UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2016

COMPARACIÓN ENTRE K MEANS Y REDES NEURONALES

Aplicación al estudio de clima escolar en colegios públicos de Bogotá.

William Cruz Ramos.[a]
wiliamcruz@usantotomas.edu.co

Directora: Dra. Luz Mary Pinzón.[b]
luzpinzon@usantotomas.edu.co

Codirector: Dr. Mauricio Restrepo.[c]
mauricio.restrepo@unimilitar.edu.co

ANTECEDENTES

El análisis de cluster ha sido un tema de investigación emergente en minería de datos debido a su variedad de aplicaciones, puede ser considerado el problema más importante de aprendizaje no supervisado que trata de encontrar una estructura en una colección de datos y que se puede definir como el proceso de organizar objetos en grupos cuyos miembros son similares en alguna forma. Con el desarrollo de muchos algoritmos de agrupamiento de datos y su amplio uso, que incluyen desde procesamiento de imágenes, biología computacional, comunicaciones móviles, hasta medicina y economía, estos algoritmos se han vuelto muy populares.

El principal problema con los algoritmos de agrupación es que no se pueden estandarizar. Un algoritmo puede dar buen resultado con un tipo de conjunto de datos, pero puede fallar o dar mal resultado con otros tipos de datos. Hasta ahora se han propuesto muchos algoritmos de agrupación, sin embargo, cada algoritmo tiene sus propios méritos y deméritos porque no funcionan bien en todas las situaciones reales.

Dentro de los algoritmos de cluster lineales no supervisados nos centramos en el K-means (Hartigan & Wong 1979) de gran utilidad para los métodos de clasificación, ya que, permite hacer particiones de grandes conjuntos de datos. Sin embargo, este algoritmo tiene dos inconvenientes: se debe dar el número de clases con sus centros iniciales y converge a óptimos locales que dependen de los puntos iniciales.
Para solucionar, al menos parcialmente, los inconvenientes de los métodos de agregación alrededor de centros móviles, se puede utilizar el método de formas fuertes o grupos estables (Lebart et al. 2006).
Los grupos estables se identifican realizando varias particiones, cada una con diferentes puntos iniciales y seleccionando aquellos grupos en los cuales los individuos se mantienen asignados a la misma clase.

En contraste encontramos aplicaciones de redes neuronales en diferentes campos siendo más frecuentes las redes neuronales supervisadas, es decir que se tiene información del atributo de decisión. Sin embargo, nuestro interés se centra en las redes no supervisadas, que son redes de auto-entrenamiento. En general este tipo de red se utiliza en reconocimiento de patrones de imágenes o comportamientos, medio ambiente y tecnología.

[a]Estudiante de estadística Universidad Santo Tomás Bogotá
[b]Profesora de la facultad de estadística de la Universidad Santo Tomás Bogotá
[c]Profesor del departamento de Matemáticas, Universidad Militar Nueva Granada

Encontramos un artículo referente al medio ambiente, que para detectar diferentes cuerpos de agua, se procedía a realizar un muestreo bajo el criterio de un experto quien proponía áreas con diferentes tipos de agua. Este proceso presentaba varias desventajas, altos costos, complejidad de la toma de muestras en terreno y todo lo que esto implica en el análisis de la información.
Al aplicar la red neuronal no supervisada, se utilizó un mapa auto-organizativo que en resumen, es una imagen satelital del lago, donde se analiza el color y distribución de los pixeles. Con esta imagen se lograron dos objetivos: un modelo simplificado de los datos de entrada y un mapa bidimensional que permite mostrar gráficamente las relaciones existentes entre los diferentes cuerpos de agua clasificados a partir de los pixeles.

En el campo de la telefonía celular se utiliza para detección de fraudes, como una herramienta con capacidad de detectar cambios de comportamiento relacionados con fraude y construir un perfil de usuario que pueda ser comparable con patrones históricos del mismo.
Adicionamente, apartir de SOM (Self Organizing Maps) se pueden construir perfiles de usuarios por medio de redes que clasifican millones de llamadas en diferentes prototipos o clases. La frecuencia con la cual el usuario hace llamadas de cada prototipo corresponde a la representación de los perfiles, perfil de comparación y perfil de Histórico. Una vez que ambos se actualizan, se comparan y se decide si la diferencia entre el consumo reciente y el histórico es lo suficientemente grande como para emitir una alarma.

Para el caso de redes neuronales supervisadas se tienen dos etapas, una de entrenamiento y otra de funcionamiento. La etapa de entrenamiento debe ser actualizada constantemente. Por ejemplo, a medida que aparecen nuevas modalidades de fraude se debe volver a iniciar el proceso de entrenamiento. Por el contrario, en las redes neuronales no supervisadas el enfoque de aprendizaje generalmente se tipifica, donde la herramienta de detección de fraude aprende por sí sola cuál es el comportamiento esperado del usuario.

Para la técnica de K-means, encontramos gran variedad de aplicaciones, haremos referencia a una de índole social: análisis de violencia sexual y a otro de percepción de la ciencia y tecnología:
Para el caso de violencia sexual se aplicó la herramienta para identificar los centros de los diferentes tipos de violencia y la proximidad entre el agresor y la victima a través de estructura de parentesco tradicional. A la vez se valida la distancia entre las clasificaciones de violencia sexual, abuso sexual y asalto sexual. La aplicación de este método es un proceso de extracción de información no trivial en grandes volúmenes de datos, cuya aplicación a la inteligencia criminal se ha constituido en un campo relativamente nuevo, para generar patrones, asociaciones, cambios y estructuras significativas que no son fáciles de observar de forma directa. Las fuentes de información son el Instituto Nacional de Medicina Legal y Ciencias Forenses, el Centro de Investigaciones Criminológicas de la Policía Metropolitana de Bogotá y la Fiscalía General de la Nación, para el estudio se tomó el registro de violencia sexual del 2007 en Bogotá.

En el segundo estudio se busca identificar la opinión y las actitudes de los colombianos sobre ciencia y la tecnología, y dar insumos para mejorar los procesos de apropiación de la ciencia y la tecnología (CTI) en Colombia.

JUSTIFICACIÓN

La idea de comparar el método K-means con las redes neuronales se basa específicamente en sus definiciones.

El concepto de red neuronal artificial está inspirado en las redes neuronales biológicas.
Una red biológica es un dispositivo no lineal altamente paralelo, caracterizado por su robustez y tolerancia a fallos. Sus principales características son:

- Aprendizaje mediante adaptación de sus pesos sinápticos a los cambios en el entorno.
- Manejo de precisión, ruido e información probabilística.
- Generalización a partir de ejemplos.

A su vez cada neurona es un procesador elemental con operaciones muy primitivas como la suma ponderada de sus pesos de entrada y la amplificación o umbralización de esta suma.

De otra parte, el método k-means utiliza dos conceptos, una distancia y un método de agrupación. Si bien se tienen diversas distancias y diversos métodos de agrupación, el problema de selección de los centros iniciales no ha tenido solución hasta el momento.

Dadas las características de las definiciones, se puede pensar que un método con fase de entrenamiento puede tener ventajas sobre otro que no la tiene.

Bajo esos preceptos veremos que sucede.

Objetivos

Objetivo General

- Analizar si las redes neuronales no supervisadas captan las formas fuertes generadas por k-means
- Caracterizar el clima escolar de los colegios oficiales del área metropolitana de Bogotá

Objetivos Específicos

- Seleccionar las variables activas que identifican patrones de clima escolar.
- Construir una clasificación jerárquica, con las variables activas.
- Analizar la clasificación anterior con el fin de caracterizar el clima escolar de los colegios oficiales.
- A partir del número de clases de la clasificación jerárquica, generar réplicas por el método k-means e identificar las formas fuertes.
- Identificar invariantes de la clasificación mediante Redes Neuronales. Comparar las Formas Fuertes de los k-means con los invariantes de las redes Neuronales

1. LOS DATOS

Se tienen los datos de la encuesta de clima escolar y victimización 2013, y los resultados Saber Pro 2013, que son de dominio público.

