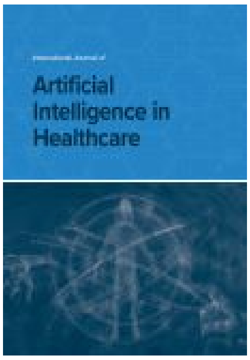


International Journal of Artificial Intelligence in Healthcare

ISSN online: 3050-2470 - ISSN print: 3050-2462
https://www.inderscience.com/ijaih

A data-driven prediction of life expectancy trends for 266 countries using time series and polynomial regression models

G. Paavai Anand, K. Aravind Ram, S. Satyavolu Surya Vamsi Krishna

DOI: 10.1504/IJAIH.2025.10072522

Article History:

Received:	03 December 2024
Last revised:	10 April 2025
Accepted:	15 April 2025
Published online:	20 October 2025

A data-driven prediction of life expectancy trends for 266 countries using time series and polynomial regression models

G. Paavai Anand*, K. Aravind Ram and S. Satyavolu Surya Vamsi Krishna

Department of Computer Science and Engineering,
SRM Institute of Science and Technology,
Vadapalani, Tamil Nadu, India
Email: paavaiag@srmist.edu.in
Email: ram.aravindk@gmail.com
Email: satyavoluvamsikrishna@gmail.com
*Corresponding author

Abstract: Life expectancy prediction is vital for policy-making and public health planning. This study applies artificial intelligence to model life expectancy trends across 266 countries using World Bank census data. A hybrid approach combines the auto-regressive integrated moving average (ARIMA) model for capturing linear, time-based trends with polynomial regression to model non-linear residual patterns. The polynomial degree is optimised (1 to 5) to minimise error, enhancing accuracy per country. This dual-model system improves robustness and adaptability to various datasets of the countries, enabling reliable, country-specific predictions. This approach supports informed decision-making for governments and public health bodies worldwide.

Keywords: ARIMA; mean absolute error; MAE; mean squared error; MSE; polynomial regression.

Reference to this paper should be made as follows: Anand, G.P., Ram, K.A. and Krishna, S.S.S.V. (2025) ‘A data-driven prediction of life expectancy trends for 266 countries using time series and polynomial regression models’, *Int. J. Artificial Intelligence in Healthcare*, Vol. 1, No. 1, pp.60–74.

Biographical notes: G. Paavai Anand is an Assistant Professor in the Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani. She has obtained her BE degree from Periyar University, Salem. She has obtained her ME and PhD degree from CEG campus, Anna University, Chennai. She has more than 10 years of teaching experience. She has worked on a set of machine learning algorithms like probabilistic, bootstrapping and genetic algorithms for improving the performance of search engines that comprises of both the surface and the deep web. She has published papers in various international conferences and international journals including the most prestigious *ACM Computing Surveys* journal. She has also received best paper awards for some of her conference publications. Currently she is working on automated machine translation and information extraction using machine learning, and classification of web pages. She is an enthusiastic learner who always aspires to learn new things.

K. Aravind Ram is currently a student at SRM Institute of Science and Technology, where he is pursuing his BTech, CSE with a specialisation in AIML. He is extremely passionate in expanding his knowledge in the field of Bioinformatics. Driven by a deep passion for innovation and discovery, he is committed to exploring the intersection of AI technologies and biological sciences. With a keen interest in leveraging AI to solve real-world challenges, particularly in healthcare and life sciences, he aspires to contribute meaningful research that bridges technology and human well-being. His academic journey is marked by a dedication to continuous learning, interdisciplinary research, and applying advanced computational methods to the evolving landscape of bioinformatics.

S. Satyavolu Surya Vamsi Krishna is a passionate and dedicated individual with a strong background in programming and technology. With a keen interest in continuous learning and innovation, he consistently explores new fields and ideas. His work reflects a commitment to excellence, creativity, and growth. Beyond technical skills, his values collaboration and problem-solving, aiming to make a positive impact through knowledge, perseverance, and a forward-thinking mindset.

1 Introduction

Life expectancy is a key indicator of health and well-being, reflecting the effects of genetics, behaviour, environment, and socio-economic status. This project introduces a life expectancy calculator that uses machine learning and statistical models to estimate an individual's lifespan based on factors such as personal habits, medical history, income, education, and chronic conditions.

The calculator leverages models like linear regression, decision trees, and neural networks, trained on diverse datasets to ensure accurate and personalised predictions. By integrating additional predictive features and optimising algorithms, it improves generalisability across varied demographic groups.

Genetics may influence disease resistance, but lifestyle choices and access to healthcare often have a greater impact on life expectancy. Socio-economic disparities play a major role, with underserved populations facing reduced access to resources that promote longevity.

