Using the Random Forest Algorithm for Searching Behavior Patterns in Electronic Health Records

C. de la Fuente, A. Urrutia, and E. Chávez

***Abstract*—The search for information associated with qualitative data is usually done using data mining algorithms, the presented research analyzes data of patients with essential hypertension (HTA), patients who have developed hypertension but there is no clear reason why it has occurred. In this research, a search of behavioral patterns was performed in the data associated with the clinical records of 8470 patients using the Random Forest algorithm. As a case study, the proposal focuses on finding the relationship between the different pathologies or factors associated to Hypertensive patients (other diseases for example). The findings validate the right use of the algorithm due to the results obtained agrees with the knowledge defined and validated in the literature. Thus, trivial knowledge can be obtained with the algorithm used. However, non-trivial knowledge was also obtained given the analysis performed on a total of 4408 data of female patients and 4062 of male patients showed a great difference between the factors or pathologies that a patient presents when classified according to their sex, thus another deep study must be carried out closely with experts in the area of the health as future research.**

***Index Terms*—Data analysis, Data mining, Health information management, Machine learning algorithms.**

I. Introducción

LAS Enfermedades Cardiovasculares (ECV) son la principal causa de muerte en todo el mundo. Cada año mueren más personas por ECV que por cualquier otra causa. Esta enfermedad hace alusión a cualquier proceso de índole vascular, incluyendo las cardiopatías congénitas, valvulopatías, endocarditis y vasculitis. La Enfermedad Coronaria (EC) debería incluirse dentro de ECV ya que afecta las arterias coronarias provocando infarto de miocardio, angina de pecho, insuficiencia renal cardiaca y muerte súbita, la EC puede ser una complicación directa de otros factores como la edad, obesidad y la Hipertensión Arterial (HTA) [1].

La hipertensión arterial esencial es compleja y no completamente conocida, ya que no hay una única causa que dé lugar a la HTA, sino que son múltiples las vías y mecanismos por los que ésta puede establecerse [2].

Parte de esta investigación fue financiada por el Proyecto DIN 11/2016 de la Universidad Católica de la Santísima Concepción.

C. de la Fuente, Universidad de Talca, Talca, Chile (e-mail:claudia.delafuente@utalca.cl).

A. Urrutia, Universidad Católica del Maule, Talca, Chile e-mail:aurrutia@ucm.cl).

E. Chávez, Universidad Católica de la Santísima Concepción, Concepción, Chile (e-mail: echavez@ucsc.cl).

En otras palabras, los pacientes con HTA esencial son aquellos pacientes a los cuales se les desconoce la causa que provocó la hipertensión; es por esto que se vuelve interesante el trabajar en la búsqueda de conocimiento en este tipo de pacientes, buscando patrones de comportamiento en su historial clínico.

Según [3], dentro de los algoritmos más utilizados para la toma de decisiones, se encuentran los de minería de datos. En [4] se menciona que la extracción de registros electrónicos de datos en salud (EHR) tiene el potencial de establecer nuevos principios de estratificación del paciente y revelar correlaciones de enfermedades desconocidas.

La presente investigación, presenta un caso que se orienta a la extracción de conocimiento respecto de la relevancia que existe entre los elementos del historial clínico de un paciente hipertenso esencial y el desarrollo de la enfermedad de Diabetes. Dado el minado de conocimiento sobre datos de pacientes, se elige el uso de una técnica predictiva para la clasificación de conocimiento. Según la revisión bibliográfica se encuentra que los arboles de decisión parecieran ser una técnica ampliamente utilizada, es por esto que se define el uso del algoritmo *Random Forest* para la extracción de conocimiento sobre los datos de la ficha clínica electrónica de pacientes hipertensos esenciales.

A continuación, se presentan los apartados respecto de la revisión de literatura en el cual se analiza el uso de algoritmos de minería de datos en la extracción de conocimiento; Se explica el funcionamiento del algoritmo *Random Forest*, la metodología, el desarrollo de la propuesta de trabajo y los resultados experimentales obtenidos.

II. Trabajos Relacionados

La investigación se centra en la búsqueda de conocimiento en pacientes con hipertensión esencial y su relación con factores de riesgo y otras enfermedades.

Para la selección de la técnica a utilizar, se realizó una revisión de la literatura, donde muy pocos de los algoritmos utilizados para la búsqueda de patrones han sido ejecutados sobre datos de pacientes hipertensos directamente, de los cuales es de destacar el uso de Árboles de Decisión [12][15][17], *Naïve Bayes* [11][18], SVM [10][11][12] y Redes Neuronales [11] para la búsqueda de conocimiento en fichas clínicas y el diagnóstico de enfermedades. De esta fase se observa un gran uso respecto de algoritmos basados en árboles de decisión. Dentro de los árboles de decisión se ha extendido el uso de

Random Forest, incluyendo comparaciones con otros algoritmos dónde supera a otros en el resultado de la predicción, diferenciándose además por incluir la importancia de cada una de las variables analizadas en cada caso [12][13][19].

