

A MACHINE LEARNING FRAMEWORK FOR REGIONAL LIFE EXPECTANCY PREDICTION

Huma Sania¹, Lubna Nausheen², Ruqiya Fatima³

¹PG Scholar, Department of CSE, Shadan Women's College of Engineering and Technology, Hyderabad,
humasania24@gmail.com

²Asst Professor, Department of CSE, Shadan Women's College of Engineering and Technology
lubnanausheen07@gmail.com

³Asst Professor, Department of CSE, Shadan Women's College of Engineering and Technology
Ruqiyafatima2716@gmail.com

ABSTRACT:

Life expectancy is a crucial indicator of a population's overall health and standard of living. This work offers an AI-driven approach for accurately predicting life expectancy using state-of-the-art machine learning algorithms. In particular, the CatBoost approach—which is renowned for its superior performance on structured data and categorical variables—was used to build a robust prediction model. The model was trained using a big dataset of health, demographic, and socioeconomic data from multiple countries. Performance evaluation results were positive, with a Mean Squared Error (MSE) of 0.18, Mean Absolute Error (MAE) of 1.15, testing R2 score of 93%, and training R2 score of 95%. These metrics confirm the model's excellent generalisation to new data and strong prediction power. By identifying key factors that affect life expectancy, this AI-powered approach can support health policy planning, improve resource allocation, and promote preventative healthcare initiatives across communities. The project shows how machine learning can be effectively used to provide insightful information on longevity prediction and public health.

Keywords: Machine Learning, CatBoost, Life Expectancy Prediction, Socioeconomic Factors, HealthCare data and Regression Models.

1. INTRODUCTION:

One important measure of a population's general health and well-being is life expectancy [8]. Numerous factors, such as lifestyle choices, healthcare access, and socioeconomic conditions, have a substantial impact on life expectancy. Public health activities are crucial since diseases like Hepatitis B have been shown to have a detrimental effect on longevity in different places [2]. Globally, life expectancies vary due to geographical variations in health-related socioeconomic factors [5]. This paper suggests a reliable and effective machine learning framework that makes use of the CatBoost algorithm in order to overcome the shortcomings of current predictive models in life expectancy estimate. While earlier research has examined methods like Random Forest, Artificial Neural Networks (ANN), Linear Regression, and most notably XGBoost, these models frequently face difficulties with managing categorical data, computational complexity, and overfitting on structured datasets.

Despite its great accuracy, XGBoost is quite sensitive to hyperparameter adjustment and necessitates a lot of data preprocessing, including the hand encoding of categorical variables. Particularly in healthcare datasets where categorical characteristics like geography, gender, smoking status, and lifestyle behaviours are crucial, these constraints may have a detrimental impact on model generalisation. Yandex created CatBoost (Categorical Boosting), a potent gradient boosting decision tree algorithm created especially to address these issues. CatBoost allows

native handling of categorical information without additional preprocessing or encoding, in contrast to conventional boosting techniques.

In order to avoid the prediction shift that is frequently introduced in standard boosting, it employs sophisticated approaches like Ordered Boosting, which guarantees that each decision tree is built during training using a carefully shuffled dataset. This method greatly enhances generalisation performance and lessens overfitting. CatBoost also uses symmetric (oblivious) trees, which offer quicker training, less memory use, and more reliable predictions between runs.

1.1 OBJECTIVE:

The objective of this research is to develop a reliable and intelligent system that accurately predicts life expectancy using state-of-the-art machine learning techniques, particularly the CatBoost algorithm. By examining a range of health-related, demographic, and lifestyle factors, such as age, gender, smoking and alcohol use, BMI, blood pressure, cholesterol, and physical activity, the research aims to identify the primary factors influencing human lifespan. By leveraging CatBoost's ability to handle categorical input and reduce overfitting, the method seeks to improve prediction accuracy relative to more traditional models like XGBoost and Random Forest. Another objective is to provide an easy-to-use web interface that enables users to input their personal health data and obtain a life expectancy estimate. This

tool is intended to help medical professionals, researchers, and patients.

2. EXISTING SYSTEM:

In order to forecast life expectancy based on health, demographic, and socioeconomic parameters, the current system used the XGBoost (Extreme Gradient Boosting) algorithm. Despite being a potent and popular machine learning method, XGBoost has a number of drawbacks. Its inability to handle categorical data directly is a significant disadvantage; it frequently necessitates extra preprocessing, such as label encoding or one-hot encoding, which adds complexity and runs the risk of introducing noise. Additionally, XGBoost may overfit on smaller or unbalanced datasets and be sensitive to hyperparameter settings. These elements could make the model less interpretable and less able to generalise, which would restrict its usefulness in real-world applications with complicated and varied data.

