IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Life Span Calculation VIA Food Consumption: A Comprehensive Analysis of Diet and Lifestyle Factors

Sankara rao L
Assistant professor
School of Computer Science and Engineering
Raghu Engineering Collage
Vizianagaram, Andhra Pradesh

Devaraju sai spandana
UG Scholar
School of Computer Science and Engineering
Raghu Engineering Collage
Vizianagaram, Andhra Pradesh

Bogipilli praveen kumar
UG Scholar
School of Computer Science and Engineering
Raghu Engineering Collage
Vizianagaram, Andhra Pradesh

Bandi sai saketh
UG Scholar
School of Computer Science and Engineering
Raghu Engineering Collage
Vizianagaram, Andhra Pradesh
saisakethbandi@gmail.com

Karreddula Roshan Yag
UG Scholar
Department of Computational Intelligence
Raghu Engineering Collage
Vizianagaram, Andhra Pradesh
karreddularoshanyag@gmail

Shaik Asif
UG Scholar
School of Computer Science and Engineering
Raghu Engineering Collage
Vizianagaram, Andhra Pradesh
shaikasif2811@gmail.com

***Abstract* — Life expectancy is a critical health indicator that reflects the overall well-being of individuals. In recent years, there has been growing interest in understanding the impact of lifestyle choices and dietary habits on life expectancy. This research paper presents a comprehensive analysis of predictive models for estimating life expectancy using a dataset that includes various lifestyle and dietary features. The study explores a range of machine learning approaches, including K-Nearest Neighbors (KNN), KNN with hyperparameter tuning, XGBoost, Gradient Boosting, and Linear Regression, as well as deep learning models with two different Multi Layered Perceptron (MLP)architectures. These models are trained on a diverse set of input variables, such as meals per day, exercise hours, stress levels, and pre-existing conditions. We preprocess the data, handle categorical features, and scale numerical attributes to ensure robust model performance. Our findings reveal that the XG Boost model outperforms the other machine learning approaches, achieving a root mean squared error (RMSE) of 4.84. The models highlight the significance of stress management, exercise, and balanced nutrition in influencing life expectancy. Additionally, we provide insights into feature importance, shedding light on the relative contributions of different factors. This research contributes to the field of health analytics by emphasizing the role of lifestyle choices in shaping longevity. Policymakers, healthcare professionals, and individuals can leverage these insights to promote healthier living and enhance overall quality of life.**

Keywords: Predictive Modeling, Life Expectancy, Diet, Lifestyle Factors, Machine Learning, K-Nearest Neighbors, Multi Layered Perceptron, Gradient Boosting, XGBoost, Linear Regression, Root Mean Squared Error, Feature Importance

I. INTRODUCTION

Life expectancy is a vital indicator of overall health and well-being within a population. It reflects a complex interplay between genetics, socioeconomic factors, environmental conditions, and, crucially, individual lifestyle choices. Understanding the specific ways in which dietary habits, exercise patterns, stress levels, and pre-existing health conditions impact life expectancy is key to promoting longevity and enhancing the quality of life. While the relationship between individual factors and life expectancy has been investigated, creating predictive models that integrate a wide range of lifestyle variables can provide a more comprehensive and illuminating perspective. Traditionally, statistical methods have been employed in this area; however, their ability to capture complex non-linear relationships and interactions between features can be limited. Modern machine learning approaches offer promising avenues for addressing these limitations. Gradient boosting algorithms, such as the Gradient Boosting Regressor and XGBoost, have demonstrated remarkable success in various predictive tasks due to their ability to model intricate patterns within data. This research aims to develop a robust predictive model for life expectancy by leveraging these advanced machine learning techniques, along with deep learning models. Drawing insights from a dataset containing diverse lifestyle and dietary features, we seek to answer the following questions: “To what extent can lifestyle choices and health status be used to accurately predict life

expectancy?", "Which specific factors demonstrate the strongest influence on life expectancy, as determined by machine learning models?", and "Can the insights gained from this analysis inform targeted interventions and health policies promoting longevity?" The findings of this study have the potential to empower both individuals and health professionals.

