Automatización
Breast cancer data set.

Contreras. Wendy, Delgado. Isabella, Henao. Luisa, Marquinez. Darwin

Resumen – El cáncer de mama es una enfermedad que afecta a una gran población mundial, sin embargo, los métodos de diagnóstico actuales suelen ser demasiado especializados lo que lleva a que sean costosos, es por ello por lo que encontrar nuevas formas de diagnóstico es importante. Por ello se han creado bases de datos que permiten crear relaciones entre diferentes variables y la presencia o ausencia de cáncer de mama. El desarrollo de este trabajo permitió ver que se puede lograr una exactitud de aproximadamente 0.76 en el diagnóstico de cáncer de mama, que si bien, no es un resultado que se considere bueno, es un resultado prometedor.

I. INTRODUCCIÓN

El cáncer de mamá es una enfermedad que causa aproximadamente 2,261,419 muertes al año (Cuevas & García, 2006) y su diagnóstico oportuno puede deducir la probabilidad de muerte(*Comportamiento Del Cáncer de Mama de La Mujer En El Período Climatérico*, n.d.). Las técnicas de diagnóstico más utilizadas en la actualidad son la mamografía, la biopsia, el ultrasonido mamario e imagen por resonancia magnética (IRM), pero estos métodos tienen limitaciones en cuanto a precio e invasividad lo cual limitan el acceso accesibilidad, fuera de la falta de concientización en el diagnóstico del mismo (Bernardes et al., 2019). Y a partir de esto, se ha analizado la posibilidad de usar biomarcadores que den indicios de cáncer de mamá, generando así pruebas de tamizaje asequibles y económicas.

Un aporte de la base de datos empleada ha sido identificar biomarcadores que puedan contribuir a la detección del cáncer de mama. Estos biomarcadores fueron determinados con base a datos antropométricos y análisis de sangre de 166 participantes. Entre las pruebas de laboratorio realizadas se encuentra la resistina, glucosa, entre otros, y como datos antropométricos se tiene la edad e índice de masa corporal.

II. METODOLOGÍA

La base de datos descargada Sethttps://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra se enfoca en el estudio del cáncer de mama mediante una información clínica, y epidemiológica. Se obtuvieron muestras de 115 personas, donde 64 fueron controles y los 51 restantes enfermos. La base de datos cuenta con atributos cuantitativos como:

- Edad,
- Índice de masa corporal (IMC)
- Glucosa
- Insulina
- Índice de resistencia a la insulina (HOMA)
- Leptina
- Adiponectina
- Resistina

- Proteína quimioatrayente de monocitos-1 (MCP-1)
-

Además, cuenta con una variable binaria la cual indica la presencia o ausencia de cáncer.
Normalización de la base de datos - Cuando no se aplica normalización a la base de datos, se generan errores en el cálculo de los pesos.

Partición de la base de datos - La evaluación del modelo de clasificación buscó determinar cuáles pacientes tienen cáncer de mama mediante las características descritas anteriormente. Para ello se empleó una normalización de la base de datos donde se buscó dejar todas las características en un rango de 0 a 1. Finalizada esta etapa se procedió a realizar la partición de la base de datos siguiendo un esquema 70-30, donde se establece que se emplean el 70 % de los datos para el entrenamiento del modelo, y el 30 % restante se emplea para la evaluación de este.

Mediante la obtención de la matriz de confusión se pudieron obtener métricas para medir la calidad del algoritmo implementado, estas se encuentran a continuación:

Verdaderos positivos (VP)	Falsos positivos (FP)
Falsos negativos (FN)	Verdaderos negativos (VN)

- Exactitud: hace referencia a la cercanía de los resultados con los valores verdaderos.

$$Exactitud = \frac{VP + VN}{VP + FP + FN + VN}$$

- Precisión: se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud.

$$Precisión = \frac{VP}{VP + FP}$$

- Sensibilidad: es el valor que nos indican la capacidad para discriminar los casos positivos, de los negativos.

$$Sensibilidad = \frac{VP}{VP + FN}$$

- Especificidad: Consiste en los casos negativos que el algoritmo ha clasificado correctamente.

$$Especificidad\ = \frac{VN}{VN + FP}$$

- F1-Score: Es de gran utilidad cuando la distribución de las clases es desigual.

$$Especificidad\ = 2 * \frac{Precisión\ *\ Sensibilidad}{Precisión\ +\ Sensibilidad}$$

III. RESULTADOS

Inicialmente se determinó cuál era el porcentaje de balanceo de la base de datos, buscando que esta presentara un alto porcentaje de balanceo frente a la cantidad de sujetos con cáncer de mama y sujetos sin cáncer de mama. Se encontró que la base de datos presenta un 79 % de balanceo entre las clases.
En la tabla 1 se puede ver información estadística de cómo se encuentran distribuidas las características evaluadas y mencionadas anteriormente.