Para esta encuesta el marco muestral se definió como:
Todos los colegios distritales, los colegios con contrato con la SED y los colegios por concesión. En cada uno de ellos se tomó una muestra al azar de estudiantes por niveles, del grado 6 al 11. Las unidades de análisis fueron el colegio y el nivel escolar (no el curso).

Ficha Técnica

Estudio	Encuesta de clima escolar y victimización 2013
Población Objetivo	Estudiantes de los grados 6 a 11 de colegios oficiales
Universo:	623 colegios Oficiales
Metodología:	Censo por colegios y jornadas
Lugar de Recolección:	Colegios ubicados en la cabecera urbana de Bogotá
Tipo de encuesta:	Auto diligenciada
Número de preguntas:	115 preguntas en 3 secciones
Periodo de Recolección:	22 de octubre a 24 de noviembre de 2013

1.1. Selección de las variables

Hay varias opciones para construir tipologías de establecimientos escolares de acuerdo con el perfil del estudiante. En la medida que el objetivo principal del proyecto de investigación es caracterizar el clima escolar de los colegios públicos del área metropolitana de Bogotá, Oberti (2012) propone seleccionar las variables que definen las características socio-demográficas de los estudiantes. A diferencia, por ejemplo, Le Donne y Rock (2010), que combinan dentro de su índice de posición escolar varias variables, unas miden las características de origen social y otras de prácticas escolares.

Según Oberti et al. (2012), otras consideraciones pueden intervenir, como la reputación de la institución acerca de los resultados en las pruebas Saber, que no son resultado de un perfil social de los estudiantes, sino de la calidad del equipo docente y de las políticas del colegio.
Por tanto, construir la tipología solo sobre las características sociales de los estudiantes nos permite observar si los establecimientos que pertenecen a un mismo grupo tienen características similares de aceptación y reputación a nivel general y escolar.

Pierre Merle, (como afirma Rubia) propone que el origen de una tipología escolar está anclado al modelo socioeconómico actual. Además que existen algunas dimensiones de este fenómeno:

- Tipología por sexo: relacionada con la separación de roles, habilidades y aptitudes.
- Tipología académica: que busca el posicionamiento de los mejores estudiantes y en razón a lo anterior lleva a los padres a seleccionar el mejor colegio para la educación de sus hijos y a las instituciones a particularizar sus currículos.
- Tipología por el orígen social: que implica una clasificación del estudiante por su orígen social y cultural, lo cual conduce a estudiantes de clases menos favorecidas o culturalmente señalados, a tener acceso a instituciones que replican las mismas condiciones y necesidades microsociales existentes en su cotidianidad. El autor añade la necesidad de analizar previamente aquellos programas que buscan corregir esta forma de segregación social y que posiblemente terminen generando mayor segregación.

1.2. Variables activas

Se siguen las recomendaciones de los profesores Préteceille y Oberti, de utilizar la preguntas relacionadas con el nivel educativo de los padres e ingresos.
La preguntas relacionadas con el nivel educativo de los padres se agruparon dado las bajas frecuencias en algunas modalidades, se tienen en cuenta 3 categorías que son:

Tabla 1: Pregunta NP 4 y 5. Educación de los padres

Código	Respuesta	Nivel
1	Primaria o menos	No ha estudiado
		Algunos años de primaria
		Terminó primaria
2	Algo de Bachiller	Algunos años de bachillerato
		Terminó bachillerato
		Estudios técnicos o tecnológicos
3	Algo de Universidad	Estudios universitarios incompletos
		Estudios universitarios completos
		Estudios de postgrado

Tabla 2: Pregunta 11

Código	Respuesta
1	No alcanzan para cubrir los gastos mínimos
2	Solo alcanzan para cubrir los gastos mínimos
3	Cubren más que los gastos mínimos

Otra variable que sugieren que entre en el análisis es el nivel sociecómico, una proxy es la pregunta: Tú crees que los ingresos de tu hogar:

1.3. Variables suplementarias

Adicionalmente, se tendrán como variables suplementarias o descriptivas de los segmentos, las siguientes variables:

Pregunta	Nivel
Sexo	Descriptores
te reconoces como	Descriptores
Estrato donde esta ubicado el colegio	Condiciones socioeconomicas
CSE (Indice de condiciones socioeconomicas) varia de 0 a 100	Condiciones socioeconomicas
Jornada	Relación con la escuela
Cómo son tus calificaciones en comparación con las de los compañeros de tu curso	Relación con la escuela
Los docentes, directivas y órganos de gobierno intervienen cuando hay agresiones o amenazas entre estudiantes?	Antecedentes sociales
El Consejo Directivo, Comité de Convivencia, otros docentes, o directivas intervienen cuando los estudiantes dicen haber sido objeto de injusticias cometidas por profesores o directivas cuando agresiones	Antecedentes sociales
Durante este año, los salones de clase (la construcción, la iluminación, los pupitres, etc),son suficientes, adecuados y permanecen en buen estado	Espacios y lugares
Los baños están bien: son suficientes, limpios y permanecen en buen estado durante todo el año	Espacios y lugares
Los baños son seguros	Espacios y lugares
Los espacios comunes como corredores y patios son suficientes, seguros y has permanecido en buen estado durante todo el año	Espacios y lugares
Confío en el colegio, es decir, creo que sus profesores (as) y sus directivas(os) hacen y deciden teniendo como propósito el beneficio de nosotros (as) los estudiantes	Confianza en la institución
El colegio en el que esta estudiando te ayuda a ser mejor persona Completamente en desacuerdo	Confianza en la institución
No se tiene por alumno, por tanto se asigna a cada colegio y por Jornada, la clasificaciń de las pruebas Saber.	Condiciones Mejor estudiante

2. ANÁLISIS DESCRIPTIVO DE LAS VARIABLES

Para ponernos en contexto con los datos a analizar hacemos el análisis descriptivo de las variables activas

2.1. Variables activas

En general el nivel educativo de los padres es medio, con un alto porcentaje de bachilleres y técnicos, y pocos con estudios universitarios o superiores

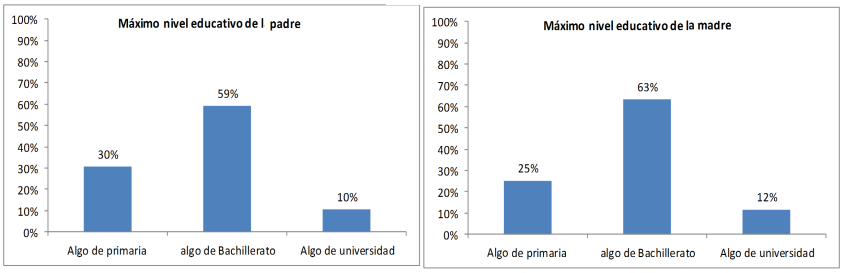


Figura 1: Nivel educativo de los padres

El nivel educativo de las madres presenta un perfil similar, sin embargo se observa un porcentaje mayor tanto de bachilleres y tecnólogas como de estudios universitarios.

La percepción del nivel de ingresos muestra que la mitad de los estudiantes tienen exactamente lo necesario y un 41 % consideran que tienen más de lo necesario.

