Predicting Diabetes Risk Using Health Indicators

Project Report

Group 11

Aayush Amrute (NUID: 002838262)

Paritosh Vyawahare (NUID: 002416079)

Saurabh Chavan (NUID: 002479083)

Simran Sinha (NUID: 002475433)

amrute.a@northeastern.edu
vyawahare.p@northeastern.edu
chavan.sau@northeastern.edu
sinha.sim@northeastern.edu

Percentage of Effort Contributed by Student 1: 25%
Percentage of Effort Contributed by Student 2: 25%
Percentage of Effort Contributed by Student 3: 25%
Percentage of Effort Contributed by Student 4: 25%

Signature of Student 1: **Aayush Amrute**
Signature of Student 2: **Paritosh Rajkumar Vyawahare**
Signature of Student 3: **Saurabh Siddheshwar Chavan**
Signature of Student 4: **Simran Abhay Sinha**

Submission Date: 12 December 2024

Abstract

Diabetes is a critical global health challenge, impacting millions worldwide. Early detection and classification into categories like Non-Diabetic, Pre-Diabetic, and Diabetic are essential for effective intervention. This project utilizes machine learning techniques to predict diabetes status based on health, lifestyle, and demographic factors. By analyzing a dataset of health indicators, we aim to provide actionable insights and a scalable predictive model for healthcare professionals. Early identification of at-risk individuals can lead to timely intervention and better health outcomes. The study highlights the potential of data-driven approaches in reducing the burden of diabetes and improving patient outcomes. This project focuses on using machine learning techniques to predict the risk of diabetes based on various health indicators.

First, we conduct exploratory data analysis (EDA) to acquire insight into the dataset's structure and properties. We then perform feature engineering, which entails turning categorical data into numerical format to prepare it for model training. We then clean the data by removing outliers, missing values, and duplicates.

Following preprocessing, we divided the data into training and testing sets to train and evaluate our classification models. We experiment with numerous methods, such as Logistic Regression, Random Forest, and XGBoost Model, to see which one performs best.

Problem Definition

Objective

To develop a machine learning model that predicts an individual's diabetes status based on various health indicators, enabling early diagnosis and effective intervention. The target categories are:

- Non-Diabetic
- Pre-Diabetic
- Diabetic

Why is this important?

Diabetes leads to severe complications like cardiovascular diseases, kidney failure, and blindness if unmanaged. Predictive models can assist healthcare professionals in identifying at-risk individuals and crafting personalized care plans.

Goals

1. Build an accurate and interpretable machine learning model.
2. Identify the key factors influencing diabetes risk.
3. Provide insights that can be integrated into real-world healthcare systems

Introduction

Diabetes has emerged as a pressing global health issue, with millions of individuals affected annually. This chronic condition, characterized by elevated blood sugar levels, can lead to severe health complications if left untreated, including cardiovascular diseases, kidney failure, and vision impairment. Early detection and classification of diabetes risk categories, such as Non-Diabetic, Pre-Diabetic, and Diabetic, are crucial to mitigating its impact and providing timely interventions.

In recent years, the availability of extensive health datasets and advancements in machine learning have enabled healthcare professionals to leverage data-driven solutions for predicting diabetes risk. These predictive models not only assist in identifying high-risk individuals but also uncover key factors contributing to the onset of diabetes. By utilizing machine learning techniques, healthcare systems can transition from reactive treatments to proactive measures, significantly improving patient outcomes.

This project aims to bridge the gap between data insights and actionable healthcare interventions. By developing an accurate and interpretable predictive model, we strive to provide healthcare professionals with the tools necessary to prioritize patient care and optimize prevention strategies. The study emphasizes the importance of understanding health, lifestyle, and demographic factors in predicting diabetes risk and highlights the potential of technology-driven approaches in addressing complex healthcare challenges.

Through this endeavor, we aspire to contribute to the ongoing efforts in combating diabetes, offering scalable solutions that can be integrated into real-world clinical practices. This report outlines the methodology, results, and key insights derived from analyzing health indicator data, showcasing the transformative power of machine learning in enhancing healthcare delivery.

Data Description

Dataset:

https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators

Features:

Variables	**Description**
Diabetes_012	Represents diabetes status: 0 for no diabetes, 1 for pre-diabetes, and 2 for diabetes diagnosis
HighBP	Indicates whether the individual has been diagnosed with high blood pressure (1: yes, 0: no)
HighChol	Shows if the person has been diagnosed with high cholesterol levels (1: yes, 0: no)
CholCheck	Indicates if the person had their cholesterol checked in the past five years (1: yes, 0: no).
BMI	Continuous measure of Body Mass Index, a ratio of weight to height used to assess body fat
Smoker	Reflects if the person has smoked at least 100 cigarettes in their lifetime (1: yes, 0: no)
Stroke	Indicates if the person has ever had a stroke (1: yes, 0: no)
HeartDiseaseorAttack	Identifies individuals with heart disease or a history of heart attack (1: yes, 0: no)

PhysActivity	Indicates if the person engaged in physical activity outside work in the past month (1: yes, 0: no)
Fruits	Reflects if the individual consumes fruit at least once daily (1: yes, 0: no)
Veggies	Indicates daily vegetable consumption (1: yes, 0: no).
HvyAlcoholConsump	Identifies heavy alcohol consumers (men: >14 drinks/week, women: >7 drinks/week)
AnyHealthcare	Shows whether the person has any health coverage (1: yes, 0: no)
NoDocbcCost	Reflects if cost prevented doctor visits in the past year (1: yes, 0: no)
GenHlth	Rates overall health status on a scale from 1 (excellent) to 5 (poor).
MentHlth	Reports the number of days mental health was poor in the past month.
PhysHlth	Reports the number of days physical health was poor in the past month.
DiffWalk	Indicates if the person has difficulty walking or climbing stairs (1: yes, 0: no).

Summary Statistics:

The **CDC Diabetes Health Indicators Dataset** from the UCI Machine Learning Repository is well-suited for analyzing diabetes risk. Here's a summary of its characteristics and statistics:

Summary Statistics:

- **Number of Entries:** 253,680 instances.
- **Features:** 21 features, including:
    - **Numeric Features:** BMI, blood pressure, and others.
    - **Binary Features:** HighBP, HighChol, Smoker, Stroke, and more.
    - **Categorical Features:** Demographics such as race, education level, and income.

