PREDICCIÓN DE FECHAS DE FALLECIMIENTO

Diana Salazar Báez, Iván Morales Cotes
Universidad Distrital Francisco José de Caldas
Email: dsalaz539@gmail.com
Universidad Distrital Francisco José de Caldas
Email: ivanalbertomorales@yahoo.com

RESUMEN

En este artículo se describe el proceso de aplicación de una técnica de minería de datos para analizar una muestra de 6000 personas en búsqueda de determinar qué aspectos de los estilos de vida actuales afectan más la cantidad de años de vida de las personas. Se busca lograr predecir qué tanto tiempo de vida le queda a cada persona teniendo en cuenta sus hábitos.

Se utilizó la metodología KDD, realizando la selección, el preprocesamiento, la transformación, la minería, la interpretación y la evaluación de resultados con el fin de establecer un modelo que permitirá con cierto grado de certeza estimar la edad a la que llegará cada persona; como consecuencia las personas podrán tener más información para tomar decisiones que posiblemente permitirán aumentar su cantidad de años de vida.

PALABRAS CLAVE

Estilos de vida, técnicas de minería de datos, descubrimiento de patrones.

ABSTRACT

This paper reflects the process of applying a data mining technique to analyze a sample of 6000 people; the analysis seeks to determine what aspects of today's lifestyles affect the number of years of people's lives.

The KDD methodology will be used, making the selection, pre-processing, transformation, data mining, evaluation and implementation in order to facilitate decision-making to increase the number of years of life of people.

KEYWORDS

Lifestyles, data mining techniques, pattern discovery.

1. INTRODUCCIÓN

La minería de datos se ha convertido en una herramienta fundamental para la toma de decisiones y junto con las metodologías que se han desarrollado puede contribuir a identificar diferentes problemas y posibles soluciones.

Este trabajo se orienta a aplicar técnicas de minería de datos y la metodología KDD para el

análisis de información proveniente de una encuesta realizada a 6000 colombianos cuyas edades están entre los 30 y los 50 años. Posteriormente con la fecha de fallecimiento de cada uno de ellos, con la herramienta Weka se establecerá como variable dependiente la cantidad de años que vivirá cada persona dependiendo de los hábitos que se hayan registrado previamente en la encuesta.

En primer lugar se realiza el planteamiento del problema para continuar con la metodología escogida; luego se presenta el desarrollo de las fases de la metodología para finalizar con las conclusiones y los trabajos futuros.

2. PLANTEAMIENTO DEL PROBLEMA

Hoy en día es muy frecuente ver personas jóvenes con problemas de salud debido a que sus hábitos no son convenientes para su cuerpo.

No existen muchas herramientas tecnológicas que faciliten a las personas validar las consecuencias de sus hábitos; muchas personas no pueden ir a las respectivas citas médicas de control.

Tal vez, si se proporciona una herramienta web de fácil acceso y de alto grado de confiabilidad (respaldada por un modelo de conocimiento), las personas utilizarán dicha herramienta y en algunos casos cambiarán sus hábitos con el fin de alcanzar una mayor edad.

La pregunta principal que se busca responder con este proyecto es:

¿Cómo afectan las variables analizadas la cantidad de años de vida de las personas en Colombia?

La respuesta se buscará aplicando minería de datos.

3. MARCO DE REFERENCIA

Clasificador REPTree:

El algoritmo de clasificación REPTree permite construir un árbol de decisión utilizando la entropía como medida de impureza y emplea la ganancia/varianza sobre la información; también basado en la reducción del error aplica una poda. Se caracteriza por ordenar los valores numéricos solo una vez con la finalidad de mejorar el rendimiento y ser más rápido [1].

"Los nodos intermedios son los atributos de entrada de los ejemplos presentados, las ramas representan valores de dichos atributos y los nodos finales son los valores de la clase [2]*."*

En los árboles obtenidos con 'REPTree' se observa que de cada nodo salen dos ramas; esta estructura representa un conjunto de sentencias if-then anidadas [3], [4].

