San Marcos

MIEMBRO DE LA RED ILUMNO

APROXIMACIÓN DE FUNCIONES CONTINUAS Y DISCRETAS

APROXIMACIÓN DE FUNCIONES CONTINUAS Y DISCRETAS

APROXIMACIÓN DE FUNCIONES CONTINUAS Y DISCRETAS CON K VECINOS CERCANOS

De acuerdo con Prada (2013):

> En la teoría de la probabilidad y en estadística, un proceso de Márkov, llamado así por el matemático ruso Andréi Márkov, es un fenómeno aleatorio dependiente del tiempo para el cual se cumple una propiedad específica: la propiedad de Márkov. En una descripción común, un proceso estocástico con la propiedad de Márkov, o sin memoria, es uno para el cual la probabilidad condicional sobre el estado presente, futuro y pasado del sistema son independientes. (p. 47)

En el ámbito de los sistemas de eventos discretos, esto implica que en cada etapa del sistema se puede actuar sobre alguna variable capaz de influir en la evolución futura del sistema, permitiendo un cierto grado de control sobre su rendimiento. Si el sistema se modela mediante una cadena de Markov, la variable de control, denominada genéricamente como u, determinará las probabilidades de la matriz de transición. Tendremos por tanto una función P(u), que devuelve una matriz de transición por cada valor de u. Tras cada transición hay que decidir qué valor debe tomar u en el estado en que nos encontramos. A estos sistemas se les denomina Procesos de Decisión de Markov (MDP).

Según Prada (2013)

> En los MDPs se deben asociar los estados (o las transiciones) a un determinado coste o beneficio. El problema abordado en un MDP consiste en encontrar el mejor valor posible de u (máximo beneficio o mínimo coste) para cada estado del sistema, lo que se denomina política óptima. La mayor dificultad consiste en que se debe optimizar no sólo el rendimiento esperado en el estado en el que tomamos la decisión, sino también el rendimiento esperado en los futuros estados a los que llegamos como consecuencia de la decisión tomada. La base de las técnicas que nos permiten encontrar la política óptima es la programación dinámica. El objetivo es diseñar las normas de funcionamiento de un sistema para que su rendimiento sea el mejor posible. (p. 48)

REGRESIÓN LOCAL MEDIANTE FUNCIONES LINEALES

El modelo de pronóstico de regresión lineal permite hallar el valor esperado de una variable aleatoria a cuando b toma un valor específico. La aplicación de este método implica un supuesto de linealidad cuando la demanda presenta un comportamiento creciente o decreciente, por tal razón, se hace indispensable que previo a la selección de este método exista un análisis de regresión que determine la intensidad de las relaciones entre las variables que componen el modelo.

La función **ESTIMACIÓN LINEAL** calcula las estadísticas de una línea con el método de los "mínimos cuadrados" para calcular la línea recta que mejor se ajuste a los datos y después devuelve una matriz que describe la línea. También puede combinar **ESTIMACIÓN LINEAL** con otras funciones para calcular las estadísticas de otros tipos de modelos que son lineales en los parámetros desconocidos, incluidas series polinómicas, logarítmicas, exponenciales y de potencias. Debido a que esta función devuelve una matriz de valores, debe ser especificada como fórmula de matriz. Encontrará las instrucciones correspondientes tras los ejemplos de este artículo. (Laudon, 2008, p. 243)
La ecuación para la línea es la siguiente:

$$y = mx + b, \text{ o bien, } y = m1x1 + m2x2 + \ldots + b$$

Si hay varios rangos de valores x, donde los valores y dependientes son función de los valores x independientes. Los valores m son coeficientes que corresponden a cada valor x, y b es un valor constante. Observe que y, x y m pueden ser vectores. La matriz que devuelve la función **ESTIMACIÓN.LINEAL** es {mn,mn-1,...,m1,b}. **ESTIMACIÓN.LINEAL** también puede devolver estadísticas de regresión adicionales.

Sintaxis

ESTIMACIÓN.LINEAL (conocido_y, [conocido_x], [constante], [estadística]).

EL MODELO DE REGRESIÓN LINEAL SIMPLE

Al tomar observaciones de ambas variables Y respuesta y X predicción o regresor, se puede representar cada punto en un diagrama de dispersión.