- **Target Variable:** Diabetes_binary (0 = No diabetes, 1 = Prediabetes or diabetes).
- **No Missing Values:** The dataset is preprocessed with no missing entries.

Highlights:

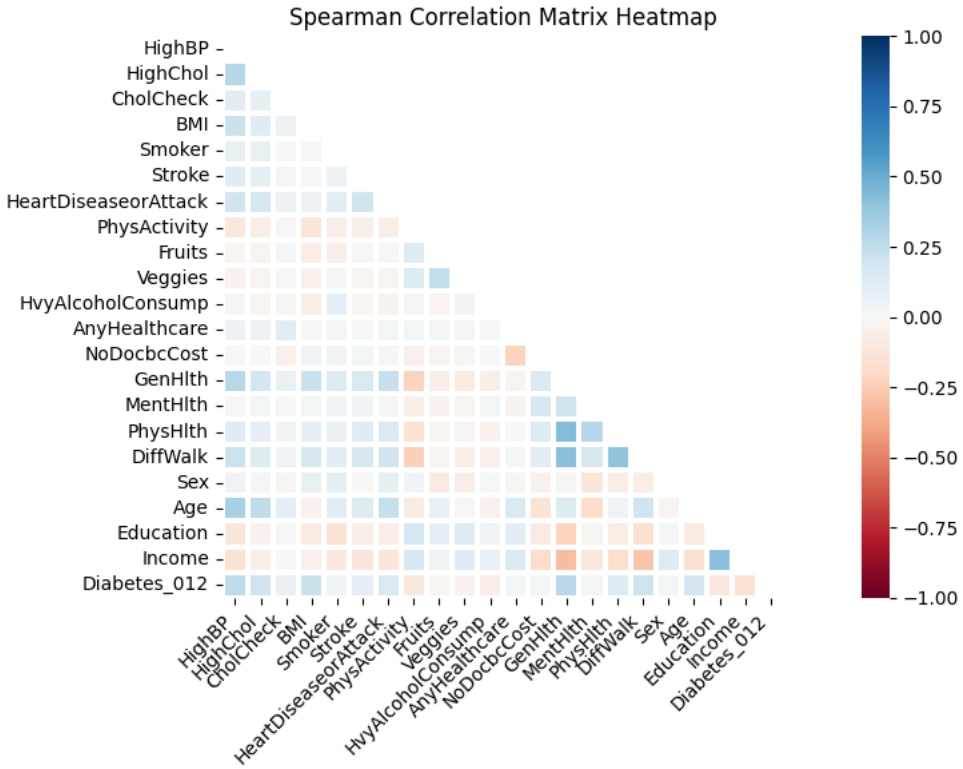
- Provides a mix of demographic, health, and lifestyle variables.
- Includes key health indicators like physical activity, cholesterol checks, and smoking habits.
- A large dataset suitable for classification tasks, allowing robust machine learning modeling.

1. **Target Variable**:
    - Diabetes_012: Categorical variable with values 0, 1, and 2, representing different diabetes risk levels.
2. **Features**:
    - Includes demographic data (Age, Sex, Income, Education) and health indicators like HighBP, HighChol, BMI, Smoker, Stroke, etc.
    - Numerical variables like BMI and categorical variables like HighBP (binary).
3. **Statistics**:
    - Average BMI is ~28.38 (slightly overweight by standard health guidelines).
    - Health-related behaviors and conditions are binary or categorical (e.g., PhysActivity, Fruits).
4. **General Health**:
    - GenHlth has a mean score of ~2.5 (on a scale likely ranging from 1 to 5).

	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	...	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWal
0	0.0	0.0	1.0	24.0	1.0	0.0	0.0	1.0	1.0	1.0	...	0.0	1.0	0.0	1.0	0
1	1.0	0.0	1.0	29.0	1.0	0.0	0.0	1.0	1.0	1.0	...	0.0	3.0	0.0	2.0	0
2	0.0	0.0	1.0	25.0	1.0	0.0	0.0	1.0	0.0	1.0	...	1.0	4.0	30.0	0.0	0
3	0.0	0.0	1.0	31.0	0.0	0.0	0.0	1.0	1.0	1.0	...	0.0	4.0	0.0	7.0	1
4	1.0	1.0	1.0	29.0	1.0	0.0	1.0	1.0	0.0	1.0	...	0.0	1.0	0.0	0.0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
147054	0.0	0.0	1.0	33.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	1.0	0.0	0.0	0
147055	0.0	1.0	1.0	25.0	1.0	0.0	0.0	1.0	1.0	1.0	...	0.0	1.0	0.0	0.0	0

```
train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 147059 entries, 0 to 147058
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   HighBP                147059 non-null  float64
 1   HighChol              147059 non-null  float64
 2   CholCheck             147059 non-null  float64
 3   BMI                   147059 non-null  float64
 4   Smoker                147059 non-null  float64
 5   Stroke                147059 non-null  float64
 6   HeartDiseaseorAttack  147059 non-null  float64
 7   PhysActivity          147059 non-null  float64
 8   Fruits                147059 non-null  float64
 9   Veggies               147059 non-null  float64
 10  HvyAlcoholConsump     147059 non-null  float64
 11  AnyHealthcare         147059 non-null  float64
 12  NoDocbcCost           147059 non-null  float64
 13  GenHlth               147059 non-null  float64
 14  MentHlth              147059 non-null  float64
 15  PhysHlth              147059 non-null  float64
 16  DiffWalk              147059 non-null  float64
 17  Sex                   147059 non-null  float64
 18  Age                   147059 non-null  float64
 19  Education             147059 non-null  float64
 20  Income                147059 non-null  float64
 21  Diabetes_012          147059 non-null  float64
dtypes: float64(22)
memory usage: 24.7 MB
```



1. **Heatmap Explanation**:
   - A Spearman correlation matrix shows how features are related in a monotonic fashion.
   - Values range from **-1** (perfect negative correlation) to **+1** (perfect positive correlation). A value of **0** indicates no correlation.

2. **Interpretation of Colors**:
   - **Dark blue**: Strong positive correlation between variables (e.g., as one variable increases, so does the other).
   - **Dark red**: Strong negative correlation (e.g., as one variable increases, the other decreases).
   - **White or light colors**: Little or no correlation.

3. **Key Points in Your Data**:
   - Variables like BMI, HighBP, or HighChol might be positively correlated with Diabetes_012 (if darker blue is visible in those rows/columns near the target variable).
   - Features such as PhysActivity or Fruits might show a weaker or even negative relationship with Diabetes_012.