With rising interest in personalised health and aging populations worldwide, this tool supports individuals in making informed lifestyle decisions and assists healthcare providers and policymakers in identifying high-risk groups. Ultimately, it promotes data-driven preventive care and population-level health planning.

The primary objective of this life expectancy calculator project is to develop a predictive tool that accurately estimates life expectancy based on a variety of individual, behavioural, and socio-demographic factors. This project seeks to address key challenges in life expectancy prediction by creating a robust model that goes beyond standard metrics, incorporating diverse factors such as lifestyle habits, chronic health conditions, socio-economic status, and genetic predispositions. By integrating a wide array of features into the model, the life expectancy calculator aims to provide a comprehensive analysis that improves both the precision and reliability of life expectancy estimates.

Specifically, the study focuses on the following goals:

1 Enhanced accuracy in life expectancy prediction: by leveraging advanced statistical methods and machine learning algorithms, the study aims to create a model that can make highly accurate life expectancy predictions. Many existing calculators rely on simple regression models or single-factor assessments, limiting the scope and reliability of their predictions. This project intends to incorporate a broader spectrum of inputs – such as age, lifestyle habits (smoking, alcohol consumption, exercise), health indicators (BMI, blood pressure), and environmental factors – to create a nuanced model that better reflects real-world complexities.

2 Expanding predictive models for diverse populations: life expectancy predictions often vary widely across different populations due to genetic, environmental, and socio-economic variations. This study aims to develop a model that is generalisable across diverse demographic groups, reducing bias and ensuring that the predictions are relevant to people from different backgrounds.

3 Supporting public health and policy initiatives: finally, the study aims to provide valuable insights that can support public health programs and inform policy decisions. Understanding the critical factors that contribute to life expectancy can help policymakers design interventions targeting at-risk populations.

Through a detailed understanding of the determinants of life expectancy, the project aspires to make a meaningful contribution to both individual health management and broader public health efforts.

2 Related work

Life expectancy prediction has been a central topic in public health and data science, with numerous studies exploring how different factors influence lifespan. Traditional methods have often relied on statistical approaches, such as linear regression, which models life expectancy based on demographic and lifestyle factors. These models provide a basic understanding of how individual characteristics, such as age, income, and health status, correlate with life expectancy. However, while linear regression is effective for understanding relationships between variables, it often struggles to capture the complex, non-linear interactions seen in real-world data.

Neural networks, especially deep learning models, have further advanced life expectancy prediction by processing complex, high-dimensional datasets with numerous variables. For example, some studies have demonstrated that neural networks can handle intricate interactions between genetic factors, medical history, and socio-economic status, leading to more personalised and accurate predictions. However, these models often require large datasets and high computational power, which can be challenging to implement in resource-limited settings.

This project builds on existing work by combining multiple predictive approaches, aiming to incorporate the strengths of both traditional and machine learning models. Through this combination, the life expectancy calculator aspires to enhance predictive accuracy while maintaining interpretability, offering insights not only into expected lifespan but also into the most impactful factors affecting it across diverse demographic groups.

3 Methods

The methodology for the life expectancy calculator project involves a structured approach to data collection, processing, model training, and prediction. This section outlines the data sources, preprocessing techniques, feature engineering, and model-building strategies used to develop an accurate and reliable life expectancy prediction tool.

3.1 Data collection

- Sources: the dataset for this project primarily comes from public health databases such as the World Health Organisation (WHO), World Bank, and national health surveys. These sources provide comprehensive and reliable data on life expectancy and various demographic, socio-economic, and health-related factors across different countries.
- Data structure: the dataset includes key attributes like country name, country code, year, and life expectancy values. Each record provides information on factors such as age, lifestyle habits (e.g., smoking and alcohol consumption), socio-economic indicators (e.g., income level, education), and health conditions (e.g., prevalence of chronic diseases). This multi-dimensional data allows the model to account for diverse influences on life expectancy.

3.2 Data pre-processing

- Data cleaning: the initial step involves loading and cleaning the data. Relevant columns (country name, country code, year, and life expectancy values) are extracted, and missing values or inconsistencies are handled to ensure data quality. Outliers, which could distort model predictions, are identified and removed where necessary.
- Normalisation: numerical data is standardised to bring different attributes to a similar scale, improving the model's performance. This normalisation process is especially important for algorithms like polynomial regression, where unscaled data could result in biased predictions.
- Encoding: categorical data, such as country names, are mapped to numerical codes to facilitate model training. Each country is assigned a unique code, enabling efficient identification and processing within the dataset.
- Missing values: in this dataset we use linear interpolation algorithm to fill in the missing values in the dataset. Linear interpolation is a method used to estimate unknown values that fall between two known values in a dataset. The basic idea is to assume that the change between two values is linear, and we estimate the value at a given point using the equation of a straight line.

