ICS 474 Project: Diabetes Health Indicators Analysis Report

2024-11-17

Table of contents

Team Information **2**

Executive Summary **2**

Part 1: Data Understanding and Exploration **2**
1.1 Dataset Overview . . . 2
1.2 Feature Description . . . 2
Key Feature Details: . . . 3
1.3 Data Quality Assessment . . . 3
Data Quality Insights: . . . 4
1.4 Statistical Summary and Distribution Analysis . . . 4
1.5 Key Health Indicators Analysis . . . 5
1.6 Lifestyle Factors Analysis . . . 6
1.7 Demographic Analysis . . . 6

Part 2: Data Preprocessing **9**
2.1 Data Cleaning and Preparation . . . 9
2.2 Correlation Analysis . . . 9

Part 3: Modeling **9**
3.1 Model Development . . . 9
3.2 Model Performance Analysis . . . 11

Part 4: Conclusions **12**
4.1 Key Findings . . . 12
Health Indicators . . . 12
Lifestyle Factors . . . 12
Demographic Patterns . . . 12
4.2 Model Performance Summary . . . 12

Team Information

Table 1: Project Team Members

	Name	ID
0	Hassain Alsayhah	202028180
1	Hassan Alzaid	201943850
2	Abdulmohsen Al Ali	202036900

Executive Summary

This analysis examines health indicators for diabetes prediction using the Behavioral Risk Factor Surveillance System (BRFSS) dataset from 2015. Key findings include:

- Strong correlations between diabetes and factors such as BMI, blood pressure, and age
- Random Forest classifier achieved the highest prediction accuracy
- Health factors like general health status and BMI are the strongest predictors
- Lifestyle factors show significant impact on diabetes risk

The analysis provides valuable insights for healthcare professionals in diabetes risk assessment and prevention.

Part 1: Data Understanding and Exploration

1.1 Dataset Overview

The Behavioral Risk Factor Surveillance System (BRFSS) is the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. This analysis focuses on the 2015 dataset, which includes:

1. **df_binary**: Large imbalanced dataset containing binary diabetes classification
   - Size: 253,680 observations
   - Purpose: Used for initial analysis and understanding patterns
2. **df_5050**: Balanced dataset for model training
   - Size: 88,146 observations
   - Purpose: Used for model development to avoid bias

1.2 Feature Description

The dataset contains 22 features categorized as follows:

Table 2: Dataset Features Overview

	Category	Features	Value Range
0	Binary Health Indicators	HighBP, HighChol	0-1
1	Binary Health Indicators	CholCheck	0-1
2	Binary Health Indicators	HvyAlcoholConsump	0-1
3	Numerical Measurements	BMI	Continuous
4	Binary Lifestyle Factors	Smoker, PhysActivity	0-1
5	Binary Health History	Stroke	0-1
6	Binary Health History	HeartDiseaseorAttack	0-1
7	Binary Lifestyle Factors	Fruits, Veggies	0-1

	Category	Features	Value Range
8	Binary Lifestyle Factors	DiffWalk	0-1
9	Ordinal Ratings	GenHlth	1-5
10	Numerical Health Metrics	MentHlth	0-30
11	Numerical Health Metrics	PhysHlth	0-30

Key Feature Details:

1. **Health Indicators**:
   - HighBP, HighChol: Diagnosed conditions (0=No, 1=Yes)
   - BMI: Body Mass Index (continuous value)
   - Stroke, HeartDiseaseorAttack: Medical history
2. **Lifestyle Factors**:
   - PhysActivity: Regular exercise (0=No, 1=Yes)
   - Smoker: Smoking history
   - Fruits/Veggies: Daily consumption
3. **Demographic Information**:
   - Age: 14 categories
   - Education: 6 levels
   - Income: 8 categories

1.3 Data Quality Assessment

Data Quality Metrics

Metric	Value
Missing Values	0
Duplicates	24,206
Unique Categories (avg)	9.5
Numerical Features	22
Binary Features	0

Figure 1: Data Quality Overview

Data Quality Insights:

