DIABETES HEALTH INDICATORS ANALYSIS!

TO

TABLE OF CONTENTS

TO

01

UNDERSTANDING AND EXPLORATION

DATASET OVERVIEW

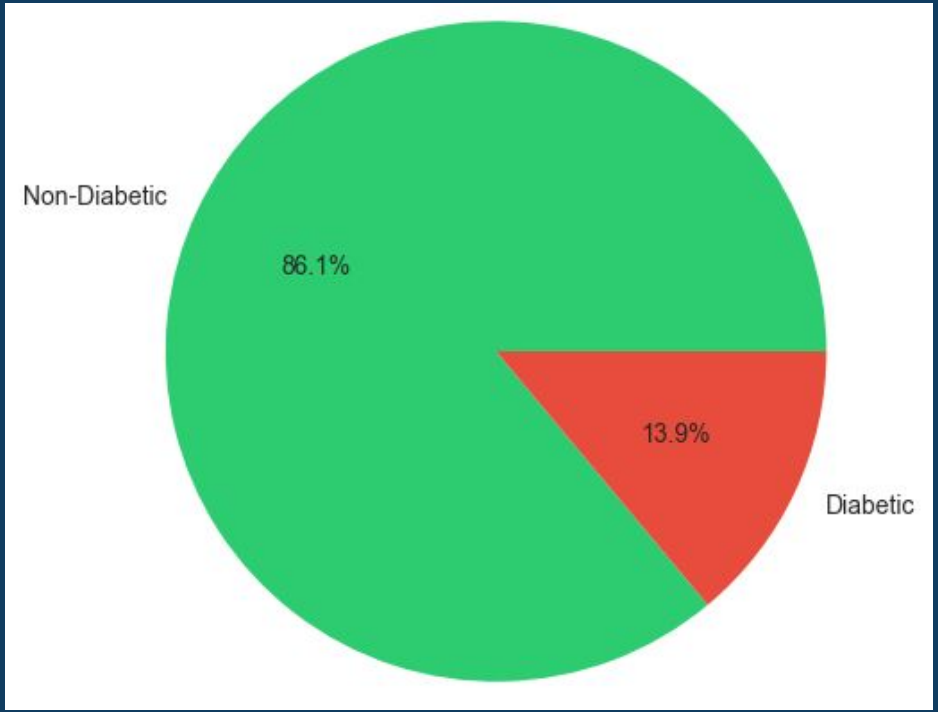
- **df_binary:** Large imbalanced dataset containing binary diabetes classification
  - • **Size:** 253,680 observations
  - • **Purpose:** Used for initial analysis and understanding patterns

- **df_5050:** Balanced dataset for model training
  - • **Size:** 88,146 observations
  - • **Purpose:** Used for model development to avoid bias

