UNIVERSIDAD DE CORDOBA

FACULTAD DE INGENIERIA

INGENIERIA DE SISTEMAS

TALLER DE CLUSTERING CON KNIME

ELECTIVA APRENDIZAJE AUTOMATICO

INTEGRANTES:

DEIMER YAMIT HERNANDEZ DE LA ROSA

DEYSER OROZCO YEPES

PROFESOR:

OSWALDO VELEZ LANGS

2020

**Desarrollo**

**1.Comprensión de los Datos**.

Identificar la estructura de los datos y su composición.

|  |  |  |
| --- | --- | --- |
| variable | descripción | tipo de dato |
| curso | curso | String |
| dia\_semana | dia\_semana | Number |
| solo\_hora | solo\_hora | Number |
| programa | programa | String |
| genero | genero | String |

**3. Compresión Datos:** Implementar un nodo “Statistics” y realizar una descripción Estadística de los datos.

**Atributos numéricos**

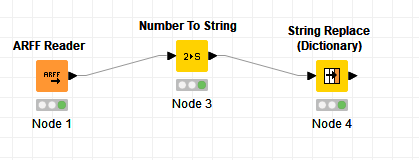
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| atributo | min | Mean | Max | Std. Dev. | varianza | histograma |
| dia\_semana | 1 | 3,5159 | 7 | 1,9282 | 3.71 |  |
| solo\_hora | 0 | 14,8368 | 23 | 5,2159 | 27.2 |  |

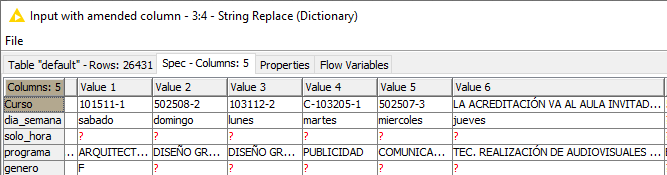
**Atributos nominales**

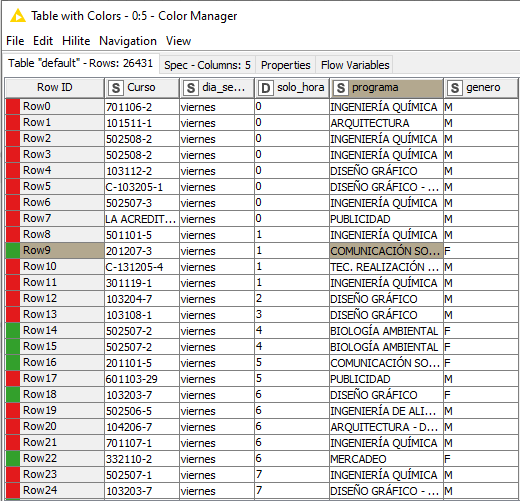
|  |  |
| --- | --- |
| Atributo | Histograma |
| curso |  |
| programa |  |
| genero |  |

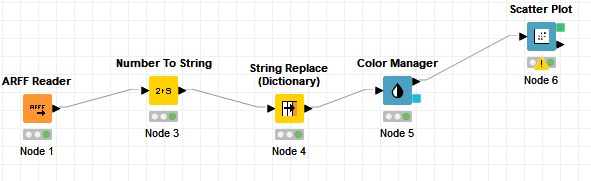
**4. Preparación de los datos:** Reemplazar los valores de la variable *Dia\_Seman,* por sus correspondientes nombres (etiquetas) de acuerdo al día de la semana. Implementar nodo “String replace (Dictionary)”. En la configuración del nodo para el reemplazo de palabras, utilizar el archivo “DictionaryDays.txt”

**R/** Para la preparación de los datos, se utilizó el nodo “Number To String”, para poder convertir el atributo de tipo Numérico *Dia\_semana* A tipo Nominal y de esta manera poder usar el Nodo “String Replace (Dictionary) ”



