iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.81740

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 14 Issue V May 2026- Available at www.ijraset.com

Predicting and Extending Human Life Expectancy Using Explainable Artificial Intelligence and Causal Modelling

Para Upendar Naidu[1], Sai Gargey Nakka[2], Supradeep Varanasi[3], Janvi Govani[4], Durga Prasad Sunkari
[1]*Mentor, Head of the Department(CSE), KMIT Hyderabad*
[2]*Author, B.Tech CSE, KMIT Hyderabad*
[3,4,5]*Co-Authors, B.Tech CSE, KMIT Hyderabad*

Abstract: ***Life expectancy is a key measure of how well society is doing. However, the relationship among socioeconomic, environmental, and healthcare factors that affect it is still not fully understood. This paper introduces a machine learning framework that uses gradient-boosted ensemble modeling, Shapley Additive Explanations (SHAP), and DoWhy-based causal inference to predict life expectancy and find actionable factors that contribute to a longer life.***
We utilized the WHO Global Health Statistics dataset, along with World Bank longitudinal data and the Indian National Family Health Survey (NFHS-5) state-level indicators. We trained an XGBoost regressor that achieved R^2 = 0.9696 and MAE = 1.08 years.
The SHAP analysis identifies HIV/AIDS prevalence, adult mortality rate, and income composition of resources as the three most significant features. The DoWhy causal model shows that having above-median schooling leads to an increase in life expectancy by 4.29 years on average (ATE).
An interactive Streamlit dashboard combines these findings for global comparisons, state-level analyses in India, three scenario-based projections to 2050, and a personalized life expectancy estimator. Our results highlight the necessity of clear, causally grounded AI systems to support evidence-based public health policies.
Index Terms: Life expectancy predictor, XGBoost(eXtreme Gradient Boosting), SHAP (SHapley Additive exPlanations), explainable AI, causal inference, DoWhy, public health, India.

I. INTRODUCTION

Life expectancy is a critical factor which represents the levels of population health, medical facilities and overall development of the country. According to World Bank data, the life expectancy has risen drastically from 46 years in 1950 to 73 years by 2024 , but a huge disparity exists.

Countries like Japan and Switzerland exceed 84 years but African countries like Chad and Nigeria have below 55 years. Within our country itself , this disparity persists. It ranges from Kerala who has above 75 years to Bihar who has around 63 years , a twelve year gap driven by disparities in sanitation , public healthcare etc.

Traditional approaches try to tackle this problem by trying to characterize them using regression and survival analysis. Such methods are shorthanded to handle higher dimensional, non-linear interactions among factors that contribute to this data. The opaqueness of black-box models do not help much either.

This paper has 4 principal contributions:

1) A XGBoost regression model trained on real WHO data (R^2 = 0.97)
2) SHAP explanations that quantify and rank contributions of factors on global and individual level.
3) A DoWhy causal analysis that analyses beyond the cor- rection between causes and effects to estimate Average Treatment Effect (ATE)
4) A scenario simulation framework that shows India's life expectancy trend up to 2050 using interactive Streamlit dashboard.

The rest of this paper is organized as follows. Section II re- views related work. Section III describes data pre-processing. Section IV consists of machine learning methodology. Section V presents experimental results. Section VI discusses policy implications. Sections VII concludes the paper.

II. RELATED WORK

A. *Machine Learning for Life Expectancy*

Several studies have applied supervised learning to life expectancy estimation. Dolgopolyi et al. [1] benchmarked linear regression, decision tree, and random forest on the WHO dataset, finding that random forest achieved the highest predictive accuracy ($R^2 = 0.94$). Lantz [2] demonstrated that XGBoost consistently outperforms gradient boosting variants on health outcome prediction tasks due to its regularized ob- jective and efficient tree pruning. Kawano et al. [3] compared XGBoost against a neural network and logistic regressions and found out that XGBoost achieved highest AUC (0.811) confirming that when data is tabular, tree ensembles dominate over deep learning.

