Understanding Diabetes Risk Factors

This presentation explores the associations between various health indicators and the prevalence of diabetes/prediabetes in U.S. adults. I will delve into the study design, data analysis, and the insights gained from our logistic regression model, highlighting key risk factors and their impact.

by Ahmed Tarek

Scientific Question and Study Design

The core scientific question investigates the associations between risk factors like high blood pressure, BMI, and smoking, and the prevalence of diabetes/prediabetes in U.S. adults. I also examine how age and sex influence these associations.

Study Design

The dataset used in this analysis is derived from the Behavioral Risk Factor Surveillance System (BRFSS) 2015 survey, which is a cross-sectional study collecting health-related data from a population sample at a single point in time. The specific dataset, diabetes_binary_5050split_health_indicators_BRFSS2015.csv, has been modified to include an equal number of individuals with diabetes (cases) and without diabetes (controls), creating a 50/50 split. This manipulation transforms the dataset to resemble a case-control study design, where cases and controls are deliberately balanced. While the original cross-sectional nature of BRFSS allows prevalence estimation.

Descriptive analysis

Dataset Overview

- **Total Observations**: 70,692
- **Target Variable**: `Diabetes_binary` (0: No Diabetes, 1: Diabetes/Prediabetes)
- **Class Distribution**:
  - Non-Diabetic: 35,346 (50.0%)
  - Diabetic/Prediabetic: 35,346 (50.0%)
- **Year Collected**: 2015
- **Source**: BRFSS (Behavioral Risk Factor Surveillance System)

Key Health Indicators

Indicator	Category	Overall %	Diabetic %	Non-Diabetic %
High Blood Pressure	No High BP	64.4%	25%	70%
	High BP	35.6%	75%	30%
High Cholesterol	No High Chol	57.8%	29.5%	69.1%
	High Chol	42.2%	70.5%	30.9%

Continuous Variables

Variable	Overall Mean	Diabetic Mean	Non-Diabetic Mean	Std Dev
BMI	29.5	32.1	25.1	6.7

Diabetes Prevalence by Age

Age Range	Age Code	Total Count	Diabetic Count	Prevalence
18-24	1	6,714	545	8.1%
25-29	2	5,231	553	10.6%
30-34	3	7,039	1,017	14.5%
35-39	4	7,559	1,497	19.8%
40-44	5	8,812	2,096	23.8%
45-49	6	9,874	3,142	31.8%
50-54	7	**10,206**	3,892	38.1%
55-59	8	**10,142**	4,576	45.1%
60-64	9	8,887	4,526	50.9%
65-69	10	6,349	3,645	57.4%
70-74	11	4,327	2,615	**60.4%**
75-79	12	2,762	1,738	**62.9%**
80+	13	1,790	1,157	**64.6%**

Data Preprocessing and Visualization

Our data preprocessing pipeline involved meticulously loading the raw dataset and recoding key categorical variables into more descriptive factors, ensuring data quality and readability. Following this, we generated a series of insightful graphs to visualize crucial distributions and relationships within the data, providing a foundational understanding before advanced modeling.

Load and Preprocess Data

Initial data handling included converting numerical codes into descriptive factors for essential variables such as `Diabetes_binary` (diabetes/prediabetes status),...

Visualize Diabetes Status Distribution

A bar plot was generated to clearly illustrate the 50/50 split of the dataset based on diabetes status, confirming the deliberate balancing of the two groups for...

Analyze BMI and Age Distributions

Histograms and comparative boxplots were used to visualize the distributions of Body Mass Index (BMI) and age. These revealed that individuals in the...

Examine HighBP and Sex Proportions

A stacked bar plot was created to vividly illustrate the prevalence of high blood pressure (HighBP) across different sexes and diabetes statuses, highlighting ho...

