



Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

K-Medias en RapidMiner

Tesina

Que para obtener el Título de:

Licenciada en Ciencias de la Computación

Presenta:

Imelda Hernández Baez

Asesora:

Dra. María de Lourdes Sandoval Solís

Puebla, Puebla.

Diciembre, 2015

A mi hija **Imelda**, siempre levantándose y siguiendo...

Agradecimientos

Gracias a **Dios** por darme la oportunidad de nacer y vivir... Gracias por darme la mejor familia:

A mis padres **Lázara** y **Florencio**, por amarme y darme todo su amor y apoyo a lo largo de mi formación personal y profesional.

A mis hermanos: **Luis**, **Roque**, **Elvia**, **Cecilia**, **Mariano** y **Guadalupe**, por su amor y apoyo incondicionales, porque no imagino mi vida sin mis seis hermanos... Gracias por darme la oportunidad de estudiar una carrera y ser orgullosamente universitaria.

A mi esposo **Jesús**, por los momentos que compartimos juntos en lo profesional y familiar... ¡Gracias por estar conmigo siempre, en todo!

A mi hija **Imelda**, Gracias por ser mi compañera, amiga, cómplice y confidente... ¡Gracias por enseñarme a levantarme y seguir!

A mi asesora, la Dra. **María de Lourdes Sandoval Solís**, por su infinita paciencia, tiempo y dedicación para dirigirme.

Admiro su entrega y pasión por el trabajo y la forma en que logra que quiera seguir aprendiendo y superándome cada día.

¡Gracias por todo!

CONTENIDO

1. INTRODUCCIÓN	6
2. MINERÍA DE DATOS	7
2.1 TÉCNICAS DE MINERÍA DE DATOS	12
3. MÉTODOS DE CLASIFICACIÓN AUTOMÁTICA	13
3.1 MÉTODO K-MEDIAS	13
3.2 MEDIDAS DE ASOCIACIÓN	20
3.2.1. Distancias de Bregman.....	20
3.2.2 Medidas Nominales.....	24
3.2.3 Medidas Numéricas.....	27
3.2.4 Medidas Mixtas.....	31
4. RAPIDMINER	32
4.1 K-medias en RapidMiner	35
5. PRUEBAS	37
5.1 Ejemplo 1: Ejemplo 1 (20 datos, 2 atributos).....	37
5.2 Ejemplo 2: ZAE (22 datos, 9 atributos)	38
5.3 Ejemplo 3: Animales (100 datos, 15 atributos).....	45
6. CONCLUSIONES	55
7. BIBLIOGRAFÍA	56

1. INTRODUCCIÓN

Actualmente, en la mayoría de las actividades cotidianas que las personas realizan cuando interactúan con grandes empresas, se generan datos que describen sus hábitos, preferencias y comportamiento; sin estar conscientes de que está ocurriendo. Por ejemplo, cuando se solicita una tarjeta de crédito al banco, se realizan compras en una tienda de autoservicio, se realiza una inscripción a la universidad o se hace ejercicio en el gimnasio, se están generando datos. Estos datos crean una gran cantidad de información, almacenada en bases de datos, que posteriormente puede ser utilizada para describir o predecir su comportamiento y ofrecer así un mejor servicio y atención por parte de las empresas con las que se tiene contacto diariamente.

Debido a esto, es de suma importancia conocer y aplicar técnicas que ayuden a “extraer” la información más relevante del comportamiento de los usuarios. Estas técnicas pueden ayudar a las empresas a clasificar, analizar o inferir la información, y se engloban en el concepto de “Minería de Datos”.

Hoy en día, existen muchas alternativas en software para desarrollar Minería de Datos, una de las más comunes y de fácil acceso es *RapidMiner*, una aplicación de software libre, con una interfaz de usuario muy sencilla y que ofrece muchas ventajas con respecto a otras herramientas.

En este trabajo se presenta un análisis de la técnica de Minería de Datos llamada *Agrupamiento*, específicamente el algoritmo *K-Medias*, en el software *RapidMiner* y las diferentes medidas de asociación que se pueden usar para su ejecución. Dependiendo del tipo de datos que se desee analizar, *RapidMiner* ofrece una serie de alternativas que son de gran ayuda para la comparación e interpretación del resultado final.

Se inicia con una introducción a la Minería de Datos, su alcance e importancia en el análisis de grandes cantidades de información, así como sus técnicas más comunes.

También se presenta una breve descripción de *RapidMiner*, su interfaz, sus principales operadores, procesos, parámetros, etc.

Finalmente, se realizan pruebas con diferentes tipos de datos y se ejecuta el algoritmo *K-Medias* con las diferentes medidas de asociación que ofrece *RapidMiner*, observando y concluyendo que la elección de dichas medidas, influye en el resultado y la interpretación final.

2. MINERÍA DE DATOS

A continuación se presentan los conceptos principales de la minería de datos, su importancia, el proceso necesario que se debe seguir para llevar a cabo un análisis de datos, la elección de la técnica correcta y la interpretación del resultado.

La **Minería de Datos (MD)** como disciplina es en gran parte transparente para los usuarios. La mayoría de las veces, ni siquiera se nota que está sucediendo. Pero cada vez que un cliente se registra para obtener una tarjeta de compras de alguna tienda de autoservicio, hace una compra utilizando una tarjeta de crédito, o navega en la web, se están creando datos. Estos datos se almacenan en conjuntos de computadoras que son propiedad de las empresas con las que se trata todos los días. [1]

Dentro de esos conjuntos de datos se encuentran los patrones o indicadores de los intereses, hábitos y comportamientos de los usuarios. La MD permite a las empresas u organizaciones localizar e interpretar esos patrones, lo que ayuda a tomar decisiones bien informadas y a mejorar el servicio a sus clientes.

Bajo el nombre de MD se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos.

Aunque las raíces de la MD se remontan a finales de 1980, aún se sigue definiendo, y refinando. Era en gran parte un conglomerado “suelto” de modelos de datos, algoritmos de análisis, y salidas especiales. En 1999, varias compañías importantes, incluyendo la automotriz Daimler-Benz, el proveedor de seguros OHRA, el fabricante de hardware y software NCR Corp. y el fabricante de software estadístico SPSS, Inc. comenzaron a trabajar juntos para formalizar y estandarizar un método para la extracción de datos. El resultado de su trabajo fue **CRISP-DM**, CRoss-Industry Standard Process for Data Mining (Proceso estándar entre la industria para minería de datos).

Aunque los participantes en la creación de CRISP-DM tenían intereses basados en ciertas herramientas software y hardware, el proceso fue diseñado independientemente de cualquier herramienta específica. Fue escrito de manera tal que tuviera una naturaleza conceptual, algo que podría aplicarse con independencia de cualquier herramienta o tipo de datos determinado. El proceso consta de seis pasos o fases, como se ilustra en la Figura 2.1. [1]

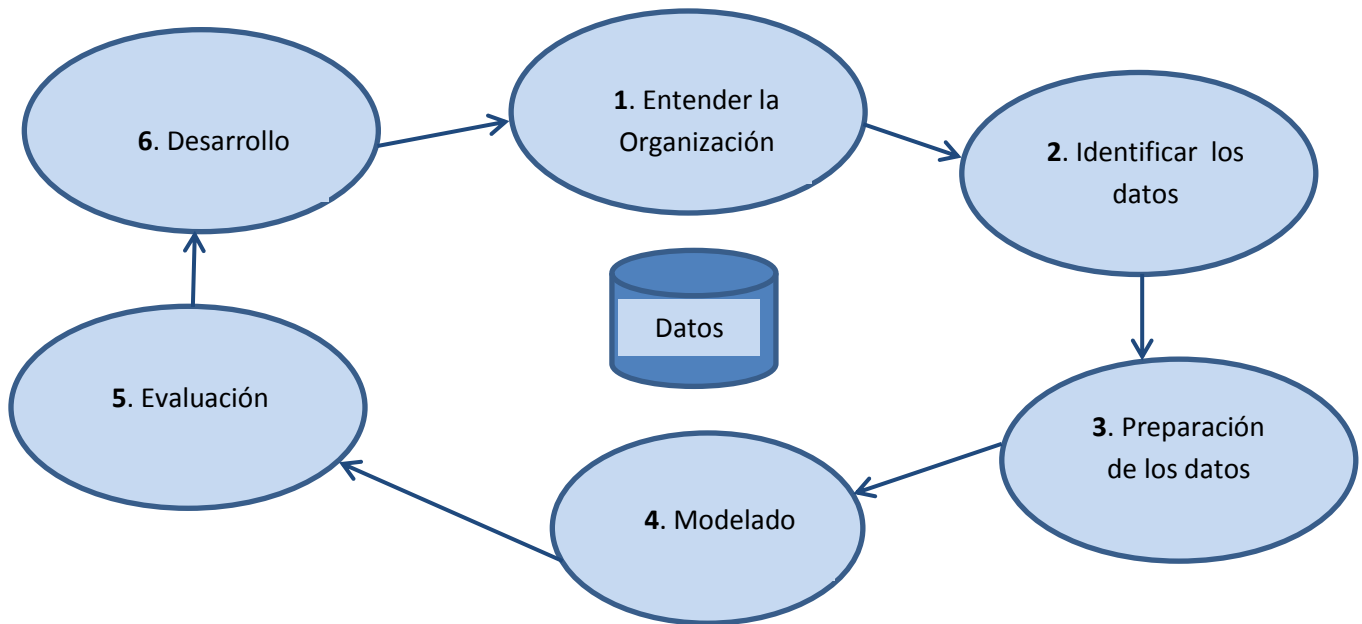


Figura 2.1: Modelo conceptual CRISP-DM

Ahora se describirá brevemente en qué consiste cada paso del modelo CRISP-DM.

PASO 1: Entender la Organización

El primer paso en CRISP-DM es “Entender la organización”, cualquier tipo de organización o negocio puede usar MD para responder preguntas y solucionar problemas. Este paso es crucial para un resultado exitoso, se debe conocer el funcionamiento de la organización y comprender perfectamente el problema que se desea resolver, ya que a menudo la gente no sabe cómo manejar la información. Se puede intentar “extraer” datos, pero si no se tiene claro qué información es relevante y qué preguntas se desean responder, el resultado de la MD no será fructífero. Se debe iniciar con algunas ideas como:

- ¿Cuáles son las quejas más comunes de los clientes?
- ¿Cuál es el producto que más demanda tiene?
- ¿Cómo se puede incrementar el margen de beneficio por unidad?
- ¿Cómo se pueden anticipar defectos de fabricación?
- ¿Cómo se deberían clasificar a los clientes para darles un mejor servicio?

PASO 2: Identificar los datos

Así como entender la organización, entender los datos es una acción de preparación.

Años atrás, cuando los trabajadores no tenían su propia computadora, los datos eran centralizados. Si se necesitaba información de la compañía, se debía solicitar un reporte a una persona que consultara la base de datos central y diera los resultados solicitados.

El uso de las computadoras personales, laptops o tabletas han logrado descentralizar la información, además, se tiene acceso a software fácil de usar como Microsoft Excel y Access.

Así, los datos están disponibles en toda la empresa, los almacenes de datos están esparcidos en cientos de dispositivos, teniendo información en hojas de cálculo de los gerentes de marketing, bases de datos de atención al cliente, y los archivos de recursos humanos.

Con esto se ha creado un problema de datos de múltiples facetas. El departamento de marketing puede tener datos importantes que podrían ser un recurso valioso para la alta dirección, pero ésta no está consciente de su existencia, ya sea debido a que marketing no desea compartir datos con otras áreas, o simplemente porque no han pensado en la importancia de la información que han recabado.

Es poco probable que la MD puede ocurrir cuando los empleados no saben qué datos tienen (o pueden tener) a su disposición, o dónde se encuentran actualmente.

Sin embargo, lograr la descentralización de los datos no es suficiente. Hay muchas preguntas que surgen una vez que se tiene acceso a la información:

¿De dónde provienen los datos?

¿Quién los recolectó?

¿Hubo algún método estándar de recolección?

¿Qué significan las distintas filas y columnas de datos?

¿Hay siglas o abreviaturas que son desconocidas o poco claras?

Se tendrán que hacer algunas reuniones con expertos en la materia en los distintos departamentos para saber de dónde provienen, cómo se recogieron, y cómo se han codificado y almacenado. Es también importante que se verifique la exactitud y fiabilidad de los datos. Datos inexactos o incompletos no son útiles en el proceso de MD, ya que las decisiones basadas en datos parciales o erróneos darán como resultado decisiones equivocadas.

Una vez que se hayan reunido, identificado y comprendido los datos, entonces se podrá iniciar el proceso de MD.

PASO 3: Preparación de los Datos

Los datos vienen en muchas formas y formatos. Algunos datos son numéricos, algunos son párrafos de texto, y otros son figuras tales como cuadros, gráficas o mapas. Existen datos que son anécdotas o narrativas, tales como la encuesta de satisfacción de un cliente o la declaración de un testigo en un juicio.

Preparar los datos implica un gran número de actividades. Estas pueden incluir juntar dos o más conjuntos de datos, reducir conjuntos de datos que contengan solamente aquellas variables que son interesantes para el problema dado, depuración de datos para dejarlos libres de anomalías, tales como observaciones aisladas (atípicas) o datos faltantes, o incluso dar un nuevo formato (re-formatear) a los datos para propósitos de consistencia.

Por ejemplo, se ha visto en hojas de cálculo o bases de datos, números de teléfono como los mostrados en la siguiente tabla:

(555) 555-5555	555/555-5555
555-555-5555	555.555.5555
555 555 5555	5555555555

Tabla 1: Diferentes formatos para números telefónicos

Cada uno de estos ofrece el mismo número telefónico, pero almacenado en diferentes formatos. El resultado final de un ejercicio de MD tiene más probabilidades de ser exitoso, cuando los datos involucrados son lo más consistentes posible.

Paso 4: Modelado

Un modelo, en MD, es una representación computarizada de observaciones del mundo real. Los modelos son la aplicación de algoritmos para hallar, identificar o desplegar cualquier patrón o mensaje en sus datos. Hay dos tipos de modelos en MD: los que **clasifican** y los que **predicen**.

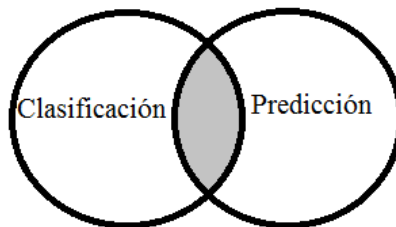


Figura 2.2. Modelos en MD

Como se puede ver en la figura 2.2, existe una intersección entre el uso de los tipos de modelos de MD.

Los modelos predictivos se usan para determinar qué atributos de los contenidos en el conjunto de datos son los más fuertes indicadores en el resultado. Este resultado es expresado como la probabilidad de que las observaciones “caigan” en cierta categoría. Los modelos que clasifican, utilizan como base los patrones que se encuentran en los datos para lograr una categorización de la información.

Los modelos pueden ser simples o complejos. Pueden contener sólo un proceso simple, una cadena de varios o incluso pueden incluir sub procesos.

Independientemente de su diseño, los modelos son la herramienta de la MD para pasar de la preparación y el entendimiento de los datos al desarrollo y la implementación.

