IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 3, May-June 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203080

Exploring Machine Learning Models for Life Expectancy Prediction Based on Socio Economic and Health

A. Lakshmipathi Rao[1], Eeduru Raju[2], Gottumukkala Jithendra Varma [3], Bandi Nithin[4]

Assistant Professor, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[1]

Student, Department of CSE, Guru Nanak Institute of Technology, Hyderabad, Telangana, India[2-4]

ABSTRACT: Life expectancy (LE) models have vast effects on the social and financial structures of many countries around the world. Numerous studies have highlighted the significant impact of life expectancy prediction on societal dynamics and the management of healthcare systems worldwide. These predictive models offer various avenues to enhance healthcare delivery and facilitate advanced care planning at the societal level. However, over time, it has become evident that existing determinants are insufficient to accurately estimate the longevity of broader population groups. Earlier models primarily relied on mortality-based data derived from specific sample populations. With the evolution of predictive technologies and the accumulation of extensive research, scholars have recognized that, beyond mortality rates, several additional factors must be considered to develop robust and reliable life expectancy prediction models. Consequently, current research increasingly incorporates variables related to education, health, economic conditions, and social welfare systems.. In the Analysis, the authors have implemented different machine learning algorithms and have achieved better accuracy on random forest based on the dataset

I. INTRODUCTION

This study explores the application of machine learning models to predict life expectancy based on a comprehensive set of socioeconomic and health factors. Traditional methods of life expectancy prediction have often relied on limited datasets and simplistic statistical models, which fail to capture the complex interplay of various determinants influencing longevity. By leveraging advanced machine learning techniques, this research aims to develop a robust predictive model that integrates diverse datasets, including environmental, genetic, and lifestyle factors. The model seeks to provide accurate and interpretable predictions, thereby offering valuable insights for healthcare planning, resource allocation, and policy-making. This introduction outlines the objectives, significance, and potential impact of the project in enhancing our understanding of life expectancy determinants and improving public health outcomes.

II.LITERATURE SURVEY

Title : Significance of Non- Academic Parameters for Predicting Student Performance Using Ensemble Learning Techniques **Year:** 2024
Author: Aggarwal, D., Mittal, S. and Bali, V
Description: Educational institutions are increasingly leveraging data mining techniques to enhance student performance. This study presents a comparison between two models—one developed solely with academic factors and the other incorporating both academic and non-academic (demographic) variables.The models are built using eight classification algorithms that are then compared to find the parameters that help to give the most appropriate model to classify a student based on his performance.

Title: Risk prediction in life insurance industry using supervised learning algorithms
Year: 2022
Author: Noorhannah Boodhun, Manoj Jayabalan.
Description: Risk assessment constitutes a fundamental aspect of the life insurance industry, enabling the classification of applicants based on their risk profiles. Insurers utilize the underwriting process to evaluate applications and determine policy pricing. With the proliferation of data and advancements in data analytics, the underwriting process can now be partially or fully automated, thereby expediting application evaluations. This research focuses on enhancing risk assessment practices within life insurance firms through the application of predictive analytics. An anonymized real-world dataset containing over one hundred attributes was employed for analysis. To improve model performance

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203080

and computational efficiency, dimensionality reduction techniques were applied to identify the most influential features. Both feature selection and feature extraction methods were used—specifically, Correlation-Based Feature Selection (CFS) and Principal Component Analysis (PCA).The predictive modeling involved the application of various machine learning algorithms, including Multiple Linear Regression, Artificial Neural Networks (ANN), Reduced Error Pruning Tree (RepTree), and Random Tree classifiers. Experimental results indicated that the RepTree algorithm outperformed other models under the CFS method, achieving the lowest mean absolute error (MAE) of 1.5285 and the lowest root mean squared error (RMSE) of 2.027. Conversely, Multiple Linear Regression yielded the best results when used with PCA, attaining the lowest MAE and RMSE values of 1.6396 and 2.0659, respectively, among the models evaluated.

