Tecnológico de Monterrey

ANALÍTICA DE DATOS Y HERRAMIENTAS DE INTELIGENCIA ARTIFICIAL II

GRUPO 101

ACTIVIDAD 2.2 VALORES NULOS - REPORTE

EQUIPO 10

Estefana Bermeo Severiano / A01367558
Miguel Saúl Fernández Ávalos / A01707491
Mariel Quetzali Fernández Montes / A0170779
Christian Jesús Soto-Vieyra Gil / A01707759

Abril 21, 2023

DataSet: Gastos y Costos 2022

Después de analizar con el método info el contenido de los datasets y sus columnas, se pudo observar que en todos los datasets existen valores nulos, por lo que será necesaria la limpieza o reemplazo de este tipo de datos.

En primer lugar se eliminarán las columnas conformadas en su totalidad por nas ya que esta composición implica que no existe información en estas columnas por lo que resultan irrelevantes para cualquier tipo de análisis. Después de este método en algunos datasets se redujeron las columnas.

También se eliminará la columna de status, ya que al analizar sus contenido, se tiene lo mismo, por lo que no es de utilidad para analizarla.

Después de realizar estos dos procedimientos con columnas, se considera que ya se contienen las columnas que pueden ser relevantes para analizar.

Ahora se eliminarán las instancias compuestas en su totalidad por nas ya que no contienen información y no tendría sentido tener instancias compuestas con valores reemplazados. Al parecer, no se tenían este tipo de filas.

Se eliminarán las instancias repetidas en folio, ya que esta columna debería contener valores únicos, por lo que si se repite en el dataset se tendría información repetida.

Continuando con el reemplazo de valores nulos, al analizar las columnas con datos nulos, se puede observar que éstas en su mayoría contienen información no numérica, por lo que se tendría que reemplazar por valores específicos.

Teniendo esto en cuenta, en el caso de columnas no numéricas se reemplazarán los valores nulos por la palabra “Checar” de modo que se pueda facilitar la visualización de instancias que requieran de un chequeo para ver si se tiene un error que se pueda solucionar. Por su parte, para las columnas con variables numéricas, se reemplazarán los valores por la mediana de la columna.

Dataset: Datos de facturación

El documento "Datos de facturación" cuenta con 3 columnas con datos nulos

- CVE_VEND
- FECHA_ENT
- FECHA_CANCELA

Para la sustitución de los datos nulos de la columna "CVE_VEND" sustituimos todos los espacios en blanco por un número "0" ya que ningún vendedor tiene esta clave y esto nos ayudaría a identificar rápidamente las facturas que no están asignadas a un vendedor.

Para la sustitución de la columna "FECHA_ENT" únicamente hay 2 valores nulos, por lo que ingresamos la frase "Sin fecha" y no fue sustituido por un valor numérico ya que la cantidad de valores nulos no tenían gran representatividad.

Por último, la columna "FECHA_CANCELA" si contaba con una gran cantidad de valores nulos, sin embargo, hace referencia a las facturas que no fueron canceladas, por lo que poner una fecha arbitrariamente dañaría el análisis de los datos, los valores nulos de esta columna fueron sustituidos por el texto “Sin cancelar”

DataSet: Detalle de Precios y Productos Fabricados 2022

En el análisis de valores nulos de este DataSet pudimos identificar que solo una columna contiene 2.

Al tratarse de una columna con valores string ya que corresponden a nombre de los vendedores, decidimos reemplazar estos dos valores por la leyenda “Sin registro” ya que no se conoce con certeza el nombre de la persona que realizó esa venta.

Por otra parte, decidimos no eliminar ninguna columna ya que consideramos que todas tienen importancia para el análisis del documento y no contienen datos repetidos.