Glucosa e Insulina de la PTOG para Clasificar Sujetos con Síndrome Metabólico Usando K-means

Miguel Altuve
Facultad de Ingeniería Electrónica
Universidad Pontificia Bolivariana
Bucaramanga, Colombia
Teléfono: +57(7)679 62 20 Ext 203
Email: miguel.altuve@upb.edu.co

Erika Severeyn
Grupo de Bioingeniería y Biofísica Aplicada
Universidad Simón Bolívar
Caracas, Venezuela
Teléfono: +58 212 906 40 14
Email: severeynerika@usb.ve

Sara Wong
Investigador Prometeo, DEET
Universidad de Cuenca
Cuenca, Ecuador
Email: swong@usb.ve

***Resumen*—En este trabajo se realiza un aprendizaje no supervisado usando el algoritmo de agrupamiento K-means para realizar una clasificación de 15 sujetos diagnosticados con síndrome metabólico de acuerdo a sus valores de HOMA–IR y de glucosa e insulina, obtenidos de la prueba de tolerancia oral a la glucosa de cinco muestras. Estos sujetos fueron diagnosticados previamente como resistentes a la insulina o no, usando el índice de HOMA–IR con un punto de corte de 2,5. El objetivo de este trabajo es observar en estos sujetos si la clasificación no supervisada tiene una relación con la resistencia a la insulina. Los resultados obtenidos sugieren la utilización de un punto de corte mucho mayor del HOMA–IR para diagnosticar el síndrome metabólico y el uso de varias variables para realizar un diagnóstico más certero.**

***Palabras clave*—K-means, aprendizaje no supervisado, resistencia a la insulina, síndrome metabólico.**

I. Introducción

La obesidad, el síndrome metabólico, la diabetes y la hipertensión son enfermedades relacionadas con el estilo de vida que se han convertido en un problema social y de salud pública, especialmente en las ciudades modernas [1]. Consumir comida chatarra, la inactividad física y llevar un estilo de vida estresante son factores que contribuyen al desarrollo de estas enfermedades [2]. Una condición asociada con la obesidad, la prediabetes y la inactividad física es la resistencia a la insulina [3], en la que la insulina producida por el cuerpo no es utilizada eficazmente por las células, incrementándose la secreción de insulina para mantener la normoglicemia y la homeostasis de los lípidos. El diagnóstico precoz de la resistencia a la insulina es de vital importancia para prevenir la diabetes y sus complicaciones; adoptar un estilo de vida saludable es fundamental para evitar el desarrollo y la progresión de la enfermedad [4].

El clamp hiperinsulinémico-euglucémico [5] es el estándar de oro para evaluar la insulinorresistencia, sin embargo el modelo de evaluación homeostática para determinar la resistencia a la insulina (HOMA–IR, por sus siglas en inglés *homeostasis model assessment-estimated insulin resistance*) [6] es mucho más conveniente ya que es menos invasivo, a pesar de sus limitaciones de precisión [7]. Entre las limitaciones del índice HOMA–IR se pueden encontrar: i) su incapacidad de detectar la insulinorresistencia en sus estadios iniciales ya que las alteraciones de la insulina y la glucosa sanguínea en ayuno se manifiestan cuando las disfunciones metabólicas ya están presente en el organismo, ii) la alta sensibilidad en pacientes que ya presentan insulinorresistencia pero no en pacientes diabéticos con disfunción beta pancreática [8], y iii) la falta de un rango de referencia para el diagnóstico de la insulinorresistencia (un valor de 2,5 es ampliamente utilizado para la estimación de la resistencia a la insulina pero otros valores de corte han sido propuestos para poblaciones específicas [9]).

