Análisis de Datos de Diabetes

García, Juan Manuel
JuanMBriones@outlook.com

Herrera, José Emiliano
eherrera1331@gmail.com

Miramontes, Fred
liloliol@hotmail.com

Román, Christopher
ferrobin34@gmail.com

Sánchez, Gabriel
178294@iberopuebla.mx

Sánchez, Ludim
ludim.anel@gmail.com

13 de agosto de 2021

Resumen

En esta práctica analizamos los datos de un dataset con registros de mujeres de la india con o sin diabetes `https://www.kaggle.com/mathchi/diabetes-data-set` recopilados del *National Institute of Diabetes and Digestive and Kidney Diseases*. Posteriormente se limpiaron los datos y se realizó un análisis exploratorio de los mismos. Se observó una fuerte relación entre la concentración de glucosa en la sangre con la probabilidad de tener diabetes, así como el peso de una mujer y la probabilidad de ccontraer dicha enfermedad. Por otro lado, se observó que ésta probabilidad aumenta cuando la mujer no ha estado embarazada. Otra observación importante se da cuando buscamos relaciones entre los factores de riesgo, ya que podemos encontrar áreas donde, a pesar de ser muy dispersas, parece que las mujeres diagnosticadas con diabetes se encuentran de un lado y las que no tienen la enfermedad se encuentran del otro.

1. Introducción

La diabetes es una enfermedad crónica en la que el cuerpo no es capaz de regular la cantidad de glucosa en la sangre. La glucosa en la sangre es la principal fuente de energía y ésta proviene de los alimentos. Con el transcurso del tiempo, el exceso de glucosa en la sangre puede causar problemas de salud, tales como enfermedades del corazón, daño a los nervios y enfermedad de los riñones.

Los principales factores de riesgo de la diabetes son:

- **Peso:** Mientras más tejido graso, más resistentes serán las células a la insulina.
- **Inactividad:** Mientras menos actividad se realice, mayor será el riesgo. La actividad física ayuda a controlar el peso, utiliza la glucosa como energía y hace que tus células sean más sensibles a la insulina.
- **Antecedentes familiares:** El riesgo se incrementa si padres o hermanos tienen diabetes tipo 2.
- **Raza o grupo étnico:** Aunque no está claro por qué, personas de ciertos orígenes, como las personas negras, hispanas, los indígenas estadounidenses y asiático-americanas, corren un mayor riesgo.
- **Edad:** El riesgo aumenta con la edad. Esto puede deberse a que la actividad física es menor, se pierde masa muscular y se aumenta de peso a medida que envejeces. Sin embargo la diabetes tipo 2 también está aumentando entre los niños, los adolescentes y los adultos jóvenes.
- **Presión arterial alta:** Una presión arterial de más de 140/90 milímetros de mercurio (mm Hg) implica un alto riesgo de desarrollar diabetes tipo 2.
- **Niveles anormales de colesterol:** Si se cuenta con niveles bajos de lipoproteínas de alta densidad o de colesterol "bueno", el riesgo de desarrollar diabetes tipo 2 será mayor. Lo triglicéridos son otro tipo de grasas que se transportan en la sangre. Las personas con niveles altos de triglicéridos afrontan un riesgo elevado de padecer diabetes tipo 2.

Desde el año 2000, la diabetes mellitus en México es la primera causa de muerte entre las mujeres y la segunda entre los hombres. En 2010, esta enfermedad causó cerca

de 83,000 muertes en el país. Es por esto que decidimos hacer un análisis sobre este problema.

2. Objetivos

Teniendo en cuenta el contexto y la importancia del estudio de esta enfermedad se plantean las siguientes preguntas:

- ¿Existe alguna relación entre la diabetes y el grosor de la piel?
- ¿Existe alguna relación de la diabetes con la edad?
- Observar si existe correlación entre los factores anteriormente mencionados y la diabetes.
- ¿Existe alguna relación entre el embarazo y la diabetes?
- ¿Quienes son más propensos a desarrollar esta enfermedad?
- ¿Qué factor de riesgo hay que tener más en cuenta para evitar contraer diabetes?

