iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** IV **Month of publication:** April 2023

DOI: https://doi.org/10.22214/ijraset.2023.50140

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

Life Expectancy using Data Analytics

Dr. Renuka Deshpande[1], Vaishnavi Uttarkar[2]
[1, 2]*Department of Computer Engineering, Shivajirao S. Jondhale College of Engineering, Dombivli*

Abstract: ***The population's life expectancy is a problem for the government or pertinent agency of a particular nation. A statistical measurement of the anticipated average lifespan of an organism is life expectancy. Data for the appropriate nation, including all required fields, must be gathered in order for the department to handle life expectancy in an efficient manner. This study analyses and forecasts trends in life expectancy using data analytics and machine learning approaches. Regression models are used in the study to find the major determinants of life expectancy as well as descriptive statistics to understand how long people live across various population groupings. The WHO data repository was used to gather the information, which was then retrieved from the Kaggle website, a reputable data science resource. The dataset contains variables including vaccination, mortality, economy, social factors, and other health-related variables. It covers the years 2000 to 2015 for 193 nations. These factors are taken into account when this study uses data analytics and machine learning to implement life expectancy. After examining numerous regression methods, the Random Forest Regressor was chosen since it generated the greatest accuracy among the models studied.***

I. INTRODUCTION

Life expectancy is a critical indicator of a population's overall health and well-being. In recent years, the use of data analytics and machine learning has emerged as a powerful tool for improving our understanding of life expectancy trends and predicting future outcomes. By collecting and analyzing large amounts of data from various sources, such as demographic data, medical records, and environmental factors, thus, we can create predictive models that can provide valuable insights into the factors that influence life expectancy using data analytics and machine learning. Life expectancy refers to the average number of years a person can expect to live, based on current age-specific mortality rates. The study of the life expectancy of a population is important for the evaluation of the degree of economic and social development of a country [1]. The residents of a country with high life standards live longer, on average, and have a small mortality ratio [1]. Data analytics and Machine learning can help identify patterns and relationships between different variables and predict future outcomes with high accuracy. This information can be used to develop targeted interventions to improve health outcomes and increase life expectancy.

II. PROBLEM DEFINITION

The purpose of this study is to understand and predict trends in life expectancy, a crucial aspect of public health, using data analytics and machine learning techniques. The study aims to identify the key predictors of life expectancy and analyze their influence on the average expected lifespan. The study is based on a dataset covering the period from 2000 to 2015 for 193 countries, collected from the WHO data repository and extracted from the Kaggle website. The dataset includes factors such as country-wise population, deaths by different age groups, diseases such as Measles and HIV/AIDS, and immunization rates for Polio and Hepatitis B, among others. The study uses regression analysis to analyze the data, with the Random Forest Regressor algorithm chosen as the best performer, achieving an accuracy of 95%. The results of this study will have significant implications for public health policy and practice, providing insight into the factors contributing to life expectancy. The study is implemented using data analytics and machine learning techniques, with the help of a dataset provided by WHO and extracted from Kaggle. Data retrieval is done using SQL, and data visualizations are created using Tableau. The Random Forest Regression is a machine learning model employed in this study to train the model and predict the life expectancy of a given country.

III. LITERATURE SURVEY

Literature survey of the system proposed by our team is as follows:

A. Literature Survey

Data analysis process that involves filling null values with mean values, building a multiple linear regression (MLR) model to explore the relationship between variables, using a stepwise regression approach to fit the data, and applying cluster analysis to identify homogeneous groups of countries[1]. The aim is to find statistically significant variables that can help explain the relationship between the variables under study[1]. A demographic methodology is employed to calculate life expectancy, utilizing a recursive smoothing technique to generate estimations of the mean duration lived by individuals belonging to a specific age

