UNIVERSITY OF NEW HAVEN

TAGLIATELA COLLEGE OF ENGINNEERING

DEPARTMENT OF COMPUTER SCIENCE

CSCI 6401 SECTION 01: DATA MINING

ACADEMIC PAPER

Project Title: **Predicting Life Expectancy: Insights from Health, Economic, and Social Factors across top two developing and developed nations**

Submitted to

Dr. Shivanjali Khare

Submitted by

Rishitha Rani Pakam – rpaka1@unh.newhaven.edu

Narasimha Reddy Padire – npadi1@unh.newhaven.edu

Lakshmi Reddy Bhavanam – lbhav2@unh.newhaven.edu

December 9, 2024

Table Of Contents

Abstract:

Life expectancy, an important measure of public health, reflects the average lifespan of a population and plays a crucial role in shaping health policies and resource allocation. This study analyzes the health, economic, and social factors influencing life expectancy in two developed nations (USA and Japan) and two developing nations (India and Nigeria) using a publicly available Kaggle dataset. Data preprocessing involved cleaning, handling missing values, addressing outliers, and scaling. Using analytical methods like K-Means Clustering, Multiple Linear Regression, Ridge Regression and Random Forest Regression, we identified GDP, BMI, total expenditure, and schooling as the key factors influencing life expectancy. The results revealed that developed nations exhibit higher life expectancy due to better socio-economic conditions, while developing nations faced challenges from limited healthcare and economic resources. These findings highlight the need for focused health and education policies in developing countries to reduce disparities, improve life expectancy, and promote global well-being.

Introduction:

In today's rapidly evolving world, predicting life expectancy is essential as it provides crucial insights into the health and well-being of populations. Life expectancy reflects the quality of healthcare, social and economic conditions, and lifestyle factors within a country. As nations face challenges in healthcare, economic development, and social inequality, understanding the factors that influence life expectancy is key to shaping effective public health policies and resource distribution. This is especially important in developing countries, where there are large gaps in life expectancy.

Although life expectancy has generally increased due to advancements in healthcare, technology, and social conditions, significant gaps still exist between developed and developing nations. Understanding these disparities is essential for addressing global health challenges. As countries work to improve their public health systems while facing growing social and economic inequalities, studying the key factors that affect life expectancy provides useful insights to help guide targeted actions and improve global health.

This study aims to identify and analyze the most significant factors impacting life expectancy, comparing their effects in the top two developed nations (USA and Japan) and two developing nations (India and Nigeria). Key factors, such as GDP, total expenditure, BMI, schooling, and other social determinants, are examined using advanced machine learning techniques to predict life expectancy patterns.

The findings are valuable for public health, healthcare policy, and social planning; they offer practical insights to help guide actions and make decisions. By identifying these factors, we can improve healthcare systems, reduce health disparities, and shape policies that promote global well-being. This research helps develop better strategies for improving health outcomes in both developed and developing countries.

Related Work:

1. Life Expectancy: Prediction & Analysis using ML. This research analyzed the factors influencing life expectancy, identified random forest as best-performing model, and highlighted the impact of adult mortality, HIV/AIDS, schooling, and BMI.
2. A Precision Public Health Study on the Divergence of Life Expectancies Over Time in United States Counties. This study examined the divergence in life expectancy, education, and income at the county level in the United States over 34 years highlighting the need for focused public health efforts.
3. Analysis on Top Factors Affecting Life Expectancy in Bangladesh. This research identified thinness among adolescents, schooling, and income as key factors affecting life expectancy in Bangladesh
4. Predicting Life Expectancy using Machine Learning Approach through Linear Regression and Decision Tree Classification Techniques. This study found that the linear regression model provided a more accurate prediction of life expectancy, with 96% accuracy compared to the decision tree model's 92%.
5. Exploring Socioeconomic Influences on Life Expectancy through Machine Learning Ensemble Regression Techniques. This study showed that XGB (Extreme Gradient Boosting) performed best in predicting life expectancy, and using ensemble methods like voting and stacking improved the results, with voting ensemble achieving an R2 of 96.7%.