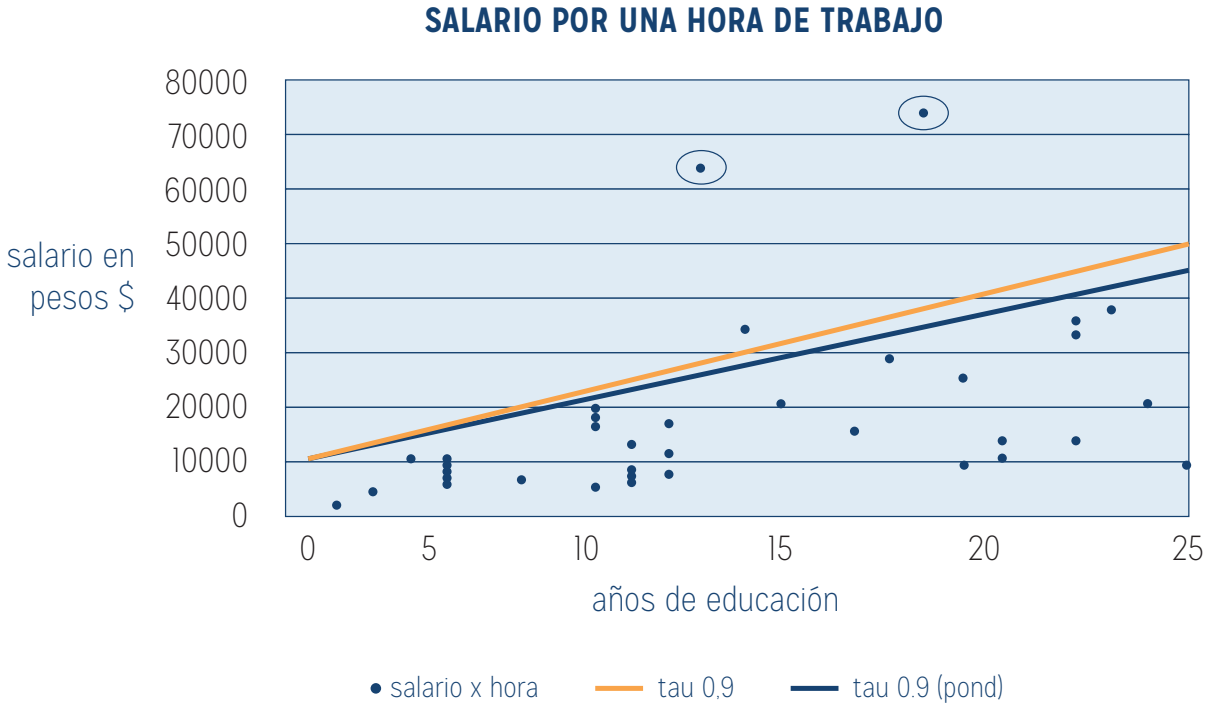


Figura 1. Rectas de regresión cuantífica y cuantílica ponderada.
Fuente. Elaboración Propia. Tomada de: https://www.google.com/search?q=regresi%C3%B3n+local+mediante+funciones+lineales+en+word&biw=1366&bih=657&source=lnms&tbm=isch&sa=X&ved=0ahUKEwjXsuTL_-bKAhXFFR4KHX-dBj8Q_AUIBygC#imgrc=8DxHKLjKOvg0JM%3A

El modelo de ajuste o modelo de regresión lineal es:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Donde los coeficientes $\beta_0 + \beta_1$ son parámetros del modelo denominados coeficientes de regresión, son constantes, a pesar de que no podemos determinarlos exactamente sin examinar todas las posibles ocurrencias de X y Y, podemos usar la información proporcionada por una muestra para hallar sus estimados b_0 , b_1. El error es difícil de determinar puesto que cambia con cada observación Y. Se asume que los errores tienen media cero, varianza desconocida σ^2 y no están correlacionados (el valor de uno no depende del valor de otro). (Laudon, 2008, p. 143)

Por esto mismo las respuestas tampoco están correlacionadas.