Opciones válidas que se pueden asignar son:

- Número mínimo de instancias por hoja (por defecto 2).
- Varianza mínima para división, proporción mínima de varianza de clase numérica de la varianza del vector para dividir.
- Número de pliegues para reducción de errores de corte (predeterminado 3).
- Semilla para barajar aleatoriamente los datos (por defecto 1).
- Sin poda.
- Profundidad máxima del árbol (por defecto -1, no máximo)

Clasificador J48:

Es un algoritmo que implementa el C4.5 y se utiliza para generar árboles de decisión para tareas de clasificación mediante particiones

realizadas recursivamente aplicando la técnica de profundidad primero; este algoritmo utiliza el factor de confianza para la poda "confidence level", siendo este parámetro el que afecta directamente el tamaño y capacidad de predicción del árbol construido [5] .

El algoritmo J48 se fundamenta principalmente en el uso de "gain ratio", esto con la finalidad de evitar que se seleccionen variables con mayor número de posibles valores [6].

Los pasos para construir el árbol consisten en:

- Determinar el atributo raíz y crear una rama con cada valor posible que pueda tomar.
- Se repite el proceso con cada rama resultante.
- En cada nodo se debe seleccionar un atributo clave que permita seguir dividiendo.

Este algoritmo se caracteriza por su rapidez a la hora de clasificar nuevos patrones y por escoger un rango de medida apropiado.

Algoritmo de Fayyad e Irani:

Para el proceso de discretización se puede utilizar El algoritmo **Fayyad e Irani,** el cual utiliza la longitud de descripción mínima (MDL) y los criterios relacionados con la entropía mínima; al ordenar los valores de las características continuas se forman barreras entre clases las cuales delimitan los puntos de corte [7].

"La mejor descripción del conjunto de datos es la que minimiza la longitud de la descripción de todo el conjunto de datos" [8].

"Fayyad e Irani para su método de discretización supervisado utiliza la siguiente notación: Si se tiene un conjunto de instancias S, un atributo X y una partición T, que divide a S en S_1 y S_2, la entropía de la clase de la información inducida por T y llamada E{X, T; S) viene dada por:" [9],

$$E(X,T;S) = \frac{|S_1|}{|S|}Ent(S_1) + \frac{|S_2|}{|S|}Ent(S_2)$$

"Donde $Ent(S_1)$ y $Ent(S_2)$ denotan respectivamente la entropía de X en los subconjuntos de instancias S_1 S_2. ;para un atributo dado X se selecciona la partición T_{min} que minimize la función de entropía sobre todas las posibles particiones como una discretización binaria. Este método es aplicado recursivamente a las particiones inducidas por T_{min} hasta que se llega a una condición de parada, creando de esta forma los múltiples intervalos del atributo X " [9].

4. METODOLOGÍA KDD (Knowledge Discovery in Databases)

Para este proyecto, se seleccionó la metodología KDD (Knowledge Discovery in Databases) debido a que es muy completa e incluye todos los pasos requeridos para tener datos con calidad y analizarlos de una manera efectiva consiguiendo identificar los patrones que luego serán utilizados durante la predicción; es decir, dada una muestra significativa de datos, se realizarán los pasos establecidos por la metodología y lo que se espera es que al aplicar el respectivo algoritmo, se logre determinar cómo unas variables independientes permiten predecir el valor de la variable dependiente (previamente seleccionada) con una alta probabilidad de acierto [10], [11].

En la figura 1 se observan los pasos para el desarrollo de la metodología KDD.

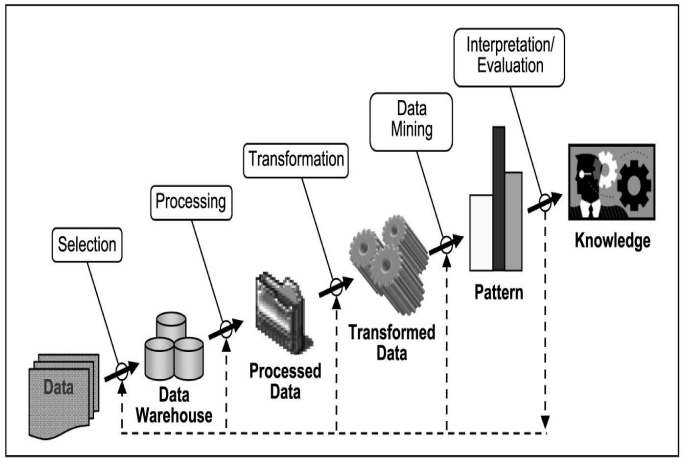


Figura 1. Etapas de la metodología KDD [12].

