Desafío Práctico DMD

Autor: BA181927 ,GM181938

2020

Universidad Don Bosco

Datawarehouse y Minería de Datos

Desafío práctico

Integrantes:

Bonilla Avilés, David Alejandro BA181927

Grande Menjivar, Willian Adonis GM181938

Catedrático: Ing. Alexander Siguenza

Soyapango, 04 de noviembre del 2020

Contenido

Autor: BA181927 ,GM181938

Parque vehicular

TIPO_PLACA	ANIO_DE_FA	CILINDRADA	CANTIDAD_D	CANTIDAD_D	VALOR_DEL_	COLORES	FECHA_DE_II	IMP_VALOR_	FECHA_DE_II	ANIO_INGRE	MES_INGRES	CLASE	F
PARTICULAR	1990	1800	0	4	4094.56	AMARILLO	16/9/1994	4094.56	17/11/1994	1994	11	AUTOMOVIL	F
PARTICULAR	1964	0	0	0	0	AMARILLO		0	16/2/1989	1989	2	AUTOMOVIL	F
ALQUILER	1984	1700	0	0	0	AMARILLO F/C B/N		0	20/3/1985	1985	3	ALQUILER	F
ALQUILER	1986	1600	0	0	0	AMARILLO F/C B/N		0	18/2/1988	1988	2	ALQUILER	F
ALQUILER	1979	0	0	0	0	AMARILLO		0	26/10/1982	1982	10	ALQUILER	F
PARTICULAR	1974	1600	0	0	0	AMARILLO		0	26/10/1982	1982	10	AUTOMOVIL	F
ALQUILER	1975	0	0	0	0	AMARILLO		0	23/5/1984	1984	5	ALQUILER	F
ALQUILER	1973	0	0	0	0	AMARILLO F/C B/N		0	3/6/1988	1988	6	ALQUILER	F
ALQUILER	1975	0	0	0	0	AMARILLO		0	4/2/1988	1988	2	ALQUILER	F
ALQUILER	1968	0	0	0	800	AMARILLO	29/7/1983	0	29/8/1983	1983	8	ALQUILER	F
ALQUILER	1977	0	0	0	0	AMARILLO		0	12/6/1984	1984	6	ALQUILER	F
ALQUILER	1965	0	0	0	0	AMARILLO		0	26/10/1982	1982	10	ALQUILER	F
ALQUILER	1983	1587	0	4	4183.76	AMARILLO F/	24/2/1995	4183.76	24/3/1995	1995	3	ALQUILER	F
ALQUILER	1978	0	0	0	0	AMARILLO		0	12/4/1989	1989	4	ALQUILER	F
PARTICULAR	1981	1770	0	0	4413.71	AMARILLO	20/9/1993	4413.71	27/10/1993	1993	10	AUTOMOVIL	F
ALQUILER	1967	0	0	0	0	AMARILLO		0	26/10/1982	1982	10	ALQUILER	F

Ilustración 1: Vista parcial de la información

Para poder dar solución a los requerimientos planteados , fue necesario analizar el archivo csv que contiene la información de los automóviles en el parque vehicular al 13 de noviembre del 201. Se opto por aplicar las estrategias de minería de datos: Agrupamiento con K-means y Árbol de decisiones.

Aplicando la estrategia árbol de decisión

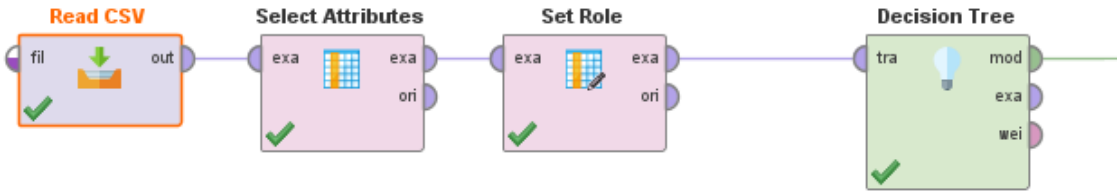


Ilustración 2: Vista general del flujo de trabajo

Para aplicar un árbol de decisión, se debe de leer el archivo csv, para ello se ocupa el operador de lectura donde se activa la opción para que marque como datos perdidos aquellos registros que no puede leer, esto se realiza porque en el millón y medio de registros, aproximadamente 15 se han dañado y no son aptos para el procesamiento.