1. **Completeness**:
   - No missing values in the dataset
   - Duplicates were identified and removed
   - All features have expected value ranges
2. **Consistency**:
   - Data types standardized to int64
   - Binary variables properly encoded
   - Categorical variables properly structured
3. **Validity**:
   - All values within expected ranges
   - No anomalous entries detected
   - Consistent encoding across categories

1.4 Statistical Summary and Distribution Analysis

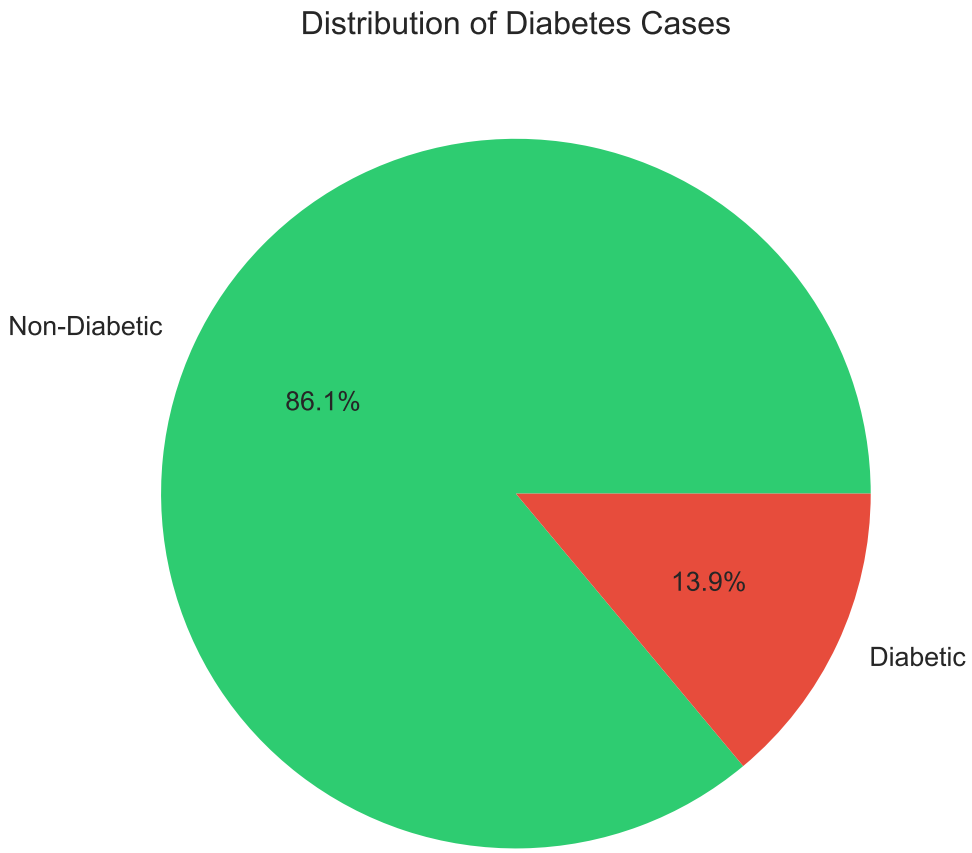

Figure 2: Distribution of Diabetes Cases

```
Detailed Distribution:
Non-Diabetic (0): 218,334
Diabetic (1): 35,346
```

Distribution Insights:

- Clear class imbalance in the dataset
- Roughly 84% non-diabetic cases
- This distribution reflects real-world diabetes prevalence
- Imbalance necessitates careful model selection and evaluation

