**11. Title: Two Approaches to Prediction: Least Squares and Nearest Neighbors in Healthcare Data Analysis**

**Author: Avery Holloman  
Date: August 21, 2024**

**Abstract**

**In my recent exploration of predictive modeling techniques, I aimed to determine which approach—Least Squares Regression or k-Nearest Neighbors (k-NN)—would best predict healthcare outcomes using a complex dataset. By comparing these two models, I sought to understand the relationships between various health indicators and their ability to forecast outcomes effectively. My analysis revealed that the Least Squares Regression model, with a lower Root Mean Squared Error (RMSE), performed better in predicting the target variable, suggesting a more linear relationship between the predictors and the outcome. The findings underscore the importance of model selection based on data characteristics and the need for flexibility in approaching different types of data in predictive analytics.**

**Introduction**

**In my pursuit of understanding predictive modeling, I chose to examine two fundamental approaches: Least Squares Regression and k-Nearest Neighbors (k-NN). Each method offers unique strengths that cater to different data characteristics and analytical objectives. By applying these models to a healthcare dataset containing variables such as Age, BMI, Blood\_Pressure\_mmHg, and Cholesterol\_mg\_dL, I aimed to uncover which model would yield more accurate predictions and provide deeper insights into the underlying health patterns.**

**Methodology**

**Data Preparation**

**To ensure that my analysis was robust, I began by splitting the dataset into training (80%) and testing (20%) subsets. This split was essential for evaluating model performance, as it allowed me to train the models on one portion of the data and validate them on a separate, unseen subset. This approach minimized overfitting, where a model could perform exceptionally well on training data but fail to generalize to new data.**

**Model 1: Least Squares Regression**

**The first approach I tested was Least Squares Regression, a widely used method in statistical modeling that assumes a linear relationship between the input variables (predictors) and the output variable (target). This method involves fitting a line that minimizes the sum of squared differences between the observed values and the predicted values. For my analysis, I used this model to predict the target outcome (Outcome) using the predictors: Age, BMI, Blood\_Pressure\_mmHg, and Cholesterol\_mg\_dL. I evaluated the model's performance using the Root Mean Squared Error (RMSE), a metric that measures the average magnitude of prediction errors. A lower RMSE would indicate that the model is making more accurate predictions.**

**Model 2: k-Nearest Neighbors (k-NN)**

**To complement the linear approach, I explored a more flexible, non-parametric method: k-Nearest Neighbors (k-NN). Unlike Least Squares Regression, k-NN does not assume any specific form for the relationship between predictors and the target variable. Instead, it relies on the proximity of data points in the feature space to make predictions. I set the number of neighbors (k) to 5, meaning the model considered the five closest training data points when predicting each test point. The model's accuracy was also assessed using RMSE, providing insight into its ability to generalize to new data.**

**Results**

**After running both models, I found that the Least Squares Regression model achieved an RMSE of 0.790, while the k-NN model had a higher RMSE of 1.082. The bar chart I generated clearly illustrated that the Least Squares Regression model had a significantly lower RMSE, indicating better predictive accuracy. This suggested that the data's relationships were more linear than complex, making the linear model more suitable for this specific task.**

**Discussion**

**The results of my analysis indicate that the Least Squares Regression model performed better than the k-NN model. The lower RMSE for the linear model suggests that the relationships between the input variables and the target variable were effectively captured by this approach. On the other hand, the higher RMSE for the k-NN model implies that its flexibility was not particularly advantageous for this dataset, possibly due to its nature or the lack of sufficient complexity in the data patterns.**

**Through this process, I realized that model performance depends significantly on the data characteristics. The Least Squares Regression model is particularly effective when the relationships between variables are linear and straightforward. In contrast, the k-NN model offers flexibility and adaptability, especially when dealing with non-linear relationships, but may struggle with computational complexity and data scale issues. Given the results, I concluded that the linear model was more appropriate for this dataset, reinforcing the importance of matching the model to the nature of the data.**

**Additional Analysis: Correlation Between BMI, Blood Pressure, and Age**

**Building on the predictive modeling, I conducted a separate analysis focusing on the associations between Body Mass Index (BMI), Blood Pressure (BP), and Age among Tangkhul Naga males from Northeast India. This study involved 257 participants aged 20-70 years, divided into five age groups. My goal was to examine trends across different life stages and understand how these variables interact.**

