Escuela Nacional de Estudios Superiores

ENES MORELIA

Proyecto Final

Evaluación Estadística de Indicadores Clínicos en la Diabetes

Alumnas
Erika Monserrat Correa Hernández.
Hannia Ashley Alvarado Galván

Profesores
María del Río Francos
César Andrés Torres Miranda
Teresa Patiño Cárdenas

Estadística Descriptiva e Inferencial.
Semestre 2024-2

Morelia, Mich,
29 de mayo de 2024

Índice

1. Introducción

La diabetes se presenta como una crisis de salud pública a nivel global, particularmente severa en México, donde amenaza el bienestar de millones.

Su incidencia, alarmantemente alta y en constante aumento, no muestra signos de desaceleración, destacando una tendencia preocupante especialmente entre los mexicanos.
A pesar de numerosos esfuerzos, las estrategias actuales para abordar la diabetes han demostrado no ser completamente efectivas. Esta situación subraya la urgente necesidad de adoptar enfoques que estén mejor adaptados a las complejidades y realidades específicas de la población mexicana afectada por esta enfermedad.

2. Objetivos

Nuestro estudio busca identificar las variables con un mayor poder predictivo sobre la incidencia de la diabetes en este grupo demográfico. Al comprender estos factores determinantes, esperamos avanzar hacia una mejor comprensión de las causas subyacentes y el desarrollo de la enfermedad.

Nuestro análisis va más allá de las variables individuales para explorar la interacción compleja entre múltiples factores que afectan la probabilidad de desarrollar diabetes, incluyendo el número de embarazos, los niveles de glucosa, la presión sanguínea, la función de pedigrí y la edad.

El objetivo final de nuestra investigación es contribuir al desarrollo de estrategias más precisas y efectivas para combatir la diabetes en México. Aspiramos a proporcionar herramientas valiosas que mejoren nuestra comprensión de la relación entre estas variables y la presencia (o ausencia) de la diabetes, y que ayuden a gestionar mejor esta enfermedad, reduciendo su impacto en las comunidades vulnerables, especialmente entre las mujeres, que enfrentan un riesgo mayor.

3. Materiales y Métodos

El set de datos con el que se trabajo fue conseguido en una plataforma en línea para ciencia de datos y aprendizaje automático.
(https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download)
Se hizo una limpieza de datos para sacar mejores resultados e interpretaciones estadísticas:

- Primero hicimos una traducción de las columnas de los datos a español para un entendimiento inmediato.
- Verificamos la cantidad de renglones y columnas (768, 9).
- Checamos la cantidad de ceros que había por cada columna.

Columnas	Cantidad de ceros
Embarazos	111
Glucosa	5
PresionSanguinea	35
GrosorPiel	227
Insulina	374
IMC	11
FuncionPedigriDiabetes	0
Edad	0
Resultado	500

Figura 1: Cantidad de ceros por cada columna.

-En embarazos es normal que no hayan tenido uno, por lo que dejamos los ceros.
-Los resultados es normal que haya ceros ya que indica la ausencia de diabetes.
-En la Glucosa, Presion Sanguínea y el IMC tienen pocos datos faltantes, por lo que rellenaremos con alguna medida de tendencia central.
Hacemos un estudio con gráficas de bigotes (boxplots) para decidir que medida de tendencia central utilizar.

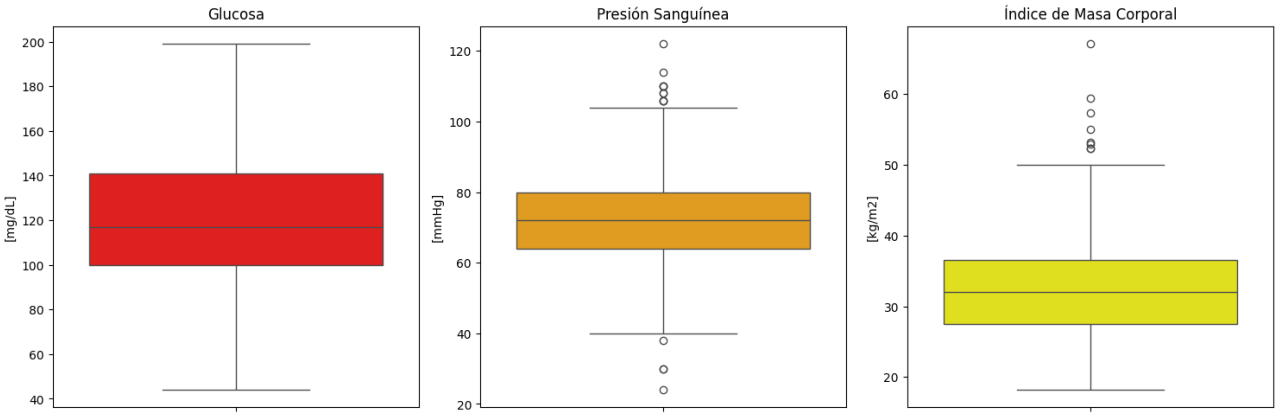


Figura 2: Gráficas de Bigotes para decidir la medida de tendencia central

Interpretación de los Boxplots: Glucosa:

La mediana parece estar alrededor de 100.
Los bigotes se extienden desde aproximadamente 60 hasta 160, indicando el rango típico.
Hay algunos valores atípicos por encima de 160.

Se puede observar que los únicos datos atípicos son los de la gráfica de la Glucosa con valor igual a 0.

Presión Sanguínea:

La mediana está cerca de 70. El rango típico (los bigotes) va desde cerca de 50 hasta 90.

Se puede observar que hay más datos atípicos que los valores igual a 0 y que se muestran más en los valores de 100 a 120 aproximadamente

Hay varios valores atípicos tanto por debajo de 50 como por encima de 90.

IMC:

La mediana está cerca de 30.
Los bigotes se extienden desde alrededor de 20 hasta 50.
Existen múltiples valores atípicos por encima de 50.

Estratificamos entre las personas que tienen diabetes y las que no, posteriormente sacamos la media de cada una de las columnas y rellenamos los datos faltantes.

-El grosor de piel tiene un faltante de el 29.55 % de los datos.
-La Insulina tiene un faltante del 48.69 % de los datos.

Por lo que optaremos por eliminar estas columnas, ya que al rellenar, podríamos sesgar el resultado de los análisis estadísticos.

Los principales métodos utilizados fueron el análisis de correlación entre las variables de nuestra base de datos apoyándonos en correlogramas para poder identificar las relaciones más fuertes en el resultado de tener diabetes o no, usando el coeficiente de correlación de Spearman. También, pruebas de Box Cox y Shapiro para determinar si había normalidad, además de pruebas de Wilcoxon y Kruskal.

En este estudio, se emplearon varios métodos estadísticos para analizar los datos de la base de datos de diabetes. Los métodos principales incluyen:

- **Análisis de correlación:** Utilizamos el análisis de correlación para explorar las relaciones entre las variables de nuestra base de datos. Nos apoyamos en correlogramas, los cuales nos permitieron identificar las relaciones más fuertes que influyen en el resultado de tener o no diabetes. Este análisis se llevó a cabo usando el coeficiente de correlación de Spearman.
- **Pruebas de normalidad:** Se aplicaron las pruebas de Box-Cox y Shapiro-Wilk para determinar la normalidad de las distribuciones de nuestras variables. Estas pruebas son esenciales para validar ciertos supuestos en pruebas estadísticas adicionales.
- **Pruebas no paramétricas:** Debido a la falta de normalidad en algunos casos, se emplearon pruebas no paramétricas como la prueba de Wilcoxon para muestras relacionadas y la prueba de Kruskal-Wallis para más de dos grupos independientes. Estas pruebas son útiles para comparar medianas entre grupos cuando los datos no siguen una distribución normal.

4. Desarrollo

4.1. Herramientas

Las herramientas que usamos para el análisis estadístico fueron Python, los conocimientos aprendidos a lo largo del semestre y algunos datos de otras páginas para poder estratificar correctamente, además de LaTeX para escribir este reporte.

Datos	Cantidad	Columnas
Originales	768	9
Limpios	768	7

Figura 3: Cantidad de datos

Variables Originales	Descripción	Tipo
Función Pedigrí	Para expresar el porcentaje de Diabetes	Continua
Edad	Para expresar la edad	Discreta
Embarazos	Para expresar el Número de embarazos	Discreta
Glucosa	Para expresar el nivel de glucosa en sangre	Discreta
Grosor de la piel	Para expresar el grosor de la piel	Discreta
IMC	Para expresar el índice de masa corporal	Continua
Insulina	Para expresar el nivel de insulina en sangre	Discreta
Presión sanguínea	Para expresar la medición de la presión arterial	Discreta
Resultado	Para expresar el resultado final, 1 es Sí y 0 es No	Discreta

Figura 4: Variables originales

Variables que utilizamos	Descripción	Tipo
Edad	Para expresar la edad	Discreta
Embarazos	Para expresar el Número de embarazos	Discreta
Glucosa	Para expresar el nivel de glucosa en sangre	Discreta
Resultado	Para expresar el resultado final, 1 es Sí y 0 es No	Discreta
IMC	Para expresar el índice de masa corporal	Continua

Figura 5: Variables utilizadas

En todos nuestros métodos decidimos optar por una significancia de 0.05 para maximizar la capacidad para detectar diferencias significativas, como lo fue en nuestras pruebas de Shapiro, donde H_0 es que existe normalidad en nuestros datos. Para la parte de Wilcoxon y Kruskal, nuestra H_0 es que no hay diferencia significativa entre las muestras.

Acerca de nuestros estadígrafos, nos enfocamos en hacer una comparativa entre personas sanas y personas enfermas. Pudimos notar que en todas las variables, las personas enfermas tienen un valor más alto en las medias, sin embargo, las más notorias fueron en el caso de la glucosa, IMC y edad.

4.2. Procedimiento

4.2.1. Estadística Descriptiva

Principalmente decidimos dividir nuestro conjunto de datos en dos estratos, con base al resultado que se tiene, 0 si no tienen diabetes y 1 si la tienen, dónde 500 no tienen la enfermedad, es decir son persoanas sanas y 268 son personas enfermas que tienen diabetes.

Hicimos un estudio de cada una de las variables originales,su correlación,dispersión ,medidas de tendencia central y asimetría.
El estudio de las variables, nos permitio identificar por cuales variables ibamos a agrupar o estratificar, y volvimos a hacer el estudio pero ya con estas variables identificadas. (Edad, Embarazos,Glucosa e Indice de Masa Corporal).
Después queríamos ver como era la distribución de nuestros datos por lo que realizamos gráficas de densidad, dispersión, y de caja y bigotes(boxplots)

Comenzamos con la prueba de normalidad o de bondad y ajuste usando Shapiro Wilk,ya que se contaba con más de 50 datos, se convirtieron los valores a tipo entero o flotante según correspondía,aplicamos la prueba BoxCox y guardamos el valor de lambda ajustado que indica como se transformaron los datos, luego usamos la prueba Shapiro Wilk en los datos ya transformados, nos dimos cuenta que no teniamos normalidad en ninguna de nuestras variables por lo que rechazamos la hipotesis nula es decir que nuestros datos no provenían de una distribución normal.

4.2.2. Pruebas no paramétricas

Iniciamos con las pruebas no paramétricas, como contabamos con más de dos muestreos realizamos la prueba de Kruskal Wallis para ver si había diferencias significativas entre las variables que elegimos, si había al menos un tratamiento diferente por lo que continuamos con pruebas pareadas de Wilcoxon verificando que el valor de p fuera menor a 0.05, es decir hay diferencias significativas, después generamos los pares utilizando la función de python 'itertools.combinations', que crea todos los pares posibles de las variables elegidas, lo que hace que podamos comparar par a par, despúes aplicamos la Prueba de Wilcoxon para cada par de columnas y como resultado nos da un valor estadístico y un valor de p, lo que nos ayuda a determinar si las diferencias son estadísticamente significativas entre las columnas elegidas, como resultado encontramos diferencias significativas entre todos los pares de columnas.