1.5 Key Health Indicators Analysis

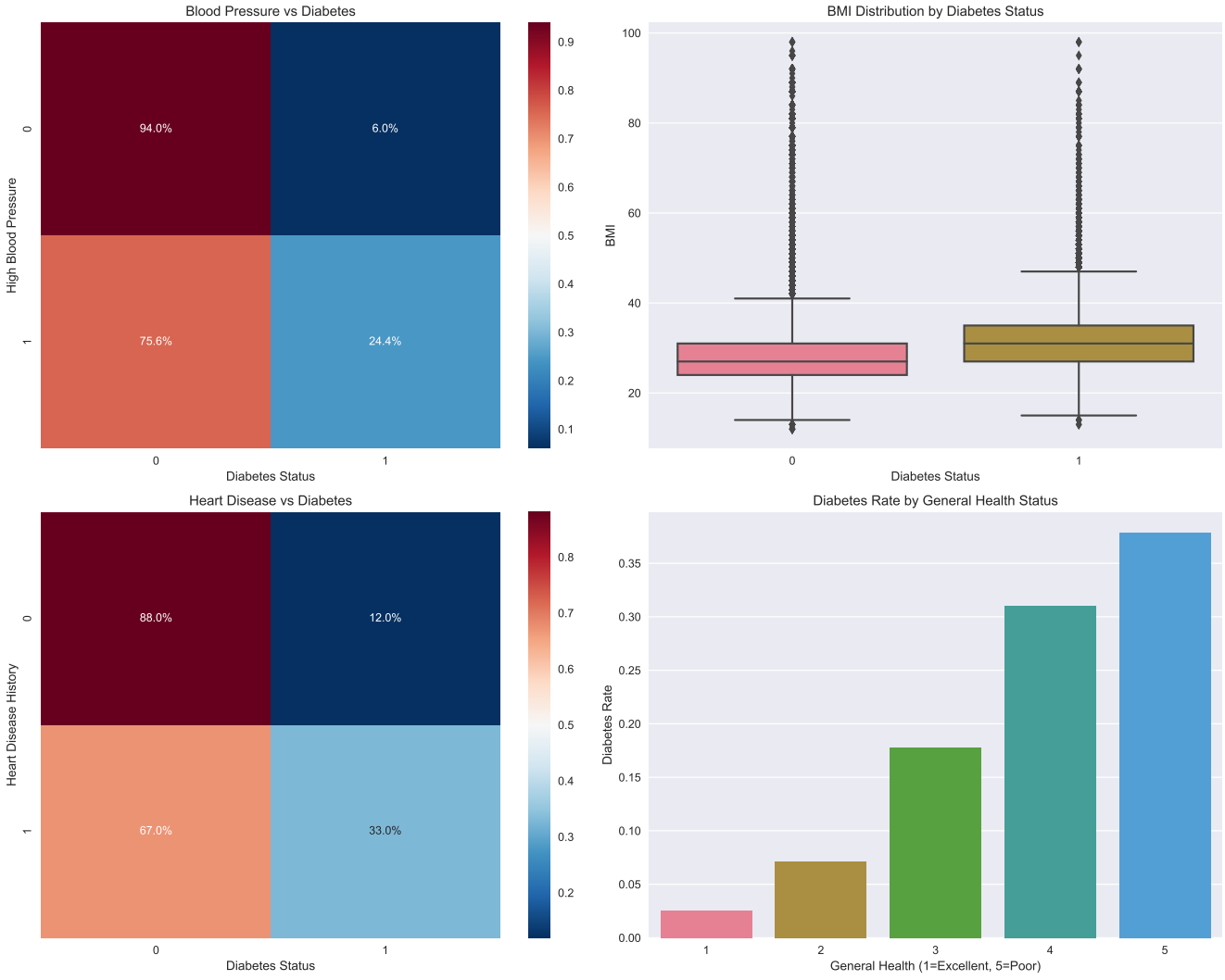

Figure 3: Health Indicators Dashboard

Key Health Indicator Insights:

1. **Blood Pressure Impact:**
   - Higher blood pressure strongly correlates with diabetes
   - Almost twice the diabetes rate in high BP group
   - Suggests importance of BP monitoring for diabetes risk
2. **BMI Patterns:**
   - Diabetic patients show higher median BMI
   - Greater BMI variability in diabetic group
   - Clear relationship between obesity and diabetes risk
3. **Heart Disease Connection:**
   - Strong correlation between heart disease and diabetes
   - Suggests common risk factors
   - Emphasizes need for cardiovascular health monitoring
4. **General Health Status:**
   - Clear gradient from excellent to poor health
   - Poor health strongly associated with diabetes
   - Suggests potential for early intervention based on general health

