JRHS 2018; 18(2): e00412

JRHS
Revista de Investigación en Ciencias de la Salud
página de inicio de la revista: www.umsha.ac.ir/jrhs

Artículo original

Una comparación entre el árbol de decisión y el bosque aleatorio para determinar los factores de riesgo asociados con la diabetes tipo 2

Habibollah Esmaily (PhD)[1], Maryam Tayefi (PhD)[2], Hassan Doosti (PhD)[3], Majid Ghayour-Mobarhan (MD, PhD)[4,5], Hossein Nezami (MSc)[6], Alireza Amirabadizadeh (MSc)[7]*

[1]*Centro de Investigación de Determinantes Sociales de la Salud, Universidad de Ciencias Médicas de Mashhad, Mashhad, Irán*
[2]*Unidad de Investigación Clínica, Universidad de Ciencias Médicas de Mashhad, Mashhad, Irán*
[3]*Departamento de Estadística, Universidad Macquarie, Sydney, NSW, Australia*
[4]*Centro de Investigación de Bioquímica de la Nutrición, Facultad de Medicina, Universidad de Ciencias Médicas de Mashhad, Mashhad, Irán*
[5]*Departamento de Ciencias y Tecnologías Modernas, Facultad de Medicina, Universidad de Ciencias Médicas de Mashhad, Mashhad, Irán*
[6]*Departamento de Ciencias Básicas, Facultad de Medicina, Universidad de Ciencias Médicas de Gonabad, Gonabad, Irán*
[7]*Centro de Investigación de Toxicología Médica y Abuso de Drogas (MTDRC), Universidad de Ciencias Médicas de Birjand, Birjand, Irán*

INFORMACIÓN DEL ARTÍCULO

Historial del artículo:
Recibió:24 diciembre 2017
Revisado:03 abril 2018 Aceptado:
17 abril 2018 Disponible en linea:
24 abril 2018

Palabras clave:
Diabetes mellitus
Árbol de decisión
Bosque aleatorio
procesamiento de datos
Irán

***Correspondencia**
Ali Reza Amirabadizadeh (MSc)
Teléfono:+98 971 785 3577
Correo electrónico:amirabadiza921@gmail.com

RESUMEN

Fondo: Nuestro objetivo fue identificar los factores de riesgo asociados de la diabetes mellitus tipo 2 (T2DM) utilizando un enfoque de minería de datos, árboles de decisión y técnicas de bosque aleatorio utilizando el programa de estudio Mashhad Stroke and Heart Atherosclerotic Disorders (MASHAD).

Diseño del estudio: Un estudio transversal.

Métodos: El estudio MASHAD comenzó en 2010 y continuará hasta 2020. Se utilizan dos herramientas de extracción de datos, a saber, árboles de decisión y bosques aleatorios, para predecir la DM2 cuando se observan otras características en 9528 sujetos reclutados de la base de datos MASHAD. Este artículo hace una comparación entre estos dos modelos en términos de precisión, sensibilidad, especificidad y el área bajo la curva ROC.

Resultados: La tasa de prevalencia de T2DM fue del 14% entre estos sujetos. El modelo de árbol de decisión tiene una precisión del 64,9 %, una sensibilidad del 64,5 %, una especificidad del 66,8 % y un área bajo la curva ROC del 68,6 %, mientras que el modelo de bosque aleatorio tiene una precisión del 71,1 %, una sensibilidad del 71,3 %, una especificidad del 69,9 % y un área bajo la curva ROC curva que mide 77.3% respectivamente.

Conclusiones: El modelo de bosque aleatorio, cuando se utiliza con mediciones demográficas, clínicas, antropométricas y bioquímicas, puede proporcionar una herramienta sencilla para identificar los factores de riesgo asociados a la diabetes tipo 2. Tal identificación puede utilizarse sustancialmente para la gestión de la política de salud para reducir el número de sujetos con DM2.

Citación: Esmaily H, Tayefi M, Doosti H, Ghayour-Mobarhan M, Nezami H, Amirabadizadeh AR. Una comparación entre el árbol de decisiones y el bosque aleatorio para determinar los factores de riesgo asociados con la diabetes tipo 2. J Res Salud Sci. 2018; 18(2): e00412.

Introducción

La diabetes mellitus tipo 2 (DM2) es un importante problema de salud pública y su mortalidad está aumentando en todo el mundo [1,2]. La OMS predice que la prevalencia de DM2 en Irán será 6.8% en 2025, y esto se traduce en 5215000 ciudadanos de Irán[3].

Los resultados de la cohorte de Teherán muestran que la prevalencia de DM2 tipo en Irán es del 11 %[4]y la cohorte de Mashhad afirma que esta prevalencia es del 14%[5].

T2DM es uno de los desafíos más serios para los países en desarrollo en los 21st siglo[6,7]. La diabetes tiene sus raíces en las interacciones entre las características genéticas, ambientales y conductuales.[8,9]. Las enfermedades cardiovasculares en particular son responsables del 80% de las muertes por DM2[10]. Los posibles factores de riesgo dominantes en el desarrollo de la DM2 son el origen étnico, la obesidad, la dieta poco saludable, la falta de actividad física, la resistencia a la insulina y los antecedentes familiares de diabetes.[11]. Las enfermedades cardíacas, los accidentes cerebrovasculares, la ceguera, las enfermedades renales y las amputaciones están asociadas con la diabetes.[12]. Por lo tanto, es fundamental identificar y diagnosticar a las personas que corren un alto riesgo de DM2[6 13].

En las últimas décadas, diferentes investigadores en Irán han utilizado métodos de minería de datos como el árbol de decisión, la red neuronal, la máquina de vectores de soporte, el bosque aleatorio para predecir los factores de riesgo asociados con la DM2.[5,14]. Una razón para no utilizar el método estadístico clásico es el número de predictores que los métodos clásicos no pueden seleccionar convenientemente. Estos dos modelos, árbol de decisión y bosque aleatorio son dos de los modelos de clasificación y no hay tantos estudios al respecto.