cohort[2]. The approach of starting with the general population figure and subtracting years for adverse factors is known as a top-down approach[3]. While it may be appropriate for situations with near-normal life expectancy, it becomes unreasonable when dealing with medical conditions that significantly affect morbidity and mortality patterns[3]. The study used Standardized Mortality Ratios (SMRs) to compare mortality rates among individuals with traumatic SCI in Great Britain[4]. SMRs were calculated using age at injury and current attained age, and the results were grouped by age at injury and current age to demonstrate significant differences[4]. Additionally, life expectancy estimates were calculated using SMRs based on current age and compared to the general population's period and cohort life tables[4]. The research explores factors contributing to low life expectancy in different countries using demographic features, income composition, and mortality rates[5]. The dataset used is specific to each country, and various regression techniques such as Simple Linear Regression, Multiple Linear Regression, Polynomial Regression, and Decision Tree are utilized to predict the impact of individual factors on life expectancy[5]. The objective is to help countries identify the primary contributing factor to low life expectancy in specific areas[5]. This study looked at how South Korea's air pollution levels impacted life expectancy and discovered that doing so considerably lowers it[6].The study looked at how South Korea's air pollution levels impacted life expectancy and discovered that doing so considerably lowers it.[7]. The paper reviewed the existing literature on the relationship between dietary patterns and life expectancy and highlighted the importance of healthy dietary habits for longer life expectancy[8].

The study examined the role of gender and family support in life expectancy by analyzing data from older adults in different countries[9]. The paper compared the relationship between life expectancy and GDP per capita in India and China, highlighting the significant differences in life expectancy between the two countries[10]. They estimated life expectancy at birth for different states and union territories in India using a Bayesian hierarchical model[11]. They investigated the impact of education on life expectancy in China by analyzing data from the China Family Panel Studies[12]. They examined the effect of health insurance on life expectancy in China by analyzing data from the New Rural Cooperative Medical Scheme[13]. They utilized the Lee-Carter model to assess life expectancy and mortality rates in different countries[14]. Analyzed the effects of alcohol consumption on life expectancy using a Markov model[15]. One paper among others studied the factors affecting life expectancy in India by analyzing data from various sources[16]. They investigated the effect of occupational health and safety on life expectancy in China's coal mining industry[17]. Analyzed the impact of environmental pollution on life expectancy in Malaysia[18]. One of the studies examined the impacts of air pollution on life expectancy in China[19]. The paper analyzed the effect of health expenditure on life expectancy in Pakistan using ARDL (autoregressive distributed lag) modeling[20]. The study analyzed the effects of medical insurance on life expectancy in China using data from China's basic medical insurance system[21]. Analyzed the factors influencing life expectancy using big data techniques[22]. They examined the relationship between socioeconomic development and life expectancy in developing countries[23]. The paper used a Bayesian model averaging approach to identify the determinants of life expectancy at birth in China[24]. The study analyzed the impact of health expenditures on life expectancy in ASEAN-5 countries using a dynamic panel data approach[25]. The paper reviewed the literature on factors affecting life expectancy in Africa[26]. The study used a Markov model to examine life expectancy, healthy life expectancy, and healthcare expenditure at birth and at age 60 in China[27]. The paper examined the life expectancy and health status of older people in Vietnam using data from a longitudinal survey[28]. The study analyzed the regional disparities and determinants of life expectancy at birth in China from 2000 to 2015 using a spatiotemporal analysis[29]. The paper examined how pollution impacts life expectancy in China using data from prefecture-level cities[30].

B. Limitations Of Literature Survey

These published papers implemented basic visualizations such as scatter plots, graphs, etc. Each factor is depicted using separate scatter plots or other kinds of graphs. They performed analysis on specific countries, for example, the United Kingdom and the United States, or just focused on disease or condition such as adjustment of current Spinal Cord Injury (SCI) in life expectancy results to reflect possible future improvements as have occurred in the general population. It focuses only on five specified diseases affecting life expectancy. Most of the aim is on predicting factors but minorly deals with visualizations.

