UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA INFORMÁTICA

Laboratorio 1

Análisis Estadístico e Inferencial

Integrantes: Diego Valdés Fernández

Valentina Campos Olguín

Curso: Análisis de Datos

Profesor: Dr. Max Chacón

Ayudante: Marcelo Álvarez

21 de octubre de 2024

Tabla de contenidos

1. Introducción

El presente estudio se centra en el análisis y la predicción de la diabetes mellitus tipo 2 en una población específica, a partir de mediciones diagnósticas. Este conjunto de datos proviene del *National Institute of Diabetes and Digestive and Kidney Diseases* y contiene información de mujeres mayores de 21 años de ascendencia Pima, una población indígena estadounidense conocida por tener una alta prevalencia de diabetes. El objetivo principal de este trabajo es desarrollar un modelo de predicción que permita diagnosticar la presencia o ausencia de diabetes tipo 2, utilizando variables médicas relevantes como el número de embarazos, índice de masa corporal (BMI), niveles de insulina, presión arterial, grosor del pliegue cutáneo, glucosa en plasma, función de pedigree de diabetes y la edad. Mediante estas variables, se espera construir un modelo predictivo robusto que apoye el diagnóstico temprano de la enfermedad (Kaggle, 2016).

1.1. Objetivos Específicos

- Realizar un análisis exploratorio de los datos para identificar patrones, distribuciones y correlaciones entre las variables.
- Aplicar técnicas de análisis estadístico descriptivo e inferencial para evaluar la significancia de cada variable en relación con el diagnóstico de diabetes.
- Construir y evaluar un modelo de regresión logística que permita predecir la diabetes en función de las variables disponibles.
- Interpretar los resultados del análisis y compararlos con la literatura existente, proporcionando recomendaciones útiles para el diagnóstico temprano de la enfermedad.

La metodología adoptada incluye la exploración detallada del conjunto de datos, la aplicación de pruebas estadísticas y la construcción de modelos de clasificación, utilizando herramientas como *R* para realizar las pruebas necesarias.

2. Descripción del problema

Se obtuvo el conjunto de datos del *National Institute of Diabetes and Digestive and Kidney Diseases*, y el principal objetivo es ayudar a predecir de manera diagnóstica si una paciente tiene o no diabetes mellitus tipo 2. Este conjunto de datos se ha hecho solo para mujeres de al menos 21 años de edad, con ascendencia Pima.

El conjunto de datos tiene de varios atributos que incluyen mediciones médicas importantes para el diagnóstico de diabetes. La variable objetivo o *outcome* indica si la paciente ha sido diagnosticada con diabetes o no, donde un valor de 1 representa un diagnóstico positivo de diabetes y un valor de 0 indica un diagnóstico negativo.

El problema de estudio se enfoca en entender cómo estas variables predictoras se relacionan entre sí y cómo se relaciona con la variable objetivo, la cual nos indica la presencia o ausencia de diabetes. Tras hacer un análisis de este conjunto de datos, se pretende descubrir cuáles son los factores más determinantes en el diagnóstico de diabetes y cómo las variables pueden influir en dicho diagnóstico.

Este estudio tiene como objetivo desarrollar un modelo estadístico que permita combinar toda la información de las variables para proporcionar una predicción precisa del diagnóstico de diabetes. Para esto, se utilizarán técnicas de análisis estadístico e inferencial que facilitarán una mejor comprensión de las relaciones entre las diferentes variables en el resultado final.

Antes de proceder con el análisis de las variables, es importante comprender sus características, sus métricas, sus valores y el cómo se relacionan con el diagnóstico de diabetes, tal cómo se detalla en la Tabla 1 y 2.

Atributo	Descripción	Relación con diabetes
Pregnancies	Número de embarazos.	Cambios hormonales pueden aumentar el riesgo de desarrollar diabetes.
Glucose	Concentración de glucosa en plasma tras 2 horas del test de tolerancia a la glucosa.	Un nivel elevado es un fuerte indicador de diabetes.
Blood Pressure	Presión arterial diastólica (mm Hg).	La hipertensión está asociada con la diabetes.
Skin Thickness	Grosor del pliegue cutáneo en milímetros.	Relacionado con la resistencia a la insulina.
Insulin	Niveles de insulina en suero (mu U/ml).	Niveles anormales indican problemas en la regulación de glucosa.
BMI	Índice de masa corporal.	Un BMI elevado es un fuerte indicador de obesidad y diabetes tipo 2.
Diabetes Pedigree Function	Estimación de predisposición genética a la diabetes.	Un valor alto indica mayor probabilidad de desarrollar diabetes.
Age	Edad de la mujer.	El riesgo de desarrollar diabetes aumenta con la edad.

Tabla 1: Atributos, descripción y relación con la diabetes

Atributo	Tipo de valor	Métrica
Pregnancies	Valor numérico discreto	Número de embarazos
Glucose	Valor numérico discreto	Glucosa en plasma (mg/dl)
Blood Pressure	Valor numérico discreto	Presión arterial diastólica (mm Hg)
Skin Thickness	Valor numérico continuo	Grosor del pliegue cutáneo (mm)
Insulin	Valor numérico continuo	Insulina en suero (mu U/ml)
BMI	Valor numérico continuo	Índice de masa corporal
Diabetes Pedigree Function	Valor numérico continuo	Función de Pedigee de Diabetes
Age	Valor numérico discreto	Edad (años)
Outcome	Valor categórico binario	Diagnóstico de diabetes (1 = Sí, 0 = No)

Tabla 2: Tipos de valores y métricas de los atributos

3. Análisis Estadístico e Inferencial

3.1. Análisis Descriptivo

Inicialmente, se realizó un análisis descriptivo de las variables del conjunto de datos. Sin embargo, se identificó la presencia de valores inválidos, específicamente valores de 0 en variables donde esto no es posible desde un punto de vista médico. Por ejemplo, no es posible tener un valor de 0 en variables como la glucosa, presión arterial, grosor del pliegue cutáneo, insulina, y el índice de masa corporal (BMI). Estas variables reflejan mediciones esenciales en la salud de los pacientes y, en condiciones normales, deberían tener valores positivos mayores a cero. Analizando el archivo, se tiene un 48.7 % de valores 0 en la variable Insulin, 29.56 % en SkinThickness y un 4.56 % en BloodPressure. Por lo tanto, se procedió a una limpieza de los datos, corrigiendo estos valores inválidos como se muestra en la Tabla 3, donde se reemplazaron los valores inválidos por la media de la columna correspondiente. Esta estrategia de imputación de valores faltantes es adecuada cuando los datos se consideran perdidos, como se discute en Little y Rubin, 2002, ya que preserva el tamaño de la muestra y evita sesgos en el análisis estadístico.

