



## Diseño de un modelo predictivo de pesquisa cardiovascular utilizando Árboles de Decisión: propensión de pacientes a presentar diabetes tipo 2, hipertensión arterial o dislipidemia: Estudio piloto, comuna de Quellón, Chiloé

Claudio Cárdenas<sup>1</sup>, Sergio González<sup>1</sup>, Rosa Nahuel<sup>2</sup>, Pablo Herrera<sup>2</sup>, Luis Ferrada<sup>3</sup>, Diego Celis<sup>4</sup>.

<sup>1</sup> Dirección Servicio de Salud Chiloé, Subdepartamento de Planificación y Control de Gestión.

<sup>2</sup> DESAM de Comuna de Quellón, Programa Cardiovascular CESFAM de Quellón

<sup>3</sup> Hospital de Castro, Médico Especialista en Salud Pública, Jefe de Ginecología

<sup>4</sup> Hospital de Castro, Médico Internista, Jefe Medicina Interna

Recibido el 19 de febrero 2018 / Aceptado el 19 agosto 2018

Rev Chil Cardiol 2018; 37: 126-133

### Resumen:

**Introducción:** En la actualidad, la Minería de Datos es cada vez más popular en el campo de la salud porque existe una necesidad de eficiencia metodológica y analítica para detectar información desconocida y valiosa en datos de salud.

**Objetivo:** Desarrollar un modelo predictivo utilizando técnicas de minería de datos, específicamente Árboles de Decisión, para pesquisar pacientes con propensión a desarrollar Diabetes Tipo II (DM II), Hipertensión Arterial (HTA) o Dislipidemia (DLP).

**Método:** Se analizó el problema de los Factores de Riesgo Cardiovascular Mayores desde una perspectiva de procesos y se estudiaron las técnicas que permiten descubrir el conocimiento del fenómeno almacenado en las bases de datos de Examen de Medicina Preventiva del Adulto (EMPA) de la Población en Control Cardiovascular que presenta DM II, HTA o DLP

**Resultados:** El Algoritmo C5, presenta un mayor poder predictivo, respecto de otros algoritmos de Árbol de Decisión. Se comprobó que las variables Edad y Circunferencia de Cintura fueron las de mayor poder de discriminación en el padecimiento de DM2, HTA o DLP. El algoritmo C5 alcanzó una precisión global de un 83,01% en la partición de prueba, luego en la misma partición el modelo logra discriminar un paciente con algunas de las patologías en el 85,25% de los casos, y uno que no presenta alguna de las patologías en un 80,27% de las oportunidades.

**Conclusión:** La Minería de Datos y en este caso, específicamente los Modelos de Árboles de Decisión son una alternativa válida para la pesquisa cardiovascular temprana.

**Palabras Clave:** pesquisa cardiovascular; minería de datos; árboles de decisión; diabetes mellitus II; hipertensión arterial; dislipidemia.



## Design of a predictive model of cardiovascular screening using decision trees: propensity of patients to present type 2 diabetes, arterial hypertension or dyslipidemia. Pilot study commune of Quellón, Chiloé.

**Introduction:** Data Mining is increasingly popular in the health field because there is a need for an efficient analytical methodology to detect unknown and valuable information of health data.

**Objective:** To develop a predictive model using data mining techniques, specifically Decision Trees, to investigate patients with a propensity to develop Type II Diabetes, Arterial Hypertension or Dyslipidemia.

The data of adult patients presenting Type II diabetes, Hypertension or Dyslipidemia being followed in a preventive cardiovascular control program were analyzed with the aim of unveiling phenomena that could help develop the prediction of these risk factors.

**Results:** With respect to other decision tree algo-

rithms, Algorithm C 5, showed a greater predictive power. The variables age and waist circumference had the greatest power of discrimination for DM2, HTA or DLP. The C 5 algorithm reached a global precision of 83.01% in the test partition. Then, in the same partition the model managed to discriminate a patient with some of the risk factors in 85.25% of cases, and to rule out any of them in 80.27% of cases.