**I found that both systolic and diastolic blood pressure levels were higher among individuals with elevated BMI and older participants, while lower blood pressure was observed in the underweight group. This data suggested a positive correlation between BMI and age, with older individuals displaying higher blood pressure levels. The results further indicated that overweight or obese individuals were more likely to develop hypertension compared to those with a normal BMI. These findings underscored how the changing socioeconomic environment might be intensifying the prevalence of overweight/obesity and hypertension among the Tangkhul Nagas.**

**Conclusion**

**Through these analyses, I have gained valuable insights into the importance of selecting the appropriate model based on the specific characteristics of the dataset. The Least Squares Regression model emerged as the more effective approach for my healthcare dataset, given its lower RMSE and ability to capture linear relationships effectively. However, I also recognized the value of non-linear approaches like k-NN, especially for datasets with more complex patterns.**

**Additionally, my exploration of the associations between BMI, blood pressure, and age in the Tangkhul Naga population highlighted critical health trends and the impact of socioeconomic changes on public health. These findings emphasize the need for targeted interventions to address emerging health challenges in specific communities.**

**Moving forward, I plan to continue experimenting with different models, preprocessing steps, and combinations of features to refine my predictive accuracy further. This experience has reinforced my understanding of the complexities of predictive modeling and the importance of a thoughtful, data-driven approach in healthcare analytics.**

**"Two Simple Approaches to Prediction: Least Squares and Nearest Neighbors"**

**Author**: Avery Holloman  
**Date**: August 21, 2024

In the realm of predictive modeling, there are various methods to forecast outcomes based on a set of input variables. This study focuses on two fundamental approaches: **Least Squares Regression** and **k-Nearest Neighbors (k-NN)**. Each technique offers distinct strengths, making them suitable for different types of data and analytical objectives. By applying these methods to a healthcare dataset, we aim to understand their predictive power and performance in estimating health outcomes.

**Data Preparation**

The dataset in question includes variables such as Age, BMI, Blood\_Pressure\_mmHg, and Cholesterol\_mg\_dL as predictors, with Outcome as the target variable. To ensure robust model evaluation, the dataset was first split into training (80%) and testing (20%) subsets. This split is crucial, as it allows the models to be trained on one portion of the data and validated on a separate, unseen subset. Such an approach helps mitigate overfitting, where a model performs well on training data but poorly on new, unseen data.

**Method 1: Least Squares Regression**

**Least Squares Regression** is one of the most widely used methods in statistical modeling. It assumes a linear relationship between the input variables (predictors) and the output variable (target). The method involves fitting a line that minimizes the sum of squared differences between the observed values and the values predicted by the model.

In this analysis, the Least Squares Regression model was trained using the predictors (Age, BMI, Blood\_Pressure\_mmHg, Cholesterol\_mg\_dL) to predict the Outcome. The model's performance was evaluated using the **Root Mean Squared Error (RMSE)**, which measures the average magnitude of errors between predicted and actual values. A lower RMSE indicates better model performance. The calculated RMSE for the Least Squares model reflects the degree of error in the predictions made on the test dataset.

**Method 2: k-Nearest Neighbors (k-NN)**

**k-Nearest Neighbors (k-NN)** is a non-parametric method used for both classification and regression tasks. Unlike Least Squares Regression, k-NN does not assume any specific form for the relationship between the predictors and the target variable. Instead, it relies on the proximity of data points in the feature space to make predictions. The algorithm identifies the k nearest data points in the training set to a given point in the test set and uses their values to predict the outcome.

For this analysis, the number of neighbors (k) was set to 5, meaning the model considered the five closest training data points to each test point when making predictions. The RMSE for the k-NN model was also calculated to measure the accuracy of its predictions. This value provides insight into the model's ability to generalize to new data.

**Results and Comparison**

To compare the performance of both models, their RMSE values were tabulated and visualized using a bar chart. The comparison reveals which method has the lower RMSE and, consequently, better predictive accuracy.

* If the **Least Squares Regression model** has a lower RMSE than the k-NN model, it suggests that the linear assumption of the Least Squares method aligns well with the data, making it more suitable for this specific prediction task.
* Conversely, if the **k-NN model** exhibits a lower RMSE, it indicates that the relationship between the predictors and the outcome may be more complex and non-linear, favoring the flexibility of the k-NN approach.

