Predictive Modeling for Assessing Stroke Risk and Life Expectancy Using Qualitative and Quantitative Approaches

Executive Summary

Introduction

In a time where we have unprecedented access to data and advancements in computational techniques, the fields of healthcare and public health have increasingly turned to predictive modeling as a powerful tool for assessing and understanding critical health indicators. The two areas we will be exploring in this report are the prediction of stroke risk and the prediction of life expectancy. The former by using qualitative methods and the latter by using quantitative methods. This report contains a comprehensive analysis of the predictive models used, shedding light on their efficacy and highlighting key findings that help us understand what factors are important in predicting these outcomes.

Stroke is the second leading cause of death and the third leading cause of disability worldwide (Pacheco-Barrios, 2022). This demands proactive strategies for risk assessment and intervention. Concurrently, Life expectancy trends are pivotal for healthcare planning and resource allocation. Implementing the potential of predictive modeling allows for the identification of individuals at higher risk for strokes and can also provide us with nuanced insights into the complex features that contribute to life expectancy.

The qualitative analysis implements seven models, each offering a unique approach to predicting stroke risk. Evaluation metrics are used to identify the strengths and weaknesses of these models and gives us an idea of their predictive accuracy. Quantitative analysis explores twelve models in the context of predicting life expectancy. The models utilize WHO data from the years 2000-2015. Evaluation of each model using MSE is performed and when possible, the importance of the predictors being used in the respective models is determined. Furthermore, a real-world prediction for U.S. life expectancy in 2023 offers further insights into the models' applicability and accuracy.

Finally, PCR is used to investigate the prediction of income composition resources, otherwise known as “Human Development Index” (HDI). HDI is a summary measure of average achievement in key dimensions of human development: a long and healthy life, being knowledgeable and having a decent standard of living (Nations, U., 2023). The performance of the model is evaluated using test MSE. A real-world prediction for current U.S. HDI is compared to the actual current U.S. HDI. Residual error is calculated to offer another perspective on model performance.

Qualitative Analysis Summary

The qualitative analysis aimed to find the best performing models for predicting stroke risk and explore the pertinent predictors at play. This was approached using classification methods, providing a binary prediction that would result in either high risk for stroke or low risk. Seven models were considered in the analysis: Logistic Regression, LDA, QDA, KNN, Classification Trees, Bagging, and Random Forests.

Key Findings:

- Among the seven qualitative models, Random Forests demonstrated the highest accuracy when predicting strokes, with an accuracy score of 0.989.

- The Random Forest Confusion Matrix indicated 12 false positives (FP), where the model predicted a stroke when it didn't occur. The model had 0 false negatives (FN). This suggests the model favors FP over FN.

- The Random Forest model also listed importance of key predictors in the model. Those of which include: Average glucose level, age, BMI, and work-type, respectively. Average glucose level having the highest importance.

- Other robust models include: Bagging, classification tree, and KNN (K=1) with accuracy scores 0.986, 0.960, and 0.976, respectively. Bagging heavily favored FPs, with 15 FPs and 0 FNs. Decision tree had balanced FPs and FNs with 21 FPs and 22 FNs, but had a harder time evaluating true positives (TP) and true negatives (TN). KNN (K=1) favored FPs over FNs, with 26 FPs and 0 FNs. While its performance was similar to random forest, KNN (K=1) did not do as well predicting TPs and TNs.

- The less robust models were as follows: Logistic regression, QDA, LDA, KNN (K=10), KNN (K=100), with accuracy scores 0.855, 0.830, 0.862, 0.873, and 0.856, respectively.

Recommendation:

Considering performance of all the models used in predicting stroke risk, the random forest model stands out as the superior choice in terms of predicting power. It's important to note that this method does favor FPs, which in the context of healthcare can be undesirable, as this could lead to unnecessary interventions or treatments. However, while random forest did favor FPs, it also had the least number of total FPs (12), compared to all the other models. While other models may have balanced FPs and FNs better, other models sacrifice accuracy.

Quantitative Analysis Summary

The quantitative analysis aimed to find the best performing models for predicting life expectancy from WHO data, collected over the years 2000-2015. Predictors for each model were explored to assess which features were the most significant. Mean Square Error (MSE) was also calculated for the models to assess their predictive performance. Finally, models were fed inputs from current U.S. 2023 data (based on each model's respective predictors), to make actual predictions on current life expectancy in the U.S. These model predictions were then compared to the actual 2023 U.S. life expectancy and residual error was calculated to further assess each model's predictive performance. The models considered in the quantitative analysis include: Linear regression, best subset selection, forward stepwise selection, backward stepwise selection, ridge regression, lasso regression, partial least squares (PLS), decision tree, bagging, random forest, boosting, and principal component regression (PCR).

Key Findings:

- Among the 12 quantitative models, Boosting, Bagging, and Random Forest models performed the best, with test MSEs of 5.018, 3.677, and 3.564, respectively. This indicates that these models made the most accurate predictions on life expectancy. All three models indicate income composition resources, HIV/AIDS, and adult mortality as the most crucial predictors for predicting life expectancy.

- Decision Tree Model also performed well, with a test MSE of 8.448. The tree was pruned to optimal tree size of 62. This model also highlights income composition resources, HIV/AIDS, and adult mortality as the predictors of highest importance.

- Lasso, Ridge, Linear Regression, PCR, Forward Subset Selection, Backward Subset Selection, and Best Subset Selection all had similar performance, with test MSE values of 13.896, 13.843, 13.913, 13.720, 13.568, 13.490, and 13.611 respectively.

- While all models showed variation in variable importances, all models seem to highlight income composition resources as either the first or second most important predictor. The only exception is PLS, which lists under-five-deaths, country population, and alcohol as the most crucial predictors. All the subset models (forward, backward, and best) also include income composition resources as a part of the subset.

- When models were given a real-world set of inputs (current U.S. feature values), all models predicted life expectancy in U.S. 2023 with relatively good accuracy, except for PCR. Current life expectancy actual in the U.S. is 79.11 years old. Lasso Regression was the closest to the actual value, with a prediction of 79.23 years old and a residual error of -0.119. PCR was the furthest from the actual value, with a prediction of 70.53 years old and a residual error of 8.580.

Recommendation:

Considering performance of all the models used in predicting life expectancy, both the random forest and bagging models stand out as the superior choices in terms of predicting power. By training multiple base models and averaging predictions, these ensemble methods provide very robust models for complex datasets and tend to be more generalized which helps avoid the issue of overfitting. These models are great for real-world data, such as this WHO dataset, which often has a lot of outliers and noise. These models are also great for interpretability, as they also lend insight into feature importance information.

Principal Component Regression Analysis Summary

The goal of the PCR analysis was to take WHO data from the years 2000-2015 and use the features of the dataset to predict income composition resources. Test MSE is used to evaluate the models performance. The model is then used to make a prediction of income composition resources, given real-world feature inputs of current U.S. statistics.

Key Findings:

- The PCR model achieved optimal performance with 15 principal components (M=15), with a test MSE of 0.0051. This suggests that the selected features effectively capture the variance in the income composition resources variable.

- Utilizing the trained PCR model to predict income composition resources in a hypothetical situation, the predicted value was 0.879. This prediction had a residual error of 0.042, indicating only a slight deviation from the actual value of 0.921.

Appendices

- Refer to Appendix 1 for a list of references.
- Refer to Appendix 2 for the complete code used in this analysis.