Random Forest es una combinación de predictores de árboles, de modo que cada árbol depende de los valores de un vector aleatorio muestreado independientemente y con la misma distribución para todos los árboles del bosque. El error de generalización de un bosque de clasificadores de árboles depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos [20]. Las estimaciones internas supervisan el error, la fuerza y la correlación, utilizando estas para mostrar la respuesta al aumento del número de características utilizadas en la división. Las estimaciones internas también se utilizan para medir la importancia variable [14]. Así, *Random Forest* propone un modelo óptimo para mantener el balance entre errores de sesgo y varianza (*ensamble*).

III. Metodologia

El caso a analizar se centra en la extracción de conocimiento trivial (definido en el conocimiento médico o de la literatura existente) y no trivial (que requiera de un estudio exhaustivo futuro por especialistas del área) a través de la búsqueda de patrones en los datos de la ficha clínica de pacientes cardiovasculares con hipertensión arterial declarada. Para ello se propone una metodología y la aplicación del algoritmo *Random Forest* utilizando la herramienta R. El análisis se realiza sobre los datos clínicos de pacientes hipertensos de manera que se pueda responder a la pregunta de investigación obtenidas en entrevistas con expertos en el área de la salud, las que son: ¿Es posible extraer conocimiento (trivial o no trivial) utilizando el algoritmo *Random Forest* en la ficha clínica electrónica de pacientes hipertensos esenciales?

La propuesta de trabajo se encuentra desarrollada según la Metodología KDD [3][16] representada en la Fig. 1. Para este caso la metodología incluye las siguientes etapas:

- Preparación y selección de datos. Se centra en conocer los datos a trabajar e identificar aquellos factores o patologías que se pueden trabajar en el análisis de pacientes con HTA esencial.
- Transformación. Se aplican métodos de limpieza, se observan y agrupan registros de pacientes de ser necesario, dejando los archivos listos para su utilización.
- Minería de datos. Se realiza la aplicación del algoritmo seleccionado a los grupos de datos creados en la fase de transformación.
- Evaluación e interpretación. Se analiza la información extraída de la aplicación del algoritmo y se generan tipos de correlaciones entre los factores analizados.
- Validación: Se realiza una validación de la información obtenida apoyada por la literatura y de ser necesario con el conocimiento de expertos en el área de la salud.

A continuación, se presenta el desarrollo de la aplicación de la metodología descrita anteriormente para el caso analizado, correspondiente a la ficha clínica de pacientes hipertensos esenciales.

Fig. 1. Metodología propuesta.

IV. Preparación y Selección De Datos

La base de datos (BD) con la cual se trabajó contenía el historial clínico de paciente (datos personales anonimizados), existiendo una gran variedad de campos que se pueden utilizar en la extracción de conocimiento. La estructura de la BD se componía por algunas columnas de datos de tipo texto de llenado libre, otras de tipo estructurado, con formato texto y numérico. Existiendo un total de 8472 registros de pacientes con hipertensión arterial esencial. Para el análisis de los datos, e identificación de cuál sería el experimento a realizar se generó una nube de datos en R (versión 3.4.3 en S.O Windows 10) con la información de la BD, (Fig. 2), lo que permitió observar que además del campo "sexo" una de las enfermedades que se repite bastante en los pacientes analizados, es la diabetes, la cual se presenta en los campos como diabetes o diabetes mellitus.

Fig. 2. Nube de datos caso clínico.

Es por lo anterior, que el caso a experimentar se orientó a la extracción de conocimiento respecto de la relevancia que existe entre los elementos del historial clínico de un paciente hipertenso y el desarrollo de la enfermedad de Diabetes.

V. Transformación de Datos

La base de datos a utilizar contenía información innecesaria para el análisis, por lo cual se realiza una etapa de selección de datos acorde al planteamiento del problema y de representación para los patrones en estudio.

La mayoría de los datos de la base de datos es de tipo binaria por lo que fue necesario transformar esto a números, por ejemplo: 0 si no tiene la enfermedad y 1 si la enfermedad es observada.

Posterior a la primera transformación, se realizó un nuevo análisis de datos, decidiendo realizar agrupaciones de datos con el objetivo de trabajar patrones de comportamiento a evaluar en el algoritmo de manera más acotada, de esta forma se disminuyó la dispersión en los resultados obtenidos, simplificando además el análisis de variables a observar.

