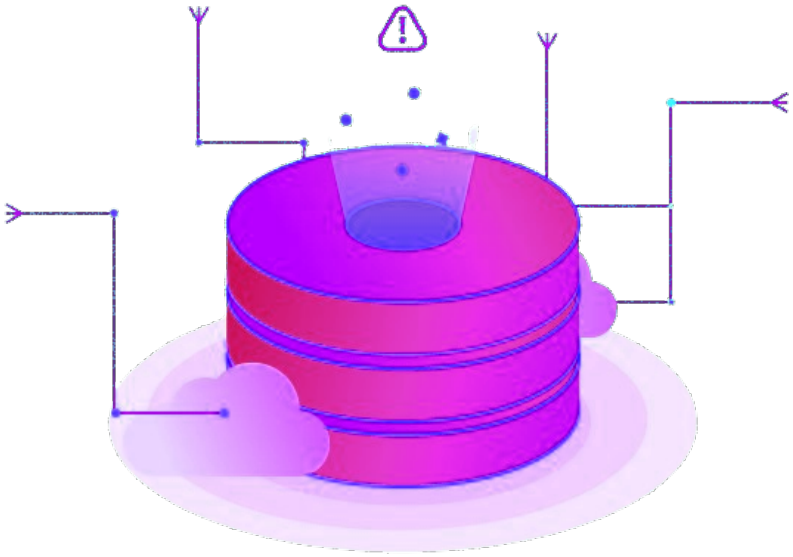
Universidad Nacional Autónoma De México
Facultad De Ciencias, 2024-I
Almacenes & Mineria de Datos



Práctica 02:

Limpieza de Datos.

Profesor:
Gerardo Avilés Rosas

Ayudante de teoría:
Gerardo Uriel Soto Miranda

Ayudante de laboratorio:
Ricardo Badillo Macías

Introducción.

Preparación de datos:

La preparación de datos es un paso obligatorio en el proceso de extracción de conocimiento. Trata de aumentar la calidad de los datos, de manera que se ajusten a algún proceso de minería de datos; sin la preparación adecuada, el algoritmo de minado no recibirá entradas adecuadas, pudiendo reportar errores antes, durante o después de su ejecución; en el mejor de los casos, el algoritmo funcionará pero es muy probable que los resultados arrojados no permitan obtener conclusiones válidas o el conocmiento generado carezca de precisión.

Las principales técnicas de esta etapa:

- **Limpieza de datos:** Permite rellenar valores perdidos, remover el ruido, resolver inconsistencia en los datos, identificar o eliminar valores atípicos.
- **Integración de datos:** Permite mezclar datos proveniente de múltiples funetes de datos heterogéneas en un repositorio coherente de datos.
- **Transformación de datos:** Permite escalar los datos y lograr que caigan en un rango pequeño (0.0 a 1.0).
- **Reducción de datos:** Permite reducir el tamaño de los datos(eliminando características redundante o agrupándolas), obteniendo una representación reducida en volumen que produce los mismo resultados analíticos (o similares).

Problemas con los datos:

Problema.	Ejemplo.
Falta de estándares	Diferentes formas de escribir nombre o fechas
Valores perdidos	No se ingresa el valor de atributos decisivos.
Información no consolidada	Múltiples identificadores de una misma entidad: un cliente con distintos códigos o identificadores.
Comparación compleja e integración	Entidades de negocio representadas de distintas formas dificulta el relacionar todas las instancias o integrar una representación única.
Datos fuera de dominio o que no cumplen con las reglas de negocio	Nombres comerciales mezclados con personales, uso inconsistente del espacio en blanco, caracteres especiales y límites de campos.
Errores en general	Errores tipográficos, faltas de ortografía, valores fuera de rango y tipos de datos incorrectos.
Homónimos	Su significado correcto depende del contexto.
Datos faltantes o invisibles	Datos con estructura y valor apropiados pueden omitir información.
Datos predeterminados	En ocaciones se ingresan valores especiales indicando que el campo tiene valor desconocido o que no se utiliza más.

Limpieza de Datos:

La **limpieza de datos** es la actividad que permite detectar y remover redundancia, errores e inconsistencias de los datos, para mejorar la calidad de los mismos.

Las tareas de la **limpieza de datos** se pueden dividir en : adquisición de datos y metadatos, reformateo, identificación de valores atípicos, suavizado de datos con ruido y correción de datos inconsistentes.

