
# Predicting Diabetes Probability using Machine Learning

Using machine learning techniques to forecast the likelihood of individuals developing diabetes based on various factors.

# Introduction

### Presentation Overview

Brief introduction to the presentation, highlighting the key topics to be covered.

### Diabetes Prediction

Using machine learning models to predict the probability of an individual developing diabetes.

### Machine Learning Approach

Outlining the specific machine learning algorithms and techniques employed in the analysis.

# Diabetes: The Challenge

Diabetes is a chronic condition characterized by high blood sugar levels, which can have serious health consequences if left unmanaged. Early detection is crucial, as it allows for timely intervention and the implementation of effective management strategies. Accurate prediction models can play a vital role in identifying individuals at risk, enabling proactive measures to prevent or delay the onset of the disease.

# Data Collection and Preprocessing

- BRFSS Data

  The Behavioral Risk Factor Surveillance System (BRFSS) is a health-related telephone survey that is collected annually by the CDC. For this project, a csv of the dataset available on Kaggle for the year 2015 was used.

- Preprocessing

  Description of the steps taken to ensure data quality, including handling missing values, removing outliers, and addressing inconsistencies in data formats and units.

- Feature Engineering

  Explanation of the process of creating new features from the raw data, such as calculating derived metrics, encoding categorical variables, and handling date and time-related features.

# Machine Learning Algorithms

## Logistic Regression

A widely used algorithm for binary classification tasks, such as predicting whether a patient has diabetes or not.

## Decision Tree

A tree-based algorithm that creates a series of rules to make predictions, useful for both classification and regression problems.

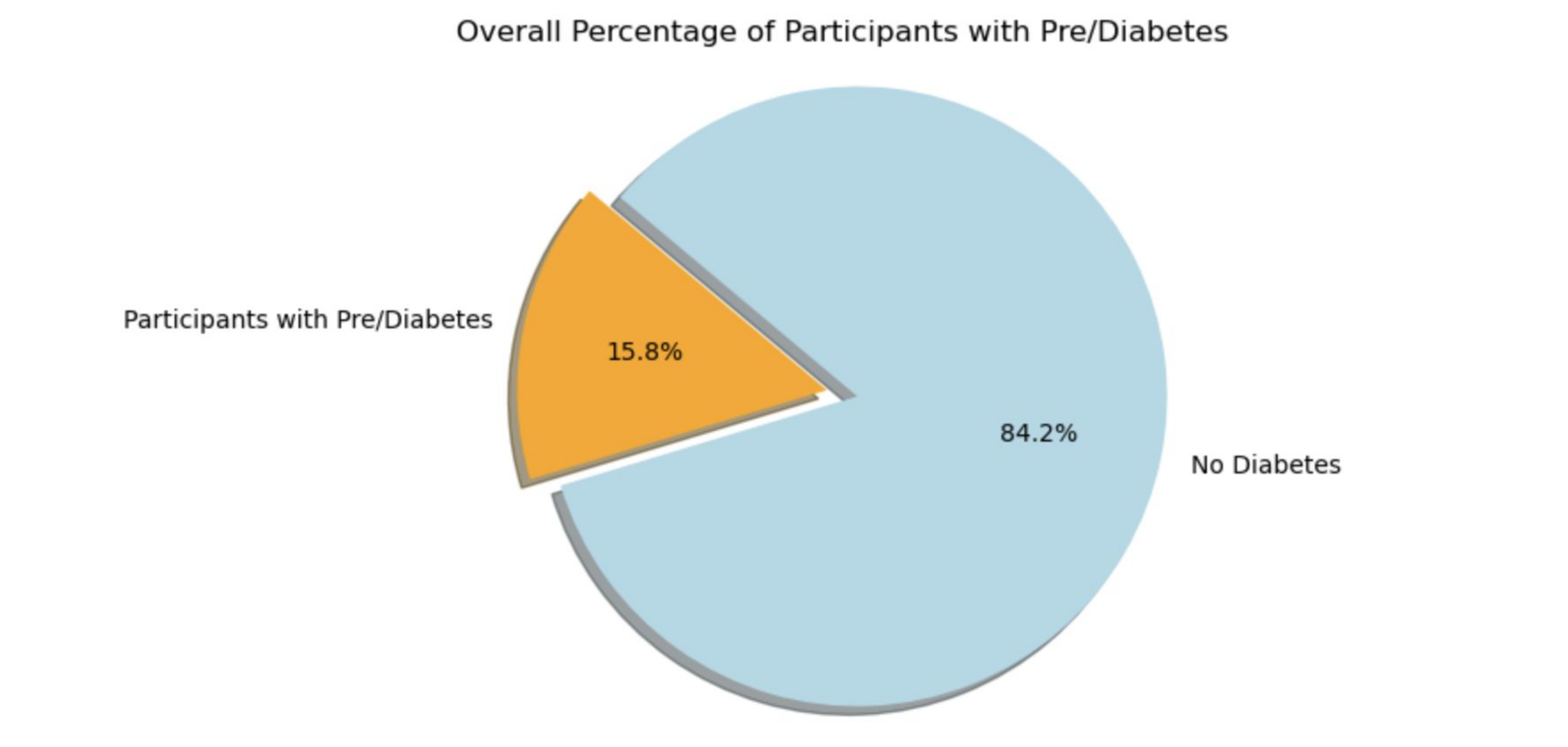
## Random Forest

An ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of the predictions.

