ACTIVIDAD 1

ANÁLISIS DE DATOS

David León Fuentes

53915949Q

ÍNDICE

1. INTRODUCCIÓN

1.1 CONTEXTO Y MOTIVACIÓN

La diabetes es una enfermedad metabólica crónica que afecta a millones de personas en todo el mundo y se caracteriza por niveles elevados de glucosa en sangre. Por ello, predecir la evolución de la diabetes y monitorear sus principales indicadores es fundamental para mejorar el tratamiento y la calidad de vida de los pacientes. Entre los principales indicadores utilizados para evaluar el estado glucémico de una persona, la hemoglobina glucosilada (HbA1c) juega un papel clave, dado que refleja el promedio de glucosa en sangre durante los últimos tres meses, permitiendo evaluar el control a largo plazo de la enfermedad. A diferencia de una medición puntual de glucosa en sangre, que puede verse afectada por múltiples factores, HbA1c proporciona una visión más estable y confiable del estado metabólico de una persona.

Además de la diabetes, la obesidad es otra condición que ha alcanzado niveles epidémicos a nivel mundial y está estrechamente relacionada con el desarrollo de enfermedades metabólicas. Un indicador clave de la obesidad es el Índice de Masa Corporal (IMC), el cual se calcula a partir del peso y la altura de una persona. Valores elevados de IMC se asocian con un mayor riesgo de resistencia a la insulina y, en consecuencia, con un peor control glucémico.

Dado que tanto la hemoglobina glucosilada como el IMC juegan un papel fundamental en la diabetes y la obesidad, es relevante desarrollar modelos de predicción que permitan estimar estas variables en función de otros factores clínicos y de estilo de vida. Esta predicción puede ser de gran utilidad para la detección temprana de la diabetes y la obesidad, la personalización de tratamientos y la toma de decisiones médicas informadas.

1.2 OBJETIVOS DEL ANÁLISIS

Este trabajo tiene como objetivo desarrollar modelos de regresión para analizar y predecir variables clave relacionadas con la diabetes y la obesidad. En particular, se abordarán los siguientes objetivos específicos:

- **Modelo de Regresión Lineal Simple:** Analizar la relación entre la glucosa en sangre y la hemoglobina glucosilada (HbA1c) y evaluar la capacidad predictiva de la glucosa en la estimación de HbA1c.
- **Modelo de Regresión Lineal Múltiple:** Identificar los factores que influyen en el Índice de Masa Corporal (IMC) y desarrollar un modelo que permita predecir el IMC en función de diversas variables clínicas.
- **Modelo de Regresión Logística:** Predecir la presencia o ausencia de diabetes en función de diversas características de los pacientes, y evaluar la precisión del modelo en la clasificación de individuos en categorías de riesgo de diabetes.

A través de estos modelos, se pretende no solo entender mejor los factores que influyen en la diabetes y la obesidad, sino también generar herramientas predictivas que puedan ser aplicadas en entornos clínicos y de investigación. La posibilidad de anticipar niveles elevados de IMC, HbA1c o la presencia de diabetes a partir de datos accesibles podría contribuir significativamente a mejorar la prevención y el tratamiento de estas enfermedades.

2. DESCRIPCIÓN DE LOS DATOS A ANALIZAR

El dataset escogido para esta actividad contiene 13 variables y más de 9000 observaciones, por lo que en primera instancia es un buen candidato para cumplir los objetivos definidos en la actividad. Dicho dataset proviene de la siguiente fuente:

https://www.kaggle.com/datasets/asinow/diabetes-dataset/data

A continuación, se detallan las características de cada una de las variables que lo componen.

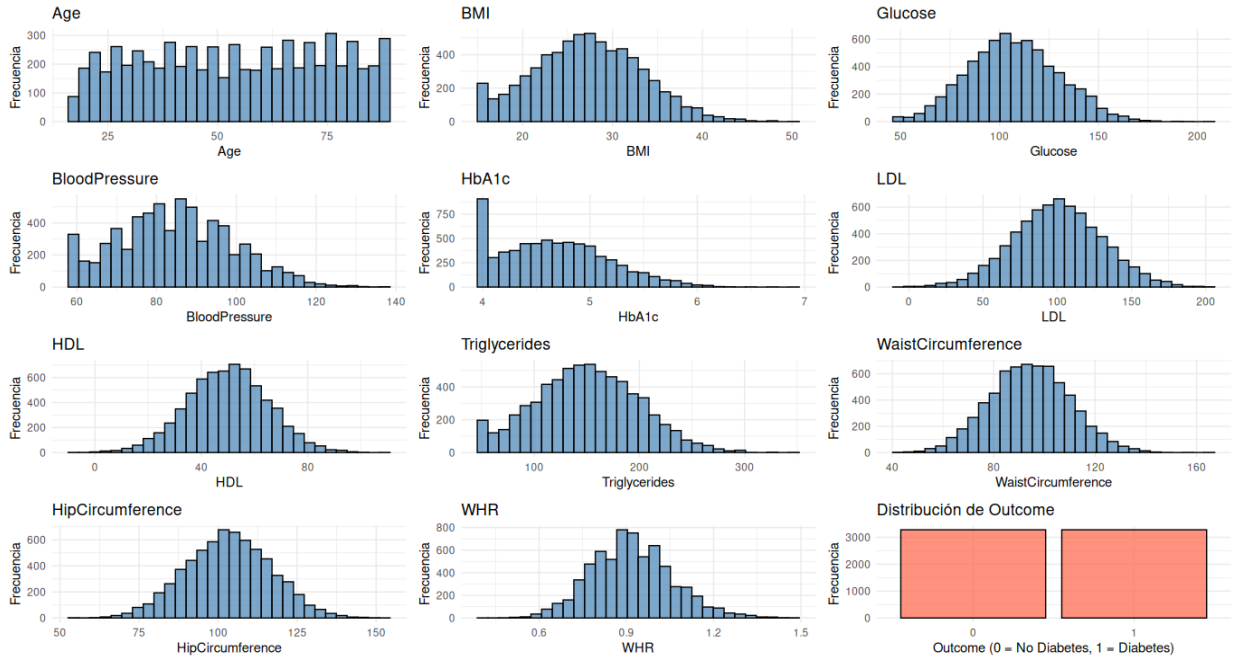
Variable	Descripción	Tipo de Variable
Age	Rango de edad de los individuos (18-90 años).	Continua (numérica)
Pregnancies	Número de veces que la paciente ha estado embarazada.	Discreta (numérica)
BMI (Body Mass Index)	Índice de masa corporal (kg/m²).	Continua (numérica)
Glucose	Concentración de glucosa en sangre (mg/dL).	Continua (numérica)
BloodPressure	Presión arterial sistólica (mmHg).	Continua (numérica)
HbA1c	Nivel de hemoglobina A1c (%).	Continua (numérica)
LDL (Low-Density Lipoprotein)	Nivel de colesterol 'malo' (mg/dL).	Continua (numérica)
HDL (High-Density Lipoprotein)	Nivel de colesterol 'bueno' (mg/dL).	Continua (numérica)
Triglycerides	Niveles de grasa en la sangre (mg/dL).	Continua (numérica)
WaistCircumference	Medida de la cintura (cm).	Continua (numérica)
HipCircumference	Medida de la cadera (cm).	Continua (numérica)
WHR (Waist-to-Hip Ratio)	Relación entre la circunferencia de la cintura/cadera.	Continua (numérica)
FamilyHistory	Indica si hay antecedentes familiares de diabetes	Binaria (categórica)
DietType	Hábitos dietéticos (0 = Desbalanceado, 1 = Balanceado, 2 = Vegano/Vegetariano).	Categórica (ordinal)
Hypertension	Presencia de hipertensión (1 = Sí, 0 = No).	Binaria (categórica)
MedicationUse	Indica si el individuo está tomando medicación (1 = Sí, 0 = No).	Binaria (categórica)
Outcome	Resultado del diagnóstico de diabetes (1 = Diabetes, 0 = No Diabetes).	Binaria (categórica)

Dado que uno de los requisitos de la actividad es que las variables sean numéricas, sólo se contemplarán para el análisis las variables **Age, BMI, Glucose, BloodPressure, HbA1c, LDL, HDL, Triglycerides, WaistCircumference, HipCircumference** y **WHR**, incluyéndose además la variable binaria **Outcome** como variable dependiente para la sección de regresión logística. Para realizar un primer análisis rápido sobre los datos, atenderemos a las estadísticas que arroja R al solicitar un resumen de las variables de interés (Anexo I). La salida que obtenemos es la reflejada en la siguiente figura.