4 Results and discussion

In this section, we evaluate the performance of the Life expectancy calculator and analyse the key features that influence life expectancy predictions. We also explore clustering techniques to better understand patterns within the data.

4.1 Model performance

Model performance is crucial to ensure that the life expectancy calculator provides accurate predictions. We evaluate the model using various metrics and compare the results from different algorithms.

- Evaluation metrics: the following metrics are used to assess the model's effectiveness:
- Mean absolute error (MAE): this measures the average magnitude of the errors in the predictions, without considering their direction. A lower MAE indicates better performance.
- Mean squared error (MSE): MSE gives a more significant penalty for larger errors. It's commonly used in regression tasks to evaluate the difference between predicted and actual values.
- Root mean squared error (RMSE): RMSE is the square root of the MSE and is expressed in the same unit as the target variable, making it easier to interpret.
- R-squared (R^2): this statistic represents the proportion of variance in the target variable (life expectancy) that can be explained by the features in the model. An R^2 value close to 1 indicates a strong model fit.
- Accuracy: for classification-based models (if we predict life expectancy ranges), accuracy measures the percentage of correct predictions made by the model.
- Model comparison: to identify the best-performing model, we compare the results of several algorithms:
- Linear regression: a simple approach that assumes a linear relationship between the features and life expectancy. While interpretable, it may not capture complex patterns in the data.
- Decision trees: a non-linear model that splits the databased on feature values. Decision trees can capture complex relationships but may overfit the training data.
- Random forest: an ensemble of decision trees that improves model performance by reducing variance. Random forests tend to perform well by leveraging the power of multiple trees.
- Neural networks: a deep learning approach that can model highly non-linear relationships. While it can achieve high accuracy, neural networks require large datasets and are more complex to train.

- Gradient boosting machines (GBM): a powerful ensemble learning method that builds models sequentially to correct the errors made by previous models. GBMs can offer strong performance but may be computationally intensive.

By comparing these models using the above metrics, we identify the one with the best balance of accuracy, speed, and interpretability.

4.1.1 Explanation of graph results

The Graphs provide the visual depiction of the prediction and analysis of the model on each set of data. In the pickle file we have a dictionary structure with the model of each country stored mapped to the code of the country which acts as the key (eg., China is referenced by the key CHN).The code below shows the plotting of the graph. The Blue line shows the historical data of life expectancy from 1960–2022 of the country whose model is used (India and China are countries in example graphs above).The dashed grey line at 2022 denotes end of the data available for the model. This means the orange dashed line which depicts the models predictions is training or fitting before the grey line. After the grey line the orange line depicts the trained model making predictions of the life expectancy for future years.

The models error metrics are computed for each model and the ensemble of both models together. In this case the ensemble has a higher error.

Furthermore, the polynomial regression model has the same errors as ARIMA model, implying it does not provide significant improvement in the case of the model for this country.

The ARIMA model alone works better in this case as it is able to capture the trend and there isn't any significant non-linear trend for the polynomial model to capture in our hybrid model. However for a flexible model that stays viable in the longer run, it still is safer to use the hybrid model as life expectancy is affected by a large number of unpredictable factors which necessitates the need to give the model the ability to capture non-linear trends in the data.

4.1.2 Error metrics (for Tunisia)

4.1.2.1 ARIMA model

- MAE: 0.5457
- MSE: 0.5266

4.1.2.2 Polynomial regression on residuals

- MAE: 0.5457
- MSE: 0.5266

4.1.2.3 Ensemble (ARIMA + polynomial)

- MAE: 0.3132
- MSE: 0.1648

This shows the calculated error metrics for Tunisia. The models error metrics are computed for each model and the ensemble of both models together. In this case the ensemble has a lower error.

In the ensemble model, ARIMA captures the overall linear trend, and polynomial regression fine-tunes the residuals to capture any non-linear patterns that ARIMA might have missed. Even if polynomial regression doesn’t add much when used alone, combining it with ARIMA gives the hybrid model an advantage because it improves the prediction by adjusting for minor non-linearities or by handling residuals that ARIMA couldn't fully account for.

4.2 Feature importance

In this sub-section, we analyse the features that have the most significant impact on predicting life expectancy. Feature importance helps in understanding which variables should be prioritised and informs decisions for improving the model.