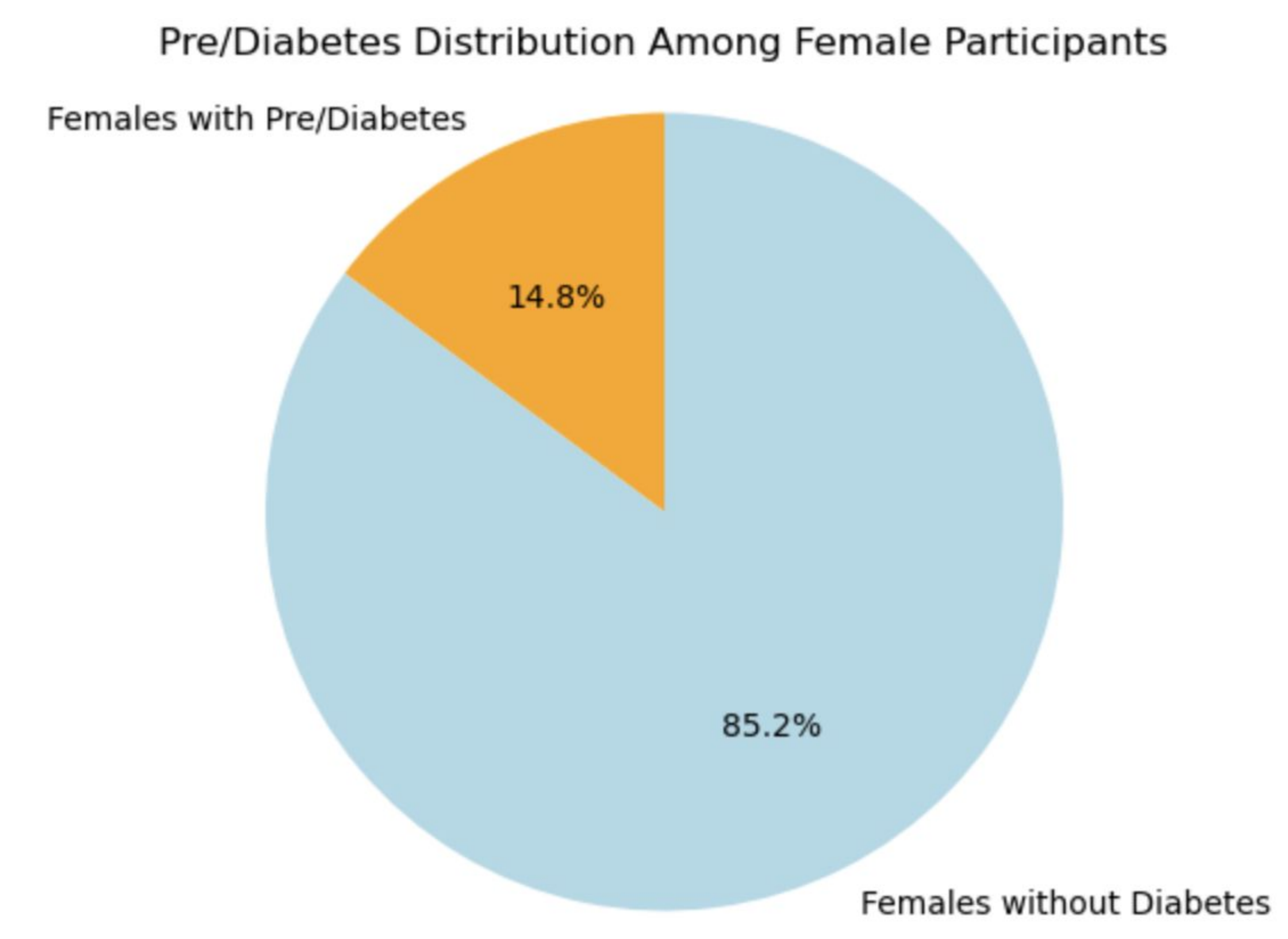
## K-Means

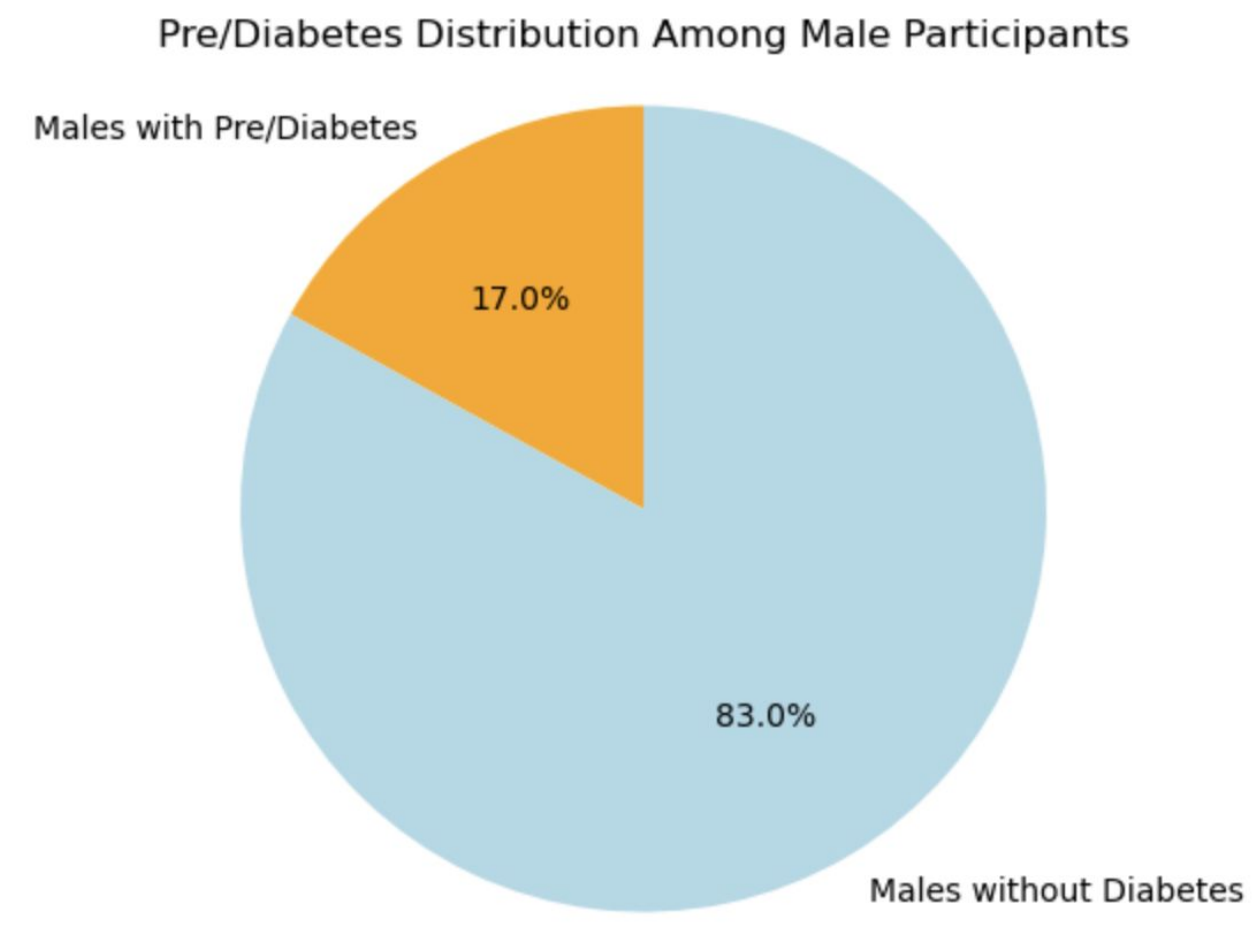
Unsupervised machine learning algorithm used for clustering data. Clustering is the task of grouping data points into clusters or groups based on their similarity.

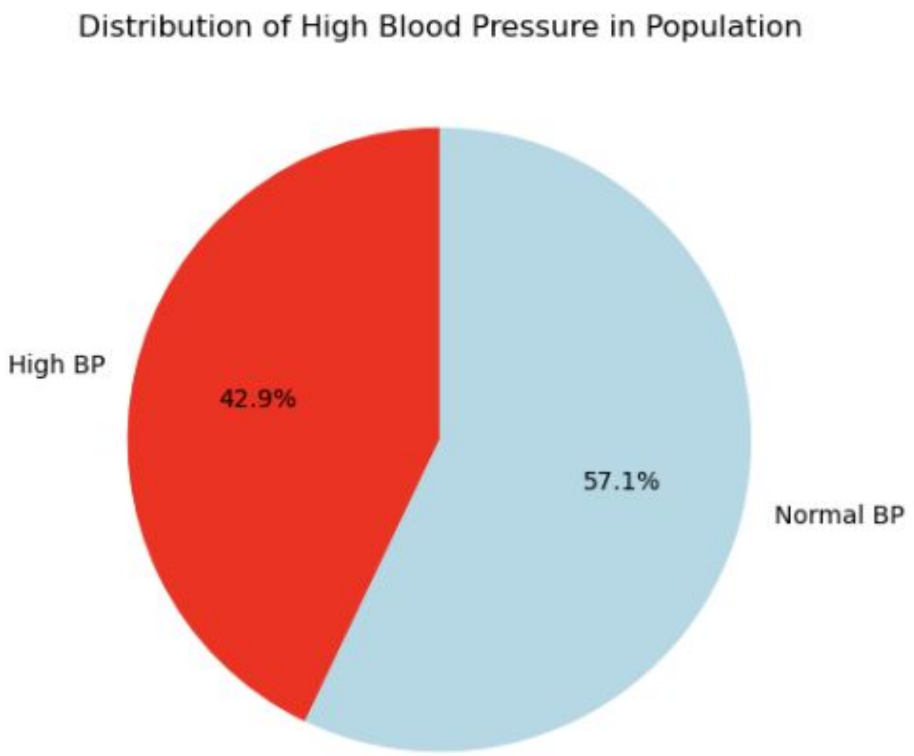
How well does the data set provide accurate predictions of whether an individual has diabetes?