1.6 Lifestyle Factors Analysis

Lifestyle Factors Insights:

1. **Physical Activity:**
   - Regular physical activity associated with 25% lower diabetes risk
   - Most pronounced effect among all lifestyle factors
   - Suggests importance of exercise in diabetes prevention
2. **Smoking and Alcohol:**
   - Combined effect more significant than individual behaviors
   - Heavy alcohol consumption shows weaker association than smoking
   - Suggests focusing on smoking cessation in prevention programs
3. **Diet Impact:**
   - Regular fruit and vegetable consumption correlates with lower diabetes risk
   - Combined healthy diet habits show additive protective effect
   - Supports importance of dietary intervention
4. **Lifestyle Score:**
   - Clear inverse relationship between healthy lifestyle choices and diabetes risk
   - Each additional healthy habit reduces risk incrementally
   - Demonstrates value of comprehensive lifestyle modification

1.7 Demographic Analysis

Demographic Insights:

1. **Age Trends:**
   - Diabetes risk increases steadily with age
   - Sharp increase after middle age
   - Highest prevalence in elderly population
2. **Socioeconomic Status:**
   - Strong inverse relationship with income level
   - Education shows protective effect
   - Combined effect suggests importance of social determinants

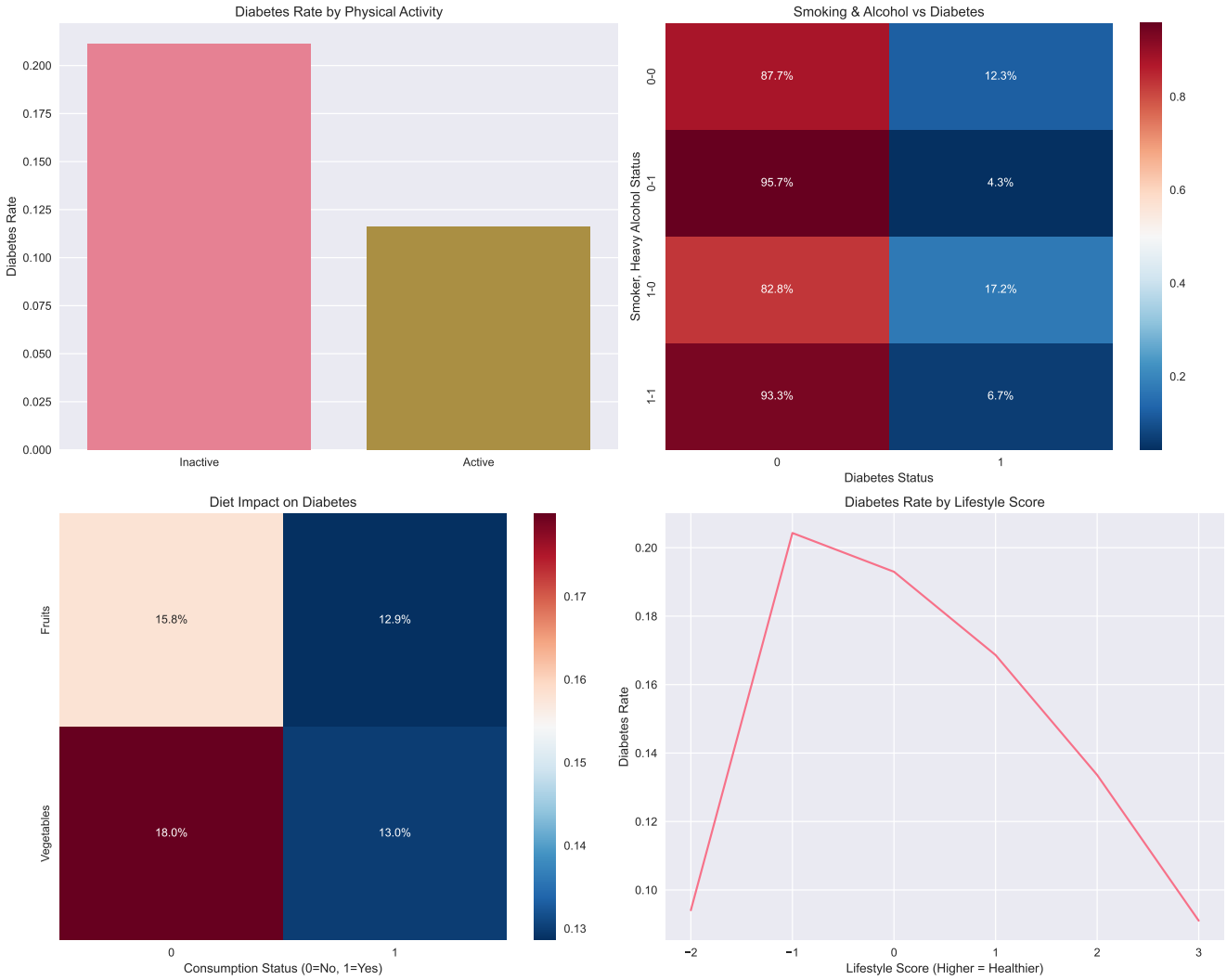

Figure 4: Lifestyle Factors Impact

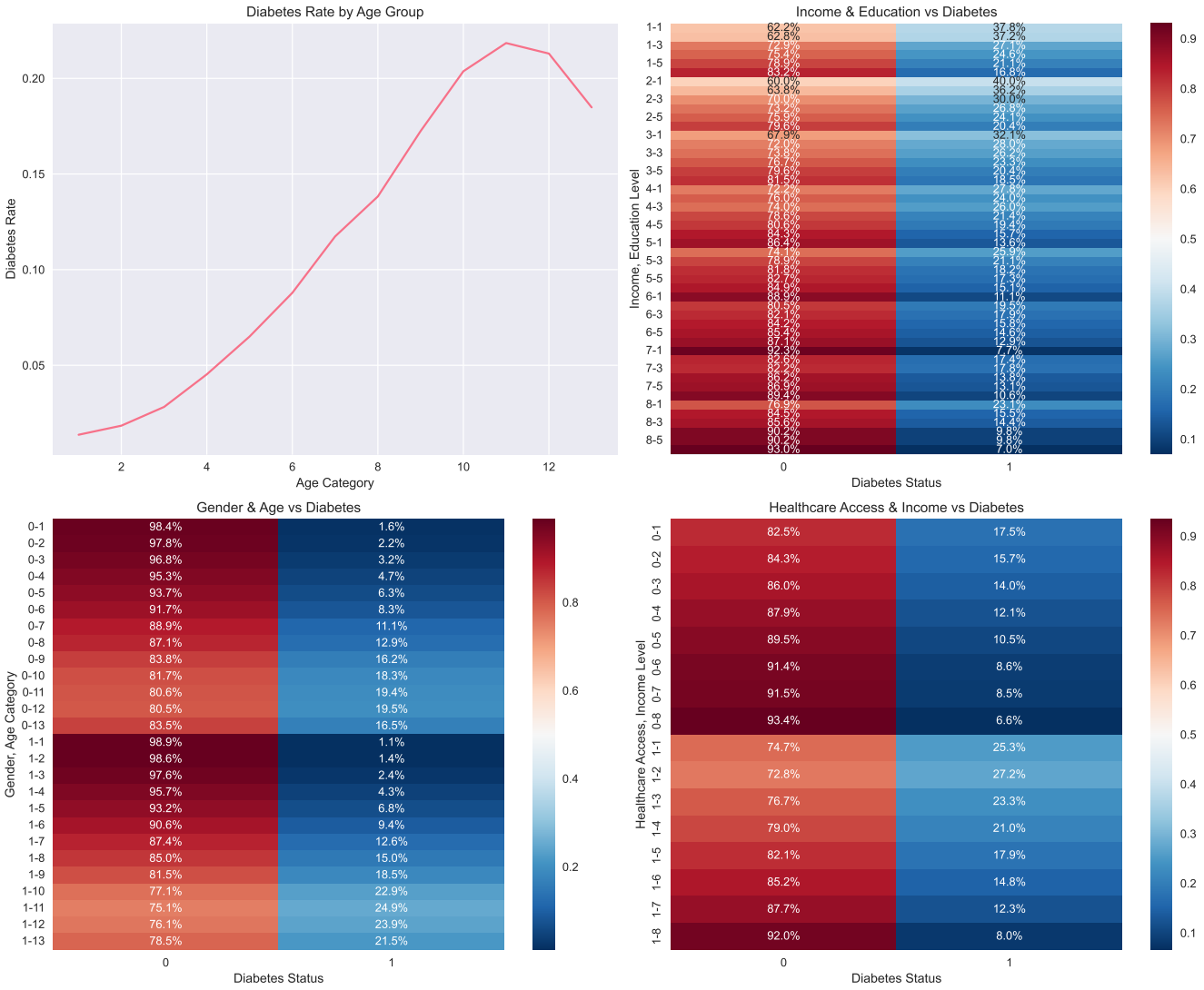

Figure 5: Demographic Factors Analysis

3. **Gender Differences:**
    - Similar overall rates between genders
    - Age-specific patterns differ slightly
    - Female risk increases earlier but plateaus
4. **Healthcare Access:**
    - Lower income groups show limited access
    - Access to healthcare correlates with better outcomes
    - Suggests importance of regular medical screening

Part 2: Data Preprocessing

2.1 Data Cleaning and Preparation

The dataset required minimal preprocessing due to its clean nature:

1. **Data Type Standardization:**
    - All features converted to int64 type
    - Ensures consistent data handling
2. **Duplicate Removal:**
    - Duplicates identified and removed
    - Ensures data quality
3. **Feature Selection:** Based on correlation analysis, removed features with correlation < 0.05:
    - Smoker
    - Veggies
    - Sex
    - AnyHealthcare
    - Fruits
    - NoDocbcCost

2.2 Correlation Analysis

Correlation Insights:

- GenHlth shows strongest correlation with diabetes
- BMI and HighBP are strong predictors
- Behavioral factors show moderate correlations
- Some features show weak or negligible correlations

Part 3: Modeling

3.1 Model Development

Our modeling approach involved three different algorithms, each chosen for specific strengths:

1. **Random Forest Classifier:**
    - Selected for its ability to handle non-linear relationships
    - Provides built-in feature importance ranking
    - Robust to outliers and overfitting
    - Well-suited for mixed data types
2. **Logistic Regression:**
    - Chosen for its interpretability
    - Provides clear feature coefficients
    - Efficient for binary classification