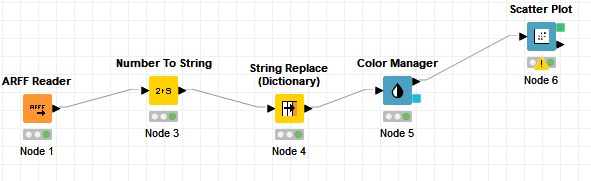


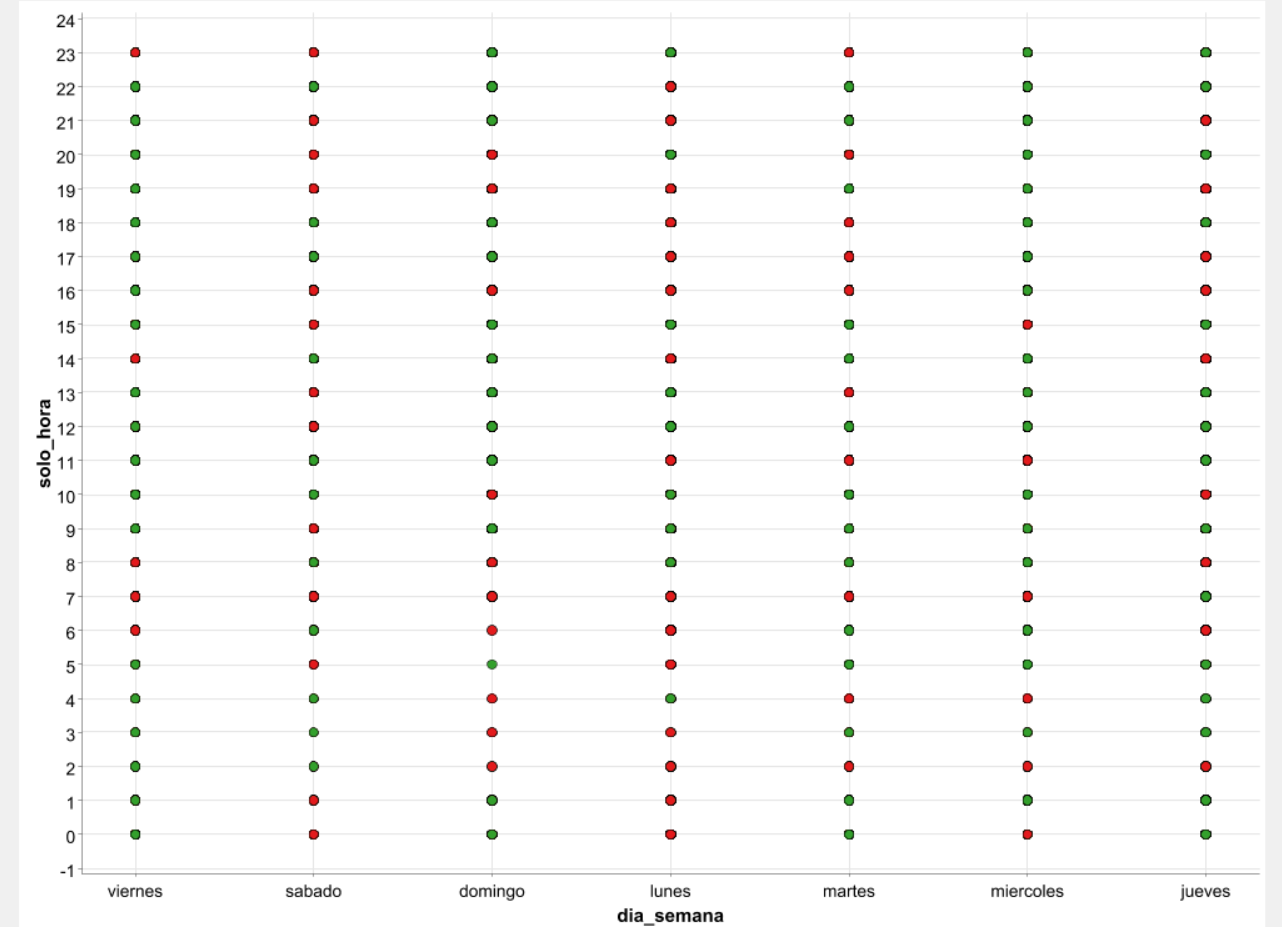
**5. Preparación de los datos.** Implementar nodo “Color Manager” para asignar “Etiquetas de color” a los registros de la base. 



**6. Análisis.** Implementar un nodo “Color manager” y conectarlo a un nodo “ScatterPlot” y reportar hallazgos relevantes en cuanto al comportamiento de los datos en su estado natural. Los datos salida del nodo “color manager” serán la entrada del nodo “Normalizer”.

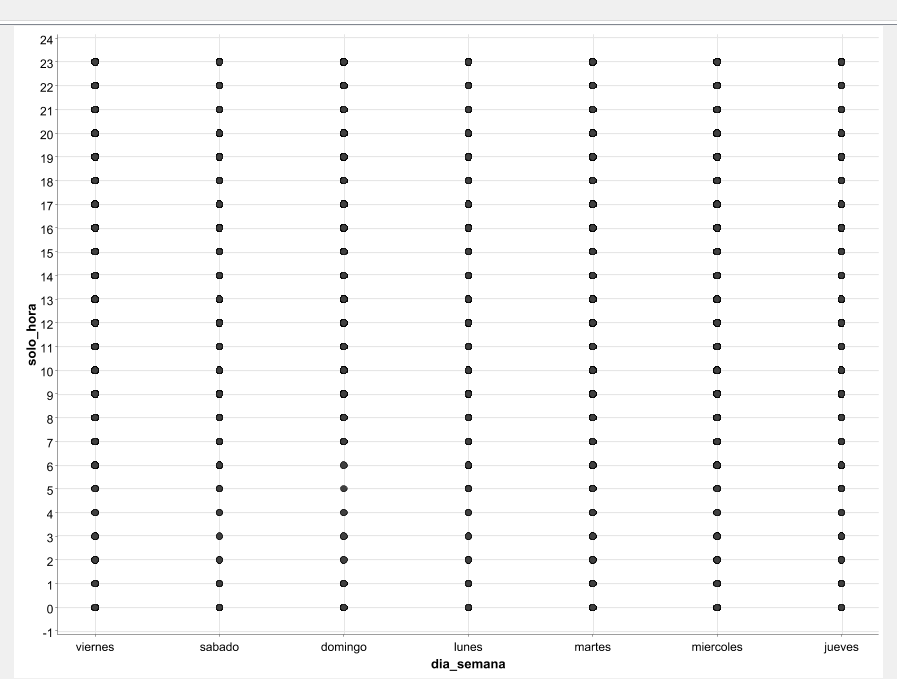
***Color manager aplicado a la columna “genero”***





Del análisis podemos concluir a simple vista que este nodo nos permite ver de manera clara datos relevantes en los datos de entrada, como puede ser la jerarquía de géneros que hay en un día en especifico a una hora en especifica. Como un ejemplo podemos ver que los sábados entre las 15hrs y las 22hrs la jerarquía de género es de Mujeres

***Metodo Scatter Plot desconectado del nodo Color Manager***



**7. Modelado.**

**7.1:** El algoritmo *“K-means”* No recibe datos nominales, por esta razón trabajamos solo con los datos de la columna “solo\_horas” que es la única después de normalizar y aplicar el nodo *“String Replace ( Dictionary )”*que sigue siendo numérica, Mientras que el algoritmo *“EM”* si recibe toda clase de atributos o tipos de datos

**7.2:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Prueba | MaxIterations | maximumNumberOfClusters | minLogLikelihoodImprovementCV: | minStdDev: | numClusters | numExecutionSlots | numFolds | Seed | Log likelihood: |
| 1 | 150 | 10 |  | 5,2159 | 3 | 8 | 10 | 25 | -14,88976 |
| 2 | 175 | 15 | -14,88976 | 5,2159 | 6 | 8 | 15 | 32 | -14,48299 |
| 3 | 200 | 20 | -14,48299 | 5,2159 | 9 | 8 | 20 | 39 | -14,48379 |
| 4 | 225 | 25 | -14,48379 | 5,2159 | 12 | 8 | 25 | 46 | -14,38166 |
| 5 | 250 | 30 | -14,38166 | 5,2159 | 15 | 8 | 30 | 53 | -14,37160 |
| 6 | 275 | 35 | -14,37160 | 5,2159 | 18 | 8 | 35 | 60 | -14,25359 |
| 7 | 300 | 40 | -14,25359 | 5,2159 | 21 | 8 | 40 | 67 | -14,24214 |

**GRAFICOS DE CLUSTERS:**

**Prueba 1:**

**Prueba 2:**

**Prueba 3:**

**Prueba 4:**

**Prueba 5:**

**Prueba 6:**

**Prueba 7:**

Uno de los análisis claros al realizar las pruebas y sacar las gráficas de cada prueba resultante, es que a mayor clusters mas exactitud en los datos se encuentra, ya que pues a medida que los parámetros van creciendo, algunos cluster simplemente dan 1 o ya no influyen en el valor de los resultados, En este caso puede decirse que un perfil de cluster basándome en las graficas antes expuestas vendría siendo uno donde la cantidad de iteraciones, y numClusters sea muy grande, además de esto que el numFold también sea bastante alto, con el fin de obtener datos más precisos con los que se puedan calcular datos a simple vista con graficas como es el caso, un ejemplo que demuestra mi análisis es tan solo ver las graficas de las pruebas 1 y 7 la diferencia en la grafica de la columna *“programa”* es abismal, ya que la cantidad de cluster, la grafica *“programa”* de la pureba 1 solo apila datos y no sabemos de que manera se pueden diferenciar, cosa diferente pasa con esta misma grafica pero en la prueba 7, ya que la cantidad de clusters nos brinda series de datos con los que podemos clasificar grupos más rápido y de ahí poder sacar una información más valiosa para dar un mejor uso de los datos