**Conclusion:** Data Mining, specifically decision tree models, is a valid alternative for early detection of cardiovascular of risk factors.

**Key Words:** cardiovascular research, datamining; decision trees; diabetes mellitus II, hypertension; dyslipidemia.



## Introducción:

Se calcula que en 2014 la prevalencia mundial de la DM II mellitus fue 9% entre los adultos mayores de 18 años. En 2012 fallecieron 1,5 millones de personas como consecuencia directa de la DM II. Más del 80% de las muertes por esta enfermedad se registra en países de ingresos bajos y medios. Según proyecciones de la OMS, la DM II será la séptima causa de mortalidad en 2030<sup>1</sup>. Según el reporte de la Organización Mundial de la Salud (OMS) del 2012, 1 de cada 3 personas en el mundo padece de HTA. La HTA es la segunda causa de muerte a nivel mundial, y se reconoce internacionalmente como "muerte silenciosa" pues en la mayoría de los casos los pacientes tienden a ser asintomáticos<sup>2</sup>. Chile, al igual que la mayoría de los países del mundo, ha experimentado cambios negativos en los estilos de vida de su población, lo que ha contribuido a una mayor prevalencia de Factores de Riesgo Cardiovascular (FRCV) modificables, siendo las Enfermedades Cardiovasculares (ECV) la causa más frecuente de mortalidad (27%)<sup>3</sup>.

La minería de datos es una de las áreas de investigación más vitales y motivadoras en el objetivo de encontrar información significativa a partir de grandes conjuntos de datos. En la actualidad, la minería de datos es cada vez más popular en el campo de la salud porque existe una necesidad de eficiencia metodológica y analítica para detectar información desconocida y valiosa en datos de salud<sup>4</sup>. En el cuidado de la salud, la minería de datos se está volviendo, no sólo cada vez más popular, sino que también cada vez más esencial. Estas herramientas pueden beneficiar enormemente a todas las partes involucradas en el cuidado de la salud. Por ejemplo, puede ayudar a las aseguradoras de salud a detectar fraudes y abusos, a que las organizaciones de la salud tomen mejores decisiones con sus usuarios, que los médicos identifiquen tratamientos efectivos y mejores prácticas, y que los pacientes reciban mejores y más accesibles servicios de salud. Las cantidades de datos generados por las transacciones de atención médica son enormes y demasiado para ser procesadas y analizadas por métodos tradicionales. La Minería de Datos proporciona la metodología y la tecnología para transformar estos cerros de datos en información útil para la toma de decisiones<sup>5</sup>.

El empleo de modelos estadísticos predictivos en las ciencias de la salud ha crecido significativamente en los últimos años. Estos emergen como un vínculo importante entre la estadística y la práctica médica; son de gran ayuda en la toma de decisiones y permiten la creación

de diversos sistemas y herramientas útiles para reducir las incertidumbres, garantizar mejores actuaciones y establecer eficaces medidas de control para la erradicación de las enfermedades<sup>6</sup>. Un ejemplo de esto, son los árboles de decisión y el conjunto de reglas derivado de éstos, que permiten predecir el riesgo de desarrollar cardiopatía hipertensiva en individuos hipertensos, por lo que su introducción en la práctica clínica mejora la evaluación del paciente con HTA<sup>7</sup>. Los árboles de clasificación o de decisión, se caracterizan por su sencillez. Su campo de acción abarca diversas áreas como el diagnóstico médico, juegos, predicciones meteorológicas y control de calidad. Son los modelos de aprendizaje inductivo supervisado, no paramétrico, más utilizado como forma de representación del conocimiento<sup>8</sup>.

