UNIVERSIDAD NACIONAL AUTONÓMA DE MÉXICO

FACULTAD DE CIENCIAS

TEMAS SELECTOS DE FÍSICA COMPUTACIONAL II. CIENCIAS DE DATOS. GRUPO 8365
SEMESTRE 2023-2

PROYECTO FINAL.

ENTREGA: 5 DE JUNIO DE 2023

ESTUDIO DE LAS RELACIONES DE MEDIDAS BIOMÉTRICAS CON UN POSIBLE DIAGNÓSTICO DE DIABETES.

Integrantes:
Priego Morales Jesús
Silva Vergara Ricardo

Profesores:
Flores Silva Pedro Arturo
Jiménez López Karen Rubi

Junio 2023

Proyecto final.
Estudio de las relaciones de medidas biométricas con un posible diagnóstico de diabetes.

Priego Morales Jesús, Silva Vergara Ricardo,
jesuspriego@ciencias.unam.mx, ricardosv@ciencias.unam.mx
5 de junio de 2023

Resumen

Se realizó un análisis detallado de algunas medidas biométricas tomadas a un conjunto de pacientes femeninos de al menos 21 años, medidas que tienen como propósito determinar si existe alguna relación de algunas mediciones con un diagnóstico de diabetes. Se pretende además, implementar un modelo de machine learning que prediga si un futuro paciente puede o no tener un diagnóstico de diabetes positivo de acuerdo a algunas medidas que se le tomen.

I. Objetivos

- Determinar si existe una correlación entre algunas medidas y un diagnóstico positivo (o negativo) de diabetes en pacientes femeninas de al menos 21 años.
- Implementar un modelo de machine learning que ayude a predecir un diagnóstico para futuros pacientes que no se encuentren en el dataset.

HIPÓTESIS GENERAL: Empleando los datos proporcionados y herramientas de estadística de datos, no se puede hacer un diagnóstico preciso de diabetes, por lo que no todos los datos dados simpre tendrán relación entre ellos o no proporcionarán información para el diagnóstico.

II. Introducción

LOS datasets nos proporcionan datos de importancia en las diferentes áreas de estudio para diferentes finalidades, por ejemplo, hacer predicciones de posibles sucesos ante un fenómeno estudiado. En este caso, utilizamos un datasets sobre relaciones de medidas biométricas con un posible diagnóstico de diabetes (figura 1).

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
...	...	...	...	...	...	...	...	...	...
763	10	101	76	48	180	32.9	0.171	63	0
764	2	122	70	27	0	36.8	0.340	27	0
765	5	121	72	23	112	26.2	0.245	30	0
766	1	126	60	0	0	30.1	0.349	47	1
767	1	93	70	31	0	30.4	0.315	23	0

768 rows × 9 columns

Figura 1. Dataset sobre datos biomédicos relacionados y diagnóstico de diabétes.

Este dataset recopila información de pacientes mujeres sobre edad, IMC, número de embarazos, presión arterial, nivel de insulina, de glucosa, espesor de piel, DPF y diagnóstico de diabetes (0 para negativo y 1 para positivo). Se observa que el número de datos registrados es demasiado, por lo que el camino para analizarlos de forma eficiente es por medio de programación. Podemos relacionar los datos dados y hacer conclusiones con respecto a hipótesis que se pueden hacer. En este caso, se hizo un análisis con pandas en Python. Incluso se puede proponer un modelo de machine learning para obtener más información de interes

III. Resultados, análisis y discusión.

Una vez hecha la lectura de donde esté guardado el dataset, se hizo una estadística global con las principales medidas descriptivas como la media, valores mínimos, máximos y algunos porcentajes (figura 2).

```
datos.describe().T
```

	count	mean	std	min	25%	50%	75%	max
Pregnancies	768.0	3.845052	3.369578	0.000	1.00000	3.0000	6.00000	17.00
Glucose	768.0	120.894531	31.972618	0.000	99.00000	117.0000	140.25000	199.00
BloodPressure	768.0	69.105469	19.355807	0.000	62.00000	72.0000	80.00000	122.00
SkinThickness	768.0	20.536458	15.952218	0.000	0.00000	23.0000	32.00000	99.00
Insulin	768.0	79.799479	115.244002	0.000	0.00000	30.5000	127.25000	846.00
BMI	768.0	31.992578	7.884160	0.000	27.30000	32.0000	36.60000	67.10
DiabetesPedigreeFunction	768.0	0.471876	0.331329	0.078	0.24375	0.3725	0.62625	2.42
Age	768.0	33.240885	11.760232	21.000	24.00000	29.0000	41.00000	81.00
Outcome	768.0	0.348958	0.476951	0.000	0.00000	0.0000	1.00000	1.00

Figura 2. Estadística global del dataset empleado.