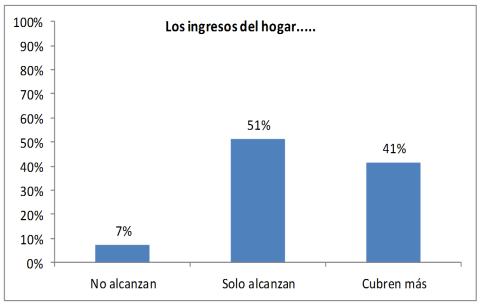


Figura 2: Nivel de ingresos

Aunque el estrato del colegio no necesariamente es el mismo del estudiante, para el caso de colegios oficiales, donde se conoce que los alumnos asisten a colegios cercanos a sus viviendas, el estrato del alumno se relaciona con el estrato del colegio.

Además, el estrato donde está ubicado el colegio y la calificación de este en las pruebas SABER presentan cierto nivel de dependencia, indicando la importancia de incluir como variable activa el nivel socioeconómico del estudiante.

Pruebas de chi-cuadrado

	Valor	gl	Sig. asintótica (bilateral)
Chi-cuadrado de Pearson	169,701[a]	30	,000
Razón de verosimilitudes	179,684	30	,000
N de casos válidos	623		

a. 25 casillas (59,5%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es ,02.

Figura 3: Chi cuadrado entre Estrato del colegio y resultado prueba SABER

2.2. Ubicación de colegios

Se presenta la ubicación de los colegios en el área urbana de la ciudad de Bogotá y la relación de los resultados en la pruebas Saber y el estrato:

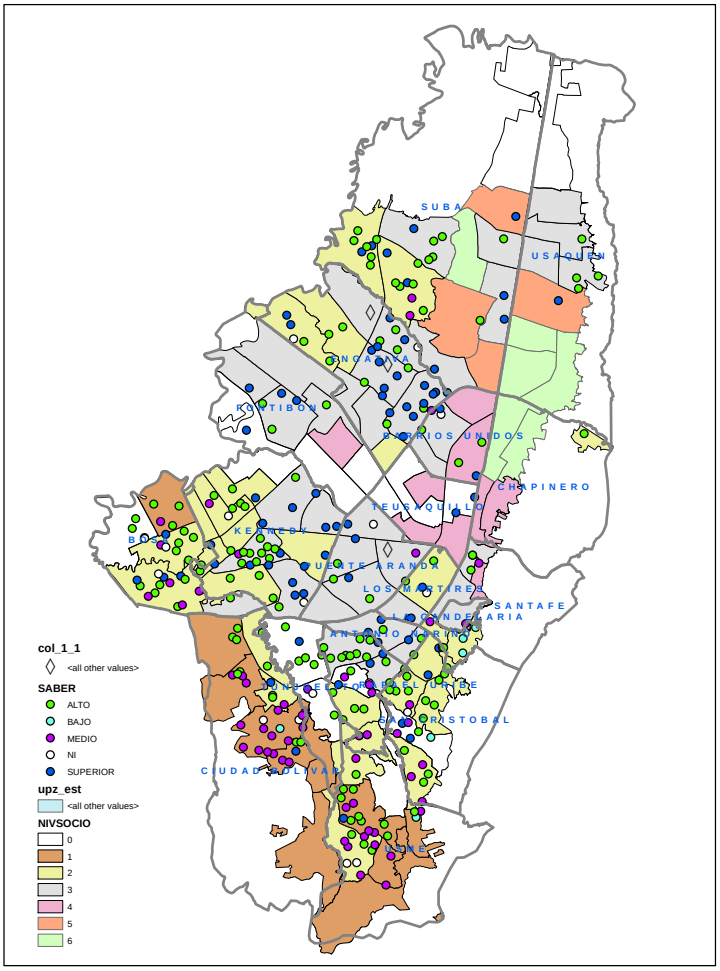


Tabla 3: Georeferenciación de los colegios oficiales por estrato y SABER PRO.

De acuerdo al mapa y teniendo en cuenta la relación entre el estrato y el resultado de la prueba SABER, podemos observar que los colegios ubicados en la zona sur de la ciudad donde prevalecen los estratos 1 y 2, principalmente en las localidades de Ciudad Bolívar, Usme y San Cristobal se presenta una proporción alta de colegios con bajo desempeño en la prueba Saber Pro. En la parte norte en los estratos 6 no se observan colegios mientras que en los estratos del 3 al 5 se ubican colegios con buen desempeño en la pruebas y los resultados están entre alto y superior.

3. LA METODOLOGÍA

Posterior a la construcción de la base de datos, que tiene un tamaño de **623 filas** y un total de **46 columnas**, se procede a la implementación de la metodología con la que se dará alcance a los objetivos planteados, esta cuenta con cuatro etapas fundamentales:

- **Clasificación Jerárquica:** En esta etapa se soluciona uno de los problemas del K-means, al encontrar el número óptimo de clases. El número de clases encontradas en está clasificación es el marco para el desarrollo de los métodos a comparar.
- **K-Means:** Conociendo el número óptimo de clases se procede a generar réplicas mediante el algoritmo de K-means.
- **Formas Fuertes:** A partir de las réplicas del K-means en la etapa anterior, se calculan las formas fuertes.
- **Red Neuronal:** A partir del número óptimo de clases se generan réplicas con este método.

A continuación se detalla cada una de las etapas que componen la metodología, se dan a conocer los resultados obtenidos y su análisis.

3.1. Clasificación jerárquica

El objetivo de la clasificación es agrupar individuos de forma que se minimice una medidad de similaridad o distancia, o en su defecto, separar grupos de individuos de forma que se maximice la medida de similaridad o distancia.

Para la clasificación jerárquica se distinguen dos métodos:

- Los Asociativos o Aglomerativos: Se parte de tantos grupos como individuos hay en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo.
- Los Disociativos o divisivos: Se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez más pequeños.

Los métodos jerárquicos permiten construir un árbol de clasificación, también llamado dendograma.

En nuestro caso utilizamos una clasificación jerárquica aglomerativa, utilizamos dos distancias, city block y euclidea con el método aglomerativo de Ward.

A continuación se explica en que consiste este método

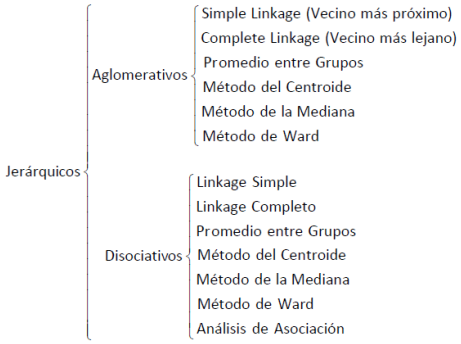


Figura 4: Métodos Analisís Jerárquicos

Método de Ward (o método de pérdida de la inercia mínima)
Cuando se unen dos conglomerados, independiente del método utilizado, la varianza aumenta. El método de Ward consigue unir los casos buscando minimizar la varianza dentro de cada grupo. Para ello se calcula, en primer lugar, la media de todas las variables en cada conglomerado. A continuación, se calcula la distancia entre cada caso y la media del conglomerado, sumando después las distancias entre todos los casos. Posteriormente se agrupan los conglomerados que generan menos aumentos en la suma de las distancias dentro de cada conglomerado. Este procedimiento crea grupos homogéneos y con tamaños similares.