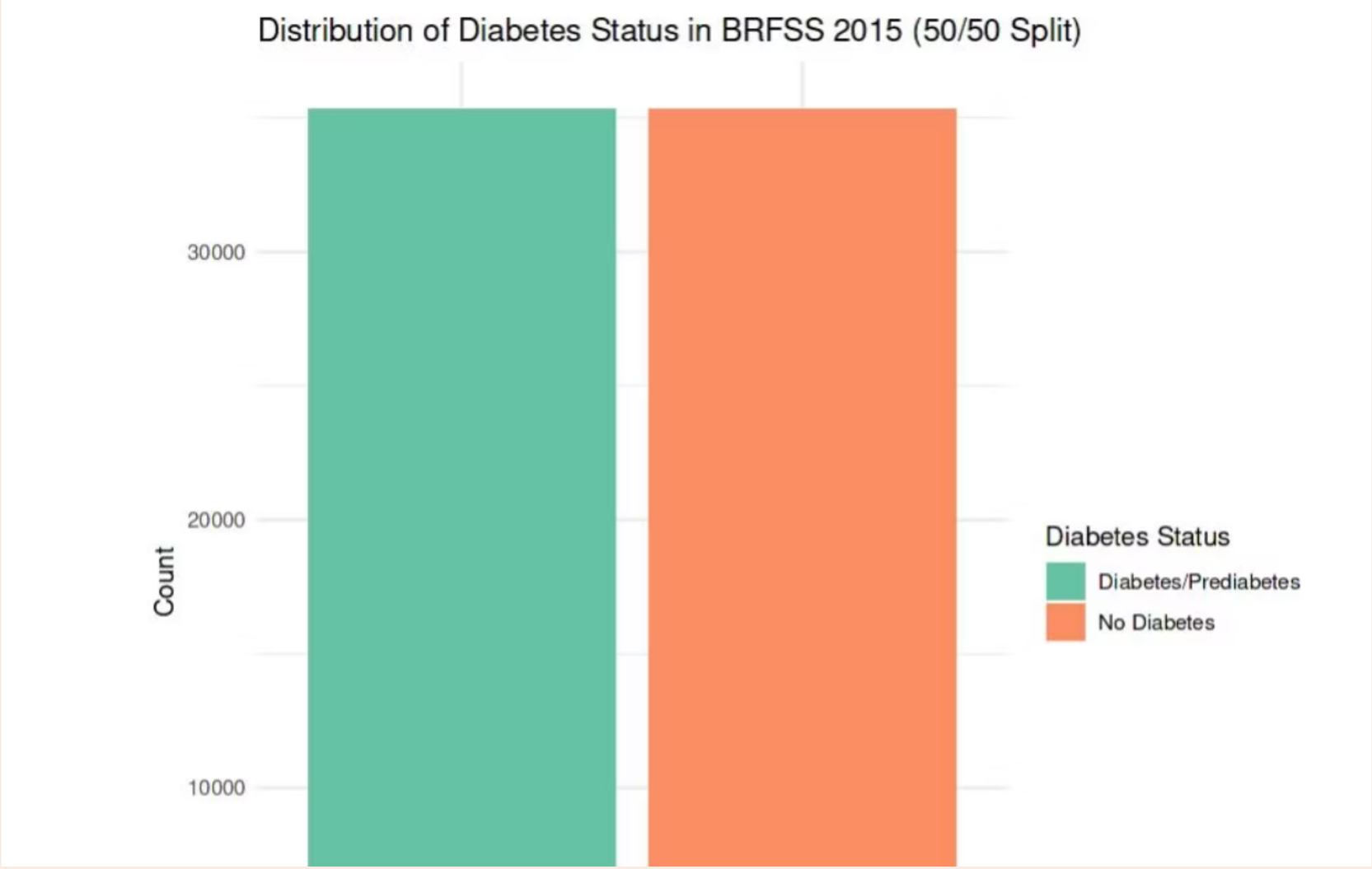
Load and Preprocess Data

Initial data handling included converting numerical codes into descriptive factors for essential variables such as `Diabetes_binary` (diabetes/prediabetes status), `Sex`, `HighBP` (high blood pressure), and `HighChol` (high cholesterol), preparing them for analysis.

```
# Step 2: Preprocess the data for visualization
cat("\n=== Data Preprocessing ===\n")
data <- data %>%  mutate(    Diabetes_binary = as.factor(dplyr::recode(as.character(Diabetes_binary),
"0" = "No Diabetes",
"1" = Diabetes/Prediabetes")),
Sex = as.factor(ifelse(Sex == 0, "Female", "Male")),
HighBP = as.factor(ifelse(HighBP == 0, "No High BP", "High BP")),
HighChol = as.factor(ifelse(HighChol == 0, "No High Chol", "High Chol"))    )
```