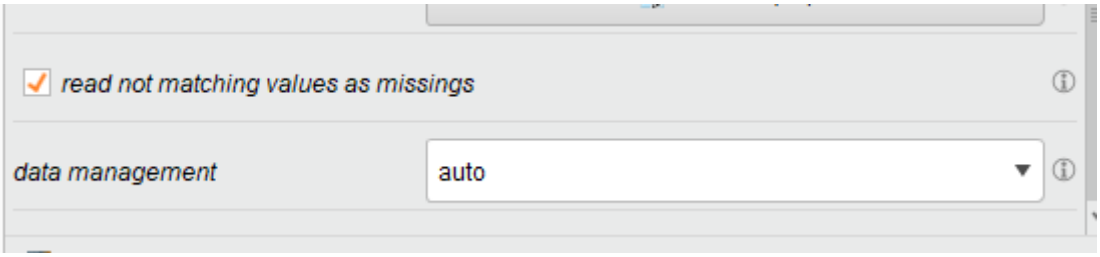


Ilustración 3: Opciones de lectura

Autor: BA181927 ,GM181938

En el operador de selección de atributos se seleccionaron aquellos que fueran descriptivos con la información brindada, pero que a su vez no genere un árbol de decisión demasiado grande. Los árboles de decisión con demasiadas partes pueden resultar incomprensibles.

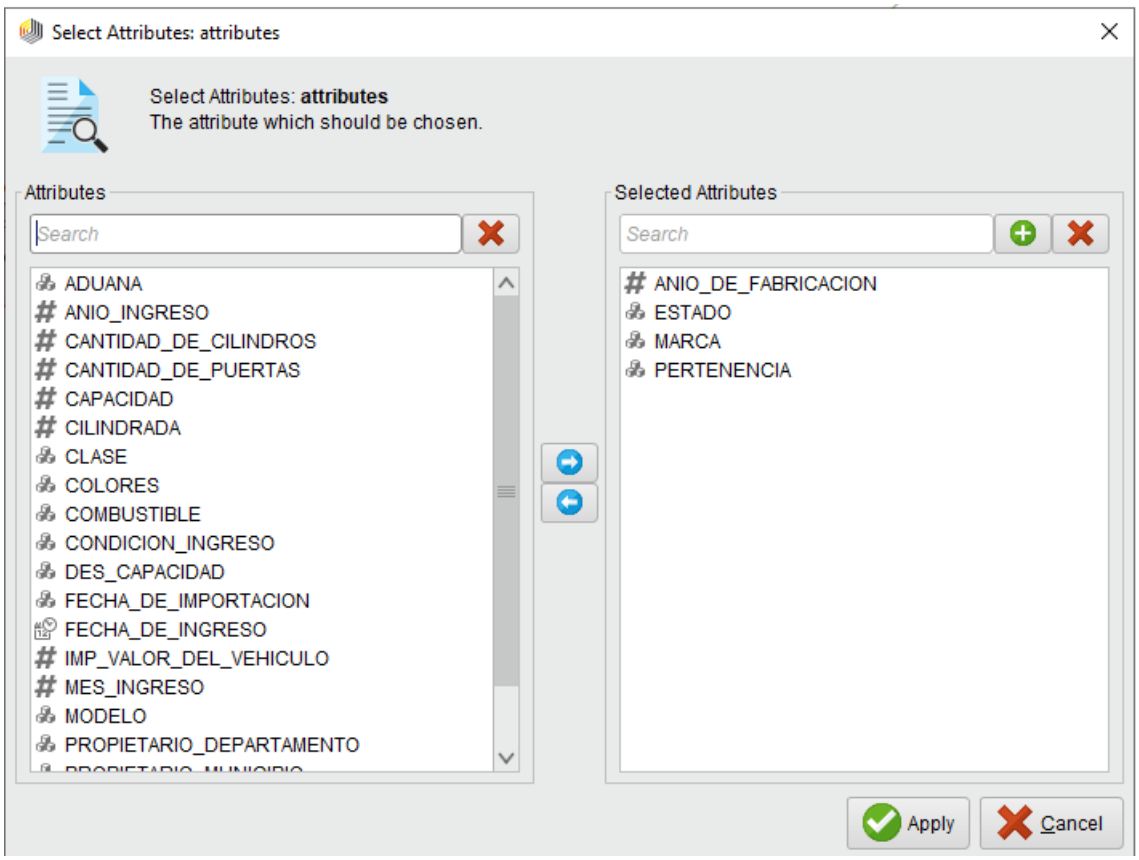


Ilustración 4: Configuración en la selección de atributos

Se le asigna el rol de label al atributo de pertenencia.