2.1 Existing System Disadvantages:

- Hard to handle text data It doesn't support category/text features directly. You have to convert them first.
- Can overfit If not tuned properly, it may give good results on training data but poor on new data.
- Needs tuning You have to adjust many settings (like depth and learning rate) to get good results.
- Takes time to train for large datasets, it can be slow and use a lot of memory.
- Not easy to explain the model is complex, so it's hard to understand how it's making predictions.

3. LITERATURE SURVAY:

CatBoost: Unbiased Boosting with Categorical Features, L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, and A. Gulin, 2023.

This paper presents the key algorithmic techniques behind CatBoost, a new gradient boosting toolkit. Their combination leads to CatBoost outperforming other publicly available boosting implementations in terms of quality on a variety of datasets. Two critical algorithmic advances introduced in CatBoost are the implementation of ordered boosting, a permutation-driven alternative to the classic algorithm, and an innovative algorithm for processing categorical features. Both techniques were created to fight a prediction shift caused by a special kind of target leakage present in all currently existing implementations of gradient boosting algorithms. In this paper, we provide a detailed analysis of this problem and demonstrate that proposed algorithms solve it effectively, leading to excellent empirical results.

DIGWO-N-BEATS: An evolutionary time series prediction method for situation prediction, H. Lin and C. Wang, 2022

Situation awareness is a key technology in many decision systems, but its performance bottleneck in the third step, namely situation prediction, has not been overcome yet. This step involves time series prediction and commonly uses deep learning methods. However, there are significant performance differences among these methods, and the tuning of their hyperparameters has not been thoroughly investigated. To address these challenges, a novel prediction method named Distributed Improved Gray Wolf Optimizer-Neural Basis Expansion Analysis for Time-Series (DIGWO-N-BEATS) is proposed for situation prediction tasks. First, an architecture based on N-BEATS, which is a cutting-edge paradigm, is formulated for modeling situation value time series. Second, a novel improved evolutionary algorithm, which can converge in parallel, is proposed to optimize thirteen hyperparameters and the model structures of N-BEATS. The experiment results demonstrate that DIGWO-N-BEATS outperforms the six most competitive baselines, reducing the average MAPE on two real-world situation awareness datasets and two time-series prediction datasets by 8.18%, 1.12%, 9.92%, and 4.98%, respectively. Furthermore, DIGWO-N-BEATS exhibits good convergence in hyperparameter optimization tasks and scalability.

4. RELATED WORK:

A deep learning-based framework for predicting life expectancy using Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) was presented by Beeksma et al. [4]. Electronic health records with patient demographics, medical history, lab results, and prescription information were used in the study. The LSTM-RNN model was chosen due to its capacity to capture temporal dependencies in patient records and analyse sequential data. The findings showed that deep learning methods can enhance prediction accuracy and efficiently model complex healthcare data.

Using Random Forest Regressor, Linear Regression, and Decision Tree Regressor models, Meshram et al. [5] created a machine learning-based prediction system. The study's goal was to compare life expectancy in developed and developing nations. Metrics including the R2 score, Mean Squared Error (MSE), and Mean Absolute Error (MAE) were used to assess the model's performance. The study highlighted the significance of socioeconomic and health-related indicators by identifying important factors impacting life expectancy, such as adult mortality rates, HIV/AIDS prevalence, and healthcare spending.

A Random Forest model was presented by Bali et al. [8] to forecast life expectancy based on socioeconomic, environmental, health, and

demographic characteristics. When the authors tested different machine learning approaches, they discovered that ensemble-based approaches performed better. Their results showed that the reliability of life expectancy calculation is improved by incorporating several predictive elements. The study also highlighted how feature selection and data pretreatment contribute to improved prediction outcomes.

An ensemble learning model based on XGBoost was presented by Lipesa et al. [10], and it outperformed Random Forest and Artificial Neural Networks (ANN) models in terms of prediction accuracy. The model's efficacy was further enhanced via hyperparameter optimisation, and accuracy and recall metrics were employed for validation. Furthermore, Chen et al. [15] suggested a linear regression method for censored data to deal with circumstances in which full survival data is not available. Their work showed the efficacy of statistical learning techniques for life expectancy projection and offered insightful information about survival analysis.

