Artículo **Revista de Análisis Cuantitativo y Estadístico**
Junio, 2016 Vol.3 No.7 1-9

Calibración y selección del modelo de aprendizaje no supervisado K-Medias, de una encuesta sobre factores de riesgo en el consumo de drogas entre estudiantes

MEDINA-VELOZ, Gricelda*†, LUNA-ROSAS, Francisco Javier, TAVAREZ-AVENDAÑO, Juan Felipe y NARVAEZ-MURILLO, René Ulises

Recibido 21 Diciembre, 2015; Aceptado 08 Marzo, 2016

Resumen

En este trabajo se aplica un método de selección y calibración del modelo de aprendizaje no supervisado de minería de datos K-Medias, para generar el diseño de un modelo que arroje los mejores resultados dado un conjunto específico de datos, a través de un proceso denominado validación cruzada ("Cross-Validation"). El objetivo que se persigue con este proyecto, es aplicar un método de calibración y selección del modelo de aprendizaje no supervisado de minería de datos K-Medias y con sus objetivos específicos se busca: 1) Promover el uso del programa estadístico de licencia libre RStudio. 2) Analizar los datos de una encuesta sobre factores de riesgo y protección en el consumo de drogas entre estudiantes a nivel universitario. Y 3) Generar un modelo de minería de datos reutilizable del método de aprendizaje no supervisado K-Medias. La creación del modelo de minería de datos involucró el desarrollo de 4 pasos principales: 1. La creación de la estructura de minería, 2. La selección del algoritmo para el modelo. 3. La elección de los datos a incluir y 4. El procesamiento de los datos. La creación de un modelo de aprendizaje no supervisado de minería de datos del método K-Medias.

Minería de datos, aprendizaje no supervisado, K-medias, factores de riesgo en drogas

Abstract

This paper presents a method of selection and a model calibration of unsupervised learning data mining K-means, to generate the design of a model that brings the best results given a specific set of data, through a process called cross-validation. The general objective of this paper is to apply a calibration method and a model selection on unsupervised learning of data mining method called K-Means. And with the specific objetives: 1) To promote the use of statistical RStudio free license program. 2) Analyze the data of a survey on risk and protective factors in drug use among students at university level. 3) Build a model reusable data mining method on unsupervised learning K-Means. The creation of data mining model involved 4 main steps: 1. The creation of the mining structure, 2. The algorithm selection for the model. 3. The choice of the data to be included and 4. The data processing. Creating an unsupervised data mining learning model of K-means method.

Data mining, unsupervised learning, K-means, drugs usage

Citación: MEDINA-VELOZ, Gricelda, LUNA-ROSAS, Francisco Javier, TAVAREZ-AVENDAÑO, Juan Felipe y NARVAEZ-MURILLO, René Ulises. Calibración y selección del modelo de aprendizaje no supervisado K-Medias, de una encuesta sobre factores de riesgo en el consumo de drogas entre estudiantes. Revista de Análisis Cuantitativo y Estadístico. 2016. 3-7: 1-9

*Correspondencia al Autor (Correo electrónico: gricelda.medina@utna.edu.mx)
† Investigador contribuyendo como primer autor.

Introducción

Existen paquetes de estadística y minería de datos que son una especie de cajas negras donde los parámetros y los algoritmos de los métodos, no son posible manipularlos para probar sus resultados con diferentes valores.

Sin embargo, el programa estadístico de R, permite que los parámetros de algunos métodos de minería de datos, se puedan programar desde la consola y manipular sus valores y algoritmos, para así, asegurarse de obtener los mejores resultados del modelo mediante este tipo de calibración.

La finalidad del proceso de calibración de modelos de minería de datos, es buscar "maximizar la Inercia Inter-Clases, minimizar el error global, o maximizar el área bajo la curva ROC" [rpubs].

Todo depende del objetivo que se está buscando con el análisis de la información en el proyecto de minería. Para efectos de este trabajo y la evaluación del modelo de minería, se utilizó la información obtenida, de la aplicación de una encuesta sobre factores de riesgo y protección en el consumo de drogas entre estudiantes.