Atributo	Min.	Q1	Media	Q3	Varianza	Max.
Pregnancies	0.000	1.000	3.845	6.000	11.354	17.000
Glucose	44.00	99.75	121.69	140.25	926.347	199.00
Blood Pressure	24.00	64.00	72.41	80.00	146.3216	122.00
Skin Thickness	7.00	25.00	29.15	32.00	77.28066	99.00
Insulin	14.00	121.50	155.50	155.50	7228.589	846.00
BMI	18.20	27.50	32.46	36.60	47.26771	67.10
Diabetes Pedigree Function	0.0780	0.2437	0.4719	0.6262	0.1097786	2.4200
Age	21.00	24.00	33.24	41.00	138.303	81.00

Tabla 3: Medidas descriptivas tras la limpieza de datos

3.2. Matriz de Correlación

Se presenta la matriz de correlación entre las variables numéricas del conjunto de datos limpios. Este análisis nos permite identificar la relación entre las variables mediante el coeficiente de correlación de Pearson. Los valores cercanos a 1 indican una correlación positiva fuerte, mientras que los valores cercanos a -1 indican una correlación negativa fuerte. Un valor cercano a 0 indica que no existe una correlación lineal significativa entre las variables.

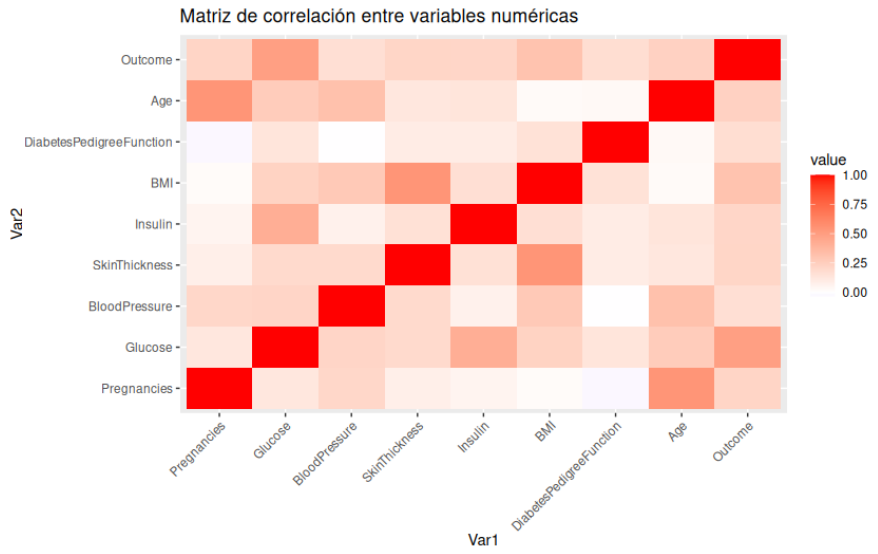


Figura 1: Matriz de correlación entre variables numéricas con datos limpios

3.2.1. Interpretación de la Matriz de Correlación

- Existe una fuerte correlación positiva de la glucosa y el BMI con el resultado de diabetes. Mientras que BloodPressure y Skin Thickness hay una débil correlación.

3.3. Visualización de la Distribución de las Variables

A continuación, se presentan los histogramas junto con las curvas de densidad para visualizar la distribución de las variables más relevantes del conjunto de datos. Los histogramas detallados se encuentran en el Anexo.

3.3.1. Distribución de Glucosa (a)

La glucosa tiene una distribución que parece cercana a una distribución normal, con una ligera asimetría positiva. La mayoría de los datos están concentrados entre 80 y 150 mg/dl. Esto podría indicar que la mayoría de las pacientes tienen niveles de glucosa dentro de un rango saludable, pero algunas presentan niveles significativamente más altos, lo que puede estar relacionado con la presencia de diabetes.

3.3.2. Distribución de la Presión Sanguínea (b)

La distribución de la presión sanguínea también parece aproximadamente normal, aunque algo sesgada hacia la izquierda. La mayoría de los datos se agrupan alrededor de 60 a 80 mm Hg. Esto podría sugerir que algunaas pacientes tienen presiones sanguíneas más bajas de lo esperado, posiblemente indicando alguna condición existente en relación a la circulación o presión arterial.

3.3.3. Distribución del Grosor de la Piel (c)

Esta distribución tiene una asimetría muy pronunciada hacia la derecha. La mayoría de valores están concentrados alrededor de lo 20 a 30 mm, pero hay casos donde el grosor de la piel es considerablemente mayor Esto puede deberse a que el grosor del pliegue cutáneo no sigue una distribución normal, y puede estar afectado por la obesidad o resistencia a la insulina.

3.3.4. Distribución de los Niveles de Insulina (d)

La distribución de insulina presenta una fuerte asimetría a la derecha, con un pico muy pronunciado en los niveles bajos de insulina. La mayoría de las pacientes tienen niveles de insulina alrededor de 0 a 100, pero existen valores muy altos en algunos casos. Esta distribución sugiere que muchas pacientes tienen niveles bajos de insulina, mientras que unas pocas tienen niveles extremadamente altos, lo cual puede estar relacionado con la resistencia a la insulina o la secreción anormal de insulina en pacientes con diabetes.

3.3.5. Distribución del Índice de Masa Corporal (BMI) (e)

El índice de masa corporal tiene una distribución relativamente normal, aunque ligeramente asimétrica hacia la derecha. Los valores están principalmente entre 20 y 40, con algunos casos de BMI muy altos. Esto indica que la mayoría de las pacientes están en rango saludable o con sobrepeso moderado, pero algunas presentan obesidad, lo que es un factor de riesgo importante para la diabetes.

3.3.6. Distribución de la Edad (f)

La edad de las pacientes presenta una distribución claramente asimétrica hacia la derecha. La mayoría de las pacientes están entre los 20 y 30 años, con menos pacientes a medida que la edad aumenta. Esto indica que la muestra tiene un sesgo hacia mujeres más jóvenes, aunque incluye a pacientes mayores de 60 años.

3.3.7. Distribución de la Cantidad de Embarazos (g)

La cantidad de embarazos presenta una distribución con una fuerte asimetría a la derecha, lo que indica que la mayoría de las mujeres en el conjunto de datos tienen pocos embarazos (0 a 2), mientras que hay algunas con un mayor número de embarazos.

3.3.8. Distribución de Predisposición Genética (h)

Esta variable también presenta una asimetría hacia la derecha. La mayoría de los valores de predisposición genética se concentran en valores bajos (cerca de 0), lo que indica que la mayoría de las mujeres en este conjunto de datos tienen una predisposición baja a la diabetes. Esta asimetría puede indicar que, aunque la predisposición genética a la diabetes está presente en algunas personas, no es extremadamente común tener una predisposición genética alta.