Data & Approach: Qualitative Analysis

Original Dataset Overview

The dataset used in this analysis originates from Kaggle. The focus of this dataset is to be used for predicting the likelihood of an individual having a stroke based on various features. These features include: age, gender, hypertension, heart disease, marital status, work type, residence type, average glucose level, body mass index (BMI), and smoking status, with the target being whether the individual has had a stroke or not.

Observations Summary

Total Number of Observations: 5425
Number of Males in the Dataset: 2271
Number of Females in the Dataset: 3154
Number of Individuals with Hypertension: 703
Number of Individuals without Hypertension: 4722
Number of Individuals with Heart Disease: 807
Number of Individuals without Heart Disease: 4618
Number of Individuals Ever Married: 3488
Number of Individuals Not Ever Married: 1937
Number of Individuals in Each Work Type:
- Private: 2812
- Self-employed: 885
- Government Job: 646
- Child-Care: 568
- Never Worked: 514

Number of Individuals Rural: 2496
Number of Individuals Urban: 2929

Number of Individuals in Each Smoking Status Category:
- Never Smoked: 3175
- Smokes: 995
- Former Smoker: 1255

Number of Individuals Who Have Had a Stroke: 788
Number of Individuals Who Have Not Had a Stroke: 4637

Data Dictionary

- id: Unique identifier
- gender: Male = 0, Female = 1
- age: Age of the patient
- hypertension: 0 if the patient doesn't have hypertension, 1 if the patient has hypertension
- heart_disease: 0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
- ever_married: Yes = 1 and No = 0
- work_type: Never_worked = 0, Self-employed = 1, Private = 2, Govt_job = 3, children = 4
- Residence_type: Rural = 0 and Urban = 1
- avg_glucose_level: Average glucose level in blood
- bmi: Body mass index
- smoking_status: Never smoked = 0, smokes = 1, formerly smoked = 2
- stroke: 1 if the patient had a stroke or 0 if not

Data Cleaning

The initial steps of the cleaning process involved handling missing/null values and addressing the values listed as 'Unknown' in the 'smoking_status' column. The rows with missing/null values were dropped, and 'Unknown' values in the 'smoking_status' column were also removed. The 'id' column, being a unique identifier, was also dropped. Additionally, a single entry with 'Other' in the ‘gender; column was removed to simplify encoding. Following data cleaning, categorical columns were encoded and mapped to numerical values (refer to Data Dictionary).

Data Exploration

While the number of observations in the original dataset is relatively high, at 5110 observations (as reported by the dataset author). Certain variables had too little datapoints to be considered significant. In the original dataset, individuals with heart disease had 307 observations, work type 0 ‘never worked’ had 14 observations, work type 4 ‘children’ had 68 observations, and individuals that have had a stroke had 288 observations. To get more datapoints in those variables, bootstrapping technique was applied to increase observations in these variables by 500. This increased the total number of observations in the dataset to 5,425 observations, thus making it significantly more robust for modeling. Refer to the Observations Summary for a complete list of final counts for the dataset.

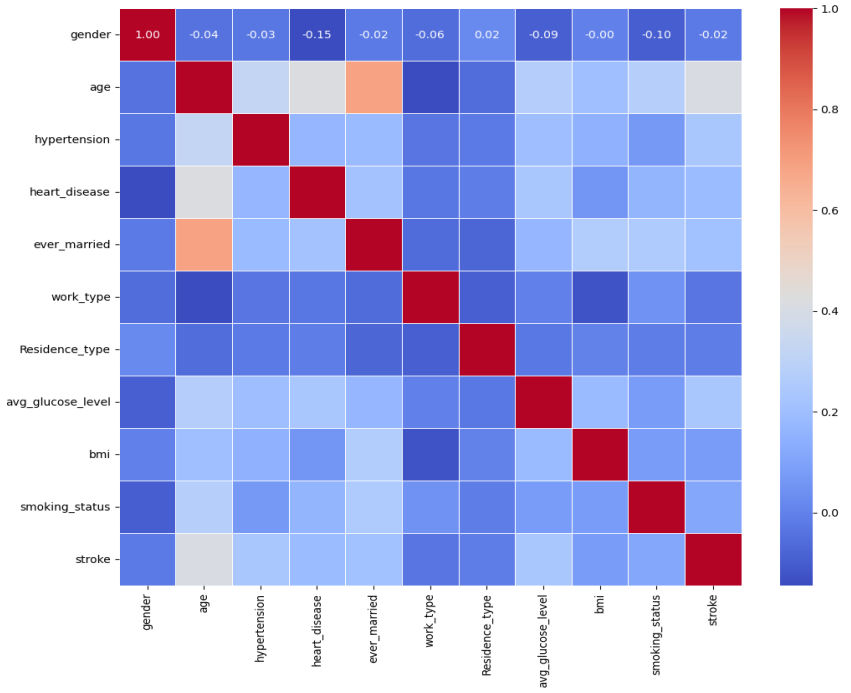
Numerical summary

	age	avg_glucose_level	bmi
count	5425.000000	5425.000000	5425.000000
mean	45.950783	112.148151	29.417954
std	22.954625	50.053917	7.262230
min	10.000000	55.120000	11.500000
25%	23.000000	78.080000	24.300000
50%	49.000000	94.040000	28.300000
75%	66.000000	125.260000	33.100000
max	82.000000	271.740000	92.000000

Numerical summary was performed on the non-categorical variables. Summary shows good range of values throughout the variables. Mean gives us an overview of central tendency and standard deviation explains variability in values among the variables.

Correlation Analysis

A correlation matrix of all variables was created to look at the relationships between variables. Positive correlations were observed between age and ever married, age and stroke, and average glucose level and stroke, with values 0.691, 0.406, and 0.239, respectively.



Heatmap shows relationship between ‘ever_married’ and ‘age’ as orange, which indicates a stronger correlation. You can also see the relationship between age and stroke, and avg_glucose_level and stroke as being less blue, indicating mild correlation.

Collinearity Evaluation

Collinearity was assessed by looking through the dataset for variables with correlations of 0.65 or greater. One pair was found, 'ever_married' and 'age'. The simple explanation for this relationship is that people tend to get married as they get older. Due to the nature of these variables, it was deemed that this correlation is reasonable.

Model Overview and Approach

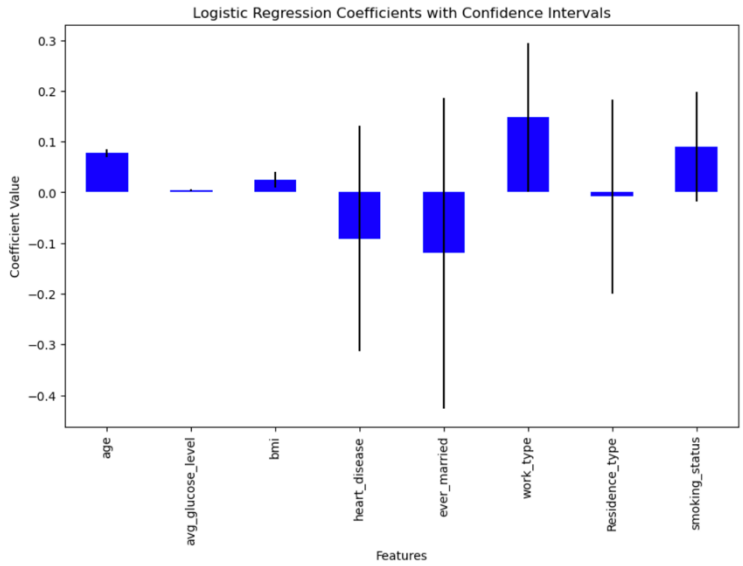
The goal with this analysis was to create models that can predict whether a patient is at high risk of having a stroke, based on the features of the dataset. To do this, the dataset was split into a training and test set. The models trained on the training set and made predictions on the test set. This analysis looks at seven different models: Logistic Regression, LDA, QDA, KNN (K=1, K=10, K=100), Classification Tree, Bagging, and Random Forest. Accuracy scores of all models were evaluated. Coefficients from Logistic Regression is examined, along with p-values. Confusion matrices were generated where applicable, which gave insight into TP, FP, TN, and FN model tendencies. A recommendation of the best performing model is given, based on the design of the models and how the models performed. Finally, taken into consideration is the context of the healthcare field, when modeling this dataset.