PASO 5: Evaluación

Todos los análisis de datos tienen el potencial de ser falsos o verdaderos. Algunos modelos pueden no encontrar patrones interesantes en los datos. Esto puede ser porque el modelo elegido no se ajusta bien al problema que se quiere resolver, porque se está usando una técnica equivocada o simplemente porque no hay ningún patrón de interés en los datos con el que el modelo pueda trabajar.

La evaluación también tiene un aspecto humano. Los usuarios experimentados en este campo, pueden interpretar el resultado desde un punto de vista que no es medible en el sentido matemático.

Una vez elegido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber usado distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema.

PASO 6: Desarrollo

Si se han identificado con éxito las preguntas que se quieren responder, se han preparado los datos que se usarán para responder a esas preguntas y se ha utilizado un modelo que pasa las pruebas de utilidad e interés, entonces se está listo para *usar* los resultados. Esta etapa es el desarrollo.

Las actividades en esta fase incluyen la automatización del modelo, reuniones con los clientes o consumidores, integración con los sistemas de información, de gestión u operacionales de la organización. Incluso alimentando un aprendizaje nuevo en el modelo para mejorar su precisión y rendimiento; llevando a cabo la medición y monitoreo de los resultados que se obtienen.

Los modelos obtenidos por técnicas de MD se aplican incorporándolos en los sistemas de análisis de información de las organizaciones, e incluso, en los sistemas transaccionales. En este sentido cabe destacar los esfuerzos del Data Mining Group, que está estandarizando el lenguaje **PMML** (Predictive Model Markup Language), de manera que los modelos de MD sean interoperables en distintas plataformas, con independencia del sistema con el que han sido construidos. Los principales fabricantes de sistemas de bases de datos y programas de análisis de la información hacen uso de este estándar. [1]

2.1 TÉCNICAS DE MINERÍA DE DATOS

Las técnicas más comunes de minería de datos se pueden clasificar en las siguientes categorías: [2]

- Clasificación y regresión
- Ponderación de atributos
- Agrupamiento y Segmentación (Clasificación automática)
- Reglas de Asociación
- Cálculo de Correlación y Dependencia
- Cálculo de semejanza

Las técnicas más representativas son:

- **Redes neuronales.**- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los seres vivos. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:
 - El perceptrón.
 - El perceptrón multicapa.
 - Los mapas auto-organizados, también conocidos como redes de Kohonen.
- **Regresión lineal.**- Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.
- **Árboles de decisión.**- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.
- **Agrupamiento.**- Es un procedimiento de agrupación de una serie de vectores según criterios de medidas de asociación (distancias o similitudes); se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos:
 - Algoritmo K-means.
 - Algoritmo K-medoids.
- **Reglas de asociación.**- Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados:

- Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.
- Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos.

3. MÉTODOS DE CLASIFICACIÓN AUTOMÁTICA

Los métodos de clasificación automática, denominados también como análisis de cluster o simplemente *agrupamiento*, tienen por objeto formar grupos generalmente de individuos o instancias, de forma que los datos de los grupos formados sean lo más similares posible y los grupos sean lo más diferentes posible unos de otros. [4]. En este apartado se describirá el algoritmo K-Medias y las diferentes medidas de asociación que se pueden usar para su ejecución.

3.1 MÉTODO K-MEDIAS

Uno de los métodos de clasificación más conocido es el K-Medias. En este algoritmo las observaciones son clasificadas en k grupos, el número total de miembros del grupo es determinado por el cálculo del centro para cada grupo y asignando cada observación al grupo con el centro más cercano. [3]

Este es el algoritmo de agrupamiento más popular. También es llamado el *algoritmo de las medias móviles* porque en cada iteración se recalculan los centros de los agrupamientos. Por esta razón se incorpora el índice t a la notación que se emplea, de manera que con $S_i(t)$ se indica el conjunto de patrones asociados al agrupamiento S_i en la iteración t y mediante $Z_i(t)$ se indica el valor de su centro en esa iteración.

El algoritmo requiere de 3 parámetros, X , que es el conjunto de observaciones de entrada, K , que es el número de agrupamientos que debe encontrar e $iter$, el número máximo de iteraciones.

El algoritmo de las K-medias se puede escribir en pseudo código de la siguiente manera:

K-medias($X, K, iter$)

Entrada:

X : un conjunto de N patrones $\{X_1, X_2, \dots, X_N\}$
 K : número de agrupamientos
 $iter$: número máximo de iteraciones

Salida:

S_1, S_2, \dots, S_k : K conjuntos de patrones
 Z_1, Z_2, \dots, Z_k : Los centros de los k agrupamientos

Algoritmo:

Paso 1: Inicializar centros

Paso 2: Asignación de Observaciones

Si $t \leq iter$

Hacer para todas las observaciones:

$$j=1:N \quad k=1:K$$

- Calcular las distancias de la observación a cada uno de los centros

$$d_{jk} = \sqrt{(X_{j1} - Z_{k1})^2 + (X_{j2} - Z_{k2})^2 + \dots + (X_{jN} - Z_{kN})^2}$$

$$d_{jk} = \sqrt{\sum_{i=1}^n (X_{ji} - Z_{ki})^2}$$

- Determinar a qué grupo pertenece la observación
 $\min = \min (d_{j1}, d_{j2}, \dots, d_{jk})$

Fin

Fin Si

Paso 3: Actualización de centros

Recalcular los centros con el promedio aritmético de los elementos que pertenecen al grupo.

A continuación se muestra un ejemplo del algoritmo.

Ejemplo 1 del algoritmo de las K-Medias

En la figura 3.1 a y b se muestra el conjunto de 20 observaciones y dos variables, X1 y X2, sobre el que se aplicará el algoritmo de las K-Medias, con $k=2$.

Datos	X1	X2
1	0	0
2	1	0
3	0	1
4	1	1
5	2	1
6	1	2
7	2	2
8	2	3
9	6	6
10	7	6
11	8	6
12	6	7
13	7	7
14	8	7
15	9	7
6	7	8
17	8	8
18	9	8
19	10	8
20	11	8

Figura 3.1a Conjunto de patrones inicial.

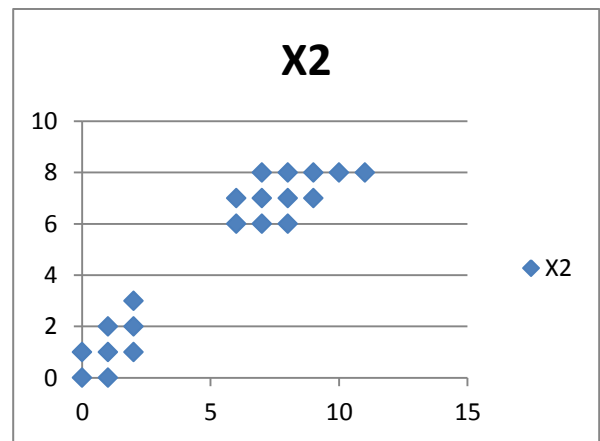


Figura 3.1b Gráfica de datos.

PASO 1: Inicialización.

Dado que $k = 2$, se tendrán dos conjuntos de patrones, ahora, supongamos que los conjuntos iniciales están formados por $S_1(0) = \{X_1\}$ y $S_2(0) = \{X_2\}$; se inicializan los centros con los valores que corresponden a X_1 y X_2 , por lo que los datos iniciales quedan de la siguiente forma:

$$S_1(0) = \{X_1\} \text{ y } X_1(0) = (0, 0)$$

$$S_2(0) = \{X_2\} \text{ y } X_1(0) = (1, 0)$$

PASO 2: Asignación y actualización de centros

Se inicia asignando los elementos más cercanos a los centros de los grupos, comparando elemento a elemento con los centros y calculando las distancias euclidianas de los diferentes puntos con ambos centros; asignando en base a la menor distancia.

Así:

$$\begin{aligned} |X_2 - Z_1| &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\ &= \sqrt{(1 - 0)^2 + (0 - 0)^2} \\ &= \sqrt{1} \end{aligned}$$

$$\begin{aligned} |X_2 - Z_2| &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\ &= \sqrt{(1 - 1)^2 + (0 - 0)^2} \\ &= \sqrt{0} \end{aligned}$$

De aquí se obtiene que la distancia mínima es la que hay del elemento X_2 al centro Z_2 , por lo tanto, este elemento se asigna al conjunto S_2 .

Ahora, hacemos la comparación para el elemento X_3 .

$$\begin{aligned} |X_3 - Z_1| &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\ &= \sqrt{(0 - 0)^2 + (1 - 0)^2} \\ &= \sqrt{1} \end{aligned}$$

$$\begin{aligned} |X_3 - Z_2| &= \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \\ &= \sqrt{(0 - 1)^2 + (1 - 0)^2} \\ &= \sqrt{2} \end{aligned}$$

La menor distancia es la que hay entre el elemento X_3 al centro Z_1 , y por tanto X_3 se asigna al conjunto.

Realizando los mismos cálculos para los demás elementos, se obtienen los siguientes datos, que se muestra en la tabla 3.2:

1a Iteración		
Centro 1: Z0(1)	0	0
Centro 2: Z0(2)	1	0
Distancia a Z1	Distancia a Z2	Asignación
0.0	1.0	Se asigna al grupo 1
1.0	0.0	Se asigna al 2
1.0	1.4	Se asigna al grupo 1
1.4	1.0	Se asigna al 2
2.2	1.4	Se asigna al 2
2.2	2.0	Se asigna al 2
2.8	2.2	Se asigna al 2
3.6	3.2	Se asigna al 2
8.5	7.8	Se asigna al 2
9.2	8.5	Se asigna al 2
10.0	9.2	Se asigna al 2
9.2	8.6	Se asigna al 2
9.9	9.2	Se asigna al 2
10.6	9.9	Se asigna al 2
11.4	10.6	Se asigna al 2
10.6	10.0	Se asigna al 2
11.3	10.6	Se asigna al 2
12.0	11.3	Se asigna al 2
12.8	12.0	Se asigna al 2
13.6	12.8	Se asigna al 2

Tabla 3.2 Ejecución del algoritmo K-Medias, primera iteración.

$$S_1(1) = \{X_1, X_3\}$$

$$S_2(1) = \{X_2, X_4, X_5, \dots X_{20}\}$$

PASO 3: Actualización de centros

Ahora, se obtienen los nuevos centros, calculando el promedio de los nuevos grupos, obteniendo:

$$Z_1(1) = (0, 0.5) \qquad Z_2(1) = (5.8, 5.3)$$

En este paso se verifica si los centros han cambiado o no. En caso de cambiar, se repite el paso 2, si no, se considera que ya se ha encontrado una buena partición y el algoritmo termina.

Para este caso,

$$Z_1(1) \neq Z_1(0) \qquad \text{y} \qquad Z_2(1) \neq Z_2(0), \text{ entonces se repite el paso 2.}$$

2ª. Iteración

Con los nuevos centros:

$$Z_1(1) = (0, 0.5) \qquad Z_2(1) = (5.8, 5.3)$$

PASO 2:

Calculando las distancias de los elementos hacia los dos centros y asignando en base a la menor distancia, se obtiene:

$$S_1(2) = \{X_1, X_2, \dots X_8\} \qquad S_2(2) = \{X_9, X_{10}, \dots X_{20}\}$$

Realizando el cálculo de los nuevos centros se obtiene:

$$Z_1(1) = (1.1, 1.3) \qquad Z_2(0) = (8.0, 7.2)$$

PASO 3:

$$\text{Como } Z_1(2) \neq Z_1(1) \qquad \text{y} \qquad Z_2(2) \neq Z_2(1), \text{ entonces se repite el paso 2.}$$

3a. Iteración

PASO 2:

Se calcula la distancia de los elementos hacia los dos centros y se asignan en base a la menor distancia. Se obtiene:

$$S_1(2) = \{X_1, X_2, \dots X_8\} \qquad S_2(2) = \{X_9, X_{10}, \dots X_{20}\}$$

Se realiza el cálculo de los nuevos centros, obteniendo:

$$Z_1(3) = (1.1, 1.3) \qquad Z_2(3) = (8.0, 7.2)$$

PASO 3:

Finalmente, como $Z_1(3) = Z_1(2)$ y $Z_2(3) = Z_2(2)$, se considera que la partición obtenida es estable y se termina el algoritmo.

OBSERVACIONES:

El algoritmo puede ofrecer variantes en los pasos de **inicialización de centros** y en el **cálculo de las distancias**. A continuación se detallarán los métodos disponibles.

INICIALIZACIÓN DE CENTROS

Se puede llevar a cabo:

- De manera aleatoria en el dominio de los datos
- Utilizando los primeros elementos de la base de datos de entrada
- De forma aleatoria de los elementos de la base de datos

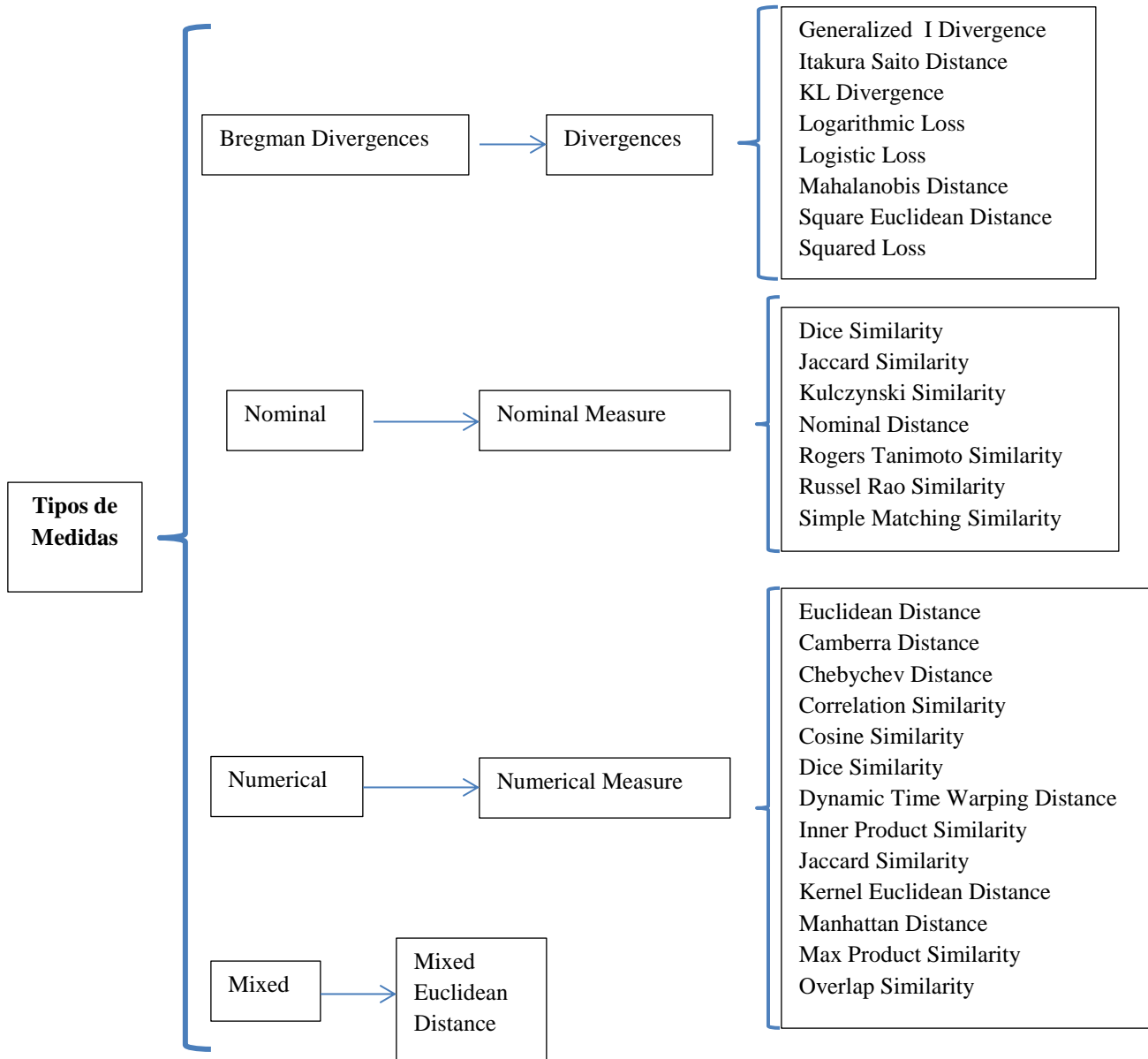
En este trabajo se usa la inicialización aleatoria en el dominio de datos, ya que es la técnica utilizada por el algoritmo K-medias implementado en RapidMiner. [2]

CÁLCULO DE LAS DISTANCIAS

RapidMiner ofrece distintas formas de calcular las medidas que caracterizan las relaciones entre las variables o los individuos. Cada medida refleja asociación en un sentido particular y es necesario elegir una apropiada para el problema concreto que se esté tratando. La medida de asociación puede ser una **distancia** o una **similaridad**.