By understanding the modifiable factors that significantly contribute to life expectancy, individuals can make informed choices for healthier living. At a broader level, the results could shape public health initiatives focused on encouraging those lifestyle behaviors that have the most profound impact on longevity. The study explores a range of machine learning approaches, including KNN, KNN with hyperparameter tuning, XGBoost, Gradient Boosting, and Linear Regression, as well as deep learning models with two different MLP architectures. The first MLP model is a dense neural network with layers of 32 and 16 neurons, respectively, and the second MLP model is an improved version with three layers of 64 neurons each. These models are trained on a diverse set of input variables, such as meals per day, exercise hours, stress levels, and pre-existing conditions. We preprocess the data, handle categorical features, and scale numerical attributes to ensure robust model performance. Our findings reveal that the XG Boost model outperforms the other machine learning approaches, achieving a root mean squared error (RMSE) of 4.84, while the improved MLP model shows enhanced predictive accuracy among the deep learning approaches. The models highlight the significance of stress management, exercise, and balanced nutrition in influencing life expectancy. Additionally, we provide insights into feature importance, shedding light on the relative contributions of different factors. This research contributes to the field of health analytics by emphasizing the role of lifestyle choices in shaping longevity. Policymakers, healthcare professionals, and individuals can leverage these insights to promote healthier living and enhance overall quality of life.

II. Related Work

In our quest to unravel the intricate relationship between diet, lifestyle, and life expectancy, we delve into existing research that sheds light on this critical topic. The following papers provide valuable insights, methodologies, and gaps that inform our own study, The study "Effects of Lifestyle-Related Risk Factors on Life Expectancy: A Comprehensive Model for Early Prevention of Premature Mortality from Noncommunicable Diseases" aimed to develop a multivariable model for predicting life expectancy based on modifiable lifestyle factors [1]. The researchers utilized a cross-sectional approach with extensive data from the National Health and Nutrition Examination Survey (1999-2014). Their analysis included key predictors such as smoking, diet, physical activity, body mass index, and the presence of chronic diseases. The study's results demonstrated the profound impact of these lifestyle-related factors on life expectancy. Importantly, it highlighted the potential for lifestyle modifications to positively influence outcomes, emphasizing the value of early intervention. However, a limitation of this work was the lack of tools specifically designed for early risk identification among young adults. Our research aims to build upon this foundation by incorporating additional lifestyle features and refining risk prediction models with a focus on younger populations.

The study "Estimating Impact of Food Choices on Life Expectancy: A Modeling Study" focused on predicting how dietary choices directly impact life expectancy [2]. Employing life table methodology and extensive data from meta-analyses and the Global Burden of Disease study, researchers estimated how sustained changes in food intake could alter life expectancy. Their findings revealed that optimized diets could potentially increase life expectancy by over a decade. Specific dietary shifts associated with these gains included increased consumption of legumes, whole grains, and nuts, along with a reduction in red and processed meats. While insightful, the study lacked tools for individualized forecasting and did not fully address the uncertainties related to the time required to achieve the full benefits of dietary changes. Our research will explore similar dietary impact modeling while considering regional variations in food preferences and striving to provide more personalized recommendations. A subsequent correction to the "Estimating Impact of Food Choices on Life Expectancy: A Modeling Study" aimed to address errors identified in the original work [3]. The researchers revised their methodology to account for limitations related to time effects, individual variation within populations, and uncertainties in their initial modeling. This resulted in updated, and likely more accurate, estimates for potential life expectancy gains achievable through dietary changes. Our research will carefully incorporate these corrections and strive to further refine our own modeling approach to ensure the highest level of rigor.

While existing research provides crucial insights into the connections between lifestyle and life expectancy, gaps remain. Our work aims to address these gaps by incorporating a broader set of lifestyle features beyond diet, such as exercise patterns and stress management techniques. Furthermore, we will focus on developing tools for early risk identification specifically tailored to young adults. Recognizing the impact of regional and cultural variations on lifestyle choices, our research will consider these factors when developing recommendations. Ultimately, our goal is to provide actionable recommendations for both individuals and policymakers, empowering them to make informed choices that promote longevity and well-being. As we embark on our study, we draw inspiration from the valuable insights of previous research while charting a path toward a more informed, healthier future.