C. Research Gap

According to the literature survey, there are basic visualizations of the datasets and each factor is visualized separately. But the relationship between the trends is not depicted in a single dashboard. In terms of prediction, none of the papers has considered each and every feature of the dataset. Thus, our project focuses to overcome these drawbacks.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

IV. METHODOLOGY

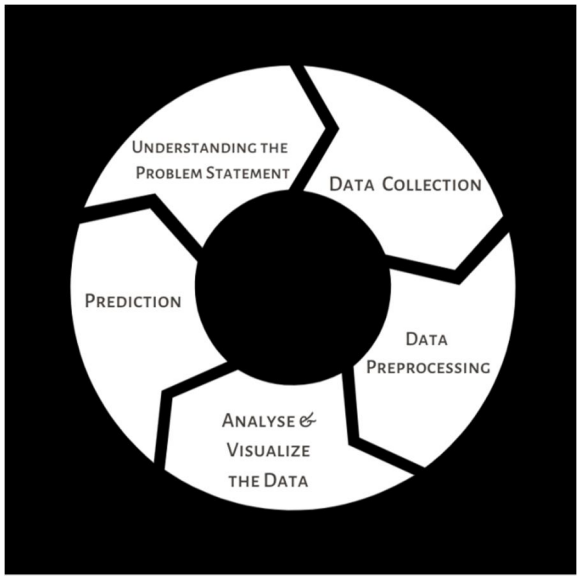


Fig 1: Procedural Design

In figure 6.1, Procedural Design is represented which depicts the whole process of this project. There are a total of five steps that define the whole process.

A. Understanding the Problem Statement

The problem statement is derived from a literature survey, while several types of analyses are done on life expectancy, none include a high degree of visualization. Most of the papers aimed to consider very few features from the dataset and then concentrated on them widely whereas in this project we had covered all features of the dataset such as a range of variables, including country-specific population figures, deaths disaggregated by age cohorts, the prevalence of diseases such as Measles and HIV/AIDS, etc and vaccination rates for Polio and Hepatitis B, etc. The count of features of the dataset is 22 columns, on the base of all these features life expectancy is predicted using a machine learning algorithm.

B. Data Collection

We require pertinent data to work with in order to do an analysis of life expectancy. As the data must be trustworthy, we must extract it from trustworthy sites, and Kaggle is one of them. The dataset which has been used in this project is the same dataset that has been used by published papers that we have used as references. The dataset contains the following variables: country name, year, life expectancy in years, GDP, BMI, alcohol consumption, population with access to improved sanitation facilities, the incidence of HIV/AIDS, measles, and other communicable diseases, as well as vaccination rates for polio, hepatitis B, and other diseases. The dataset also includes variables that reflect various areas and income levels to aid in analyzing how these factors affect life expectancy in diverse socioeconomic contexts and geographic locations. The information is useful for studying and analyzing factors that affect life expectancy rates across various nations and regions. Hence, this dataset has been an asset to our project to conclude all the factors and insights.

C. Data Preprocessing

Data cleaning has a position in data preprocessing. Raw data is a disaster in the real world. It is frequently lacking and lacks a regular, consistent design, and it may not just be inaccurate and inconsistent. As the original dataset contains null values and has a preset format. As the preset format was up to mark and didn't need any changes but there were many null values also called missing values for better understanding in normal words. For assigning appropriate values in the missing place we substituted with a mean value of the whole column of that particular column, this method was for numerical columns. For string features, mode is used to find the appropriate value for missing values. This preprocessing is done by machine learning in Python.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

D. Analyze & Visualize the Data

The data that have been gathered are summarized in this step. It's important to prepare the data correctly before you analyze it. Finding patterns, correlations, or trends serves the specific function. Data curation into a more comprehensible form, emphasizing patterns and outliers, and data visualization all assist in telling tales and this has been implemented using Tableau in a single dashboard. A dashboard is frequently constructed when there is a lot of material to be represented through data visualization. Some Features among the dataset are converged to visualize in order to showcase meaningful insights. Points which are visualized are such as Life Expectancy Overview which represents a world map where each country is shaded according to the range of life expectancy, Different age group deaths vs Diseases affecting life Expectancy, and General government expenditure on health. These all points are represented in a single dashboard. For a record, the dataset limits to only from the year 2000 to the year 2015 and therefore, the visualization is depicted within the mentioned range only.

E. Prediction

In the final phase of the analysis, machine learning algorithms were tested and compared to make predictions based on the analyzed data. After trying several algorithms, it was found that the random forest algorithm provided a high level of accuracy (95%) compared to other models. This algorithm was chosen as the final model for making predictions on life expectancy. The predicted result is impacted by all 22 columns of the dataset which have been described earlier. Predictive modeling helps in forecasting the life expectancy of any given country for future years and can be useful for decision-making and resource allocation.