Notando:

- X_{ij} al valor de la j-ésima variable sobre el i-ésimo individuo del k-ésimo grupo, suponiendo que dicho grupo posee n_k individuos.
- m_k al centroide del grupo k, con componentes m_k^j.
- E_k a la suma de cuadrados de los errores del grupo k, o sea, la distancia euclídea al cuadrado entre cada individuo del grupo k a su centroide

$$E_k = \sum_{i=1}^{nk}\sum_{j=1}^{n}(x_{ij}^k - m_j^k)^2 = \sum_{i=1}^{nk}\sum_{j=1}^{n}(x_{ij}^k)^2 - n_k\sum_{j=1}^{n}(m_j^k)^2$$

- E a la suma de cuadrados de los errores para todos los grupos, o sea, si suponemos que hay h grupos

$$E = \sum_{k=1}^{h} E_k$$

El proceso comienza con m grupos, cada uno de los cuales está compuesto por un solo individuo, por lo que cada individuo coincide con el centro del grupo y por lo tanto en este primer paso se tendrá $E_k = 0$ para cada grupo y con ello, $E = 0$. El objetivo del método de Ward es encontrar en cada etapa aquellos dos grupos cuya unión proporcione el menor incremento en E que es la suma total de errores.
Supongamos ahora que los grupos C_p y C_q se unen resultando un nuevo grupo C_t. Entonces el incremento de E será:

$$\Delta E_{pq} = E_t - E_p - E_q =$$
$$\left[\sum_{i=1}^{n_t}\sum_{j=1}^{n}(x_{ij}^t)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2\right] - \left[\sum_{i=1}^{n_p}\sum_{j=1}^{n}(x_{ij}^p)^2 - n_p\sum_{j=1}^{n}(m_j^p)^2\right] - \left[\sum_{i=1}^{n_q}\sum_{j=1}^{n}(x_{ij}^q)^2 - n_t\sum_{j=1}^{n}(m_j^q)^2\right]$$
$$= n_p\sum_{j=1}^{n}(m_j^p)^2 + n_q\sum_{j=1}^{n}(m_j^q)^2 - n_t\sum_{j=1}^{n}(m_j^t)^2$$

El método busca los C_p y C_q tales que ΔE_{pq} sea mínimo.

El Dendograma: Es una representación gráfica en forma de árbol que resume el proceso de agrupación en un análisis jerárquico.
Los objetos similares se conectan mediante enlaces cuya posición en el diagrama está determinada por el nivel de similitud/disimilitud entre los objetos.

Ejemplo: Veamos cómo funciona método de Ward con un caso sencillo, de 5 individuos sobre los cuales se miden dos variables.

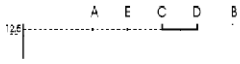
Tabla 4: Tabla de datos.

Individuo	J1	J2
A	10	5
B	20	20
C	30	10
D	30	15
E	5	10

Para este caso, se quiere agrupar los 5 individuos de acuerdo a sus similitudes en las dos variables. El proceso se inicia evaluando el número de particiones que agrupan 2 individuos, realiza dichas las particiones, calcula los centros, evalúa E_k y ΔE_k y agrega lo dos conjuntos que conserven el mínimo ΔE_k, es decir, la mínima inercia.

Nivel 1

En este primer se calcula las posibles combinaciones de individuos, el número es $\binom{5}{2} = 10$



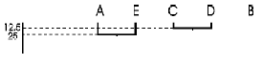
Partición	Centroides	E_k	E	ΔE
$(A,B),C,D,E$	$C_{AB}=(15,12.5)$	$E_{AB}=162.5$ $E_C=E_D=E_E=0$	162,5	162,5
$(A,C),B,D,E$	$C_{AC}=(20,7.5)$	$E_{AC}=212.5$ $E_B=E_D=E_E=0$	212,5	212,5
$(A,D),B,C,E$	$C_{AD}=(20,10)$	$E_{AD}=250$ $E_B=E_C=E_E=0$	250	250
$(A,E),B,C,D$	$C_{AE}=(7.5,7.5)$	$E_{AE}=25$ $E_B=E_C=E_D=0$	25	25
$(B,C),A,D,E$	$C_{BC}=(25,15)$	$E_{BC}=100$ $E_A=E_D=E_E=0$	100	100
$(B,D),A,C,E$	$C_{BD}=(25,17.5)$	$E_{BD}=62.5$ $E_A=E_C=E_E=0$	62,5	62,5
$(B,E),A,C,D$	$C_{BE}=(12.5,15)$	$E_{BE}=162.5$ $E_A=E_C=E_D=0$	162,5	162,5
$(C,D),A,B,E$	$C_{CD}=(30,12.5)$	$E_{CD}=12.5$ $E_A=E_B=E_E=0$	12,5	12,5
$(C,E),A,B,D$	$C_{CE}=(17.5;10)$	$E_{CE}=312.5$ $E_A=E_B=E_D=0$	312,5	312,5
$(D,E),A,B,C$	$C_{DE}=(17.5;12.5)$	$E_{EE}=325$ $E_A=E_B=E_C=0$	325	325

Figura 5: Cálculos a partir del primer nivel.

Observamos que la partición $(C,D),A,B,E$ es la que genera el menor ΔE_k, esto significa que son los individuos más cercanos en las dos variables y con la distancia euclidea. Este resultado da inicio al dendograma, donde se agrupan los individuos CyD. Esto de nota con líneas y la altura de la línea esta definida por ΔE_k.

Nivel 2

En este nivel se considera el grupo (C,D) como un elemento, lo que implica que hay 4 elementos. Se repite el procedimiento del paso anterior, ahora con 4 elementos, donde se tienen $\binom{4}{2}=6$ combinaciones posibles.



Partición	Centroides	E_k	E	ΔE
$(A,C,D),B,E$	$C_{ACD}=(23,33,10)$	$E_{ACD}=316,66$ $E_B=E_E=0$	316,66	304,16
$(B,C,D),A,E$	$C_{BCD}=(25,66,15)$	$E_{BCD}=116,66$ $E_A=E_E=0$	116,66	104,16
$(C,D,E),A,B$	$C_{CDE}=(21,66,11,66)$	$E_{CDE}=433,33$ $E_A=E_B=0$	433,33	420,83
$(A,B),(C,D),E$	$C_{AB}=(15,12,5)$ $C_{CD}=(30,12,5)$	$E_{AB}=162,5$ $E_{CD}=12,5$ $E_E=0$	175	162,5
$(A,E),(C,D),B$	$C_{AE}=(7,5,7,5)$ $C_{CD}=(30,12,5)$	$E_{AE}=25$ $E_{CD}=12,5$ $E_B=0$	37,5	25
$(B,E),(C,D),A$	$C_{BE}=(12,5,15)$ $C_{CD}=(30,12,5)$	$E_{BE}=162,5$ $E_{CD}=12,5$ $E_A=0$	175	162,5

Figura 6: Cálculos a partir del segundo nivel.

Revisamos el ΔE_k y tenemos que es menor en el caso $(A,E),(C,D),B$, lo que significa que ya tenemos un segundo grupo con una altura de 25 que corresponde al ΔE_k.

Nivel 3

Retomando las agrupaciones anteriores nos quedan 3 elementos $(A,E),(C,D),B$, y las posibles combinaciones en 2 grupos son $\binom{3}{2}=3$.

Partición	*Centroides*	E_k	E	ΔE
$(A,C,D,E),B$	$C_{ACDE} = (18{,}75, 10)$	$E_{ACDE} = 568{,}75$ $E_B = 0$	568,75	531,25
$(A,B,E),(C,D)$	$C_{ABE} = (11{,}66, 11{,}66)$ $C_{CD} = (30, 12{,}5)$	$E_{ABE} = 233{,}33$ $E_{CD} = 12{,}5$	245,8	208,3
$(A,E),(B,C,D)$	$C_{AE} = (7{,}5, 7{,}5)$ $C_{BCD} = (26{,}66, 15)$	$E_{AE} = 25$ $E_{BCD} = 116{,}66$	141,66	104,16

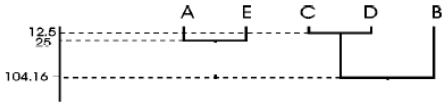


Figura 7: Cálculos a partir del tercer nivel.

Después de ver los cálculos escogemos la agrupación $(A,E),(B,C,D)$. En el dendograma equivale a unir el grupo C,D con el elemento B a una altura de 104,16.

