Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, April : 2024

LIFE EXPECTANCY ANALYSIS

Dr. K.Chaitanya ,Associate Professor,

K. BHARGAVI, J.RANI , K. N .PAVANI, D. SRAVYA,

Department of CSE (Data Science), SRK Institute of Technology, Vijayawada, A.P, India.

ABSTRACT

This study aims to predict life expectancy using various data analysis and visualization techniques in Python. The analysis utilizes popular libraries such as Pandas, Matplotlib, NumPy, Seaborn, Plotly, SciPy, and Pandas Profiling, along with scientific Python tools. The dataset is preprocessed using Pandas to handle missing values and outliers are treated using the Winsorize method. Exploratory data analysis is conducted to understand the Distribution and relationships between variables, aided by visualization tools like Matplotlib, Seaborn, and Plotly. Additionally, Pandas Profiling is employed for comprehensive data profiling. Life expectancy prediction models are built using SciPy and other statistical methods. Finally, the results are visualized using Plotly Express for interactive data visualization, providing insights into factors influencing life expectancy and the predictive accuracy of the models. This comprehensive approach facilitates understanding and predicting life expectancy while demonstrating the capabilities of Python libraries in data analysis and visualization.

INTRODUCTION

Life Expectancy is an analytical as well as a statistical measure of the longevity of the population depending upon distinct factors. Over the years, Life expectancy

observations are being vastly used in medical, healthcare planning, and pension-related services, by concerned government authorities and private bodies. Advancements in forecasting, predictive analysis techniques, and data- science technologies have now made it possible to develop accurate predictive models. In many countries, itis a matter of political debate about how to decide the retirement age and how to manage the financial issues related to the public matter. Life expectancy predictions provide solutions related to these issues in many developed countries. With the advancement in new systematic, accurate, efficient, and result-oriented techniques in the field of Data Science, now predictions of the Life Expectancy of the selected region are becoming more prominent in demand of the government authorities and the private bodies and their policy- making. Fig 1 represents the Data Analysis Architecture.

There have been many vast improvements in the field of data science and analytical techniques, which explains the rise in life expectancy around the world. These significant improvements in the predictive analysis techniques have also led us to more ways so that authors can improve the life expectancy of the distinct population. These improvements were solely dependent upon specific indicators.

The extensive research into the prior life expectancy models has suggested us the inclusion of many more indicators than expected, such as; GDP (Gross Domestic Product), healthcare expenditure, family income, educational expenditure, infant mortality rate, adult mortality rate, healthcare plans, and population of the selected region. Recent studies have also revealed the impact of geographical factors, climate conditions on life expectancy. Implicitly, the educational background of people, health plans, economic stability, and the burden of diseases, BMI, and environmental variables

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 4, April : 2024

also affect the lifestyle of the people.

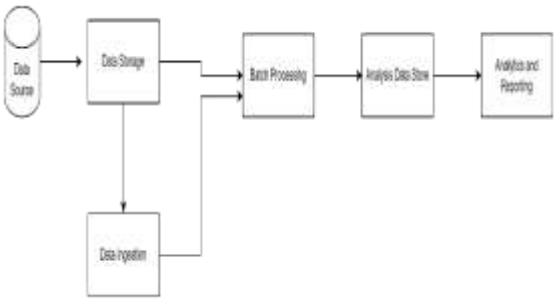

Fig 1: Data Analysis Architecture

LITERATURE SURVEY

Aggarwal D. et al. (2017) Proposed a Life Expectancy using Machine Learning, in their paper, presented a 5-year predictive tool of patients from a recent one or two hospital visits. In recent years, this is one of the most accurate predictive models, providing results up to 5 years. Their work was based upon the ensemble of different Machine Learning models. The Ensemble technique is a Machine Learning method in which various basic models are combined to achieve the final predictive model. This was the reason behind the deployment of the final model with precise implementation because it was an ensemble with 75 individual models. To attain such a level precision, they need to be specific about patients and need to have access to each individual's EMR data. It performed far better than the previous models

N. Kerdprasop et al. (2017) In their paper, suggested that there is an association of environmental as well as economical factors to life expectancy. They used the Chi Square Automatic Interaction Detector (CHAID) method for categorical values and regression on continuous and numerical values. The underlying principle under the CHAID algorithm is a Decision tree. It is a tool used to discover the relationship between the variables. Using this technique, they attempted to relate between economic growth & resources of a country to the life expectancy of its people. They have done the prediction of the life expectancy of people living near the Mekong River. Eventually, their results revealed a strong relationship between GDP growth &

environment to the life expectancy of people.