Conviene ver al regresor o predictor ***X*** como la variable controlada por el analista y evaluada con el mínimo error, mientras que la variable de respuesta ***Y*** es una variable aleatoria, es decir que existe una distribución de ***Y*** con cada valor de ***X.***

Este tipo de modelo sus aplicaciones se pueden subdividir en:

I. Predicción de nuevas observaciones

Esta es otra de las aplicaciones del modelo de regresión, predecir nuevas observaciones Y correspondientes a un nivel específico de la variable regresora X. La banda de predicción es más ancha dado que depende tanto del error del modelo de ajuste y el error asociado con observaciones futuras $(Y_0 - \bar{Y}_0)$. El intervalo es mínimo en $X_0 = \bar{X}_0$ y se amplia conforme se incrementa la diferencia $X_0 = \bar{X}$.

La variable aleatoria, $\varphi = Y_0 - \bar{Y}_0$

Está normalmente distribuida con media cero y varianza:

$$V(\varphi) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right]$$

Si se usa $\bar{Y}_0$ para predecir a Y_0, entonces el error estándar de $\varphi = (Y_0 - \bar{Y}_0)$, es el estadístico apropiado para establecer un intervalo de predicción probabilístico, en el caso de un intervalo 100 (1 - α) % sobre una observación futura en x_0 se tiene:

$$\hat{Y}_0 - t_{\alpha/2,n-2} \sqrt{MSE \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right]} \leq Y_0 \leq -\hat{Y}_0 + t_{\alpha/2,n-2} \sqrt{MSE \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right]}$$

Se puede generalizar para encontrar un intervalo de predicción del 100 (1-à)φporciento para la media de m observaciones futuras en X = Xo. Sea Ymedia la media de las observaciones futuras en X = Xo. El intervalo de predicción estimado es:

$$\hat{Y}_0 - t_{\alpha/2,n-2} \sqrt{MSE \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right]} \leq Y_0 \leq -\hat{Y}_0 + t_{\alpha/2,n-2} \sqrt{MSE \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{S_{xx}} \right]}$$

PRUEBAS DE HIPÓTESIS PARA LA PENDIENTE E INTERSECCIÓN

2. Prueba de hipótesis para Ho: $\beta_0 = \beta_{10}$ contra H_i: $\beta_0 \neq \beta_{10}$

Calculando el estadístico t, considerando que $\beta_{10} = 0$, se tiene:

$$t = \frac{b_0}{se(b_0)}$$

Probar la hipótesis para b0 no tiene interés práctico.

Ahora para probar la significancia de b_1 se tiene:

$$t_0 = \frac{b_1}{\sqrt{MSE / S_{xx}}} \qquad \text{para } (\alpha/2, n-2) \text{ grados de libertad}$$

Si $t_0 > t_{\alpha/2,n-2}$ se rechaza la hipótesis nula, indicando que b_1 es significativo y se tiene regresión lineal.

Del ejemplo:

$$t = \frac{b_1}{se(b_1)} = \frac{-0.798}{0.0105} = -7.60$$

Como t= 7.60 excede el valor crítico de t=2.069, se rechaza Ho (o sea el valor de p << 0.05) .Por tanto este coeficiente es significativo.

Es importante notar que el valor de $F = t_2$.

La salida del Minitab es como sigue:

Predictor	Coef	SE Coef	T	P
Constant = b0	13.6230	0.5815	23.43	0.000
C2 = b1	-0.07983	0.01052	-7.59	0.000

ADMINISTRACIÓN DEL CONOCIMIENTO Y MANTENIMIENTO DE MODELOS DE MINERÍA DE DATOS EN EL TIEMPO

De acuerdo con Greenberg (2001):

Es una de las técnicas que está influyendo en las actividades de negocio de las empresas y en la que están involucrados un extenso y creciente número de investigadores a nivel mundial, por las implicaciones, estrategias y beneficios que arroja. Como resultado de esta primera aproximación se presentan algunas de las consideraciones más resaltantes derivadas de los planteamientos teóricos, en torno a la minería de datos y su impacto en la toma de decisiones en los negocios:

- Nuestra capacidad para almacenar datos ha crecido exponencialmente los últimos años, pero la capacidad de procesarlos no ha ido a la par. Por tal motivo, es necesario contar con técnicas que tengan la capacidad de procesar y entender datos tanto estructurados como no estructurados, para apoyar la toma de decisiones en cualquier ámbito del conocimiento.
- La minería de datos ha tenido una reciente inclusión en los negocios, debido a la enorme preocupación de las empresas por conocer más allá de los datos que éstos manejan.
- La minería de datos, bien empleada, se convierte en una herramienta estratégica que eleva los niveles de competencia en el cambiante mundo de los negocios. La toma de decisiones efectivas depende de la rapidez con que se identifica y analiza información importante. La existencia de metodologías innovadoras para desarrollar el proceso de identificación y análisis, debe necesariamente mejorar la ventaja competitiva para incrementar el mayor número de clientes.
- Entre las ventajas de la minería de datos está su facilidad de uso y la aplicabilidad de un conocimiento adecuado de los distintos tipos de algoritmos empleados, ya que éstos brindan los mismos resultados y cada uno, con una eficiencia diferente. Como desventaja destaca que hay que dedicar mucho más esfuerzo al establecimiento de medidas de evaluación del resultado derivado de la aplicación del Data Mining. (Hansen y Mouritsen, 1999, p. 122)

- No todos los datos son apropiados para la minería. La búsqueda de patrones debe centrarse en aquellos que tengan un impacto significativo en el negocio. Si bien los datos de poca utilización se encuentran mezclados con los de alta utilización, contar con un motor de consultas que permita realizar ordenamientos y selección de datos ayuda a determinar cuáles serán aquéllos que se extraerán. (Hansen y Mouritsen, 1999, p. 122)
- "Un Data Warehouse está diseñado para realizar procesamientos veloces de consultas, lo cual representa una herramienta de suma utilidad en la tarea de identificación del subconjunto de datos requerido." (Hansen y Mouritsen, 1999, p. 122)
- Hacer *Data Mining* sobre datos que se actualizan a menudo es un desafío. Por consiguiente, hay varios problemas que necesitan ser investigados extensamente, antes de que se pueda llevar a cabo lo que se conoce como Data Mining en tiempo real. El uso de indicadores en pro de medir la bondad, aplicabilidad, la relevancia y la novedad, de los resultados de la minería de datos, pueden resultar en ocasiones muy subjetivos, pero los negocios necesitan contar con algún medio que les permita medir el interés y el impacto del conocimiento que se puede obtener al aplicar minería de datos. De igual manera, la intervención y experiencia del tomador de decisiones es relevante para establecer algunas medidas y poder calcular los indicadores antes mencionados. (Hansen y Mouritsen, 1999, p. 123)
- "Para el aprovechamiento de la gran cantidad de conocimiento en la minería de datos es necesario reducir la cantidad de datos, quedándonos sólo con la información mínima necesaria, para disminuir el esfuerzo computacional y humano. El resto de la información se vuelve redundante, trayendo consigo ruido y dependencias que deben tratar de evitarse; esto se basa en un axioma fundamental: la hipótesis más simple." (Hansen y Mouritsen, 1999, p. 122)
- La tenencia de datos no es el elemento esencial en una toma de decisión acertada. Al convertir dichos datos en información evaluada y ésta en conocimiento para la acción, se proporciona el apoyo necesario para la toma de una decisión argumentada, que oriente a la empresa hacia el cumplimiento de sus metas y objetivos. (Hansen y Mouritsen, 1999, p. 122)
- Con *Data Mining*, las organizaciones cuentan con una nueva forma de ver sus datos, prometiendo beneficios a la solución de una gran variedad de problemas como: planeación económica, inteligencia empresarial, finanzas, análisis de mercados y análisis de perfiles de clientes. (Hansen y Mouritsen, 1999, p. 123)

