

.....

Proyecto: **Análisis para la
detección de grupos
vulnerables de Diabetes**

Agosto 13, 2021


.....

Equipo 5





¿Quiénes somos?

- Reporte en LaTeX: José Emiliano Herrera Velázquez
 - Estructura de notebook: Gabriel Sánchez
 - Visualización de datos: Fred Jiordi Miramontes Arias
 - Presentación: Ludim Sánchez
 - Notebook API: Christopher Román Jaimes
 - Vídeo: Juan Manuel Garcia Briones
- 

Repositorio:

<https://github.com/emilianoel/>

[Proyecto BEDU Equipo 5](#)

https://raw.githubusercontent.com/emilianoel/Proyecto_BEDU_Equipo_5/main/Reporte/Diabetes.pdf

https://github.com/emilianoel/Proyecto_BEDU_Equipo_5/blob/main/Proyecto.ipynb



Identificación de un problema

Desde el año 2000, la diabetes mellitus en México fue 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.



Investigación & Búsqueda de soluciones anteriores

[Sign In](#) [Register](#)

Dataset

Diabetes Data Set

This dataset is originally from the N. Inst. of Diabetes & Diges. & Kidney Dis.

Mehmet A.

• updated a year ago (Version 1)

Data

Tasks (1)

Code (28)

Discussion

Activity

Metadata

Download (23 KB)

New Notebook

Usability 10.0

License

CC0: Public Domain

Tags

education, health, diabetes

Description

Context

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

Content

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

es, analyze web traffic, and improve your experience on the site. By using Kaggle, you agree to our use of cookies.

Got it

Learn more

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?



.....

Limpieza de datos


.....



Observación de variables

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. 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.
 - Outcome: 0 o 1.
- 

Una vez que generamos nuestras preguntas, obtuvimos una base de datos correspondiente a una muestra poblacional de la India, donde los participantes son en su totalidad mujeres.

```
import pandas as pd
import matplotlib.pyplot as plt
```

- ### LECTURA DE DATA SET

Ya con nuestros datos, pasamos a hacer uso de pandas para leer los valores y obtener nuestro Data Frame. Imprimimos nuestra tabla y hacemos un análisis rápido de los valores que tenemos y sus columnas.

```
df = pd.read_csv("/content/drive/MyDrive/diabetes2.csv")
df
```

	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

768 rows × 9 columns

Limpieza

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

Estos valores NaN se sustituyeron por 0, resultando:

Debido a que nuestro data set no es muy amplio, queremos conservar todos los datos para poder analizar mejor, por ello en una nueva variable pasaremos el mismo Data Frame y llenaremos los NAN con 0.

```
df2 = df.fillna(0)
df2
```

	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

768 rows x 9 columns


Agregar nuevo campo

Por último, al dataset se le agregó un nuevo campo llamado “ComposicionCorporal”, denotado como “cc”, la cual nos servirá para hacer las predicciones. Esta nueva variables categórica, utiliza el nivel del índice de masa corporal (BMI) para asignar su valor.

	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

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

Figura 3: Dataset completo.



.....

Análisis Exploratorio

.....

Separando muestras

Lo primero que se observa es que, del total de la muestra, hay 500 mujeres “sanas” (0-sanas, 1-diabetes) y 268 mujeres con diabetes detectada.