V. RESULTS

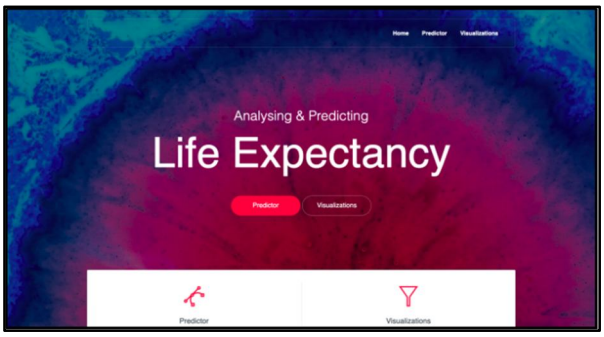


Fig. 2 Home Page

Figure 6.1 depicts the homepage of the project, which offers both a predictor page and a visualization page.

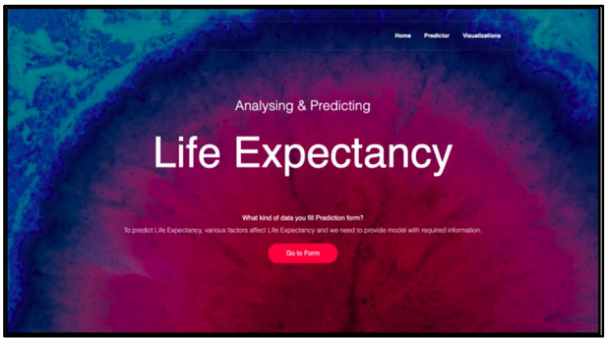


Fig. 3 Prediction Page

This figure is about the prediction page, predictor page allows you to navigate to the prediction form. input specific variables and receive a prediction or estimate based on the algorithm.

Fig. 4 Prediction Form

This figure is about the prediction form where you have to fill in the required 22 inputs and receive a prediction or estimate based on the help of an algorithm. For example, we have to find the life expectancy of India for the year 2021 then we will fill all the required fields with accurate data such as Country will be India, the year would be 2021, country's status that is India's status was developing as India was and still a developing country. The adult mortality rate in 2021 was 9.42 per 1000 population and the infant death rate of India was 26 per 1000 population, etc in this way we will fill all the required remaining fields

Fig. 5 Result Page

Based on the data you have provided, the Random Forest regression algorithm has calculated your predicted life expectancy i.e 66.039 years in age. This is used to estimate that you will live for approximately years.

A. Points To Visualise

1) Life Expectancy Overview:

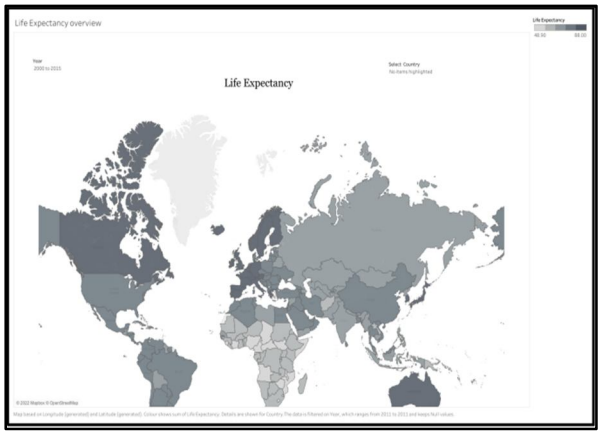


Fig. 6 Life Expectancy Overview

Life Expectancy Overview represents a world map where each country is shaded according to the range of life expectancy and can be varied by changing years in the range of 2000 to 2015 as the reliable data limits with this range only.