5. PROPOSED SYSTEM:

By including the CatBoost algorithm, a gradient boosting method designed especially for managing structured data and categorical features, the suggested solution overcomes these difficulties. In contrast to XGBoost, CatBoost handles categorical variables natively, removing the requirement for significant data modification and maintaining the input data's original structure. Additionally, it incorporates sophisticated regularisation strategies to reduce overfitting and enhance model stability. With a reduced Mean Squared Error (0.18), a smaller Mean Absolute Error (1.15), and higher R^2 scores for both training (95%) and testing (93%), the CatBoost-based model thus performs better than the prior system. These enhancements show how well the model can generalise and has a high predicted accuracy. As a result, the suggested system improves life expectancy prediction's overall performance and dependability, providing a more effective and precise means of assisting with policy planning and healthcare decision-making.

5.1 Proposed System Advantages:

- **Handles categorical data directly** No need to convert text or category columns manually.
- **Less overfitting** Uses a special technique called *ordered boosting* to reduce overfitting.
- **Easy to use** Works well with default settings, so you don't have to tune too many parameters.
- **Fast and efficient** Trains quickly and uses less memory compared to some other boosting methods.

6. MODULES:

• Data Collection Module:

Purpose: To collect a reliable dataset containing essential health, demographic, and lifestyle information.

Explanation: This module collects information about life expectancy from international health sources. It takes into account factors including age, gender, blood pressure, BMI, cholesterol, smoking, drinking, diabetes, and degree of physical activity.

• Data Preprocessing Module:

Purpose: To clean and prepare the raw data for machine learning modelling.

Explanation: This module guarantees that all data is in an appropriate format for model training, manages missing values, identifies outliers, and encodes categorical features (although CatBoost handles them natively). If necessary, it also incorporates standardisation or normalisation.

• Exploratory Data Analysis (EDA) Module:

Purpose: To understand data distributions and relationships between variables.

Explanation: This module makes it easier to see how different factors affect life expectancy by using statistical graphs, charts, and correlation heatmaps. Additionally, it determines which predictors are most important.

• Model Building Module (CatBoost Algorithm):

Purpose: To develop and train the life expectancy prediction model using CatBoost.

Explanation: This is the main module that trains a regression model on the dataset using the CatBoost algorithm. CatBoost was chosen because it performs exceptionally well on structured data and can handle category variables directly with less overfitting.

• Model Evaluation Module:

Purpose: To assess the model's accuracy and performance.

Explanation: Metrics like Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 score are used in this module to assess the trained model. It guarantees that the model works properly on training and testing data.

• Prediction Module:

Purpose: To generate life expectancy predictions based on user inputs.

Explanation: This module uses the trained CatBoost model to forecast an individual's predicted life span based on additional input data, such as age, BMI, and health practices.

• Web Interface Module (Flask Integration):

Purpose: To provide an easy-to-use interface for user interaction.

Explanation: Users can enter their personal information and receive real-time life expectancy projections through a Flask-based web application. This enhances the system's usability and accessibility.

7. TECHNIQUE OR ALGORITHM USED:

7.1 EXISTING TECHNIQUE:

• XGBoost Algorithm:

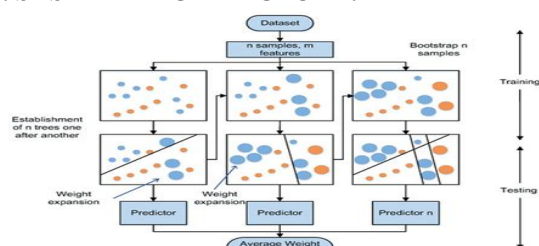
Gradient boosting is a machine learning technique that generates an ensemble of decision trees in a sequential manner. XGBoost is an enhanced implementation of this technique. In XGBoost, every new tree attempts to correct the mistakes caused by the earlier trees. Gradient descent, which modifies the tree splits to gradually lower prediction errors, is used by the algorithm to minimise a loss function (such as mean squared error). To speed up training and lessen overfitting, XGBoost employs strategies like tree trimming, regularisation, and parallel processing. Nevertheless, categorical data is not directly handled by XGBoost. Label encoding or one-hot encoding must be used to transform categories into numbers, which occasionally adds complexity or noise.