4.2.3. Inferencias Bivariadas

¿Qué significa que haya normalidad?

Cuando se dice que los datos tienen distribución normal, se refiere a que su distribución Es simétrica alrededor de la media y que además ésta coincide con la mediana. También conocida como distribución de Gauss o curva de campana.

Correlación no paramétrica (si no hay normalidad)

Para los datos que se tenía de diabetes, no se tenía normalidad en ninguna de las variables, en esta situación se optó por correlaciones no paramétricas, y con ello, se utiliza el coeficiente de correlación de Spearman, esto nos ayuda a medir la fuerza y la dirección de la asociación entre dos variables. Este coeficiente puede tomar valores de entre -1 y 1, siendo -1 una correlación perfecta negativa entre las variables y 1 una correlación perfecta positiva.

Sintetizar las correlaciones estadísticamente significativas mediante un correlograma

Para poder visualizar de mejor manera y más rápida las correlaciones entre las variables se hace uso de un correlograma que se puede generar utilizando R, Seaborn o matplotlib, esto es de gran utilidad para identificar patrones entre variables. Un ejemplo puede ser la correlación entre Edad y Embarazos.

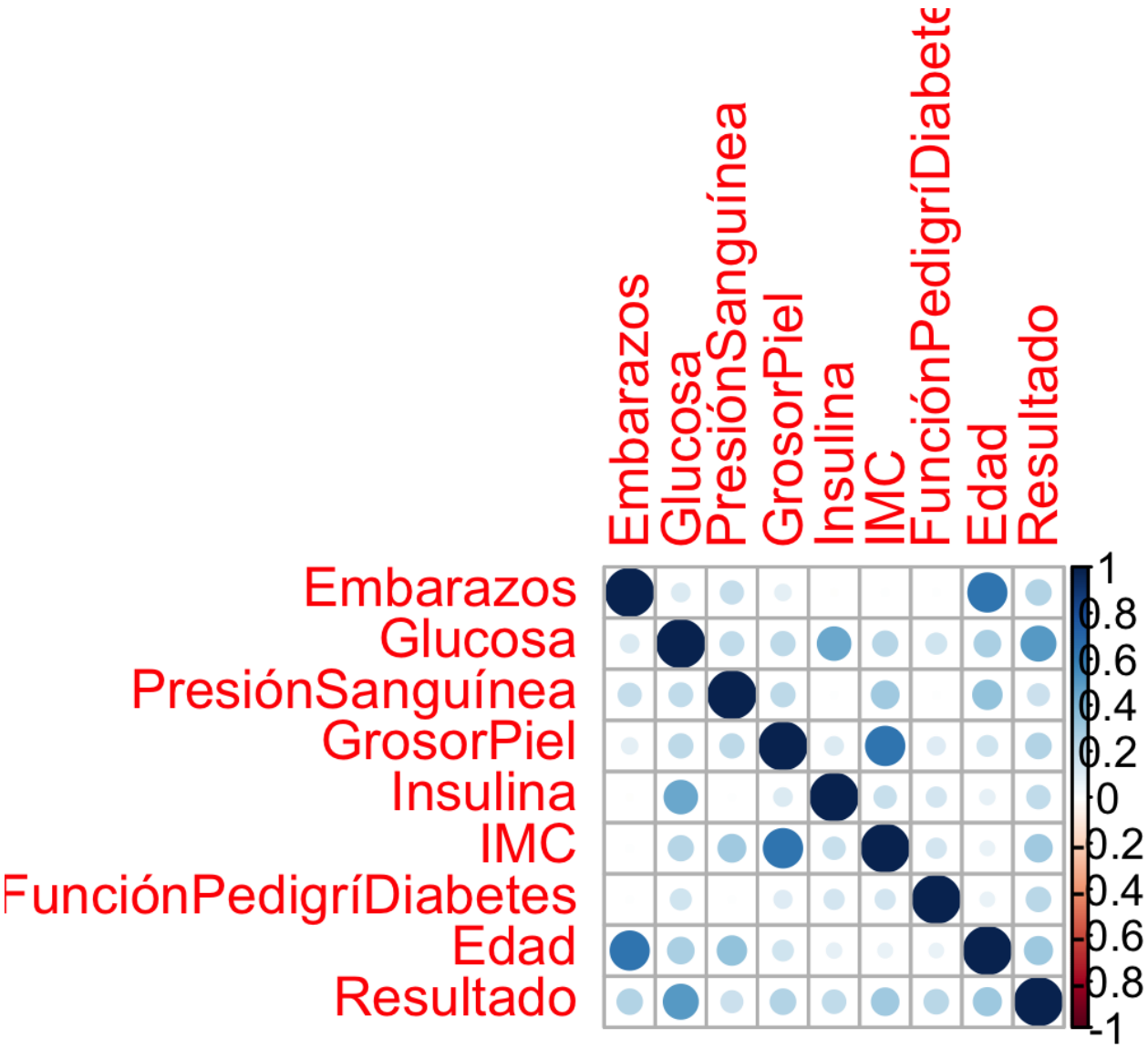


Figura 6: Correlograma general de las variables.

Para medir si realmente hay una correlación significativa, se hace una matriz de p-valores de las correlaciones entre las columnas de una matriz de datos, esto para identificar las variables donde hay un efecto más fuerte entre las variables y de igual manera se puede hacer un correlograma para identificar visualmente las correlaciones significativas utilizando el método spearman por falta de normalidad. Indicando el nivel de significancia.

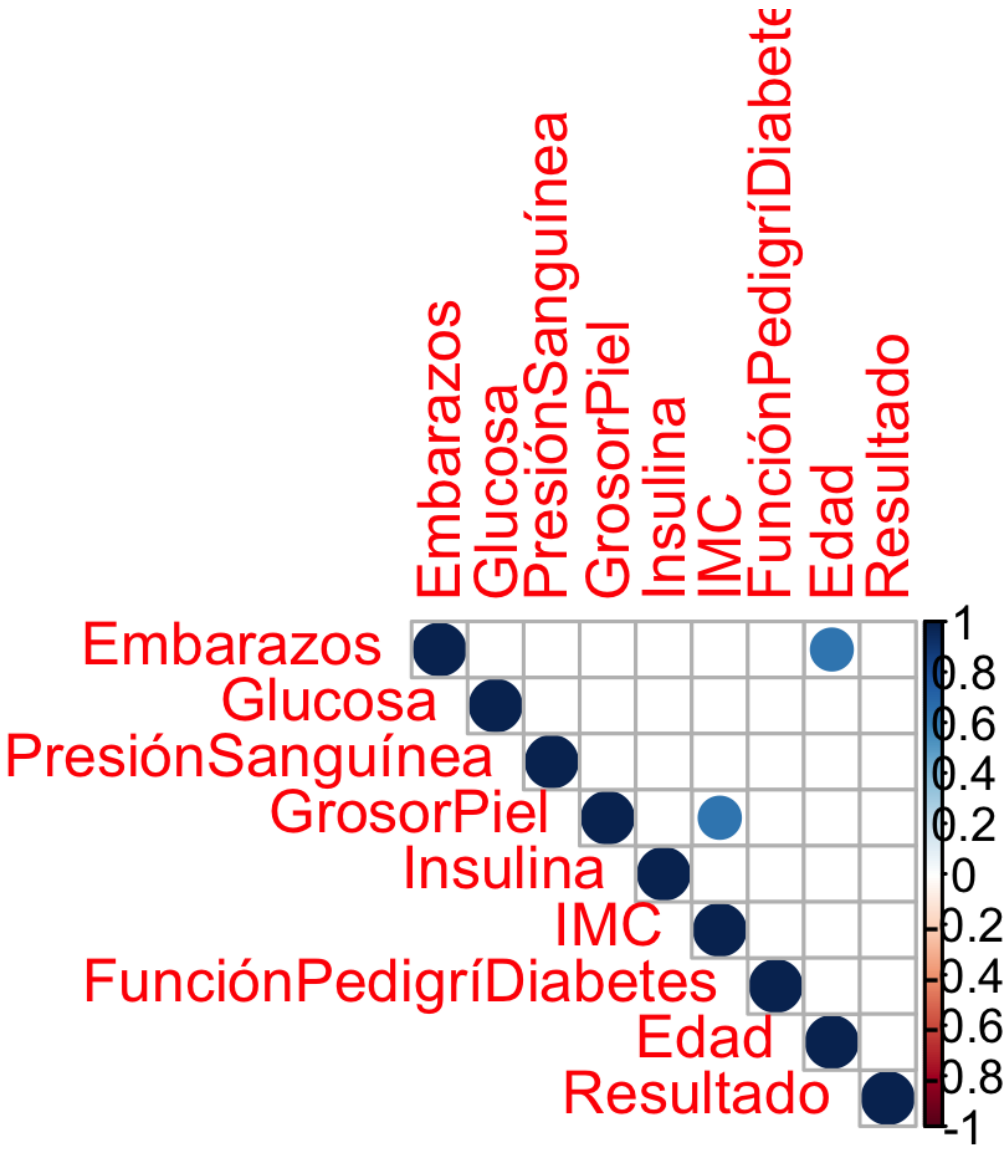


Figura 7: Correlograma de variables con significancia de 0.01.

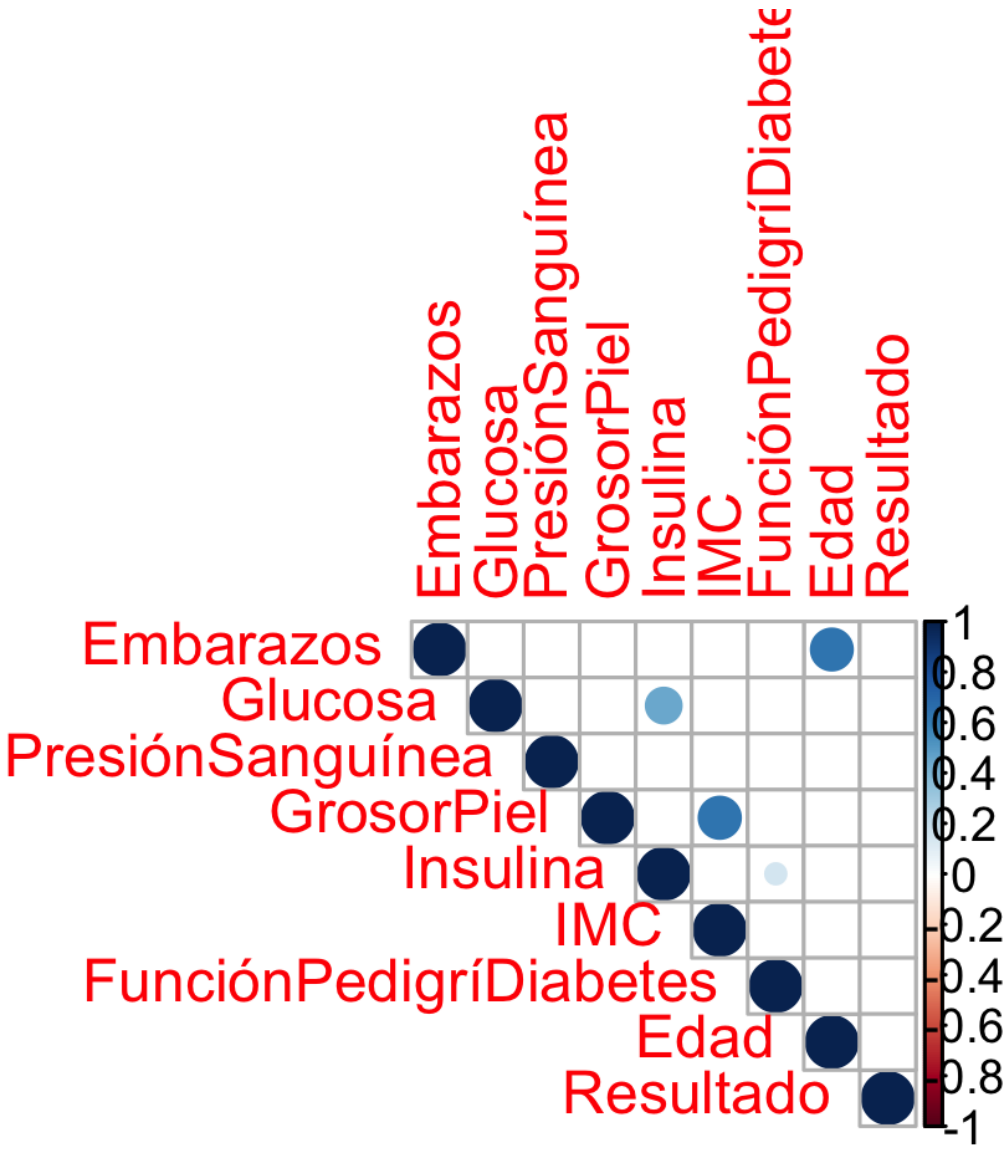


Figura 8: Correlograma de variables con significancia de 0.05.