- Cuando se elige una **distancia** como medida de asociación (por ejemplo la distancia Euclidiana) los grupos formados contendrán individuos parecidos de forma que la distancia entre ellos ha de ser pequeña.
- Cuando se elige una medida de **similaridad** los grupos formados contendrán individuos con una similaridad alta entre ellos.

RapidMiner ofrece distintos medios para realizar el algoritmo K-Medias, dependiendo del tipo de datos que se tienen en la base de datos inicial, estos pueden ser: Numéricos, Nominales, Mixtos y una categoría especial llamada Bregman Divergences. A continuación se muestra un cuadro resumen con los diferentes tipos de medidas.



Cuadro C-3.1 Medidas disponibles en RapidMiner para el algoritmo K-Medias.

3.2 MEDIDAS DE ASOCIACIÓN

Como se observa en el cuadro C-3.1, RapidMiner ofrece 4 tipos de medidas para llevar a cabo el algoritmo K-Medias: Distancias de Bregman, Medidas Nominales, Medidas Numéricas y Medidas Mixtas.

Se presentará una breve descripción de cada una de ellas, para ello se requiere conocer las siguientes definiciones:

Definición 3.2.a. [6]. Sea U un conjunto finito o infinito de elementos. Una función $d: U \times U \rightarrow \mathbb{R}_0^+$, se dice **métrica** si y sólo si $\forall x, y \in U$ se cumple:

1. $d(x, y) \geq 0$
2. $d(x, y) = 0 \leftrightarrow x = y$
3. $d(x, y) = d(y, x)$ (Propiedad simétrica)
4. $d(x, z) \leq d(x, y) + d(y, z), \forall z \in U$ (Desigualdad del triángulo)

Las métricas se usan para medir distancias.

Definición 3.2.b. [6]. Sea U un conjunto finito o infinito de elementos. Una función $s: U \times U \rightarrow \mathbb{R}$ se llama **similaridad** si cumple las siguientes propiedades: $\forall x, y \in U$

1. $s(x, y) \geq s_0$
2. $s(x, x) = s_0$
3. $s(x, y) = s(y, x)$

donde s_0 es un número real finito arbitrario, mayor que cero.

A continuación, se expondrán algunas de las distancias y similaridades más usuales en la práctica, incluidas en RapidMiner.

3.2.1. Distancias de Bregman

Las distancias de Bregman fueron nombradas así después de que L. M. Bregman introdujo el concepto en 1967. Recientemente investigadores en algoritmos geométricos han mostrado que muchos algoritmos importantes pueden ser generalizados sustituyendo métricas Euclidianas por distancias definidas por Bregman.

Para definir las distancias de Bregman, se usará la siguiente notación:

$\mathbf{x}, \mathbf{y}, \mathbf{z}$: vectores

S : Conjunto

\mathbb{R} : El conjunto de los números Reales,

\mathbb{R}_0^+ : Reales no negativos,

\mathbb{R}^+ : Reales positivos

\mathbb{R}^d : Espacio vectorial Real de dimensión d .

ϕ : función $\phi: \text{dom}(\phi) \rightarrow \text{range}(\phi)$

$\text{dom}(\phi)$: Dominio de la función ϕ

$\text{range}(\phi)$: Rango de la función ϕ

$ri(S)$: conjunto abierto de S

x : Producto cartesiano

Para $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\|\mathbf{x}\|$ denota la norma L_2

$\langle \mathbf{x}, \mathbf{y} \rangle$: producto interno.

\log : logaritmo natural

p, q : funciones de distribución de probabilidad.

Definición 3.2.c. [5]. Sea $\phi : S \rightarrow \mathbb{R}$, $S = \text{dom}(\phi)$, una función estrictamente convexa, definida en un conjunto convexo no vacío $S \rightarrow \mathbb{R}$, tal que ϕ es diferenciable en $ri(S)$.

La **distancia de Bregman**, $d\phi : S \times ri(S) \rightarrow [0, \infty)$ está definida como

$$d\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle$$

donde $\nabla\phi(\mathbf{y})$ representa el vector gradiente de ϕ evaluado en \mathbf{y} .

Ejemplo 1: La **Distancia Euclidiana Cuadrada** es quizá la distancia de Bregman más simple y la más ampliamente usada.

La función $\phi(x) = \langle x, x \rangle$ es estrictamente convexa, diferenciable en \mathbb{R}^d y

$$\begin{aligned} d\phi(\mathbf{x}, \mathbf{y}) &= \langle x, x \rangle - \langle y, y \rangle - \langle x - y, \nabla\phi(y) \rangle \\ &= \langle x, x \rangle - \langle y, y \rangle - \langle x - y, 2yx \rangle \\ &= \langle x - y, x - y \rangle = \|\mathbf{x} - \mathbf{y}\|^2 \end{aligned}$$

Ejemplo 2: Otra distancia de Bregman ampliamente usada es la **Distancia Kullback-Leibler** o **KL**. Si p es una distribución de probabilidad discreta, tal que: $\sum_{j=1}^d p_j = 1$, la entropía negativa $\phi(p) = \sum_{j=1}^d p_j \log_2 p_j$ es una función convexa. La distancia correspondiente de Bregman es

$$\begin{aligned} d\phi(p, q) &= \sum_{j=1}^d p_j \log_2 p_j - \sum_{j=1}^d q_j \log_2 q_j - \langle p - q, \nabla\phi(q) \rangle \\ &= \sum_{j=1}^d p_j \log_2 p_j - \sum_{j=1}^d q_j \log_2 q_j - \sum_{j=1}^d (p_j - q_j)(\log_2 q_j + \log_2 e) \\ &= \sum_{j=1}^d p_j \log_2 \left(\frac{p_j}{q_j} \right) - \log_2 e \sum_{j=1}^d (p_j - q_j) \\ &= \text{KL}(p \parallel q) \end{aligned}$$

la distancia KL entre las dos distribuciones dado que $\sum_{j=1}^d q_j = \sum_{j=1}^d p_j = 1$

La tabla 3.3 muestra un resumen con las distancias de Bregman más comunes. [5]

Dominio	$\phi(\mathbf{x})$	$d\phi(\mathbf{x}, \mathbf{y})$	Distancia
\mathbb{R}	x^2	$(x-y)^2$	<i>Squared Loss</i>
\mathbb{R}^+	$x \log x$	$x \log \left(\frac{x}{y}\right) - (x - y)$	
$[0, 1]$	$x \log x + (1-x) \log(1-x)$	$x \log \left(\frac{x}{y}\right) + (1-x) \log \left(\frac{1-x}{1-y}\right)$	<i>Logistic Loss</i>
\mathbb{R}^+	$-\log x$	$\frac{x}{y} - \log \left(\frac{x}{y}\right) - 1$	<i>Itakura-Saito</i>
\mathbb{R}	e^x	$e^x - e^y - (x - y)e^y$	
\mathbb{R}^d	$\ \mathbf{x}\ ^2$	$\ \mathbf{x} - \mathbf{y}\ ^2$	<i>Euclidiana Cuadrada</i>
\mathbb{R}^d	$\mathbf{x}^T A \mathbf{x}$	$(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})$	<i>Mahalanobis</i>
<i>d-Simplex</i>	$\sum_{j=1}^d x_j \log_2 x_j$	$\sum_{j=1}^d x_j \log_2 \left(\frac{x_j}{y_j}\right)$	<i>KL</i>
\mathbb{R}^d	$\sum_{j=1}^d x_j \log_2 x_j$	$\sum_{j=1}^d x_j \log_2 \left(\frac{x_j}{y_j}\right) - \sum_{j=1}^d (x_j - y_j)$	<i>Generalized I</i>

Tabla 3.3: Distancias de Bregman generadas de algunas funciones convexas.

Propiedades de las distancias de Bregman

Sea $\phi : S \rightarrow \mathbb{R}$, $S = \text{dom}(\phi)$, una función estrictamente convexa, definida en un conjunto convexo no vacío $S \rightarrow \mathbb{R}$, tal que ϕ es diferenciable en $ri(S)$.

La **distancia de Bregman**, $d\phi : S \times ri(S) \rightarrow [0, \infty)$ está definida como

$$d\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla \phi(\mathbf{y}) \rangle$$

Entonces las siguientes propiedades son verdaderas:

1. **No negatividad:** $d\phi(\mathbf{x}, \mathbf{y}) \geq 0$, $\forall \mathbf{x} \in S, \mathbf{y} \in ri(S)$, y la igualdad se mantiene si y sólo si, $\mathbf{x} = \mathbf{y}$.

2. **Convexidad:** $d\phi$ es siempre convexa en el primer argumento, pero no necesariamente en el segundo. La distancia Euclidiana Cuadrada y la distancia KL son ejemplos de distancias de Bregman que son convexas en ambos argumentos, pero la distancia de Bregman correspondiente a la función estrictamente convexa $\phi(x) = x^3$, definida en \mathbb{R}^+ , dada por $d\phi(x, y) = x^3 - y^3 - 3(x-y)y^2$ es un ejemplo de distancia que no es convexa en y .

3. **Linealidad:** Las distancia de Bregman es un operador lineal, es decir, $\forall \mathbf{x} \in S, \mathbf{y} \in ri(S)$,

$$\begin{aligned} d_{\phi_1 + \phi_2}(\mathbf{x}, \mathbf{y}) &= d_{\phi_1}(\mathbf{x}, \mathbf{y}) + d_{\phi_2}(\mathbf{x}, \mathbf{y}), \\ d_{c\phi}(\mathbf{x}, \mathbf{y}) &= c d_{\phi}(\mathbf{x}, \mathbf{y}) \text{ (para } c \geq 0) \end{aligned}$$

4. Clases de Equivalencia

Las distancias de Bregman de funciones que difieren en un término afín, son idénticas. Es decir, si $\phi(\mathbf{x}) = \phi_0(\mathbf{x}) + \langle \mathbf{b}, \mathbf{x} \rangle + c$, donde $\mathbf{b} \in \mathbb{R}^d$ y $c \in \mathbb{R}$, entonces $d_{\phi}(\mathbf{x}, \mathbf{y}) = d_{\phi_0}(\mathbf{x}, \mathbf{y})$, $\forall \mathbf{x} \in S, \mathbf{y} \in ri(S)$. Por consiguiente, el conjunto de todas las funciones convexas diferenciables, en un conjunto convexo S , pueden ser particionadas en clases de equivalencia de la forma:

$$[\phi_0] = \{ \phi \mid d_{\phi}(\mathbf{x}, \mathbf{y}) = d_{\phi_0}(\mathbf{x}, \mathbf{y}) \forall \mathbf{x} \in S, \mathbf{y} \in ri(S) \}$$

5. Separación Lineal

El lugar geométrico de todos los puntos $\mathbf{x} \in S$, que son equidistantes desde dos puntos fijos $\mu_1, \mu_2 \in ri(S)$, en términos de una divergencia de Bregman, es un hiperplano. Es decir, las particiones inducidas por las divergencias de Bregman, tienen separadores lineales dados por:

$$\begin{aligned}
 & d_\phi(\mathbf{x}, \mu_1) = d_\phi(\mathbf{x}, \mu_2) \\
 \rightarrow & \phi(\mathbf{x}) - \phi(\mu_1) - \langle \mathbf{x} - \mu_1, \nabla\phi(\mu_1) \rangle = \phi(\mathbf{x}) - \phi(\mu_2) - \langle \mathbf{x} - \mu_2, \nabla\phi(\mu_2) \rangle \\
 \rightarrow & \langle \mathbf{x}, \nabla\phi(\mu_2) - \nabla\phi(\mu_1) \rangle = (\phi(\mu_1) - \phi(\mu_2)) - (\langle \mu_1, \nabla\phi(\mu_1) \rangle - \langle \mu_2, \nabla\phi(\mu_2) \rangle)
 \end{aligned}$$

3.2.2 Medidas Nominales

En ocasiones se encuentran variables que pueden tomar dos valores (blanco-negro, si-no, hombre-mujer, verdadero-falso, etc.). En tales casos se emplea el convenio de usar los valores binarios 1 y 0 para ambos valores.

Al relacionar dos variables binarias, se forma una tabla de contingencia 2×2 , que se puede esquematizar de la forma: [7]

X_i/X_j	1	0	Totales
1	A	b	$a+b$
0	C	d	$c+d$
Totales	$a+c$	$b+d$	$a+b+c+d$

Tabla 3.4. Tabla de contingencia para datos binarios

En la anterior tabla se tiene:

1. a representa el número de individuos que toman el valor 1 en cada variable de forma simultánea.
2. b indica el número de individuos de la muestra que toman el valor 1 en la variable X_i y 0 en la X_j .
3. c es el número de individuos de la muestra que toman el valor 0 en la variable X_i y 1 en la X_j .
4. d representa el número de individuos que toman el valor 0 en cada variable, al mismo tiempo.
5. $a+c$ muestra el número de veces que la variable X_j toma el valor 1, independientemente del valor tomado por X_i .
6. $b+d$ es el número de veces que la variable X_j toma el valor 0, independientemente del valor tomado por X_i .
7. $a+b$ es el número de veces que la variable X_i toma el valor 1, independientemente del valor tomado por X_j .
8. $c+d$ es el número de veces que la variable X_i toma el valor 0, independientemente del valor tomado por X_j .

Medidas basadas en coincidencias

Una forma intuitiva de medir la similaridad en variables binarias es contar el número de veces que ambas variables toman el mismo valor de forma simultánea. Con ello dos variables serían más parecidas en tanto cuanto mayor fuera el número de coincidencias a lo largo de los individuos.

No obstante, algunos factores influyen en las medidas que se pueden definir. Por ejemplo, una primera cuestión es qué hacer con las parejas del tipo 0-0, ya que si las dicotomías son del tipo presencia-ausencia, los datos de la casilla d no poseen ningún atributo y no deberían tomar parte en la medida de asociación.

Otra cuestión que surge es cómo ponderar las coincidencias y cómo las no coincidencias, o lo que es lo mismo, una diagonal u otra de la tabla 3.4.