Nivel 4

Evidentemente en este paso se unirán los dos grupos existentes. Los valores del centroide y de los incrementos de las distancias serán los siguientes:

Partición	*Centroide*	E	ΔE
(A,B,C,D,E)	$C_{ABCDE} = (19, 12)$	650	508,34

Figura 8: Resultado final

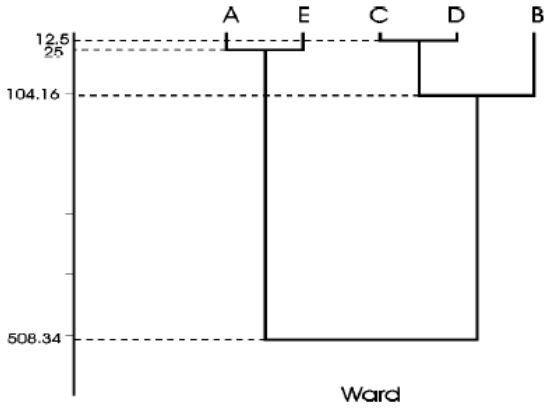


Figura 9: Dendograma

El dendograma es una gráfica de mucha utilidad porque permite visualizar las distancias entre grupos que se unen y nos da idea de el número óptimo de clases.

Número de clases a analizar

Como se ha mencionado, se quiere obtener clases lo más homogéneas posibles y tal que estén suficientemente separadas. Este objetivo se puede concretar numéricamente a partir de la siguiente propiedad:

Supóngase que tenemos una partición $P = (C_1, C_2, ..., C_k)$ del conjunto Ω, donde $g_1, g_2, ..., g_k$ son los centros de gravedad de las clases que corresponden a:

$$g_k = \frac{1}{|C_k|} \sum_{i \epsilon C_k} x_i$$

El centro de gravedad total es:

$$g = \frac{1}{n} \sum_{i=1}^{n} x_i$$

La inercia total del conjunto de puntos, se calcula como:

$$I = \frac{1}{n} \sum_{i=1}^{n} \|x_i - g\|^2$$

Se define :
Inercia inter-clases, como:

$$B(P) = \sum_{n=1}^{N} \frac{|C_k|}{n} \|g_k - g\|^2$$

es decir, la inercia de los centros de gravedad respecto al centro de gravedad total.
Inercia intra-clases, como:

$$W(P) = \sum_{k=1}^{K} I(C_k) = \sum_{k=1}^{K} \frac{1}{n} \sum_{i \epsilon C_k} \|x_i - g_k\|^2$$

Es decir, la inercia el interior de la clase.

EL teorema de igualdad de Fisher, dice que:

$$I = B(P) + W(P)$$

$$\text{Inercia total} = \text{Inercia inter-clases} + \text{Inercia Intra-clases}$$

Entonces como el objetivo de la clasificación es separar las clases, se quiere que $B(P)$ sea máxima. Como la inercia total I es fija para el conjunto dado, maximizar $B(P)$ implica minimizar $W(P)$.

Por tanto los dos objetivos, homogeneidad al interior de las clases y separación entre las clases, se alcanzan al mismo tiempo .

Además:

$$1 = \frac{B(P)}{I} + \frac{W(P)}{I}$$

Entonces un criterio para identificar el número óptimo K de clases, es hallar para una partición en K, clases el índice:

$$\frac{B(P)}{I}$$

y comparar con los demás particiones. Entre más alto el índice mejor la partición.

Para este análisis se realizaron 3 particiones, una con 3, otra con 5 y otra con 7 clases.

Cluster	inter/total Before	Individuos 1	2	3	4	5	6	7
3	0,5594	211	317	95				
5	0,6637	179	226	141	65	12		
7	0,7107	145	66	202	133	65	1	11

Figura 10: Tabla número de grupos vs inercia retenida

Observando el cociente de la varianza inter sobre la total, la opción de 5 grupo retiene el 66 % de varianza y tiene grupos con número suficiente de individuos para ser analizados y generar políticas públicas.
La opción de 6 grupos aunque retiene más inercia, genera grupos con pocos elementos, situación que dificulta generar algún tipo de política.

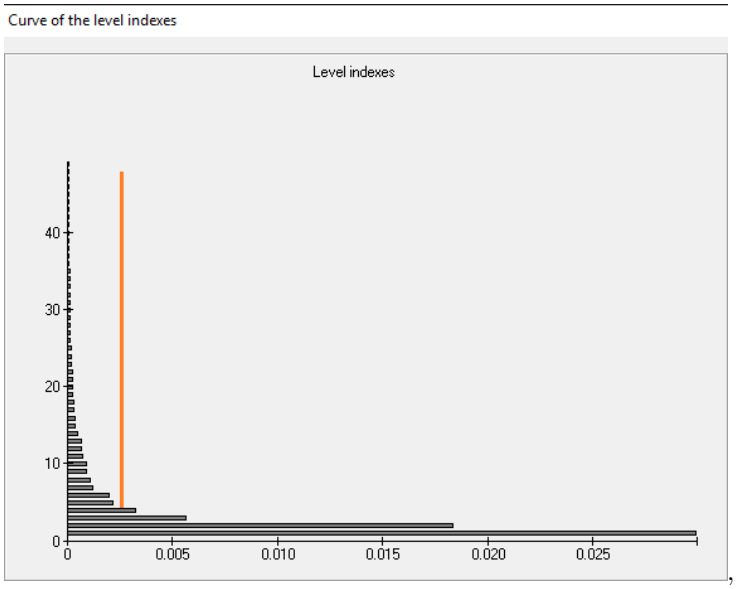


Figura 11: Curva de niveles

Otro criterio para identificar la partción óptima es a partir de la gráfica de los índices de nivel, se puede observar saltos entre niveles que hacen referencia a particiones o grupos, esta no es definitiva en la decisión pero nos puede brindar información adicional sobre el posible número de grupos. Se observa que el número s óptimos para un buen análisis, teniendo en cuenta las diferencias en tamaño entre cada barra denominada nivel, estas diferencias son los saltos y nos da una idea gráficamente si puede haber diferencia entre elegir el nivel 2 o el 3.

Para nuestro análisis tenemos saltos en los niveles 1, 2, 3 ,4 y 6, y los mayores saltos se encuentran en los dos primeros, que seria elegir 2 o 3 particiones, pero este numero de particiones no son apropiadas para nuestro método, ya que son muy pocos grupo. para poder utilizar la información gráfica podemos tener en cuenta la la información obtenida anteriormente, que nos dice que las particiones apropiadas son 3,5 y 7, y de aquí podemos reafirmar que 5 esta dentro de las mejores opciones.

Resultado de los 5 grupos

Se describen los 5 grupos de acuerdo a la variables activas y a las suplementarias que generan diferencias significativas.

El primer grupo, denominado Educación baja, Ingreso bajo se nota, **EbIb**, se caracteriza por:
-Tiene 179 colegios -Padres con los más bajos niveles académicos y donde los ingresos no cubren lo necesario.
-Más colegios de la jornada de la tarde.
-Se auto clasifican como estudiantes con calificaciones inferiores a los demás y en las pruebas saber tienen evaluaciones entre medio y bajo.
-Se ubican en estrato 1 y 2. El promedio del CSE es de 34,3
-Consideran que los baños del colegio no son un lugar seguro, los espacios comunes no son suficientes
-Donde siempre intervienen las directivas y docentes cuando hay agresiones.
-No confían que el colegio los ayude a ser mejores personas.

El segundo grupo, Educación media, Ingreso bajo se nota **EmIb**
-Con 226 colegios, correponde al grupo más grande. -Los padres con nivel de escolaridad entre primaria y algo de bachillerato, donde los ingresos no cubren lo necesario.
-47 % de colegios de la jornada de la mañana y 47 % de jornada de la tarde.
-Son colegios de estrato 2 con resultados medio o alto en las pruebas saber. Con CSE promedio de 49,3 ubicándose en la mitad de la escala.
-Consideran que el colegio donde estudian los ayuda a ser mejores personas.
-Las aulas de clase y espacios comunes son suficientes.
-Los docentes y directivas actúan siempre que hay peleas o agresiones.