Beeksma et al. (2019) presented this paper in BMC Medical Informatics. It is one of the recent researches in the concerned subject area. They proposed the idea of using supervised machine learning using recurrent neural networks on deceased patients by using their medical records. They approached the task with supervised machine learning. Then, they trained and tested the data on LSTM (Long Short-Term Memory) recurrent neural networks. The LSTM method is a form of RNN(Recurrent Neural Networks). In the RNN method, the output from the calculation is taken as input within the current phase. LSTM method was formulated to beat the matter of long-term dependencies of RNN. It is mostly used in the time-series data due to the lags between the important unknown events in the time series. Their model was based on non-text and text features of medical records. The first one was the base model (containing non-text features), and another one was the key word model (containing text features in EMR). The Keyword model proposed a better accuracy of 29% than compared to 20% of the base model. But, some limitations of this were the data availability and not generalized in predictions**.**

Rana Sabry Proposed Utilizing Artificial Neural Networks (ANNs) on the Life Expectancy WHO dataset offers a sophisticated approach to understanding the complex relationships between various factors and life expectancy outcomes. ANNs, a type of deep learning algorithm inspired by the structure and function of the human brain, can capture intricate nonlinear relationships and interactions within the data. By training an ANN on the dataset, researchers can uncover hidden patterns and associations that may not be apparent through traditional statistical methods. ANNs have the capacity to consider a multitude of variables

simultaneously, including healthcare access, socioeconomic factors, environmental conditions, and health indicators, allowing for a comprehensive analysis of the determinants of life expectancy. While ANNs require more computational resources and expertise to implement compared to linear regression, they offer the potential for superior predictive accuracy and the ability to uncover nuanced insights that can inform targeted interventions and policy decisions aimed at improving population health and extending life expectancy.

Youssif Shaaban Qzamel Proposed The Life Expectancy WHO dataset, analyzed using the Random Forest regressor algorithm, achieved an impressive accuracy score of 98%. Random Forest is a powerful ensemble learning technique that combines multiple decision trees to make predictions, offering robustness and flexibility for modeling complex relationships in the data. With its ability to handle nonlinearities and interactions among variables, the Random Forest algorithm effectively captures the diverse factors influencing life expectancy, including healthcare access, socioeconomic conditions, and environmental factors. By achieving a high accuracy score of 98%, the Random Forest model demonstrates its capability to provide reliable predictions of life expectancy, offering valuable insights for public health policy-making and interventions aimed at improving population health outcomes**.**

Abhijeeth. Proposed Analyzing the Life Expectancy WHO dataset using regression algorithms provides a versatile approach to understanding the multitude of factors influencing life expectancy across diverse populations. Regression algorithms, ranging from traditional linear regression to more complex techniques like random forest regression and support vector regression, Kneighbors, Regressor, Decision Tree Regressor offer a robust framework for modeling the relationship

between independent variables and life expectancy outcomes. These algorithms allow researchers to identify significant predictors such as healthcare access, socioeconomic status, environmental factors, and health indicators, and quantify their impact on life expectancy. By employing regression algorithms, policymakers and public health practitioners can gain insights into the determinants of life expectancy and develop targeted interventions to improve population health outcomes. While each regression algorithm has its strengths and limitations, collectively, they offer valuable tools for uncovering patterns, making predictions, and informing evidence-based strategies to enhance overall well-being and extend life expectancy

EXISTING SYSTEM

The existing system of life expectancy analysis encompasses a multifaceted approach integrating demographic, epidemiological, and statistical methods. At its core lies the collection of comprehensive data from vital statistics registration systems, censuses, surveys, and health records. This data serves as the foundation for constructing life tables, which provide invaluable insights into mortality rates and survivorship probabilities across different age groups. Concurrently, epidemiological studies explore the intricate interplay of factors influencing life expectancy, ranging from lifestyle choices to socio-economic disparities and healthcare accessibility. Statistical techniques such as survival analysis further refine understanding, modeling the time until specific events like death occur. Trends analysis tracks shifts in life expectancy over time, informing policymakers and healthcare stakeholders about the effectiveness of interventions and emerging health challenges. Ultimately, this holistic approach not only elucidates population health dynamics but also guides

the formulation of targeted strategies to promote longevity and well-being

Disadvantages:

- Reliance on Historical Data
- Disparities in Data Quality
- Complex Determinants
- Limited Predictive Power
- Inability to Account for Emerging Health Threats
- Incomplete Vital Registration Systems
- Challenges in Addressing Social Determinants
- Difficulty in Adapting to Changing Demographics

PROPOSED SYSTEM

Analyzing the Life Expectancy WHO dataset using pure data analysis techniques offers several advantages over using machine learning algorithms. Firstly, pure data analysis techniques provide a straightforward and interpretable approach to understanding the relationships between various factors and life expectancy outcomes. By focusing on statistical methods such as correlation analysis, regression, and hypothesis testing, researchers can gain deeper insights into the determinants of life expectancy without the complexity introduced by machine learning models. Additionally, pure data analysis techniques are often computationally less intensive and require fewer resources compared to machine learning algorithms, making them more accessible and easier to implement for researchers with limited computational expertise. Moreover, pure data analysis techniques allow for rigorous hypothesis testing and model validation, ensuring the reliability and robustness of findings. Overall, leveraging pure data analysis techniques enables researchers to uncover valuable insights into the drivers of life expectancy variations and inform evidence-based interventions to improve

population health outcomes effectively. Fig 2 represents the Activity Diagram.