7.2 PROPOSED TECHNIQUE USED OR ALGORITHM USED:

• CatBoost Algorithm:

Yandex created the gradient boosting algorithm CatBoost with the express purpose of effectively managing datasets with categorical features. It constructs trees one after the other, fixing the mistakes of the previous one, just as XGBoost. But what makes CatBoost unique is that it handles categorical data natively, which means it uses statistical techniques like target-based encoding to automatically translate categorical features into numbers without the need for human preprocessing. Additionally, it presents ordered boosting, a method that ensures that each new tree is trained using a carefully randomised version of the data in order to prevent overfitting. Furthermore, CatBoost performs exceptionally well on structured/tabular datasets, requires less parameter adjustment, and is generally more stable.

8. SYSTEM ARCHITECTURE:



9. RESULTS:



SIGN UP

USERNAME

PASSWORD

LOGIN



LIFE EXPECTANCY PREDICTOR

Enter Input Features

Age

Gender

Smoker?

Alcohol?

Diabetes?

Blood Pressure

Cholesterol

BMI (e.g., 22.5)

Glucose Level (e.g., 90.0)

Exercise Hours/Week (e.g., 5.0)

Sleep Hours/Night (e.g., 7.0)

Predict

Enter Input Features

25

Female

No

No

No

Normal

Normal

40

80

4

5

Predict

SIGN UP

USERNAME

admin

PASSWORD

Login

[HOME](#) [ABSTRACT](#) [PERFORMANCE](#) [CONTACT](#) [LOGOUT](#)



PREDICTION

HOME / PREDICT

 Predicted Life Expectancy: 74.52 years

10. DISCUSSION AND CONCLUSION:

The CatBoost algorithm, a sophisticated gradient boosting method created especially to handle categorical information, is used in this study to propose a reliable and intelligent machine learning model for life expectancy prediction. The suggested model has demonstrated notable gains in predicting accuracy and generalisation by integrating CatBoost's inherent support for categorical data and its ordered boosting technique with a large dataset that includes health, lifestyle, and demographic characteristics. The model outperformed previous approaches, such as XGBoost, with a training R^2 score of 95%, a testing R^2 score of 93%, and a low MSE of 0.18 and MAE of 1.15. These outcomes confirm CatBoost's effectiveness and accuracy in challenging health data settings. Furthermore, real-time prediction is made possible by the incorporation of an intuitive web interface, which makes this system extremely useful in clinical, research, and health policy contexts. Future improvements are anticipated to significantly increase model performance and usefulness in personalised healthcare, such as the addition of real-time sensor data, environmental variables, and mental health indicators.

11. FUTURE ENHANCEMENT:

Future improvements to this project could concentrate on incorporating dynamic behavioural trends and real-time health data to further increase the precision and dependability of life expectancy prediction. Vital signs including heart rate, sleep quality, physical activity, and stress levels can be continuously monitored with wearable device data (e.g., smartwatches and fitness trackers), providing deeper

and more immediate health information. A fuller insight of individual health profiles can also be obtained by combining genetic factors, nutrition data, and mental health indicators. Because they have a big impact on long-term health outcomes, environmental factors such as the air quality index (AQI), climate, and pollution exposure should also be taken into account. Furthermore, prediction models for various populations can be improved by using geographical data to comprehend regional healthcare availability and lifestyle variations. To help healthcare professionals and policymakers better understand the model's predictions, sophisticated methods like explainable AI (XAI) can be implemented. With these improvements, the system will be more flexible, comprehensive, and customised for practical application in contemporary healthcare settings.

12. REFERENCES:

- [1] L.-L. Liu, H.-D. Yin, T. Xiao, L. Huang, and Y.-M. Cheng, "Dynamic prediction of landslide life expectancy using ensemble system incorporating classical prediction models and machine learning," *Geosci. Frontiers*, vol. 15, no. 2, Mar. 2024, Art. no. 101758.
- [2] N. Ali, D. Srivastava, A. Tiwari, A. Pandey, A. K. Pandey, and A. Sahu, "Predicting life expectancy of hepatitis B patients using machine learning," in *Proc. IEEE Int. Conf. Distrib. Comput. Electr. Circuits Electron. (ICDCECE)*, 2022, pp. 1–4.
- [3] N. S. Pisal, S. Abdul-Rahman, M. Hanafiah, and S. I. Kamarudin, "Prediction of life expectancy for Asian population using machine learning algorithms," *Malaysian J.*, vol. 7, no. 2, pp. 1150–1161, 2022.
- [4] M. Beeksmas, S. Verberne, A. Van Den Bosch, E. Das, I. Hendrickx, and S. Groenewoud, "Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 1–15, Dec. 2019.
- [5] S. S. Meshram, "Comparative analysis of life expectancy between developed and developing countries using machine learning," in *Proc. IEEE Bombay Sect. Signature Conf. (IBSSC)*, Dec. 2020, pp. 6–10.
- [6] Z. Guo, Y. Que, and X. Li, "Life expectancy prediction based on particle swarm optimization support vector machine regression," in *Proc. IEEE 3rd Int. Conf. Electr. Eng., Big Data Algorithms (EEBDA)*, Feb. 2024, pp. 629–635.
- [7] R. Roy, R. Tanwar, and V. Singh, "Evaluating models for better life expectancy prediction," in *Proc. Comput. Intell. Data Anal.* Singapore: Springer, Sep. 2022, pp. 389–404.
- [8] V. Bali, D. Aggarwal, S. Singh, and A. Shukla, "Life expectancy: Prediction & analysis using ML," in *Proc. 9th Int. Conf. Rel., INFOCOM Technol.*

