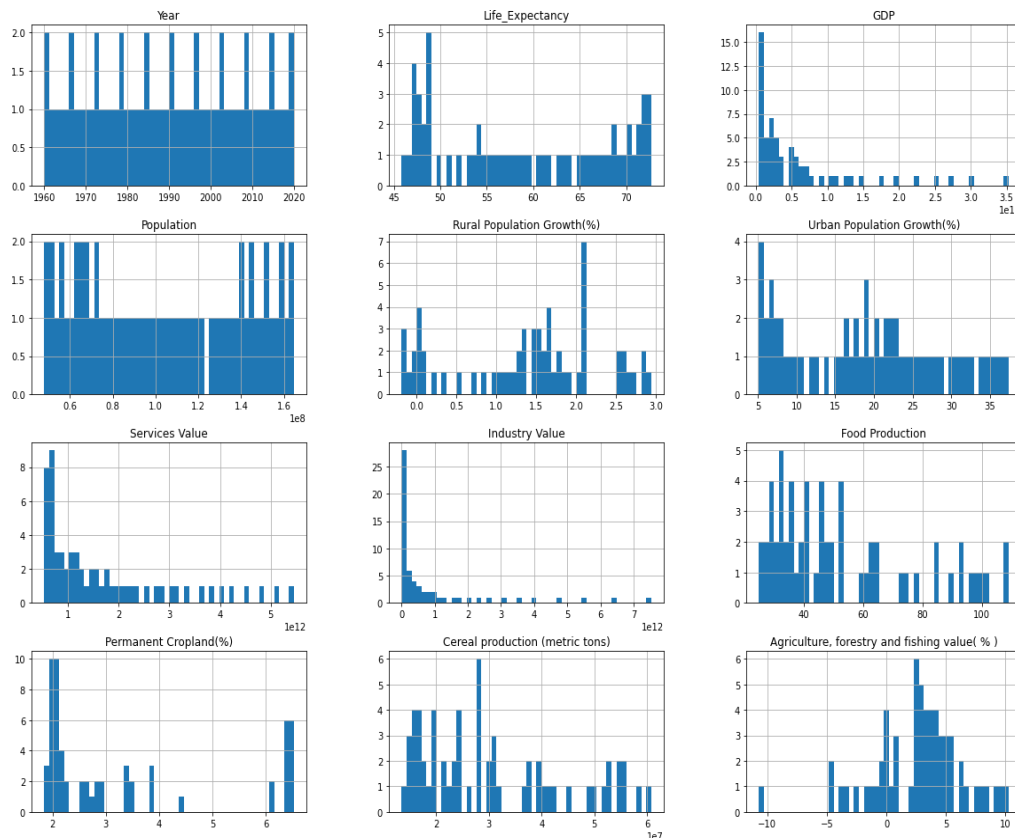


Figure: 4.1.2. Visualize of All Feature

Figure: 4.1.3 and Figure: 4.1.4 shows there are two types of population growth: rural and urban in 1990's century urban population percent was more than rural and year by year rural population growth was decrease and urban population growth was increase. The level of living improves as more people move to the city.

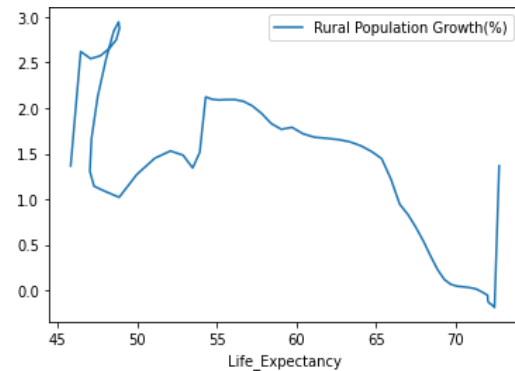
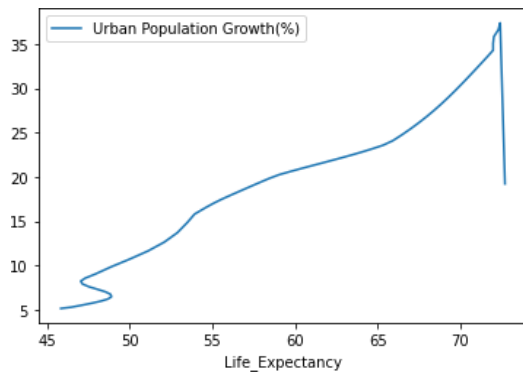