III. Methodology

In our study, we have meticulously crafted a methodology that encompasses data preparation, preprocessing, model selection, and training. Each step is critical to the integrity and success of our research. Here's an overview of our approach.

A. Data preparation

The dataset serves as the foundation of our predictive modeling. We have compiled a comprehensive dataset by referencing the World Health Organization's (WHO) extensive health databases, alongside datasets available on platforms like Kaggle and Hugging Face. Additionally, we conducted surveys to enrich our dataset with firsthand information. This multifaceted approach ensures a rich and diverse collection of data points, encompassing various aspects of diet and lifestyle.

B. Data Preprocessing

Data preprocessing is a crucial step that transforms raw data into a format amenable to analysis. We began by cleaning the data, addressing missing values, and

standardizing the format for consistency. Categorical variables were encoded to facilitate computational handling, and all numerical features were normalized to ensure they contribute equally to the model's performance. This process not only refines the data but also enhances the model's ability to learn from it effectively.

C. Models used

n this research, we have applied a variety of machine learning and deep learning models to address the complex challenge of predicting life expectancy. Our machine learning arsenal included K-Nearest Neighbors (KNN) and Linear Regression. KNN was particularly chosen for its pattern recognition capabilities in complex datasets without the need for an underlying data distribution assumption. Linear Regression acted as a baseline model, providing a comparative standard and demonstrating the potential of even simple models in predicting life expectancy. On the deep learning front, we designed two dense neural network architectures. The first model incorporated ReLU activation functions within its hidden layers, leveraging ReLU's known effectiveness in nonlinear regression problems. The second model was an improved version that expanded the network's depth with additional hidden layers and utilized a linear activation function for the output layer, enabling it to directly predict life expectancy. These deep learning models were pivotal for their ability to uncover intricate relationships between lifestyle factors and life expectancy, offering insights beyond the reach of simpler models. The nuanced understanding they provide could be instrumental in guiding individuals towards healthier lifestyle choices and informing public health policies aimed at increasing longevity.

a) K-Nearest Neighbors (KNN)

The K-Nearest Neighbors (KNN) algorithm stands out for its simplicity and adaptability in both classification and regression problems. Its central principle rests on the assumption that data points located in close proximity within a feature space tend to exhibit similar characteristics or outputs. In the context of life expectancy prediction, KNN aims to predict a new individual's life expectancy by examining the lifespans of their 'nearest neighbors' in the dataset – individuals whose recorded lifestyle features are most similar. To determine these neighbors, KNN relies on a distance metric. Commonly used distance measures include:

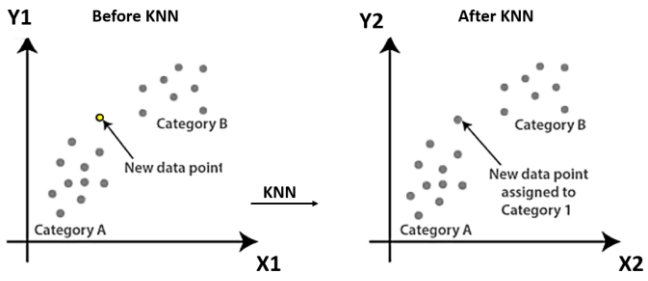

Fig. 1. Why KNN Algorithm

Euclidean distance: The intuitive 'straight-line' distance between two points in a multidimensional space. For points X and Y with coordinates $(x1, x2, \ldots xn)$ and $(y1, y2, \ldots yn)$ respectively, it's calculated as:

$$\sqrt{[(x1 - y1)^2 + (x2 - y2)^2 + \ldots + (xn - yn)^2]} \quad (1)$$

Manhattan distance: Considers distances along axes and is calculated as:

$$\sum|xi - yi| \quad (i = 1 \; to \; n) \quad (2)$$

The choice of distance metric can influence model performance. Once a metric is chosen, KNN identifies the 'K' closest data points (where 'K' is a hyperparameter you select) in the training dataset to the new, unseen query point representing an individual's lifestyle data. KNN then proceeds to calculate the predicted life expectancy for the query point. In regression tasks like yours, this is often done by simply averaging the life expectancy values of the 'K' nearest neighbors. This average becomes the prediction for the new individual. Importantly, the value of 'K' plays a significant role in the model's performance. Too small of a 'K' can make the model susceptible to noise and overfitting, while too large of a value might lead to the inclusion of dissimilar neighbors and less accurate predictions. [3]