These research papers on life expectancy prediction have explored various factors, including health, economic, and social influences, using machine learning models like linear regression, decision trees and ensemble regression. Most of these studies focused on specific countries or factors to understand their impact on life expectancy. Our proposed study is similar in approach but aims to compare life expectancy predictions by considering social, health, and economic factors across the top two developed and developing nations. By offering a comparative perspective and using advanced machine learning techniques, our study will improve prediction accuracy and help policymakers address health disparities more effectively.

The proposed method:

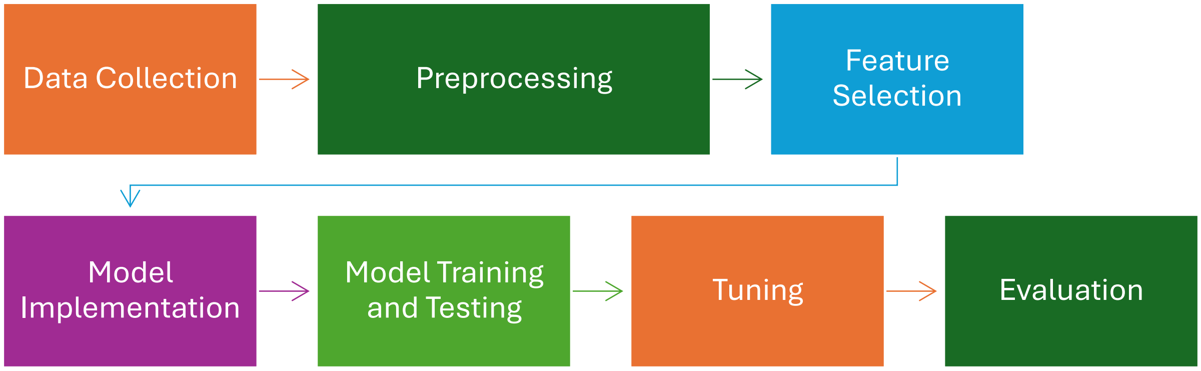


Figure: Methodology for Predicting Life Expectancy

We followed a systematic approach, as outlined in the figure, to predict life expectancy based on health, social, and economic factors. The process involved several steps: data collection, cleaning, feature selection, model application, training, testing, tuning, and evaluation. This methodology helped us to analyze the data effectively and build a reliable model for accurate predictions.

Data Collection:

For this research, we used the publicly available Life Expectancy (World Health Organization) 2024 dataset from Kaggle. It includes 22 attributes and 2938 records from 193 countries. The dataset contains health, economic, and social factors that influence life expectancy, such as Country, Year, Life expectancy, Adult Mortality, GDP, Alcohol consumption, and Schooling. The data consists of both numerical and categorical variables, with 2 object-type attributes, 11 decimal-type attributes, and 9 integer-type attributes.

Preprocessing:

After collecting the data, we moved on to preprocessing. This included cleaning the dataset by filling the **missing values** with the median and handling **outliers**. We used **Winsorization** to handle outliers by adjusting them to a specific range. We also scaled the numerical features with MinMaxScaler to ensure that all features contributed equally to the model. These steps helped enhance the data quality, making it suitable for analysis and model training.

Feature Selection:

We selected key features such as GDP, BMI, alcohol consumption, adult mortality, schooling, and healthcare, as these factors significantly influence life expectancy.

Model Implementation:

The following data mining techniques are applied in our research for predicting life expectancy:

1. K-means Clustering
2. Multiple Linear Regression
3. Ridge Regression
4. Random Forest Regression

K-means Clustering: We used K-means Clustering to group countries into developed and developing categories based on key features like GDP, total expenditure, adult mortality, and schooling to analyze life expectancy differences.

Multiple Linear Regression: We implemented Multiple Linear Regression to model the relationship between life expectancy and factors such as GDP, alcohol consumption, schooling, and adult mortality to identify their influence on life expectancy.

