Un modelo predictivo para reducir la tasa de ausentismo en atenciones médicas programadas

Ing. Juan Miguel Moine
Ing. Cristian Germán Bigatti
Ing. Guillermo Leale
Est. Graciela Carnevali
Est. Esther Francheli

Grupo de Investigación en Minería de Datos
U.T.N. Facultad Regional Rosario

Resumen
El presente trabajo tiene por objeto describir la forma en que las técnicas de minería de datos pueden eficientizar la gestión de turnos médicos, analizando la historia de turnos otorgados y las distintas variables que pueden influir a la hora de decidir la asistencia a dicho turno. Se ha trabajado sobre un caso de estudio real, implementando técnicas de minería de datos a través de la metodología CRISP-DM.

Palabras clave
Minería de Datos, gestión de turnos médicos, CRISP-DM, árboles de decisión.

Introducción

Un problema muy frecuente al que día a día se enfrentan los centros de salud es la saturación de las agendas para las atenciones médicas programadas. En algunos casos, cuando un paciente necesita ser atendido, no logra conseguir un turno para una fecha a corto plazo, debiendo esperar varios días para poder asistir a una consulta. Esta situación se genera, fundamentalmente, debido al reducido horario de atención de algunos profesionales y a la gran demanda de atenciones médicas especializadas. El problema crece cuando existe una alta tasa de ausentismo, ya sea porque el paciente no asiste a la consulta, o bien porque cancela su turno sin anticipación, el mismo día de la atención.

El caso en estudio es el de un centro médico ubicado en la provincia de Santa Fe, Argentina. En el mismo trabajan 6 médicos, todos especialistas en Clínica Médica. Los directivos del centro médico están interesados en reducir la tasa de ausentismo para aprovechar mejor los horarios de atención de los profesionales, y de esta forma brindar al paciente una fecha de atención lo más pronta posible.

La problemática del centro médico fue abordada utilizando Minería de Datos. La Minería de Datos (Data Mining), es el proceso de extraer conocimiento útil, comprensible y novedoso de grandes volúmenes de datos. Su principal objetivo es encontrar información oculta o implícita, que no es posible obtener mediante métodos estadísticos convencionales [1,2].

Se estableció como objetivo del trabajo la optimización de la gestión de turnos mediante la construcción de un modelo predictivo que permita detectar aquellos pacientes que posiblemente no asistan a su consulta.

Elementos de trabajo y metodología

Para el desarrollo del trabajo se ha aplicado la metodología CRISP-DM (CRoss Industry Standard Process for Data Mining) [3,4]. Esta metodología fue desarrollada desde un enfoque práctico por las empresas SPSS, Daimler Chrysler y NCR y no está ligada a ningún producto comercial.

CRISP-DM propone seis fases para el proceso de minería de datos (Fig.1), cada una de las cuales se descompone en un conjunto de tareas. Para cada tarea la metodología propone los entregables que surgen como resultado de la ejecución de la misma.