Lo que se utiliza es el análisis de varianza (ANOVA), que es una técnica estadística para comparar las medias de dos variables y poder determinar si hay diferencias significativas entre ellos.

Variable	Df	Sum Sq	Mean Sq	F value	Pr(>F)	Shapiro p-value
Embarazos	1	8.59	8.591	39.67	5.07e-10	< 2.2e-16
IMC	1	17.34	17.343	84.54	< 2e-16	< 2.2e-16
Glucosa	1	42.92	42.92	249.9	< 2e-16	6.498e-16
Edad	1	9.91	9.913	46.14	2.21e-11	< 2.2e-16

Figura 9: Tabla de regresión y normalidad de los residuos

En este caso, se hizo el ANOVA con cada variable y la prueba de normalidad de los residuos. Por desgracia, no había normalidad en ninguno de los residuos de las ANOVAs, ya que el valor p era menor a 0.05.

4.3. Atributos

- **La Función Pedigrí**

  La función pedigrí para la diabetes (FPD) es una herramienta que se utiliza para estimar el riesgo de desarrollar diabetes tipo 2 en función de los antecedentes familiares de la enfermedad. Se basa en la idea de que la diabetes tipo 2 es una enfermedad compleja que está influenciada tanto por factores genéticos como ambientales.

  La FPD se calcula asignando un valor ponderado a cada miembro de la familia en función de su estado de diabetes. Los miembros de la familia con diabetes reciben un valor más alto, mientras que los que no la tienen reciben un valor más bajo. Los valores ponderados se suman a continuación para obtener una puntuación total de FPD.

  Una puntuación de FPD más alta indica un mayor riesgo de desarrollar diabetes tipo 2. Sin embargo, es importante tener en cuenta que la FPD es solo una herramienta de estimación y no puede predecir con certeza si una persona desarrollará o no la enfermedad. Otros factores, como el estilo de vida y el peso corporal, también juegan un papel importante en el riesgo de diabetes tipo 2.

  Al comenzar a hacer el estudio de esta variable hicimos una división en tres intervalos: coment describir porque los decidimos -bajo -moderado -alto

  Nos percatamos de que la mayoría de las personas tenían un indice de pedigrí bajo, y la distribución de la DPF en la población no era suficientemente variada como para proporcionar una clara distinción entre los individuos con y sin diabetes, por lo que decidimos no usar esta variable.

  - **Edad**

  La edad es un factor crucial en la estratificación de los riesgos y en la interpretación de los resultados del estudio. Los riesgos asociados con la diabetes, así como las recomendaciones de tratamiento, pueden variar significativamente con la edad del individuo.
  Por lo que decidimos usar esta variable, además en la esta variable tiene un indice alto de correlación (0.54 en un rango de 0 a 1.)

  - **Embarazos**

  Esta variable ayuda a identificar a las mujeres que pueden tener un mayor riesgo de diabetes gestacional o tipo 2 después del parto,y de acuerdo a la cantidad de embarazos que tuvo permitiendo estudios específicos sobre la ausencia o presencia de la enfermedad en este grupo.Por lo que decidimos usar esta variable, además en la esta variable tiene un indice alto de correlación (0.54 en un rango de 0 a 1.)

- **Glucosa**

El nivel de glucosa en sangre es la medida directa más importante para diagnosticar la diabetes. Los niveles elevados de glucosa en ayunas o en respuestas a las pruebas de tolerancia oral a la glucosa son indicativos de disfunciones en el metabolismo de la glucosa del cuerpo y a decidir de manera acertada la presencia de la enfermedad. Por lo que decidimos usar esta variable, además en la esta variable tiene un indice alto de correlación (0.50 en un rango de 0 a 1.)

- **Grosor de la piel**

El grosor de la piel puede ser un indicador indirecto de la resistencia a la insulina. Un grosor de pliegue cutáneo más grande puede estar asociado con mayores cantidades de tejido adiposo subcutáneo, lo cual a su vez se ha vinculado con resistencia a la insulina.

La resistencia a la insulina es un precursor clave de la diabetes tipo 2. El grosor de la piel tenia una cantidad considerable de datos faltantes que al rellenarlos hubieramos sesgado las estadísticas por lo que decidimos no usarlos.

- **Indice de Masa Coporal (IMC)**

El IMC es una medida simple que relaciona el peso con la altura de una persona (kg/m^2) y se usa para categorizar a los individuos en diferentes rangos de peso: bajo peso, normal, sobrepeso y obesidad. Estas categorías ayudan a los profesionales médicos a identificar rápidamente a las personas que pueden tener un riesgo elevado de problemas de salud crónicos como la diabetes tipo 2. Por lo que decidimos usarlo aunque su indice de correlación no es tan alto.

- **Insulina**

El nivel de insulina en la sangre es un indicador crítico en la medicina, especialmente en el contexto del diagnóstico y manejo de la diabetes. Pero en nuestro conjunto de datos faltaba un porcentaje demasiado alto y no quisimos sesgar el estudio, por lo que eliminamos estas variable.

- **Presión sanguinea**

La presión arterial es un indicador vital de la salud cardiovascular y tiene implicaciones significativas en diversas condiciones médicas, incluyendo enfermedades cardíacas, accidentes cerebrovasculares y la diabetes. En nuestro conjunto descubrimos que la mayoría de las personas con diabetes tenían una presión baja o normal.La razón por la que no la usamos es porque no tenía una correlación tan alta como las demás.

Por el análisis previo consideramos que las variables más relevantes para nuestro estudio son Edad, Embarazos, Indice de Masa Corporal y Glucosa.

5. Gráficas

5.1. Matriz de Correlación

5.1.1. ¿Qué son las matrices de correlación?

- Las matrices de correlación son herramientas matemáticas que se utilizan para resumir la fuerza y la dirección de la relación lineal entre dos variables. Se representan en forma de tabla cuadrada, donde cada celda contiene un coeficiente de correlación que representa la correlación entre dos variables específicas.

5.1.2. ¿Cómo se interpretan las matrices de correlación?

Los valores en las matrices de correlación van de -1 a 1. Un valor de 1 indica una correlación positiva perfecta, lo que significa que las dos variables aumentan o disminuyen juntas en la misma proporción. Un valor de -1 indica una correlación negativa perfecta, lo que significa que a medida que una variable aumenta, la otra disminuye en la misma proporción. Un valor de 0 indica que no hay correlación lineal entre las variables.

5.1.3. ¿Cómo se leén las matrices de correlación?

Para leer una matriz de correlación, primero debes identificar las variables que se representan en cada fila y columna. Luego, busca el valor correspondiente a las dos variables que te interesan en la celda correspondiente. Este valor representa la fuerza y la dirección de la correlación lineal entre esas dos variables.

5.2. Correlograma

5.2.1. ¿Qué es un Correlograma?

Un correlograma (también llamado Gráfica ACF de función de correlación automática o Gráfica de autocorrelación) es una forma visual de mostrar la correlación serial en datos que cambian con el tiempo (es decir , datos de series temporales). La correlación en serie (también llamada autocorrelación) es donde un error en un punto en el tiempo viaja a un punto posterior en el tiempo.

5.2.2. ¿Cómo se leé un correlograma?:

Tamaño del Punto: El tamaño de los círculos en cada celda de la matriz refleja la magnitud de la correlación entre las variables correspondientes. Un círculo más grande sugiere una correlación más fuerte (ya sea positiva o negativa). Color del Punto: Los colores varían desde el azul hasta el rojo. El azul representa correlaciones negativas, mientras que el rojo indica correlaciones positivas. La intensidad del color se incrementa con la fuerza de la correlación.

5.3. Gráficas de Densidad

5.3.1. ¿Qué son las gráficas de Densidad?

Las gráficas de densidad, también conocidas como gráficos de densidad de kernel, son una herramienta visual utilizada en estadísticas para estimar la distribución de probabilidad de una variable numérica continua. Estos gráficos son especialmente útiles para visualizar la forma de la distribución de los datos, comparar distribuciones y observar características como la simetría, la asimetría, la presencia de modas y la dispersión.

5.3.2. ¿Cómo se leén?

Hay que identificar los componentes de la Gráfica de Densidad:
Eje X: Representa la variable para la que se estima la densidad.
Eje Y: Muestra la densidad estimada en cada punto, aunque no siempre se interpreta como una probabilidad directa. Es más bien un valor proporcional que depende del ancho de banda del kernel utilizado.
Curva: La línea en la gráfica que muestra la densidad en cada punto a lo largo del eje X.
Identificar Modas: La gráfica de densidad puede mostrar picos, conocidos como modas, que indican dónde se concentran los valores. Una distribución puede ser unimodal (un pico), bimodal (dos picos) o multimodal (múltiples picos).
Evaluar la Asimetría: Si la curva es asimétrica, puede indicar que la distribución de los datos es sesgada. Una curva que se extiende más hacia la derecha indica asimetría positiva (cola derecha más larga), mientras que una extensión hacia la izquierda indica asimetría negativa.
Comparar Distribuciones: Al superponer gráficas de densidad de diferentes grupos, puedes comparar visualmente sus distribuciones. Esto es útil para ver diferencias en la centralidad, dispersión y forma de las distribuciones.
Estimar Intervalos: Observando la amplitud de la curva, puedes estimar dónde cae la mayoría de los datos, lo que puede ser útil para determinar intervalos de confianza o identificar valores atípicos.

5.4. Gráficas de dispersión

5.4.1. ¿Qué son las gráficas de dispersión?

Los gráficos de dispersión, o diagramas de dispersión, son una herramienta fundamental en estadísticas y análisis de datos para visualizar y analizar la relación entre dos variables numéricas. Estos gráficos utilizan puntos cartesianos para representar los valores de las variables, permitiendo observar cómo una variable se comporta en relación con la otra, identificar tendencias, patrones, y posibles correlaciones.

5.4.2. ¿Cómo se leén los gráficos de dispersión?