5. DESARROLLO

a. Selección de datos

El conjunto de datos seleccionado corresponde al resultado de la aplicación de una encuesta en línea. En dicha encuesta se tienen preguntas sobre el estado civil de la persona, su género, su edad, su estatura, si fuma y si consume multivitamínicos, como se puede ver en el siguiente enlace:

https://docs.google.com/forms/d/e/1FAIpQLSev1GjUgBTg9FqfEoEYfPU-XePkTGSngup87UT3dLRQPvYH3w/viewform?c=0&w=1

Dado que para este proyecto no era viable esperar mucho tiempo hasta que se respondieran las 6000 encuestas y que fallecieran las personas encuestadas, se realizó una generación aleatoria que nos permitió probar el proceso que luego será reproducido cuando se tengan los verdaderos datos.

A través de una ETL con SQL Server Data Tools, los datos se almacenaron en una base de datos (SQL Server) en una tabla llamada “datos”.

Cabe mencionar que algunos datos fueron modificados con el fin de generar ciertas casuísticas a analizar posteriormente.

b. Preprocesamiento de datos

Se encontraron datos duplicados, inconsistentes e incompletos, por lo cual en la base de datos se ejecutaron unos scripts que se encargaron de revisar cada uno de los registros con el fin de mover a otra tabla llamada “inconsistencias” todos aquellos que no cumplieron con ciertas validaciones como se explica a continuación.

El primer script ejecutado en SQL Server correspondió a un conjunto de consultas que permitieron identificar datos atípicos extremos con respecto a la edad en la que se realizó la encuesta.

El segundo script ejecutado correspondió al borrado de aquellas filas con estatura mayor a 210 centímetros, aquellas sin género y con respecto a los datos atípicos previamente identificados, solo se borraron aquellas filas con edad menor a uno mientras que se mantuvieron aquellas con edad mayor a 78, debido a que se consideró que sí se pueden presentar casos de edades superiores; es importante resaltar que antes de borrar dichas filas, las mismas fueron insertadas en una tabla llamada “inconsistencias” para que posteriormente se les pueda realizar un análisis diferente.

Por otro lado, se detectó que faltaban algunos datos correspondientes a los atributos "multivitaminico" y "estatura"; para dar solución a dicha situación se aplicó en Weka el algoritmo correspondiente a "ReplaceMissingValues" (de "unsupervised.attribute") con el fin de "rellenar" los datos faltantes sin afectar significativamente los resultados.

Teniendo en cuenta que se utilizaron también datos provenientes de una encuesta realizada en el año 2014 (128 filas), y que se detectó que habían datos correspondientes al atributo "estado civil" desactualizados, se consumió un servicio web expuesto por la Registraduría Nacional con el fin de actualizar dichos datos; dicha actualización se reflejó en 17 registros.

c. Transformación de datos

Posteriormente se llevó a cabo una discretización de atributos como se explica a continuación:

En SQL Server se ejecutó una discretización del atributo "edad de fallecimiento" utilizando "simple binning" con 4 "bins".

Luego se exportaron los datos de SQL Server a un archivo plano que posteriormente fue modificado (cabecera y formato) para tener un archivo arff compatible con Weka 3.8.1.

Después de cargar el archivo arff en Weka se realizó una discretización de los atributos "estatura" (cms) y "edad al realizar encuesta" utilizando el algoritmo de Fayyad e Irani.