Reformatear datos:

Las tareas de reformateo de datos consisten en imputar valores perdidos, unificar los formatos de fecha y convertir valores nominales a numéricos.

Para realizar la imputación de datos los principales métodos son:

i. **Ignorar la tupla:** Se utiliza cuando la tupla contiene varios atributos con valores perdidos que es casi inexistente.

ii. **Llenar el valor manualmente:** Es un proceso ineficiente ya que dependemos de que tan grande sean los volúmenes de datos con valores perdidos, ademas de que se pueden cometer errores al rellenarlos.

iii. **Utilizar una constante global:** Se reemplazan los valores perdidos con un mismo valor constante.

iv. **Utilizar una medida de de tendencia central:** Se utiliza una medida que indique el valor "medio" de la distribución de datos. Para una distribución normal se puede utilizar la media. Si la distribución está sesgada se debe usar la mediana.

v. **Utilizar la media o mediana, para las tuplas que pertenezcan a la misma clase:** Se realiza un imputación en grupos de tuplas que comparten alguna característica.

vi. **Utilizar el valor más probable:** En este caso, el valor puede ser determinado con un análisis de regresión, o algún método de minería de datos.

El último método es la mejor opçión ya que utiliza la mayor cantidad de información de los datos para predecir el valor perdido, en cambio los métodos anteriores producen valores sesgados en los datos.

Convertir valores nominales a númericos

Varias tecnicas para minería de datos requieren trabajar sólo con atributos numéricos. En estos casos es necesario convertir los valores nominales a númericos. Las principales técnicas son:

i. **Valores binarios a numéricos:** En el caso de datos como genero, que manejan 2 valores, se puede convertir el atributo con valores 1 y 0.

ii. **Atributos ordenados a numéricos:** Podemos convertir los datos preservando el orden natural por ejemplo: las calificaciones MB -¿10; B -¿8; S -¿6; NA -¿5.

iii. **Nominales(pocos atributos):** Atributos multivaluados, no ordenados, con un número pequeño de valores (no más de 20 opciones). Para cada valor v, se puede crear un marcador binario en donde, si pertenece al valor es 1, y si no es 0.

iv. **Nominal(muchos valores):**Por ejemplo, códigos para los estados o para profesiones, se deben ignorar los campos ID, ya que estos valores son únicos para cada registro, y siempre que sea posible, formar grupos seleccionando los más frecuentes; se crean marcadores binarios para los valores seleccionados.

Suavizar datos con ruido:

El **ruido** es un error aleatorio o variación en una medida que se puede deber a: los instrumentos que se encargan de recolectar los datos, limitaciones tecnológicas o inconsistencia en convenciones de nomenclatura. Algunos métodos para suavizar los datos con ruido son: binning y regresión lineal.

Binning

Es un método que atenúa el ruido en un valor ordenado, consultando los valores alrededor de él. Los valores son distribuidos en "bins" realizando un suavizado local.

- **Igual ancho(distancia):** En este caso, se divide el rango de valores en N intervalos de igual tamaño. Se trata del método más directo; sin embargo, los valores atípicos pueden dominar la presentación y los datos asimétricos no se manejan bien. Si A y B son el menor y el mayor valor del atributo, el ancho del intervalo se calcula como $W = \frac{(B-A)}{N}$. Una vez que se tienen los bins generados, se puede sustituir cada valor por la media.

- **Igual profundidad(frecuencia):** En este caso, se divide el rango en N intervalos, cada uno conteniendo aproximadamente el mismo número de instancias. Se trata de un método que tiene un buen escalamiento de datos, pero administrar atributos categóricos puede ser complicado. Se puede sustituir cada valor, por la media o bien por el valor extremo del bin más cercano.