Tener en cuenta las variables antropométricas al momento de evaluar la sensibilidad a la insulina, podría ayudar en el diagnóstico precoz de la resistencia a la insulina [10]. Varios factores, como el índice de masa corporal (IMC), la edad, la presión arterial diastólica, colesterol HDL y LDL, y el índice HOMA–IR, han sido empleados para predecir la incidencia de de años del síndrome metabólico en una muestra masculina japonesa [11]. Además, utilizando máquinas de soporte vectorial y regresión logística Bayesiana, el IMC, el índice cintura/cadera, la glucosa en ayuno, los lípidos plasmáticos, las enzimas hepáticas y la hipertensión son factores asociados con un elevado índice HOMA–IR en una muestra hispana adulta [9].

El presente trabajo representa la continuación de estos trabajos previos, pero se centra en la clasificación no supervisada de pacientes con síndrome metabólico utilizando la información proporcionada por diez variables cuantitativas (cinco variables de insulina y cinco variables de glucosa) obtenidas de la prueba de tolerancia oral a la glucosa (PTOG). El objetivo es investigar si, utilizando los datos clínicos de insulina y glucosa en sangre, el agrupamiento automático obtenido con el algoritmo K-means está asociado a la resistencia a la insulina de los individuos, lo que pudiera conducir a una mejor comprensión de esta enfermedad.

El resto del artículo está organizado de la siguiente manera. En la siguiente sección se describen la metodología empleada para hacer el aprendizaje no supervisado usando el algoritmo de agrupamiento K-means y los datos de glucosa e insulina empleados en la estrategia de clasificación. Luego se presentan y se analizan los resultados de la clasificación obtenidos. Finalmente, las conclusiones del trabajo se detallan en la última

sección del manuscrito.

II. Metodología

II-A. Clasificación usando K-means

En el aprendizaje supervisado los objetos son asignados a una clase que ha sido previamente etiquetada teniendo en cuenta el conocimiento previo de las agrupaciones de los objetos, mientras que en el aprendizaje no supervisado los objetos son agrupados de acuerdo a una medida de similitud sin contar con un conocimiento acerca de cómo los datos pudieran ser agregados a los grupos. Este último enfoque de clasificación es empleado cuando los datos no están etiquetados y hay un interés en encontrar subgrupos en el conjunto de datos [12].

El aprendizaje no supervisado es generalmente el primer enfoque a ensayar en tareas de análisis exploratorio de datos, en clasificación y en predicción. Éste es particularmente útil hoy en día ya que el análisis manual de la gran cantidad de datos multidimensionales que son generados constantemente en diferentes aplicaciones y por diferentes medios, particularmente en el cuidado de la salud (véase, por ejemplo Microsoft HealthVault, PatientsLikeMe, IBM Watson), es particularmente complicado. En el campo de la ingeniería biomédica, el aprendizaje no supervisado ha sido empleado con éxito en la clasificación de las etapas del sueño utilizando señales electroencefalográficas [13], en la agrupación de los datos de expresión génica [14], en el descubrimiento de patrones de fenotipo [15], entre otros.

En este trabajo se utiliza el algoritmo K-means [16], uno de los algoritmos de agrupamiento más simples y de uso común, para realizar el aprendizaje no supervisado. El objetivo del algoritmo K-means es dividir M observaciones de N dimensiones (variables) en K grupos, de modo que la suma de cuadrados dentro de los grupos ($SDDG$) es minimizada. El procedimiento general consiste en mover de forma iterativa los puntos de un grupo a otro hasta encontrar una partición K con una $SDDG$ localmente óptima [17].

Dos experimentos se llevaron a cabo en este trabajo:

1. Experimento E1: se aplicó algoritmo K-means con $K = 2$ y $K = 3$ grupos usando el índice HOMA–IR índice como variable para dividir las M observaciones ($N = 1$ dimensión).
2. Experimento E2: se aplicó algoritmo K-means con $K = 2$ y $K = 3$ grupos usando los cinco datos de insulina y cinco de glucosa como variables para dividir las M observaciones ($N = 10$ dimensiones). Los datos de glucosa e insulina fueron normalizados para tener media cero y varianza unitaria, antes de ejecutar el algoritmo de K-means.