Indentificamos los componentes de un Gráfico de Dispersión:

Eje X y Eje Y: Los ejes representan las variables que se están examinando. Por ejemplo, el eje X podría representar la edad de individuos, y el eje Y sus ingresos.
Puntos: Cada punto en el gráfico corresponde a un registro en el conjunto de datos, donde las coordenadas del punto representan los valores de las dos variables.

5.5. Gráficas de caja y bigotes (boxplots)

5.5.1. ¿Qué son las gráficas de caja y bicotes?

Las gráficas de caja y bigotes, conocidas también como diagramas de caja o boxplots, son un tipo de representación gráfica utilizada en estadística descriptiva. Estos gráficos ofrecen una visualización eficaz de la distribución de los datos numéricos a través de sus cuartiles y son útiles para identificar valores atípicos, la simetría de los datos, y comparar distribuciones entre varios grupos.

5.5.2. ¿Cómo se leén las gráficas de caja y bigotes?

Identificamos los componentes de una Gráfica de Caja y Bigotes:

Caja: Borde Inferior (Q1): Representa el primer cuartil, el cual contiene el 25 % de los datos.
Borde Superior (Q3): Representa el tercer cuartil, el cual contiene el 75 % de los datos.
Línea dentro de la caja (Mediana, Q2): Divide los datos en dos mitades iguales.
La mediana es el valor que se encuentra justo en el medio del conjunto de datos cuando están ordenados.
Bigotes:

Los bigotes de la caja se extienden desde el primer y tercer cuartil hasta el valor mínimo y máximo dentro de un intervalo calculado (normalmente 1.5 veces el rango intercuartílico por encima de Q3 y por debajo de Q1). Estos representan la variabilidad fuera de los cuartiles centralizados.

Puntos (Valores Atípicos):

Los puntos que se encuentran fuera de los bigotes son considerados valores atípicos y son destacados como puntos individuales en el gráfico.

Comparar Mediana: Observar la línea de la mediana en cada caja para comparar la centralidad de las distribuciones entre grupos o a través del tiempo.

Evaluación de la Dispersión: La altura de la caja (la distancia entre Q1 y Q3) indica el rango intercuartílico (IQR), que es una medida de la dispersión estadística. Una caja más alta indica una mayor variabilidad.
Simetría de los Datos: Si la mediana no está en el centro de la caja, sugiere que los datos son asimétricos. Si la línea de la mediana está cerca de Q1, los datos son sesgados hacia la derecha y viceversa.
Identificación de Valores Atípicos: Los puntos fuera de los bigotes indican valores atípicos que pueden necesitar atención especial o indicar variabilidad extrema en los datos.
Comparaciones entre Grupos: Al colocar múltiples boxplots lado a lado, se puede comparar rápidamente la distribución de los datos entre diferentes grupos o condiciones.

6. Resultados

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	3.298	110.649488	70.9124	30.859674	0.429734	31.19
mediana	2.0	107.5	71.8	30.4	0.336	27.0
std	3.017185	24.702421	11.928757	6.501303	0.299085	11.667655
var	9.103403	610.209589	142.295237	42.266937	0.089452	136.134168
ran	13	153.0	98.0	39.1	2.251	60
sesgo	positivo	positivo	negativo	positivo	positivo	positivo

Figura 10: Tabla descriptiva personas sanas

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	4.865672	142.325437	75.541045	35.406767	0.5505	37.067164
mediana	4.0	140.5	76.0	34.3	0.449	36.0
std	3.741239	29.488211	11.957564	6.590161	0.372354	10.968254
var	13.99687	869.554584	142.983328	43.430217	0.138648	120.302588
ran	17	121.0	84.0	44.2	2.332	49
sesgo	positivo	positivo	negativo	positivo	positivo	positivo

Figura 11: Tabla descriptiva personas con Diabetes

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	3.267943	87.971292	68.884211	30.632243	0.444823	29.354067
mediana	2.0	90.0	68.0	30.4	0.37	26.0
std	3.042168	10.143639	11.264401	6.604006	0.282394	9.303587
var	9.254785	102.893403	126.886721	43.612896	0.079746	86.556726
ran	14	56.0	98.0	36.8	1.615	47
sesgo	positivo	negativo	positivo	positivo	positivo	positivo

Figura 12: Tabla descriptiva Glucosa normal

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	3.267943	87.971292	68.884211	30.632243	0.444823	29.354067
mediana	2.0	90.0	68.0	30.4	0.37	26.0
std	3.042168	10.143639	11.264401	6.604006	0.282394	9.303587
var	9.254785	102.893403	126.886721	43.612896	0.079746	86.556726
ran	14	56.0	98.0	36.8	1.615	47
sesgo	positivo	negativo	positivo	positivo	positivo	positivo

Figura 13: Tabla descriptiva Glucosa prediabetes

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	2.413655	116.471374	69.799197	32.130289	0.474715	25.905622
mediana	2.0	111.290698	70.0	31.6	0.3775	25.0
std	2.24316	27.933376	11.913126	7.196041	0.341021	3.956647
var	5.031765	780.273481	141.922575	51.783008	0.116295	15.655059
ran	11	143.0	98.0	48.9	2.342	14
sesgo	positivo	positivo	negativo	positivo	positivo	positivo

Figura 14: Tabla descriptiva Edad Jovenes Adultos

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	6.708995	127.784217	76.827513	33.886772	0.451693	41.714286
mediana	7.0	126.0	76.0	33.3	0.341	41.0
std	3.551292	32.531224	10.787231	6.073791	0.312644	3.960289
var	12.611674	1058.280505	116.364345	36.890941	0.097747	15.683891
ran	17	153.0	62.0	31.5	1.696	14
sesgo	negativo	positivo	positivo	positivo	positivo	positivo

Figura 15: Tabla descriptiva Adultos mediana edad

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	6.147059	140.176471	79.220588	31.386864	0.503574	56.617647
mediana	6.0	138.0	78.0	31.6	0.4355	56.5
std	3.469546	30.221047	11.436842	6.222603	0.329213	4.253285
var	12.037752	913.311677	130.801361	38.720783	0.108381	18.09043
ran	13	107.0	60.0	26.9	1.32	14
sesgo	positivo	positivo	positivo	negativo	positivo	positivo

Figura 16: Tabla descriptiva Edad Adultos Mayores

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	3.845052	121.703075	72.527604	32.44642	0.471876	33.240885
mediana	3.0	117.0	72.0	32.05	0.3725	29.0
std	3.369578	30.462152	12.133545	6.87897	0.331329	11.760232
var	11.354056	927.942727	147.222914	47.320221	0.109779	138.303046
ran	17	155.0	98.0	48.9	2.342	60
sesgo	positivo	positivo	positivo	positivo	positivo	positivo

Figura 17: Tabla descriptiva Edad Ancianos

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	4.494673	121.483959	72.277017	31.994021	0.463604	34.193303
mediana	4.0	116.0	72.0	31.6	0.37	30.0
std	3.217291	30.652901	12.164722	6.435562	0.31385	11.818653
var	10.350962	939.600341	147.980462	41.416464	0.098502	139.680565
ran	16	155.0	98.0	39.1	2.245	60
sesgo	positivo	positivo	positivo	positivo	positivo	positivo

Figura 18: Tabla descriptiva personas que han tenido Embarazos

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	0.0	123.0	74.010811	35.124135	0.520838	27.603604
mediana	0.0	119.0	75.0	34.6	0.381	25.0
std	0.0	29.408719	11.893217	8.631368	0.418568	9.688118
var	0.0	864.872727	141.448609	74.500513	0.175199	93.859623
ran	0	141.0	70.0	48.7	2.342	46
sesgo	simétrico	positivo	negativo	positivo	positivo	positivo

Figura 19: Tabla descriptiva personas que no han tenido embarazos

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	0.75	95.25	69.5	18.25	0.413	24.0
mediana	1.0	97.0	69.0	18.2	0.4405	24.0
std	0.5	8.80814	5.0	0.1	0.228644	3.464102
var	0.25	77.583333	25.0	0.01	0.052278	12.0
ran	1	21.0	12.0	0.2	0.477	6
sesgo	negativo	negativo	positivo	positivo	negativo	simétrico

Figura 20: Tabla descriptiva IMC bajo

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	3.845052	121.703075	72.527604	32.44642	0.471876	33.240885
mediana	3.0	117.0	72.0	32.05	0.3725	29.0
std	3.369578	30.462152	12.133545	6.87897	0.331329	11.760232
var	11.354056	927.942727	147.222914	47.320221	0.109779	138.303046
ran	17	155.0	98.0	48.9	2.342	60
sesgo	positivo	positivo	positivo	positivo	positivo	positivo

Figura 21: Tabla descriptiva IMC Normal

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	3.845052	121.703075	72.527604	32.44642	0.471876	33.240885
mediana	3.0	117.0	72.0	32.05	0.3725	29.0
std	3.369578	30.462152	12.133545	6.87897	0.331329	11.760232
var	11.354056	927.942727	147.222914	47.320221	0.109779	138.303046
ran	17	155.0	98.0	48.9	2.342	60
sesgo	positivo	positivo	positivo	positivo	positivo	positivo

Figura 22: Tabla descriptiva IMC sobrepeso

	Embarazos	Glucosa	PresionSanguinea	IMC	FuncionPedigriDiabetes	Edad
media	4.016563	126.343268	74.567702	36.440892	0.497153	33.73499
mediana	3.0	123.0	74.0	35.2	0.396	30.0
std	3.551012	31.658787	11.906961	5.231215	0.348257	11.013956
var	12.609684	1002.27881	141.775718	27.365606	0.121283	121.307217
ran	17	142.0	84.0	37.1	2.335	49
sesgo	positivo	positivo	positivo	positivo	positivo	positivo