2) Factors like Different age groups and diseases affecting life Expectancy according to their country's status:

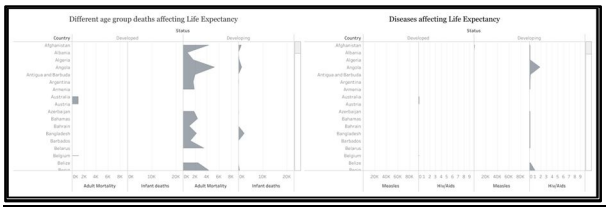


Fig. 7 Different Age group deaths vs Diseases affecting Life Expectancy

This figure portrays that different age group deaths occur more in developing countries whereas in contrast developed countries hold rare age group death cases and diseases affecting life expectancy.

3) General government expenditure on health

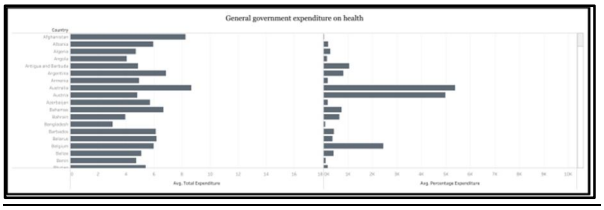


Fig. 8 General government expenditure on health

This figure shows insights into general government expenditure on health.