Ridge Regression: We implemented Ridge Regression to address multicollinearity and overfitting in the Multiple Linear Regression model by adding a penalty term to the size of coefficients(L2 Regularization), which improved stability and generalization.

Random Forest Regression: We implemented Random Forest Regression, using multiple decision trees to enhance prediction accuracy and provide more reliable life expectancy predictions.

Model Training and Testing:

We trained the models using **80%** of the data, allowing them to learn the patterns and relationships between the features and life expectancy. After training, the remaining **20%** of the data was used for testing to evaluate the accuracy and performance of the models on unseen data.

Tuning:

We fine-tuned the hyperparameters of each model to enhance their performance. For K-means Clustering, we used **GridSearchCV** and **RandomizedSearchCV** to find the optimal number of clusters and other parameters. In Multiple Linear Regression, we applied Ridge Regression to adjust the regularization strength, improving stability and reducing overfitting. For Random Forest Regression, we fine-tuned the number of trees, maximum depth, and minimum samples per leaf using **GridSearchCV** to improve prediction accuracy. These adjustments helped the models to achieve optimal performance in predicting life expectancy.

Evaluation:

We evaluated all the models using specific evaluation metrics. For K-means Clustering, we used **Silhouette Score, Davies-Bouldin Score,** and **Inertia** to assess the clustering quality. Multiple Linear Regression and Ridge Regression were evaluated based on **Mean Absolute Error (MAE), Mean Squared Error (MSE),** and **R-squared (R^2)** to measure prediction accuracy. Random Forest Regression was also evaluated using MAE, MSE, and R^2, along with **cross-validation**, to assess its prediction reliability. These metrics helped determine the best-performing model for predicting life expectancy.

The experimental results:

We evaluated the performance of our models and fine-tuned their hyperparameters to achieve optimal results:

Data Mining Technique	Hyperparameter	Values Used	Best Values Found (GridsearchCV)	Best Values Found (RandomSearchCV)
K-Means Clustering	➢ n_clusters	➢ different k values	➢ 10	➢ 15
	➢ init	➢ k-means++, random	➢ random	➢ random
	➢ n_init	➢ [10,20]	➢ 10	➢ 15
	➢ max_iter	➢ [300, 500]	➢ 300	➢ 300
	➢ tol	➢ [1e-3, 1e-4,1e-5]	➢ 0.0001	➢ 1e-5
	➢ random_statt	➢ [0,10,20,30, 40,50]	➢ 50	➢ 0

To evaluate the performance of K-Means clustering, we compared key metrics before and after tuning hyperparameters

Metric	Before Tuning	After GridSearchCV	After RandomSearchCV
Silhouette Score	0.465	0.627	0.515
Davies-Bouldin Score	0.936	0.488	0.524
Inertia	1628.816	0.707	0.334

Below is the comparison of K-Means clustering performance metrics, including accuracy, precision, and F1 score, before and after hyperparameter tuning:

Metric	Value (Before Tuning)	Value (After Tuning)
Accuracy	100%	100%
Precision	100%	100%
F1 Score	100%	100%

We tested various hyperparameter values during tuning and identified the best ones, which are listed below for each regression model:

Data Mining Technique	Hyperparameters	Values Used	Best Values Found
Multiple Linear Regression	Not Applicable (Baseline Model)	Not Applicable (Baseline Model)	Not Applicable (Baseline Model)
Ridge Regression	ridge_alpha	[1e-5, 1e-4, 1e-3, 0.01, 0.1, 1, 10, 100, 1000, 10000]	1
Random Forest Regression	➢ n_estimators ➢ max_depth ➢ min_samples_split ➢ min_samples_leaf	➢ [100, 200, 300] ➢ [5, 10, 15] ➢ [2, 5, 10] ➢ [1, 2, 4]	➢ 100 ➢ 5 ➢ 2 ➢ 1

Below is the comparison of performance metrics for the three regression models after hyperparameter tuning:

Metric	Multiple Linear Regression	Ridge Regression (After Tuning)	Random Forest Regression (After Tuning)
Mean Squared Error (MSE)	11.841	3.01	0.243
Mean Absolute Error (MAE)	2.90	1.66	0.347
R-squared Score	0.92(92%)	0.94(94%)	0.998(99.84%)
CV R^2 Score	0.85	0.95	0.9564