Figura 23: Tabla descriptiva IMC obesidad

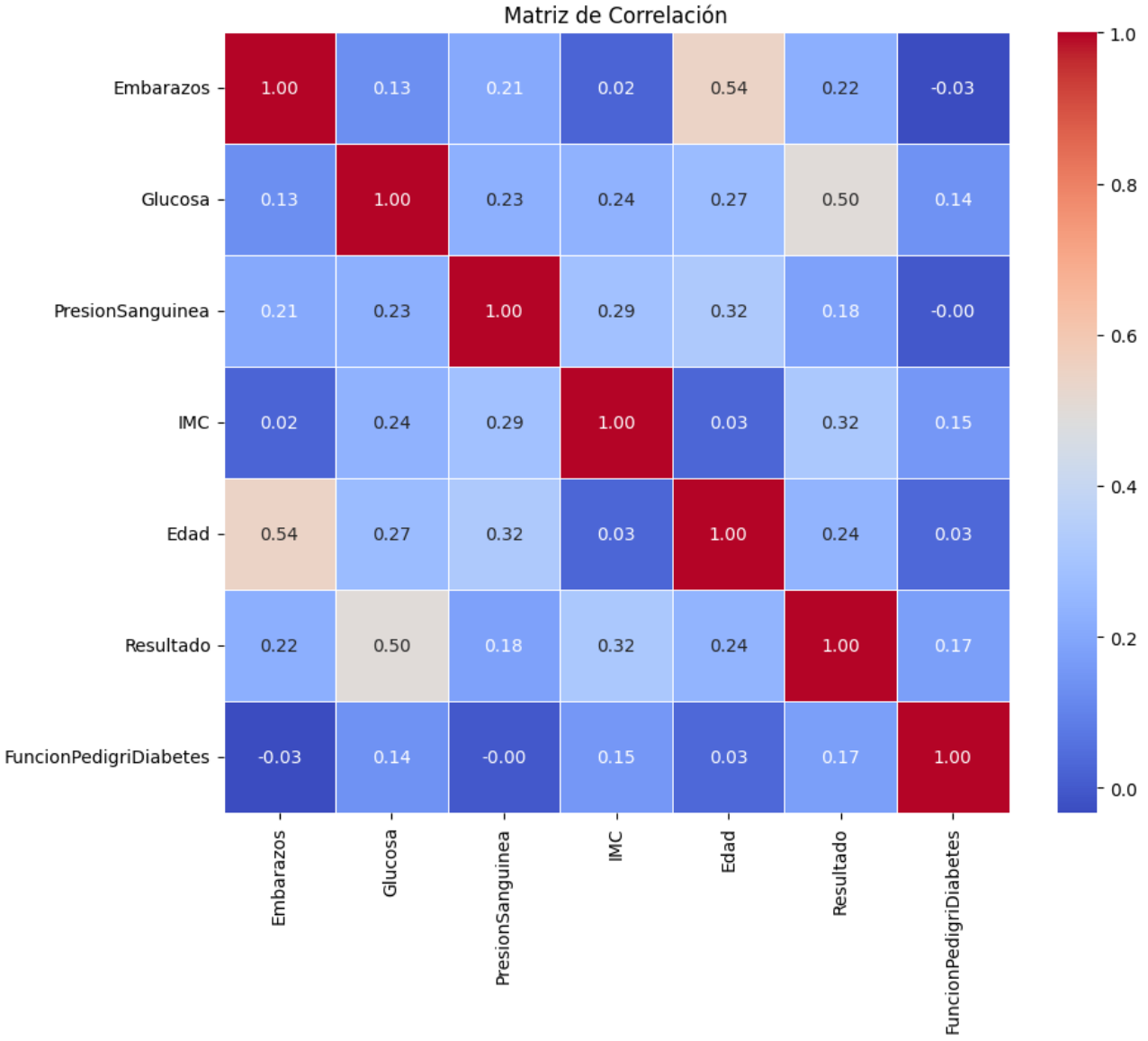
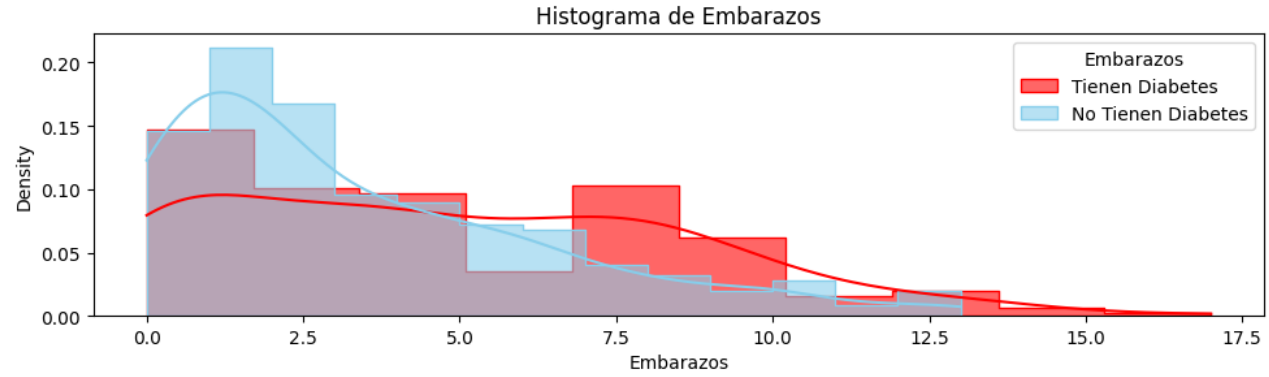
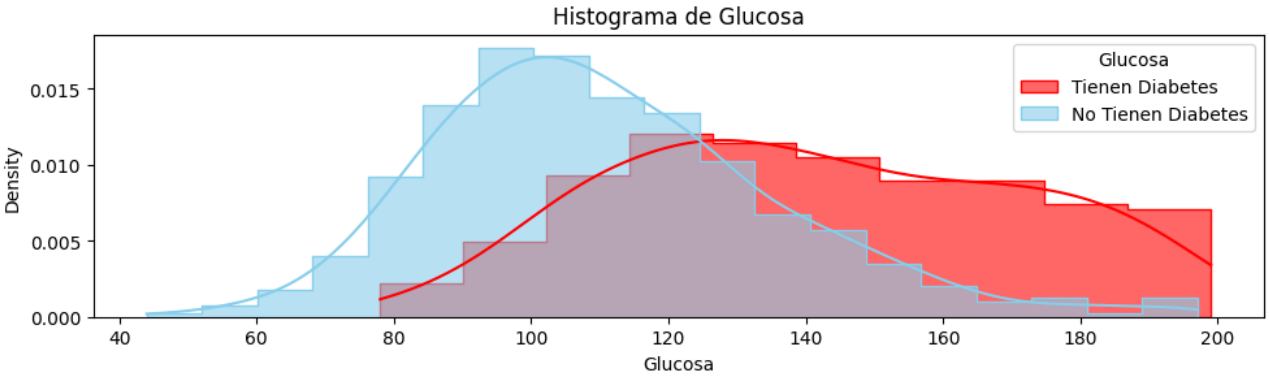
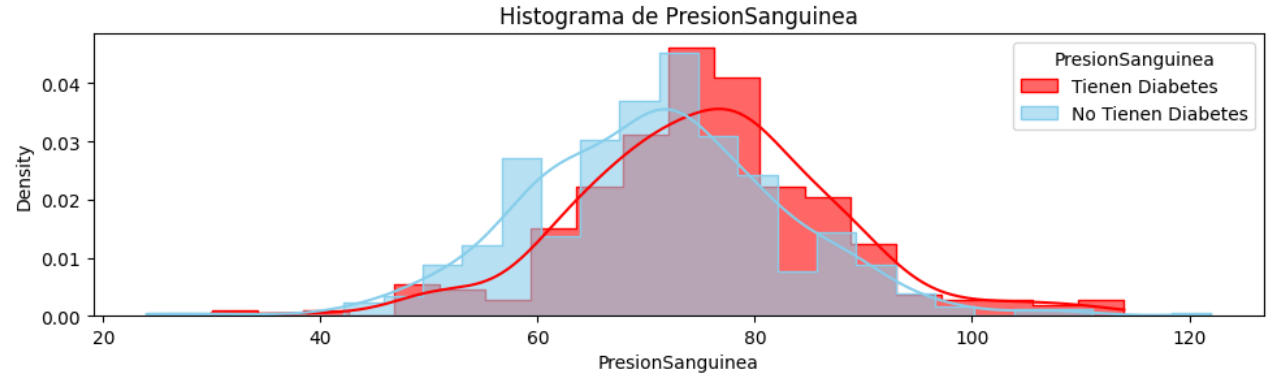


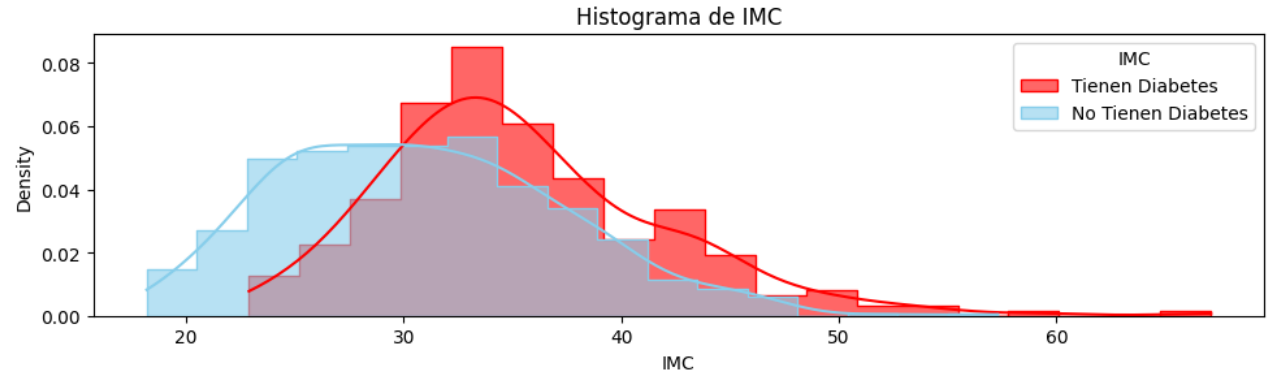
Figura 24: Matriz de correlación

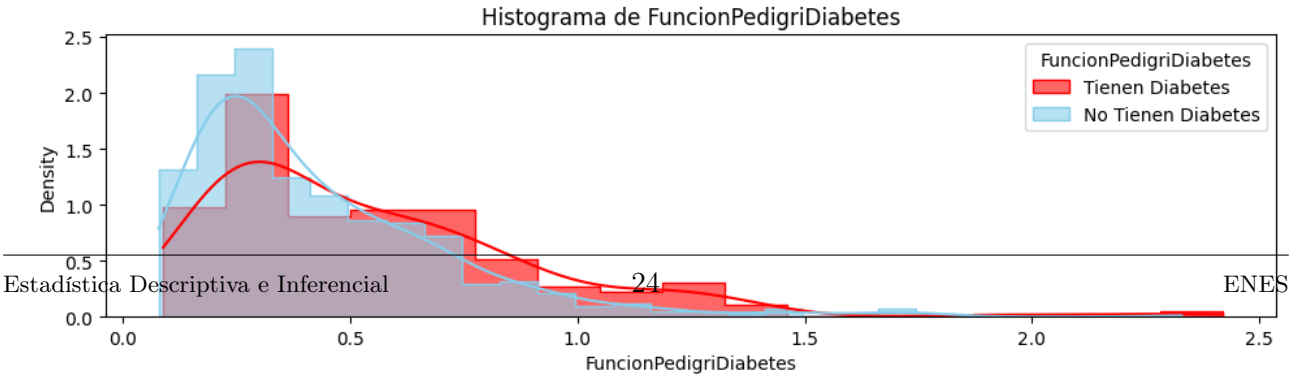
Figura 25: Correlograma


Histograma de Embarazos
Embarazos
Tienen Diabetes
No Tienen Diabetes
Density
Embarazos


Histograma de Glucosa
Glucosa
Tienen Diabetes
No Tienen Diabetes
Density
Glucosa


Histograma de PresionSanguinea
PresionSanguinea
Tienen Diabetes
No Tienen Diabetes
Density
PresionSanguinea


Histograma de IMC
IMC
Tienen Diabetes
No Tienen Diabetes
Density
IMC


Histograma de FuncionPedigriDiabetes
FuncionPedigriDiabetes
Tienen Diabetes
No Tienen Diabetes
Density
FuncionPedigriDiabetes