La minería de datos es una nueva colección de métodos estadísticos utilizados para características significativamente asociadas con T2DM[15,16]. La minería de datos puede descubrir nuevos factores y también encontrar relaciones entre factores que pueden revelar patrones y desarrollar predicciones basadas en nuevos factores asociados con DMT2[17,18].

No hay muchos estudios sobre los factores de riesgo asociados a la DM2 utilizando algoritmos de minería de datos en el este de Irán hasta el momento. En este estudio, desarrollamos el modelo predictivo para

identificar los factores de riesgo asociados de T2DM como complemento en la detección y la salud pública en el este de Irán.

Métodos

Participantes

El estudio MASHAD comenzó en 2010 y continuará hasta 2020. La ciudad de Mashhad se encuentra en la parte nororiental de Irán. La población total de Mashhad se estimó utilizando el censo nacional iraní de 2006, por lo que el tamaño de la muestra se determinó en consecuencia. Los participantes procedían de tres regiones de Mashhad. Cada región se dividió en nueve sitios centrados en las divisiones del Mashhad Healthcare Center. En general, 9528 sujetos se inscribieron como parte del estudio MASHAD[19].

Este protocolo fue aprobado por el Comité de Ética de MUMS y se obtuvo un consentimiento informado por escrito de cada participante.

Se recogieron de todos los sujetos características demográficas como edad, sexo, estado civil, educación, hábito de fumar cigarrillos, nivel de actividad física (PAL), antecedentes familiares de diabetes (FHD) y puntuación de depresión. Se utilizó el inventario de depresión de Beck-II (BDI-II) para evaluar la depresión. Se obtuvo información antropométrica incluyendo peso, talla, circunferencia de cintura y cadera. Las presiones arteriales sistólica y diastólica se midieron como se describió anteriormente.[19]. Los parámetros bioquímicos incluyeron: triglicéridos séricos en ayunas (TG), colesterol total (TC), colesterol HDL y colesterol LDL, glucosa en sangre en ayunas (FBG) y hs-CRP se midieron como se describió anteriormente[19]. La DM2 diagnosticada se identificó en base a glucosa en sangre en ayunas (FBG) ≥126 mg/dl[20].

Variables de entrada

Los datos finales contienen 9528 registros y 18 variables, divididos en 17 variables predictoras y una variable de resultado o objetivo. La variable objetivo tiene dos estados posibles, a saber, ocurrencia de T2DM o no ocurrencia de T2DM. Las características demográficas incluyeron edad, sexo, índice de masa corporal (IMC), estado civil, nivel de educación y marcadores bioquímicos, nivel de actividad física (PAL), hábitos de tabaquismo, antecedentes familiares de diabetes (FHD) y puntuación de depresión se consideraron como predictores (Tabla 1-2).

Modelo de árbol de decisión

Un árbol de decisión es un método no paramétrico nombrado de acuerdo con la naturaleza de la variable objetivo. Se llama árbol de clasificación si la variable objetivo es categórica y árbol de regresión si la variable objetivo es continua. El propósito de un árbol de decisión es desarrollar un modelo predictivo en términos de variables predictoras. El árbol se forma dividiendo sucesivamente los datos según una de las variables predictoras. Un árbol de decisión consta de tres tipos de nodos: nodo raíz, nodos internos y nodos hoja.[21-23]. Los algoritmos de árboles de decisión desarrollan criterios de división en los nodos internos del árbol. La división de un nodo intenta minimizar la impureza del nodo. Si una división no puede lograr ninguna mejora en términos de reducción de impurezas, el nodo no se divide y se declara como nodo hoja. Si una división es capaz de reducir la impureza, entonces se selecciona la división que proporciona la máxima reducción de impurezas y se forman dos ramas, formando dos nuevos nodos. Los criterios populares de división son la ganancia de información, el índice de Gini y la relación de ganancia. CART es uno de los algoritmos de árbol de decisión que construya un árbol binario utilizando el índice de Gini para seleccionar la variable de división en cada nodo interno. El índice de Gini en un nodo D está dado por

$$\text{Gini}(D) = 1 - \sum_{i=1}^{\text{metro}} P_i^2$$

donde p_i es la probabilidad de que una observación en D pertenezca a la clase Ci y se estima mediante |Ci, D|/|D|[24, 25]. La suma se toma sobre las clases posibles. El árbol comienza con todas las observaciones que forman el nodo raíz y las divisiones sucesivas determinan el orden de importancia de las variables predictoras.

Tabla 1: Comparación de las características basales entre los grupos de diabetes y no diabetes

Variables	Número	Por ciento	Número	Por ciento	*PAGS*valor
Sexo					0.040
Masculino	518	38.1	3277	40.1	
Femenino	843	61,9	4890	59,9	
Nivel educacional					0.001
Alto	109	8.0	936	11.5	
Moderado	374	27.5	2912	35.7	
Bajo	878	64.5	4319	52,9	
Estatus laboral					0.001
empleado	400	29.4	3114	38.1	
Jubilado	178	13.1	755	9.2	
Estudiantes	0	0.0	20	0.2	
desempleado	783	57.5	4278	52.4	
Estado civil					0.001
Casado	1239	91.0	7636	93.5	
Único	5	0.4	54	0.7	
Viuda	96	7.1	366	4.5	
Divorciado	21	1.5	111	1.4	
Tabaquismo					0.050
Sí	272	20.0	1775	21.7	
No	1089	80.0	6392	78.3	
Antecedentes familiares de diabetes					0.001
Sí	647	47.5	1994	24.4	
No	714	52.5	6173	75,6	
Depresión					0.001
Sí	461	33,9	2226	27.3	
No	900	66.1	5941	27.7	

Tabla 2: Comparación de marcadores bioquímicos entre grupos diabéticos y no diabéticos