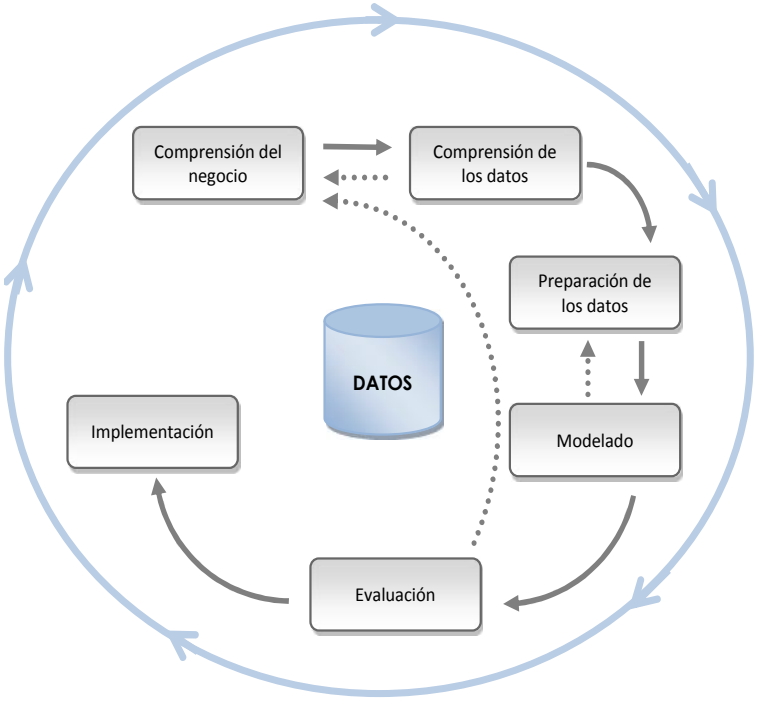

Fig. 1. *Metodología CRISP-DM*

En la fase de "Comprensión del negocio" se determinan los objetivos y requerimientos del proyecto desde una perspectiva organizacional, definiendo el problema de minería y el plan de trabajo.

La fase de "Comprensión de los datos" consiste en la recolección de datos que se utilizarán en el proyecto y la familiarización con los mismos. En esta etapa es posible el surgimiento de las primeras hipótesis acerca de la información que podría estar oculta.

La etapa de "Preparación de los datos" comprende aquellas actividades de limpieza y transformación de los datos. El resultado de esta fase será una vista minable[1] sobre la cual se aplicarán las técnicas de minería.

En la fase de "Modelado" se aplican las diversas técnicas y algoritmos sobre el conjunto de datos para obtener la información oculta y los patrones implícitos en ellos.

[1] Vista minable: tabla plana donde las filas corresponden a observaciones o individuos a analizar y las columnas a las variables seleccionadas como entrada para la técnica de minería

En este trabajo, por la naturaleza del problema que se aborda, se han seleccionado las siguientes técnicas de clasificación:

- Árboles de decisión: consisten en una serie de decisiones o condiciones organizadas en forma jerárquica, a modo de árbol, de tal manera que la decisión final a tomar se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta alguna de sus hojas [5].
- Regresión logística: permite determinar a partir de un conjunto de variables denominadas independientes la probabilidad de pertenencia de una observación futura a una clase determinada denominada variable dependiente [6].
- Método del vecino más cercano: en este método cada nuevo caso se compara con los existentes utilizando una métrica de distancia. Se asigna a la nueva instancia la clase mayoritaria entre los casos más próximos (vecinos más cercanos) [7].
- Clasificador Naive Bayes: es un clasificador probabilístico basado en el teorema de Bayes. Asume que las variables predictoras (o explicativas) son independientes, conocida la variable de respuesta [8].

En la fase de "Evaluación" se analizan las tareas realizadas al momento y los patrones obtenidos en función de los objetivos de negocio. En esta etapa se debería determinar si se ha omitido algún objetivo importante del negocio y si el nuevo conocimiento será implementado, es decir, si se pasará a la próxima etapa.

Finalmente, en la fase de "Implementación" se incorpora el nuevo conocimiento a los procesos de la organización, o bien se comunica a los directivos para dar soporte a la toma de decisiones estratégicas.

Resultados

Se presenta a continuación un breve resumen de los resultados obtenidos en cada una de las fases del trabajo.

Modelado del Negocio

En el caso de estudio, se determinó que el problema a abordar es la alta tasa de ausentismo de pacientes en atenciones médicas programadas. Este ausentismo genera que el turno no pueda ser aprovechado por otros pacientes.

Se estableció como objetivo del proyecto mejorar la gestión de turnos médicos, mediante la detección de pacientes que probablemente no asistan a su turno. Se dispuso que se trabajara con algoritmos de clasificación (árboles de decisión, regresión logística, Naive Bayes y vecinos más cercanos) para construir un modelo predictivo, que permita estimar si un paciente asistirá o no a la consulta. Con la información que brindará el modelo, se podrán tomar acciones preventivas como confirmaciones telefónicas con aquellos pacientes que presentan una alta probabilidad de ausentarse.

Comprensión de los datos

En el caso de estudio, los datos fueron extraídos de la base de datos del sistema operacional de turnos. El periodo en estudio fue desde el 01/01/2011 al 31/07/2011, donde

se ha recolectado información de 8810 turnos. Cada turno constituyó la unidad observacional del problema.