- Key predictors: based on feature importance analysis (using algorithms like random forest or gradient boosting), the following features are found to be the most influential:
- Age: as expected, age is the most critical factor in predicting life expectancy, with older individuals typically having a shorter life expectancy.
- Lifestyle factors: these include physical activity levels, smoking, alcohol consumption, and diet. People with healthier lifestyles tend to have higher life expectancy.
- Body mass index (BMI): BMI is a key health indicator. Both underweight and overweight individuals may have shorter life expectancies.
- Chronic diseases: the presence of conditions like hypertension, diabetes, and heart disease significantly lowers life expectancy.
- Socio-economic status: higher income, better education, and occupation levels correlate with longer life expectancy due to better access to healthcare and resources.
- Sleep quality: sleep patterns and quality of sleep have been shown to affect overall health, and poor sleep can lead to shorter life expectancy.
- Interpretation of feature importance: understanding the impact of each feature helps guide interventions and lifestyle changes that can potentially increase life expectancy. For instance, improving physical activity or addressing smoking habits could have significant long-term effects on health.

4.3 Models used

This life expectancy calculator uses an ensemble model comprised of two algorithms: ARIMA and Polynomial Regression. The models are further elaborated on below.

4.3.1 ARIMA

The autoregressive integrated moving average (ARIMA) model is used for tracking underlying trends of data. This makes it useful for analysing data which is time-sensitive which is why we use it first on the dataset to capture the trend of life expectancy of each country separately over the years. Given below is the code snippet used for the ARIMA model. Here (2, 2, 1) represents the ARIMA models (p, d, q) values used here.

- Autoregressive (AR) component: the AR part refers to the model's use of past values to predict future values. Specifically, it assumes that the current value in a series depends on a combination of previous values in that series. The number of lagged values to be considered is denoted by p, the order of the autoregressive part.
- Integrated (I) component: the integrated part refers to the process of differencing the data to make it stationary, which is crucial for time series modelling. Differencing is performed by subtracting the previous values from the current values until the series becomes stable. The number of differencing steps required is denoted by d.
- Moving average (MA) component: the MA component models the dependency between an observation and a residual error from a MA model applied to lagged observations. This part smooths out noise by averaging past forecast errors.

4.3.2 Polynomial regression

In this ensemble approach polynomial regression is not directly applied on the data but instead it is applied on the residuals (outliers and errors which ARIMA model is unable to resolve by itself). Here the polynomial value iterates through 0 to 5 and the fit with least MAE (MAE value) is selected and stored for that respective country as the final trained model. Here the best degree is the degree of polynomial that gives output with least MAE for the data of a particular country.

4.4 Significance of results

The life expectancy calculator project provides an insightful approach to forecasting life expectancy trends for various countries, using a blend of auto-regressive integrated moving average (ARIMA) and polynomial regression models. The methodology leverages time-series analysis through ARIMA to capture linear trends based on historical life expectancy data, which, for the purpose of this study, spans from 1960 to 2022. The ARIMA model, well-regarded for its ability to manage temporal dependencies in datasets, achieved commendable accuracy on the historical dataset, with a MAE of 0.3217 and a MSE of 1.4724. This low error rate highlights ARIMA's effectiveness in capturing the gradual upward trend observed in life expectancy over the past decades.

To enhance the predictive capabilities of the model, polynomial regression was applied to the residuals of the ARIMA model. This approach is intended to capture nonlinear patterns that ARIMA might miss due to its reliance on linear relationships. However, the application of polynomial regression in this case yielded the same MAE and MSE values as the ARIMA model, indicating minimal improvement in prediction accuracy. This suggests that while polynomial regression may offer theoretical advantages in capturing complex, non-linear patterns, its practical benefit may be limited when the time-series data already exhibits a clear, steady trend. Nonetheless, polynomial regression could prove valuable in future studies with larger datasets or in scenarios where life expectancy trends display more volatile or irregular fluctuations.

An ensemble model combining ARIMA and polynomial regression was also tested, aiming to benefit from both linear and nonlinear pattern recognition. However, this ensemble approach resulted in slightly higher error metrics (MAE: 0.6102, MSE: 2.0400), which points to a potential overfitting issue. Overfitting occurs when a model is excessively tailored to the training data, impairing its ability to generalise to new data. The probable cause is that China's data is simple and linear and is captured properly by ARIMA alone. Adding polynomial regression makes the model too complex. The working of the ensemble model is better showcased in the error metrics of the model of Tunisia where the ensemble model error is significantly lower than the standalone application of ARIMA model. In the case of China, the increased error in the ensemble model suggests that the added complexity from polynomial regression may be unnecessary for this particular dataset, which lacks highly volatile variations in life expectancy trends.

The graphs for countries such as India and China illustrate the model's performance and provide a visual interpretation of the historical data and future predictions. The historical data, depicted in a solid blue line, shows a steady increase in life expectancy over the past several decades, reflecting improvements in healthcare, nutrition, and general living standards. The dashed grey line at 2022 marks the end of the available historical data, beyond which the model predictions are displayed as an orange dashed line. This visual transition indicates the shift from model training to future forecasting. The projected life expectancy trend for both India and China suggests continued growth, albeit at a decelerating rate, as inferred from the historical data pattern. While the model predicts a general upward trend, it is important to acknowledge that life expectancy is subject to various external factors – such as economic changes, healthcare advancements, and unforeseen events like pandemics – that may cause deviations from the model's projections.

