Informe aprendizaje no supervisado

Diego Hernández Jiménez

1. Introducción

Hasta ahora, todos los análisis que se han realizado estaban orientados a resolver un problema de aprendizaje supervisado, a saber, clasificar los individuos que sufren, o no, ictus, sabiendo a qué categoría pertenecen para poder valorar la precisión en la clasificación. Sin embargo, también es posible plantear otro de tipo de problemas. Sin tener en consideración la clase a que pertenece cada sujeto, ¿podemos formar a partir de las variables predictoras dos grupos que se correspondan con los "perfiles" de individuos que sufren ictus e individuos que no sufren ictus?

El objetivo de este trabajo, por tanto, es identificar la existencia de agrupaciones o *clusters* que se forman de manera natural en los datos. Se presupone la existencia de dos grupos diferenciados, uno que corresponde a las personas que han sufrido ictus y otro que identifica a personas que no han sufrido ictus. Una elevada coincidencia entre estas agrupaciones y las categorías reales permitirá utilizar con cierta confianza los *clusters* como "perfiles de riesgo".

A priori, teniendo en cuenta estudios como el de Boehme et al (2017), los individuos deberían diferenciarse más en la edad, el padecimiento de hipertensión, diabetes, afecciones cardíacas y sobrepeso y obesidad (cuya influencia probablemente sea indirecta, al afectar al resto de factores). Son, por ello, las variables que se incluirán en el análisis. Más específicamente se incluyen *age*, *hypertension*, *avg_glucose_level* (nivel de glucosa en sangre, no se trata de la variable discretizada para categorizar sujetos diabéticos y no diabéticos), *heart_disease* y *bmi* (valor de índice de masa corporal, no se trata de la variable discretizada) de la base de datos que se ha estado utilizando en trabajos anteriores. Para el análisis se propone el empleo del algoritmo K-medias. Puede resultar apropiado por varios motivos. Por un lado, se conocde de manera previa el número de *clusters* que se pretende encontrar. Por otro lado, a pesar de que la técnica utilice medidas de similitud que requieren el uso de variables cuantitativas, no resulta muy difícil incorporar variables dicotómicas (Johnson y Wichern, 2014). Finalmente, frente a técnicas como el análisis de clúster jerárquico, resulta computacionalmente más eficiente, sobre todo con muestras grandes como ocurre en este caso.

2. Métodos y herramientas

El procedimiento llevado a cabo puede describirse en 3 fases. Todo el análisis se realiza con Rapidminer, salvo que se mencione lo contrario.

1. La primera parte consiste en el preprocesamiento de los datos para poder cumplir los requisitos del algoritmo de K-medias. Ello implica la normalización de las variables cuantitativas (normalización de rango) para evitar sesgos como función de la métrica de las variables (unidades mayores conllevan distancias mayores). También se eliminan los valores perdidos, que únicamente se producen en la variable *bmi*. La muestra queda reducida así en un 3% aproximadamente. Por último, para poder calcular la similitud mediante la distancia euclídea en las variables categóricas, se convierten en variables indicador o *dummy*. Con esta codificación, en Rapidminer puede utilizarse la distancia euclídea, que pasa a ser un recuento del número de no coincidencias (*mismatches*) (Johnson y Wichern, 2014).

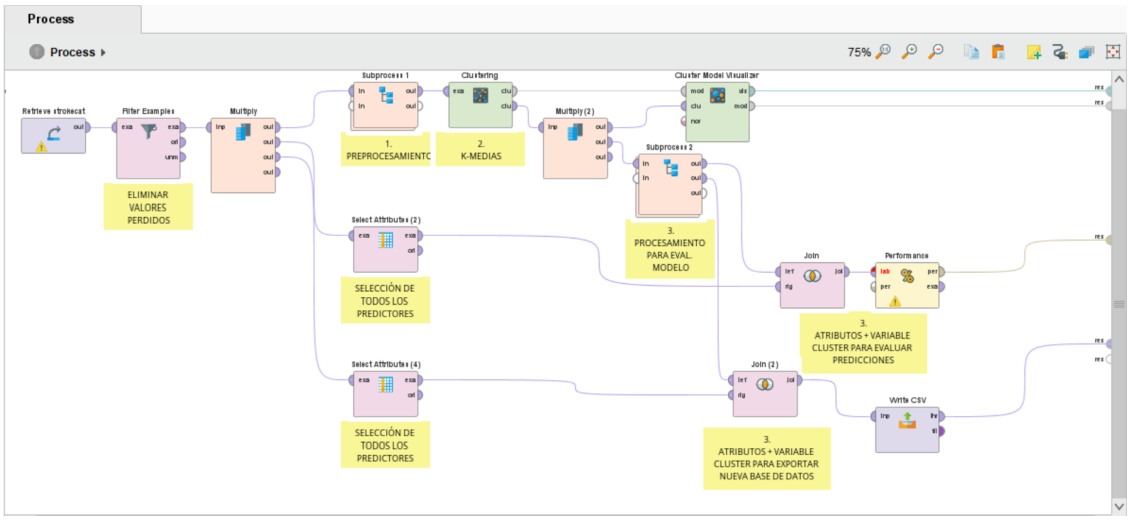

Figure 1: Visión general del procedimiento

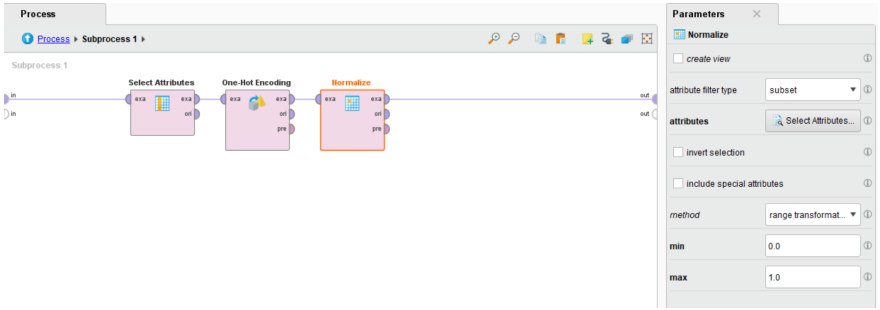

Figure 2: Subproceso 1: preprocesamiento

2. En la segunda fase se ajusta el modelo de K-medias, siendo $k = 2$ y usando la distancia euclídea como medida de similitud. Es bien sabido que el resultado del algoritmo K-medias (localización de los centroides al finalizar el proceso) es sensible a la localización inicial de los centroides (James et al, 2013). Se tiene en cuenta, pero no requiere ninguna modificación, pues Rapidminer implementa de manera oculta el algoritmo con distintas coordenadas de partida y selecciona finalmente el modelo que minimiza las distancias intraclúster y maximiza las distancias interclúster, es decir, el modelo más apropiado.

3. La tercera fase concierne la evaluación de los resultados. La tabla de centroides proporcionada por Rapidminer contiene la información necesaria para comprobar si los *clusters* formados son consistentes con lo esperado. Sin embargo, dado que se ha realizado una normalización de las variables, para facilitar el análisis y la comparación de los *clusters*, se utiliza la base de datos original, pero con la variable *cluster* añadida, que identifica a cada caso como miembro de uno de los conglomerados. Además, dado que se dispone de las etiquetas que identifican a los sujetos en los grupos de interés (*stroke=Yes* y *stroke=No*), es posible comparar directamente la asignación que se realiza con el procedimiento de K-medias y las clases reales y evaluar así la capacidad predictiva como en el contexto de aprendizaje supervisado.

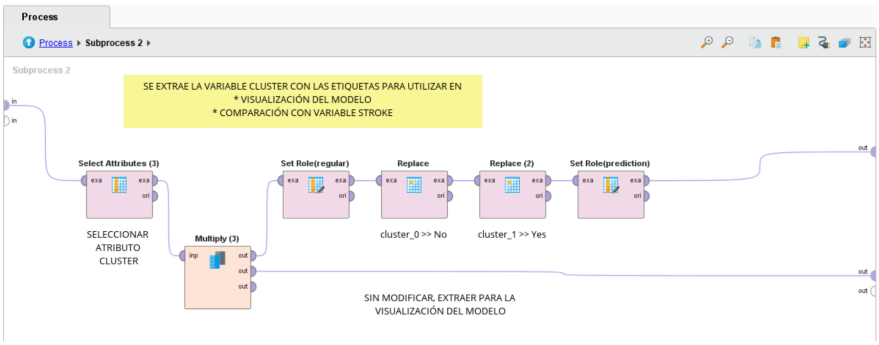

Figure 3: Subproceso 2: extracción variable cluster

3. Resultados y discusión

Los resultados generales se resumen en el siguiente gráfico.

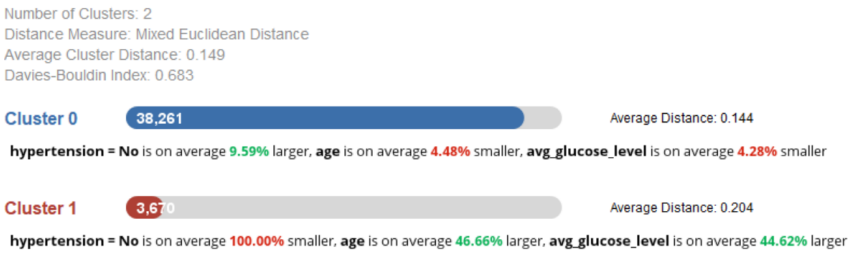

Figure 4: Características de los clusters finales

Mediante inspección visual podemos comprobar que se asemeja a lo que cabría esperar teóricamente. La gran mayoría de sujetos pertenecen al *cluster* 0, que podemos denominar de bajo riesgo, y un porcentaje menor se categoriza como de alto riesgo. Puede observarse en las características de este *cluster* 1 que tanto la edad como el nivel de glucosa es más de un 40% mayor, lo cual concuerda con lo que se conoce. Además, entre estos sujetos el padecimiento de hipertensión es claramente superior, aunque al tratarse de una variable categórica, la interpretación es más útil en términos de proporciones. Por último, cabe destacar que del promedio de distancias puede inferirse que los sujetos del *cluster* 0 son más similares entre sí de lo que lo son los individuos del *cluster* 1. Si se utilizan los datos originales (métrica original, por tanto) y se obtienen los promedios por *cluster* obtenemos la siguiente tabla, que va en línea de lo comentado:

Table 1: Promedios agrupados por cluster

Cluster	*age*	*avg_glucose_level*	*bmi*	*hypertension=Yes*	*heart_disease=Yes*
0	39.97	101.55	28.22	0	0.03
1	61.33	125.33	32.65	0.22	0.12

Además, los contrastes de medias para las variables cuantitativas y de proporciones para las variables dicotómicas (realizados en R y JASP) revela que las diferencias entre grupos son estadísticamente significativas y con un tamaño del efecto sustantivo (el estadístico d de Cohen para *age*, *avg_glucose_level*, *bmi* es -1.153,-0.478 y -0.573 respectivamente).

En general puede decirse que los grupos formados se asemejan a los grupos naturales de *stroke=No* y *stroke=Yes*. Pero no necesariamente tiene por qué haber una buena correspondencia. Examinando las proporciones casos *Yes* por *cluster* se observa que existen pocos en cada grupo. En el *cluster* 1, que cabría esperar que la mayoría fuesen individuos que han sufrido ictus, solo el 4.69% aproximadamente lo ha sufrido. Por otro lado, un 1,23% del *cluster* 0 son casos *Yes*. Esta discrepancia se refleja en el rendimiento del modelo como clasificador. Se pronostican más casos positivos de los que realmente hay, siendo la sensibilidad, por ello, de un 26.75% por ciento. La medida de F_1 también es reducida, de un 7.98%.

Table 2: Matriz de confusión

	Predicción=No (cluster_0)	Predicción=Yes (cluster_1)
Verdadero=No	37790	3498
Verdadero=Yes	471	172

4. Conclusiones

Del análisis de K-medias puede concluirse que se han encontrado dos grupos muy parecidos a los grupos de ictus y no ictus. No obstante, la correspondencia está lejos de ser perfecta, y en el grupo que en principio corresponde a individuos que sufren ictus hay muchos más casos de los que se dan en la muestra. A pesar de ello, los clusters encontrados pueden tener utilidad para identficar patrones de riesgo. No cabe duda de que, aunque no suponga necesariamente que se vaya a sufrir ictus, la avanzada edad, nivel alto de azúcar, hipertensión... suponen mayor propensión a sufrirlo.

5. Referencias bibliográficas

Boehme, A. K., Esenwa, C., y Elkind, M. S. (2017). Stroke risk factors, genetics, and prevention. *Circulation research*, 120(3), 472-495.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). Unsupervised learning En G. James, D. Witten, T. Hastie y R. Tibshirani, *An introduction to statistical learning* (pp. 373-401). Springer.

Johnson, R. A., y Wichern, D. W. (2014). Chapter 12. Clustering, Distance Methods and Ordination. En R. A. Johnson y D. W. Wichern, *Applied multivariate statistical analysis* (pp.696-703). Pearson Education Limited.