	Diabetes		no diabéticos		
Variables	Significar	Dakota del Sur	Aguamiel	Dakota del Sur	*PAGS*valor
Presión arterial sistólica (mmHg)	128,8	18,4	121,1	18.2	0.001
Presión arterial diastólica (mmHg)	81,4	10,4	78,9	11,1	0.001
Colesterol total (mg/dl)	205.5	46,3	189,7	37.8	0.001
Lipoproteína de baja densidad (mg/dl)	122.5	39,1	115,7	34.6	0.001
Lipoproteína de alta densidad (mg/dl)	41.8	9.6	42,7	9,9	0.004
Triglicéridos (mg/dl)	160,0	122,0	117,0	83,0	0.001
Alta sensibilidad -CRP	2,7	4,34	1,6	2,3	0.002

Bosque aleatorio

Random forest es un método de aprendizaje conjunto. Genera muchos árboles de clasificación seleccionando subconjuntos del conjunto de datos dado y seleccionando aleatoriamente subconjuntos de variables predictoras, agregando finalmente los resultados de todos los modelos para obtener un bosque aleatorio. Se obtienen múltiples árboles de clasificación a partir de muestras de arranque para llegar a las reglas finales de clasificación "mayoritarias". Los parámetros de entrenamiento de árboles usados en el presente estudio son (i) ntree=500, el número de árboles generados (ii) ntry=17, el número de variables predictoras usadas en cada árbol, y (iii) tamaño de nodo=5, el mínimo número de observaciones en un nodo hoja. Los algoritmos de aprendizaje automático supervisados dividen los datos en dos partes, a saber, datos de entrenamiento y datos de prueba.

Una de las características más importantes del bosque aleatorio y el árbol de decisión es la salida de la importancia de la variable. La importancia de la variable mide el grado de asociación entre una determinada variable y el resultado de la clasificación. El bosque aleatorio y el árbol de decisiones tienen cuatro medidas para la importancia variable: puntuación de importancia bruta para la clase 0, puntuación de importancia bruta para la clase 1, disminución de la precisión y el índice de Gini[26].

Los análisis estadísticos se realizaron utilizando los paquetes R rpart (para árboles de decisión), Random Forest (para Random Forest) y Caret. La muestra completa contenía 1361 individuos con DM2 y los restantes 8167 individuos sin DM2. El presente estudio adoptó un método de validación cruzada de 10 veces para evaluar el árbol de decisión y el modelo de bosque aleatorio. El método de validación cruzada de 10 veces implica la separación aleatoria de los conjuntos de datos adquiridos en 10 conjuntos de datos que tienen el mismo tamaño de muestra. Los modelos de árbol de decisión y bosque aleatorio se construyen sobre la base de un conjunto de datos de entrenamiento. El resto de los nueve conjuntos de datos se utilizaron como datos de prueba para verificar la eficacia del modelo. Se realizaron diez pruebas empíricas repetidas, donde cada subconjunto se utilizó como datos de prueba.

El árbol de decisión desarrollado sobre los datos de entrenamiento se usó para obtener los criterios de división para diferentes nodos y luego se aplicó a la observación en los datos de prueba. El árbol resultante se utiliza para medir la sensibilidad, la especificidad y la precisión del modelo. Si los valores de estas medidas son altos para los datos de entrenamiento y bajos para los datos de prueba, se considera un caso de sobreajuste. Estas medidas deben obtenerse en datos de entrenamiento así como en datos de prueba para establecer la validez del modelo. Los modelos informados en este documento han sido validados y los resultados de los datos de prueba se informan aquí.

Los modelos se evalúan mediante la construcción de la matriz de confusión para los datos de prueba. Además, la precisión, la sensibilidad y la especificidad también se miden para cada modelo. La precisión, sensibilidad y especificidad de un modelo de clasificación se definen de la siguiente manera[27].

- Precisión=(TP+TN)/(TP+FP+TN+FN)
- Sensibilidad=TP/(TP+FN)
- Especificidad= TN/(FP+TN)

Aquí TP, TN, FP y FN son verdadero positivo, verdadero negativo, falso positivo y falso negativo respectivamente.

La curva característica operativa del receptor (ROC) es el gráfico que muestra la imagen completa de la compensación entre la sensibilidad y (1-especificidad) a través de una serie de puntos de corte. El área bajo la curva ROC se considera una medida eficaz de la validez inherente de una prueba diagnóstica.

Resultados

Las características antropométricas y bioquímicas se resumen en la Tabla 1 y 2, respectivamente. En general, 1361 (14,3%) personas tenían DM2. De 1361 diabéticos, 843 (61,9%) eran mujeres, 1239 (91%) estaban casados y 783 (57,5%) estaban desempleados. Los sujetos con DM2 mostraron una presión arterial sistólica, triglicéridos, hs-CRP, diastólica significativamente más alta. presión arterial, colesterol total sérico y colesterol LDL, mientras que mostraron un colesterol HDL significativamente más bajo que los sujetos sin DM2. La edad media de los diabéticos fue mayor que la de los no diabéticos (52,01 ±7,2 vs 47,70 ±8,1, *PAGS*<0,001). El IMC medio de los pacientes diabéticos fue de 28,78 ± 4,4 y de los no diabéticos de 27,76 ± 4,7. Los resultados de la independiente *t*-la prueba mostró que el IMC en los diabéticos era significativamente más alto que en las personas no diabéticas (*PAGS*<0,001). La PAL media de los diabéticos fue menor que la de los no diabéticos (1,59 ±0,86 frente a 1,60 ±0,64, *PAGS*=0,040).

Según los resultados del modelo de bosque aleatorio, TG, hs-CRP, SBP, LDL, TC, FHD, edad, BMI y PAL fueron los factores de riesgo más importantes relacionados con la DM2 (Figura 1). en un subgrupo con TG>204.5 y PCR-hs≤1.32 y ocupación=empleo, 79,2% fue la probabilidad de no ocurrencia de DM2. En el subgrupo con TG>204,5 y PCR-hs<1,32 y ocupación=desempleo y PCR-hs>4,66, la probabilidad de ocurrencia de DM2 es del 90% (tabla 3).

