Fase de Minería de Datos

Ubicación en el proceso KDD

La minería de datos es un paso dentro del proceso de descubrimiento de conocimiento en bases de datos (KDD por sus siglas en inglés). En la figura 1 se puede ver este proceso y la ubicación de la minería de datos.

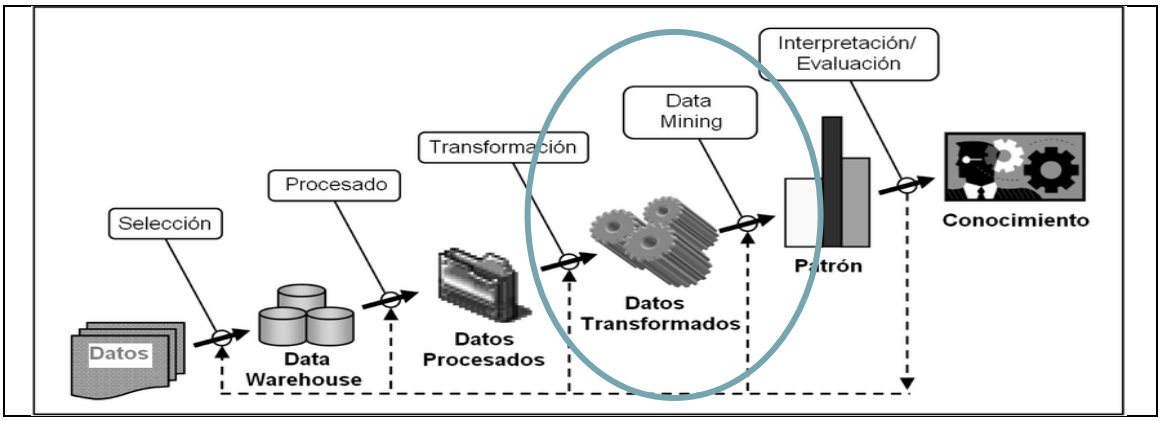


Figura 1. Proceso KDD.

Entonces, para poder aplicar técnicas de minería de datos, necesitamos el insumo principal: los datos. Hemos visto distintas fuentes de datos que podemos acceder, pero no solo es el acceso a los datos crudos lo que nos va a permitir aplicar algún tipo de técnica. Primero, hay que adaptar los datos originales a nuestras necesidades: selección, limpieza y transformación.

Una vez que tenemos los datos en la forma que necesitamos, debemos determinar qué tipo de tarea de minería de datos debemos realizar. Luego definir con qué modelo la vamos a llevar a cabo y por último elegir el mejor algoritmo para ello.

Tareas de la minería de datos

Una tarea de minería de datos es un *problema que se debe resolver* y ésta puede ser de dos tipos: predictiva o descriptiva (figura 2).

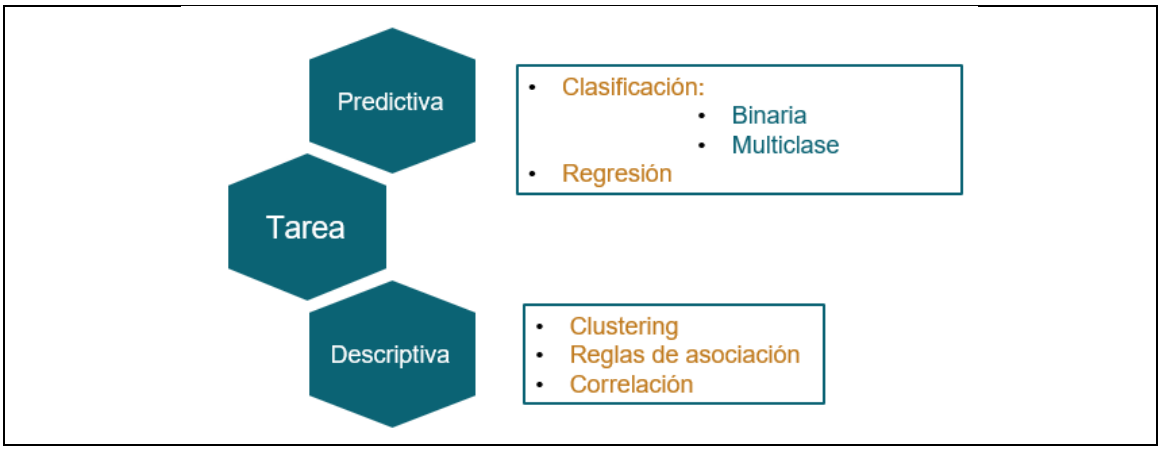


Figura 2. Tareas de la minería de datos.

Cada tipo de tarea tiene sus técnicas asociadas y los algoritmos correspondientes para aplicarlas. Los algoritmos que se utilizan se dividen a su vez en dos tipos, de acuerdo con el tipo de "aprendizaje" del mismo: de **aprendizaje supervisado** y de **aprendizaje no supervisado**.

Esta clasificación depende de si los valores de la variable target están etiquetados o no, es decir que para cada elemento se conoce el valor del atributo objetivo para el conjunto de datos del que se dispone. Esto le permitirá al algoritmo poder "aprender" y ser capaz de predecir la variable objetivo para un conjunto de datos nuevo. Los dos grupos de algoritmos de aprendizaje supervisado son: los algoritmos de regresión (variable objetivo numérica) y los algoritmos de clasificación (variable objetivo categórica). Ejemplos: modelos de regresión lineal y logística, los árboles de decisión, las redes neuronales y K-vecinos más cercanos entre otros.

Por otra parte, los algoritmos de aprendizaje no supervisado basan su proceso de entrenamiento en un conjunto de datos sin etiquetas o clases previamente definidas: a priori no se conoce ningún valor objetivo o de clase, ya sea categórico o numérico. El aprendizaje no supervisado está dedicado a las tareas de agrupamiento, también llamadas clustering o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos de acuerdo con las características de éstos. Ejemplos: k-medias y apriori.

Evaluación e interpretación

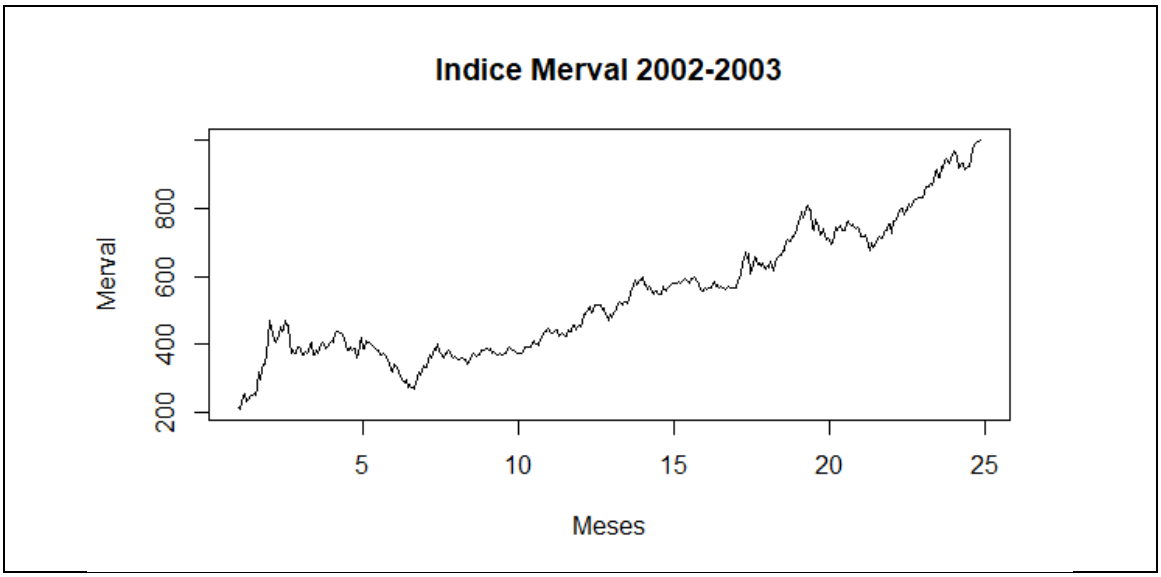
El modelo que se seleccionó para la tarea que resuelve nuestro problema, debe ser entrenado y probado. Para esto se utilizan los métodos de validación conocidos (validación simple, k-fold cross validation, etc.), que permitirán obtener resultados más precisos y disminuir el sobreajuste o sobreentrenamiento (overfitting) del modelo.

Para evaluar nuestro modelo, se utilizan distintas métricas. Por ejemplo, en el caso de una regresión lineal se puede utilizar el error cuadrático medio; o si se trata de una clasificación se puede utilizar la precisión o el AUC (área bajo la curva ROC), dependiendo si es una clasificación binaria o multiclase.

Independientemente de las métricas mencionadas anteriormente, en muchos casos hay que evaluar también el contexto donde el modelo se va a utilizar. Aquí se pueden utilizar la matriz de confusión o la curva ROC para observar el resultado de nuestro modelo.

Actividades

1 – El índice MERVAL es una de las herramientas que permite analizar el desarrollo de la economía nacional. Se dispone de un archivo de valores separados por coma con los datos de su evolución diaria entre los años 2002 y 2003 (merval.csv).



Explore los datos disponibles y explique (sin implementar código) que pasos seguiría para conseguir predecir el valor del índice para el día siguiente, a partir de los valores del índice de los últimos 5 días.

Indicar el tratamiento propuesto para los datos, que tipo de tarea utilizaría, el modelo a aplicar y con qué funciones de R lo implementaría.

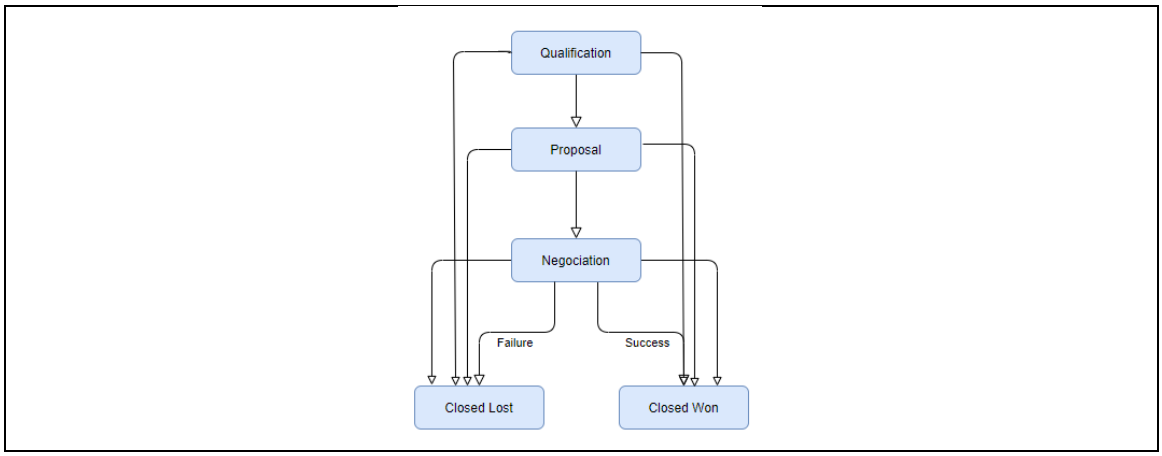
Por último, indique que validación utilizaría y que métrica utilizaría para evaluar su modelo.

2 – Una empresa dedicada a la venta e instalación de equipos de aire acondicionado para grandes superficies quiere predecir la probabilidad de éxito de sus oportunidades comerciales[1].

Una “oportunidad” consiste en un proyecto de venta o instalación de equipos para un cliente. La venta se estructura alrededor de TRF (Toneladas de refrigeración) y puede estar compuesta por varios productos distintos. El "pipeline" hace referencia al flujo de oportunidades prospecto que la empresa está desarrollando.

El equipo comercial asigna a distintos momentos, para cada oportunidad, un estado en la negociación. En la Ilustración se muestran los estados que las oportunidades tienen dentro del pipeline.

[1] Modificado de “Predicción de éxitos en oportunidades comerciales”



Pipeline [2].

Pregunta del negocio: ¿Cuál es la probabilidad de que la oportunidad se convierta en un caso Closed Won?

Los datos disponibles cuentan con información histórica de oportunidades de los últimos cuatro meses, como por ejemplo información sobre el vendedor a cargo de la venta, información geográfica de los clientes, TRS pedidas, fecha prevista de entrega de los equipos, etc.

Se pretende utilizar los resultados para mejorar el rendimiento y optimizar el esfuerzo de los vendedores.

Utilizando solamente los datos provistos en el archivo Entrenamiento_ECI_2020.csv indique de qué tipo son las variables disponibles, cuáles serían útiles y cuáles no. ¿Cuál es la variable objetivo?

Identifique la tarea de minería de datos necesaria para resolver el problema y comente con qué modelos se podría resolver, aclarando qué variables utilizaría en cada caso.

¿Qué tipo de validación utilizaría para el/los modelos elegidos?

Referencias

1. Orallo, J et all. “Introducción a la Minería de Datos” (2004).
2. Pipeline. Disponible en: https://rawcdn.githack.com/lfernandezpiana/hello-world/18389790db75e9913af3a8564391eeadcf8c4c13/ilustracion1_alix_partners.png