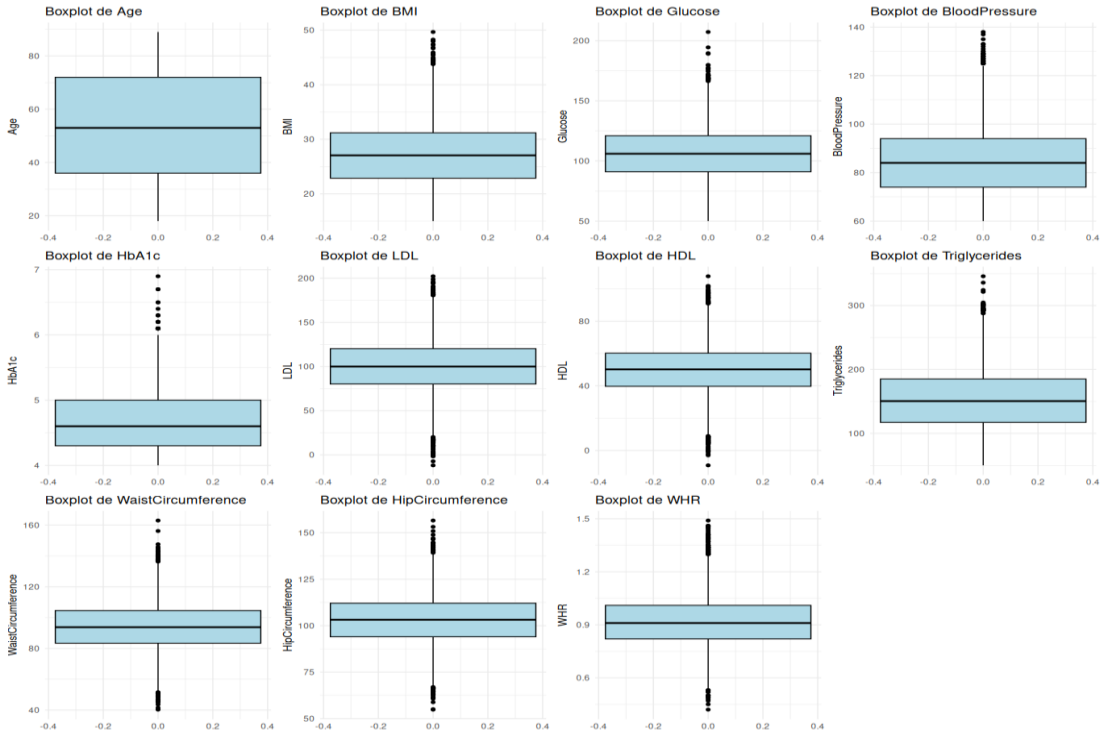
```
      Age         Pregnancies          BMI            Glucose      BloodPressure        HbA1c            LDL        
 Min.   :18.00   Min.   : 0.000   Min.   :15.00   Min.   : 50.0   Min.   : 60.00   Min.   :4.000   Min.   :-12.0  
 1st Qu.:36.00   1st Qu.: 4.000   1st Qu.:22.87   1st Qu.: 91.0   1st Qu.: 74.00   1st Qu.:4.300   1st Qu.: 80.1  
 Median :53.00   Median : 8.000   Median :27.05   Median :106.0   Median : 84.00   Median :4.600   Median : 99.9  
 Mean   :53.58   Mean   : 7.986   Mean   :27.05   Mean   :106.1   Mean   : 84.48   Mean   :4.651   Mean   :100.1  
 3rd Qu.:72.00   3rd Qu.:12.000   3rd Qu.:31.18   3rd Qu.:121.0   3rd Qu.: 94.00   3rd Qu.:5.000   3rd Qu.:120.2  
 Max.   :89.00   Max.   :16.000   Max.   :49.66   Max.   :207.2   Max.   :138.00   Max.   :6.900   Max.   :202.2  

      HDL         Triglycerides   WaistCircumference HipCircumference      WHR            Outcome      
 Min.   : -9.20   Min.   : 50.0   Min.   : 40.30     Min.   : 54.8    Min.   :0.4200   Min.   :0.0000  
 1st Qu.: 39.70   1st Qu.:117.2   1st Qu.: 83.40     1st Qu.: 94.0    1st Qu.:0.8200   1st Qu.:0.0000  
 Median : 50.20   Median :150.6   Median : 93.80     Median :103.2    Median :0.9100   Median :0.0000  
 Mean   : 49.95   Mean   :151.1   Mean   : 93.95     Mean   :103.1    Mean   :0.9174   Mean   :0.3441  
 3rd Qu.: 60.20   3rd Qu.:185.1   3rd Qu.:104.60     3rd Qu.:112.1    3rd Qu.:1.0100   3rd Qu.:1.0000  
 Max.   :107.80   Max.   :345.8   Max.   :163.00     Max.   :156.6    Max.   :1.4900   Max.   :1.0000  
```

El análisis inicial muestra que la muestra tiene una distribución de edad amplia, centrada en adultos mayores (~53 años). Además, el 34% de los pacientes tienen diabetes, lo que genera un desequilibrio en los datos y puede sesgar los modelos predictivos. Para corregirlo, se ajusta la muestra eliminando aleatoriamente observaciones de no diabéticos hasta alcanzar un 50% en ambas clases, sin afectar la calidad del dataset. Posteriormente, se analizan las distribuciones mediante histogramas y un gráfico de barras para la variable **Outcome**, verificando el balance de la población.



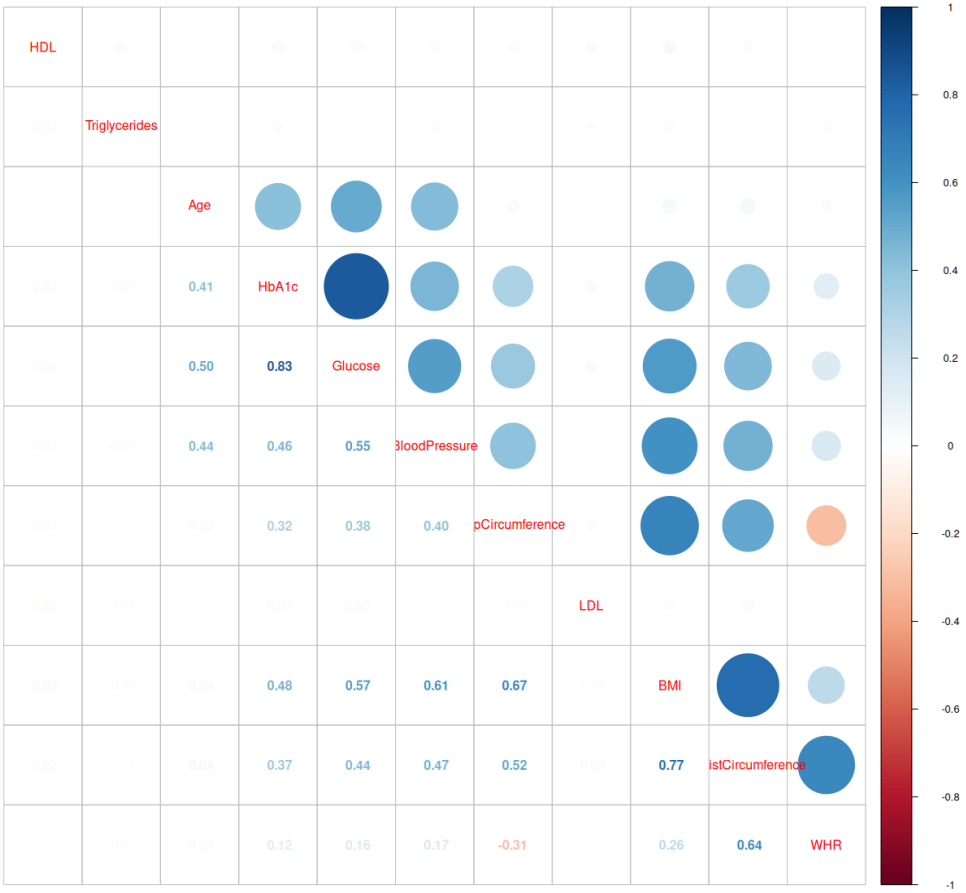
Además, graficamos también los diagramas de cajas y bigotes asociados a cada una de las variables continuas del dataset, quedando dichos diagramas reflejados en la siguiente figura.



Analizando los gráficos, se extraen las siguientes conclusiones:

- **Balance de la Muestra:** La muestra está equilibrada, con igual número de personas con y sin diabetes.
- **Rango de Edad:** La distribución etaria abarca de 18 a 89 años sin sesgos aparentes.
- **Distribución de Variables:** LDL, HDL, HipCircumference y WHR siguen una distribución normal simétrica, indicando ausencia de sesgos extremos.
- **Análisis de Otras Variables:** La mayoría presentan distribuciones similares a una campana, aunque algunas están desplazadas. Los diagramas de cajas revelan valores atípicos, pero estos podrían reflejar características propias de la población.
- **Necesidad de Escalado:** Dada la variabilidad en las escalas de las variables, es necesario normalizar los datos para evitar problemas en los modelos de regresión.

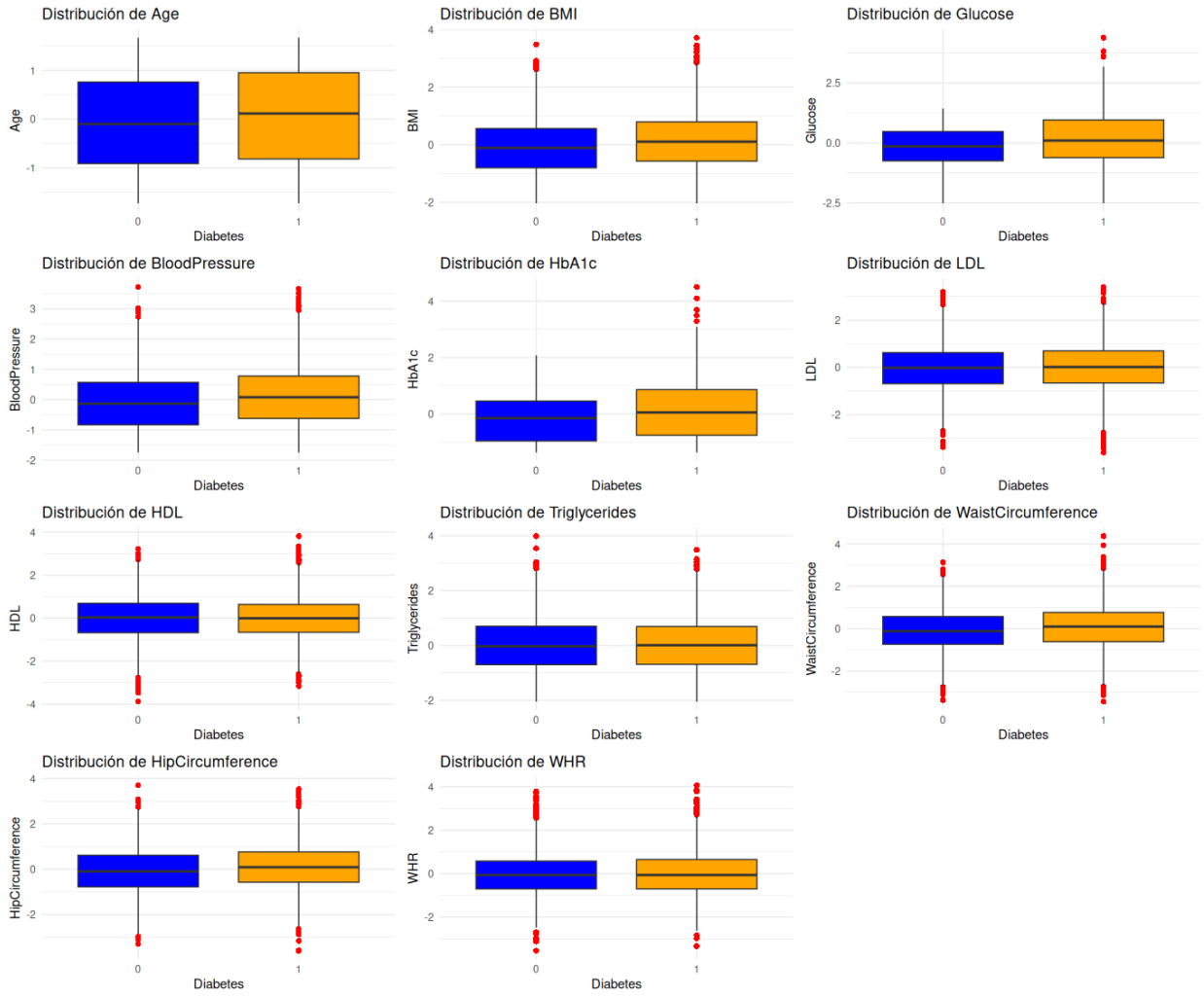
A continuacion, se grafica el mapa de correlación de las variables continuas, como se muestra en la siguiente figura.