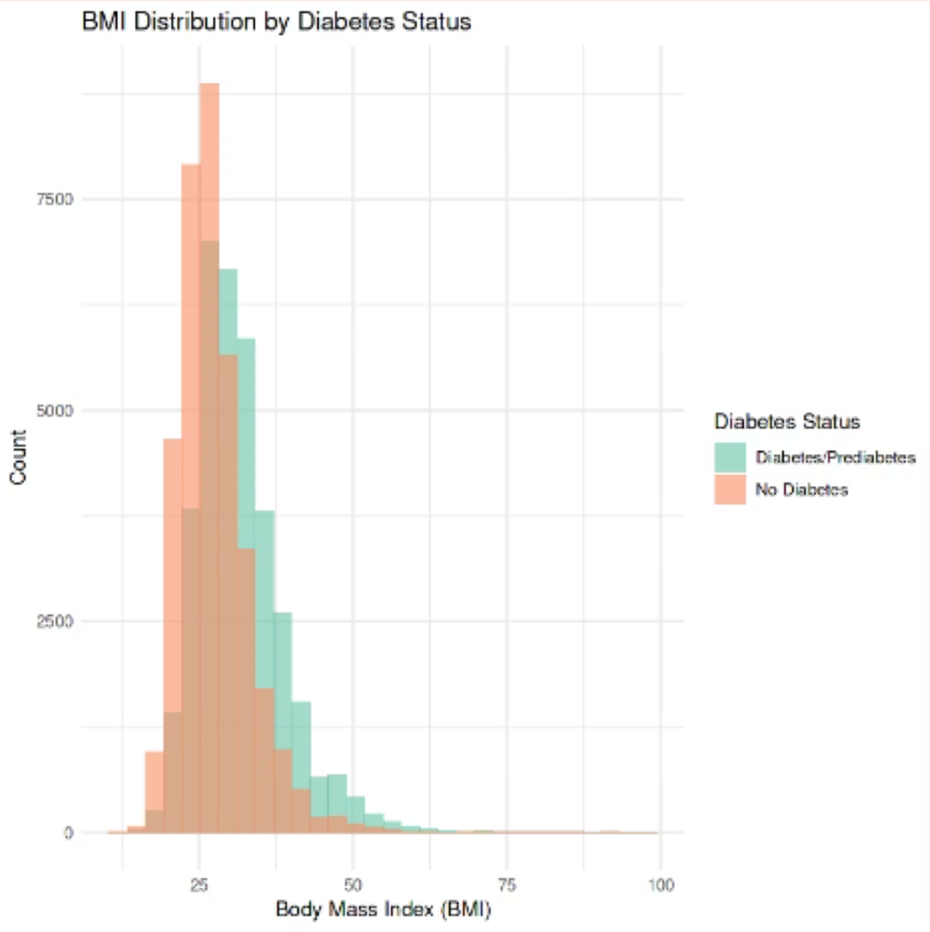
Visualize Diabetes Status Distribution

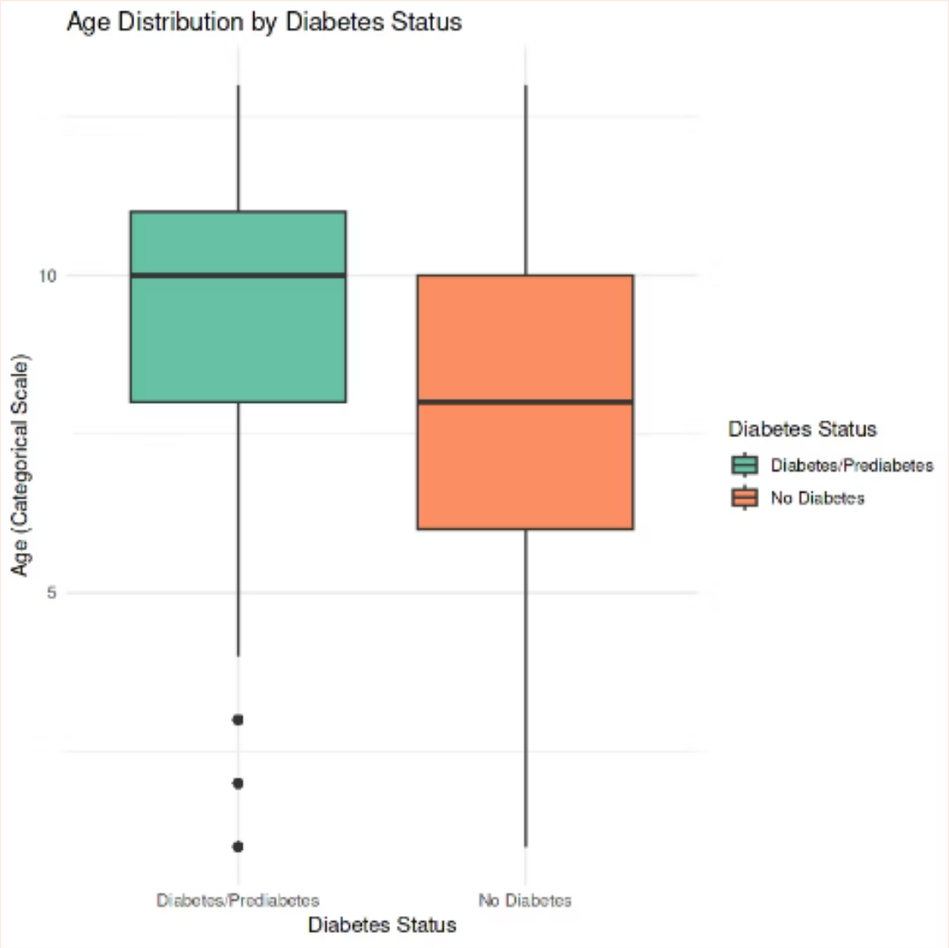
A bar plot was generated to clearly illustrate the 50/50 split of the dataset based on diabetes status, confirming the deliberate balancing of the two groups for effective model training.



