[Pregunta 1] (25 puntos). Definir la información y el método de presentación a utilizar.

• Enumera y describe los bloques de información o entidades que se pueden deducir del caso presentado y las ya nombradas. (5 puntos)

Antes de definir las entidades, creo interesante analizar un poco el modelo ya que dichas entidades se van a basar en las tablas del modelo y en sus atributos. El archivo JPG nos deja entrever la existencia de las siguientes tablas: VENTAS, TABLA_UNION, GEO_TIENDA, Encuesta, Geolocaliación y TIPO_ENCUESTA, cada una con sus consiguientes atributos. Voy a describir brevemente estas tablas o bloques de información (no entraré en todos los atributos para no hacerlo excesivamente largo):

- VENTAS: Se describen las ventas de cada tienda en una fecha concreta, indicando su objetivo y el año anterior (entiendo que este atributo se refiere a las ventas del año anterior)
- TABLA_UNION: Unido a la tabla Ventas a través. Esta tabla, como indicaré más adelante, es una tabla de paso que sirve para unir diferentes tablas, en este caso a través de la clave Fecha-Tienda (formada a partir de esos dos atributos). Ofrece información muy redundante, ya que con sólo ése campo ya podríamos deducir todo lo demás a través de un query (el día, el mes, el número de mes, el año y la variable Año-Mes), lo que no saturaría tanto nuestro Data Warehouse. No obstante, entiendo que si están ahí será por un propósito.
- GEO_TIENDA: Relacionado con TABLA_UNION a través de la clave “Tiendas”. Ofrece información como el CP, ciudad, país o la propia tienda de referencia.
- GEOLOCALIZACION: Relacionado con GEO_TIENDA a través de la clave principal “Código Geo”. Ofrece más características de localización, tales como las coordenadas, latitud, longitud, código del país o Estado/Comunidad. Esta tabla ofrece también información redundante, ya que Código_pais puede deducirse fácilmente con los dos primeros dígitos de Cógido Geo. Al margen de esto, no veo la utilidad de saber la latitud y la longitud de la tienda teniendo sus coordenadas GPS…. El atributo Precisión lo considero irrelevante debido a la gran cantidad de información faltante.
- Encuesta: Relacionado a través de la variable “Fecha-Tienda” con TABLAUNION, los atributos que se aquí se representan indican el resultado de las encuestas a través de la variable Puntuación, siendo un 0 como negativo y un 1 como puntuación positiva.
- TIPO_ENCUESTA: Relacionado con la tabla anterior a partir de la variable “Medición”, nos indica los diferentes ítems que valoran los clientes, jerarquizados por Zona, Servicio y Medición.

Este es el escenario que describe el modelo inicial del que vamos a partir. Para empezar a describir mi modelo de entidades, me voy a basar en la siguiente definición de Entidad dentro del modelo Entidad-Relación:

(Una entidad) representa una “cosa”, "objeto" o "concepto" del mundo real con existencia independiente, es decir, se diferencia únicamente de otro objeto o cosa, incluso siendo del mismo tipo, o una misma entidad (Wikipedia)

Bajo este enfoque, determinamos que cada entidad tiene sus atributos o características propios y que son éstos los que le dan un carácter diferenciador. ¿Podemos, por tanto, asociar de forma inequívoca las tablas definidas con las entidades? En mi opinión, no. Voy a explicar por qué. En primer lugar, existen lo que en el diseño de bases de datos relacionales llamamos *tablas de paso*, de las cuales -al menos, siendo estrictos con la definición de Wikipedia anterior- podría discutirse

si realmente se ajustan del todo la definición de entidad. En segundo lugar, hay entidades cuyos atributos son redundantes o, como podemos apreciar viendo la información en Excel, irrelevantes debido a la gran cantidad de atributos vacíos (debido a la dificultad para medir los datos, para exportarlos, para introducirlos, etc.). Por último, creo que las tablas de MODELO-TABLAS.PNG no representan de forma integral el potencial de nuestro modelo respecto a las entidades, ya que hay entidades que creo que aportarían información útil y robustez al modelo.

Empezamos por lo que yo eliminaría o modificaría teniendo en cuenta únicamente los objetivos que se nos piden al principio de la PEC (es decir, obviando otros posibles objetivos, a fin de facilitar su comprensión). Después de analizar la información contenida en las tablas GEO_TIENDA y GEOLOCALIZACIÓN, no veo la utilidad (repito: para los objetivos planteados en esta PEC) de tener dos tablas diferenciadas, ya que ambas ofrecen información de ubicación física. Por tanto, pienso que los atributos de ambas tablas podrían perfectamente fusionarse en una única tabla.

Respecto a cómo mejorar el modelo, creo que nos falta una entidad que creo que sería vital para nuestros objetivos: la entidad CLIENTES. Habría muchas más posibilidades de añadir entidades (Proveedores, Empleados, etc.) pero creo que CLIENTES sería importante. Recordemos que el enunciado nos indica que son "encuestas que realizan a los clientes sobre ciertas preguntas que son iguales para todos los centros". En mi opinión, es un error de diseño que una entidad tan importante no quede reflejada de ninguna manera. Estamos hablando de una compañía de retail, donde conocer algunos datos básicos sobre los clientes que cubren las encuestas nos permitiría multitud de posibilidades, ya que conocer el perfil de nuestros clientes nos posibilitaría-por ejemplo, a través de un proceso de clustering- emprender acciones *ad hoc* hacía determinados grupos de interés. Soy consciente de que el cliente quiere una encuesta rápida y fácil de contestar, por lo que podremos tener pocos campos adicionales más y estos han de estar bajo el marco del RGPD (conocer el teléfono nos sería de gran valor, pero tendríamos que hilar mucho más fino para informar al cliente de que tiene que dar su consentimiento claro y expreso, los fines para los que se piden, etc.). Los atributos que elegiría (obviando la clave que conectaría a esta tabla de clientes con la de las encuestas) serían cuatro:

- Franja de edad: Menos de 25 años, de 26 a 50, de 51 a 70, más de 70.
- Frecuencia de compra: Diaria, varias veces a la semana, semanal, mensual.
- Día de la semana: (marcar con un círculo una de las 7 opciones dadas)
- Código postal: (campo numérico)

La franja de edad es útil para emprender acciones de marketing selectivas, encaminadas a *targets* diferentes. La frecuencia de compra creo que es un atributo útil para ayudar a diseñar con más precisión la frecuencia de las promociones. El día de la semana nos podría permitir -por ejemplo, a través un algoritmo de Business Analytics- entender algún patrón por el que, por ejemplo, los empleados pueden ser más descuidados al principio de semana con su apariencia, pero menos descuidados con la limpieza al final de semana, lo que posibilitaría acciones mucho más precisas. Por ejemplo, también podría determinar aspectos como relacionar el tiempo en una zona determinada (recordemos que tenemos precisión de localización total) en relación a la limpieza del local, determinando -por ejemplo- si existe algún patrón entre que haya llovido el día de la encuesta y se haya puntuado como sucia la limpieza del local. El código postal me parece muy interesante porque, además de ser un campo rápido de introducir para el cliente, nos permite cotejar dicho CP con el CP de la tienda y estudiar la proximidad del cliente a la tienda, lo que ayudaría también en otros departamentos como el de expansión a la hora de abrir

nuevas tiendas, al saber -por ejemplo- si muchos clientes de una zona lejana concreta vienen a una tienda.