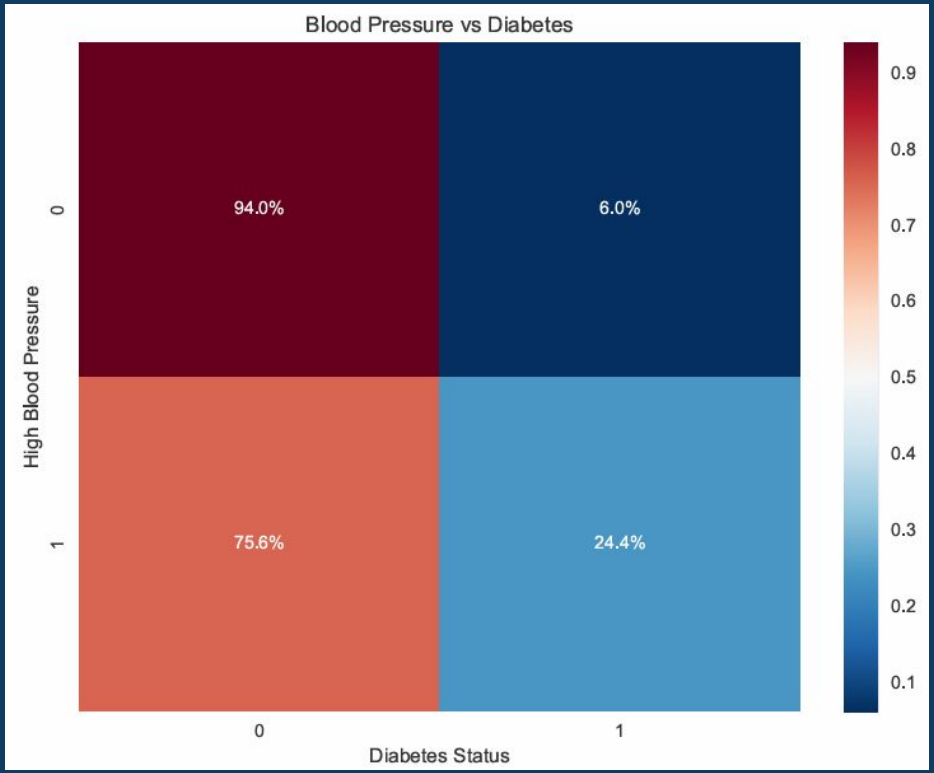
KEY FEATURE DETAILS

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

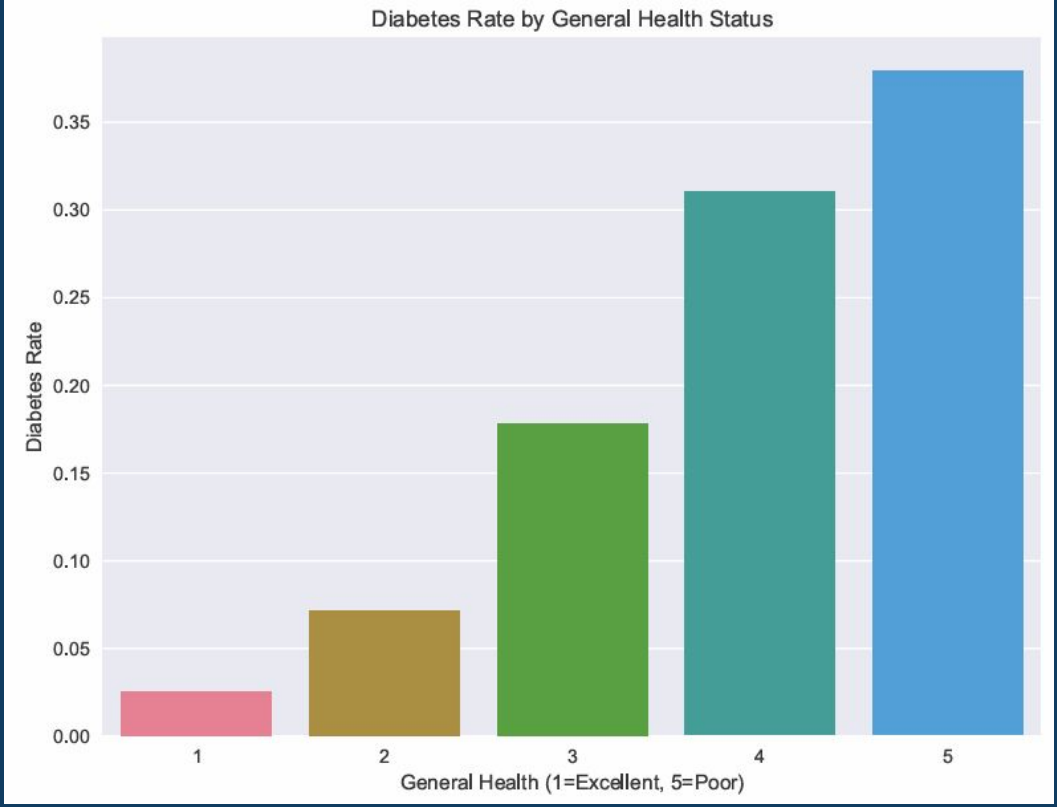
STATISTICAL SUMMARY AND DISTRIBUTION ANALYSIS



