Diabetes Risks Prediction

Diabetes Risks Prediction

Seif H. Kungulio

December 15, 2025

Contents

Business Understanding

Business objective

The primary business objective of this project is to support the Texas Department of State Health Services in enhancing diabetes prevention efforts across the state. By leveraging self-reported health behavior and demographic data, the goal is to develop a predictive model by 2026 that identifies Texas adults with a 30% or higher risk of developing diabetes within the next five years. This model aims to facilitate early detection, reduce long-term healthcare costs, promote health equity, and optimize resource allocation in vulnerable communities.

Problem statement

How can public health agencies in Texas use self-reported behavioral and demographic data to identify adults at 30%+ risk of developing diabetes within 5 years, using BMI (>30) marker, and implement targeted interventions by 2026?.

Business success criteria

The project will be considered successful if it:

- Produces a validated predictive model with practical thresholds for decision-making
- Demonstrates improved identification of high-risk populations compared to current screening protocols
- Supports targeted outreach that aligns with state health equity goals
- Provides actionable insights for stakeholders through dashboards or visual summaries

Data Understanding

Data collection

The primary dataset used is the **Diabetes Health Indicators Dataset** sourced from **Kaggle**, based on the 2015 Behavioral Risk Factor Surveillance System (BRFSS). This dataset includes over **253,000 observations** and **22 attributes**, offering a rich foundation for predictive modeling using demographic, behavioral, and clinical health indicators.

Load the data

```
diabetes <- read.csv("resources/diabetes_health_indicators_BRFSS2015.csv")
```

Sanity check

Sanity check for expected columns

```
stopifnot(all(c("Diabetes_binary","HighBP","HighChol","CholCheck","BMI",
                "Smoker","Stroke","HeartDiseaseorAttack","PhysActivity",
                "Fruits","Veggies","HvyAlcoholConsump","AnyHealthcare",
                "NoDocbcCost","GenHlth","MentHlth","PhysHlth","DiffWalk",
                "Sex","Age","Education","Income") %in% names(diabetes)))
```

Check for dimension and structure

Check for dimension and structure

```
dim(diabetes)
```

```
## [1] 253680     22
```

```
str(diabetes)
```

```
## 'data.frame':    253680 obs. of  22 variables:
##  $ Diabetes_binary     : num  0 0 0 0 0 0 0 0 1 0 ...
##  $ HighBP              : num  1 0 1 1 1 1 1 1 1 0 ...
##  $ HighChol            : num  1 0 1 0 1 1 0 1 1 0 ...
##  $ CholCheck           : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ BMI                 : num  40 25 28 27 24 25 30 25 30 24 ...
##  $ Smoker              : num  1 1 0 0 0 1 1 1 1 0 ...
##  $ Stroke              : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ HeartDiseaseorAttack: num  0 0 0 0 0 0 0 0 1 0 ...
##  $ PhysActivity        : num  0 1 0 1 1 1 0 1 0 0 ...
##  $ Fruits              : num  0 0 1 1 1 1 0 0 1 0 ...
##  $ Veggies             : num  1 0 0 1 1 1 0 1 1 1 ...
##  $ HvyAlcoholConsump   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ AnyHealthcare       : num  1 0 1 1 1 1 1 1 1 1 ...
##  $ NoDocbcCost         : num  0 1 1 0 0 0 0 0 0 0 ...
```

```
## $ GenHlth : num 5 3 5 2 2 2 3 3 5 2 ...
## $ MentHlth : num 18 0 30 0 3 0 0 0 30 0 ...
## $ PhysHlth : num 15 0 30 0 0 2 14 0 30 0 ...
## $ DiffWalk : num 1 0 1 0 0 0 0 1 1 0 ...
## $ Sex : num 0 0 0 0 0 1 0 0 0 1 ...
## $ Age : num 9 7 9 11 11 10 9 11 9 8 ...
## $ Education : num 4 6 4 3 5 6 6 4 5 4 ...
## $ Income : num 3 1 8 6 4 8 7 4 1 3 ...
```