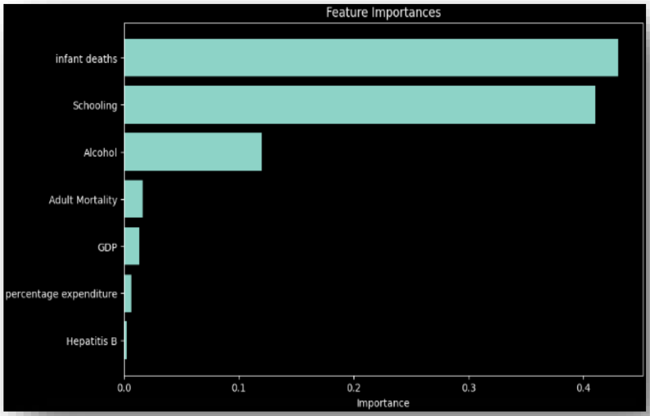
Discussion:

In the discussion of our results, we found that the **Multiple Linear Regression** (MLR) model, while providing a good baseline, struggled with multicollinearity and couldn't capture non-linear relationships, which affected its accuracy. **Ridge Regression** addressed multicollinearity by regularizing the coefficients (used L2 regularization), but it still couldn't fully handle the complex interactions between features. On the other hand, **Random Forest Regression** outperformed both models by capturing non-linear relationships and interactions between features, especially in developing countries where factors like **Infant Deaths** had a stronger impact on life expectancy. This shows that Random Forest is better suited for handling complex data and providing more accurate predictions compared to simpler models.

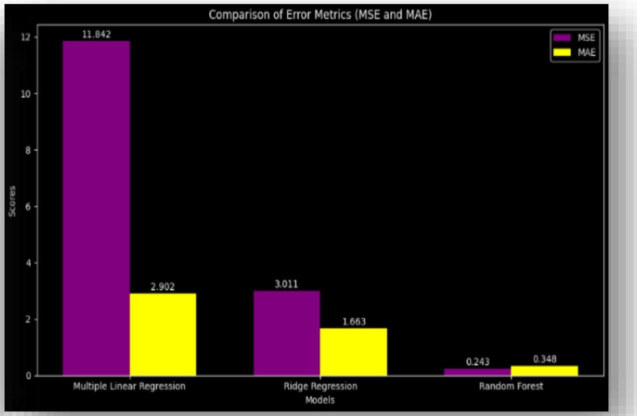
Additionally, **K-Means clustering** provided valuable insights by grouping countries based on similar health, economic, and social features, offering a clearer understanding of the global distribution of life expectancy. While the clustering model achieved perfect accuracy, precision, and F1 score, its performance benefited significantly from hyperparameter tuning, which improved cluster separation and the overall quality of the groupings.

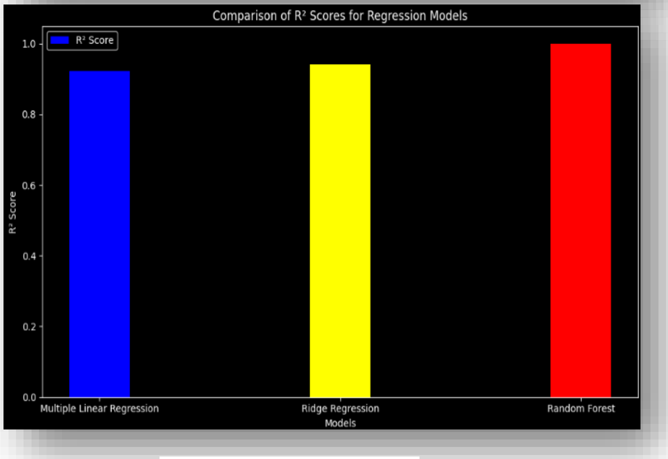
We also included a feature importance bar chart and a learning curve for the Random Forest model. The feature importance chart highlights the contributions of various inputs to the predictions, while the learning curve demonstrates that the model is well-balanced and not overfitting.