3. Dataset

El dataset utilizado contiene datos recopilados del *National Institute of Diabetes and Digestive and Kidney Diseases* y se puede consultar en la siguiente liga `https://www.kaggle.com/mathchi/diabetes-data-set`. Cada registro contiene ciertos parámetros de mujeres de la India de, al menos 21 años. Los campos del dataset son los siguientes:

- **Pregnancies:** Número de veces que se ha embarazado.
- **Glucose:** Concentración de glucosa en plasma a dos horas en un test oral de tolerancia a la glucosa.
- **BloodPressure:** Presión sanguínea.
- **SkinThicness:** Grosor de la piel del tricep.
- **Insulin:** Suero de insulina.
- **BMI:** Índice de masa corporal.
- **DiabetesPedigreeFunction:** Diabetes pedigree function.
- **Age:** Edad.
- **Outcome:** diabetes 0 o no diabetes (*sana*) 1.

4. Limpieza de los datos

En general, el dataset se encontraba muy limpio, a excepción de unos valores NaN.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	0	33.6	0.627	50	1
1	8	183.0	64.0	NaN	0	23.3	0.672	32	1
2	0	137.0	40.0	35.0	168	43.1	2.288	33	1
3	3	78.0	50.0	32.0	88	31.0	0.248	26	1
4	2	197.0	70.0	45.0	543	30.5	0.158	53	1
...	...	...	...	...	...	...	...	...	...
763	9	89.0	62.0	NaN	0	22.5	0.142	33	0
764	10	101.0	76.0	48.0	180	32.9	0.171	63	0
765	2	122.0	70.0	27.0	0	36.8	0.340	27	0
766	5	121.0	72.0	23.0	112	26.2	0.245	30	0
767	1	93.0	70.0	31.0	0	30.4	0.315	23	0

Figura 1: Dataset con el que se va a trabajar.

Estos valores NaN se sustituyeron por 0, resultando:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	0	33.6	0.627	50	1
1	8	183.0	64.0	0.0	0	23.3	0.672	32	1
2	0	137.0	40.0	35.0	168	43.1	2.288	33	1
3	3	78.0	50.0	32.0	88	31.0	0.248	26	1
4	2	197.0	70.0	45.0	543	30.5	0.158	53	1
...	...	...	...	...	...	...	...	...	...
763	9	89.0	62.0	0.0	0	22.5	0.142	33	0
764	10	101.0	76.0	48.0	180	32.9	0.171	63	0
765	2	122.0	70.0	27.0	0	36.8	0.340	27	0
766	5	121.0	72.0	23.0	112	26.2	0.245	30	0
767	1	93.0	70.0	31.0	0	30.4	0.315	23	0

Figura 2: Dataset ya "limpio"

Por último, al dataset se le agregó un nuevo campo llamado "ComposicionCorporal", denotado como "cc" en la ecuación [1], con variables categóricas que indican el nivel del índice de masa corporal, es decir:

$$cc = \begin{cases} \text{Bajo, si } BMI \leq 18.85 \\ \text{Normal, si } 18.5 < BMI \leq 24.9 \\ \text{Elevado, si } 24.9 < BMI \leq 29.9 \\ \text{Obesidad, si } 29.9 < BMI \end{cases} \quad (1)$$

Resultando:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome	ComposicionCorporal
0	6	148.0	72.0	35.0	0	33.6	0.627	50	1	Obesidad
1	8	183.0	64.0	20.5	0	23.3	0.672	32	1	Normal
2	0	137.0	40.0	35.0	168	43.1	2.288	33	1	Obesidad
3	3	78.0	50.0	32.0	88	31.0	0.248	26	1	Obesidad
4	2	197.0	70.0	45.0	543	30.5	0.158	53	1	Obesidad
...	...	...	...	...	...	...	...	...	...	...
763	9	89.0	62.0	20.5	0	22.5	0.142	33	0	Normal
764	10	101.0	76.0	48.0	180	32.9	0.171	63	0	Obesidad
765	2	122.0	70.0	27.0	0	36.8	0.340	27	0	Obesidad
766	5	121.0	72.0	23.0	112	26.2	0.245	30	0	Elevado
767	1	93.0	70.0	31.0	0	30.4	0.315	23	0	Obesidad

Figura 3: Dataset completo.