Por lo tanto, como resultado de este agrupamiento, para el caso clínico se distinguieron tres agrupaciones de datos posibles de realizar, tomando como referencia la información entregada por un experto en el área de la salud y la revisión de la literatura:

- Agrupación por enfermedad o condición cardiovascular: incluye todos los pacientes que han tenido un episodio o enfermedad cardiovascular o coronaria.
- Agrupación por otro tipo de enfermedades presentes.
- Agrupación por factores de riesgo cardiovascular, definidas de la siguiente manera:
  a. Grupo 1 (Factores de Riesgo): tabaco, sedentarismo, alcoholismo.
  b. Grupo 2 (enfermedad cardiovascular): hiperlipidemia, miocardiopatía, Enfermedad congénita, arritmia, angioplastia, infarto, angina, enfermedad arterial no coronaria, cardiopatía, EPOC.
  c. Grupo 3 (Otras enfermedades): enfermedad reumática, insuficiencia renal, epilepsia.

Otra agrupación a considerar y que al mirar la nube de datos (Fig. 2) resulta interesante realizar, es la separación de cada uno de los grupos mencionados por el campo sexo.

Al término de la etapa, luego de la limpieza y transformación de los datos se identifican 8470 registros a utilizar en el proceso de minería de datos, clasificándose en 4408 datos de pacientes de sexo Femenino y 4062 de pacientes del sexo Masculino.

VI. Mineria de Datos

Para la aplicación del algoritmo *Random Forest* se generaron 2/3 de los registros para el set de datos de entrenamiento y 1/3 para el conjunto evaluador. El algoritmo separa el set de datos para entrenamiento en pequeños conjuntos de individuos con sus registros de diferentes grupos de patologías elegidos al azar, pero siempre manteniendo la enfermedad de diabetes dentro de los conjuntos. Por cada conjunto de individuos se crea un árbol de decisión. Una vez creado sus árboles se comienza a trabajar con el conjunto de datos evaluador, donde para cada nuevo individuo cada árbol evalúa y entrega su predicción.

Un método bastante particular que se puede integrar al desarrollo de este algoritmo es el cálculo de la tabla de importancia. Esta tabla consiste en identificar la variable que más se parece o concuerda con la clasificación de la variable objetivo; es decir entre más arriba en posición se encuentre la variable en la tabla de importancia, mayor es la correlación hacia la clasificación de la variable objetivo. Para poder obtener esta clasificación de la tabla, el método revisa que variable o factor predomina en las cabeceras de los ensambles (el nodo más alto del ensamble) y respecto a esta revisión genera la tabla e importancia.

A continuación, se presenta la aplicación del algoritmo a través del conjunto de experimentos realizados aplicando el algoritmo *Random Forest*. Se realizaron 6 experimentos incluyendo la separación por sexo de los pacientes (femenino, masculino), y por las tres agrupaciones mencionadas con anterioridad (por enfermedad o condición cardiovascular, por otro tipo de enfermedad y por factor de riesgo cardiovascular). Cada experimento permite observar el número de ensambles que permitieron generar la convergencia del algoritmo en su predicción a través de los gráficos y los resultados arrojados por el algoritmo, respecto de las correlaciones encontradas. Cabe destacar, que cada experimento se repitió 5 veces, presentando sólo cambios mínimos en el valor del porcentaje de importancia no en las correlaciones.

A. Experimento 1

El objetivo de este experimento fue establecer la correlación entre el grado de importancia de factores relacionados con conjunto de enfermedades coronarias, en pacientes con hipertensión de género Femenino y su tendencia a desarrollar la enfermedad de diabetes. Las Fig. 3 y Fig. 4 presentan los resultados de dicho experimento.

El gráfico correspondiente a la Fig. 3, muestra el error obtenido al intentar predecir un valor de la clase “sí” (línea verde) o “no” (línea roja) y el OOB error (línea negra), indicando que para llegar a una convergencia entre los errores asociados a las predicciones de cada individuo tuvo que superar los 70 ensambles. Es decir, que al finalizar la ejecución el error al intentar predecir un valor de la clase “sí” o “no” y el OOB error llega a un mismo valor, en este caso 0%.

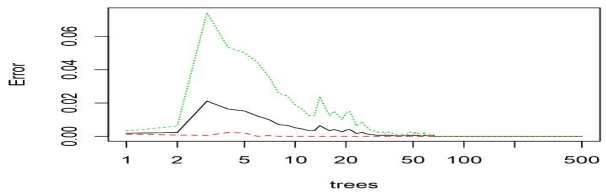


Fig. 3. Convergencia en ensambles del conjunto de enfermedades coronarias en pacientes femeninos.