Detailed Findings: Qualitative Analysis

Logistic Regression Model

The linear regression model highlighted some key features, with age and avg_glucose_level having p-values close to 0, bmi having a p-value of 0.002, and work-type having a p-value of 0.049. Lesser significant variables: heart_disease, ever_married, Residence_type, and smoking_status, had p-values 0.419, 0.445, 0.935, 0.104, respectively. Coefficients of significant predictors: age, avg_glucose_level, bmi, and work_type, had the values: 0.078, 0.005, 0.025, and 0.148 respectively. These coefficients help describe the relationship between these predictors and the target variable (See plot below). Confusion matrix indicates 901 instances of TN, 124 instances of FN, 33 instances of false positive, and 27 instances of true positive.

Figure 1



The bars represent coefficient level, either positive (up) or negative (negative). Lines indicate confidence intervals. Confidence intervals including 0 generally indicate that the variable is not a significant predictor. This plot confirms age, avg_glucose_level, bmi, and work type as significant predictors in this model. Work type and age in particular have large relative coefficients.

LDA, QDA and KNN Models

Confusion matrix for LDA indicates 902 instances of TN, 118 instances of FN, 32 instances of false positive, and 33 instances of true positive, with an accuracy score of 0.862.

Confusion matrix for QDA indicates 810 instances of TN, 60 instances of FN, 124 instances of false positive, and 91 instances of true positive, with an accuracy score of 0.830.

Confusion matrix for KNN (K=1) indicates 908 instances of TN, 0 instances of FN, 26 instances of false positive, and151 instances of true positive, with an accuracy score of 0.976.

Confusion matrix for KNN (K=10) indicates 888 instances of TN, 92 instances of FN, 46 instances of false positive, and 59 instances of true positive, with an accuracy score of 0.873.

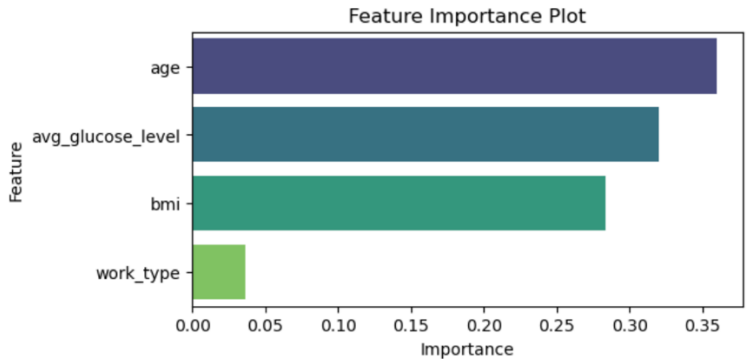
Confusion matrix for KNN (K=100) indicates 919 instances of TN, 141 instances of FN, 15 instances of false positive, and 10 instances of true positive, with an accuracy score of 0.856.

Classification Tree Model

Confusion matrix for the pre-pruned tree indicates 913 instances of TN, 22 instances of FN, 21 instances of false positive, and 129 instances of true positive, with an accuracy score of 0.976. Prior to cross-validation, tree contains 547 nodes.

Confusion matrix for the pruned tree indicates 908 instances of TN, 0 instances of FN, 26 instances of false positive, and 151 instances of true positive, with an accuracy score of 0.960. After cross-validation, optimal tree contains 367 nodes.

Figure 2

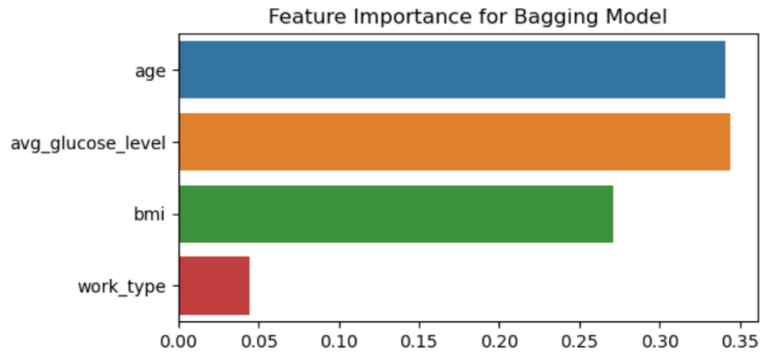


This plot takes the best tree (pruned tree, 367 nodes) and presents significant predictors and their respective importance. We see age, avg_glucose_level, bmi, and work_type come up as significant predictors in this model, with age seeming to have highest importance in predicting stroke.

Bagging Model

Confusion matrix for the bagging model indicates 919 instances of TN, 0 instances of FN, 15 instances of false positive, and 151 instances of true positive, with an accuracy score of 0.986.

Figure 3

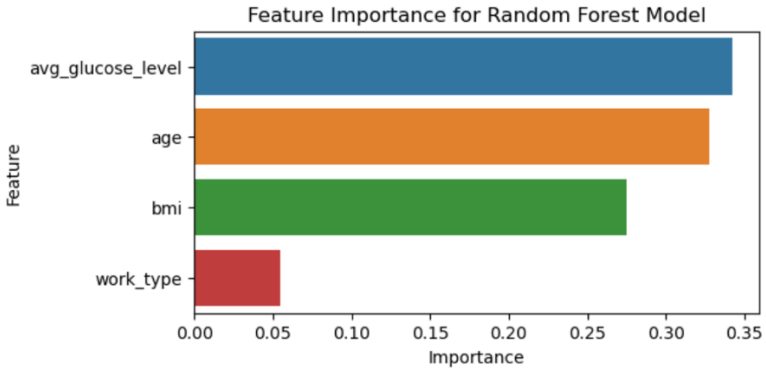


This plot presents significant features and respective importance of the bagging model. We see age and avg_glucose_level as having strong importances in this model, with avg_glucose_level having slightly higher importance than age.

Random Forest Model

Confusion matrix for the random forest model indicates 922 instances of TN, 0 instances of FN, 12 instances of false positive, and 151 instances of true positive, with an accuracy score of 0.989.

Figure 4



This plot highlights feature importances within the random forest model. We see avg_glucose_level as the most important predictor in this model, with age closely behind.