3.4. Conclusiones generales

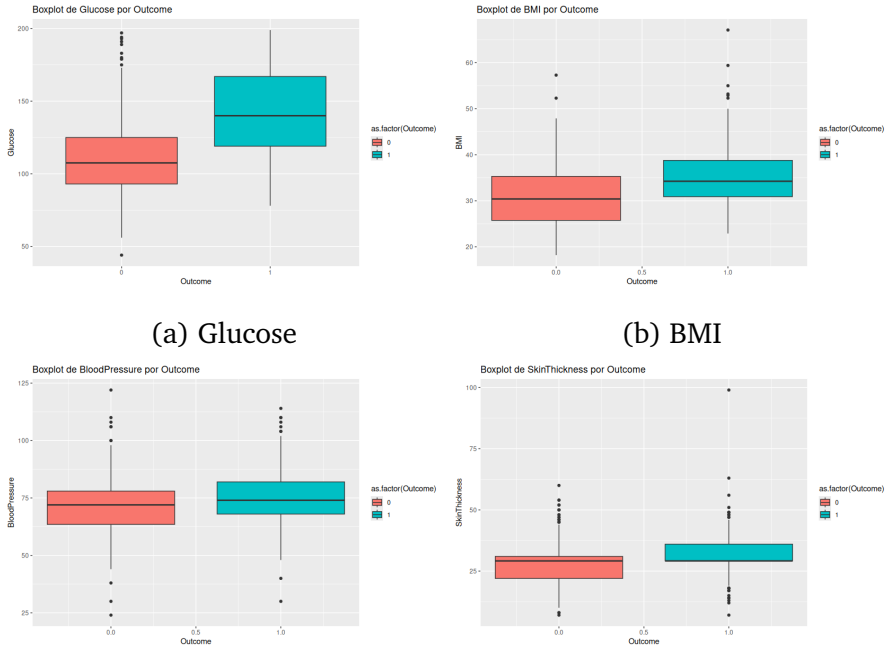
La asimetría de algunas distribuciones, como las de insulina y grosor de piel, y la normalidad de otras como las de glucosa y BMI, pueden indicar la necesidad de tratar ciertos datos con transformaciones antes de aplicar modelos estadísticos. Las variables como cantidad de embarazos y predisposición genética muestran distribuciones sesgadas a la derecha, lo que sugiere que a pesar de que la mayoría de las pacientes tienen pocos embarazos o baja predisposición genética, existen algunos casos que presentan valores significativamente mayores, lo que podría influir en los resultados del análisis.

La variabilidad de los niveles de insulina, grosor de piel y predisposición genética sugiere que estos podrían ser factores importantes para dividir a los pacientes con y sin diabetes, especialmente cuando se comparan con otras variables como la cantidad de embarazos.

Finalmente, las distribuciones de glucosa y BMI, al ser más cercanas a una distribución normal, podrían ser variables predictoras clave en los modelos para predecir la diabetes.

3.5. Boxplots de Variables por Outcome

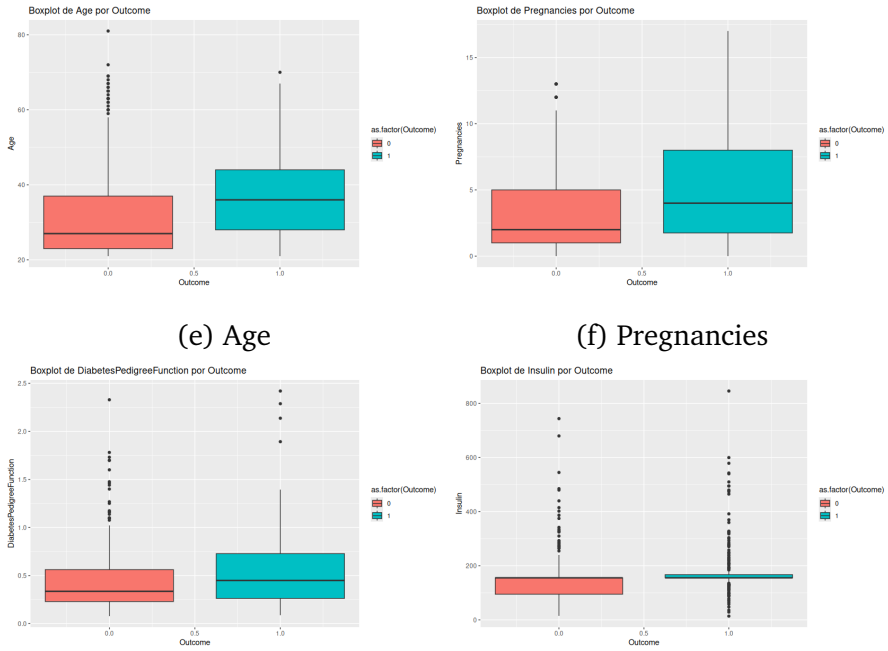
En esta sección se muestran los boxplots de diferentes variables divididos por la variable de resultado (*Outcome*), lo cual permite visualizar cómo varían las distribuciones de las variables en función de si los pacientes fueron diagnosticados con diabetes (Outcome = 1) o no (Outcome = 0).



(a) Glucose (b) BMI

(c) BloodPressure (d) SkinThickness

Figura 2: Boxplots de diferentes variables en función del Outcome (Primera parte).



(e) Age (f) Pregnancies

(g) DiabetesPedigreeFunction (h) Insulin

Figura 2: Boxplots de diferentes variables en función del Outcome (Continuación).

3.5.1. Glucosa (a)

Las personas con un outcome igual a 1 tienden a tener niveles de glucosa más altos que aquellas con outcome igual a 0. El valor mediano de la glucosa es considerablemente mayor en las personas diagnosticadas con diabetes, lo que tiene sentido con lo que se espera de una condición relacionada con la regulación deficiente de los niveles de glucosa en la sangre.

3.5.2. Índice de masa corporal (b)

El grosor de la piel muestra una mediana similar entre ambos grupos, a pesar de que el rango intercuartil es mayor en el grupo de personas diagnosticadas con diabetes. A pesar de no haber gran diferencia en la mediana, las personas con outcome igual a 1 tienden a tener más variabilidad en el grosor de la piel, lo cual podría estar asociado a una mayor presencia de tejido adiposo.

3.5.3. Presión arterial (c)

Las personas de ambos grupos tienden a tener una presión arterial bastante similar, con medianas y rangos intercuartiles similares. Esto quiere decir que no parece haber una gran diferencia en los niveles de presión arterial entre las personas con y sin diabetes.

3.5.4. Grosor de piel (d)

El grosor de la piel muestra una mediana similar entre ambos grupos, aunque el rango intercuartil es mayor en el grupo de las personas diagnosticadas con diabetes.

3.5.5. Edad (e)

Las personas con outcome igual a 1 tienden a ser mayores, con una mediana ligeramente superior en comparación con el grupo outcome igual a 0. Esto podría indicar que la edad está correlacionada con la aparición de la diabetes, ya que las personas de edad avanzada tienden a tener mayor riesgo de desarrollar esta enfermedad.

3.5.6. Embarazos (f)

Las mujeres diagnosticadas con diabetes tienden a haber tenido un número mayor de embarazos en comparación con el grupo outcome igual a 0. Esto podría indicar que el número de embarazos puede estar relacionado con un mayor riesgo de diabetes.

3.5.7. Predisposición genética para la diabetes (g)

Este factor es ligeramente mayor en el grupo de las mujeres con outcome igual a 1, aunque no parece haber una gran diferencia en las medianas. Según esto, las personas con diabetes tienen una ligera tendencia a tener mayor predisposición genética, lo cual es consistente con la naturaleza hereditaria de la diabetes.

3.5.8. Insulina (h)

Los valores de insulina son más altos y dispersos en el grupo no diagnosticado con diabetes, mientras que el grupo que sí lo está presenta valores más bajos. Esto puede ser debido a que la resistencia a la insulina es característica en la diabetes tipo 2.

3.6. Pruebas de Normalidad

Para evaluar si las variables siguen una distribución normal, se aplicó la prueba de normalidad de Shapiro-Wilk. Esta prueba es apropiada para la muestra de 768 mujeres, ya que se recomienda su uso para tamaños de muestra pequeños o moderados. Los resultados se presentan a continuación. Un valor p menor a 0.05 indica que la variable no sigue una distribución normal.