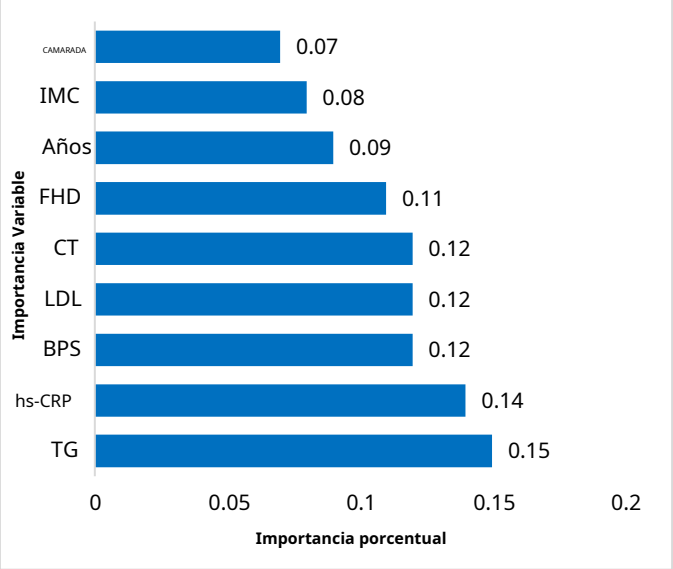
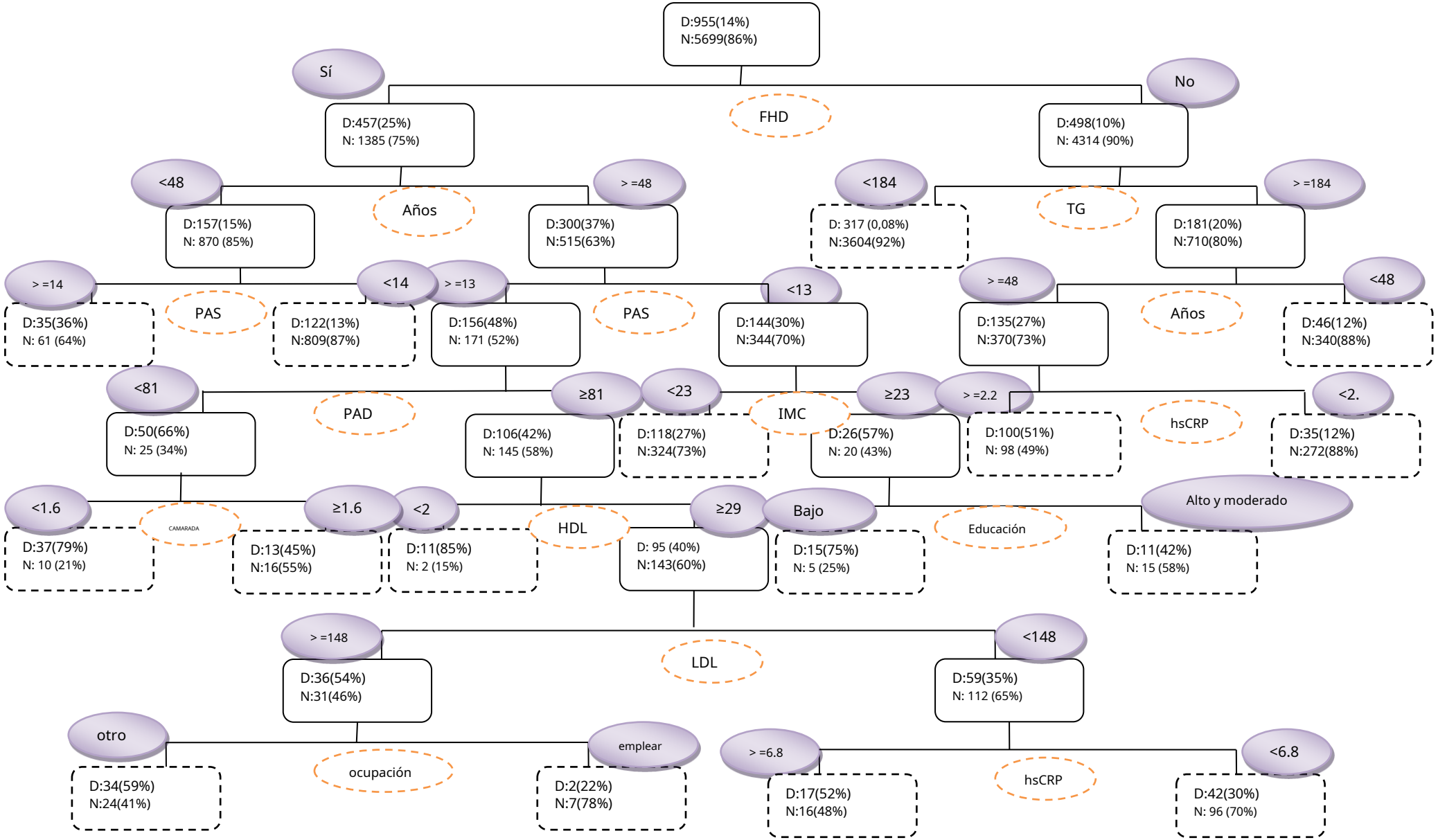


Figura 1: La importancia de las variables de entrada en el modelo Random Forest. El eje X muestra el porcentaje de importancia y el eje Y representa la importancia de las variables

Según los resultados del modelo de árbol de decisiones, FHD, edad, TG, SBP, hs-CRP, IMC y DBP fueron los factores de riesgo más importantes relacionados con T2DM. La Figura 2 muestra el árbol completo producido por CART. El árbol de decisión mostró que en un subgrupo con FHD=no y TG<184, el 92% es la probabilidad de no ocurrencia de DM2. En otro subgrupo, si FHD=sí, edad<48 y PAS<140, la DM2 no ocurrirá con una probabilidad del 87% (Tabla 3).