Este método forma grupos bien definidos y mutuamente excluyentes, cuyos miembros presentan características similares que ayudan a determinar la participación y el comportamiento relacionado con la salud. Ello sirve de base para intervenciones en salud pública al utilizar variables demográficas y clínicas para identificar grupos de mayor riesgo con respecto a un resultado específico<sup>9</sup>. El objetivo de este estudio fue desarrollar un modelo predictivo utilizando técnicas de minería de datos, específicamente, árboles de decisión, para pesquisar pacientes con propensión a desarrollar DM II Tipo II, HTA o DLP. Los resultados de este estudio permitirán pesquisar pacientes que tengan propensión de desarrollar algunas de estas tres patologías, con el objetivo de buscar la detección precoz de la enfermedad, y su tratamiento oportuno, por medio de una atención personalizada en APS.

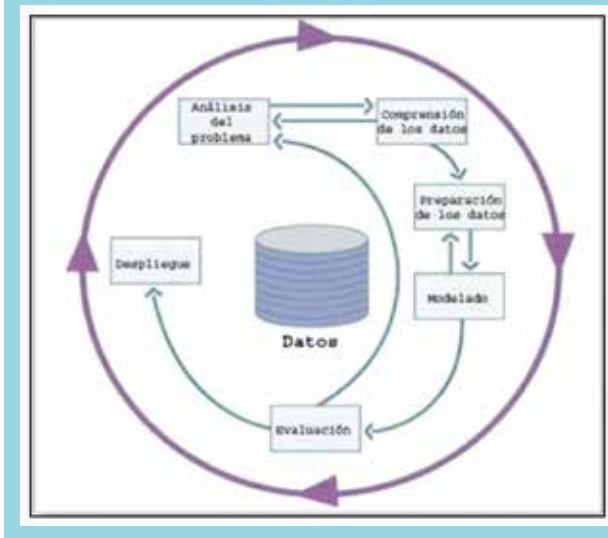
## Métodos:

Se analizó el problema de los Factores de Riesgo Cardiovascular Mayores desde una perspectiva de procesos y se estudiaron las técnicas que permiten descubrir el conocimiento del fenómeno almacenado en las bases de datos de la Población en Control cardiovascular que presenta DM II, HTA o DLP en exámenes de medicina preventiva del adulto (EMPA). Se identificaron patrones contenidos en los datos para determinar las variables predictivas y seleccionar los algoritmos que se utilizaron en el desarrollo del modelo predictivo. Se desarrolló un prototipo funcional del modelo de árbol de decisión para, finalmente, evaluar la calidad de predicción del prototipo y corregir los posibles errores del modelo. Metodología de Minería de Datos. Para el desarrollo del proyecto se utilizó la metodología CRISP-DM, que



es de libre distribución y puede trabajar con cualquier herramienta para desarrollar cualquier proyecto. Esta metodología estructura el ciclo de vida de un proyecto de Minería de Datos en seis fases, que interactúan entre ellas de forma iterativa durante el desarrollo del proyecto<sup>2</sup> (Figura 1). La primera fase pretende conocer los objetivos del negocio, valorar la situación actual y realizar un plan del proyecto; la segunda fase pretende conocer los datos iniciales de los que disponemos para la realización del proyecto; en la tercera fase se preparan los datos recogidos en la fase anterior, para que en la cuarta fase se aplique una técnica de modelado que se adapte a los objetivos del negocio y genere información nueva y relevante. En la quinta fase se analizan los resultados obtenidos valorando su calidad, para que, por último, se indiquen los planes de distribución del proyecto<sup>10</sup>.

Figura 1: Fases de metodología CRISP-DM.



**Algoritmos de Árboles de Decisión.** Un árbol de decisión es un modelo de predicción cuyo objetivo principal es el aprendizaje inductivo a partir de observaciones y construcciones lógicas. Son muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la solución de un problema<sup>11</sup>. Constituyen probablemente el modelo de clasificación más utilizado y popular. El conocimiento obtenido durante el proceso de aprendizaje inductivo se representa mediante un árbol<sup>12</sup>. Los árboles de decisión poseen ventajas sobre otros métodos de clasificación

como, por ejemplo, las redes neuronales en cuanto a requerimientos en la capacidad de cómputo y al tiempo de generación de resultados. Específicamente, en las redes neuronales el tiempo de aprendizaje es elevado; además, no pueden interpretar lo que han aprendido y se requiere una amplia cantidad de datos para el entrenamiento. En tanto, en los árboles de decisión se identifica el orden de verificación de las condiciones y las acciones que se deben llevar a cabo y se plantea el problema para que todas las opciones sean analizadas, con lo que se facilita la interpretación de la decisión tomada<sup>13</sup>.