Variable	p-valor	Distribución Normal
Glucosa	1.777e-11	No es normal
Presión Sanguínea	6.463e-6	No es normal
Grosor de Piel	2.2e-16	No es normal
Insulina	$<$ 2.2e-16	No es normal
BMI	6.526e-9	No es normal
Edad	$<$ 2.2e-16	No es normal
Embarazos	$<$ 2.2e-16	No es normal
Función de Pedigree de Diabetes	$<$ 2.2e-16	No es normal

Tabla 4: Resultados de la prueba de normalidad de Shapiro-Wilk para las variables

3.6.1. Interpretación de los resultados de normalidad

Se aplicó la prueba de Shapiro-Wilk para determinar si las variables siguen una distribución normal. Como se puede observar en la tabla 4, el valor *p* para todas las variables es menor a 0.05, lo que indica que ninguna de las variables sigue una distribución normal. Aunque esto podría sugerir la necesidad de técnicas no paramétricas en algunos casos, la regresión logística es un método paramétrico que no requiere la suposición de normalidad en las variables independientes. Por lo tanto, la falta de normalidad no representa un impedimento para continuar con este modelo, que resulta adecuado para analizar y predecir la variable dependiente categórica.

3.7. Pruebas de Hipótesis

En este apartado, se realizaron pruebas de hipótesis para evaluar las diferencias entre los grupos diagnosticados con diabetes y aquellos que no. El objetivo principal fue determinar si existen diferencias significativas en las variables más relevantes y confirmar la relación entre las variables predictoras.

3.7.1. Pruebas de Hipótesis para Comparación de Medias

Para comparar las medias de las variables entre ambos grupos de pacientes, se plantearon las siguientes hipótesis:

- **Hipótesis Nula (H_0):** No existe una diferencia significativa en las medias de las variables entre los grupos de pacientes con y sin diabetes.
- **Hipótesis Alternativa (H_A):** Existe una diferencia significtiva en las medias de las variables entre los grupos de pacientes con y sin diabetes.

Dado que los resultados de la prueba de normalidad indicaron que las variables no siguen una distribución normal, se utilizó la prueba de Mann-Whitney U para comparar las medianas entre ambos grupos. Se seleccionó un nivel de significancia de $\alpha = 0{,}05$. Los resultados mostraron que todas las variables presentan diferencias significativas en sus distribuciones entre los pacientes con diabetes y aquellos sin diabetes.

Tabla 5: Resultados de las Pruebas de Mann-Whitney U para cada variable

Variable	**p-value**
Glucose	2.2e-16
BMI	2.2e-16
Blood Pressure	1.997e-06
Skin Thickness	2.878e-09
Insulin	2.276e-12
Age	2.2e-16
Pregnancies	3.745e-08
Diabetes Pedigree Function	1.197e-06

Estos resultados sugieren que todas las variables seleccionadas tienen una diferencias significativas en relación con la presencia de diabetes. Las pruebas recalcaron la importancia de estas variables como factores asociados a la enfermedad, lo cual justifica la implementación de un modelo de regresión logística para identificar cuáles de ellas contribuyen de manera más efectiva a la clasificación y predicción de la existencia de diabetes.

3.8. Resultados del Modelo de Regresión Logística

Se presenta el resumen de los coeficientes obtenidos mediante un modelo de regresión logística para predecir el diagnóstico de diabetes en función de las variables predictoras.

Variable	**Estimación**	**Error Estándar**	**Valor z**	**Valor p**	**Significancia**
(Intercept)	-9.0968	0.8126	-11.195	¡2e-16	***
Pregnancies	0.1250	0.0324	3.860	0.000113	***
Glucose	0.0374	0.0039	9.630	¡2e-16	***
BloodPressure	-0.0088	0.0086	-1.028	0.3038	
SkinThickness	0.0035	0.0131	0.265	0.7910	
Insulin	-0.0008	0.0012	-0.671	0.5022	
BMI	0.0931	0.0178	5.219	1.8e-07	***
DiabetesPedigreeFunction	0.8661	0.2963	2.923	0.0035	**
Age	0.0131	0.0095	1.382	0.1671	

Tabla 6: Resumen de los coeficientes del modelo de regresión logística con niveles de significancia.

3.8.1. Interpretación de los resultados del modelo lógistico

En el modelo de regresión logística, las variables de interés que muestran un impacto significativo en el diagnóstico de diabetes son Pregnancies, Glucose, BMI y Diabetes Pedigree Function, ya que las variables ya dichas tienen un coeficiente altamente significativo ($p < 0{,}01$), lo que nos dice que estas variables son determinantes claves en la probabilidad de desarrollar la enfermedad.

3.9. Importancia de las Variables en el Modelo de Regresión Logística

La siguiente figura muestra la importancia de las variables predictoras en el modelo de regresión logística basado en los coeficientes obtenidos.

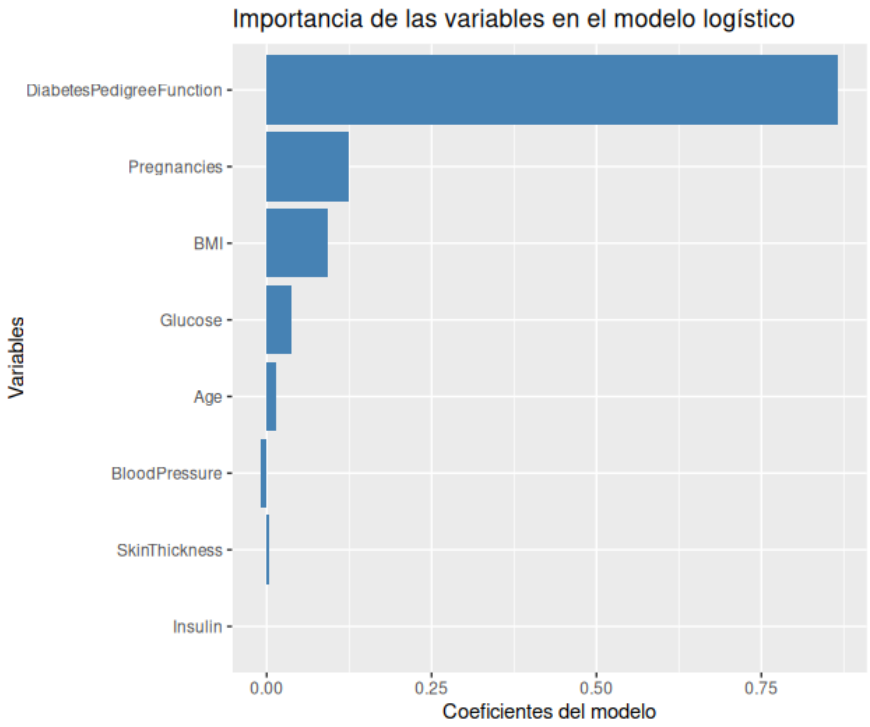


Figura 3: Importancia de las variables en el modelo de regresión logística.

En la gráfica, podemos observar que la variable DiabetesPedigreeFunction tiene el mayor impacto en el modelo, seguida por variables como Pregnancies, BMI, y Glucose.

3.10. Matriz de Confusión

Se presenta la matriz de confusión resultante del modelo de regresión logística aplicado.