A continuación se exponen algunas de las medidas que han ido surgiendo, atendiendo a varios criterios como los anteriores. Estas medidas están incluidas en RapidMiner.

Medida de Russell y Rao

$$\frac{a}{a + b + c + d}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar tenga el valor 1 en ambas variables. Notemos que este coeficiente excluye la pareja 0 – 0, al contar el número de coincidencias pero no lo hace así al contar el número de posibles parejas. Asimismo, esta medida proporciona igual peso a las coincidencias y a las no coincidencias.

Medida de parejas simples (Simple Matching)

$$\frac{a + d}{a + b + c + d}$$

Este coeficiente mide la probabilidad de que un individuo elegido al azar presente una coincidencia de cualquier tipo, pesando de igual forma las coincidencias y las no coincidencias.

Medida de Jaccard

$$\frac{a}{a + b + c}$$

Esta medida mide la probabilidad condicionada de que un individuo elegido al azar presente un 1 en ambas variables, dado que las coincidencias del tipo 0–0 han sido descartadas primero y por lo tanto han sido tratadas de forma irrelevante.

Medida de Dice

$$\frac{2a}{2a + b + c}$$

Esta medida excluye el par 0–0 de forma completa, pesando de forma doble las coincidencias del tipo 1 – 1. Se puede ver este coeficiente como una extensión de la medida de Jaccard, aunque su sentido probabilístico se pierde.

Medida de Rogers-Tanimoto

$$\frac{a + d}{a + d + 2(b + c)}$$

Este coeficiente puede interpretarse como una extensión de la medida de parejas simples, pesando con el doble valor las no coincidencias.

Medida de Kulczynski

$$\frac{a}{b + c}$$

Esta medida muestra el cociente entre coincidencias y no coincidencias, excluyendo los pares 0 – 0.

A continuación se muestra un ejemplo del cálculo de similitud entre dos especies E1 y E2, usando ocho rasgos funcionales binarios, R1 a R8.

Especie/Rasgos	R1	R2	R3	R4	R5	R6	R7	R8
E1	1	0	0	1	0	0	0	0
E2	0	1	0	1	1	0	0	0

La tabla de contingencias asociada se muestra a continuación:

		E1		Totales
		Presente (1)	Ausente (0)	
E2	Presente (1)	1	2	3
	Ausente (0)	1	4	5
	Totales	2	6	8

Para este ejemplo, la medida de Jaccard es:

$$\frac{a}{a+b+c} = \frac{1}{1+2+1} = \frac{1}{4} = 0.25$$

La medida de parejas simples es:

$$\frac{a+d}{a+b+c+d} = \frac{1+4}{1+2+1+4} = \frac{5}{8}$$

La de Dice es:

$$\frac{2a}{2a+b+c} = \frac{2(1)}{2(1)+2+1} = \frac{2}{5}$$

3.2.3 Medidas Numéricas

A continuación se describirán brevemente las medidas numéricas con las que cuenta RapidMiner para ejecutar el algoritmo K-medias.

Distancia Euclidiana

La distancia euclidiana o euclídea es la distancia "ordinaria" (que se mediría con una regla) entre dos puntos de un espacio euclidiano, la cual se deduce a partir del teorema de Pitágoras.

Definición 3.2.3.a. En general, la distancia euclidiana entre los puntos $p=(p_1, p_2, \dots, p_n)$ y $q=(q_1, q_2, \dots, q_n)$, del espacio euclidiano n -dimensional, se define como: [8]

$$d_E(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}$$

Distancia Camberra

La distancia Canberra es una medida numérica de la distancia entre pares de puntos en un espacio vectorial, introducida en 1966 y refinada en 1967 por GN Lance y WT Williams. Es una versión ponderada de la distancia de Manhattan. La distancia Canberra se ha utilizado como una métrica para comparar las listas clasificadas y para la detección de intrusiones en seguridad computacional.

Definición 3.2.3.b. [9] La distancia de Canberra entre dos vectores $p=(p_1, p_2, \dots, p_n)$ y $q=(q_1, q_2, \dots, q_n)$, en un espacio vectorial real de dimensión n , se define como:

$$d_{CAMBERRA}(p, q) = \sum_{i=1}^n \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

Distancia Chebychev

La distancia de Chebychev (o Tchebichev), también conocida como la norma L_∞ , es una métrica definida sobre un espacio vectorial donde la distancia entre dos vectores es la mayor diferencia en cualquiera de las coordenadas del espacio.

Definición 3.2.3.c. [10] La distancia de Chebyshev entre dos vectores o puntos p y q , con coordenadas estándar p_i y q_i , respectivamente, es:

$$D_{Chebyshev}(p, q) = \max_i(|p_i - q_i|)$$

Similaridad por Correlación

Está basada en el coeficiente de correlación.

Definición 3.2.3.d. [11] El coeficiente de correlación ρ para dos variables aleatorias x y y es:

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

Donde $Cov(x, y)$ es la covarianza de x, y . Y σ_x, σ_y son las desviaciones estándar de x y y , respectivamente.

Similaridad Coseno

Es una medida de la similaridad existente entre dos vectores en un espacio que posee un producto interior, con el que se evalúa el valor del coseno del ángulo comprendido entre ellos. Esta función trigonométrica proporciona un valor igual a 1 si el ángulo comprendido es cero, es decir si ambos vectores apuntan a un mismo lugar. Para cualquier ángulo existente entre los vectores, el coseno arrojaría un valor inferior a uno. Si los vectores fuesen ortogonales el coseno se anularía, y si apuntasen en sentido contrario su valor sería -1. De esta forma, el valor de esta similitud se encuentra en el intervalo cerrado $[-1, 1]$.

Esta distancia se emplea en la búsqueda y recuperación de información representando las palabras en un espacio vectorial. La similaridad coseno no debe ser considerada como una métrica debido a que no cumple la desigualdad del triángulo.

Definición 3.2.3.e. [12] Dados dos vectores a y b , la similaridad coseno se define usando el producto punto y la norma como:

$$\text{cosine}(a, b) = \frac{a \cdot b}{\|a\| \|b\|} = \frac{\sum_{i=1}^N a_i b_i}{\sqrt{\sum_{i=1}^N a_i^2} \sqrt{\sum_{i=1}^N b_i^2}}$$

Similaridad Dice

Está basada en el coeficiente de Dice para valores numéricos.

Distancia Dynamic Time Warping (DTW)

La “deformación” en tiempo dinámico (DTW, por sus siglas en inglés) es un algoritmo para medir la similitud entre dos secuencias temporales que pueden variar en el tiempo o la velocidad. DTW se ha aplicado a las secuencias temporales de vídeo, audio y datos de gráficos - de hecho, todos los datos que se pueden convertir en una secuencia lineal se pueden analizar con DTW. Una aplicación bien conocida ha sido el reconocimiento automático del habla, para hacer frente a diferentes velocidades de habla. Otras aplicaciones incluyen el reconocimiento del altavoz y el reconocimiento de la firma en línea. [13]

En general, DTW es un método que calcula una coincidencia óptima entre dos secuencias dadas con ciertas restricciones. Este método de alineación de secuencias se utiliza a menudo en la clasificación de series de tiempo.

La distancia DTW está mapeada a una medida de similitud, usando la función: [14]

$$f(x) = 1 - \left(\frac{x}{1+x}\right)$$

Similitud del Producto Interno

Está basada en el cálculo del producto interno entre dos vectores.

Definición 3.2.3.f. [15] Sean los vectores $P=(p_1, p_2, \dots, p_n)$ y $Q=(q_1, q_2, \dots, q_n)$, el producto interno se define como:

$$P \cdot Q = p_1q_1 + p_2q_2 + \dots + p_nq_n$$

Similitud de Jaccard

Es una variante del coeficiente de Jaccard para valores numéricos.

Distancia Euclidiana Kernel

Definición 3.2.3.g. [16] Sean P y Q vectores en R^d , La distancia Kernel entre P y Q es:

$$D_K^2(P, Q) \triangleq \sum_{p \in P} \sum_{p' \in P} K(p, p') + \sum_{q \in Q} \sum_{q' \in Q} K(q, q') - 2 \sum_{p \in P} \sum_{q \in Q} K(p, q)$$

Distancia Manhattan

La distancia Manhattan o “Taxicab”, d_1 , entre dos vectores $p=(p_1, p_2, \dots, p_n)$ y $q=(q_1, q_2, \dots, q_n)$ en un espacio vectorial real n -dimensional, con un sistema de coordenadas cartesianas fijo, es la suma de las longitudes de las proyecciones del segmento de línea entre los puntos sobre el sistema de ejes coordenados.

Definición 3.2.3.h. [17] La distancia Manhattan entre dos vectores $p=(p_1, p_2, \dots, p_n)$ y $q=(q_1, q_2, \dots, q_n)$ se define como:

$$d_1(p, q) = \|p - q\|_1 = \sum_{i=1}^n |p_i - q_i|$$

Similaridad del Producto Máximo

Es una similaridad especializada que toma el máximo producto de dos valores característicos. Si este valor es cero, la similaridad es indefinida. Esta medida de similaridad se utiliza principalmente con características extraídas de modelos de clúster. [14]

Similaridad Overlap

Es una variante de “Simple Matching Similarity” para atributos numéricos. [14]

Esta similaridad está relacionada con la probabilidad condicional. La idea principal es calcular el grado en el cual los conjuntos D y Q se traslapan entre sí (cuanto más cerca de 1, es mejor). Se calcula como el tamaño de la intersección de D y Q , entre el mínimo tamaño de los dos conjuntos:

Definición 3.2.3.i. [18] La similaridad Overlap se obtiene:

$$O(D, Q) = \frac{|D \cap Q|}{\min(|D|, |Q|)}$$

3.2.4 Medidas Mixtas

Para trabajar con datos que incluyen variables numéricas y nominales, Rapidminer ofrece las medidas mixtas, específicamente la Distancia Euclidiana Mixta.

Distancia Euclidiana Mixta

Es la distancia Euclidiana para valores numéricos y nominales. Para valores nominales, una distancia se cuenta si los valores son diferentes. En muchos casos, se deben normalizar los valores numéricos para obtener buenos resultados.

4. RAPIDMINER

RapidMiner es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico. Se usa en investigación y en aplicaciones empresariales. [2]

El desarrollo de RapidMiner inició bajo el nombre de “Yet Another Learning Environment” (YALE) en el departamento de Inteligencia Artificial de la Universidad de Dortmund, Alemania, bajo la dirección de la Dra. Katharina Morik. El software se volvió más y más robusto conforme pasó el tiempo, más de un millón y medio de descargas se han registrado desde que inició su desarrollo, en 2001. [2]

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, pre procesamiento de datos y visualización. Puede ser descargado desde el sitio: <http://www.rapidminer.com>

CARACTERÍSTICAS

- Desarrollado en Java
- Multiplataforma
- Representación interna de los procesos de análisis de datos en archivos XML
- Permite el desarrollo de programas a través de un lenguaje script
- Puede usarse de diversas maneras:
 - A través de un GUI
 - En línea de comandos
 - En batch
 - Desde otros programas a través de llamadas a sus bibliotecas
- Extensible
- Incluye gráficos y herramientas de visualización de datos
- Dispone de un módulo de integración con R

RapidMiner es el líder mundial de código abierto para la minería de datos debido a la combinación de su tecnología de primera calidad y su rango de funcionalidad. Cubre un amplio rango de conceptos de minería de datos, además de ser una herramienta flexible para aprender y explorar, la interfaz gráfica de usuario tiene como objetivo simplificar el uso para las tareas complejas de esta área.

PREPARACIÓN DE LOS DATOS

Los datos de entrada en RapidMiner deben almacenarse en la carpeta *Local Repository*, para que el sistema les dé el formato adecuado para su procesamiento. Se pueden importar datos en diferentes formatos: Excel, CSV, XML, binarios; así como tablas de bases de datos (*MySQL*, *PostgreSQL*, *Ingress*, *Oracle*, etc.)

En este trabajo, los datos se manejarán en formato *Excel*.

LA INTERFAZ DE RapidMiner 5.3

RapidMiner proporciona una interfaz gráfica de usuario muy sencilla. Si se desea crear un nuevo proceso, basta con seleccionar el icono *New Process*, y se abrirá la ventana de diseño. Se describirán a continuación sus principales elementos.

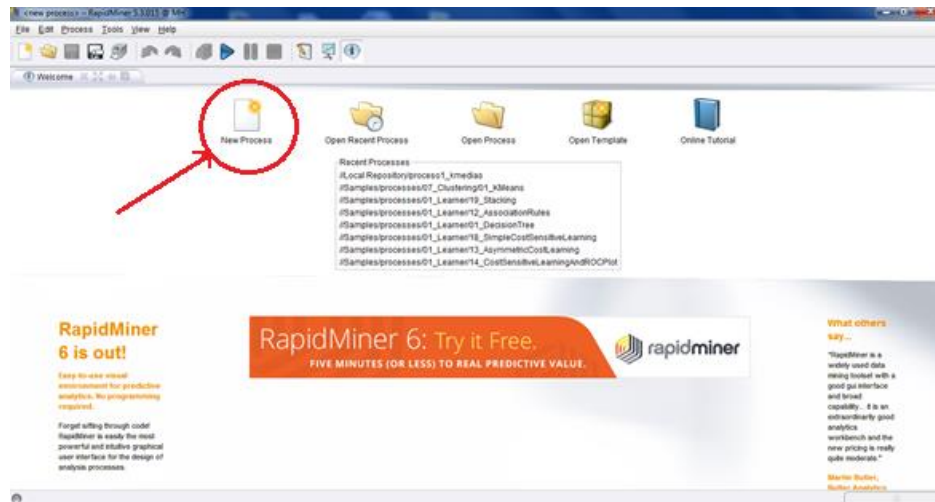


Figura 4.1. La interfaz principal de RapidMiner

En la ventana para crear nuevos procesos, se pueden observar diferentes pestañas: Operadores, Repositorios, Procesos, Parámetros, Ayuda, Comentarios, etc. (Ver figura 4.1)

En esta ventana se llevará a cabo el diseño de los procesos de minería de datos, desde la preparación de los datos hasta la selección y configuración de la técnica que se desee usar.

El diseño se lleva a cabo en la ventana principal, llamada “Main Process”

K-Medias en RapidMiner

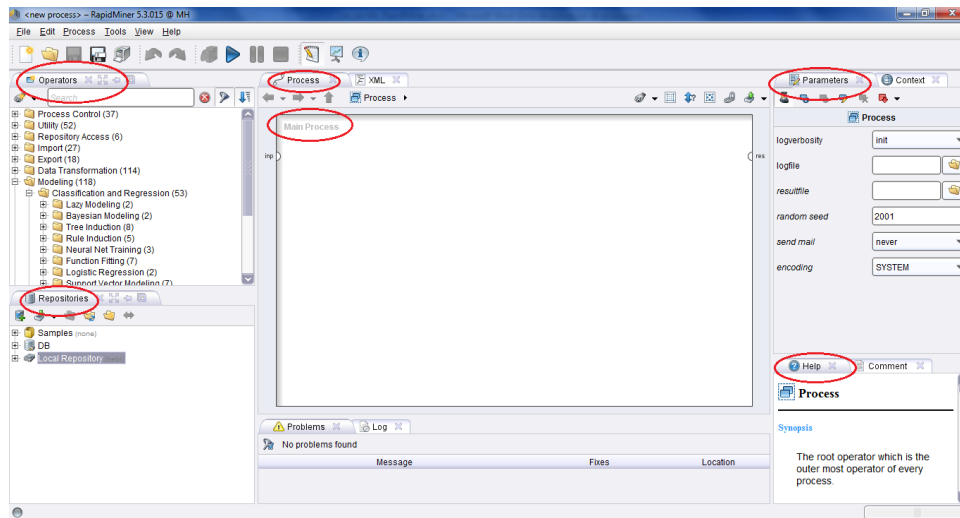


Figura 4.2 Principales componentes de RapidMiner

Para iniciar un proceso, se necesita “abrir” la base de datos, esto se hace en la ventana *Operators*, dentro de la carpeta llamada *Repository Access*, se elige el operador *Retrieve* y se arrastra a la ventana *Main Process*.

