UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE COMPUTACIÓN
MINERÍA DE DATOS

TAREAS Y PROCESO DE MINERÍA DE DATOS

Profesor:
Wilmer González

Integrantes:
Heider Delgado López 24981800
Katherine Colina 19499302
Adolfo Adrián 18913118

Caracas, Junio 2019

Proceso de minería de datos

Planteamiento del problema

Se quiere crear un sistema de recomendación con el objetivo de que dado el conjunto de datos de entrada y tomando en consideración los gustos del usuario (contenido visto con anterioridad) el sistema recomiende contenido aun no visto por el usuario que pudiera ser de su interés.

Para cumplir nuestro objetivo el conjunto de data a utilizar está disponible en: https://www.kaggle.com/CooperUnion/anime-recommendations-database#anime.csv

Nuestro negocio será https://myanimelist.net/ es para ellos que se creara el sistema de recomendación basada en la data obtenida.

Nos plantearemos los siguientes objetivos basados en la data

- Crear un sistema de recomendación de animes para el usuario, de esta forma el usuario que visite la página se sentirá más atraído a quedarse en la página.

- Con un sistema de recomendaciones de anime los usuarios que se queden en la página pueden hacer opiniones/criticas de estos animes además de agregarle valor de usuarios comprometidos.

- Para dar solución a los objetivos se pueden responder las siguientes preguntas:

1. ¿Que un usuario vea series de acción también vera series de drama?
2. ¿Cuáles son los géneros más relevantes y más visto, así como también cuales son los más votados
3. ¿Qué tanto influye que el anime sea una película o una serie de televisión en su puntuación?
4. ¿Es más óptimo recomendar series o películas?
5. ¿Si esta persona vio estas series cuales serían las mejor para recomendarle?
6. ¿Si esta persona coloca varias puntuaciones positivas a ciertas series entonces cuales sería mejor recomendarle?

Generación de datos

Los datos se encuentra disponibles en formato .csv (comma separated values) desde la pagina kaggle en el siguiente URL: https://www.kaggle.com/CooperUnion/anime-recommendations-database#anime.csv

No se requiere de mas acciones para la obtención de los datos. Estos datos fueron recolectados en el 2016 utilizando la API de myanimelist.net actualmente al momento de hacer este trabajo la API de MAL se encuentra fuera de servicio mas sobre el tema: https://myanimelist.net/forum/?topicid=1740204, sin embargo esta disponible una API no oficial llamada Jikan:

https://myanimelist.net/forum/?topicid=1616529&show=50#msg55620086 donde todavía se pueden hacer recolecciones de datos.

Conjunto de datos

El contenido de nuestros archivos de data son los siguientes: se tiene información de 73,516 usuarios y 12,294 animes, estos datos están distribuidos en dos archivos Anime.csv y Rating.csv

Contenido de los archivos:

Anime.csv

Atributo	Tipo	Valores Posibles	Descripción
anime_id	numérico	$\mathbb{N}$	Id único de myanimelist.net para identificar un anime.
Name	texto	[A-Z][a-z]*	Nombre completo del anime.
Genre	categórico	[Drama, Romance, School, Supernatural, Action, Adventure, Fantasy, Magic, Military, Shounen, ...]	Géneros separados por comas del anime.
Type	categórico	Película, Televisión, OVA(Video Original Anime), etc...	Tipo de anime según su formato de publicación.
episodes	numérico	$\mathbb{N}$	Cuantos episodios tiene la serie (1 si es película)
Rating	numérico	$[1,10] \in \mathbb{Q}$	Puntaje medio de un anime en escala de 1 al 10
members	numérico	$\mathbb{N}$	Número de usuarios de la comunidad que perteneces al “grupo” de ese anime.

Rating.csv

Atributo	Tipo	Valores Posibles	Descripción
user_id	numérico	$\mathbb{N}$	Id generado aleatoriamente para un usuario
anime_id	numérico	$\mathbb{N}$	El anime que este usuario a calificado.