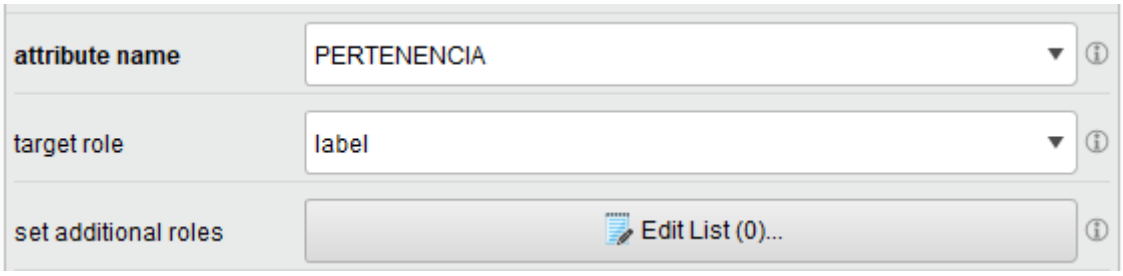


Ilustración 5 : Asignación de rol a un atributo

Autor: BA181927 ,GM181938

El árbol resultante es el siguiente:

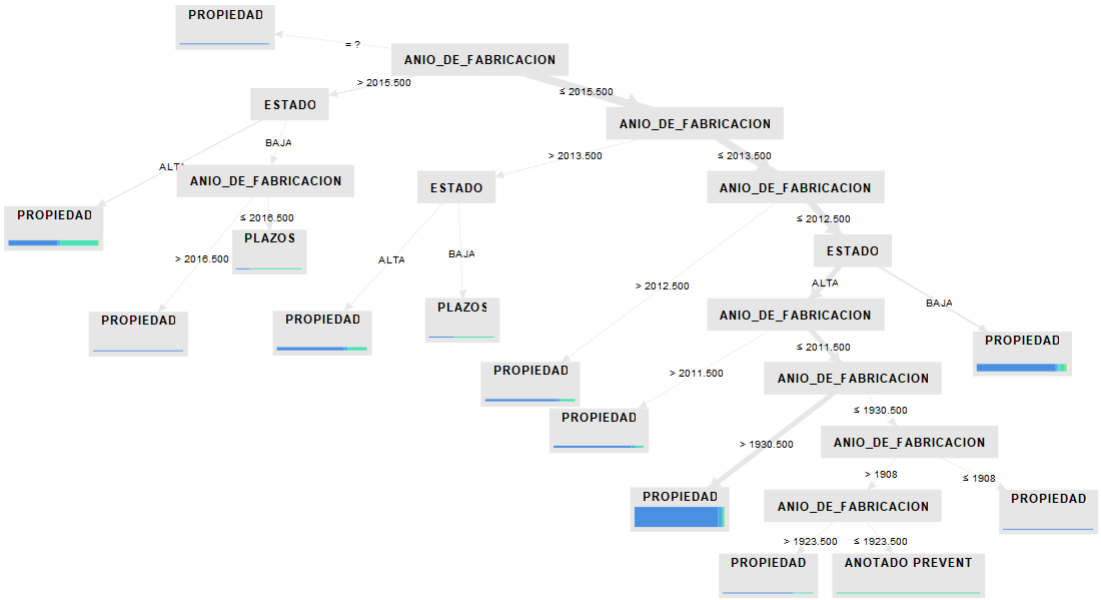


Ilustración 6: Descripción grafica de datos, toma en cuenta el año de fabricación, el estado del mismo (alta,baja) y la marca, para poder describir el estado de pertenencia del vehículo (Propiedad o estado preventa).

Aplicando la estrategia agrupamiento K-Means

En este caso para aplicar de manera correcta el agrupamiento K Means, se optó por utilizar el software de minería de datos “KNIME Analytics Platform”, ya que se necesita la capacidad de procesar grandes cantidades de datos y a su vez, personalizar el modo de lectura de los mismos.

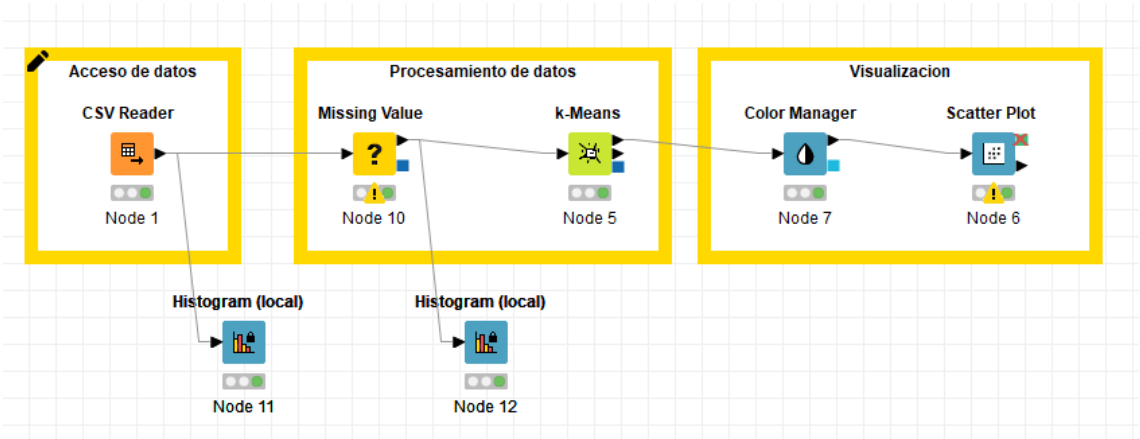


Ilustración 7: Vista general del flujo de trabajo

Como se describe en la ilustración 7, los flujos de trabajo dentro de KNIME se pueden clasificar en tres partes, el acceso a datos, el procesamiento de los mismos y la visualización.