Una vez que se tiene el *Retrieve* en la ventana de diseño, se procede a localizar el archivo que contiene los datos, dando click en la carpeta *Repository entry*, ubicada en el panel derecho de la ventana principal. (Ver figura 4.3)

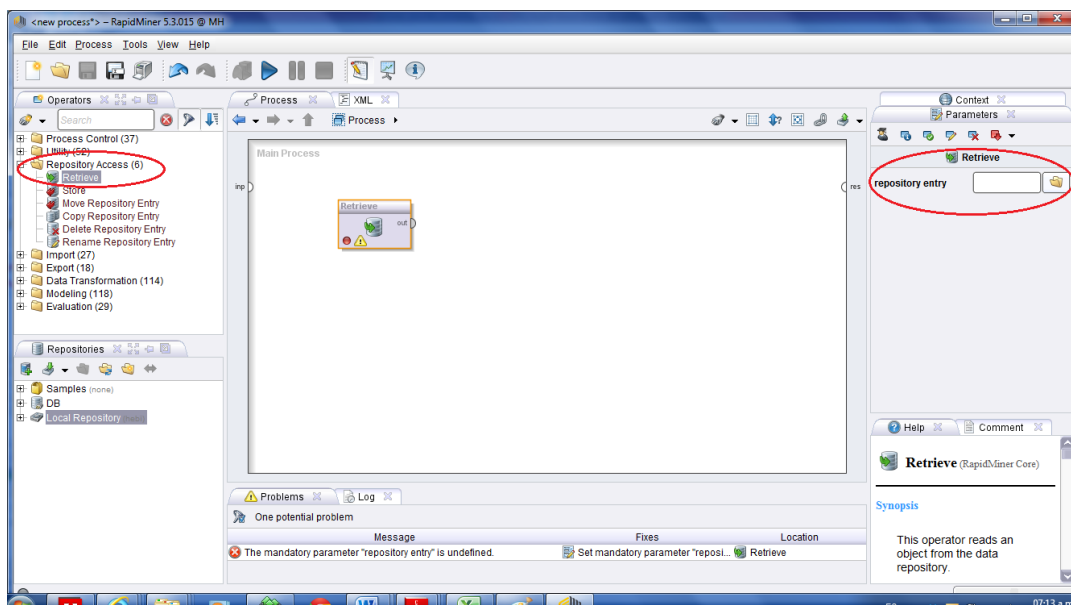


Figura 4.3 El operador Retrieve


A partir de este paso, ya se puede elegir el operador que se desee para realizar minería de datos.

4.1 K-medias en RapidMiner

Ahora se describirá el funcionamiento del algoritmo K-medias en RapidMiner, usando el ejemplo 1 (pág. 12) analizado anteriormente.

Los datos de entrada de Rapidminer se pueden importar de diferentes aplicaciones. Particularmente, se trabajará con datos importados desde Excel. En la opción *File* de la barra de menús, seleccionar la opción *Import Data* → *Import Excel Sheet*; seleccionar el archivo en formato Excel y seguir los pasos para completar la importación. Es muy importante guardar el archivo resultante en la carpeta *Local Repository*, ya que RapidMiner sólo considera como archivos de entrada listos para usarse aquellos que están alojados en esta ubicación

Una vez que se tiene en la carpeta *Local Repository* el archivo con el que se va a trabajar, se realizan los siguientes pasos: [2]

1. Agregar el operador **Repository Access** → **Retrieve** a la zona de trabajo y localizar el archivo */Local Repository/ejemplo1_kmedias* con el navegador del parámetro *repository entry*.
2. Agregar el operador **Modeling** → **Clustering and Segmentation** → **k-Means**. El parámetro *k* indicará el número de particiones que se desea obtener. En este caso $k=2$. Conectar la salida del operador **Retrieve** a la entrada **exa** de este operador y la salida **clu** (cluster model) de este último al conector **res** del panel.
3. Ejecutar el proceso dando click en el icono  de la barra de herramientas.

En la figura 4.4 se muestra el proceso.

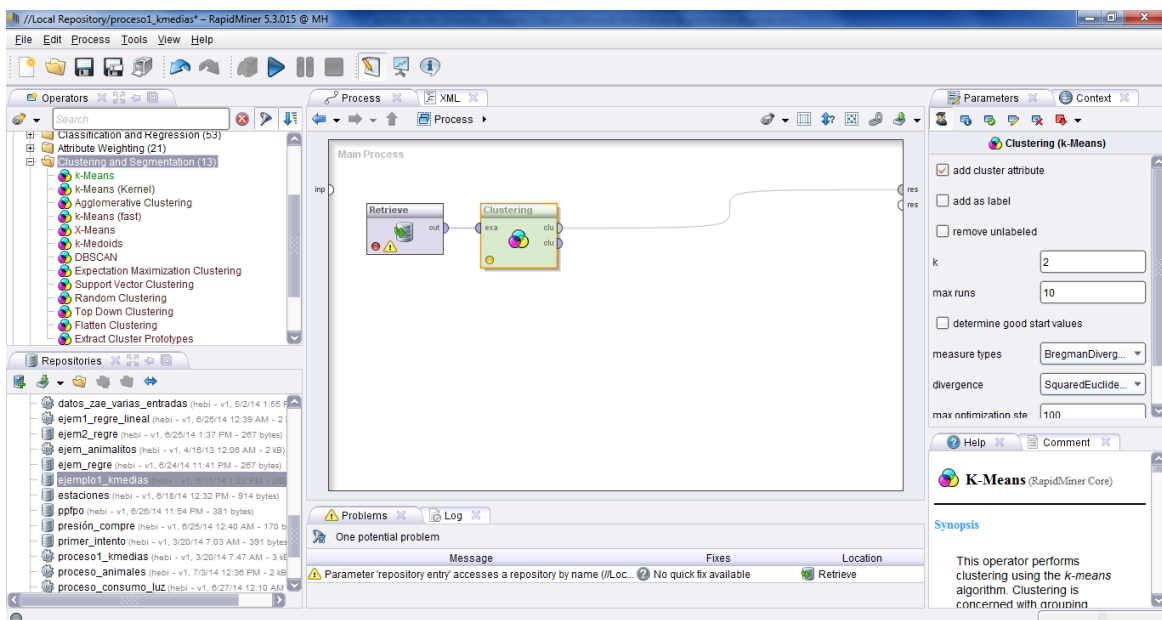


Figura 4.4 K-medias en RapidMiner

4. Observar e interpretar los resultados. (Ver figura 4.4. a-c)

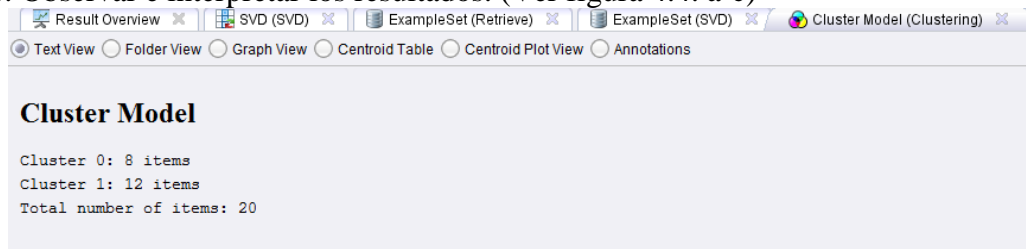


Figura 4.4.a. Resultado final de asignación de cluster para el ejemplo 1.

The screenshot shows the 'Data View' window with a table of 20 rows. The columns are 'Row No.', 'id', 'cluster', 'X1', and 'X2'. The data is as follows:

Row No.	id	cluster	X1	X2
1	1	cluster_0	0	0
2	2	cluster_0	1	0
3	3	cluster_0	0	1
4	4	cluster_0	1	1
5	5	cluster_0	2	1
6	6	cluster_0	1	2
7	7	cluster_0	2	2
8	8	cluster_0	2	3
9	9	cluster_1	6	6
10	10	cluster_1	7	6
11	11	cluster_1	8	6
12	12	cluster_1	6	7
13	13	cluster_1	7	7
14	14	cluster_1	8	7
15	15	cluster_1	9	7
16	16	cluster_1	7	8
17	17	cluster_1	8	8
18	18	cluster_1	9	8
19	19	cluster_1	10	8

Figura 4.4.b. Lista de asignación de cada elemento del ejemplo 1.

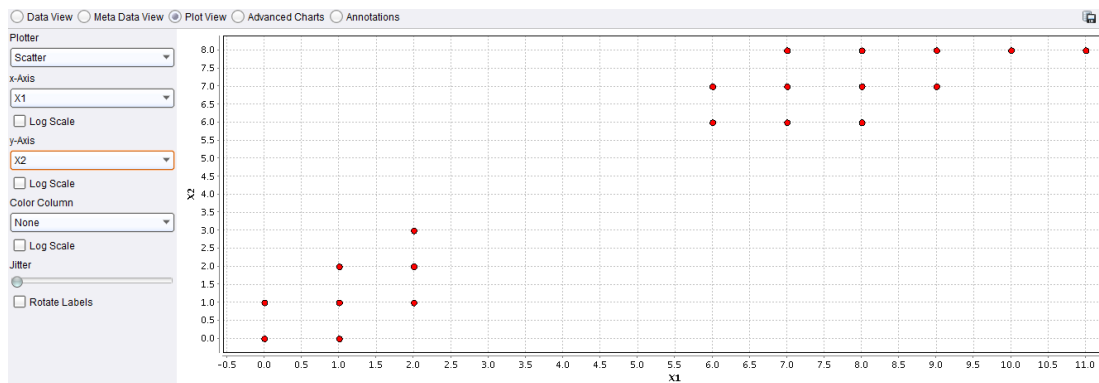


Figura 4.4.c. Gráfica de salida para el ejemplo 1.

Se puede observar que se obtuvieron los mismos resultados de asignación que en ejemplo 1, el cluster 0 tiene 8 elementos y el cluster 1 tiene 12.

5. PRUEBAS

En este apartado se presentan las pruebas realizadas en RapidMiner. Los dos primeros ejemplos son datos numéricos y se llevó a cabo la ejecución con Distancias de Bregman y con Medidas Numéricas. El ejemplo tres consiste en un conjunto de datos nominales y mixtos.

5.1 Ejemplo 1: Ejemplo 1 (20 datos, 2 atributos)

Se muestra a continuación el resultado de las pruebas en RapidMiner para el ejemplo 1 (pág. 12), con $k=2$.

Distancias de Bregman

a) Squared Euclidean Distance

TABLA DE CENTROIDES	cluster 0	cluster 1
X1	1.125	8
X2	1.25	7.166667

Clusters obtenidos	N° de elementos	ELEMENTOS
CLUSTER 0	8	1, 2, 3, 4, 5, 6, 7, 8
CLUSTER 1	12	9, 10 ..., 20
total de ítems	20	

b) Mahalanobis Distance

TABLA DE CENTROIDES	cluster 0	cluster 1
X1	1.125	8
X2	1.25	7.16666667

Clusters obtenidos	N° de elementos	ELEMENTOS
CLUSTER 0	8	1, 2, 3, 4, 5, 6, 7, 8
CLUSTER 1	12	9, 10 ..., 20
Total de ítems	20	

Las restantes Distancias de Bregman no son aplicables para este conjunto de datos.

5.2 Ejemplo 2: ZAE (22 datos, 9 atributos)

Zonificación Agro Ecológica

La Zonificación Agro Ecológica (ZAE) es la división de un área en unidades más pequeñas, que tienen similares características relacionadas con su aptitud y potencial de producción.

El propósito de zonificar es separar áreas con similares potencialidades y limitantes para el desarrollo y la planificación del uso de recursos rurales. Así, se pueden formular programas específicos de desarrollo para proporcionar el apoyo más efectivo a cada zona.

Para este ejemplo, se considera la zona de estudio que abarca la región sur del estado de Puebla, con una extensión de 8775.8 km² y abarca 39 municipios de la entidad. Para la caracterización climática, se localizaron las estaciones meteorológicas que se encontraron dentro de la zona de estudio y que tuviesen información completa de los registros climatológicos por más de diez años. Se determinaron 22 estaciones, de las cuales se analizaron 9 variables. [4]

Los datos se muestran en la tabla siguiente:

	Estaciones	Temp_media	Temp_max	Temp_min	Precipitación	Evapotraspiración	No. Días lluvia	Niebla	Granizo	Altitud
1	Jojalpan	25.9	35.7	16.1	796.6	1727.28	66.9	22.7	0.7	840
2	Teotlalco	23.8	33.4	14.3	600.9	1528.91	45.4	0	0	806
3	Ixcamilpa	26.1	34.4	17.8	777	1750.34	76.3	39.1	0	921
4	Axutla	24.1	32.7	15.4	691.3	1556.05	57.5	28.8	0.5	996
5	Anonas	23.4	30.6	16.2	625	1491.05	50.2	2.2	0	1070
6	Piaxtla	24.3	32.8	15.6	1067	1579.84	72.9	0.6	0.2	1190
7	Las Peñas	23.8	31.9	15.7	624	1531.1	47.9	0.1	0	1190
8	Acatlán	23.8	31.4	16.1	608.2	1530.44	57.2	2.8	0.3	1200
9	Acatlán_1	23.8	32.7	14.9	658.7	1526.56	49.2	3.4	0.3	1230
10	Tonahuixtla	22.1	29.3	14.9	237.4	1376.54	22.2	0	0	1140
11	Santa Ana Tepejillo	20.2	28.8	11.6	337.7	1211.99	37.7	3.6	0	950
12	Zapotitlán	20.5	29.8	11.2	401.2	1235.79	42.6	0	0	1090
13	Zapotitlán_1	20.7	29.4	12	422	1251.21	47	0	0	1275
14	Caltepec	18.4	26.7	10.2	394.6	1062.72	44.6	0	0.1	1441
15	Acatepec	16.4	23.1	9.8	558.2	901.48	48.6	81.1	0	1450
16	Altepeixi	22.7	31.1	14.4	304.7	1436.19	43.5	0	0.3	1500
17	Zinacatepec	23.7	32.7	14.7	249.1	1523.24	26.1	0	0	1972
18	Axusco	23.3	31.4	15.2	361.6	1489.86	61.1	77.8	0.4	2016
19	Calipán	22.4	29.9	14.9	384.7	1405.39	48.8	4.3	0	2035
20	Alcomunga	11.3	17.7	4.9	2496.7	528.11	136.4	3.8	0.8	2520
21	Zoquitlán	15.3	20.7	9.8	1650.3	812.97	115.8	8.3	0.1	2260
22	Tlacotepec Díaz	23.9	29.8	18	3100.9	1544.61	162.6	3.6	0.1	390

A continuación se muestran los resultados obtenidos aplicando el algoritmo de las K-Medias, con $k=5$, utilizando las Distancias de Bregman y las Medidas Numéricas en RapidMiner.