Se puede obtener mucha información de interés, por ejemplo, que el rango de edad de las pacientes va de 21 a 81, que hay más casos de diagnóstico negativo que positivo de diabetes debido al promedio obtenido de 0.3489 y que hay mucha diferencia entre los valores de insulina registrados debido a la desviación estándar. Con esta información basta para hacer algunas hipótesis, por ejemplo, que el caso de datos de nivel de insulina nos puede dar de forma más acertada un diagnóstico de diabetes. Sin embargo, podemos intentar relacionar las

demás medidas biométricas con el diagnóstico y observar si la información es favorable. Por ejemplo, si relacionamos la edad con el diagnóstico obtenemos lo mostrado en la Figura 3:

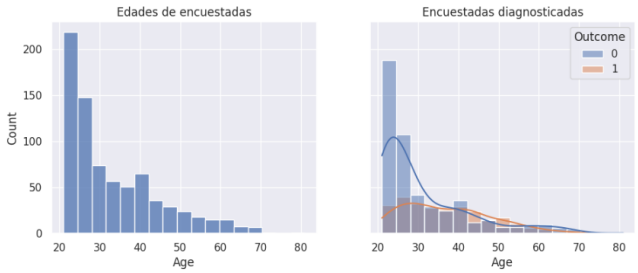

Figura 3. Relación edad y diagnóstico por medio de histogramas.

Podemos observar que hay más registros de personas jóvenes y, comparando las barras con respecto a los casos positivos de diabetes, se marcan los valores más altos en la población joven, lo cual implica que hay un sesgo con respecto a la edad y por lo tanto no nos basta con saber la edad para obtener un diagnóstico preciso, aunque sí podemos comparar los casos negativos y positivos, obteniendo que hay más casos negativos que positivos en la población menor a 30 años.

Por otro lado, podemos ver si hay valores atípicos, por ejemplo, en el apartado de niveles de insulina podemos encontrar valores anormales mayores a 300 (Figura 4). y 'estos los podemos utilizar para encontrar una relación con el diagnóstico.

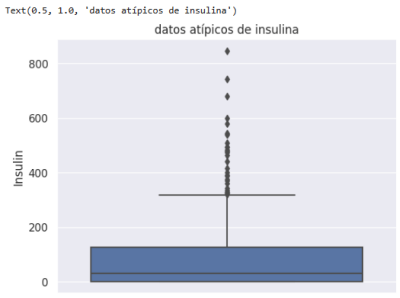

Figura 4. Valores atípicos de niveles de insulina.

Posteriormente, de las figuras 5 y 6, se observa que los valores atípicos proporcionan información relevante, puesto que la tendencia de diagnósticos positivos es mayor en este apartado que en los valores típicos.

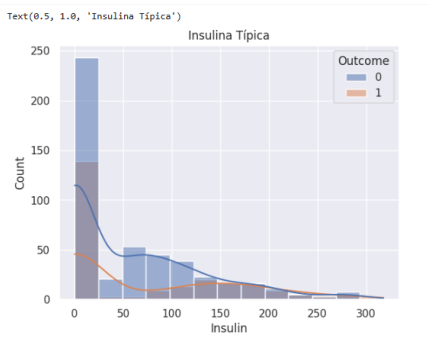

Figura 5. Valores típicos y diagnóstico de diabetes.

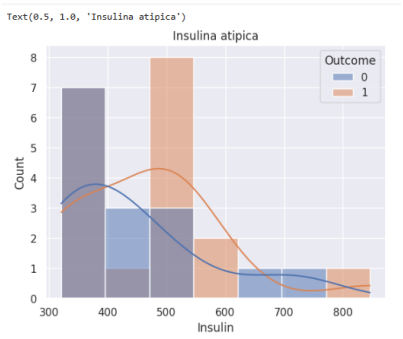

Figura 6. Valores atípicos y diagnóstico de diabetes. Obsérvese que la curva que está por encima corresponde a los diagnósticos positivos.

Por otro lado, al observar que no hay muchos valores atípicos en la sección de nivel de glucosa, basta con relacionar directamente con el diagnóstico (Figura 7).

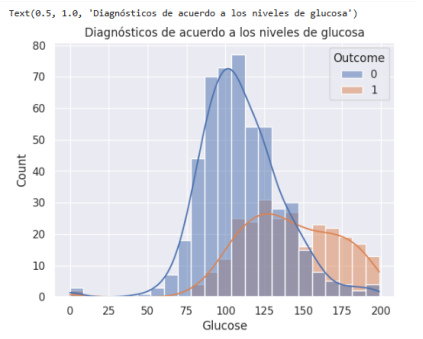

Figura 7. Valores de niveles de glucosa y diagnóstico de diabetes.

Notemos que hay una clara tendencia de que, entre mayores sean los niveles de glucosa, mayor es el riesgo de dar positivo a la enfermedad, donde la proporción de casos positivos se hace mayor a partir de valores por encima de 150.

Para las demás medidas biométricas podemos relacionarlas directamente con el diagnóstico (Figura 8) y observar que los casos positivos están por debajo de los negativos, aunque para el caso del índice de masa corporal y los antecedentes familiares DPF tienen una relación alta con los diagnósticos positivos para los cuales el índice tiene un valor por encima de 40 y la función pedigrí de la diabetes por arriba de 0.8.