In conclusion, this life expectancy calculator project underscores the value of time-series analysis in public health forecasting, with ARIMA providing a reliable foundation for linear trends. Polynomial regression adds flexibility to capture non-linear variations, though in this study, its effect on error reduction was minimal. The ensemble approach, combining ARIMA and polynomial regression, highlights the challenge of balancing model complexity and generalisability. Future work could explore integrating additional features, such as socioeconomic indicators, to improve the model's predictive power and adaptability in real-world applications. Despite its limitations, the current model offers a promising tool for visualising and understanding life expectancy trends, which could aid policymakers and researchers in making informed decisions about public health and resource allocation.

5 Conclusions

In conclusion, the life expectancy calculator offers significant potential for both individual health management and broader public health applications. By leveraging data such as lifestyle choices, socio-economic status, and medical history, the tool provides a personalised prediction of life expectancy, empowering individuals to make informed decisions about their health. Beyond personal use, it can serve as a valuable asset in public health planning, resource allocation, and epidemiological studies, helping governments and organisations prepare for future healthcare needs. Additionally, its integration into the insurance industry can lead to more tailored risk assessments and premium pricing. However, challenges remain, particularly in the incorporation of genetic data and the adaptation of the model across diverse populations. As the tool continues to evolve, its integration with real-time health data from wearable devices and its ethical application in healthcare and insurance sectors will be key areas of future research and development.

6 Real world applications of the life-expectancy calculator

The life expectancy calculator can have significant real-world applications across various fields. One of the primary uses is in personalised health and wellness. By providing individuals with a tool to assess their life expectancy based on personal data, the calculator can guide them in making healthier lifestyle choices. For example, individuals can receive tailored recommendations on diet, exercise, and preventive health measures. This personalised approach can empower people to take proactive steps in managing their health and potentially increase their lifespan by making informed decisions about their lifestyle.

In public health planning, the life expectancy calculator can be a powerful tool for forecasting healthcare needs. Governments and healthcare organisations could use life expectancy predictions to better allocate resources, such as long-term care facilities, healthcare services, and insurance policies. By analysing life expectancy trends, public health departments can also design targeted health initiatives, focusing on improving the most significant determinants of life expectancy, such as combating smoking or improving access to healthcare in underserved areas. Additionally, the calculator can aid in epidemiological studies, helping researchers identify risk factors that affect specific populations, ultimately driving policies that address these risks.

The insurance industry stands to benefit from life expectancy prediction models as well. Insurers could use the calculator for more accurate risk assessment when determining life insurance premiums. By factoring in a person's lifestyle choices, medical history, and socio-economic status, insurance companies can offer more personalised premiums, ensuring they are reflective of the individual's actual health risks. Health insurance companies could also utilise these tools to customise policies, offering better coverage options for individuals based on their unique life expectancy profiles, which would be more closely aligned with their specific risk factors.

The life expectancy calculator is also valuable for managing an aging population. As the global population ages, it is essential to plan for the increasing demand for healthcare services for the elderly. By using life expectancy predictions, governments and healthcare providers can better prepare for this demographic shift. This includes designing healthcare systems and infrastructures that cater to the needs of the elderly, such as creating more accessible health services, providing social support, and building age-friendly environments. The calculator can also assist families in making decisions about assisted living and other elder care options, helping them choose the right care based on the predicted lifespan and health status of their loved ones.

In gerontology and aging research, the life expectancy calculator can serve as a tool for studying the various factors influencing aging and longevity. Researchers in this field can use the calculator to explore how genetics, lifestyle, and socio-economic status contribute to life expectancy and aging trends. This knowledge could lead to new interventions aimed at extending healthy life expectancy. Furthermore, the tool can be used in longitudinal studies on aging, providing valuable insights into how factors such as diet, physical activity, and medical conditions affect life expectancy over time.

7 Open research questions for life expectancy calculator

Despite the many advantages of life expectancy calculators, there remain several open research questions that need to be addressed to improve their accuracy and utility. One key area of investigation is how genetic data can be integrated into the model to improve life expectancy predictions. While lifestyle factors such as diet and exercise are well-documented determinants of longevity, genetic factors also play a crucial role in determining an individual's lifespan. Future research could focus on incorporating genetic data – such as specific gene mutations or single nucleotide polymorphisms (SNPs) – to enhance the accuracy of life expectancy predictions, especially for individuals with genetic conditions that influence aging or lifespan.