Feature Importance

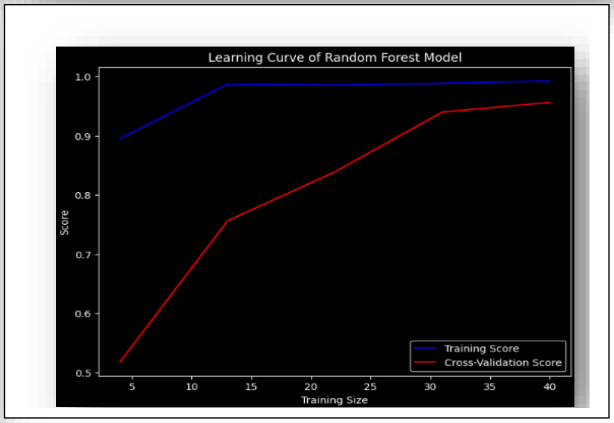


MSE AND MAE Comparison





R^2 Value Comparison



Learning Curve

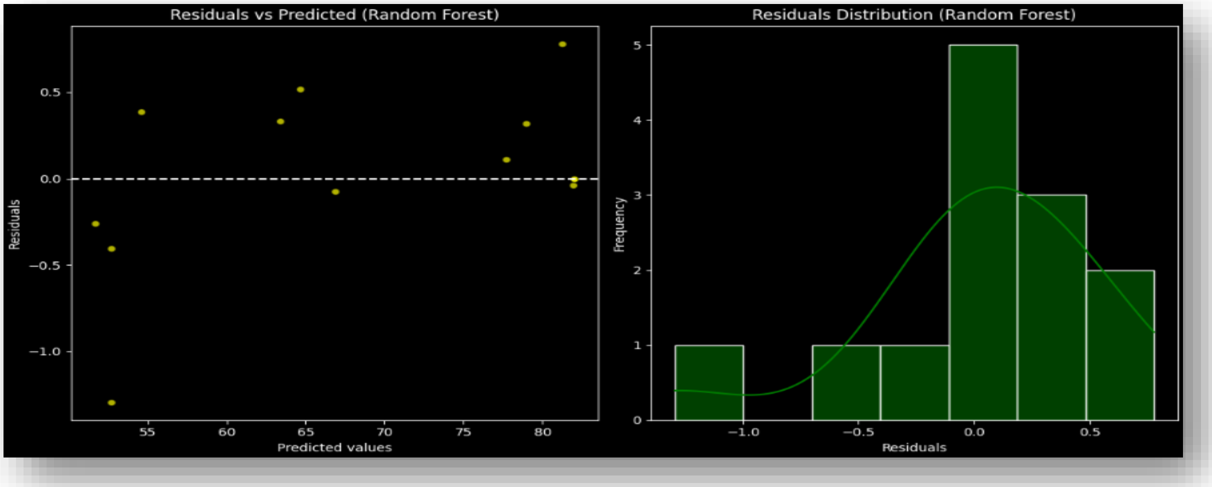


Figure: Residuals vs Predicted Distributions

The residuals vs. predicted plot for the Random Forest model was generated to evaluate its performance. The plot shows the residuals (differences between actual and predicted values) on the y-axis and the predicted values on the x-axis. A random scatter of points around the zero line indicates that the model captures the data patterns accurately, with no obvious biases or systematic errors.