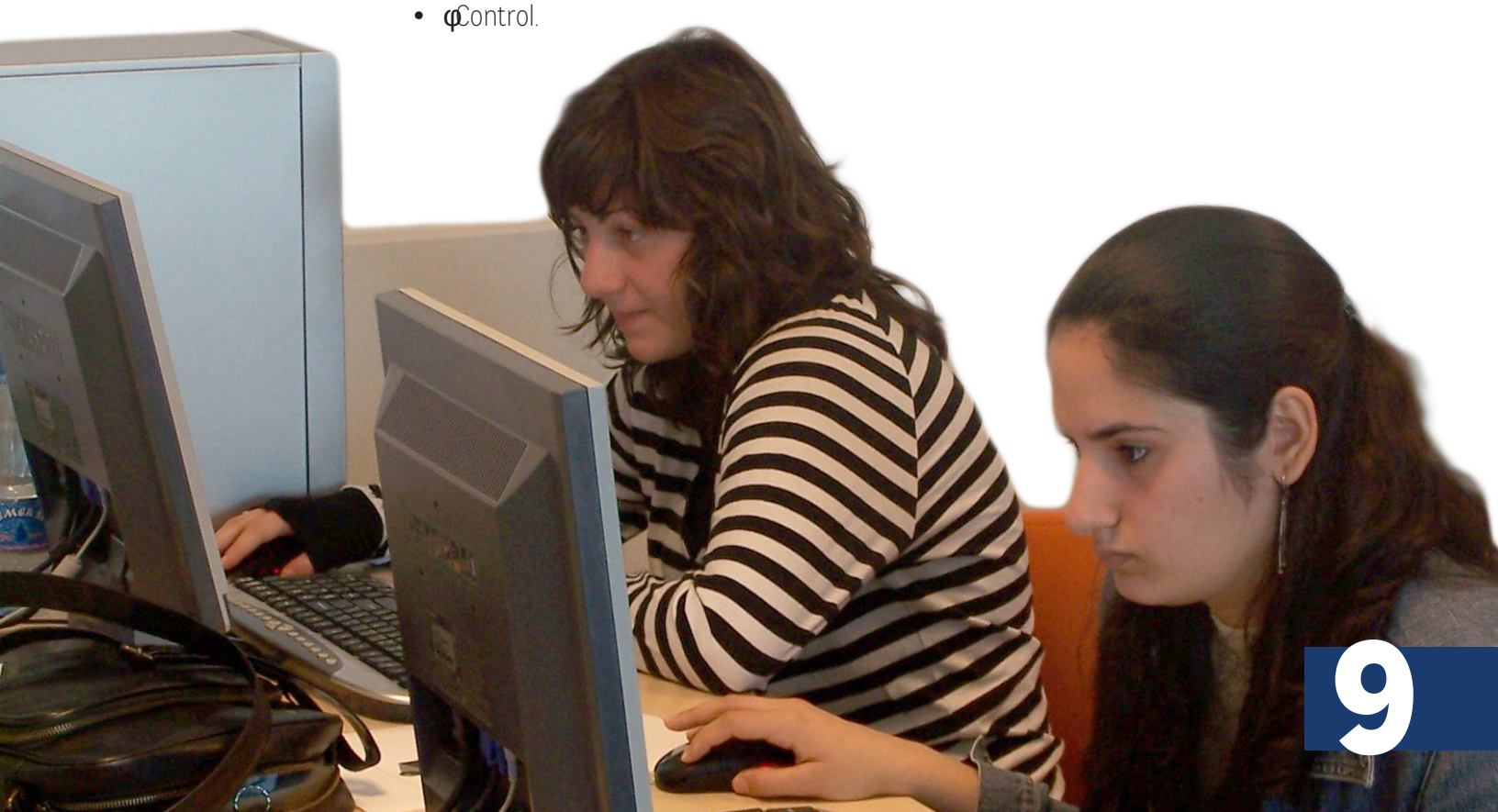
Aunque todavía queda mucho trabajo por hacer en esta temática, se necesita crear más y mejores procesos para generar resultados eficientes en los negocios, y más aún, desarrollar aplicaciones reales que pongan en práctica todos los principios relacionados con la minería de datos. En espera de que se alcance esta meta, se han desarrollado y probado ideas a niveles específicos, abriendo nuevos campos para la innovación, suficientemente interesantes para motivar la investigación en esta área.

El método de análisis llamado análisis de regresión, investiga y modela la relación entre una variable Y dependiente o de respuesta en función de otras variables de predicción X's, a través del método de mínimos cuadrados (Hansen y Mouritsen, 1999, p. 124)

Como ejemplo supóngase que un ingeniero industrial de una embotelladora está analizando la entrega de producto y el servicio requerido por un operador de ruta para surtir y dar mantenimiento a maquinas dispensadoras. El ingeniero visita 25 locales al azar con máquinas dispensadoras, observando el tiempo de entrega en minutos y el volumen de producto surtido en cada uno.