Check summary

Basic summary

```r
skim(diabetes)
```

Table 1: Data summary

Name	diabetes
Number of rows	253680
Number of columns	22
Column type frequency:	
numeric	22
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Diabetes_binary	0	1	0.14	0.35	0	0	0	0	1	
HighBP	0	1	0.43	0.49	0	0	0	1	1	
HighChol	0	1	0.42	0.49	0	0	0	1	1	
CholCheck	0	1	0.96	0.19	0	1	1	1	1	
BMI	0	1	28.38	6.61	12	24	27	31	98	
Smoker	0	1	0.44	0.50	0	0	0	1	1	
Stroke	0	1	0.04	0.20	0	0	0	0	1	
HeartDiseaseorAttack	0	1	0.09	0.29	0	0	0	0	1	
PhysActivity	0	1	0.76	0.43	0	1	1	1	1	
Fruits	0	1	0.63	0.48	0	0	1	1	1	
Veggies	0	1	0.81	0.39	0	1	1	1	1	
HvyAlcoholConsump	0	1	0.06	0.23	0	0	0	0	1	
AnyHealthcare	0	1	0.95	0.22	0	1	1	1	1	
NoDocbcCost	0	1	0.08	0.28	0	0	0	0	1	
GenHlth	0	1	2.51	1.07	1	2	2	3	5	
MentHlth	0	1	3.18	7.41	0	0	0	2	30	
PhysHlth	0	1	4.24	8.72	0	0	0	3	30	
DiffWalk	0	1	0.17	0.37	0	0	0	0	1	
Sex	0	1	0.44	0.50	0	0	0	1	1	
Age	0	1	8.03	3.05	1	6	8	10	13	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Education	0	1	5.05	0.99	1	4	5	6	6	
Income	0	1	6.05	2.07	1	5	7	8	8	

Statistical summary

```
summary(diabetes)
```

```
##  Diabetes_binary      HighBP         HighChol        CholCheck
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.0000
##  Median :0.0000   Median :0.000   Median :0.0000   Median :1.0000
##  Mean   :0.1393   Mean   :0.429   Mean   :0.4241   Mean   :0.9627
##  3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.000   Max.   :1.0000   Max.   :1.0000
##       BMI            Smoker           Stroke         HeartDiseaseorAttack
##  Min.   :12.00   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:24.00   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :27.00   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :28.38   Mean   :0.4432   Mean   :0.04057   Mean   :0.09419
##  3rd Qu.:31.00   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :98.00   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##   PhysActivity        Fruits           Veggies        HvyAlcoholConsump
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.0000   1st Qu.:0.0000
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :0.0000
##  Mean   :0.7565   Mean   :0.6343   Mean   :0.8114   Mean   :0.0562
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  AnyHealthcare      NoDocbcCost         GenHlth          MentHlth
##  Min.   :0.0000   Min.   :0.00000   Min.   :1.000   Min.   : 0.000
##  1st Qu.:1.0000   1st Qu.:0.00000   1st Qu.:2.000   1st Qu.: 0.000
##  Median :1.0000   Median :0.00000   Median :2.000   Median : 0.000
##  Mean   :0.9511   Mean   :0.08418   Mean   :2.511   Mean   : 3.185
##  3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:3.000   3rd Qu.: 2.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :5.000   Max.   :30.000
##     PhysHlth         DiffWalk           Sex              Age
##  Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   : 1.000
##  1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 6.000
##  Median : 0.000   Median :0.0000   Median :0.0000   Median : 8.000
##  Mean   : 4.242   Mean   :0.1682   Mean   :0.4403   Mean   : 8.032
##  3rd Qu.: 3.000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:10.000
##  Max.   :30.000   Max.   :1.0000   Max.   :1.0000   Max.   :13.000
##    Education        Income
##  Min.   :1.00   Min.   :1.000
##  1st Qu.:4.00   1st Qu.:5.000
##  Median :5.00   Median :7.000
##  Mean   :5.05   Mean   :6.054
##  3rd Qu.:6.00   3rd Qu.:8.000
##  Max.   :6.00   Max.   :8.000
```