The following figure displays four-line charts that show trends in life expectancy:

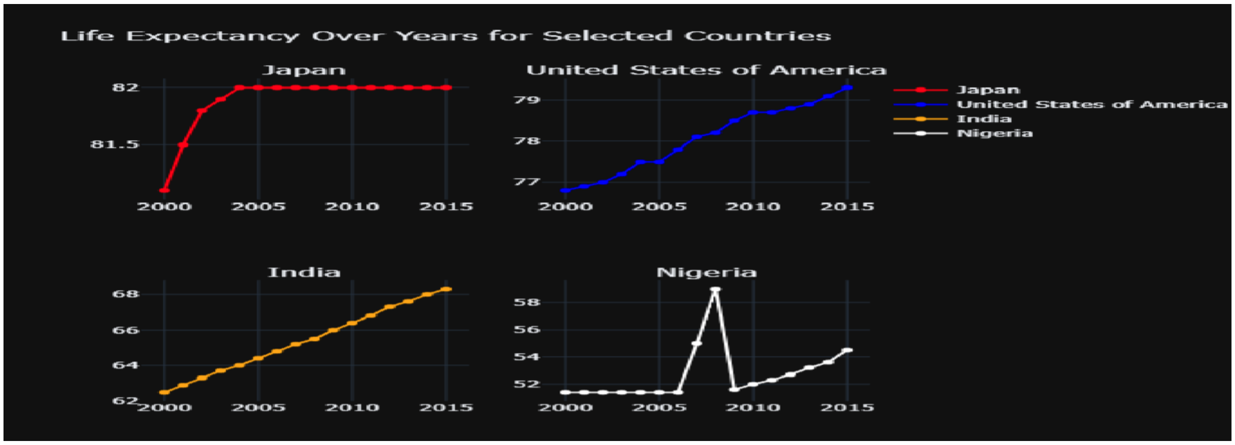
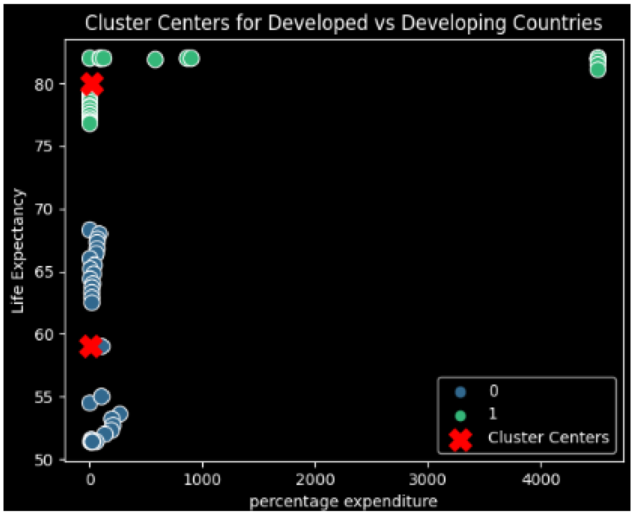
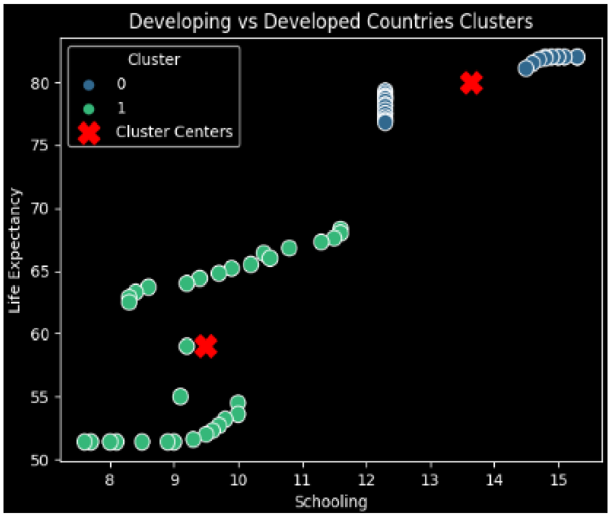


Figure: Life Expectancy Trends in Developing and Developed Countries

From these results we observed that Japan has seen a steady increase in life expectancy, reaching one of the highest levels among the four countries. The United States has shown a gradual increase, but at a slower rate than Japan. India has improved significantly in recent years, though its life expectancy is still lower than Japan and the USA. Nigeria's life expectancy fluctuated, with a sharp decline around 2005, followed by a gradual rise.