La sensibilidad (95 % IC) del árbol de decisión y el modelo de bosque aleatorio son, respectivamente, 64,5 % (62,9, 86,7) y 71,3 % (65,3, 74,4), y su tasa de especificidad (95 % IC) es 66,8 % (58,3, 70,8) y 69,9% (65,4, 77,1) respectivamente, y su precisión (IC 95%) es 64,9% (63,6, 80,4) y 71,1% (66,8, 73,5). Usamos el área bajo la curva ± error estándar (95% IC) para comparar estos dos modelos. El valor relacionado en el caso del árbol de decisión ascendió a 68,6 ± 1,39 (65,8-71,3) y 77,3 ± 0,001 (73,8, 78,8) para el modelo de bosque aleatorio (Figura 3). El árbol de decisión y el modelo de bosque aleatorio (D=6.53, *PAGS*<0,001).



NORTE:no diabéticos;**D:**Diabetes;**FHD:**antecedentes familiares de diabetes;**TG:**triglicéridos;**PAD:**presión arterial diastólica;**PCR-hs:**proteína C reactiva de alta sensibilidad;

Figura 2: El árbol de decisión CART con conjunto de datos de entrenamiento

Tabla 3: Las reglas extraídas a través de modelos aleatorios de bosques y árboles de decisión.

Modelo de bosque aleatorio
R1: SI TG>204,5 y hs-CRP≤1,32 y ocupación=empleo, ENTONCES clase: persona sin diabetes (187/236 o 79,2%) R2: SI
TG>204,5 y hs-CRP<1,32 y ocupación=jubilado y TC≤ 257, clase ENTONCES: persona sin diabetes (43/72 o 59,7%)
R3: SI TG>204,5 y hs-CRP<1,32 y ocupación=jubilado y TC>257 y LDL≤110,9, ENTONCES clase: persona sin diabetes (21/23 o 91,3%) R4: SI
TG>204,5 y hs-CRP< 1,32 y ocupación=jubilado y TC>257 y LDL>110,9, ENTONCES clase: persona con diabetes (5/9 o 55,5%)
R5: SI TG>204.5 y hs-CRP<1.32 y ocupación=desempleo y hs-CRP>4.66, ENTONCES clase: persona con diabetes (9/10 o 90%)
R6: SI TG>204,5 y hs-CRP<1,32 y ocupación=desempleo y hs-CRP≤4,66 y BPD<57,9, ENTONCES clase: persona sin diabetes (138/199 o 69,3%)
R7: SI TG>204,5 y hs-CRP<1,32 y ocupación=desempleo y hs-CRP≤4,66 y BPD>57,9 y FHD=sí, ENTONCES clase: persona con diabetes (14/16 o 87,5%)
R8: SI TG>204,5 y hs-CRP<1,32 y ocupación=desempleo y hs-CRP≤4,66 y BPD>57,9 y FHD=no, ENTONCES clase: persona sin diabetes (25/32 o 78,1%)
R9: SI TG≤204.5 y hs-CRP>1.81 y edad≤46.10, ENTONCES clase: persona sin diabetes (569/753 o 79.1%)
R10: SI TG≤204,5 y hs-CRP>1,81 y edad>46,10 y HDL>67,5 y TG>227 e IMC>24,61, ENTONCES clase: persona con diabetes (8/9 o 88,8%)
R11: SI TG≤204.5 y hs-CRP>1.81 y edad>46.10 y HDL>67.5 y TG>227 e IMC≤24.61, ENTONCES clase: persona sin diabetes (5/9 o 55.5%)
R12: SI TG≤204.5 y hs-CRP>1.81 y edad>46.10 y HDL>67.5 y TG≤227, ENTONCES clase: persona sin diabetes (11/12 o 91.6%)
R13: SI TG≤204,5 y hs-CRP>1,81 y edad>46,10 y HDL≤67,5 y PAL>2,18, ENTONCES clase: persona sin diabetes (129/136 o 94,8%)
R14: SI TG≤204.5 y hs-CRP>1.81 y edad>46.10 y HDL≤67.5 y PAL≤2.18 y BPS≤128.16, ENTONCES clase: persona sin diabetes (4/8 o 50%)
R15: SI TG≤204.5 y hs-CRP>1.81 y edad>46.10 y HDL≤67.5 y PAL≤2.18 y BPS>128.16, ENTONCES clase: persona con diabetes (8/12 o 66.6%)
Modelo de árbol de decisión
R1: SI FHD=no y TG<184, ENTONCES clase: persona sin diabetes (3604/3921 o 92%)
R2: SI FHD=no, TG≥184 y edad<48, ENTONCES clase: persona sin diabetes (340/386 o 88%)
R3: SI FHD=no, TG≥184, edad≥48 y hs-CRP<2.2, ENTONCES clase: persona sin diabetes (272/307 o 88%)
R4: SI FHD=no, TG≥184, edad≥48 y hs-CRP≥2.2, ENTONCES clase: persona con diabetes (100/198 o 51%)
R5: SI FHD=sí, edad<48 y PAS<140, ENTONCES clase: persona sin diabetes (809/894 o 90%)
R6: SI FHD=sí, edad<48 y PAS≥140, ENTONCES clase: persona con diabetes (72/133 o 54%)
R7: SI FHD=sí, edad≥48, PAS≥130, PAD<81 y PAL≥1,6, ENTONCES clase: persona sin diabetes (16/29 o 55%)
R8: SI FHD=sí, edad≥48, PAS≥130, PAD<81 y PAL<1,6, ENTONCES clase: persona con diabetes (37/47 o 79%)
R9: SI FHD=sí, edad≥48, PAS≥130, PAD≥81, HDL<29, ENTONCES clase: persona con diabetes (11/13 o 85%)
R10: SI FHD=sí, edad≥48, PAS≥130, PAD≥81, HDL≥29, LDL<148 y hs-CRP<6,8, ENTONCES clase: persona sin diabetes (96/138 o 70%)
R11: SI FHD=sí, edad≥48, PAS≥130, PAD≥81, HDL≥29, LDL<148 y hs-CRP≥6.8, ENTONCES clase: persona con diabetes (17/33 o 52%)
R12: SI FHD=sí, edad≥48, PAS≥130, PAD≥81, HDL≥29, LDL≥148 y ocupación=empleado, ENTONCES clase: persona sin diabetes (7/9 o 78%)
R13: SI FHD=sí, edad≥48, PAS≥130, PAD≥81, HDL≥29, LDL≥148 y ocupación=otro, ENTONCES clase: persona con diabetes (34/58 o 59%)
R14: SI FHD=sí, edad≥48, SBP<130, IMC<23, ENTONCES clase: persona sin diabetes (324/442 o 73%)
R15: SI FHD=sí, edad≥48, PAS<130, IMC≥23 y educación=baja, ENTONCES clase: persona con diabetes (15/20 o 75%)
R16: SI FHD=sí, edad≥48, PAS<130, IMC≥23 y educación=alta y moderada, ENTONCES clase: persona sin diabetes (15/26 o 58%)