Para evitar mínimos locales, el algoritmo K-means se ejecutó 10 veces en cada experimento usando diferentes inicializaciones de los centroides de los grupos. Luego, se seleccionó la realización que arrojó la $SDDG$ total más baja. En cada ejecución, los grupos fueron inicializados utilizando el algoritmo K-means++ [18] y las distancias del punto al centroide fueron calculadas utilizando el cuadrado de la distancia Euclídea. Para cada ejecución, el número máximo de iteraciones del algoritmo K-means se estableció en 100. El coeficiente de silueta (CS) fue utilizado para evaluar la asignación de los datos a los grupos (mientras más alto el CS mejor) [19].

II-B. Datos de insulina y glucosa

Los datos fueron obtenidos de $M = 15$ sujetos (observaciones o instancias) a los que se les realizó una PTOG de cincos muestras en el Hospital Universitario de Caracas, Venezuela. Durante la PTOG, los niveles de insulina (I) y glucosa (G) fueron determinados en cinco muestras diferentes de sangre: una muestra en ayuno (I_0 y G_0) y cuatro muestras más ($I_1, \ldots, I_4$ y $G_1, \ldots, G_4$) después de la ingesta oral de 75 g de glucosa, a intervalos de 30 minutos cada una (minutos 30, 60, 90 y 120) [7].

El protocolo del estudio se adhirió a los principios de la Declaración de Helsinki y fue aprobado por el Comité Ético del Hospital Universitario de Caracas; todos los sujetos dieron su consentimiento informado por escrito. Los sujetos del estudio fueron diagnosticados con síndrome metabólico de acuerdo a [20].

La sensibilidad a la insulina se determinó usando el método HOMA–IR que se describe en la ecuación 1, en donde I_0 (μUI/ml) es la insulina plasmática en ayuno y G_0 (mg/dl) es la glucosa plasmática en ayuno.

$$\mathrm{IS_H} = \frac{G_0\, I_0}{405} \tag{1}$$

Los sujetos fueron clasificados como resistencia a la insulina VERDADERO/FALSO usando el índice HOMA–IR: VERDADERO para $\mathrm{Is_H} > 2{,}5$. De acuerdo con esto, de los 15 sujetos, 8 ($53{,}33\,\%$) fueron diagnosticados como $\mathrm{IR_V}$ (resistencia a la insulina VERDADERO) y 7 ($46{,}66\,\%$) como $\mathrm{IR_F}$ (resistencia a la insulina FALSO). Esta información no fue utilizada por el algoritmo de K-means, sino durante la interpretación de los resultados de la agrupación automática.

La tabla I muestra la media, la mediana y el rango de valores de las variables estudiadas en este trabajo.

III. Resultados

Los resultados se dividen en dos partes, de acuerdo con los experimentos detallados en la sección II-A.

III-A. Experimento E1

La figura 1 muestra la asignación de los individuos a los grupos y la disposición de los centroides, para $K = 2$ y $K = 3$ grupos, usando el índice HOMA–IR como observación.

Usando $K = 2$ grupos (CS $= 0{,}8247 \pm 0{,}15$), el agrupamiento automático con K-means dividió la muestra en once sujetos en el G_1 (círculos de color rojo en la figura 1) y cuatro sujetos en el G_2 (círculos de color azul en la figura 1). De acuerdo al punto de corte usado (HOMA–IR$> 2{,}5$), cuatro sujetos de G_1 tenían resistencia a la insulina y 7 no, mientras que todos los sujetos de G_2 tenían resistencia a la insulina. Para $K = 3$ grupos (SC $= 0{,}844 \pm 0{,}129$), dos sujetos conformaron G_1

Tabla I
CARACTERÍSTICAS DE REFERENCIA DE LOS SUJETOS.