Autor: BA181927 ,GM181938

Dentro de las configuraciones del lector CSV, se especificó que no existe con un carácter de comas definido, y que nuestro archivo no cuenta con un encabezado de filas, es decir no cuenta con un id definido cada registro.

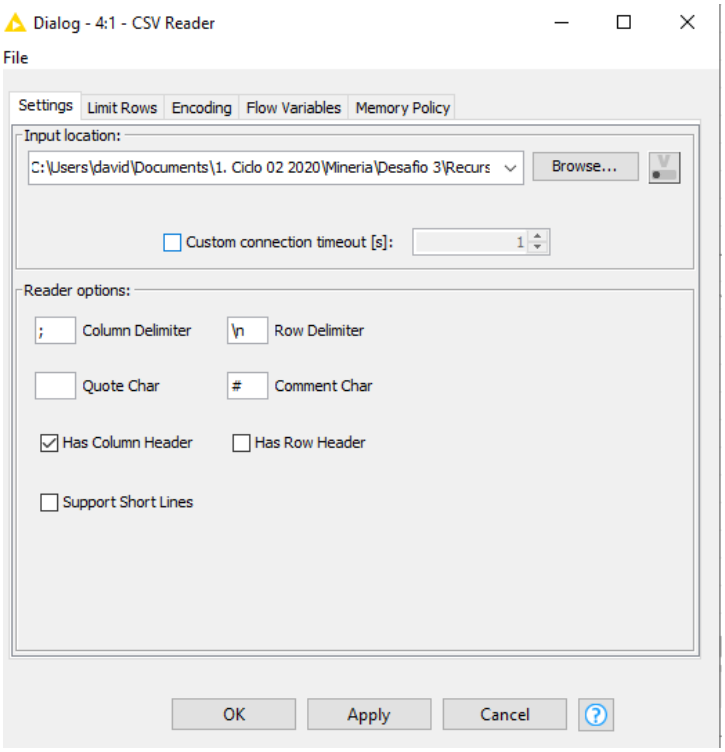


Ilustración 8: Configuración del lector CSV

Se pueden observar dentro del flujo de trabajo nodos de visualización “Histogramas”, estos nos ayudan para poder ver de manera gráfica los diferentes datos de entrada, El nodo 11 nos permite ver de manera gráfica los datos después de ser leídos directamente del archivo csv. Se pueden observar que existen datos perdidos, celdas sin registros en ciertas filas de los datos. Esto representa un problema ya que el procesamiento por agrupación de datos K Means no puede manejar datos perdidos. Por esto mismo se agrega un nodo nombrado “Missing Value”.

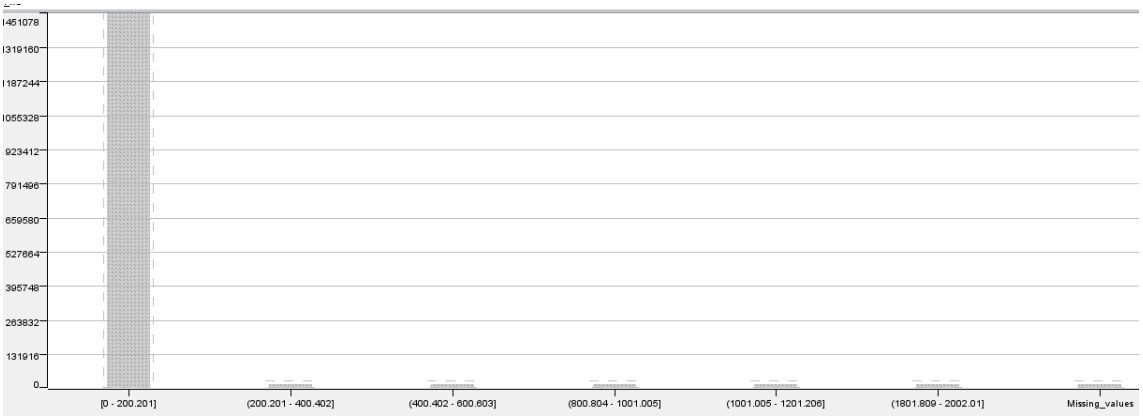


Ilustración 9: Vista del histograma del nodo 11

El nodo “Missing Value” permite tratar los datos perdidos , para este caso en especifico se optó por eliminar todas aquellas filas que contengan datos perdidos.