Figure: 4.1.3 Urban Population Growth % Figure:4.1.4 Rural Population Growth %

MODELS	MAE	MSE	RMSE	ACCURACY
Multiple Linear Regression	1.46	8.82	2.97	88.07%
K-Neighbors Regressor	0.96	4.17	2.04	94.35%
Decision Tree Regressor	2.63	15.30	3.91	79.32%
Random Forest Regressor	1.06	4.28	2.06	94.21%
Stacking Regressor	1.02	3.90	1.97	94.72%
Gradient Boosting Regressor	0.94	2.43	1.55	96.71%
Extreme Gradient Boosting Regressor	0.58	0.44	0.66	99.39%
Light Gradient Boosting Machine Regressor	2.62	9.57	3.09	87.06%

Table: 4.1.1 Error and Accuracy Comparison

Figure: 4.1.1 shows that life expectancy and rural population growth have a negative relationship. We can see how these characteristics are intertwined with life expectancy and have an influence on how we live our lives. It worth has fluctuated over time. Its value has fluctuated in the past, increasing at times and decreasing at others. We drop Rural population growth and Agriculture, Forestry and Fishing value as it was having negative correlation and less correlation between life expectancy.

Table: 4.1.1 Shows that we utilize eight different regression models to determine which models are the most accurate. Among all the models, the Extreme Gradient Boosting Regressor has the best accuracy and the least error. It was 99 percent accurate. The accuracy of K-Neighbors, Random Forest, and Stacking Regressor was 94 percent. Among them, Slightly Stacking had the highest accuracy. We utilized three models for the stacking regressor: K-Neighbors, Gradient Bosting, and Random Forest Regressor, and Random Forest for the meta regressor. Among all the models, Decision Tree has the lowest accuracy at 79 percent. With 96 percent accuracy, the Gradient Boosting Regressor comes in second. 88 percent and 87 percent for Multiple Linear Regression and Light Gradient Boosting Machine Regressor, respectively.