Variable	Media	Mediana	Rango
Edad (años)	31,40	33	18–44
HOMA–IR	3,15	2,58	0,53–6,19
G_0 (mg/dL)	105,66	103	95–119
G_1 (mg/dL)	165	161	126–202
G_2 (mg/dL)	166,93	165	101–231
G_3 (mg/dL)	149,06	146	116–173
G_4 (mg/dL)	129,46	131	99–164
I_0 (μUI/mL)	12,13	11	2–25
I_1 (μUI/mL)	100,93	70	21–300
I_2 (μUI/mL)	117,46	70	15–300
I_3 (μUI/mL)	111,13	75	27–300
I_4 (μUI/mL)	100,60	94	13–300

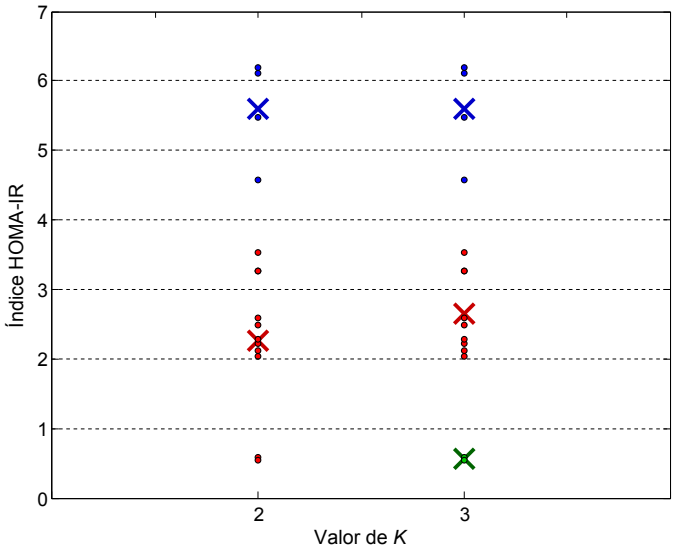

Figura 1. Asignación de los individuos (círculos) a los grupos para $K = 2$ y $K = 3$, usando el índice HOMA–IR como observación. Para $K = 2$, los círculos de color rojo conforman el grupo G_1 y los círculos de color azul conforman el G_2, mientras que para $K = 3$, los círculos de color verde conforman el grupo G_1, los de color rojo el G_2 y los de color azul el G_3. Los centroides de los grupos están representados por el carácter X.

(círculos de color verde en la figura 1), nueve sujetos el G_2 (círculos de color rojo en la figura 1) y cuatro sujetos el G_3 (círculos de color azul en la figura 1). Los sujetos de G_1 no tenía insulinorresistencia de acuerdo al punto de corte usado, el G_2 agrupó a cinco sujetos sin insulinorresistencia y cuatro con insulinorresistencia, y los cuatro sujetos de G_3 presentaron insulinorresistencia. Estos resultados se resumen en la tabla II.

III-B. Experimento E2

Usando las observaciones multidimensionales, el agrupamiento con K-means con $K = 2$ grupos (CS $= 0{,}545 \pm 0{,}149$) dividió a los sujetos en G_1 con once sujetos de los cuales cuatro tenían resistencia a la insulina y en G_2 con cuatro sujetos, todos con resistencia a la insulina. Para $K = 3$ grupos (SC $= 0{,}570 \pm 0{,}191$), un sujeto conformó el G_1,

Tabla II
RESULTADOS DEL APRENDIZAJE NO SUPERVISADO UTILIZANDO EL ÍNDICE HOMA–IR ($N = 1$ DIMENSIÓN). EL VALOR DE CS SE EXPRESA COMO MEDIA $\pm$ DESVIACIÓN ESTÁNDAR, m_{G_i} CORRESPONDE AL NÚMERO DE SUJETOS EN EL GRUPO G_i.