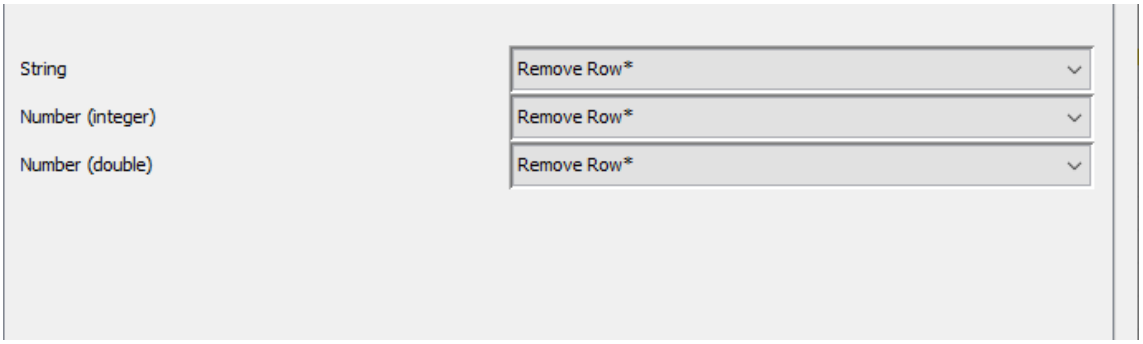


Ilustración 10: Configuración del nodo Missing Value.

Si se observa los resultados dentro del histograma en el nodo 12, se observa que algunos datos han sido removidos, así como los datos perdidos, ya que se removieron todos las filas que contengan datos perdidos.

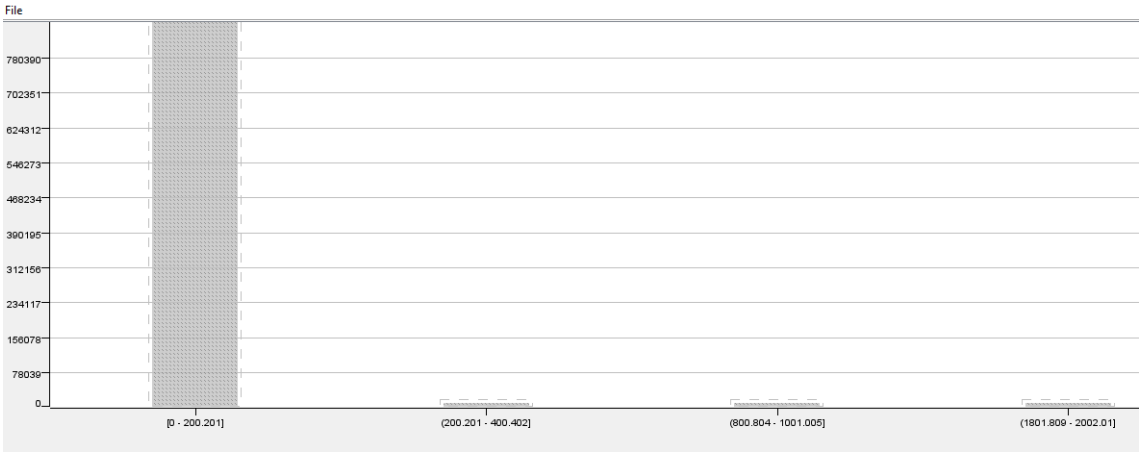


Ilustración 11: Vista del histograma del nodo 12

Dentro del nodo donde se aplica el agrupamiento K Means se realiza la siguiente configuración:

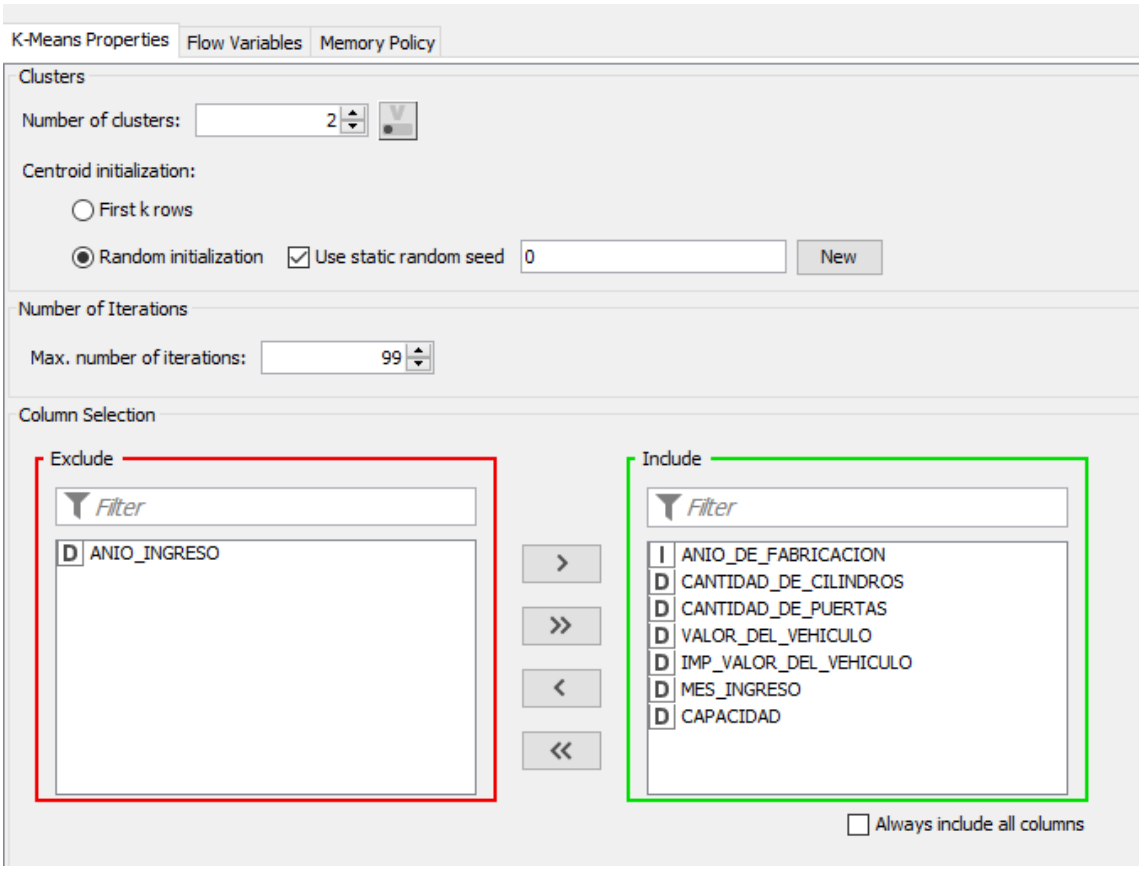


Ilustración 12: Propiedades nodo K Means, se observan las columnas que se han incluido para aplicar el agrupamiento K Means.

Con el nodo de Color manager se asigna un color diferente para cada cluster, y el nodo 6 Scatter Plot nos *permite* observar los resultados del agrupamiento.

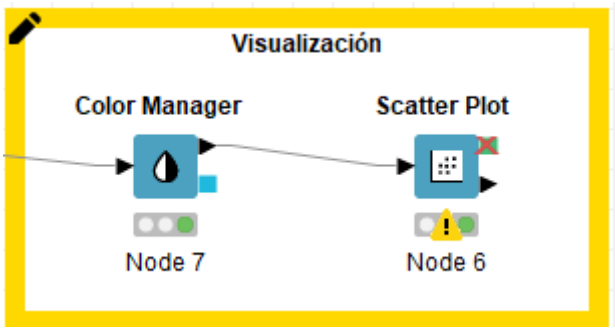


Ilustración 13: Visualización de datos

Dentro de los resultados se pueden observar dos clusters bastante marcados.