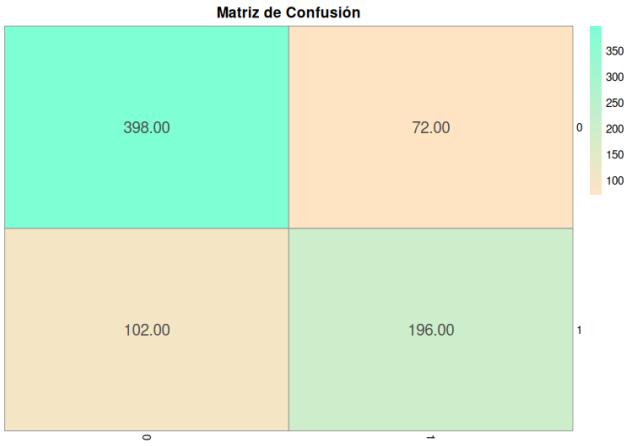


Figura 4: Matriz de Confusión del modelo de regresión logística.

La matriz de confusión se interpreta de la siguiente manera:

- **Verdaderos Positivos (VP):** 196 casos en los que el modelo predijo correctamente que el paciente tiene diabetes (Outcome = 1).
- **Verdaderos Negativos (VN):** 398 casos en los que el modelo predijo correctamente que el paciente no tiene diabetes (Outcome = 0).
- **Falsos Positivos (FP):** 102 casos en los que el modelo predijo diabetes, pero el paciente no la tiene.
- **Falsos Negativos (FN):** 72 casos en los que el modelo no predijo diabetes, pero el paciente sí la tiene.

3.10.1. Resultados de Sensibilidad, Especificidad y Precisión

A partir de la matriz de confusión, se obtienen los siguientes valores:

- **Sensibilidad:** $\frac{VP}{VP+FN} = \frac{196}{196+72} = 0{,}7313$. La sensibilidad refleja la capacidad del modelo para identificar correctamente a los pacientes con diabetes. En este caso, el 73.13 % de los pacientes con diabetes fueron correctamente identificados.
- **Especificidad:** $\frac{VN}{VN+FP} = \frac{398}{398+102} = 0{,}796$. La especificidad mide la capacidad del modelo para identificar correctamente a los pacientes sin diabetes. En este caso, el 79.6 % de los pacientes sin diabetes fueron correctamente identificados.

- **Precisión del modelo:** $\frac{VP+VN}{Total} = \frac{196+398}{768} = 0{,}7734$. La precisión indica qué tan bien predijo el modelo en general, siendo el 77.34 %.

3.11. Curva ROC y AUC

Para evaluar el rendimiento del modelo de clasificación, se utilizó la curva ROC (Receiver Operating Characteristic) y el cálculo del AUC (Area Under the Curve). La curva ROC muestra la relación entre la sensibilidad y la especificidad, lo que permite visualizar el desempeño del modelo en diferentes umbrales de decisión.

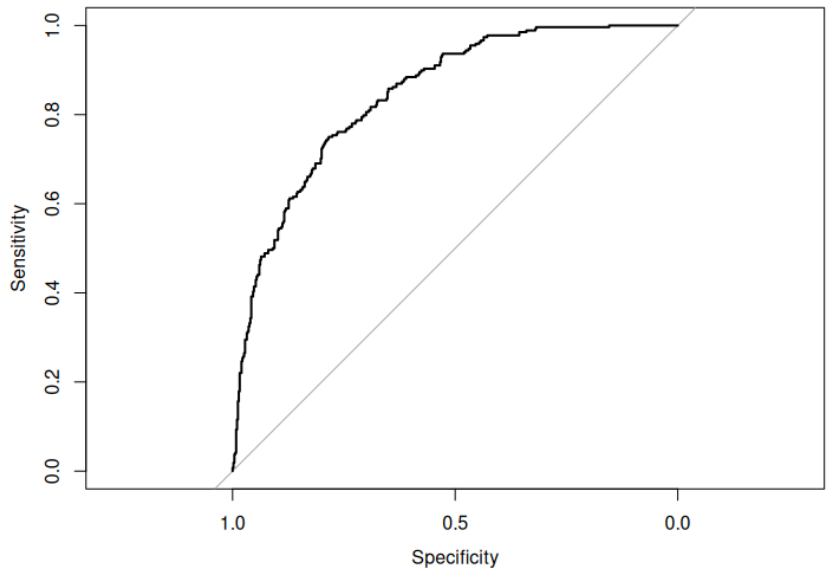


Figura 5: Curva ROC para el modelo de regresión logística.

La curva ROC presentada en la Figura 5 muestra un AUC de 0.8446, lo cual indica que el modelo tiene un buen rendimiento en la clasificación de los pacientes con y sin diabetes. Un AUC de 0.8446 significa que, en promedio, el modelo es capaz de discriminar correctamente entre un paciente con diabetes y uno sin diabetes el 84.46 % de las veces. Para este modelo, se utilizó el umbral de 0.35.

Luego, se utilizó la curva ROC para encontrar el mejor umbral de decisión (*threshold*) que maximiza tanto la sensibilidad como la especificidad del modelo. El mejor umbral encontrado fue de 0.3401, lo cual significa que a partir de una probabilidad mayor a este

valor, se predice que un paciente tiene diabetes.

3.11.1. Análisis de la Curva ROC

La curva ROC se acerca a la esquina superior izquierda, lo que indica que el modelo tiene una alta capacidad de discriminación. El AUC de 0.8446 es un buen indicador del rendimiento del modelo, siendo cercano a 1.

3.12. Matriz de Confusión y Curva ROC con Umbral Óptimo

Se presenta la matriz de confusión y la curva ROC para el modelo de regresión logística, ajustado con un umbral óptimo de 0.3401. Este umbral se obtuvo mediante el análisis de la curva ROC, maximizando la sensibilidad y especificidad del modelo.

3.12.1. Matriz de Confusión con Umbral Óptimo

La matriz de confusión resultante del nuevo umbral se presenta a continuación:

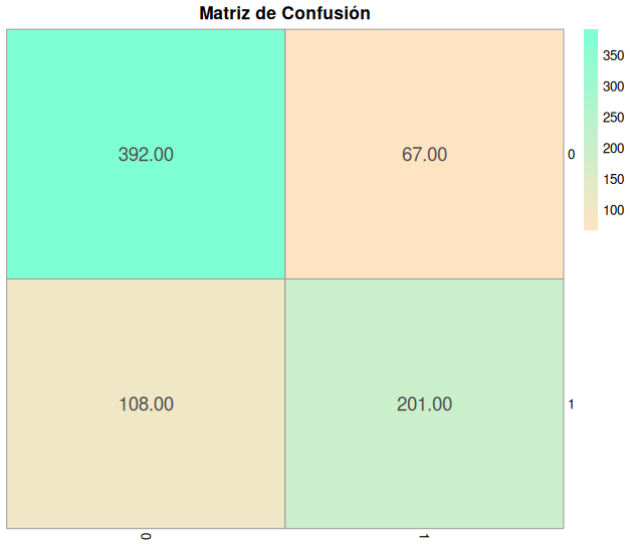


Figura 6: Matriz de Confusión con umbral óptimo (0.3401).

La matriz de confusión se interpreta de la siguiente manera:

- **Verdaderos Positivos (VP)**: 201 casos en los que el modelo predijo correctamente que el paciente tiene diabetes (Outcome = 1).