**Software estadístico y de Minería de Datos.** Para las etapas de comprensión y preparación de datos se utiliza Microsoft Excel, STATA 15.0 e IBM SPSS Modeler 16.0, y para las etapas de modelación, evaluación e implementación IBM SPSS Modeler 16.0.

**Variable Objetivo.** Esta variable se definió con la denominación “CONTROL”, compuesta por dos grupos: GRUPO SI = Grupo de pacientes en control del Programa Cardiovascular, que presenta al menos una de las tres patologías, DM HTA o DLP. GRUPO NO = Grupo de Pacientes EMPA (2016 a junio 2017) y que no están en control en Programa Cardiovascular, al corte de junio del 2017 y, eventualmente, no presenta ninguna de las tres patologías señaladas. Luego, este grupo servirá para poder discriminar y encontrar aquellos patrones en los datos que caracterizan a las personas con algunas de las tres patologías del grupo “SI” y las diferencian de aquellos en el grupo “NO”.

**Data utilizada.** Base de Datos de la población en control, en la Comuna de Quellón, con corte a junio 2017, con un total de 2.865 registros y 41 campos y Base de datos EMP del año 2016 y a junio 2017, extraídos del sistema RAYEN (apartada por el Subdepartamento de Tecnología de la Información del Servicio de Salud Chiloé), con un total de 2.436 registros y 68 campos. Registros totales entre ambas bases de datos fueron 5.301. Posterior a las primeras etapas de la metodología CRISP-DM, se llegó a la fase de modelado con una matriz depurada que contenía nueve campos predictores. Estos fueron: edad, circunferencia de cintura (CC), presión arterial sistólica, colesterol, talla, presión arterial diastólica, peso, sexo, y tabaquismo. Se dispuso de 3.586 registros, 2.006 correspondientes al grupo que presenta al menos una de las tres patologías estudiadas (Grupo SI) y 1.580 a la categoría que eventualmente no presenta patología (Gr-



Tabla 1: Algoritmos de Árboles de Decisión, utilizados en la etapa de Modelado.

Algoritmo de Árbol de Decisión	Porcentaje de Precisión General
C 5 (Árbol de máxima ganancia de información)	83,017%
CHAID (Árbol de detección automática de interacciones mediante chi-cuadrado)	78,834%
CRT (Árbol de Clasificación y Regresión)	78,779%
QUEST (Árbol estadístico rápido, insesgado y eficiente)	77,496%

po NO), siendo estos divididos en dos subconjunto: una muestra del 70% de ellos, que fue utilizada para entrenar el modelo seleccionado, y otra muestra del 30% del total que fue utilizada únicamente para probar (validar) la estabilidad del mismo en datos inéditos para el modelo generado.

**Seguridad de la Información.** Para cumplir con la política de seguridad de la información del Servicio de Salud Chiloé, se firmó, entre el DESAM de la Comuna de Quellón y la Dirección de Servicio de Salud Chiloé, un convenio de acuerdo de cautela de registros personales que define el procedimiento de traspaso de información, a través de canales institucionales, los accesos y respon-