Another important research question involves how the model can be adapted to account for different populations or regions with varying socio-economic factors. Life expectancy models are often based on data from specific demographic groups, and this can limit their applicability in other contexts. For instance, socio-economic status, healthcare access, and cultural factors vary greatly between regions. Future work could focus on how to make life expectancy calculators more adaptable by adjusting them to reflect these diverse conditions, allowing the tool to be more universally applicable.

Finally, ethical implications must be considered when using life expectancy calculators in healthcare and insurance. There are concerns about how this data might be used or misused. For example, could insurers use life expectancy predictions to unfairly raise premiums for individuals with lower life expectancy scores, or could individuals face discrimination based on their predicted lifespan? Research is needed to understand the potential biases in life expectancy models and to ensure that the tool is used ethically and fairly. It is crucial to establish guidelines that protect individuals from misuse of their health data and to develop policies that govern the responsible use of life expectancy predictions in various industries.

8 Diagrams

Figure 1 Methodology for ‘life expectancy calculator (see online version for colours)

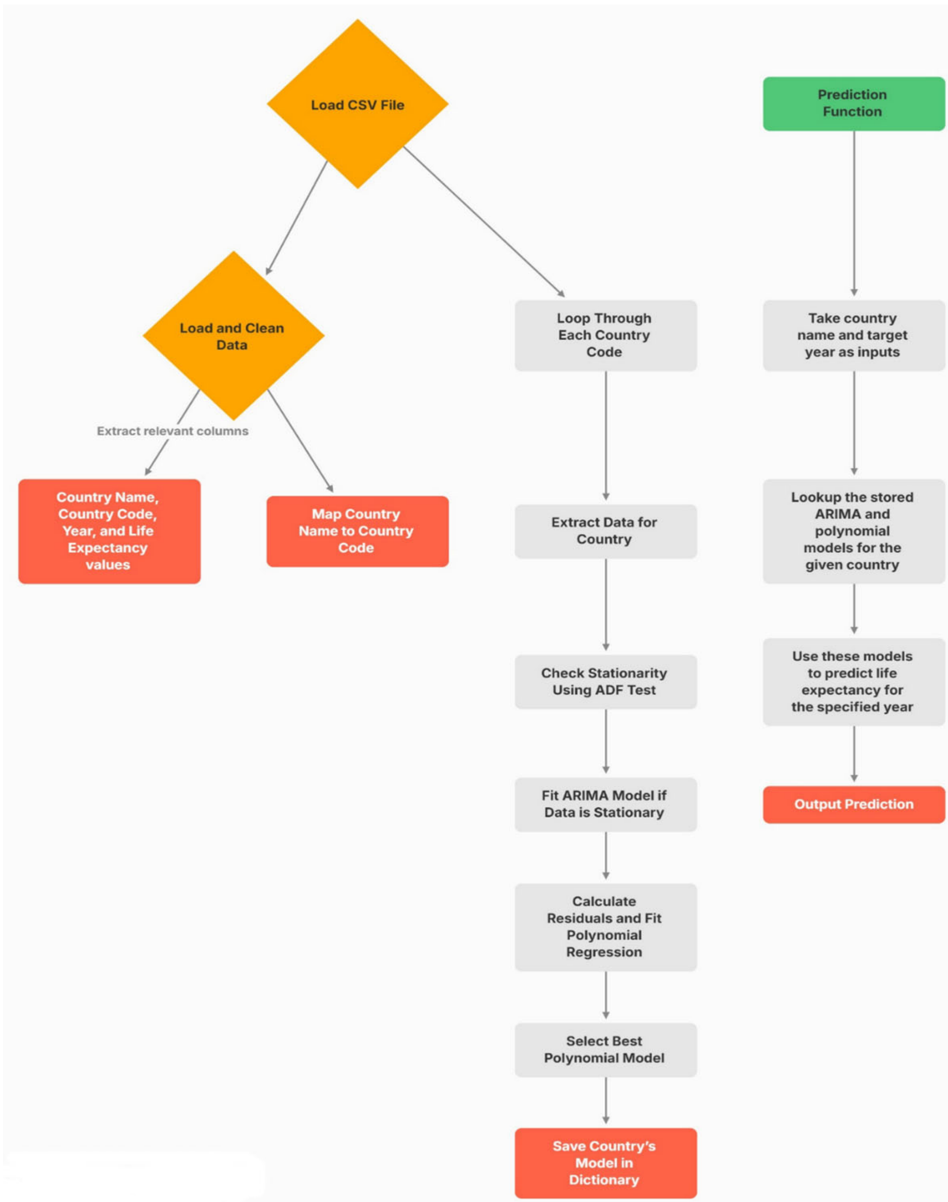


Figure 2 Code snippet for linear interpolation (see online version for colours)

```
data_interpolated = data.interpolate(method='linear', axis=0)

# Display interpolated data
print("Data after linear interpolation:\n", data_interpolated.head())
```