Parámetro	$K = 2$	$K = 3$
CS	$0{,}8247 \pm 0{,}154$	$0{,}844 \pm 0{,}129$
m_{G_1}	11 ($4 \in \text{IR}_\text{V}$)	2 ($0 \in \text{IR}_\text{V}$)
m_{G_2}	4 ($4 \in \text{IR}_\text{V}$)	9 ($4 \in \text{IR}_\text{V}$)
m_{G_3}		4 ($4 \in \text{IR}_\text{V}$)

Tabla III
RESULTADOS DEL APRENDIZAJE NO SUPERVISADO UTILIZANDO LOS DATOS DE INSULINA Y GLUCOSA ($N = 10$ DIMENSIONES). EL VALOR DE CS SE EXPRESA COMO MEDIA $\pm$ DESVIACIÓN ESTÁNDAR, m_{G_i} CORRESPONDE AL NÚMERO DE SUJETOS EN EL GRUPO G_i.

Parámetro	$K = 2$	$K = 3$
CS	$0{,}545 \pm 0{,}149$	$0{,}570 \pm 0{,}191$
m_{G_1}	11 ($4 \in \text{IR}_\text{V}$)	1 ($1 \in \text{IR}_\text{V}$)
m_{G_2}	4 ($4 \in \text{IR}_\text{V}$)	10 ($3 \in \text{IR}_\text{V}$)
m_{G_3}		4 ($4 \in \text{IR}_\text{V}$)

diez sujetos el G_2 y cuatro sujetos el G_3. El sujeto del G_1 tenía insulinorresistencia, el G_2 agrupó a siete sujetos sin insulinorresistencia y tres con insulinorresistencia, y los cuatro sujetos de G_3 presentaron insulinorresistencia. Estos resultados se resumen en la tabla III.

IV. DISCUSIÓN

Los resultados obtenidos en este trabajo son bastante interesantes. Se pudo observar que el aprendizaje no supervisado con K-means para $K = 2$ grupos, usando como observaciones cuantitativas tanto el índice HOMA–IR (unidimensional) como los datos de glucosa e insulina de la PTOG (multidimensional), no divide a los sujetos con síndrome metabólico como insulinorresistentes VERDADERO/FALSO de la misma manera que el índice HOMA–IR con un punto de corte de 2,5 sugiere. Este resultado está en concordancia con el trabajo presentado por Qu *et al*. quienes encontraron experimentalmente un punto de corte mayor (HOMA–IR $> 3{,}8$) para la detección de la resistencia a la insulina en Mexicanos [9].

De hecho, para $K = 2$ grupos, los sujetos fueron clasificados de manera idéntica en ambos experimentos, tal como se observa en la tablas II y III. Este resultado pone en evidencia la fortaleza del algoritmo K-means para realizar un aprendizaje no supervisado usando tanto datos unidimensionales como multidimensionales y resalta el hecho de que en la muestra hay cuatro individuos (pertenecientes a $G2$) que no comparten las mismas características que el resto de los individuos, y estos individuos efectivamente presentan la resistencia a la insulina. Igualmente, la clasificación usando datos unidimensionales y multidimensionales evidencia que los valores de glucosa e insulina de la PTOG caracterizan de la misma manera que usar una relación de glucosa e insulina basal, como es el caso del índice HOMA–IR, por tanto para este tipo de población,

en donde ya existe una enfermedad metabólica pre-existente (síndrome metabólico), el uso de todos los valores de glucosa e insulina de la PTOG pareciera ser redundante cuando se trata de clasificar los sujetos con resistencia a la insulina. Caso contrario ocurre cuando se estudian poblaciones donde no hay una enfermedad metabólica aparente, podría presentarse una resistencia a la insulina incipiente que solo es posible detectar en los valores de glucosa e insulina a los 120 minutos de la PTOG.