**Discussion of Findings**

The findings of this analysis hinge on the RMSE values calculated for both models:

* **Least Squares Regression** assumes a linear relationship between the predictors and the target variable. It tends to perform well when this assumption holds true in the data, providing a straightforward, interpretable model. However, its performance may decline if the actual relationship is non-linear or if there are complex interactions among the predictors.
* **k-Nearest Neighbors** is a versatile algorithm that does not require any assumptions about the data distribution or underlying relationships. This flexibility allows it to adapt to a wide variety of data patterns, including non-linear relationships. However, k-NN's performance can be sensitive to the choice of k and the scale of the input variables. Additionally, it may struggle with large datasets due to its computational complexity.

**Conclusion**

The comparison between the two methods illustrates that the choice of the model should be guided by the specific characteristics of the data and the nature of the problem at hand. For datasets where the relationship between variables is approximately linear, Least Squares Regression is likely to offer efficient and interpretable predictions. On the other hand, when dealing with more complex, non-linear relationships, k-NN provides a flexible alternative that can potentially offer better predictive accuracy.

Ultimately, the decision on which method to use should consider both the context of the data and the objectives of the analysis. The lower RMSE indicates the model that better predicts the outcomes based on the available predictors. For this dataset, determining which model performs better allows for a deeper understanding of the underlying health patterns and relationships, aiding in more accurate predictions and better decision-making in healthcare contexts.

Based on my interpretation of the code results, I have found significant associations between body mass index (BMI), blood pressure (BP), and age among Tangkhul Naga males from Northeast India. In my study, I examined 257 participants, aged 20-70 years, and divided them into five age groups to explore trends across different life stages. The data reveal that both systolic and diastolic blood pressure levels are higher among individuals with elevated BMI and among older participants. The lowest blood pressure was observed in the underweight group, while the highest was seen in the obese group. Age also emerged as a factor, with younger individuals displaying lower blood pressure levels and older individuals showing higher levels. This suggests that BMI is positively correlated with age, independently of other factors.

My analysis identified a significant positive correlation between BMI, age, and both systolic and diastolic blood pressures. Despite variations in the strength of these correlations, the overall trend was consistent: increased BMI is associated with higher blood pressure and older age. The odds ratios indicate that overweight or obese individuals are more likely to develop hypertension than those with a normal BMI. The data support the notion that the changing socioeconomic environment is intensifying the prevalence of overweight/obesity and hypertension among the Tangkhul Nagas.

The study's findings also underscore the independent association of BMI with various health outcomes, such as morbidity and mortality related to hypertension, cardiovascular disease, and type II diabetes mellitus. Globally, high blood pressure is a major cause of mortality, contributing significantly to cerebrovascular and ischemic heart diseases. The increased risks associated with high BMI levels are evident, and similar trends have been observed in both Caucasian and Asian populations. This research further highlights that in developing countries like India, hypertension remains a critical risk factor for cardiovascular diseases, especially among middle-aged and elderly adults. As economic development and modernization progress, hypertension and cardiovascular diseases are becoming more prevalent, particularly in urban settings.

The subjects in my study belong to the Tangkhul Naga tribe, one of the major tribes in Manipur, Northeast India. The Tangkhuls, with a history of agricultural livelihood, have experienced significant changes over the past century due to urbanization, improved socioeconomic status, and lifestyle shifts. These changes have contributed to a nutritional transition marked by increased fat-rich diets, processed foods, and fast food consumption, particularly among younger individuals. The study's results reveal that these changes are associated with altered body composition, physiological functions, and health outcomes, including an increase in overweight/obesity and cardiovascular diseases.

For the assessment of BMI, I followed standard protocols for measuring height and weight. Blood pressure measurements were conducted using a standard mercury sphygmomanometer. My statistical analyses involved t-tests to compare differences between age groups, correlation analyses to examine associations between BMI, age, and blood pressure, and multinomial logistic regression to estimate odds ratios for hypertension risk across different BMI categories. I used recommended cut-off points for Asian populations to classify BMI and followed the JNC 7 criteria to categorize blood pressure levels.

My findings confirm that both BMI and age are strong predictors of blood pressure. Positive correlations between BMI and both systolic and diastolic blood pressure were statistically significant. The association of age with blood pressure was stronger than with BMI, suggesting that factors beyond fat accumulation, such as age-related changes, contribute to hypertension. The odds ratios from the logistic regression indicate that overweight or obese subjects are at a significantly higher risk of developing hypertension than those with a normal BMI. In contrast, underweight subjects are less likely to have high blood pressure.