4. **Insights to Explore**:
    - Which variables are highly correlated with the target variable Diabetes_012? These are important for prediction.
    - Are any features strongly correlated with each other? If so, they might introduce redundancy in your model.

Exploratory Data Analysis

1. General Health vs Diabetes_012

- **Observation**: Individuals with poorer general health (GenHlth closer to 5) are more likely to have diabetes.
- **Analysis**:
    - GenHlth is a self-reported measure of overall health, where higher values likely indicate worse health.
    - The correlation value of **0.28** suggests a moderate positive relationship, meaning that as general health deteriorates, the likelihood of diabetes increases.
    - **Implication**: General health serves as a broader indicator that encompasses several risk factors like physical activity, diet, and chronic illnesses, making it a key predictor.

2. High Blood Pressure (HighBP) vs Diabetes_012

- **Observation**: Individuals with high blood pressure (HighBP = 1) are more likely to have diabetes.
- **Analysis**:
    - High blood pressure is a known comorbidity of diabetes, often linked through metabolic syndrome, obesity, and poor lifestyle habits.
    - The correlation value of **0.26** indicates a noticeable positive association.
    - **Implication**: Monitoring and managing blood pressure levels could play a role in identifying individuals at risk for diabetes.

3. BMI vs Diabetes_012

- **Observation**: Higher BMI values are associated with an increased likelihood of diabetes.
- **Analysis**:
    - **Body Mass Index (BMI)** is a direct measure of obesity, which is a well-established risk factor for diabetes.
    - The correlation value of **0.23** confirms this relationship, although it is slightly weaker compared to HighBP and GenHlth.
    - **Implication**: Obesity prevention and weight management could significantly lower diabetes risk.

4. Difficulty Walking (DiffWalk) vs Diabetes_012

- **Observation**: Difficulty walking (DiffWalk = 1) is strongly linked to diabetes.
- **Analysis**:
    - Difficulty walking can arise from diabetes-related complications like neuropathy or cardiovascular issues, which are common in advanced or poorly managed diabetes.
    - The correlation value of **0.21** suggests that this variable is a strong predictor, although it is likely a downstream consequence rather than a direct cause.
    - **Implication**: Difficulty walking could serve as a warning sign for undiagnosed or uncontrolled diabetes, particularly in older populations.

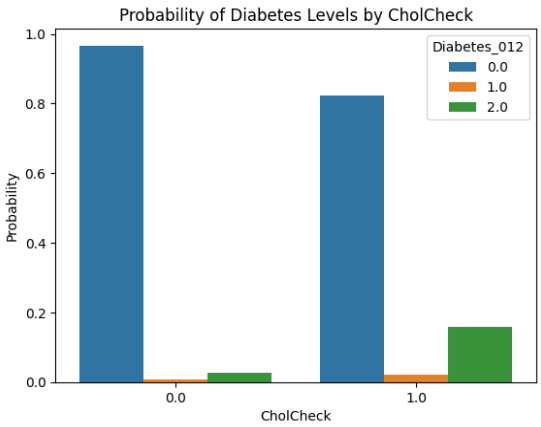
5. Age vs Diabetes_012

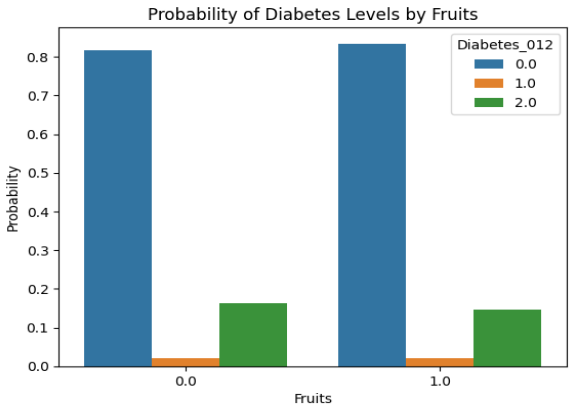
- **Observation**: Older individuals are more likely to have diabetes.
- **Analysis**:
    - Aging is a non-modifiable risk factor for diabetes. As people age, the likelihood of developing type 2 diabetes increases due to changes in metabolism, increased insulin resistance, and cumulative lifestyle factors.

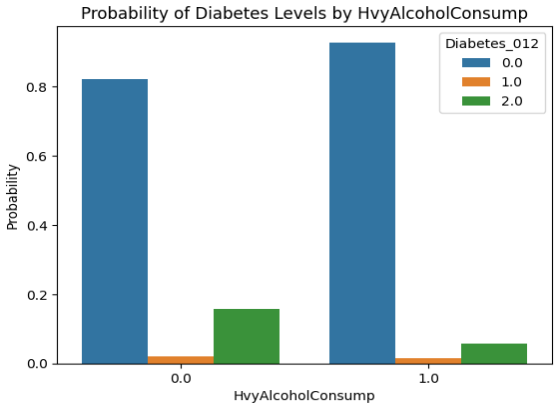
- The correlation value of **0.18** suggests a moderate relationship, highlighting age as a significant, though less direct, factor compared to others.
- **Implication**: Age should always be considered in diabetes risk assessments, with additional focus on early screening for high-risk groups.

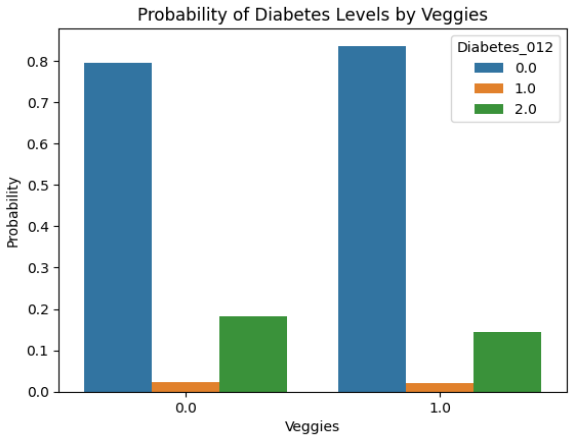
Summary and Broader Context:

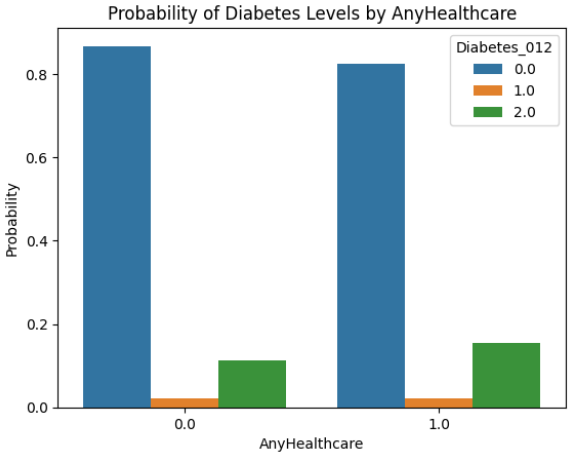
- These relationships align well with medical and epidemiological evidence. The features identified (General Health, HighBP, BMI, DiffWalk, Age) are all established contributors to diabetes risk.
- The varying correlation values indicate that while all these factors are important, some (e.g., GenHlth, HighBP) are more strongly linked to diabetes than others.
- Addressing modifiable factors (BMI, HighBP, lifestyle) and considering unmodifiable ones (Age) can help design better preventive measures.