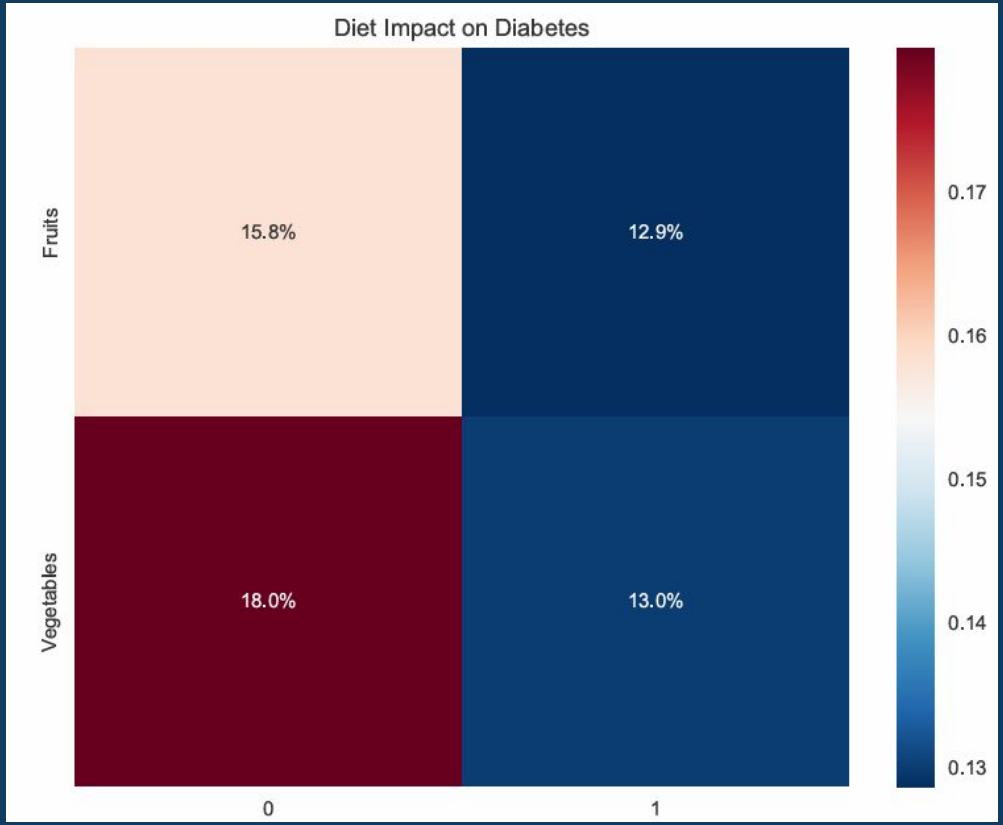
KEY HEALTH INDICATORS ANALYSIS



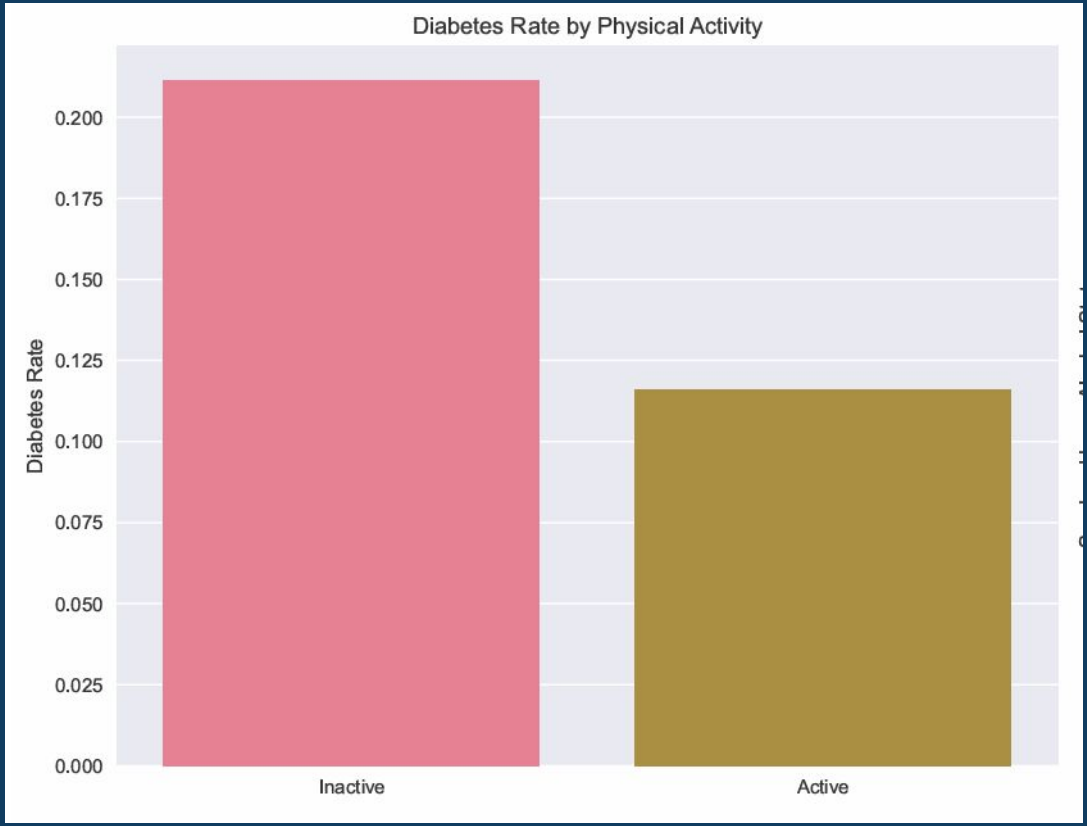
KEY HEALTH INDICATORS ANALYSIS



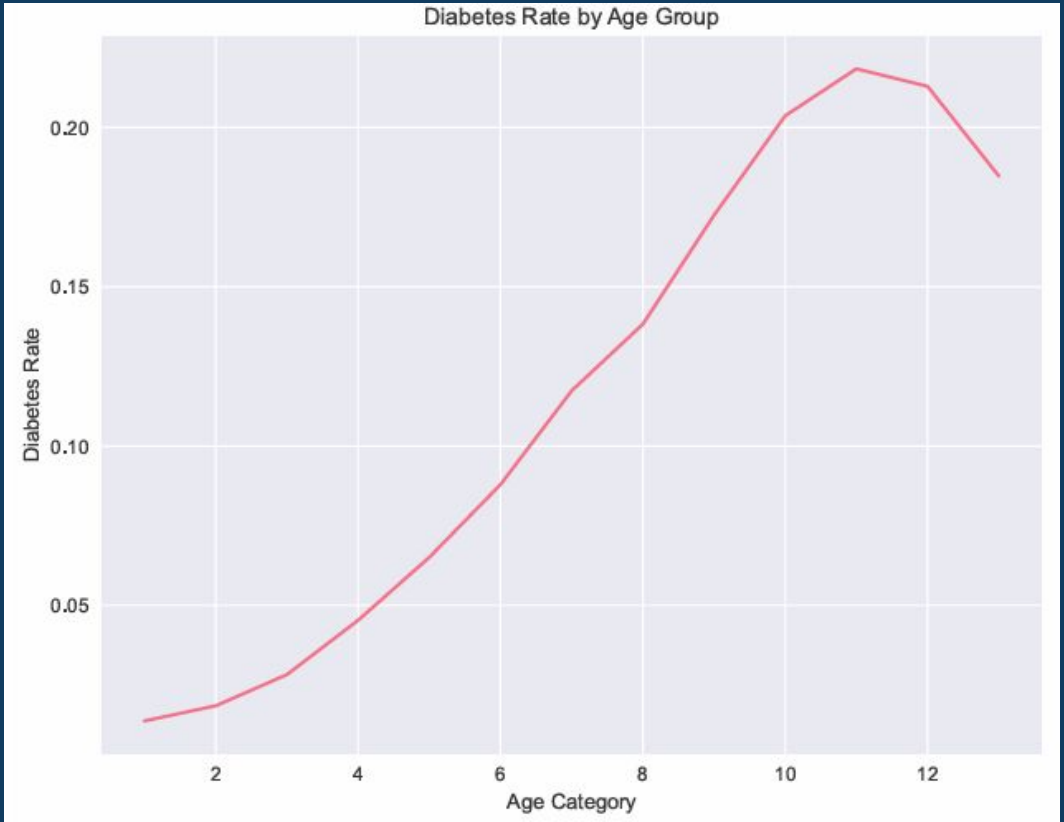
LIFESTYLE FACTORS ANALYSIS