El tema fue elegido, debido a su relevancia actual, en los ambientes educativos, y a los resultados encontrados en la Encuesta Nacional de Consumo de Drogas en Estudiantes aplicada en el año 2014 y publicada en el año 2015.

Consumo de drogas entre estudiantes

El consumo del alcohol, de solventes inhalables, de productos de tabaco y de drogas ilegales representa un complejo fenómeno originado por un amplio entramado de factores de riesgo cuyo abordaje requiere de información epidemiológica veraz y actualizada.

Se publicó en el año 2015, la Encuesta Nacional de Consumo de Drogas en Estudiantes 2014, coordinada por la Comisión Nacional Contra las Adicciones, el Centro Nacional para la Prevención y Control de las Adicciones y el Instituto Nacional de Psiquiatría "Ramón de la Fuente Muñiz".

En el que se encontraron resultados alarmantes del consumo de estas sustancias en estudiantes de primaria, secundaria y preparatoria. Sus datos ofrecen un amplio panorama de la situación actual y las variables asociadas, como la oportunidad de exposición, las edades de inicio y las prevalencias de consumo.

La información se obtuvo mediante un cuestionario estandarizado, y las secciones que cubría eran datos sociodemográficos, datos sobre el consumo de drogas, la conducta antisocial, el ámbito social, y el ámbito interpersonal.

Y aunque la problemática no es igual en todos los estados, ni en todos los niveles educativos o edades, sirvió para analizar los resultados obtenidos específicamente en el estado de Aguascalientes. Con los cuales se logró conocer que la prevalencia del consumo de drogas, de al menos una vez en estudiantes de secundaria es del 12.3% y en bachillerato es del 22.7%.

De los alumnos encuestados de secundaria y bachillerato, 8192 de ellos, tienen apoyo o tratamiento debido al consumo de drogas y que el 43.7% inició el consumo de drogas entre los 13 y 14 años de edad [Comisión Nacional Contra las Adicciones], por lo que se optó, por analizar a nivel universitario los factores de riesgo y protección a los cuales se enfrentan los estudiantes aún a este nivel académico.

ISSN 2410-3438

Modelos de Minería de Datos

El propósito general de crear modelos de minería de datos, es para buscar con ello la descripción o la predicción de la información analizada [Pang-Ning], [Hand]. Con los modelos descriptivos se particionan o segmentan un conjunto de datos en grupos o conglomerados, basándose en la similaridad de ciertas variables de la información que se analiza, descubriendo de forma autónoma, correlaciones y categorías similares entre ellos, buscando que, los integrantes de cada grupos o conglomerado que se forma sean lo más acoplados o parecidos entre sí, y que los grupos sean lo más separados o diferentes posible entre ellos.

En cambio los modelos predictivos o de aprendizaje supervisado, pretenden predecir valores futuros, o desconocidos, de las variables, y su objetivo es crear una función capaz de predecir el valor correspondiente a una variable, después de haber analizado una serie de ejemplos [Hand], [Han].

Un modelo de minería de datos, se crea mediante la aplicación de un algoritmo a los datos, que después se podrá aplicar a nuevos proyectos de minería para crear predicciones y deducir relaciones y/o patrones entre ellos.

El modelo de minería, recibe los datos de una estructura de minería, los analiza utilizando el algoritmo más adecuado, y que ofrezca los mejores resultados para ese conjunto de información, para luego almacenar los resultados derivados de su procesamiento estadístico, como son los patrones encontrados y el resultado del análisis [López], [Fayad]. Para crear un modelo de minería de datos, se siguen los pasos generales descritos en la Tabla 1.

Paso	Actividad
1	Se Crea la estructura de minería de datos
2	Se selecciona el algoritmo más adecuado para la tarea analítica.
3	Se eligen las columnas que se incluirán en el modelo y se especifica su uso
4	Se rellena el modelo con datos procesando la estructura y el modelo.

Tabla 1 Pasos para crear un modelo de minería de datos

La metodología utilizada para la creación de este modelo descriptivo de minería de datos, se describe a continuación.