Tabla 1 Información sobre la base de datos

	Edad	IMC	Glucosa	Insulina	HOMA	Leptina	Adiponectina	Resistina	MCP
Media	57.3	27.58	97.79	10.01	2.69	26.62	10.18	14.73	534.65
Desviación Estándar	16.11	5.02	22.53	10.07	3.64	19.18	6.84	12.39	345.91
Mínimo	24.0	18.37	60.0	2.43	0.47	4.31	1.66	3.21	45.84
Máximo	89.0	38.58	201.0	58.46	25.05	90.28	38.04	82.1	1698.44

La Tabla 2 y 3 presentan los resultaos obtenidos mediante la implementación del perceptrón considerando toda la base de datos, ya sea normalizada o sin normalizar, de esta se destaca que los resultados obtenidos mediante la implementación de la normalización presentaron mejores resultados que los que se obtuvieron sin la normalización. Frente al tiempo medido, se podría decir que las diferencias en ellos no son significativas, sin embargo, se debe considerar que la cantidad de iteraciones varia. Cabe mencionar que en algunos de los casos no se pudieron evaluar todas las combinaciones de parámetros ya que el gradiente no encontraba los pesos adecuados, y por ende el algoritmo no se ejecutaba correctamente.

Tabla 2 resultados con todas las características normalizados

α	Iteraciones	Exactitud	Precisión	Sensibilidad	F1 score	Tiempo
1.0	10	0.46	0.46	1.0	0.63	0.0029861927032470703
0.1	10	0.46	0.46	1.0	0.63	0.0009989738464355469
0.01	10	0.51	0.45	0.31	0.37	0.0010001659393310547
0.001	10	0.43	0.42	0.62	0.5	0.0010085105895996094
1.0	50	0.46	0.46	1.0	0.63	0.0010039806365966797
0.1	50	0.46	0.46	1.0	0.63	0.0010128021240234375
0.01	50	0.6	0.56	0.56	0.56	0.0009968280792236328
0.001	50	0.57	0.67	0.12	0.2	0.0009999275207519531
1.0	100	0.46	0.46	1.0	0.63	0.001999378204345703
0.1	100	0.46	0.46	1.0	0.63	0.0010004043579101562
0.01	100	0.63	0.6	0.56	0.58	0.0010008811950683594
0.001	**100**	**0.69**	**1.0**	**0.31**	**0.47**	**0.0019998550415039062**
0.1	200	0.46	0.46	1.0	0.63	0.0019981861114501953
0.01	200	0.66	0.64	0.56	0.6	0.001996755599975586
0.001	200	0.63	0.62	0.5	0.55	0.002504587173461914

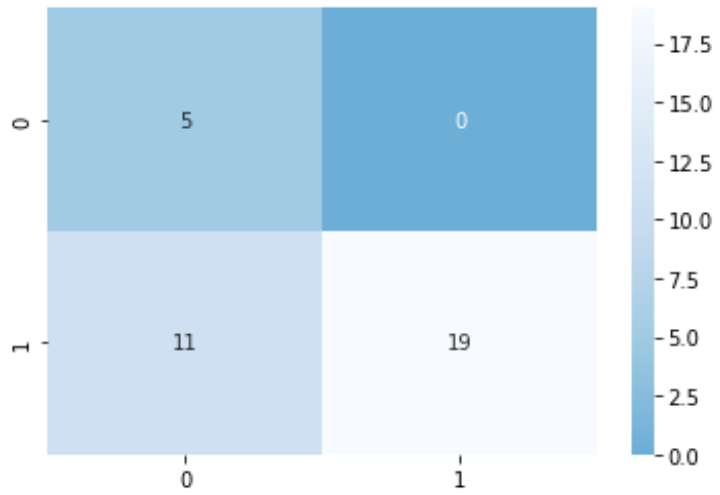

Ilustración 1 matriz de confusión (todas las características normalizadas)

Las ilustraciones1 y 2 muestran las matrices de confusión obtenidas de la implementación del perceptrón y el gradiente descendente considerando los datos normalizados o sin la normalización, se puede evidenciar que los algoritmos implementados sin la normalización de los datos presentan problemas en la clasificación, especialmente detectando los verdaderos negativos, lo cual supone un problema. Contrario a lo que se ve en con la normalización de los datos.