5. Análisis Exploratorio

Una vez se obtuvo el dataset [3] procedemos a realizar un análisis exploratorio de los datos.

Se obtiene una descripción general de los campos para mujeres con diabetes, tabla [4], y mujeres *sanas*, tabla [5].

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.000000	268.0
mean	142.158955	74.950000	28.895522	100.335821	35.380597	0.550500	37.067164	1.0
std	29.546454	12.016731	10.300190	138.689125	6.597094	0.372354	10.968254	0.0
min	78.000000	30.000000	7.000000	0.000000	22.900000	0.088000	21.000000	1.0
25%	119.000000	68.000000	20.500000	0.000000	30.900000	0.262500	28.000000	1.0
50%	140.000000	74.000000	27.000000	0.000000	34.250000	0.449000	36.000000	1.0
75%	167.000000	82.000000	36.000000	167.250000	38.775000	0.728000	44.000000	1.0
max	199.000000	114.000000	99.000000	846.000000	67.100000	2.420000	70.000000	1.0

Figura 4: Descripción del *subdataframe* con los registros de las mujeres con diabetes.

	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.000000	500.0
mean	110.704800	70.809800	25.363000	68.792000	30.878400	0.429734	31.190000	0.0
std	24.714788	11.932299	9.036096	98.865289	6.502777	0.299085	11.667655	0.0
min	44.000000	24.000000	7.000000	0.000000	18.200000	0.078000	21.000000	0.0
25%	93.000000	63.500000	20.500000	0.000000	25.750000	0.229750	23.000000	0.0
50%	107.500000	70.000000	21.000000	39.000000	30.400000	0.336000	27.000000	0.0
75%	125.000000	78.000000	31.000000	105.000000	35.300000	0.561750	37.000000	0.0
max	197.000000	122.000000	60.000000	744.000000	57.300000	2.329000	81.000000	0.0

Figura 5: Descripción del *subdataframe* con los registros de las mujeres *sanas*.

Lo primero que se observa es que, del total de la muestra, hay 500 mujeres "sanas"[1] y 268 mujeres con diabetes detectada.

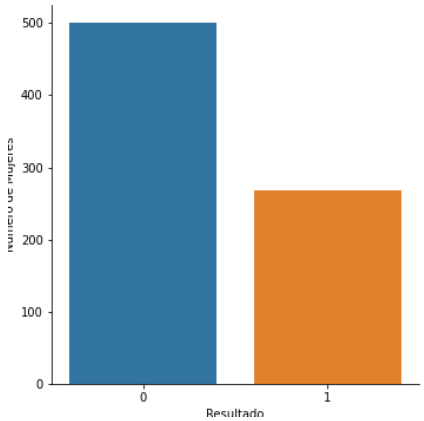

Figura 6: Gráfica de barras que muestra la incidencia de diabetes en la muestra de población tomada.

Cómo se muestra en la figura [6], el número 0 representa a una mujer *sana* y 1 a una mujer con diabetes.

[1]Entiéndase por sanas a mujeres que salieron negativas al examen de detección de la diabetes.

Por otro lado, y hablando en términos generales de la muestra, se observa que la mayoría de los registros corresponden a mujeres jóvenes entre 21 y 36 años, figura [7].

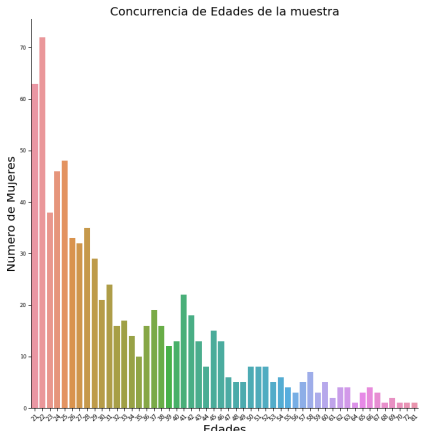

Figura 7: Edad de mujeres de la muestra.

