Taller Certificación D&A

Objetivo del taller

El objetivo del taller es utilizar los conceptos que transmitimos a lo largo de todas las clases y diferentes temáticas del taller. La obtención del certificado en introducción a Data & Analytics estará condicionado por la aprobación de este taller.

Formato de entrega

Fecha de entrega: lunes 12 de abril 8 am.

Defensa: 14 de abril, agenda entre 9 a 11 am.

La defensa consta de una presentación de 10 minutos en la cual se debe exponer las conclusiones a las que se llegó. 5 minutos para consultas del cuerpo docente.

Se espera la entrega de un documento, con la respuesta a cada pregunta, indicando qué elementos y/o medidas se utilizaron para dar respuesta.

En la herramienta Cognos Analytics, dentro de la sección “Contenido del equipo” se encontrará una carpeta por cada equipo. A dicha carpeta únicamente tendrán acceso los integrantes del equipo y los profesores.

Todos los objetos creados en Cognos Analytics (módulo de datos, dashboards o reportes) deben ser guardados en la carpeta de cada equipo para poder ser corregidos y evaluados.

Problema a resolver

Nuestro cliente es Pronto Cycle Share system. Pronto está ubicado geográficamente en Seattle y provee 500 bicicletas distribuidas en 54 estaciones. La información que tenemos disponible es:

- Stations: contiene la información relativa a las estaciones.
- Trip: contiene la información de cada uno de los viajes realizados por los usuarios.
- Weather: contiene información relativa al clima.

Business Intelligence:

Utilizando IBM Cognos Analytics, se debe crear un módulo que permita la exploración de la información antes mencionada. En la carpeta de contenidos del equipo, encontrarán cargados los 3 archivos csv (Stations, Trip y Weather).

Cada equipo contará con una carpeta personalizada y restringida por seguridad. A cada carpeta podremos entrar los profesores y los integrantes del equipo.

Utilizando la información provista, debemos ayudar a nuestro cliente a responder las siguientes preguntas:

- ¿Dónde están ubicadas geográficamente las estaciones?
- ¿Cuáles son las 10 estaciones más utilizadas? ¿y las menos utilizadas?
- ¿El clima afecta al uso de las bicicletas? ¿De que manera?
- ¿Se están utilizando las bicicletas? ¿Cuál es la duración promedio de los viajes?
- ¿Cuál fue la bici más utilizada?
- ¿Cuántos viajes promedio tiene cada bici?
- ¿Cuántos viajes promedio por día?
- ¿Hay recorridos más frecuentes que otros?
- ¿Podemos conocer mejor a nuestros usuarios? ¿Qué público es el que nos utiliza?
- ¿Cómo evolucionó a través de los años la cantidad de usuarios registrados (member)?

Big Data y ML:

Utilizando Python con pandas o Python con Spark, en base a los datos de trip.csv, stations.csv y weather.csv realizar los siguientes análisis. En particular, las tareas de E-L-T deben ser resueltas con Spark.

Ejercicio 1 – Limpieza y exploración

1) Limpiar duplicados y generar los atributos derivados que se consideran apropiados.

2) Analizar los valores nulos para todas las variables y proponer estrategias para completarlas cuando sea pertinente. En cada caso utilizar la estrategia que consideren más apropiada y justificar brevemente.

3) Análisis de variables:

- ¿Cuáles son numéricas? ¿Cuáles son categóricas? ¿Existe otro tipo de información?
- Para las variables numéricas, analice su distribución, extraiga conclusiones. Utilice la mayor cantidad posible de herramientas dadas en clase para presentar sus conclusiones.
- Para las variables categóricas, analice su distribución y proponga métodos para codificarlas (OHE, ordinal, dummy, etc).
- ¿Cómo trabajaría las fechas? Justifique brevemente.

Luego de concluida la limpieza y exploración de los datos, realice un join entre los tres conjuntos de datos. Guarde el archivo final en el formato de preferencia y el notebook con el código ejecutado y los resultados desplegados.

Ejercicio 2 - Modelado

Tomando los datos generados en el ejercicio anterior, se plantean los siguientes problemas de negocio:

1) Para ampliar el conocimiento de nuestros clientes nos gustaría tenerlos agrupados de acuerdo a su comportamiento, en base a los datos que podemos obtener proponer dos alternativas de clusterización y justificar su valor para el negocio.

2) Se detectó en el último tiempo desajustes en el volumen de inventario en el año y la demanda, para mejorar esta situación se quiere implementar un modelo de ML que pueda predecir la cantidad de viajes esperados a lo largo del año.

Sobre los resultados del mismo realice dos recomendaciones (cantidad de bicis por estación en cada momento, necesidad de aumentar stock disponible, etc).

- Para la variable a predecir definir cuantiles para la cantidad de viajes por mes y trabajar como un problema de clasificación.

- La variable clima está a nivel diario, proponer una forma de considerar el clima de todo el mes.
- Pueden utilizar gráficas para comunicar los resultados o para explorar si lo desea.
- Use toda la información que considere apropiada y justifique su elección.

Luego de realizado el análisis guarde el o los notebooks que generaron para ambos problemas con el código ejecutado y los resultados desplegados.

Descripción de los datos

Stations.csv

station_id	Identificador de la estación.
Name	Nombre de la estación.
Lat	Latitud de la estación.
Long	Longitud de la estación.
install_date	Fecha en la cuál comenzó a funcionar la estación.
install_dockcount	Cantidad de docks en la estación el día de la instalación.
modification_date	Fecha en la que se modifico la estación, por ejemplo de ubicación o cantidad de docks.
current_dockcount	Cantidad de docks actuales.
decommission_date	Fecha en la cual se desinstaló la estación

Trip.csv

trip_id	Identificador del viaje.
starttiime	Comienzo del viaje.
stoptime	Fin del viaje.
bikeid	Identificador del viaje.
tripduration	Duración del viaje en segundos.
fromstationname	Nombre de la estación en la cual comienza el viaje.
tostationname	Nombre de la estación en la cual finaliza el viaje.
fromstationid	Identificador de la estación en la cual comienza el viaje.
tostationid	Identificador de la estación en la cual finaliza el viaje.
usertype	Tipo de usuario.
gender	Genero del usuario
birthyear	Año de nacimiento