***Validity & Reliability Assessment*: Qualitative Analysis**

Model Comparison and Analysis

Random Forest, Bagging, Classification Tree and KNN (K=1) models seemed to significantly outperform the LDA, QDA, Logistic Regression, and KNN (K=10, K=100) models, in terms of accuracy. Random Forest and Bagging models involve ensembling multiple decision trees, allowing them to capture more complex relationships in the data. The combination of diverse trees helps mitigate overfitting and enhances generalization of the models. The Classification Tree model, when pruned effectively (as was performed in this analysis with cross-validation) can capture intricate patterns in the data. This provides a more flexible model when compared to linear models such as Logistic Regression or simpler models such as KNN. Another factor is Random Forest, Bagging, and Classification Tree models' ability to handle non-linear relationships. These models can capture patterns and interactions that linear models, such as logistic regression cannot. Random Forest and Bagging models are also super robust in handling outliers and noise in the data. Outliers will have less impact on the overall performance due to the nature of ensemble methods. We do however, see KNN (K=1) perform really well, with an accuracy score of 0.976, which compares with the Random Forest, Classification Tree, and Bagging models' performance. This is most likely due to the ability of KNN (K=1) to handle outliers. With K=1, only the nearest neighbor is considered. We see that as we increase K, the performance of the model decreases. While we do see good performance of the K=1 model, we can only extract very limited information from the model. Whereas with Random Forest,

Bagging, and Classification Tree models, we can extract feature importances and get an idea of how much each of the significant predictors weighs into the model.

Feature Importance

Assessing all the models, there is a clear set of predictors that seem to be crucial in predicting stroke. Logistic Regression, Bagging, Random Forest, and Classification Tree models all point to age, avg_glucose_level, bmi, and work_type as the most significant predictors. Logistic Regression, a linear model, appears to place the most weight on work_type and age, based on their coefficients. While the more complex models, Random Forest, Classification Tree, and Bagging, appear to place more significance on age and avg_glucose_level. It confirms real-world outcomes that all models have designated age as a significant predictor when it comes to predicting stroke. However, age gives very little insight into lifestyle factors that can increase stroke risk, which is where the interesting information is. This is because patients can't control aging, as this is an inevitable part of life, but patients can make lifestyle changes to reduce their risk of stroke.

Implications and Recommendation

After assessment of the predictive models and considering implications of the healthcare field with the analysis of this data, the Random Forest model is the recommendation for prediction of stroke. The Random Forest models indicates avg_glucose_level as the most significant predictor when determining stroke risk. This implies that if there's a single lifestyle change a patient can make to reduce their risk of stroke, it would be to keep their levels of sugar intake down and to perform regular exercise. These lifestyle changes can potentially increase cellular sensitivity to glucose, thus having less free glucose in the blood. This reduction in blood glucose level also has other related benefits, such as lower incidence of type II diabetes, and lower BMI, which will also reduce risk of stroke. It's important to note, that the Random Forest model did favor false positives (FPs). In healthcare this can lead to unnecessary interventions or treatments (Dresselhaus, 2002). However, while the Random Forest model did favor FPs, it also had the lowest amount of FP predictions, when compared with other models. Solidifying the Random Forest model as the best model for predicting stroke with this dataset.

Future Research and Limitations

Future research of this subject is dictated by the limitations of the dataset. Filling a more robust dataset could greatly increase the implications taken from the analysis. For example, instead of a hypertension binary variable (1 = yes 0 = no) format, a column of systolic and diastolic blood pressures could be used for analysis. This could further tighten up the analysis, such as looking at stroke risk at varying degrees of blood pressure (pre-hypertensive, moderately-hypertensive, severely-hypertensive). It would also be useful to have data regarding more comorbidities, such as diabetes, cancer, renal disease, etc. Additional lifestyle data would also be useful, such as time spent exercising per week, diet/nutrition data, etc. Addressing these limitations in the data could provide future analysis with more nuanced information on factors that predict stroke risk and offer more implications in terms of lifestyle changes individuals can make to reduce their risk of stroke.

Data & Approach: Quantitative Analysis

Original Dataset Overview

The Global Health Observatory (GHO) data repository under the World Health Organization (WHO), keeps track of the health status as well as many other related factors for all countries. These datasets are made available to public for the purpose of data analysis. This life expectancy dataset contains health factors for 193 different countries and was collected from the WHO data repository. The corresponding economic data was collected from United Nation website. Among all health-related factors only critical factors were chosen to be included in this dataset. The final dataset contains 20 predicting variables and 2938 rows. Most analysis that currently exists was done considering multiple linear regression based on datasets of one year for all the countries, whereas this analysis considers modeling of all countries over the period of 2000-2015.

Data Dictionary

- status: Developing or Developed nation (Developing = 0 and Developed = 1)
- adult_mortality: Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
- alcohol: Recorded per capita (ages 15+) consumption (in litres) of pure alcohol per year
- percentage_expenditure: Expenditure on health as a percentage of Gross Domestic Product per capita(%)
- hepatitis_b: Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
- bmi: Average Body Mass Index of the entire country population
- under_five_deaths: Number of under-five deaths per 1000 population
- polio: Pol3 immunization coverage among 1-year-olds (%)
- total_expenditure: General government expenditure on health as a percentage of total government expenditure (%)
- diphtheria: DTP3 immunization coverage among 1-year-olds (%)
- hiv_aids: Deaths per 1000 live births HIV/AIDS (0-4 years)
- country_gdp: Gross Domestic Product per capita (in USD)
- country_population: Population of the country
- thinness_5_to_19_years: Prevalence of thinness among children and adolescents for Age 5 to 19 (%)
- income_composition_resources: Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
- schooling: Number of years of Schooling
- life_expectancy: Life Expectancy in age

Data Cleaning

The first step in data cleaning was to rename columns to appropriate naming conventions for working in Python. For example, ‘Adult Mortality’ was renamed to ‘adult_mortality’.

Removing capitalization helps mitigate issues when coding down the road. Removing the space between words and utilizing an underscore instead, formats the variable in a way that's more useable in Python. After renaming, column removable was performed. 'Country' and 'Year' columns were removed, as they weren't necessary for this analysis. Infant deaths was also removed due it closely mirroring the variable under-five deaths, and they were highly correlated (0.99 correlation coefficient). The 'measles' column was also dropped because the dictionary states that the data is recorded per 1000 individuals, however, a significant number of values in the column were greater 1000. Additionally, the original data had split thinness into thinness ages 5-9 and thinness ages 10-19. The decision was made to combine this data to thinness ages 5-19, as this was easier to work with and encompassed pre-adult thinness more completely. Finally, all null and missing values were removed from the dataset. Prior to moving on to exploration, the categorical variable 'status' was encoded numerically (refer to Data Dictionary), which establishes the final working dataset for analysis.