Analyze BMI and Age Distributions

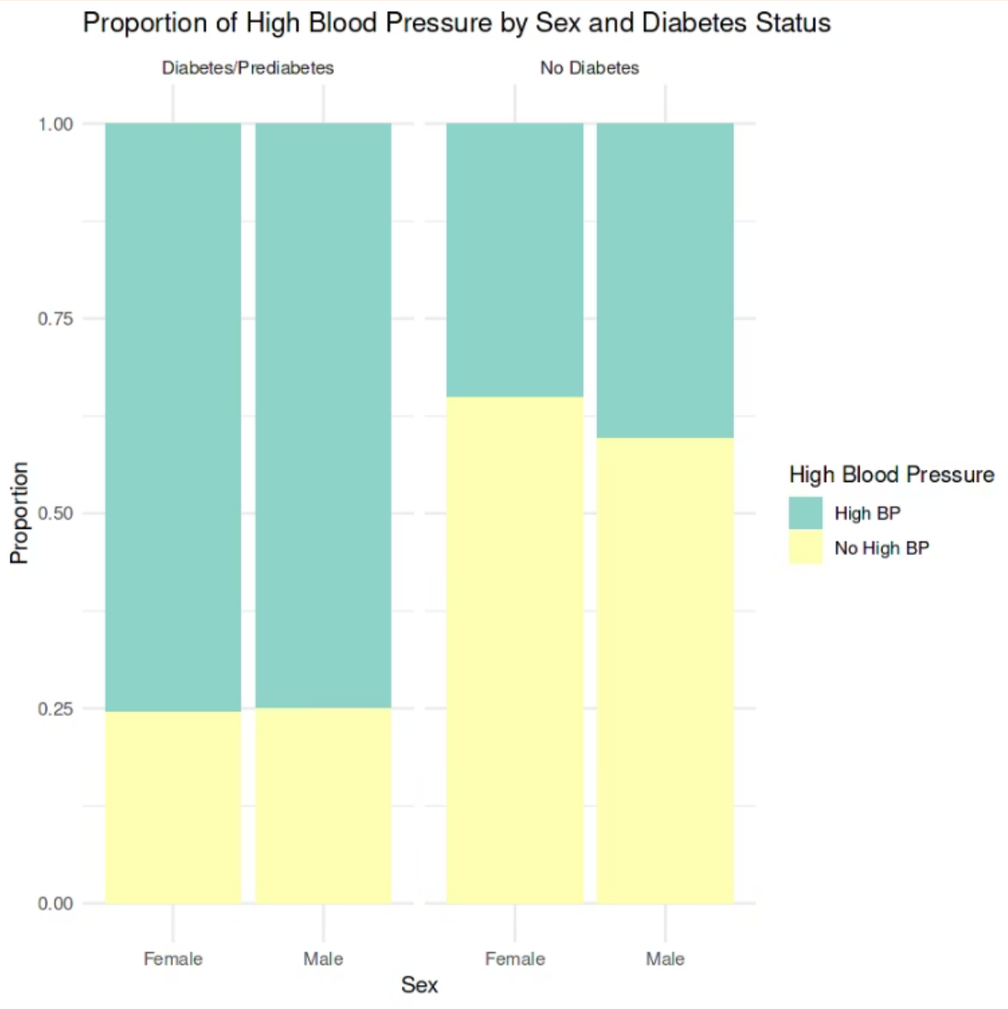
Histograms and comparative boxplots were used to visualize the distributions of Body Mass Index (BMI) and age. These revealed that individuals in the diabetes/prediabetes group generally exhibit higher BMI values and are older compared to the non-diabetes group.