Se escogió dicho algoritmo teniendo en cuenta que al revisar su documentación (en parte explicada previamente) se consideró que cumple con lo requerido para realizar la discretización de los respectivos campos (en el procesamiento posterior se trabajará con variables categóricas).

El resultado de la discretización realizada en Weka se aplicó sobre los datos en SQL Server con un script; luego se creó y ejecutó una vista para exportar a un archivo plano los datos que posteriormente también se cargaron en Weka (archivo arff).

En el siguiente punto se hace referencia al modelo aplicado sobre los datos recién cargados en Weka.

d. Selección y aplicación del modelo

Para aplicar minería de datos sobre nuestra base de datos (encuestas) se escogió el clasificador "REPTree" en Weka.

Para una primera ejecución se tomó como conjunto de datos de entrenamiento (utilizado por el algoritmo) el 75% del total de los datos, mientras que el conjunto de datos de pruebas (para verificar la eficacia del modelo) correspondió al 25%.

Al aplicar el clasificador se seleccionó como atributo a predecir (clase) la cantidad de años aproximada (rango de edad) que las personas vivirán, dado que es ese el atributo que las personas desean conocer sabiendo de antemano sus hábitos y características físicas.

Como se puede observar en la Figura 2, se obtuvo un árbol de decisión que excluyó algunos atributos por su irrelevancia en el contexto de la predicción de las edades que alcanzarán las personas.

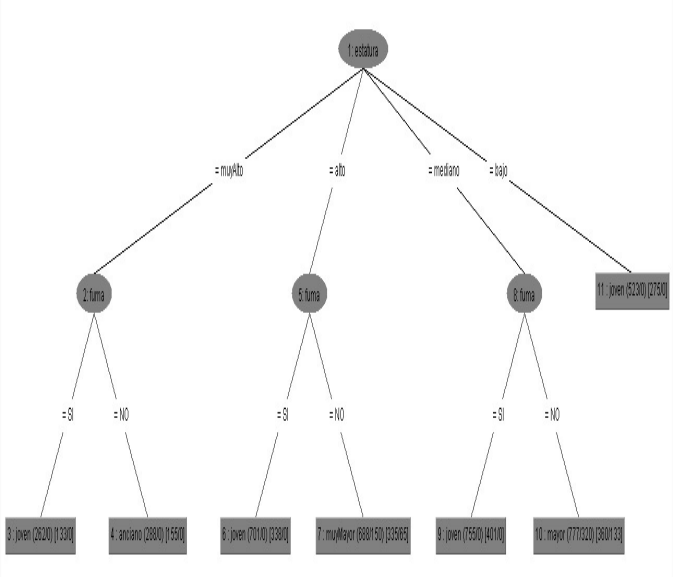
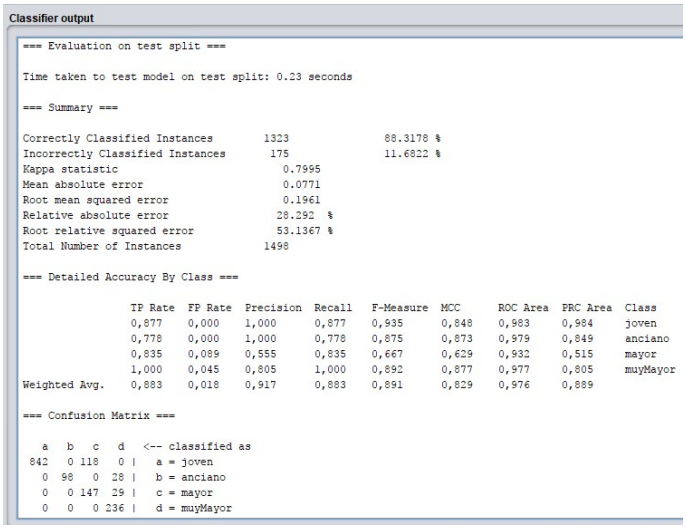


Figura 2. Árbol de decisión aplicando el clasificador REPTree

Posteriormente, se realizó una segunda ejecución pero esta vez en vez de utilizar "percentage split" se utilizó la opción "use training set", la cual en vez de realizar el entrenamiento con un porcentaje predefinido de los datos (ejemplo: 75%), lo hace con el conjunto total. De igual manera ocurre con los datos de pruebas. Para dicha ejecución el porcentaje de aciertos fue 88.8499%.