Title: Novel Approach for Wind Speed Forecasting Using LSTM-ARIMA Deep Learning Models
Year: 2021
Author: Bali V. , Kumar, A. and Gangwar ,S.
Description: The term which is used to predict wind speed to produce wind power is wind speed forecasting. Deep learning, is a form of AI, basically indulging in artificial intelligence and thus can greatly increase the precision rate on larger datasets. In this research paper, the two techniques are being used together to obtain the better forecasting results. Both the techniques are forecasting based and combining LSTM. The forecasting accuracy can be enhanced by leveraging the pattern retention capability of LSTM, which excels at learning long-term dependencies. Integrating the ARIMA model further strengthens the prediction by estimating the probability of future values falling within a defined range. Therefore, combining both techniques into a hybrid model is likely to produce more accurate results than using either method independently. The primary objective of this research is to explore and evaluate the effectiveness of such a hybrid forecasting approach.

III. EXISTING SYSTEM

Estimating life expectancy is essential for enabling well-informed end-of-life decisions. Accurate predictions aid in determining appropriate treatment strategies, optimizing healthcare resource allocation, and supporting structured Advance Care Planning. In the existing approach, life expectancy prediction is formulated as a supervised machine learning problem. A Long Short-Term Memory (LSTM) recurrent neural network was trained and validated using medical records of individuals who had passed away. The model was developed using a ten-fold cross-validation method, ensuring robustness and generalizability, and its performance was evaluated on an independent test dataset.
Two models were considered for comparison: a baseline model utilizing only structured clinical data, and an enhanced model that incorporated additional features extracted from unstructured clinical notes (referred to as the "keyword model"). The predictive performance of both models was benchmarked against physicians' prognostic accuracy, as reported in prior scholarly studies.

EXISTING SYSTEM DISADVANTAGES

- It takes longer to train & require more memory to train.
- Dropout is much harder to implement and are sensitive to different random weight initializations.

IV. PROPOSED SYSTEM

Life expectancy is a statistical measure of the average time an organism is expected to live. The analysis incorporates statistical data related to year of birth, current age, and various demographic factors. As defined by the World Bank, *life expectancy at birth* represents the average number of years a newborn is projected to live, assuming that current mortality patterns remain unchanged throughout their lifetime. In essence, it reflects the age-specific mortality rates of a given year, offering a snapshot of the population's overall mortality profile. This constitutes a retrospective or back-end analysis. In contrast, the objective of this study is to forecast life expectancy at birth through a forward-looking approach. We propose a front-end analysis that estimates life expectancy based on prevailing living conditions and the availability of regional resources, employing diverse machine learning techniques for predictive modeling.

PROPOSED SYSTEM ADVANTAGES:

- It can perform both regression and classification tasks.
- It can produce good predictions that can be understood easily.
- It can handle large datasets efficiently

