LABORATORIO EN CLASE
COMPRENSIÓN DE LOS DATOS

Profesor: Elías Buitrago B.

Última actualización: 30-Noviembre-2023

OBJETIVOS DE LA CLASE

- Experimentar con las actividades relevantes que propone IBM para la fase "data understanding" de la metodología CRISP-DM en un caso de estudio definido.

INTRODUCCIÓN

Existen diversas guías con el paso a paso de actividades a ejecutar durante la segunda fase de la metodología CRISP-DM. Sin embargo, se podría afirmar que se han propuesto cuatro grupos de actividades principales: Recolectar datos iniciales, describir, explorar y verificar la calidad de los datos. Para mayor profundidad al respecto, por favor revisar en el material adjunto de clase la lectura relacionada; también puede consultar la lectura directamente en línea en la fuente original. Partiendo de los conceptos que se explican detalladamente en la lectura citada, se recomienda continuar detallando cada uno de los cuatro grupos de actividades. Para efectos de esta clase, se seguirán las guías metodologías c propuestas por IBM. A continuación, se presenta una síntesis con los enlaces correspondientes a las fuentes originales, recordando que cada lectura se podrá consultar en el material adjunto a la clase.

Recolectar datos iniciales

Según IBM, se deben tener en cuenta una serie de preguntas en el momento de realizar la recolección inicial de datos. En esta sección se listan dichas preguntas. Puede consultar la publicación detallada mediante este hipervínculo.

- ✓ ¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?
- ✓ ¿Qué variables parecen irrelevantes y pueden ser excluidos?
- ✓ ¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?
- ✓ ¿Hay demasiadas variables para el método de modelado de su elección?
- ✓ ¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?
- ✓ ¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?

Describir los datos

Según IBM, es recomendable "enfocarse en la cantidad y calidad de los datos, y realizar un reporte de descripción de los datos; puede consultar la publicación detallada mediante este hipervínculo. De igual manera, las preguntas que sugiere IBM para responder en esta sección son las siguientes:

- ✓ ¿Cuál es el formato de los datos?
- ✓ ¿Cuál es el método utilizado para capturar los datos?
- ✓ ¿Qué tamaño tiene la base de datos (en número de filas y columnas)?
- ✓ ¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?
- ✓ ¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?
- ✓ ¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?
- ✓ ¿Es capaz de priorizar las variables relevantes? Si no es así, ¿hay analistas de negocio disponibles para proporcionar más información?

Explorar los datos

Según IBM, es recomendable "utilizar herramientas de visualización" para explorar los datos. En la práctica esto significa aplicar estadística descriptiva buscando gráficos para visualizar de manera resumida los hallazgos. De esta manera el equipo de analítica de datos podrá entender mejor la naturaleza de los datos. Por ejemplo, definir si los datos se corresponden con una distribución de probabilidad determinada. Además, tendrán insumos valiosos para realizar entregar un reporte de descripción de los datos, en caso de que sea requerido. Una guía detallada sobre este punto se puede consultar aquí. De igual manera, las preguntas sugeridas para responder en esta sección son las siguientes:

- ✓ ¿Qué tipo de hipótesis se ha formado sobre los datos?
- ✓ ¿Qué variables parecen prometedoras para un análisis más profundo?
- ✓ ¿Sus exploraciones han revelado nuevas características sobre los datos?
- ✓ ¿Cómo han cambiado estas exploraciones su hipótesis inicial?
- ✓ ¿Considera que debería reformular el alcance del proyecto?
- ✓ ¿Esta exploración ha alterado los objetivos?
- ✓ ¿Puede identificar subconjuntos particulares de datos para su uso posterior?

Verificar la calidad de los datos

Según IBM, es recomendable verificar la calidad de los datos con un enfoque en los siguientes aspectos:

- Identificar datos faltantes
- Identificar errores tipográficos en los datos
- Identificar errores de medición (en las unidades de medida)
- Identificar inconsistencias en la codificación
- Identificar deficiencias en los metadatos

Complementariamente, IBM sugiere dar respuesta a las siguientes preguntas:

- ✓ ¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?
- ✓ ¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?
- ✓ ¿Ha explorado las desviaciones para determinar si son "ruido" o fenómenos que vale la pena analizar más a fondo?