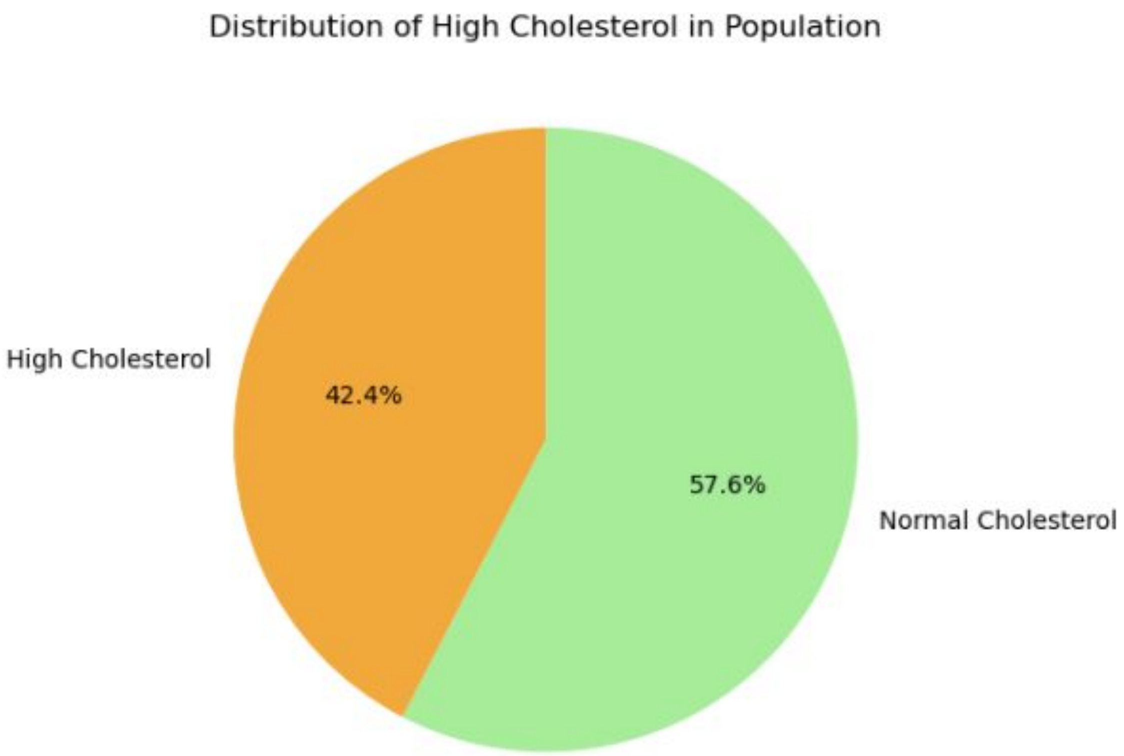


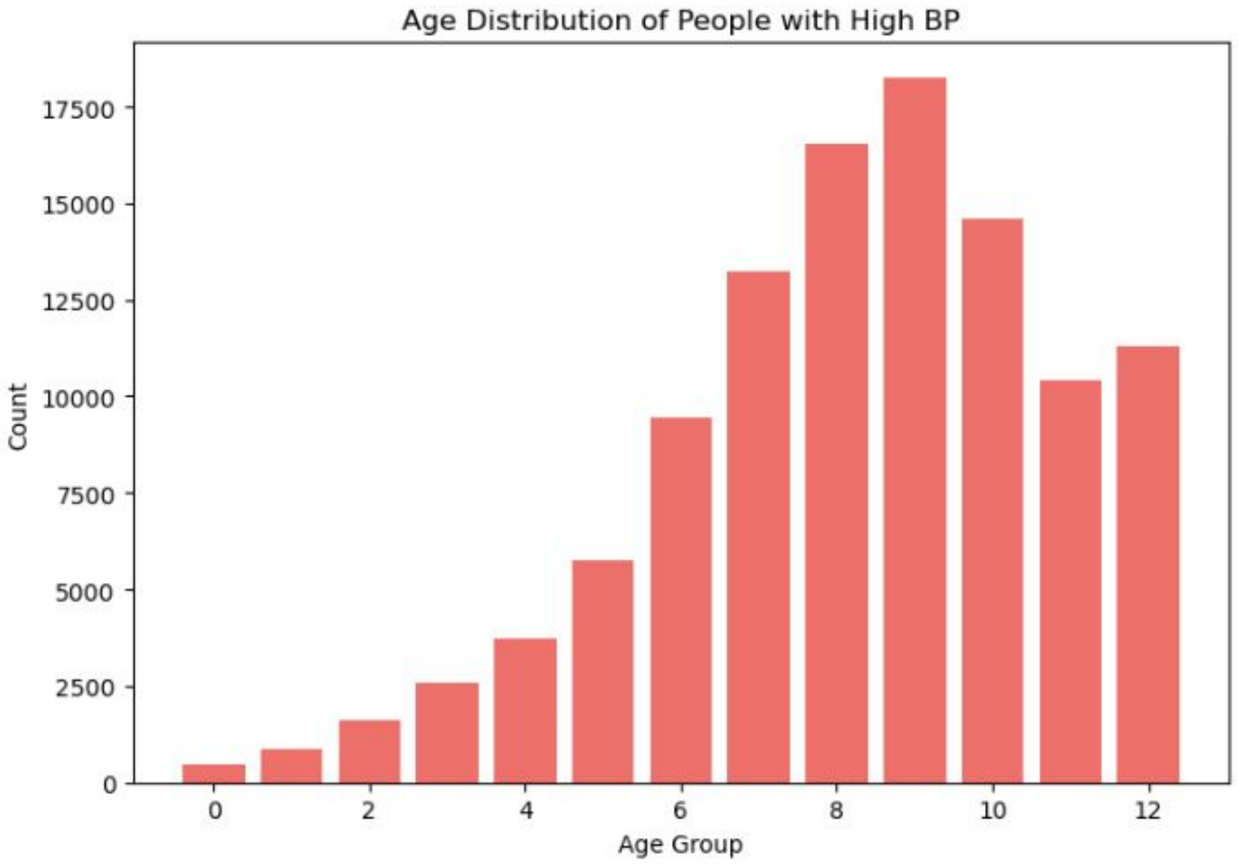
Overall Percentage of Participants with Pre/Diabetes
Participants with Pre/Diabetes
15.8%
84.2%
No Diabetes


Pre/Diabetes Distribution Among Female Participants
Females with Pre/Diabetes
14.8%
85.2%
Females without Diabetes

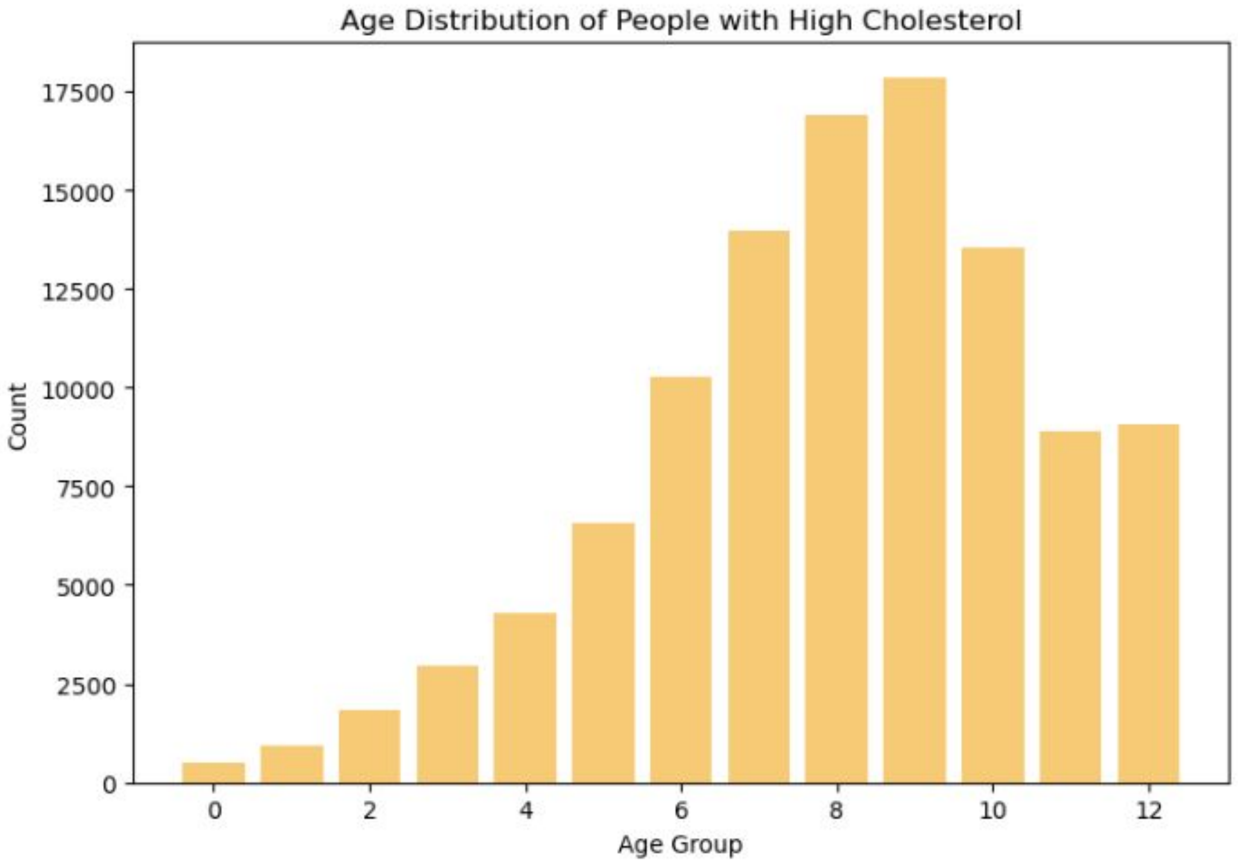

Pre/Diabetes Distribution Among Male Participants
Males with Pre/Diabetes
17.0%
83.0%
Males without Diabetes


Distribution of High Blood Pressure in Population
High BP
42.9%
57.1%
Normal BP

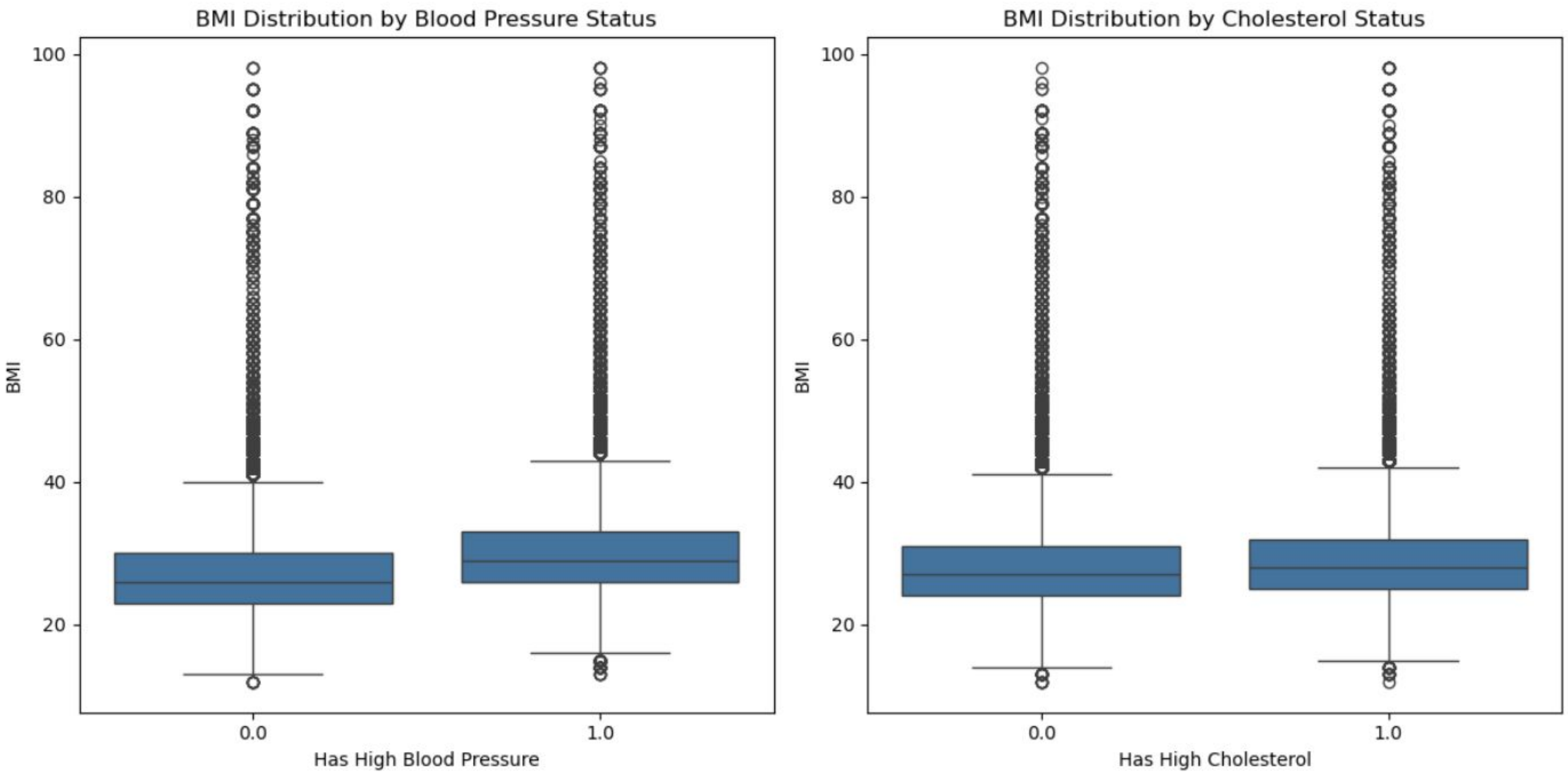

Distribution of High Cholesterol in Population
High Cholesterol
42.4%
57.6%
Normal Cholesterol