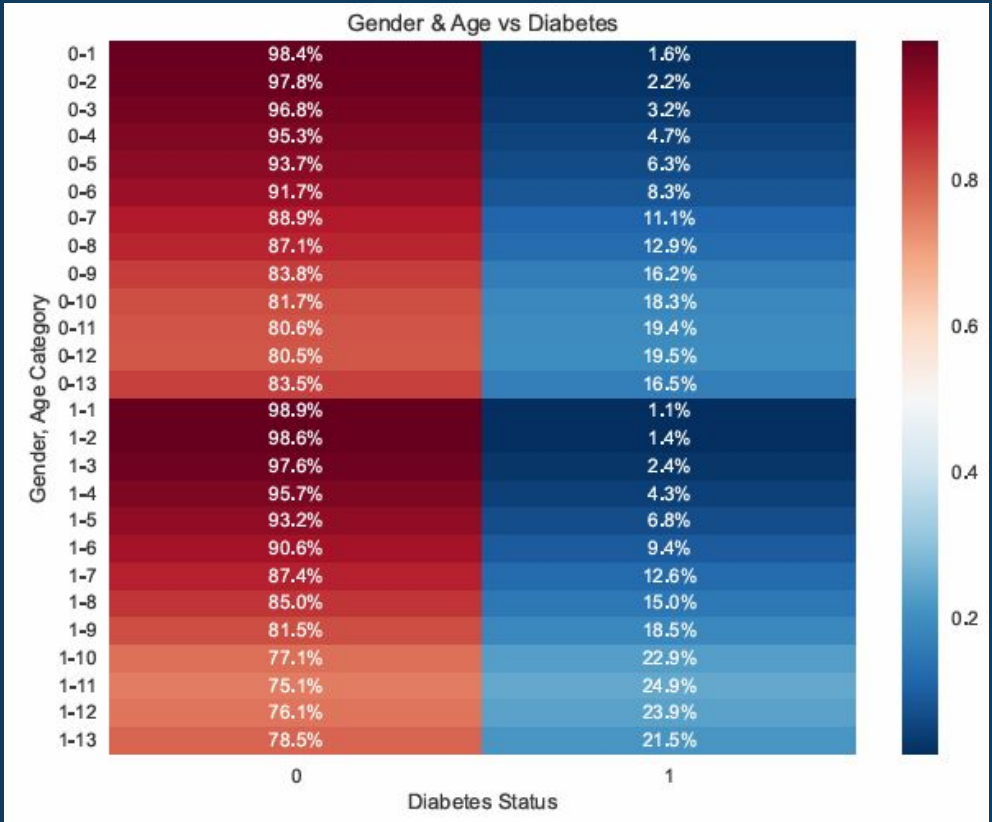
LIFESTYLE FACTORS ANALYSIS



DEMOGRAPHIC ANALYSIS



DEMOGRAPHIC ANALYSIS



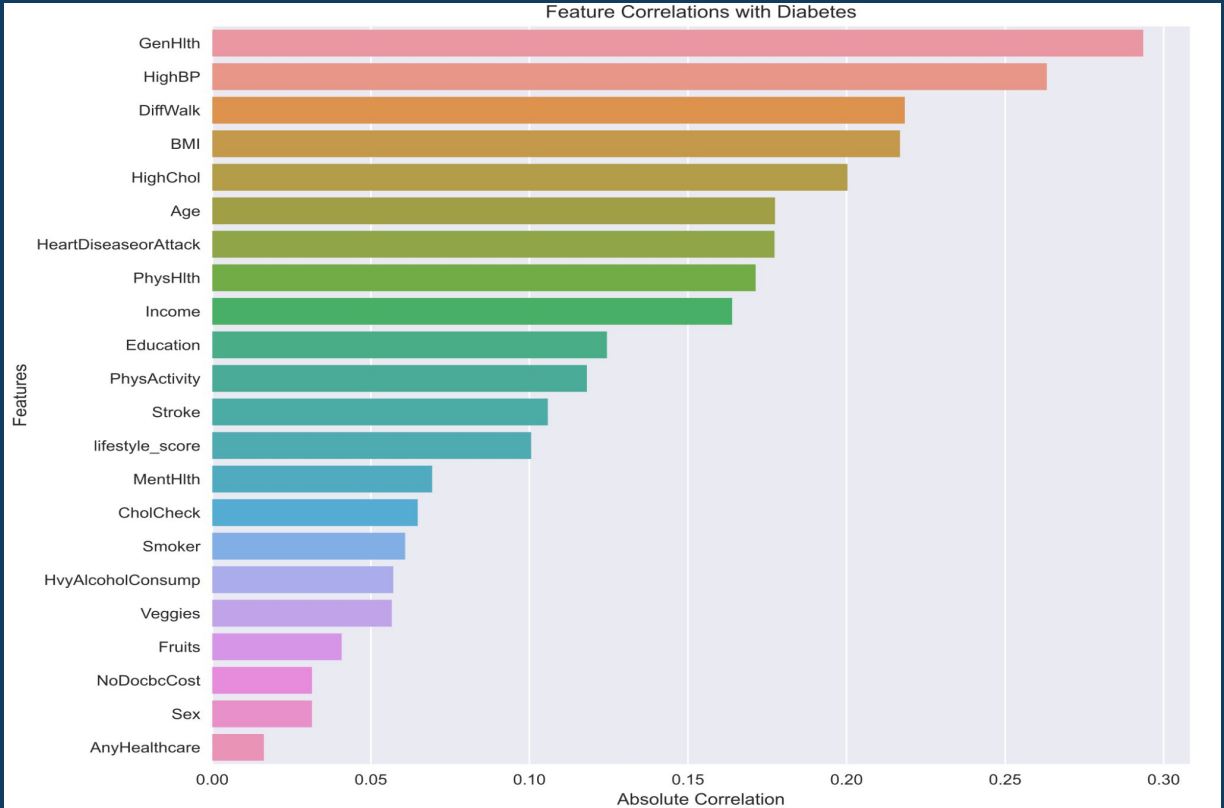
02

DATA PREPROCESSING

DATA CLEANING AND PREPARATION

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

CORRELATION ANALYSIS



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

03

MODELING

MODEL DEVELOPMENT

1. Random Forest Classifier:
    a. Selected for its ability to handle non-linear relationships
    b. Provides built-in feature importance ranking
    c. Robust to outliers and overfitting
    d. Well-suited for mixed data types