Graficando de nuevo la concurrencia de edades pero ahora separadas según su diagnóstico

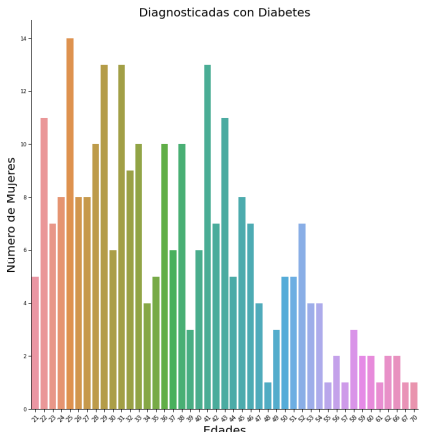

Figura 8: Gráfica de la concurrencia de edades de mujeres con diabetes.

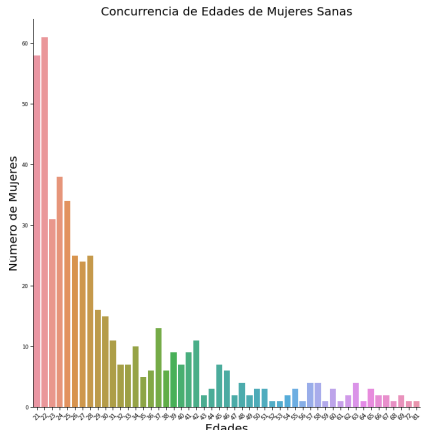

Figura 9: Gráfica de la concurrencia de edades de mujeres sanas.

Como se observa en la figura [8] la distribución de mujeres con diabetes según su edad parece tener una distribución distinta a la mostrada en la figura [9]. Para la primera gráfica [8], parece que los casos se distribuyen normalmente[2] y para la gráfica [9] se tiene lo que parece ser un decaimiento exponencial. Esto se debe a que, como ya se mencionó, la mayoría de las mujeres registradas son jóvenes.

Como se menciona en la introducción, uno de los factores de riesgo es la concentración de glucosa en sangre. Por esta razón se obtiene la siguiente gráfica de cajas que se muestra en la figura [10].

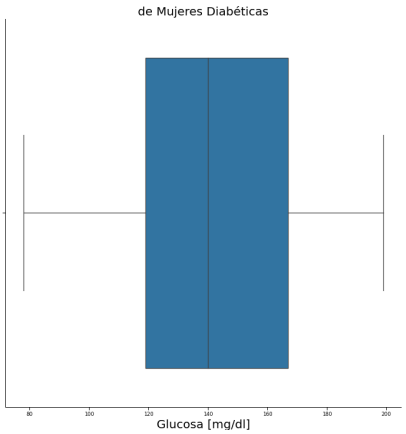

Figura 10: Gráfica de cajas para la concentración de glucosa en sangre de mujeres con diabetes.

Por otro lado se muestra la gráfica de cajas para mujeres *sanas*, figura [11].

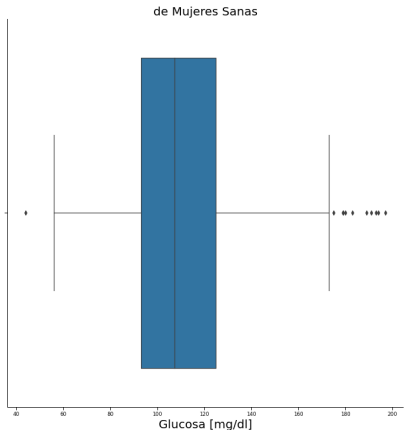

Figura 11: Gráfica de cajas para la concentración de glucosa en sangre de mujeres *sanas*.

Nótese que, para las mujeres sanas, el promedio de la glucosa en sangres es de 110.705 ± 24.74 $[mg/dl]$ mientras que para las mujeres con diabetes el promedio de glucosa en sangre es 142.159 ± 29.546 $[mg/dl]$. Sin embargo, algo interesante a resaltar es que, considerando las desviaciones estándar de ambas cantidades, hay mujeres que tienen la glucosa elevada y, sin embargo, no tienen diabetes, lo que quiere decir que, definitivamente, la concentración de diabetes en la sangre, a pesar de ser un factor de riesgo, no es un factor suficiente para el diagnóstico de la enfermedad. Esta hipótesis toma aún más fuerza cuando se observan las siguientes gráficas:

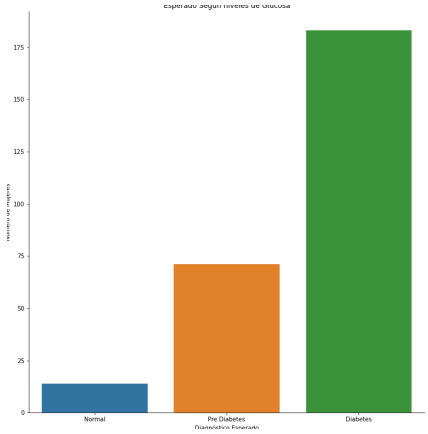

Figura 12: Gráfica de barras de los niveles de glucosa las mujeres diagnosticadas con diabetes.

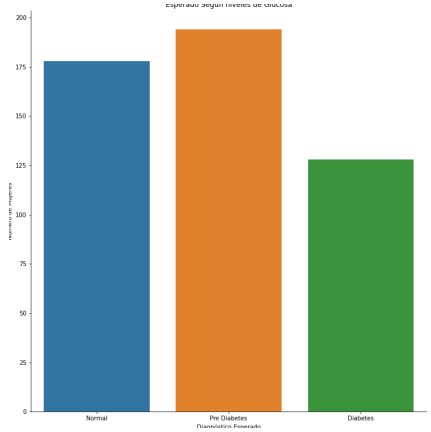

Figura 13: Gráfica de barras de los niveles de glucosa las mujeres *sanas*.

Para realizar las gráficas, figuras [12] y [13] se agregó otra columna al *dataframe* llamada "TipoDiabetes", denotado como "td" en la ecuación [2], con variables categóricas que indican el diagnóstico esperado.

$$td = \begin{cases} \text{Normal, si } glucosa < 100 \\ \text{Prediabetes, si } 100 \leq glucosa < 125 \\ \text{Diabetes, si } 125 \leq glucosa \end{cases} \quad (2)$$

Resulta ser que 128 mujeres *sanas* deberían tener diabetes por su alto nivel de glucosa en la sangre, sin embargo salieron negativas en la prueba.

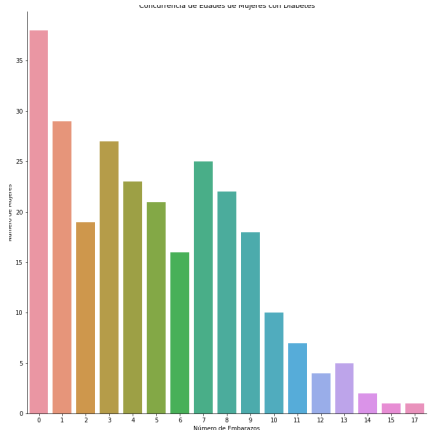

Figura 14: Gráfica de barras que muestra la cantidad de mujeres con diabetes según su número de embarazos.

Otro resultado a destacar lo muestra la matriz de correlación, tabla [15].

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
Pregnancies	1.000000	0.127970	0.208984	0.013103	-0.073535	0.021542	-0.033523	0.544341	0.221898
Glucose	0.127970	1.000000	0.219675	0.160581	0.331203	0.231583	0.137112	0.266608	0.492906
BloodPressure	0.208984	0.219675	1.000000	0.133895	-0.038133	0.281218	0.000376	0.326743	0.162981
SkinThickness	0.013103	0.160581	0.133895	1.000000	0.287041	0.535602	0.155117	0.026022	0.174811
Insulin	-0.073535	0.331203	-0.038133	0.287041	1.000000	0.185483	0.185071	-0.042163	0.130548
BMI	0.021542	0.231583	0.281218	0.535602	0.185483	1.000000	0.153528	0.025794	0.312317
DiabetesPedigreeFunction	-0.033523	0.137112	0.000376	0.155117	0.185071	0.153528	1.000000	0.033561	0.173844
Age	0.544341	0.266608	0.326743	0.026022	-0.042163	0.025794	0.033561	1.000000	0.238356
Outcome	0.221898	0.492906	0.162981	0.174811	0.130548	0.312317	0.173844	0.238356	1.000000

Figura 15: Tabla de correlación.

Números cercanos a 1 indican una correlación lineal positiva entre los campos, números cercanos a -1 indican una correlación lineal negativa entre los campos y números cercanos a 0 indican poca o nula correlación.