Metodología

1) Creación de la estructura de minería de datos: El primer paso a desarrollar en este trabajo, fue el análisis de los parámetros del método descriptivo de Clustering K-Medias, en el programa estadístico de R a través de la interfaz gráfica del programa de RStudio [Gómez], [Artime], [Chambers], [Venables], [Hothorn], con el objetivo de analizar los valores del modelo en R, y poder medir los grados de libertad o número de parámetros que se manejan en el algoritmo, los cuales se podrán calibrar y así, de esa manera crear la estructura del modelo de minería. Como se puede observar en la Figura 1, uno de los parámetros más importantes del modelo es el algoritmo a usar en el método, donde R propone el uso de 4 posibles algoritmos ("Hartigan-Wong", "Lloyd", "Forgy","MacQueen"), para su calibración.

ISSN 2410-3438

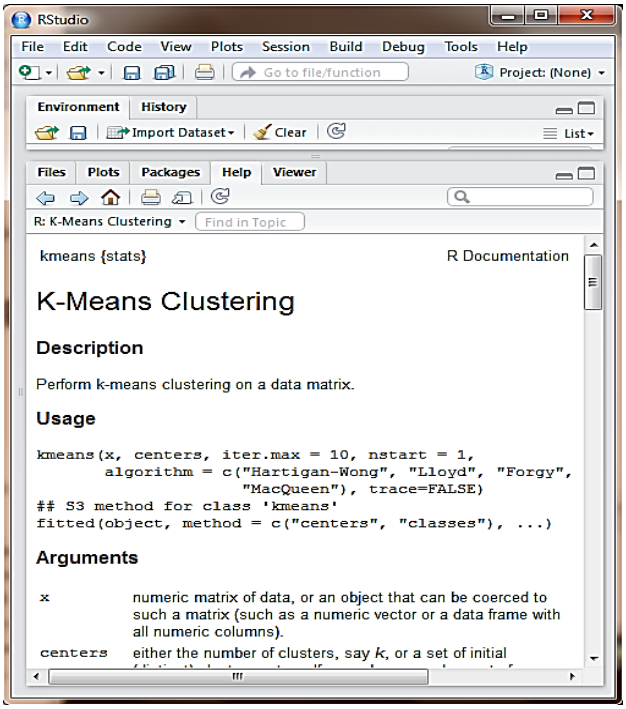


Figura 1 Estructura del modelo de minería de datos k-medias en el programa RStudio

La descripción general de cada uno de estos algoritmos se describe a continuación.

A. Método Lloyd.

Es un algoritmo propuesto por Lloyd en 1957, toma un conjunto de observaciones o casos y los clusteriza en grupos tratando de minimizar la distancia inter clusters mediante la suma de sus cuadrados [Lloyd].

$$\sum_{i=1}^{k} \sum_{x_j \in S_i} \left(x_j - \mu_i\right)^2 \quad (1)$$

B. Método Forgy.

Forgy, en 1965, sugiere un algoritmo simple consistente en la siguiente secuencia de pasos: 1) Comenzar con una configuración inicial. Ir al paso 2 si se comienza con un conjunto de puntos semilla.

Ir al paso 3, si se comienza con una partición de los casos. 2) Colocar cada individuo en el cluster con la semilla más próxima. Las semillas permanecen fijas para cada ciclo completo que recorra el conjunto de datos. 3) Calcular los nuevos puntos semilla como los centroides de los clusters. 4) Alternar los pasos, segundo y tercero hasta que el proceso converja, es decir, continuar hasta que ningún individuo cambie de cluster en el paso segundo [Forgy].

C. Método MacQueen.

MacQueen, en 1972, emplea el término K-Medias para denotar el proceso de asignar cada individuo al cluster (de los K prefijados), con el centroide más próximo. La clave de este procedimiento radica en que el centroide se calcula a partir de los miembros del cluster, tras cada asignación y no al final de cada ciclo, como ocurre en los métodos de Forgy y Jancey.