MODEL DEVELOPMENT

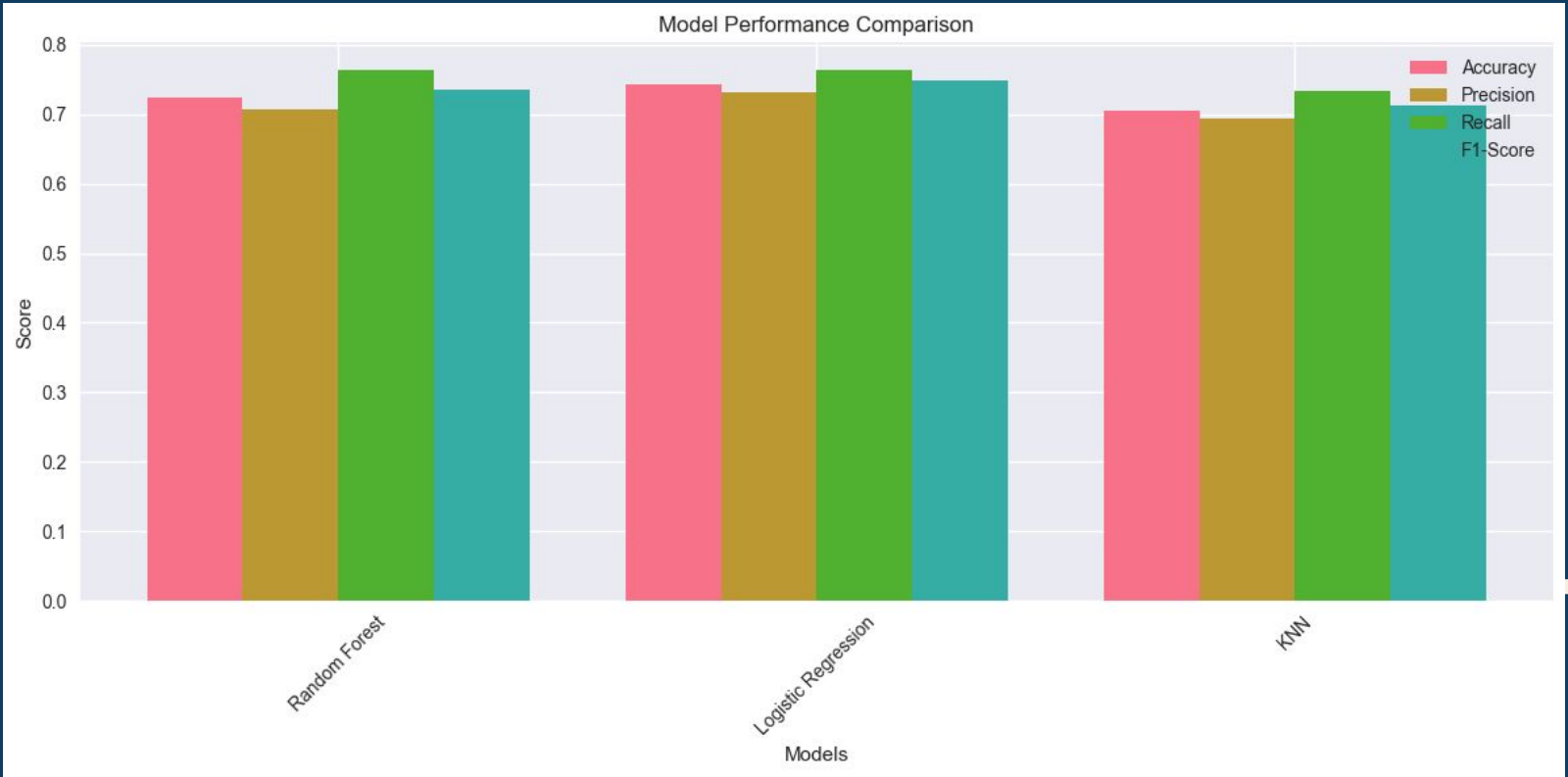
2. Logistic Regression:

a. Chosen for its interpretability
b. Provides clear feature coefficients
c. Efficient for binary classification
d. Good baseline model for comparison