Distancias de Bregman

a) Squared Euclidean Distance

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	23.13333333	24.33333333	13.3	23.9	20.14285714
Temp_max	31.33333333	32.84444444	19.2	29.8	28.31428571
Temp_min	14.93333333	15.78888889	7.35	18	12.01428571
Precipitación	331.8	716.5222222	2073.5	3100.9	379.4
Evapotranspiración	1472.83	1580.174444	670.54	1544.61	1210.845714
No. Días lluvia	45.33333333	58.16666667	126.1	162.6	40.88571429
Niebla	27.36666667	11.07777778	6.05	3.6	12.1
Granizo	0.13333333	0.22222222	0.45	0.1	0.057142857
Altitud	2007.666667	1049.222222	2390	390	1263.714286

CLUSTERS OBTENIDOS	N°. Elementos	Elementos
Cluster 0	3	17, 18, 19
Cluster 1	9	1, 2, 3, 4, 5, 6, 7, 8, 9
Cluster 2	2	20, 21
Cluster 3	1	22
Cluster 4	7	10, 11, 12, 13, 14, 15, 16
Total de items	22	

b) Mahalanobis Distance

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	22.56	26.1	11.3	23.9	21.44444444
Temp_max	30.88	34.4	17.7	29.8	29.47777778
Temp_min	14.26	17.8	4.9	18	13.37777778
Precipitación	559.98	777	2496.7	3100.9	597.0444444
Evapotranspiración	1425.054	1750.34	528.11	1544.61	1325.341111
No. Días lluvia	52.42	76.3	136.4	162.6	51.22222222
Niebla	21.88	39.1	3.8	3.6	1.877777778
Granizo	0.26	0	0.8	0.1	0.033333333
Altitud	1254.9	921	2520	390	1455.777778

CLUSTERS OBTENIDOS	N°. Elementos	Elementos
Cluster 0	10	1, 2, 4, 5, 8, 9, 14, 15, 16, 18
Cluster 1	1	3
Cluster 2	1	20
Cluster 3	1	22
Cluster 4	9	6, 7, 10, 11, 12, 13, 17, 19, 21
Total de items	22	

Las restantes Distancias de Bregman no son aplicables para este conjunto de datos.

Medidas Numéricas

a) Distancia Euclidiana

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	23.13333333	24.33333333	13.3	23.9	20.1428571
Temp_max	31.33333333	32.84444444	19.2	29.8	28.3142857
Temp_min	14.93333333	15.78888889	7.35	18	12.0142857
Precipitación	331.8	716.5222222	2073.5	3100.9	379.4
Evapotranspiración	1472.83	1580.174444	670.54	1544.61	1210.84571
No. Días lluvia	45.33333333	58.16666667	126.1	162.6	40.8857143
Niebla	27.36666667	11.07777778	6.05	3.6	12.1
Granizo	0.13333333	0.22222222	0.45	0.1	0.05714286
Altitud	2007.666667	1049.222222	2390	390	1263.71429

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	3	17, 18, 19
Cluster 1	9	1, 2, 3, 4, 5, 6, 7, 8, 9
Cluster 2	2	20, 21
Cluster 3	1	22
Cluster 4	7	10, 11, 12, 13, 16
Total items:	22	

b) Distancia Camberra

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	23.4545455	?	13.3	23.9	21.425
Temp_max	31.8727273	?	19.2	29.8	29.65
Temp_min	15.0272727	?	7.35	18	13.2125
Precipitación	640.745455	?	2073.5	3100.9	381.4625
Evapotranspiración	1505.27727	?	670.54	1544.61	1319.74125
No. Días lluvia	56.9545455	?	126.1	162.6	39.9
Niebla	23.3090909	?	6.05	3.6	1.2625
Granizo	0.24545455	?	0.45	0.1	0.0125
Altitud	1212.63636	?	2390	390	1371.625

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	11	1, 2, 3, 4, 6, 7, 8, 9, 15, 16, 18
Cluster 1	0	
Cluster 2	2	20, 21
Cluster 3	1	22
Cluster 4	8	5, 10, 11, 12, 13, 14, 17, 19
Total items:	22	

c) Distancia Chebychev

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	23.13333333	23.9	13.3	23.92	20.13333333
Temp_max	31.33333333	29.8	19.2	32.44	28.23333333
Temp_min	14.93333333	18	7.35	15.37	12.08333333
Precipitación	331.8	3100.9	2073.5	678.64	386.35
Evapotranspiración	1472.83	1544.61	670.54	1543.36	1210.655
No. Días lluvia	45.33333333	162.6	126.1	56.12	41.4166667
Niebla	27.36666667	3.6	6.05	10.33	13.5166667
Granizo	0.13333333	0.1	0.45	0.2	0.06666667
Altitud	2007.666667	390	2390	1039.3	1316

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	3	17, 18, 19
Cluster 1	1	22
Cluster 2	2	20, 21
Cluster 3	10	1, 2, 3, 4, 5, 6, 7, 8, 9, 11
Cluster 4	6	10, 12, 13, 14, 15, 16
Total items:	22	

d) Similitud por Correlación

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	25.2666667	21.15	23.9333	16.8333	22.1285714
Temp_max	34.5	29.15	32.0333	22.7333	30.4714286
Temp_min	16.0666667	13.2	15.7333	10.9	13.7714286
Precipitación	724.833333	375.483333	794.433	2415.97	469.885714
Evapotranspiración	1668.84333	1303.14667	1542.31	961.897	1380.51857
No. Días lluvia	62.8666667	45.45	60.2	138.267	43.4
Niebla	20.6	27.2	10.5333	5.23333	1.41428571
Granizo	0.23333333	0.13333333	0.23333	0.33333	0.08571429
Altitud	855.666667	1735.66667	1085.33	1723.33	1153.57143

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	3	1, 2, 3
Cluster 1	6	14, 15, 16, 17, 18, 19
Cluster 2	3	4, 5, 6
Cluster 3	3	20, 21, 22
Cluster 4	7	7, 8, 9, 10, 11, 12, 13
Total items:	22	

e) Similitud Coseno

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	24.84	22.2875	13.3	23.9	21.15
Temp_max	33.8	30.4875	19.2	29.8	29.15
Temp_min	15.84	14.075	7.35	18	13.2
Precipitación	786.56	489.275	2073.5	3100.9	375.483333
Evapotranspiración	1628.484	1394.335	670.54	1544.61	1303.14667
No. Días lluvia	63.8	44.25	126.1	162.6	45.45
Niebla	18.24	1.5125	6.05	3.6	27.2
Granizo	0.28	0.075	0.45	0.1	0.13333333
Altitud	950.6	1143.125	2390	390	1735.66667

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	5	1, 2, 3, 4, 6
Cluster 1	8	5, 7, 8, 9, 10, 11, 12, 13
Cluster 2	2	20, 21
Cluster 3	1	22
Cluster 4	6	14, 15, 16, 17, 18
Total items:	22	

f) Similitud de Dice

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	21.8136364	?	?	?	?
Temp_max	29.8181818	?	?	?	?
Temp_min	13.8045455	?	?	?	?
Precipitación	788.536364	?	?	?	?
Evapotranspiración	1363.71227	?	?	?	?
No. Días lluvia	61.8409091	?	?	?	?
Niebla	12.8272727	?	?	?	?
Granizo	0.17272727	?	?	?	?
Altitud	1340.09091	?	?	?	?

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	22	
Cluster 1	0	
Cluster 2	0	
Cluster 3	0	
Cluster 4	0	
Total items:	22	

g) Distancia DTW

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	23.025	23.9	13.3	24.3333	19.7166667
Temp_max	31.275	29.8	19.2	32.8444	27.85
Temp_min	14.8	18	7.35	15.7889	11.6166667
Precipitación	325.025	3100.9	2073.5	716.522	391.85
Evapotranspiración	1463.67	1544.61	670.54	1580.17	1173.28833
No. Días lluvia	44.875	162.6	126.1	58.1667	40.45
Niebla	20.525	3.6	6.05	11.0778	14.1166667
Granizo	0.175	0.1	0.45	0.22222	0.01666667
Altitud	1880.75	390	2390	1049.22	1224.33333

CLUSTERS OBTENIDOS	N°. Elementos	Elementos
Cluster 0	4	16, 17, 18, 19
Cluster 1	1	22
Cluster 2	2	20, 21
Cluster 3	9	1, 2, 3, 4, 5, 6, 7, 8, 9
Cluster 4	6	10, 11, 12, 13, 14, 15
Total items:	22	

h) Similaridad del Producto Interno

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	?	?	21.8136	?	?
Temp_max	?	?	29.8182	?	?
Temp_min	?	?	13.8045	?	?
Precipitación	?	?	788.536	?	?
Evapotranspiración	?	?	1363.71	?	?
No. Días lluvia	?	?	61.8409	?	?
Niebla	?	?	12.8273	?	?
Granizo	?	?	0.17273	?	?
Altitud	?	?	1340.09	?	?

CLUSTERS OBTENIDOS	N°. Elementos	Elementos
Cluster 0	0	
Cluster 1	0	
Cluster 2	22	
Cluster 3	0	
Cluster 4	0	
Total items:	22	

i) Similaridad de Jaccard

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	24.975	?	18.7	?	21.8
Temp_max	34.05	?	25.25	?	29.914286
Temp_min	15.9	?	12.075	?	13.7
Precipitación	716.45	?	2078.7	?	440.50714
Evapotranspiración	1640.645	?	1116.4	?	1355.2543
No. Días lluvia	61.525	?	121.93	?	44.764286
Niebla	22.65	?	4.075	?	12.521429
Granizo	0.3	?	0.3	?	0.1
Altitud	890.75	?	1590	?	1397.0714

CLUSTERS OBTENIDOS	N°. Elementos	Elementos
Cluster 0	4	1, 2, 3, 4
Cluster 1	0	
Cluster 2	4	6, 20, 21, 22
Cluster 3	0	
Cluster 4	14	5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19
Total items:	22	

j) Distancia Euclidiana Kernel

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	21.794118	26.1	11.3	23.9	24.3
Temp_max	29.964706	34.4	17.7	29.8	32.8
Temp_min	13.629412	17.8	4.9	18	15.6
Precipitación	546.01176	777	2496.7	3100.9	1067
Evapotranspiración	1356.9218	1750.34	528.11	1544.61	1579.84
No. Días lluvia	50.847059	76.3	136.4	162.6	72.9
Niebla	13.823529	39.1	3.8	3.6	0.6
Granizo	0.1588235	0	0.8	0.1	0.2
Altitud	1368.8824	921	2520	390	1190

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	18	1, 2, 4, 5, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21
Cluster 1	1	3
Cluster 2	1	20
Cluster 3	1	22
Cluster 4	1	6
Total items:	22	

k) Distancia Manhattan

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	23.13333333	23.9	13.3	24.333	20.142857
Temp_max	31.33333333	29.8	19.2	32.844	28.314286
Temp_min	14.93333333	18	7.35	15.789	12.014286
Precipitación	331.8	3100.9	2073.5	716.52	379.4
Evapotranspiración	1472.83	1544.61	670.54	1580.2	1210.8457
No. Días lluvia	45.33333333	162.6	126.1	58.167	40.885714
Niebla	27.36666667	3.6	6.05	11.078	12.1
Granizo	0.13333333	0.1	0.45	0.2222	0.0571429
Altitud	2007.666667	390	2390	1049.2	1263.7143

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	3	17, 18, 19
Cluster 1	1	22
Cluster 2	2	21, 22
Cluster 3	9	1, 2, 3, 4, 5, 6, 7, 8, 9
Cluster 4	7	10, 11, 12, 13, 14, 15, 16
Total items:	22	

l) Similitud del Producto Máximo

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	21.8136364	?	?	?	?
Temp_max	29.8181818	?	?	?	?
Temp_min	13.8045455	?	?	?	?
Precipitación	788.536364	?	?	?	?
Evapotranspiración	1363.71227	?	?	?	?
No. Días lluvia	61.8409091	?	?	?	?
Niebla	12.8272727	?	?	?	?
Granizo	0.17272727	?	?	?	?
Altitud	1340.09091	?	?	?	?

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	22	
Cluster 1	0	
Cluster 2	0	
Cluster 3	0	
Cluster 4	0	
Total items:	22	

m) Similaridad Overlap

TABLA DE CENTROIDES	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Temp_media	24.583333	22.3833	23.9	14.333	21.866667
Temp_max	33.133333	30.7833	29.8	20.5	30.2
Temp_min	16	13.9667	18	8.1667	13.566667
Precipitación	692.63333	544.867	3100.9	1568.4	352.78333
Evapotranspiración	1603.6283	1404.02	1544.6	747.52	1361.435
No. Días lluvia	59.166667	45.1667	162.6	100.27	45.183333
Niebla	16.15	1.06667	3.6	31.067	13.683333
Granizo	0.3	0.03333	0.1	0.3	0.133333
Altitud	1062.8333	1041	390	2076.7	1706.5

CLUSTERS OBTENIDOS	Nº. Elementos	Elementos
Cluster 0	6	1, 3, 4, 7, 8, 9
Cluster 1	6	2, 5, 6, 10, 11, 12
Cluster 2	1	22
Cluster 3	3	15, 20, 21
Cluster 4	6	13, 14, 16, 17, 18, 19
Total items:	22	

En este ejemplo se pudo observar que con la Distancia Euclidiana Cuadrada y la distancia de Mahalanobis se obtuvieron resultados muy parecidos a los mostrados en [4], la elección de los clusters iniciales puede hacer la diferencia en el resultado final.

Para el caso de medidas numéricas, los clusters varían en tamaño y contenido, pero se mantiene el resultado esperado por el experto, agrupando a los elementos “más alejados” del resto en un solo cluster.

