Health Metrics Analysis and Disease Prediction Report

1. Introduction

This report aims to analyze health metrics of patients and identify potential health risks based on their measurements, including blood pressure, cholesterol levels, and Body Mass Index (BMI). The analysis also includes predictive modeling to assess the likelihood of patients developing hypertension based on their current health status.

2. Objective:

- Explore and understand the features of the dataset, including key health metrics such as blood pressure, cholesterol levels, and BMI.
- Perform data preprocessing, addressing missing values and outliers to ensure data integrity.
- Identify key factors influencing health risks using statistical analysis techniques.
- Develop predictive models to forecast the likelihood of hypertension based on historical health data.

3. Scope:

1. Data Exploration: Understanding the dataset and its features (blood pressure, cholesterol levels, BMI).
2. Data Preprocessing: Cleaning the dataset by handling missing values and detecting/removing outliers.
3. Feature Analysis: Identifying significant features that impact health outcomes.
4. Data Visualization: Creating visual representations to analyze relationships between health metrics and risk factors.
5. Model Building: Developing and evaluating predictive models for health risk assessment.
6. Reporting: Documenting findings and providing recommendations to stakeholders.

4. Methodology:

- Data Collection
    - Describe the data sources and how the data was obtained.
    - Example: "The data for this analysis was collected from a MySQL database, which includes patient records and health metrics. The database contains tables for patient demographics, health metrics, and medical history."

- Data Preprocessing
    - Explain any preprocessing steps taken to clean and prepare the data for analysis.
    - Example: "Data preprocessing involved several steps:

- Missing Values: Missing values were identified and handled using imputation techniques. For continuous variables, mean imputation was applied, while categorical variables were filled with the mode."
- Data Type Conversion: Data types for various columns were converted to appropriate formats (e.g., categorical variables were encoded)."
- Outlier Removal: Outliers were detected using the IQR method and were removed to ensure data integrity."

- Exploratory Data Analysis (EDA)

  Outline the techniques used to explore the data.Example: "EDA was conducted to gain insights into the data distribution and relationships between variables. This included:

  - Descriptive statistics to summarize key metrics (e.g., mean, median, standard deviation).
  - Visualizations using Matplotlib and Seaborn, including histograms, box plots, and scatter plots to visualize distributions and correlations."

- Feature Engineering
  - Discuss any new features created to improve the model's predictive power.
  - Example: "Feature engineering involved creating new variables based on existing data, such as categorizing BMI into 'Underweight,' 'Normal weight,' 'Overweight,' and 'Obese' based on standard BMI ranges."

- Model Selection
  - Describe the machine learning models chosen for predicting health metrics and why.
  - Example: "The following machine learning models were selected for this analysis:
  - Random Forest: Chosen for its robustness and ability to handle nonlinear relationships.
  - Logistic Regression: Used for binary classification tasks such as predicting the risk of hypertension.
  - K-Means Clustering: Applied for segmenting patients based on health metrics."
  - Model Training and Evaluation

5. Modeling:

1. Linear Regression
   a. Linear Regression was utilized as a foundational model due to its simplicity and interpretability. The following steps were involved:

b. Data Preparation: We standardized the input features to ensure they were on a similar scale, which is essential for the performance of linear models. Missing values were handled appropriately to maintain the integrity of the dataset.
c. Model Training: We split the dataset into training and testing sets using an 80-20 split. The training set was used to fit the Linear Regression model, with the target variable being the hypertension risk indicator.
d. Evaluation: The model's performance was evaluated using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared values. These metrics provided insight into the model's accuracy and predictive capability.

2. Random Forest
    a. To capture more complex relationships within the data, we also implemented a Random Forest model. The methodology included:
    b. Hyperparameter Tuning: We performed grid search for hyperparameter tuning to optimize the number of trees and the maximum depth of the forest, which enhanced the model's predictive power.
    c. Model Training: Similar to the Linear Regression approach, we trained the Random Forest model using the training dataset. Random Forest's ensemble nature allowed it to handle non-linear relationships effectively.
    d. Feature Importance: After training, we analyzed the feature importance scores to understand which variables had the most significant impact on the prediction of hypertension risk. This analysis aids in deriving insights from the model regarding key health metrics.
    e. Evaluation: We assessed the Random Forest model using similar evaluation metrics (MAE, MSE, R-squared) and compared its performance against the Linear Regression model to determine which provided better predictive accuracy.

7. Conclusion

In summary:

- The analysis provided insights into the health metrics influencing the risk of hypertension.
- Predictive models developed can assist healthcare providers in identifying patients at risk and tailoring interventions.
- Future work could explore additional health features and advanced machine learning models for improved predictions.