Tecnológico de Monterrey

ITESM

MAESTRÍA EN INTELIGENCIA ARTIFICIAL APLICADA

SEGUNDO TRIMESTRE

SEMANA 3 ACTIVIDAD 1

MODULO 5

NOMBRE: VILLALPANDO GUERRERO JIRAM CESAR

Parte1. Fundamentos de bases de datos y para ciencia de datos.

Los datos son el insumo principal para la en la ciencia de datos pues necesitamos de una cantidad considerable para poder realizar un análisis determinado. Por ello las bases de datos están implícitas en la ciencia de datos ya que todo el tiempo será indispensable. Una base de datos puede estar en cualquier área: finanzas, marketing, rrhh, producción, tecnología, etc, pues todas las organizaciones hacen uso de datos empresariales, y estas pueden ser de distintos tipos, ej: SQL o NoSQL. Asimismo, es importante tener saber cómo realizar importación de todas las distintas bases, ya que, al momento de tener nuestro código, cada una se diferenciará en una pequeña manera técnica.

Las bases de datos relacionales son un tipo en donde los datos están bien organizados y se almacenan en tablas que se vinculan entre sí para establecer relaciones. Algunas características son: flexibilidad, facilidad de uso, no redundancia y confiabilidad.

Por otro lado, las bases de datos no relacionales son un tipo que se caracteriza porque los datos se conectan por categorías en lugar de relaciones. En este tipo de base de datos los datos se organizan en algo pareció a un documento. Algunas características son: Coherencia, disponibilidad y tolerancia a la partición

Fundamentos de almacenes de datos (Data Warehouse) para ciencia de datos.

Un Data Warehouse es un tipo de gestión que está enfocado en habilidades de Business Intelligence y su análisis, generalmente estos almacenes solo se hacen para consultas y análisis con gran cantidad de datos. Los almacene de datos son:

Que están orientado al tema, están integrados, no son volátiles y son variantes de tiempo.

La arquitectura de los almacenes de datos cumple con los siguientes puntos: son simples, cuentan con un área de preparación sencilla, los datos se agregarán al repositorio central y por último que los sandboxes son áreas privadas y seguras.

Al igual que el desarrollo de todas las tecnologías los almacenes de datos han evolucionado y cuentan con distintas características. Por ejemplo, en la actualidad existen almacenes de datos en la nube, y estas tienen algunas ventajas como: compatibilidad elástica y escalable, facilidad de uso para la mayoría de las personas, facilidad de manejo pues se tiene a proveedores y por lo tanto se tiene un ahorro de costos finales.

Asimismo, es importante destacar la arquitectura de datos moderna, estos cumplen con las siguientes características: base de datos convergentes, servicios de transformación e ingesta de datos de autoservicio, compatibilidad sql, múltiples opciones de análisis y lo más interesante que tiene una gestión automatizada.

Parte 3: Preparación de los datos

¿Qué datos considero más importantes? ¿Por qué?

Consideré más importantes a todos aquellos que no contenían algún nulo, pues de esta manera se trató de que no afectara la tendencia de todos los demás datos. Porque el objetivo era la limpieza de datos, entones primeramente se buscaron a aquellos que eran nulos a través de df.inst().values.any(), esto arrojó los 42 valores que tenían nulos. Habría quizá otros datos más importantes a estos si se tuviera que hacer un análisis específico con alguna variable determinada o en su defecto si se tuviera algún objetivo específico para realizar este análisis.

¿Se eliminaron o reemplazaron datos nulos? ¿Qué se hizo y por qué?

Se decidió eliminar los datos nulos porque, se determinó que no eran una cantidad considerable ya que del total formaban tan solo 42, y la base de datos cuenta con 30mil, por lo tanto, una pequeña cantidad no afectaría de manera profunda al análisis final. Cabe destacar que sí se analizaron otras opciones como implementar la media en los casos faltantes o realizar una interpolación ente los dos datos en los que se encontraba en valor nulo.

¿Es necesario limpiar los datos para el análisis? Sí / No / ¿Por qué?

Considero que sí es necesario realizar una limpieza de los datos pues es un proceso en que se corrigen las entradas que tienen un error en un macro de datos, por lo tanto, de esta manera se detectar partes incompletas o que no beneficiarán en el análisis de datos que se está buscando. Si bien en este caso no fueron muchos datos nulos, es una buena práctica para estas actividades en general ya que en otras fuentes pueden existir problemas más profundos y un sesgo debido a esto.

¿Existen problemas de formato que deban solucionar antes del proceso de modelado? Sí / No / Por qué.

En este caso no encontré problemas ya que la base de datos se encontraba muy bien organizada y tenía la mayoría de los campos llenados, por ende, no se tuvo que solucionar algún problema más que la limpieza de datos de los pocos campos que se encontraban con nulos. Cabe señalar que la base de datos recuperada no tuvo problemas al ser llamada ya que es tomada de un sitio web.

¿Qué ajustes se realizaron en el proceso de limpieza de datos (agregar, integrar, eliminar, modificar registros (filas), cambiar atributos (columnas)?

Para realizar la limpieza decidí hacer eliminación de filar pues considero que al no ser tantos registros que en este caso fueron 42, no tiene un impacto fuerte al final del análisis. Al comienzo pensaba que quizá sería mejor usar la media para sustituir estos valores, pero reitero que al ser pocos valores la mejor decisión fue la eliminación. Además, el método de eliminación es útil y se recomienda para bases de datos grandes, como fue el caso de esta, pues no produce algún sesgo y además los resultados son conservadores.