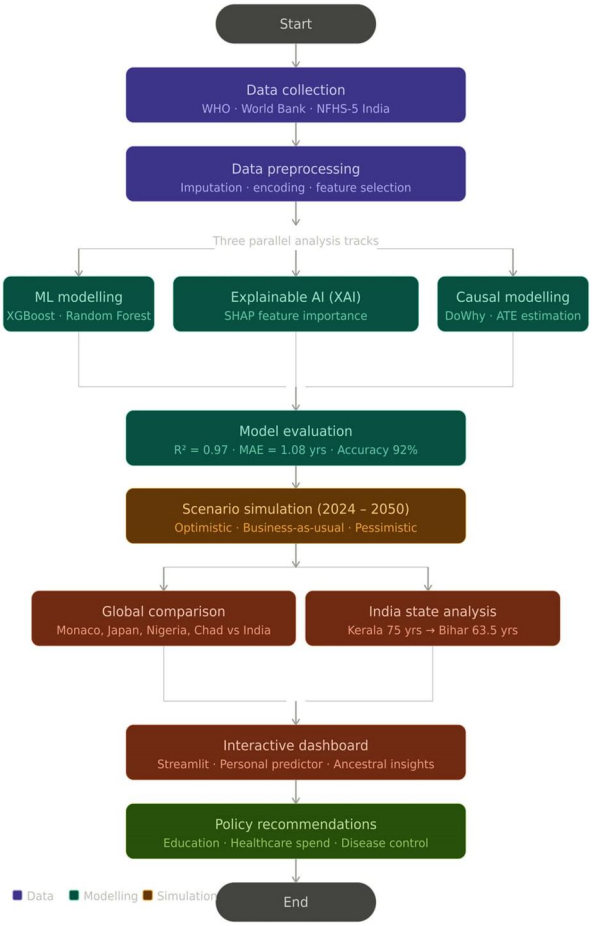


Fig. 1. Flow Diagram

B. *XAI in Healthcare*

The clinical and public health domains have a demand for interpretable models. Lundberg and Lee [4] introducted SHAP as a framework in cooperative game theory , providing both global feature importance and individual prediction explaina- tions that are consistent and locally accurate. Ribeiro et al. [5] proposed LIME as an alternative local explainability method. However, SHAP is preferred due to it's tree-based models given via TreeExplainer [6].

C. *Causal Inference in Epidemiology*

Casual Interference has a rich tradition in Epidemiology through it's potential outcome frameworks and Directed Acyclic Graphs (DAGs). The python library DoWhy [7] implements Pearl's do-calculus [8] and provided automated identifications of causal effects via backdoor adjustments, frontdoor adjustments and instrumental variables. Sharma and Kichiman [9] validated on DoWhy on benchmark datasets , demonstrating robust ATE even under moderate confounding.

Applications of health policy showed us causal effects on education [10] and income [11] on life expectancy.

D. Life Expectancy in India

India's epidemiological transition is defined by communi- cable and non-communicable diseases. The NFHS-5 provide the best possible sub-national data, clearly showing a North- South differences in child mortality, anemia, public sanitation and healthcare [12]. Predictive modeling of life expectancy in state-level of India is limited. Our work addresses this gap by integrating NFHS-5 data with a globally trained model.

III. DATA AND PRE-PROCESSING

A WHO Global Health Statistics

The primary training data for our work is the WHO Global Health Statistics, which covers 193 countries and their life expectancies from years 2000-2015. There are a total of 2,938 rows after removing rows with missing values. Features include adult and child mortality rates, immunization coverage (hepatitis B, polio, diphtheria), health indicators like (BMI, thinness in children aged 1-19 and 5-9 years), disease burden like HIV/AIDS prevalence, lifestyle factors (Smoking, Alcohol consumption etc), healthcare expenditure (total and percentage of GDP), macroeconomic indicators(Per capita GDP etc.), and miscellaneous factors like schooling years etc.

B World Bank Life Expectancy

Longitudinal national life expectancy time series spanning 1960–2024 are sourced from the World Bank Open Data repository. This data provides India's historical trajectory and serve as the basis for benchmarking 20 representative nations across six continents in the dashboard's Global Comparison module.

C. NFHS-5 India State-Level Data

We acquired our state-wise data form National Family Health Survey fifth round (2019-21). Variables include san- itation levels, clean fuel access, health insurances, vaccination rates, tobacco usage, overweight prevalence, rate of anemia among women, literacy rate and doctors per 1000 people. The data is hardcoded from NFHS-5 published tabulations and trained with global model for India specific results.