Tabla 3 resultados todas las características sin normalizar

α	Iteraciones	Exactitud	Precisión	Sensibilidad	F1 score	Tiempo
1.0	10	0.49	0.49	1.0	0.66	0.0010001659393310547
0.1	10	0.49	0.49	1.0	0.66	0.0009975433349609375
0.01	10	0.49	0.49	1.0	0.66	0.0010025501251220703
0.001	10	0.49	0.49	1.0	0.66	0.0
0.01	50	0.49	0.49	1.0	0.66	0.0010001659393310547
0.001	50	0.49	0.49	1.0	0.66	0.0009996891021728516

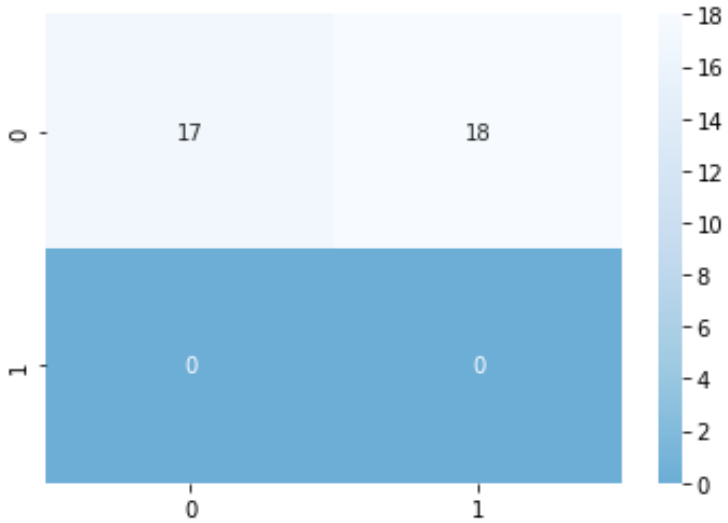

Ilustración 2 matriz de confusión (todas las características sin normalizar)

En las Tablas 4, 5 y 6 se pueden evidenciar los resultados obtenidos posterior a la selección visual de características, es de destacar, que se percibe una aparente disminución en los tiempos de ejecución, y una leve mejoría en algunos de los resultados obtenidos, al evaluar HOMA y resistin.

Tabla 4 Resultados obtenidos con la edad y el índice de masa corporar normalizando las características

α	Iteraciones	Exactitud	Precisión	Sensibilidad	F1 score	Tiempo
1.0	10	0.49	0.49	1.0	0.66	0.0010008811950683594
0.1	10	0.49	0.49	1.0	0.66	0.0
0.01	**10**	**0.69**	**0.8**	**0.47**	**0.59**	**0.001001119613647461**
0.001	10	0.51	0.5	0.76	0.6	0.00099945068359375
1.0	50	0.49	0.49	1.0	0.66	0.0009996891021728516
0.1	50	0.49	0.49	1.0	0.66	0.0010001659393310547
1.0	100	0.49	0.49	1.0	0.66	0.0009992122650146484
0.1	100	0.49	0.49	1.0	0.66	0.0009996891021728516
0.1	200	0.49	0.49	1.0	0.66	0.0019998550415039062

Tabla 5 resultados obtenidos con la insulina y la leptina normalizando las características

α	Iteraciones	Exactitud	Precisión	Sensibilidad	F1 score	Tiempo
1.0	10	0.34	0.34	1.0	0.51	0.0009996891021728516
0.1	10	0.34	0.34	1.0	0.51	0.0
0.01	10	0.46	0.18	0.17	0.17	0.0010004043579101562
0.001	10	0.34	0.34	1.0	0.51	0.0009989738464355469
1.0	50	0.34	0.34	1.0	0.51	0.0009999275207519531
0.1	50	0.34	0.34	1.0	0.51	0.0010001659393310547
0.01	50	0.69	0.52	0.92	0.66	0.002001047134399414
0.001	50	0.34	0.13	0.17	0.15	0.002000093460083008
1.0	100	0.34	0.34	1.0	0.51	0.00099945068359375
0.1	100	0.34	0.34	1.0	0.51	0.0010013580322265625
0.01	100	0.69	0.52	0.92	0.66	0.0010001659393310547
0.001	100	0.6	0.38	0.25	0.3	0.0009999275207519531
0.1	200	0.34	0.34	1.0	0.51	0.0029997825622558594
0.01	200	0.69	0.52	0.92	0.66	0.0030002593994140625
0.001	**200**	**0.71**	**0.58**	**0.58**	**0.58**	**0.0019998550415039062**

Tabla 6 resultados obtenidos con la HOMA y resistin normalizando las características