Una vez determinados estos planteamientos, podríamos definir los bloques de información en dos o tres (podríamos, incluso, simplificar la siguiente lista y hacer un gran bloque de información denominado "tiendas", donde figuraran tanto datos de localización como datos administrativos o de facturación). Contestando a la pregunta, podríamos hablar de las siguientes cuatro entidades:

- Tiendas: Datos de identificación y localización de las tiendas
- Ventas: Datos de ventas y objetivos de las tiendas
- Encuestas: Encuestas realizadas en cada tienda
- Clientes: Clientes que contestan a la encuesta

Por supuesto, cada una de estas entidades tiene sus propios atributos, tal y como he definido en la entidad que he añadido CLIENTES (por supuesto, en la tabla en la que se basa esta entidad habría que añadir un campo autonumérico que identifique cada cliente y los campos con los que queremos que se relacione con otras tablas, en este caso, con las encuestas que hacemos). No obstante, dado que únicamente se pide identificar las entidades, no entraré a valorar atributos.

• Encuentra las posibles relaciones de dichas entidades. (10 puntos)

Voy a partir del esquema inicial de tablas y atributos para valorar el escenario inicial, y luego veremos la relación entre entidades. El esquema inicial sería de tablas y atributos sería:

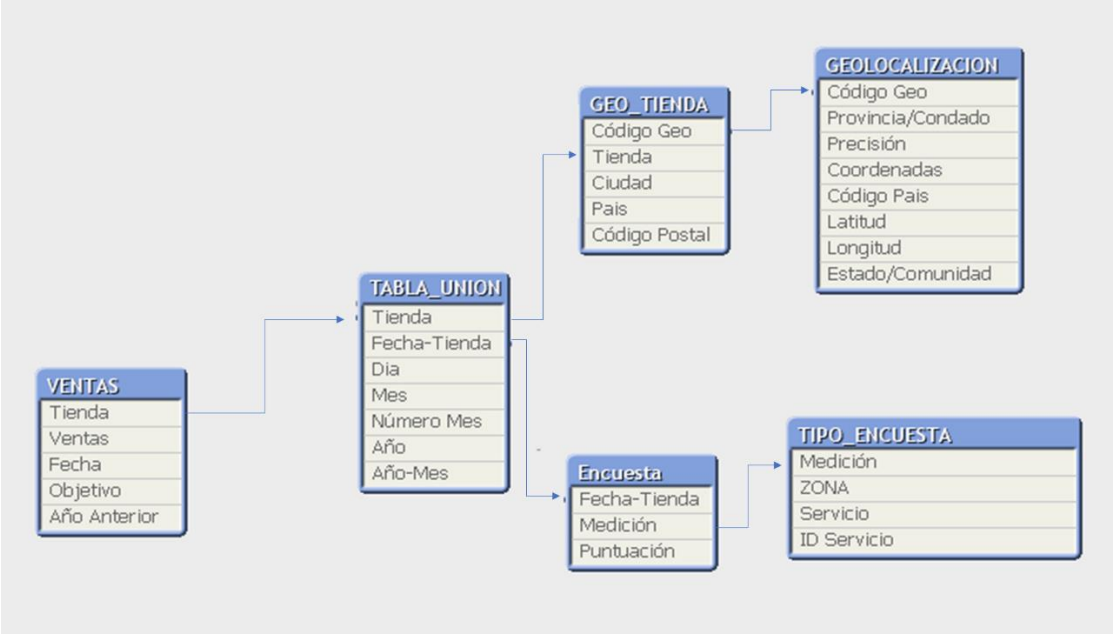
La entidad "ventas" relaciona las tablas de VENTAS y TABLA_UNION (lo que definíamos antes como una tabla de paso), así como las tablas TABLA_UNION con GEO_TIENDA.

La entidad "Código Geo" relaciona las tablas GEO_TIENDA y GEO_LOCALIZACION.

La entidad "Fecha_tienda" relaciona las tablas TABLA_UNION con ENCUESTA.

La entidad "Medición" relaciona las tablas ENCUESTA con TIPO_ENCUESTA.

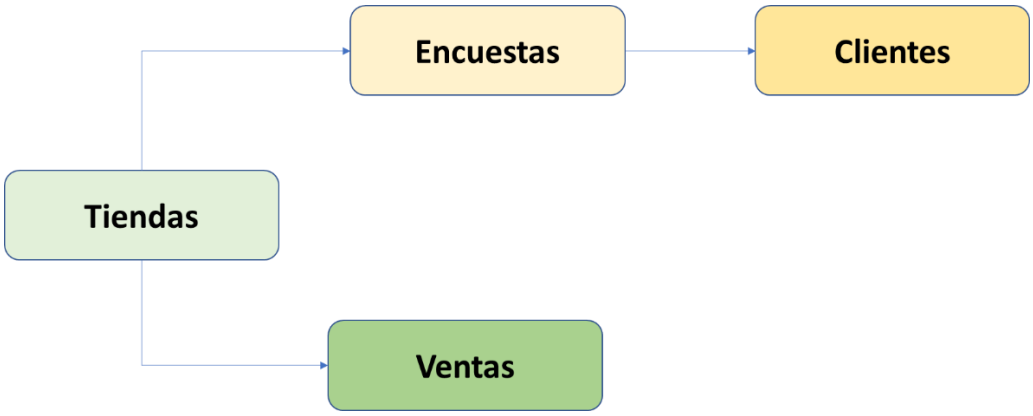
De forma gráfica, quedaría:



No obstante al gráfico anterior, me parece que los datos son inconsistentes en algunos aspectos. Por ejemplo, parece que las ventas se asignan sólo al día 1, apareciendo hasta 11 registros en enero con la misma fecha en varias tiendas, mientras que el resto de fechas no parece seguir un patrón lógico (a no ser que los datos se hayan generado aleatoriamente, claro). Parece, incluso, como si a cada registro de ventas le correspondiera un registro de la encuesta (a excepción, como dije, de enero), ya que parecen tener una distribución parecida… si esto fuese así, habría una relación entre Ventas y Encuestas aunque, sinceramente, no me tiene mucho sentido. Sólo quería mencionar el hecho de que los propios datos (no su estructura) pueden dificultar el establecimiento de relaciones. No es un punto importante, pero analizando los datos quería comentarlo.

Visto el escenario inicial, yo propondría añadir la tabla clientes, la cual iría relacionada con Encuesta.

En cuanto a relación de entidades, las relaciones serían:



Las tiendas serían la entidad central. Cada tienda, tiene unos datos de ventas (donde se incluirían las fechas de cada venta, los objetivos, posibles desviaciones, etc.). Asimismo, cada tienda efectúa varias encuestas, cada una de las cuales está formada por información de los clientes.