It's crucial to remember that before applying KNN, feature scaling or standardization is often necessary. This ensures that no single feature dominates the distance calculations simply due to its scale. Moreover, despite its simplicity, KNN can become computationally expensive with larger datasets, as distances need to be calculated to every point in the training set. KNN offers several advantages. It's easy to understand and implement, it makes no assumptions about the underlying data distribution (making it effective for nonlinear relationships), and it offers a degree of interpretability by allowing you to examine the 'neighbors' influencing a prediction. However, it's important to be aware of its sensitivity to noise and potential performance degradation in high-dimensional datasets. [4]

b) KNN with Hyperparameter Tuning

While the underlying principle of KNN – predicting based on the similarity of neighbours – remains consistent, the key difference between a standard KNN model and a KNN model with hyperparameter tuning lies in how the optimal value of 'K' is determined. In a standard KNN, you likely select a value for 'K' based on intuition or convention. This might work reasonably well, but there's no guarantee it's the most suitable 'K' for your specific dataset.

This is where KNN with hyperparameter tuning shines. Instead of relying on a fixed 'K', it utilizes a systematic, data-driven approach. It tests a range of 'K' values and evaluates their performance on a validation set. The 'K' value yielding the best results (for example, the lowest RMSE) is then selected as the optimal hyperparameter. This data-driven optimization makes a significant difference. The tuned KNN model is meticulously tailored to the patterns within your dataset. It helps mitigate the risks of overfitting (model being too sensitive to noise) or underfitting (model being too simplistic). Ultimately, this often leads to enhanced accuracy and robustness when making predictions on unseen data.

c) Linear Regression

Linear regression stands as a cornerstone of statistical modeling, establishing a relationship between a continuous dependent variable (such as life expectancy) and a set of independent variables or features (such as diet, exercise, stress, etc.). Its core principle lies in discovering the best-fitting line, or hyperplane in the case of multiple features, that captures the trend within your dataset. This relationship is expressed through a linear equation, where the predicted output is a weighted sum of the input feature values plus a bias term:

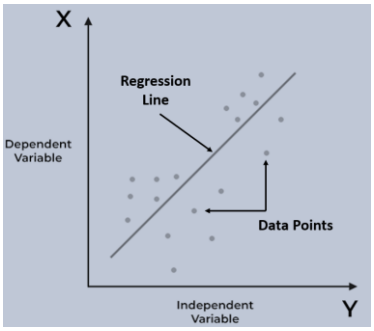

Fig. 2. Linear Regression Architecture

$$y = w1 * x1 + w2 * x2 + \cdots + wn * xn + b \quad (3)$$

- y = predicted output (life expectancy)
- $x1, x2, \ldots xn$ = feature values
- $w1, w2, \ldots wn$ = weights associated with each feature
- b = bias term

The weights associated with each feature signify their relative importance in determining the output, while the bias helps adjust the line's overall position. During training, a linear regression model learns the optimal values for these weights and the bias. This iterative learning process is guided by a loss function, with Mean Squared Error (MSE) being a common choice, measuring the average squared difference between predicted and actual values. Optimization algorithms, such as gradient descent, strive to minimize this loss function. They progressively update the model's parameters, refining the line's fit on the training data. Once trained, the linear regression model can effectively predict the target variable for new, unseen data points based on their corresponding set of features. [6]

d) XG Boost

XGBoost, an acronym for Extreme Gradient Boosting, is a sophisticated machine learning algorithm renowned for its performance in predictive modeling tasks. It operates on the principle of gradient boosting, where multiple decision trees are constructed in a sequential manner, with each tree being built to correct the residuals of the previous ones. The objective function that XGBoost optimizes is composed of two parts: the loss function, which measures how well the model's predictions align with the actual data, and the regularization term, which penalizes the complexity of the model to avoid overfitting.