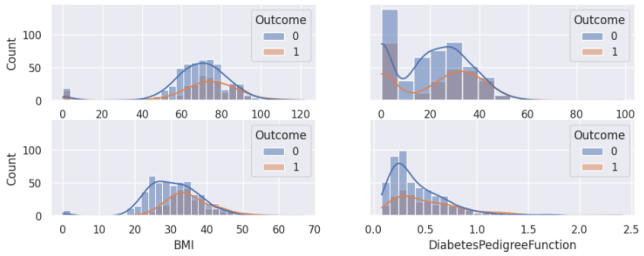

Figura 8. Valores de Indice de masa corporal y DPF en diagnóstico de diabetes.

Por otro lado, implementando un modelo de machine learning en el algoritmo de Naive Bayes, el cual considera que existe independencia entre los datos empleados, aunque existen correlaciones entre algunas de nuestras variables, las cuales se puede reducir usando la reducción de dimensionalidad por PCA. Estas correlaciones las podemos apreciar en la Figura 9.

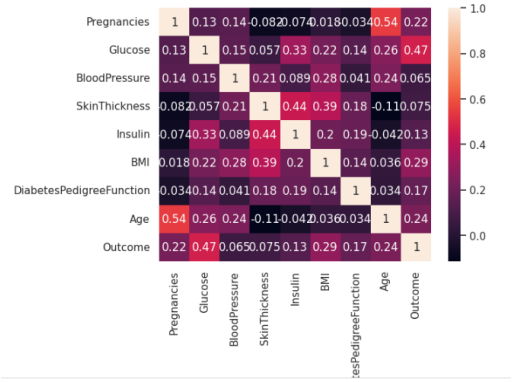

Figura 9. Correlaciones entre variables usadas.

Se observan que los datos con mejor correlación, dadas a un valor mayor a 0.3 según la Figura 9, corresponden a: Embarazos y edad, glucosa e insulina, grosor de piel e insulina y grosor de piel e índice de masa corporal.

Una vez hecho esto, podemos hacer una reducción de dimensionalidad, esto con el fin de no considerar las variables correlacionadas a nuestra consideración (según lo encontrado anteriormente en nuestro análisis). Y al separar el dataset de manera que éste quede proporcionado con respecto a los casos positivos y negativos, se obtiene la variable que queremos predecir y los vectores de características para estandarizar el vector de características para al fin aplicar el modelo de machine learning.

Entonces, implementando el modelo de Naive Bayes para primer caso se obtiene: 24 falsos negativos, 9 falsos positivos, 91 verdaderos negativos y 30 verdaderos postivos, por lo que nos deja con un error del 21.4 % de nuestro modelo, este error es algo elevado, esto puede deberse a que, como mencionamos, nuestros datos son desproporcionales, pues contamos con una mayor proporción de diagnósticos negativos. Adicional a lo anterior mencionado, también el hecho de reducir la dimensionalidad sólo eliminando columnas quizá no es la mejor forma de hacerlo pues estamos eliminando mucha información importante que no se perdería con un método mejor empleado. Aplicando las variables anteriormente excluidas y PCA, obtenemos: 6 falsos positivos, 39 falsos negativos, 4 verdaderos positivos y 15 verdaderos positivos, por lo que nos deja con un error del 29.2 % lo cual es mayor en comparación a nuestro anterior método para reducir la dimensionalidad, por lo cual podemos mencionar que aplicar PCA para TODOS nuestros datos este caso no fue la mejor decisión.

IV. Conclusiones

Con el dataset ocupado se tiene que, en general, los datos están desproporcionados, por ejemplo, se puede apreciar que la cantidad de gente encuestada fue en su mayoría gente muy joven, dejando sesgo en los datos. Aún con este sesgo se pueden apreciar ciertas variables que son determinantes para un diagnóstico positivo, las cuales son: Niveles de Glucosa, Niveles de insulina en la sangre, Índice de masa corporal y los antecedentes familiares.

Al implementar el método de machine learning mediante 2 caminos diferentes se obtuvo que al desechar un par de columnas, gracias a las predicciones hechas, para reducir la dimensión fue más eficiente que usar PCA puesto que no todos los datos tienen una alta correlación, por lo cual se pierde menos información eliminando columnas que obteniendo los componentes principales de todos los datos. Además, no se pudo obtener un menor error en las predicciones debido a que hay una desproporción en los datos que hay entre diagnósticos positivos y negativos, por lo cual es complicado que la máquina prediga casos positivos con tan poco entrenamiento y entonces los clasifica como casos negativos.

Referencias

[1] Kagle. Datasets (2022). ”RELACIONES DE MEDIDAS BIOMÉTRICAS CON UN POSIBLE DIAGNÓSTICO DE DIABETES.” en *http://bitly.ws/Hq87*