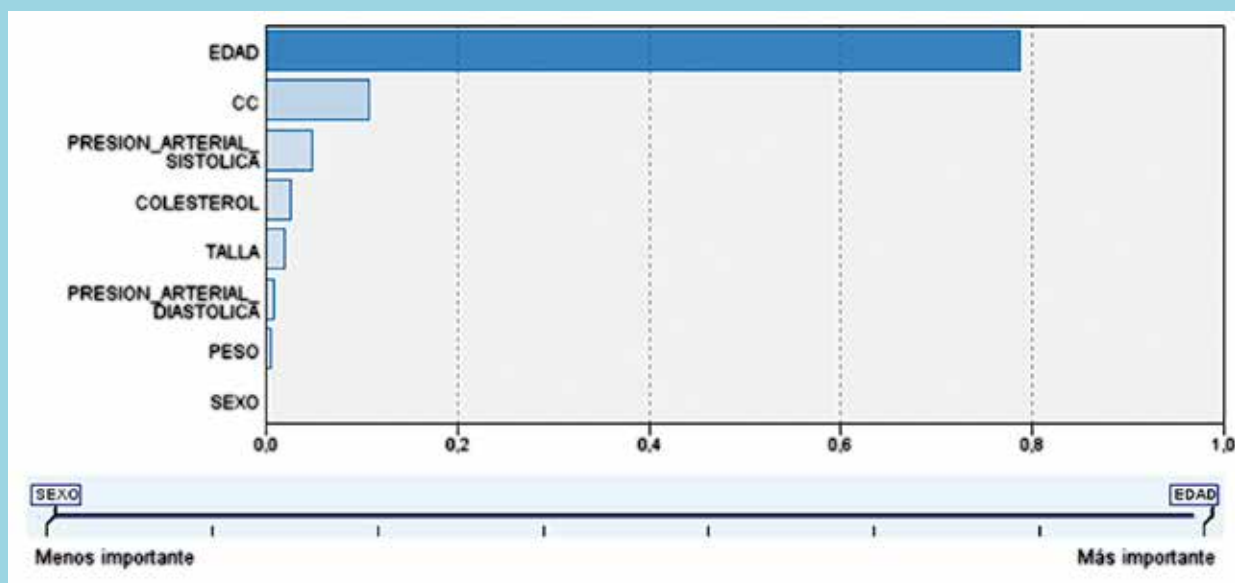
sabilidades de los funcionarios implicados directamente en este proyecto.

El trabajo fue analizado y aprobado por el Comité de Ética Asistencial de la Dirección del Servicio de Salud Chiloé.

### Resultados:

Como resultado de la integración de las bases de datos referidas sólo resultaron concordantes los siguientes campos predictores: edad, circunferencia de Cintura (CC) presión arterial sistólica, colesterol, talla, presión arterial diastólica, peso, sexo, y tabaquismo. Las demás variables no resultaron integrables entre ambas bases de datos o presentaban demasiados datos nulos, que hacían

Figura 2: Predictores (variables) utilizados por el algoritmo seleccionado para evaluación (C 5), según orden de importancia.





impráctica su imputación. Luego, estas variables fueron usadas como predictores en cuatro diferentes algoritmos de árboles de decisión (Tabla 1).

Según lo observado en la tabla anterior, se selecciona el Algoritmo de C5 o Árbol de máxima ganancia de información, para utilizarlo en la etapa de evaluación, entregando esta etapa, según el modelo seleccionado, los predictores según su grado de importancia en la discriminación de un paciente con patología de uno que no la presenta (Figura 2).

Según la Figura 2, el algoritmo seleccionado, C5, destaca como variables determinantes de la presencia de menos una de las tres patologías, DM II, HTA o DLP a la variable edad, luego la circunferencia de cintura (CC) y, en menor medida, las presiones arteriales, colesterol, talla, peso y sexo. Quedaron fuera del modelo la variable tabaquismo, que posiblemente debido al método de medición cualitativa nominal, no aporta información suficiente. Al pasar el modelo árbol de decisión C5 a la etapa de evaluación y aplicarlo a la muestra de entrenamiento (formación) y posteriormente a la Muestra de Prueba (validación), se obtienen excelentes resultados en ambos grupos de datos, observándose una aproximación razonable, sin presencia de sobre aprendizaje en la muestra de entrenamiento, y estable en la muestra de validación. Esto se observa claramente en el gráfico de evaluación de ganancias (Figura 3), en donde la curva \$C-CONTROL, correspondiente al comportamiento del modelo que predice la categoría SI se aproxima adecuadamente a la curva \$BEST-CONTROL, que muestra el compor-

tamiento del modelo ideal prediciendo la categoría SI de la variable objetivo (CONTROL).

