Universidad Don Bosco

Desafío Práctico DMD

**Autor: BA181927 ,GM181938**

**2020**



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

[Parque vehicular 4](#_Toc55863955)

[Aplicando la estrategia árbol de decisión 4](#_Toc55863956)

[Aplicando la estrategia agrupamiento K-Means 6](#_Toc55863957)

[Esquelas de infracción de tránsito 10](#_Toc55863958)

# Parque vehicular

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.

Ilustración 1: Vista parcial de la información

### 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.

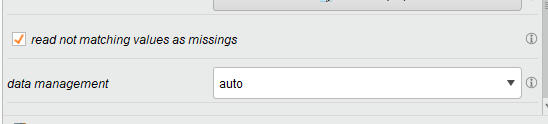


Ilustración 3: Opciones de lectura

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.

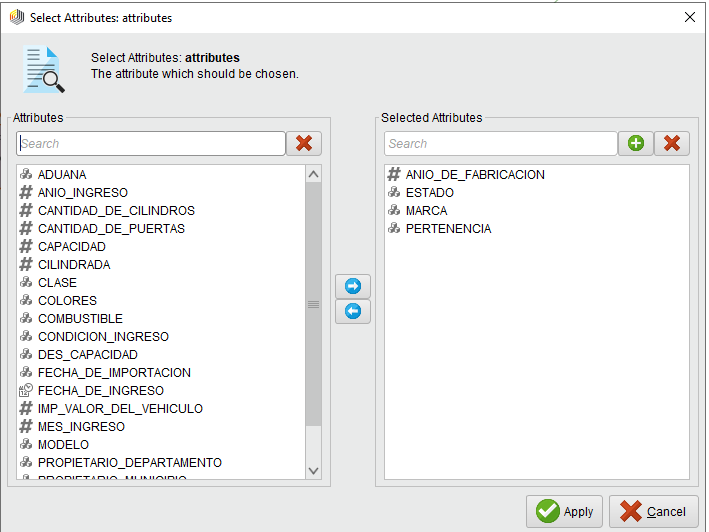


Ilustración 4: Configuración en la selección de atributos

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

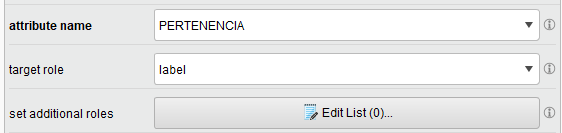


Ilustración 5 : Asignación de rol a un atributo

El árbol resultante es el siguiente:

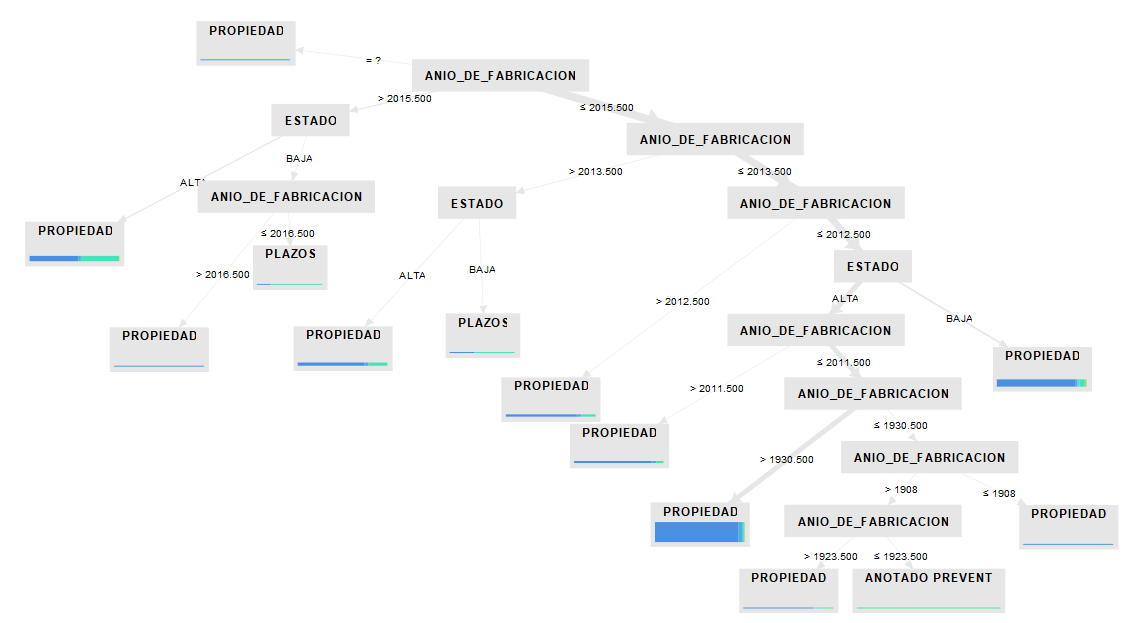


Ilustración 6: Descripción grafica de datos, toma en cuenta el año de fabricación, el estado del mismo (alta,baja) y la marca, para poder describir el estado de pertenencia del vehículo (Propiedad o estado preventa).

### 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.

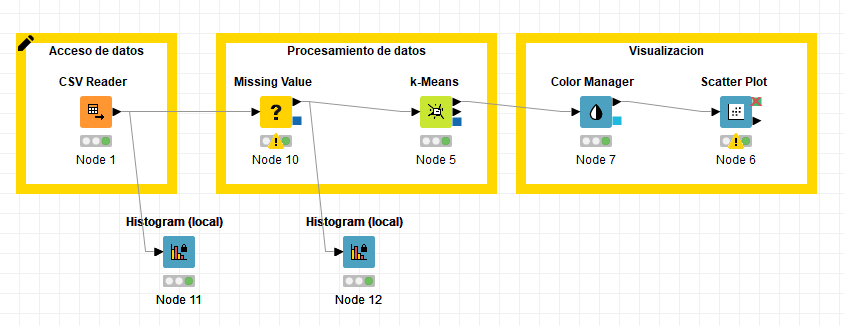


Ilustración 7: Vista general del flujo de trabajo

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.

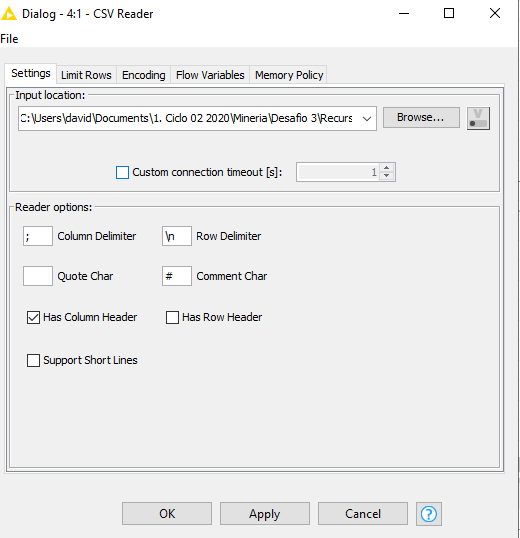
Dentro de las configuraciones del lector CSV, se especificó que no existe con un carácter de comas definido, y que nuestro archivo no cuenta con un encabezado de filas, es decir no cuenta con un id definido cada registro.

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”.