5.3 Ejemplo 3: Animales (100 datos, 15 atributos)

El siguiente ejemplo es una base de datos que contiene animales, descritos a través de características que se pueden evaluar como falsas o verdaderas. Incluye variables numéricas y nominales. [19] Los datos se muestran en la siguiente tabla:

K-Medias en RapidMiner

Num.	pele	plumas	huevos	leche	vuela	acuatico	depredador	dentado	columna	respira	venenoso	aletas	patas	cola	doméstico
1	true	false	false	true	false	false	true	true	true	true	false	false	4	false	false
2	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
3	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
4	true	false	false	true	false	false	true	true	true	true	false	false	4	false	false
5	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
6	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
7	true	false	false	true	false	false	false	true	true	true	false	false	4	true	true
8	false	false	true	false	false	true	false	true	true	false	false	true	0	true	true
9	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
10	true	false	false	true	false	false	false	true	true	true	false	false	4	false	true
11	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
12	false	true	true	false	true	false	false	false	true	true	false	false	2	true	true
13	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
14	false	false	true	false	false	false	true	false	false	false	false	false	0	false	false
15	false	false	true	false	false	true	true	false	false	false	false	false	4	false	false
16	false	false	true	false	false	true	true	false	false	false	false	false	6	false	false
17	false	true	true	false	true	true	false	true	true	true	false	false	2	true	false
18	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
19	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
20	false	false	false	true	false	true	true	true	true	true	false	true	0	true	false
21	false	true	true	false	true	false	false	false	true	true	false	false	2	true	true
22	false	true	true	false	true	true	false	false	true	true	false	false	2	true	false
23	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
24	false	true	true	false	true	false	false	false	true	true	false	false	2	true	false
25	false	false	true	false	false	false	false	false	false	true	false	false	6	false	false
26	false	false	true	false	false	true	true	true	true	true	false	false	4	false	false
27	true	false	false	true	true	false	false	true	true	true	false	false	2	true	false
28	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
29	true	false	false	true	false	false	true	true	true	true	false	false	2	false	true
30	false	false	true	false	true	false	false	false	false	true	false	false	6	false	false
31	true	false	false	true	false	false	false	true	true	true	false	false	4	true	true
32	true	false	false	true	false	false	false	true	true	true	false	false	2	false	false
33	false	true	true	false	true	true	true	false	true	true	false	false	2	true	false
34	false	false	true	false	false	true	false	true	true	false	false	true	0	true	false
35	true	false	false	true	false	false	false	true	true	true	false	false	4	true	true
36	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
37	false	true	true	false	true	false	true	false	true	true	false	false	2	true	false
38	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
39	true	false	true	false	true	false	false	false	false	true	true	false	6	false	true
40	true	false	true	false	true	false	false	false	false	true	false	false	6	false	false
41	false	true	true	false	false	false	true	false	true	true	false	false	2	true	false
42	false	false	true	false	true	false	true	false	false	true	false	false	6	false	false
43	false	true	true	false	true	false	false	false	true	true	false	false	2	true	false
44	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
45	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
46	false	false	true	false	false	true	true	false	false	false	false	false	6	false	false
47	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
48	true	false	false	true	false	true	true	true	true	true	false	false	4	true	false
49	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
50	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false

K-Medias en RapidMiner

Num.	peló	plumas	huevos	leche	vuela	acuatico	depredador	dentado	columna	respira	venenoso	aletas	patas	cola	doméstico
51	true	false	true	false	true	false	false	false	false	true	false	false	6	false	false
52	false	false	true	false	false	true	true	true	true	true	false	false	4	true	false
53	false	false	true	false	false	true	true	false	false	false	false	false	8	false	false
54	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
55	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
56	false	true	true	false	false	false	false	false	true	true	false	false	2	true	false
57	false	true	true	false	true	false	false	false	true	true	false	false	2	true	true
58	false	true	true	false	false	true	true	false	true	true	false	false	2	true	false
59	false	true	true	false	true	false	false	false	true	true	false	false	2	true	false
60	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
61	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
62	false	false	true	false	false	false	true	true	true	true	true	false	0	true	false
63	true	false	true	true	false	true	true	false	true	true	false	false	4	true	false
64	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
65	true	false	false	true	false	false	false	true	true	true	false	false	4	true	true
66	false	false	false	true	false	true	true	true	true	true	false	true	0	true	false
67	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
68	true	false	false	true	false	false	true	true	true	true	false	false	4	true	true
69	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
70	true	false	false	true	false	false	false	true	true	true	false	false	4	true	true
71	false	true	true	false	false	false	true	false	true	true	false	false	2	true	false
72	false	false	false	false	false	false	true	false	false	true	true	false	8	true	false
73	false	false	true	false	false	true	false	true	true	false	false	true	0	true	false
74	true	false	false	true	false	true	true	true	true	true	false	true	0	false	false
75	true	false	false	true	false	true	true	true	true	true	false	true	2	true	false
76	false	false	false	false	false	true	true	true	true	false	true	false	0	true	false
77	false	false	true	false	true	true	true	false	false	false	true	false	0	false	false
78	false	true	true	false	true	true	true	false	true	true	false	false	2	true	false
79	false	true	true	false	true	true	true	false	true	true	false	false	2	true	false
80	false	false	true	false	false	false	true	true	true	true	false	false	0	true	false
81	false	false	true	false	false	false	false	false	false	true	false	false	0	false	false
82	false	false	true	false	false	true	false	true	true	false	false	true	0	true	false
83	false	true	true	false	true	false	false	false	true	true	false	false	2	true	false
84	true	false	false	true	false	false	false	true	true	true	false	false	2	true	false
85	false	false	true	false	false	true	true	false	false	false	false	false	5	false	false
86	false	false	true	false	false	true	true	true	true	false	true	true	0	true	false
87	false	true	true	false	true	true	false	false	true	true	false	false	2	true	false
88	false	false	true	false	false	false	false	false	false	true	false	false	6	false	false
89	false	false	true	false	false	true	false	true	true	true	false	false	4	false	false
90	false	false	true	false	false	false	false	false	true	true	false	false	4	true	false
91	false	false	true	false	false	false	true	true	true	true	false	false	4	true	false
92	false	false	true	false	false	true	true	true	true	false	false	true	0	true	false
93	true	false	false	true	true	false	false	true	true	true	false	false	2	true	false
94	true	false	false	true	false	false	false	true	true	true	false	false	4	true	false
95	false	true	true	false	true	false	true	false	true	true	false	false	2	true	false
96	true	false	false	true	false	false	false	true	true	true	false	false	2	true	false
97	true	false	true	false	true	false	false	false	false	true	true	false	6	false	false
98	true	false	false	true	false	false	true	true	true	true	false	false	4	true	false
99	false	false	true	false	false	false	false	false	false	true	false	false	0	false	false
100	false	true	true	false	true	false	false	false	true	true	false	false	2	true	false

Inicialmente se usaron las medidas nominales, cambiando el atributo “patas” como nominal, en lugar de entero.

A continuación se muestran los resultados obtenidos aplicando el algoritmo de las K-Medias, con $k=7$.

Medidas Nominales

a) Similaridad de Dice

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	0.4	0.1176471	0.23809524	0.875	1	1	1
plumas	0	0	0	0.8333333	0	0	0
huevos	0.6666667	0	0.23809524	0.9166667	0.933333	0	1
leche	0.6666667	0	0.23809524	0.875	1	1	1
vuela	0.4	0	0	0.75	0	0	0
acuatico	0	0.2941176	0.23809524	0.2916667	0.8	0	0.8571429
depredador	0.7333333	0	0.95238095	0.5833333	0	0	0
dentado	0.6666667	0	0	0.9166667	0	1	1
columna	0.6666667	0	0	0	0	1	1
respira	0	0	0.19047619	0	0.666667	0	1
venenoso	0.1333333	0	0	0	0.2	1	0.1428571
aletas	0	0.2352941	0.19047619	0	0.6	0	0
patas	0.1333333	0.1764706	0.19047619	0	0.8	0	0.2857143
cola	0	0.9411765	0.95238095	1	0.933333	1	0
doméstico	0.2	0	0.3333333	0.125	0	0	0

Clusters	Nº. Elementos	Elementos
Cluster 0:	15	1, 4, 10, 25, 29, 30, 32, 39, 40, 42, 51, 81, 88, 97, 99
Cluster 1:	17	5, 11, 20, 44, 45, 47, 48, 49, 50, 54, 64, 66, 67, 69, 74, 75, 98
Cluster 2:	21	2, 6, 7, 8, 18, 23, 28, 31, 34, 35, 36, 55, 65, 68, 70, 73, 82, 84, 89, 94, 96
Cluster 3:	24	12, 17, 21, 22, 24, 27, 33, 37, 41, 43, 56, 57, 58, 59, 63, 71, 78, 79, 83, 87, 90, 93, 95, 100
Cluster 4:	15	3, 9, 13, 19, 26, 38, 52, 60, 61, 62, 76, 80, 86, 91, 92
Cluster 5:	1	72
Cluster 6:	7	14, 15, 16, 46, 53, 77, 85
Total Items:	100	

b) Similaridad de Jaccard

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	0.5	0.9	0	0.9545455	1	?	1
plumas	0	0	0	0.9090909	0	?	0
huevos	0.8333333	0.8	0	1	1	?	0.8181818
leche	0.8333333	0.8	0	0.9545455	1	?	1
vuela	0.5	0	0.0588235	0.7272727	0	?	0
acuatico	0	0.9	0.0294118	0.3181818	1	?	0.7272727
depredador	0.75	0.2	0.5588235	0.5454545	1	?	0
dentado	0.8333333	0	0	1	0	?	0.7272727
columna	0.8333333	0	0	0	0	?	0.7272727
respira	0	0.55	0	0	1	?	0.8181818
venenoso	0.1666667	0	0	0	0	?	0.4545455
aletas	0	0.75	0	0	1	?	0.0909091
patas	0.1666667	0.75	0	0	1	?	0.4545455
cola	0	0.85	0.9117647	1	1	?	0.3636364
doméstico	0.0833333	0	0.2352941	0.1363636	1	?	0

Clusters	Nº. Elementos	Elementos
Cluster 0:	12	1, 4, 25, 30, 39, 40, 42, 51, 81, 88, 97, 99
Cluster 1:	20	3, 9, 13, 19, 20, 26, 34, 38, 52, 60, 61, 66, 73, 74, 75, 80, 82, 89, 91, 92
Cluster 2:	34	2, 5, 6, 7, 10, 11, 18, 23, 27, 28, 29, 31, 32, 35, 36, 44, 45, 47, 48, 49, 50, 54, 55, 64, 65, 67, 68, 69, 70, 84, 93, 94, 96, 98
Cluster 3:	22	12, 17, 21, 22, 24, 33, 37, 41, 43, 56, 57, 58, 59, 71, 78, 79, 83, 87, 90, 95, 100
Cluster 4:	18	
Cluster 5:	0	
Cluster 6:	11	14, 15, 16, 46, 53, 62, 72, 76, 77, 85, 86
Total Items:	100	

c) Similaridad de Kulczynski

“No se puede ejecutar el operador”

d) Distancia Nominal

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	0.57	?	?	?	?	?	?
plumas	0.2	?	?	?	?	?	?
huevos	0.58	?	?	?	?	?	?
leche	0.59	?	?	?	?	?	?
vuela	0.24	?	?	?	?	?	?
acuatico	0.35	?	?	?	?	?	?
depredador	0.45	?	?	?	?	?	?
dentado	0.4	?	?	?	?	?	?
columna	0.18	?	?	?	?	?	?
respira	0.21	?	?	?	?	?	?
venenoso	0.07	?	?	?	?	?	?
aletas	0.17	?	?	?	?	?	?
patas	0.23	?	?	?	?	?	?
cola	0.75	?	?	?	?	?	?
doméstico	0.13	?	?	?	?	?	?

Clusters	Nº. Elementos	Elementos
Cluster 0:	100	
Cluster 1:	0	
Cluster 2:	0	
Cluster 3:	0	
Cluster 4:	0	
Cluster 5:	0	
Cluster 6:	0	
Total Items:	100	

e) Similitud de Rogers Tanimoto

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	?	0.425531915	0.6	?	0.7857143	?	?
plumas	?	0	0	?	0.7142857	?	?
huevos	?	0.404255319	0.52	?	0.9285714	?	?
leche	?	0.425531915	0.52	?	0.9285714	?	?
vuela	?	0	0	?	0.8571429	?	?
acuatico	?	0.255319149	0.68	?	0.2142857	?	?
depredador	?	0.361702128	0.4	?	0.6428571	?	?
dentado	?	0.29787234	0	?	0.9285714	?	?
columna	?	0.255319149	0	?	0.2142857	?	?
respira	?	0.170212766	0.52	?	0	?	?
venenoso	?	0.085106383	0.04	?	0.0714286	?	?
aletas	?	0	0.68	?	0	?	?
patas	?	0.14893617	0.64	?	0	?	?
cola	?	0.659574468	0.88	?	0.7857143	?	?
doméstico	?	0	0.36	?	0.1428571	?	?

Clusters	Nº. Elementos	Elementos
Cluster 0:	0	
Cluster 1:	47	1, 2, 4, 5, 6, 11, 14, 15, 16, 18, 23, 25, 26, 28, 32, 36, 44, 45, 46, 47, 48, 49, 50, 52, 53, 54, 55, 62, 63, 64, 67, 69, 72, 76, 77, 80, 81, 84, 85, 88, 89, 90, 91, 94, 96, 98, 99
Cluster 2:	25	3, 7, 8, 9, 10, 13, 19, 20, 29, 31, 34, 35, 38, 60, 61, 65, 66, 68, 70, 73, 74, 75, 82, 86, 92
Cluster 3:	0	
Cluster 4:	28	12, 17, 21, 22, 24, 27, 30, 33, 37, 39, 40, 41, 42, 43, 51, 56, 57, 58, 59, 71, 78, 79, 83, 87, 93, 95, 97, 100
Cluster 5:	0	
Cluster 6:	0	
Total Items:	100	

f) Similitud de Russel Rao

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	0.06666667	0	0.6666667	0.78125	1	?	1
plumas	0	0	0	0.625	0	?	0
huevos	0	0	0.5757576	0.9375	1	?	1
leche	0.06666667	0	0.6060606	0.90625	1	?	1
vuela	0	0	0	0.75	0	?	0
acuatico	0	0	0.6666667	0.21875	0	?	0.8571429
depredador	0	1	0.3333333	0.65625	1	?	0
dentado	0.06666667	0	0	0.9375	1	?	1
columna	0.06666667	0	0	0.25	1	?	1
respira	0	0	0.4242424	0	0	?	1
venenoso	0.06666667	0	0.0909091	0.0625	0	?	0.1428571
aletas	0	0	0.5151515	0	0	?	0
patas	0	0	0.5757576	0	1	?	0.2857143
cola	0.86666667	0.90909091	0.8484848	0.75	0	?	0
doméstico	0	0	0.2727273	0.125	0	?	0

Clusters	Nº. Elementos	Elementos
Cluster 0:	15	1, 4, 5, 11, 44, 45, 47, 49, 50, 54, 64, 67, 69, 72, 98
Cluster 1:	11	2, 6, 18, 23, 28, 32, 36, 55, 84, 94, 96
Cluster 2:	33	3, 7, 8, 9, 10, 13, 19, 20, 26, 29, 31, 34, 35, 38, 48, 52, 60, 61, 62, 65, 66, 68, 70, 73, 74, 75, 76, 80, 82, 86, 89, 91, 92
Cluster 3:	32	12, 17, 21, 22, 24, 25, 27, 30, 33, 37, 39, 40, 41, 42, 43, 51, 56, 57, 58, 59, 63, 71, 78, 79, 83, 87, 88, 90, 93, 95, 97, 100
Cluster 4:	2	81, 99
Cluster 5:	0	
Cluster 6:	7	14, 15, 16, 46, 53, 77, 85
Total Items:	100	

g) Similitud de Simple Matching

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	?	0.42553191	0.6	?	0.785714	?	?
plumas	?	0	0	?	0.714286	?	?
huevos	?	0.40425532	0.52	?	0.928571	?	?
leche	?	0.42553191	0.52	?	0.928571	?	?
vuela	?	0	0	?	0.857143	?	?
acuatico	?	0.25531915	0.68	?	0.214286	?	?
depredador	?	0.36170213	0.4	?	0.642857	?	?
dentado	?	0.29787234	0	?	0.928571	?	?
columna	?	0.25531915	0	?	0.214286	?	?
respira	?	0.17021277	0.52	?	0	?	?
venenoso	?	0.08510638	0.04	?	0.071429	?	?
aletas	?	0	0.68	?	0	?	?
patas	?	0.14893617	0.64	?	0	?	?
cola	?	0.65957447	0.88	?	0.785714	?	?
doméstico	?	0	0.36	?	0.142857	?	?