- Optim., Trends Future Directions (ICRITO), Sep. 2021, pp. 1–8.
- [9] A. Lakshmanarao, A. Srisaila, T. S. R. Kiran, G. Lalitha, and K. V. Kumar, “Life expectancy prediction through analysis of immunization and HDI factors using machine learning regression algorithms,” *Int. J. Online Biomed. Eng.*, vol. 18, no. 13, pp. 73–83, Oct. 2022.
- [10] B. A. Lipesa, “Machine learning based prediction of life expectancy,” Ph.D. dissertation, Inst. Math. Sci., Strathmore Univ., Nairobi, Kenya, 2022.
- [11] M. Ayuso, J. M. Bravo, and R. Holzmann, “Getting life expectancy estimates right for pension policy: Period versus cohort approach,” *J. Pension Econ. Finance*, vol. 20, no. 2, pp. 212–231, 2021.
- [12] O. G. Aje, S. B. Akanni, I. A. Abdualazeez, R. A. Ibrahim, and A. A. O. Adebayo, “Forecasting of male life expectancy in nigeria: Boxjenkins approach,” *Int. J. Develop. Math.*, vol. 1, no. 1, pp. 258–268, 2024.
- [13] T. Freeman, H. A. Gesesew, C. Bamba, E. R. J. Giugliani, J. Popay, D. Sanders, J. Macinko, C. Musolino, and F. Baum, “Why do some countries do better or worse in life expectancy relative to income? An analysis of Brazil, ethiopia, and the United States of America,” *Int. J. Equity Health*, vol. 19, no. 1, pp. 1–19, Dec. 2020.
- [14] M. Trpkova-Nestorovska and N. Levkov, “Determinants of life expectancy: Analysis of southeastern European countries,” *Knowl. Int. J.*, vol. 31, no. 1, pp. 193–198, Jun. 2019.
- [15] H. Wang, M. Naghavi, C. Allen, R. M. Barber, Z. A. Bhutta, A. Carter, D. C. Casey, F. J. Charlson, A. Z. Chen, and M. M. Coates, “Global, regional, and national life expectancy, all-cause mortality, and causespecific mortality for 249 causes of death, 1980–2015: Asystematic analysis for the global burden of disease study 2015,” *Lancet*, vol. 388, no. 10053, pp. 1459–1544, 2016.
- [16] C. L. Szwarcwald, P. R. B. D. Souza Jr., A. P. Marques, W. D. S. D. Almeida, and D. E. R. Montilla, “Inequalities in healthy life expectancy by Brazilian geographic regions: Findings from the national health survey, 2013,” *Int. J. Equity Health*, vol. 15, no. 1, pp. 1–9, Dec. 2016.
- [17] T. Girum, E. Muktar, and M. Shegaze, “Determinants of life expectancy in low and medium human development index countries,” *Med. Stud. /Studia Medyczne*, vol. 34, no. 3, pp. 218–225, 2018.
- [18] H. Lin and C. Wang, “DIGWO-N-BEATS: An evolutionary time series prediction method for situation prediction,” *Inf. Sci.*, vol. 664, Apr. 2024, Art. no. 120316.
- [19] H. Lin, C. Wang, and Q. Hao, “A novel personality detection method based on high-dimensional psycholinguistic features and improved distributed gray wolf optimizer for feature selection,” *Inf. Process. Manage.*, vol. 60, no. 2, Mar. 2023, Art. no. 103217.
- [20] Kaggle. Life Expectancy (WHO). Accessed: Jul. 30, 2020. [Online]. Available: <https://www.kaggle.com/kumarajarshi/life-expectancy-who/data>