El algoritmo C5 alcanzó una precisión global de un 83,01% en la partición de prueba (validación). Posteriormente, en la misma partición el modelo logra discriminar un paciente con algunas de la patología en el 85,25% de los casos, y uno que no presenta alguna de las patologías en un 80,27% de las oportunidades (Tabla 2). Por tanto, es claro que el modelo predictivo es más certero para reconocer casos positivos (pacientes enfermos) que negativos (pacientes sanos).

### Discusión:

Los resultados extraídos de este estudio indican que el Algoritmo C5, o Árbol de máxima ganancia de información, presenta un mayor poder predictivo, comparado con algoritmos como CHAID CRT y QUEST. Los modelos C5.0 dividen la muestra en función del campo que ofrece la máxima ganancia de información. Las distintas submuestras definidas por la primera división se vuelven a dividir, por lo general basándose en otro campo, y el proceso se repite hasta que resulta imposible volver a dividir las submuestras. Por último, se vuelven a examinar las divisiones del nivel inferior, y se eliminan, o podan, las que no contribuyen significativamente con el valor del modelo<sup>15</sup>. Para la obtención del modelo predictivo mediante el algoritmo C5, se comprobó que las variables edad (79% de importancia predictiva) y circunferencia de cintura<sup>9</sup> (11% de importancia predictiva) fueron las de mayor discriminación para DM2, HTA o

Figura 3: Gráficos de ganancias, para muestra de entrenamiento (formación) y de validación (prueba).

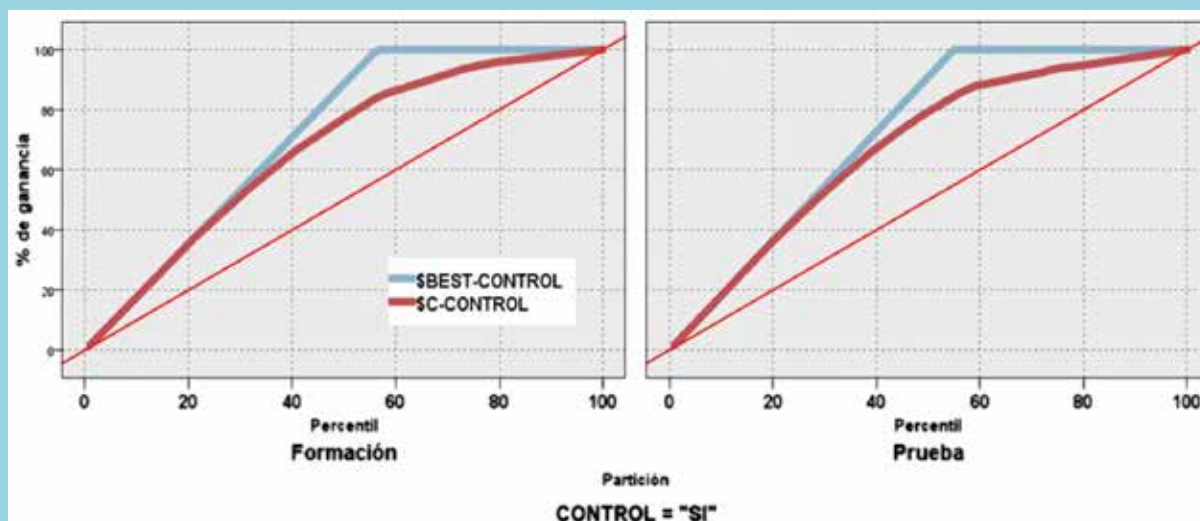




Tabla 2: Tabla de Clasificación del Algoritmo C 5 o Árbol de máxima ganancia de información, sobre la partición de prueba (Validación).