Data Exploration

Numerical Summary

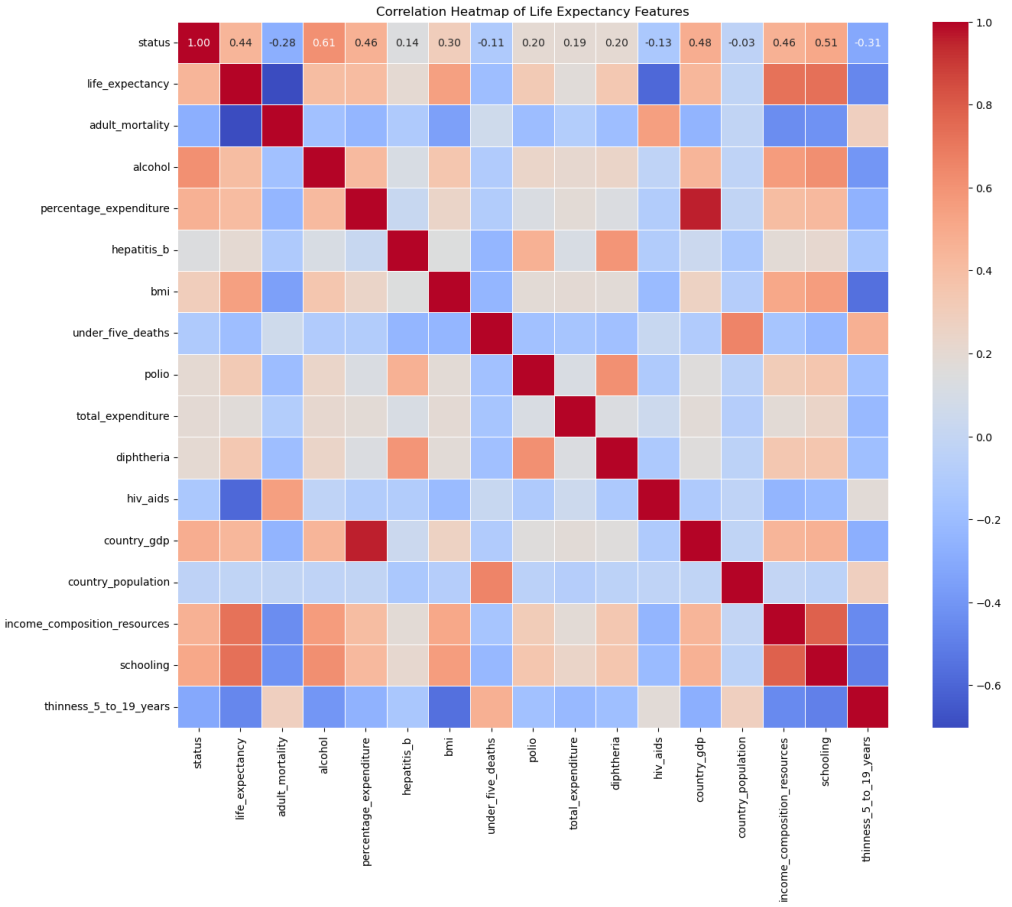
Variable	Mean	Standard Deviation	Minimum	Maximum
life_expectancy	69.3023	8.79683	44.000	89.0000
adult_mortality	168.215	125.310	1.0000	723.000
alcohol	4.5332	4.02919	0.0100	17.8700
percentage_expenditure	698.9736	1759.229	0.0000	18961.3486
hepatitis_b	79.2177	25.60466	2.0000	99.0000
bmi	38.1286	19.75425	2.0000	77.1000
under_five_deaths	44.2201	162.898	0.0000	2100.0000
polio	83.5646	22.45056	3.0000	99.0000
total_expenditure	5.9559	2.29939	0.7400	14.3900
diphtheria	84.1552	21.57919	2.0000	99.0000
hiv_aids	1.9839	6.03236	0.1000	50.6000
country_gdp	5566.0319	11475.900	1.6814	119172.7418
country_population	1.465363E+07	7.046039E+07	3.400000E+01	1.293859E+09
income_composition_resources	0.63155	0.18309	0.0000	0.9360
schooling	12.1199	2.79539	4.2000	20.7000
thinness_5_to_19_years	9.7584	9.08471	0.2000	55.4000

Numerical summary was performed on the non-categorical variables. Summary shows good range of values throughout the variables. Mean indicates an overview of central tendency and standard deviation explains variability in values among the variables.

A count of the variable 'status' was performed, to see how many developed versus non-developed nations were present in the dataset. Count shows 1407 undeveloped countries and 242 developed countries.

Correlation Analysis

A correlation matrix of all variables was created to look at the relationships between variables. Highly correlated pairs (threshold = 0.65) include:
Life Expectancy and Adult Mortality (0.703)
Life Expectancy and Income Composition Resources (0.721)
Life Expectancy and Schooling Years (0.728)
Percentage Expenditure and Country GDP (0.959)
Under-Five Deaths and Country Population (0.659)
Income Composition Resources and Schooling Years (0.785)



Heatmap of life expectancy correlations visually summarizes correlations between variables, with closer to dark red indicating strong positive correlation and closer to dark blue indicating stronger negative correlation.

Collinearity Evaluation

Collinearity was assessed by looking through the dataset for variables with correlations of 0.65 or greater. Percentage Expenditure and Country GDP, Under-Five Deaths and Country Population, and Income Composition Resources and Schooling Years were highly correlated.

However, upon removing these variables from analysis, no significant change in the models' performance was observed, so, they were kept in the dataset.

Model Overview and Approach

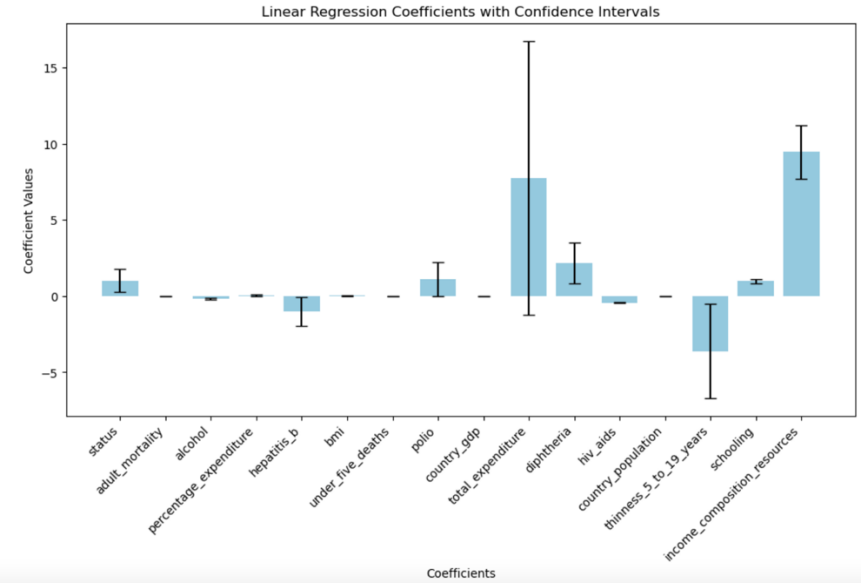
The goal with this analysis was to implement models capable of predicting life expectancy based on the features of the dataset. To do this, the dataset was split into a training and test set. The split sets were then normalized to resolve percentage columns to decimals. Then the models were trained on the training set and made predictions on the test set. This analysis looks at 12 different models and their ability to predict life expectancy. Mean Square Error (MSE), of each model is evaluated and when applicable, feature importances are also evaluated. Finally, a recommendation of the best performing model is given, based on the design of the models and how the models performed.

Detailed Findings: Quantitative Analysis

Linear Regression Model

The linear regression model highlighted some key features of the dataset. The significant features being Status, Adult Mortality, Alcohol Consumption, Hepatitis-b Immunization, BMI, Under-Five Deaths, Diphtheria Immunization, HIV/AIDS, Youth Thinness, Schooling Years, and Income Composition Resources, indicated by p-values >0.05. The strongest coefficient amongst significant features being Income Composition Resources at 9.4645. The variables Total Expenditure, Percentage Expenditure, Polio Vaccination, Country GDP, and Country Population, all had p-values <0.05 and were thus considered not significant in this model.