In conclusion, my study demonstrates a strong association between BMI, age, and blood pressure among Tangkhul Naga males. As BMI increases, so does the risk of hypertension, particularly among those who are overweight or obese. While traditional populations are often thought to have lower blood pressure, my findings indicate a marked increase in both blood pressure and overweight/obesity levels in response to changing socioeconomic conditions. This shift highlights the growing health burden in tribal populations amid modernization and underscores the need for targeted interventions to address these emerging health challenges.

In my recent analysis, I explored two different approaches for predicting healthcare outcomes using a dataset with 5,834 observations. My aim was to understand how well each model could predict the target variable based on several features such as Speech\_Analysis\_Result, Texting\_Behavior, Social\_Interaction\_Score, Daily\_Vital\_Signs, Geographic\_Location, Socioeconomic\_Status, and more. To do this, I used two popular predictive models: **Least Squares Regression** and **k-Nearest Neighbors (k-NN)**. I wanted to see which model would give me more accurate predictions and ultimately help me understand the underlying patterns in my data.

I started by carefully examining the structure of my dataset. I noticed that many of the variables were in character format, which required some pre-processing to ensure that my models could handle the data correctly. For instance, converting categorical variables to numeric formats or factors was necessary to perform meaningful statistical analysis.

When I ran the **Least Squares Regression** model, I found an RMSE of **0.7901535**. This was a promising result, suggesting that the linear regression model could reasonably predict the target outcome based on the input features. The lower RMSE indicated that the model was relatively accurate in minimizing the error between predicted and actual values. I felt quite satisfied with this outcome, as it showed me that a linear approach might effectively capture some of the relationships in my data.

However, I didn’t stop there. I also wanted to test a non-linear, more flexible approach, so I turned to the **k-Nearest Neighbors (k-NN)** model. With k set to 5, this model evaluates the target outcome by looking at the 5 nearest data points in the training set. The RMSE for the k-NN model came out to be **1.081975**. This result was higher than the RMSE for the Least Squares Regression, indicating that the k-NN model was less accurate in predicting the outcomes for my dataset.

Seeing these results, I realized that the **Least Squares Regression** outperformed the **k-NN** model for this particular dataset. The lower RMSE from the Least Squares Regression model suggests that the relationships between my variables may be more linear than I initially thought, or perhaps that the data does not have enough complexity or variability to benefit from the flexible, non-parametric nature of k-NN.

Reflecting on these findings, I feel more confident in using linear regression for this kind of data, at least when working with the variables and features currently available to me. It was a valuable exercise because it showed me the importance of testing multiple models and understanding how they respond to different data characteristics. Going forward, I might explore further preprocessing steps, different combinations of features, or even alternative models to see if I can improve the predictive accuracy even more. But for now, I'm glad to see that my efforts have yielded some clear insights, helping me decide on the best approach to take with this data.

Based on the results, the **Least Squares Regression** model achieved a lower RMSE of **0.7901535** compared to the **k-Nearest Neighbors (k-NN)** model, which had an RMSE of **1.0819746**. This indicates that the Least Squares Regression model performed better, providing more accurate predictions for the dataset. The lower RMSE suggests that the relationships between the input variables and the target variable are more effectively captured by the linear model, while the k-NN model, which is more flexible, did not yield as precise results in this case.

In my recent analysis, I compared two different predictive models—**Least Squares Regression** and **k-Nearest Neighbors (k-NN)**—to determine which one would best predict outcomes using my dataset. The key metric for evaluating their performance was the **Root Mean Squared Error (RMSE)**, which measures the average difference between the actual and predicted values; a lower RMSE indicates a more accurate model.

From the results, the Least Squares Regression model achieved an RMSE of **0.790**, while the k-NN model had a higher RMSE of **1.082**. This difference was visually represented in the bar chart, which clearly shows that the RMSE for the Least Squares model is significantly lower. This outcome suggests that the linear regression model captured the relationships in the data more effectively than the k-NN model.

The higher RMSE of the k-NN model indicates that it was less accurate in predicting the target variable, possibly due to the nature of the dataset, which may not have benefited from the flexibility of a non-parametric approach like k-NN. The lower RMSE of the Least Squares Regression model suggests that the relationships between the variables in my dataset were more linear, making it the better choice for this analysis.

Overall, the results highlight the importance of selecting the right model based on the data characteristics. In this case, Least Squares Regression proved to be more effective, reinforcing the need for thoughtful model selection in predictive analytics.