La Fig. 4 presenta la existencia de correlaciones entre las enfermedades coronarias en pacientes femeninos, observándose una mayor correlación a la diabetes, que un paciente presente una cardiopatía, y aun menor nivel de correlación el resto de las enfermedades como por ejemplo de hiperlipidemia o arritmia.

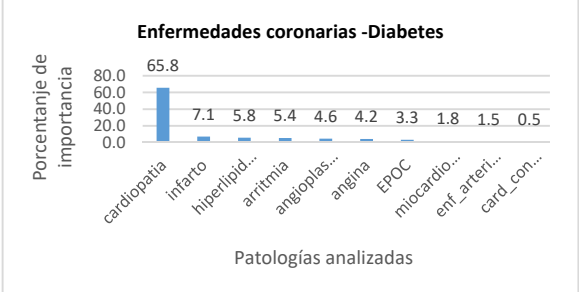


Fig. 4. Correlación del conjunto de enfermedades coronarias en pacientes femeninos.

B. *Experimento 2*

Correlaciona el grado de importancia de factores relacionados con enfermedades no coronarias, en pacientes con hipertensión de género femenino, y su tendencia a desarrollar la enfermedad de diabetes (Véase Fig. 5 y Fig. 6).

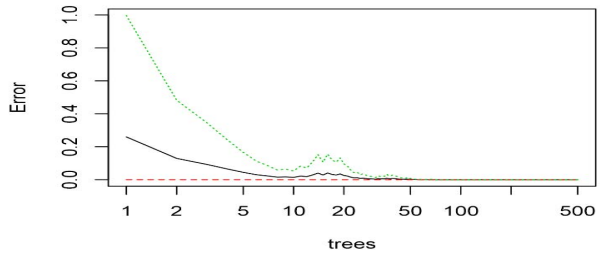


Fig.5. Convergencia en ensambles del conjunto de enfermedades no coronarias en pacientes femeninos.

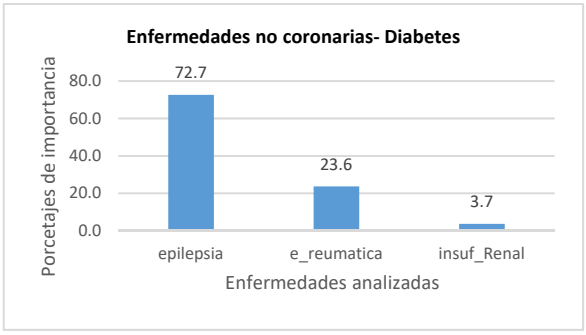


Fig. 6. Correlación del conjunto de enfermedades no coronarias en pacientes femeninos.

Al analizar la gráfica de la Fig. 6, se observa que una de las enfermedades no coronarias, en pacientes femeninos, importantes de analizar es la epilepsia, ya que tiene una posible correlación con el factor diabetes. La enfermedad reumática en este tipo de pacientes también tiene una correlación con el factor diabetes.

La insuficiencia renal no tiene una correlación directa con el factor diabetes, en pacientes hipertensos femeninos.

C. *Experimento 3*

Grado de importancia de factores relacionados con enfermedades denominadas factores de riesgo, en pacientes con hipertensión y de género Femenino y su tendencia a desarrollar la enfermedad de diabetes (Véase Fig. 7 y Fig. 8).

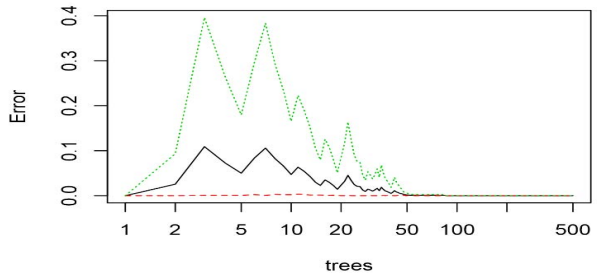


Fig. 7. Convergencia en ensambles del conjunto de factores de riesgo en pacientes femeninos.

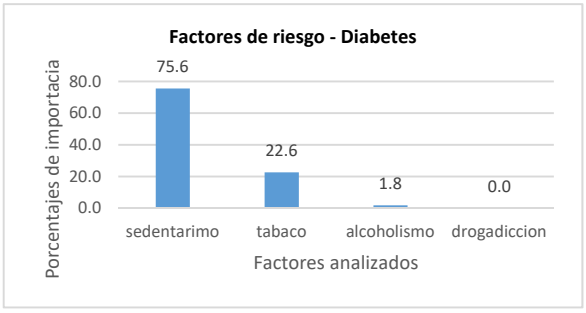


Fig. 8. Correlación del conjunto de factores de riesgo, pacientes femeninos.

