Estadística y Minería de Datos: Similitudes y Diferencias

Javier Alagón **Estadística Aplicada**

Cada vez es más frecuente encontrarse en el mundo de la Investigación de Mercados, tanto en la literatura relacionada como en discusiones entre investigadores de mercado, con el término *"Minería de Datos"* (Data Mining, en Inglés). La asociación del término con el mundo de la Estadística, es por lo general, inmediata. Sin embargo, tal y como revisaremos en este artículo, los dos términos, Minería de Datos y Estadística, se refieren a **dos disciplinas** que tienen mucho en común, pero que también presentan fuertes contrastes y diferencias, que hacen precisamente que tengamos dos áreas del conocimiento distintas, y no a una (la Minería de Datos), como un simple subconjunto de la otra (la Estadística).

El término **Minería de Datos** conlleva una fuerte promesa latente, **la promesa de exploración y el posible encuentro de relaciones subyacentes en los datos,** que arrojen luz sobre el fenómeno estudiado, que comprueben empíricamente hipótesis o que se planteen otras nuevas. Dicha promesa puede resultar sumamente seductora para el investigador o analista, que quizás se haya sentido en ocasiones anteriores, abrumado por el rigor y formalismo matemático de la Estadística convencional.

La esencia de la Minería de Datos se encuentra en la **posibilidad del descubrimiento de información insospechada, pero sumamente valiosa.** Esto significa que la naturaleza de la Minería de Datos es exploratoria, en contraste con la naturaleza confirmatoria de muchas áreas de la Estadística ("Confirmatoria" en el sentido de confirmar hipótesis).

Ahora bien, el Análisis Exploratorio de Datos **NO** es nuevo en Estadística; sin embargo, las técnicas convencionales de dicho Análisis Exploratorio, no son suficientes para conjuntos de datos **ENORMES**, como los que maneja la Minería de Datos hoy en día, con la potencia y rapidez de las computadoras actuales.

Lo anterior resulta lógico, si tomamos en cuenta que el cuerpo central de la Estadística se desarrolló en épocas en donde no había computadoras. De esta manera, un juego de datos de 1,000 puntos sería considerado como "grande" para la Estadística convencional, pero **NO** se puede comparar con las más de 50 millones de transacciones con tarjeta de crédito que realizamos los mexicanos anualmente y que conforman bases de datos sumamente interesantes para el mundo financiero.

El ejemplo anterior ilustra una diferencia fundamental entre la Estadística y la Minería de Datos: **con Estadística se pueden hacer manipulaciones de los datos de manera directa; no así con la Minería de Datos.** Esta fuerte necesidad computacional, se traduce en un gran énfasis en **algoritmos numéricos** para la Minería de Datos.

Por supuesto que las computadoras son herramienta fundamental para la Estadística. Existen una gran variedad de técnicas estadísticas convencionales que no funcionarían sin computadoras. Sin embargo, la gran diferencia estriba en el **TAMAÑO** de las bases de datos manejadas por ambas disciplinas.

Los expertos en Minería de Datos pueden encontrar relaciones insospechadas en sus datos, pero también puede ocurrir, que varias de dichas relaciones no son tan congruentes con la realidad o se deben a desviaciones extrañas en los datos. Por ello, se debe ejercer doble dosis de precaución en interpretaciones hechas con Minería de Datos. Y por supuesto, que también puede ocurrir, que muchas de las relaciones y estructuras subyacentes en los datos, y descubiertas con la Minería de Datos, resultan sumamente obvias y no aportan conocimiento alguno sobre el fenómeno bajo estudio.

Finalmente, otra diferencia fundamental entre las dos disciplinas, es que a la Minería de Datos no le concierne **la selección de la información.** A diferencia de la Estadística, en donde una cuestión fundamental es **cómo seleccionar** la muestra de la mejor manera, de tal forma que sea representativa de la población, la Minería de Datos supone esencialmente que los datos **YA han sido recolectados y se aboca al descubrimiento de sus secretos.**

En el futuro, ambas disciplinas se tocarán irremediablemente y como nos gusta un final feliz para esta pequeña nota, necesariamente serán complementarias en el entendimiento de fenómenos dentro del mundo de la investigación de mercados.