El algoritmo de MacQueen considera la siguiente secuencia de pasos: 1. Toma los K primeros casos como clusters unitarios. Luego 2. Asigna cada uno de los m − K individuos restantes al cluster con el centroide más próximo, y después de cada asignación, recalcula el centroide del cluster obtenido.

Finalmente, 3. Tras la asignación de todos los individuos en el paso segundo, toma los centroides de los clusters existentes como puntos semillas fijos, y hace una pasada más sobre los datos asignando cada dato al punto semilla más cercano.

El último paso, es el mismo que el del método de Forgy, excepto que la recolocación se efectúa una vez más sin esperar a que se produzca la convergencia [MacQueen].

ISSN 2410-3438

D. Método Hartigan and Wong.

El objetivo del algoritmo K-medias, es dividir M puntos en N dimensiones dentro de K conglomerados, de modo que se minimiza la suma dentro de los cuadrados. Pero esto no es práctico, excepto cuando M y N son pequeños y K = 2. En cambio, con el algoritmo de Hartigan se busca un óptimo local, que solucione que no se muevan de un punto de un grupo a otro, reduciendo la suma inter-clusters de los cuadrados [Hartigan].

$$Sum(k) = \sum_{i=0}^{n} \sum_{j=0}^{p} (x(i,j) - x(k,j))^2 \quad (2)$$

Considerando entonces, que el método k-medias en el programa estadístico de R, puede ser calibrado con cada uno de estos algoritmos. La estructura del modelo de minería se construyó, probando el método con cada uno de ellos, para comparar sus resultados y elegir aquel que arrojó mejores resultados para este conjunto específico de datos.

Factor	Elemento de riesgo o protección
1	Malestar emocional
2	Satisfacción en las relaciones personales
3	Concepto y valoración de las drogas
4	Espiritualidad
5	Permiso social y acceso a las drogas
6	Habilidad social y auto control

Tabla 2 Factores de riesgo y protección en el consumo de drogas

Los datos utilizados para hacer las pruebas del modelo de minería, se obtuvieron de la aplicación de un cuestionario a los estudiantes, y éste se estructuró considerando como base el trabajo propuesto por un grupo de sicólogos colombianos sobre la elaboración de un sondeo de factores de riesgo y de protección para el consumo de drogas en jóvenes universitarios, publicado en el 2006, por la revista de la Universidad Católica de Colombia.

El cuestionario se aplicó a 60 alumnos de la carrera de Tecnologías de la Información y la Comunicación, de la Universidad Tecnológica del Norte de Aguascalientes. Los factores que se analizaron con el cuestionario, se muestran en la Tabla 2.

2) Selección del algoritmo: En la selección del algoritmo más adecuado, y que ofrece los mejores resultados para este conjunto de datos, se utilizaron las técnicas estadísticas de cross validation, técnicas propuestas por los doctores Bradley Efron y Rob Tibshirani de la Universidad de Standford [Efron], [Friedman]. Ellos proponen que, para obtener una adecuada calibración del modelo, primeramente se debe iniciar con la selección del número de clusters (k = número de grupos que se crearán). Este proceso es automático, cuando se utiliza en minería de datos el método de clasificación jerárquica, debido a que, el árbol binario que se genera, sugiere automáticamente cuantos clusters se deben de elegir. Pero para el caso de k-medias, existe un cálculo llamado el codo de Jambú, que es el que ayuda a determinar, el número de clusters que se deben elegir para el método [Jambú], [Tibshirani]. El problema con esto es, que su cálculo requiere de muchas operaciones, ya que implica ir graficando la inercia intra clases, del proceso de calibración del modelo. El nombre de codo, lo recibe en base a la forma de la gráfica que se genera y, el punto donde comienza su estabilización o disminuye la variación, indica donde se encuentra el k ideal, que hace que la inercia se estabilice, lo que implica, qua ya no sea necesario tratar de mejorar porque la inercia intra clases se ha estabilizado.