Uno de los factores de riesgo importantes de analizar en pacientes femeninos ya que tiene una mayor correlación con la diabetes es el sedentarismo, sin dejar de observar el factor tabaco. El alcoholismo y la drogadicción en este tipo de pacientes no tiene una relevancia o correlación con el factor diabetes.

D. *Experimento 4*

Grado de importancia de factores relacionados con el conjunto de enfermedades coronarias, en pacientes con hipertensión y de género masculino y su tendencia a desarrollar la enfermedad de diabetes (Véase Fig. 9 y Fig.10).

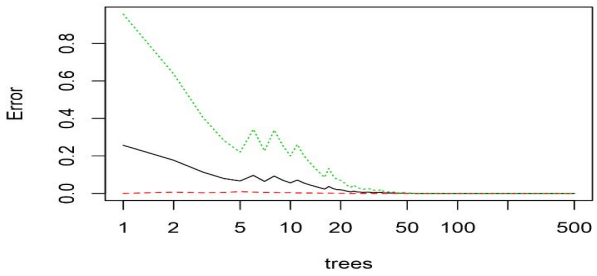


Fig. 9. Convergencia en ensambles del conjunto de enfermedades coronarias en pacientes masculinos.

La Fig. 10, presenta la existencia de correlaciones entre varias enfermedades coronarias en pacientes masculinos, observándose una mayor correlación con la diabetes, la enfermedad angina, cardiopatía e infartos, y aun menor nivel de correlación, pero de igual forma relevante hiperlipidemia, angioplastia, arritmia, enfermedad arterial y EPOC, y a un cero nivel de correlación miocardiopatía y cardiopatía congénita.

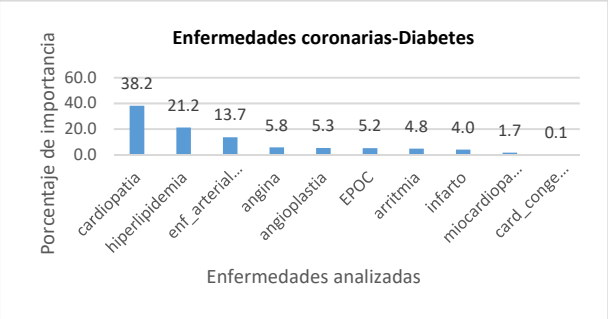


Fig.10. Correlación del conjunto de enfermedades coronarias en pacientes masculinos.

E. Experimento 5

Grado de importancia de factores relacionados con enfermedades no coronarias, en pacientes con hipertensión y de género masculino y su tendencia a desarrollar la enfermedad de diabetes (Véase Fig.11 y Fig.12).

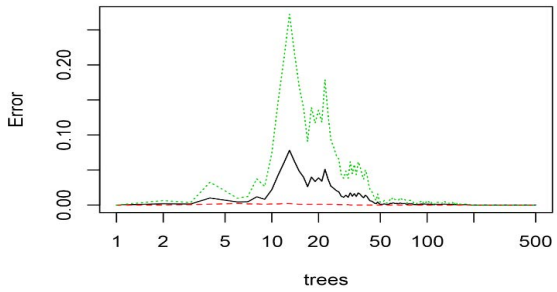


Fig. 11. Convergencia en ensambles del conjunto de enfermedades no coronarias en pacientes masculinos.

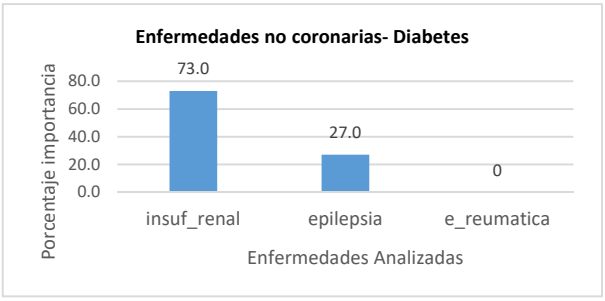


Fig. 12. Correlación del conjunto de enfermedades no coronarias pacientes masculinos.

Al analizar la gráfica se observa que una de las enfermedades no coronarias en pacientes masculinos, importantes de analizar es la insuficiencia renal, ya que tiene una posible correlación con el factor diabetes.

La enfermedad reumática en este tipo de pacientes no tiene una correlación con el factor diabetes.

F. *Experimento 6*

Grado de importancia de factores relacionados con enfermedades denominadas factores de riesgo; en pacientes con hipertensión y de género masculino y su tendencia a desarrollar la enfermedad de diabetes (Véase Fig. 13 y Fig.14).