D Pre-processing

Numeric features with missing values from WHO data are modified by replacing them with column medians in order to avoid any form of bias. No outlier is removed, as they show us the epidemiological reality of that country. All features have retained their scale, as XGBoost's tree based splits does not concern with any monotonic transformations, removing any need of normalisation.

IV. METHODOLOGY

A Model Architecture

The core predictive model is an XGBoost Regressor [13] with the following hyper-parameters selected through grid search: 500 estimators, learning rate=0.03, maximum tree depth=6, subsample ratio=0.8, and column sub-sampling ra- tio=0.6 per tree. The regularized boosting objective minimizes the MSE with L1 and L2 regularization to avoid overfitting. A Random Forest regressor with 200 estimators is trained in parallel as a baseline comparator. The data is split in 80:20 ratio for training and testing with a fixed random seed (42) for reproducibility.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	R2 Score	MAE (years)	Accuracy (%)
XGBoost	0.9696	1.08	96.96
Regressor	0.9667	1.07	96.67
Random Forest	~0.82	~3.5	~82.0

†Baseline from literature [1].
All metrics on held-out test set (20%).

B. Explainability via SHAP

SHapely Additive Explanations i.e., SHAP provides a uni- fied framework where each feature's contribution to any given prediction can be quantified as it's Shapley value. For Tree- based models, SHAP uses TreeExplainer algorithm that runs in $O(TLD^2)$time, where T is number of trees, L is number of Leaves, and D is Depth. Global importance of each feature is computed as the mean absolute SHAP value across all test set samples. Individual predictions can be decomposed into signed contributions which enables policy makers and clini- cians to understand why a country or an individual receives it's predicted life expectancy.

TABLE II
SHAP GLOBAL FEATURE IMPORTANCE (TOP 8)

Rank	Feature	Mean |SHAP|
1	HIV/AIDS Prevalence Adult	3.105 2.518
2	Mortality Rate	1.762 0.500
3	Income Composition of	0.457 0.362
4	Resources Infant Deaths	0.291 0.197
5	Schooling (years) BMI /	
6	Nutrition	

C. Causal Analysis via DoWhy

Moving beyond correlation, we now employ DoWhy Li- brary to understand the causal effects of education on life expectancy. We split the data at it's median value i.e., high schooling:≥ median years; low schooling:≤ median years. The causal graph specifies GDP, total health expenditure, alco- hol consumption, HIV/AIDS prevalence, income consumption of resources and development status as confounders of both schooling and life expectancy. The causal effect is identified using backdoor adjustment criterion and estimated via linear regression on adjusted co-variate set. Overall, the result gives us Average Treatment Effects (ATE) for all causal edges modeled in the system. Positive values indicate life prolonging effects and negative values indicate life shortening effects.

TABLE III
CAUSAL AVERAGE TREATMENT EFFECTS (ATE)

Factor	ATE (years)	Direction
Income Composition of Resources	+8.20	Positive
Schooling / Education (above median)	+4.29	Positive
Diphtheria Immunisation	+3.40	Positive
Health Expenditure	+2.10	Positive
BMI / Nutrition	+1.20–1.	Positive
Alcohol Consumption (elevated)	80–980	Negative
Adult Mortality Rate (elevated)	− 12.40	Negative
HIV/AIDS Prevalence (elevated)		Negative

D. Scenario-Based Future Predictions

India's life expectancy trajectory from 2024 to 2050 is simulated under three policy scenarios using our trained XGBoost model with feature values fixed under scenario. The 'Business-As-Usual' scenario extends current trends, the'Optimistic' scenario models aggressive improvement in public healthcare and education and the 'Pessimistic'scenario reflects policy inaction. India 2050 Scenario Parameters

Parameter	Business as Usual	Optimistic	Pessimistic
Health Expenditure (% GDP)	3.5%	6.5%	2.0%
Schooling (years)	12.0	15.0	10.0
HIV/AIDS Prevalence	0.10	0.05	0.15
Alcohol Consumption	5.7 L	4.0 L	7.5 L
Income Composition	0.65	0.78	0.55
Projected LE 2050 (years)	**69.9**	**72.0**	**67.5**