- **Verdaderos Negativos (VN)**: 392 casos en los que el modelo predijo correctamente que el paciente no tiene diabetes (Outcome = 0).
- **Falsos Positivos (FP)**: 108 casos en los que el modelo predijo diabetes, pero el paciente no la tiene.
- **Falsos Negativos (FN)**: 67 casos en los que el modelo no predijo diabetes, pero el paciente sí la tiene.

3.12.2. Resultados de Sensibilidad, Especificidad y Precisión con el Umbral Óptimo

A partir de esta nueva matriz de confusión, se obtienen los siguientes valores:

- **Sensibilidad:** $\frac{VP}{VP+FN} = \frac{201}{201+67} = 0{,}75$. La sensibilidad refleja la capacidad del modelo para identificar correctamente a los pacientes con diabetes. En este caso, el 75 % de los pacientes con diabetes fueron correctamente identificados.
- **Especificidad:** $\frac{VN}{VN+FP} = \frac{392}{392+108} = 0{,}784$. La especificidad mide la capacidad del modelo para identificar correctamente a los pacientes sin diabetes. En este caso, el 78.4 % de los pacientes sin diabetes fueron correctamente identificados.
- **Precisión del modelo:** $\frac{VP+VN}{Total} = \frac{201+392}{768} = 0{,}7721$. La precisión indica que el modelo predijo correctamente en el 77.21 % de los casos.

3.12.3. Curva ROC con Umbral Óptimo

Se utilizó la curva ROC para visualizar el rendimiento del modelo con el nuevo umbral. La curva se presenta a continuación:

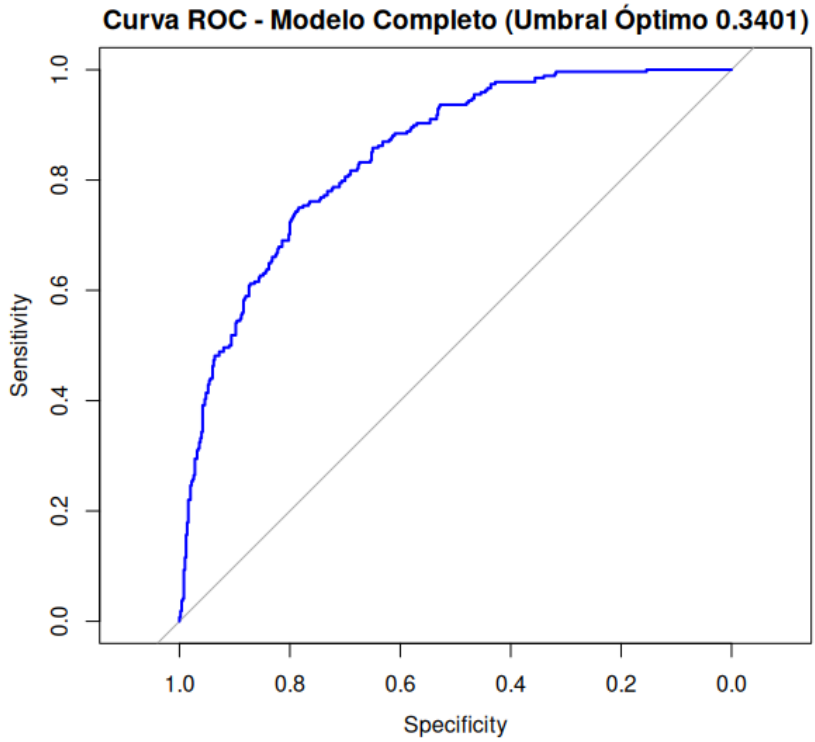


Figura 7: Curva ROC para el modelo completo con umbral óptimo (0.3401).

El AUC con el nuevo umbral es de 0.8446, lo que indica un buen rendimiento del modelo. La curva ROC muestra una alta capacidad de discriminación entre pacientes con y sin diabetes.

3.12.4. Análisis de la Curva ROC

La curva ROC continúa mostrando una buena capacidad discriminativa, acercándose a la esquina superior izquierda. El AUC de 0.8446 sugiere que el modelo tiene una alta precisión para predecir el diagnóstico de diabetes. El nuevo umbral de 0.3401 mejora la especificidad, reduciendo los falsos positivos, pero disminuye ligeramente la sensibilidad, aumentando los falsos negativos.

3.13. Ajuste del Modelo Reducido

Se ajustó un modelo de regresión logística reducido utilizando tres variables predictoras: *Glucose*, *BMI* y *DiabetesPedigreeFunction*. El ajuste de este modelo reducido busca

evaluar la capacidad predictiva de estas variables clave para diagnosticar diabetes.

A continuación, se presenta el gráfico de residuos vs valores ajustados para el modelo original, lo cual permite analizar visualmente la calidad del ajuste.

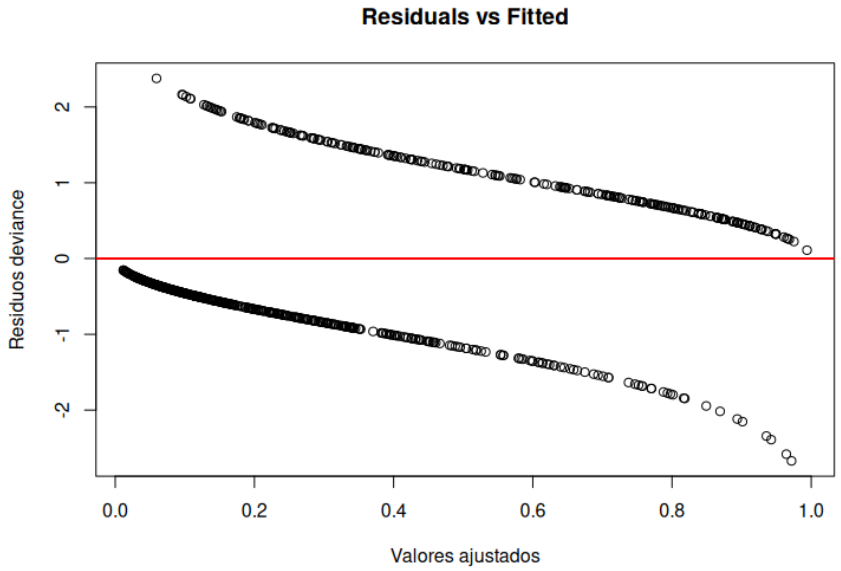


Figura 8: Gráfico de Residuos vs Valores Ajustados para el Modelo Original.

En este gráfico, los residuos se grafican en función de los valores ajustados del modelo. La línea roja indica el nivel en el cual los residuos no presentan desviación significativa. Aunque algunos puntos se alejan de la línea, la mayoría se distribuyen de manera homogénea, lo que sugiere un ajuste razonable del modelo original.

El próximo paso será comparar el desempeño del modelo reducido con el modelo original para evaluar su precisión y efectividad en la predicción del diagnóstico de diabetes.

3.14. Comparación de las Curvas ROC: Modelo Completo vs. Modelo Reducido

La Figura 9 muestra la comparación entre las curvas ROC del modelo completo y el modelo reducido, ambos evaluados con el mismo umbral óptimo (0.3401). El área bajo la curva (AUC) es un indicador clave para evaluar el rendimiento de los modelos de clasificación, ya que representa la capacidad del modelo para discriminar entre las clases positivas

y negativas (en este caso, pacientes con y sin diabetes).