The general form of the objective function that XGBoost seeks to minimize can be represented as:

$$Obj(\theta) = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i(\theta)\right) + \sum_{k=1}^{K} \Omega(f_k(\theta)) \quad (4)$$

Here, (l) denotes the loss function that assesses the discrepancy between the predicted ($\hat{y}_i$) and the actual (y_i) values, while (Ω) represents the regularization term that controls the model's complexity through parameters (θ), and (f_k) are the individual trees in the model.

XGBoost's architecture is engineered for high efficiency and scalability, which allows it to handle large datasets effectively. It incorporates several system-level optimizations, such as cache-aware access patterns for computing gradients and a column block structure for parallel learning, which significantly enhance its speed and performance.

The algorithm also features a unique regularization component, which includes both L1 (Lasso regression) and L2 (Ridge regression) regularization terms. This dual regularization approach helps in reducing overfitting and improving the model's generalization capabilities. [7, 8]

e) Gradient Boost Regressor

Gradient boosting is a machine learning technique that constructs a predictive model in the form of an ensemble of weak predictive models, typically decision trees. The method builds the model in a stage-wise fashion and generalizes them by allowing optimization of an arbitrary differentiable loss function.

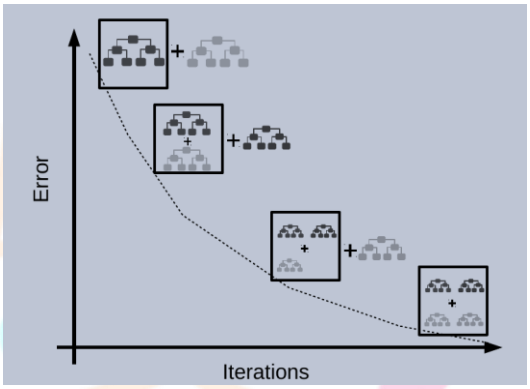

Fig. 3 Gradient Boost Regressor Algorithm.

The algorithm begins with a simple model, ($F_0(x)$), which could be the mean of the target values. It then proceeds to improve this model by iteratively adding decision trees, ($L_m(x)$), which are fit to the negative gradient of the loss function, ($L(y, F(x))$), with respect to the model's predictions, ($F(x)$). At each stage (m), the following steps are performed:

Compute the negative gradient of the loss function, known as the pseudo-residuals:

$$r_{jm} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad (5)$$

Fit a decision tree to the pseudo-residuals, r_{jm}), to obtain the leaf regions, (R_{jm}), (j=1,2,…, jm).

For each leaf region, compute the output value that will minimize the loss function:

$$\gamma_{jm} = a\gamma gmin\gamma \sum_{x \in R_{jm}} L(y_i, F_{m-1}(x_i) + \gamma)\, m \quad (6)$$

Update the model with a shrinkage factor (\nu), also known as the learning rate:

$$F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{Jm} \gamma_{jm} I\left(x \in R_{jm}\right) \quad (7)$$

This iterative process continues for (M) stages or until a stopping criterion is met, resulting in the final model:

$$F_M(x) = F_0(x) + \sum_{m=1}^{M} l \sum_{j=1}^{Jm} \gamma_{jm} I\left(x \in R_{jm}\right) \quad (8)$$

The learning rate (\nu) is a crucial hyperparameter that scales the contribution of each tree. It requires careful tuning as it controls the speed at which the model learns. Smaller values

of (\nu) typically require more iterations but can lead to a more robust model by preventing overfitting.
Gradient boosting has been successfully applied to a variety of problems, ranging from regression to classification tasks. Its effectiveness has been demonstrated in numerous machine learning competitions and real-world applications. [9, 10]

f) Dense Network

The architecture of the improved dense model, a type of multi-layer perceptron, is designed to process input data through a series of layers, each composed of nodes or "neurons". This particular model consists of an input layer followed by three hidden layers and an output layer. The input layer receives the data, which is then passed through the hidden layers, each containing 64 neurons. These neurons use the rectified linear unit (ReLU) activation function, which introduces non-linearity into the model, allowing it to learn complex patterns in the data. The ReLU activation function is chosen for its efficiency and effectiveness in deep learning models, as it helps to mitigate the vanishing gradient problem that can occur with other activation functions. The output layer consists of a single neuron with a linear activation function, making the model suitable for regression tasks where the prediction is a continuous value.