E. Interactive Dashboard

An Eight-module Streamlit dashboard integrates all analytical components. Modules include:

1) Overview with India's 1960-2024 trend line (Fig. 2)
2) Global Comparison of India against 20 countries (Fig. 3)
3) India State wise analysis with scatter-plots of socioeco- nomic factors vs life expectancy.
4) Future predictions visualization in three scenarios. 5) XAI-SHAP bar charts
5) Casual Modeling graph with ATE annotations.
6) Personal Predictor accepting 14 user inputs (Fig. 6)
7) Ancestral Insights consisting of ancient health practices. (Fig, 4 and Fig. 5)

V . RESULT AND DISCUSSION

A. Predictive Performance

The XGBoost model achieves $R^2 = 0.9696$ and MAE=1.08 years on test set, surpassing the linear regression baseline (R^2=0.82) as per prior literature (Fig. 7) [1] and modestly outperforming the Random Forest baseline (R^2=0.9667

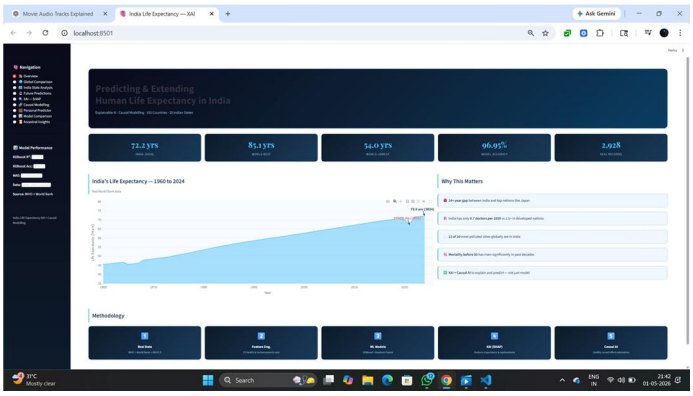


Fig. 2. Dashboard

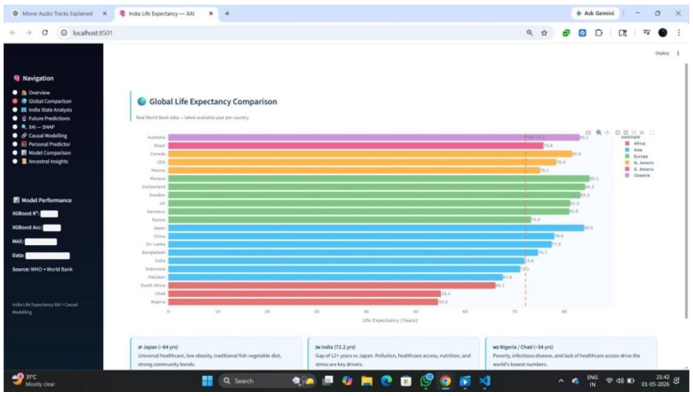


Fig. 3. Global comparison against 20 countries

and MAE=1.07 years). The marginal MAE advantage is well within measurement noise, but XGBoost's superior R^2 value shows us better global variance capture.

B. SHAP Feature Importance

Prevalence of HIV/AIDS is the dominant predictor (Mean—SHAP—=3.11 Years), reflecting the dangerous impact of AIDS on life expectancy , especially in Sub-Saharan countries during 2000-2015. Adult mortality rate ranks second (2.52 years), functioning partly as a proxy for unobserved confounders like famine and healthcare collapse. Income composition of resources (income, education and health) ranks third (1.76), showcasing that multi-dimensional development is superior to income alone. (Fig. 8)

C. Causal Inference

The DoWhy backdoor adjustment estimate yields ATE=+4.29 years above-median schooling after conditioning the factors GDP, HIV/AIDS, income composition and development status. This is consistent with education on health, preventive behavior and economic attainment [11]. The negative ATE of AIDS (-12.4 years) and adult mortality (-9.8 years) confirm these as top priorities to intervene and work on. (Fig. 9)