El tercer grupo que llamamos Educación media, Ingreso medio, notado **EmIm**, se identifica por:
-Con 141 colegios. -Padres con algo de bachillerato hasta universidad, con ingresos que cubren hasta lo mínimo y en ocasiones un poco más de lo necesario.
-Se ubican en estrato 3 y tiene promedio de 59,3 en el índice de CSE -Se auto clasifican como estudiantes que obtienen calificaciones superior a los demás y obtienen resultados saber entre alto y superior.
-No confían en que los profesores y directivas actúen en pro de los estudiantes, ni confían en que el colegio donde estudian los haga ser mejores personas.
-Las directivas no actúan cuando se presentan injusticias por parte de los profesores, ni intervienen cuando hay peleas o agresiones entre los alumnos.

El cuarto grupo, denominado Educación alta, Ingreso alto, lo notamos **EaIa**, se caracteriza por:
- Tiene 65 colegios. -Padres con algo de bachillerato hasta universidad, aumentando la proporción de bachilleres en comparación con el grupo anterior, con ingresos que alcanzan a cubrir más de lo necesario.
-Se ubican estrato 3 y tiene promedio de 66,9 en el índice de CSE.
-Se auto clasifican como estudiantes que obtienen calificaciones superior a los demás y obtienen resultados saber con nota superior.
-Consideran que los baños del colegio no son seguros y las aulas no se encuentran en buen estado y los espacios comunes no son suficientes.
-Los docentes y directivas intervienen con frecuencia cuando hay peleas o agresiones entre estudiantes.

El quinto grupo con Educación más alta e Ingreso alto, notado **EaaIa** , corresponde a
- El grupo más pequeño con 12 colegios. -Alta proporción de Padres con universidad, con ingresos que alcanzan a cubrir más de lo necesario.
-Se ubican estrato 3 y tiene promedio de 77,26 en el índice de CSE.
-Se auto clasifican como estudiantes que obtienen calificaciones superior a los demás y obtienen resultados saber con nota superior.
-Consideran que los baños del colegio son seguros, las aulas se encuentran en buen estado y los espacios comunes son suficientes.

-Los docentes y directivas intervienen siempre que hay peleas o agresiones entre estudiantes.

Ubicación de los 5 clústers

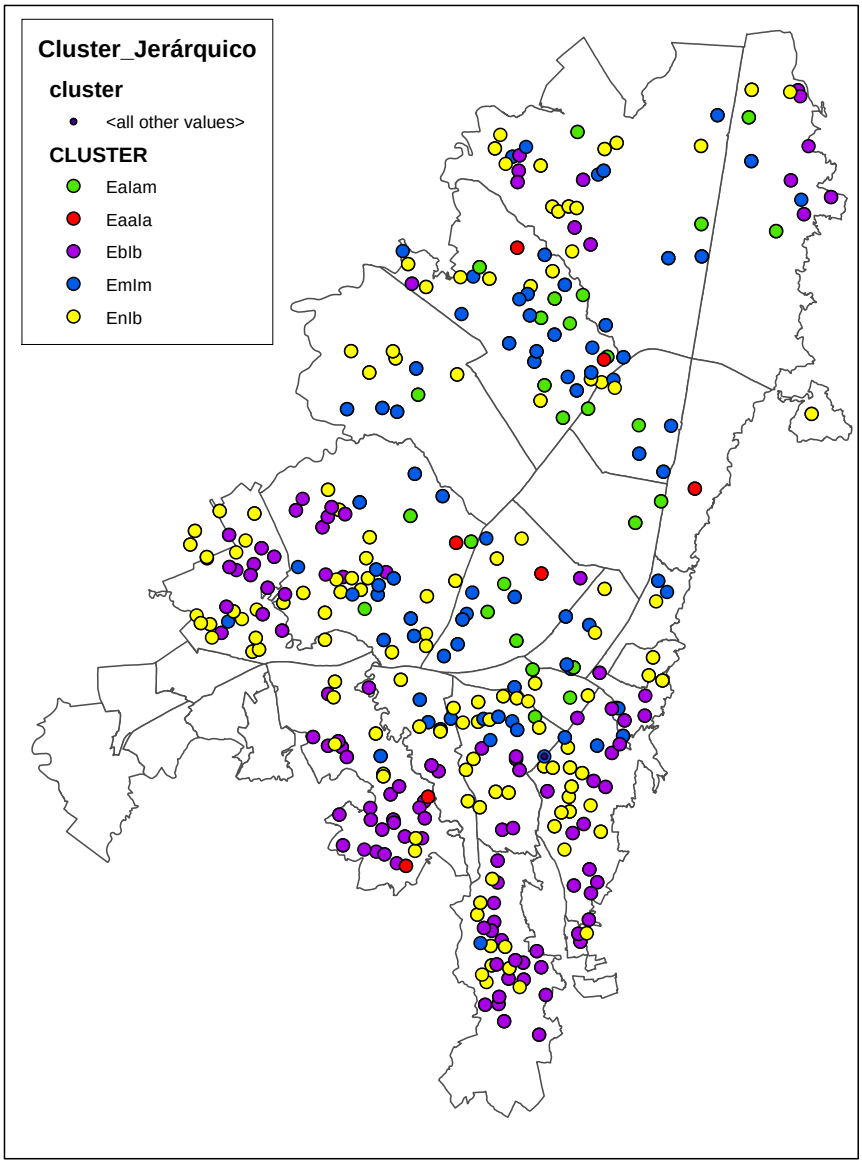


Figura 12: Ubicación de colegios oficiales caracterizados por la clasificaón Jerárquica.

Se construye un mapa con las variables clúster segun clasificación jerárquica, para ubicar los cluster dentro del mapa de Bogotá, donde se observa que:

- Los colegios pertenecientes al clúster EbIb(Educación baja, Ingreso bajo) se ubican en su mayoria en el sur de la ciudad, donde predomina el estrato 1 y 2, pertenecientes a las localidades de San Cristobál, Ciudad Bolivar, Usme y Bosa. En la parte norte vemos algunos de estos colegios que se ubican en la periferia de Usaquen, y barrios de estrato bajo en Suba.
- Los colegios pertenecientes al clúster EmIb(Educación media, Ingreso bajo) tienden a estar combinados con el anterior cluster en su ubicación dentro de la ciudad, a diferencia de el cluster EbIb se observan algunos colegios en la parte central de Bogotá en las localidades de Puente Aranda y Barrios Unidos.
- Los colegios pertenecientes al clúster EmIm(Educación media, Ingreso medio) se ubican en los barrios de estrato medio y algunos pocos en barrios de estrato bajo, su mayor concentración esta en las localidades de Antonio Nariño y Kenedy.
- Los colegios pertenecientes al clúster EaIa(Educación alta, Ingreso alto) su ubicaón esta muy cercana al anterior pero a diferencia de este no se ubica en barrios de estrato bajo y su concentración esta en las localidades de Barrios Unidos y Kennedy.
- Los colegios pertenecientes al clúster EaaIa(Educación más alta, Ingreso alto) son pocos y se ubican en la localidad de Kennedy, Teusaquillo, Barrios unidos y Ciudad Bolivar.

3.2. K-Means

El objetivo de K-means es encontrar una partición P de Ω y representantes de las clases, tales que $W(P)$ sea mínima.

Existe un poco de confusión en la literatura acerca del método k-means, ya que hay dos métodos distintos que son llamados con el mismo nombre. Originalmente, Forgy propuso en 1965 un primer método de reasignación recentraje que consiste básicamente en la iteración sucesiva, hasta obtener convergencia, de las dos operaciones siguientes:

- Representar un a clase por su centro de gravedad , es decir, por su vector de medias.
- Asignar dos objetos a la clase del centro de gravedad más cercano.