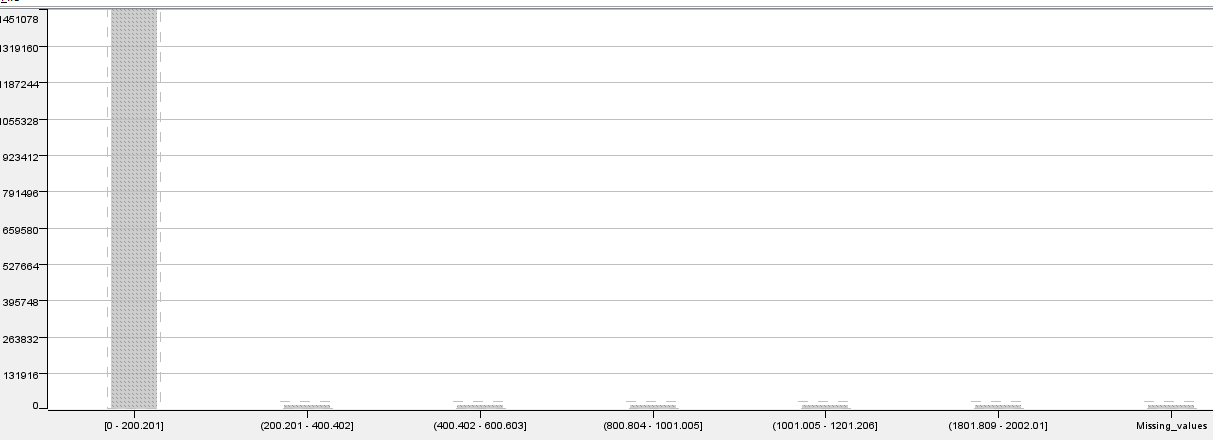


Ilustración 9: Vista del histograma del nodo 11

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.

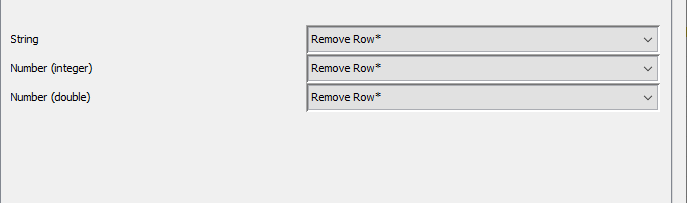


Ilustración 10: Configuración del nodo Missing Value.

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.

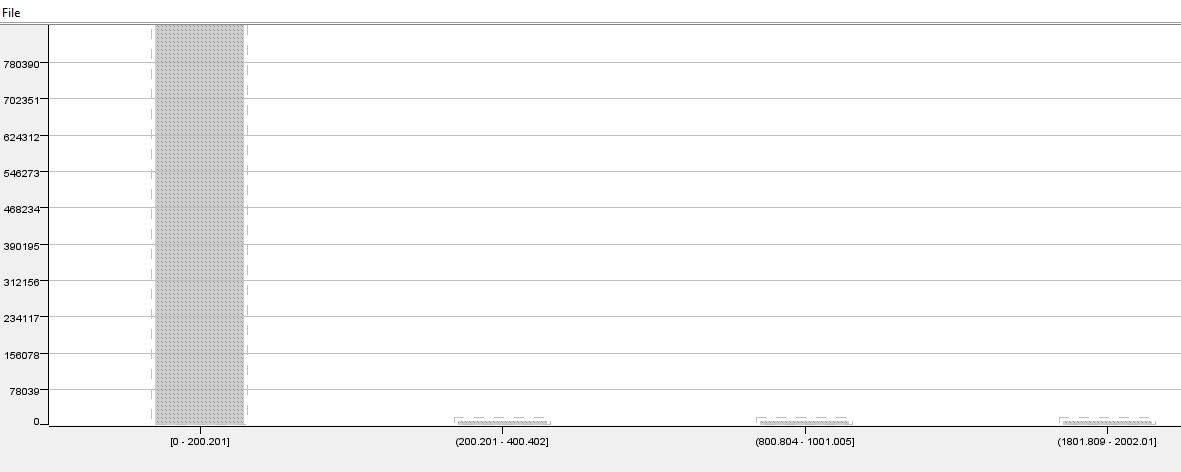


Ilustración 11: Vista del histograma del nodo 12

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

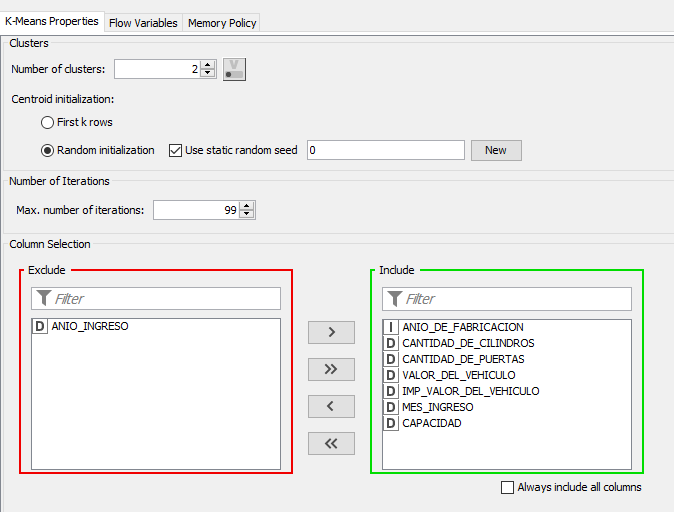


Ilustración 12: Propiedades nodo K Means, se observan las columnas que se han incluido para aplicar el agrupamiento K Means.

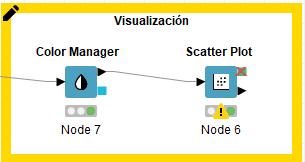
Con el nodo de Color manager se asigna un color diferente para cada cluster, y el nodo 6 Scatter Plot nos permite observar los resultados del agrupamiento.

Ilustración 13: Visualización de datos

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

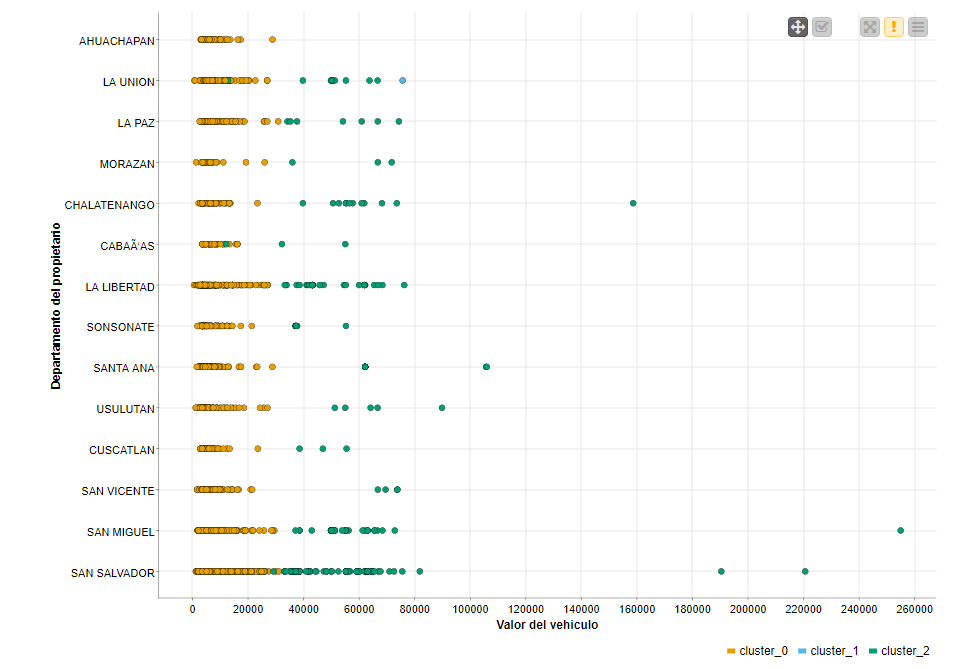


Ilustración 14: Vista de los resultados del procesamiento K Means

**Procedimiento alcanzado: 100%**

# Esquelas de infracción de tránsito

**Procedimiento alcanzado: 100%**