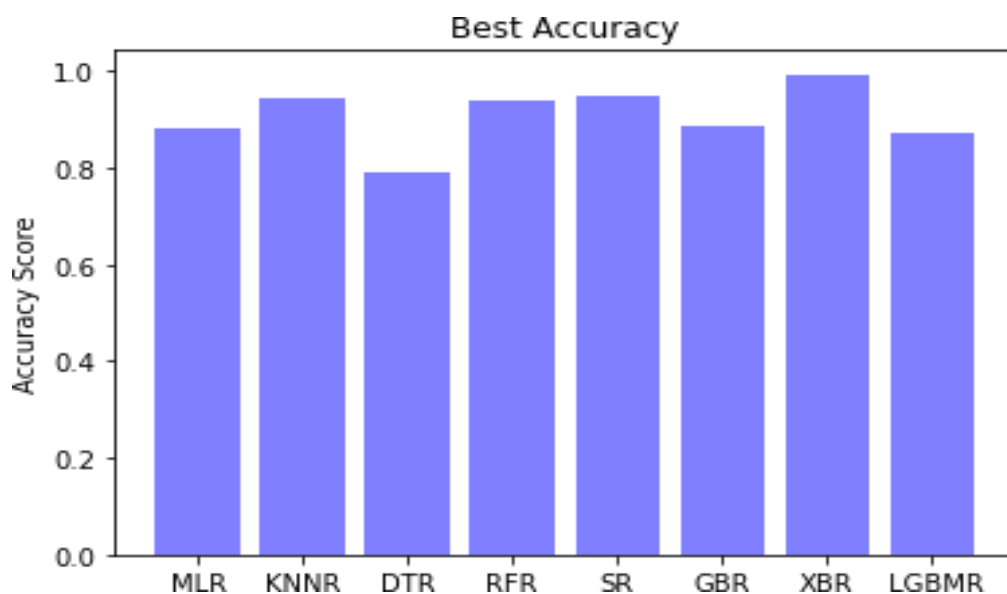


Figure: 4.1.5 Accuracy Among All the Models

Figure: 4.1.5 shows the accuracy among all the model. The Extreme Gradient Boosting Regressor has the best accuracy.

CHAPTER 5 CONCLUSION

5.1 FINDINGS AND CONTRIBUTIONS

A country's life expectancy is affected by a variety of factors. showed the pairwise relationship between life expectancy and a number of independent variables. Make a prediction with the help of a machine learning model. Extreme Gradients Boosting Regressors in general forecast better than other Regressors. Our findings lead us to conclude that life expectancy may be predicted using GDP, urban population growth (percentage), services value, industry value, food production, permanent cropland (percentage), and cereal output (metric tons). Larger data sets may result in more accurate predictions

5.2 FUTURE WORK

In future we can use deep learning model. Also we can take more data for make prediction better. In the future, more data and new machine learning models will be used to enhance prediction.

REFERENCES

- [1] Beeksmā, M., Verberne, S., van den Bosch, A. et al. Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Med Inform Decis Mak* 19, 36 (2019). <https://doi.org/10.1186/s12911-019-0775-2>
- [2] Andrea Nigri, Susanna Levantesi & Mario Marino (2021) Life expectancy and lifespan disparity forecasting: a long short-term memory approach, *Scandinavian Actuarial Journal*, 2021:2, 110-133, DOI: 10.1080/03461238.2020.1814855
- [3] Tareque, M. I., Begum, S., & Saito, Y. (2013). Gender Differences in Disability-Free Life Expectancy at Old Ages in Bangladesh. *Journal of Aging and Health*, 25(8), 1299–1312. <https://doi.org/10.1177/0898264313501388>
- [4] Tareque, M., Hoque, N., Islam, T., Kawahara, K., & Sugawa, M. (2013). Relationships between the Active Aging Index and Disability-Free Life Expectancy: A Case Study in the Rajshahi District of Bangladesh. *Canadian Journal on Aging / La Revue Canadienne Du Vieillissement*, 32(4), 417-432. doi:10.1017/S0714980813000494
- [5] Tareque, M.I., Saito, Y. & Kawahara, K. Healthy life expectancy and the correlates of self-rated health in Bangladesh in 1996 and 2002. *BMC Public Health* 15, 312 (2015). <https://doi.org/10.1186/s12889-015-1640-6>
- [6] TAREQUE, M., ISLAM, T., KAWAHARA, K., SUGAWA, M., & SAITO, Y. (2015). Healthy life expectancy and the correlates of self-rated health in an ageing population in Rajshahi district of Bangladesh. *Ageing and Society*, 35(5), 1075-1094. doi:10.1017/S0144686X14000130
- [7] Zaman SB, Hossain N, Mehta V, Sharmin S, Mahmood SAI. An Association of Total Health Expenditure with GDP and Life Expectancy. *J Med Res Innov.* 2017;1(2):AU7-AU12. DOI: 10.15419/jmri.72
- [8] Khan, Hasinur Rahaman and Asaduzzaman, Md., Literate Life Expectancy in Bangladesh: A New Approach of Social Indicator (January 2007). *Journal of Data Science*, Vol. 5, pp. 131-142, 2007, Available at SSRN: <https://ssrn.com/abstract=2021511>