- ✓ ¿Ha realizado una comprobación de plausibilidad de los valores? Tome notas sobre cualquier conflicto aparente (como adolescentes con altos niveles de ingresos).
- ✓ ¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?
- ✓ ¿Los datos se almacenan en archivos planos? Si es así, ¿Son los delimitadores coherentes entre los archivos?
- ✓ ¿Cada registro contiene el mismo número de campos?

DESCRIPCIÓN GENERAL

Se conformarán grupos de máximo 3 estudiantes asignados por cada computador disponible en la sala de sistemas asignada. El taller en clase se apoya en las siguientes actividades:

1. Leer con detalle la lectura 6.
2. Leer con detalle la lectura 7.
3. Desarrollar los lineamientos de IBM en lo relacionado con "data understanding" para un el caso de estudio que se plasma mediante el conjunto de datos "*housing_fincaraiz.csv*". Esto implica hacer limpieza de los datos iniciales, describirlos y explorarlos, así como realizar la respectiva verificación de calidad de estos. Opcional, podrían utilizar un dataset que descarguen durante la actividad de web scraping.
4. Deben preparar un informe en PDF que evidencie el trabajo realizado. Así mismo, debe quedar evidencia del código en su perfil de GitHub.

Seminario Big Data, analítica de datos y sistemas de información

INFORME DE LABORATORIO 3, COMPRENSIÓN DE LOS DATOS

Docente:
Elías Buitrago Bolívar

Jhoan Sebastian Riaño Herrera
38525

Brayan Camilo Salazar González
70058

Facultad de Ingeniería
Ingeniería Biomédica
Bogotá DC
2024

DESARROLLO

1. **Recolección datos iniciales**

   Después de ejecutar el código planteado sobre Webscrapping, se obtuvo una base de datos con registros del portal web de Tucarro.com, estos contenían unas variables ya predeterminadas que conformarían toda la base de datos, dando respuesta a las preguntas que respectan a esta sección del laboratorio a continuación:

   - ✓ ¿Qué variables (columnas, atributos) de la(s) tabla(s) o base(s) de datos parecen más prometedores?
     Para poder predecir el valor de un vehículo, los atributos que parecen mas prometedores son el año del vehículo y el kilometraje.
   - ✓ ¿Qué variables parecen irrelevantes y pueden ser excluidos?
     Se pueden excluir las variables del precio de venta y el modelo o referencia del vehículo, dado que nos centraremos en el Volkswagen Jetta sin importar la referencia.
   - ✓ ¿Hay suficientes datos para sacar conclusiones generalizables o hacer predicciones precisas?
     En un primer acercamiento si pueden ser suficientes, basados únicamente en el kilometraje u el año del vehículo, para un mejor modelo de predicción se podría implementar variables que sean un poco más subjetivas, tales como, el color del vehículo, estado general, modificaciones o tipo de combustible que utiliza.
   - ✓ ¿Hay demasiadas variables para el método de modelado de su elección?
     Pensamos que para un primer modelo de predicción, son suficientes.
   - ✓ ¿Está fusionando varias fuentes de datos? Si es así, ¿hay áreas que podrían plantear un problema al fusionar?
     No, hasta el momento únicamente se está utilizando una fuente de datos.
   - ✓ ¿Ha considerado cómo se manejan los valores que faltan en cada uno de sus orígenes de datos?
     Inicialmente, pensamos que tenemos que analizar cuales son los datos que faltan y en qué proporción puede afectar el modelo, de este modo, se piensa que se pueden omitir dichos registros.