McQueen propone un método muy similar, donde también se representan las clases por su centro de gravedad, y se examina cada individuo para asignarlo a la clase más cercana. La diferencia con el método de Forgy es que inmediatamente después de asignar un individuo a una clase, el centro de ésta es recalculado, mientras que Forgy primero hace todas las asignaciones y luego recalcula los centros. Variantes del método de Forgy son propuestas en Francia como Método de Nubes Dinámicas por E. Diday a partir de 1967. McQueen es quien propone el nombre de “k -means”, que se usa hasta la fecha, estos métodos también reciben nombres como nubes dinámicas, centros móviles, o reasignación-recentraje.

Algoritmo K-means

0. Inicialización: Escoger al azar K objetos de Ω,que servirán cómo núcleos o centros iniciales. Significa escoger al azar $g_1, g_1, ..., g_k$ en Ω; sean :

$$C_1 = \emptyset, ..., C_k = \emptyset$$

1. Asignación: Asignar cada objeto a la clase del centro de gravedad más cercano. es decir, para todo i $\epsilon\Omega$, hacer:
Si $d(x_i, g_{k^*}) \leq \{d(x_i, g_k)$ para todo $k = 1, ..., K\}$ entonces asignar x_i a la clase C_{k^*}; si el mínimo se alcanza para dos clases diferentes entonces asignarlo a la clase de índice menor, es decir, con el menor número.

2. Representación: Calcular los centros de gravedad de la partición. Así, para todo $k\epsilon\{1, ..., K\}$ hacer: $g_k = \frac{1}{|C_k|} \sum_{x_i \epsilon C_k} x_i$. Calcular el criterio $W = \sum_{k=1}^{K} \sum_{x_i \epsilon C_k} \|x_i - g_k\|^2$

3. Control de parada: Si la variación en el criterio W entre la iteración precedente y la presente es menor que un umbral dado, o si se sobrepasa el número máximo de iteraciones entonces detenerse, de lo contrario ir al paso 1.

El resultado de la aplicación del método k-means, depende de la escogencia de los núcleos.

Formas fuertes. Este método surge como una solución a la debilidad que tiene el algoritmo s K-means. La técnica de formas fuertes consiste, básicamente, en hacer réplicas de las particiones del conjunto inicial, utilizando cada vez diferentes centros. Luego, revisar en todas las réplicas los individuos que permanecieron juntos.
Estos grupos se consideran estables, lo que significa que independientemente de la selección de los centros, la distancia entre ellos es pequeña lo que hace que sean similares.

3.3. Análisis de las formas Fuertes con K-means

Conociendo que el número óptimo de clases es 5, se procedió a realizar réplicas del método de K-means. Cada conjunto de réplicas las denominamos ciclos, y en las casillas nombradas inicial se escribe el número de individuos que forman cada clase en la primera réplica del ciclo. Hacemos los ciclos utilizando la distancia euclídea.

Distancia euclidea:

$$d_{ij} = \sqrt{\sum_{r=1}^{t} (X_{ir} - X_{jr})^2}$$

Réplicas	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
5	188	117	141	140	219	159	7	7	68	63
		45		1		52				2
		24				8				2
		2								1
Réplicas										
10	235	130	7	7	66	55	149	128	166	101
		52				7		12		45
		28				2		8		15
		22				1		1		2
		2				1				2
		1								1
Réplicas										
20	205	140	15	7	191	103	137	137	75	55
		53		7		46				13
		7		1		18				3
		3				16				2
		1				6				1
		1				2				1
Réplicas										
30	188	101	68	55	7	7	220	130	140	128
		46		7				53		9
		15		3				22		3
		9		1				6		
		7		1				4		
		3		1				3		
		3						1		
		2								
		2								

Tabla 5: Tabla de resultados con 5, 10, 20 y 30 réplicas,distancia euclidea

Analizando los 4 ciclos, de 5, 10, 20 y 30 réplicas con el método k-means y la distancia euclidea(tabla 12), obtenemos que clase más pequeña es estable, se mantiene a lo largo de los ciclos. La segunda clase, no tan estable como la menor, mantiene unido un buen número de elementos la más estable, manteniendo en los ciclos analizados. Las otras dos clases son las menos estables porque se subdividen en más de una forma fuerte, mostrando posibles subclases.

Réplicas	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
5	152	152	169	169	121	121	111	111	70	1
										69

Réplicas	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
10	112	99	169	16	152	9	69	67	121	3
		1		153		135		2		118
		12				8				

Réplicas	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
20	152	5	121	1	167	2	113	2	70	2
		2		2		150		1		1
		3		118		1		49		40
		5				2		49		27
		128				6		7		
		1,1,1				3		5		
		6				3				

Réplicas	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
30	118	118	163	94	164	146	67	67	111	62
				26		6				20
				11		4				16
				9		2				4
				8		2				3
				4		2				2
				3		1				1-1-1-1
				1-1-1-1		1				

Tabla 6: Tabla de resultados con 5, 10, 20 y 30 réplicas, distancia city block

Dado que el método K-means tiene como criterio principal la distancia entre elementos, hacemos las réplicas utilizando la distancia del city block, con el fin de identificar cual distancia, en nuestro caso, puede ser mas robusta. Entendiendo robusta, que conserva la mayor proporción de individuos en cada clase. Distancia del city block, también llamada distancia absoluta o de Manhattan:

$$d_{ij} = \sum_{r=1}^{t} |X_{ir} - X_{jr}|$$

Hay 3 clases que mantienen una forma fuerte a lo largo de las réplicas y dos clases que se subdividen mostrando tres subclases en cada caso (tabla 13).

Comparando las formas fuertes que se obtienen en los diferentes ciclos y con las las dos distancias, tenemos que la distancia euclidea aún con pocas réplicas(5) encuentra formas que se mantienen a lo largo de los ciclos. Es decir que las identifica rápidamente.
La distancia del city block necesita un número elevado de réplicas para encontrar divisiones en las formas fuertes.
Era de esperarse que el número de individuos en las clases, usando una u otra distancia, fueran diferentes.

Formas fuertes con distancia Euclidea				
Réplicas	5	10	20	30
Cluster 1	117, 45	101, 45	103, 46	101, 46
Cluster 2	140	128	137	128
Cluster 3	159, 52	130, 52	140, 53	130, 53
Cluster 4	7	7	7	7
Cluster 5	63	55	55	55

Formas fuertes con distancia City Block				
Réplicas	5	10	20	30
Cluster 1	152	135	128	94,26,11
Cluster 2	169	153	150	146
Cluster 3	121	118	118	118
Cluster 4	111	99	49-49	62-20
Cluster 5	69	67	40-27	67

Tabla 7: Reúmen de formas fuertes utilizando distancia euclídea y distancia city block

3.4. Redes Neuronales

Los sistemas nerviosos son los sistemas de procesamiento de información más complejos que se conocen y están formados por un gran número de células llamadas neuronas. Las neuronas contienen energía potencial eléctrica que se utiliza para enviar señales a otras neuronas. Esta energía proviene de la diferencia de potencial generada por las concentraciones de iones de Sodio (Na^+) y Potasio (H^+), tanto al interior como al exterior de la célula.

3.4.1. Un modelo Biológico

Las neuronas se comunican con otras mediante la transmisión de impulsos nerviosos a través de las sinapsis. Si la suma de los impulsos que recibe una neurona supera un cierto valor umbral b la neurona es capaz de enviar un nuevo impulso. Este comportamiento se describe en términos matemáticos, como se explica a continuación.