Age Distribution of People with High BP
Count
Age Group

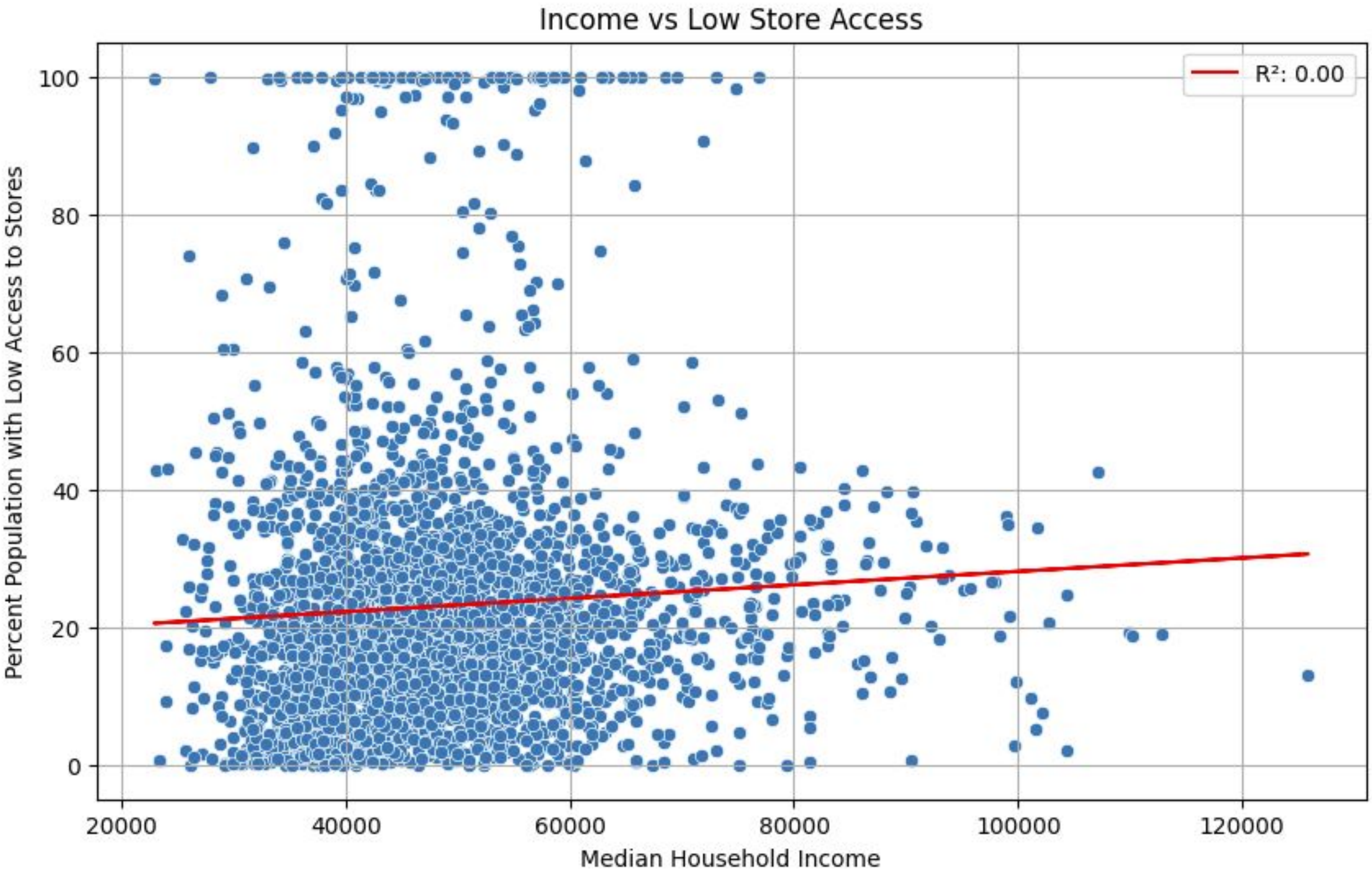



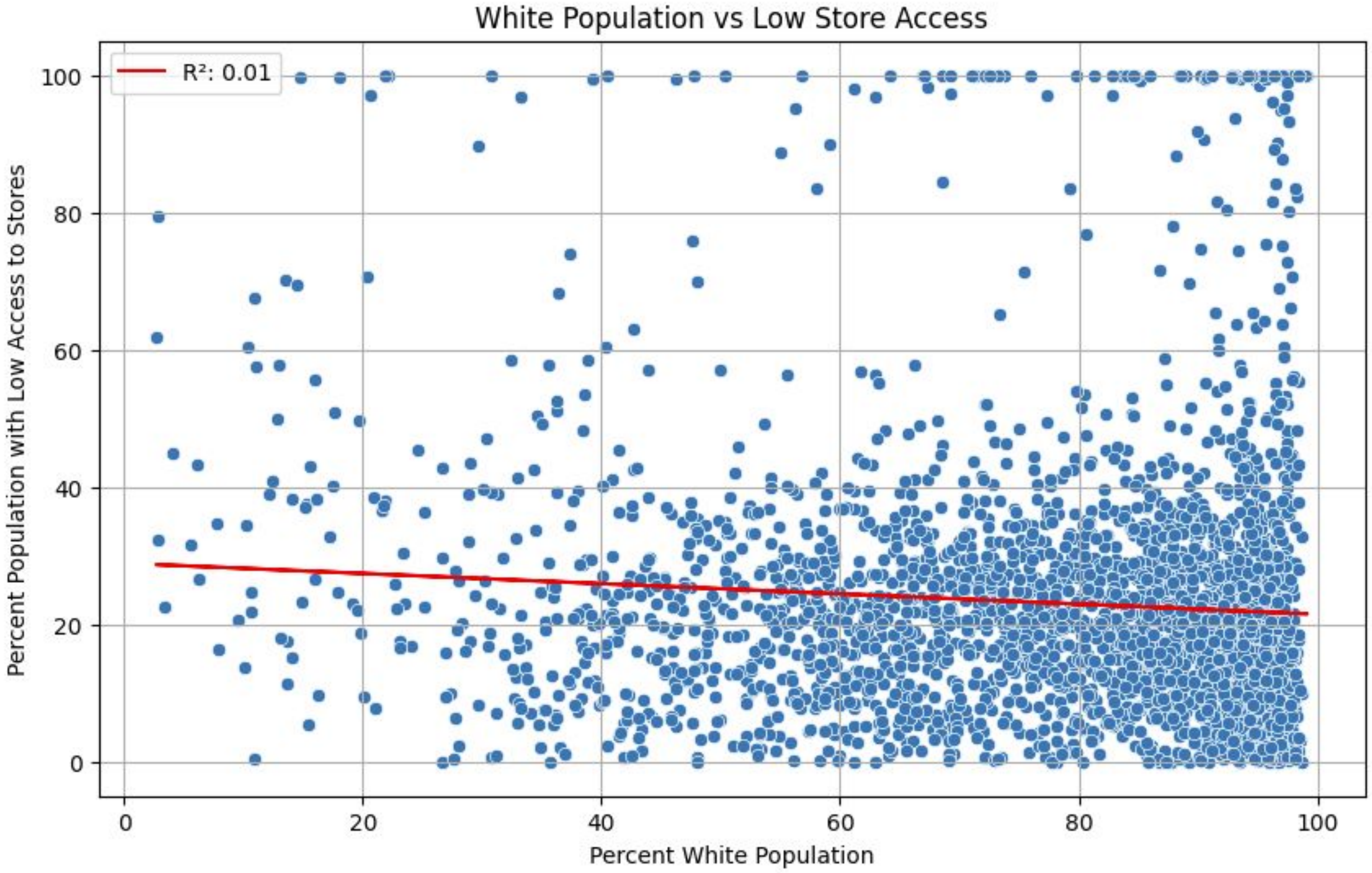
Age Distribution of People with High Cholesterol
Count
Age Group



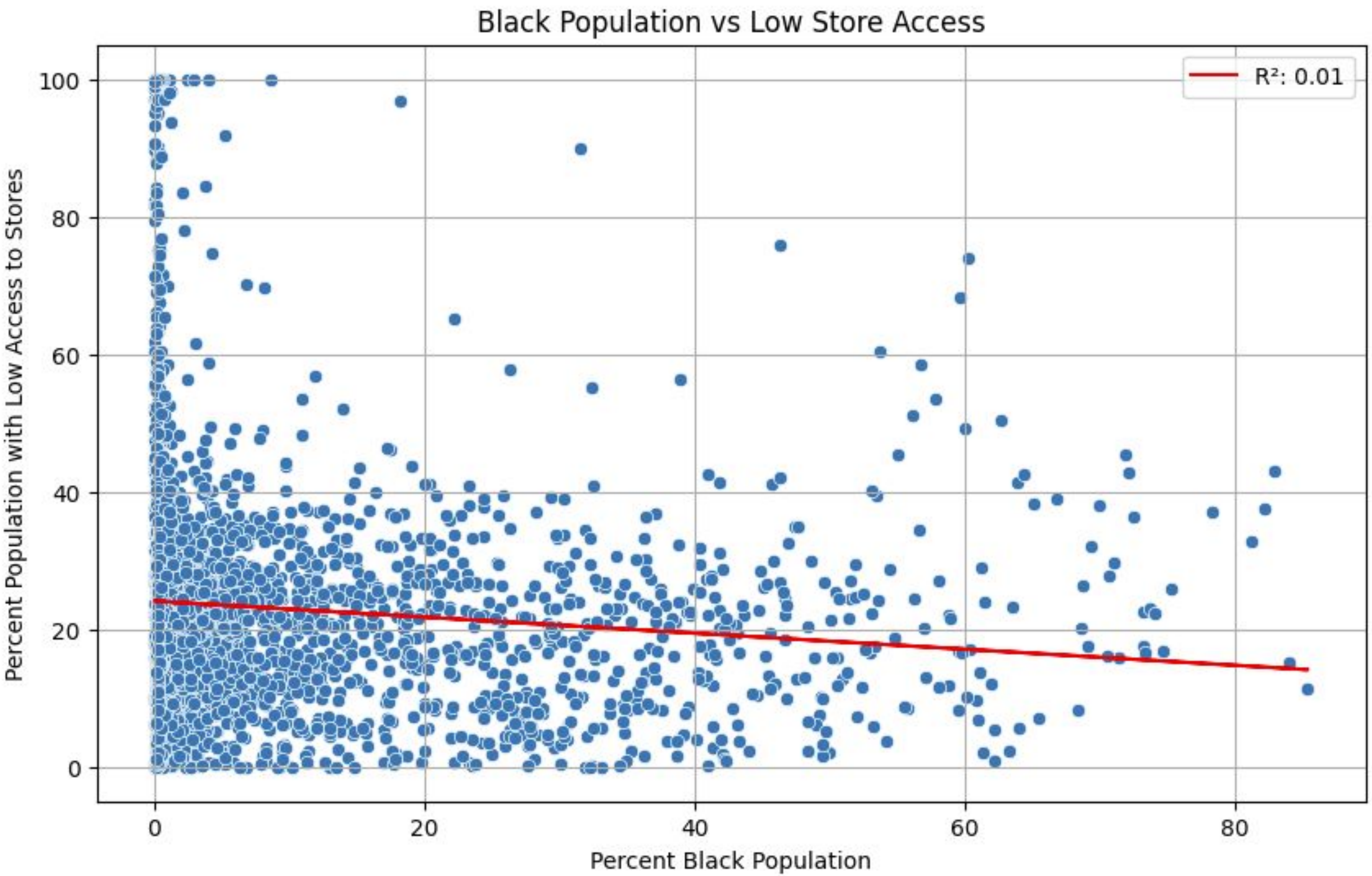
BMI Distribution by Blood Pressure Status
BMI
100
80
60
40
20
0.0
1.0
Has High Blood Pressure
BMI Distribution by Cholesterol Status
BMI
100
80
60
40
20
0.0
1.0
Has High Cholesterol