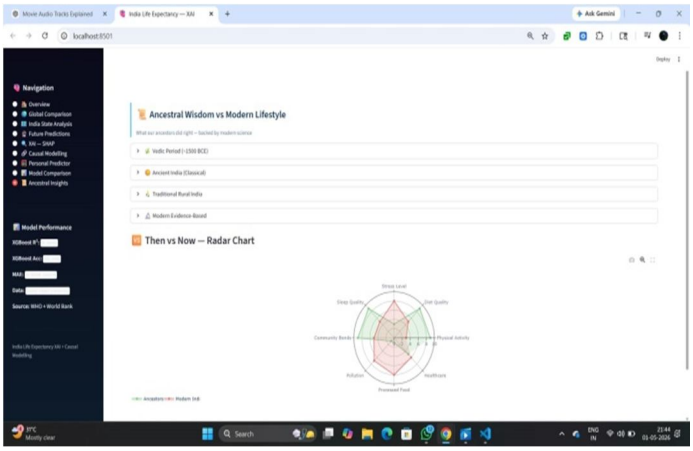


Fig. 4. Ancestral Wisdom-I

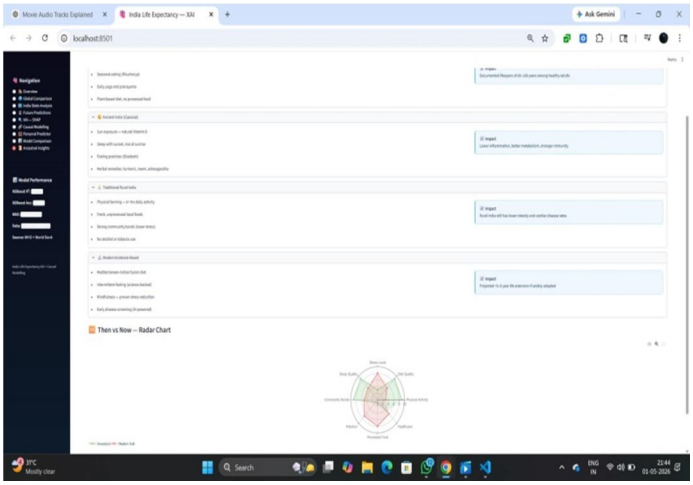


Fig. 5. Ancestral Wisdom-II

D. State-level analysis on India

Among our Indian states, Kerala leads the chart with 75.0 years life expectancy, reflected by their highest physician density (2.1 per 1000) and high vaccination rate (89.3%).

Bihar records the lowest with 63.5 years of life expectancy, and sadly associated with country's lowest health insurance penetration (9.8%), high tobacco use (48.6%) and physician density of only 0.4 per 1000. The Pearson correlation between physician density and life expectancy of state is $r = 0.87 (p \leq 0.001)$, confirming public healthcare access is the driving factor of life expectancy and quality of health among masses. (Fig. 10)

E. Future Projections

Under the Business-as-usual scenario, India's life expectancy halts at 69.9 years till 2050, reflecting a steady prediction when parameters are held constant. The Optimistic scenario projects an increase to 72 years by 2050, driven primarily by investment in public health expenditure from 3.5% to 6.5% of GDP and raising mean schooling from 12 to 15 years - showing how important these factors are. The Pessimistic scenario however, projects a decline to 67.5 years, emphasizing the risks of alcohol consumption and policy inaction. The 4.5 years gap between Optimistic and

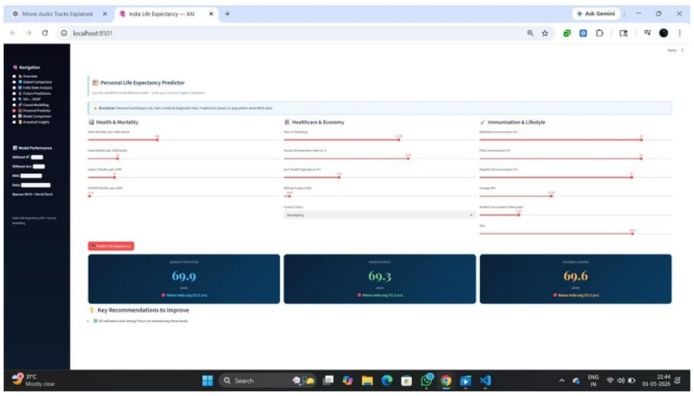


Fig. 6. Personal Life Expectancy Predictor (14 user inputs)