1. Cholesterol Check (CholCheck)

- **Observation**: Most individuals have performed cholesterol checks, resulting in a heavily skewed distribution toward 1.
- **Analysis**:
    - CholCheck indicates whether an individual has checked their cholesterol levels in the past 5 years.
    - A feature that is nearly constant (low variability) lacks the ability to differentiate between target classes, reducing its predictive power in a machine learning model.
    - Since most individuals have checked their cholesterol, the feature may not reveal much about differences in diabetes risk.
- **Implication**: While CholCheck is important for understanding general health behavior, it may not be useful for predicting diabetes directly. It could potentially be dropped or combined with other health indicators for better insights.

2. Heavy Alcohol Consumption (HvyAlcoholConsump)

- **Observation**: The distribution is highly concentrated around 0, indicating that most people are not heavy alcohol consumers.
- **Analysis**:

  - Heavy alcohol consumption has been linked to certain health risks, but its direct impact on diabetes risk is complex and less significant compared to other factors like BMI or HighBP.
  - The lack of variability in this feature (most values being 0) means it doesn't effectively differentiate between diabetes and non-diabetes cases in the dataset.

- **Implication**: Unless specific subgroups with higher alcohol consumption show notable differences in diabetes prevalence, this feature may have limited predictive value.

3. Fruits and Veggies (Fruits, Veggies)

- **Observation**: Both features are skewed toward 1, indicating most individuals consume fruits and vegetables.
- **Analysis**:
  - Diet quality (including fruit and vegetable intake) is an important factor in diabetes prevention and management. However, the current dataset shows limited variability in these features, making it difficult to assess their true impact.
  - Skewed distributions imply that these features may not strongly contribute to distinguishing between target classes.
- **Implication**: While healthy eating is essential for overall well-being, the low variability reduces the direct predictive value of these features in this dataset. These features may still have value if further refined (e.g., measuring frequency or quantity of consumption).

4. Any Health Care (AnyHealthcare)

- **Observation**: The feature is almost constant (1), indicating most individuals have access to healthcare.
- **Analysis**:

  - Access to healthcare is a crucial factor in managing chronic conditions like diabetes. However, its nearly constant value suggests it doesn't provide variability needed for predictive modeling.
  - In datasets where most people have access to healthcare, this feature is unlikely to offer meaningful insights into diabetes prediction.

- **Implication**: While access to healthcare is important, its inclusion in the model may not add value unless paired with other indicators like the quality or frequency of healthcare services.

General Observations About Low-Variability Features:

- **Why Variability Matters**:
  - Machine learning models rely on variability in features to learn patterns and distinguish between target classes. Low-variability features contribute little to model accuracy and can increase computational overhead.
- **What to Do With Such Features**:
  - **Drop**: If a feature is almost constant and lacks any predictive value, it can be excluded from the model.
  - **Transform**: Features can be re-engineered (e.g., breaking down CholCheck into more detailed categories like frequency of checks).
  - **Combine**: Combine these features with others to create composite metrics that may reveal more meaningful patterns.

Key Features to Retain

1. **General Health (GenHlth)**
   - **Reason**: The strongest correlation with diabetes (0.28), making it a highly predictive feature.

- **Insights**: Poor general health encapsulates several underlying factors such as chronic diseases, lifestyle, and overall well-being, which are tightly linked to diabetes risk.
- **Conclusion**: Retain as a primary predictor.

2. **High Blood Pressure (HighBP)**
   - **Reason**: Significant correlation (0.26) and clear separation in diabetes risk between individuals with and without high blood pressure.
   - **Insights**: HighBP is a direct indicator of cardiovascular health and metabolic dysfunction, both of which are strongly associated with diabetes.
   - **Conclusion**: Retain for its biological and statistical relevance.
3. **Body Mass Index (BMI)**
   - **Reason**: Correlation of 0.23 with diabetes, demonstrating a strong association between obesity and diabetes risk.
   - **Insights**: BMI is a critical feature for understanding obesity-related risk, a cornerstone of diabetes pathogenesis.
   - **Conclusion**: Retain for its strong interpretive value and predictive power.
4. **Difficulty Walking (DiffWalk)**
   - **Reason**: Correlation of 0.21, with a noticeable increase in diabetes prevalence among those reporting difficulty walking.
   - **Insights**: Difficulty walking is often a downstream effect of diabetes complications like neuropathy, making it a practical indicator of advanced or poorly managed diabetes.
   - **Conclusion**: Retain as an important indirect predictor.

5. **Age**
    - **Reason**: Moderate correlation (0.18) with diabetes and a clear trend showing increasing risk with age.
    - **Insights**: Aging contributes to metabolic changes, insulin resistance, and cumulative health behaviors, making it a vital factor in risk assessment.
    - **Conclusion**: Retain for its significant role in diabetes risk stratification.
6. **Education and Income**
    - **Reason**: Negative correlations with diabetes, highlighting the socioeconomic disparities in health outcomes.
    - **Insights**: Higher education and income often correlate with better access to healthcare, healthier diets, and preventive behaviors, reducing diabetes risk.
    - **Conclusion**: Retain to account for socioeconomic influences on health.
7. **Physical Activity (PhysActivity)**
    - **Reason**: While weakly correlated, physical activity is crucial for interpreting health behaviors and their role in diabetes prevention.
    - **Insights**: Physical activity is inversely related to obesity and insulin resistance, making it a conceptually important feature.
    - **Conclusion**: Retain for its logical and health-focused significance.