Se puede notar que la concentración de glucosa en la sangre es guarda una correlación moderada con los demás campos. Esto quiere decir que, a pesar de no ser un factor de riesgo determinante en el diagnóstico de la diabetes, éste resulta aumentar mucho la probabilidad de esta enfermedad. De igual manera el índice de masa corporal, "BMI" en el dataset, muestra una estrecha correlación con los demás factores de riesgo.

A continuación se muestran unas gráficas de la concentración de glucosa en la sangre con los demás campos de la tabla, diferenciando con color azul los casos de mujeres *sanas* y en naranja los casos de mujeres con diabetes.

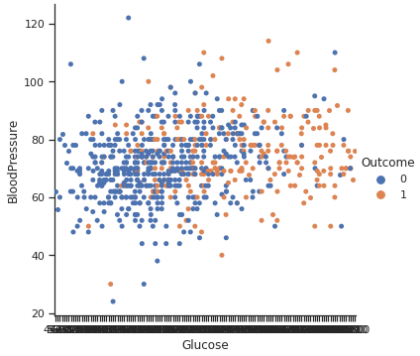

Figura 16: Gráfica de la presión sanguínea y la concentración de la glucosa en la sangre.

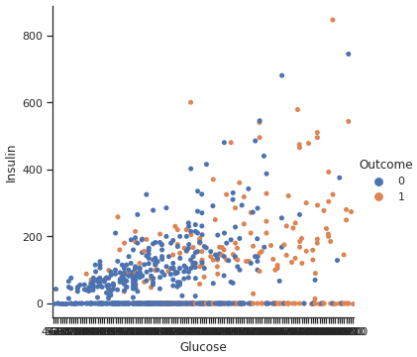

Figura 17: Gráfica de la cantidad de insulina y la concentración de la glucosa en la sangre.

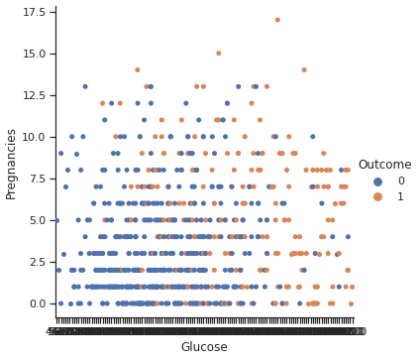

Figura 18: Gráfica del número de embarazos y la concentración de la glucosa en la sangre.

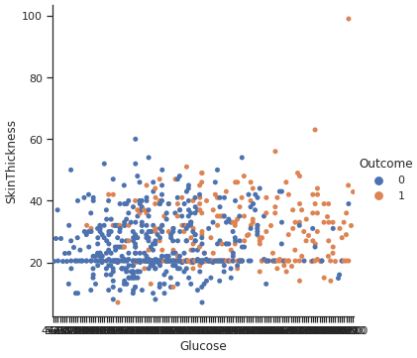

Figura 19: Gráfica del grosor de la piel y la concentración de la glucosa en la sangre.

En las gráficas anteriores podemos observar claramente que existen regiones diferenciadas para valores de Glucosa y Presión Sanguínea, figura [16], Glucosa e Insulina, figura [17], Glucosa y número de embarazos, figura [18] y Glucosa y Grosor de la Piel, figura [19], en las que, las mujeres tienden a desarrollar la Diabetes, zonas naranjas.

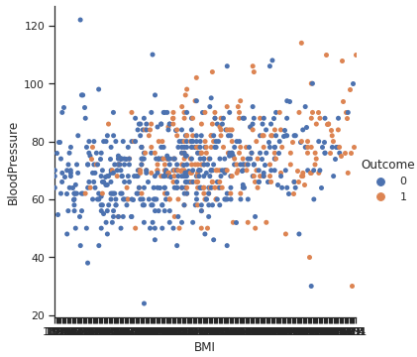

Figura 20: Gráfica de la presión sanguínea y el índice de masa corporal.