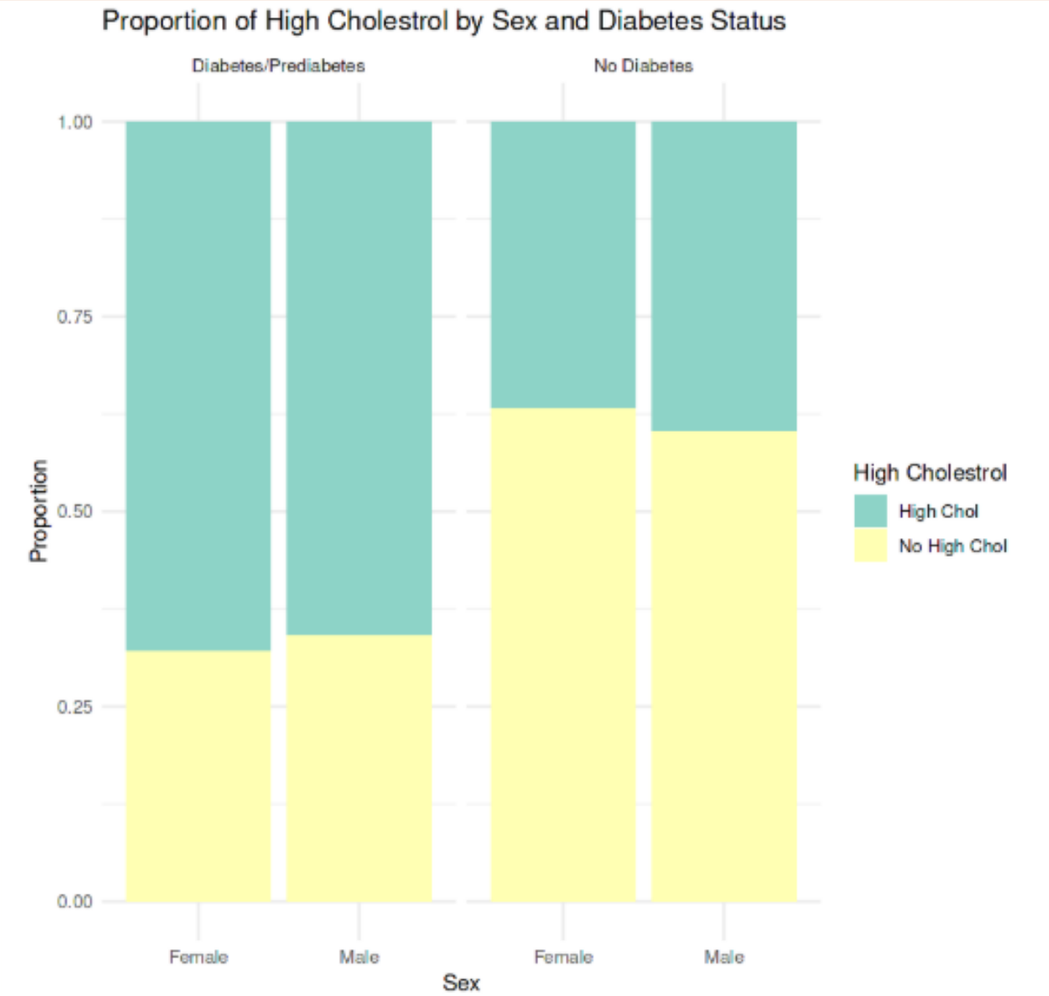




Examine HighBP, HighChol and Sex Proportions

A stacked bar plot was created to vividly illustrate the prevalence of high blood pressure (HighBP) and high cholesterol (HighChol) across different sexes and diabetes statuses, highlighting how this significant risk factor varies within demographic segments.





Data Loading and Initial Exploration

This section details the foundational steps of our study, including the ingestion of the dataset and a preliminary examination of its structure and content.

R Code for Data Loading and Summary

```
# Step 1: Load the dataset
data <- read_csv("/kaggle/input/diabetes-health-indicators-
dataset/diabetes_binary_5050split_health_indicators_BRFSS2015.csv",
show_col_types = FALSE)

# Step 2: Explore and preprocess the data
cat("\n=== Data Summary ===\n")

# Check for missing values
cat("Missing Values Check:\n")
print(colSums(is.na(data)))
# No missing values expected in this dataset

# Convert categorical variables to factors
data <- data %>%
         mutate(
         Diabetes_binary = as.factor(Diabetes_binary),
         Sex = as.factor(Sex),    Smoker = as.factor(Smoker),
HighBP =  as.factor(HighBP),    HighChol = as.factor(HighChol),
PhysActivity =     as.factor(PhysActivity)  )

# Summary statistics
cat("\nDescriptive Statistics of Variables:\n")
print(summary(data))
```

Output: Data Summary and Statistics

```
=== Data Summary ===
Missing Values Check:
Diabetes_binary            HighBP           HighChol
       0                    0                   0
CholCheck                  BMI               Smoker
       0                    0                   0
Stroke         HeartDiseaseorAttack         PhysActivity
  0                    0                   0
Fruits              Veggies    HvyAlcoholConsump
  0                    0                   0
AnyHealthcare          NoDocbcCost              GenHlth
          0                 0                    0
MentHlth             PhysHlth             DiffWalk
    0                    0                   0
Sex                  Age             Education
0                    0                   0
Income
0

Descriptive Statistics of Variables:

Diabetes_binary HighBP    HighChol    CholCheck           BMI        Smoker
0:35346         0:30860   0:33529   Min.   :0.0000   Min.   :12.00   0:37094   1:35346
1:39832   1:37163   1st Qu.:1.0000   1st Qu.:25.00   1:33598
Median :1.0000   Median :29.00                                                    Mean
:0.9753   Mean   :29.86                                                 3rd Qu.:1.0000
3rd Qu.:33.00        Max.   :1.0000   Max.   :98.00         ...
```

Logistic Regression Model

To answer our scientific question, we fitted a logistic regression model. This model predicts diabetes status based on age, sex, BMI, high blood pressure, high cholesterol, smoking, and physical activity.

```
# Step 3: Fit logistic regression model
cat("\n=== Logistic Regression Model ===\n")
cat("Model: Predicting Diabetes with Health Indicators\n")

# Predictors chosen based on health indicators relevant to diabetes
model <- glm( Diabetes_binary ~ Age + Sex + BMI + HighBP + HighChol + Smoker +PhysActivity, family = binomial(link = "logit"), data =
data)

# Step 4: Model summary
cat("\n=== Model Coefficients and Odds Ratios ===\n")
print(summary(model))
```

Model Diagnostics and Performance

We performed several diagnostic checks on our model. Model Coefficients and Odds Ratios,Effect modification for interaction between Age and BMT and The model's predictive performance was evaluated using the Area Under the ROC Curve (AUC), which showed a moderate predictive ability.

Model Coefficients and Odds Ratios

Calculating Coefficients and Odd ratios for all factors in our data

Effect Modification

An interaction term between Age and BMI was tested, revealing a significant effect modification, suggesting their combined impact on diabetes risk is more than additive.

ROC Curve and AUC

The ROC curve visually represents the model's performance, and the AUC value of approximately 0.79 indicates a moderate ability to distinguish between individuals with and without diabetes/prediabetes.