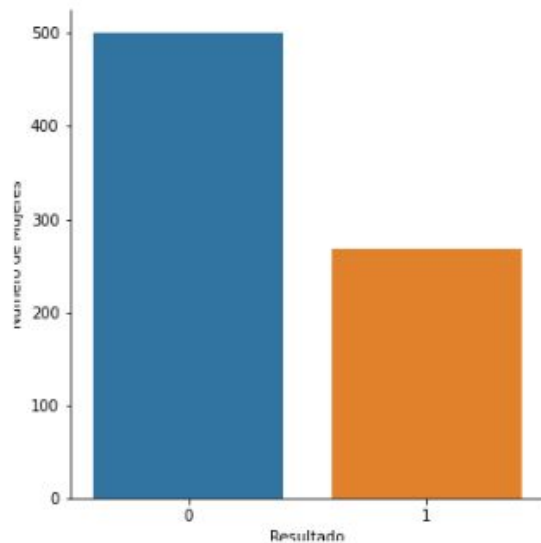
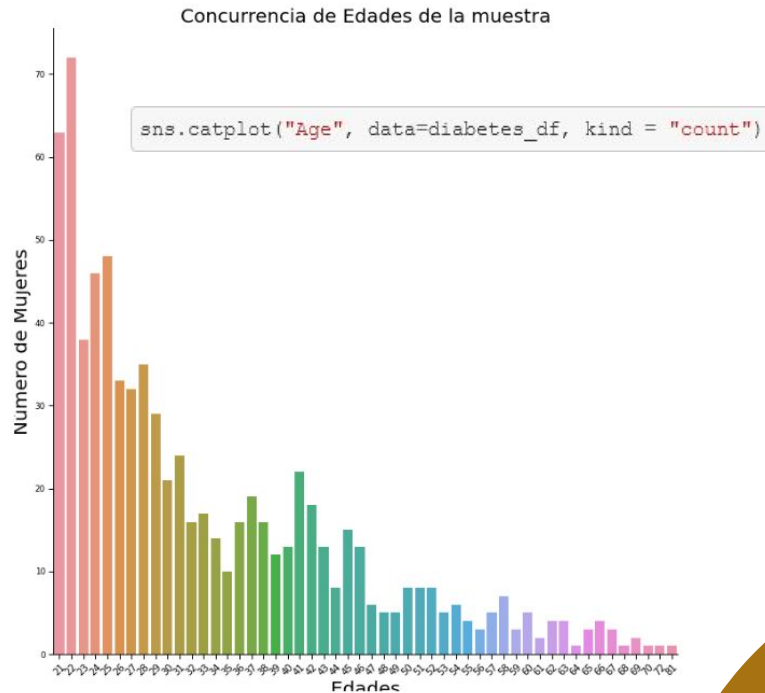
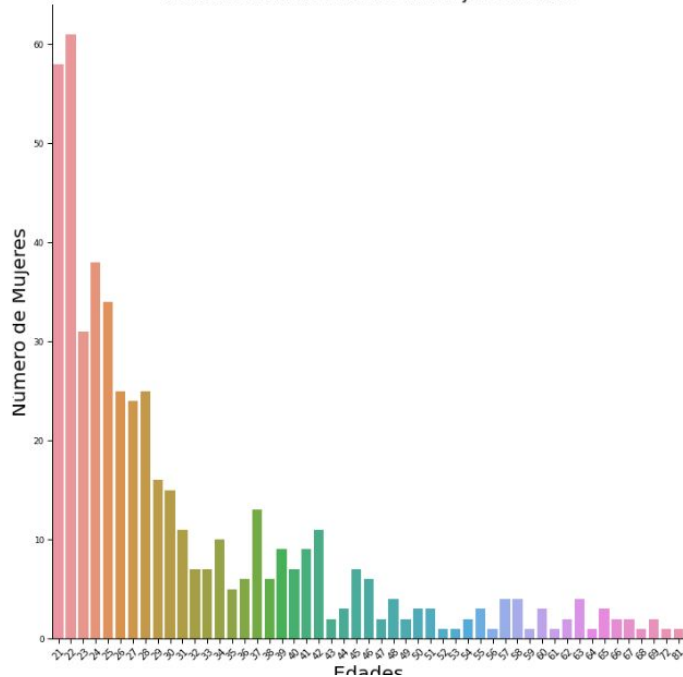


Figura 6: Gráfica de barras que muestra la incidencia de diabetes en la muestra de población tomada.