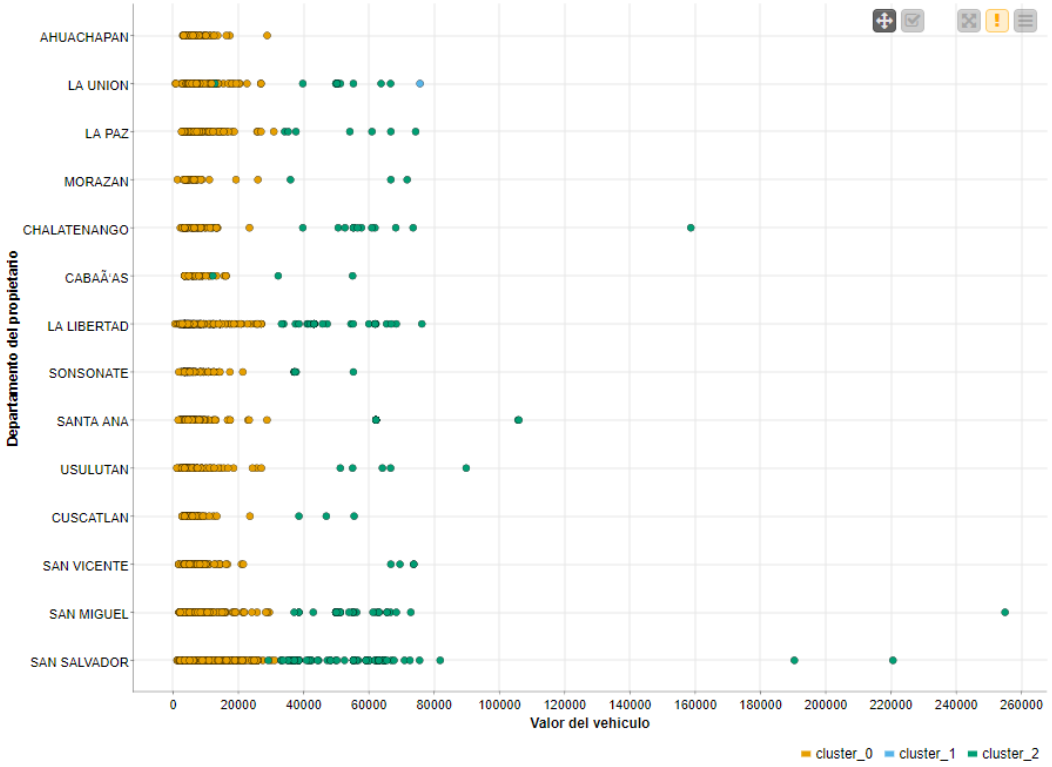


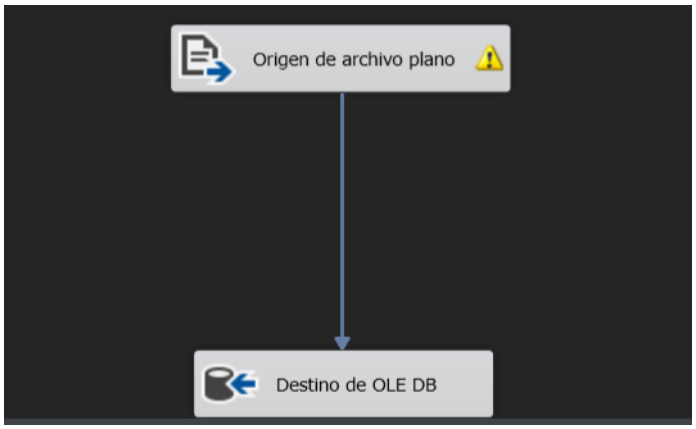
Ilustración 14: Vista de los resultados del procesamiento K Means

Procedimiento alcanzado: 100%

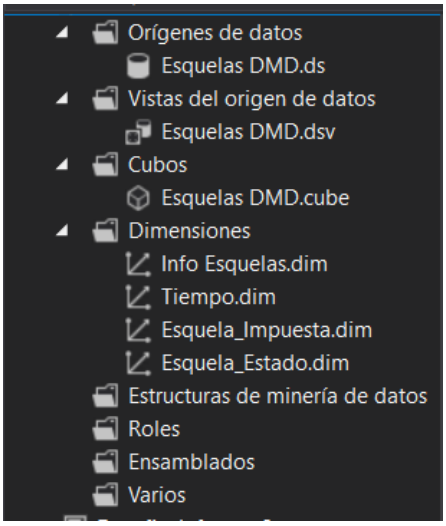
Esquelas de infracción de tránsito

Procedimiento alcanzado: 100%

Las 2 técnicas de minería de datos que se utilizaron en este desafío fueron los cubos OLAP y el reporting services, por lo que en primer lugar se realizó un ETL sencillo en el que simplemente se necesitaba transferir los datos a una base de datos, tal como se muestra en la siguiente imagen. (Se respetaron los tipos de datos para cada campo, para evitar errores en el proceso ETL).