Usando $K = 2$ grupos, el valor mínimo observado del índice HOMA–IR en los individuos del grupo 2 fue de $4{,}58$, mientras que el valor máximo del HOMA–IR en los individuos del grupo 1 fue de $3{,}52$. Sin embargo, dada la diferencia entre estos dos valores y el tamaño de la muestra ($M = 15$) del presente estudio, no podemos indicar cuál sería el valor de corte apropiado del índice HOMA–IR para diagnosticar la resistencia a la insulina en esta población. No obstante se puede observar que, en sujetos con síndrome metabólico, un HOMA–IR $> 3{,}6$ pudiera ser útil para diagnosticar estos pacientes con resistencia a la insulina. Analizar un mayor número de individuos pudiera ser de gran ayuda para encontrar el punto de corte óptimo del HOMA–IR en esta población y en poder diferenciar los individuos de manera distinta usando tanto datos unidimensionales (como los del experimento E1) como multidimensionales (como los del experimento E2).

En el caso de $K = 3$ los resultados de los experimentos E1 y E2 no fueron idénticos. En el caso del experimento E1, en comparación con $K = 2$ el algoritmo de agrupamiento creó un grupo adicional para agrupar a los dos sujetos con más alta sensibilidad a la insulina (HOMA–IR más bajo), tal como se muestra en la figura 1, mientras que en el caso del experimento E2, los resultados del agrupamiento no parecieran tener relación con el índice HOMA–IR, ya que de hecho los individuos con resistencia a la insulina fueron repartidos en los tres grupos (G_1, G_2 y G_3). El individuo del grupo G_1 presenta una curva de glucosa normal, pero la curva de insulina se mantiene en ascenso durante toda la PTOG. A pesar de que el HOMA–IR no pareciera indicar una baja sensibilidad a la insulina en este individuo, el hecho de tener una insulina muy elevada al final de la PTOG es signo de una resistencia a la insulina en sus estadios iniciales. Por otro lado, el agrupamiento de los cuatro individuos en G_3 en este experimento corresponden a los mismos individuos obtenidos en el experimento E1 para $K = 3$ y usando $K = 2$ en los experimentos E1 y E2. Este hecho confirma que esos cuatro individuos tienen características diferentes y que son separados del resto de la muestra tanto usando datos unidimensionales como multidimensionales empleando tanto dos como tres grupos.

V. Conclusiones

En este trabajo se utilizó el algoritmo K-means para agrupar sujetos con síndrome metabólico en dos y tres grupos, usando como observaciones tanto el índice HOMA–IR como los valores de glucosa e insulina de la PTOG, con el objetivo de relacionar los grupos con la resistencia a la insulina.

Los resultados obtenidos en este trabajo muestran que es probable que el diagnóstico de resistencia a la insulina usando un valor de corte de HOMA–IR $> 2{,}5$ no sea el adecuado en estos individuos ya que la separación automática de los individuos arrojó un valor de corte que se pudiera encontrar en rango 3.5–4.5, sin embargo, es necesario realizar un estudio con una muestra mucho mayor para encontrar el valor apropiado.

Además, sería interesante agrupar a los sujetos en tres grupos distintos y que uno de esos grupos pudiera indicar la predisposición a desarrollar la resistencia a la insulina. Para ello sería necesario repetir el experimento con $K = 3$ sobre una muestra mucho mayor.

Es interesante observar que los resultados del agrupamiento con el algoritmo K-means fueron obtenidos empleando cinco datos de glucosa y cinco de insulina para cada persona mientras que el diagnóstico de resistencia a la insulina usando el HOMA–IR solo toma en cuenta los valores de glucosa e insulina en ayuno. El contar con más datos de un mismo individuo permite tener una mejor información del sistema que se está observando y sirve de ayuda en la toma de decisiones, sin embargo tomar una decisión usando datos multidimensionales puede ser una tarea complicada para un médico especialista pero no para un algoritmo computacional como el K-means. Efectivamente será el médico quien tendrá la decisión de aceptar o rechazar cualquier diagnóstico automático.