A partir de la información disponible se han extraído las siguientes variables (Tabla 1):

- Fecha del turno: fecha para la cual se le otorgó un turno al paciente.
- Hora: hora del turno.
- Día: día de la semana del turno.
- Médico: médico que asistirá al paciente.
- Tiene HC: indica si el paciente tiene historia clínica al momento que solicita el turno.
- Es sobreturno: indica si el turno fue otorgado excepcionalmente por no haber más turnos disponibles.
- Edad del paciente.
- Sexo paciente.
- Es de la ciudad: indica si el paciente vive en la ciudad donde se ubica el centro médico.
- Fecha de emisión: fecha en la que el turno fue otorgado.
- Atención por obra social: indica si el paciente se atendió por obra social.
- Asistió: indica si el paciente asistió a la consulta.

Atributo	Tipo	Valores posibles
Fecha turno	Fecha	[01/01/2011 – 31/07/2011]
Hora	Entero	[8-20]
Día	Categórica	[Lunes – Viernes]
Médico	Categórica	M1, M2, M3, M4, M5, M6
Tiene HC	Booleana	Si/no
Es sobreturno	Booleana	Si/no
Edad	Entero	
Sexo	Categórica	M / F
Es de la ciudad	Booleana	Si/no
Fecha de emisión	Fecha	
Atención OS	Booleana	Si/no
Asistió	Booleana	Si/no

Tabla 1. *Descripción de las variables extraídas.*

La Figura 2 muestra un primer análisis exploratorio realizado con el software WEKA [9], donde se puede analizar la distribución de cada una de las variables en estudio. Las instancias en color rojo representan los turnos de pacientes ausentes.

Analizando la distribución de la variable de respuesta (asistió) se puede observar que del total de turnos otorgados, en 1226 de ellos el paciente no ha asistido a la consulta (14% ausentismo).

