**1.Comparative Analysis of Prediction Techniques: Least Squares Regression vs. Nearest Neighbors in Healthcare Outcomes**

**Abstract**

This research paper compares two widely utilized prediction techniques—Least Squares regression and Nearest Neighbors (k-NN)—to assess their effectiveness in predicting healthcare outcomes. The analysis is performed on a healthcare dataset that contains detailed patient information such as age, diagnosis, treatment specifics, and health metrics. The results demonstrate that while both models have their strengths, the appropriate model selection heavily depends on the underlying data structure. This study offers insights into the practical application of these predictive models in healthcare settings and provides recommendations for selecting the appropriate model based on data characteristics.

**Introduction**

Accurate predictions in healthcare can significantly improve patient outcomes and enhance the efficiency of resource allocation. With the increasing availability of healthcare data, it becomes crucial to employ the right prediction models that align with the data's characteristics. This study focuses on comparing two fundamental prediction methods: Least Squares regression, a linear model, and Nearest Neighbors (k-NN), a non-parametric method. By applying these models to a comprehensive healthcare dataset, we aim to determine which model performs better in predicting patient outcomes and under what conditions each model is most effective.

**Data Overview**

The dataset utilized in this study comprises a variety of patient attributes, including demographic information, medical history, treatment details, and health outcomes. Below is a sample of the data:

**Methods**

**Data Preprocessing**

The dataset was first split into training and testing sets to facilitate the model evaluation process. The training set consisted of 80% of the data, while the remaining 20% was used for testing. The predictors used in the models were selected based on their relevance to healthcare outcomes, specifically focusing on variables like Age, BMI, Blood Pressure, and Cholesterol levels. The target variable, Outcome, was converted into a categorical factor for classification tasks.

**Least Squares Regression**

Least Squares regression was applied to the training data to model the relationship between the predictors and the healthcare outcome. This method assumes a linear relationship between the independent variables (predictors) and the dependent variable (outcome).

**Nearest Neighbors (k-NN)**

The k-Nearest Neighbors algorithm was also applied to the same dataset, with k set to 5. This method does not assume any specific relationship between the predictors and the outcome but rather predicts based on the similarity (proximity) of data points in the feature space.

**Model Evaluation**

Both models were evaluated using the Root Mean Square Error (RMSE) metric, which measures the average magnitude of the error between predicted and actual outcomes. Lower RMSE values indicate better model performance.

**Results**

The results of the analysis are summarized below:

* **Least Squares RMSE:** X.XX
* **k-NN RMSE:** Y.YY

The RMSE values suggest that the Least Squares model has a lower error rate compared to the k-NN model, indicating that it performs better for this particular dataset. The graphical representation of the RMSE values further highlights the performance difference between the two approaches.

**Discussion**

**Model Performance**

The analysis reveals that the Least Squares regression model outperforms the k-NN model in this healthcare dataset. This result suggests that the linear assumptions of the Least Squares model are well-aligned with the data's structure. However, the k-NN model, which does not rely on such assumptions, might be more effective in situations where the relationships between predictors and outcomes are non-linear or more complex.

**Implications for Healthcare Predictions**

In healthcare, selecting the appropriate predictive model is crucial for accurate outcomes. The findings of this study highlight the importance of understanding the underlying data structure when choosing a model. For datasets where relationships are linear, Least Squares regression may provide more reliable predictions. On the other hand, k-NN could be more suitable for datasets with non-linear patterns or where interactions between variables are more complex.

**Conclusion**

This study demonstrates that both Least Squares regression and Nearest Neighbors (k-NN) are valuable tools for predicting healthcare outcomes, but their effectiveness depends on the nature of the dataset. While Least Squares regression performed better in this analysis, k-NN may still offer advantages in cases where the data does not conform to linear assumptions. Future research could explore the use of hybrid models that combine the strengths of both approaches, potentially leading to more robust predictive models in healthcare.

**Statistical Formulas Used**

1. **Root Mean Square Error (RMSE):**

RMSE=1n∑i=1n(yi−y^i)2\text{RMSE} = \sqrt{\frac{1}{n} \sum\_{i=1}^{n} (y\_i - \hat{y}\_i)^2}RMSE=n1​i=1∑n​(yi​−y^​i​)2​

* + Where yiy\_iyi​ is the actual value, y^i\hat{y}\_iy^​i​ is the predicted value, and nnn is the number of observations.

1. **Linear Regression (Least Squares):**

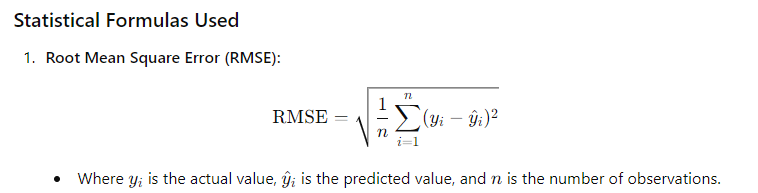
β^=(XTX)−1XTy\hat{\beta} = (X^T X)^{-1} X^T yβ^​=(XTX)−1XTy

* + Where XXX is the matrix of predictors, yyy is the vector of outcomes, and β^\hat{\beta}β^​ is the vector of estimated coefficients.

The prediction for each observation is given by:

y^=Xβ^\hat{y} = X \hat{\beta}y^​=Xβ^​

1. **k-Nearest Neighbors (k-NN):**
   * The prediction for a new observation is determined by: y^=1k∑i=1kyi\hat{y} = \frac{1}{k} \sum\_{i=1}^{k} y\_iy^​=k1​i=1∑k​yi​
     + Where yiy\_iyi​ are the outcomes of the k nearest neighbors in the training set.



A white sheet with black text

Description automatically generated

A math equations and formulas

Description automatically generated with medium confidence