Figure 5



The bars represent coefficient level, either positive (up) or negative (negative). Lines indicate confidence intervals. Confidence intervals including 0 generally indicate that the

variable is not a significant predictor. This plot confirms Status, Adult Mortality, Alcohol Consumption, Hepatitis-b Immunization, BMI, Under-Five Deaths, Diphtheria Immunization, HIV/AIDS, Youth Thinness, Schooling Years, and Income Composition Resource as significant predictors in this model. Income Composition Resource in particular, has a large relative coefficient compared to the other features.

Best, Forward, Backward Subset Selection Models

The Best Subset Selection model selected variables 'adult_mortality', 'percentage_expenditure', 'bmi', 'under_five_deaths', 'polio', 'diphtheria', 'hiv_aids', 'schooling', and 'income_composition_resources' as the best set of predictors. Best Subset Selection model MSE was 13.610.

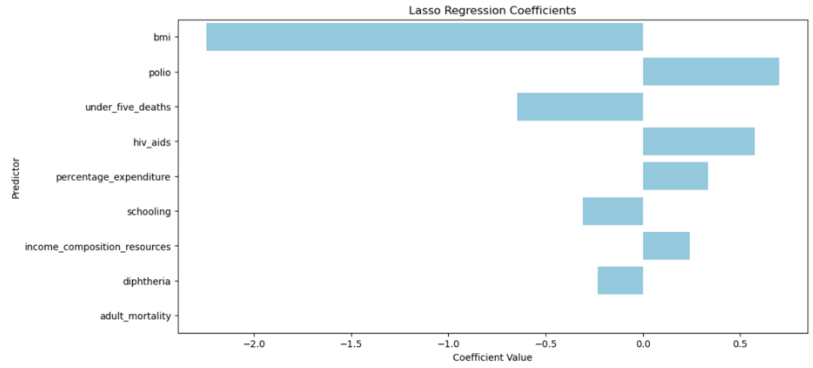
The Forward Selection model selected variables 'income_composition_resources', 'hiv_aids', 'adult_mortality', 'bmi', 'diphtheria', 'country_gdp', 'polio', and 'percentage_expenditure' as the best predictors. Forward Selection MSE was 13.568.

The Backward Selection model selected variables 'country_population', 'diphtheria', 'status', 'hiv_aids', 'under_five_deaths', 'income_composition_resources', 'adult_mortality', 'polio', 'schooling', 'country_gdp', and 'bmi' as the best set of predictors. Backward Selection MSE was 13.639.

LASSO Regression

The optimal regularization parameter (alpha) identified through cross-validation was 0.01. The Lasso Regression model achieved a Test Mean Squared Error (MSE) of 13.90. The Lasso Regression assigns different weights (coefficients) to each predictor, indicating their impact on the target variable. Income composition resources demonstrated the highest positive coefficient (2.747), suggesting a strong positive correlation with life expectancy. On the other hand, BMI exhibited the strongest negative coefficient, implying an inverse relationship with life expectancy.

Figure 6

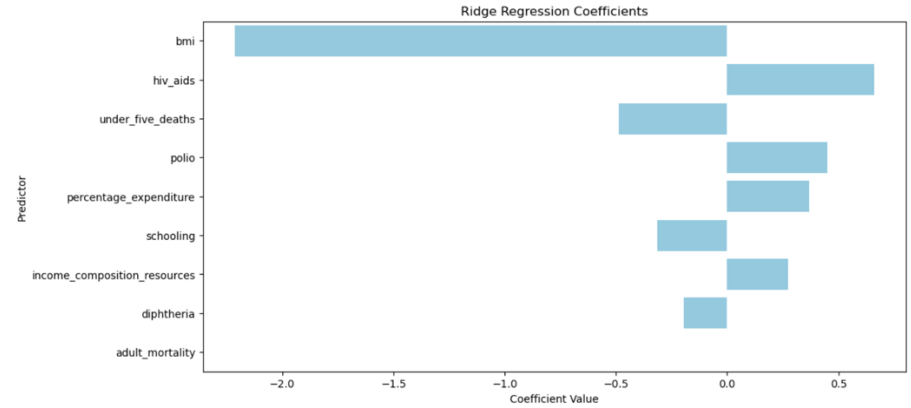


BMI is noted as the highest weight coefficient, having an inverse relationship with life expectancy. A strong positive relationship between polio and hiv_aids with life expectancy, is also observed in the plot. Interestingly, under_five_deaths is inversely correlated with life expectancy in this model.

RIDGE Regression

Ridge Regressions was trained using cross-validation and incorporating feature scaling to enhance predictive performance. The model identified the optimal regularization strength (alpha) through cross-validation, which was 100.0. The mean squared error (MSE) on the test set was calculated to be 13.84. The Ridge Regression model assigns different weights (coefficients) to each predictor variable. Top negative coefficients include BMI (-2.21) and Under-Five Deaths (-0.49). Top positive coefficients include Polio Immunization (0.45) and Percentage Expenditure (0.37). Interestingly, HIV/AIDS is also is a top positive predictor at 0.66.

Figure 7

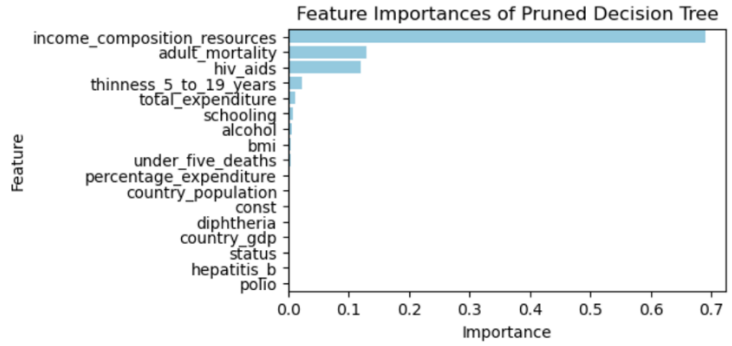


We see BMI as the strongest predictor in the Ridge Regression model, with an inverse relationship to life expectancy. We also see a strong positive relationship between life expectancy and HIV/AIDS, and Polio Immunization.

Decision Tree Model

The initial tree, prior to pruning, yielded an MSE of 8.63 on the test set. The unpruned Decision Tree had 2,351 nodes and a depth of 29, indicating a very complex structure. Using grid search and cross-validation, the optimal tree size was found to be 62 nodes, reducing the complexity of the model. Pruning resulted in a more interpretable tree with 123 nodes and a depth of 13. The pruned Decision Tree highlighted key predictors contributing to life expectancy. The most influential feature was "income_composition_resources" (0.690 importance), followed by "adult_mortality" and "hiv_aids". The pruned model achieved a slightly improved MSE of 8.45 on the test. Noteworthy splits in the tree include the impact of "income_composition_resources," "hiv_aids," and "adult_mortality" on life expectancy.

Figure 8



The importance plot shows how heavily Income Composition Resources is weighted in the model compared to other predictors. We also see strong importance in Adult Mortality and HIV/AIDS in this model.