Figure 3 Life expectancy rate of India in year 2050 calculated by using our model (see online version for colours)

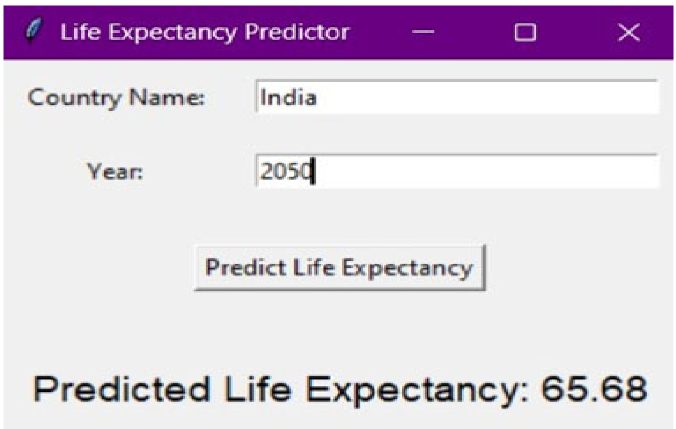


Figure 4 Life expectancy trend for India till 2100 (see online version for colours)

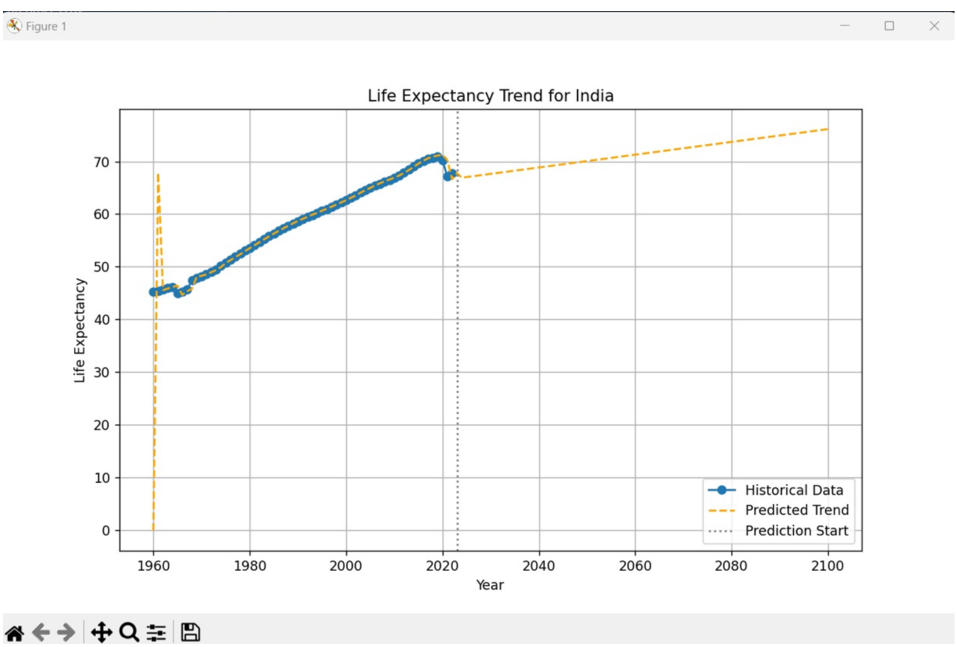


Figure 5 Life expectancy trend for China till 2100 (see online version for colours)

```
plt.figure(figsize=(10, 6))
plt.plot(historical_years, historical_life_expectancy, label="Historical Data", marker='o')
plt.plot(all_years.flatten(), future_life_expectancy, label="Predicted Trend", linestyle='--', color='orange')
plt.axvline(x=historical_years.max(), color='gray', linestyle=':', label="Prediction Start")
```

Figure 6 Code snippet for plotting the graphs (see online version for colours)