Class balance of target

```
diabetes %>%
  count(Diabetes_binary) %>%
  mutate(pct = n/sum(n)*100)
```

```
##   Diabetes_binary      n     pct
## 1               0 218334 86.0667
## 2               1  35346 13.9333
```

Data description

The dataset contains 253,680 records and 22 variables relevant to diabetes risk prediction. The target variable is binary (Diabetes_binary), while the predictors include behavioral, clinical, and demographic factors such as blood pressure, cholesterol, BMI, physical activity, general health, and income. Most variables (17) are binary categorical, with the remaining 5 being numeric or ordinal. This structure supports classification modeling and offers strong coverage of key health indicators needed to identify individuals at risk for diabetes.

Data dictionary

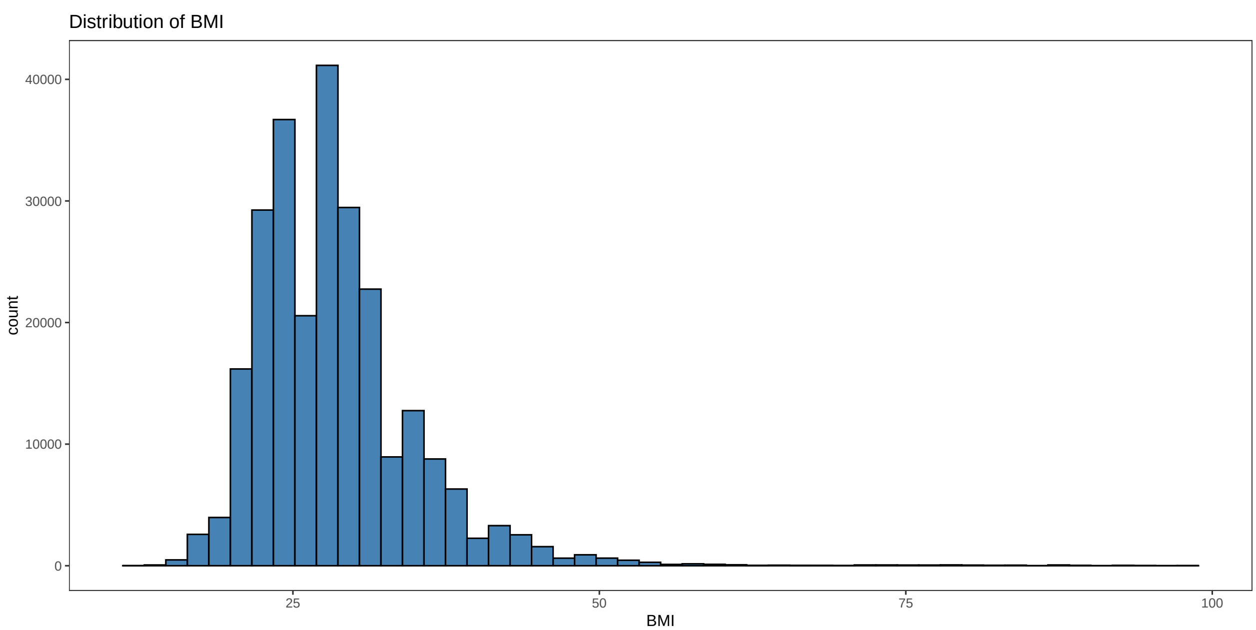
Variable	Data Type	Descriptions	Constraints
Diabetes_binary	Number	Diabetes status includes prediabetes.	Values: 0 or 1
HighBP	Number	Ever told you have high blood pressure.	Values: 0 or 1
HighChol	Number	Ever told you have high cholesterol.	Values: 0 or 1
CholCheck	Number	Cholesterol checked within the past 5 years.	Values: 0 or 1
BMI	Number	Body Mass Index (kg/m^2), calculated from self-reported height & weight.	Ranges: ~ 12 to 100
Smoker	Number	Smoked at least 100 cigarettes in lifetime.	Values: 0 or 1
Stroke	Number	Ever told you had a stroke.	Values: 0 or 1
HeartDiseaseorAttack	Number	Ever told you had coronary heart disease (CHD) or myocardial infarction (MI).	Values: 0 or 1
PhysActivity	Number	Any physical activity or exercise in past 30 days, not including job.	Values: 0 or 1
Fruits	Number	Consume fruit 1 or more times per day.	Values: 0 or 1
Veggies	Number	Consume vegetables 1 or more times per day.	Values: 0 or 1
HvyAlcoholConsump	Number	Heavy alcohol consumption (men >14 drinks/week; women >7 drinks/week).	Values: 0 or 1
AnyHealthcare	Number	Have any kind of health care coverage.	Values: 0 or 1
NoDocbcCost	Number	Could not see a doctor in the past 12 months because of cost.	Values: 0 or 1
GenHlth	Number	Self-rated general health (1 = Excellent, 2 = Very good, 3 = Good, 4 = Fair, 5 = Poor).	Ranges: 1 to 5
MentHlth	Number	Number of days mental health not good in past 30 days.	Ranges: 0 to 30
PhysHlth	Number	Number of days physical health not good in past 30 days.	Ranges: 0 to 30