En el siguiente punto se explicará el resultado obtenido, y se dará un ejemplo de aplicación futura. De igual manera se explicará la Figura 3, la cual muestra entre otras cosas el porcentaje de acierto del modelo sobre la respectiva muestra.



Classifier output

```
=== Evaluation on test split ===

Time taken to test model on test split: 0.23 seconds

=== Summary ===

Correctly Classified Instances        1323               88.3178 %
Incorrectly Classified Instances       175               11.6822 %
Kappa statistic                          0.7995
Mean absolute error                      0.0771
Root mean squared error                  0.1961
Relative absolute error                 28.292  %
Root relative squared error             53.1367 %
Total Number of Instances             1498

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,877    0,000    1,000      0,877    0,935      0,848    0,983     0,984     joven
                 0,778    0,000    1,000      0,778    0,875      0,873    0,979     0,849     anciano
                 0,835    0,089    0,555      0,835    0,667      0,629    0,932     0,515     mayor
                 1,000    0,045    0,805      1,000    0,892      0,877    0,977     0,805     muyMayor
Weighted Avg.    0,883    0,018    0,917      0,883    0,891      0,829    0,976     0,889

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 842   0 118   0 |   a = joven
   0  98   0  28 |   b = anciano
   0   0 147  29 |   c = mayor
   0   0   0 236 |   d = muyMayor
```

Figura 3. Resultado al aplicar REPTree

e. Interpretación del resultado

Lo que se logró aplicando el clasificador "REPTree" fue generar un árbol de decisión que permite estimar la cantidad de años que alcanzarán a vivir las personas dados sus hábitos y características.

Analizando dicho árbol se puede ver por ejemplo que si una persona tiene alta estatura y no fuma, probablemente llegará a ser anciana (mayor a 75 años), pero si otra persona también de alta estatura fuma, esta última posiblemente fallecerá "joven" (menor a 58 años).

Dado lo anterior, se espera que con el modelo, cuando una nueva persona llene la encuesta, un sistema de información podrá informarle si continuando con sus hábitos logrará vivir cierta cantidad de años.

f. Evaluación del modelo

Al aplicar el modelo sobre 1500 registros correspondientes al 25% de los datos (destinados para pruebas), se llegó a la siguiente matriz de confusión:

		Predicción del modelo			
		Joven	Anciano	Mayor	muyMayor
Datos reales	Joven	842	0	118	0
	Anciano	0	98	0	28
	Mayor	0	0	147	29
	Muy Mayor	0	0	0	236

Tabla 1. Resultados de la predicción del modelo.

Interpretando la matriz anterior podemos decir que para la muestra de 1498 registros, según la predicción generada a partir del árbol, hubo 842 casos cuya edad de fallecimiento fue "joven" y el resultado del modelo fue acertado, mientras que para otros 118 casos el modelo "falló" creyendo que la edad de fallecimiento era "mayor" (entre 59 y 66 años).

De manera análoga, según la predicción generada a partir del árbol, hubo 147 casos cuya edad de fallecimiento fue "mayor" y el resultado del modelo fue acertado, mientras que para otros 29 casos el modelo falló creyendo que la edad de fallecimiento era "muyMayor" (entre 67 y 74 años).

Teniendo en cuenta los resultados generados en Weka, se podría estimar que el porcentaje de aciertos del modelo al predecir es aproximadamente 88.3178%.

Se realizó la ejecución con el clasificador J48 obteniendo el mismo porcentaje de aciertos.

Lo anterior nos muestra como el clasificador REPTree para ciertos problemas nos lleva a resultados satisfactorios.