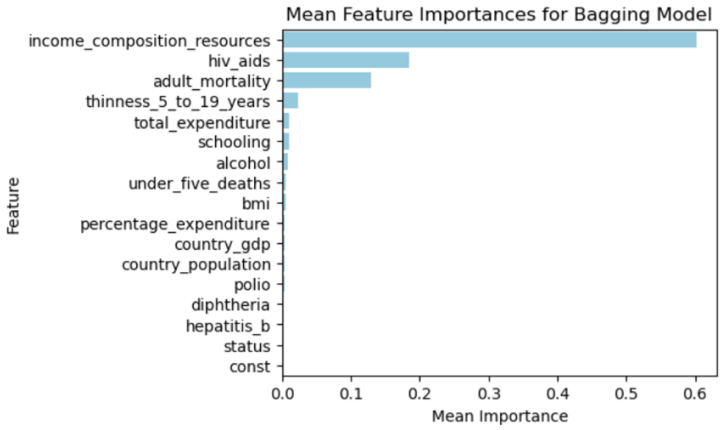
Random Forest Model

The Random Forest model exhibited strong predictive performance, as evidenced by an MSE of 3.56. This indicates that the model's predictions were, on average, very close to the actual values in the test dataset. Income Composition Resources surfaced as the most critical predictor, contributing significantly to the model's predictions. This emphasizes the substantial impact of a country's income distribution, highlighting the socioeconomic aspect of life expectancy. This was followed by HIV/AIDS and Adult Mortality.

Bagging Model

The bagging model, utilized a decision tree regressor as the base estimator, this demonstrated strong predictive performance. The MSE for the test set is recorded at 3.68. Feature importances were extracted from each base estimator in the bagging model and averaged to provide a comprehensive view of variable importance across all iterations.

Figure 9

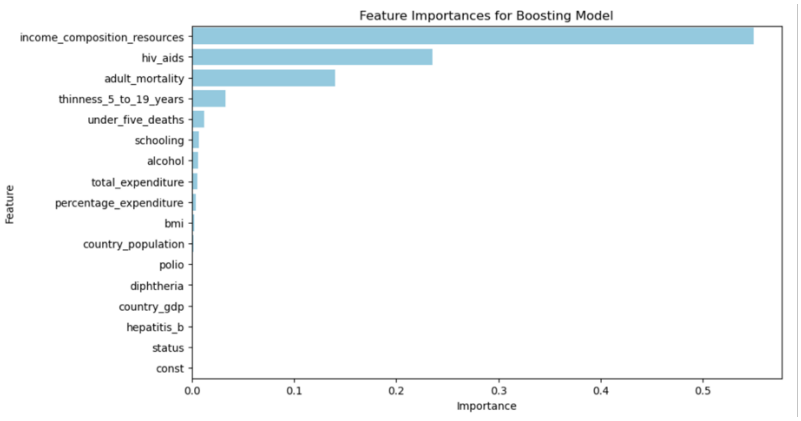


Once again, Income Composition Resources is a strong predictor in the model. We observe HIV/AIDS and Adult Mortality as having fairly strong significance in the model as well.

Boosting Model

The Boosting model, trained with 100 estimators and a learning rate of 0.1, exhibited an MSE of 5.0175.

Figure 10



Income Composition Resources again is a very strong predictor of life expectancy in the model. HIV/AIDS and Adult Mortality are once again seen as strong predictors as well.

Partial Least Squares Model (PLS)

The PLS model, with three components, yields an MSE of 13.8476. The analysis of PLS loadings provided insight into the features contribution to the model's predictions.

Component 1: Features with the highest loadings in Component 1 include: "under_five_deaths," "country_population," and "alcohol." These features negatively contribute to life expectancy.

Component 2: "hiv_aids," "adult_mortality," and "total_expenditure" have the highest loadings in Component 2. Higher values in these features are associated with decreased life expectancy.

Component 3: "under_five_deaths," "country_population," and "alcohol" have impactful loadings in Component 3. Higher values in these features are associated with reduced life expectancy.

It's also noted that "under_five_deaths," "country_population," and "alcohol" have the highest absolute mean loadings, indicating their strong influence on the model's predictions.

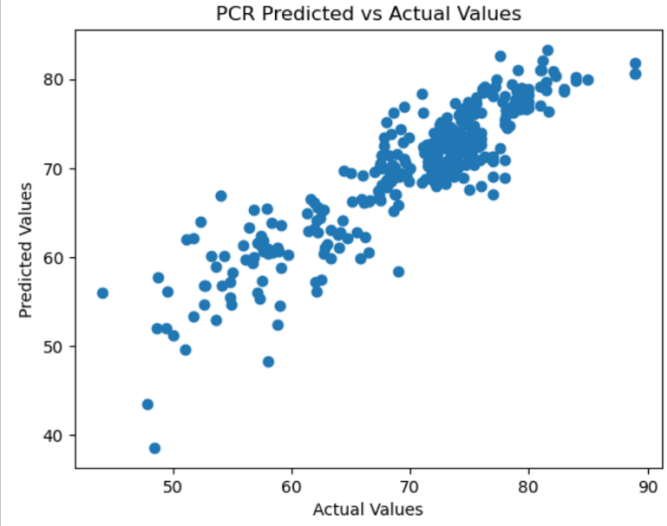
	Absolute_Mean
under_five_deaths	0.353222
country_population	0.334550
alcohol	0.308225
hiv_aids	0.282012
adult_mortality	0.253004
total_expenditure	0.237601
status	0.234559
schooling	0.195338
country_gdp	0.186077
percentage_expenditure	0.183444
bmi	0.177468
polio	0.163861
income_composition_resources	0.163083
diphtheria	0.159750
hepatitis_b	0.144973
thinness_5_to_19_years	0.133261
const	0.000000

Table highlights the weight of each predictor's influence on the model's ability to predict life expectancy.

Principal Components Regression (PCR)

The PCR analysis involved creating a predictive model using principal components. Cross-validation was utilized to tune the number of principal components through a range of values. The best-performing PCR model, as determined by cross-validation, utilized 14 principal components. The calculated MSE for the PCR model was 13.72.

Figure 11



Overall, the points appear to adhere closely to a straight diagonal line, which indicates the model performed well at making predictions when compared to actual values. The model seems to perform better at higher values than at lower values, which can be seen by the clustering of points at higher value levels and by what seems to be more outliers at the lower end of values.

Using the Models to make a Real-world Prediction

Linear Regression

The Linear Regression model predicted a life expectancy of 79.24 for the USA in 2023. The residual error is minimal at -0.13, indicating a close fit to the actual life expectancy.

Best Subset Selection

The Best Subset Selection model predicted a life expectancy of 76.64, with a residual error of 2.47. This model, while accurate, shows a slight underestimation compared to the Linear Regression.

Lasso Regression

The Lasso Regression model predicted a life expectancy of 79.23, closely aligned with Linear Regression. The residual error is negligible at -0.12.

Ridge Regression

Ridge Regression yielded a higher prediction of 81.01, indicating a more optimistic outlook on life expectancy. The residual error is -1.90, suggesting a slight overestimation.

Decision Tree, Random Forest, and Bagging

These ensemble methods provided predictions ranging from 75.39 to 76.3, with residual errors of 3.72, 3.62, and 2.81, respectively. These models seem to underestimate life expectancy.

Boosting

The Boosting model produced a prediction of 78.29, with a small residual error of 0.82. This suggests a balanced prediction close to the actual life expectancy.

Partial Least Squares (PLS)

The PLS model predicted a higher life expectancy of 81.37, with a residual error of -2.26. This model tends to overestimate life expectancy compared to other models.