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203080

V. SYSTEM ARCHITECTURE

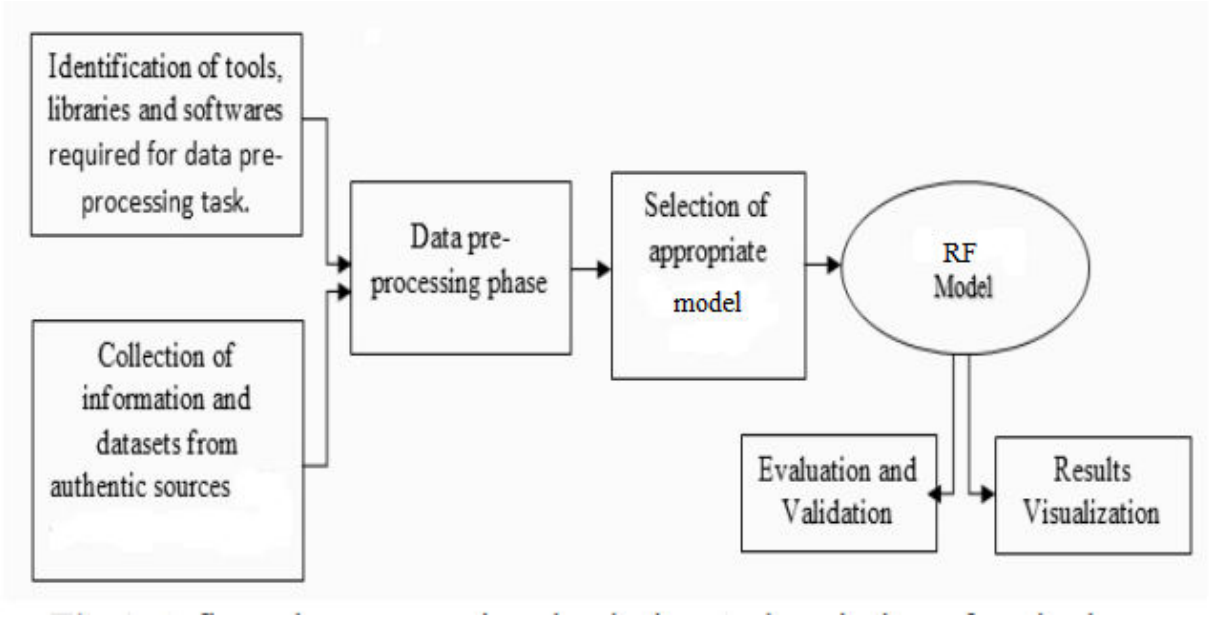

Figure 1: System Architecture

VI. METHODOLOGIES

MODULES NAME AND EXPLANATION:

Data Collection:

This is the first real step towards the real development of a learning model, collecting data. This step is critical, as the quality and comprehensiveness of the data significantly influence the model's performance; more accurate and extensive datasets typically yield more reliable and precise predictive outcomes.

There are several techniques to collect the data, like web scraping, manual interventions and etc.

- Life Expectancy Prediction depending on Health Data

Dataset:

The dataset consists of individual data in that there are 2939 rows & 22 columns in the dataset, which are described below:

1. Country
2. Year
3. Status
4. Life expectancy
5. Adult Mortality
6. infant deaths
7. Alcohol
8. percentage expenditure
9. Hepatitis B
10. Measles
11. BMI
12. under-five deaths
13. Polio
14. Total expenditure
15. Diphtheria
16. HIV/AIDS
17. GDP
18. Population
19. Thinness 1-19 years

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203080

20. Income composition of resources
21. Schooling

Data Preparation:
Wrangle data and prepare it for training. Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc. Randomizing data removes any biases caused by the specific sequence in which the data was collected or processed. Visualizing the data aids in uncovering important relationships between variables, identifying class imbalances (potential bias), and conducting exploratory analysis.
Split into training and evaluation sets.

Model Selection:
We used Random Forest model created our Life Expectancy Prediction algorithm, we got an accuracy of 97.6% so we implemented this algorithm for testing / prediction.

Analyze and Prediction:
In the actual dataset, we chose some main features like:
1. Health data - detailed description of columns for life health status data.
2. Consumption of any drug content - indicates whether the person is taking any kid of alcohol and so on etc.,
3. Score – what is the life Expectancy of a person depending on health data and other data.

Accuracy on test set:
We got an accuracy of 96.04% on test set.

Saving the Trained Model:
After thoroughly training and testing your model, the first step toward deploying it in a production environment is to save the model into a .h5 or. pkl file using a library like pickle.
Make sure you have pickle installed in your environment.
Next, let's import the module and dump the model into.pkl file

VII. ALGORITHM USED

Random Forest Algorithm:
Introduction:
Random Forest, an ensemble learning technique, works by building multiple decision trees during training and making predictions based on the majority vote for classification tasks or the average output for regression tasks.. It is designed to improve the predictive accuracy and control over-fitting compared to single decision trees.

Key Concepts: Ensemble Learning: Combines multiple models to solve a particular computational problem.