How does socioeconomic status affect access to healthy food and healthcare?

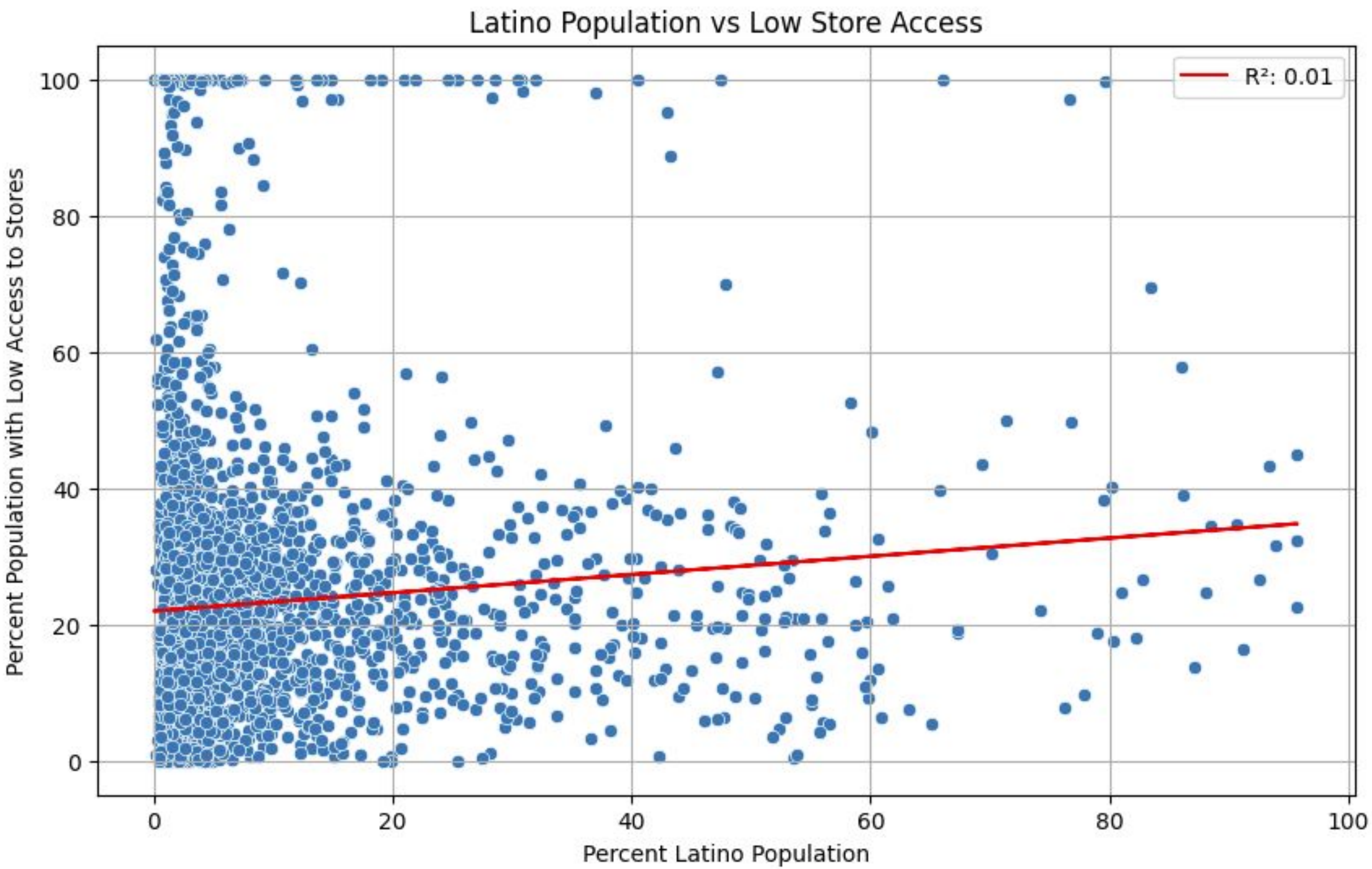

Income vs Low Store Access
R²: 0.00
Percent Population with Low Access to Stores
Median Household Income
0
20
40
60
80
100
20000
40000
60000
80000
100000
120000



White Population vs Low Store Access
R²: 0.01
Percent Population with Low Access to Stores
Percent White Population
0
20
40
60
80
100



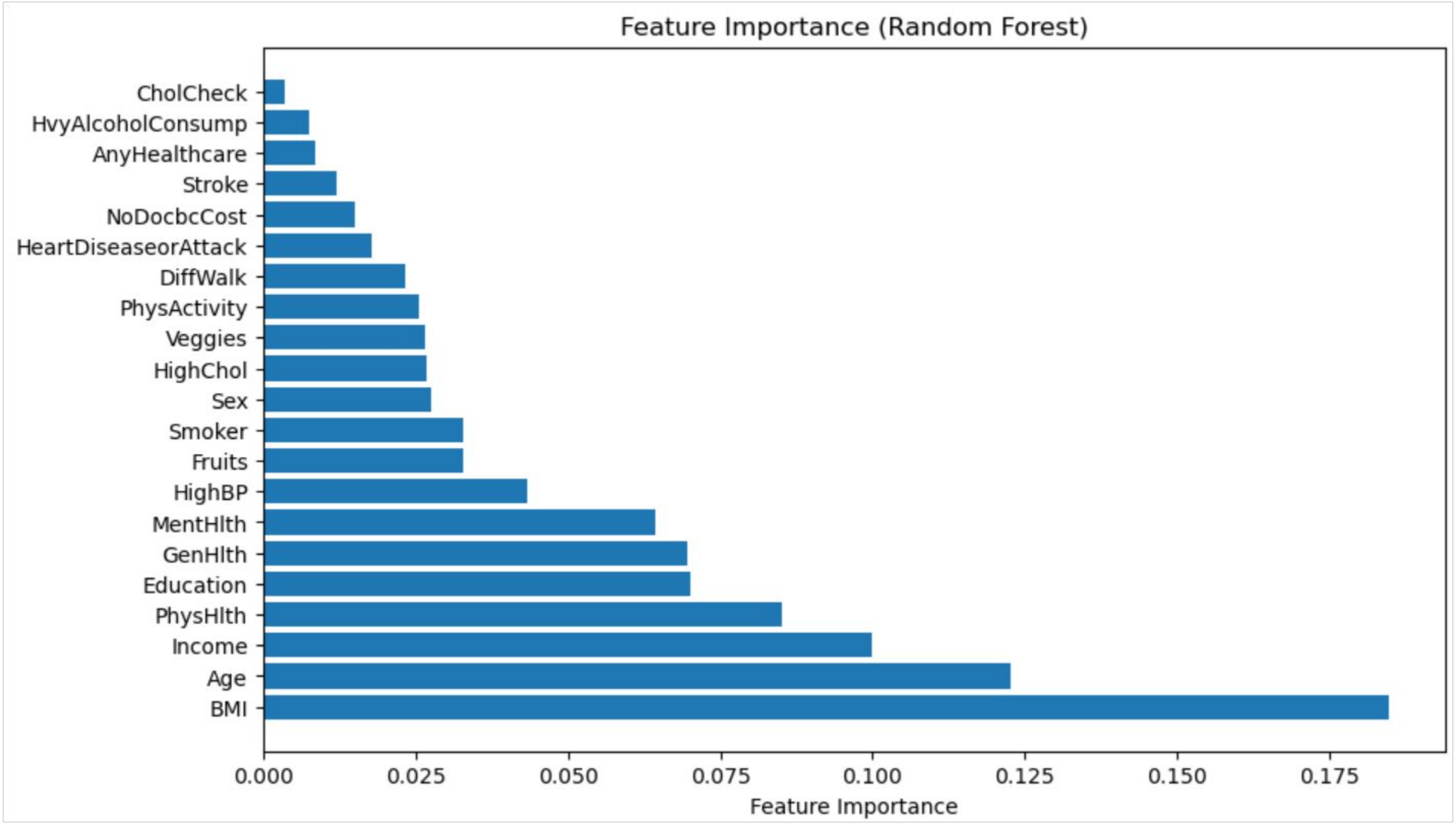
Black Population vs Low Store Access
R²: 0.01
Percent Population with Low Access to Stores
Percent Black Population
0
20
40
60
80
100