El tipo de relación entre entidades sería el siguiente:

- Entre Tiendas y Ventas, la relación sería uno a varios, ya que cada tienda tiene varios registros de venta y cada registro de venta pertenece únicamente a una tienda.
- Entre Tiendas y Encuestas, la relación sería uno a varios, ya que cada tienda puede tener asociadas varias encuestas (por ejemplo, las encuestas pueden ser semanales, mensuales, etc.).
- Entre Encuestas y Clientes, podría haber dos posibilidades. Por un lado, podría ser una relación uno a varios, ya que cada encuesta estaría podría estar contestada por varios clientes. Otra posibilidad sería que tuviera una relación uno a uno, en cuyo caso asociaríamos cada encuesta con un único cliente. Dado que es una entidad nueva que no está recogida en el modelo inicial, dejo abiertas ambas posibilidades de relación.

• Define algunos de los posibles elementos cuantitativos de la información, o métricas. (5 puntos)

- **Puntuación**, dentro de la tabla encuesta: Es una variable binaria que puede adoptar los resultados de 1 y 0. Entiendo que 1 hace referencia a que supera el análisis del ítem detallado en “Medición” y que 0 indica que no lo superaría.

- **Ventas**, dentro de la tabla ventas: Hace referencia a la cifra de ventas, expresadas en unidades monetarias (no tenemos información para determinar la moneda ni si están expresadas en miles o en millones).

- **Objetivo**, dentro de la tabla ventas, hace referencia al objetivo fijado por la compañía para dicha tienda. Al igual que los dos ejemplos anteriores, estamos hablando de métricas ya que miden o cuantifican algo.

• Define alguno de los elementos cualitativos o descriptores de la información, características o entidades dimensionales del supuesto. Completa las relaciones del esquema que ya ha aportado el departamento de BI (ver PEC-Modelo-Tablas.png). Un ejemplo que te puede ayudar en la definición (ver tabla 1). (5 puntos)

- **Provincia**, **Código de país** o **Estado/Comunidad** (entre otras), todas dentro de la tabla geolocalización, son ejemplos de variables cualitativas que sirven para obtener la localización de la tienda con distintos grados de granularidad.

- **Zona, Servicio** o **Medición**, todas dentro de la tabla Tipo_encuesta, son ejemplos de variables cualitativas usadas para clasificar los distintos ítems de la encuesta. De esta forma, en una visión de arriba abajo, el servicio sería la variable de partida, estando formado cada servicio por una zona que, a su vez, se detalla en una o más mediciones.

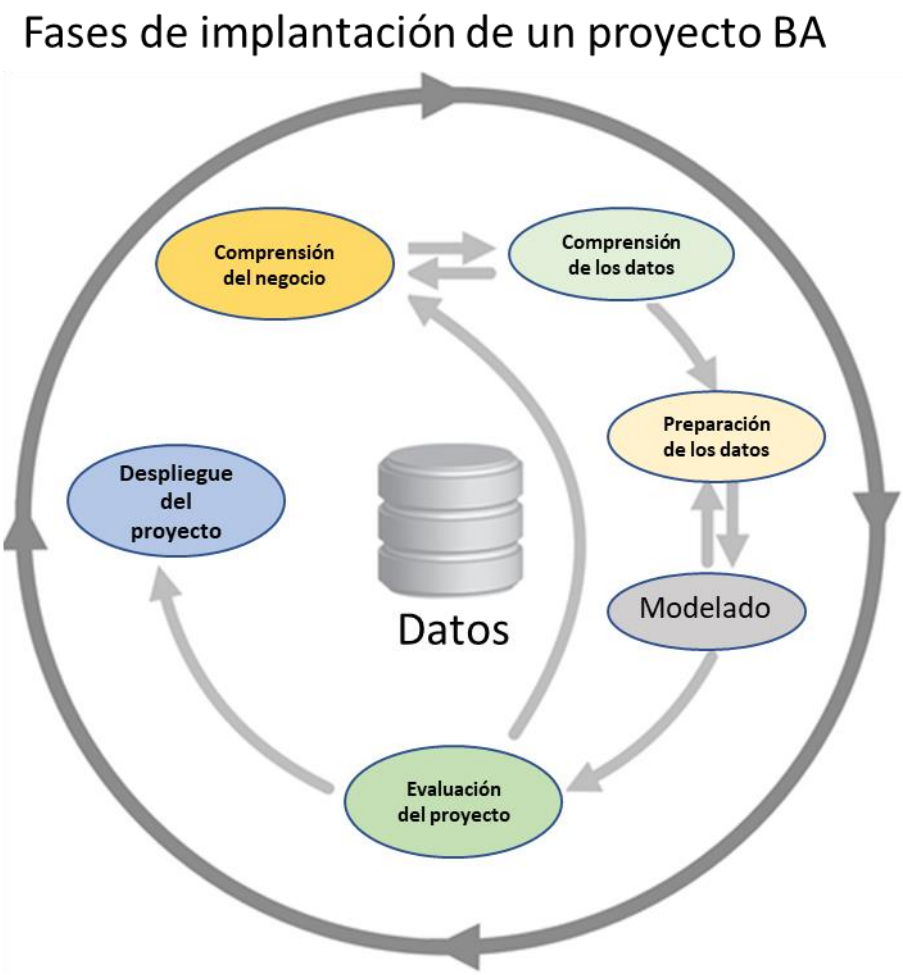
- **País** o **Ciudad**, ambas dentro de la tabla geo_tienda, son ejemplos de variables cualitativas utilizadas para saber, con distintos grados de granularidad, la localización física de una tienda. Al igual que las variables anteriores, estamos hablando de atributos que especifican una característica de la entidad que queremos analizar (en este caso, la localización de la tienda)

[Pregunta 2] (10 puntos). Define el esquema de las fases del proyecto que va desde los orígenes de los datos hasta el despliegue del proyecto. (5 puntos por las definiciones y 5 por el uso de gráficos).

Con la ayuda de gráficos y de definiciones de las fases, se tendrá que explicar de forma esquemática las fases que constituyen la implementación de un proyecto de Operations Analytics.

Aunque, ciertamente, podríamos basarnos en un modelo BI ya que el recorrido y los flujos utilizados son muy similares, voy a optar por un enfoque BA en cuanto a fases. Por tanto, en base a esta consideración y las argumentaciones dadas en el foro, voy a basarme en las fases de un proyecto desde la óptica de Business Analytics, basadas en CRISP-DM (Chapman, 1999).

Vamos a partir, por tanto, del siguiente modelo, del cual iremos desgranando y detallando cada una de sus fases:



Fuente: Elaboración propia, basada en recursos de la asignatura (Curto, Josep. 2018. Qué es Operations Analytics: Conceptos, Técnicas y Tecnologías)

Es interesante recordar que la metodología CRISP-DM se basa en cuatro estadios o niveles de abstracción, jerarquizados desde lo más general a lo más específico. De esta forma, empezaremos detallando las fases (objetivo de esta pregunta), con lo que seguiríamos definiendo las tareas generales, las cuales se desgranarían a su vez en tareas específicas para, finalmente, dar lugar a las distintas instancias de proceso. En definitiva, se basará en proyectar un modelo genérico a un modelo específico.