A partir del mapa de correlación, se extraen las siguientes conclusiones:

- **Regresión Lineal Simple:** La variable **HbA1c** muestra su mayor correlación con **Glucose**, lo cual es esperable, ya que ambas reflejan el control glucémico, aunque en diferentes escalas temporales.
- **Regresión Lineal Múltiple:** Para predecir el Índice de Masa Corporal *(***BMI***)*, las variables más correlacionadas son **WaistCircumference**, **HipCircumference**, **BloodPressure**, **Glucose** y **HbA1c**. La relación con las medidas de cintura y cadera es coherente, ya que el **BMI** suele estar asociado con mayores dimensiones corporales.

Finalmente, para la exploración inicial de las variables asociadas a la **regresión logística**, se analizan las relaciones entre la presencia de diabetes (**Outcome**) y las demás variables mediante diagramas de cajas y bigotes, observando la distribución de los valores en los grupos de diabéticos y no diabéticos. Estos gráficos se presentan en la siguiente figura.



A partir de los gráficos, se identifican las variables con mayores diferencias entre grupos y que podrían ser buenos predictores: **Glucose**, **HbA1c**, **BMI**, **WaistCircumference** y **WHR**. Los pacientes con diabetes muestran niveles más altos de glucosa y HbA1c, reflejando su impacto en el control glucémico. Además, presentan un BMI, circunferencia de cintura y relación cintura-cadera mayores, lo que sugiere una posible influencia de la distribución de grasa corporal en la enfermedad.

Por otro lado, algunas variables muestran diferencias menos marcadas y podrían no ser predictoras fuertes. **BloodPressure** tiene solo una ligera variación y sus distribuciones se solapan. **HDL** (colesterol bueno) no muestra una distinción clara entre grupos, y aunque los diabéticos presentan una **HipCircumference** ligeramente mayor, la diferencia no parece determinante.

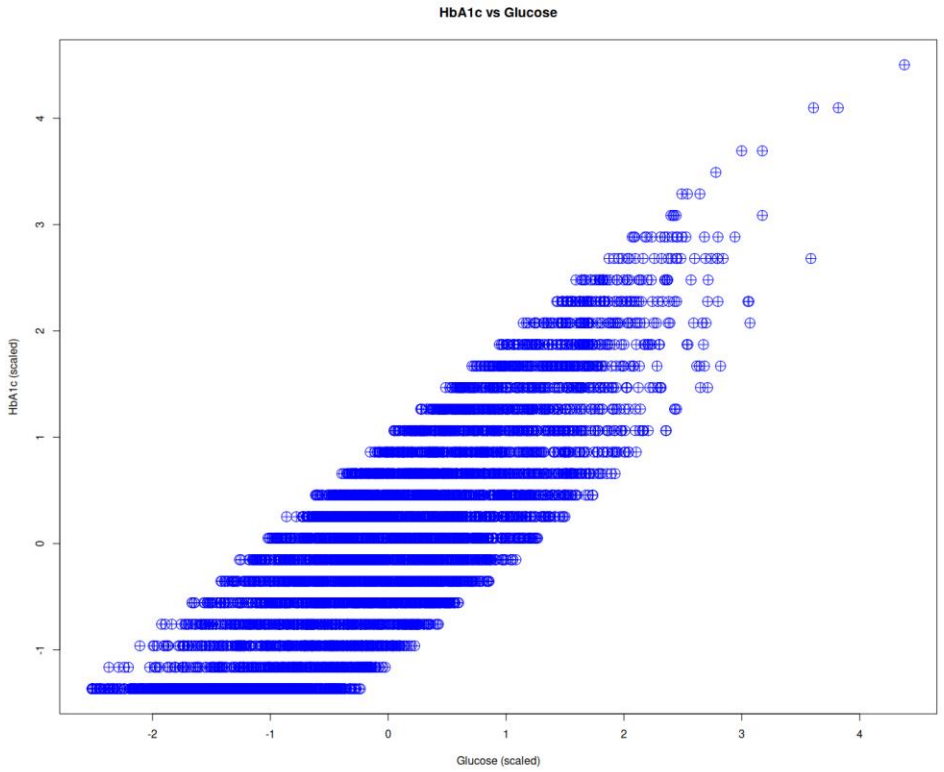
Este análisis confirma la coherencia de las variables seleccionadas para los modelos de regresión, ya que presentan asociaciones significativas con los indicadores clave.

3. ANÁLISIS

3.1. REGRESIÓN LINEAL / POLINÓMICA

3.1.1 Creación del modelo

El objetivo de esta sección es analizar la relación entre la glucosa en sangre y la hemoglobina glucosilada (**HbA1c**), evaluando la capacidad predictiva de la glucosa en la estimación de **HbA1c** mediante un modelo de regresión lineal o polinómica. En la siguiente figura se observa que la relación entre ambas variables parece ser lineal, por lo que se optará por un modelo de regresión lineal para su construcción.



Aunque la variable **HbA1c** es numérica continua, en los gráficos parece discreta. Esto se debe a un efecto visual causado probablemente por la resolución limitada del aparato de medición utilizado en la creación del dataset, lo que no afecta su naturaleza continua. Tras determinar el tipo de relación entre **Glucose** y **HbA1c**, se procede a la construcción del modelo, cuyas estadísticas se presentan a continuación.

```
Call:
lm(formula = HbA1c ~ Glucose, data = data_balanced)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16982 -0.45665  0.01145  0.46227  1.21982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.491e-15  6.870e-03     0.0        1
Glucose     8.308e-01  6.871e-03   120.9   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5566 on 6562 degrees of freedom
Multiple R-squared:  0.6902,    Adjusted R-squared:  0.6902
F-statistic: 1.462e+04 on 1 and 6562 DF,  p-value: < 2.2e-16
```

Atendiendo a las estadísticas del modelo, se sacan las siguientes conclusiones:

- **Significancia de las variables**: La variable **Glucose** tiene un $p_{valor} < 2 * 10^{-16} < \alpha = 0.05$, lo que indica que es altamente significativa en la predicción de **HbA1c**. El **Intercept** tiene un p-valor de 1, lo que sugiere que no es significativamente diferente de cero, indicando que la función cruza el origen.
- **Dirección y magnitud del coeficiente**: El coeficiente de **Glucose** (0.8308) indica que, por cada aumento de 1 unidad en la glucosa, el valor estimado de **HbA1c** aumenta en 0.83 unidades, manteniendo todo lo demás constante.
- **Calidad del ajuste del modelo**: El R^2 es 0.6902, lo que significa que el modelo explica el 69.02% de la variabilidad en **HbA1c** a partir de **Glucose**. El error estándar de los residuos es 0.5566, lo que indica que las predicciones tienen una desviación promedio de 0.56 unidades de **HbA1c** respecto a los valores reales.
- **Significación global del modelo**: El modelo tiene un $p_{valor} < 2.2 * 10^{-16} < \alpha = 0.05$, lo que indica que es estadísticamente significativo, y que **Glucose** tiene un impacto real en la predicción de **HbA1c**.

Por tanto, la función del modelo de regresión lineal simple resultante es la siguiente:

$$\text{HbA1c} = 2.491 \times 10^{-15} + 0.8308 \times \text{Glucose}$$

Sin embargo, dado que las variables han sido previamente escaladas, la fórmula obtenida es una relación en unidades estándar, es decir, las unidades de **Glucose** y **HbA1c** han sido transformadas en desviaciones estándar. Por ello, para usar esta fórmula con los valores originales (sin escalar), es necesario deshacer el escalado, con lo que la función del modelo de regresión simple resultaría de la siguiente forma:

$$\text{HbA1c} = \left[\left(2.491 \times 10^{-15} + 0.8308 \times \frac{(\text{Glucose} - \mu_{\text{Glucose}})}{\sigma_{\text{Glucose}}}\right) \times \sigma_{\text{HbA1c}}\right] + \mu_{\text{HbA1c}}$$

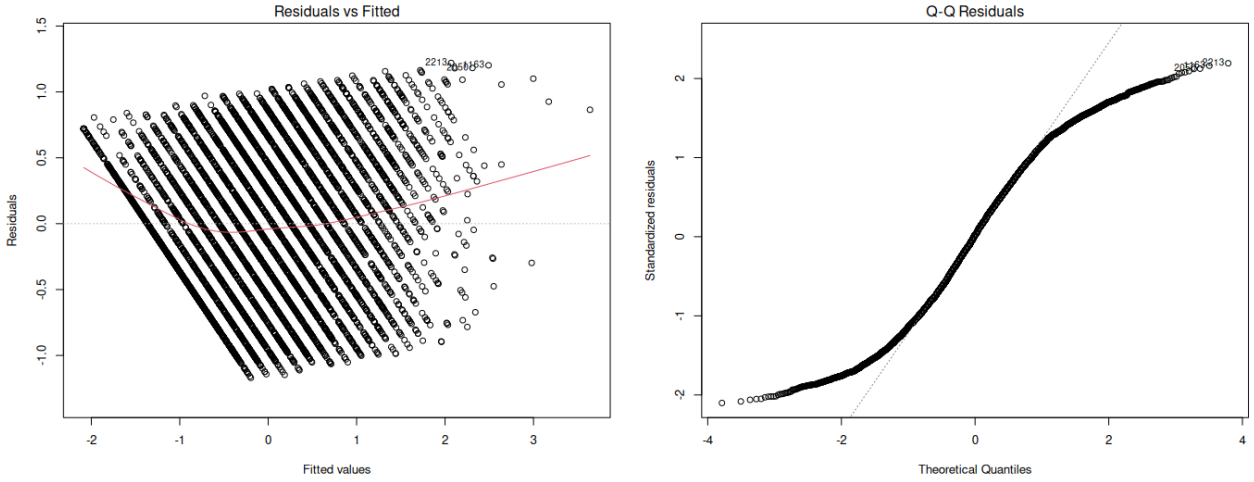
Donde:

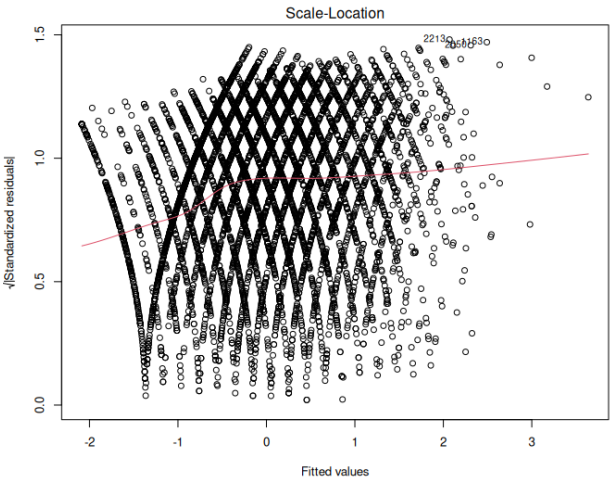
- μ_x es la media de la variable x.
- σ_x es la desviación estándar de la variable x.

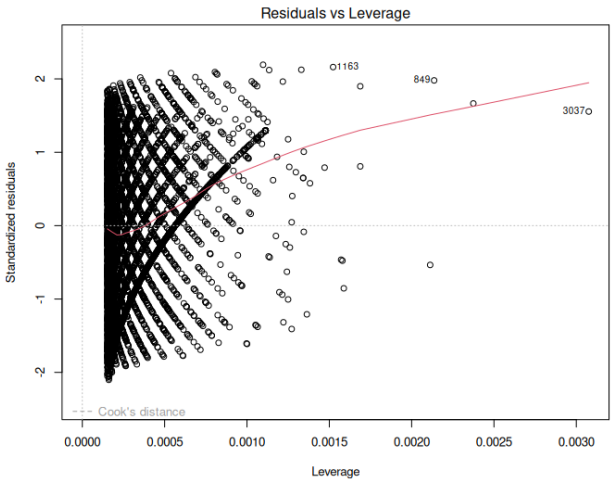
Esta ecuación permite aplicar el modelo a los valores originales de las variables sin necesidad de transformarlas previamente a unidades estándar.

3.1.2 Estudio de los criterios de aplicabilidad

Tras valorar el modelo, es necesario estudiar si se cumplen los criterios de aplicabilidad. Para ello, atenderemos a los gráficos de las siguientes figuras.







En primer lugar, el gráfico **Residuals vs. Fitted Values** muestra una tendencia en la distribución de los residuos, lo que indica que la varianza no es constante, lo que sugiere la posible presencia de relaciones no lineales entre las variables. Por otro lado, en el **Q-Q Plot** de los residuos, la mayoría de los puntos siguen la línea diagonal, pero se observan desviaciones significativas en los extremos. Esto indica que los residuos no siguen una distribución normal. De manera similar, el gráfico **Scale-Location** también muestra una tendencia en la distribución de los residuos, apoyando la hipótesis de la falta de homocedasticidad, es decir, la varianza de los errores no es constante. En cuanto al gráfico **Residuals vs. Leverage**, no se observan residuos fuera del rango [-3,3], lo que sugiere que no hay puntos atípicos con una influencia significativa en el modelo.

Finalmente, para verificar estos hallazgos, se realizaron los tests de **Ramsey RESET**, **Breusch-Pagan** y **Shapiro-Wilk**, los cuales evalúan la linealidad, homocedasticidad y normalidad de los residuos, respectivamente. Los resultados obtenidos fueron los siguientes:

Criterio	Resultados de la prueba	Conclusión
Linealidad	RESET test data: simple_model RESET = 123.86, df1 = 2, df2 = 6560, p-value < 2.2e-16	Basándonos en los resultados del test ($p_{valor} < 2.2 * 10^{-16} < \alpha = 0.05$), rechazamos la hipótesis nula de que el modelo está correctamente especificado. Esto indica que probablemente el modelo omita variables importantes o hay relaciones no lineales entre las variables.
Homocedasticidad	studentized Breusch-Pagan test data: simple_model BP = 223.16, df = 1, p-value < 2.2e-16	Basándonos en los resultados del test ($p_{valor} < 2.2 * 10^{-16} < \alpha = 0.05$), rechazamos la hipótesis nula de homocedasticidad. Esto sugiere que hay heterocedasticidad en los residuos del modelo. Es decir, la varianza de los residuos no es constante a lo largo de los valores ajustados, lo que podría indicar que el modelo no se ajusta igual de bien a todos los rangos de los datos.
Normalidad en los residuos	Shapiro-Wilk normality test data: sample(simple_model$residuals, 5000) W = 0.97077, p-value < 2.2e-16	Basándonos en los resultados del test ($p_{valor} < 2.2 * 10^{-16} < \alpha = 0.05$), rechazamos la hipótesis nula de normalidad de los residuos, lo que sugiere que los residuos no siguen una distribución normal.

3.2. REGRESIÓN MULTILINEAL

3.2.1 Creación del modelo

El objetivo de esta sección es analizar la relación entre el **IMC** (**BMI**) y las demás variables continuas del dataset, para evaluar su capacidad predictiva en la estimación del **IMC** mediante un modelo de regresión lineal múltiple.

Para la creación del modelo, se empleará una estrategia de **stepwise mixto**, comenzando con un modelo que incluya todas las variables y eliminando iterativamente aquellas menos relevantes según el criterio **AIC**, hasta obtener el mejor modelo posible.

En la salida del algoritmo (adjunta en Anexo II), se identificó que la variable **HDL** no era significativa, por lo que se eliminó del modelo final. Las estadísticas del mejor modelo de regresión múltiple obtenido se recogen en la figura de la derecha.

```
Call:
lm(formula = BMI ~ Age + Glucose + BloodPressure + WaistCircumference +
    HipCircumference, data = data_balanced)

Residuals:
     Min       1Q   Median       3Q      Max
-1.55698 -0.31931 -0.00073  0.30812  1.97581

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.181e-16  5.731e-03    0.00        1
Age               -2.351e-01  7.292e-03  -32.24   <2e-16 ***
Glucose            2.587e-01  7.954e-03   32.53   <2e-16 ***
BloodPressure      2.770e-01  7.739e-03   35.79   <2e-16 ***
WaistCircumference 4.018e-01  7.476e-03   53.75   <2e-16 ***
HipCircumference   2.542e-01  7.013e-03   36.24   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4643 on 6558 degrees of freedom
Multiple R-squared:  0.7846,    Adjusted R-squared:  0.7844
F-statistic:  4778 on 5 and 6558 DF,  p-value: < 2.2e-16
```

Atendiendo a las estadísticas del modelo, se sacan las siguientes conclusiones:

- **Significancia de las variables**: Las variables **Age**, **Glucose**, **BloodPressure**, **WaistCircumference** y **HipCircumference** tienen un $p_{valor} < 2 * 10^{-16} < \alpha = 0.05$, lo que indica que son estadísticamente significativas para explicar la variabilidad del **BMI**. El **Intercept** tiene un $p_{valor} = 1 > \alpha = 0.05$, lo que sugiere que no es significativamente diferente de cero, indicando que la función cruza el origen.
- **Dirección y magnitud de los coeficientes**: **WaistCircumference** tiene el coeficiente más alto (0.4018), siendo la variable con mayor impacto en **BMI**. Un aumento de 1 unidad en la circunferencia de la cintura se asocia con un incremento promedio de 0.40 en **BMI**. **HipCircumference** también muestra una fuerte relación con **BMI** (0.2542), lo cual es esperable debido a su relación con el peso y la composición corporal. **BloodPressure** y **Glucose** tienen coeficientes positivos (0.2770 y 0.2587, respectivamente), sugiriendo que un aumento en estos valores está asociado con un mayor **BMI**, en línea con la relación entre obesidad, hipertensión y metabolismo de la glucosa. **Age** es la única variable con un coeficiente negativo (-0.2351), indicando que, manteniendo constantes las demás variables, el **BMI** tiende a disminuir ligeramente con la edad, reflejando cambios en la composición corporal.
- **Calidad del ajuste del modelo**: El R^2 ajustado es 0.7844, lo que indica que el modelo explica el 78.44% de la variabilidad en **BMI**, sugiriendo un buen ajuste. El error estándar de los residuos es 0.4643, indicando que las predicciones se desvían en promedio 0.46 unidades de **BMI** respecto a los valores reales.
- **Significación global del modelo**: El modelo tiene un $p_{valor} < 2.2 * 10^{-16} < \alpha = 0.05$, lo que confirma es significativo, es decir, al menos una de las variables predictoras tiene un efecto real sobre **BMI**.

Por tanto, la función resultante del modelo de regresión lineal múltiple es la siguiente:

$$BMI = -1.181 \times 10 - 16 - 0.2351 \cdot Age + 0.2587 \cdot Glucose + 0.2770 \cdot BloodPressure +$$

$$0.4018 \cdot WaistCircumference + 0.2542 \cdot HipCircumference$$

Sin embargo, al igual que en el caso del modelo de regresión simple, dado que las variables han sido previamente escaladas, la ecuación obtenida representa una relación en unidades estándar. Para obtener la ecuación en valores originales (sin escalar), es necesario deshacer el escalado utilizando la media y la desviación estándar de cada variable:

$$BMI = \big[(-1.181 \times 10 - 16 - 0.2351 \times (Age - \mu_{Age}) \times \sigma_{Age} + 0.2587 \times (Glucose - \mu_{Glucose}) \times \sigma_{Glucose} + 0.2770 \times (BloodPressure - \mu_{BloodPressure}) \times \sigma_{BloodPressure} + 0.4018 \times (WaistCircumference - \mu_{WaistCircumference}) \times \sigma_{WaistCircumference} + 0.2542 \times (HipCircumference - \mu_{HipCircumference}) \times \sigma_{HipCircumference}) \times \sigma_{BMI}\big] + \mu_{BMI}$$

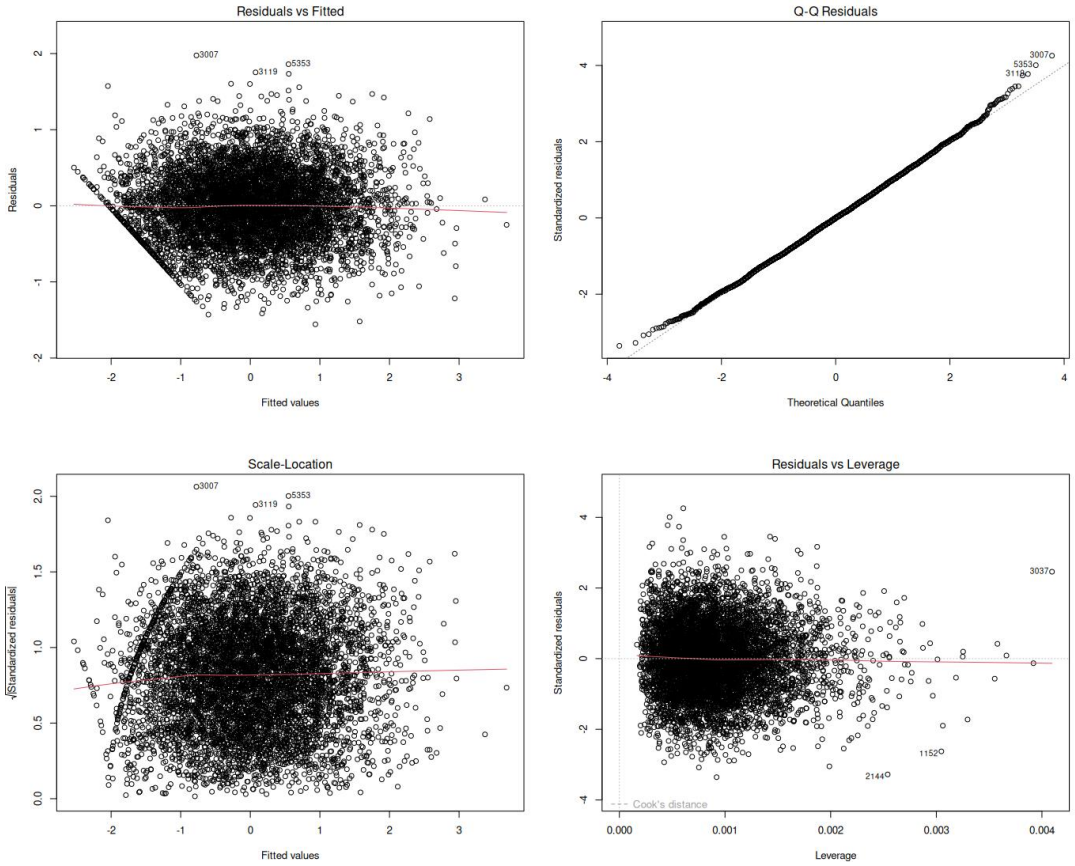
Donde:

- μ_x es la media de la variable x.
- σ_x es la desviación estándar de la variable x.

Esta ecuación permite aplicar el modelo a los valores originales de las variables sin necesidad de transformarlas previamente a unidades estándar.

3.2.2 Estudio de los criterios de aplicabilidad

Tras valorar el modelo, es necesario estudiar si se cumplen los criterios de aplicabilidad. Para ello, atenderemos a los gráficos de las siguientes figuras.



En primer lugar, el gráfico **Residuals vs. Fitted Values** muestra un patrón no completamente aleatorio en la distribución de los residuos, lo que sugiere que la varianza no es constante y podría haber relaciones no lineales entre las variables predictoras y la variable objetivo. Además, se observa una acumulación de puntos en la parte izquierda, lo que podría indicar la presencia de heterocedasticidad o una mala especificación del modelo. En el **Q-Q Plot**, la mayoría de los puntos siguen la línea diagonal, lo que sugiere que los residuos siguen aproximadamente una distribución normal. El gráfico **Scale-Location** refuerza la sospecha de heterocedasticidad, ya que los residuos no están distribuidos de manera uniforme, mostrando cierta tendencia a medida que aumentan los valores ajustados. Finalmente, en el gráfico **Residuals vs. Leverage**, no se observan puntos con una influencia excesiva en el modelo, aunque algunos valores muestran mayor leverage y podrían requerir un análisis más detallado para determinar si afectan significativamente la estimación de los coeficientes.

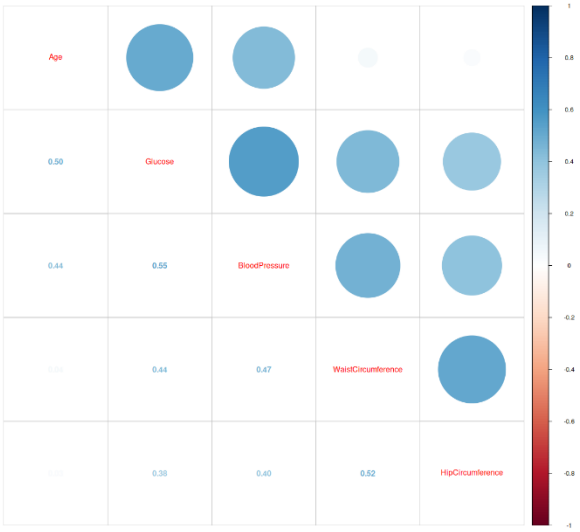
Para verificar estas hipótesis basadas en pruebas visuales, se detallan a continuación los resultados de los tests de Ramsey RESET, Breusch-Pagan y Shapiro-Wilk, los cuales evalúan la linealidad, homocedasticidad y normalidad de los residuos, respectivamente. Los resultados obtenidos fueron los siguientes:

Criterio	Resultados de la prueba	Conclusión
Linealidad	RESET test data: stepped_multiple_model RESET = 5.9517, df1 = 2, df2 = 6556, p-value = 0.002616	Basándonos en los resultados del test RESET ($p_{valor} = 0.003 < \alpha = 0.05$), rechazamos la hipótesis nula de que el modelo está correctamente especificado. Esto indica que el modelo podría estar omitiendo variables importantes o que existen relaciones no lineales entre las variables.
Homocedasticidad	studentized Breusch-Pagan test data: stepped_multiple_model BP = 17.014, df = 5, p-value = 0.004473	Basándonos en los resultados del test de Breusch-Pagan ($p_{valor} = 0.004 < \alpha = 0.05$), rechazamos la hipótesis nula de homocedasticidad. Esto indica la presencia de heterocedasticidad en los residuos del modelo, es decir, la varianza de los errores no es constante a lo largo de los valores ajustados.
Normalidad en los residuos	Shapiro-Wilk normality test data: sample(stepped_multiple_model$residuals, 5000) W = 0.99923, p-value = 0.0261	Basándonos en los resultados del test de Shapiro-Wilk aplicado a una muestra de los residuos ($p_{valor} = 0.03 < \alpha = 0.05$), rechazamos la hipótesis nula de normalidad de los residuos, lo que sugiere que los residuos no siguen una distribución normal.

Finalmente, se analiza la presencia de multicolinealidad entre las variables del modelo mediante la matriz de correlación y el cálculo del factor de inflación de varianza (VIF), cuyos resultados se presentan a continuación.

```
              Age          Glucose
         1.618961         1.926177

BloodPressure WaistCircumference    HipCircumference
      1.823584          1.701429            1.497351
```



Dado que la matriz de correlación no muestra valores extremadamente altos (<0.6) y ninguna variable presenta un VIF elevado, se concluye que no hay problemas de colinealidad entre las variables predictoras del modelo.

3.3 REGRESIÓN LOGÍSTICA

3.3.1 Creación del modelo

El objetivo de esta sección es analizar la relación entre la presencia de diabetes (**Outcome**) y las demás variables continuas del dataset, para evaluar su capacidad predictiva mediante un modelo de regresión logística.

En la exploración inicial de los datos se identificó que las variables más significativas para estimar la presencia de obesidad eran **Glucose**, **HbA1c**, **BMI**, **WaistCircumference** y **WHR**. Sin embargo, para confirmar esta hipótesis, el modelo inicial incluyó todas las variables de interés, refinándose mediante una estrategia **stepwise basada en AIC**.

En la salida del algoritmo (adjunta en Anexo III), se identificó que las variables más relevantes para el modelo eran **Glucose**, **HbA1c**, **WaistCircumference** y **WHR**, validando en gran parte la hipótesis inicial. No obstante, se observó que **HbA1c** no era significativa según su p-valor, por lo que fue descartada como variable predictora.

Finalmente, a la derecha se presentan las estadísticas del mejor modelo obtenido para estimar la presencia de obesidad en la población.

```
Call:
glm(formula = Outcome ~ Glucose + WaistCircumference + WHR, family = binomial,
    data = data_balanced)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.0005272  0.0250740   0.021  0.98323
Glucose           0.3203873  0.0289673  11.060  < 2e-16 ***
WaistCircumference  0.1008553  0.0367857   2.742  0.00611 **
WHR              -0.0700660  0.0334043  -2.098  0.03595 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9099.6  on 6563  degrees of freedom
Residual deviance: 8895.2  on 6560  degrees of freedom
AIC: 8903.2

Number of Fisher Scoring iterations: 4
```

De acuerdo con las estadísticas del modelo, se pueden extraer las siguientes conclusiones:

- **Significancia de las variables predictoras**:
  - **Glucose** tiene un $p_{valor} < 2 * 10^{-16} < \alpha = 0.05$, indicando una fuerte relación con **Outcome**, lo que sugiere que niveles más altos de glucosa están significativamente asociados con un mayor riesgo de diabetes.
  - **WaistCircumference** presenta un $p_{valor} = 0.006 < \alpha = 0.05$, confirmando su significancia.
  - **WHR** tiene un $p_{valor} = 0.04 < \alpha = 0.05$, lo que indica un efecto marginalmente significativo.
  - El **Intercept** tiene un $p_{valor} = 0.98 > \alpha = 0.05$, lo que sugiere que no es significativamente diferente de cero, indicando que la probabilidad base de diabetes cuando todas las variables son 0 no es distinta de 0.5.
- **Dirección y magnitud de los coeficientes**:
  - **Glucose** es la variable con el coeficiente más alto (0.3204), indicando que un aumento en los niveles de glucosa incrementa la probabilidad de diabetes.
  - **WaistCircumference** también tiene un efecto positivo (0.1009), aunque menor, en la probabilidad de diabetes, en línea con la relación entre grasa abdominal y diabetes tipo 2.
  - **WHR** tiene un coeficiente negativo (-0.0701), sugiriendo que, manteniendo constantes las demás variables, un aumento en la relación cintura-cadera estaría asociado con una menor probabilidad de diabetes. Este resultado es contraintuitivo y podría deberse a correlaciones con otras variables no incluidas en el modelo.

La función logística del modelo que permite calcular la probabilidad de que una persona tenga diabetes en función de las variables de **Glucose**, **WaistCircumference** y **WHR** es la que sigue a continuación. En este caso, no es necesario desescalar las variables, ya que su escala no afecta la interpretación de la probabilidad predicha.

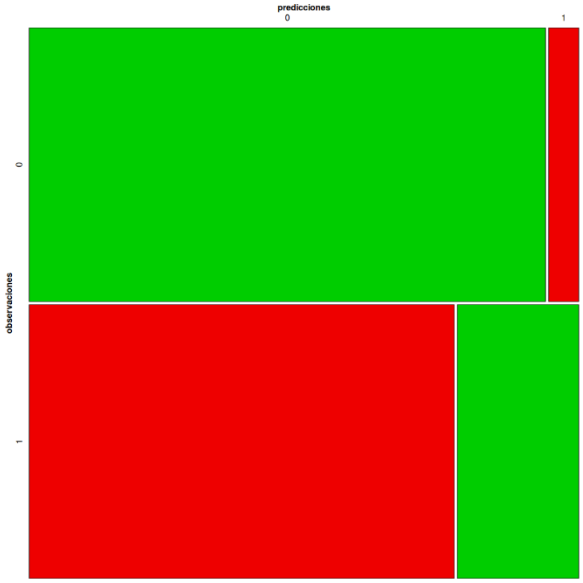
$$P(\text{Outcome} = 1) = \frac{1}{1 + e^{-(0.0005272 + 0.3204 \cdot \text{Glucose} + 0.1009 \cdot \text{WaistCircumference} - 0.0701 \cdot \text{WHR})}}$$

3.3.2 Estudio de la significancia global y evaluación del rendimiento del modelo

Para evaluar la significancia global del modelo, se analiza el p_{valor} obtenido mediante la distribución chi-cuadrado, considerando la diferencia de residuos y los grados de libertad del modelo. Este p-valor contrasta la hipótesis nula de que el modelo ajustado no mejora significativamente el modelo nulo (detalles en Anexo I).

Dado que el modelo alcanza un $p_{valor} = 4.56 * 10^{-44} < \alpha = 0.05$, se concluye que el modelo con predictores es significativamente mejor que el modelo sin ellos, lo que valida su utilidad y la relevancia de las variables seleccionadas.

Para evaluar el rendimiento del modelo, se utiliza una matriz de confusión, a partir de la cual se calcula la exactitud y se analizan los errores de clasificación (falsos positivos y falsos negativos). Se ha determinado que el threshold óptimo es 0.6 (cálculo detallado en Anexo I). El mosaico correspondiente a la matriz de confusión del modelo queda reflejado en la figura de la derecha.



Se tiene por tanto una matriz de confusión con los siguientes valores:

	Predicción: No diabético (0)	Predicción: Diabético (1)
Real: No diabético (0)	3101 (TN)	181 (FP)
Real: Diabético (1)	2553 (FN)	729 (TP)

A continuación, se calculan las métricas correspondientes para evaluar los valores de la matriz de confusión.

Métrica	Cálculo
Accuracy	$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{729 + 3101}{729 + 3101 + 181 + 2553} = \frac{3830}{6564} \approx 0.5835$
Precision	$\text{Precision} = \frac{TP}{TP + FP} = \frac{729}{729 + 181} = \frac{729}{910} \approx 0.8011$
Recall	$\text{Recall} = \frac{TP}{TP + FN} = \frac{729}{729 + 2553} = \frac{729}{3282} \approx 0.2221$
F1-Score	$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{0.8011 \times 0.2221}{0.8011 + 0.2221} = 2 \times \frac{0.1779}{1.0232} \approx 0.3478$

En función de las métricas anteriores, se observa que el modelo tiene un buen desempeño al identificar a los no diabéticos, con una alta cantidad de verdaderos negativos (3101), pero presenta problemas significativos al detectar a los diabéticos, con un bajo recall del 22.22%. Esto indica que muchos diabéticos son clasificados erróneamente como no diabéticos, lo que genera un alto número de falsos negativos. Aunque la precisión para la clase diabético es buena (80.11%), esto se debe a que cuando el modelo predice que alguien es diabético, acierta, pero lo hace pocas veces. El F1-score de 34.77% refleja un bajo desempeño en la detección de la clase minoritaria, lo que indica que el modelo está sesgado hacia predecir la clase no diabético y no es confiable para detectar diabetes.

4. CONCLUSIONES

El **modelo de Regresión Lineal Simple** muestra una relación fuerte entre **glucosa** y **HbA1c**, explicando el 69% de su variabilidad. Esto valida el uso de la **glucosa** como proxy para estimar el control glucémico en pacientes diabéticos, especialmente donde la medición de **HbA1c** no sea inmediata. Sin embargo, la heterocedasticidad y la no normalidad de los residuos indican que el modelo no captura toda la complejidad biológica. Aun así, su simplicidad lo hace útil para seguimientos generales.

El **modelo de Regresión Lineal Múltiple** explica el 78.4% de la variabilidad del **IMC**, con **circunferencia de cintura** y **cadera** como predictores claves. Esto refuerza la relación entre la **grasa abdominal** y los **riesgos metabólicos**, validando su uso en prevención de la obesidad. La relación negativa entre **edad** e **IMC** podría reflejar pérdida de **masa muscular** en adultos mayores, un aspecto no medido en el dataset. Aunque el modelo es sólido para análisis poblacionales, la **heterocedasticidad** sugiere que subestima el **IMC** en individuos con medidas extremas, lo que limita su uso en casos clínicos complejos.

El **modelo de Regresión Logística** confirma la **glucosa** y la **circunferencia de cintura** como predictores significativos de **diabetes**, alineándose con los criterios diagnósticos actuales. No obstante, su **baja sensibilidad (22.2%)** lo hace poco fiable para detección temprana, con alto riesgo de **falsos negativos**. Además, el coeficiente negativo de **WHR** contradice la literatura, posiblemente por colinealidad o sesgos en la muestra. Aunque su **precisión (80%)** es adecuada para confirmar diagnósticos, no sustituye pruebas clínicas específicas en ningún caso.

En definitiva, estos modelos son un primer paso hacia herramientas predictivas, pero su implementación clínica requerirá refinamientos que aseguren equidad, precisión y adaptación a diversidades poblacionales.

5. LIMITACIONES Y TRABAJO FUTURO

La primera limitación encontrada es que los datos del dataset usado en la actividad provienen de una **fuente única** (**Kaggle**), sin información sobre la **procedencia geográfica** o **étnica** de los **pacientes**. Esto limita la **generalización**, ya que factores **genéticos** y **culturales** influyen en la **diabetes** y la **obesidad**.

Por otro lado, los modelos creados no incluyen **indicadores de estilo de vida** (dieta detallada, actividad física, tabaquismo), esenciales para modelos predictivos precisos en enfermedades metabólicas. Esto se debe a que esta actividad se ha limitado únicamente al uso de variables cuantitativas, dejando fuera algunas variables categóricas influyentes que podrían haber mejorado el rendimiento de los modelos.

Para mejorar el modelo predictor de HbA1c (regresión lineal simple), sería interesante explorar **regresiones no lineales** o incluir **variables temporales** que aporten más contexto sobre la **variabilidad glucémica** de los **pacientes**. Esto permitiría capturar mejor la **dinámica temporal** de la **diabetes** y mejorar la precisión del modelo.

En cuanto al modelo predictor de **IMC** (regresión lineal múltiple), se podría incorporar interacciones con variables categóricas que no fueron consideradas en este análisis, como **DietType** (hábitos alimenticios), lo que permitiría evaluar su impacto en la **obesidad** con mayor precisión.

Finalmente, para el modelo detector de **pacientes diabéticos** (regresión logística), sería recomendable probar **algoritmos de clasificación** más avanzados, así como aplicar técnicas de **oversampling** (**SMOTE**) para mejorar la detección de pacientes con **diabetes**. Estas estrategias permitirían reducir el **sesgo** del modelo hacia la **clase mayoritaria** y mejorar su **capacidad de generalización** en entornos clínicos.

ANEXO I

CÓDIGO DE R EMPLEADO EN EL DESARROLLO DE LA ACTIVIDAD

```r
# Instalación de librerías
install.packages("dplyr")
install.packages("corrplot")
install.packages("ggplo2")
install.packages("tidyr")
install.packages("patchwork")
install.packages("lmtest")
install.packages("car")
install.packages("MASS")
install.packages("vcd")
install.packages("pROC")
install.packages("conflicted")

# Importación de librerías
library(corrplot)
library(ggplot2)
library(tidyr)
library(patchwork)
library(lmtest)
library(car)
library(MASS)
library(vcd)
library(pROC)
library(dplyr)
library(conflicted)
conflicts_prefer(dplyr::select)
conflicts_prefer(dplyr::filter)

# Para reproducibilidad
set.seed(123)

# Exploración de los datos
wd <- getwd();
file <- paste(wd, "diabetes_dataset.csv",sep='/')
data <- read.csv(file,sep=',',header = TRUE)
head(data)
summary(data)

# Variables de interés continuas
vars_continuas <- c("Age", "BMI", "Glucose", "BloodPressure", "HbA1c",
                    "LDL", "HDL", "Triglycerides", "WaistCircumference", "HipCircumference", "WHR")

# Selección solo de las variables de interés (Outome indica la presencia de diabetes)
data <- data %>%
 select(all_of(c(vars_continuas, "Outcome")))
summary(data)

# Balanceo de la muestra para obtener un 50% de diabéticos y un 50% de no diabéticos en el dataset
diabetes_1 <- data %>% filter(Outcome == 1)  # Pacientes con diabetes
diabetes_0 <- data %>% filter(Outcome == 0)  # Pacientes sin diabetes
diabetes_0_sampled <- diabetes_0 %>% sample_n(nrow(diabetes_1))
data_balanced <- bind_rows(diabetes_1, diabetes_0_sampled)

# Gráficos de los histogramas y diagrama de barras de las variables de interés
histogramas <- lapply(vars_continuas, function(var) {
 ggplot(data_balanced, aes(x = .data[[var]])) +
   geom_histogram(bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +
   theme_minimal() +
   labs(title = var, x = var, y = "Frecuencia")
})
barplot_outcome <- ggplot(data_balanced, aes(x = factor(Outcome))) +
 geom_bar(fill = "tomato", color = "black", alpha = 0.7) +
 theme_minimal() +
 labs(title = "Distribución de Outcome", x = "Outcome (0 = No Diabetes, 1 = Diabetes)", y = "Frecuencia")

todos_graficos <- c(histogramas, list(barplot_outcome))
wrap_plots(todos_graficos, ncol = 3)

# Gráficos de cajas y bigotes para las variables continuas de interés
boxplots <- lapply(vars_continuas, function(var) {
```

```
 ggplot(data, aes_string(y = var)) +
 geom_boxplot(fill = "lightblue", color = "black") +
 ggtitle(paste("Boxplot de", var)) +
 theme_minimal()
})
wrap_plots(boxplots, ncol = 4)

# Correlación entre las variables continuas de interés
M = cor(data_balanced[vars_continuas])
corrplot.mixed(M,order='AOE')

# Escalado de las variables
data_balanced[vars_continuas] <- scale(data_balanced[vars_continuas])

# Gráfico de HbA1c vs Glucose
plot(HbA1c~Glucose, data=data_balanced,xlab='Glucose (scaled)',ylab='HbA1c (scaled)', main='HbA1c vs Glucose',
   pch=10,cex=2,col='blue')

# Creación del modelo de regresión lineal simple
simple_model <- lm(HbA1c ~ Glucose, data = data_balanced)
summary(simple_model)

# Gráficos para comprobar los criterios de aplicabilidad del modelo de regresión lineal simple
par(mfrow=c(2, 2))
plot(simple_model, which=1)
plot(simple_model, which=2)
plot(simple_model, which=3)
plot(simple_model, which=5)
par(mfrow=c(1, 1))

# Tests para comprobar los criterios de aplicabilidad del modelo de regresión lineal simple
bptest(simple_model)
resettest(simple_model)
shapiro.test(sample(simple_model$residuals,5000))

# Creación del modelo de regresión lineal múltiple
multiple_model <- lm(BMI ~
Age+HbA1c+Glucose+BloodPressure+LDL+HDL+Triglycerides+WaistCircumference+HipCircumference+WHR,
               data=data_balanced)
summary(multiple_model)

# Iteración para encontrar el mejor modelo posible de regresión lineal múltiple
step(object = multiple_model, direction = "both", trace = 1)

# Creación del mejor modelo posible de regresión lineal múltiple (elimando HDL del modelo devuelto po el algoritmo anterior)
stepped_multiple_model <- lm(formula = BMI ~ Age + Glucose + BloodPressure + WaistCircumference +
   HipCircumference, data = data_balanced)
summary(stepped_multiple_model)

# Gráficos para comprobar los criterios de aplicabilidad del modelo de regresión lineal múltiple
par(mfrow=c(2, 2))
plot(stepped_multiple_model, which=1)
plot(stepped_multiple_model, which=2)
plot(stepped_multiple_model, which=3)
plot(stepped_multiple_model, which=5)
par(mfrow=c(1, 1))

# Tests para comprobar los criterios de aplicabilidad del modelo de regresión lineal múltiple
bptest(stepped_multiple_model)
resettest(stepped_multiple_model)
shapiro.test(sample(stepped_multiple_model$residuals,5000))

# Correlación entre variables predictoras del modelo de regresión múltiple
M_multiple <- cor(data_balanced[, c("Age", "Glucose", "BloodPressure", "WaistCircumference", "HipCircumference")])
corrplot.mixed(M_multiple,order='AOE')

# Análisis de inflación de varianza del modelo de regresión múltiple
vif(stepped_multiple_model)

# Gráficas de boxplots para ver la distribución de datos de cada variable en cada uno de los grupos (diabéticos o no)
```

```r
boxplots <- lapply(vars_continuas, function(var) {
 ggplot(data_balanced, aes(x = factor(Outcome), y = .data[[var]], fill = factor(Outcome))) +
   geom_boxplot(outlier.colour = "red", outlier.shape = 16, outlier.size = 2) +
   scale_fill_manual(values = c("blue", "orange"), labels = c("No Diabetes", "Diabetes")) +
   labs(x = "Diabetes", y = var, title = paste("Distribución de", var)) +
   theme_minimal() +
   theme(legend.position = "none")
})
wrap_plots(boxplots, ncol = 3)

# Creación del modelo logístico inicial
logistic_model <- glm(Outcome ~ ., data = data_balanced, family = binomial)
summary(logistic_model)

# Iteración para encontrar el mejor modelo posible de regresión logística
stepAIC(object = logistic_model, direction = "both", trace = 1)

# Creación del mejor modelo de regresión logística (eliminando HbA1c)
optimal_logistic_model <- glm(formula = Outcome ~ Glucose + WaistCircumference + WHR, family = binomial, data =
data_balanced)
summary(optimal_logistic_model)

# Evalación del p-valor de la distribución chi-cuadrado de la diferencia de residuos del modelo y el número de grados de libertad
dif_residuos <- optimal_logistic_model$null.deviance - optimal_logistic_model$deviance
df <- optimal_logistic_model$df.null - optimal_logistic_model$df.residual
optimal_logistic_model_p_value <- pchisq(q = dif_residuos, df = df, lower.tail = FALSE)
print(optimal_logistic_model_p_value)

# Evaluación del rendimiento del modelo de regresión logística
umbrales <- seq(0, 1, by = 0.05)
resultados <- sapply(umbrales, function(threshold) {
 predicciones <- ifelse(optimal_logistic_model$fitted.values > threshold, 1, 0)
 cm <- table(optimal_logistic_model$model$Outcome, predicciones)
 accuracy <- sum(diag(cm)) / sum(cm)
 return(accuracy)
})
umbral_optimo <- umbrales[which.max(resultados)]

predicciones <- ifelse(test = optimal_logistic_model$fitted.values > umbral_optimo, yes = 1, no = 0)
matriz_confusion <- table(optimal_logistic_model$model$Outcome, predicciones, dnn = c("observaciones", "predicciones"))
matriz_confusion
mosaic(matriz_confusion, shade = T, colorize = T, gp = gpar(fill = matrix(c("green3", "red2", "red2", "green3"), 2, 2)))
```

ANEXO II

SALIDA DE R-STUDIO DEL ALGORITMO STEPWISE PARA EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

```
> step(object = multiple_model, direction = "both", trace = 1)
Start:  AIC=-10062.43
BMI ~ Age + HbA1c + Glucose + BloodPressure + LDL + HDL + Triglycerides +
    WaistCircumference + HipCircumference + WHR

                  Df Sum of Sq    RSS      AIC
- LDL                1     0.029 1412.4 -10064.3
- WHR                1     0.062 1412.4 -10064.1
- Triglycerides      1     0.335 1412.7 -10062.9
- HbA1c              1     0.343 1412.7 -10062.8
<none>                           1412.4 -10062.4
- HDL                1     0.486 1412.9 -10062.2
- HipCircumference   1     9.259 1421.6 -10021.5
- WaistCircumference 1    19.500 1431.9  -9974.4
- Glucose            1    96.645 1509.0  -9630.0
- Age                1   223.055 1635.4  -9101.9
- BloodPressure      1   275.231 1687.6  -8895.8

Step:  AIC=-10064.3
BMI ~ Age + HbA1c + Glucose + BloodPressure + HDL + Triglycerides +
    WaistCircumference + HipCircumference + WHR

                  Df Sum of Sq    RSS      AIC
- WHR                1     0.064 1412.5 -10066.0
- Triglycerides      1     0.338 1412.8 -10064.7
- HbA1c              1     0.342 1412.8 -10064.7
<none>                           1412.4 -10064.3
- HDL                1     0.482 1412.9 -10064.1
+ LDL                1     0.029 1412.4 -10062.4
- HipCircumference   1     9.245 1421.7 -10023.5
- WaistCircumference 1    19.524 1431.9  -9976.2
- Glucose            1    96.617 1509.0  -9632.0
- Age                1   223.046 1635.5  -9103.9
- BloodPressure      1   275.513 1687.9  -8896.6

Step:  AIC=-10066
BMI ~ Age + HbA1c + Glucose + BloodPressure + HDL + Triglycerides +
    WaistCircumference + HipCircumference

                  Df Sum of Sq    RSS      AIC
- Triglycerides      1      0.34 1412.8 -10066.4
- HbA1c              1      0.35 1412.8 -10066.4
<none>                           1412.5 -10066.0
- HDL                1      0.49 1413.0 -10065.7
+ WHR                1      0.06 1412.4 -10064.3
+ LDL                1      0.03 1412.4 -10064.1
- Glucose            1     96.61 1509.1  -9633.7
- Age                1    223.22 1635.7  -9104.9
- BloodPressure      1    275.49 1688.0  -8898.4
- HipCircumference   1    283.14 1695.6  -8868.8
- WaistCircumference 1    622.67 2035.1  -7670.7

Step:  AIC=-10066.43
BMI ~ Age + HbA1c + Glucose + BloodPressure + HDL + WaistCircumference +
    HipCircumference

                  Df Sum of Sq    RSS      AIC
- HbA1c              1      0.36 1413.2 -10066.8
<none>                           1412.8 -10066.4
- HDL                1      0.47 1413.3 -10066.2
+ Triglycerides      1      0.34 1412.5 -10066.0
+ WHR                1      0.06 1412.8 -10064.7
+ LDL                1      0.03 1412.8 -10064.6
- Glucose            1     96.55 1509.4  -9634.5
- Age                1    223.63 1636.4  -9103.9
- BloodPressure      1    275.90 1688.7  -8897.5
- HipCircumference   1    282.88 1695.7  -8870.5
- WaistCircumference 1    622.72 2035.5  -7671.4

Step:  AIC=-10066.77
```

```
BMI ~ Age + Glucose + BloodPressure + HDL + WaistCircumference +
    HipCircumference

                  Df Sum of Sq    RSS      AIC
<none>                         1413.2 -10066.8
- HDL                1      0.48 1413.7 -10066.5
+ HbA1c              1      0.36 1412.8 -10066.4
+ Triglycerides      1      0.35 1412.8 -10066.4
+ WHR                1      0.07 1413.1 -10065.1
+ LDL                1      0.03 1413.1 -10064.9
- Age                1    223.82 1637.0  -9103.7
- Glucose            1    227.70 1640.9  -9088.2
- BloodPressure      1    276.01 1689.2  -8897.7
- HipCircumference   1    283.04 1696.2  -8870.4
- WaistCircumference 1    622.52 2035.7  -7672.9

Call:
lm(formula = BMI ~ Age + Glucose + BloodPressure + HDL + WaistCircumference +
    HipCircumference, data = data_balanced)

Coefficients:
      (Intercept)                Age            Glucose       BloodPressure                HDL  WaistCircumference
      -1.196e-16         -2.350e-01          2.585e-01          2.769e-01         -8.559e-03          4.017e-01
 HipCircumference
       2.541e-01

> stepped_multiple_model <- lm(formula = BMI ~ Age + Glucose + BloodPressure + HDL + WaistCircumference +
+                                HipCircumference, data = data_balanced)
> summary(stepped_multiple_model)

Call:
lm(formula = BMI ~ Age + Glucose + BloodPressure + HDL + WaistCircumference +
    HipCircumference, data = data_balanced)

Residuals:
    Min      1Q  Median      3Q     Max
-1.56025 -0.31989 -0.00012  0.30896  1.99054

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -1.196e-16  5.730e-03   0.000    1.000
Age               -2.350e-01  7.292e-03 -32.226   <2e-16 ***
Glucose            2.585e-01  7.954e-03  32.504   <2e-16 ***
BloodPressure      2.769e-01  7.739e-03  35.786   <2e-16 ***
HDL               -8.559e-03  5.733e-03  -1.493    0.135
WaistCircumference 4.017e-01  7.475e-03  53.744   <2e-16 ***
HipCircumference   2.541e-01  7.012e-03  36.239   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4642 on 6557 degrees of freedom
Multiple R-squared:  0.7847,    Adjusted R-squared:  0.7845
F-statistic:  3982 on 6 and 6557 DF,  p-value: < 2.2e-16
```

ANEXO III

SALIDA DE R-STUDIO DEL ALGORITMO STEPWISE PARA EL MODELO DE REGRESIÓN LOGÍSTICA

```
> stepAIC(object = logistic_model, direction = "both", trace = 1)
Start:  AIC=8912.61
Outcome ~ Age + BMI + Glucose + BloodPressure + HbA1c + LDL +
    HDL + Triglycerides + WaistCircumference + HipCircumference +
    WHR

                   Df Deviance    AIC
- HipCircumference    1   8888.6 8910.6
- HDL                 1   8888.6 8910.6
- WHR                 1   8888.8 8910.8
- Triglycerides       1   8888.9 8910.9
- WaistCircumference  1   8889.0 8911.0
- BloodPressure       1   8889.5 8911.5
- BMI                 1   8889.9 8911.9
- LDL                 1   8890.1 8912.1
<none>                    8888.6 8912.6
- HbA1c               1   8890.8 8912.8
- Age                 1   8890.9 8912.9
- Glucose             1   8920.1 8942.1

Step:  AIC=8910.61
Outcome ~ Age + BMI + Glucose + BloodPressure + HbA1c + LDL +
    HDL + Triglycerides + WaistCircumference + WHR

                   Df Deviance    AIC
- HDL                 1   8888.6 8908.6
- Triglycerides       1   8888.9 8908.9
- BloodPressure       1   8889.5 8909.5
- BMI                 1   8889.9 8909.9
- LDL                 1   8890.1 8910.1
<none>                    8888.6 8910.6
- HbA1c               1   8890.8 8910.8
- Age                 1   8890.9 8910.9
+ HipCircumference    1   8888.6 8912.6
- WHR                 1   8893.1 8913.1
- WaistCircumference  1   8893.5 8913.5
- Glucose             1   8920.1 8940.1

Step:  AIC=8908.62
Outcome ~ Age + BMI + Glucose + BloodPressure + HbA1c + LDL +
    Triglycerides + WaistCircumference + WHR

                   Df Deviance    AIC
- Triglycerides       1   8888.9 8906.9
- BloodPressure       1   8889.5 8907.5
- BMI                 1   8889.9 8907.9
- LDL                 1   8890.1 8908.1
<none>                    8888.6 8908.6
- HbA1c               1   8890.8 8908.8
- Age                 1   8890.9 8908.9
+ HDL                 1   8888.6 8910.6
+ HipCircumference    1   8888.6 8910.6
- WHR                 1   8893.1 8911.1
- WaistCircumference  1   8893.5 8911.5
- Glucose             1   8920.1 8938.1

Step:  AIC=8906.93
Outcome ~ Age + BMI + Glucose + BloodPressure + HbA1c + LDL +
    WaistCircumference + WHR

                   Df Deviance    AIC
- BloodPressure       1   8889.8 8905.8
- BMI                 1   8890.2 8906.2
- LDL                 1   8890.4 8906.4
<none>                    8888.9 8906.9
- HbA1c               1   8891.1 8907.1
- Age                 1   8891.2 8907.2
+ Triglycerides       1   8888.6 8908.6
+ HDL                 1   8888.9 8908.9
+ HipCircumference    1   8888.9 8908.9
```

```
- WHR                1   8893.4 8909.4
- WaistCircumference  1   8893.9 8909.9
- Glucose            1   8920.5 8936.5

Step:  AIC=8905.77
Outcome ~ Age + BMI + Glucose + HbA1c + LDL + WaistCircumference +
    WHR

                Df Deviance    AIC
- BMI                1   8890.5 8904.5
- LDL                1   8891.2 8905.2
- Age                1   8891.2 8905.2
<none>                   8889.8 8905.8
- HbA1c              1   8891.9 8905.9
+ BloodPressure      1   8888.9 8906.9
+ Triglycerides      1   8889.5 8907.5
+ HDL                1   8889.8 8907.8
+ HipCircumference   1   8889.8 8907.8
- WHR                1   8894.3 8908.3
- WaistCircumference  1   8894.7 8908.7
- Glucose            1   8921.5 8935.5

Step:  AIC=8904.49
Outcome ~ Age + Glucose + HbA1c + LDL + WaistCircumference +
    WHR

                Df Deviance    AIC
- Age                1   8891.6 8903.6
- LDL                1   8891.9 8903.9
<none>                   8890.5 8904.5
- HbA1c              1   8892.6 8904.6
+ BMI                1   8889.8 8905.8
+ Triglycerides      1   8890.2 8906.2
+ BloodPressure      1   8890.2 8906.2
- WHR                1   8894.3 8906.3
+ HDL                1   8890.5 8906.5
+ HipCircumference   1   8890.5 8906.5
- WaistCircumference  1   8896.1 8908.1
- Glucose            1   8922.1 8934.1

Step:  AIC=8903.58
Outcome ~ Glucose + HbA1c + LDL + WaistCircumference + WHR

                Df Deviance    AIC
- LDL                1   8893.0 8903.0
<none>                   8891.6 8903.6
- HbA1c              1   8893.7 8903.7
+ Age                1   8890.5 8904.5
+ BMI                1   8891.2 8905.2
+ Triglycerides      1   8891.3 8905.3
+ BloodPressure      1   8891.6 8905.6
+ HDL                1   8891.6 8905.6
+ HipCircumference   1   8891.6 8905.6
- WHR                1   8895.9 8905.9
- WaistCircumference  1   8899.0 8909.0
- Glucose            1   8923.4 8933.4

Step:  AIC=8903.04
Outcome ~ Glucose + HbA1c + WaistCircumference + WHR

                Df Deviance    AIC
<none>                   8893.0 8903.0
- HbA1c              1   8895.2 8903.2
+ LDL                1   8891.6 8903.6
+ Age                1   8891.9 8903.9
+ BMI                1   8892.7 8904.7
+ Triglycerides      1   8892.7 8904.7
+ BloodPressure      1   8893.0 8905.0
+ HDL                1   8893.0 8905.0
+ HipCircumference   1   8893.0 8905.0
```

```
- WHR                1   8897.4 8905.4
- WaistCircumference  1   8900.5 8908.5
- Glucose            1   8925.0 8933.0

Call:  glm(formula = Outcome ~ Glucose + HbA1c + WaistCircumference +
    WHR, family = binomial, data = data_balanced)

Coefficients:
      (Intercept)            Glucose             HbA1c  WaistCircumference               WHR
        0.0007968          0.2659595          0.0656147          0.1005363         -0.0695379

Degrees of Freedom: 6563 Total (i.e. Null);  6559 Residual
Null Deviance:      9100
Residual Deviance: 8893     AIC: 8903
```