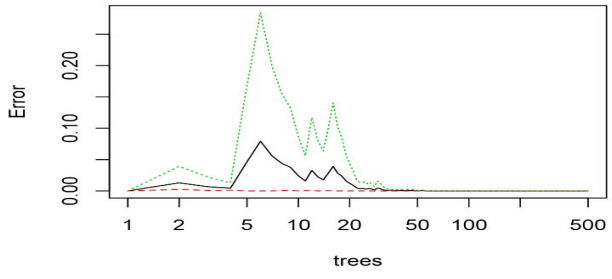


Fig. 13. Convergencia en ensambles del conjunto de factores de riesgo en pacientes masculinos.

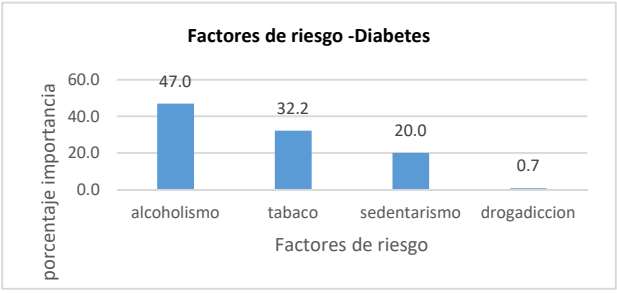


Fig. 14. Correlación del conjunto de factores de riesgo en pacientes masculinos.

Se observa que todos los factores analizados tienen algún tipo de correlación con el factor diabetes, pero uno de los factores de riesgo que sobresale en pacientes masculinos es el alcoholismo. La drogadicción en este tipo de pacientes no tiene una correlación muy notoria, sin embargo, un número pequeño de ensambles lo han elegido como un factor decisivo (Fig. 14).

VII. Evaluación y Validación de Resultados

Al analizar los experimentos realizados se observa que se han encontrado diferentes patrones entre los pacientes con hipertensión esencial, el algoritmo *Random Forest* logra identificar a través de sus ensambles variables de importancia al momento de tomar una decisión respecto de los elementos a utilizar para lograr predecir a un nuevo candidato.

Uno de los principales resultados indica la variación entre patrones respecto del género de un paciente (femenino o masculino). Por ejemplo, en el experimento 2 y experimento 5 o entre el experimento 3 y experimento 6, se observan diferencias de patrones respecto del género del paciente por categoría de agrupación.

La Tabla 1 presenta el listado de patrones encontrados por el algoritmo *Random Forest* en pacientes con HTA respecto a la presencia de la enfermedad de diabetes, es decir aquellos factores o patologías que fueron definidos por el algoritmo con una mayor correlación con la diabetes, clasificados por experimento y género.

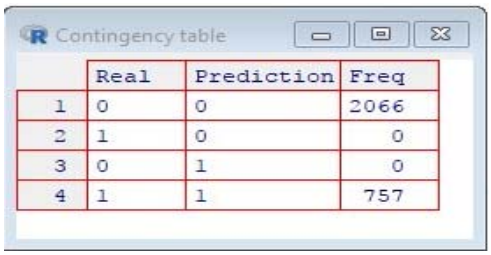
Los resultados obtenidos, son validados con el mismo algoritmo, incluyendo una tabla de contingencia, la cual va almacenando y contando los falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos, al ir evaluando a cada individuo.

TABLA I
PATRONES ENCONTRADOS POR EL ALGORITMO

N° Patrón	Experimento	Sexo	Factor
1	1	F	Cardiopatía
2	2	F	Epilepsia
3	3	F	Sedentarismo
4	4	M	Cardiopatía
5	4	M	Hiperlipidemia
6	5	M	Insuficiencia Renal
7	6	M	Alcoholismo
8	6	M	Tabaco

En la Fig. 15, se presenta un ejemplo correspondiente a una de las tablas de contingencia entregada por el algoritmo RF, al ser ejecutada la búsqueda de patrones.

La tabla de contingencia en el desarrollo de los experimentos entregó valores para todas las ejecuciones realizadas sin falsos positivos o negativos.


Contingency table

	Real	Prediction	Freq
1	0	0	2066
2	1	0	0
3	0	1	0
4	1	1	757

Fig.15. Tabla de contingencia entregada en un experimento.

Para la validación de la información encontrada a través de los resultados entregados por el algoritmo *Random Forest*, se realizó una búsqueda de información trivial en la literatura, lo que no fue validado por la literatura deberá ser validado por expertos a largo plazo pues pudiese ser conocimiento no trivial.