Variable	Data Type	Descriptions	Constraints
DiffWalk	Number	Serious difficulty walking or climbing stairs.	Values: 0 or 1
Sex	Number	Sex (0 = Female, 1 = Male).	Values: 0 or 1
Age	Number	Age category (1=18 to 24, 2=25 to 29, 3=30 to 34, 4=35 to 39, 5=40 to 44, 6=45 to 49, 7=50 to 54, 8=55 to 59, 9=60 to 64, 10=65 to 69, 11=70 to 74, 12=75 to 79, 13=80+).	Values: 0 or 1
Education	Number	Education level ranges (1=Never attended/Kindergarten only, 2=Grades 1 to 8, 3=Grades 9 to 11, 4=Grade 12 or GED, 5=Some college or technical school, 6=College 4 years or more).	Ranges: 1 to 6
Income	Number	Household income ranges (1=<$10,000, 2=$10,000 to $15,000, 3=$15,000 to $20,000, 4=$20,000 to $25,000, 5=$25,000 to $35,000, 6=$35,000 to $50,000, 7=$50,000 to $75,000, 8=$75,000+)..	Ranges: 1 to 8

Pre-analysis visualization

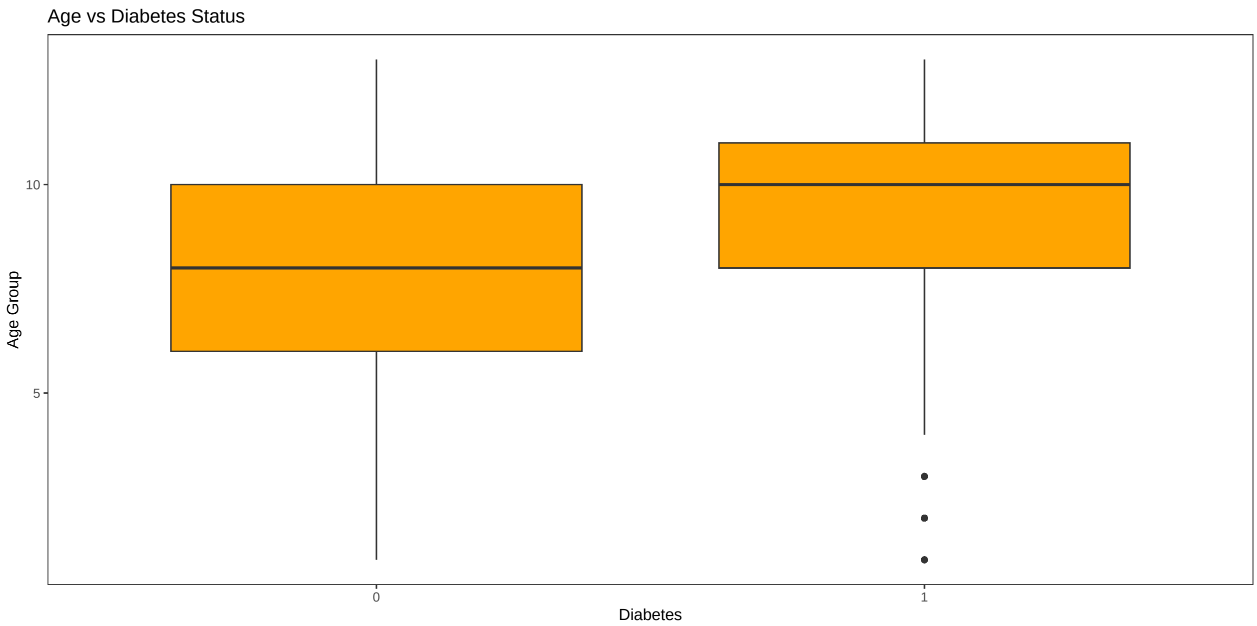
BMI Histogram

```r
ggplot(diabetes, aes(x = BMI)) +
  geom_histogram(bins = 50, fill = "steelblue", color = "black") +
  theme_test() + labs(title = "Distribution of BMI")
```



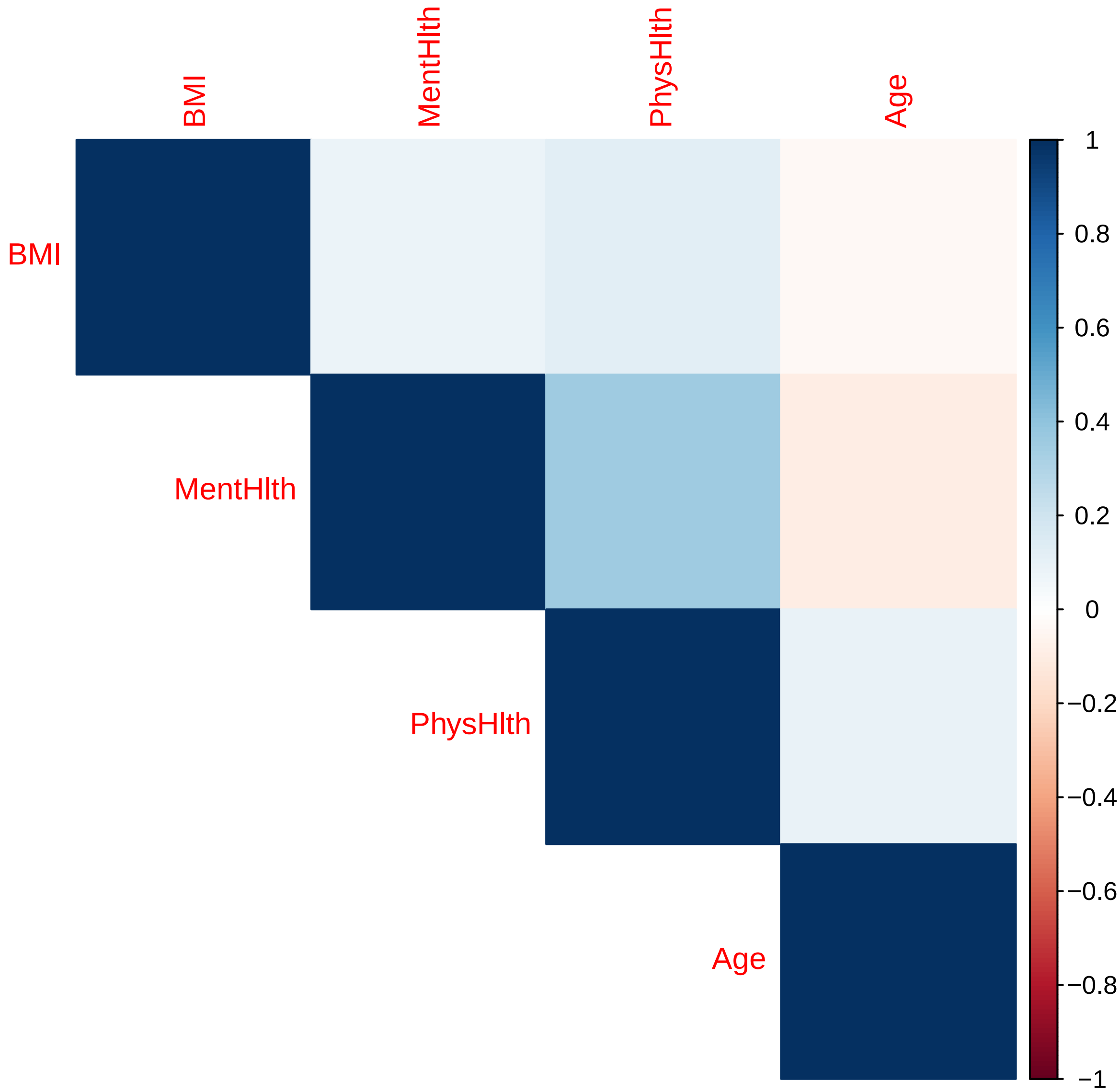
Age Boxplots

```r
ggplot(diabetes, aes(x = as.factor(Diabetes_binary), y = Age)) +
  geom_boxplot(fill = "orange") +
  labs(title = "Age vs Diabetes Status", x = "Diabetes", y = "Age Group") +
  theme_test()
```