The above two plots illustrate the clustering of countries based on life expectancy, percentage expenditure, and schooling. In the first graph, life expectancy is plotted against percentage expenditure, while the second graph shows it against schooling. The cluster centers are distinctly marked, and the clusters are well-separated, highlighting clear distinctions between developed and developing countries.

Conclusion:

Our research successfully identified key factors affecting life expectancy, focusing on two developed and developing countries. By analyzing health, economic, and social variables, we found that GDP, schooling, and healthcare play a crucial role in life expectancy. Developed countries like Japan and the USA showed strong correlations with these factors, while countries like India and Nigeria were more influenced by factors like infant mortality and healthcare access. The Random Forest model, with an R-squared of 0.998, performed the best in predicting life expectancy, while the Multiple Linear Regression model had a lower R-squared value of 0.923, it still provided reliable predictions, making it a useful backup. K-Means clustering effectively grouped countries into developed and developing categories, offering valuable insights for improving health outcomes globally.

Future Work:

In the future, this research could expand the dataset to include more diverse socio-economic indicators, which would improve prediction accuracy. Exploring advanced machine learning techniques like neural networks could further enhance model performance. Additionally, incorporating temporal data to examine life expectancy trends over time would provide valuable insights into the effectiveness of health policies and interventions.

Appendix for link to the GitHub repository:

https://github.com/rishitharani24/Data-Mining---Team-Synergy

References:

[1] https://www.kaggle.com/datasets/sonialikhan/life-expectancy-who-2024

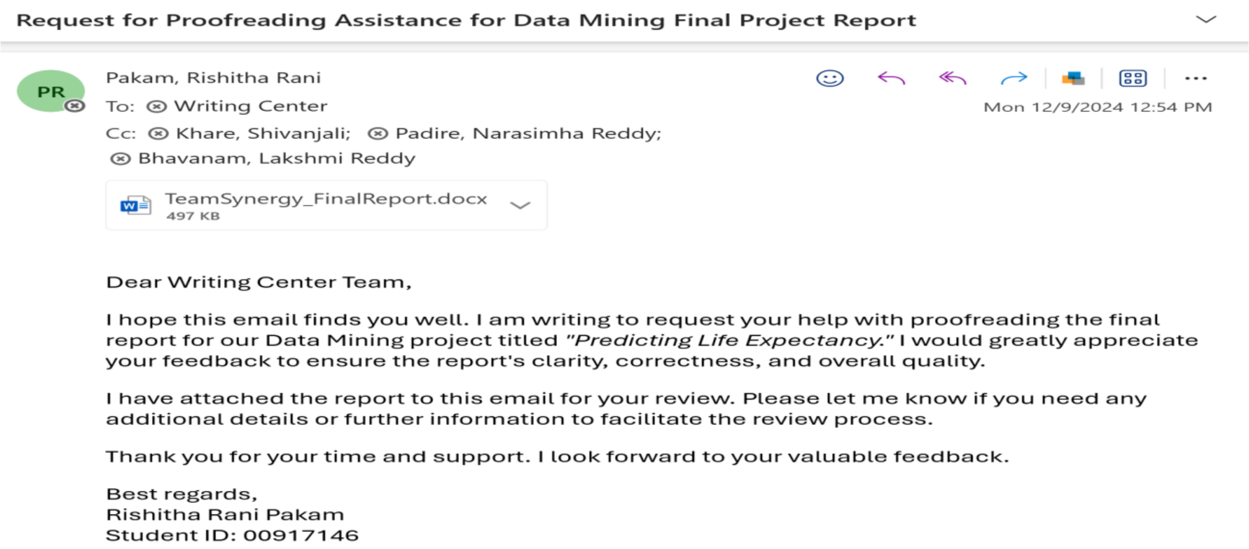
[2] https://www.kaggle.com/code/narasimhareddypadire/optimization-phase6-teamsynergy/edit

[3] https://ieeexplore.ieee.org/document/9579594

[4] https://www.kaggle.com/code/pkdarabi/forecasting-life-expectancy-by-regression-models

[5] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9596123

Proofreading with an email to Writing Center:



Request for Proofreading Assistance for Data Mining Final Project Report

Pakam, Rishitha Rani

To: Writing Center

Cc: Khare, Shivanjali; Padire, Narasimha Reddy; Bhavanam, Lakshmi Reddy

Mon 12/9/2024 12:54 PM

TeamSynergy_FinalReport.docx
497 KB

Dear Writing Center Team,

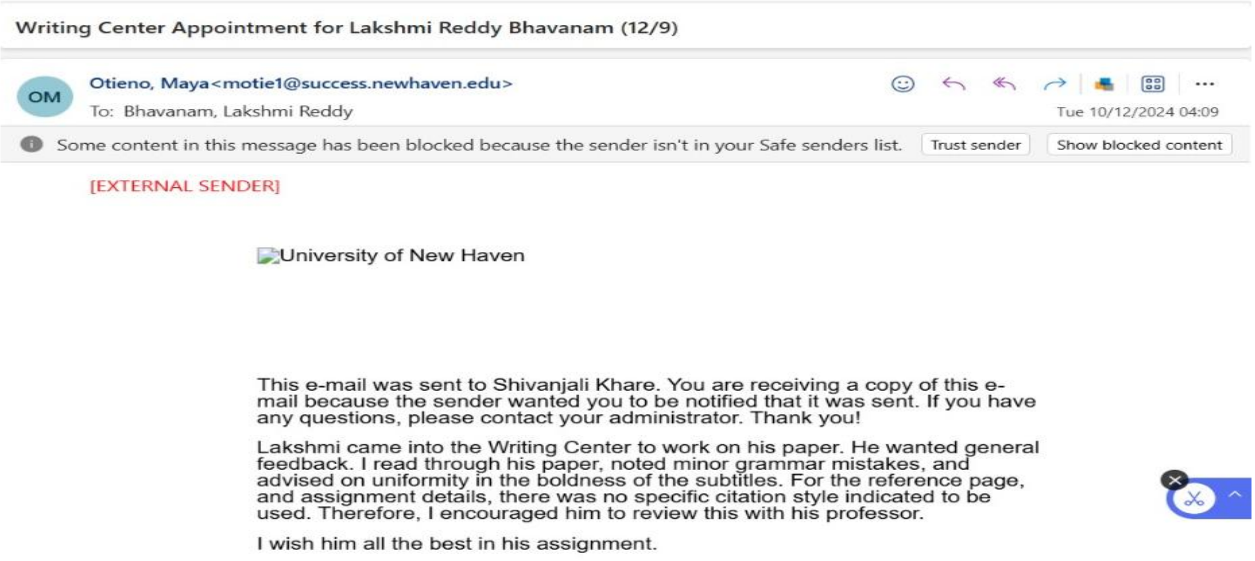
I hope this email finds you well. I am writing to request your help with proofreading the final report for our Data Mining project titled *"Predicting Life Expectancy."* I would greatly appreciate your feedback to ensure the report's clarity, correctness, and overall quality.

I have attached the report to this email for your review. Please let me know if you need any additional details or further information to facilitate the review process.

Thank you for your time and support. I look forward to your valuable feedback.

Best regards,
Rishitha Rani Pakam
Student ID: 00917146

Proofreading with an email from Writing Center:



Writing Center Appointment for Lakshmi Reddy Bhavanam (12/9)

Otieno, Maya<motie1@success.newhaven.edu>

To: Bhavanam, Lakshmi Reddy

Tue 10/12/2024 04:09

Some content in this message has been blocked because the sender isn't in your Safe senders list.

[EXTERNAL SENDER]

University of New Haven

This e-mail was sent to Shivanjali Khare. You are receiving a copy of this e-mail because the sender wanted you to be notified that it was sent. If you have any questions, please contact your administrator. Thank you!

Lakshmi came into the Writing Center to work on his paper. He wanted general feedback. I read through his paper, noted minor grammar mistakes, and advised on uniformity in the boldness of the subtitles. For the reference page, and assignment details, there was no specific citation style indicated to be used. Therefore, I encouraged him to review this with his professor.

I wish him all the best in his assignment.