Clusters	Nº. Elementos	Elementos
Cluster 0:	0	
Cluster 1:	47	1, 2, 4, 5, 6, 11, 14, 15, 16, 18, 23, 25, 26, 28, 32, 36, 44, 45, 46, 47, 48, 49, 50, 52, 53, 54, 55, 62, 63, 64, 67, 69, 72, 76, 77, 80, 81, 84, 85, 88, 89, 90, 91, 94, 96, 98, 99
Cluster 2:	25	3, 7, 8, 9, 10, 13, 19, 20, 29, 31, 34, 35, 38, 60, 61, 65, 66, 68, 70, 73, 74, 75, 82, 86, 92
Cluster 3:	0	
Cluster 4:	28	12, 17, 21, 22, 24, 27, 30, 33, 37, 39, 40, 41, 42, 43, 51, 56, 57, 58, 59, 71, 78, 79, 83, 87, 93, 95, 97, 100
Cluster 5:	0	
Cluster 6:	0	
Total Items:	100	

Medidas Mixtas

Finalmente, se probó el algoritmo con medidas mixtas, $k=7$, obteniendo el siguiente resultado:

a) Distancia Euclidiana Mixta

Centroides	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
pelo	0.69230769	?	0.95652174	?	0.16216216	0.74074074	?
plumas	0	?	0	?	0	0.74074074	?
huevos	0.92307692	?	0.82608696	?	0.18918919	0.74074074	?
leche	1	?	0.86956522	?	0.16216216	0.74074074	?
vuela	0.46153846	?	0	?	0	0.66666667	?
acuatico	0.30769231	?	0.7826087	?	0.16216216	0.25925926	?
depredador	0.53846154	?	0.26086957	?	0.43243243	0.59259259	?
dentado	1	?	0.17391304	?	0.08108108	0.74074074	?
columna	1	?	0.17391304	?	0.02702703	0	?
respira	0.30769231	?	0.69565217	?	0.02702703	0	?
venenoso	0.23076923	?	0.17391304	?	0	0	?
aletas	0	?	0.69565217	?	0	0.03703704	?
patas	6.23076923	?	0	?	4	2	?
cola	0.07692308	?	0.7826087	?	0.83783784	0.92592593	?
doméstico	0.07692308	?	0.04347826	?	0.18918919	0.14814815	?

Clusters obtenidos	Nº. Elementos	Elementos
Cluster 0:	13	16, 25, 30, 39, 40, 42, 46, 51, 53, 72, 85, 88, 97
Cluster 1:	0	
Cluster 2:	23	3, 8, 9, 13, 14, 20, 19, 34, 38, 60, 61, 62, 66, 73, 74, 76, 77, 80, 81, 82, 86, 92, 99
Cluster 3:	0	
Cluster 4:	37	1, 2, 4, 5, 6, 7, 10, 11, 15, 18, 23, 26, 28, 31, 35, 36, 44, 45, 47, 48, 49, 50, 52, 54, 55, 63, 64, 65, 67, 68, 69, 70, 89, 90, 91, 94, 98
Cluster 5:	27	12, 17, 21, 22, 24, 27, 29, 32, 33, 37, 41, 43, 56, 57, 58, 59, 71, 75, 78, 79, 83, 84, 87, 93, 95, 96, 100
Cluster 6:	0	
Total ítems	100	

Se muestra a continuación una tabla resumen con los errores totales que hubo en cada ejecución, y con las diferentes medidas. Para cada animal, se considera 0 si se asigna al cluster correcto y 1 en caso contrario.

K-Medias en RapidMiner

NUM	TIPO	NOMBRE	MEDIDAS NOMINALES					MEDIDAS MIXTAS
			DICE	JACCARD	ROG-TANIMOTO	RUSSEL-RAO	SIMPLE MATCHING	MIXED
1	mamífero	oso hormigero	1	1	0	0	0	0
2	mamífero	antílope	0	0	0	0	0	0
3	pez	robalo	0	0	0	0	0	0
4	mamífero	oso	1	1	0	0	0	0
5	mamífero	jabalí	0	0	0	0	0	0
6	mamífero	búfalo	0	0	0	0	0	0
7	mamífero	temero	0	0	1	1	1	0
8	pez	carpa	1	0	0	0	0	0
9	pez	bagre	0	0	0	0	0	0
10	mamífero	conejo de indias	0	0	1	0	1	0
11	mamífero	chita	0	0	0	0	0	0
12	ave	pollo	0	0	0	0	0	0
13	pez	chub	0	0	0	0	0	0
14	invertebrado	almeja	0	0	1	0	1	1
15	invertebrado	cangrejo	0	0	1	0	1	1
16	invertebrado	cangrejo de río	0	0	1	0	1	0
17	ave	cuervo	0	0	0	0	0	0
18	mamífero	venado	0	0	0	0	0	0
19	pez	cazón	0	0	0	0	0	0
20	mamífero	delfín	0	1	1	1	1	1
21	ave	paloma	0	0	0	0	0	0
22	ave	pato	0	0	0	0	0	0
23	mamífero	elefante	0	0	0	0	0	0
24	ave	flamingo	0	0	0	0	0	0
25	insecto	pulga	0	0	1	1	1	0
26	anfibio	sapo	1	1	1	1	1	1
27	mamífero	fruitbat	1	1	1	1	1	1
28	mamífero	jirafa	0	0	0	0	0	0
29	mamífero	girl	0	0	1	1	1	1
30	insecto	mosquito	0	0	1	1	1	0
31	mamífero	cabra	0	0	1	1	1	0
32	mamífero	gorila	1	0	0	0	0	1
33	ave	gaviota	0	0	0	0	0	0
34	pez	haddock	1	0	0	0	0	0
35	mamífero	hamster	0	0	1	1	1	0
36	mamífero	liebre	0	0	0	0	0	0
37	ave	halcón	0	0	0	0	0	0
38	pez	arenque	0	0	0	0	0	0
39	insecto	abeja	1	0	1	1	1	0
40	insecto	mosca	0	0	1	1	1	0
41	ave	kiwi	0	0	0	0	0	0
42	insecto	catarina	0	0	1	1	1	0
43	ave	alondra	0	0	0	0	0	0
44	mamífero	leopardo	0	0	0	0	0	0
45	mamífero	león	0	0	0	0	0	0
46	invertebrado	langosta	0	0	1	0	1	0
47	mamífero	lince	0	0	0	0	0	0
48	mamífero	mink	0	0	0	1	0	0
49	mamífero	topo	0	0	0	0	0	0
50	mamífero	mangosta	0	0	0	0	0	0

K-Medias en RapidMiner

NUM	TIPO	NOMBRE	MEDIDAS NOMINALES					MEDIDAS MIXTAS
			DICE	JACCARD	ROG-TANIMOTO	RUSSEL-RAO	SIMPLE MATCHING	MIXED
51	insecto	polilla	0	0	1	1	1	0
52	anfibio	tritón	1	1	1	1	1	1
53	invertebrado	pulpo	0	0	1	0	1	0
54	mamífero	zarigüeya	0	0	0	0	0	0
55	mamífero	orix	0	0	0	0	0	0
56	ave	aveztruz	0	0	0	0	0	0
57	ave	perico	0	0	0	0	0	0
58	ave	pingüino	0	0	0	0	0	0
59	ave	faisán	0	0	0	0	0	0
60	pez	lucio	0	0	0	0	0	0
61	pez	piraña	0	0	0	0	0	0
62	reptil	crótalo	1	1	1	1	1	1
63	mamífero	ornitorrinco	1	1	0	1	0	0
64	mamífero	turón	0	0	0	0	0	0
65	mamífero	pony	0	0	1	1	1	0
66	mamífero	porpoise	0	1	1	1	1	1
67	mamífero	puma	0	0	0	0	0	0
68	mamífero	minino	0	0	1	1	1	0
69	mamífero	mapache	0	0	0	0	0	0
70	mamífero	reno	0	0	1	1	1	0
71	ave	ñandú	0	0	0	0	0	0
72	invertebrado	escorpión	0	0	1	1	1	0
73	pez	caballito de mar	0	0	0	0	0	0
74	mamífero	foca	0	1	1	1	1	1
75	mamífero	león marino	0	1	1	1	1	1
76	reptil	serpiente de mar	1	1	1	1	1	1
77	invertebrado	avispa de mar	0	0	1	0	1	1
78	ave	skimmer	0	0	0	0	0	0
79	ave	skua	0	0	0	0	0	0
80	reptil	gusano lento	1	1	1	1	1	1
81	invertebrado	baboso	1	1	1	0	1	1
82	pez	sole	1	0	0	0	0	0
83	ave	gorrión	0	0	0	0	0	0
84	mamífero	ardilla	0	0	0	0	0	1
85	invertebrado	estrella de mar	0	0	1	0	1	0
86	pez	stingray	0	1	0	0	0	0
87	ave	cisne	0	0	0	0	0	0
88	insecto	termita	0	0	1	1	1	0
89	anfibio	toad	1	1	1	1	1	1
90	reptil	tortuga	1	1	1	1	1	1
91	reptil	tuatara	1	1	1	1	1	1
92	pez	atún	0	0	0	0	0	0
93	mamífero	vampiro	1	0	1	1	1	1
94	mamífero	vole	0	0	0	0	0	0
95	ave	buitre	0	0	0	0	0	0
96	mamífero	wallaby	0	0	0	0	0	1
97	insecto	avispa	0	0	1	1	1	0
98	mamífero	lobo	0	0	0	0	0	0
99	invertebrado	gusano	1	1	1	0	1	1
100	ave	wren	0	0	0	0	0	0
		TOTAL ERRORES:	20	19	40	32	40	23

Se puede notar que la selección de la medida de similitud afecta en gran medida el resultado obtenido en la distribución de los clusters.

Para este caso, la distancia nominal no dio un resultado favorable, al colocar todos los elementos en un solo cluster. Las medidas que mejor se ajustaron a estos datos son Jaccard, Dice y la Mixta, con 19, 20 y 23 errores, respectivamente. La ejecución con las medidas Rogers-Tanimoto y Simple Matching dieron exactamente la misma distribución de clusters.

La interpretación final del resultado depende también del punto de vista del experto en el área. Él debe decidir qué medida usar con base en su conocimiento y experiencia. Aun así, si se considera que el agrupamiento no es una buena técnica de Minería de Datos, se deberá probar otro modelo para un mejor resultado.

6. CONCLUSIONES

Al elaborar este trabajo se puede concluir que:

El agrupamiento es una de las técnicas de Minería de Datos más común. Se puede aplicar a una gran cantidad de información para obtener una clasificación exitosa y casi sin errores de los datos. Uno de los algoritmos más usado, por su sencillez y fácil implementación es el algoritmo de las K-medias, el cual funciona calculando la media promedio de cada cluster, hasta encontrar la menor distancia entre cada elemento del cluster y su centro, logrando que cada uno de ellos sea lo más diferente posible a los demás.

RapidMiner es una herramienta de gran ayuda para la aplicación de técnicas de Minería de Datos. Ofrece grandes ventajas con respecto a otras aplicaciones, al tener una interfaz de usuario muy simple, aceptar datos de entrada en diferentes formatos, una gran cantidad de operadores y modelos de Minería de Datos y además una combinación de opciones y parámetros para la ejecución de cada una de las técnicas ofrecidas.

El algoritmo de las K-medias en RapidMiner se puede implementar con diversas medidas de asociación. La elección de esta medida dependerá en gran parte, primero del tipo de datos que se tenga como entrada, y después del punto de vista del experto en el área que se esté trabajando. El experto es quien tendrá la última palabra en el éxito de la aplicación de la técnica de Minería de Datos, ya que es él quien domina la información de entrada y por supuesto posee la experiencia para la interpretación del resultado final obtenido. Si el resultado no es el esperado, se puede proceder a la elección de otra técnica que describa o infiera mejor el resultado final.

7. BIBLIOGRAFÍA

- [1] North, Matthew. 2012, Data mining for the masses. A Global Text Project Book. Disponible en: <http://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf>
- [2] <http://rapidminer.com/documentation/>
- [3] Rodríguez R. Oldemar. 2009. Maestría en Administración de la Tecnología de la Información. Escuela de Informática. Universidad Nacional de Costa Rica.
- [4] Vásquez García Beatriz. 2009, Tesis de Maestría en Ciencias de la Computación: Estudio comparativo de métodos de clasificación automática en la Zonificación Agro Ecológica del sur del estado de Puebla.
- [5] Banerjee, Merugu, Dhillon and Ghosh. 2005, Clustering with Bregman divergences. Journal of Machine Learning Research 6. 1705–1749
- [6] Análisis Cluster. http://www.ugr.es/~bioestad/_private/cpfund7.pdf (consultado el 3 de abril de 2015)
- [7] Casanoves, Fernando ; Pla, Laura; Di Rienzo, Julio A. 2011. Valoración y análisis de la diversidad funcional y su relación con los servicios eco sistémicos. Centro Agronómico Tropical de Investigación y Enseñanza, CATIE, Turrialba, Costa Rica.
- [8] Bourbaki, Nicolas (1987). Capítulos 1–5. Topological vector spaces. Springer. ISBN 3-540-13627-4.
- [9] Schulz, Jan. October 2011. Canberra distance. Code 10. Retrieved 18. http://www.code10.info/index.php?option=com_content&view=article&id=49:article_canberra-distance&catid=38:cat_coding_algorithms_data-similarity&Itemid=57. (Consultado el 12 de julio de 2015.)
- [10] <https://lyfat.wordpress.com/2012/05/22/euclidean-vs-chebyshev-vs-manhattan-distance/> (Consultado 13 de mayo de 2015.)

[11] Mendenhall, William. Probabilidad y Estadística para Ingeniería y Ciencias. 4ª. Edición. Prentice Hall. Cap. 6. Pág. 279.

[12] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto, 2014, Soft Similarity and Soft Cosine Measure: Similarity of Features in Vector Space Model. Disponible en: <http://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2043/1921>

[13] Al-Naymat, G., Chawla, S., & Taheri, J. 2012. SparseDTW: A Novel Approach to Speed up Dynamic Time Warping.

[14] Package com.rapidminer.tools.math.similarity.numerical

<http://www-ai.cs.uni-dortmund.de/LEHRE/VORLESUNGEN/MLRN/WS0809/rm-api/com/rapidminer/tools/math/similarity/numerical/package-summary.html> (Consultado el 16 de marzo de 2015)

[15] Lang, Serg. Algebra Lineal. Addison Wesley Iberoamericana. Cap. 1, pág. 10.

[16] Phillips, Jeff M. – Venkatasubramanian, Suresh (2010) A gentle Introduction to the Kernel Distance. Disponible en: <http://www.cs.utah.edu/~suresh/papers/kerneld/kerneld.pdf>

[17] Paul E. Black, 31 May 2006. Manhattan distance, in *Dictionary of Algorithms and Data Structures*, Vreda Pieterse and Paul E. Black. Consultado el 25 de febrero de 2015. Disponible en: <http://www.nist.gov/dads/HTML/manhattanDistance.html>

[18]

<http://www.dia.fi.upm.es/~ocorcho/Asignaturas/ModelosRazonamiento/PresentacionesClases/02%20-%20SimMeasures.pdf> (Consultado el 13 de julio de 2015)

[19] UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Zoo>. (Consultado el 16 de mayo de 2015)