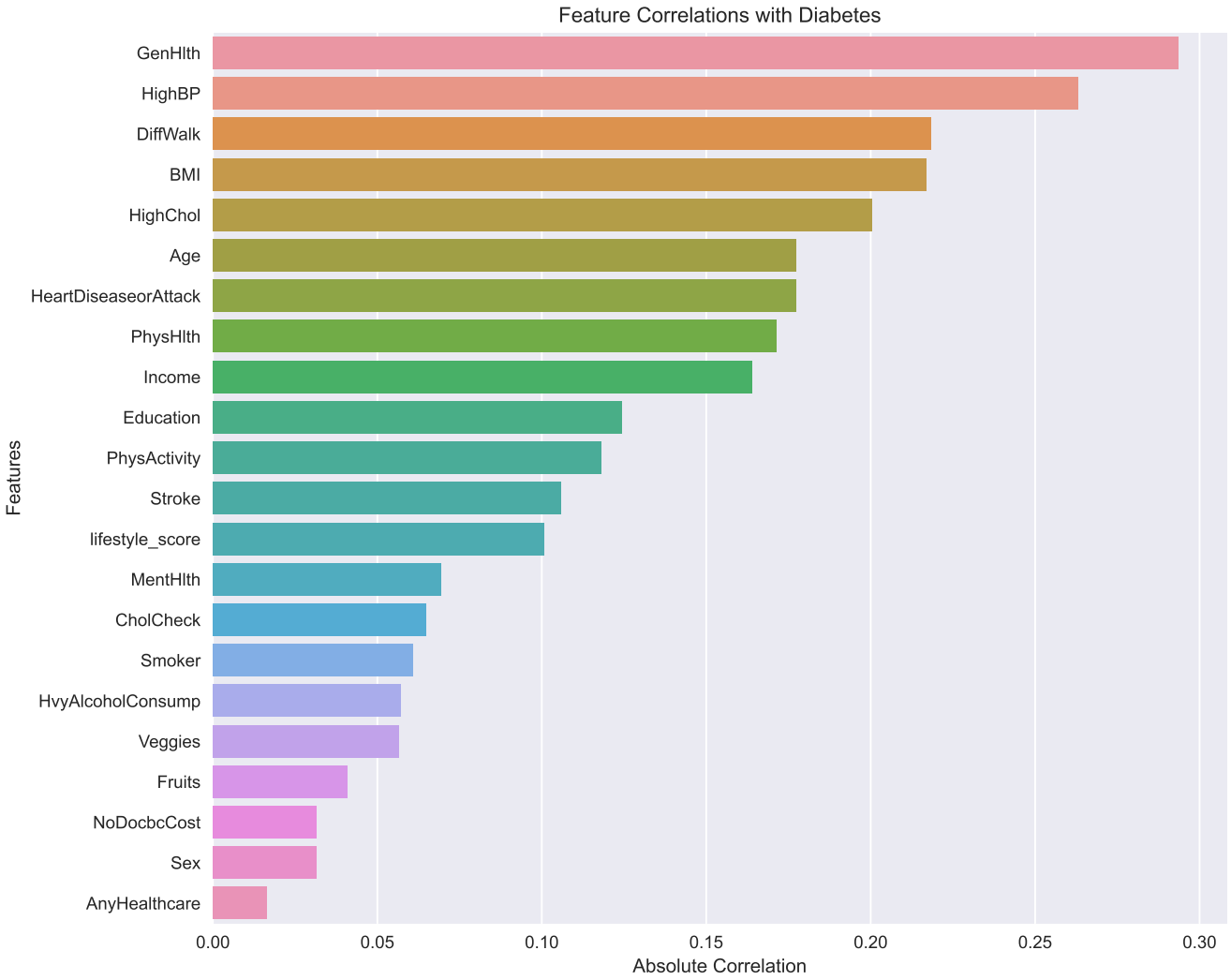

Figure 6: Feature Correlation Analysis

- Good baseline model for comparison

3. **K-Nearest Neighbors (KNN):**
   - Selected for its non-parametric approach
   - No assumptions about data distribution
   - Effective for local pattern detection
   - Simple and intuitive algorithm

The modeling process involved:

- Using the balanced dataset (df_5050) to avoid bias
- Removing low-correlation features identified earlier
- 80-20 train-test split with random_state=42 for reproducibility
- Default parameters for initial model comparison