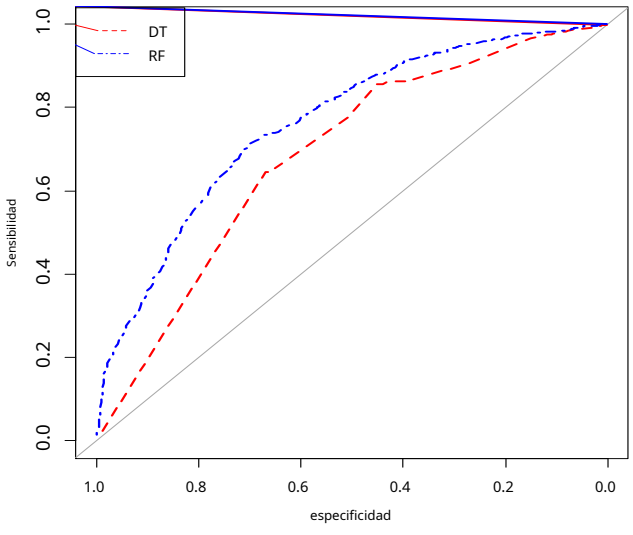


figura 3 :Curva Roc del modelo DT y RF en el conjunto de datos de prueba

Discusión

Desarrollamos un modelo de predicción basado en un estudio transversal para predecir los factores de riesgo de T2DM según los modelos de árbol de decisión y bosque aleatorio.

El modelo de bosque aleatorio mostró que TG, hs-CRP, SBP, LDL, TC, FHD, edad, BMI y PAL estaban fuertemente asociados con T2DM. El modelo de árbol de decisiones encontró que FHD, edad, TG, SBP, hs-CRP, BMI y DBP estaban fuertemente asociados con la aparición de T2DM. Al juntar los dos resultados, TG, FHD, hs-CRP, SBP y BMI son factores de riesgo asociados comunes de T2DM en los dos modelos. En un estudio de cohortes mediante el uso de un árbol de decisión, TG, antecedentes familiares de DM2, IMC, PAS, nivel educativo y ocupación fueron los factores de riesgo asociados a DM2[4].

El algoritmo del árbol de decisión es un modelo de clasificación basado en diferentes variables predictoras y está siendo ampliamente utilizado en medicina.[28-30]. RF crea clasificación múltiple y regresión

(CART), cada uno entrenado en una muestra de arranque de los datos de entrenamiento originales y busca en un subconjunto seleccionado al azar de variables de entrada para determinar la división[31]. Las variables como antecedentes familiares de diabetes, edad, triglicéridos, colesterol LDL, índice de masa corporal y nivel de actividad física ya han sido identificadas como importantes factores de riesgo asociados a la diabetes.[32-34]. El presente estudio ha encontrado hs-CRP como un factor de riesgo asociado importante de T2DM, pero no se ha informado hasta ahora.[28, 33].

Los resultados de nuestro estudio mostraron que los antecedentes familiares de diabetes y triglicéridos fueron los factores de riesgo más importantes relacionados con la DM2 en los modelos de árbol de decisión y bosque aleatorio. En otros estudios también, los antecedentes familiares de diabetes y TG fueron los factores de riesgo asociados más importantes para la DM2[4,30].

Los árboles de decisión son una de las herramientas más sencillas para los sistemas de decisión y fáciles de entender. Los árboles de decisión se pueden convertir fácilmente en reglas si-entonces. Los programas basados en estas reglas se pueden crear y usar en computadoras personales para análisis de decisiones, y se pueden usar fácilmente con médicos y personal de atención médica para concluir los resultados.[4, 35-38].

En este estudio, la comparación de los modelos de árbol de decisión y bosque aleatorio mostró que los valores de sensibilidad y especificidad del bosque aleatorio eran más altos que los del árbol de decisión, lo que contradecía los estudios anteriores.[31, 39]. Por otro lado, la sensibilidad del algoritmo C4.5 fue mayor que el bosque aleatorio, pero la especificidad del bosque aleatorio fue mayor que el árbol de decisión (C4.5)[39]. La razón de ser la diferencia entre la sensibilidad de ellos es usar un algoritmo diferente.

La curva ROC es una técnica para visualizar, organizar y elegir la clasificación en función del rendimiento de la clasificación. El área bajo la curva (AUC) es un índice de qué modelo funciona mejor y tiene un alto nivel de precisión. Este índice, que compara el rendimiento de los verdaderos positivos y los falsos positivos de dos extremos de decisión diferentes, a menudo se usa para evaluar la precisión predictiva de los modelos de clasificación.[40].