A continuación, se presenta el listado de patrones encontrados por el algoritmo RF en pacientes con HTA esencial respecto de presentar la enfermedad de diabetes que pudieron ser validados por literatura especializada en el tema, es decir aquellos factores o patologías que fueron definidos por el algoritmo con una mayor correlación con la diabetes y que concuerda con lo establecido por la revisión de la literatura en el tema. Estos patrones validan conocimiento trivial en el tema (Ver Tabla II).

TABLA II
PATRONES ENCONTRADOS POR EL ALGORITMO Y SU VALIDACIÓN

	Experimento	Sexo	Factor	Validación
1	1	F	Cardiopatía	Literatura [6]
2	2	F	Epilepsia	Literatura [5]
3	3	F	Sedentarismo	Literatura [7]
4	4	M	Cardiopatía	Literatura [6]
5	4	M	Hiperlipidemia	Literatura [15]
6	5	M	Insuficiencia Renal	Literatura [9]
7	6	M	Alcoholismo	Literatura [8]
8	6	M	Tabaco	Literatura [15]

Esta tabla demuestra la validación de los patrones encontrados a través de una búsqueda en la literatura, pero estas no presentan una separación entre pacientes de tipo femenino y masculino. Cada bibliografía presenta al paciente hipertenso como un todo, sin separación de género ni diferencias entre tipos de hipertensión.

Es por lo anterior que esta separación por genero requiere de una investigación profunda realizada por especialistas del área (cardiólogos). Luego de entrevistas de validación con 3 expertos se determina que los patrones identificados por genero corresponden a conocimiento no trivial.

VIII. Conclusiones

Respecto de la pregunta de investigación inicial: ¿Es posible extraer conocimiento trivial y no trivial utilizando el algoritmo *Random Forest* en la ficha clínica electrónica de pacientes hipertensos esenciales?, se observa que, al aplicar y analizar los resultados obtenidos respecto de las correlaciones de factores con la diabetes, se ha logrado identificar lo siguiente:

- La investigación realizada corresponde a resultados de información de tipo trivial, el algoritmo *Random Forest* fue capaz de validar la teoría en relación a la hipertensión, se pudo validar el conocimiento extraído en cada experimento comprobando todas las aseveraciones realizadas. Sin embargo, la información presente en la literatura no genera una separación entre los grupos femeninos o masculinos.
- El algoritmo *Random Forest* ha respondido presentando información trivial, pero también dentro de la información obtenida se han observado una diferencia entre los factores o patologías que presenta un paciente según su género, lo que se podría tratar como información no trivial y que debiese ser validado por expertos en el área (trabajo futuro).

Las asociaciones encontradas por el algoritmo *Random Forest* pueden tener mucha relación con el tipo de estilo de vida que tenga un paciente hipertenso en el cuidado de su salud; punto no menos importante ya que para un profesional del área de la salud que debe tratar a este tipo de pacientes hipertensos, el género del paciente enfermo pasa a ser relevante al momento de entregar el tratamiento a seguir, ya que según el algoritmo de *Random Forest* es mucho más probable que un paciente femenino o masculino se mantenga dentro de los patrones encontrados; por ejemplo: es mucho más probable que un hombre que tiene HTA y diabetes desarrolle una insuficiencia renal y una mujer la enfermedad de epilepsia, entre otros casos.

Como conclusión final se observa que el algoritmo *Random Forest* es un buen indicador como base para el desarrollo de un sistema experto en el tema, es decir un sistema de soporte a las decisiones como apoyo al diagnóstico clínico de enfermedades cardiovasculares como la hipertensión arterial esencial.

Referencias

[1] J. S. Ruiz, Epidemiología de la Enfermedad Cardiovascular, Madrid: Díaz de Santos, 2012.

[2] J. M. Alcazar, A. Oliveras, L. M. Orte, S. Jimenez y J. Segura, "Nefrología al Día", Sociedad Española de Nefrología, 16 septiembre 2016. [En línea]. Available: http://revistanefrologia.com/es-monografias-nefrologia-dia-articulo-hipertension-arterial-esencial-23.

[3] J. Hernández, M. J. Ramírez y C. Ferri, Introducción a la Minería de Datos, Madrid: Pearson educación, 2004.

[4] P. B. Jensen, L. Jensen y S. Brunak, "Mining Electronic Health Records: Towards Better Research Applications and Clinical Care", PubMed, 2012.

[5] I. Larrañagaa, I. Conget, C. Viñals, A. Pané, I. Vinagre y M. Giménez, «Diabetes Tipo 1 y Epilepsia. ¿Más que Mera Coincidencia?,» Elsevier, vol. Endocrinología y Nutrición, 2016.