Rating	numérico	$\mathbb{Z}$	Puntaje que el usuario a calificado en escala de 1 al 10 (-1 si lo ha visto pero no calificado)

Cabe destacar que un usuario se convierte en miembro de un anime (información reflejada en Anime.csv) cuando añadir anime a su lista de animes visto completamente, adicionalmente a esto los usuarios pueden añadir un puntaje a ese anime esta es la información contenida en el archivo Rating.csv

Preparación de los datos

- Para rating.csv en el atributo rating el valor -1 de serie no calificada deberá ser reemplazado por null para que afecte cálculos como el promedio o media del rating del anime
- Realizar el proceso de EDA (Exploratory Data Analisys)
- Para anime.csv los animes de genero desconocido se les puede asignar un género utilizando alguna técnica de predicción.
- Se pudiera reducir el número de usuarios y generar nuevo atributo de rating basado en los ratings otorgados por miembros que realicen más aportes en la comunidad
- El uso de PCA seria de utilidad para visualizar cuales son los géneros de animes más populares usando rating y miembros del anime.
- Uso de PCA para verificar cual es la relación entre rating y tipo de anime (serie y película)

Vista minable

Una Vista Minable es la consolidación en una única tabla de todas las observaciones y los atributos sobre los que se aplicarian los algoritmos de minería de datos.

Después de realizar el proceso de preparación de los datos, podemos obtener una única tabla con otros conjuntos de datos a partir de PCA para terminar haciendo el procedimiento de minería de datos.

Generación de modelos

Se enfocará en una tarea descriptiva de asociación, filtrado colaborativo, para esto se pueden usar redes neuronales, K-vecinos, regresión lineal o arboles de decisiones con esta puedo obtener las relaciones más cercanas basándome en el histórico de animes y ver las relaciones entre atributos respectivamente.

Patrones modelos

Con estos modelos esperamos obtener graficas llamadas “word clouds” donde tomaremos algunos de los géneros más representativos y los agruparemos con el resto géneros donde el género más representativo este más grande y los géneros con que este se relaciona estén más pequeños

correspondiendo el tamaño del resto de los géneros su relación con el género principal, es decir, mientas más grande sea el tamaño de la palabra más relación tendrá con el género principal

Evaluación e Interpretación

Si los géneros explorados guardan mucha relación, es decir, su tamaño en los “word clouds” es grande, entonces son géneros que deberían ser recomendados entre si. Además de predecir el siguiente género o anime que podría ver el usuario.

Conclusiones:

Sería pertinente añadir en animes recomendados en las páginas de animes para mostrar los distintos animes que el usuario puede estar interesado basado en los géneros, esto puede traer valor a la página al obtener un buen sistema de recomendación.

TDSP

Entendiendo el negocio

Nuestro cliente es https://myanimelist.net/ ellos manejan una página web de animes donde se puede puede consultar duración del anime, género, rating otorgado por los usuarios, miembros suscritos a ese anime, cantidad de episodios entre otros elementos.

El objetivo es crear un sistema de recomendación con filtrado colaborativo, basándose en el contenido visto y por otros usuarios con contenido similar mostrarle animes que el usuario aun no ha visto y que pudieran ser de su interés.

El conjunto de datos a utilizar se encuentra disponible en: https://www.kaggle.com/CooperUnion/anime-recommendations-database#anime.csv

Ese conjunto de datos tiene la siguiente descripción:

El contenido de nuestros archivos de data son los siguientes: se tiene información de 73,516 usuarios y 12,294 animes, estos datos están distribuidos en dos archivos Anime.csv y Rating.csv

Contenido de los archivos:
Anime.csv

Atributo	Tipo	Valores Posibles	Descripción
anime_id	numérico	$\mathbb{N}$	Id único de myanimelist.net para identificar un anime.
Name	texto	[A-Z][a-z]*	Nombre completo del anime.
Genre	categórico	[Drama, Romance, School, Supernatural, Action, Adventure, Fantasy, Magic, Military, Shounen, ...]	Géneros separados por comas del anime.
Type	categórico	Película, Televisión, OVA(Video Original Anime), etc...	Tipo de anime según su formato de publicación.
episodes	numérico	$\mathbb{N}$	Cuantos episodios tiene la serie (1 si es película)
Rating	numérico	$[1,10] \in \mathbb{Q}$	Puntaje medio de un anime en escala de

			1 al 10
members	numérico	ℕ	Número de usuarios de la comunidad que perteneces al “grupo” de ese anime.

Rating.csv

Atributo	Tipo	Valores Posibles	Descripción
user_id	numérico	ℕ	Id generado aleatoriamente para un usuario
anime_id	numérico	ℕ	El anime que este usuario a calificado.
Rating	numérico	ℤ	Puntaje que el usuario a calificado en escala de 1 al 10 (-1 si lo ha visto pero no calificado)

Nuestro objetivo sobre el conjunto de datos para generar las recomendaciones será:

-Crear un sistema de recomendación de animes personalizado para cada usuario con esto se busca que el usuario se mantenga más tiempo utilizando la pagina

-Aumentar el número de usuarios que dan puntaje y hacen críticas sobre los animes agregando así valor a usuarios comprometidos.

Para cumplir con estos objetivos podemos empezar respondiendo las siguientes preguntas:

1. ¿Cuál es la relación entre los géneros de anime que ve un usuario?
2. ¿Cuáles son los géneros más vistos y que tienen más puntaje otorgado por los usuarios?
3. ¿Cuál es la influencia del tipo de anime y su rating?
4. ¿Qué tipo de serie es de más utilidad recomendar?
5. ¿Cuántos miembros hay por tipo de anime?

Obtención de data y entendimiento

En esta fase se realizan las actividades de validación, limpieza y visualización de la data, tenemos en este caso:

- Lo primero a realizar en esta etapa es un proceso EDA (Exploratory Data Analisys) con ello podríamos obtener:

- ✓ promedio de episodios para más adelante analizar si esto influye con el rating
- ✓ se puede obtener el promedio de members y comparar esto con la cantidad de usuarios de nuestra data.
- ✓ promedio de rating del anime con ello se podría inferir que tan bueno es el rating general de todos los animes hasta la fecha.

- ❖ se hace necesario en el atributo rating realizar una limpieza de los registros con valor -1 pues esto genera ruido en los cálculos de promedio, se sustituye este valor por NULL.
- ❖ En el atributo episodes existen registros con valor Unknown, se le debe colocar un valor numérico como mínimo 1, por asociación si se toma en cuenta el type del anime se puede llegar a la decisión de que los type con valor TV tendrán 1 solo episodio, y para los demás types se le puede colocar el valor del promedio de episodes de ese type.
- ❖ Se puede utilizar un gráfico de barras el promedio de rating por type de anime.
- ❖ Crear un atributo promedio, que sea el promedio de rating otorgado por un usuario.
- ❖ Uso de PCA para visualizar las relaciones entre género, rating, members, type.
- ❖ Para el filtrado colaborativo, tendremos que crear una tabla dinámica de usuarios en un eje y nombres de programas de anime en el otro. Esta tabla será de ayuda para validar la relación entre usuarios y los animes y con ello predecir con mejor precisión los posibles animes que serían de interés para un usuario.

Modelado

- ❖ De la fase anterior se genera la necesidad de transformar los datos, se procede a estandarizar restando la media de cada rating no se debe tomar en cuenta los usuarios que han realizado una sola calificación en el rating o que a todos los animes calificados les colocara el mismo valor, esto nos lleva a la generación de una matriz esparcida
- ❖ Luego de los análisis de los pasos anteriores se procede seleccionar anime, rating, episodes y members con esto puedo realizar un conteo de la frecuencia de los géneros de anime y despues de aplicar un algoritmo de entrenamiento como arboles de decisiones o K-vecinos, para obtener los géneros mas representativos, con los géneros mas representativos puedo saber que tan relacionados están entre ellos mediante un wordcloud.
- ❖ Se toma de los datos que generaron el wordcloud un 80% para entrenamiento y un 20% para pruebas
- ❖ Se pudieran realizar varios ajustes sobre la cantidad de datos para entrenar y probar

Implementación

En esta fase consideramos que las fases anteriores transcurrieron con éxito y generaron un modelo lo bastante bueno como para recomendar un anime a un usuario usando el filtrado colaborativo, por ello se procede a utilizarlo dentro de la página del cliente https://myanimelist.net.

Se espera que con la puesta en uso del modelo cuando un usuario este visualizando información de un anime particular se le recomendaran otros animes del mismo género que le podría gustar.

El monitoreo de esta parte es necesario, si un usuario visualiza información de un anime recomendado y posteriormente lo agrega a su lista y le da un rating quiere decir que el anime es muy probable que le gustara, pero es necesario almacenar esta información y validar si el rating otorgado al anime que se le recomendó es un valor bueno.

- Por ejemplo, se podría establecer un rating mayor que 7 como que al usuario le gusto la recomendación, de lo contrario sería necesario analizar los datos almacenados que se basan en los datos obtenidos al dar una recomendación y realizar nuevamente la parte de modelado para corregir el modelo.

Si el sistema genera recomendaciones consideradas buenas entonces el **cliente aceptara** el producto y seria el fin del proceso de minería.

Crisp-DM

Comprensión del negocio

El negocio será https://myanimelist.net/, ellos requieren un sistema de recomendación.

El objetivo es generar un sistema de recomendación basado en los gustos de los usuarios tomando en cuenta los animes que ha visto y los animes que han visto otros usuarios los cuales tienen gustos de animes en común usando para esto filtrado colaborativo.

Para este sistema se busca:

- Crear un sistema de recomendación de animes para el usuario, de esta forma el usuario que visite la página se sentirá más atraído a quedarse en la página.

- mantener a los usuarios navegando en la página y que realicen opiniones/criticas de estos animes además de agregarle valor de usuarios comprometidos.

- Para dar solución a los objetivos del sistema se pueden responder las siguientes preguntas:

1. ¿Que un usuario vea series de acción también vera series de drama?
2. ¿Cuáles son los géneros más relevantes y más visto, así como también cuales son los más votados
3. ¿Qué tanto influye que el anime sea una película o una serie de televisión en su puntuación?
4. ¿Es más óptimo recomendar series o películas?
5. ¿Si esta persona vio estas series cuales serían las mejor para recomendarle?
6. ¿Si esta persona coloca varias puntuaciones positivas a ciertas series entonces cuales sería mejor recomendarle?

Comprensión de los datos.

Para cumplir los objetivo planteados en la fase anterior se tiene el conjunto de data que está disponible en: https://www.kaggle.com/CooperUnion/anime-recommendations-database#anime.csv

El contenido del conjunto de datos es el siguiente, se tiene información de 73,516 usuarios y 12,294 animes, estos datos están distribuidos en dos archivos Anime.csv y Rating.csv

- Se deben analizar los campos que no tengan valores, es decir, si un rating no está cargado, por ejemplo, determinar si eliminar estos datos, o asignarle un valor por defecto
- En el proceso de Análisis Exploratorio de los datos podemos hacer:

- ✓ Realizar graficas de los datos (ejemplo, genero vs rating) y analizar la dispersión de los mismos con esta información podemos visualizar que datos seleccionar o limpiar en las fases posteriores.

Preparación de los datos

- ❖ Selección de los datos: tienes como finalidad generar una vista minable más representativa para el problema planteado. Para ello se pude buscar reducir la dimensionalidad de los datos otorgados para quedarnos con el conjunto de data de mayor significancia, esta selección se puede hacer basada en los da tos obtenidos en el análisis exploratorio, por ejemplo, tomar solo los animes con rating mayor a 6 para que las recomendaciones sean más eficientes.

- ❖ Para la limpieza de los datos se puede remover los acentos, caracteres especiales para un tratamiento más sencillo de los campos que sean texto, eliminar valores que puedan generar dispersión de la data como lo son valores negativos.

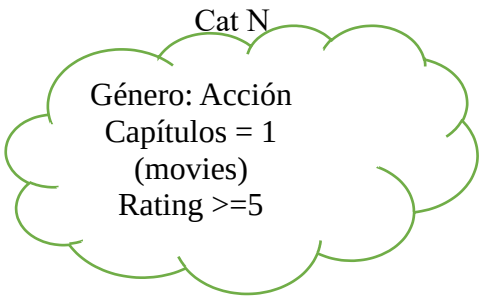
- ❖ Creación de nuevos atributos: puedes crearse atributos que nos permitan relacionar columnas de nuestra base de datos que nos permitan determinar una tendencia o patrón a recomendar programas de una clase en específico. Para la generación de nuevos atributos se puede usar PCA, para crear y estudiar relaciones que podrían ser, por ejemplo. (atributoNew1: relación type- episodes, atributoNew2: Relación name-rating……).

  Por ejemplo, el atributoNew1 podría ser entre la columna genere y la columna members para establecer la relación entre ambas. El atributoNew2 podríamos tener como se relacionas las columnas episodios y rating. De esta forma vamos creando data que permita el entendimiento que a simple vista los datos no pueden otorgar.

Tendríamos algo de la forma:

Clases definidas	Atributos Generados por PCA
Cat 1 Género: Terror Capítulos > 100 Ratin > 7	+ atributoNew1 = Recomendación 1
Cat 2 Género: Si-Fy Capítulos < 15 Rating >=5	

+ atributoNew1 + atributoNew2 = Recomendación 2



+ atributoNewN = Recomendación N

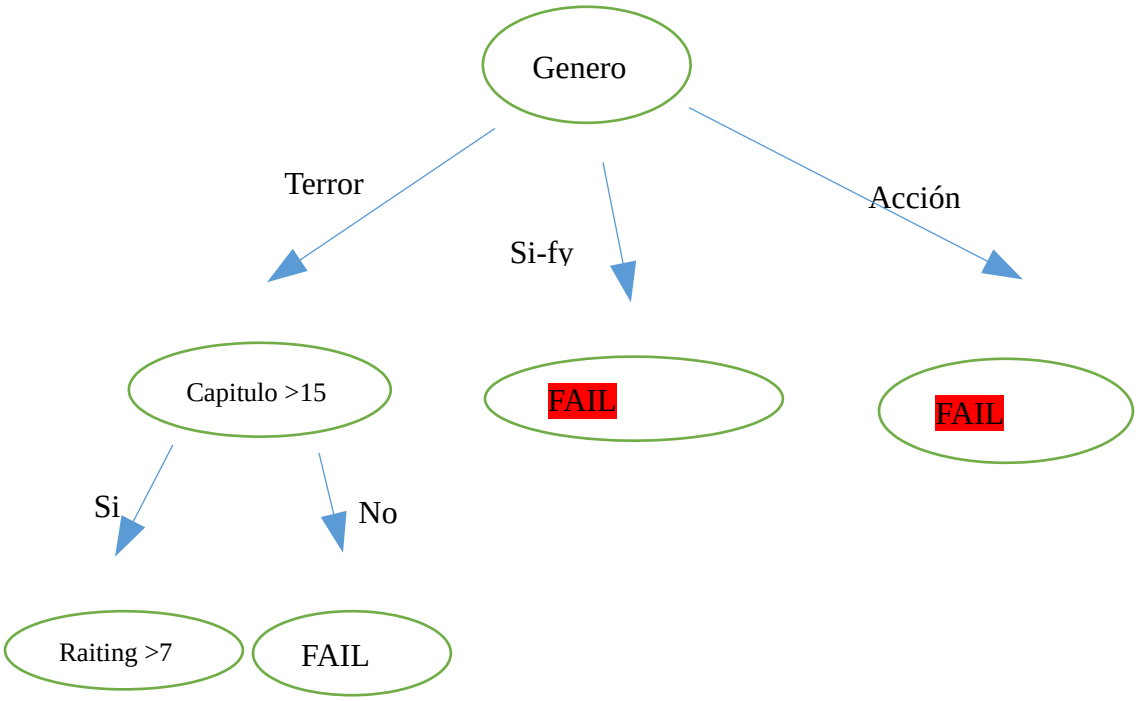
Aquí se nota que las clases están creadas, por ejemplo, con un rating ya determinado (esto debido al estudio realizado en la preparación de datos, para elegir datos con mayor significancia).

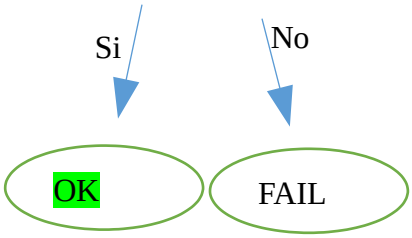
Este paso es de suma importancia en el modelo d CRISP DM ya que quizá se pueda observar que según los tipos de recomendaciones que se generen, sea o no necesario retornar a la fase de generación de preparación de la data y estudiar distintas combinaciones de PCA con su varianza asociada, parar crear otros "atributoNew#", es decir estas dos fases (pre procesamiento y modelado) se pueden retroalimentar.

Modelado

Para poder elaborar la clasificación anteriormente mencionada, se puede utilizar la técnica de arboles de decisiones. Donde según las características de las transiciones se decidirá en que grupo catalogar cierto animé o película. A continuación, un breve ejemplo (parte) de lo previamente estructurado.

Para determinar si pertenece a la Cat 1 descrita anteriormente:





Por otro lado, para realizar las asociaciones de atributoNew# y la categoría correspondiente se puede utilizar el algoritmo de a priori. Donde, ya teniendo los atributos generados por PCA y las categorías de los arboles debemos ver que ítem son frecuentes, según un umbral definido. Para ello se debe realizar la "Itemsets Laticce", que representará todas las posibles formas de combinar los atributosNew con las categorías (categorías y atributoNew, ambos vistos como ítems en esta representación), esto dará como resultado 2^N-1 posibles combinaciones (pero se reducen por el principio monotonicidad), sin embargo, será un número elevado por lo que deben tenerse un número controlado de ítems para luego poder realizar la recomendación correspondiente.

Evaluación

Para la fase de evaluación bastaría con realizar pruebas con ciertos usuarios, donde se podría evaluar si le el show recomendado fue visto y/o calificado, tomando como un parámetro de correctitud de la recomendación el tiempo transcurrido desde la recomendación hasta el día donde este lo visualiza.

Este paso es el más importante uno de los más relevantes ya que desde este, se puede agregar información relevante, que permita un mejor entendimiento del negocio, que será clave para realizar ajustes al modelo.

Es aquí donde se decidirá qué se debe hacer en los siguientes pasos, si como bien se dijo, modificar el modelado, basado en la información obtenida durante la evaluación. O por otra parte plantearse las metas para la fase del deployment, si se realizará la anexión del modelo a una página web, una aplicación móvil, etc.

Explotación

En esta fase se deben realizar tareas de "control", por ejemplo, si todo el sistema esta automatizado debe llevar estadísticas de que tanto los shows (ya sean películas o series) son vistas realmente, cuáles de ellas terminan siendo calificadas por el usuario "target". Basado en este tipo de pensamiento se pude aprender a realizar otro tipo de atributosNew# para mejorar la certeza de las recomendaciones. También se en esta fase es donde realmente se terminará de entender todo el funcionamiento real del negocio (tanto para el diseñador del modelo, como para el negocio mismo), se documentarán los resultados para generar una retroalimentación en todo el proceso

Comparativa entre metodologías

Las metodologías TDSP y CRISP-DM especifican las tareas a realizar en cada fase del proceso de minería de datos, asignando tareas concretas y definiendo el resultado que es deseable de obtener tras cada fase.

CRISP-DM utiliza un proceso cíclico donde algunas fases son dependientes de otras, es decir, el resultado de una fase es la que da generación a la siguiente pues utiliza la información obtenida en ella para cumplir con sus objetivos.

TDSP utiliza una metodología ágil dentro de su ciclo de procesamiento lo cual permite hacer análisis más inteligentes pues las relaciones de dependencia entre las fases más importantes del ciclo son bidireccionales lo cual permite redefinir más fácilmente objetivos particulares de cada fase que pueden afectar a otras.

Al realizar el análisis de nuestro caso de estudio se evidencio que las metodologías TDSP y CRISP-DM tiene ciertas diferencias en la definición de objetivos de cada fase de su ciclo de vida,

tenemos que en la fase de entendimiento del negocio para CRISP-DM solo se establecen los objetivos del negocio y se genera un plan para elaborar el proyecto realizándose preguntas que puedan dar respuesta al objetivo del negocio.

En TDSP esta fase además de ser en la cual se definen los objetivos del negocio también se identifican cual es el origen de los datos y se describen, la fase de identificar los datos ocurre en CRISP-DM en la fase de comprensión de los datos donde se describen y exploran los datos, además de verificar la calidad de los mismos.

En TDSP la fase de adquisición de los datos se realiza un análisis exploratorio de los datos, se valida la calidad de la data y se busca comprender los patrones que son inherentes a los datos y se realiza limpieza de los datos de ser necesario, para CRISP-DM esta fase la fase de comprensión de los datos agrega a lo anterior identificar el origen de los datos y describir los mismo.

En CRIP-DM la fase de preparación de los datos corresponde con las tareas para obtener la vista minable como lo son selección, limpieza, construcción, integración y normalización de los datos, esto en TDSP ocurren la fase de modelado, donde se seleccionan características, se transforma la data, se entrena el modelo con los datos seleccionados, se construyen nuevas características, se evalúa el modelo usando validación cruzada.

La fase de modelado para CRISP-DM corresponde a seleccionar la técnica de modelados, diseñar un modelo de datos, construir el modelo y evaluarlo, es evidente que las fases de modelado para ambas metodologías comparten elementos en común, pero CRISP-DM tiene ciertos elementos como lo es

construcción de atributos en la fase de preparación de los datos que para TDSP se realiza en la fase de modelado.

La fase de evaluación de CRIPS-DM se contempla dentro de la fase de modela de TDSP.

La fase de despliegue en ambas metodologías implica monitorear el modelo validar que esté cumpliendo con los objetivos planteados al inicio del proceso, esto no implica que acá termine el uso de la metodología pues se pudiera ampliar los objetivos del negocio o redefinir los objetivos que se tenían originalmente, pero basándose en el modelo que ya está en uso.