6. CONCLUSIONES

Dados los resultados obtenidos con el proceso metodológico de minería de datos, se podría concluir que el modelo puede aplicarse por ejemplo en un portal web (como el de Colpensiones) de tal forma que cualquier persona que tenga entre 30 y 50 años de edad en Colombia, podría llenar la encuesta e inmediatamente obtener un resultado que le permita predecir (con cierto margen de error) cuántos años (rango de edad) le quedan de vida si continúa con sus hábitos.

Con lo anterior, una persona que esté interesada en vivir muchos años, podría determinar si debería cambiar sus hábitos o no (los datos ingresados en la encuesta se podrán ajustar en un "simulador" para ver cómo podría lograrse cambiar el resultado).

7. TRABAJOS FUTUROS

Se espera que en un futuro cercano se analicen los resultados del modelo actual y se busque la manera de obtener un porcentaje mayor de aciertos posiblemente utilizando otro algoritmo y/o aumentando significativamente la cantidad de datos durante el entrenamiento. Se espera que una empresa como por ejemplo Colpensiones (la cual tiene acceso a las fechas de fallecimiento de las personas y está informada sobre este artículo) pueda aplicar el modelo en un portal web y permitirle a los colombianos estimar su fecha de fallecimiento teniendo en cuenta sus hábitos y características físicas.

Sería conveniente que en un futuro cercano, se desarrolle una aplicación web que inicialmente reciba un archivo "arff" y a partir del mismo

genere y muestre un árbol de decisión (como lo hace Weka con el clasificador REPTree); el árbol como tal deberá ser almacenado en una base de datos de tal forma que la misma aplicación también generará automáticamente una interfaz web que podrá ser utilizada por diferentes usuarios, ingresando datos reales con el fin de ver el resultado de la predicción (dada por el árbol).

REFERENCIAS

[1]"REPTree", *Weka.sourceforge.net*, 2011. [Online]. Available: http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/REPTree.html. [Accessed: 29- Nov- 2017].

[2] S. Álvarez Teruelo and I. Fernández Pacheco, "Análisis de Predicción de Terremotos". 2015. Available: en: http://www.it.uc3m.es/jvillena/irc/practicas/07-08/PrediccionTerremotos.pdf [Accessed: 30- Oct- 2017].

[3] Universidad Carlos III de Madrid, "Evaluación de Modelos para predicción meteorológica", Available:http://www.it.uc3m.es/jvillena/irc/practicas/04-05/21mem.pdf [Accessed: 30- Oct- 2017].

[4] Dr. B. Srinivasan, P.Mekala, "Mining Social Networking Data for Classification Using REPTree", International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 10, October 2014 pp-155-160.

[5] B. Sierra Araujo. *Aprendizaje Automático: conceptos básicos y avanzados. Aspectos piráticos utilizando el software WEKA*. s.l. : Pearson, Prentice Hal, 2006.

[6] M. García Jiménez and A. Álvarez Sierra, *"Análisis de Datos en WEKA – Pruebas de Selectividad"*, 2010. Disponible en http://www.it.uc3m.es/jvillena/irc/practicas/06-07/28.pdf [Accessed: 15- Oct- 2017].

[7] A. Troncoso Lora, *"Técnicas de Preprocesado",* Series Temporales, Máster en Computación, Universitat Politècnica de Catalunya, 2015.

[8] U. Fayyad and K. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the 13th International Conference on Artificial Intelligence, pages 1022-1027, 1993.

[9] V. Robles Porcada, "Clasificación Supervisada basada en Redes Bayesianas. Aplicación en Biología Computacional", Licenciado en Informática, thesis, Universidad Politécnica de Madrid, Facultad de Informática, 2003.

[10] U. Fayyad and G. Piatetsky-Shapiro, *"From Data Mining to Knowledge Discovery in Databasesand Padhraic Smyth"*, American Association for Artificial Intelligence, 1996

[11] Data Mining of Qur'an - Mining the Quran, Sites.google.com (2015), [online] Available at: https://sites.google.com/site/miningthequran/text-mining-of-qur-an [Accessed 14 Oct. 2017].

[12] U. DBD, "KDD Process/Overview", Www2.cs.uregina.ca, 2013. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html. [Accessed: 30- Oct- 2017].