[6] C. Yáñez, «Más de la Mitad de los Casos de Diabetes en Chile se Deben al Sedentarismo,» La Tercera, 19 Junio 2017.

[7] University of California, San Francisco, «Diabetes Educación Online,» UCSF, [En línea]. Available: https://dtc.ucsf.edu/es/la-vida-con-diabetes/dieta-y-nutricion/la-diabetes-y-el-alcohol-2/. [Último acceso: noviembre 2017].

[8] MedlinPlus, «Diabetes y Enfermedad Renal,» Biblioteca Nacional de Medicina de los EE. UU. [En línea]. [Último acceso: noviembre 2017].

[9] H. Polat, H. Mehr y A. Cetin. Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods. *Journal of Medical Systems*, vol.41, no.4, pp. 55, 2017.
[10] W. Khannara, N. Iam-On, y T. Boongoen. Predicting Duration of CKD Progression in Patients with Hypertension and Diabetes. In *Intelligent and Evolutionary Systems*. Springer. pp. 129-141, 2016.
[11] M. Khalilia, S. Chakraborty, y M. Popescu. Predicting Disease Risks from Highly Imbalanced Data Using Random Forest. *BMC Medical Informatics and Decision Making*, vol.11, no. 1, pp. 51, 2011.
[12] M. J. Kane, N. Price, M. Scotch y P. Rabinowitz. "Comparison of ARIMA and Random Forest Time Series Models for Prediction of Avian Influenza H5N1 Outbreaks". BMC Bioinformatics, vol.15, no. 1, 2014.
[13] L. Breiman. (2001). "Random Forests. Machine Learning", vol. 45, no. 1, pp.5-32.
[14] W.N. Kelley. Medicina interna, Volumen 1. Ed. Médica Panamericana, 1993.
[15] E. Valdés Ramos y N. Rodríguez. "Frecuencia de la Hipertensión Arterial y su Relación con Algunas Variables Clínicas en Pacientes con Diabetes Mellitus Tipo 2". Revista Cubana de Endocrinología, vol.20, no.3, pp.77-88, 2009.
[16] C. Santos, B. Pedroso, A. Guimarães, D. Carvalho y L. Pilatti. "Forecasting of Human Development Index of Latin American Countries Through Data Mining Techniques". IEEE Latin America Transactions, vol. 15, no.9, 2017.
[17] J. Kao, H. Chen, F. Lai, L. Hsu y H. Liaw. (2015) "Decision Tree Approach to Predict Lung Cancer the Data Mining Technology". In: Park J., Pan Y., Chao HC., Yi G. (Eds) Ubiquitous Computing Application and Wireless Sensor. Lecture Notes in Electrical Engineering, vol.331. Springer.
[18] N. Kureshi, S. Abidi y C. Blouin, "A Predictive Model for Personalized Therapeutic Interventions in Non-small Cell Lung Cancer," in IEEE Journal of Biomedical and Health Informatics, vol. 20, no.1, pp. 424-431, 2016.
[19] I. Soares, J. Dias, H. Rocha, et al. "Predicting Xerostomia After IMRT Treatments: a Data Mining Approach", Health Technology, vol. 8, no. 1, pp. 159–168, 2018.
[20] R. Medina and C. Ñique, "Bosques Aleatorios como Extensión de los Árboles de Clasificación con los Programas R y Python," Interfases, pp. 165–189, 2017.

Claudia de la Fuente, Licenciada en Ciencias de la Ingeniería y Magister en Ciencias de la Computación de la Universidad Catolica del Maule (Chile). Actualmente, es académico de la Universidad de Talca (Chile), sus principales áreas de investigación son: bases de datos, análisis de datos e innovación docente.

Angélica Urrutia, Licenciada en Matematica y Computacion de la Universidad de Santiago de Chile (Chile), Magister en Ciencias de la Computación de la Universidad de Concepción (Chile), Doctora de la Universidad de Castilla La Mancha (España). Actualmente, es profesor Titular del Departamento de Computación e Informática de la Universidad Católica del Maule (Chile), sus principales áreas de investigación son: bases de datos, calidad de datos, inteligencia de negocio, análisis y extracción de datos.

Emma Chávez, Licenciada en Ciencias de la Ingenieria de la Universidada Católica de la Santisima Concepción, Magister en Gestion de Tecnologías de Información de la Universidad del Biobio (Chile), Master Honours and Doctor of Philosophy (PhD) de la Universidad de Bond (Australia). Actualmente profesor Asistente del Departamento de Ingeniería Informática de la Universidad Católica de la Santísima Concepción (Chile), sus principales áreas de investigación son: gestión de datos,calidad de datos en salud , innovacion tecnológica e informática médica.