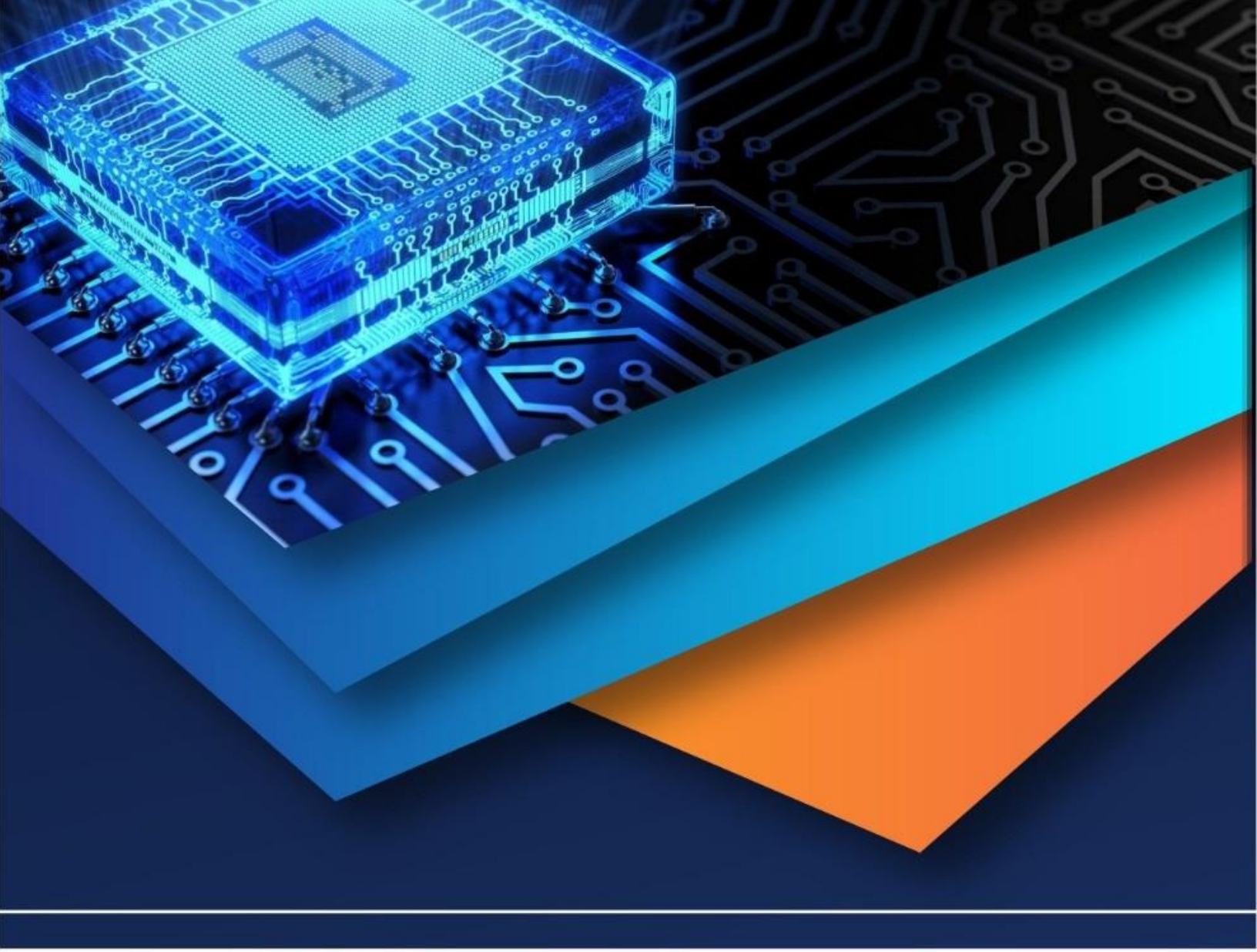
VI. CONCLUSION

The analysis of life expectancy was performed by collecting reliable data from Kaggle, preprocessing it to remove errors and inconsistencies, and then visualizing the trends and patterns. Predictive modeling was then performed using the random forest algorithm, which provided a high level of accuracy (95%) in forecasting future trends in life expectancy. The results were visually represented and offered valuable insights into the trends and patterns in life expectancy. The overall process of data collection, preprocessing, visualization, and prediction helps understand the current state of life expectancy and provides a platform for decision-making and resource allocation. This analysis serves as a valuable contribution to the field of data science and can be used for further research and development.

REFERENCES

[1] Abhinaya. V, Dharani. B. C, Vandana. A, Dr. Velvadivu. P, Dr. Sathya. C., “Statistical Analysis On Factors Influencing Life Expectancy”, Coimbatore, India, e-ISSN: 2395-0056, July 2021.
[2] Jessica Y Ho, Arun S Hendi, “Recent trends in life expectancy across high income countries: retrospective observational study”, USA, 10.1136/bmj.k2562, 15 August, 2018.
[3] David Strauss, Lewis Rosenbloom, Jordan Brooks, Robert Shavelle, “Life expectancy in cerebral palsy: an update”, USA, DOI: 10.1111/j.1469-8749.2008.03000.
[4] Michael J DeVivo, Gordana Savic, Hans L Frankel, Bakulesh M Soni, “Comparison of statistical methods for calculating life expectancy after spinal cord injury”, 10.1038/s41393-018-0067-1, February 2018.
[5] Palak Agarwal, Navisha Shetty, Kavita Jhajharia, Gaurav Aggarwal, Neha V Sharma, “Machine Learning For Prognosis of Life Expectancy and Diseases”, ISSN: 2278-3075, August 2019.
[6] Li, Y., Liang, Y., Liu, R., Li, Y., & Li, Y. (2020). Analysis of factors influencing the life expectancy of the elderly based on machine learning. PeerJ, 8, e8365. doi: 10.7717/peerj.8365
[7] Kim S., Lee S. Y., “Effects of air pollution on life expectancy in South Korea”, Environmental Science and Pollution Research, vol. 28, no. 10, pp. 12603-12613, 2021.
[8] Lassale C., “Dietary patterns and life expectancy”, Current Opinion in Clinical Nutrition and Metabolic Care, vol. 23, no. 4, pp. 239-244, 2020.
[9] Li J., Chen J., “Gender, family support, and life expectancy: a cross-national analysis of older adults”, BMC Public Health, vol. 21, no. 1, pp. 1-10, 2021.
[10] Chauhan P., “Life Expectancy and GDP per capita: A Comparative Analysis of India and China”, International Journal of Scientific Research, vol. 10, no. 2, pp. 61-65, 2021.
[11] Doshi R., Patel P., Shah B., “Estimation of Life Expectancy at Birth for Indian States and Union Territories by a Bayesian Hierarchical Model”, Journal of Health Management, vol. 22, no. 2, pp. 201-215, 2020.
[12] Huang Y., Chen S., Zhang X., Wu J., “Effects of education on life expectancy in China: Evidence from the China Family Panel Studies”, PLoS One, vol. 15, no. 3, pp. 1-12, 2020