The model employs the Adam optimizer, a popular choice for training deep learning models due to its adaptive learning rate capabilities, which can lead to faster convergence. The loss function used is the mean squared error (MSE), which measures the average of the squares of the errors between predicted and actual values, making it appropriate for regression. Additionally, the model tracks the mean absolute error (MAE) as a metric, providing an understanding of the average magnitude of errors in the predictions. This architecture is effective for regression tasks due to its ability to capture complex relationships within the data, adaptively adjust learning rates, and minimize prediction errors through iterative optimization. The model's structure, with multiple layers of neurons and the use of ReLU activation, enables it to approximate non-linear functions, which is often required for accurate predictions in real-world scenarios. [11, 12]

IV. Dataset Description

This carefully designed dataset, meticulously assembled to predict life expectancy, incorporates a wide range of elements known to significantly impact overall health. With nearly 10,000 records, it provides a substantial repository of individual profiles, each outlining specific dietary preferences. These preferences include details like the frequency of meals, consumption of legumes (a vital source of plant-based protein), intake of carbohydrates such as rice, overall protein consumption, reliance on processed foods, and the amount of alcohol consumed. Additionally, crucial lifestyle factors like weekly exercise patterns, typical sleep duration, and smoking habits are carefully documented. Furthermore, the dataset delves into individual factors such as age, self-reported stress levels (categorized as low, medium, or high), and any history of chronic health conditions (such as diabetes or heart disease). This meticulously collected and preprocessed data, with life expectancy as the primary focus, will lay the groundwork for predictive models. The ultimate objective is to meticulously analyze these diverse features, seeking to understand the intricate connections between lifestyle choices and longevity, and ultimately provide valuable guidance for health-conscious decision-making.

V. Results and Discussion

Our exploration began with a KNN regressor. With a default of 5 neighbours, the initial KNN model achieved a Mean Squared Error (MSE) of 60.85. This translates to our life expectancy predictions being off, on average, by roughly 7.8 years. While this established a baseline, the initial KNN's performance suggested possible overfitting or susceptibility to noise. Seeking improvement, we employed GridSearchCV to tune the 'k' hyperparameter. This optimization led to a decrease in the KNN's MSE to 55.2, a notable improvement. This demonstrates the importance of balancing bias and variance by carefully selecting the number of neighbours. Nonetheless, we remain mindful of KNN models' potential reliance on specific data points as 'k' changes.

Table I. Models Performance

Name of the Model	MSE	RMSE
XGBoost (Revised)	23.45	4.84
Gradient Boost	30.79	5.55
KNN	60.85	7.8
KNN (Tuned)	55.2	7.43
MLP Improved	47.93	6.92
MLP	50.27	7.09
Linear Regression	48.71	6.98

Next, we introduced XGBoost, a robust tree-based ensemble method. This model achieved an MSE of 23.45. Interestingly, XGBoost outperformed our tuned KNN model. XGBoost's inherent ability to handle non-linearity and complex interactions between features likely contributed to this performance difference. Our analysis then ventured into the realm of deep learning with a Multi-Layer Perceptron (MLP) Regressor. The achieved MSE of 50.27 showcases the ability of neural networks to model intricate relationships. The MLP's performance compared to our previous models was significantly better. Further experimentation with deeper architectures or alternative activation functions could shed light on potential performance gains.

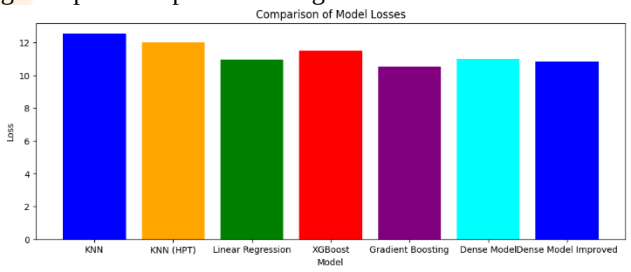

Fig. 9. GRU Accuracy vs epoch for second dataset

To investigate the impact of model complexity on performance, we experimented with a Multi-Layer Perceptron (MLP) architecture using various numbers of hidden layers. This exploration revealed that increasing the number of layers in the MLP led to an improved Mean Squared Error (MSE) of 47.93. This finding suggests that a more complex architecture was better suited to capturing the nuanced relationships within our data, achieving superior performance compared to the previous models explored.