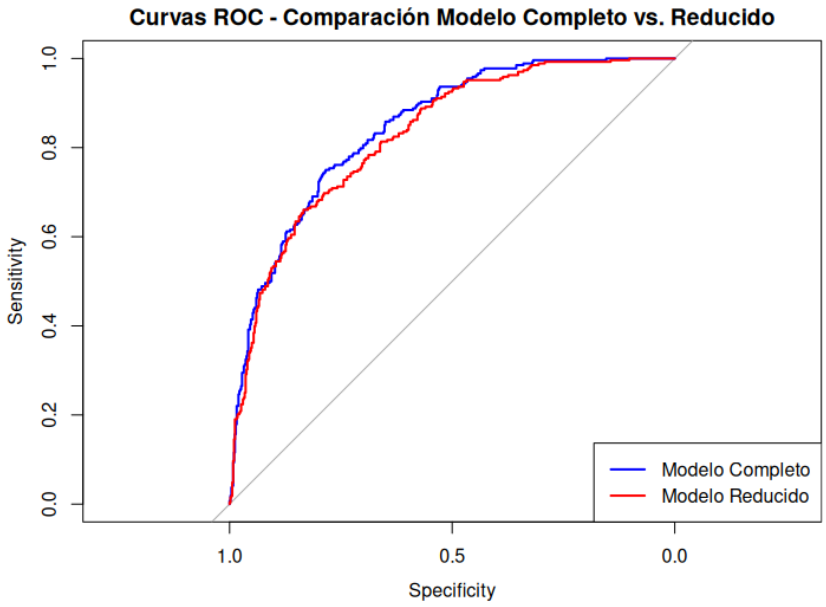


Figura 9: Curvas ROC - Comparación entre el Modelo Completo y el Modelo Reducido.

3.14.1. Análisis de las Curvas ROC

Como se observa en la Figura 9, el AUC del modelo completo (curva azul) es ligeramente superior al del modelo reducido (curva roja). Esto indica que el modelo completo tiene un mejor rendimiento en la clasificación de los pacientes en comparación con el modelo reducido.

Los resultados clave de esta comparación son:

- **Modelo Completo:** La curva ROC se muestra en azul, con un AUC mayor que el del modelo reducido, lo que sugiere una mayor capacidad para discriminar entre pacientes con y sin diabetes.
- **Modelo Reducido:** La curva ROC en rojo, aunque sigue una trayectoria similar a la del modelo completo, tiene un AUC ligeramente menor, lo que indica que su rendimiento es inferior en términos de clasificación.

En general, la diferencia entre ambos modelos es visible en el gráfico, donde el modelo completo logra una mejor sensibilidad y especificidad en comparación con el modelo

reducido. Sin embargo, el modelo reducido sigue ofreciendo una aproximación aceptable con un menor número de variables predictoras, lo que puede ser útil para simplificar el modelo sin sacrificar demasiado el rendimiento.

4. Conclusiones

El modelo de regresión logística reducido ha mostrado una buena capacidad predictiva con un AUC de 0.8277, lo que indica un desempeño eficaz en la clasificación de pacientes con y sin diabetes. Al optimizar el umbral a 0.3401, se observó una mejora en la sensibilidad del modelo, alcanzando un 75 %, aunque con una ligera disminución en la especificidad, lo que sugiere un buen equilibrio en la identificación de pacientes con diabetes. Sin embargo, el modelo podría beneficiarse de ajustes adicionales para mejorar la especificidad y reducir los falsos positivos.

El modelo reducido, que solo considera a las variables de glucosa, índice de masa corporal (BMI) y predisposición genética, también demostró ser un buen predictor. Este modelo simplificado ofrece una opción práctica cuando se dispone de menos datos, sin sacrificar en sobremanera la precisión del diagnóstico. Estos resultados concuerdan con estudios previos (Ozery-Flato et al., 2013) y (González Sánchez & Serrano Ríos, 2011), que demostraron que la combinación de glucosa en plasma en ayunas, índice de masa corporal (BMI) y hemoglobina glucosilada es un predictor preciso de la aparición de diabetes tipo 2 en sujetos con síndrome metabólico. En dicho estudio, el área bajo la curva (AUC) del modelo predictivo fue de 0.92, destacando la importancia de estas variables como predictores confiables para el riesgo de diabetes (National Institute of Diabetes and Digestive and Kidney Diseases, 2024). Cabe mencionar que, aunque los estudios citados utilizaron bases de datos diferentes, los factores de riesgo identificados fueron consistentes en distintas poblaciones, lo que subraya la relevancia de estas variables en la predicción de la diabetes.

Para mejorar este modelo en futuros estudios, podría ser útil emplear el uso de técnicas de machine learning más avanzadas, como árboles de decisión o redes neuronales, que han mostrado ser eficaces en otros estudios para mejorar la sensibilidad y especificidad en la detección de diabetes tipo 2 (Xie et al., 2019). Además, la validación cruzada con di-

ferentes subconjuntos de datos podría optimizar aún más el rendimiento y generalización del modelo. Ampliar el conjunto de datos con una mayor diversidad de poblaciones podría aumentar la robustez y aplicabilidad del modelo en diferentes contextos médicos.

Finalmente, se contrastaron los resultados obtenidos con un análisis basado en la misma base de datos Pima Indians Diabetes Database, en el cual se alcanzó un AUC de 0.86 utilizando un pipeline completo de machine learning (pouryaayria, 2024). Esta consistencia de resultados sugiere que los modelos utilizados son acertados y que las variables seleccionadas son efectivas para la detección de diabetes en este conjunto de datos.

Referencias

González Sánchez, J. L., & Serrano Ríos, M. (2011). Genética de la diabetes mellitus [Accessed: 20-10-2024]. *Nefrología*, *31*(6), 686-695. https://www.revistanefrologia.com/es-genetica-diabetes-mellitus-articulo-X2013757511002452

Kaggle. (2016). Pima Indians diabetes database [Consultado el 18 de octubre, 2024].

Little, R. J., & Rubin, D. B. (2002). Statistical analysis with missing data [Consultado el 18 de octubre, 2024]. *John Wiley & Sons*.

National Institute of Diabetes and Digestive and Kidney Diseases. (2024). Factores de riesgo para la diabetes tipo 2 [Accessed: 20-10-2024]. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/factores-riesgo-tipo-2

Ozery-Flato, M., Parush, N., El-Hay, T., Visockienė, Ž., Ryliškytė, L., Badarienė, J., Solovjova, S., Kovaitė, M., Navickas, R., & Laucevičius, A. (2013). Predictive models for type 2 diabetes onset in middle-aged subjects with the metabolic syndrome [Consultado el 18 de octubre, 2024]. *Diabetology & Metabolic Syndrome*, *5*(36). https://doi.org/10.1186/1758-5996-5-36

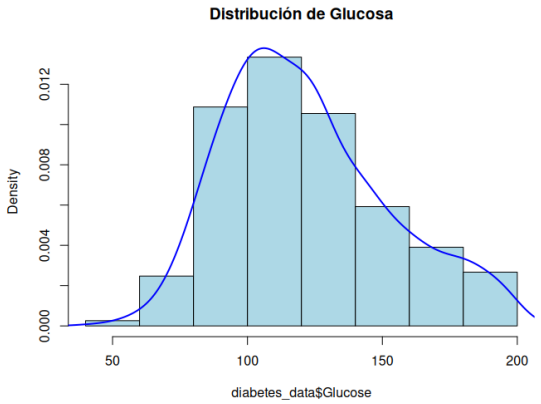
pouryaayria. (2024). A complete ML pipeline tutorial with AUC 86 [Consultado el 18 de octubre, 2024].