Histograma de Edad

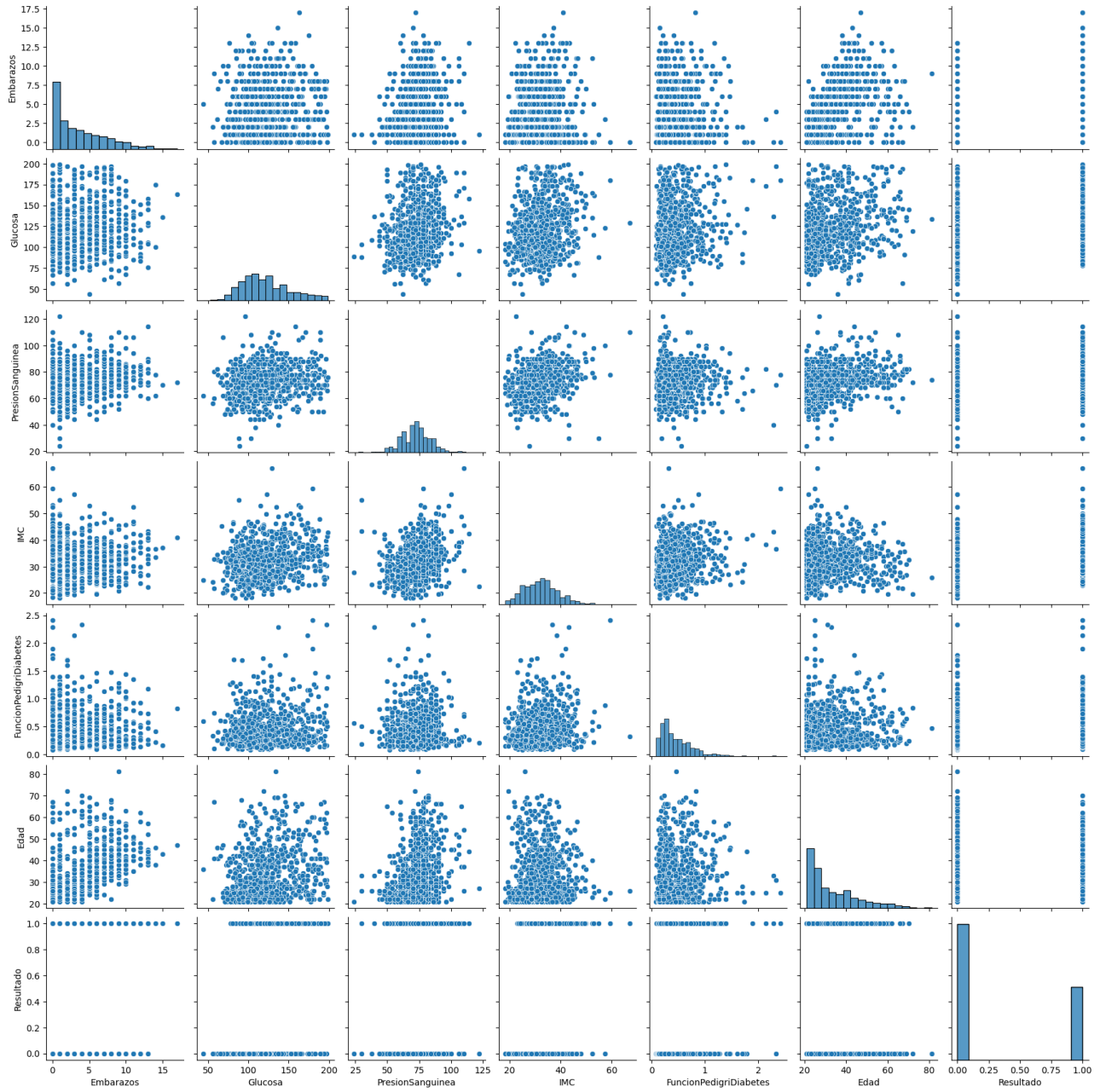


Figura 27: Dispersión con Hitogramas

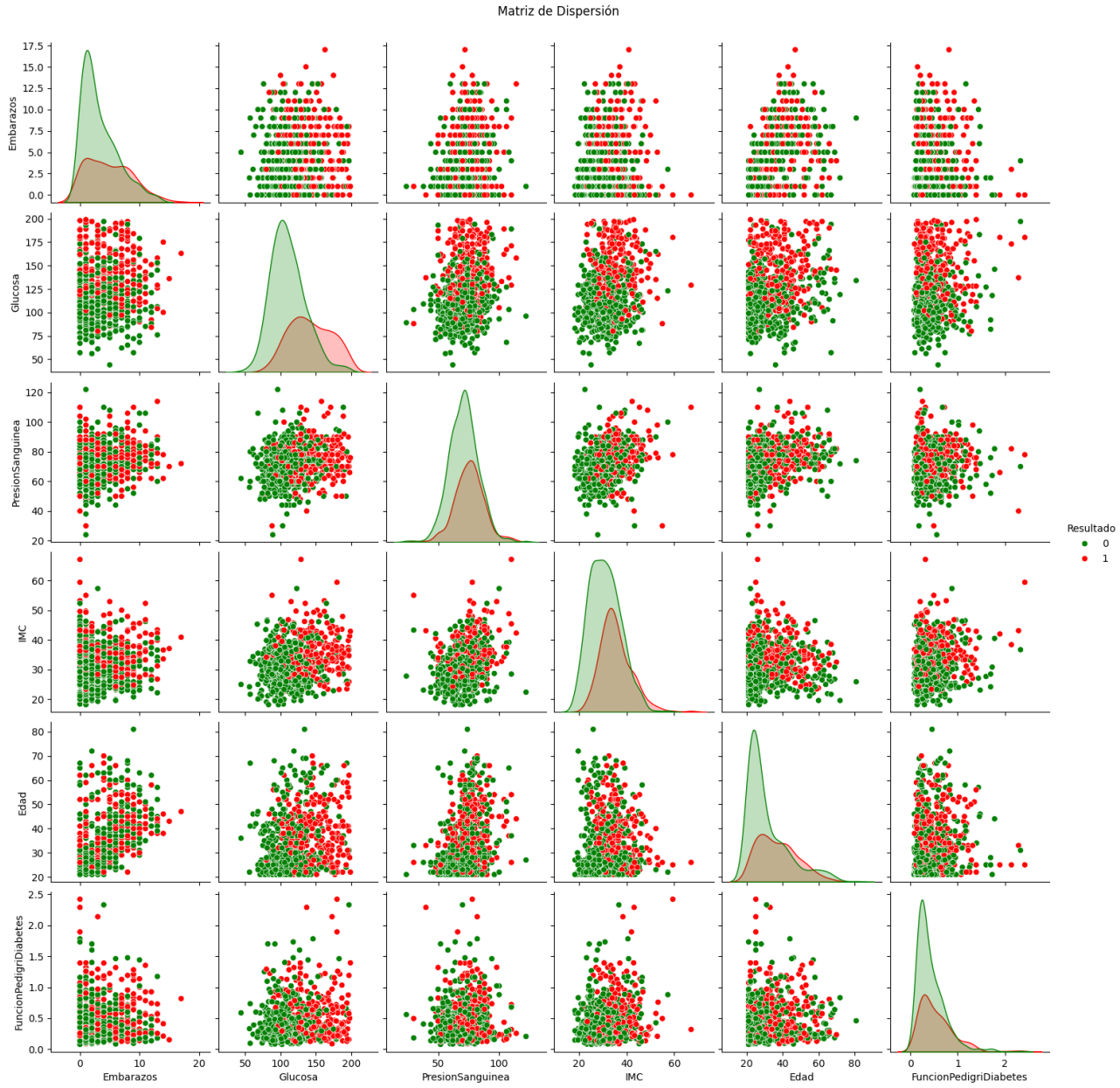


Figura 28: Matriz de dispersión personas sanas verdes, personas con dabetes rojas

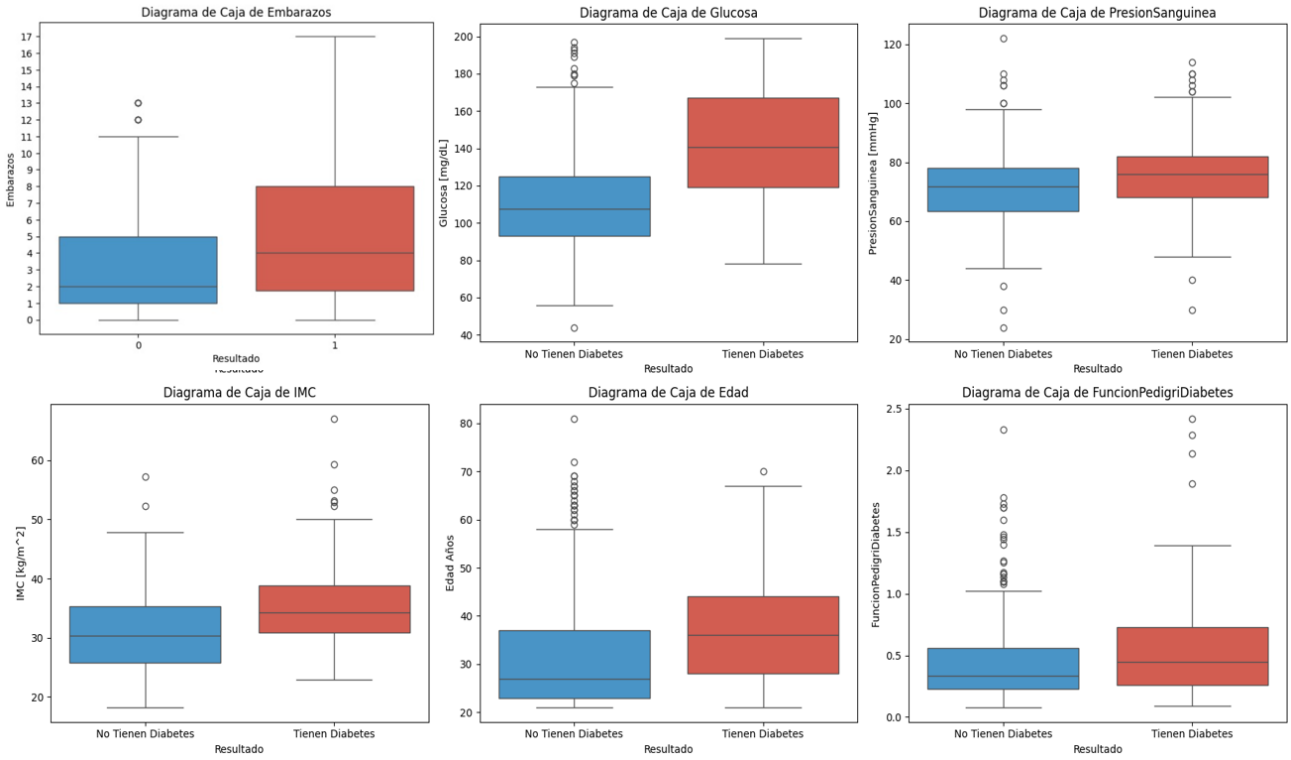


Figura 29: Gráficas de caja y bigotes (boxplots)

7. Discusión

Interpretación de resultados:

7.1. Estadística Descriptiva

7.1.1. Tablas descriptivas de personas con diabetes y sin diabetes

Al hacer las medidas de tendencia central y asimetría de todos los datos en general, observamos que las estadísticas sugieren diferencias significativas entre los individuos con y sin diabetes en términos de glucosa, presión sanguínea, IMC, y edad. Los pacientes con diabetes tienden a tener valores más altos en promedio, lo que es consistente con los conocimientos médicos sobre estas condiciones, y el sesgo en varias de las variables indica que, aunque los valores medios puedan parecer normales, existen variabilidades y extremos que podrían ser clínicamente importantes.

7.1.2. Tabla descriptiva de Glusosa Estratificada

Estas tablas descriptivas muestran las estadísticas de varias variables biométricas y clínicas, agrupadas por categorías de glucosa: normal, prediabetes y diabetes. Cada tabla proporciona la media, mediana, desviación estándar (std), varianza (var), rango (ran) y sesgo de las variables. Aquí está el análisis de cada grupo:

1. Glucosa Normal
Glucosa Media: 87.97 mg/dL, indicando niveles normales de glucosa.
Presión Sanguínea Media: 68.84 mmHg, dentro de rangos normales, sugiriendo ausencia de hipertensión.
IMC Media: 30.63, lo cual es en el límite superior del sobrepeso según los estándares de IMC.
Función Pedigrí de Diabetes Media: 0.444, relativamente baja, indicando menor predisposición genética a la diabetes.
Edad Media: 29.35 años, lo cual sugiere una población más joven.