Overall, our analysis identifies Gradient Boost and XGBoost as leading contenders for this life expectancy prediction task. While a simpler model like KNN can aid interpretability, tree-based and deep learning methods appear to capture the nuances required for higher accuracy. The optimal model choice must balance this trade-off between interpretability and potential accuracy gains. Additionally, it would be insightful to compare training times and

computational resource requirements, as these factors become important in real-world applications.

VI. Conclusion

In light of the updated results, our conclusion has been refined to reflect the performance of the various predictive models for life expectancy. The XGBoost model, with its ensemble learning approach, has demonstrated exceptional performance, achieving the lowest Mean Squared Error (MSE) of 23.45 and Root Mean Squared Error (RMSE) of 4.84. This robust model adeptly captures complex patterns and relationships within the data, which is crucial for accurate life expectancy predictions. The Gradient Boost model also showcased impressive results, with an MSE of 30.79 and an RMSE of 5.55, indicating its strength in handling non-linear interactions and diverse data features. However, it is the XGBoost model that stands out for its superior predictive power and efficiency, making it the most effective tool in our study for this task. The success of these models, particularly XGBoost, can be attributed to their sophisticated algorithms that are capable of iterative learning and feature importance ranking, which are essential for nuanced data interpretation. The deep learning models, such as the improved Multi-Layer Perceptron (MLP), also performed well, with an MSE of 47.93 and an RMSE of 6.92, underscoring the potential of neural networks in complex predictive tasks.

As we conclude, it is evident that the XGBoost model's balance of accuracy and complexity makes it an ideal choice for predicting life expectancy. Its ability to outperform other models confirms its suitability for applications that require a nuanced understanding of data. Looking ahead, the potential for extending this work is vast. Future research could explore the integration of additional data sources, the application of more advanced machine learning techniques, and the deployment of these models in real-world scenarios. The insights gained from this study pave the way for further advancements in the field, with the ultimate goal of enhancing the accuracy and applicability of life expectancy predictions.

VII. References

[1] Fadnes, L. T., Økland, J. M., Haaland, Ø. A., & Johansson, K. A. (2022). Estimating impact of food choices on life expectancy: A modeling study. PLoS medicine, 19(2), e1003889.

[2] Nirmalya Chowdhury and Diganta Saha. An efficient method of feature selection for text document classification. In R M K Sinha and V N Shukla, editors, Proceedings of ICSLT-O-COCOSDA - iSTRANS 2004 International Conference - Vol 1, pages 264–267. Tata McGraw-Hill Publishing Company Ltd, 2004.

[3] Kramer, O., & Kramer, O. (2013). K-nearest neighbors. Dimensionality reduction with unsupervised nearest neighbors, 13-23.

[4] Laloë, T. (2008). A k-nearest neighbor approach for functional regression. Statistics & probability letters, 78(10), 1189-1193.

[5] Gundapu, Sunil & Mamidi, Radhika. (2021). Multichannel LSTM-CNN for Telugu Technical Domain Identification.

[6] Naga Sudha, D., & Madhavee Latha, Y. (2021). Semi Supervised Multi Text Classifications for Telugu Documents. Turkish Journal of Computer and Mathematics Education, 12(12), 644–648

[7] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

[8] Developers, X. (2021). XGBoost documentation.

[9] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54, 1937-1967.

[10] Li, C. (2016). A gentle introduction to gradient boosting. URL: http://www. ccs. neu. edu/home/vip/teach/MLcourse/4_ boosting/slides/gradient_boosting. pdf, 59.

[11] Riedmiller, M., & Lernen, A. (2014). Multi layer perceptron. Machine Learning Lab Special Lecture, University of Freiburg, 24.

[12] Popescu, M. C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, 8(7), 579-588.