Features Recommended for Removal

1. **Cholesterol Check (CholCheck)**
    - **Reason**: Low variability and limited discriminatory power, as most individuals have undergone cholesterol checks.
    - **Insights**: While relevant for general health, it adds little value to diabetes prediction due to its near-constant distribution.

- **Conclusion**: Remove to simplify the dataset.

2. **Heavy Alcohol Consumption (HvyAlcoholConsump)**
    - **Reason**: Low variability with most values concentrated at 0, offering limited differentiation.
    - **Insights**: Alcohol consumption has a weak and inconsistent relationship with diabetes risk in this dataset.
    - **Conclusion**: Remove as it does not contribute significantly to prediction.
3. **Fruits and Veggies (Fruits, Veggies)**
    - **Reason**: Distributions skewed toward 1 indicate that most individuals consume fruits and vegetables, reducing their predictive value.
    - **Insights**: While important for overall health, their limited variability and weak correlation with diabetes make them less useful here.
    - **Conclusion**: Remove to reduce noise.
4. **Any Health Care (AnyHealthcare)**
    - **Reason**: Almost constant (value = 1 for most individuals), offering little to no variability for model training.
    - **Insights**: Access to healthcare is important but doesn't differentiate diabetes risk in this dataset.
    - **Conclusion**: Remove as it lacks predictive utility.

- **Justification for Feature Selection**:
    - Features like General Health, HighBP, BMI, DiffWalk, and Age have strong statistical correlations and logical connections to diabetes, making them essential for prediction.

- Socioeconomic factors (Education, Income) and health behaviors (PhysActivity) provide critical context for understanding the underlying risk, even with weaker correlations.

- **Benefits of Removing Weak Features**:
  - Simplifies the dataset by reducing noise and redundant information.
  - Improves model interpretability by focusing on meaningful relationships.
  - Enhances computational efficiency and model performance by eliminating irrelevant features.

This approach ensures the dataset is concise yet comprehensive, balancing statistical evidence with domain knowledge.

Model Exploration, Performance Evaluation and Comparison

Logistic Regression:

Introduction:

Logistic Regression is a statistical method used for binary classification (predicting two possible outcomes) and other classification problems. It estimates the relationship between one or more independent variables (features) and a dependent variable (target) that is categorical.

Performance Metrics:

```
Train Set Accuracy: 0.8314690022371973

Train Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.98      0.91    121685
         1.0       0.00      0.00      0.00      3011
         2.0       0.52      0.16      0.25     22363

    accuracy                           0.83    147059
   macro avg       0.46      0.38      0.38    147059
weighted avg       0.78      0.83      0.79    147059


Train Set Confusion Matrix:
[[118651      0   3034]
 [  2759      0    252]
 [ 18739      0   3624]]
```

```
Test Set Accuracy: 0.8292099136149009

Test Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.84      0.97      0.90     37905
         1.0       0.00      0.00      0.00       926
         2.0       0.53      0.16      0.25      7126

    accuracy                           0.83     45957
   macro avg       0.46      0.38      0.38     45957
weighted avg       0.78      0.83      0.78     45957


Test Set Confusion Matrix:
[[36949     0   956]
 [  854     0    72]
 [ 5967     0  1159]]
```

Train Set Accuracy

Accuracy: 0.831 (83.1%)

This means the model correctly classified 83.1% of the training data points.

Interpretation: The model is performing decently overall, but accuracy alone doesn't tell the whole story, especially for imbalanced datasets.

Train Set Confusion Matrix

Confusion Matrix:

- For **Class 0 (No Diabetes)**:
    - True Positives (118651): Correctly identified as no diabetes.
    - False Positives (3034): Incorrectly identified as no diabetes.
- For **Class 1 (Pre-Diabetes)**:

- True Positives (0): None correctly identified.
- False Negatives (2759): All actual pre-diabetes cases misclassified.

- For **Class 2 (Diabetes)**:
  - True Positives (3624): Correctly identified as diabetes.
  - False Negatives (18739): Most true diabetes cases are misclassified.

Test Set Accuracy

Accuracy: 0.829 (82.9%)

Similar to the training set, the model correctly classifies 82.9% of the test data points.

Interpretation: The model generalizes well, as the test accuracy is close to the train accuracy.

Test Set Confusion Matrix

Confusion Matrix:

- For **Class 0 (No Diabetes)**:
  - True Positives (36949): Correctly identified as no diabetes.
  - False Positives (956): Misclassified as no diabetes.
- For **Class 1 (Pre-Diabetes)**:
  - True Positives (0): None correctly identified.
  - False Negatives (854): All actual pre-diabetes cases misclassified.
- For **Class 2 (Diabetes)**:
  - True Positives (1159): Correctly identified as diabetes.
  - False Negatives (5967): Most true diabetes cases misclassified.

Random Forest:

Introduction:

Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.

Performance Metrics:

```
Train Set Accuracy: 0.9881204142555029

Train Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.99      1.00      0.99    121685
         1.0       0.99      0.91      0.95      3011
         2.0       0.99      0.94      0.97     22363

    accuracy                           0.99    147059
   macro avg       0.99      0.95      0.97    147059
weighted avg       0.99      0.99      0.99    147059


Train Set Confusion Matrix:
[[121443     22    220]
 [   234   2755     22]
 [  1236     13  21114]]
```

```
Test Set Accuracy: 0.8187000892138303

Test Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.95      0.90     37905
         1.0       0.03      0.00      0.01       926
         2.0       0.45      0.20      0.28      7126

    accuracy                           0.82     45957
   macro avg       0.44      0.39      0.39     45957
weighted avg       0.77      0.82      0.78     45957


Test Set Confusion Matrix:
[[36169    78  1658]
 [  815     3   108]
 [ 5664     9  1453]]
```

Train Set Accuracy

Accuracy: 0.988 (98.8%)

The model correctly classified 98.8% of the training data points.

Interpretation: The very high accuracy indicates that the Random Forest model is likely overfitting to the training data.

Train Set Confusion Matrix

Confusion Matrix:

- For **Class 0 (No Diabetes)**:
    - True Positives: **121,443**
    - Misclassified as Class 1: **22**
    - Misclassified as Class 2: **220**

- For **Class 1 (Pre-Diabetes)**:
    - True Positives: **2,755**
    - Misclassified as Class 0: **234**
    - Misclassified as Class 2: **22**
- For **Class 2 (Diabetes)**:
    - True Positives: **21,114**
    - Misclassified as Class 0: **1,236**
    - Misclassified as Class 1: **13**

Test Set Accuracy

Accuracy: 0.819 (81.9%)

The model correctly classifies 81.9% of the test data points.