- [9] Sidey-Gibbons, J., Sidey-Gibbons, C. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 19, 64 (2019). <https://doi.org/10.1186/s12874-019-0681-4>
- [10] Ho J Y, Hendi A S. Recent trends in life expectancy across high income countries: retrospective observational study *BMJ* 2018; 362 :k2562 doi:10.1136/bmj.k2562
- [11] S. S. Meshram, "Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning," 2020 IEEE Bombay Section Signature Conference (IBSSC), 2020, pp. 6-10, doi: 10.1109/IBSSC51096.2020.9332159.
- [12] Matsuo K, Purushotham S, Moeini A, Li G, Machida H, Liu Y, Roman LD. A pilot study in using deep learning to predict limited life expectancy in women with recurrent cervical cancer. *Am J Obstet Gynecol.* 2017 Dec;217(6):703-705. doi: 10.1016/j.ajog.2017.08.012. Epub 2017 Aug 24. PMID: 28843741; PMCID: PMC7534808.
- [13] Olshansky SJ, Antonucci T, Berkman L, Binstock RH, Boersch-Supan A, Cacioppo JT, Carnes BA, Carstensen LL, Fried LP, Goldman DP, Jackson J, Kohli M, Rother J, Zheng Y, Rowe J. Differences in life expectancy due to race and educational differences are widening, and many may not catch up. *Health Aff (Millwood)*. 2012 Aug;31(8):1803-13. doi: 10.1377/hlthaff.2011.0746. PMID: 22869659.
- [14] Alam, MS, Islam, MS, Shahzad, SJH, Bilal, S. Rapid rise of life expectancy in Bangladesh: Does financial development matter? *Int J Fin Econ.* 2020; 1– 14. <https://doi.org/10.1002/ijfe.2046>
- [15] Husain, Abhar Rukh. "Life Expectancy in Developing Countries: A Cross-Section Analysis." *The Bangladesh Development Studies*, vol. 28, no. 1/2, 2002, pp. 161–178. JSTOR, www.jstor.org/stable/40795653. Accessed 11 Sept. 2021.
- [16] M. A. Rubi, H. I. Bijoy and A. K. Bitto, "Life Expectancy Prediction Based on GDP and Population Size of Bangladesh using Multiple Linear Regression and ANN Model," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579594.
- [17] "Bangladesh - Gross Domestic Product (GDP) 2026 | Statista." Statista. N.p, 2021.

Web. 31 May 2021. <https://www.statista.com/statistics/438219/gross-domestic-product-gdp-in-bangladesh/>

[18] "Bangladesh - Total Population 2016-2026 | Statista." Statista. N.p., 2021. Web. 31 May 2021. <https://www.statista.com/statistics/438167/total-population-of-bangladesh/>

PLAGARISM REPORT

1/25/22, 3:29 PM

Turnitin

Turnitin Originality Report

Processed on: 25-Jan-2022 15:27 +06

ID: 1747759202

Word Count: 5007

Submitted: 1

181-35-2440 By Fatema Tuj Jannat

Similarity Index

22%

Similarity by Source

Internet Sources: 14%

Publications: 19%

Student Papers: 8%

3% match (publications)

[Maksuda Akter Rubi, Hasan Imam Bijoy, Abu Kowshir Bitto. "Life Expectancy Prediction Based on GDP and Population Size of Bangladesh using Multiple Linear Regression and ANN Model", 2021 12th International Conference on Computing, Communication and Networking Technologies \(ICCCNT\), 2021](#)

2% match (Internet from 19-Oct-2017)

<http://content.healthaffairs.org/content/31/8/1803.full.html>

1% match (Internet from 09-Nov-2021)

<https://bmcmmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0681-4>

1% match (student papers from 22-Nov-2017)

[Submitted to Andrews University on 2017-11-22](#)

1% match (publications)

[Hasinur Rahaman Khan, Md, A. M. Azharul Islam, and Faisal Ababneh. "Substantial gender gap reduction in Bangladesh explained by the proximity measure of literacy and life expectancy", Journal of Applied Statistics, 2016.](#)

1% match (student papers from 07-Apr-2018)

Class: Article 2018

Assignment: Journal Article

Paper ID: [942523876](#)