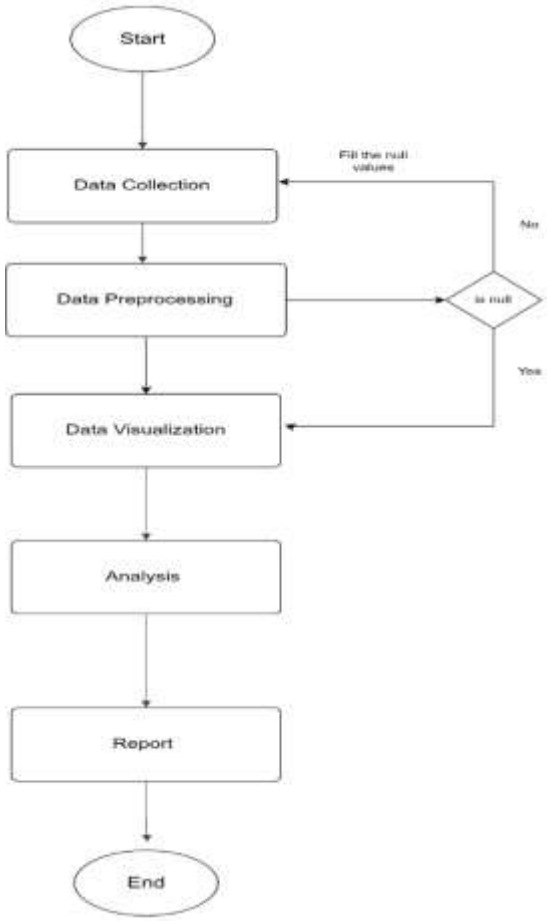

Fig 2: Activity Diagram

Advantages:

- Transparency and Interpretability
- Understanding Trends
- Identification of Correlations
- Descriptive Statistics
- Hypothesis Testing
- Visualization
- Resource Efficiency
- Interdisciplinary Collaboration

.

CONCLUSION

Based on the analysis of the WHO dataset on life expectancy, it can be concluded that life expectancy varies significantly across different regions and countries. Factors such as access to healthcare, economic development, education, and public health initiatives play crucial roles in determining life expectancy. Additionally, trends over time indicate improvements in life expectancy globally, but disparities still exist between developed and developing nations. Further research and targeted interventions are needed to address these disparities and improve overall global life expectancy

FUTURE SCOPE: The future scope of life expectancy research using the WHO dataset holds immense potential for advancing our understanding of global

health trends and informing targeted interventions to improve population health outcomes. With ongoing advancements in data science, machine learning, and epidemiology, researchers can leverage the vast amount of data within the WHO dataset to uncover nuanced insights into the determinants of life expectancy across diverse populations. Additionally, emerging technologies such as artificial intelligence and predictive analytics offer opportunities to develop sophisticated models that can forecast future life expectancy trends, identify at-risk populations, and prioritize resource allocation for preventive healthcare interventions. Furthermore, the integration of multi-disciplinary approaches, including genetics, environmental science, and social determinants of health, can provide a more comprehensive understanding of the complex factors shaping life expectancy. By harnessing the wealth of data available through the WHO dataset and leveraging cutting-edge methodologies, future research endeavors have the potential to drive transformative changes in public health policies and practices, ultimately contributing to extended life expectancy and improved well-being for populations worldwide.

REFERENCES

[1] Aggrawl, D., Mittal, S., Bali V., "Prediction Model for Classifying Students Based on Performance using Machine Learning Techniques," International Journal of Recent Technology and Engineering, vol.8 p.2S7, 496-503, 20

[2] Kerdprasop, N. and Foreman, K. J., "Association of economic and environmental factors to life expectancy of people in the Mekong basin," IEEE Conference on Industrial Electronics and Applications (ICIEA), pp. 1984-1989, 2017.

[3] Beekshma., "A neural-network analyzer for mortality forecast," ASTIN Bulletin:

The Journal of the IAA, vol. 48, no. 2, pp. 481- 508,

[4] Rana Sabry Proposed Utilizing Artificial Neural Networks (ANNs) on the Life Expectancy WHO dataset offers a sophisticated approach to understanding the complex relationships between various factors and life expectancy outcomes. “https://www.kaggle.com/code/ranasabrii/life-expectancy-regression-with-ann”

[5] Youssif Shaaban Qzamel Proposed the Life Expectancy WHO dataset. The Life Expectancy WHO dataset, analyzed using the Random Forest regressor algorithm, achieved an impressive accuracy score of 98%. “https://www.kaggle.com/code/youssifshaabanqzamel/life-expectancy-98-score”

[6] Abhijeeth Proposed Analyzing the Life Expectancy WHO dataset using regression algorithms provides a versatile approach to understanding the multitude of factors influencing life expectancy across diverse populations. “https://www.kaggle.com/code/vstacknocopyright/modeling-life-expectancy-using-regression-algo-s”

[7] Dataset from "https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who"