Predicting Life Expectancy in the US using Naïve Bayes Classifier, Linear Regression and Nearest Neighbor Classifer

*Abstract*—Life expectancy predictions play a crucial role in public health planning and policies, where statistical regression algorithms offer a robust approach to forecasting. This project explores various regression models to predict life expectancy. Comparative analysis of these algorithms contribute to a deeper understanding of how statistical models can be employed in predicting life expectancy.

# Introduction

# Life expectancy in the United States varies throughout all the states and many other factors come into play affecting a person's life expectancy. Life expectancy can vary on factors such as gender, race, health, occupation, and even tobacco/alcohol use. Using these factors our goal was to determine which statistical classifier model was most applicable to use to predict life expectancy of a person living in the United States. The classifier models we decided to use to predict life expectancy were Naive Bayes’ Classifier, Nearest Neighbor Classifier, and a Regression Model. We examined these models using various accuracy scores to determine which model is best for predicting a person’s life expectancy in the United States.

# Dataset

The dataset for this project came from the CDC population data from 2020 which contains life expectancy statistics based on U.S. states as well as life expectancy in the U.S based on gender. The CDC dataset helps estimate the U.S. life expectancy at birth for each state and sex and the estimates for this dataset were predicted for the whole 50 states in the U.S. and the data used to estimate the life expectancy for the year 2020 are state specific final numbers for the number of deaths in that year. There were other datasets from the year 2019 and 2018 but we decided to use 2020 since it was the most recent dataset available. Other sources that were used for our dataset included data from the Country Health Rankings and Roadmaps that included data of life expectancy based on race in the U.S. which also included data from the year 2020. For this project the CDC dataset was given in csv type format so we decided to download the provided csv file and copy it into our own csv file we used for out project, for the Country Health Rankings and Roadmaps dataset, the data was given in a visual map of each U.S. state and when clicked gave information of each demographic life expectancy, so we decided to manually put each state demographic life expectancy numbers and place it into our csv file for our project. Each dataset provided us with attributes that we used for our project which included state, race, which included the races of White, Black, Hispanic, Asian, Indian, and Alaska Native and gender, Male and Female. The attributes helped us in creating our classifiers of Naive Bayes, Nearest Neighbor and Regression model in helping us determine which statistical method is most applicable in calculating life expectancy for a person living in the United States.

# Methods and Measures Used

This project primarily utilizes three different statistical methods to determine which of the models most accurately predicts life expectancy based on the given attributes. These methods were employed using R, and included the Naïve Bayes Classifier, Nearest, Neighbor Classifier and Linear Regression model, each chosen for their distinct approach to analyzing and predicting life expectancy. A set of measures were used to evaluate the accuracy and performance of these models, namely Leave One Out Cross Validation (LOOCV), 5-NN Classifier, Mean Square Error (MSE), Root Mean Square Error (RSME), and R^2. This framework promoted a nuanced analysis of each algorithm’s effectiveness in predicting life expectancy.

# Results and Interpretations

## Naïve Bayes Classifier

The first step in using Naive Bayes’ Classifier was data processing. Since this classifier can only predict a class and not an actual value, we had to split life expectancy in five separate groups. After the data processing is complete the conditional probability tables are generated for each feature against the life expectancy groups, we created. The next step was to create a prediction function that worked by finding the posterior probability from multiplying the prior with the likelihood feature. The function then selects the prediction based on the highest posterior probability. The method we used to test the accuracy of this classifier was Leave-One-Out Cross Validation. LOOCV determines the accuracy by training the model on all but one data point at a single time and then testing it on the excluded point. We chose to use this testing method because the dataset was large enough to create a seventy thirty training testing split. The accuracy we received was twenty percent which is quite low. The reason this score is low is because the model predicts a group which is an age range instead of an actual age concrete value.

## Linear Regression

The linear regression model aims to predict life expectancy using the ‘State’, ‘Race’, and ‘Gender’ attributes. The process first involved manipulating and filtering the data. A ‘Feature’ column and ‘Life Expectancy’ column were created and contained values from the gathered variables. These columns iterated through the dataset to assign attributes to either Feature or Life Expectancy, and the demographical attributes were then converted to categorical factors to facilitate categorical analysis. After the data was properly manipulated, the linear regression algorithm was then used to calculate life expectancy based on these attributes. Following model fitting, predictions were generated from the model and residuals were calculated. The RSME and R^2 values were computed to evaluate the model’s performance, however the model receiving an R^2 of 1, and an RSME of 3.503177e-14 prompted a large consideration of overfitting.