Principal Components Regression (PCR)

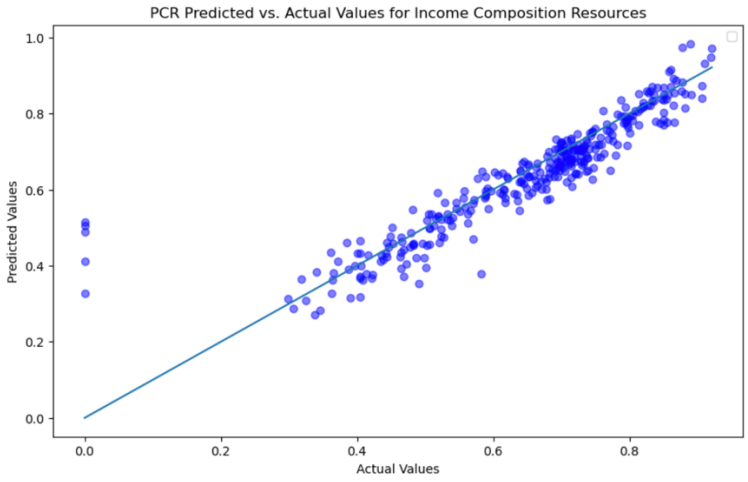
PCR provided the lowest prediction at 70.53, indicating a significant deviation from the actual life expectancy, with a large residual error of 8.58.

Utilizing PCR Model to Predict Income Composition Resources (HDI)

The PCR model, with an optimal number of principal components (M) equal to 15, demonstrated strong predictive accuracy on the test set, with an MSE of 0.005. The optimal number of principal components (M = 15) was determined through cross-validation. This suggests that the first 15 principal components capture the majority of the variability in the original predictors.

The model was used to make a real-world prediction of Income Composition Resources, using 2023 U.S. data inputs. The model predicted Income Composition Resources for the U.S. in 2023 at 0.879. This result indicates a strong alignment with the actual value of 0.921. The residual error was found to be 0.0421.

Figure 12



The model shows some outliers at the low-end of "actual value", but performs well throughout, which is indicated by the tight adherence of points to the diagonal line.

***Validity & Reliability Assessment*: Qualitative Analysis**

Model Comparison and Analysis

In assessing the validity and reliability of the models, Mean Squared Error (MSE) served as the primary evaluation of performance for these models. Comparing Linear Regression, Best Subset Selection, Forward Subset Selection, Backward Subset Selection, Lasso, Ridge, PLS, and PCR to Decision Tree, Bagging, and Boosting models indicated lower MSE values in the latter group of models. This indicates superior performance in predicting life expectancy compared to the linear regression-based models. This is most likely since the decision tree models handle complex relationships in the data better than the linear-based models. The decision tree models are also more robust at handling outliers and noise.

Adding to the assessment, PCR model was also utilized specifically for predicting Income Composition Resources. The model demonstrated robust validity and reliability with an MSE of 0.005. Optimal number of principal components (M = 15), determined by cross-validation, ensured that the model effectively captured the variability in the original predictors. Real-world prediction for the U.S. in 2023 aligned strongly with the actual value, with a low residual error of 0.0421.

Feature Importance

Across the models, "Income Composition Resources", "HIV/AIDS", and "Adult Mortality" consistently emerged as highly influential predictors. The agreement amongst the diverse modeling techniques reinforced the reliability of these features in predicting life expectancy. Specifically in the Decision Tree, Bagging, and Boosting models, "Income

Composition Resources" consistently stands out, suggesting its robust influence on life expectancy.

The additional PCR analysis of Income Composition Resources further showcases the importance of Income Composition Resources across various models, by highlighting how closely related this variable is the other health metric features of the dataset.

Implications and Recommendations

Research in the field of public health has made clear conclusions that indicate how substantial an impact even basic access to publicly funded care can make in improving life expectancy (Galvani-Townsend, 2022). The consistent identification of certain features as influential across various models provides actionable insights. Policymakers and public health professionals can utilize these findings to prioritize interventions that address factors such as income composition resources (Human Development Index HDI), HIV/AIDS prevention, and adult mortality to improve life expectancy. Socioeconomic interventions such as investments in healthcare infrastructure, education, and disease prevention programs may yield positive outcomes. While not the most significant predictors in these models, immunization also seemed to play a large part in life expectancy, so, vaccinating children is another way to improve overall life expectancy.

Based on assessment of each model's performance in predicting life expectancy, the recommended model is the Random Forest Model, with the lowest MSE of 3.56. This model is great at handling complex relationships between variables, handling outliers, and handling noise, which are all common issues found in public health/healthcare datasets.

Future Research and Limitations

For future research, it would be interesting to follow the trend in each variable up to a future date, such as 2027. This dataset contains data from 2000-2015, which could be enough data to expand on the trend to a future year. These future data points could be used as inputs in the models to predict future life expectancy. The inputs could be changed based on varying hypothetical scenarios as well. Such as, a hypothetical situation in which a country invests in their HDI, what kind of life expectancy could we expect out of such an investment?
It would also be interesting to run predictions from other countries to see how each model performs with different countries' inputs.

Limitations include the reliability of available data, multicollinearity, and assumptions inherent in the models. Sensitivity analyses, external validation with new datasets, and robustness checks could address some of these limitations. Moreover, understanding the contextual applicability of findings is crucial, as the models can perform differently in different settings.

Appendix 1

References

Centers for Disease Control and Prevention. (2023, June 13). *FastStats - Immunization*. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/fastats/immunize.htm

Dresselhaus, T. R., Luck, J., & Peabody, J. W. (2002). The ethical problem of false positives: a prospective evaluation of physician reporting in the medical record. *Journal of medical ethics*, *28*(5), 291–294. https://doi.org/10.1136/jme.28.5.291

Fedesoriano. (2020). Stroke Prediction Dataset. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Galvani-Townsend, S., Martinez, I., & Pandey, A. (2022). Is life expectancy higher in countries and territories with publicly funded health care? Global analysis of health care access and the social determinants of health. *Journal of global health*, *12*, 04091. https://doi.org/10.7189/jogh.12.04091

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An introduction to statistical learning. *Springer Texts in Statistics*. https://doi.org/10.1007/978-3-031-38747-0

KumarRajarshi. (2017). Life Expectancy (WHO) Dataset. Kaggle. https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who?select=Life+Expectancy+Data.csv

Nations, U. (2023, November 28). *Human development index*. Human Development Reports. https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

Pacheco-Barrios, K., Giannoni-Luza, S., Navarro-Flores, A., Rebello-Sanchez, I., Parente, J., Balbuena, A., de Melo, P. S., Otiniano-Sifuentes, R., Rivera-Torrejón, O., Abanto, C., Alva-Diaz, C., Musolino, P. L., & Fregni, F. (2022). Burden of stroke and population-attributable fractions of risk factors in Latin America and the Caribbean. *Journal of the American Heart Association*, *11*(21). https://doi.org/10.1161/jaha.122.027044

U.S. Department of Health and Human Services. (n.d.). *Alcohol use in the United States: Age groups and demographic characteristics*. National Institute on Alcohol Abuse and Alcoholism. https://www.niaaa.nih.gov/alcohols-effects-health/alcohol-topics/alcohol-facts-and-statistics/alcohol-use-united-states-age-groups-and-demographic-characteristics#:~:text=According%20to%20the%202022%20NSDUH%2C%20215.6%20million%20adults%20ages%2018,some%20point%20in%20their%20lifetime.&text=1%2C2-,This%20includes%3A,86.2%25%20in%20this%20age%20group)&text=108.1%20million%20women%20ages%2018,81.9%25%20in%20this%20age%20group)

US Health Statistics and data trends: Life expectancy, health insurance, and more. USAFacts. (2023, January 11). https://usafacts.org/topics/health/

Appendix 2

A complete record of the code used follows this report.