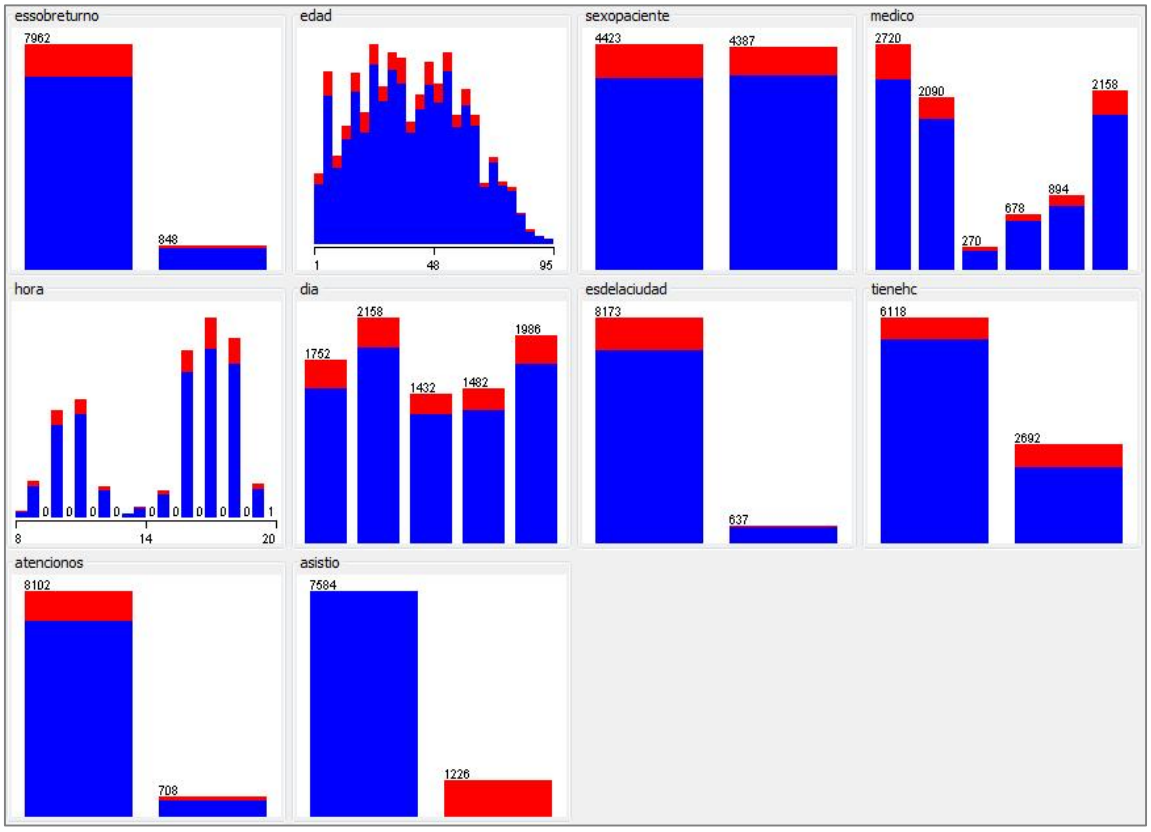

Fig. 2. *Distribución de las variables.*

Durante el estudio descriptivo se ha descubierto que la tasa de ausentismo resulta menor si discriminamos por las variables "es sobreturno" y "es de la ciudad". Para los pacientes que solicitan sobreturno la tasa de ausentismo fue de 9.3%. Para los pacientes que son de otra ciudad, la tasa de ausentismo se vio reducida al 6.3%.

Debido a que el clima es un factor muy importante que podría influir en el ausentismo de un paciente, se ha recolectado información externa acerca de las condiciones climáticas de cada día de atención [10]. A partir de estos datos se ha creado una nueva variable llamada "llovió", que toma los valores "si" o "no" en función de si hubo precipitaciones en el momento del turno.

Se estima que en el 6% de los turnos registrados ha llovido (543 instancias). Analizando la variable en función del ausentismo, se ha detectado que en los días en que llovió la tasa de ausentismo fue de 17%, resultando 3% mayor a la tasa de ausentismo general.

Preparación de los datos

Luego de un proceso de depuración y transformación de las variables originales, se ha llegado a la construcción de una vista minable formada por 8810 observaciones (turnos) caracterizadas por las variables descriptas en la Tabla 2. La última variable, "asistió", representa la salida del modelo (variable explicada).

Atributo	Tipo	Valores posibles
Horario	Categórica	[mañana-tarde]
Día	Categórica	[Lunes – Viernes]
Médico	Categórica	M1, M2, M3, M4, M5, M6
Tiene HC	Booleana	Si/no
Es sobreturno	Booleana	Si/no
Edad	Entero	
Sexo	Categórica	M / F
Es de la ciudad	Booleana	Si/no
Días antelación	Entero	[0-30]
Atención OS	Booleana	Si/no
Llovió	Booleana	Si/no
Asistió	**Booleana**	**Si/no**

Tabla 2. *Variables del conjunto de datos final o vista minable.*

En la Figura 3 se presenta un resumen del conjunto de datos final. La variable de respuesta ("asistió") no está balanceada, ya que la cantidad de turnos a los que el paciente no asistió es mucho menor.

```
essobreturno      edad       sexopaciente medico      horario           dia
f:7962       Min.   : 1.00   F:4423       m1: 894   mañana:2860   FRIDAY   :1432
t: 848       1st Qu.:23.00   M:4387       m2:2158   tarde :5950   MONDAY   :1752
             Median :38.00                m3:2720                 THURSDAY :1986
             Mean   :39.23                m4: 678                 TUESDAY  :2158
             3rd Qu.:55.00                m5:2090                 WEDNESDAY:1482
             Max.   :95.00                m6: 270


 esdelaciudad tienehc  atencionos llovio   diasantelacionsolicitud asistio
 f: 637       f:2692   f: 708     f:8267   Min.   : 0.000          NO:1226
 t:8173       t:6118   t:8102     t: 543   1st Qu.: 0.000          SI:7584
                                          Median : 2.000
                                          Mean   : 3.577
                                          3rd Qu.: 6.000
                                          Max.   :30.000
```

Fig. 3. *Resumen del conjunto de datos final.*

Modelado

Durante la fase de modelado, se aplicaron cuatro técnicas de clasificación sobre el conjunto de datos: árboles de decisión, Naive Bayes, método del vecino más próximo (KNN) y regresión logística.

