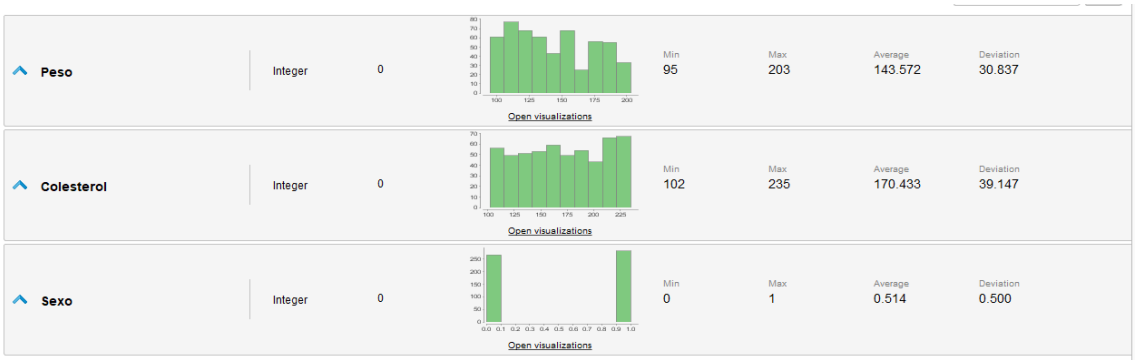
UT5 TA 1

El data set no tiene outliers, la variable sexo toma los valores 0 y 1 por tanto se considera pertinente evaluar su inclusión en el modelo o no. En caso de incluirla se debería considerar como una variable numérica lo que podría tener algunas connotaciones en cómo se interpretan los valores 0 y 1.



Analizando los atributos se observa que ni peso ni colesterol tienen distribuciones normales, el peso tiene un rango de 95 a 203 libras y el colesterol de 102 a 235 mg/dl.

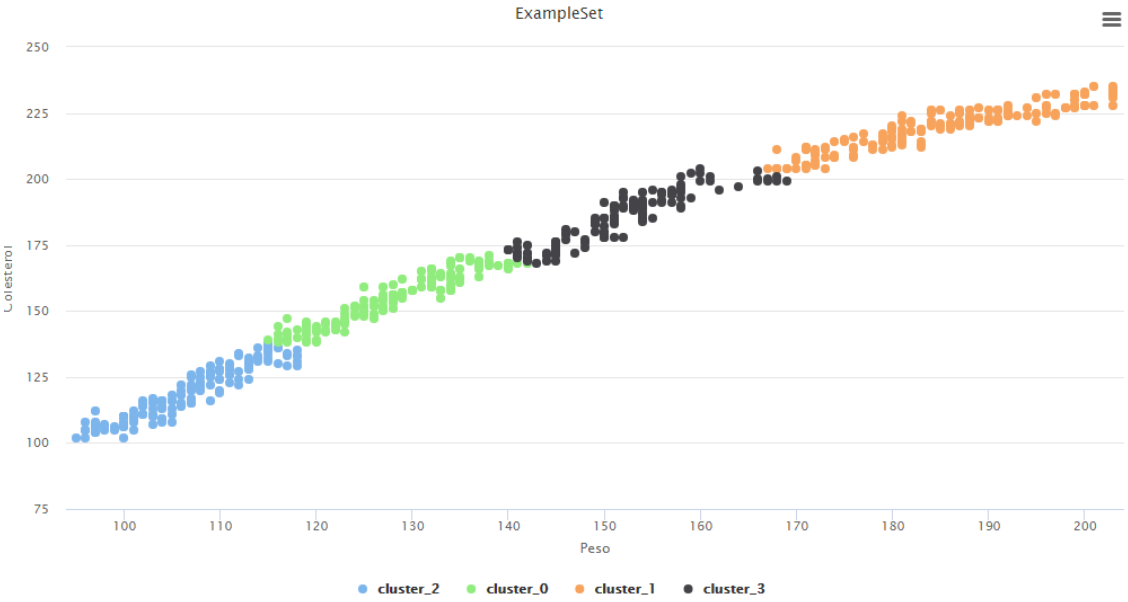
También dadas las variaciones en los datos, se considera pertinente normalizar los datos.

Modelado

Cluster Model

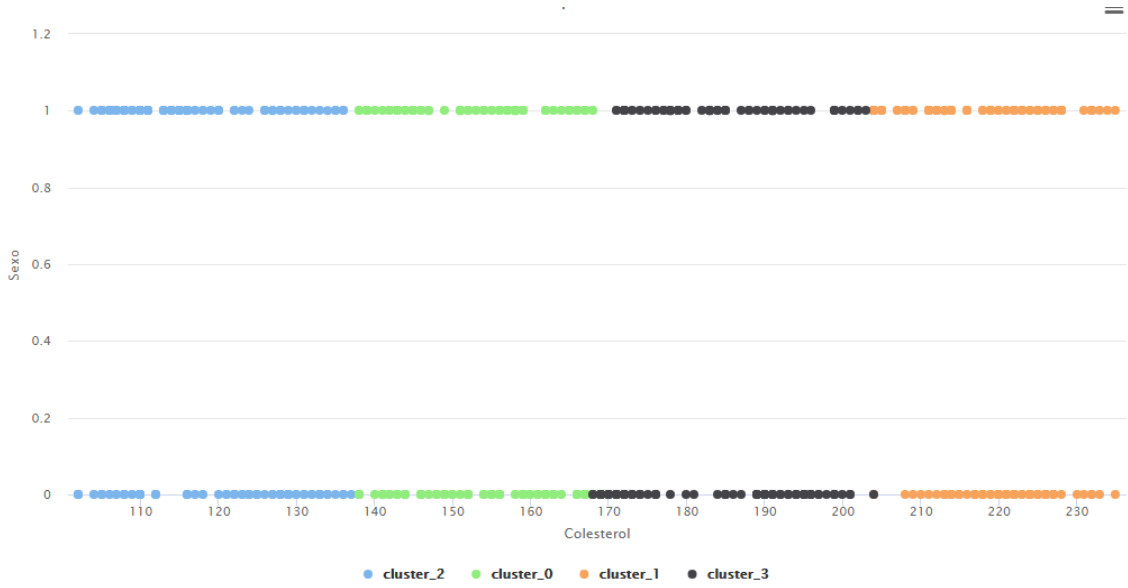
```
Cluster 0: 135 items
Cluster 1: 154 items
Cluster 2: 140 items
Cluster 3: 118 items
Total number of items: 547
```

El modelo arroja 4 cluster, en el siguiente gráfico se observa la relación peso colesterol según cluster:

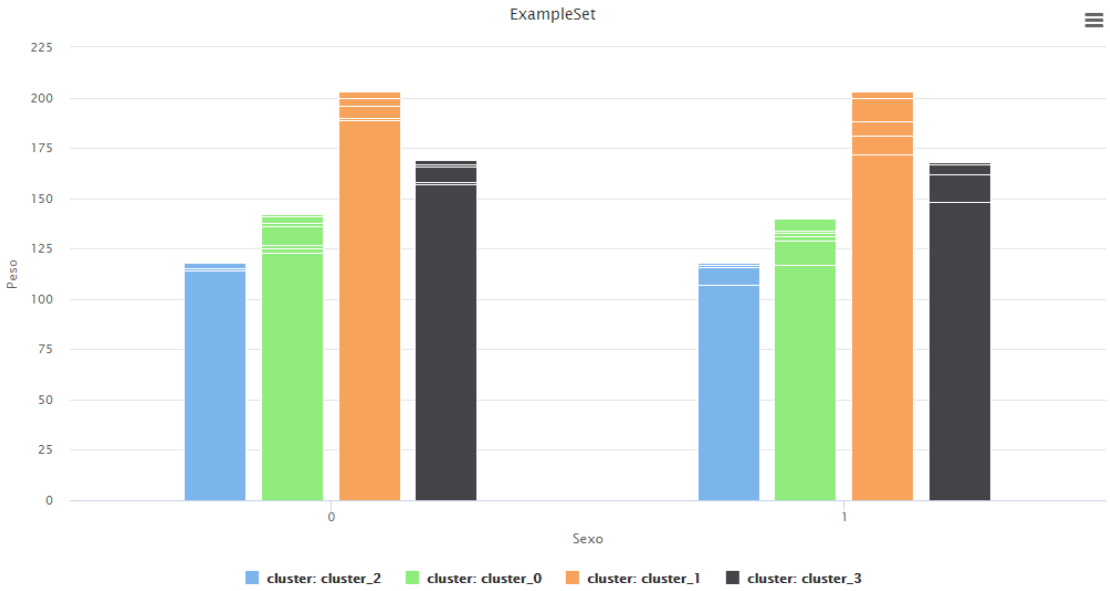


Se puede ver que el cluster anaranjado, situado en la esquina superior derecha de la pantalla agrupa a los iindividuos con niveles más altos de colesterol y peso.

En el siguiente gráfico se observan los cluster según niveles de colesterol, coloreando el sexo



También se puede observar según cluster (coloreado) niveles de colesterol según sexo observando en el cluster_1 que presenta en el gráfico anterior mayores niveles de colesterol y peso, que hay más hombres(1) que mujeres (0).



A su vez se observa la gráfica de peso y sexo según cluster, es posible ver que en el cluster_1 hay individuos con mayor peso, mientras que el cluster_2 presenta los pesos menores

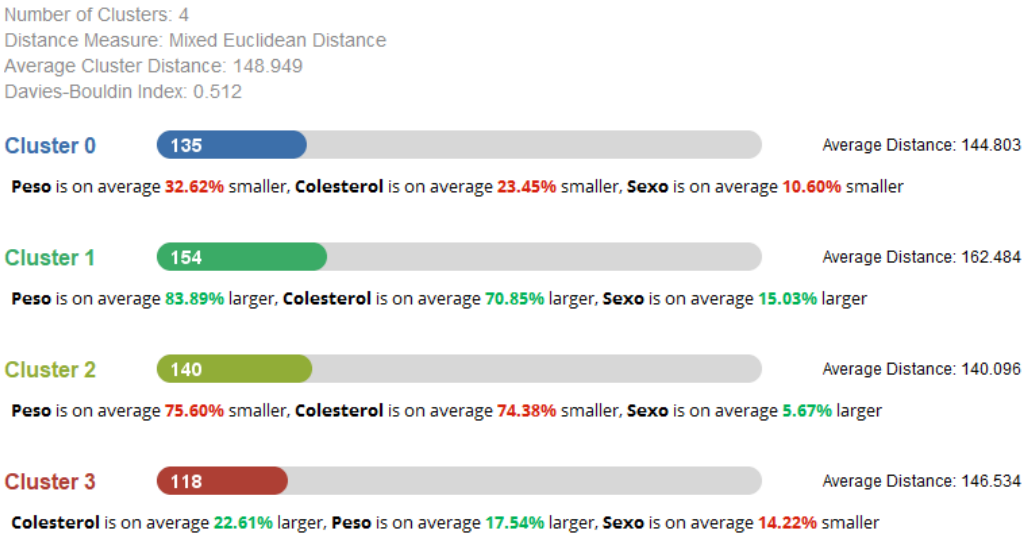
Evaluación

En el presente caso, el cluster_1 es el que tiene mayores valores de peso y colesterol.

En folder view podemos identificar los individuos de dicho cluster, y podeoms filtrar los ejemplos de dicho cluster utilizando el operador filter atributes

Row No.	id	cluster	Peso	Colesterol	Sexo
1	6	cluster_1	198	227	1
2	9	cluster_1	191	223	0
3	10	cluster_1	186	221	1
4	12	cluster_1	188	222	1
5	16	cluster_1	178	213	0
6	18	cluster_1	168	204	1
7	23	cluster_1	199	228	1
8	26	cluster_1	183	218	0
9	28	cluster_1	190	222	0
10	29	cluster_1	174	208	1
11	31	cluster_1	169	204	1
12	35	cluster_1	178	213	0
13	37	cluster_1	195	225	1
14	41	cluster_1	197	225	1

El modelo de clustering visualization da como salida:



Lo que refleja lo evaluado anteriormente, el cluster_1 presenta los valores más altos de peso y colesterol, mientras que el cluster_2 los más bajos. Se observa que en cluster_1 hay más hombres que en los otros cluster, porque la variable sexo es “más grande” lo que representa más 1.

Despliegue

Para los demás cluster debería fijar los valores de peso y colesterol :

Cluster_0 115 a 142 peso y 139 a 168 colesterol

Cluster_1 167 a 203 y 204 a 234

Cluster_2 95 a 118 peso y 102 a 135 colesterol

Cluster_3 140 a 169 peso y 173 a 199 colesterol