α	Iteraciones	Exactitud	Precisión	Sensibilidad	F1 score	Tiempo
1.0	10	0.43	0.43	1.0	0.6	0.001001119613647461
0.1	10	0.43	0.43	1.0	0.6	0.0009961128234863281
0.01	10	0.49	0.2	0.07	0.1	0.0009996891021728516
0.001	10	0.26	0.13	0.13	0.13	0.0
1.0	50	0.43	0.43	1.0	0.6	0.0009982585906982422
0.1	50	0.43	0.43	1.0	0.6	0.0020008087158203125
0.01	50	0.71	0.67	0.67	0.67	0.0009992122650146484
1.0	100	0.43	0.43	1.0	0.6	0.0029997825622558594
0.1	100	0.43	0.43	1.0	0.6	0.0020008087158203125

0.01	**100**	**0.74**	**0.67**	**0.8**	**0.73**	**0.0020003318786621094**
0.1	200	0.43	0.43	1.0	0.6	0.0029973983764648438
0.01	200	0.71	0.63	0.8	0.7	0.002998828887939453
0.001	200	0.66	1.0	0.2	0.33	0.0070002079010009766

Es importante destacar que se evidencia que los resultados obtenidos tienen una fuerte de los parámetros que fueron
sincronizados, mostrando que este proceso es fundamental a la hora de evaluar los algoritmos.

En la tabla 7 se pueden ver los resultados que se obtuvieron al realizar la ejecución repetitiva del algoritmo tomando los parámetros que se consideraron como los mejores (ver resultados de la tabla 2). Se evidencia que los resultados no fueron los mismos en cada una de las iteraciones, esto se debe a que el algoritmo se inicia de forma aleatoria, lo cual indica que los pesos que encuentra el gradiente descendente no siempre son los mismos, lo cual produce diferentes fronteras de decisión, y por ende diferentes resultados.

Tabla 7 resultados obtenidos con 20 iteraciones con la base de datos completa normaliza y los mejores hiperparametros

Iteración	Exactitud	Precisión	Sensibilidad	F1 score	Tiempo
0	0.63	0.62	0.33	0.43	0.0019991397857666016
1	0.57	0.5	0.4	0.44	0.0020017623901367188
2	0.46	0.36	0.33	0.34	0.002000570297241211
3	0.66	0.6	0.6	0.6	0.0009987354278564453
4	0.51	0.4	0.27	0.32	0.0020012855529785156
5	0.57	0.5	0.47	0.48	0.0010001659393310547
6	0.63	0.58	0.47	0.52	0.0019998550415039062
7	0.69	0.67	0.53	0.59	0.0010001659393310547
8	0.6	0.53	0.53	0.53	0.002000093460083008
9	0.66	0.62	0.53	0.57	0.0010006427764892578
10	0.63	0.56	0.6	0.58	0.002000570297241211
11	0.6	0.56	0.33	0.42	0.0010011119613647461
12	0.57	0.5	0.4	0.44	0.001999378204345703
13	0.63	0.57	0.53	0.55	0.0010004043579101562
14	0.63	0.55	0.73	0.63	0.0020003318786621094
15	0.6	0.6	0.2	0.3	0.0010001659393310547
16	0.66	0.6	0.6	0.6	0.002000570297241211
17	0.6	0.53	0.6	0.56	0.0010013580322265625
18	0.6	0.54	0.47	0.5	0.001999378204345703
19	0.6	0.55	0.4	0.46	0.0010001659393310547

IV. CONCLUSIÓN

Mediante la aplicación de algoritmos de perceptrón simple y su respectivo gradiente descendente para la identificación de los mejores pesos, se pudo evidenciar que se ven beneficiados de la aplicación de algoritmos de normalización y de la selección visual de características. Los mejores resultados se obtuvieron con una exactitud del 0.69 haciendo uso de toda la base de datos y la respectiva normalización. Igualmente se evidencio que al seleccionar HOMA y resistin la exactitud máxima obtenida fue de 0.74, mejorando los resultados obtenidos previamente.

Bibliografía

Bernardes, N. B., Sá, A. C. F. de, Facioli, L. de S., Ferreira, M. L., Sá, O. R. de, Costa, R. de M., Last Name1, F. N., & Last Name2, F. N. (2019). Câncer de Mama X Diagnóstico / Breast Cancer X Diagnosis. *ID on Line. Revista de Psicologia*, *13*(44), 877–885. https://doi.org/10.14295/IDONLINE.V13I44.1636

Comportamiento del cáncer de mama de la mujer en el período climatérico. (n.d.). Retrieved November 30, 2022, from http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=s0138-600x2006000300005

Cuevas, S. A. R., & García, M. C. (2006). Epidemiología del cáncer de mama. *Ginecología y Obstetricia de México*, *74*(11), 585–593. www.revistasmedicasmexicanas.com.mx