El siguiente gráfico expresa de forma clara el enfoque que sigue esta metodología:

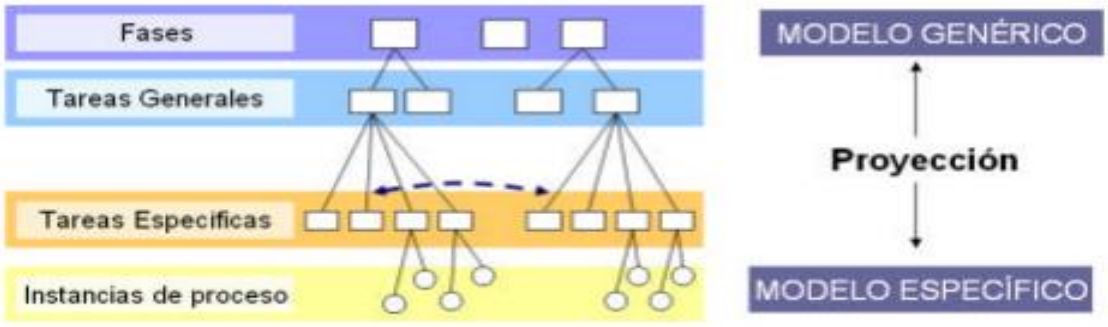


Figura 4: Esquema de los cuatro niveles de abstracción de la metodología CRISP-DM

Fuente: Metodologías para la realización de proyectos de Data Mining (Rodríguez Montequín, Mª Teresa; Álvarez Cabal, J. Rodríguez Montequín, Mª Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo)

Vamos, pues, a detallar lo que se pide en esta pregunta: las fases de implantación de un proyecto.

Comprensión del negocio (Business Understanding)

Este es el punto de partida adecuado para un correcto despliegue del resto de fases. En este punto, tendremos que situar un contexto sobre el cuál establecer los objetivos. Es decir, contestar a las preguntas ¿dónde estamos? y ¿qué queremos conseguir?

Asimismo, habría que analizar cuál es nuestra situación en cuanto a recursos, requerimientos, etc., a fin de ver -entre otras cosas- si nuestros recursos son suficientes para acometer el proyecto. Esta tarea podía responder a las preguntas ¿de cuántos recursos disponemos? ¿es suficiente nuestra infraestructura -física, tecnológica, formativa, etc.- para acometer el proyecto?

Una vez realizadas estas tareas dentro de esta primera fase inicial, tendremos que elaborar un plan de proyecto, donde debería de quedar perfectamente reflejado. En dicho plan, tendrían que quedar especificadas las distintas fases del plan, así como el equipo que lo llevará a cabo y la forma que tendrá de hacerlo. En definitiva, ¿qué hoja de ruta seguiremos? ¿qué técnicas emplearemos?

Esta fase se retroalimentará, a su vez, de la siguiente, con la que mantiene una relación de feedback recíproco.

Comprensión de los datos (Data Understanding)

Una vez tenemos claro el punto en el que nos encontramos, hacia dónde queremos ir y cómo lo haremos, es el momento de profundizar en nuestros datos. Empezaremos por hacer una recopilación inicial de los datos para, posteriormente, describirlos. En estas primeras etapas, es muy útil hacer un análisis exploratorio que nos permita comprender mejor dichos datos, a fin de establecer un primer contacto en donde se pueden localizar problemas, a la par que podemos dictaminar las relaciones más evidentes que posibiliten el establecimiento de hipótesis de trabajo.

Por último, antes de preparar los mismos, es vital que verifiquemos la calidad de los datos. Este proceso nos ahorrará muchos problemas ya que minimiza los riesgos en el proyecto, ahorrará tiempo y recursos a través de un mejor uso de la infraestructura tecnológica y mejorará la toma de decisiones, ya que éstas se basarán en información extraída de datos validados y confiables.

En definitiva, en esta fase podremos contestar a preguntas como ¿de qué datos disponemos? ¿cuán fiables son estos?

Al igual definía antes, el feedback ha de ser recíproco con la primera fase ya que, por ejemplo, la comprensión de los datos o la fiabilidad de estos pueden cambiar aspectos de la fase anterior, como la modificación del objetivo.

Preparación de datos (Data Preparation)

Una vez situados en contexto, podemos acometer la fase de preparación de datos. Este punto se antoja, una vez más, vital, ya que de él dependerá que los datos que se modelicen sean de la máxima calidad posible (de hecho, estas dos fases deberían interactuar de forma sistemática dado que procesaremos los datos de una forma u otra en relación a la técnica de modelado que usemos en la fase posterior).

El primer paso sería la selección de los datos, ya que no todos los datos brutos que tengamos, por muy confiables que sean, nos serán de utilidad. Una vez seleccionados, tendremos que hacer un proceso en que se encontrarían tareas como la limpieza de los datos, la generación de variables que nos puedan hacer falta, la integración de distintos orígenes de datos o el formateo de los mismos.

La idea, por tanto, es conseguir que los datos no sólo sean fiables, sino que estén perfectamente limpios (por ejemplo, sin campos N/A), con el formato adecuado (por ejemplo, un campo de ventas ha de ser numérico para poder operar con él) y con todas las fuentes de datos perfectamente integradas (recordemos que no todos los datos tienen que estar en el mismo sitio).

En definitiva, podríamos responder a la pregunta ¿están los datos preparados para su modelización?

Modelización (Modeling)

Este es el punto donde seleccionaremos la técnica de modelado que mejor se adapte a lo que queremos (recordemos que los objetivos deben de estar perfectamente definidos en la fase 1). La elección de esta técnica se antoja fundamental, por lo que deberemos de tener en cuenta consideraciones como ¿es apropiada la técnica seleccionada a nuestro problema? ¿disponemos de datos adecuados para llevarla a cabo? ¿cumple los requerimientos del problema? ¿cuánto tiempo necesitaremos para obtener el modelo? o ¿tenemos los conocimientos adecuados para llevar a cabo dicha técnica de una forma precisa?

Una vez seleccionada la técnica de modelado, tendremos que hacer un diseño de la evaluación -el cuál posibilitará establecer el grado de bondad de los modelos- a fin de poder construir un modelo válido.

Evaluación (Evaluation)

Llegados a este punto, ya tenemos los datos perfectamente preparados y con un modelado adecuado, procederemos a la generación y evaluación del modelo. En este punto, evaluaremos

tanto los resultados obtenidos con el modelo como el propio proceso. Creo que es importante especificar un poco más el concepto "evaluar el modelo". En este caso, esta evaluación se llevaría a cabo desde una óptica de si el modelo cumple -o no- los criterios de éxito que hemos definido para el problema; por tanto, el enfoque de evaluación no es respecto al punto de vista de los propios datos, sino de lo que hemos definido para el problema. Como digo, también es importante revisar todo el proceso, a fin de intentar localizar algún error, eliminar alguna redundancia que ralentice o dificulte el proceso, etc.