REAL	PREDICCIÓN	
	NO	SI
NO	80,27%	19,72%
SI	14,74%	85,25%

DLP, presentado en conjunto un 80% de valor predictivo. No obstante, también influyó la presión arterial sistólica (5%), colesterol (3%), talla (2%) y presión arterial diastólica, peso y sexo bajo el 1% de valor predictivo. Luego, al interpretar las reglas de negocios del árbol de decisión, se puede observar que el algoritmo realiza una primera estratificación por edad identificando dos rangos claramente definidos: pacientes de 53,86 o menos años y pacientes mayores a esta edad. El primer grupo presenta una probabilidad de 32% de presentar al menos una de las tres patologías. Por su parte, el segundo grupo presenta una probabilidad de 86,94% de presentar al menos una de estas patologías. Por tanto, claramente la variable edad resulta determinante en estratificar los pacientes respecto del riesgo de presentar o no presentar al menos unas de las patologías. Al analizar el árbol dentro del estrato definido como bajo riesgo, se observan perfiles importantes que describen las reglas que definen como un paciente en este estrato puede presentar igualmente, al menos, una de las patologías. Es así como un paciente menor de 41 años con una circunferencia de cintura (CC) mayor a 106,5 centímetros y una talla menor o igual 150 centímetros tendría, según el modelo predictivo, una probabilidad sobre 83% de presentar, al menos, unas de las tres patologías. Ahora, si la talla es mayor a 150 centímetros esta probabilidad se reduce a 32%, aproximadamente. Del patrón anterior es destacable cómo la circunferencia de cintura, así como la talla del paciente, indica un bajo riesgo de presentar al menos una de las patologías. Otro perfil interesante en este mismo segmento se da en pacientes de entre 46 y 53 años, con una circunferencia de cintura mayor a 105 centímetros, una presión arterial diastólica entre 69 mmHg y 70 mmHg, los cuales tienen una probabilidad de un 83,5% de presentar al menos una de las patologías. Finalmente, y en el segmento que agrupa la mayor cantidad de

pacientes del estrato (20 % del total de la muestra de entrenamiento, 742 pacientes), se define como personas menores de 41 años y una circunferencia de cintura menor o igual a 106,5 centímetros presentan una probabilidad de solo 12,5% de padecer al menos una de las tres patologías. Por otra parte, al analizar el estrato de Alto Riesgo, los pacientes mayores a 65 años presentan una probabilidad de un 100% de presentar al menos una de las patologías. Otro perfil importante en este estrato son los pacientes que tienen una edad de entre 53 y 65 años y que además tienen una circunferencia de cintura mayor a 105 centímetros quienes tienen una probabilidad del 88,4% de presentar patología, mientras que los pacientes en el mismo tramo de edad, pero con una circunferencia de cintura menor a 105 centímetros, la probabilidad se reduce a 67,6%. Ahora, continuando en este mismo patrón se observa como aquellos pacientes que tienen una edad de entre 53 y 65 años, con una circunferencia de cintura menor a 105 centímetros, y un colesterol total menor a 199 mg/dL presentan una probabilidad de 79,8% de presentar las patologías; por el contrario, los que presentan un colesterol total mayor a 199 mg/dL disminuyen su probabilidad de presentar patología a un 56,7%. Este último fenómeno es muy interesante, ya que el colesterol total parece actuar como un factor protector, lo cual sin duda se podría aclarar si se dispusiera de mediciones de tipos de colesterol específicos que en los exámenes EMPA no están disponibles.