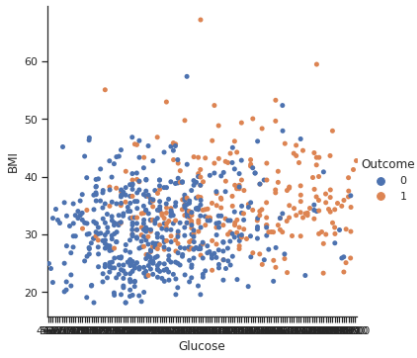

Figura 21: Gráfica de la concentración de glucosa en la sangre y el índice de masa corporal.

En las imágenes anteriores, figuras [20] y [21], también se pueden diferenciar regiones en el espacio donde, claramente, existe una diferenciación entre el Índice de Masa Corporal y la Presión Sanguínea y el Índice de Masa Corporal y la Concentración de la Glucosa en la Sangre, para mujeres con o sin diabetes.

Los resultados anteriores hablan de una segmentación de la población de mujeres de la India que son más propensas a desarrollar esta enfermedad.

6. Conclusiones

Claramente los factores que pueden derivar al desarrollo de la diabetes son varios y en general dependen de cada cuerpo, sin embargo, se pudo notar cierta inclinación de ciertos factores de riesgo que incrementan esta posibilidad, siendo los más considerables el peso, tabla [22] y la concentración de glucosa en la sangre, tabla [23].

	Numero_de_personas	%_del_total_de_personas
Normal	7.0	2.611940
Elevado	40.0	14.925373
Obesidad	221.0	82.462687

Figura 22: Probabilidad de desarrollar diabetes según el peso.

	Numero_de_personas	%_del_total_de_personas
Tipo_Normal	14.0	5.223881
Tipo_Pre-Diabetes	71.0	26.492537
Tipo_Diabetes	183.0	68.283582

Figura 23: Probabilidad de desarrollar diabetes según la concentración de glucosa en la sangre.

Además se puede observar cierta relación entre algunos factores de riesgo que se relacionan con la diabetes. Por ejemplo, el grosor de la piel no guarda una relación aparente con tener o no diabetes, sin embargo, cuando a esta se le relaciona con la glucosa en la sangre, claramente se aprecian regiones en donde se distingue a las personas que tienen diabetes de las que no, figura [19].

En el caso de la edad y la diabetes, figura [8], y considerando que la NIDDK [3] dice que la probabilidad de contraer esta enfermedad aumenta con el paso de los años, resulta extraño tener esta distribución, sin embargo, puede explicarse con el hecho de que la mayoría de mujeres registradas se encuentran en la juventud.

Algo curioso a señalar es que, mientras menos embarazos tenga una mujer, mayor es la probabilidad de que esta desarrolle esta enfermedad, figura [14].

En México, según la Encuesta Nacional de Salud y Nutrición, ENSANUT, en 2006, la prevalencia de la diabetes sea mayor en las mujeres se delinea más claramente mientras mayor es la edad, siendo que en el grupo de edad de 50 a 59 años, la prevalencia llegó a un 14.2 % en mujeres y 12.7 % en hombres, mientras que en el grupo de 60 a 69 años, la prevalencia fue de 21.3 % en mujeres y 16.8 % en hombres[1]. Por lo que, al menos en nuestro país la mujer es más susceptible de contraer la enfermedad.

Por último señalar que, por medio del análisis de nuestros datos, logramos responder las preguntas que nos planteamos como objetivo principal en este proyecto, además de que generamos nuevas preguntas a partir de observar el comportamiento de nuestros datos.

Referencias

[1] S. de Salud. Género y salud. http://cnegsr.salud.gob.mx/contenidos/descargas/EquidadGenero/MayAgo10.pdf, Mayo-2010. Accedido 13-08-2021.

[2] C. de Wikipedia. Distribución normal. https://es.wikipedia.org/wiki/Distribuci\%C3\%B3n_normal, 20-02-2021. Accedido 13-08-2021.

[3] N. I. of Diabetes, Digestive, and K. Diseases. Diabetes tipo 2. https://www.niddk.nih.gov/health-information/informacion-de-la-salud/diabetes/informacion-general/que-es/diabetes-tipo-2#:~:text=Sin\%20embargo\%2C\%20la\%20diabetes\%20tipo,diabetes\%20o\%20sobrepeso\%20u\%20obesidad., 2020. Accedido 13-08-2021.