Model Coefficients and Odds Ratios

Calculating Coefficients and Odd ratios for all factors in our data

R Code

```
cat("\n=== Model Coefficients and Odds Ratios ===\n")
# Calculate odds ratios

odds_ratios <- exp(coef(model))


results <- cbind(odds_ratios)

print(results)
```

Output

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.995821   0.063625 -78.520  < 2e-16 ***
Age            0.170507   0.003528  48.327  < 2e-16 ***
BMI            0.090052   0.001507  59.756  < 2e-16 ***
Sex1           0.180890   0.017612  10.271  < 2e-16 ***
HighBP1        0.959728   0.018615  51.556  < 2e-16 ***
HighChol1      0.709013   0.017874  39.668  < 2e-16 ***
Smoker1        0.135470   0.017635   7.682 1.57e-14 ***
PhysActivity1 -0.344072   0.019408 -17.729  < 2e-16 ***
```

```
=== Odds Ratios ===
               odds_ratios
(Intercept)   0.006766162
Age           1.185905990
BMI           1.094231269
Sex1          1.198282976
HighBP1       2.610986828
HighChol1     2.031985408
Smoker1       1.145074843
PhysActivity1 0.708877866
```

Effect Modification

$$\mathrm{OR}_{\text{BMI at Age A}} = \exp(\beta_{\text{BMI}} + \beta_{\text{Age:BMI}} \times A)$$

R Code

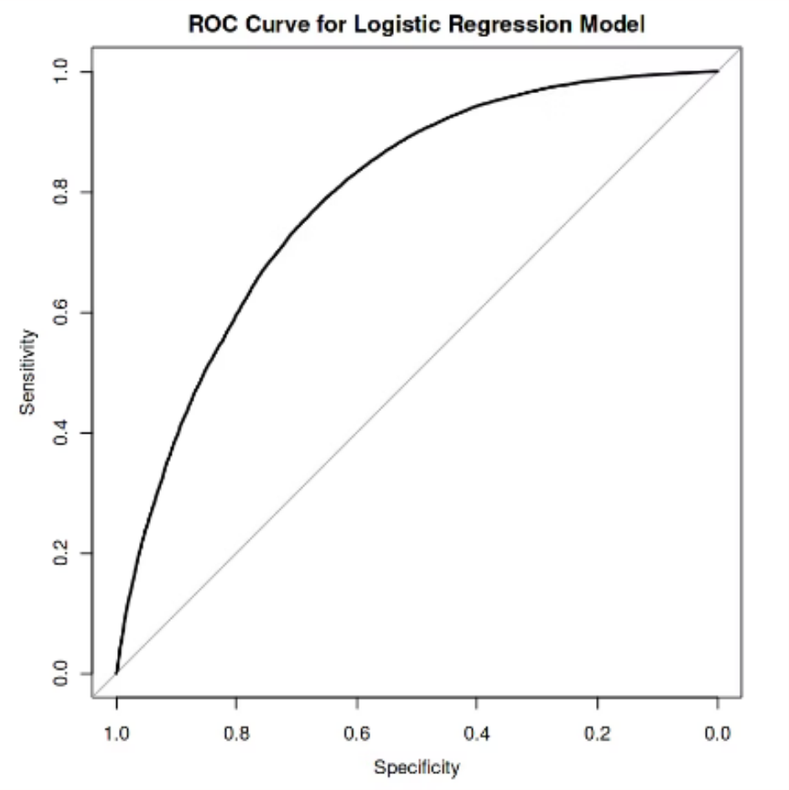
```
# Fit logistic regression model with interaction Term

model_interaction <- glm(
          Diabetes_binary ~ Age * BMI + Sex + HighBP + HighChol +
          Smoker + PhysActivity,  family = binomial(link = "logit"),
          data = data)

# Display summary to check interaction term

significancesummary(model_interaction)


# Calculate OR for BMI at different ages

beta_bmi <- log(1.094231269)  # From main effect

beta_interaction <- 0.0045964  # Hypothetical interaction term

ages <- c(30, 50,70)

or_bmi <- exp(beta_bmi + beta_interaction * ages)

names(or_bmi) <- paste("Age", ages)

print("Odds Ratios for BMI at different ages:")

print(or_bmi)
```

Output

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.9057326  0.1351352 -28.902  < 2e-16 ***
Age            0.0370070  0.0152472   2.427   0.0152 *
BMI            0.0529973  0.0043445  12.199  < 2e-16 ***
Sex1           0.1749476  0.0176345   9.921  < 2e-16 ***
HighBP1        0.9591559  0.0186338  51.474  < 2e-16 ***
HighChol1      0.7084835  0.0178969  39.587  < 2e-16 ***
Smoker1        0.1360606  0.0176518   7.708 1.28e-14 ***
PhysActivity1 -0.3400024  0.0194415 -17.488  < 2e-16 ***
Age:BMI        0.0045964  0.0005126   8.967  < 2e-16 ***
```

```
[1] "Odds Ratios for BMI at different ages:"
  Age 30   Age 50   Age 70
1.256015 1.376952 1.509533
```

ROC Curve and AUC



The ROC curve visually represents the model's performance, and the AUC value of approximately 0.79 indicates a moderate ability to distinguish between individuals with and without diabetes/prediabetes.