En el estudio actual, el AUC del bosque aleatorio del conjunto de datos de prueba fue significativamente más alto que el del árbol de decisión, lo cual fue consistente con estudios previos.[31, 39]. El modelo de bosque aleatorio es un modelo preciso para la investigación de nuevos marcadores predictores, que está en línea con el modelo anterior.[14, 31].

La fortaleza del estudio radica en su gran tamaño de muestra que lo hace aplicable a la población general. Una posible limitación de este estudio es que se basa en datos transversales y no puede obtener resultados obtenidos a partir de datos longitudinales o de cohortes.

Conclusiones

Los modelos de bosques aleatorios pueden proporcionar buenos modelos de predicción debido a su eficacia, sensibilidad y especificidad. Según el modelo de bosque aleatorio, TG y hs-CRP son los factores de riesgo asociados más importantes para T2DM. Este estudio también ha identificado algunos factores de riesgo nuevos asociados con la DM2 que indican la necesidad de una mayor evaluación de la aplicabilidad clínica de este modelo.

Agradecimientos

Este estudio fue apoyado financieramente por el Centro de Investigación de Bioquímica de la Nutrición de la Universidad de Ciencias Médicas de Mashhad, Mashhad, Irán.

Declaracion de conflicto de interes

Los autores declaran que no existe conflicto de intereses.

Fondos

Este estudio fue apoyado por la Universidad Mashhad de Ciencias Médicas.

Reflejos

-Según el modelo RF, TG, hs-CRP, SBP y FHD son los factores de riesgo asociados más importantes para la DM2.

-Basado en el modelo DT FHD, TG, edad y hs-CRP son los factores de riesgo asociados más importantes para T2DM.

-El modelo RF demostró un mejor poder discriminatorio en comparación con el modelo DT.

Referencias

1.Hu D, Fu P, Xie J, Chen CS, Yu D, Whelton PK, et al. Prevalencia creciente y baja concienciación, tratamiento y control de la diabetes mellitus entre adultos chinos: el estudio InterASIA. Diabetes Res Clin Práctica. 2008; 81(2): 250-7.

2.Nathan DM, Buse JB, Davidson MB, Ferrannini E, Holman RR, Sherwin R, et al. Manejo médico de la hiperglucemia en la diabetes tipo 2: un algoritmo de consenso para el inicio y ajuste de la terapia una declaración de consenso de la Asociación Americana de Diabetes y la Asociación Europea para el Estudio de la Diabetes. Cuidado de la diabetes. 2009; 32(1): 193-203.

3.Farzianpour F, Fouroshani AR, Hosseini S, Hosseini S, Hosseini SS. Una comparación entre dos métodos educativos de pacientes diabéticos en Irán. 2ª Conferencia Internacional sobre Economía, Educación y Gestión; 1 de junio - 2 de junio; Shanghái 2012.

4.Ramezankhani A, Pournik O, Shahrabi J, Khalili D, Azizi F, Hadaegh F. Aplicación del árbol de decisión para la identificación de una población de bajo riesgo de diabetes tipo 2. Estudio de lípidos y glucosa de Teherán. Diabetes Res Clin Práctica. 2014; 105(3): 391-8.

5.Esmaeily H, Tayefi M, Ghayour-Mobarhan M, Amirabadizadeh A. Comparación de tres algoritmos de minería de datos para identificar los factores de riesgo asociados a la diabetes tipo 2. Iran Biomed J. 2018 (en prensa).

6.Whiting DR, Guariguata L, Weil C, Shaw J. Atlas de diabetes de la FID: estimaciones globales de la prevalencia de diabetes para 2011 y 2030. Diabetes Res Clin Pract. 2011; 94(3): 311-21.

7.Hemmati M, Zohoori E, Mehrpour O, Karamian M, Asghari S, Zarban A, et al. Potencial antiaterogénico de azufaifo, azafrán y agracejo: acciones antidiabéticas y antioxidantes. Excli J. 2015; 14(4): 908-15.

8.Booth GL, Kapral MK, Fung K, Tu JV. Relación entre la edad y la enfermedad cardiovascular en hombres y mujeres con diabetes en comparación con personas no diabéticas: un estudio de cohorte retrospectivo basado en la población. La lanceta. 2006; 368 (9529): 29-36.

9.Alberti KGM, Zimmet P, Shaw J. Federación Internacional de Diabetes: un consenso sobre la prevención de la diabetes tipo 2. Diabetes Med. 2007; 24(5): 451-63.

10Investigadores ESPERANZAS. Efectos del ramipril sobre los resultados cardiovasculares y microvasculares en personas con diabetes mellitus: resultados del estudio HOPE y el subestudio MICRO-HOPE. La lanceta. 2000; 355 (9200): 253-9.

11Barr EL, Zimmet PZ, Welborn TA, Jolley D, Magliano DJ, Dunstan DW, et al. Riesgo de mortalidad cardiovascular y por todas las causas en personas con diabetes mellitus, alteración de la glucosa en ayunas y alteración de la tolerancia a la glucosa The Australian Diabetes, Obesity, and Lifestyle Study (AusDiab). Circulación. 2007; 116(2): 151-7.

12Pi-Sunyer FX. ¿Qué tan efectivos son los cambios en el estilo de vida en la prevención de la diabetes mellitus tipo 2? Nutr Rev. 2007; 65(3): 101-10.

13Norris SL, Kansagara D, Bougatsos C, Fu R. Detección de diabetes tipo 2 en adultos: una revisión de la evidencia para el grupo de trabajo de servicios preventivos de EE. UU. Ann Intern Med. 2008; 148(11): 855-68.

14Tapak L, Mahjub H, Hamidi O, Poorolajal J. Comparación de datos reales de métodos de extracción de datos en la predicción de la diabetes en Irán. Salud Informar Res. 2013; 19(3): 177-85.