Finalmente, una vez evaluado el modelo, estableceremos los siguientes pasos. Si la evaluación -en los términos descritos- es positiva, procederemos a la explotación del mismo, como veremos en la siguiente fase

Despliegue (Deployment)

Llegamos a la fase final, donde se producirá el despliegue y explotación del modelo. En esta última fase, deberemos de tener en cuenta las siguientes etapas:

- Hacer una planificación del despliegue, a fin de determinar una estrategia para que el modelo pueda estar operativo. Es importante determinar todos los pasos, estableciendo quién y cuándo los ejecutará.

- Planificación de la monitorización y mantenimiento, lo que nos permitirá minimizar el riesgo de algunos efectos indeseados (como un uso incorrecto de los resultados).

- Informe final del proyecto, en el cual deberán estar referidos los distintos entregables que hemos generado en el proyecto. Es usual que, a mayores de dicho informe, se establezcan las conclusiones de forma personal en una última reunión.

- Revisión del proyecto, donde trataremos de sintetizar la experiencia que hemos adquirido en el proyecto, evaluando qué aspectos y elementos han ido según lo planificado, y cuales no. Deberían de estar incluidos, asimismo, los puntos de mejora, a fin de ser tenidos en cuenta para otros proyectos u otros problemas de índole similar que se puedan aprovechar de nuestra experiencia.

[Pregunta 3] (15 puntos). Preparación del dato. (10 puntos generación del atributo Fecha-Tienda, 5 puntos modificación del formato del campo)

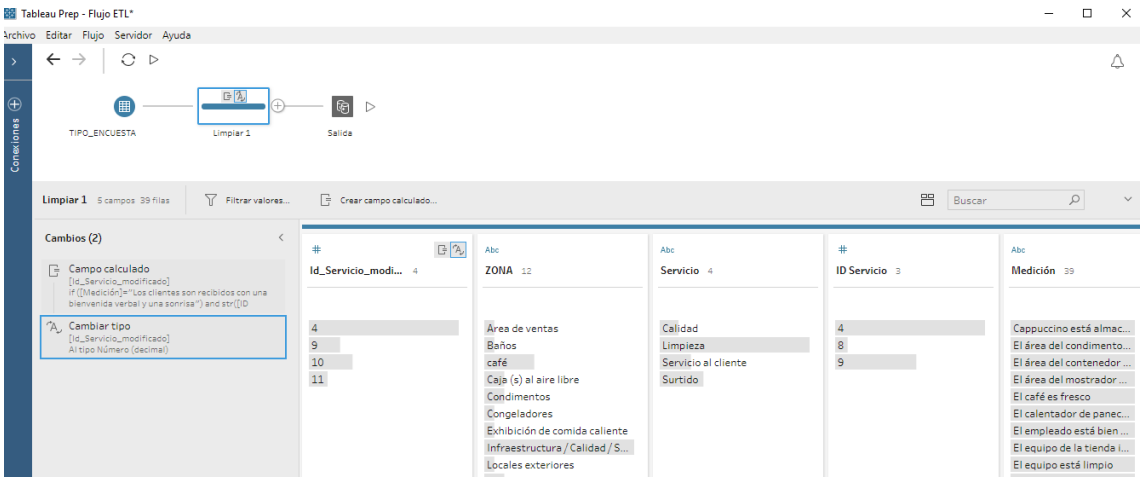
En primer lugar, voy a focalizar este punto de preparación del dato en la tabla TIPO_ENCUESTA, donde no hace falta un análisis demasiado profundo para ver que sus datos tienen inconsistencias. Por ejemplo, ID_SERVICIO y SERVICIO son campos muy habituales en cualquier base de datos, donde el primero identifica numéricamente un registro (siendo dicho identificador, único), mientras que el segundo campo describe en formato texto dicho identificador numérico. En base a esto, existe una inconsistencia con ID_SERVICIO y SERVICIO, dado que para las mediciones:

- Los clientes son recibidos con una bienvenida verbal y una sonrisa
- El empleado está bien arreglado en uniforme con etiqueta de nombre

El identificador único (ID_SERVICIO) es 8, cuando -en realidad- debería de ser consistente con el resto de campos dependientes de dicho indicador. En base a esto, debería de ser 9 y no 8. Este es, pues, el primer cambio que haré con Tableau Prep.

Tambien se nos indica que todas las ocurrencias relacionadas con el campo Servicio = Surtido tiene que tener un ID de servicio 10 y que todas las ocurrencias relacionadas con el campo Servicio = Calidad tiene que tener un ID de servicio 11. Asimismo, he pasado a numéricos el campo de ID_Servicio y el campo calculado (que es el de ID_Servicio modificado, tal y como veremos ahora).

El flujo, en Tableau Prep, es el siguiente:



Como vemos, he añadido un campo calculado llamado ID_Servicio_modificado donde he hecho todos los cambios que se pedían:

- Cambiar el ID_Servicio de 8 a 9 cuando la medición es “Los clientes son recibidos con una bienvenida verbal y una sonrisa”
- Cambiar el ID_Servicio de 8 a 9 cuando la medición es “El empleado está bien arreglado en uniforme con etiqueta de nombre”
- Cambiar el ID_Servicio de “Surtido” a 10
- Cambiar el ID_Servicio de “Calidad” a 11

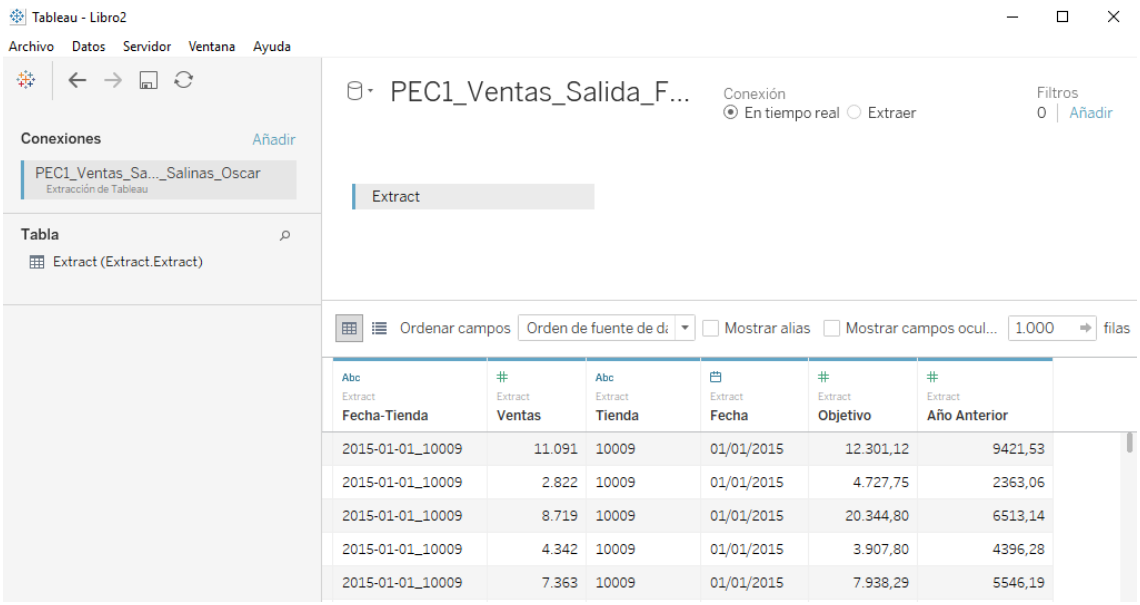
El código utilizado en el campo calculado es:

```
IF([Medición]="Los clientes son recibidos con una bienvenida verbal y una sonrisa") and str([ID Servicio])="8" THEN "9"

ELSEIF ([Medición]="El empleado está bien arreglado en uniforme con etiqueta de nombre") and str([ID Servicio])="8" THEN "9"

ELSEIF ([Servicio]= "Surtido") then "10"

ELSEIF ([Servicio]= "Calidad") then "11"

ELSE str([ID Servicio])

END
```

Quiero añadir que no he borrado el anterior campo, el de ID_Servicio, y he dejado ambos porque no lo pide el ejercicio. No obstante, lo adecuado sería eliminar este ID_Servicio y renombrar el nuevo campo calculado con ID_Servicio

Vayamos ahora con el fichero de ventas

El equipo de BI requiere de un modelo que contenga la concatenación de la Fecha que viene en el fichero de ventas más la concatenación de un símbolo de barra baja '_' y el campo Tienda. La solución, nuevamente, pasa por crear un campo calculado, el cual he denominado Fecha-Tienda. En esta ocasión, cabe decir que he respetado el formato americano de aaaa-mm-dd, ya que creo que es más útil para ordenar (y, además, no he visto ninguna instrucción que sugiera cambiar el formato, lo cual sería fácil con otro campo calculado). He asegurado, asimismo, el formato numérico del campo de ventas resultante. El resultado es:



Abc Extract Fecha-Tienda	Extract Ventas	Abc Extract Tienda	Extract Fecha	Extract Objetivo	Extract Año Anterior
2015-01-01_10009	11.091	10009	01/01/2015	12.301,12	9421,53
2015-01-01_10009	2.822	10009	01/01/2015	4.727,75	2363,06
2015-01-01_10009	8.719	10009	01/01/2015	20.344,80	6513,14
2015-01-01_10009	4.342	10009	01/01/2015	3.907,80	4396,28
2015-01-01_10009	7.363	10009	01/01/2015	7.938,29	5546,19

… donde se puede apreciar en la salida del flujo que los cambios pedidos se han llevado a cabo. El campo calculado ha sido Fecha-Tienda, el cual servirá para relacionar esta tabla con Tabla_Union a través de dicho campo.

[Pregunta 4] (20 puntos) Caso de negocio de Open Data. (10 puntos caso de negocio, 10 puntos descripción)

Del listado de conjuntos de datos abiertos del gobierno de España "datos.gob.es" elige uno e inventa un caso de negocio relacionado con Business Analytics Operacional. Define las principales entidades, medidas, tipos de representación de la información del conjunto de datos, su utilidad y posibles aplicaciones.

Mi caso de negocio es el de una empresa, PlayKids Ltd., situada en el sur de España, dedicada a la fabricación y distribución de mobiliario para parques infantiles, y que está pensando en expandirse. Esta empresa es muy competitiva para determinado tipo de parques (con al menos 100m2 de área, con al menos 45m2 de perímetro y con al menos 5 elementos)... y poco competitiva para el resto de parques infantiles que no cumplan estos requisitos. Se va a abrir un concurso público en Madrid y queremos saber las posibilidades de ser competitivos para emprender acciones *ad hoc* (acciones comerciales, reuniones con los técnicos municipales, publicidad, etc.) y aumentar nuestras posibilidades, ya que sabemos que en este tipo de concursos de pequeña cuantía existe un "componente subjetivo" (sin comentarios) en las contrataciones. ¿Debemos pues, empezar a trabajar en estas acciones antes de que salga el concurso en Madrid o escogemos otra capital de provincia?

El fichero Open Data que he elegido para mi caso de negocio es:

Del enlace http://datos.gob.es/es/catalogo/l01280796-areas-infantiles-municipales1

Dicho fichero, descargado en formato Excel, incluye los datos de 1972 parques infantiles de Madrid, a través de 19 atributos de cada uno de ellos. Las distintas variables que conforman este fichero son

- Tipo: Siempre es área infantil

- Dirección: Calle o plaza donde está el parque

- Identificador: Código que identifica al parque (único por parque)

- Código: Otro código que identifica al parque (único por parque)

- CODNDP,N,19,0: Otro código que identifica al parque (único por parque)

- Tipo de valla: Metálica, de madera, sin valla…

- Tipo de suelo: Arena, caucho…

- Localización: En plaza, en jardines…

- Acceso: Valla retranqueada, acceso con Tramex…

- Nº de elementos: Nº de elementos del parque

- Área: Área en m2

- Perímetro: Perímetro en m2

- Dirección auxiliar: Lo mismo que dirección, pero con otros atributos adicionales (pendiente ascendente o descendente, etc.)

- Distrito: Distrito de Madrid donde se ubica

- Bar rio: Barrio de Madrid donde se ubica

- Latitud: Latitud del parque

- Longitud: Longitud del parque

- UTMX: Otro tipo de coordenadas de latitud

- UTMY: Otro tipo de coordenadas de longitud

Las métricas que tenemos son muy heterogéneas: m2, nº de elementos, coordenadas o identificadores son algunas de ellas. Asimismo, tenemos también variables categóricas como Distrito, barrio, Localización o Dirección. Como vemos, existen elementos muy interesantes para nosotros (para empezar, los tres que nos hacen ser competitivos en un concurso público), aunque también existen elementos que no nos aportarían nada para nuestro propósito.

Vamos a empezar definiendo las entidades de mi caso de negocio:

- Localización física del parque

- Identificación del parque

- Tipo de valla del parque

Óscar Olcina Salinas

- Tipo de acceso del parque

- Nº de elementos del parque

- Área del parque

- Perímetro del parque

Una vez identificadas las identidades, empezamos a tratar los datos. Lo he hecho en Excel, aunque podía también haberlo hecho con Tableau Prep al igual que en la pregunta anterior. Para empezar, he eliminado aquellas variables que no son útiles a nuestro modelo y he incorporado tres variables de paso nuevas y la variable objetivo (identificadas en color salmón):

PERIMETRO	N.ELEMEN	AREA	Perimetro_adecuado	Mas_de_5_elementos	Area_adecuada	**Competitividad**
86,16	7	532,55	Si	Si	Si	**Somos competitivos**
53,92	6	179,466	Si	Si	Si	**Somos competitivos**
64,15	7	225,27	Si	Si	Si	**Somos competitivos**
61,85	6	163,165	Si	Si	Si	**Somos competitivos**
47,52	4	139,554	Si	No	Si	No somos competitivos
64,99	5	266,675	Si	Si	Si	**Somos competitivos**
61,11	4	165,251	Si	No	Si	No somos competitivos
46,54	5	143,939	Si	Si	Si	**Somos competitivos**
60,26	5	201,082	Si	Si	Si	**Somos competitivos**
39,73	8	79,041	No	Si	No	No somos competitivos
70,21	9	263,153	Si	Si	Si	**Somos competitivos**
73,78	9	328,501	Si	Si	Si	**Somos competitivos**
37,73	3	87,249	No	No	No	No somos competitivos
63,67	2	225,271	Si	No	Si	No somos competitivos
51,29	1	181,091	Si	No	Si	No somos competitivos
18,47	2	22,935	No	No	No	No somos competitivos

- **Perímetro_adecuado** se encarga de leer la variable PERIMETRO y dar como salida un "Sí" si ésta tiene 45m2 o más, o "No" si no lo tiene. La fórmula utilizada para el primer registro es:

  =SI(A2>=45;"Si";"No")

- **Mas_de_5_elementos** se encarga de leer la variable N.ELEMENTOS y dar como salida un "Sí" si el parque tiene 5 elementos más, o "No" si no los tiene. La fórmula utilizada para el primer registro es:

  =SI(B2>=5;"Si";"No")

- **Area_adecuada** se encarga de leer la variable AREA y dar como salida un "Sí" si ésta tiene 100m2 o más, o "No" si no la tiene. La fórmula utilizada para el primer registro es:

  =SI(C2>=100;"Si";"No")

- **COMPETITIVIDAD** se encarga de dar como salida un resultado binario: "Somos competitivos" en caso de que las tres celdas que representan las variables anteriores reflejen un resultado de "Sí", indicando "No somos competitivos" si falla un criterio, dos o los tres. Ésta será la variable objetivo que vamos a estimar con el algoritmo, y el fin

último de las tres variables de paso anteriores. La fórmula utilizada para el primer registro es:

=SI(Y(D2="Si";E2="Si";F2="Si"); "Somos competitivos";"No somos competitivos")

Una vez tenemos la etiqueta (resultado) de cada registro a través de la variable COMPETITIVIDAD, las variables de paso sobran. Para ello, las eliminamos y nos quedamos sólo con las variables numéricas y la variable objetivo, por lo que los primeros registros quedarían así:

PERIMETRO	N.ELEMENTOS	AREA	Competitividad
86,16	7	532,55	Somos competitivos
53,92	6	179,466	Somos competitivos
64,15	7	225,27	Somos competitivos
61,85	6	163,165	Somos competitivos
47,52	4	139,554	No somos competitivos
64,99	5	266,675	Somos competitivos

Sólo queda grabar los datos en formato CSV y ya quedaría todo listo para la pregunta siguiente. No obstante, ¿qué pretendemos con esto? Esta nueva variable (Competitividad) tiene una importancia vital ya que a través de ella tenemos clasificado todo el conjunto de datos. ¿Qué utilidad tiene esto? Dado que tenemos etiquetado cada registro (somos competitivos, o no lo somos), podríamos entrenar un algoritmo de aprendizaje supervisado. Ya que hablamos de un atributo categórico, la idea sería construir un algoritmo de clasificación que sea capaz de predecir nuestro nivel de competitividad en un concurso público para un parque infantil en base al entrenamiento del modelo, clasificándolo como "somos competitivos" o "no somos competitivos", algo que desarrollaré en la pregunta siguiente.

La utilidad de la preparación de los datos y, en definitiva, del uso del modelo para el desarrollo de un algoritmo supervisado de clasificación, parece -en mi opinión- clara. Otra cosa será el grado de exactitud que consiga el modelo, algo que veremos en el punto siguiente.

[Pregunta 5] (30 puntos). Predicción en el caso de negocio. (15 puntos explicación, 15 script)

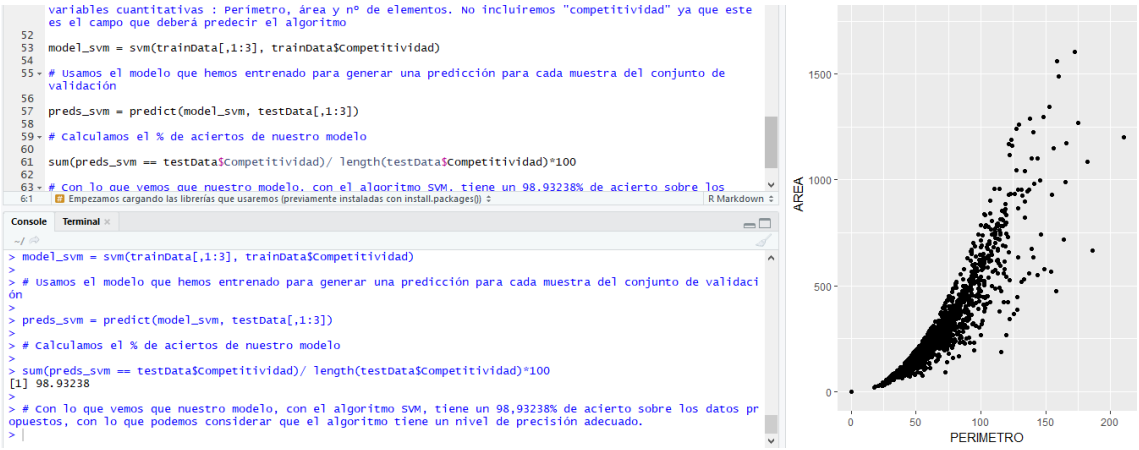
Realmente, he contestado en gran parte a esta pregunta en la exposición de mi caso de negocio en el ejercicio anterior. La utilidad reside en tratar de predecir, en base a los argumentos que hemos calculado dentro de la empresa (en base a las ofertas presentadas en estos años, cuántas hemos ganado, sus atributos, etc.), si podremos ser o no competitivos en este caso.

Esta decisión es importante ya que, para introducirnos en un nuevo mercado como es Madrid, la inversión de recursos será considerable, por lo que conviene tener toda la información posible a fin de determinar si invertimos en Madrid (publicidad, reuniones con los técnicos del Ayuntamiento para que nos conozcan, etc.) o si, realmente, no tenemos muchas posibilidades.

Como dije en la pregunta anterior, tenemos información que, a mi entender, bien utilizada y explotada nos debería de servir para ayudar en la toma de decisiones. Como dije, el hecho de etiquetar cada registro del fichero Open Data nos permitirá utilizar un algoritmo supervisado de clasificación, para lo que he decidido utilizar el modelo SVM.

Este modelo se utiliza tanto en entornos de clasificación como de regresión, aunque -como dije- en nuestro caso, con una variable objetiva categórica, lo utilizaremos como un método de clasificación. El funcionamiento es igual de otros muchos algoritmos, dividiendo el dataset en dos sub conjuntos de datos: un data set de entrenamiento, y uno de test. El SVM representará a los puntos de nuestra muestra (dataset de entrenamiento) en el espacio, por lo que separará las clases en dos espacios lo más amplios posibles a través de un hiperplano de separación, que es lo que llamamos vector soporte; por tanto, cuando se vayan probando las muestras en el modelo, en función de si pertenecen a uno u otro espacio, se clasificarán de una u otra manera (en nuestro caso, si somos competitivos o no lo somos).

El desarrollo del algoritmo se ha hecho en R, tal y como se pedía, aunque -en mi opinión- Python trabaja también de forma excelente este tipo de modelos. Como se puede ver, el resultado es más que satisfactorio, ya que con el entrenamiento del modelo (confrontado, posteriormente, con los datos de test), conseguimos una precisión de más del 98% (en concreto, 98.93238%):



Por tanto, consideramos la utilidad del modelo debido a su alta capacidad de predicción aunque, como indico en el script, existe una correlación altísima (se puede ver en el gráfico anterior) entre dos de las tres variables, lo que desvirtúa el modelo. Al margen de este hecho, creo que puede constituir un buen ejemplo de como funciona un algoritmo supervisado y qué utilidad podemos sacarle.

El script completo es el siguiente:

```
# Empezamos cargando las librerías que usaremos (previamente instaladas con install.packages())

library("e1071")

library("ggplot2")

# Seleccionamos directorio de trabajo

setwd("C:")

# Ahora cargamos el fichero que usaremos para esta pregunta. Previamente he guardado el Excel modificado como fichero CSV

parques<-read.csv(file="c:/parques_definitivo.csv", header=TRUE, dec=",", sep=";", row.names=NULL)
```

 # Sacamos los descriptivos básicos de las variables, de que tipo son y vemos los cinco primeros registros

```
summary(parques)

str(parques)

head(parques)
```

 # Ahora tendremos que cambiar formatos, ya que las tres primeras variables tienen que ser numéricas ("competitividad" está bien como factor) y después comprobamos si ha hecho bien los cambios

```
parques$PERIMETRO <- as.numeric(parques$PERIMETRO)

parques$N.ELEMENTOS <- as.numeric(parques$N.ELEMENTOS)

parques$AREA <- as.numeric (parques$AREA)
```

str(parques)

 # Una vez hemos comprobado que los cambios de formato son los que queremos, estudiamos la relación entre variables mediante gráficos de dispersión. Vemos, de entrada, que hay una correlación clara entre perímetro y área

```
plot(parques)
```

 # …y ahora de forma más exacta

```
cor(parques[,c("PERIMETRO","N.ELEMENTOS","AREA")], use="complete")
```

 # Vemos que la correlación del número de elementos con el perímetro y el área es pequeña (0,49 y 0,45), pero que área y perímetro están muy fuertemente correlacionados (0,9173). De forma gráfica lo veríamos asi:

```
ggplot(parques, aes(x=PERIMETRO, y=AREA)) + geom_point()
```

 # Vamos a empezar ahora a entrenar el modelo. Dividimos el fichero en el 70% para entreno y el 30% para validación

 # En primer lugar, hacemos el set.seed() y dividimos el dataset

```
set.seed(1234)
```

 # Seed inicializa el generador de números aleatorios que usaremos para separar los datos en de entrenamiento (train) y validación (test). Usando un seed fijo, nos aseguramos de que todos generamos los mismos conjuntos y los resultados son reproducibles

```
ind <- sample(2, nrow(parques), replace=TRUE, prob=c(0.7, 0.3))

trainData <- parques[ind==1,]

testData <- parques [ind==2,]
```

 # Aplicamos el modelo SVM, pasándole como parámetros la matriz de entrenamiento compuesta por las 3 variables cuantitativas: Perímetro, área y nº de elementos. No incluiremos "competitividad" ya que este es el campo que deberá predecir el algoritmo

```
model_svm = svm(trainData[,1:3], trainData$Competitividad)

# Usamos el modelo que hemos entrenado para generar una predicción para cada muestra del conjunto de validación

preds_svm = predict(model_svm, testData[,1:3])

# Calculamos el % de aciertos de nuestro modelo

sum(preds_svm == testData$Competitividad)/ length(testData$Competitividad)*100

# Con lo que vemos que nuestro modelo, con el algoritmo SVM, tiene un 98,93238% de acierto sobre los datos propuestos, con lo que podemos considerar que el algoritmo tiene un nivel de precisión adecuado.
```

Notas:

- Soy consciente de que este grado de correlación entre dos de los tres elementos del modelo afectarán muy mucho a los resultados y, por tanto, a la eficacia del modelo, pero creo que valdrá como ejemplo de aplicación de BA a un caso de negocio, recogiendo tanto la operativa a seguir y como los objetivos a buscar.

- Me he ceñido a poner un ejemplo con SVM, pero podría haber utilizado otros algoritmos. Asimismo, a fin de no complicar el script en exceso, tampoco he hecho pruebas con distintos KERNELs, que es algo que sí haría en un ejercicio real.

Bibliografía y webgrafía utilizada

- **Academia.EDU**: "Monografía de la base de datos" (Michael Ravelo). URL: http://www.academia.edu/19636870/Monografia_De_Base_De_Datos
- **Aeiopro.Com**: "Metodología para la gestión de proyectos en Data Mining" (Rodríguez Montequín, Mª Teresa; Álvarez Cabal, J. Valeriano; Mesa Fernández, José Manuel; González Valdés, Adolfo). URL: https://www.aeipro.com/files/congresos/2003pamplona/ciip03_0257_0265.2134.pdf
- **Blog de Luis Pizarro**: "Tecnologías de Información". URL: http://luispizarroconce.blogspot.com/2015/10/23-desarrollo-de-cubo-olap-para-excel.html
- El Blog de Mikel Niño: "CRISP-DM: Fase de "Despliegue" (Deployment)". URL: http://www.mikelnino.com/2016/11/crisp-dm-metodologia-data-mining-despliegue-deployment.html
- Power DATA: "El valor de la gestión de datos". URL: https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/368784/introducci-n-a-la-calidad-de-datos-definici-n-control-y-beneficios
- **Singular Data & Analytics**: "CRISP-DM Fase V. Evaluation. Evaluación (obtención de resultados)". URL: https://data.sngular.com/es/art/30/crisp-dm-fase-v-evaluation-evaluacion-obtencion-de-resultados

- **UOC**: "Aplicaciones Analíticas en Operaciones y Logística" (Cabanillas, David. 2018). URL: https://s3.eu-west-1.amazonaws.com/pilots-elearn-public-assets/B2.589/rmd-materials/2._Aplicaciones_Analiticas_en_Operaciones_y_Logistica.nb.html
- **UOC:** "Fundamentos de Inteligencia de Negocio" (Curto Díaz, Josep. 2016). URL: http://cv.uoc.edu/cgibin/uocapp?s=d9580920baea57ba257967184c4bd8515ea121ecf00cc9d3e4b181281f0f6f18b16e52d4fc3b5efd3c88d3b6af095ede901fd92f602529ad533df3e2b4f0db2f&ticket=ST-334392-J00deF3AwjFA1CKxGdFP-cv.uoc.edu
- **UOC**: "Qué es Operations Analytics: Conceptos, Técnicas y Tecnologías" (Curto, Joseph. 2018). URL: https://s3.eu-west-1.amazonaws.com/pilots-elearn-public-assets/B2.589/rmd-materials/1._Que_es_Operations_Analytics.nb.html
- **Wikipedia:** "Máquinas de vectores de soporte". URL: https://es.wikipedia.org/wiki/M%C3%A1quinas_de_vectores_de_soporte
- **Wikipedia**: "Modelo Entidad-Relación". URL: https://es.m.wikipedia.org/wiki/Modelo_entidad-relaci%C3%B3n