R Code

```
# Model diagnosticscat

("\n=== Model Performance: ROC and AUC ===\n")

# Predict probabilities

data$predicted_prob <- predict(model, type = "response")

# ROC curve and Area Under the Curve (AUC)

roc_obj <- roc(data$Diabetes_binary, data$predicted_prob)

auc_value <- auc(roc_obj)

cat("Area Under the ROC Curve (AUC):", auc_value, "\n")

plot(roc_obj, main = "ROC Curve for Logistic Regression Model")
```

```
=== Model Performance: ROC and AUC ===
Setting levels: control = Diabetes/Prediabetes, case = No Diabetes

Setting direction: controls < cases

Area Under the ROC Curve (AUC): 0.7902809
```

Model Evaluation

I tried to apply train-test split and k-fold cross validation methods to improve the accuracy of the model, 0.79 is normal but still not the best

Train Test Split

splitting the data into train and test, train data with 0.75 of the main data, and test is 0.25 of the main data

K-fold cross validation

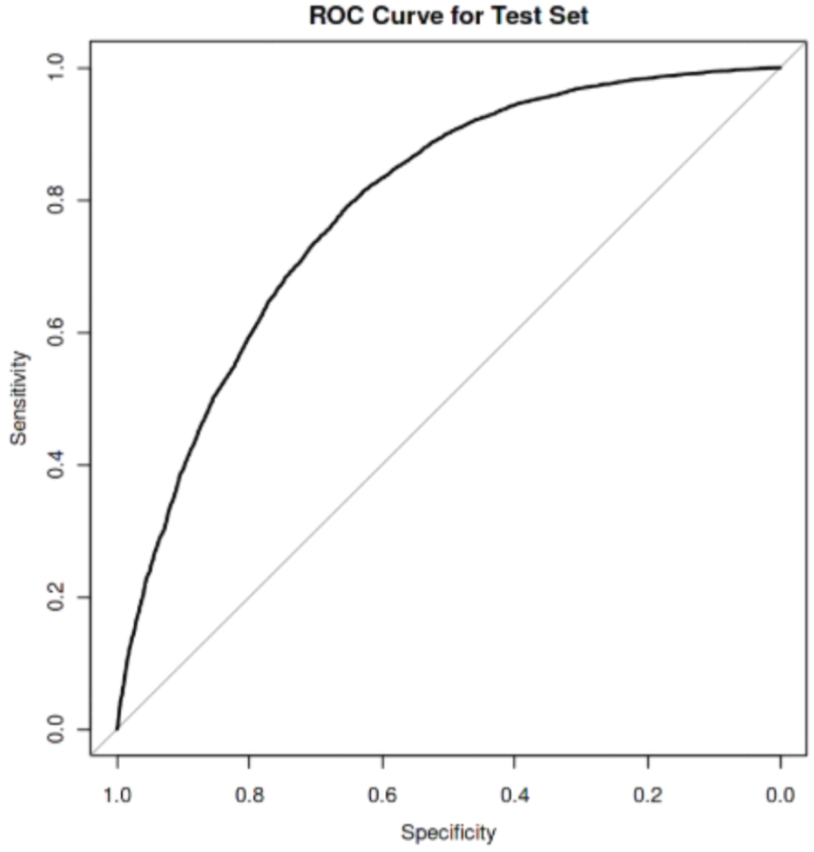
K fold is a method to split the data into k parts and use k-1 for training and 1 for testing, and repeating this process with all k parts

Train Test split

R Code

```
# Load additional libraries for evaluation

library(caret)
library(pROC)
library(lattice)

# Set seed for reproducibility

set.seed(123)

# Train-Test Split

cat("\n=== Model Evaluation: Train-Test Split ===\n")

train_index <- createDataPartition(data$Diabetes_binary, p = 0.75, list =
FALSE)

train_data <- data[train_index, ]test_data <- data[-train_index, ]

# Fit model on training data (same predictors as your original model)

model_train <- glm(Diabetes_binary ~ Age + Sex + BMI + HighBP + HighChol +
Smoker + PhysActivity,                    family = binomial(link = "logit"),
data = train_data)

# Predict on test datatest_data$predicted_prob <- predict(model_train, newdata
= test_data, type = "response")

# ROC and AUC for test data
roc_test <- roc(test_data$Diabetes_binary, test_data$predicted_prob)
auc_test <- auc(roc_test)cat("Test Set AUC:", auc_test, "\n")plot(roc_test,
main = "ROC Curve for Test Set")
```