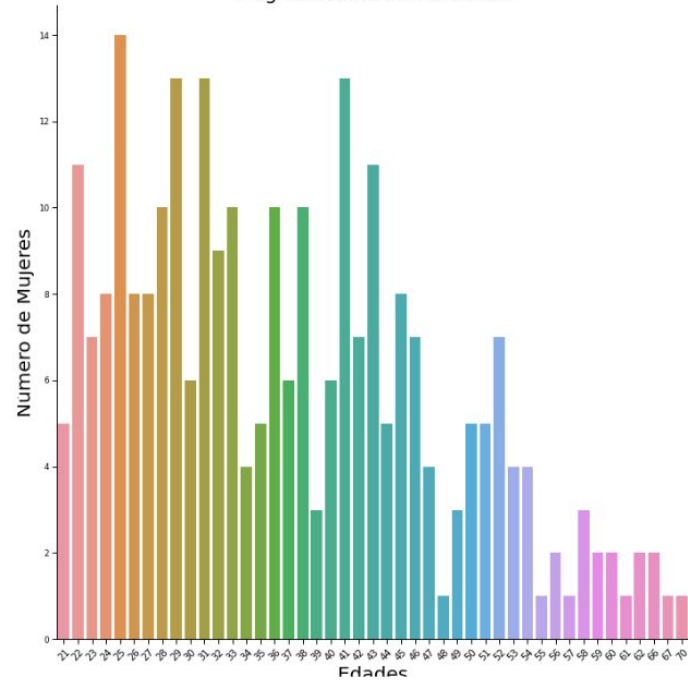


Mujeres sin diabetes vs Mujeres con diabetes

Concurrencia de Edades de Mujeres Sanas



Diagnosticadas con Diabetes



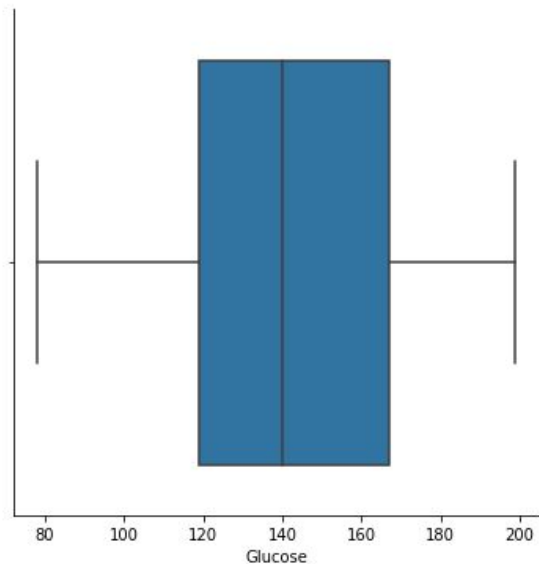
¿La glucosa?

Aquellas que han dado positivo mantienen una tendencia en su nivel de glucosa entre 120 y 165, teniendo su media en 140

```
sns.catplot("Glucose", data=diabetes_confirmado_df, kind = "box")
```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning
m version 0.12, the only valid positional argument will be `data`,
t in an error or misinterpretation.

<seaborn.axisgrid.FacetGrid at 0x7f84a3d2f550>



¿La glucosa?

Mientras que los pacientes que dieron negativo se mantienen en un rango 90 a 125, con una media de 105.