1. Actual v. Predicted Life Expectancy

A graph with a line

Description automatically generated

Figure 1 demonstrates our data as displayed nearly perfectly accurate. Given that our dataset was overall small, this offers a possible explanation as to why our R^2 and RSME values were returning nearly pristine.

## Nearest Neighbor Classifier

In operating our Nearest Neighbor classifier, we first placed our data in a csv file for the classifier to read and made sure there were unique column names, we then we split the dataset by placing it into training and testing sets. The dataset was then separated into labels, ‘TotalLifeExpectany’, and features for both the training and testing sets with the columns ‘State’ and ‘TotalLifeExpectancy’ being excluded from the features. The features were then scaled, more specifically the features were scaled to have zero mean and unit variance and helped standardize the data to prevent features with larger scales from interfering or dominating the algorithm. The labels were then converted to a numeric format which is necessary for this model as the nearest neighbor classifier relies on numeric labels or variables to calculate its mathematical operations. We then performed the K-NN model training in which we helped define and trained using the KNN function in R. the formula ‘train\_label’ used in the R code for our classifier is used for the model to predict ‘train\_label’ based on all the other features. The k is then set to 5, meaning that the model considers the 5 nearest neighbors. The final step in our Nearest Neighbor classifier was the model evaluation in which the Mean Absolute Error or MAE is calculated as a measure of how well the model performs on the test set, with smaller MAE values indicating better performance and bigger MAE values indicating worser performance. The Mean Absolute Error of this classifier with our data came out to be about 1.353.

# Key Takeaways and Future Improvements

Regression analysis stands out as the most effective method for calculating the life expectancy of a person living in the United States. This is primarily because regression models excel in capturing direct relationships between independent and dependent variables. In the context of life expectancy, various factors such as lifestyle, environmental conditions, healthcare access, and genetic predispositions can be considered as independent variables that directly influence the dependent variable, which is the life expectancy. Unlike classifiers, which predict categories, regression models are adept at predicting actual numeric values. This capability is crucial in the case of life expectancy prediction, where the outcome is a specific age or a range of ages, rather than a mere category. Therefore, the precision and direct correlation handling of regression models make them the most suitable tool for this kind of predictive analysis.

In improving our project, we would try and add more data in our dataset to help reduce the chances of overfitting in our models. We also wished to improve upon better organization of attributes for our dataset in our project as we had a hard time working around unorganized dataset when modeling the regression model. Another challenge we faced was finding datasets that contained life expectancy attributes other than gender, state, and race.

# Team Contributions

The contributions for the team project are as follows. Barr Mohammed researched the dataset for life expectancy for race in each state of the U.S. to be used as the overall dataset for the classifiers in the project. He worked on the dataset/attribute slide and the Nearest Neighbor slide for the presentation and lastly worked on completing the Nearest Neighbor classifier code in R studio. Vlad Pavlovich also researched the dataset on life expectancy per state and per gender and constructed the main dataset that holds the entire data. He created a Naive Bayes’ Classifier that predicts life expectancy in R studio, worked on adapting the data set to work with the linear regression predictor and partially worked on creating the linear regression model. He also worked on the Naive Bayes Classifier slide, Improvement and Challenges slide. McKayla Widener helped with background research and organizing team meetings and work schedules and partially worked on the linear regression model construction. She created the linear regression model slide and explanation, and drew overall conclusions from the research, and assisted in formatting the final report in IEE style. Top of Form

# ReferencesBottom of Form

1. National Center for Health Statistics, “U.S. State Life Expectancy,” Centers for Disease Control and Prevention, 2020. Available: <https://www.cdc.gov/nchs/data-visualization/state-life-expectancy/index_2020.htm>.
2. Forbes Advisor, "Average Salary by State," Forbes. Available: <https://www.forbes.com/advisor/business/average-salary-by-state/>
3. Centers for Disease Control and Prevention (CDC), "State Life Expectancy," National Center for Health Statistics, 2020. Available: <https://www.cdc.gov/nchs/data-visualization/state-life-expectancy/index_2020.htm>. [Accessed: Dec. 13, 2023].
4. Our World in Data, "Life Expectancy vs. GDP per Capita," Available: <https://ourworldindata.org/grapher/life-expectancy-vs-gdp-per-capita?tab=table&time=1925..latest&country=~USA>.