Agradecimientos

S. Wong agradece el patrocinio del Proyecto Prometeo de la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador.

Referencias

[1] F. B. Hu, "Globalization of Diabetes The role of diet, lifestyle, and genes," *Diabetes care*, vol. 34, no. 6, pp. 1249–1257, 2011.

[2] B. M. Popkin, L. S. Adair, and S. W. Ng, "Global nutrition transition and the pandemic of obesity in developing countries," *Nutrition reviews*, vol. 70, no. 1, pp. 3–21, 2012.

[3] S. M. Grundy, "Pre-diabetes, metabolic syndrome, and cardiovascular risk," *Journal of the American College of Cardiology*, vol. 59, no. 7, pp. 635–643, 2012.

[4] F. Abbasi, B. W. Brown, C. Lamendola, T. McLaughlin, and G. M. Reaven, "Relationship between obesity, insulin resistance, and coronary heart disease risk," *Journal of the American College of Cardiology*, vol. 40, no. 5, pp. 937–943, 2002.

[5] M. Greenfield, L. Doberne, F. Kraemer, T. Tobey, and G. Reaven, "Assessment of insulin resistance with the insulin suppression test and the euglycemic clamp," *Diabetes*, vol. 30, no. 5, pp. 387–392, 1981.

[6] D. Matthews, J. Hosker, A. Rudenski, B. Naylor, D. Treacher, and R. Turner, "Homeostasis model assessment: insulin resistance and β-cell function from fasting plasma glucose and insulin concentrations in man," *Diabetologia*, vol. 28, no. 7, pp. 412–419, 1985.

[7] M. Altuve, E. Severeyn, and S. Wong, "Adaptation of five indirect insulin sensitivity evaluation methods to three populations: Metabolic syndrome, athletic and normal subjects," in *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*. IEEE, 2014, pp. 4555–4558.

[8] R. Muniyappa, S. Lee, H. Chen, and M. J. Quon, "Current approaches for assessing insulin sensitivity and resistance in vivo: advantages, limitations, and appropriate usage," *Am J Physiol Endocrinol Metab*, vol. 294, no. 1, pp. E15–E26, 2008.

[9] H.-Q. Qu, Q. Li, A. R. Rentfro, S. P. Fisher-Hoch, and J. B. McCormick, "The definition of insulin resistance using HOMA-IR for Americans of Mexican descent using machine learning," *PloS one*, vol. 6, no. 6, p. e21041, 2011.

[10] E. Severeyn, S. Wong, H. Herrera, and M. Altuve, "Anthropometric measurements for assessing insulin sensitivity on patients with metabolic syndrome, sedentaries and marathoners," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015.

[11] H. Hirose, T. Takayama, S. Hozawa, T. Hibi, and I. Saito, "Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin," *Computers in biology and medicine*, vol. 41, no. 11, pp. 1051–1056, 2011.

[12] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2001.

[13] S. Güneş, K. Polat, and c. Yosunkaya, "Efficient sleep stage recognition system based on EEG signal using k-means clustering based feature weighting," *Expert Systems with Applications*, vol. 37, no. 12, pp. 7922–7928, 2010.

[14] L. J. van't Veer, H. Dai, M. J. Van De Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[15] T. A. Lasko, J. C. Denny, and M. A. Levy, "Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data," *PloS one*, vol. 8, no. 6, p. e66341, 2013.

[16] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[17] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied statistics*, pp. 100–108, 1979.

[18] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[19] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[20] S. M. Grundy, H. B. Brewer, J. I. Cleeman, S. C. Smith, and C. Lenfant, "Definition of metabolic syndrome report of the National Heart, Lung, and Blood Institute/American Heart Association Conference on scientific issues related to definition," *Circulation*, vol. 109, no. 3, pp. 433–438, 2004.