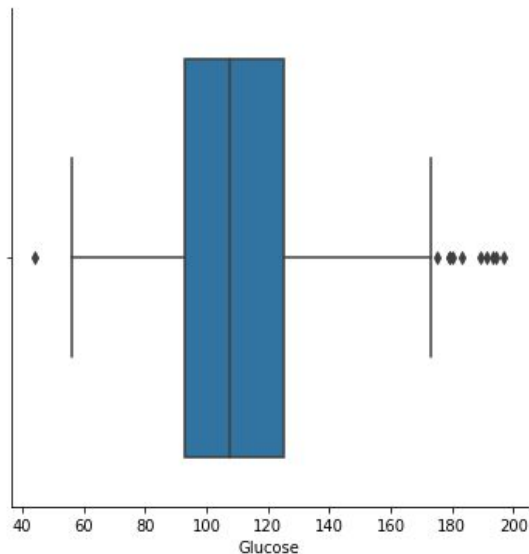
Si bien, estos datos tienen congruencia dado que este padecimiento se relaciona con la cantidad de glucosa en la sangre, hay también unos pocos pacientes que tuvieron niveles de glucosa superiores a la media de diabetes y no la padecen, lo que indica que de no controlarla en un futuro podrían llegar a padecerla.

```
sns.catplot("Glucose", data=diabetes_negado_df, kind = "box")
```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py: m version 0.12, the only valid positional argument will be `data` in an error or misinterpretation.

FutureWarning

<seaborn.axisgrid.FacetGrid at 0x7f84a3c793d0>






Observaciones

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].

Algo interesante a resaltar es que, considerando las desviaciones estándar de ambas cantidades, hay mujeres que tienen la glucosa elevada, sin embargo, no tienen diabetes, lo que quiere decir que, definitivamente, *la concentración de glucosa 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:



- ##### COMPOSICIÓN CORPORAL

```
diabetes_confirmado_df.groupby("ComposicionCorporal").size()
```

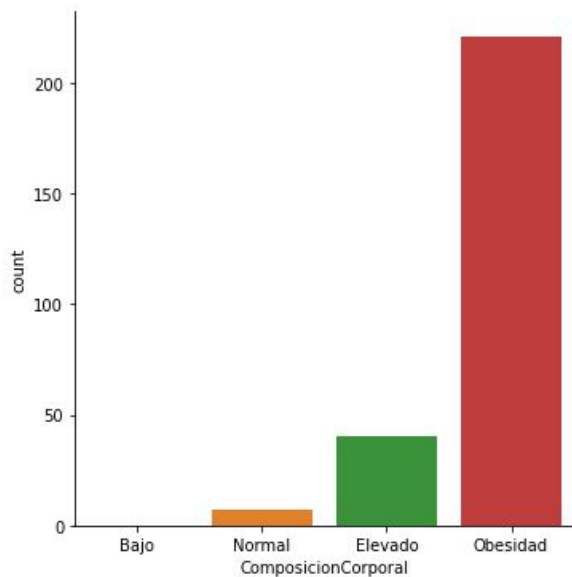
```
ComposicionCorporal
Elevado          40
Normal           7
Obesidad        221
dtype: int64
```

```
sns.catplot("ComposicionCorporal", data=diabetes_confirmado_df, kind = "count",order=["Bajo","Normal","Elevado","Obesidad"])
```

```
/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg
m version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword w
t in an error or misinterpretation.
```

```
FutureWarning
```

```
<seaborn.axisgrid.FacetGrid at 0x7f17f300a5d0>
```



```
diabetes_negado_df.groupby("ComposicionCorporal").size()
```

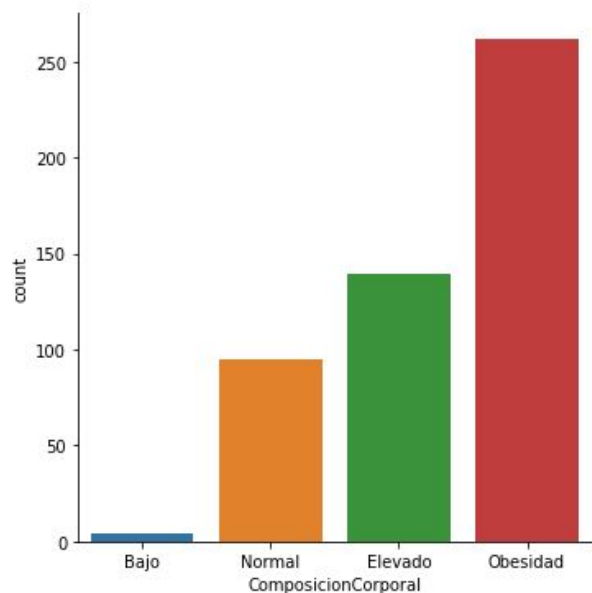
```
ComposicionCorporal
Bajo                4
Elevado            139
Normal              95
Obesidad           262
dtype: int64
```

```
sns.catplot("ComposicionCorporal", data=diabetes_negado_df, kind = "count", order=["Bajo", "Normal", "Elevado", "Obesidad"])
```

