

Análisis de Datos de Diabetes

García, Juan Manuel first1.last1@xxxxx.com	Herrera, José Emiliano eherrera1331@gmail.com	Miramontes, Fred liloliol@hotmail.com
Román, Christopher ferrobin34@gmail.com	Sánchez, Gabriel 178294@iberopuebla.mx	Sánchez, Ludim sanchezludim.anel@gmail.com

13 de agosto de 2021

Resumen

En esta práctica se identificaron los productos, partículas resultantes, de una colisión de un haz de protones con un gas de SF_6 , mejor conocido como *hexafluoruro de azufre*. Dichos productos fueron los siguientes iones: F^+ , SF^+ , S^+ , SF_2^{++} , SF^+ , SF_2^+ , SF_3^+ , SF_4^+ y SF_5^+ , resultando éste último el más abundante. Los iones anteriores fueron identificados a través de su *Relación Masa/Carga*, sin embargo no se notó presencia de iones SF_6 , esto puede deberse a su configuración electrónica poco estable. Las colisiones se llevaron a cabo en el *acelerador lineal* de bajas energías (1 keV a 10 keV) que se encuentra en el edificio *Tlahuizcalpan* de la Facultad de Ciencias de la Universidad Nacional Autónoma de México. Además, en ésta práctica se discute el funcionamiento de las partes del acelerador lineal, así como la razón de la ausencia del ión SF_6 .

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. Los 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 la correlación entre los factores anteriormente mencionados y la diabetes.
- ¿Existe alguna relación entre el embarazo y la diabetes?
- ¿Quiénes son más propensos a contraer esta enfermedad, hombres o mujeres?
- ¿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:** 0 o 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:

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

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.

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

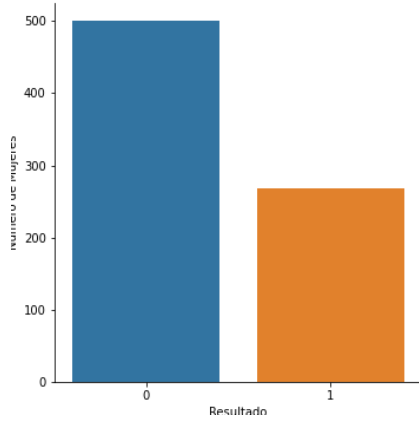


Figura 4: Gráfica de barras que muestra la incidencia de diabetes en la muestra de población tomada.

Cómo se muestra en la figura [4], el número 0 representa a una mujer *sana* y 1 a una mujer con 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 [5].

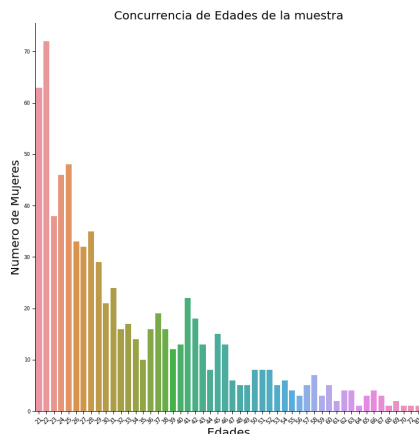


Figura 5: Edad de mujeres de la muestra.

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

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

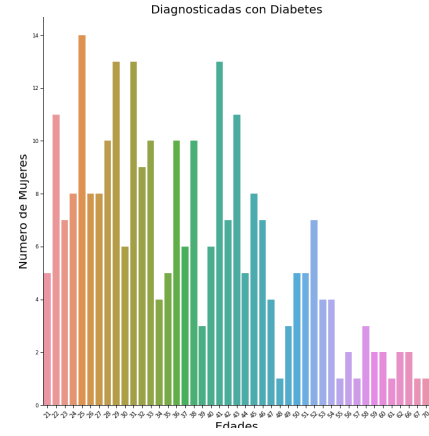


Figura 6: Gráfica de la concurrencia de edades de mujeres con diabetes.

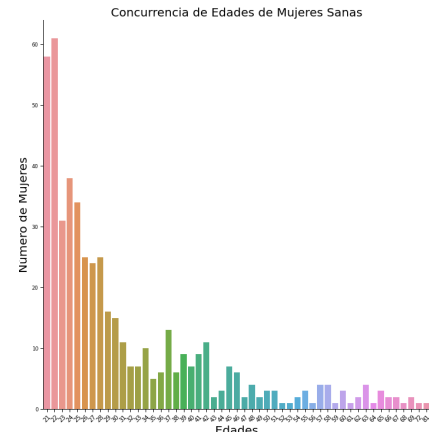


Figura 7: Gráfica de la concurrencia de edades de mujeres sanas.

Como se observa en la figura [6] la distribución de mujeres con diabetes según su edad parece tener una distribución distinta a la mostrada en la figura [7]. Para la primera gráfica [6], parece que los casos se distribuyen normalmente[1] y para la gráfica [7] 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.

Referencias

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