En general los modelos de regresión tienen varios propósitos:

- Descripción de datos a través de ecuaciones
- Estimación de parámetros para obtener una ecuación modelo
- Predicción y estimación.
- Control.



RECALIBRAR MODELOS

La información relativa a clientes y su entorno se ha convertido en fuente de prevención de riesgos de crédito. En efecto, existe una tendencia general en todos los sectores a recoger, almacenar y analizar información crediticia como soporte a la toma de decisiones de análisis de riesgos de crédito.

Los avances en la tecnología de *Data Warehouse* hacen posible la optimización de los sistemas de análisis de riesgo de crédito de acuerdo con Groth (2000):

Para la gestión del riesgo de crédito los sistemas operacionales han ofrecido:

- Sistemas de Información para Gerencia (MIS) e informes de Soporte a la Decisión de Problemas (DSS) estáticos y no abiertos a nuevas relaciones y orígenes de datos, situación en la que la incorporación de nuevas fuentes de información ha sido un problema en lugar de una ventaja.
- Exploraciones de datos e informes cerrados y estáticos.
- Análisis sin inclusión de consideraciones temporales lo que imposibilita el análisis del pasado y la previsión del futuro.
- Herramientas de *credit-scoring* no flexibles, construidas sobre algoritmos difícilmente modificables, no adaptados al entorno de la empresa, o exclusivamente basados en la experiencia personal no contrastada, con lo que los sistemas han ayudado a repetir los errores en vez de a corregirlos.

Pero estos sistemas tradicionales se enfrentan a una problemática difícil de resolver para acomodarse a las necesidades analíticas de los sistemas de análisis del riesgo, necesidades que se pueden cubrir mediante el uso de tecnologías de *Data Warehouse*. Dentro de la prevención de impagados, utilizando sistemas OLAP se puede obtener el grado interno de concentración de riesgos con el cliente, y almacenar la variedad de fuentes internas o externas de información disponibles sobre el mismo. Ello nos permite obtener sin dificultad la posición consolidada respecto al riesgo del cliente. El análisis se puede realizar asimismo por las diferentes características de la operación para la que se realiza el análisis, en cuanto al plazo y la cuantía de la misma, la modalidad de crédito elegida, la finalidad de la operación o las garantías asociadas a la misma. Usando las mismas capacidades es fácil el establecer una segmentación ABC de la cartera de clientes potenciales o reales que nos optimicen el nivel de esfuerzo en el análisis de riesgos.

En el soporte al proceso de anticipación al riesgo se puede dar un adecuado soporte a la correcta generación y consideración de señales de alerta, teniendo en cuenta las pautas y condicionantes diferenciados dependiendo del tipo de cliente y producto usando *Data Mining*. (Hansen y Mouritsen, 1999, p. 422)

Para el caso del seguimiento del ciclo de impagados, de nuevo el uso de sistemas OLAP, simplifican el análisis la diversidad de los diferentes parámetros que intervienen en el mismo, tales como la jerarquía de centros de recobro a contemplar, la diferente consideración dependiendo de la antigüedad del impago, del cliente o del importe impagado. Un sistema de *Data Mining* puede aconsejar la mejor acción en caso de impagados, litigio, precontencioso, etc. frente a los parámetros de importe, antigüedad, zona geográfica, etc.

Estos sistemas hacen que el analista se dedique con más intensidad al análisis de la información, que es donde aporta su mayor valor añadido, que a la obtención de la misma. No obstante, estos sistemas deben de huir de las automatizaciones completas sin intervención del analista: es él el que mejor sabe lo que quiere descubrir. (Hansen y Mouritsen, 1999, p. 128)

EVALUACIÓN DE RESULTADOS

LA EVALUACIÓN DEL RESULTADO DETERMINA LA APLICACIÓN SATISFACTORIA DE LAS HERRAMIENTAS DE MINERÍA DE DATOS, REDUCIÉNDOSE LA UTILIDAD DE ESTAS TÉCNICAS EN LA MEDIDA EN QUE NO SE EVALÚE ADECUADAMENTE LA INFORMACIÓN QUE GENERAN.