Regresión lineal:

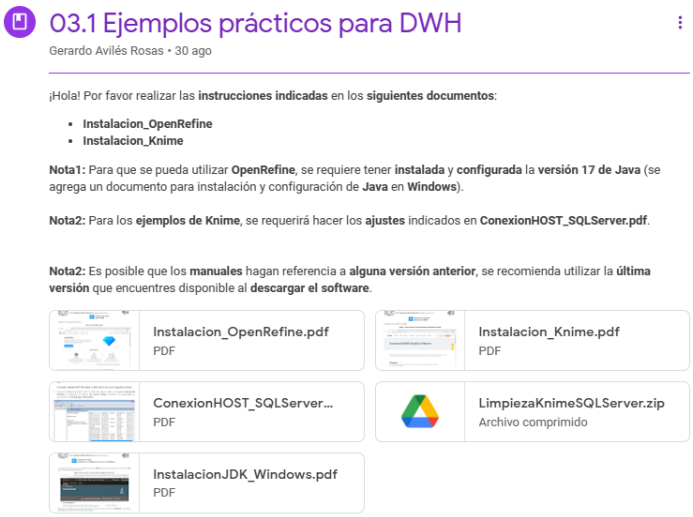
El análisis de **regresión lineal** es una técnica estadística que se utiliza para explorar y cuantificar la relación entre 2 variables, una llamada dependiente (Y) y otra llamda independiente o predictora (X), a partir de una ecuación lñineal. Esta técnica también puede ayudar a reducir el ruido: dado un conjunto de tuplas representado por dos variables, se puede encontrar la línea recta que mejor se ajuste a esos datos a través de una expresión matemática con la que se puedan reproducir los datos y eliminar el error en la medida de lo posible.

Tools:

- **OpenRefine**: Es una aplicación de escritorio de código abierto, para la limpieza y transformación de datos a otros formatos. Es similar a las aplicaciones de hoja de cálculoy puede manejar formatos de archivo como CSV, pero se comporta más como una base de datos.
  Enlace de descarga: `https://openrefine.org/download/`

- **KNIME**: Es una plataforma de minería de datos que permite el desarrollo de modelos en un entorno visual. Está construido bajo la plataforma Eclipse. **Enlace de descarga:** `https://www.knime.com/downloads/`

La explicación de la instalación junto a la configuración de ambos programas se encuentra en el classroom, en la parte de Almacenes de Datos.



Actividades.

Dadas los datasets movies.csv y laptopData.csv deberan hacer lo siguiente:

i. Deberan hacer limpieza de datos utilizando **openrefine** sobre movies.csv, y el dataset resultante lo llamaran **moviesClean.csv**.

ii. Deberan hacer limpieza de datos utilizando **KNIME** sobre laptopData.csv, en este caso deberan insertar en una base de datos de **SQL Server** llamada **laptop** la información despues de aplicar limpieza de datos. El proyecto generado por KNIME, deberan llamarlo **LimpiezaLaptop.knwf** y ademas deberan agregar los diccionarios que utilizaron para hacer los cambios. El archivo csv generado despues de la limpieza lo llamaran **laptopClean.csv**.

iii. Deberan crear un archivo PDF llamado **Practica02.pdf**, en donde incluyan el analisis que realizaron y las cosas que hicieron tanto para laptopData.csv como para movies.csv.

Figura 1: *Actividades.*

Entregables.

Deberán subir un archivo con formato *zip* a *Google Classroom*, de acuerdo a lo indicado en los lineamientos de entrega. Debe de estar organizado de la siguiente manera, (suponiendo que el nombre del equipo que está entregando es *Dream Team* y los integrantes del equipo son los profesores del curso).

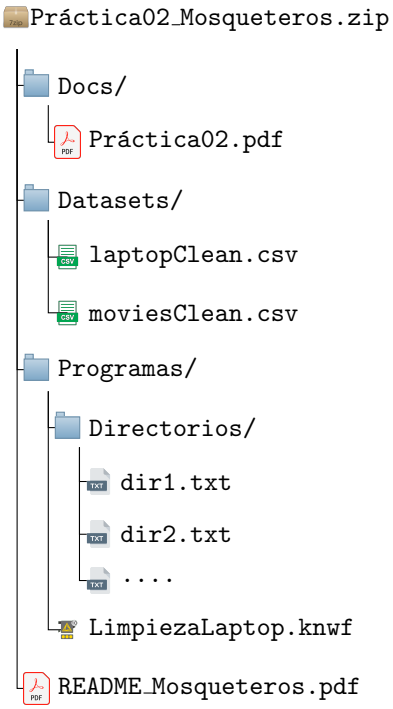


Figura 2: *Entregables.*

Nota.

Para cualquier duda o comentario que pudiera surgirles al hacer este trabajo, recuerden que cuentan con la asignación de este entregable en el grupo de *Classroom*, en donde seguramente encontrarás las respuestas que necesites.

Figura 3: *Nota.*