Correlation among numeric variables

```r
numeric_vars <- diabetes %>% select(BMI, MentHlth, PhysHlth, Age)
corrplot(cor(numeric_vars), method = "color", type = "upper")
```



Data quality assessment

- **Completeness**: No traditional missing values; however, several variables use domain-specific placeholders (e.g., `88`, `77`, `99`) that denote `None`, `Don't know`, or `Refused`.
- **Duplicates**: 24,206 duplicate rows were removed.
- **Validity**: No string-based categorical variables; all attributes are encoded numerically.
- **Outliers**: Variables such as `BMI` have values up to `98`, suggesting the need for outlier treatment or binning.
- **Skewness**: Continuous variables like `BMI`, `MentHlth`, and `PhysHlth` are skewed, potentially impacting model performance.

Data Preparation

Remove duplicates

Check and remove duplicates if exist

```
diabetes <- distinct(diabetes)
```

Handle BRFSS placeholders

Handle BRFSS placeholders for MentHlth, PhysHlth, and GenHlth special codes/ placeholders. In some BRFSS releases, 88 can mean zero **0** days. Hence I will convert 77, 88, and 99 to NA.

```
placeholder_vals <- c(77, 88, 99)
diabetes <- diabetes %>%
  mutate(
    MentHlth = ifelse(MentHlth %in% placeholder_vals, NA, MentHlth),
    PhysHlth = ifelse(PhysHlth %in% placeholder_vals, NA, PhysHlth),
    GenHlth  = ifelse(GenHlth  %in% placeholder_vals, NA, GenHlth)
  )
```

Remove outliers

Remove outliers from BMI using IQR method

```
# Calculate IQR bounds for BMI
Q1 <- quantile(diabetes$BMI, 0.25, na.rm = TRUE)
Q3 <- quantile(diabetes$BMI, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Remove rows with BMI outliers
diabetes <- diabetes %>%
  filter(BMI >= lower_bound & BMI <= upper_bound)
```

Feature engineering

```
diabetes <- diabetes %>%
  mutate(
    # Target as factor with positive class "Yes"
    Diabetes_binary = factor(ifelse(Diabetes_binary == 1, "Yes", "No"),
                             levels = c("Yes","No")),

    # Engineered features
    Chronic_Risk_Load = HighBP + HighChol + Stroke + HeartDiseaseorAttack,
    # AnyHealthcare: 1=has coverage, 0=no coverage
    Healthcare_Barrier_Index = (1 - AnyHealthcare) + NoDocbcCost,
```

```r
    # Binned mental & physical health (reduce skew)
    MentHlth_bin = cut(MentHlth, breaks = c(-Inf, 0, 10, 20, 30),
                       labels = c("None","Low","Moderate","High"), right = TRUE),
    PhysHlth_bin = cut(PhysHlth, breaks = c(-Inf, 0, 10, 20, 30),
                       labels = c("None","Low","Moderate","High"), right = TRUE),

    # Age life-stage buckets from BRFSS age codes (1..13)
    AgeGroup3 = dplyr::case_when(
      Age %in% 1:4  ~ "18-34",
      Age %in% 5:8  ~ "35-54",
      Age %in% 9:13 ~ "55+",
      TRUE ~ NA_character_
    )
  ) %>%
  mutate(
    across(c(Smoker, PhysActivity, Fruits, Veggies, HvyAlcoholConsump,
             AnyHealthcare, NoDocbcCost, DiffWalk, Sex,
             HighBP, HighChol, Stroke, HeartDiseaseorAttack), ~ factor(.x)),
    GenHlth      = factor(GenHlth, levels = 1:5,
                          labels = c("Excellent","VeryGood","Good","Fair","Poor"),
                          ordered = TRUE),
    Education    = factor(Education, levels = 1:6, ordered = TRUE),
    Income       = factor(Income, levels = 1:8, ordered = TRUE),
    Age          = factor(Age, levels = 1:13, ordered = TRUE),
    AgeGroup3    = factor(AgeGroup3, levels = c("18-34","35-54","55+")),
    MentHlth_bin = factor(MentHlth_bin),
    PhysHlth_bin = factor(PhysHlth_bin)
  )
```

Train/Test split

```r
set.seed(123)
train_idx <- caret::createDataPartition(diabetes$Diabetes_binary,
                                        p = 0.8, list = FALSE)
train <- diabetes[train_idx, ]
test  <- diabetes[-train_idx, ]
```

Near-zero variance prune

```r
predictor_candidates <- setdiff(names(train), "Diabetes_binary")
nzv_info <- caret::nearZeroVar(train[, predictor_candidates], saveMetrics = TRUE)
keep_cols <- rownames(nzv_info)[!nzv_info$nzv]
train <- train[, c("Diabetes_binary", keep_cols)]
test  <- test [, c("Diabetes_binary", keep_cols)]
```

Scale numeric features

```
num_cols <- names(train)[sapply(train, is.numeric)]
if (length(num_cols)) {
  pp <- caret::preProcess(train[, num_cols, drop = FALSE],
                          method = c("center", "scale"))
  train[num_cols] <- predict(pp, train[num_cols, drop = FALSE])
  test [num_cols] <- predict(pp, test [num_cols,  drop = FALSE])
}
```

```
## Warning in `[.data.frame`(train, num_cols, drop = FALSE): 'drop' argument will
## be ignored

## Warning in `[.data.frame`(test, num_cols, drop = FALSE): 'drop' argument will
## be ignored
```

Remove incomplete cases

```
train <- tidyr::drop_na(train)
test  <- tidyr::drop_na(test)
```

Helper objects

```
predictor_cols <- setdiff(names(train), "Diabetes_binary")
pos_class <- "Yes"
threshold <- 0.5
```

Modeling

Evaluation

Deployment

Conclusion

References