Cada vez más se están utilizando herramientas de minería de datos en los negocios, en general, y en la investigación de marketing en particular. Entre sus principales ventajas se encuentra la elevada capacidad predictiva manifestada por los resultados alcanzados, mientras que entre sus puntos débiles está la escasez de medidas para evaluar el resultado alcanzado.

Este trabajo trata de recoger y organizar en 4 facetas (bondad de ajuste, relevancia, novedad y aplicabilidad del resultado), las medidas que, de manera dispersa, aparecen en la literatura. El esquema que aquí se propone de 4 grupos de indicadores permitirá al profesional una toma de decisiones más eficiente.

Además, también es útil para identificar áreas donde se requieren mayores esfuerzos en el desarrollo de medidas de evaluación del resultado en minería de datos.

La evaluación del resultado determina la aplicación satisfactoria de las herramientas de minería de datos, reduciéndose la utilidad de estas técnicas en la medida en que no se evalúe adecuadamente la información que generan. La propuesta que se realiza en este trabajo viene a reunir indicadores y establecer un esquema de referencia en la evaluación de los resultados en minería de datos.

Se deduce de lo comentado, en primer lugar, que es necesario evaluar 4 facetas del resultado: bondad de ajuste, relevancia, novedad y aplicabilidad. El cálculo de estas medidas permitirá cumplir con las promesas que realiza la minería de datos a través de su definición. La particular característica en minería de datos de descubrimiento automático de información exige una evaluación más amplia de la información obtenida, más allá de la bondad del resultado. (Hansen y Mouritsen, 1999, p. 129)

Desde nuestro punto de vista resulta más conveniente un empleo conjunto de todos los coeficientes presentados, para conocer el posible interés e impacto de un proceso de minería de datos. El decisor tiene la oportunidad de filtrar las reglas y resultados obtenidos en el proceso de análisis por cada uno de los coeficientes calculados en la fase de evaluación.

En segundo lugar, la escasez de indicadores de evaluación sugiere la necesidad de dedicar en el futuro esfuerzos adicionales para el desarrollo de medidas de evaluación del resultado en minería de datos que permitan la comparación entre aquellos generados por distintas herramientas.

En particular, desde una perspectiva pragmática resulta nece necesario desarrollar indicadores objetivos y operativos referidos a la novedad y aplicabilidad del resultado. Por último, no conviene olvidar que es la prueba del modelo generado en el negocio el paso final en la validación, y el que realmente establecerá la valía. (Hansen y Mouritsen, 1999, p. 132)

REFERENCIAS BIBLIOGRÁFICAS

Benzecri, J. (2015). Gestión del conocimiento y minería de datos. París: Dunod.

Berry, M. y Linoff, G. (2014). Data Mining Techniques for Marketing Sales and Customer Support. USA: John Wile& Sons

Chatfield, C. y Collins, A.J. (1999). Introduction to multivariate analysis. London: Chapman and Hall.

Kamber M. (2006). Data mining: concepts and techniques. Morgan Kaufmann.

Hoaglin D.C., Mosteller F., Tukey J.W. (2005). Exploring Data tables. Trends and Shapes, Wiley, N.Y.

Ester, M., Kriegel, H., y Sander, J (1999). Knowledge discovery in spatial databases. KI-99. Advanc Artif Intellig.

Jambu, M. (2000). Classification Automatique pour l'Analyse des données. París: Dunod.
Johnson Dallas, E. (2013). Métodos multivariados aplicados al análisis de datos. México: Thomson editores.

Johnson, R.A. y Wichern Dean, W. (2010). Applied Multivariare Statistical Analysis. 3rd De. USA: Prentice Hall Inc.

Wichern Dean, W. (2008). Sistemas de información gerencial: Administración de la información digital. México: Pearson.

Lebart, L., Morineau, A. y Tabard, N. (2000). Téchniques de la description statistique. París: Dunod.

Lebart, L., Morineau, A. y Piron, M. (2005). Statistique exploratoire multidimensionelle. París: Dunod.