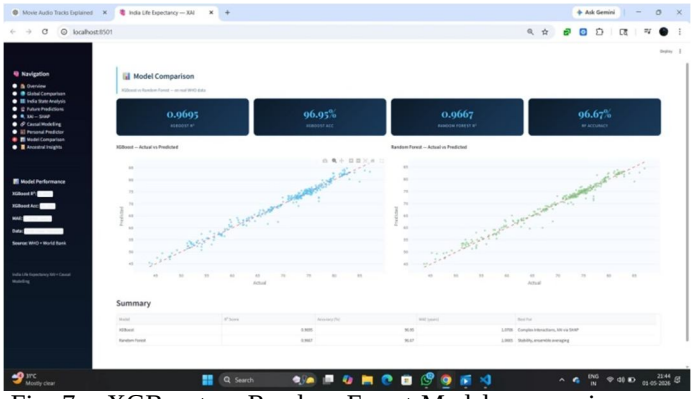


Fig. 7. XGBoost vs Random Forest Models comparison

Pessimistic scenarios showcase the vulnerability of life expectancy even by small changes. (Fig. 11)

F. Policy Implications

The integrated findings carry actionable implications on India and Global health policy. First, education investment yields the most effective causal gain of +4.29 years, particularly for girls given the anemia and literacy gaps observed in NFHS-5. Second, the 12 year gap for Bihar shows us that National averages masks inequalities and targeted resources transfers to Bihar, MP and Chhattisgarh are necessary. Third, disproportionate SHAP weight of AIDS (globally) and tobacco consumption (India-specific) calls for education to public on these matters and necessary steps to be taken by Government in these regards. Fourth, by simply doubling the expenditure on public healthcare services from 3.5% to 6.5% is expected to add 2 years of life expectancy under causal model. Lastly, the personal predictor module in dashboard helps an individual to track his health by himself and take accountable and measurable steps using the insights of translation of population level outputs to individual risk assessment.

VI. LIMITATIONS AND FUTURE WORK

The WHO training data spans from 2010 to 2015 so it may not fully reflect the post-pandemic trend in mortality

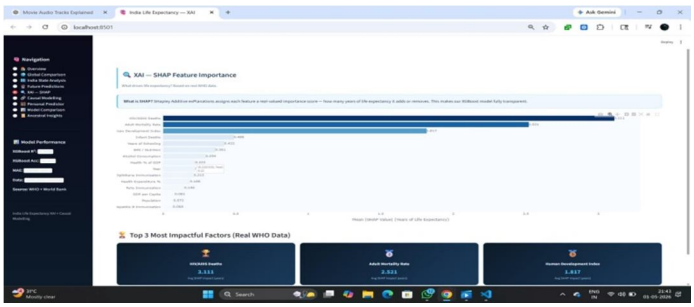


Fig. 8. XAI-SHAP bar graph

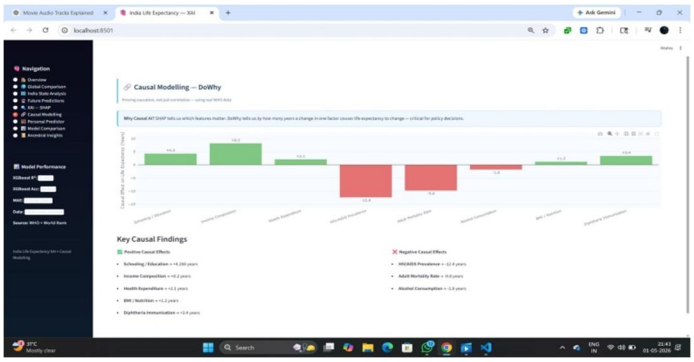


Fig. 9. Causal modeling using DoWhy

patterns, which caused a global decline by 1.8 years between 2019 and 2021 [1]. Future iterations should include post pandemic data as it becomes available. The India state-level dataset is small (n = 25) and relies on NFHS-5 cross-sectional tabulations rather than longitudinal panel data; panel modeling would enable stronger causal identification at the sub-national level. Future work will explore federated learning to incorporate hospital-level electronic health records, deep learning on longitudinal cohort data, and reinforcement learning for dynamic policy optimization.

