PROJECT TITLE: LIFE EXPECTANCY

DATA SET

Name

Life Expectancy Data

Link

https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who/download?datasetVersionNumber=1

Description

The data was collected from WHO and United Nations website with the help of Deeksha Russell and Duan Wang. This dataset focuses on factors influencing life expectancy in 193 countries between 2000 and 2015. It combines health, economic, and social data from sources like the World Health Organization and the United Nations. The aim is to explore the impact of various factors on life expectancy, such as immunization, mortality rates, economics, and social aspects. By analyzing this data, researchers can determine the critical variables that affect life expectancy and offer insights into healthcare expenditure, mortality rates, lifestyle choices, education, and more. This dataset is a valuable resource for understanding and improving life expectancy in different countries.

WORK DONE

In this project, an analysis of life expectancy was conducted using the dataset obtained. The data preprocessing involved imputing missing values, normalization and column encoding. Missing values were imputed by mean imputation and using Hotdeck imputation. Four different algorithms were used: Decision Tree, Random Forest, Elastic Net Regression and Multivariate Linear Regression. The decision tree model was trained twice, without pruning, and then with pruning. The Random Forest model was trained with 500 trees, and the same evaluation metrics were computed. In the case of Elastic Net Regression, optimal lambda was determined and then metrics were obtained. Also, feature selection was performed in the case of Elastic Net Regression and Multivariate Linear Regression.

LITERATURE SURVEY

In the field of life expectancy prediction using machine learning, several research papers have explored different algorithms and approaches. These studies consistently found that Random Forest is an effective model for predicting life expectancy, achieving high accuracy.
They also identified key factors affecting life expectancy, such as adult mortality, BMI, schooling, income composition of resources, and immunization coverage.

Some studies focused on specific populations, like the Asian population, while others considered a broader range of countries. It's essential to note that preprocessing steps, like handling missing data and feature selection, play a vital role in improving model performance. These findings suggest that when conducting research on life expectancy using machine learning, it's important to consider the choice of algorithm, preprocessing techniques, and the specific factors that have the most significant impact on life expectancy.

Moreover, these studies highlight the potential of machine learning to inform public health policy and decision-making by understanding and predicting life expectancy more accurately.

DATA PRE-PROCESSING

1. Handled missing values (NA) in various columns by
    a) Mean value imputation
    b) Hotdeck imputation
2. Normalized the numerical features to ensure uniform scales.
3. Checked and displayed the count of missing values in each column.
4. Encoded categorical columns like "Country," "Status," and "Year" into factor variables for use in machine learning models.
5. Split the dataset into training and testing sets using an 80/20 split ratio.

ALGORITHMS IMPLEMENTED

1. **DECISION TREE**
   A decision tree model was created using the 'rpart' library and the recursive partitioning (RPART) algorithm was used. The 'Country' variable has been excluded in the formula. The minsplit parameter is set to 4. It defines the minimum number of observations in a node that is required to split a node in the decision tree.

2. **RANDOM FOREST**
   A Random Forest model was trained on the dataset using the 'randomForest' library. The ntree parameter is set to 500. It determines the number of trees to be used in the Random Forest ensemble.

3. **ELASTIC NET REGRESSION**
   The project performed Elastic Net Regression using the 'glmnet' library. Cross-validation was used to find the optimal lambda, and coefficients for the optimal lambda were obtained. The alpha parameter is set to 0.5. This parameter controls the mix between L1 (Lasso) and L2 (Ridge) regularization. A value of 0.5 indicates an equal mix of L1 and L2 regularization, which is characteristic of Elastic Net.

4. **MULTIVARIATE LINEAR REGRESSION**
   Multivariate Linear Regression (MLR) was performed using 'lm' function. The Lasso feature selection process involves optimizing the lambda parameter through cross-validation to find the best regularization strength for Lasso regression. The code calculates R-squared, RMSE, and MSE metrics for both the MLR without feature selection and the MLR with Lasso feature selection to evaluate their respective performance. The choice of hyperparameters (such as minsplit, ntree, alpha and lambda) can significantly impact the performance of the models.

EVALUATION METRICS

The model's performance (in all the algorithms) was assessed using R-squared, Root Mean Squared Error (RMSE), and Mean Squared Error (MSE).

RESULTS AND DISCUSSION

Quantitative Findings (Mean Imputation):

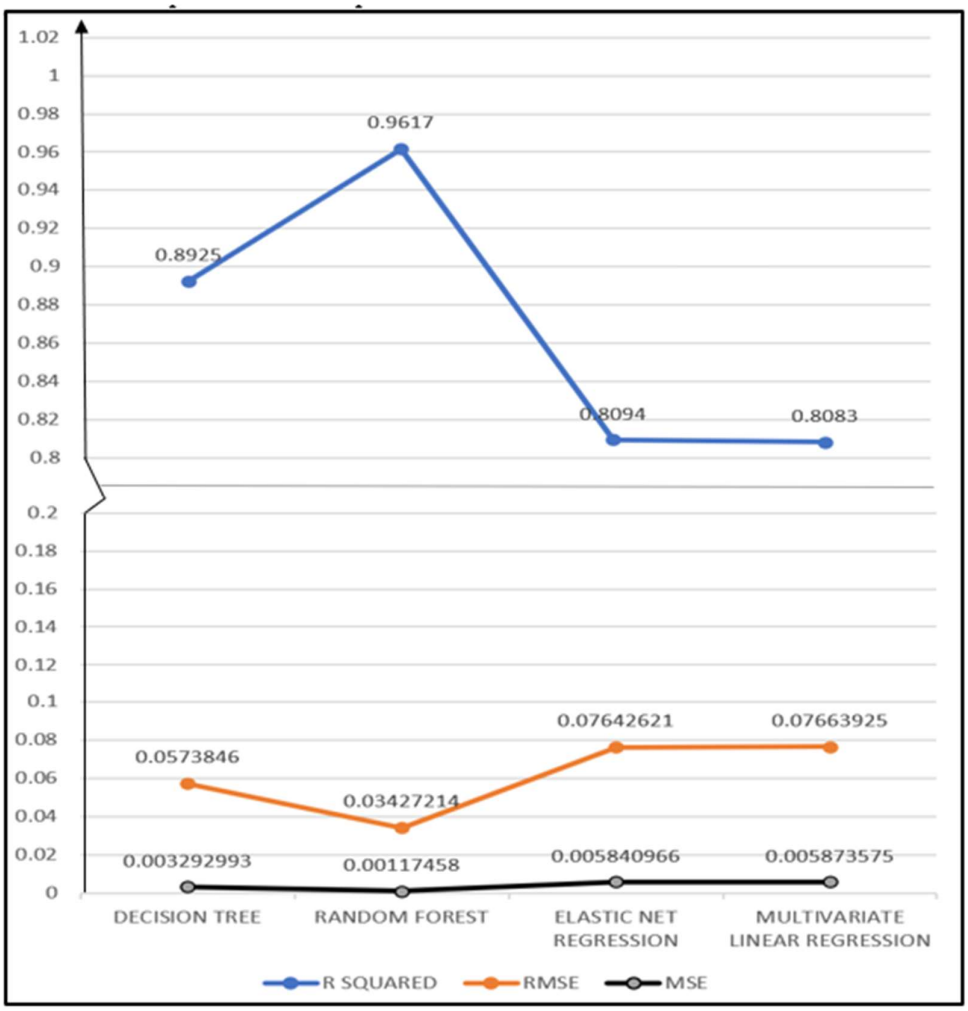
METRICS	DECISION TREE	RANDOM FOREST	ELASTIC NET REGRESSION	MULTIVARIATE LINEAR REGRESSION
R SQUARED	0.8925	0.9617	0.8094	0.8083
RMSE	0.0573846	0.03427214	0.07642621	0.07663925
MSE	0.003292993	0.00117458	0.005840966	0.005873575

Quantitative Findings (Hotdeck Imputation):

METRICS	DECISION TREE	RANDOM FOREST	ELASTIC NET REGRESSION	MULTIVARIATE LINEAR REGRESSION
R SQUARED	0.8934	0.9619	0.8056	0.8053
RMSE	0.05717086	0.03420654	0.07721803	0.07727689
MSE	0.003268508	0.001170088	0.005962624	0.005971717

Hence, "Mean imputation" consistently resulted in better R-squared and RMSE values for most of the models compared to "Hotdeck Imputation".

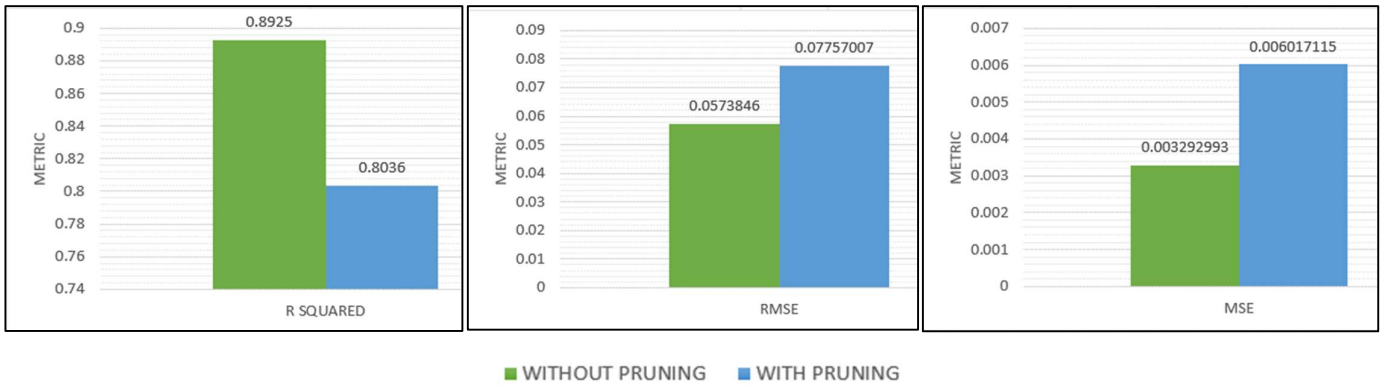
The combined plot for mean imputation can be drawn as shown:



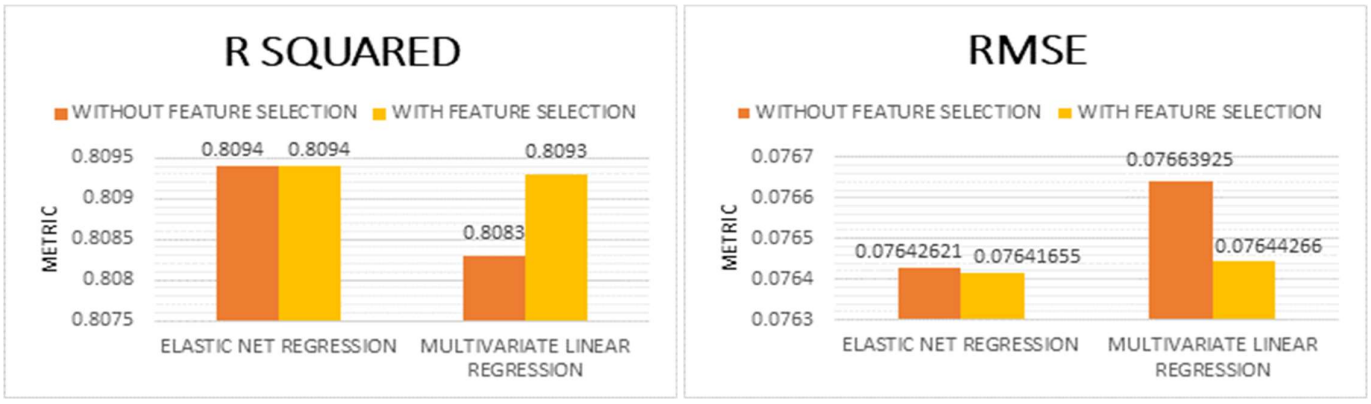
NOTE- Above is a line plot with a discontinuity due to widely varying values of the metrics.

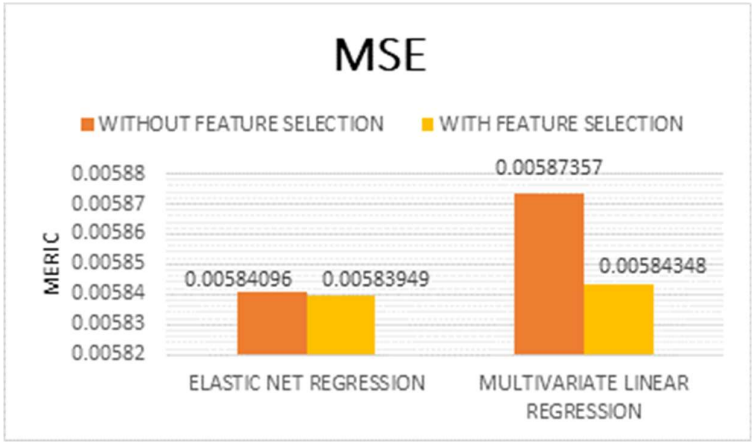
FEATURE SELECTION

In the case of both Random Forest and Decision Trees, feature selection didn't seem to have a positive impact on the model's performance. Instead, when we used feature selection, the root mean squared error (RMSE), mean squared error (MSE), and R-squared (R2) values actually worsened.
For Decision Trees, it's important to note that there is an in-built feature selection mechanism, so it's not necessary to perform additional feature selection. Instead, pruning has been performed for decision trees. On the other hand, Random Forest produced excellent results without any feature selection.
However, when it comes to Elastic Net Regression and Multivariate Linear Regression, feature selections (wrapper techniques and lasso) led to an improvement in the model's performance.
Below is the comparison graph for decision tree with and without pruning:



Below are the comparisons for elastic net regression and multivariate linear regression with and without the respective feature selection:





CONCLUSION

In this project, we conducted an analysis of life expectancy. After extensive data preprocessing, including imputing missing values and normalizing the data, we applied three different machine learning algorithms: Decision Tree, Random Forest, Elastic Net Regression and Multivariate Linear regression. The models were evaluated based on R-squared, RMSE, and MSE.

Findings reveal that the Random Forest algorithm achieved the highest R-squared value of 0.9617, indicating the best overall performance in predicting life expectancy. It also exhibited the lowest RMSE and MSE values among the three models, which signifies superior accuracy and precision.

To conclude, depending on the specific goals of the analysis, the choice of the best model may vary. Further analysis and domain knowledge are essential for making more informed decisions.

REFERENCES

[1] Xu, T., Tsui, M., & Chiu, D. M. (2022). Why Hong Kong Ranks Highest in Life Expectancy: Looking for Answers from Data Science and Social Sciences. Journal of Social Computing, 3(3), 250–261. https://doi.org/10.23919/jsc.2022.0009

[2] Ronmi, A. E., Prasad, R., & Raphael, B. A. (2023). How can artificial intelligence and data science algorithms predict life expectancy - An empirical investigation spanning 193 countries. International Journal of Information Management Data Insights, 3(1), 100168. https://doi.org/10.1016/j.jjimei.2023.100168

[3] Deshpande, R., & Uttarkar, V. (2023). Life Expectancy using Data Analytics. International Journal for Research in Applied Science and Engineering Technology, 11(4), 972–978. https://doi.org/10.22214/ijraset.2023.50140

[4] Vydehi, K. (2020). Machine learning techniques for life expectancy prediction. International Journal of Advanced Trends in Computer Science and Engineering, 9(4), 4503–4507. https://doi.org/10.30534/ijatcse/2020/45942020

[5] Meshram, S. S. (2020). Comparative Analysis of Life Expectancy between Developed and Developing Countries using Machine Learning. 2020 IEEE Bombay Section Signature Conference (IBSSC). https://doi.org/10.1109/ibssc51096.2020.9332159

[6] Pisal, N. S., Hanafiah, M., Rahman, S. A.-, & Kamarudin, S. I. (2022). PREDICTION OF LIFE EXPECTANCY FOR ASIAN POPULATION USING MACHINE LEARNING ALGORITHMS. Malaysian Journal of Computing, 2600-8238 online. https://doi.org/10.24191/mjoc.v7i2.18218

[7] Lakshmanarao, A., Srisaila, A., T, S. R. K., Lalitha, G., & K, V. K. (2022). Life Expectancy Prediction through Analysis of Immunization and HDI Factors using Machine Learning Regression Algorithms. International Journal of Online and Biomedical Engineering, 18(13), 73–83. https://doi.org/10.3991/ijoe.v18i13.33315

[8] Lipesa, B. A., Okango, E., Omolo, B., & Omondi, E. O. (2023). An application of a supervised machine learning model for predicting life expectancy. SN Applied Sciences, 5(7). https://doi.org/10.1007/s42452-023-05404-w