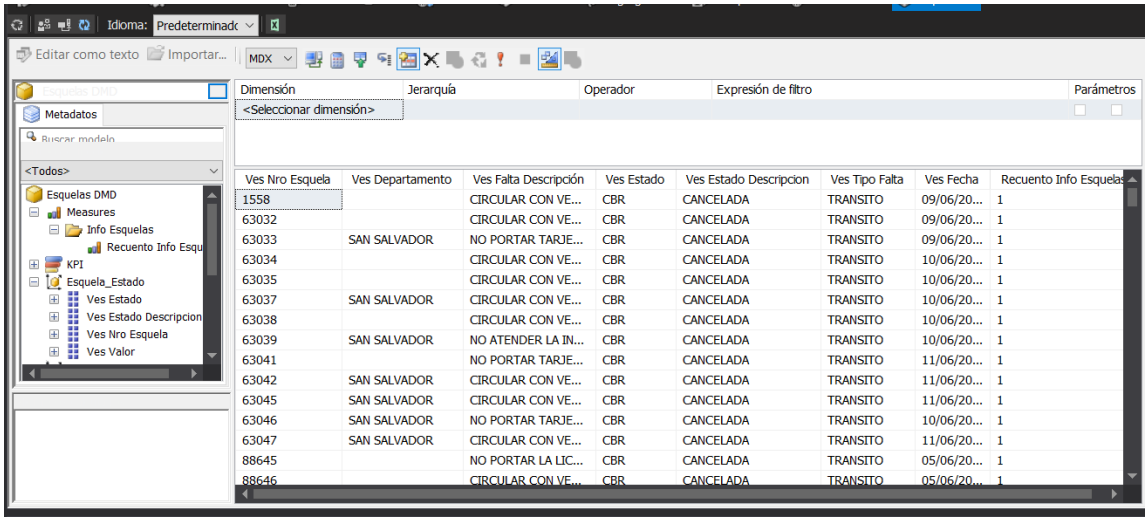


De esta manera, se realizó por medio de la herramienta de Analysis services un cubo con sus respectivas dimensiones, tal como se muestra a continuación:



Por lo que se creó un origen de datos para la información almacenada anteriormente con el proceso de ETL, de igual manera se creó una vista para tener una vista amplia de los datos y poder explorar los datos, en este caso no fue necesario crear un calculo generado, pero pudo haber sido una opción.

Con respecto a las dimensiones creadas se generaron 3, en este caso la de tiempo que almacena el ID de la esquela y la fecha en la que se impuso, luego la dimensión esquela impuesta, en la que se almacena todo con respecto al porque se impuso y el tipo de esquela; y por ultimo la dimensión Esquela estado en la cual se muestra el estado y el valor de la esquela, junto con el departamento en el que se impuso. El cubo quedó de la siguiente manera al poner todos los elementos:



Ves Nro Esquela	Ves Departamento	Ves Falta Descripción	Ves Estado	Ves Estado Descripcion	Ves Tipo Falta	Ves Fecha	Recuento Info Esquelas
1558		CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	09/06/20...	1
63032		CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	09/06/20...	1
63033	SAN SALVADOR	NO PORTAR TARJE...	CBR	CANCELADA	TRANSITO	09/06/20...	1
63034		CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	10/06/20...	1
63035		CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	10/06/20...	1
63037	SAN SALVADOR	CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	10/06/20...	1
63038		CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	10/06/20...	1
63039	SAN SALVADOR	NO ATENDER LA IN...	CBR	CANCELADA	TRANSITO	10/06/20...	1
63041		NO PORTAR TARJE...	CBR	CANCELADA	TRANSITO	11/06/20...	1
63042	SAN SALVADOR	CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	11/06/20...	1
63045	SAN SALVADOR	CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	11/06/20...	1
63046	SAN SALVADOR	NO PORTAR TARJE...	CBR	CANCELADA	TRANSITO	10/06/20...	1
63047	SAN SALVADOR	CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	11/06/20...	1
88645		NO PORTAR LA LIC...	CBR	CANCELADA	TRANSITO	05/06/20...	1
88646		CIRCULAR CON VE...	CBR	CANCELADA	TRANSITO	05/06/20...	1

Y de esta manera ya se podrían hacer los reportes, en este caso se hicieron 2 reportes, el primero sobre la esquela, su descripción y la fecha en la que se colocó para así poder revisar, cuales son las esquelas que más se colocaron en las distintas fechas; y en el segundo reporte se muestra el Estado, la descripción, el departamento y la descripción de como fue la falta de dicha esquela, tal como se muestra en las siguientes imágenes:

Reporte de esquelas

Ves Nro Esquela	Ves Fecha	Ves Falta Descripción	Recuento Info Esquelas
1558	09/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63032	09/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63033	09/06/2003 00:00	NO PORTAR TARJETA DE CIRCULACION	1
63034	10/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63035	10/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63037	10/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63038	10/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63039	10/06/2003 00:00	NO ATENDER LA INDICACION DE PARADA DEL AGENTE DE LA POLICIA NACIONAL CIVIL	1
63041	11/06/2003 00:00	NO PORTAR TARJETA DE CIRCULACION	1
63042	11/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63045	11/06/2003 00:00	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63046	10/06/2003 00:00	NO PORTAR TARJETA DE CIRCULACION	1

Reporte del Estado de la esquela

Ves Nro Esquela	Ves Estado	Ves Estado Descripcion	Ves Departamento	Ves Falta Descripción	Recuento Info Esquelas
1558	CBR	CANCELADA		CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63032	CBR	CANCELADA		CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63033	CBR	CANCELADA	SAN SALVADOR	NO PORTAR TARJETA DE CIRCULACION	1
63034	CBR	CANCELADA		CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63035	CBR	CANCELADA		CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63037	CBR	CANCELADA	SAN SALVADOR	CIRCULAR CON VEHICULOS NO MATRICULADOS	1
63038	CBR	CANCELADA		CIRCULAR CON VEHICULOS	1