Algorithm Steps:
1. Bootstrap Sampling: Randomly select subsets of data from the training set with replacement. Each subset is used to train an individual decision tree.
2. Feature Selection: For each split in a decision tree, randomly select a subset of features. The best feature from this subset is used to make the split.
3. Decision Tree Construction: Construct decision trees for each bootstrap sample using the selected features. Trees are grown to the maximum depth and are not pruned.
4. Aggregation: For classification tasks, the final prediction is determined by majority voting across all trees. For regression tasks, the final prediction is the average of the predictions from all trees.

Advantages
- Accuracy: Generally provides high predictive accuracy due to the ensemble nature.
- Robustness: Handles missing values and maintains accuracy even when a portion of the data is missing.
- Versatility: Can be used for both classification and regression tasks.

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in | | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203080

VIII. EXPERIMENTAL RESULTS

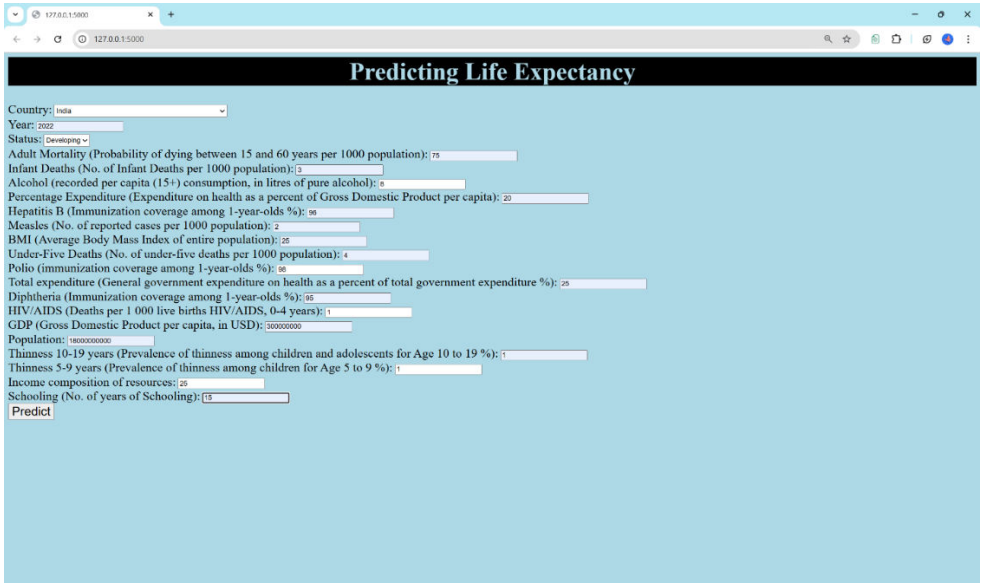

Figure 2: User Interface of the project

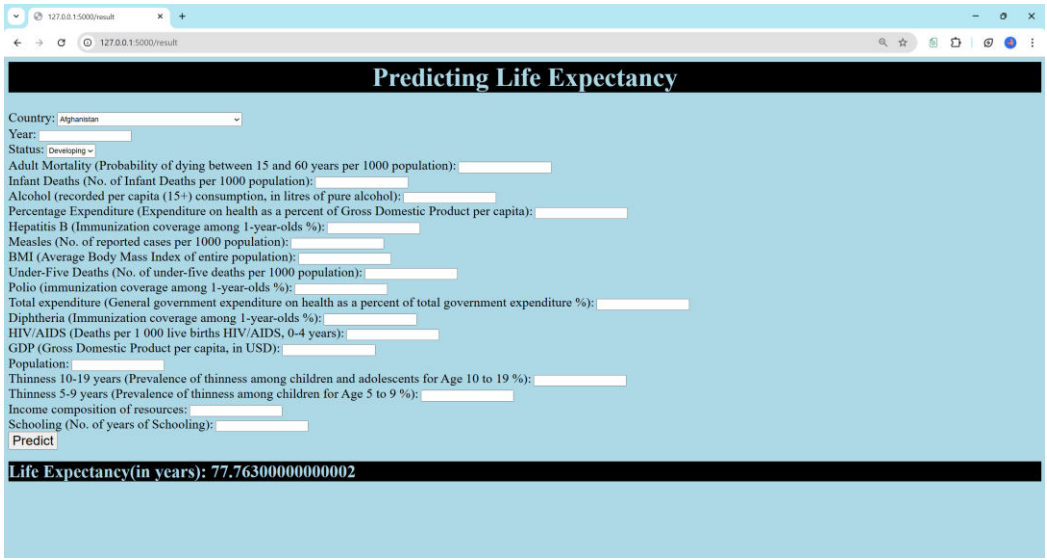

Figure 3: Life expectancy output

IX. CONCLUSION

Initially, authors have dropped features such as year, country, and status. The main aim was to analyze the impact of features on the outcome and how it varies. The first task was to find the best-performing model. Among different models, random forest performs best with an MAE of 1.27 and an R2 score of 96% on the test set. Adult mortality, HIV/AIDS, schooling, and BMI are the most impacting factors on life expectancy among the features. Schooling, Income Composition, and BMI have positively correlated to the outcome. Surprising thing was that some features such as GDP, total expenditure, and infant deaths were not that impactful on the final result. But the initial assumption is proven wrong here about these features.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 3, May - June 2025 ||

DOI:10.15680/IJARETY.2025.1203080

X. FUTURE ENHANCEMENT

The ongoing evolution of data science and machine learning offers significant opportunities for enhancing life expectancy prediction models. Future enhancements will focus on several key areas:

1. Advanced Data Integration and Quality Assurance : Prioritize the integration of diverse datasets, including environmental, genetic, and lifestyle factors.
2. Enhanced Feature Engineering and Selection : Develop advanced techniques for automated feature selection and extraction. Utilize cutting-edge algorithms to identify impactful predictors and sophisticated methods for analyzing feature importance.
3. Model Optimization and Hybridization : Explore the integration of Random Forest with other advanced models, such as neural networks and support vector machines. Create hybrid models and employ techniques like Bayesian optimization for hyperparameter tuning.
4. Interpretability and Explainability : Implement rule extraction, decision path analysis, and explainable AI (XAI) techniques to enhance model

REFERENCES

[1] Yoneda, S., Švábenský, V., Li, G., Deguchi, D., & Shimada, A. (2025). Ranking-Based At-Risk Student Prediction Using Federated Learning and Differential Features.
[2] Junejo, N. U. R., Nawaz, M. W., Huang, Q., Dong, X., Wang, C., & Zheng, G. (2024). Accurate Multi-Category Student Performance Forecasting at Early Stages of Online Education Using Neural Networks.
[3] EL Habti, F. E., Hiri, M., Chrayah, M., Bouzidi, A., & Aknin, N. (2025). Enhancing Student Performance Prediction in e-Learning Ecosystems Using Machine Learning Techniques. International Journal of Information and Education Technology, 15(2), 301–311.
[4] Abdulkadir, U. I., & Fernando, A. (2024). A Deep Learning Model for Insurance Claims Predictions. Journal on Artificial Intelligence, 6(1), 71–83.
[5] Shou, Z., Xie, M., Mo, J., & Zhang, H. (2024). Predicting Student Performance in Online Learning: A Multidimensional Time-Series Data Analysis Approach. Applied Sciences, 14(6), 2522.
[6] Chen, Y., Sun, J., Wang, J., Zhao, L., Song, X., & Zhai, L. (2025). Machine Learning-Driven Student Performance Prediction for Enhancing Tiered Instruction. arXiv preprint arXiv:2502.03143.
[7] Jimenez Martinez, A. L., Sood, K., & Mahto, R. (2024). Early Detection of At-Risk Students Using Machine Learning. arXiv preprint arXiv:2412.09483.
[8] Leelaluk, S., Tang, C., Švábenský, V., & Shimada, A. (2024). Knowledge Distillation in RNN-Attention Models for Early Prediction of Student Performance.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com