ISSN 2410-3438

```
setwd("C:/Gris/MD")
datos <- read.csv("Datos.csv",header=TRUE, sep=";", dec=",", row.names=1)
dim(datos)
InerciaIC.Hartigan=rep(0,30)
InerciaIC.Lloyd=rep(0,30)
InerciaIC.Forgy=rep(0,30)
InerciaIC.MacQueen=rep(0,30)
for(k in 1:30) {
  grupos=kmeans(datos,k,iter.max=100,algorithm = "Hartigan-Wong")
  InerciaIC.Hartigan[k]=grupos$tot.withinss
  grupos=kmeans(datos,k,iter.max=100,algorithm = "Lloyd")
  InerciaIC.Lloyd[k]=grupos$tot.withinss
  grupos=kmeans(datos,k,iter.max=100,algorithm = "Forgy")
  InerciaIC.Forgy[k]=grupos$tot.withinss
  grupos=kmeans(datos,k,iter.max=100,algorithm = "MacQueen")
  InerciaIC.MacQueen[k]=grupos$tot.withinss
}

plot(InerciaIC.Hartigan,col="blue",type="b")
points(InerciaIC.Lloyd,col="red",type="b")
points(InerciaIC.Forgy,col="green",type="b")
points(InerciaIC.MacQueen,col="magenta",type="b")
legend("topright",legend = c("Hartigan","Lloyd","Forgy","MacQueen"),
       col = c("blue","red","green","magenta"), lty = 1, lwd = 1)
```

Figura 2 Calibración del modelo k-medias en el programa RStudio [rpubs]

Para generar el codo de Jambú de este proyecto, se corrieron 30 veces el algoritmo k-medias, como una forma de asegurarse que el resultado obtenido sea el más adecuado, recordando que si se corre el proceso de k-medias una vez y luego se corre una vez más, el resultado puede ser considerablemente diferente, ya que el método arranca con datos aleatorios elegidos automáticamente, Gráfico 1.

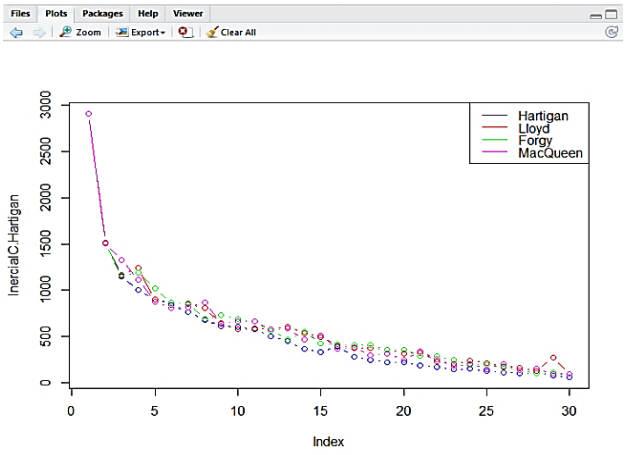


Gráfico 1 Codo de Jambú

En el proceso se creó un ciclo que se ejecutó 30 veces, y en cada corrida del ciclo, el método k-medias se programó con un algoritmo diferente ("Hartigan-Wong", "Lloyd", "Forgy", "MacQueen").

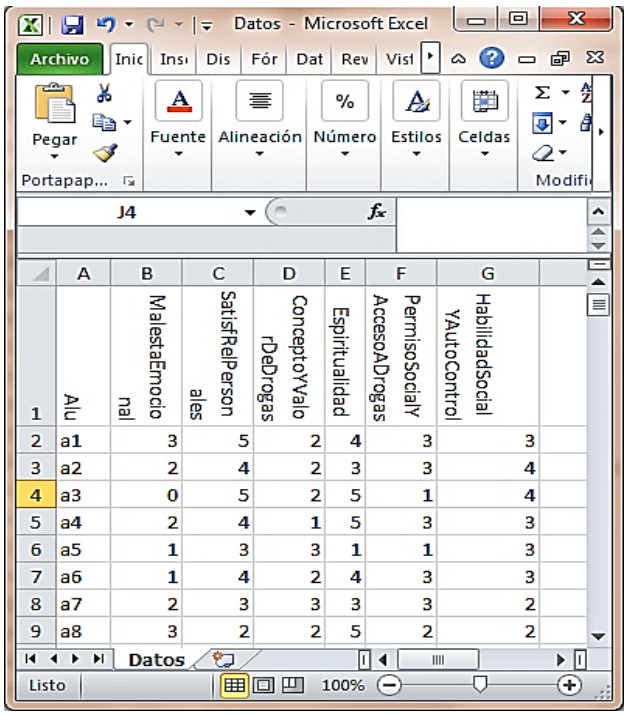
El resultado de la inercia se guardó en su respectivo vector y posición, por medio de la variable grupos$tot.withnss que almacena la inercia intra clases Figura 2.

```
> Hartigan<-0
> Lloyd<-0
> Forgy<-0
> MacQueen<-0
> for(i in 1:50) {
+   grupos<-kmeans(datos,5,iter.max=100,algorithm = "Hartigan-Wong")
+   Hartigan<-Hartigan+grupos$betweenss
+   grupos<-kmeans(datos,5,iter.max=100,algorithm = "Lloyd")
+   Lloyd<-Lloyd+grupos$betweenss
+   grupos<-kmeans(datos,5,iter.max=100,algorithm = "Forgy")
+   Forgy<-Forgy+grupos$betweenss
+   grupos<-kmeans(datos,5,iter.max=100,algorithm = "MacQueen")
+   MacQueen<-MacQueen+grupos$betweenss
+ }
>
> Hartigan/50
[1] 2006.719
> Lloyd/50
[1] 1963.549
> Forgy/50
[1] 1964.973
> MacQueen/50
[1] 1969.767
>
```

Figura 3 Selección del algoritmo en el modelo k-medias del programa RStudio [rpubs]

Después se plotearon los vectores de las inercias de cada algoritmo para producir el gráfico que se muestra en la Gráfico 1, y que representa el codo de Jambú, con el cual se buscó, el punto de estabilización para encontrar el mejor valor de k, y encontrar el número de clusters que se recomienda utilizar en el modelo de minería de datos para este juego específico de datos. Después de seleccionar, mediante la ayuda del codo de Jambú, el número de clusters que generaría el modelo, se hizo una nueva calibración, ahora para elegir cuál de los algoritmos produce los mejores resultados. Para ello se definieron variables que sirvieron como acumuladores de inercias intra clases en cada uno de los algoritmos del modelo. Luego se creó el algoritmo que se corrió 50 veces y en cada corrida se almacenó en el acumulador las inercias intra clases. Finalmente se promedió cada acumulador entre el número de iteracciones que corrió el ciclo. Para luego elegir el algoritmo que dio una inercia inter clases mayor. Como puede observarse en la Figura 3 el algoritmo que arrojó los mejores resultados para este juego específico de datos es el algoritmo de Hartigan y Won.

ISSN 2410-3438



1	Alu	MalestaEmocional	SatisfRelPersonales	ConceptoYValorDeDrogas	Espiritualidad	PermisoSocialYAccesoADrogas	HabilidadSocialYAutoControl
2	a1	3	5	2	4	3	3
3	a2	2	4	2	3	3	4
4	a3	0	5	2	5	1	4
5	a4	2	4	1	5	3	3
6	a5	1	3	3	1	1	3
7	a6	1	4	2	4	3	3
8	a7	2	3	3	3	3	2
9	a8	3	2	2	5	2	2

Figura 4 Estructura del archivo .csv, resultado de la encuesta aplicada sobre factores de riesgo y protección en el consumo de drogas

3) Elección de las columnas a incluir en el modelo. Para la elección del número de columnas a incluir en el modelo, se determinó que, todas las columnas son necesarias para la evaluación y construcción del mismo, ya que el archivo de datos que se usó para probar el modelo, es un archivo .csv, con solo 7 columnas, donde la columna 1 representa los alumnos a los cuales les fue aplicada la encuesta, y las columnas 2-7 muestran cada uno de los factores que se analizaron del tema objeto de estudio, los cuales se mencionaron en la Tabla 2 del documento. La estructura de dicho archivo, se muestra en la Figura 4.

4) Llenado del modelo y procesamiento de la estructura. El archivo del modelo se almacenó con el nombre de Calibracion_K_Medias.R, para su uso posterior en la evaluación de nuevos análisis descriptivos de datos.

Con los que solo será necesario cargar el nuevo archivo .csv con los nuevos datos y el modelo generará automáticamente los resultados que arrojan cada uno de los algoritmos de calibración del método k-medias. Los cuales indicarán, cuál de los cuatro algoritmos se recomienda usar, para la construcción del modelo de minería, que ofrezca los mejores resultados para ese nuevo juego específico de datos.

Resultados obtenidos

En base a los resultados obtenidos en el proceso de calibración del modelo de minería, y como se puede observar en la Figura 3, el algoritmo con el cual se obtuvieron mejores resultados, fue el algoritmo propuesto por Hartigan, ya que este algoritmo reportó una mayor inercia inter clases, lo que significa que los grupos o clusters formados con este algoritmo tienen más semejanzas y similitudes entre sus integrantes, y que los grupos entre ellos muestran mayores diferencias que los grupos generados por los otros algoritmos. El Gráfico 2, muestra la gráfica de araña que generó el modelo minería, y en la que se puede observar, las características que identifican a cada cluster creado.

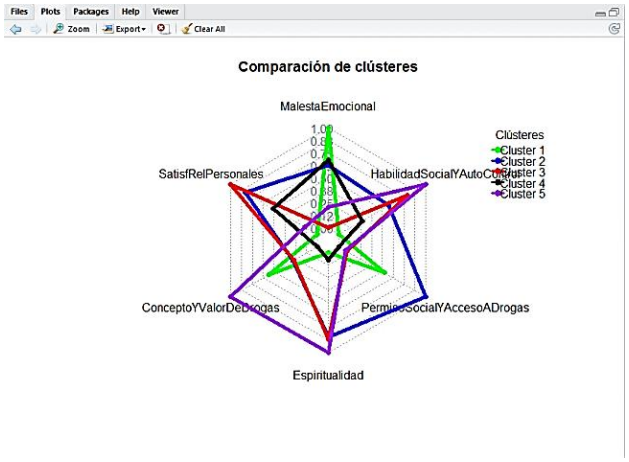


Gráfico 2 Comparativa de clusters en forma de araña, resultado del modelo k-medias

ISSN 2410-3438

En base a esta gráfica se construyó la Tabla 3 de resumen, donde se indican en la columna 3 las características que identifican a cada cluster, en base al concepto analizado en la columna 2. Por ejemplo, el factor de malestar emocional está más presente en los clusters 1, 2 y 4.

Factor	Concepto	Cluster
1	Malestar emocional	1,2,4
2	Satisfacción en las relaciones personales	3,2,4
3	Concepto y valoración de las drogas	5,1
4	Espiritualidad	5,2,3
5	Permiso social y acceso a las drogas	1,2
6	Habilidad social y auto control	5,3,4

Tabla 3 Identificación de clusters en base a los factores de riesgo y protección en el consumo de drogas

Los clusters que piensan que las drogas son permitidas socialmente y tienen un mayor acceso a ellas, son los integrantes de los clusters 1 y 2. Y así sucesivamente se pueden analizar y valorar las características de cada cluster formado por el modelo de minería. Esto nos puede dar una referencia para la elaboración de programas y/o acciones a considerar, los cuales se pueden preparar para ayudar y disminuir el alto grado consumo de drogas entre los estudiantes, a cualquier nivel académico.

Otra de las gráficas que arroja el modelo de minería, es el plot de características de clusters, Gráfico 3, el cual puede ser útil para hacer comparaciones del total de clusters con mayor facilidad. Por ejemplo se puede observar fácilmente, que el cluster 3, es el cluster con menor nivel de malestar emocional contrario a lo que reporta el cluster 1, con el mayor nivel en este aspecto. El cluster 2 es el que reporta más permiso social y acceso a drogas, y el cluster 5 es el que reporta un mayor nivel de espiritualidad, habilidad social y auto control, etc.

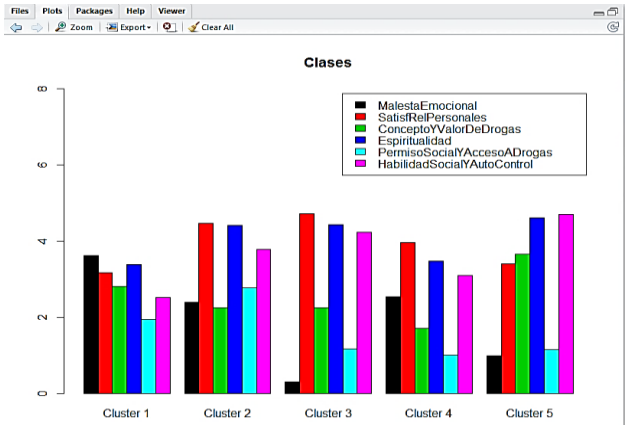


Gráfico 3 Plot comparativo de clusters, resultado del modelo k-medias

Conclusiones

En este trabajo se mostró, cómo con el uso de la técnica de validación cruzada (cross validation) se puede calibrar un método exploratorio de minería de datos, (este caso el método k-medias), para construir y seleccionar el mejor modelo de minería que arroje los mejores resultados, para un juego específico de datos. Además de que, el modelo continúa siendo útil para la evaluación de nuevos datos de minería que utilicen el método de k-medias. Solo faltaría la preparación adecuada de los datos, cargar los datos en el modelo y este automáticamente generará los resultados, calibrados con cada uno de los algoritmos posibles con los que se puede trabajar el método en el programa de RStudio. Con lo cual se asegura que el modelo está construido con los mejores parámetros que el método puede ofrecer y por ende, generar mejores resultados con un mayor índice de confiablidad.

Referencias

[Artime] Artime, C. C., & Blanco, N. C. (2013). Paquetes estadísticos con licencia libre (I)/Free statistical software (I). REMA, 18(2), 12-33.

[Chambers] Chambers, J. (2008). Software for data analysis: programming with R. Springer Science & Business Media.

ISSN 2410-3438

[Comisión Nacional Contra las Adicciones] Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz; Comisión Nacional Contra las Adicciones, Secretaría de Salud. Encuesta Nacional de Consumo de Drogas en Estudiantes 2014: Reporte de Drogas. Villatoro-Velázquez JA, Oliva Robles, N., Fregoso Ito,D., Bustos Gamiño, M., Mujica Salazar, A., Martín del Campo Sánchez, R., Nanni Alvarado,R. y Medina-Mora ME. México DF, México: INPRFM; 2015.

[Efron] Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. The American Statistician, 37(1), 36-48.

[Fayyad] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[Forgy] Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics, 21, 768-769.

[Friedman] Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1). Springer, Berlin: Springer series in statistics.

[Gómez] Gómez, A. A. (2008). Estadística básica con R y R-Commander. Servicio Publicaciones UCA.

[Han] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[Hothorn] Hothorn, T., & Everitt, B. S. (2014). A handbook of statistical analyses using R. CRC press.

[Hand] Hand, D. J., Mannila, H., & Smyth, P. (2001). Principles of data mining. MIT press.

[Hartigan] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100-108.

[Jambu] Jambu, M., Tan, S. H., & Stern, D. N. (1989). Exploration informatique et statistique des données. Dunod.

[Jambu] Jambu, M. (1991). Exploratory and multivariate data analysis. Elsevier.

[López] López, C. P. (2007). Minería de datos: técnicas y herramientas. Editorial Paraninfo.

[Lloyd] Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.

[MacQueen] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[Pang-Ning] Pang-Ning, T., Steinbach, M., & Kumar, V. (2006). Introduction to data mining. In Library of congress (Vol. 74).

[rpubs] http://rpubs.com/orodriguez/13318

[Tibshirani] Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63(2), 411-423.

[Venables] Venables, W. N., & Smith, D. M. (1990). R: A Programming Environment for Data Analysis and Graphics. Technical Report, Department of Statistics, University of Adelaide (Dec.).

ISSN 2410-3438