2. Glucosa Prediabetes
Glucosa Media: 112.86 mg/dL, que está en el rango de prediabetes.
Presión Sanguínea Media: 71.24 mmHg, ligeramente superior comparado con el grupo de glucosa normal.
IMC Media: 31.74, que cae en la categoría de obesidad.
Función Pedigrí de Diabetes Media: 0.436, similar al grupo anterior, indicando una predisposición genética comparable.
Edad Media: 32.07 años, ligeramente mayor que el grupo de glucosa normal.

3. Glucosa Diabetes
Glucosa Media: 152.96 mg/dL, claramente en el rango de diabetes.
Presión Sanguínea Media: 76.19 mmHg, la más alta entre los tres grupos, mostrando un riesgo incrementado de hipertensión.
IMC Media: 34.33, que indica obesidad .
Función Pedigrí de Diabetes Media: 0.523, la más alta entre los grupos, reflejando una mayor predisposición genética a la diabetes.
Edad Media: 36.97 años, siendo el grupo de mayor edad en comparación con los otros dos.

Hay una progresión clara en los niveles de glucosa, presión sanguínea, IMC, y edad a medida que se avanza de glucosa normal a diabetes. Esto sugiere que estos factores están correlacionados con el empeoramiento del estado metabólico que es consecuencia de la diabetes.
Todos los grupos muestran un sesgo positivo en varias medidas, lo que indica que los valores tienden a estar más distribuidos hacia el extremo superior de la escala.
El aumento en la edad y el IMC con el deterioro del estado glucémico subraya la importancia de la gestión del peso y la monitorización de la glucosa como medidas preventivas contra la diabetes.
Aunque la función de pedigrí incrementa con peores estados glucémicos, su moderada magnitud sugiere que otros factores, como el estilo de vida y la dieta, también son cruciales para la prevención y manejo de la diabetes.

7.1.3. Tabla descriptiva de Edad Estratificada

Jóvenes Adultos Embarazos: La media de embarazos es baja, lo cual es esperado en un grupo más joven.
Glucosa: La media está cerca del límite superior de lo normal, lo cual podría indicar un riesgo elevado de prediabetes en este grupo.
Presión Sanguínea: La media está dentro de rangos normales.

IMC: La media está en el rango de obesidad, lo cual es preocupante para la salud a largo plazo.
Función de Pedigrí de Diabetes: Moderada, lo que sugiere una predisposición genética a la diabetes en este grupo.
Edad: Promedio aproximado de 25 años.

Adultos de Mediana Edad Embarazos: Mayor número de embarazos en promedio, acorde con la edad.
Glucosa: Elevada significativamente, promediando en niveles de prediabetes.
Presión Sanguínea: Ligeramente más alta que en los jóvenes adultos, pero aún dentro de un rango normal.
IMC: Promedio también en el rango de obesidad.
Función de Pedigrí de Diabetes: Similar a la de los jóvenes adultos, indicando un riesgo genético continuo.
Edad: Promedio aproximado de 41 años.

Adultos Mayores Embarazos: El número de embarazos es más alto, lo cual refleja la acumulación de eventos reproductivos a lo largo de la vida.
Glucosa: La más alta entre los grupos, claramente en el rango diabético.
Presión Sanguínea: Elevada, posiblemente indicando hipertensión, común en este grupo de edad.
IMC: Similar a los otros grupos, también en el rango de obesidad.
Función de Pedigrí de Diabetes: La más alta, indicando la mayor predisposición genética. Edad: Promedio aproximado de 56 años.

A medida que aumenta la edad, también lo hacen los niveles de glucosa y la presión sanguínea, ambos indicadores clave del deterioro metabólico asociado con el riesgo de diabetes.
El IMC elevado en todos los grupos etarios es una preocupación constante, subrayando la necesidad de intervenciones en dieta y ejercicio.
La función de pedigrí de diabetes sugiere un componente hereditario significativo que podría ser crucial para la intervención temprana.

7.1.4. Tabla descriptiva de Embarazos Estratificada

Personas que Han Tenido Embarazos

Embarazos: La media de embarazos es aproximadamente 4.5, lo que es esperado dado el criterio de inclusión para este grupo.
Glucosa Media: 121.48 mg/dL, lo cual sugiere un nivel elevado de glucosa y un posible riesgo de diabetes.
Presión Sanguínea Media: 72.27 mmHg, dentro de rangos normales.
IMC Media: 31.99, indicativo de obesidad, aumentando el riesgo de complicaciones metabólicas incluyendo diabetes.
Función Pedigrí de Diabetes: 0.46, lo cual sugiere una predisposición genética moderada a la diabetes.

Edad Media: 34 años, indicando que este grupo está en sus años de mediana edad, un factor de riesgo para el desarrollo de diabetes tipo 2.

Personas que No Han Tenido Embarazos

Embarazos: Como se esperaría, el número de embarazos es 0.
Glucosa Media: 123.0 mg/dL, ligeramente más alta que el grupo que ha tenido embarazos, también indicando un riesgo elevado de diabetes.
Presión Sanguínea Media: 74.01 mmHg, también dentro de lo normal pero ligeramente más alta que el grupo anterior.
IMC Media: 35.12, significativamente más alto y cae en la categoría de obesidad grado II, lo cual es una gran preocupación para riesgos de salud relacionados.
Función Pedigrí de Diabetes: 0.52, más alta que en el grupo que ha tenido embarazos, reflejando una predisposición genética aún más fuerte.
Edad Media: 27 años, considerablemente más joven que el grupo que ha tenido embarazos.

Ambos grupos muestran niveles elevados de glucosa y un IMC alto, dos indicadores clave de riesgo para la diabetes.
A pesar de ser más jóvenes, las personas sin embarazos muestran un mayor riesgo genético y un IMC más alto, lo cual podría ser indicativo de peores hábitos de salud o de una predisposición genética más fuerte hacia problemas metabólicos.
El hecho de haber tenido embarazos no parece influir significativamente en los niveles de glucosa entre los grupos, pero sí se observa un IMC más bajo y una edad promedio más alta en aquellos que han tenido embarazos. Esto podría ser indicativo de diferentes etapas de vida y responsabilidades que podrían influir en la salud general.

7.1.5. Tabla descriptiva de Índice de Masa Corporal Estratificado

1. IMC Bajo
Glucosa: La media es 95.25 mg/dL, dentro de un rango normal pero cercano al límite superior.
Presión Sanguínea: Promedio de 69.5 mmHg, indicando valores normales.
IMC: Promedio extremadamente bajo de 18.25, lo que podría indicar desnutrición o una constitución corporal naturalmente delgada.
Función Pedigrí de Diabetes: Relativamente alta (0.413) para un IMC bajo, sugiriendo una predisposición genética significativa.
Edad: Joven, con un promedio de 24 años.
2. IMC Normal
Glucosa: 121.7 mg/dL, que es bastante elevado para estar en la categoría de IMC normal.
Presión Sanguínea: 72.6 mmHg, dentro de un rango saludable.
IMC: Promedio de 32.05, que en realidad está en el rango de obesidad según los estándares generales, lo que indica un posible error en la categorización o un error tipográfico en la tabla.
Función Pedigrí de Diabetes: 0.4718, indicando un riesgo genético moderado.
Edad: Media de 33 años.
3. IMC Sobrepeso
Glucosa: Promedio alto de 126.34 mg/dL, indicando un riesgo de prediabetes.

Presión Sanguínea: Alta, con un promedio de 74.6 mmHg
. IMC: 36.44, clasificado incorrectamente y debería estar en la categoría de obesidad.
Función Pedigrí de Diabetes: 0.4971, la más alta entre los grupos, resaltando un riesgo genético significativo. Edad: Promedio de 33 años, similar al grupo con IMC normal.
4. IMC Obesidad
Glucosa: Muy alta, con un promedio de 126.34 mg/dL, reflejando un riesgo significativo de diabetes.
Presión Sanguínea: Alta, promediando 74.6 mmHg.
IMC: Extremadamente alto con un promedio de 36.44.
Función Pedigrí de Diabetes: Alta, con un promedio de 0.4971.
Edad: Promedio de 33 años.

Las clasificaciones de IMC parecen estar erróneamente asignadas en las tablas, ya que los valores de IMC para los grupos 'Normal' y 'Sobrepeso' están claramente en el rango de obesidad. Esto podría deberse a un error de etiquetado en las tablas.
Los niveles de glucosa están elevados en todos los grupos, indicando un riesgo generalizado de trastornos metabólicos en la población estudiada.
La presión sanguínea aumenta ligeramente con el aumento del IMC, lo que es consistente con el riesgo cardiovascular asociado con un mayor peso corporal.
La función de pedigrí de diabetes es relativamente alta en todos los grupos, sugiriendo una fuerte influencia genética en el desarrollo de diabetes en esta población.

7.2. Matriz de correlación

Correlación entre Embarazos y Edad:

Observamos una correlación moderada a fuerte entre el número de embarazos y la edad (aproximadamente 0.54). Esto indica que las mujeres mayores tienden a haber tenido más embarazos, lo cual es un resultado esperado.
Correlación entre Glucosa y Resultado:

Hay una correlación significativa (aproximadamente 0.50) entre los niveles de glucosa y el resultado de diabetes. Esto confirma que la glucosa es un predictor crucial para la diabetes, dado que niveles altos son indicativos de la enfermedad.
Correlación entre Presión Sanguínea e IMC:

Existe una correlación moderada (aproximadamente 0.29) entre la presión sanguínea y el IMC. Esto sugiere que un IMC más alto puede estar asociado con una presión sanguínea más elevada, un factor de riesgo conocido para muchas condiciones de salud, incluida la diabetes.
Correlación entre Edad y Resultado:

La edad muestra una correlación moderada (aproximadamente 0.24) con el resultado de diabetes. Esto puede indicar que el riesgo de desarrollar diabetes aumenta con la edad.

Función Pedigrí de Diabetes:

La función de pedigrí de diabetes muestra correlaciones bajas con todas las otras variables y con el resultado, lo que sugiere que mientras tiene algún impacto genético sobre la predisposición a la diabetes, no es un predictor tan fuerte como otros factores como la glucosa.

7.3. Correlograma

Correlaciones Clave:

Embarazos y Edad: Una correlación positiva moderada indicada por un círculo de tamaño mediano y color rojo. Esto implica que a medida que aumenta la edad, también tiende a aumentar el número de embarazos.
Glucosa y Resultado: Un círculo grande y rojo muestra una fuerte correlación positiva, lo que indica que niveles más altos de glucosa están significativamente asociados con la presencia de diabetes.
Presión Sanguínea e IMC: El tamaño mediano y el color rojo sugieren una correlación positiva moderada, indicando que un IMC más alto podría estar relacionado con una presión arterial más alta.

Menores Correlaciones:

Función de Pedigrí de Diabetes: Los círculos pequeños y de color menos intenso en relación con esta variable indican correlaciones bajas con otras métricas clínicas y biométricas.

7.4. Gráficas de Densidad

textbfHistograma de Embarazos:

Las personas con diabetes tienden a tener un mayor número de embarazos en comparación con las que no tienen diabetes,esto podría sugerir que un mayor número de embarazos está asociado con un riesgo incrementado de diabetes, posiblemente debido a factores como la diabetes gestacional o cambios hormonales y metabólicos relacionados con el embarazo.

Histograma de Glucosa:

La densidad de personas con diabetes muestra concentraciones más altas de glucosa comparado con quienes no tienen diabetes, esta es una representación visual de cómo los niveles elevados de glucosa en sangre están directamente relacionados con la diabetes, reafirmando la glucosa como un marcador principal para el diagnóstico de la diabetes.