Latino Population vs Low Store Access
R²: 0.01
Percent Population with Low Access to Stores
Percent Latino Population
0
20
40
60
80
100
0
20
40
60
80
100

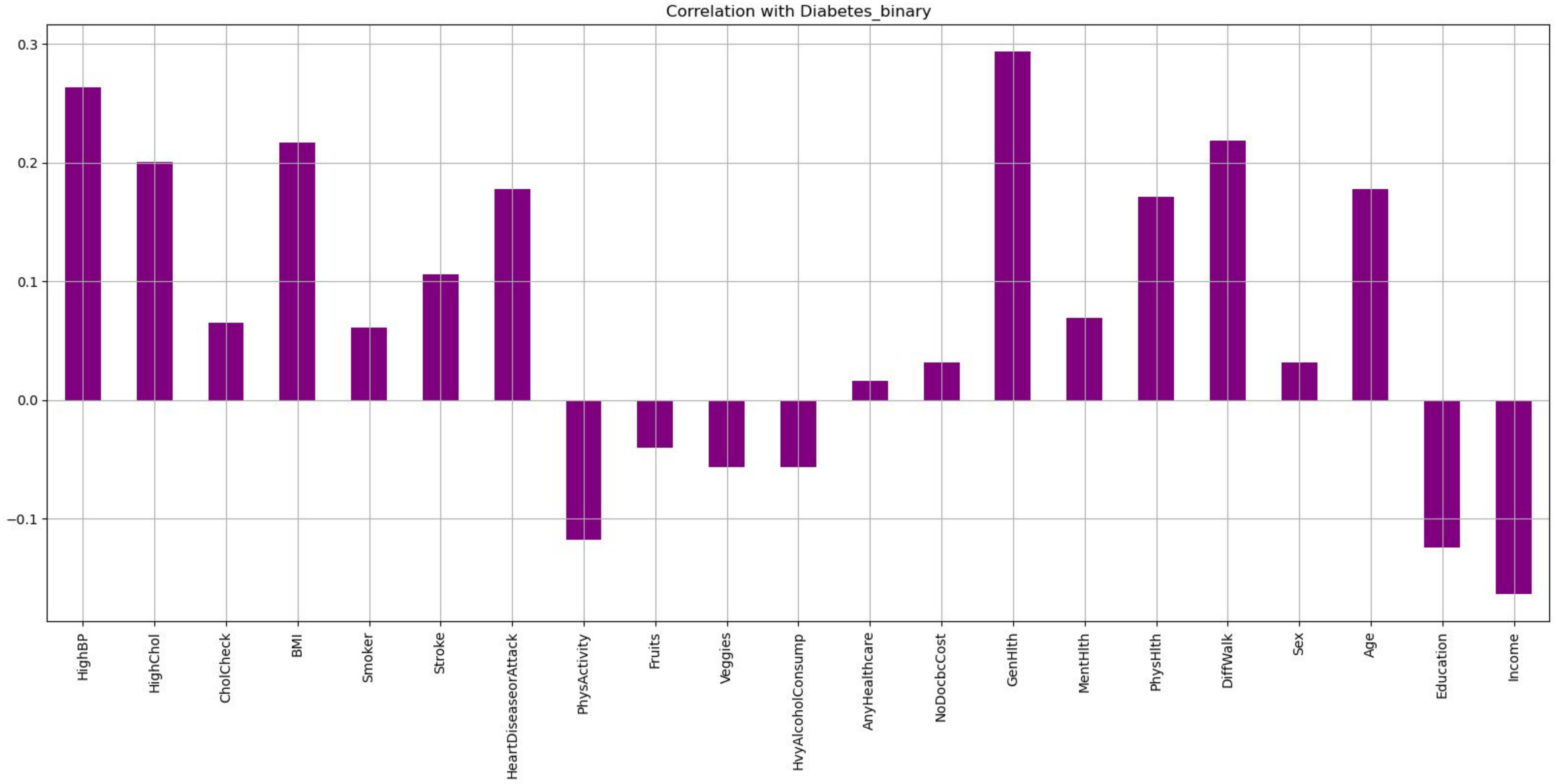
# What features are most indicative of someone getting diabetes?

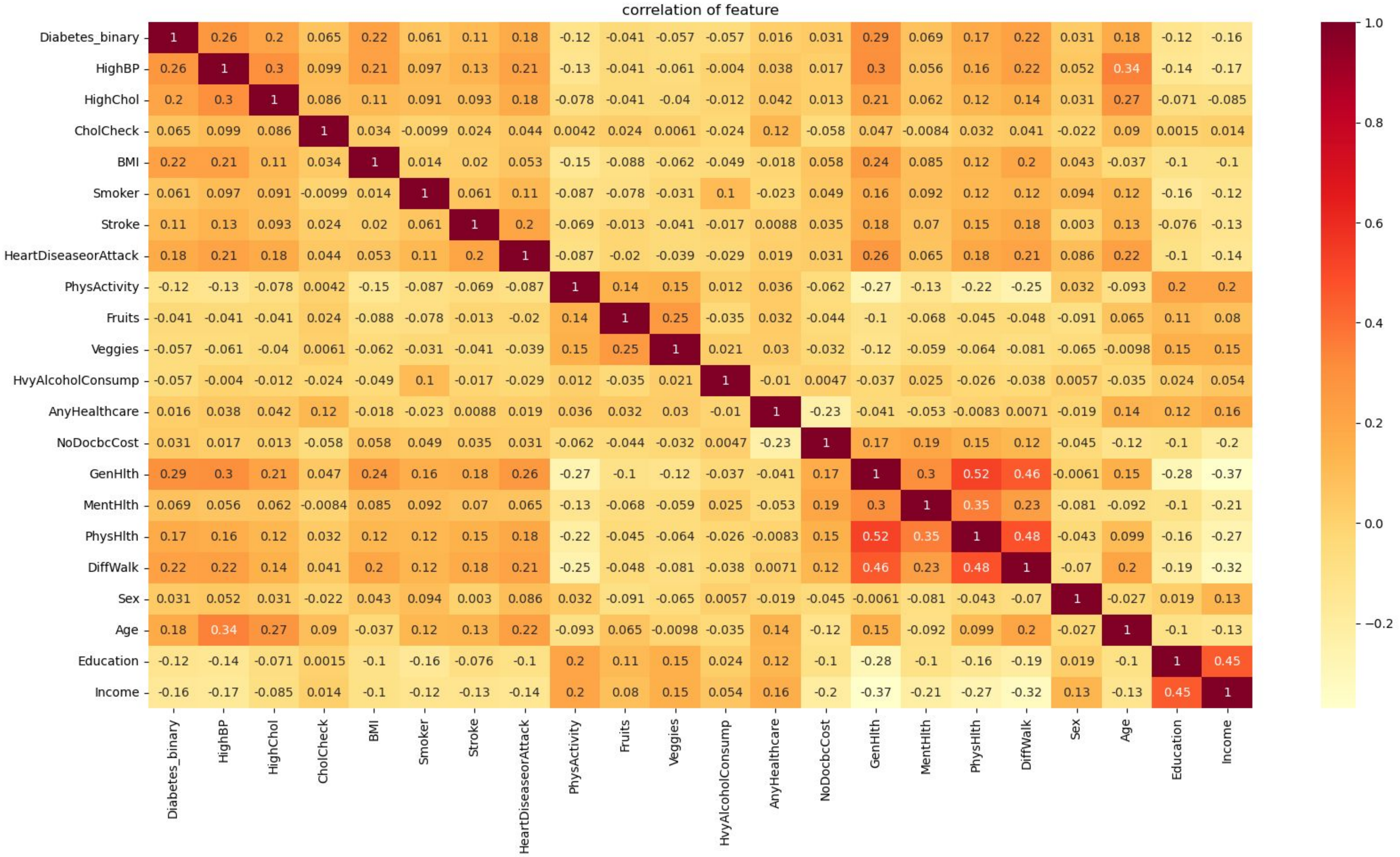
**Below are is a comparison of the accuracy and recall for ‘1s’. Random Forest performed the best.**

| **Model** | **Accuracy** | **Recall for 1s** |
|---|---|---|
| Logistic Regression | 0.74 | 0.76 |
| Decision Trees | 0.77 | 0.82 |
| Random Forest | 0.78 | 0.82 |



Feature Importance (Random Forest)
CholCheck
HvyAlcoholConsump
AnyHealthcare
Stroke
NoDocbcCost
HeartDiseaseorAttack
DiffWalk
PhysActivity
Veggies
HighChol
Sex
Smoker
Fruits
HighBP
MentHlth
GenHlth
Education
PhysHlth
Income
Age
BMI
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
Feature Importance


Correlation with Diabetes_binary
0.3
0.2
0.1
0.0
−0.1
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income


correlation of feature
Diabetes_binary
HighBP
HighChol
CholCheck
BMI
Smoker
Stroke
HeartDiseaseorAttack
PhysActivity
Fruits
Veggies
HvyAlcoholConsump
AnyHealthcare
NoDocbcCost
GenHlth
MentHlth
PhysHlth
DiffWalk
Sex
Age
Education
Income
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

# Select Best features using KBest method

```
#using SelectKBest class to extract top 10 best features
BestFeatures = SelectKBest(score_func=chi2, k=10)
fit = BestFeatures.fit(X_train_fi,y_train_fi)

df_scores = pd.DataFrame(fit.scores_)
df_columns = pd.DataFrame(X_train_fi.columns)

#concatenating two dataframes for better visualization
f_Scores = pd.concat([df_columns,df_scores],axis=1)               # feature scores
f_Scores.columns = ['Feature','Score']

f_Scores = f_Scores.sort_values('Score', ascending=False)
f_Scores
```