Se han tomado el 65% de los datos para entrenar el modelo y el 35% restante para testearlo.

Se evaluó la capacidad de cada modelo en función de su matriz de confusión y la matriz de costos de la Tabla 3.

	Clasificó ASISTIÓ	Clasificó NO ASISTIÓ
ASISTIÓ	0	1
NO ASISTIÓ	2,5	0

Tabla 3*. Matriz de costos.*

Como se puede observar, el costo más alto existe cuando el paciente no asiste a la consulta, y el modelo predice que asistirá. El costo de que el paciente asista cuando el modelo predice que no asistirá es menor, ya que el turno será igualmente ocupado.

Los resultados al aplicar las distintas técnicas arrojaron modelos poco precisos (especialmente al predecir los pacientes que no asistirán a la consulta) por el desbalance que existe entre las clases de la variable de respuesta. Ante esta situación se procedió a realizar un muestreo de la clase mayoritaria, reduciendo la cantidad de instancias a 2452 turnos (contra 7584 que tenía el conjunto de datos original) [11].

Los modelos obtenidos a partir del conjunto de datos muestreado han resultado ser mucho mejores que aquellos construidos sobre los datos originales. La capacidad predictiva de los modelos ha aumentado considerablemente, ya que la desproporción entre la clase mayoritaria y la minoritaria se ha reducido.

La Tabla 4 muestra un resumen de los valores obtenidos al aplicar las diferentes técnicas de minería de datos. Para cada modelo se ha evaluado su capacidad predictiva y el costo total del mismo.

Técnica	Capacidad predictiva	Costo total del modelo
Naive Bayes	66%	727
Vecinos más próximos	59%	784.5
Regresión logística	67%	961.5
Árboles de decisión	68%	657

***Tabla 4**. Resultados obtenidos con las diferentes técnicas.*

Evaluación

El modelo de menor costo se ha obtenido con la técnica de árboles de decisión (Figura 4), el cual posee una capacidad predictiva general del 68%.