Histograma de Presión Sanguínea:

Los individuos con diabetes parecen tener en general una distribución más amplia y ligeramente superior en los niveles de presión arterial comparado con aquellos sin diabetes. La hipertensión es común en personas con diabetes, lo cual puede ser reflejo de una relación entre la resistencia a la insulina y el aumento de la presión arterial.

Histograma de IMC:

Las personas con diabetes muestran un pico más alto y amplio en el rango de IMC más elevado, indicando una mayor prevalencia de sobrepeso u obesidad.
Esto apoya la noción de que un IMC alto es un factor de riesgo significativo para la diabetes, probablemente debido a la asociación entre el exceso de grasa corporal y la resistencia a la insulina.

Histograma de Función Pedigrí de Diabetes:

La densidad para aquellos con diabetes es ligeramente mayor para valores más altos de la función de pedigrí, aunque las distribuciones son relativamente similares. Aunque hay una tendencia a valores más altos de pedigrí en diabéticos, lo que sugiere una influencia genética, la superposición considerable indica que por sí sola, esta medida no puede predecir con precisión el riesgo de diabetes.

Histograma de Edad:

Las personas con diabetes tienden a ser mayores en comparación con aquellas sin diabetes, mostrando una distribución que se desplaza hacia la derecha. IEste gráfico subraya la relación entre la edad avanzada y el riesgo incrementado de desarrollar diabetes, probablemente debido a cambios metabólicos asociados con el envejecimiento y la acumulación de factores de riesgo a lo largo del tiempo.

7.5. Gráficas de dispersión

Histogramas (Diagonal):

Cada histograma en la diagonal muestra la distribución de una variable específica, como .ᴱᵐbarazos", "Glucosa", "PresionSanguinea", etc. Por ejemplo, el histograma de .ᴱᵐbarazos"muestra que muchos registros tienen pocos embarazos, y pocos registros tienen muchos embarazos, indicando una distribución sesgada a la derecha.

Gráficos de Dispersión (Fuera de la Diagonal):

Embarazos y Edad: Existe una tendencia de que a mayor número de embarazos, mayor es la edad de la persona, lo que tiene sentido biológicamente.
Glucosa y Resultado: Se puede observar que altos niveles de glucosa están asociados con el resultado positivo de diabetes (Resultado = 1), lo que es consistente con la patología de la diabetes.

Presión Sanguínea e IMC: Existe una agrupación que indica una relación positiva entre la presión sanguínea y el IMC, sugiriendo que un IMC más alto puede estar asociado con una presión sanguínea más alta.

Función de Pedigrí de Diabetes y Resultado: No parece haber una correlación clara en el gráfico, lo que sugiere que la función de pedigrí por sí sola no es un fuerte predictor del resultado de diabetes, aunque hay una ligera tendencia a tener valores más altos de función de pedigrí en los diabéticos.

Edad y Resultado: La edad parece estar distribuida de manera relativamente uniforme entre los resultados de diabetes positivos y negativos, aunque hay una concentración ligeramente mayor de casos positivos en edades más avanzadas.

Histogramas y Kernel Density Estimation (Diagonal):

Cada histograma muestra la distribución de una variable, con el KDE superpuesto para suavizar la distribución. Esto permite visualizar la forma general de la distribución de cada variable, facilitando la identificación de sesgos, picos y la distribución general.

Notablemente, las distribuciones para variables como la glucosa muestran diferencias claras entre los dos grupos, donde los individuos con diabetes tienden a tener valores más altos.

Gráficos de Dispersión (Fuera de la Diagonal):

Cada gráfico de dispersión compara dos variables, con puntos en rojo representando a los diabéticos y puntos en verde a los no diabéticos. Estos gráficos son cruciales para visualizar la relación entre pares de variables y cómo cada grupo se agrupa dentro de estas relaciones.
Por ejemplo, la relación entre la glucosa y la presión sanguínea muestra una dispersión que sugiere que niveles más altos de glucosa están frecuentemente asociados con niveles más altos de presión sanguínea, especialmente en los individuos con diabetes.

Diferencias entre Grupos: Hay claras diferencias en cómo se distribuyen ciertas variables entre aquellos con y sin diabetes. Variables como glucosa, presión sanguínea, y BMI (IMC) muestran patrones distintivos que podrían ser utilizados para identificar riesgos de diabetes o para segmentar intervenciones.

Correlaciones entre Variables: Los gráficos de dispersión muestran relaciones que pueden ser claves para entender cómo diferentes factores de salud interactúan entre sí en el contexto de la diabetes. Por ejemplo, la relación entre la glucosa y la presión sanguínea podría ser indicativa de cómo el manejo de una podría influir en la otra.

Implicaciones para la Prevención y el Manejo: Conocer estas correlaciones y distribuciones ayuda a informar estrategias para la prevención y manejo de la diabetes. Intervenciones dirigidas a reducir la glucosa y la presión sanguínea podrían ser efectivas para controlar o prevenir la diabetes en poblaciones de riesgo.

Identificación de Factores de Riesgo: Estos gráficos pueden ayudar a identificar factores de riesgo importantes que están fuertemente asociados con la diabetes, proporcionando un fundamento para pruebas diagnósticas o programas de intervención basados en la evidencia observada.

7.6. Gráficas de caja y bigotes

Cada boxplot compara la distribución de los valores para una variable específica entre los dos grupos, ofreciendo una visión clara sobre cómo se diferencian.

Diagrama de Caja de Embarazos

Los individuos que tienen diabetes tienden a tener un mayor número de embarazos en comparación con aquellos que no tienen diabetes. Este resultado puede ser indicativo de un vínculo entre la frecuencia de embarazos y el riesgo de desarrollar diabetes, probablemente debido a la relación entre diabetes gestacional y diabetes tipo 2.

Diagrama de Caja de Glucosa

Los niveles de glucosa son notablemente más altos en personas con diabetes. Esta es una distinción clara que refleja cómo la hiperglucemia es un marcador central de la diabetes.

Diagrama de Caja de Presión Sanguínea

Hay una diferencia leve en la presión sanguínea entre los dos grupos, con los individuos diabéticos mostrando niveles ligeramente más altos. Esto sugiere una posible correlación entre la hipertensión y la diabetes, lo cual es conocido en la literatura médica como parte del síndrome metabólico.

Diagrama de Caja de IMC

Los individuos con diabetes tienen un IMC más alto en promedio, indicando una relación entre la obesidad o sobrepeso y la diabetes. Esto está en línea con la comprensión de que un IMC elevado es un factor de riesgo para el desarrollo de diabetes tipo 2.

Diagrama de Caja de Función Pedigrí de Diabetes

La función de pedigrí de diabetes es ligeramente mayor en promedio en aquellos con diabetes. Este es un indicador genético que refleja la predisposición hereditaria a desarrollar la enfermedad.

Diagrama de Caja de Edad

La distribución de la edad muestra que los individuos con diabetes son generalmente mayores. Esto refleja el aumento del riesgo de diabetes con la edad y está en línea con la epidemiología de la diabetes tipo 2, que frecuentemente se desarrolla más tarde en la vida.

8. Conclusión

Las conclusiones que podemos extraer de nuestro análisis son las siguientes:

- 1. Existe una relación entre el Índice de Masa Corporal (IMC) y la edad en el desarrollo de la diabetes.
- 2. La presión sanguínea no parece influir significativamente en la obtención de diabetes, ya que la mayoría de los diabéticos presentan presión sanguínea baja o normal.
- 3. El índice de pedigrí no tiene un impacto considerable, ya que muchas personas con un índice de pedigrí alto no padecen la enfermedad.
- 4. Podemos concluir que el cuidado personal, incluyendo hábitos y estilo de vida, tiene más influencia en el desarrollo de la diabetes que la herencia genética.
- 5. Como se predecía, la mayoría de los individuos con diabetes presentan sobrepeso u obesidad.

En esta sección, hemos analizado la pertinencia de la base de datos y los métodos empleados en nuestras clases y aprendizaje de esta materia. La base de datos utilizada se mostró adecuada para los objetivos del estudio, proporcionándonos datos relevantes y fiables. Sin embargo, reconocemos que la elección de una base de datos más extensa o diversificada podría haber enriquecido los resultados y proporcionado una perspectiva más amplia.

En cuanto a los métodos empleados, estos fueron seleccionados por su idoneidad para el análisis específico que realizamos. No obstante, somos conscientes de que la implementación de técnicas adicionales, como el uso de algoritmos de machine learning más avanzados, podría haber mejorado la profundidad y precisión de nuestros hallazgos.

Una de las limitaciones que enfrentamos fue el tiempo disponible para la realización del trabajo. Este factor influyó en la capacidad para explorar métodos alternativos y en la extensión del análisis. A pesar de estas limitaciones, consideramos que el trabajo realizado ofrece una visión valiosa y fundamentada sobre el tema investigado.

Creemos que herramientas que usamos como software de análisis estadístico ayudo a la visualización de datos añadiendo un valor significativo a nuestro estudio. Estas herramientas facilitaron para una comprensión más clara y una presentación más efectiva de los resultados.

En cuanto al desempeño, consideramos que nuestra comprensión y aplicación de los conceptos aprendidos durante el curso fueron satisfactorios. Sin embargo, reconocemos que hay áreas en las que podríamos mejorar, especialmente en la integración de metodologías más innovadoras y en la interpretación crítica de los resultados obtenidos.

Este trabajo ha creado expectativas respecto a los posibles usos de los métodos empleados. Por ejemplo, la metodología podría aplicarse en otros contextos o con diferentes conjuntos de datos para evaluar su efectividad y adaptabilidad. Además, este estudio nos ha proporcionado una base sólida para futuros trabajos, donde podremos aplicar los aprendizajes y mejorar nuestras técnicas de investigación.

En conclusión, aunque estamos satisfechos con los resultados obtenidos y con nuestro desempeño, identificamos áreas de mejora y oportunidades para enriquecer futuros trabajos mediante la inclusión de nuevas herramientas y métodos. Este proceso de reflexión crítica nos permite reconocer nuestras fortalezas y debilidades, preparándonos para abordar proyectos futuros con una perspectiva más amplia y un enfoque más robusto.

9. Referencias

Referencias

[1] Apolonio. (2024). *Todo lo que necesitas saber sobre el diagrama de caja y bigotes: conceptos clave y ejemplos prácticos*. Disponible en: https://apolonio.es/diagrama-de-caja-y-bigotes/ [Accedido: 28 Mayo 2024].

[2] Data Viz Catalogue. (2024). *Gráfico de Densidad*. Disponible en: https://datavizcatalogue.com/methods/density_plot.html [Accedido: 28 Mayo 2024].

[3] (2022, 6 octubre). Kaggle. Disponible en: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download [Accedido: 28 Mayo 2024].

[4] Khan Academy. (2024). *Correlation coefficient intuition examples*. Disponible en: https://es.khanacademy.org/math/ap-statistics/bivariate-data-ap/correlation-coefficient-r/v/correlation-coefficient-intuition-examples [Accedido: 28 Mayo 2024].

[5] Tu Dashboard. (2024). *Gráfica de dispersión. Qué es y cuáles son sus características*. Disponible en: https://tudashboard.com/grafica-de-dispersion/ [Accedido: 28 Mayo 2024].

[6] Wikipedia. (2024). *Matriz de correlación.* Disponible en: https://es.wikipedia.org/wiki/Matriz_de_correlaci%C3%B3n [Accedido: 28 Mayo 2024].

[7] YouTube. (2024). *R | Gráfico de correlaciones (correlograma / corrplot).* Disponible en: https://m.youtube.com/watch?v=LJ6SX7PBreg [Accedido: 28 Mayo 2024].