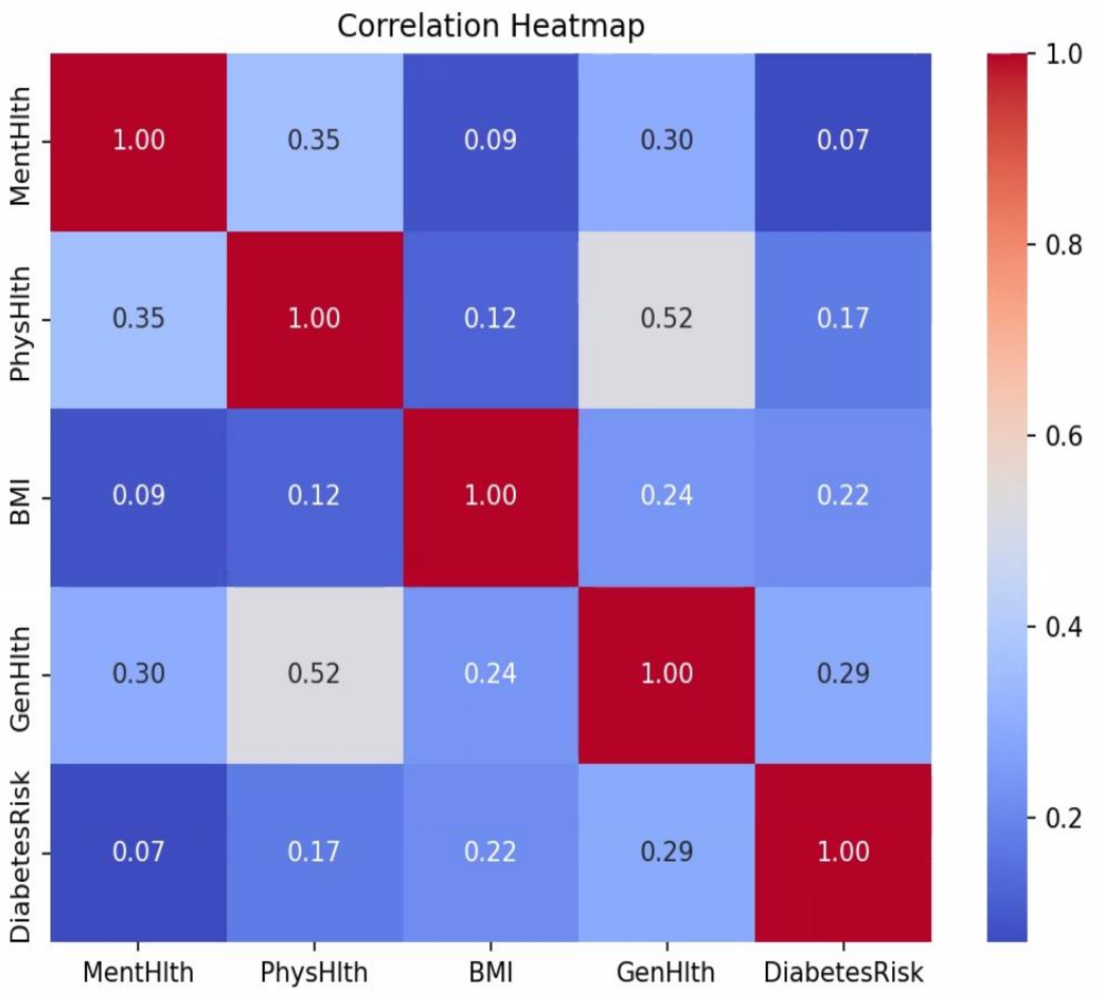
KBest was also used to see another method for determining features importance ranking. Results are sorted by highest chi score.

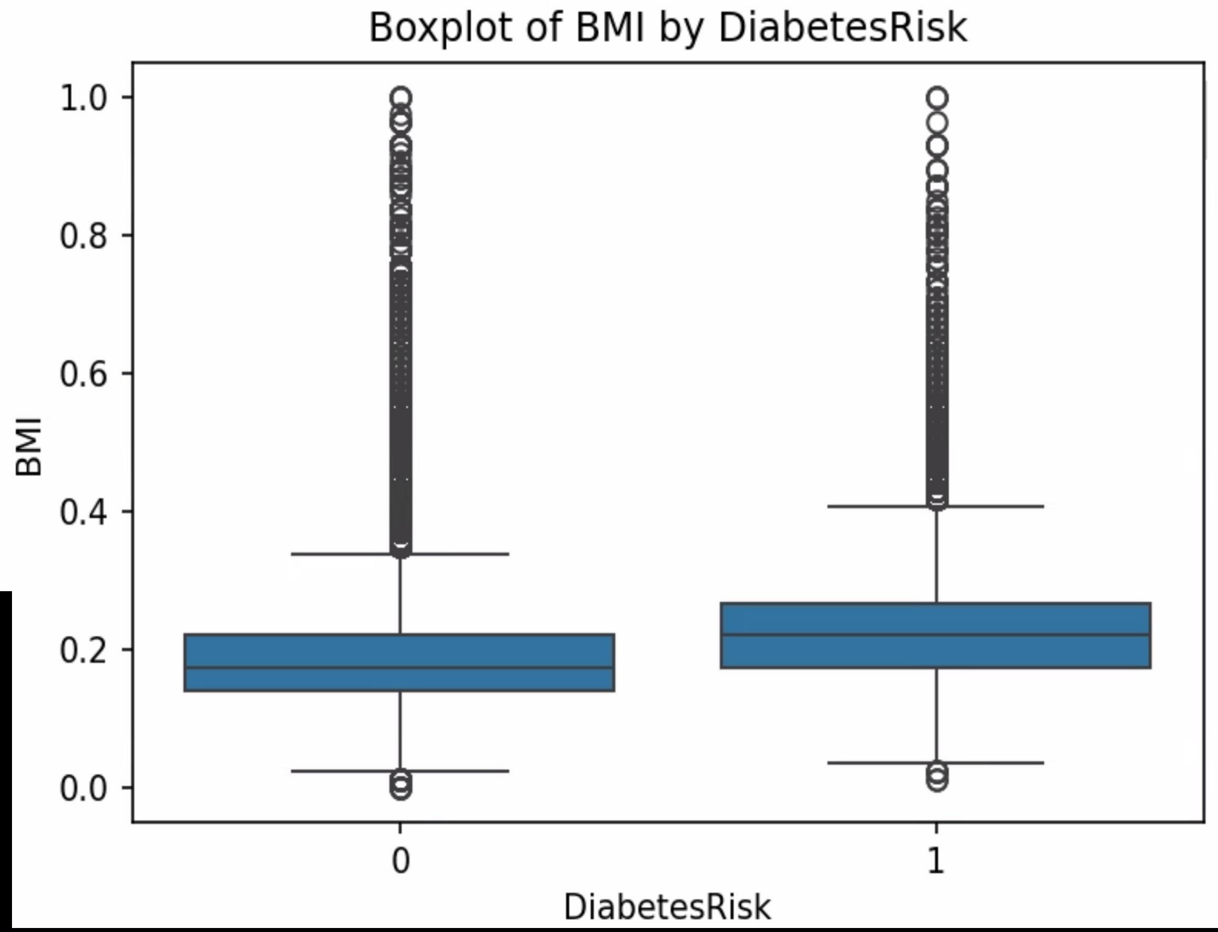
| | Feature | Score |
|---|---|---|
| 15 | PhysHlth | 133424.406534 |
| 14 | MentHlth | 21029.632228 |
| 3 | BMI | 18355.166400 |
| 16 | DiffWalk | 10059.506391 |
| 0 | HighBP | 10029.013935 |
| 13 | GenHlth | 9938.507776 |
| 18 | Age | 9276.141199 |
| 6 | HeartDiseaseorAttack | 7221.975378 |
| 1 | HighChol | 5859.710582 |
| 20 | Income | 4829.816361 |
| 5 | Stroke | 2725.225194 |
| 7 | PhysActivity | 861.887532 |
| 10 | HvyAlcoholConsump | 779.424807 |
| 19 | Education | 756.035496 |
| 4 | Smoker | 521.978858 |
| 12 | NoDocbcCost | 229.542412 |

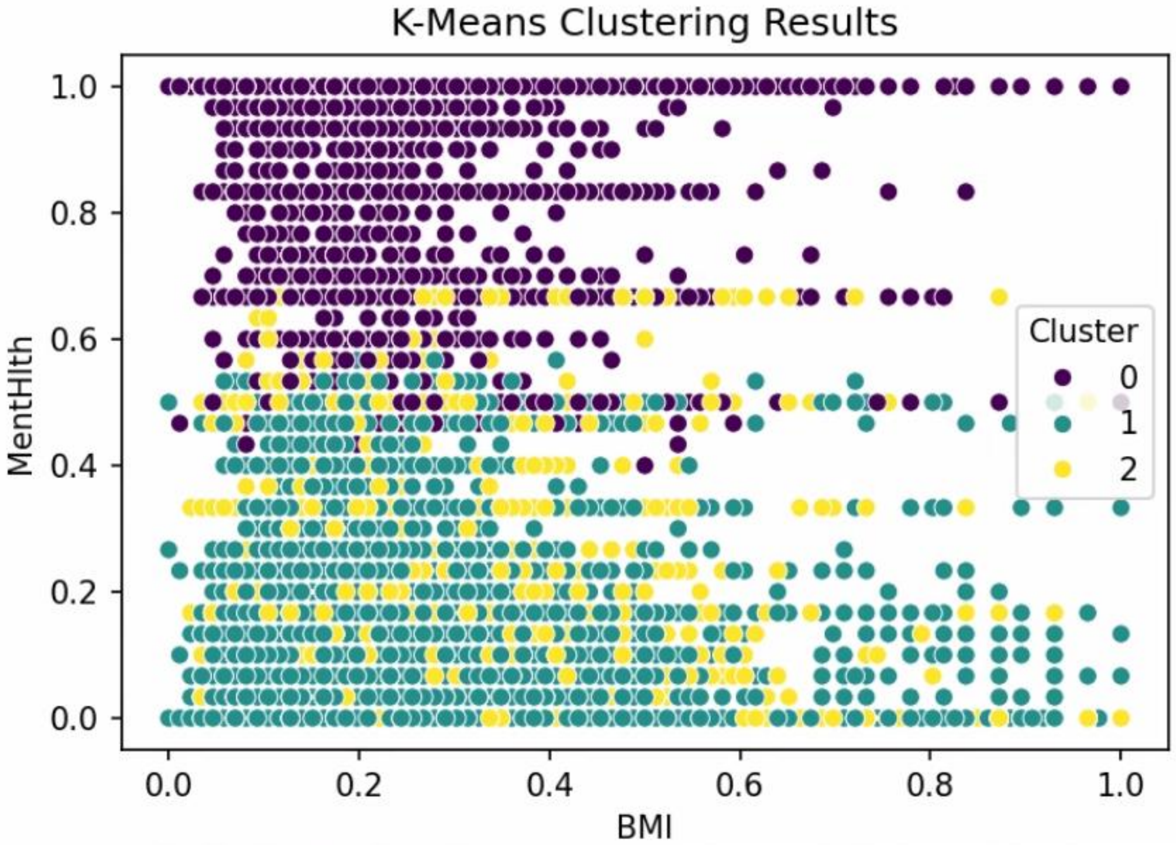
# How does Mental Health and Insulin resistance strongly influence diabetes risk?