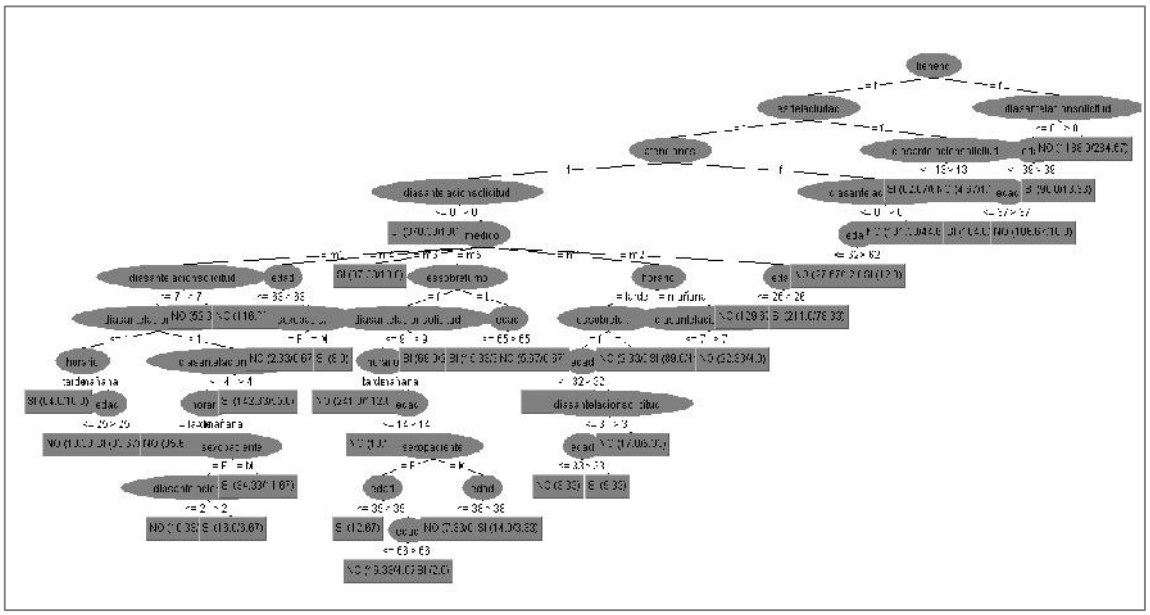

***Fig 4**. Árbol de decisión resultante.*

Como se puede observar en la Figura 5, la matriz de confusión refleja la correcta capacidad predictiva del modelo. Para los turnos donde el paciente no asistió, se han clasificado correctamente 300 de 464 turnos (64%). En aquellos casos donde el paciente asistió, se han clasificado correctamente 576 de 823 turnos (70%).

```
=== Confusion Matrix ===

   a   b   <-- classified as
 576 247 |   a = SI
 164 300 |   b = NO
```

***Fig. 5**. Matriz de confusión del árbol de decisión*

Implementación

Se ha efectuado una planificación acerca de cómo el modelo resultante podría ser implementado en el sistema operacional de turnos. Esta planificación incluye la especificación de un programa de monitoreo y control para la evaluación continua del desempeño del modelo.

Discusión

En el presente trabajo se ha demostrado cómo la minería de datos puede utilizarse como herramienta para optimizar la gestión de turnos médicos. Utilizando la metodología CRISP-DM y diferentes técnicas de clasificación, se logró la construcción de un modelo predictivo que permite estimar la asistencia de un paciente a su turno.

A pesar de haber obtenido un buen modelo con los datos que disponía el centro médico, el mismo podría mejorarse notablemente si se dispusiera de mayor cantidad de variables.

Está en estudio la elección de otras variables pertinentes que puedan mejorar el ajuste del modelo para lo que se están llevando a cabo estudios pilotos que apoyen la decisión de aumentar el número de variables en relación al costo – beneficio de registrarlas.

Referencias

1. Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). *The KDD process for extracting useful knowledge from volumes of data*. Communications of the ACM, 39(11), 27-34.
2. Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). *Knowledge discovery and data mining: Towards a unifying framework*. Knowledge Discovery and Data Mining, 82-88.
3. Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*.
4. Wirth, R., & Hipp, J. (2000). *CRISP-DM: Towards a standard process model for data mining*. IV International Conference on the Practical Applications of Knowledge Discovery and Data Mining (pp. 29-39).
5. Witten, I. H., Frank, E., & Hall, M. A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.
6. Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example* (Vol. 607). Wiley-Interscience.
7. Orallo, J. H., Quintana, M. J. R., & Ramírez, C. F. (2004). *Introducción a la Minería de Datos.* Pearson Prentice Hall.
8. Han, J., & Kamber, M. (2006). *Data mining: concepts and techniques*. Morgan Kaufmann.

9. University of Waikato. *WEKA, Data Mining Software in JAVA*. Recuperado el 15 de agosto de 2011, de http://www.cs.waikato.ac.nz/ml/weka.
10. TuTiempo Network. *Clima Mundial, Datos Históricos Climáticos*. Recuperado el 02 de septiembre de 2011, de http://www.tutiempo.net/clima.
11. Garcia, V. (2010). *Distribuciones de Clases No Balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje*. (Tesis Doctoral). Departament de llenguatges i Sistemes Informàtics, Universitat Jaume I. España.

Datos de Contacto

Ing. Juan Miguel Moine, U.T.N. Facultad Regional Rosario, juanmiguelmoine@gmail.com
Ing. Cristian Germán Bigatti, U.T.N. Facultad Regional Rosario, cristianbigatti@gmail.com
Ing. Guillermo Leale, U.T.N. Facultad Regional Rosario, guillermoleale@gmail.com
Est. Graciela Carnevali, U.T.N. Facultad Regional Rosario, greisc2002@yahoo.com.ar
Est. Esther Francheli, U.T.N. Facultad Regional Rosario, efranchelli@hotmail.com