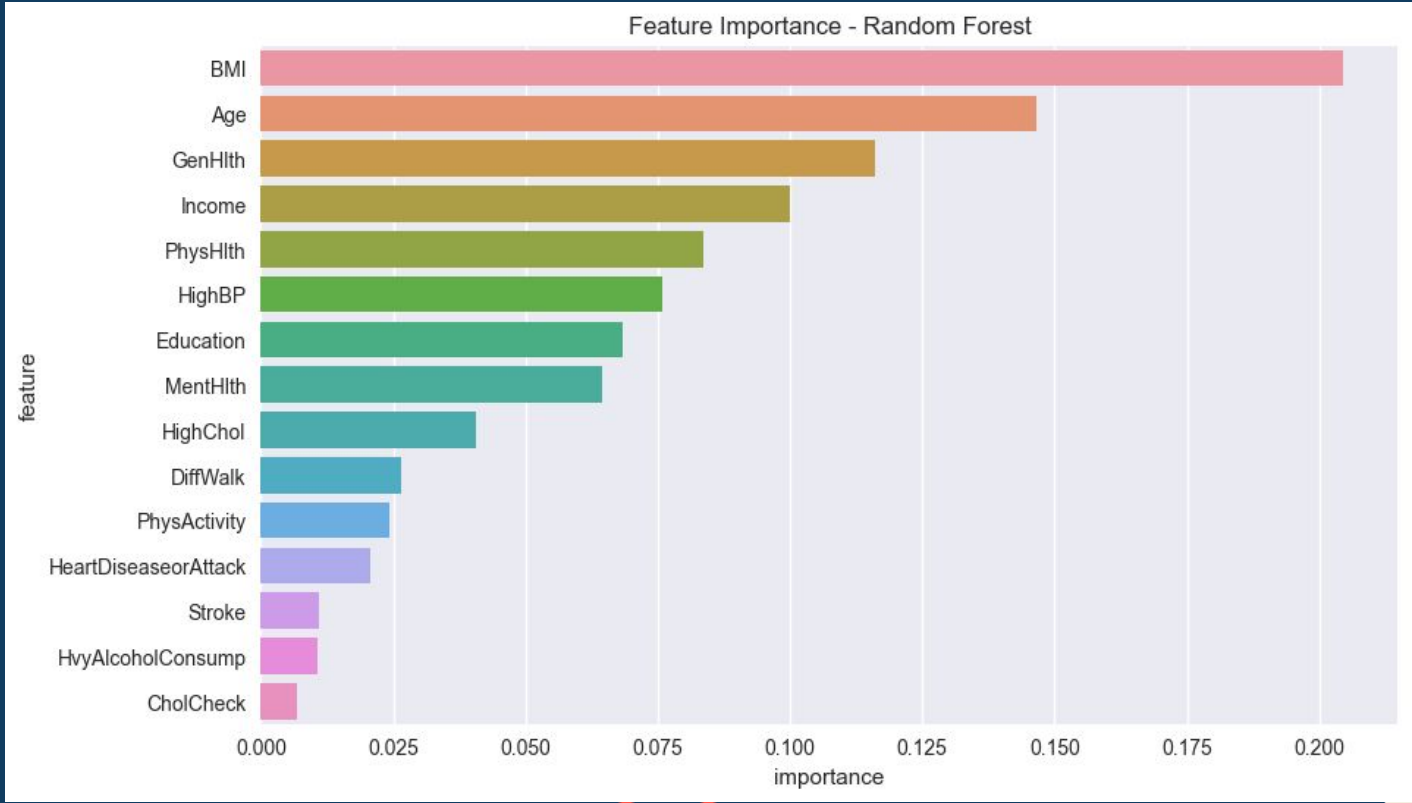
MODEL DEVELOPMENT

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

MODEL PERFORMANCE ANALYSIS



MODEL PERFORMANCE ANALYSIS



04

CONCLUSIONS

KEY FINDINGS

HEALTH INDICATORS

1. Cardiovascular Health:
    a. High blood pressure increases diabetes risk by over 25%.
    b. Heart disease patients have double the diabetes rate.
    c. Combined BP and cholesterol issues significantly increase risk.
2. Body Mass Index (BMI):
    a. Higher BMI correlates strongly with diabetes risk.
    b. Emphasizes the importance of weight management.
3. General Health Status:
    a. Strong predictor of diabetes.
    b. Progressive increase in risk with declining health.
    c. Highlights potential for early intervention.

KEY FINDINGS

LIFESTYLE FACTORS

1. Physical Activity:
    a. Reduces diabetes risk by 25%.
    b. Most significant modifiable factor.

2. Diet and Nutrition:
    a. Healthy diets, particularly fruits and vegetables, lower diabetes risk.
    b. Combined dietary habits show additive protective effects.

3. Behavioral Factors:
    a. Smoking has a moderate correlation with diabetes.
    b. Alcohol consumption less significant but still relevant.

KEY FINDINGS

DEMOGRAPHIC PATTERNS

1. Age and Gender:
   a. Risk increases steadily with age, highest in elderly populations.
   b. Gender differences are minimal but age-specific patterns vary.

2. Socioeconomic Factors:
   a. Higher income and education reduce risk.
   b. Better healthcare access leads to improved outcomes.

MODEL PERFORMANCE SUMMARY

1. **Best Performing Model:**
   a. Random Forest Classifier achieved 75% accuracy.
   b. Balanced precision and recall make it suitable for diabetes risk screening.

2. **Feature Importance:**
   a. General Health Status: Most significant predictor.
   b. BMI and Age: Strong predictors.
   c. Cardiovascular factors (BP, cholesterol) also highly relevant.

THANK YOU