UNIDAD TEMÁTICA 4: Algoritmos No Lineales

Trabajo de Aplicación 2 Arboles de Decisión en RapidMiner

EJERCICIO 1

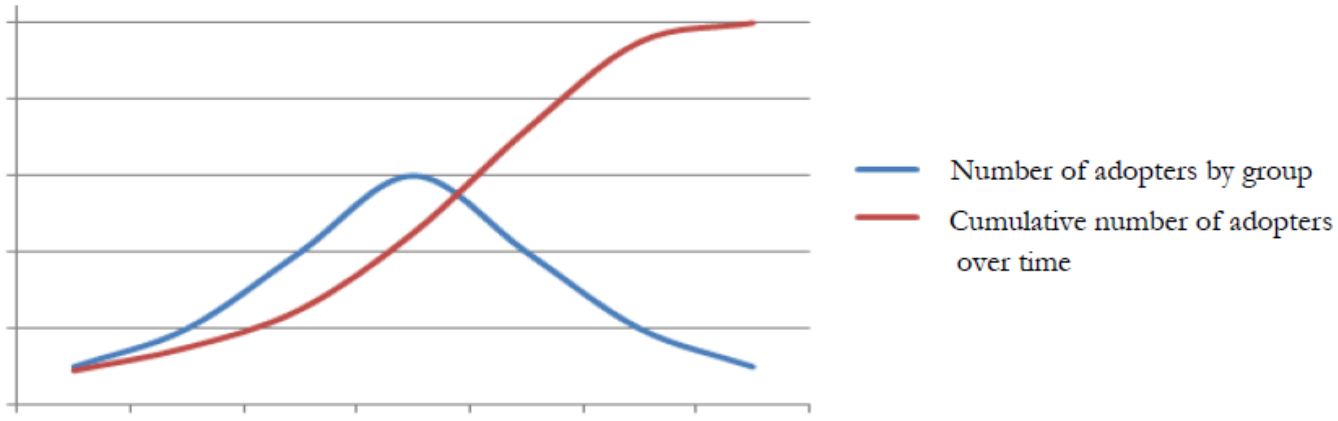
ESCENARIO

Martín trabaja en una gran compañía que vende dispositivos electrónicos en línea. La empresa está a punto de lanzar un nuevo eReader y desea maximizar la efectividad de su marketing para el producto. Tienen muchos clientes, algunos de los cuales ya han comprado productos similares, de versiones anteriores. Martín se pregunta qué es lo que hace que alguna gente desee comprar algo tan pronto como se lanza al mercado, mientras que otras personas no se apuran tanto.

Para impulsar la venta del eReader, la compañía también ofrece una serie de otros productos y servicios específicos para el eReader, a través de su sitio web (revistas digitales, periódicos, libros, música, etc.). También vende muchos otros tipos de productos como libros impresos y otros dispositivos electrónicos.

Martín cree que realizar minería de sobre los datos de los clientes referidos a sus comportamientos en el sitio web, podría ayudarle a entender qué clientes comprarían el nuevo eReader apenas sea lanzado, quienes en un tiempo corto y quiénes lo harían más tarde. Con esta información espera poder ajustar su estrategia de marketing.

También ha estudiado las teorías vigentes sobre adopción de nuevas tecnologías (ej.: artículo sobre teoría de difusión del investigador y sociólogo Everett Rogers). Rogers indica en su publicación que la adopción de una nueva tecnología o innovación sigue una curva con forma de "S": un pequeño grupo de clientes, los más innovadores y emprendedores, adoptan la tecnología al principio, seguidos de grandes grupos que la adopta en la etapa media, y finalmente seguidos por los que la adoptan en forma tardía.



Martín cree que, basado en esta teoría, podría categorizar a los clientes de la compañía en alguno de los cuatro grupos que eventualmente podrían comprar el nuevo eReader: innovadores, adoptantes tempranos, mayoría temprana y mayoría tardía.

Datos disponibles – comprensión

La compañía cuenta con un valioso conjunto de datos de información de cada cliente, que incluye las cosas que ha observado, y las que ha finalmente comprador.
Martín ha preparado dos datasets para este trabajo. El dataset de entrenamiento contiene las actividades en el sitio web de los clientes que compraron la versión anterior del eReader, y el momento (con respecto al lanzamiento) en

que lo hicieron. El segundo dataset está compuesto por los atributos de los clientes actuales que Martín espera que compren el nuevo producto. Espera prever en qué categoría estaría cada cliente contenido en el dataset de evaluación, basándose en los perfiles y tiempos de compra que aparecen en el dataset de entrenamiento.
Al analizar sus datasets Martín encontró que las actividades de los clientes en las áreas de medios y libros digitales, así como su actividad general referente a los dispositivos electrónicos a la venta en el sitio web, parecen tener mucho en común con el momento en que una persona compra un eReader. Teniendo en cuenta esto, hemos compilado la información en los datasets, con los siguientes atributos:

- **ID**: identificador único del cliente, numérico
- **Edad**: la edad en años redondeada al entero más cercano.
- **EstadoCivil**: "C" para los casados, "S" para todas las otras alternativas.
  Sexo: F = femenino; M= masculino.
  ActividadWebsite: refleja el nivel de actividad en el sitio web: Escasa, Regular o Frecuente
- **MiroElectronicos12**: indica si la persona hay mirado o no productos electrónicos en el sitio de la compañía (SI / NO) en el último año
- **ComproElectronicos12**: indica si la persona ha comprado o no productos electrónicos en el sitio de la compañía en el último año (SI / NO)
- **ComproMedios18**: indica si la persona ha comprado o no productos digitales (ej: MP3) en el sitio de la compañía en el último año y medio (SI / NO). Este atributo NO incluye libros digitales
- **ComproLibrosDigitales**: Martín cree que este atributo puede ser un muy buen indicador del comportamiento de compra para el nuevo eReader, y por ello se lo ha separado de los demás atributos que refieren a compras. En este caso se indica si el cliente *alguna vez* compró libros digitales, no se restringe sólo al último año.
- **MetodoPago**: la forma más frecuente en que el cliente ha efectuado sus pagos:
  - Transferencia bancaria
  - CuentaWebsite – el cliente ha dispuesto una tarjeta de crédito o cuenta bancaria para débito automático en el sitio
  - TarjetaCredito – el cliente ingresa los datos de la tarjeta y autorización en cada compra
  - DebitoMensual – el cliente realiza compras regularmente y recibe una factura que puede abonar mensualmente
- **AdopcionEReader**: este atributo existe sólo en el dataset de entrenamiento. Tiene los datos de los clientes que han comprado eReaders de generaciones anteriores. Los que compraron dentro de una semana del lanzamiento son registrados como "Innovadores". Los que compraron entre una y tres semanas luego del lanzamiento, se registran como "AdoptanteTemprano". Luego de tres semanas, pero dentro de los primeros 2 meses, se consideran "MayoriaTemprana" y los demás, "MayoríaTardía". Este atributo servirá como etiqueta al aplicar el modelo al datset de evaluación.