/usr/local/lib/python3.7/dist-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword argument in version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword argument in an error or misinterpretation.

FutureWarning


<seaborn.axisgrid.FacetGrid at 0x7f17f2fda150>

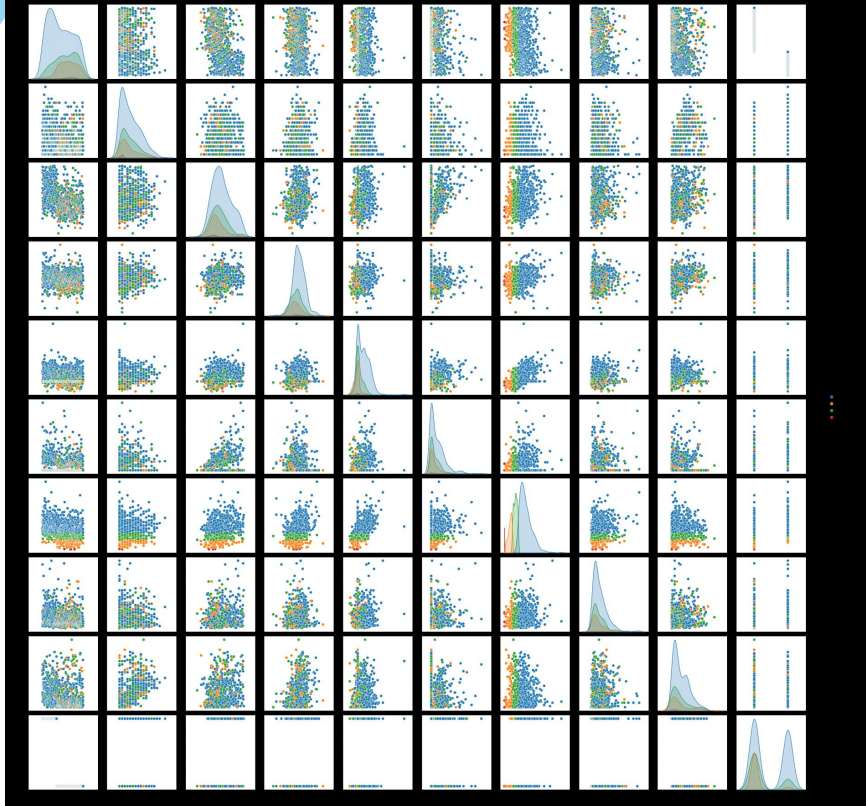




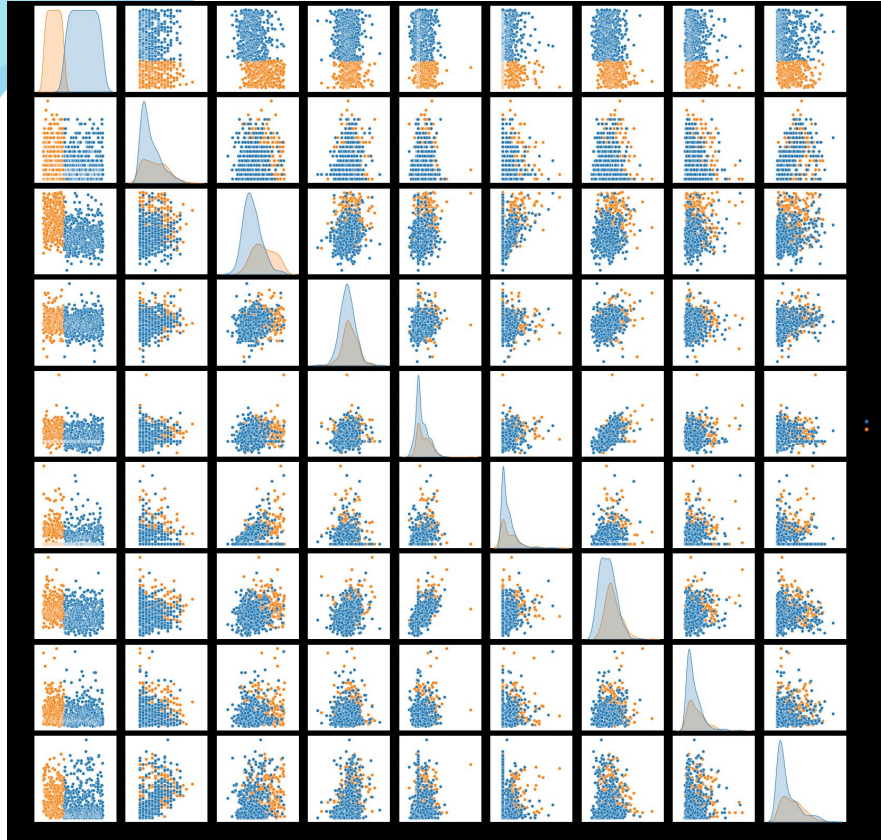
Observaciones

Graficamos la composición corporal de pacientes positivos y negativos a diabetes, con ello encontramos que aquellos que sufren *obesidad era el 82% de los pacientes que tenían diabetes*. Mientras que en los *pacientes sin diabetes la obesidad equivalía al 52%*, además de presentar pacientes bajos de peso, cuyo grupo no estuvo presente en los pacientes positivos a diabetes.





Podemos observar que existe una *tendencia lineal creciente entre la relación del índice de masa corporal (BMI) y el grosor de la piel de los pacientes (SkinThickness)*. Cuando mayor fue BMI mayor grosor presentaron en su piel, de igual manera, al ir aumentando el grosor de piel se comenzó a observar comportamientos corporales elevados y obesidad.



Verificando con las tablas anteriores observamos que algunas personas podrían tener un *nivel alto de glucosa sin padecer diabetes, sin embargo, en esto influye el historial familiar* de cada paciente respecto a esta enfermedad (DiabetesPedigreefunction).

Cuando este índice es mayor, las personas tienen una probabilidad mayor de contraer el padecimiento a menores niveles de azúcar.

Mientras que los pacientes con valores bajos en su historial familiar pudieron presentar cantidades altas de glucosa en la sangre sin contraer diabetes.

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.

	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.



Conclusiones

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.


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





Conclusiones

Este proyecto nos ayudó a realizar y entender los temas vistos en clase, ya que logramos obtener un dataset acorde a nuestro proyecto, revisar nuestros datos y desarrollar una estrategia óptima para la limpieza de nuestros datos, así como de crear nuevas columnas que nos ayudarán a inferir otros datos y/o nos facilitara la obtención e interpretación de los mismos.



¿Nuestro retos?

Internos

- Organización para trabajo en equipo de forma remota y con distintos horarios
- Trabajar a marcha forzada (hasta última hora)

Externos

- Integrante enferma
- No todos podían contribuir la misma cantidad de tiempo
- No se encontraron datos abiertos de México

Mejores prácticas: Canal en Discord, respuesta de compañeros por grupo de WhatsApp, participación de equipo en sesiones en vivo.