15. Chen HY, Chuang CH, Yang YJ, Wu TP. Exploración de los factores de riesgo del parto prematuro mediante la minería de datos. Aplicación de sistema experto 2011; 38(5): 5384-7.

dieciséis. Aslan K, Bozdemir H, Sahin C, Noyan Ogulata S. ¿Puede la red neuronal estimar el pronóstico de los pacientes con epilepsia según los factores de riesgo? J Med Syst. 2010; 34(4): 541-50.

17Bellazzi R, Zupan B. Minería de datos predictivos en medicina clínica: cuestiones y directrices actuales. Int J Med Inform. 2008; 77(2): 81-97.

18Delen D, Walker G, Kadam A. Predicción de la supervivencia al cáncer de mama: una comparación de tres métodos de extracción de datos. Artif Intel Med. 2005; 34(2): 113-27.

19Ghayour-Mobarhan M, Moohebati M, Esmaily H, Ebrahimi M, Parizadeh SMR, Heidari-Bakavoli AR, et al. Estudio Mashhad sobre accidente cerebrovascular y trastorno aterosclerótico cardíaco (MASHAD): diseño, características basales y estimación del riesgo cardiovascular a 10 años. Int J Salud Pública. 2015; 60(5): 561-72.

20Asociación Americana de Diabetes. Informe del comité de expertos en diagnóstico y clasificación de la diabetes mellitus. Cuidado de la diabetes. 1997; 20: 1183-97.

21Fawcett T. Una introducción al análisis ROC. Patrón Recognit Lett. 2006; 27(8): 861-74.

22Shi G. Minería de datos y descubrimiento de conocimientos para geocientíficos. Oxford: Elsevier; 2013.

23Tayefi M, Tajfard M, Saffar S, Hanachi P, Amirabadizadeh AR, Esmaeily H, et al. hs-CRP está fuertemente asociado con la enfermedad coronaria (CHD): un enfoque de extracción de datos que utiliza un algoritmo de árbol de decisión. Informática Métodos Programas Biomed. 2017; 141: 105-9.

24Han J, Kamber M, Pei J. Minería de datos: conceptos y técnicas: conceptos y técnicas. 3.ª edición Elsevier; 2011.

25Tayefi M, Esmaeili H, Saberi Karimian M, Amirabadi Zadeh A, Ebrahimi M, Safarian M, et al. La aplicación de un árbol de decisión para establecer los parámetros asociados a la hipertensión. Informática Métodos Programas Biomed. 2017; 139: 83-91.

26Juan Lu ZQ. Los elementos del aprendizaje estadístico: minería de datos, inferencia y predicción. JR Stat Soc Serie A Stat en Sco. 2010; 173(3): 693-4.

27Lavrač N. Técnicas seleccionadas para la minería de datos en medicina. Artif Intel Med. 1999; 16(1): 3-23.

28Fayyad, UM, Wierse A, Grinstein, GG. Visualización de información en minería de datos y descubrimiento de conocimiento. Morgan Kaufman. 2002.

29Kammerer JS, McNabb SJ, Becerra JE, Rosenblum L, Shang N, Iademarco MF, et al. Transmisión de la tuberculosis en entornos no tradicionales: un enfoque de árbol de decisiones. Am J Previo Med. 2005; 28(2): 201-7.

30Amirabadizadeh A, Nezami H, Vaughn MG, Nakhaee S, Mehrpour O. Identificación de factores de riesgo para el uso de drogas en una muestra de tratamiento iraní: un enfoque de predicción mediante árboles de decisión. Sust Uso Mal uso. 2018; 53(6): 1030-40.

31Maroco J, Silva D, Rodrigues A, Guerreiro M, Santana I, de Mendonça A. Métodos de minería de datos en la predicción de la demencia: una comparación de datos reales de la precisión, sensibilidad y especificidad del análisis discriminante lineal, regresión logística, redes neuronales , admite máquinas de vectores, árboles de clasificación y bosques aleatorios. Notas BMC Res. 2011; 4(1): 299-313.

32. Abbasi A, Peelen LM, Corpeleijn E, van der Schouw YT, Stolk RP, Spijkerman AM, et al. Modelos de predicción del riesgo de desarrollar diabetes tipo 2: búsqueda bibliográfica sistemática y estudio de validación externo independiente. BMJ. 2012; 345: e5900.

33. Guariguata L, Whiting D, Hambleton I, Beagley J, Linnenkamp U, Shaw J. Estimaciones globales de prevalencia de diabetes para 2013 y proyecciones para 2035. Diabetes Res Clin Pract. 2014; 103(2): 137-49.

34. Dallo FJ, Weller SC. Eficacia de las recomendaciones de cribado de diabetes mellitus. Proc Natl Acad Sci US A. 2003; 100(18): 10574-9.

35. Podgorelec V, Kokol P, Stiglic B, Rozman I. Árboles de decisión: una descripción general y su uso en medicina. J Med Syst. 2002; 26(5): 445-63.

36. Musen MA, Middleton B, Greenes RA. Sistemas de apoyo a la decisión clínica. Springer Science+ Medios comerciales; 2014.

37. Wang CJ, Li YQ, Wang L, Li LL, Guo YR, Zhang LY, et al. Desarrollo y evaluación de un enfoque de predicción simple y eficaz para identificar a las personas con alto riesgo de dislipidemia en residentes rurales adultos. Más uno. 2012;7(8): e43834.

38. Tayefi M, Saberi-Karimian M, Esmaeili H, Zadeh AA, Ebrahimi M, Mohebati M, et al. Evaluación de los factores de riesgo asociados al síndrome metabólico mediante el árbol de decisión. Comp Clin Path. 2018; 27(1); 215-23.

39. Ge G, Wong GW. Clasificación de datos de espectrometría de masas de cáncer de páncreas premaligno mediante conjuntos de árboles de decisión. BMC Bioinformática. 2008; 9: 275-87.

40Ke Ws, Hwang Y, Lin E. Farmacogenómica de la eficacia de los fármacos en el tratamiento con interferón de la hepatitis C crónica mediante algoritmos de clasificación. Aplicación avanzada Bioinform Chem. 2010; 3: 39-44.