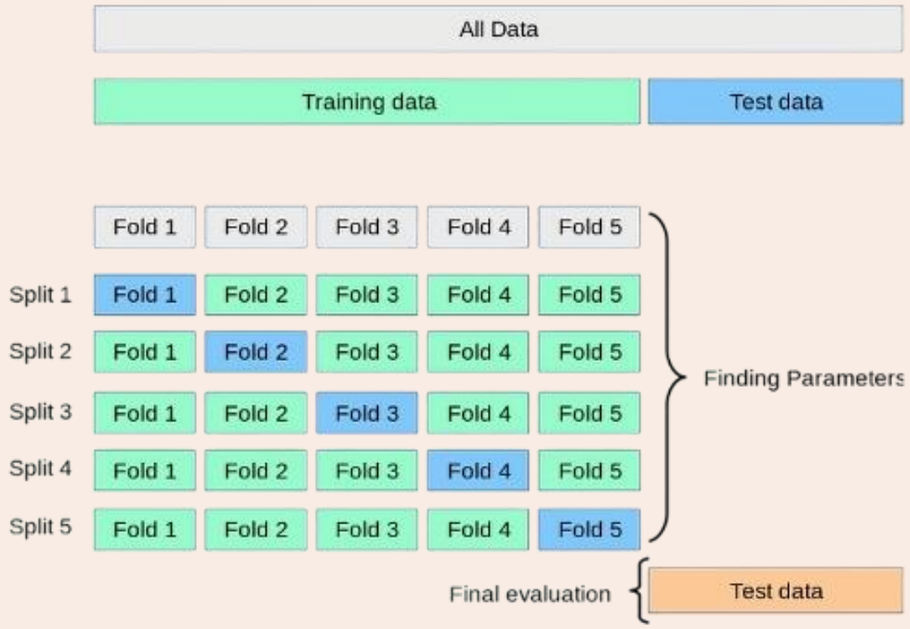
```
=== Model Evaluation: Train-Test Split ===
Setting levels: control = 0, case = 1

Setting direction: controls < cases

Test Set AUC: 0.7901041
```



K-fold cross validation

R Code

```
# Load necessary libraries (already in your code, repeated for
completeness)
library(caret)
library(pROC)

# Fix factor levels for Diabetes_binary to be valid R variable names

cat("\n=== Fixing Factor Levels for Diabetes_binary ===\n")
data$Diabetes_binary <-
as.factor(data$Diabetes_binary)levels(data$Diabetes_binary) <-
make.names(levels(data$Diabetes_binary))

# Convert to valid names

cat("Updated levels for Diabetes_binary:\n")
print(levels(data$Diabetes_binary))

# 5-Fold Cross-Validation (k=5) with classProbs = TRUE

cat("\n=== Model Evaluation: 5-Fold Cross-Validation ===\n")
set.seed(123)

# For reproducibility
cv_control <- trainControl(  method = "cv",  number = 5,
summaryFunction = twoClassSummary,  classProbs = TRUE

# Enable class probabilities for ROC)
cv_results <- train(  Diabetes_binary ~ Age + Sex + BMI + HighBP +
HighChol + Smoker + PhysActivity,  data = data,  method = "glm",
family = "binomial",  trControl = cv_control,  metric = "ROC")

# Extract and display cross-validated AUC
cv_auc <- cv_results$results$ROC
cat("Cross-Validated AUC:", cv_auc, "\n")
print(cv_results$results)
```



Output

```
=== Model Evaluation: 5-Fold Cross-Validation ===
Cross-Validated AUC: 0.7902278
  parameter       ROC      Sens      Spec      ROCSD      SensSD      SpecSD
1      none 0.7902278 0.6922705 0.7475244 0.00435297 0.005036996 0.005268072
```

Test the model

Trying to input sample data and see the results

R Code

```
trial_data <- data.frame(
  Age = 20,  Sex = factor(1, levels = c(0, 1)),
# No labels if model uses numeric factors
BMI = 19,
HighBP = factor(0, levels = c(0, 1)),  # Match model's format
HighChol = factor(0, levels = c(0, 1)),
Smoker = factor(0, levels = c(0, 1)),
PhysActivity = factor(1, levels = c(0, 1)))

# Ensure column names EXACTLY match training

datacolnames(trial_data) <- c("Age", "Sex", "BMI", "HighBP",
"HighChol", "Smoker", "PhysActivity")

# Predict

prob <- predict(model_train, newdata = trial_data, type =
"response")

# Converts to 0/1
binary_prediction <- ifelse(prob > 0.5, 'Diabetes', 'No diabetes')

cat(binary_prediction)
```

Output

First trial with:

age = 20 , BMI = 19 , HighBP = 0, HighChol = 0, Smoker = 0 , PhysActivity = 1

Output: **No diabetes**

Second trial with:

age = 35 , BMI = 19 , HighBP = 1, HighChol = 0, Smoker = 0 , PhysActivity = 1

Output: **Diabetes**

Conclusion and Limitations

Our analysis revealed significant associations between high blood pressure, high cholesterol, and higher BMI with increased diabetes/prediabetes risk. Older age and male sex also correlated with higher prevalence. While the model shows moderate predictive ability.

Key Findings

Strong link between HighBP, HighChol, BMI and diabetes risk.

Older age and male sex correlate with higher prevalence.

The Age:BMI interaction indicates that BMI's impact on diabetes risk is stronger in older individuals.

Model AUC of ~0.79 indicates moderate predictive ability.

Evaluation model techniques didn't improve the model

Thank You!

We appreciate your attention and engagement throughout this presentation. We hope our insights on diabetes risk factors have been valuable.

Questions?

Please feel free to ask any questions you may have. We are happy to discuss our findings further.

Contact Us

For more information or collaboration opportunities, please reach out:

- Email: ahmedtarek2632@gmail.com