[13] Jiang Y., Li Y., Li X., Ma Z., Wang R., Li J., "The Impact of Health Insurance on Life Expectancy: Evidence from China's New Rural Cooperative Medical Scheme", International Journal of Environmental Research and Public Health, vol. 17, no. 5, pp. 1-14, 2020.
[14] Alkhamis A. A., "Assessment of life expectancy and mortality rates using the Lee-Carter model", Applied Economics, vol. 53, no. 3, pp. 358-373, 2021.
[15] Li Q., Li X., Li J., Li M., Cui Y., "Analysis of the Effects of Alcohol Consumption on Life Expectancy Based on a Markov Model", Risk Analysis, vol. 40, no. 3, pp. 542-555, 2020.
[16] Patel K. J., Gajera H. P., Parmar P. N., "A Study of Factors Affecting Life Expectancy in India", Journal of Critical Reviews, vol. 7, no. 2, pp. 119-124, 2020.
[17] Cao X., "The Effect of Occupational Health and Safety on Life Expectancy: Evidence from China's Coal Mining Industry", International Journal of Environmental Research and Public Health, vol. 18, no. 1, pp. 1-12, 2021.
[18] Hashim K., "Effects of Environmental Pollution on Life Expectancy in Malaysia", Environmental Science and Pollution Research, vol. 28, no. 9, pp. 11624-11630, 2021.
[19] Liu Y., Zhou H., "Impacts of air pollution on life expectancy in China", International Journal of Environmental Research and Public Health, vol. 17, no. 5, pp. 1-14, 2020.
[20] Saleem N., Ahmad N., Shaikh T. A., Khan M. S., "Analysis of the effect of health expenditure on life expectancy in Pakistan using ARDL", BMC Health Services Research, vol. 21, no. 1, pp. 1-10, 2021.
[21] Wu Y., Hou G., Liu W., "The effects of medical insurance on life expectancy: a study of China's basic medical insurance", International Journal of Health Economics and Management, vol. 21, no. 1, pp. 47-62, 2021.
[22] Zhao L., Zhu L., Lu Z., Qin L., Huang Y., Zhang J., "Analysis of Factors Influencing Life Expectancy Based on Big Data", Journal of Healthcare Engineering, vol. 2020, pp. 1-11, 2020.
[23] Lin H., Li H., Li X., "Socioeconomic Development and Life Expectancy in Developing Countries", Social Indicators Research, vol. 150, no. 1, pp. 219-239, 2020.
[24] Liu M., Zhou C., Li J., Li Y., Li X., Wang Y., "Determinants of life expectancy at birth in China: A Bayesian model averaging approach", Journal of Health Economics, vol. 79, pp. 1-11, 2021.
[25] Nguyen T. H., Vu T. K., Nguyen H. L., Vu H. T., "The impact of health expenditures on life expectancy in ASEAN-5 countries: A dynamic panel data approach", Journal of Health Economics and Outcomes Research, vol. 8, no. 2, pp. 153-164, 2020.
[26] Samik-Ibrahim I., Abuhayat A. M., Oyefuga O. H., "Factors affecting life expectancy in Africa: A review of literature", Archives of Medicine and Health Sciences, vol. 9, no. 2, pp. 313-318, 2021.
[27] Sun R., Zhang Q., Liu W., Wang Y., "Life Expectancy, Healthy Life Expectancy, and Healthcare Expenditure at Birth and at Age 60 in China: A Markov Model", Risk Analysis, vol. 41, no. 4, pp. 719-733, 2021.
[28] Trung H. V., Tuan A. V., Hoang L. T., "Life expectancy and health status of older people in Vietnam: Evidence from a longitudinal survey", BMC Public Health, vol. 21, no. 1, pp. 1-13, 2021.
[29] Wang C., Luo L., Huang Y., Hu Y., Wang C., "Regional disparities and determinants of life expectancy at birth in China, 2000-2015: A spatiotemporal analysis", PLoS One, vol. 15, no. 11, pp. 1-15, 2020.
[30] Wang J., Jia S., "How does pollution impact life expectancy in China? Evidence from prefecture-level cities", Environmental Science and Pollution Research, vol. 28, no. 15, pp. 19456-19465, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY