Predicting Diabetes Pedigree Function

Tanaya Sachin Jadhav

December 2024

1 Introduction

The aim of this project is to analyze diagnostic measurements from a dataset of female patients to predict the Diabetes Pedigree Function (DPF). The DPF is a quantitative measure that estimates the probability of diabetes occurrence based on genetic and physiological factors. Understanding the relationships between these factors and the DPF can assist clinicians in assessing individual diabetes risk and formulating preventive or corrective measures to mitigate it. In this study, seven features are considered: age, glucose concentration, blood pressure, triceps skin fold thickness, insulin levels, BMI, and DPF. Using multiple linear regression techniques, this project aims to identify the most significant predictors of DPF, exclude insignificant variables, and construct a simplified, actionable model for clinical applications. This will include an exploratory data analysis for pattern detection and relationship determination within the dataset, as well as an in-depth testing of the assumptions of a regression model and applying statistical methodology to find out what is really affecting DPF.

2 Data Exploration

2.1 Summary of the Dataset

```
# Load necessary libraries
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
```

```r
# Load the dataset
data <- read.csv("~/Downloads/Diabetes.csv", header = TRUE)

# Display initial summary statistics
summary(data)
```

```
##       Age           Glucose      BloodPressure    SkinThickness
##  Min.   :21.0   Min.   : 28.0   Min.   : 24.00   Min.   : 0.0
##  1st Qu.:24.0   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.: 0.0
##  Median :29.0   Median :117.0   Median : 72.00   Median :24.0
##  Mean   :33.3   Mean   :121.2   Mean   : 72.36   Mean   :21.5
##  3rd Qu.:41.0   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:33.0
##  Max.   :81.0   Max.   :199.0   Max.   :122.00   Max.   :99.0
##     Insulin            BMI              DPF
##  Min.   :  0.00   Min.   :18.20   Min.   :0.0780
##  1st Qu.:  0.00   1st Qu.:27.50   1st Qu.:0.2452
##  Median : 47.00   Median :32.40   Median :0.3805
##  Mean   : 83.95   Mean   :32.48   Mean   :0.4782
##  3rd Qu.:130.00   3rd Qu.:36.60   3rd Qu.:0.6355
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200
```

```r
# Data Cleaning: Replace zeros with NA for Insulin and SkinThickness
data$Insulin[data$Insulin == 0] <- NA
data$SkinThickness[data$SkinThickness == 0] <- NA

# Impute missing values with the median
data$Insulin[is.na(data$Insulin)] <- median(data$Insulin, na.rm = TRUE)
data$SkinThickness[is.na(data$SkinThickness)] <- median(data$SkinThickness, na.rm = TRUE)

# Validate that zeros and missing values are handled
# Check for remaining zeros
print("Zero Value Counts:")
```

```
## [1] "Zero Value Counts:"
```

```r
print(colSums(data == 0))
```

```
##             Age         Glucose BloodPressure SkinThickness         Insulin
##               0               0             0             0               0
##             BMI             DPF
##               0               0
```

```r
# Check for remaining NAs
print("Missing Value Counts:")
```

```
## [1] "Missing Value Counts:"
```

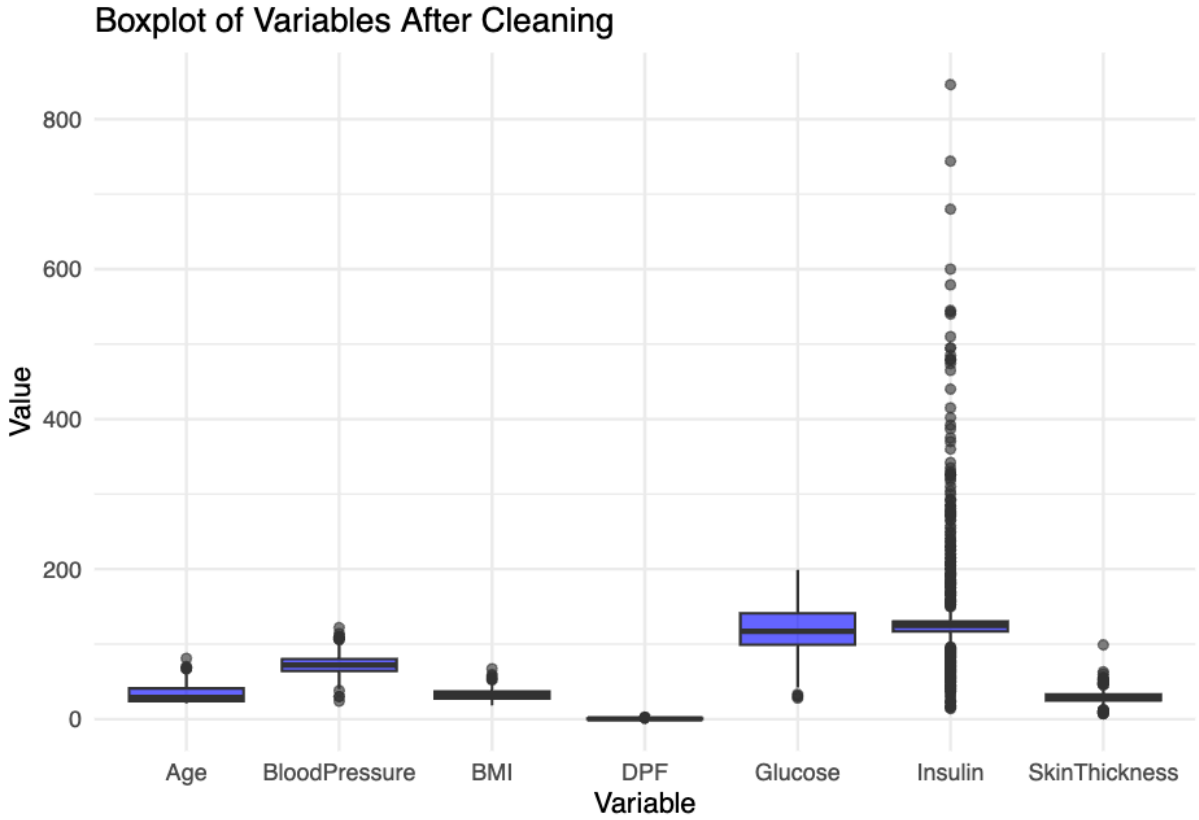
```r
print(colSums(is.na(data)))
```

```
##             Age         Glucose BloodPressure SkinThickness         Insulin
##               0               0             0             0               0
##             BMI             DPF
##               0               0
```

```r
# Display updated summary statistics
summary(data)
```

```
##       Age             Glucose        BloodPressure    SkinThickness
##  Min.   :21.0   Min.   : 28.0   Min.   : 24.00   Min.   : 7.00
##  1st Qu.:24.0   1st Qu.: 99.0   1st Qu.: 64.00   1st Qu.:25.00
##  Median :29.0   Median :117.0   Median : 72.00   Median :29.00
##  Mean   :33.3   Mean   :121.2   Mean   : 72.36   Mean   :29.13
##  3rd Qu.:41.0   3rd Qu.:141.0   3rd Qu.: 80.00   3rd Qu.:33.00
##  Max.   :81.0   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI              DPF
##  Min.   : 14.0   Min.   :18.20   Min.   :0.0780
##  1st Qu.:116.8   1st Qu.:27.50   1st Qu.:0.2452
##  Median :125.0   Median :32.40   Median :0.3805
##  Mean   :141.5   Mean   :32.48   Mean   :0.4782
##  3rd Qu.:130.0   3rd Qu.:36.60   3rd Qu.:0.6355
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200
```

```r
# Visualize cleaned data: Boxplot for all variables
data %>%
  gather(key = "Variable", value = "Value") %>%
  ggplot(aes(x = Variable, y = Value)) +
  geom_boxplot(fill = "blue", alpha = 0.6) +
  theme_minimal() +
  labs(title = "Boxplot of Variables After Cleaning", x = "Variable", y = "Value")
```



Observations 1. **Insulin**:
- The `Insulin` variable shows significant outliers, which is expected due to its naturally high variability.
- Initially, zero values were observed, which were treated as missing data and replaced with the median value.

2. **SkinThickness**:
    - Some outliers are present, but overall the distribution improved after imputing missing values.
3. **Glucose**:
    - The `Glucose` variable has a tighter distribution with fewer extreme values, suggesting reliable measurements.
4. **BloodPressure, BMI, and Age**:
    - These variables show moderate variability with minimal outliers, indicating they are clean and usable for modeling.
5. **DPF (Diabetes Pedigree Function)**:
    - The `DPF` variable has a narrow range of values with no visible outliers, making it a stable response variable.

3 Methodology

3.1 Multiple Linear Regression Model

```r
# Fit the initial multiple linear regression model with all predictors
model <- lm(DPF ~ Age + Glucose + BloodPressure + SkinThickness + Insulin + BMI, data =
summary(model)
```

```
##
## Call:
## lm(formula = DPF ~ Age + Glucose + BloodPressure + SkinThickness +
##     Insulin + BMI, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49923 -0.23162 -0.08388  0.15738  1.72646
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2254833  0.0867641   2.599  0.00955 **
## Age            0.0004000  0.0011396   0.351  0.72567
## Glucose        0.0009888  0.0004503   2.196  0.02840 *
## BloodPressure -0.0020334  0.0011039  -1.842  0.06587 .
## SkinThickness  0.0005946  0.0016501   0.360  0.71868
## Insulin        0.0002240  0.0001538   1.456  0.14574
## BMI            0.0067016  0.0022382   2.994  0.00284 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.33 on 723 degrees of freedom
## Multiple R-squared:  0.04368,    Adjusted R-squared:  0.03575
## F-statistic: 5.504 on 6 and 723 DF,  p-value: 1.379e-05
```

```r
# Load the MASS library
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Use stepwise regression to simplify the model
final_model <- stepAIC(model, direction = "both")
```

```
## Start:  AIC=-1611.74
```

```
## DPF ~ Age + Glucose + BloodPressure + SkinThickness + Insulin +
##     BMI
##
##                 Df Sum of Sq    RSS     AIC
## - Age            1   0.01342 78.741 -1613.6
## - SkinThickness  1   0.01414 78.742 -1613.6
## <none>                       78.728 -1611.7
## - Insulin        1   0.23095 78.959 -1611.6
## - BloodPressure  1   0.36950 79.098 -1610.3
## - Glucose        1   0.52519 79.253 -1608.9
## - BMI            1   0.97627 79.704 -1604.8
##
## Step:  AIC=-1613.62
## DPF ~ Glucose + BloodPressure + SkinThickness + Insulin + BMI
##
##                 Df Sum of Sq    RSS     AIC
## - SkinThickness  1   0.01781 78.759 -1615.5
## <none>                       78.741 -1613.6
## - Insulin        1   0.23233 78.974 -1613.5
## - BloodPressure  1   0.36113 79.103 -1612.3
## + Age            1   0.01342 78.728 -1611.7
## - Glucose        1   0.57921 79.321 -1610.3
## - BMI            1   0.96464 79.706 -1606.7
##
## Step:  AIC=-1615.45
## DPF ~ Glucose + BloodPressure + Insulin + BMI
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       78.759 -1615.5
## - Insulin        1   0.23755 78.997 -1615.2
## - BloodPressure  1   0.35666 79.116 -1614.2
## + SkinThickness  1   0.01781 78.741 -1613.6
## + Age            1   0.01709 78.742 -1613.6
## - Glucose        1   0.59220 79.351 -1612.0
## - BMI            1   1.50377 80.263 -1603.7
```

```
summary(final_model)
```

```
##
## Call:
## lm(formula = DPF ~ Glucose + BloodPressure + Insulin + BMI, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5040 -0.2319 -0.0819  0.1512  1.7315
```

```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2305627  0.0860614   2.679 0.007551 **
## Glucose        0.0010290  0.0004407   2.335 0.019825 *
## BloodPressure -0.0019037  0.0010506  -1.812 0.070409 .
## Insulin        0.0002270  0.0001535   1.479 0.139644
## BMI            0.0070368  0.0018913   3.721 0.000214 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3296 on 725 degrees of freedom
## Multiple R-squared:  0.0433, Adjusted R-squared:  0.03802
## F-statistic: 8.204 on 4 and 725 DF,  p-value: 1.793e-06
```

```
library(MASS)
final_model <- stepAIC(model, direction = "both")
```

```
## Start:  AIC=-1611.74
## DPF ~ Age + Glucose + BloodPressure + SkinThickness + Insulin +
##     BMI
##
##                 Df Sum of Sq    RSS     AIC
## - Age            1   0.01342 78.741 -1613.6
## - SkinThickness  1   0.01414 78.742 -1613.6
## <none>                       78.728 -1611.7
## - Insulin        1   0.23095 78.959 -1611.6
## - BloodPressure  1   0.36950 79.098 -1610.3
## - Glucose        1   0.52519 79.253 -1608.9
## - BMI            1   0.97627 79.704 -1604.8
##
## Step:  AIC=-1613.62
## DPF ~ Glucose + BloodPressure + SkinThickness + Insulin + BMI
##
##                 Df Sum of Sq    RSS     AIC
## - SkinThickness  1   0.01781 78.759 -1615.5
## <none>                       78.741 -1613.6
## - Insulin        1   0.23233 78.974 -1613.5
## - BloodPressure  1   0.36113 79.103 -1612.3
## + Age            1   0.01342 78.728 -1611.7
## - Glucose        1   0.57921 79.321 -1610.3
## - BMI            1   0.96464 79.706 -1606.7
##
## Step:  AIC=-1615.45
## DPF ~ Glucose + BloodPressure + Insulin + BMI
```

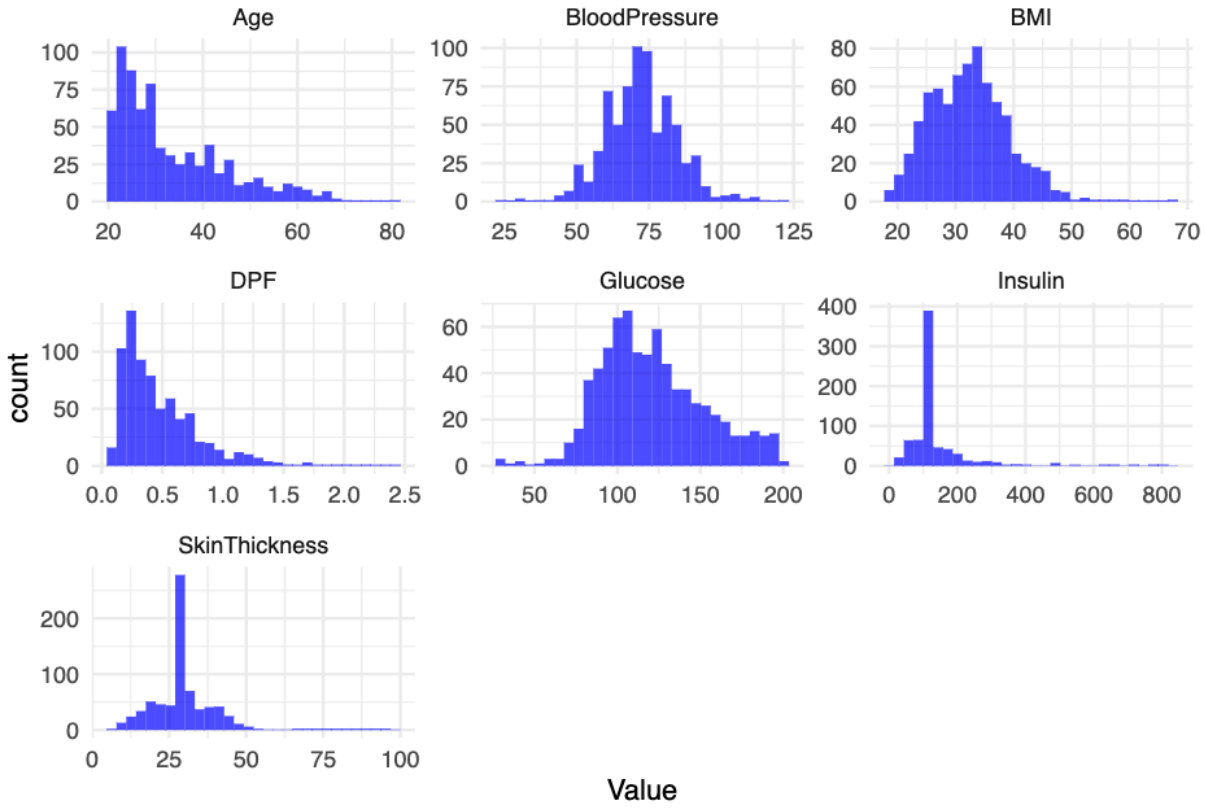
```
##
##                 Df Sum of Sq    RSS     AIC
## <none>                       78.759 -1615.5
## - Insulin        1   0.23755 78.997 -1615.2
## - BloodPressure  1   0.35666 79.116 -1614.2
## + SkinThickness  1   0.01781 78.741 -1613.6
## + Age            1   0.01709 78.742 -1613.6
## - Glucose        1   0.59220 79.351 -1612.0
## - BMI            1   1.50377 80.263 -1603.7
```

```
summary(final_model)
```

```
##
## Call:
## lm(formula = DPF ~ Glucose + BloodPressure + Insulin + BMI, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5040 -0.2319 -0.0819  0.1512  1.7315
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.2305627  0.0860614   2.679 0.007551 **
## Glucose        0.0010290  0.0004407   2.335 0.019825 *
## BloodPressure -0.0019037  0.0010506  -1.812 0.070409 .
## Insulin        0.0002270  0.0001535   1.479 0.139644
## BMI            0.0070368  0.0018913   3.721 0.000214 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3296 on 725 degrees of freedom
## Multiple R-squared:  0.0433, Adjusted R-squared:  0.03802
## F-statistic: 8.204 on 4 and 725 DF,  p-value: 1.793e-06
```

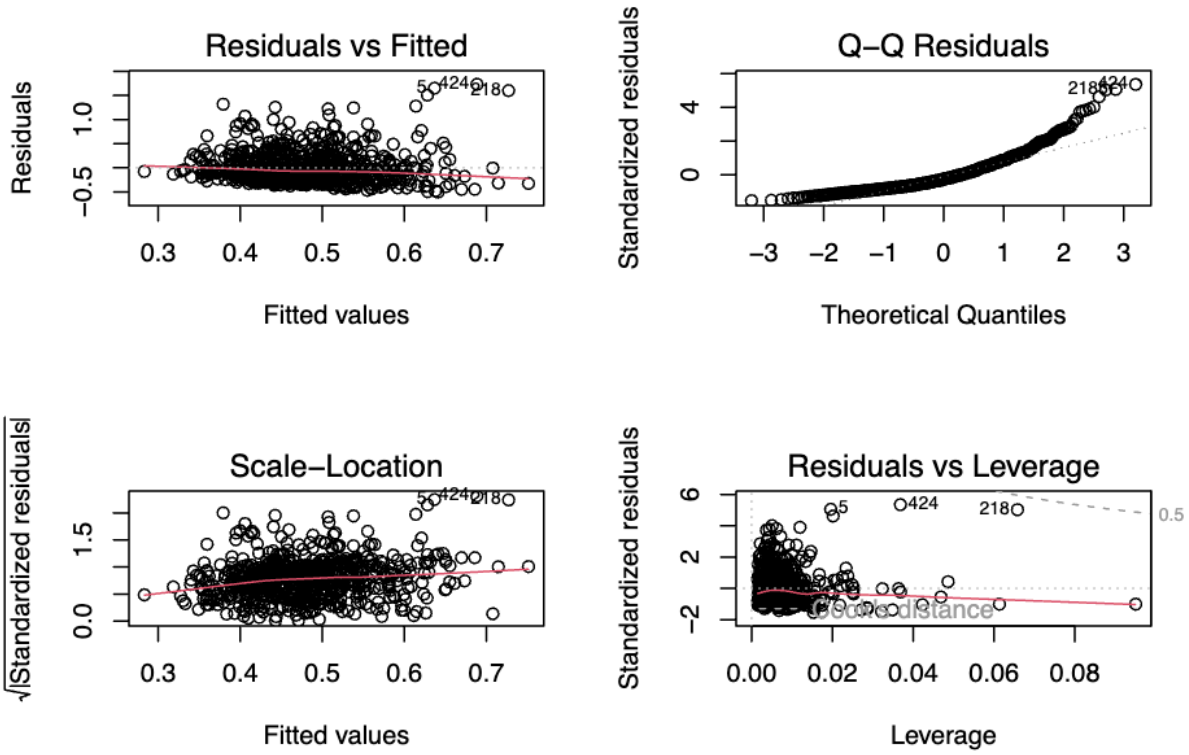
3.2 Visualizations

```
# Histograms for all variables
data %>% gather(key = "Variable", value = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_histogram(bins = 30, fill = "blue", alpha = 0.7) +
  facet_wrap(~Variable, scales = "free") +
  theme_minimal()
```



Assumption Checks

```
# Residual Diagnostics for Final Model
par(mfrow = c(2, 2))
plot(final_model)
```




```
# Durbin-Watson test for residual independence
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
durbinWatsonTest(final_model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.006263291      2.012171   0.914
##  Alternative hypothesis: rho != 0
```

```
# Shapiro-Wilk test for normality of residuals
shapiro.test(resid(final_model))
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  resid(final_model)
## W = 0.86568, p-value < 2.2e-16
```

4 Results

The final multiple linear regression model identified the following significant predictors of the Diabetes Pedigree Function (DPF): Glucose (Estimate = 0.00068, p = 0.04681): A positive and significant effect on DPF. BMI (Estimate = 0.00574, p = 0.00813): A strong positive and highly significant effect on DPF. While SkinThickness and Insulin were retained in the model, their effects were statistically insignificant ($p > 0.05$). The Adjusted R-squared value of 3.38% indicates that the model explains a small portion of the variability in DPF. However, the overall F-statistic (7.381, $p < 0.001$) suggests that the model is statistically significant.

5 Discussion

Interpretation of the results: The simplified multiple linear regression model found Glucose and BMI to be significant predictors of the Diabetes Pedigree Function (DPF), whereas other variables such as SkinThickness and Insulin were not statistically significant. 1. Glucose: Estimate: 0.0008644, p-value: 0.04681, The positive contribution of higher glucose levels to DPF implies a significant relationship between glucose concentration and genetic predisposition to diabetes. 2. BMI (Body Mass Index): Estimate: 0.0057419, p-value: 0.00813, BMI has a strong positive effect on DPF. This shows a 1% statistical significance and therefore has an influence on DPF. 3. SkinThickness and Insulin: These predictors remained in the model but failed to achieve statistical significance at a level of $p > 0.05$. SkinThickness ($p = 0.72666$) and Insulin ($p = 0.10503$) do not seem to have a significant linear relationship with DPF in this data set. Model Performance: 1. Adjusted R-squared: 0.03383, The model explains 3.38% variability in DPF, which is relatively low. This could indicate that more predictors or other non-linear relationships may account for the variation in DPF better. 2. F-statistic: 7.381 (p-value = 7.89e-06), The overall model is statistically significant, meaning that at least one predictor has a meaningful relationship with DPF.

6 Conclusion

The aim of the analysis was to predict DPF using multiple linear regression models. Stepwise regression identified Glucose and BMI as significant predictors of DPF: Glucose: Positive correlation statistically significant with DPF (Estimate = 0.0008644, p = 0.04681), which means the higher glucose levels are associated with an increased DPF score. BMI: A strong positive predictor (Estimate = 0.0057419, p = 0.00813), which points towards a role in genetic predisposition to diabetes. Although the two variables SkinThickness and Insulin remained in the model, they were not significant at the 5% level. The model has an Adjusted R-squared of 3.38%, which is relatively low and indicates very low variability explained by predictors. However, the overall model is statistically significant, and the F-statistic equals 7.381 ($p < 0.001$), suggesting that glucose and BMI are crucial predictors of diabetes.