VII. CONCLUSION

We have presented a compatible, comprehensive, interpretable and casually grounded data driven machine learning framework for life expectancy prediction. The XGBoost model achieves $R^2 = 0.97$ on WHO data. SHAP analysis identifies HIV/AIDS prevalence, adult mortality and income composition as high impact predictors. DoWhy Causal modeling shows us that above median education increases the life expectancy by 4.29 years. When applied to India, the framework reveals a 12 year sub-national gap and projects a aggressive policy reform could yield a 2 year increase by 2050.

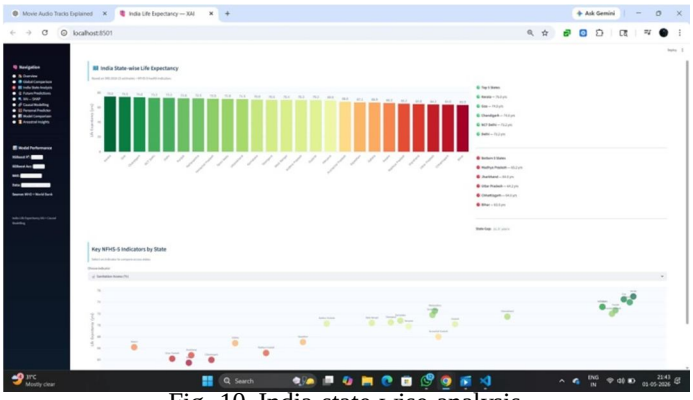


Fig. 10. India state wise analysis

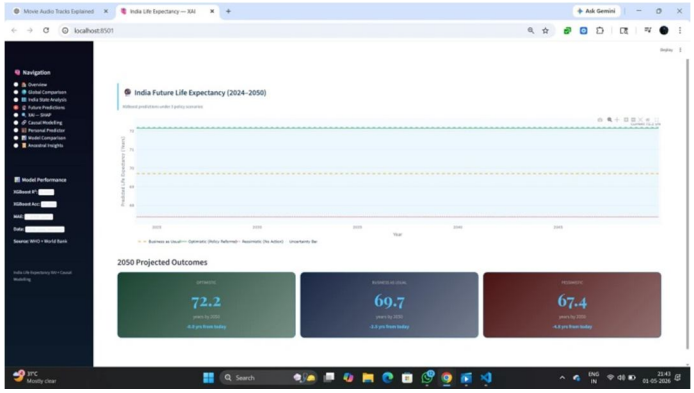


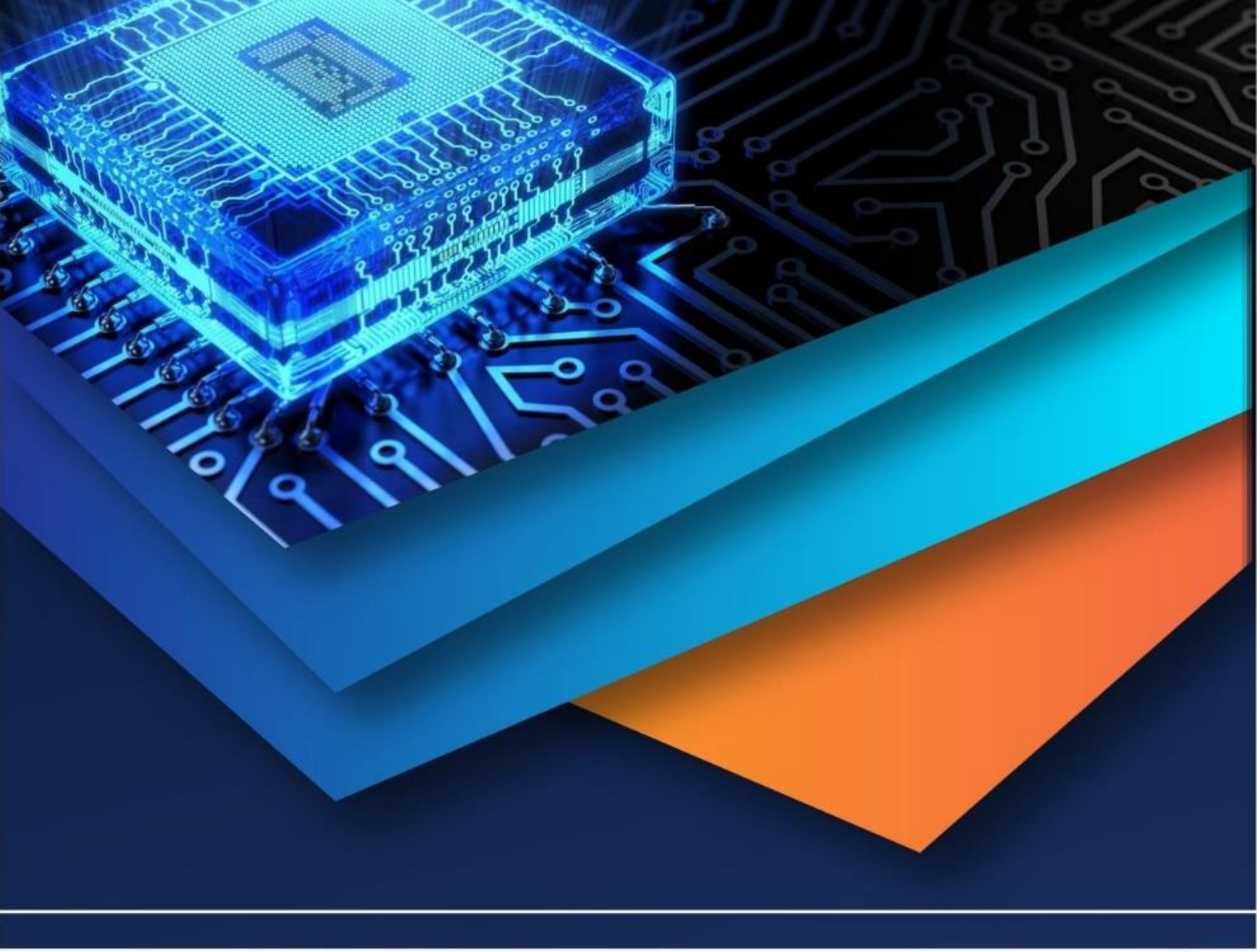
Fig. 11.India future life expectancy prediction

VIII. ACKNOWLEDGMENT

We want to thank our mentor Mr. Para Upender Naidu for guiding us at every step and providing his insights which helped us develop this project.

REFERENCES

[1] R. Dolgopolyi, I. Amaslidou, and A. Margaritou, "Interpretable ma- chine learning for life expectancy prediction: A comparative study of linear regression, decision tree, and random forest," arXiv preprint arXiv:2510.00542, 2025.
[2] B. Lantz, Machine Learning with R, 3rd ed. Birmingham: Packt Publishing, 2019.
[3] K. Kawano, Y. Otaki, N. Suzuki, S. Fujimoto, K. Iseki et al., "Prediction of mortality risk of health checkup participants using machine learning- based models: the J-SHC study," Scientific Reports, vol. 12, p. 14113, 2022.
[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
[5] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 1135–1144.
[6] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, no. 1, pp. 56–67, Jan. 2020.
[7] A. Sharma and E. Kiciman, "DoWhy: An end-to-end library for causal inference," arXiv preprint arXiv:2011.04216, 2020.
[8] J. Pearl, "The do-calculus revisited," in Proc. 28th Conf. Uncertainty in Artificial Intelligence (UAI), Catalina Island, CA, 2012, pp. 3–11.
[9] A. Sharma and E. Kiciman, "DoWhy: Addressing challenges in express- ing and validating causal assumptions," in Proc. Workshop on CausalML, NeurIPS, 2019.
[10] R. Chetty, M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler, "The association between income and life expectancy in the United States, 2001–2014," JAMA, vol. 315, no. 16, pp. 1750–1766, Apr. 2016.
[11] A. Lleras-Muney, "The relationship between education and adult mor- tality in the United States," Review of Economic Studies, vol. 72, no. 1, pp. 189–221, Jan. 2005.
[12] International Institute for Population Sciences (IIPS) and ICF, "National family health survey (NFHS-5), 2019–21: India," IIPS, Mumbai, Tech. Rep., 2022.
[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system,"in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, 2016, pp. 785–794.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY