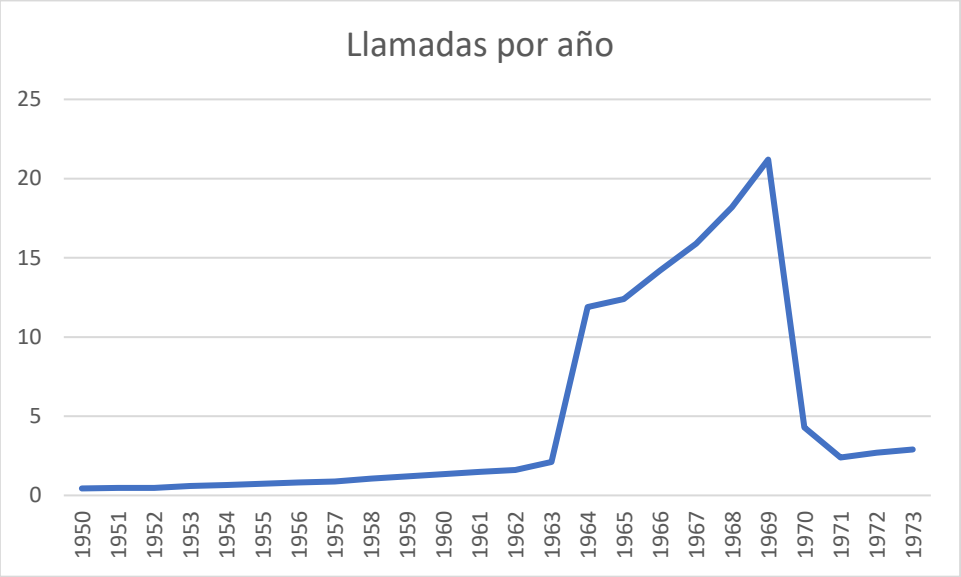
UT2 – TA03

Ejercicio 1

Gráfico de llamadas por año:



De acuerdo a los datos recabados de telephones.csv y la gráfica adjunta se observa un incremento del número de llamadas de 1964 a 1964, aumentando de 5 a 10 veces en comparación a 1963.

Luego de analizar el contexto se observa que las llamadas a la encuesta estadística Belga publicada por el Ministerio de Economía. Resulta que en esos años ocurrió un error y se registró la duración (en minutos) de las llamadas en lugar de la cantidad de llamadas. También los años 1963 y 1970 se ven parcialmente afectados.

Identificados como datos anómalos podríamos descartarlos, si tenemos suficientes datos pero nos quedaría un salto en la serie anual, también podríamos ignorarlos.

Ejercicio 2:

Detect Outlier (Distances)

Este operador identifica n outliers dados en el set de ejemplo basado en la distancia a los k vecinos más cercanos. Las variables k y n pueden ser especificadas mediante parámetros.

Cada punto es rankeado sobre la base de su distancia a los k vecinos más cercano y el top de n points en este ranking son declarados outliers.

Parámetros que toma:

Número de vecinos: Es un integer que especifica el k valor a ser analizado de los k vecinos más cercanos. El mínimo valor de este parámetro es 1 y máximo 1 millón.

Número de outliers: Este parámetro es de tipo integer y especifica el número máximo de outliers a ser identifiados. El exampleset de resultado tendrá n ejemplos considerados outliers. Tiene un rango de 2 a 1 millón.

Función de distancia: este parámetro es de selección y especifica la función de distancia que será usada para calcular la distancia entre dos ejemplos.

Detect Outlier (LOF)

Este operador identifica valores atípicos en el conjunto de ejemplos dado en función de factores locales atípicos (Local Outlier Factors - LOF). El LOF se basa en un concepto de densidad local, donde la localidad está dada por los k vecinos más cercanos, cuya distancia se usa para estimar la densidad. Al comparar la densidad local de un objeto con las densidades locales de sus vecinos, uno puede identificar regiones de densidad similar y puntos que tienen una densidad sustancialmente menor que sus vecinos. Se consideran valores atípicos.

MinPts es un parámetro usado para especificar los 'k'vecinos y usa los máximos LOFs para los objetos dentro de un rango de MinPts.

Parámetros que toma:

puntos mínimos para límite inferior: identifica el límite inferior de MinPts para el test de outlier

puntos mínimos para el límite superior: identifica el límite superior de MinPts para el test de outlier

función de distancia: Parámetro de selección, especifica la función de distancia que se usará para calcular la distancia entre dos objetos.

Detect Outlier (Densities)

Este operador identifica valores atípicos en el conjunto de ejemplos dado en función de la densidad de datos. El operador de detect outlier densities calcula los DB (usando p y D), para el set de ejemplo utilizado. Un DB(p,D) outlier es un objeto que tiene al menos una distancia D de al menos un proporción p del total de objetos. Los dos valores reales p y D pueden ser identificados como parámetros de proporción y distancia respectivamente.

Parámetros que toma:

Distancia: de tipo real especifica el parámetro D de sdistancia para calcular el DB(p,D)

Proporción: de tipo real especifica la proporación p para calcular el DB(p,D)

Función de distancia: parámetro que especifica la función de distancia que será usada para calcular la distancia entre dos ejemplos

Detect Outlier (COF)

Identificado outliers basado en el Class Outlier Factor. El principal concepto del algoritmo es rankear cada distancia del set de ejemplo dados los parámetros N (top de clases de outliers) y K (número de vecinos más cercanos).

El ranking de cada instancia se calcula utilizando la fórmula:

COF = PCL(T,K) - norm(deviation(T)) + norm(kDist(T))

Donde PCL(T,K) es la probabilidad de la etiqueta de clase de la instancia T con respecto a la etiqueta de clase de sus k vecinos más cercanos

norm(Deviation(T)) and norm(KDist(T)) son los valores normalizados de Deviation(T) y KDist(T) y sus valores van de 0 a 1.

Deviation(T) es cuánto la instancia T se devía de instancias de la misma clase.

KDist(T) es la sumatoria de las distancias entre T y sus k vecinos más cercanos.

Parámetros que toma:

Número de vecinos: tipo integer especifica el k valor para los k vecinos más cercanos a ser analizados, su rango va de 1 a 1 millón.

Número de clases outliers: especifica el número máximo de Class Outliers a ser identificados. El rang ova de 2 a 1 millón.

Tipos de medida: usado para seleccionar el tipo de medida a ser usada para las distancia entre los puntos.

Medida mixta: disponible cuando el tipo de medida es seteado en medida mixta la única opción disponible es 'Mixed Euclidean Distance'.

Medida nominal: disponible cuando el tipo de medida es seteado en 'nominal measures'. No puede ser aplicado si el set de ejemplo tiene atributos numéricos.

Medida numérica: **disponible cuando el tipo de medida** es seteado en 'numerical measures'. No puede ser aplicado si el set de ejemplo tiene atributos nominales.

Divergencia: disponible cuando el tipo de medida es seteado en 'Bregman divergences'.

Tipo de kernel: disponible cuando el tipo de medida es seteado en 'Kernel Euclidean Distance'.

Ejercicio 3:

Atributos del data set iris:
Todos los atributos son de tipo numérico (real en rapid miner):

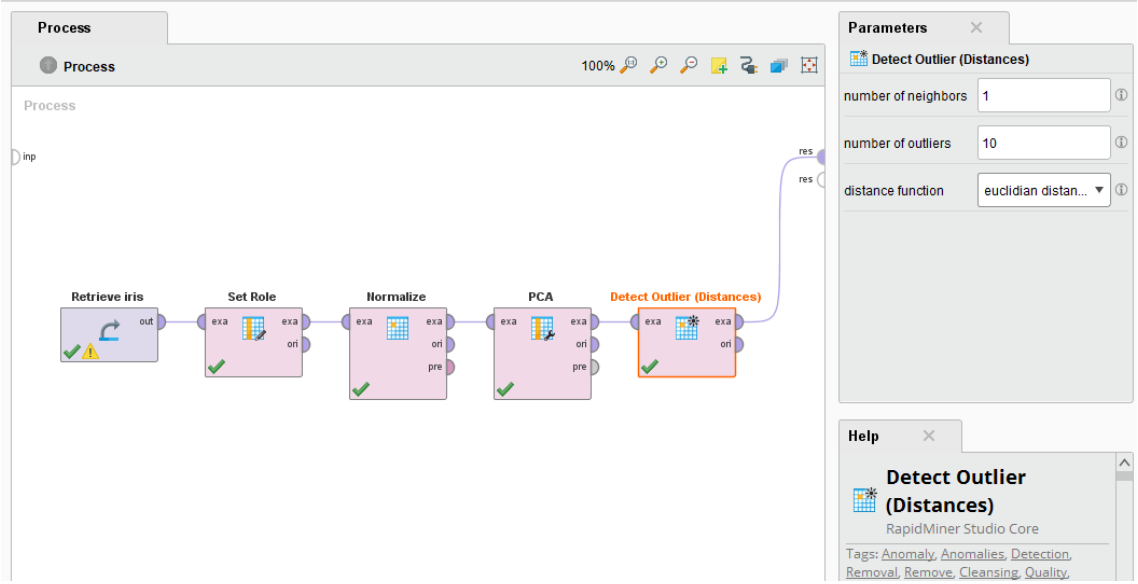
ATRIBUTO:	RANGO:	DISTRIBUCIÓN:
Largo del sépalo en cm:	4,3 a 7,9	Distribución sesgada a la derecha.
Ancho del sépalo en cm:	2 a 4,4	Pareciera asemejarse a una distribución normal, quizás

		sesgada levemente a la derecha
Largo del pétalo en cm:	1 a 6,9	No tiene distribución reconocida simple vista.
Ancho del pétalo en cm:	0,10 a 2,5	No tiene distribución reconocida simple vista.

Clasificación:

-Iris Setosa

- Iris Versicolour
- Iris Virginica



Row No.	att5 ↑	outlier	pc_1	pc_2
13	Iris-setosa	false	-0.662	0.112
14	Iris-setosa	false	-0.753	0.167
15	Iris-setosa	false	-0.599	-0.384
16	Iris-setosa	true	-0.550	-0.519
17	Iris-setosa	false	-0.576	-0.298
18	Iris-setosa	false	-0.603	-0.111
19	Iris-setosa	false	-0.519	-0.291
20	Iris-setosa	false	-0.611	-0.223
21	Iris-setosa	false	-0.558	-0.106
22	Iris-setosa	false	-0.578	-0.185
23	Iris-setosa	true	-0.737	-0.095
24	Iris-setosa	false	-0.506	-0.031
25	Iris-setosa	false	-0.608	-0.033
26	Iris-setosa	false	-0.591	0.091
27	Iris-setosa	false	-0.561	-0.059