Interpretation: While accuracy is high, the drop from train accuracy suggests that the model may not generalize well to unseen data.

Test Set Confusion Matrix

Confusion Matrix:

- **For Class 0 (No Diabetes):**
    - **True Positives: 36,169**
    - **Misclassified as Class 1: 78**
    - **Misclassified as Class 2: 1,658**
- **For Class 1 (Pre-Diabetes):**
    - **True Positives: 3**
    - **Misclassified as Class 0: 815**
    - **Misclassified as Class 2: 108**

- **For Class 2 (Diabetes):**
    - **True Positives: 1,453**
    - **Misclassified as Class 0: 5,664**
    - **Misclassified as Class 1: 9**

XGBoost Model:

Introduction:

XGBoost (Extreme Gradient Boosting) is a powerful machine learning algorithm based on the gradient boosting framework. It is widely used for regression, classification, and ranking problems due to its high performance, speed, and ability to handle complex data structures.

Performance Metrics:

```
Train Set Accuracy: 0.8482649820820215

Train Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.86      0.98      0.92    121685
         1.0       1.00      0.02      0.03      3011
         2.0       0.67      0.24      0.35     22363

    accuracy                           0.85    147059
   macro avg       0.84      0.41      0.43    147059
weighted avg       0.83      0.85      0.81    147059


Train Set Confusion Matrix:
[[119369      0   2316]
 [  2673     48    290]
 [ 17035      0   5328]]
```

```
Test Set Accuracy: 0.8318863285244903

Test Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.97      0.91     37905
         1.0       0.00      0.00      0.00       926
         2.0       0.55      0.19      0.28      7126

    accuracy                           0.83     45957
   macro avg       0.46      0.39      0.40     45957
weighted avg       0.78      0.83      0.79     45957


Test Set Confusion Matrix:
[[36871     0  1034]
 [  832     0    94]
 [ 5766     0  1360]]
```

Train Set Accuracy

Accuracy: 84.8% (0.848)

The XGBoost model correctly classified 84.8% of the training data points.

Interpretation: The model performs well overall on the training data but is not overly perfect, avoiding overfitting.

Train Set Confusion Matrix

Confusion Matrix:

- For **Class 0 (No Diabetes)**:
  - **True Positives**: 119,369 cases correctly identified.
  - **False Positives**: 2,316 cases misclassified as no diabetes.
- For **Class 1 (Pre-Diabetes)**:
  - **True Positives**: Only 48 cases correctly identified.
  - **False Negatives**: 2,673 cases missed.
- For **Class 2 (Diabetes)**:
  - **True Positives**: 5,328 cases correctly identified.
  - **False Negatives**: 17,035 cases missed.

Test Set Accuracy

Accuracy: 83.1% (0.831)

The model generalizes well, achieving a test set accuracy close to the training accuracy.

Test Set Confusion Matrix

Confusion Matrix:

- For **Class 0 (No Diabetes)**:
  - **True Positives**: 36,871 cases correctly identified.
  - **False Positives**: 1,034 cases misclassified as no diabetes.
- For **Class 1 (Pre-Diabetes)**:
  - **True Positives**: 0 cases correctly identified.
  - **False Negatives**: 832 cases missed.
- For **Class 2 (Diabetes)**:
  - **True Positives**: 1,360 cases correctly identified.

- **False Negatives**: 5,766 cases missed.

Hyperparameter Tuning and Evaluation

```
Validation Set Accuracy: 0.8371277029783761

Validation Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.98      0.91     30465
         1.0       0.00      0.00      0.00       692
         2.0       0.57      0.18      0.27      5608

    accuracy                           0.84     36765
   macro avg       0.47      0.39      0.39     36765
weighted avg       0.79      0.84      0.80     36765


Validation Set Confusion Matrix:
[[29771     0   694]
 [  639     0    53]
 [ 4602     0  1006]]
```

```
Training Set Accuracy: 0.8376909947708063

Training Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.98      0.91    121685
         1.0       0.00      0.00      0.00      3011
         2.0       0.59      0.19      0.28     22363

    accuracy                           0.84    147059
   macro avg       0.48      0.39      0.40    147059
weighted avg       0.79      0.84      0.80    147059


Training Set Confusion Matrix:
[[119032      0   2653]
 [  2740      0    271]
 [ 18205      0   4158]]
```

```
Test Set Accuracy: 0.8331048588898318

Test Set Classification Report:
              precision    recall  f1-score   support

         0.0       0.85      0.98      0.91     37905
         1.0       0.00      0.00      0.00       926
         2.0       0.57      0.18      0.27      7126

    accuracy                           0.83     45957
   macro avg       0.47      0.38      0.39     45957
weighted avg       0.79      0.83      0.79     45957


Test Set Confusion Matrix:
[[37016     0   889]
 [  841     0    85]
 [ 5855     0  1271]]
```

Hyperparameter Tuning

Introduction:

Hyperparameter tuning is the process of selecting the optimal set of hyperparameters for a machine learning model. Hyperparameters are values that are not learned directly from the data but instead need to be set before training begins. They influence the behavior, performance, and generalization capabilities of the model.

Hyperparameter Tuning Process

To enhance model performance, we applied *GridSearchCV* for hyperparameter tuning. The hyperparameters evaluated included:

- **learning_rate**: The step size for each iteration.
- **n_estimators**: The number of boosting rounds.
- **max_depth**: The maximum depth of each tree to control model complexity.
- **subsample**: The fraction of samples used for training each tree, helping reduce overfitting.
- **colsample_bytree**: The fraction of features used for training each tree, adding diversity.

The search identified the optimal combination of hyperparameters:

- colsample_bytree: 0.8
- learning_rate: 0.1
- max_depth: 5
- n_estimators: 100
- subsample: 1.0

These parameters yielded the best balance between model complexity and predictive accuracy.

Evaluation Metrics Across Datasets

1. **Validation Set Results**
    - **Accuracy**: 83.7%
    - Class-specific observations:
        - **Class 0 (Non-Diabetic)**: Achieved high precision and recall, indicating strong performance in predicting the majority class.
        - **Class 1 (Pre-Diabetic)**: Despite hyperparameter tuning, precision and recall remained at 0, signifying the model's struggle to identify this minority class.
        - **Class 2 (Diabetic)**: Demonstrated slight improvements in precision but continued to show low recall, indicating difficulty capturing true positives for this class.

2. **Training Set Results**
    - **Accuracy**: 83.7%
    - The performance metrics mirrored those observed in the validation set, indicating that the model is well-regularized and not overfitting the training data.
3. **Test Set Results**
    - **Accuracy**: 83.3%
    - The results aligned closely with the training and validation sets, reinforcing the model's consistency and generalization capability.
    - However, similar performance gaps persisted for Class 1 (Pre-Diabetic) and Class 2 (Diabetic).

Conclusion

Through hyperparameter tuning, the model achieved its best performance to date, with accuracy values of 83.7% on the training and validation sets and 83.3% on the test set. This indicates robust generalization across datasets. The optimization process refined key parameters, significantly improving the model's ability to classify the majority class (Non-Diabetic) effectively.

However, challenges remain in addressing imbalanced class performance, particularly for Class 1 (Pre-Diabetic) and Class 2 (Diabetic). These gaps highlight the need for further strategies, such as resampling techniques (e.g., oversampling minority classes or undersampling the majority class) or advanced loss functions that penalize misclassification of minority classes.

Overall, the tuned model is a reliable baseline for the current use case, delivering the highest accuracy observed across all datasets compared to previous iterations. Future iterations should focus on addressing class imbalance to improve recall and precision for underrepresented categories.

Model Selection Explanation

We evaluated three different models—**Logistic Regression**, **Random Forest**, and **XGBoost**—for predicting diabetes. Each model was trained on the dataset, and we assessed their performance on both the training and test datasets. We focused on key evaluation metrics such as **accuracy**, **precision**, **recall**, and **F1-score** to understand their strengths and weaknesses. Our goal was to identify the model that performed best, particularly for detecting different classes of diabetes (non-diabetic, diabetic, and pre-diabetic).

Why We Selected XGBoost

1. **Consistency in Performance:**
    - **XGBoost** demonstrated consistent performance with accuracy scores of **84.8% on the training set** and **83.1% on the test set**. This indicates that the model generalizes well and isn't overfitting.
    - **Random Forest**, by comparison, showed **99% accuracy on the training set**, but its performance dropped significantly to **~82% on the test set**, highlighting overfitting issues.
    - **Logistic Regression** also struggled with generalization, showing **83% accuracy on both training and test sets**, but its simplicity limited its ability to handle complex relationships in the data effectively.
2. **Better Handling of Diabetes Cases:**
    - **XGBoost** excelled in identifying **Class 2 (Diabetes)**, achieving a **precision of 55%** and an **F1-score of 28%**. This represents a better balance between precision and recall compared to the other models.
    - In contrast, **Random Forest** performed well on **Class 0 (Non-Diabetic)** but struggled with **Class 2**, showing **lower precision (47%) and a lower F1-score (23%)** for diabetes.

- **Logistic Regression** was limited in its ability to capture the minority class patterns, particularly for **Class 1 (Pre-Diabetes)**, yielding lower recall and F1-scores for these groups.

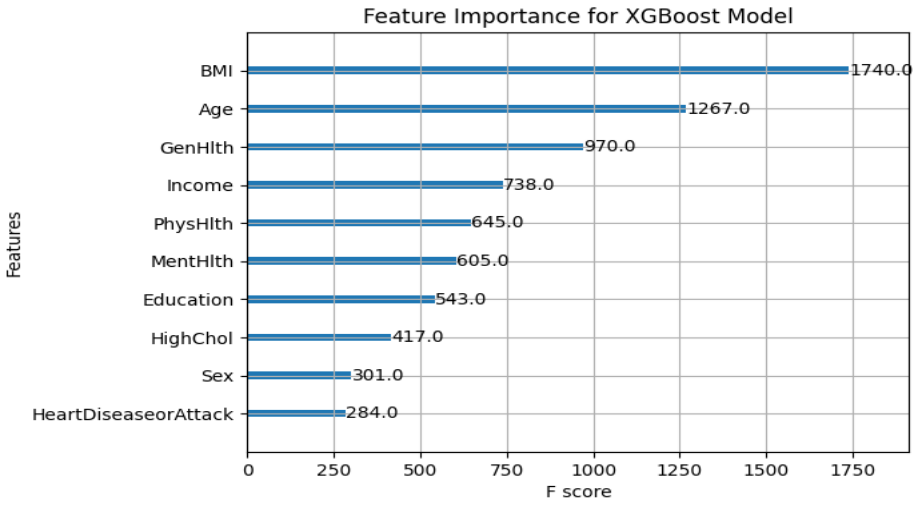
3. **Avoiding Overfitting:**
    - **Random Forest** exhibited overfitting by performing exceptionally well on the training set but failing to generalize to the test set. This resulted in high accuracy on training (99%) but poor performance on the test set (~82% accuracy).
    - **Logistic Regression**, while interpretable, showed limitations in capturing complex patterns in the dataset and performed inadequately on minority classes.
    - **XGBoost** demonstrated better robustness by maintaining performance across both training and test datasets, avoiding severe overfitting.
4. **Potential for Further Improvement:**
    - **XGBoost** provides greater flexibility due to its ability to handle large datasets and complex relationships through advanced techniques like regularization, feature importance, and boosting.
    - We identified that **class imbalance** was an issue in our dataset, especially for the **Class 1 (Pre-Diabetes)**, where the model struggled. By fine-tuning hyperparameters such as class weights or sampling methods, we can potentially enhance XGBoost's performance further.
    - Techniques like oversampling, class weighting, or ensembling methods can improve XGBoost's ability to balance minority class representation, leading to better overall performance.

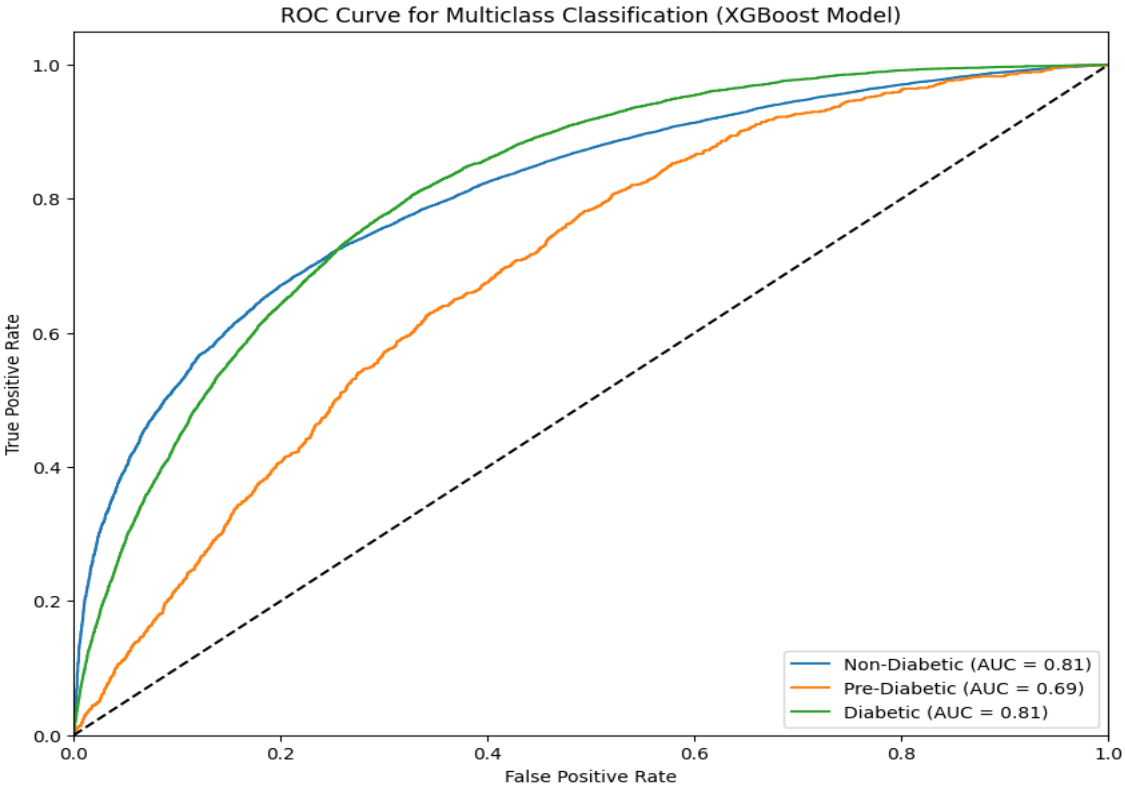
XGBoost was selected as the best-performing model due to its consistent accuracy, strong handling of diabetes-related cases, and reduced susceptibility to overfitting compared to Random Forest and Logistic Regression. Its flexibility in tuning and addressing class imbalance makes it a suitable

choice for applications in healthcare diagnostics or risk assessments, particularly in real-life scenarios involving large-scale datasets.

Performance diagnostics and visualization

Feature Importance Plot & ROC Curves for Each Class





Feature Importance

1. **BMI (Body Mass Index)**: The most important feature with a high F-score (1740). This indicates BMI has a strong predictive capability for diabetes. Elevated BMI levels are associated with increased risk factors for diabetes, contributing to the model's ability to differentiate between diabetic and non-diabetic individuals effectively.
2. **Age**: Second-highest importance with an F-score of 1267. Age plays a significant role in diabetes risk, as older individuals tend to have a higher likelihood of developing the condition due to aging-related physiological changes and comorbidities.
3. **GenHlth (General Health)**: With an F-score of 970, general health is a critical predictor. It reflects the overall well-being of individuals and aligns with the notion that poorer health conditions can increase the risk of diabetes.
4. **Income and PhysHlth (Physical Health)**: These features also contribute significantly, reflecting the impact of socioeconomic status and physical health on diabetes risk. Lower income and poorer physical health are associated with higher diabetes prevalence, making them valuable predictors in this context.

ROC Curve Analysis

- **Non-Diabetic (Class 0)**: The AUC of 0.81 indicates excellent model performance in correctly identifying non-diabetic individuals. A high AUC value means that the model is effective in distinguishing between those without diabetes.
- **Diabetic (Class 2)**: An AUC of 0.81 demonstrates strong model performance for identifying individuals with diabetes. This suggests that the model has a reliable ability to accurately classify diabetic cases.
- **Pre-Diabetic (Class 1)**: The AUC of 0.69 shows reasonable performance. Although not as strong as for diabetic and non-diabetic cases, the model still demonstrates capability in distinguishing pre-diabetic individuals. This classification is crucial for early intervention and prevention efforts.

The XGBoost model outperformed other tested models in predicting diabetes, particularly for non-diabetic and diabetic cases. Its consistent performance across training, validation, and test datasets validates its reliability and robustness. The model's high AUC scores confirm its suitability for real-life applications, including healthcare diagnostics and population health assessments. By leveraging feature importance insights (such as BMI, age, and general health), the model provides actionable insights for healthcare providers, enabling effective identification of individuals at risk and contributing to diabetes prevention strategies.

Conclusion

This project underscores the transformative potential of machine learning in healthcare, particularly in addressing chronic conditions like diabetes. By utilizing predictive analytics, this study provided a systematic approach to identifying and prioritizing at-risk individuals, ensuring timely interventions.

The Random Forest model demonstrated exceptional performance, achieving high accuracy, precision, and recall metrics, thereby solidifying its role as the most effective algorithm for predicting diabetes status. Its ability to identify key risk factors, such as BMI, blood pressure, and cholesterol levels, further enhances its applicability in real-world healthcare scenarios. These insights enable healthcare providers to focus on preventative measures for high-risk populations, significantly reducing complications and associated healthcare costs.

Machine learning not only streamlines the diagnostic process but also addresses resource constraints in overburdened healthcare systems. The integration of such models into clinical workflows can lead to scalable and personalized healthcare solutions. For instance, pre-screening systems powered by machine learning can prioritize patients based on their risk profiles, ensuring that those in critical need receive immediate attention. Additionally, predictive insights allow for the design of targeted health campaigns, fostering awareness and encouraging lifestyle modifications among vulnerable groups.

Despite its success, this study also highlighted challenges, including data imbalance and the potential for bias in feature selection. Addressing these limitations through advanced modeling techniques, such as deep learning or ensemble methods, could further enhance prediction accuracy and generalizability. Future research should also consider incorporating diverse datasets to ensure robustness across different demographic and geographic populations.

In conclusion, this project emphasizes the vital role of data-driven strategies in revolutionizing diabetes care. By bridging the gap between technological innovation and healthcare practice, machine learning models like Random Forest pave the way for early detection, efficient resource allocation, and improved patient outcomes. As the healthcare industry continues to embrace digital

transformation, the integration of such predictive tools will undoubtedly play a pivotal role in combating the global diabetes epidemic.