**Heart Disease Prediction Report**

**Introduction:**

According to the CDC, heart disease is a leading cause of death across various races in the U.S., primarily influenced by risk factors such as high blood pressure, high cholesterol, and smoking. Additional important indicators include diabetes status, obesity (high BMI), insufficient physical activity, and excessive alcohol consumption. Advances in computing enable the application of machine learning techniques to identify patterns in data, which can significantly aid in predicting heart disease.

**Dataset Overview:**

The dataset utilized for this analysis includes a range of health metrics and lifestyle factors, with the target column being "HeartDisease," which indicates whether an individual has heart disease.

**Key Features:**

* BMI (Body Mass Index): Reflects body fat based on height and weight.
* Smoking Status: Indicates if the individual smokes.
* Alcohol Consumption: Measures drinking habits.
* Diabetes Status: Indicates whether the individual has diabetes.
* Physical Activity: Represents exercise frequency and intensity.
* Blood Pressure and Cholesterol Levels: Key indicators of heart health.

**Dataset Characteristics:**

The dataset is unbalanced, with a higher number of individuals not having heart disease compared to those who do. This imbalance can negatively impact model performance, leading to a bias towards the majority class.

**Methodology:**

**K-Nearest Neighbors (KNN) Model**

To predict heart disease, we employed the K-Nearest Neighbors (KNN) algorithm, which classifies instances based on the closest data points in the feature space.

**1. Data Preprocessing:**

* Categorical variables were encoded appropriately.
* Numerical features were standardized to ensure uniformity in distance calculations.

**2.Feature Engineering:**

To improve the model's predictive capability, several new features were created based on the existing health metrics and lifestyle factors:

**Health Composite Score:**

Description: A combined score reflecting overall well-being based on physical and mental health.

* Calculation:

HealthCompositeScore=(PhysicalHealth+MentalHealth)/2

**BMI Categories:**

* Description: Categorizing individuals into BMI categories (e.g., Underweight, Normal, Overweight, Obese).
* Categories:

Underweight: BMI < 18.5

Normal: 18.5 ≤ BMI < 25

Overweight: 25 ≤ BMI < 30

Obese: BMI ≥ 30

**3. Addressing Class Imbalance:**

* To counter the effects of the unbalanced dataset, we applied the Synthetic Minority Over-sampling Technique (SMOTE). This technique generates synthetic samples for the minority class (individuals with heart disease), helping to create a more balanced dataset for training.

**4. Model Training:**

* The dataset was divided into training and test sets.
* The KNN model was trained on the balanced dataset using an optimal value of K determined through cross-validation.

**5. Model Evaluation:**

* Model performance was evaluated using metrics such as accuracy, precision, recall, and F1 score on the test dataset.

**Results:**

**KNN Model Performance Metrics After SMOTE:**

The KNN model produced the following performance metrics after training on the balanced dataset:

* **Precision**: 0.80
* **Recall**: 0.93
* **F1 Score**: 0.86

These results indicate that the model effectively identifies individuals at risk of heart disease while mitigating the challenges posed by the original unbalanced dataset.

**Interpretation of Results:**

* **Precision (0.80)**: This indicates that 80% of the individuals identified as having heart disease truly have the condition. This high precision suggests that the model is reliable in its predictions.
* **Recall (0.93)**: A recall of 93% means that the model successfully identified 93% of actual heart disease cases. This high recall is critical in healthcare settings, where missing a diagnosis can have serious implications.
* **F1 Score (0.86)**: The F1 score is the harmonic mean of precision and recall, balancing the two metrics. A score of 0.86 reflects a strong overall performance of the model, making it a valuable tool for predicting heart disease.

**Conclusion:**

The application of SMOTE significantly improved the model's ability to predict heart disease by creating a more balanced dataset. This, in turn, led to enhanced performance metrics, demonstrating the effectiveness of using machine learning techniques in healthcare for risk assessment.

**Recommendations:**

To reduce the risk of heart disease, it is essential to:

* Regularly monitor and manage blood pressure and cholesterol levels.
* Maintain a healthy weight through diet and exercise.
* Avoid smoking and limit alcohol consumption.
* Stay informed about diabetes status and manage it effectively.

This report highlights the importance of data-driven approaches in healthcare, emphasizing the need for targeted interventions to mitigate heart disease risk. Future research could explore more advanced modeling techniques and additional features to further improve prediction accuracy.