2. **Descripción de los datos**

   Ya teniendo la base de datos descargada, se procede a organizar con el fin de esclarecer cuales son los datos y variables que se tienen. De esta manera siguiendo las indicaciones de **IBM** se identifica que tenemos variables de tipo numérico, ya que son, el kilometraje, el año del vehículo y el precio de venta propuesto por el vendedor. Las preguntas planteadas en esta sección del laboratorio se responden a continuación:

   - ✓ ¿Cuál es el formato de los datos?
     Los datos vienen en un formato de archivo *.csv

- ✓ ¿Cuál es el método utilizado para capturar los datos?
  El método escogido para la captura de los datos fue Webscrapping.
- ✓ ¿Qué tamaño tiene la base de datos (en número de filas y columnas)?
  La base de datos tiene un tamaño de 629 filas y 4 columnas.
- ✓ ¿Incluyen los datos una o más variables relevantes para la pregunta de negocio?
  Si, incluyen 2 variables que consideramos relevantes, las cuales son, el kilometraje y el año del vehículo.
- ✓ ¿Qué tipos de datos están presentes (simbólicos, numéricos, etc.)?
  En el archivo original, los datos tenían datos alfanuméricos en el nombre del vehículo, simbólico en el precio del vehículo y en el kilometraje se encontraba el uso de coma (,) como separador de miles.
- ✓ ¿Ha calculado estadísticas básicas para las variables clave? ¿Qué información le ha proporcionado sobre la cuestión de negocio?
  No, hasta el momento no se han calculado estadísticas básicas para las variables.
- ✓ ¿Es capaz de priorizar las variables relevantes? Si no es así, ¿hay analistas de negocio disponibles para proporcionar más información?
  Pensando firmemente que las variables mas relevantes en el modelo predictivo serán el kilometraje y el año del vehículo, están priorizadas con el fin de entender el posible comportamiento de los registros.

3. **Exploración de los datos.**

Ya teniendo la base de datos, se procedió a aplicar filtros de búsqueda que permitieron establecer ciertas características de los registros, por ejemplo, que la mayor cantidad de vehículos en venta son aquellas referencias que su cilindraje está por encima de los 2.0L. Respondiendo las preguntas planteadas:

- ✓ ¿Qué tipo de hipótesis se ha formado sobre los datos?
  La mayor cantidad de vehículos disponibles para venta son aquellos que su cilindrada sobrepasa los 2000cc, también que los factores que mas influyen en el precio de venta es el kilometraje y el año del vehículo, pues se encuentran registros del mismo vehículo, pero que por alguna variación en cualquiera de las opciones afecta directamente su valor.
- ✓ ¿Qué variables parecen prometedoras para un análisis más profundo?
  Seria interesante hacer un próximo acercamiento con variables tales como, variaciones del modelo, si tiene o no modificaciones, turbo, etc.
- ✓ ¿Sus exploraciones han revelado nuevas características sobre los datos?
  No
- ✓ ¿Cómo han cambiado estas exploraciones su hipótesis inicial?
  No
- ✓ ¿Considera que debería reformular el alcance del proyecto?
  No
- ✓ ¿Esta exploración ha alterado los objetivos?
  No
- ✓ ¿Puede identificar subconjuntos particulares de datos para su uso posterior?
  No

4. **Verificación de los datos.**

Después de realizar una exploración minuciosa de los datos, se logró comprobar cuales eran aquellos datos que no ofrecían información relevante.

- ✓ ¿Ha identificado variables faltantes y campos en blanco? Si es así, ¿Hay algún significado detrás de tales valores faltantes?
  R: Cuando se descargaron los datos y se inició el proceso de exploración se encontraron registros en los cuales en la variable "precio" decía "publicado", esto se debe a que el vendedor no describió el precio del vehículo en la etiqueta correspondiente, estos datos se eliminaron.
- ✓ ¿Hay inconsistencias ortográficas que puedan causar problemas en fusiones o transformaciones posteriores?
  R: Si, nosotros hicimos la labor de limpieza de datos por medio de Excel, entonces al cargar el archivo *.csv directamente a Excel, todas las palabras que contenían tilde (Automático, clásico, mecánico…), el carácter donde se situaba la tilde era cambiado automáticamente por "??", entonces fue necesario filtrar los datos para poder reemplazar esas palabras y que posteriormente no generaran conflicto.
- ✓ ¿Ha considerado excluir datos que no tienen impacto en sus hipótesis?
  R: Si, hemos considerado descartar los vehículos que en la variable kilometraje tienen valor cero, ya que su valor se compara con el de un concesionario.
- ✓ ¿Cada registro contiene el mismo número de campos?
  R: Si, después de aplicar toda la metodología y tener finalmente los datos limpios, cada registro poseen la misma cantidad de campos.