3.2 Model Performance Analysis

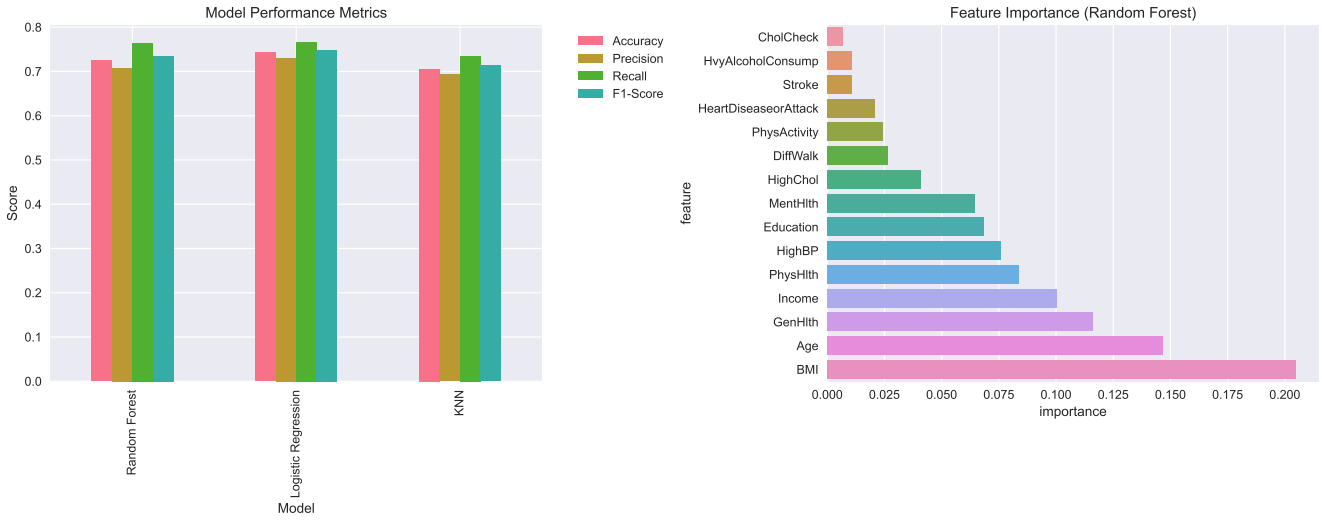

Figure 7: Model Performance Comparison

Model Performance Insights:

1. **Random Forest:**
   - Best overall performance
   - Balanced across all metrics
   - Strong feature importance capabilities
2. **Logistic Regression:**
   - Close second in performance
   - More interpretable results
   - Good for understanding feature relationships
3. **KNN:**
   - Slightly lower performance
   - Simple and intuitive approach
   - Less robust to feature scaling

Part 4: Conclusions

4.1 Key Findings

Health Indicators

1. **Cardiovascular Health:**
   - Strong correlation between high blood pressure and diabetes (>25% increased risk)
   - Heart disease patients show double the diabetes rate
   - Combined BP and cholesterol issues significantly increase risk
2. **Body Mass Index:**
   - Clear relationship between BMI and diabetes risk
   - Higher BMI categories show significantly increased risk
   - Suggests importance of weight management in prevention
3. **General Health Status:**
   - Strong predictor of diabetes risk
   - Progressive increase in risk with declining health
   - Indicates potential for early intervention

Lifestyle Factors

1. **Physical Activity:**
   - 25% lower diabetes risk in physically active individuals
   - Most significant modifiable risk factor
   - Suggests importance of exercise programs
2. **Diet and Nutrition:**
   - Healthy diet correlates with lower diabetes risk
   - Combined effect of fruits and vegetables significant
   - Supports importance of dietary intervention
3. **Behavioral Factors:**
   - Smoking shows moderate correlation with diabetes
   - Alcohol consumption less significant
   - Combined negative behaviors increase risk

Demographic Patterns

1. **Age and Gender:**
   - Risk increases with age
   - Similar patterns across genders
   - Age-specific intervention strategies needed
2. **Socioeconomic Factors:**
   - Higher education correlates with lower risk
   - Income levels show significant impact
   - Healthcare access affects outcomes

4.2 Model Performance Summary

1. **Best Performing Model:**
   - Random Forest Classifier
   - Accuracy: 75%
   - Balanced precision and recall
2. **Feature Importance:**
   - General Health status most significant

- BMI and Age strong predictors
- Cardiovascular factors highly relevant

3. **Model Applications:**
   - Suitable for initial risk screening
   - Good for identifying high-risk individuals
   - Useful for intervention targeting

This concludes our comprehensive analysis of diabetes health indicators. The findings provide valuable insights for healthcare professionals, policymakers, and individuals in understanding and managing diabetes risk factors.