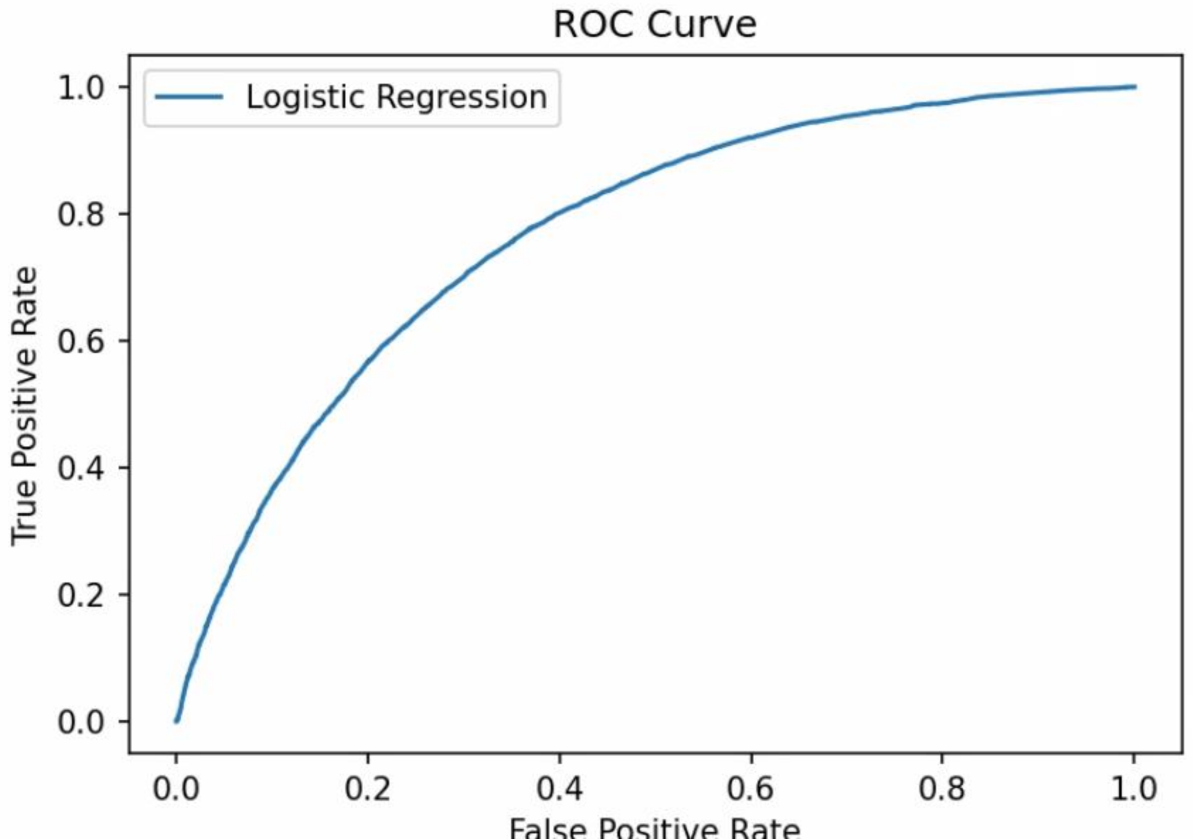
To explore these relationships

- Correlation analysis,
- Applied clustering methods
- Evaluated predictive performance using logistic regression.

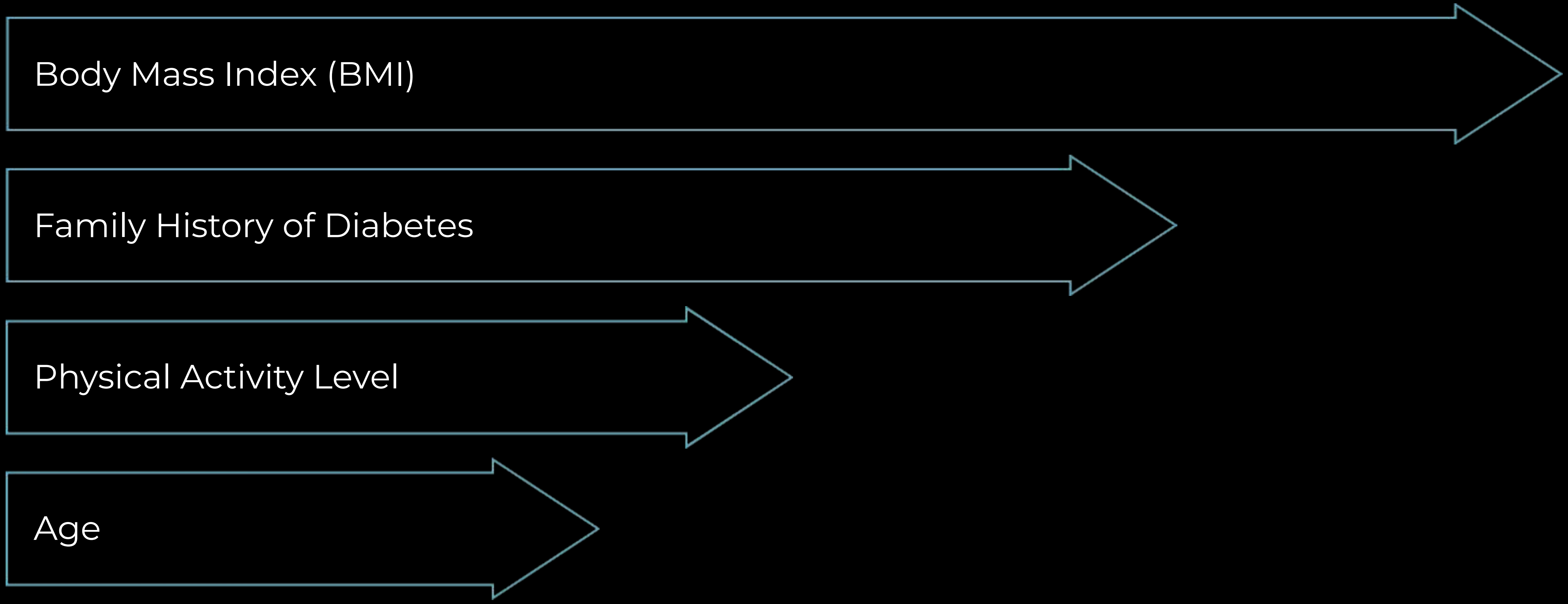

Correlation Heatmap
MentHlth
PhysHlth
BMI
GenHlth
DiabetesRisk
1.00 0.35 0.09 0.30 0.07
0.35 1.00 0.12 0.52 0.17
0.09 0.12 1.00 0.24 0.22
0.30 0.52 0.24 1.00 0.29
0.07 0.17 0.22 0.29 1.00
1.0
0.8
0.6
0.4
0.2


Boxplot of BMI by DiabetesRisk
BMI
1.0
0.8
0.6
0.4
0.2
0.0
0
1
DiabetesRisk


K-Means Clustering Results
MentHlth
BMI
Cluster
0
1
2
0.0
0.2
0.4
0.6
0.8
1.0


ROC Curve
Logistic Regression
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

# Feature Importance



# Can we use a subset of the risk factors to accurately predict whether an individual has diabetes?

The results of the logistic regression model shows that we can correlate BMI with diabetes with a sufficient degree of accuracy

- The random forest results also indicates that BMI, income and age are important factors
- Therefore creating a predictive model is possible and is a good avenue for future work.

# Clinical Validation

### Establish Partnerships

Collaborate with healthcare providers, hospitals, and medical research institutions to gain access to real-world patient data for model validation.

### Data Collection and Curation

Work with healthcare professionals to collect and curate a diverse patient dataset, ensuring it represents the target population and clinical scenarios.

### Model Evaluation

Assess the model's performance on the real-world patient data, evaluating metrics such as accuracy, precision, recall, and F1-score to validate its clinical utility.

### Clinical Feedback and Iteration

Engage with healthcare professionals to gather their feedback on the model's performance and incorporate their insights to refine and improve the model.

# Future Considerations

### Explore Natural Language Processing (NLP) techniques

Investigate the use of NLP to extract insights from unstructured data sources, such as customer reviews or social media posts.

### Integrate Geospatial Data

Incorporate location-based data, such as GPS or satellite imagery, to enhance the model's understanding of spatial relationships and patterns.

### Experiment with Transfer Learning

Leverage pre-trained models and transfer learning approaches to accelerate model development and improve performance on specific tasks.

### Enhance Multimodal Modeling

Explore the integration of various data modalities, such as images, audio, and text, to gain a more comprehensive understanding of the problem domain.

### Investigate Causal Inference

Employ causal inference techniques to better understand the underlying relationships and mechanisms driving the observed patterns in the data.