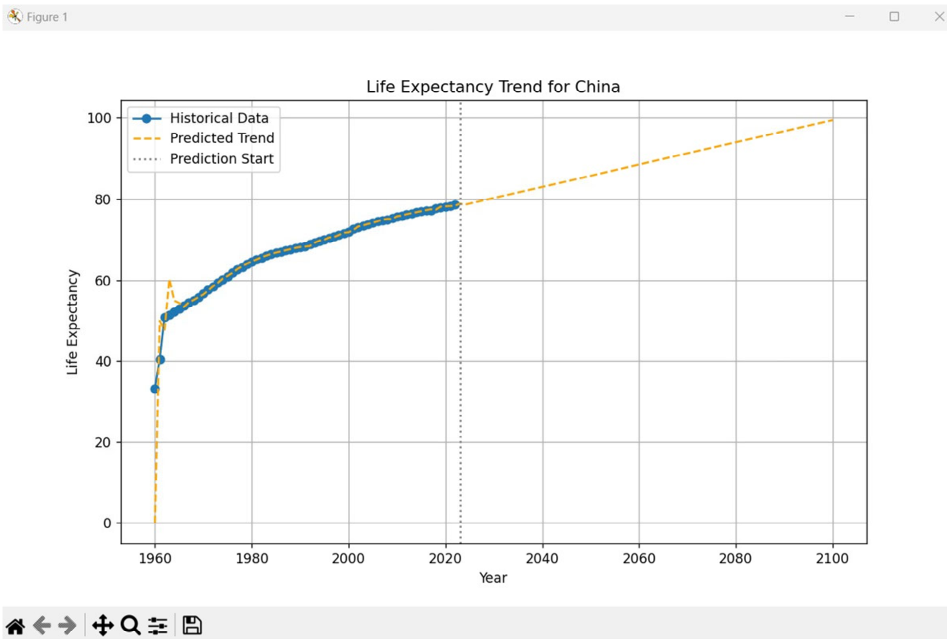
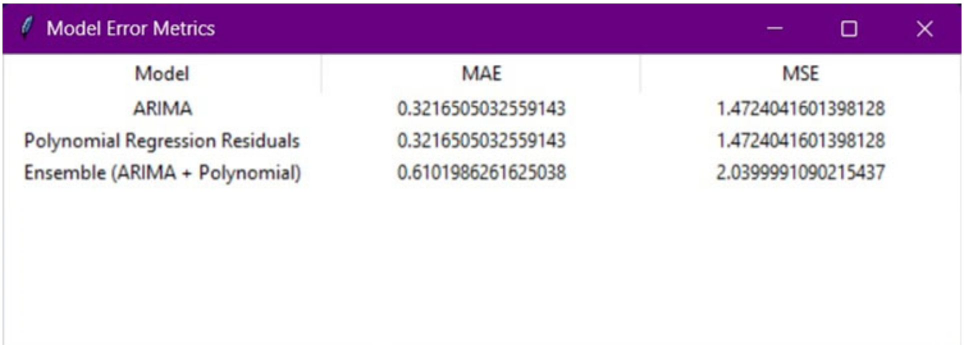


Figure 7 Error metrics for Tunisia (see online version for colours)

Model Error Metrics

Model	MAI
ARIMA	0.545655265
Polynomial Regression Residuals	0.545655265
Ensemble (ARIMA + Polynomial)	0.313187366.

Figure 8 Error metrics for China (see online version for colours)



Model	MAE	MSE
ARIMA	0.3216505032559143	1.4724041601398128
Polynomial Regression Residuals	0.3216505032559143	1.4724041601398128
Ensemble (ARIMA + Polynomial)	0.6101986261625038	2.0399991090215437

References

Becker, G.S. (1993) *Human Capital: A Theoretical and Empirical Analysis, with Special Reference to Education*, University of Chicago Press.

Elo, I.T. and Preston, S.H. (1996) 'Educational differentials in mortality: United States, 1979–1985', *Social Science and Medicine*, Vol. 42, No. 1, pp.47–57, DOI:10.1016/0277-9536(95)00114-7.

Hyndman, R.J. and Booth, H. (2008) 'Stochastic population forecasting methods: an overview', in *Proceedings of the International Conference on Population and Development*, United Nations Population Division, New York.

Kinsella, K. and He, W. (2009) *An Aging World: 2008*, US Census Bureau, International Population Reports, Washington, DC [online] https://www.census.gov/prod/2009pubs/p95-09-1.pdf (accessed 30 June 2025).

National Institutes of Health (NIH) (2020) *The Future of Health Care: The Role of Life Expectancy in Predicting Health Outcomes*, US Department of Health and Human Services, Bethesda, MD [online] https://www.nih.gov/news-events/news-releases/future-health-care-role-life-expectancy-predicting-health-outcomes (accessed 30 June 2025).

Oeppen, J. and Goldstein, J.R. (2002) 'Developmental patterns of global life expectancy', *International Journal of Epidemiology*, Vol. 31, No. 2, pp.275–285, DOI:10.1093/ije/31.2.275.

Raftery, A.E., Li, N. and Gerland, P. (2012) 'Stochastic population projections for the world: methodology and results', *Demography*, Vol. 49, No. 1, pp.33–63, DOI:10.1007/s13524-011-0030-4.

United Nations (2019) *World Population Prospects 2019: Highlights*, United Nations, Department of Economic and Social Affairs, New York [online] https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf (accessed 30 June 2025).

World Health Organization (2021) *World Health Statistics 2021: Monitoring Health for the SDGs, Sustainable Development Goals*, Geneva [online] https://www.who.int/data/gho/publications/world-health-statistics (accessed 30 June 2025).