Xie, Z., Nikolayeva, O., Luo, J., & Li, D. (2019). Building risk prediction models for type 2 diabetes using machine learning techniques [Consultado el 18 de octubre, 2024]. *Preventing Chronic Disease*, *16*(E130). https://doi.org/10.5888/pcd16.190109

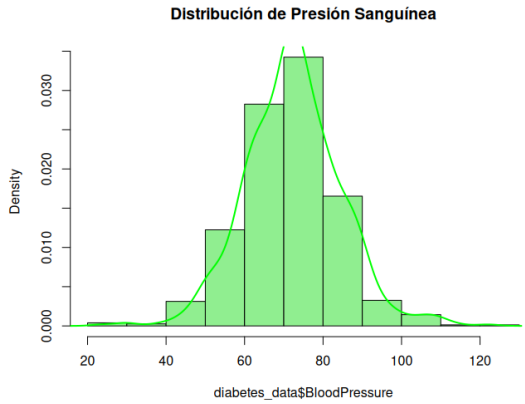
5. Anexo

Visualización de Histogramas

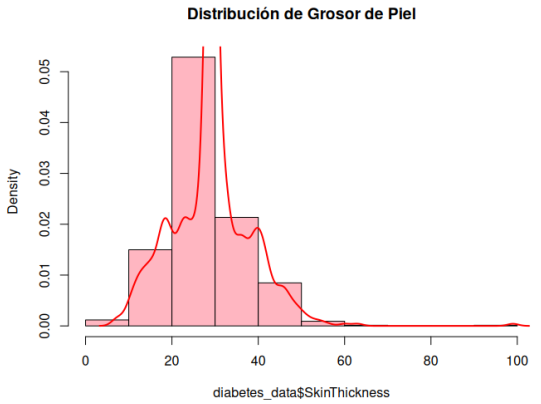
Aquí se presentan los histogramas de las variables relevantes para una visualización más detallada.



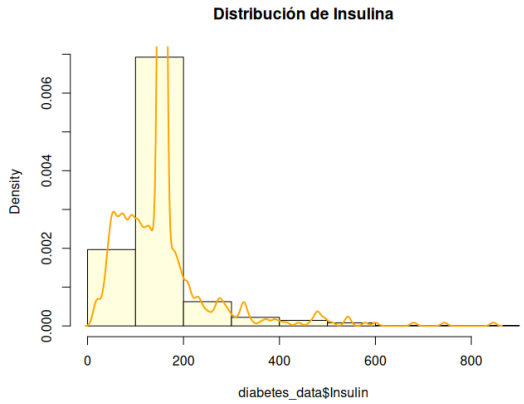
(a) Distribución de Glucosa



(b) Distribución de la Presión Sanguínea

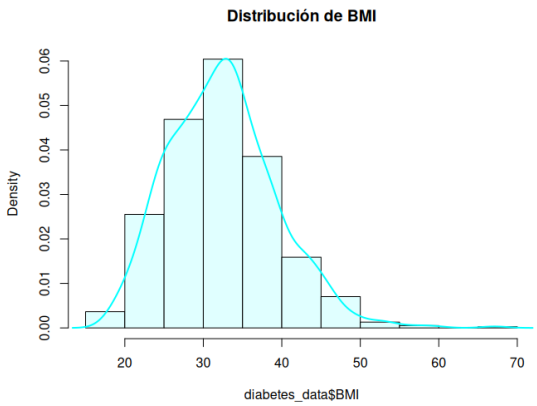


(c) Distribución del Grosor de la Piel

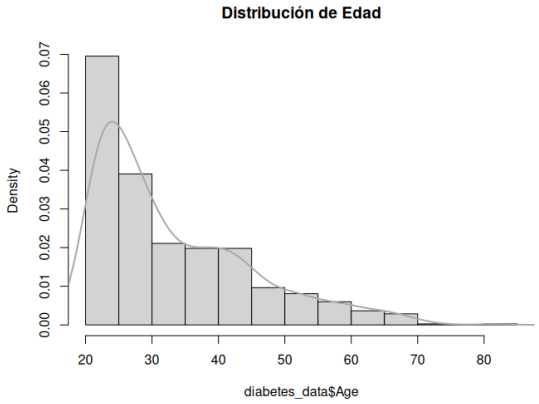


(d) Distribución de los Niveles de Insulina

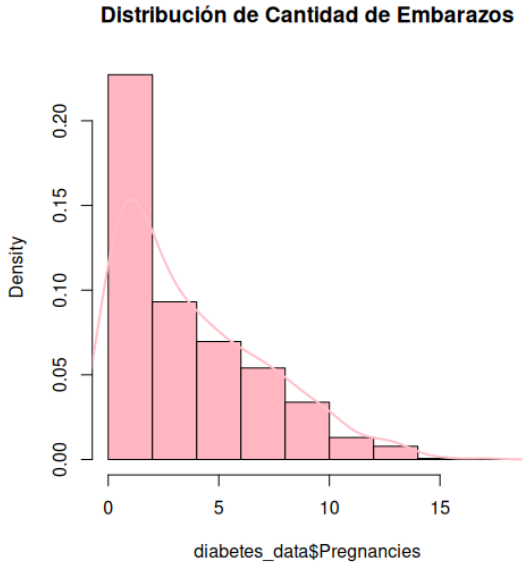
Figura 10: Histogramas con curvas de densidad para las variables Glucose, BloodPressure, SkinThickness y Insulin.



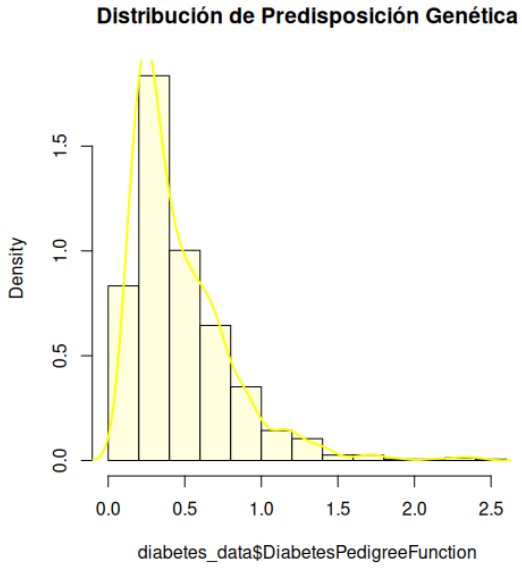
(a) Distribución del Índice de Masa Corporal (BMI)



(b) Distribución de la Edad



(c) Distribución de Cantidad de Embarazos



(d) Distribución de la Predisposición Genética

Figura 11: Histogramas con curvas de densidad para las variables BMI, Age, Pregnancies y DiabetesPedigreeFunction.