### Conclusión:

La Minería de Datos y, en este caso, específicamente los modelos de árboles de decisión son una alternativa válida para la pesquisa temprana de DM2, HTA o DLP. Su aplicación permitirá extender los beneficios de las acciones de prevención de estas enfermedades y su diagnóstico temprano, sin aumentar considerablemente los costos y con la utilización de los datos capturados mediante el examen de Medicina Preventiva del Adulto (EMPA) o cualquier otra instancia de contacto con el paciente que permita obtener las variables utilizadas por el modelo predictivo.

**Limitaciones del trabajo.** Eventualmente, puede haber un sesgo en catalogar pacientes en quienes se realizó el EMPA y que no están en control, como sin riesgo de desarrollar DM, HTA o DLP. Ello a pesar de la detección y exclusión de datos anómalos, ya que, eventualmente, podría haber falsos negativos (pacientes que creímos no estaban en riesgo de desarrollar DM, HTA o DLP, y sí lo están).



## Referencias

1. WORLD HEALTH ORGANIZATION. Global Health Estimates: Deaths by Cause, Age, Sex and Country, 2014. Geneva: WHO.
2. DAVILA F, SANCHEZ C. Técnicas de minería de datos aplicados al diagnóstico de entidades clínicas. RCIM. 2012; 4: 174-183.
3. MARTÍNEZ M, LEIVA A, SOTOMAYOR C, VICTORIANO T, VON CHRISMAR A, PINEDA S. Factores de riesgo cardiovascular en estudiantes de la Universidad Austral de Chile. Rev. Méd. Chile. 2012; 140: 426-435.
4. DIVYA T, SONALI A. A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology. 2013; 5: 241-266.
5. HIAN CHYE K, GERALD T. Data Mining Applications in Healthcare. Journal of Healthcare Information Management. 2011; 19: 64- 72
6. LÓPEZ FR, YANES SR, SUÁREZ SPR, AVELLO MR, GUTTIÉRREZ EM, ALVARADO F. Modelo estadístico predictivo para el padecimiento de pie diabético en pacientes con diabetes mellitus tipo II. Medisur. 2016; 14.
7. ÁLVAREZ A, GONZÁLEZ J, MACEO L, FRÓMETA A, BÁRZAGA S, CERVANTES A. Árbol para predecir el desarrollo de la cardiopatía hipertensiva Rev. Cubana Med. 2014; 53: 266-281
8. SEGURA A. Aplicación de árboles de decisión en la salud pública. Revista CES Salud Pública. 2012; 3: 94-103.
9. BARRIENTOS R, CRUZ N, ACOSTA H, RABATTE I, GOGASCOECHEA MC, PAVÓN P, et al. Árboles de decisión como herramienta en el diagnóstico médico. Rev Med UV, julio - diciembre, 2009: 19-24.
10. SOLARTE G, OCAMPO C. Técnicas de clasificación y análisis de presentación del conocimiento para problemas de diagnóstico. Scientia et Technica Año XV, No 42 agosto de 2009. Universidad Tecnológica de Pereira. ISSN 0122-1701.
11. 9. BARRIENTOS R, CRUZ N, ACOSTA H, RABATTE I, GOGASCOECHEA MC, PAVÓN P, ET AL., et al. Árboles de decisión como herramienta en el diagnóstico médico. Rev Med UV, 2009: 19-24.
12. BASTIDAS M, CABELLO M, RODRIGUEZ A, GARCÍA J. Sistema de apoyo para la toma de decisiones en termografía de glándulas mamarias. Revista Mexicana de Ingeniería Biomédica. 2017; 38: 166-187.
13. HERNANDEZ, H. Aplicación de Minería de Datos a Información de Pacientes Prediabéticos. Revista Ibero Americana de Producción Académica y Gestión Educativa. Vol. 1, No. 1. 2014.
14. NISHIZAKI, R. Análisis de datos sanitarios aplicando metodología CRISP-DM. UAM. Departamento de Ingeniería Informática. Universidad Autónoma de Madrid. Trabajo Fin de Grado. 2017.
15. IBM. Nodos de modelado de IBM SPSS. Modeler 16. Copyright IBM Corp. 1994, 2013