Si $x = (x_1, x_2, \ldots, x_n)$ es el vector que representa cada uno de los impulsos que recibe la neurona y $w = (w_1, w_2, \ldots, w_n)$ los pesos de cada una de las conexiones, el producto escalar $x \cdot w = x_1 w_1 + \cdots x_n w_n$ representa la suma total de las sinapsis. Si este valor supera el valor umbral b, la neurona se activa de acuerdo con la función de salida: $o(x \cdot w - b) = \begin{cases} 1 & x \cdot w - b \geq 0 \\ 0 & x \cdot w - b < 0 \end{cases}$.

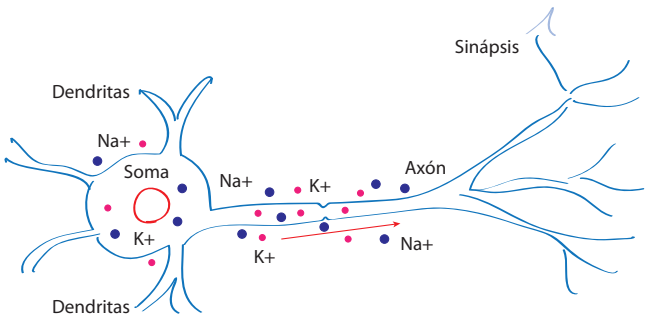


Figura 13: Neurona biológica.

Este modelo biológico fue simplificado y escrito en términos matemáticos.

3.4.2. Funciones de activación

En un sentido estricto la activación de una neurona funciona con la idea de todo o nada, lo cual se puede modelar mediante la función escalón unitario: $o(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$ que se muestra en la figura 14.

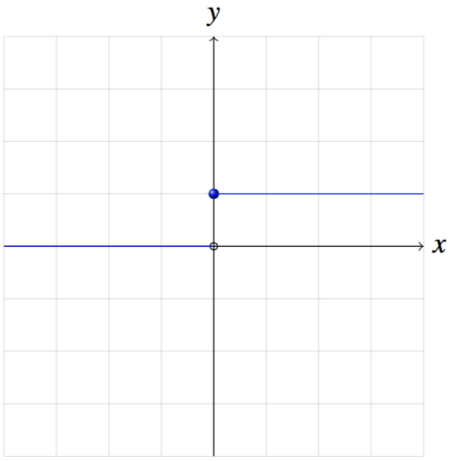


Figura 14: Gráfica de la función $o(x)$.

Desde el punto de vista matemático la función de salida puede modelarse con otras funciones como las que se listan a continuación (los nombres han sido usados de acuerdo con la plataforma de MATLAB).

1. Hardlim: $o(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$
2. Hardlims: $o(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}$
3. Lineal: $o(x) = x$
4. Logsig: $o(x) = \dfrac{1}{1+e^{-x}}$
5. Tansig: $o(x) = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$

Las funciones 3 a 5 son funciones continuas y diferenciables, cuyas derivadas se expresan en términos de las mismos valores de la función, por ejemplo, para la función Logsig $o(x)$ tenemos que $o'(x) = o(1-o)$, lo cual es conveniente en el proceso de minimización del error.

3.4.3. Redes Supervisadas

Dentro de los diferentes tipos de redes supervisadas se encuentra el perceptrón de varias capas, el cual se emplea en modelos de clasificación. Un perceptrón multicapa tiene la topología que se muestra en la figura 15.

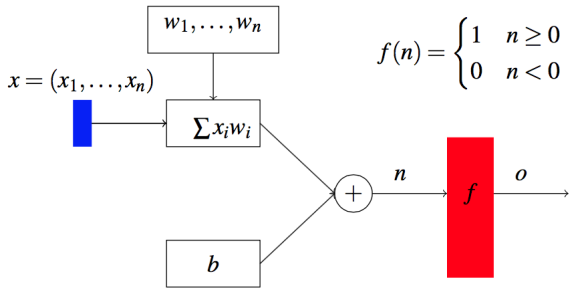


Figura 15: Modelo matemático de una Neurona.

El entrenamiento de una red consiste en la determinación de los valores (pesos) w_i tal que la salida de la red $o(x \cdot w)$ coincidan con los valores deseados y_i o en su defecto minimice el error, el cual está dado por:

$$E = \frac{1}{2} \sum_{i=1}^{n} (y_i - o_i)^2.$$

Se presenta el algoritmo de entrenamiento para redes supervisadas, a partir de un conjunto de entrenamiento: $P = \{(x_1, y_1), \ldots, (x_K, y_K)\}$. Los vectores x_i representan las entradas a la red y los valores y_i las salidas esperadas.

Se elige un parámetro $\eta > 0$ y de determinan los pesos w_i en forma aleatoria. Para cada entrada x_k se calcula la salida 0_k y se compara con el valor esperado y_k. De acuerdo con el error $|y_k - o_k|$ los pesos w_i se modifican. Se acumula el error producido durante un ciclo, que consiste en presentar todos los datos. El entrenamiento termina cuando el error sea 0, o cuando se cumpla un determinado número de ciclos.

Un esquema del algoritmo de entrenamiento del perceptrón se muestra a continuación.

Algorithm 1 Entrenamiento perceptrón

1: $\eta > 0$;
2: $w_i := rand$;
3: **while** $k \leq K$ **do**
4: $x \leftarrow x_k, y \leftarrow y_k, o \leftarrow o(x)$
5: $w_i \leftarrow w_i + \eta(y - o)x$
6: $E \leftarrow E + \frac{1}{2}(y - o)^2$
7: $k \leftarrow k + 1$
8: **end while**
9: **if** $E = 0$ **then**
10: Termina entrenamiento
11: **else**
12: $E = 0, k = 1$, inicia nuevo entrenamiento.
13: **end if**

3.4.4. Redes no supervisadas

Son redes para los cuales su entrenamiento solo tiene datos de estrada y no hay datos de salida, con los que se pueda evaluar su desempeño. Son redes apropiadas para aplicarlas en problemas de *grupoing*. Es

decir, este tipo de redes puede agrupar los objetos similares y separar los que no lo son.

Redes competitivas Las redes competitivas son redes no supervisadas, donde las neuronas compiten entre sí, para determinar quién está más cerca de la entrada x.

Suponemos que un conjunto de datos en $\mathbb{R}^n$ deber ser segmentado en k grupos diferentes. Dada una entrada $x \in \mathbb{R}^n$, la idea es elegir entre un conjunto de pesos $\{w_1, w_2, \cdots, w_k\}$ el vector de pesos w_r que esté más cerca a la entrada x, es decir, hallar el w_r tal que:

$$\|x - w_r\| = \min_i \|x - w_i\|.$$

El índice r representa la neurona ganadora, el cual representa el vector que está más cercano a la entrada x. En general se supone que los pesos son vectores normales, es decir, $\|w_i\| = 1$ para todo i.

Usando la igualdad $\|x - w_r\|^2 = (x - w_i) \cdot (x - w_i) = \|x\|^2 - 2x \cdot w_i + \|w_i\|^2$ se puede ver que hallar el mínimo de las m distancias equivale a hallar el máximo de los m productos escalares:

$$x \cdot w_r = \max_i \{x \cdot w_i\}.$$

Algoritmo Kohonen (el ganador toma todo) Dado el conjunto de vectores: $\{x_1, \ldots, x_K\}$, para $x_i \in \mathbb{R}^n$.

1. $\eta > 0$
2. Se definen los w_i aleatorios, con $\|w_i\| = 1$.
3. $w_r = w_r + \eta(x - w_r)$
4. $w_r = w_r / \|(w_r\|$ (normalización)
5. $w_i = w_i$ para $i \neq r$

De acuerdo con la regla de actualización $w_r = w_r + \eta(x - w_r) = \eta x + (1 - \eta) w_r$ se puede ver que el nuevo valor de w_r es una combinación lineal convexa de w_r y x, por tanto estará dentro del segmento negro de la figura 16.

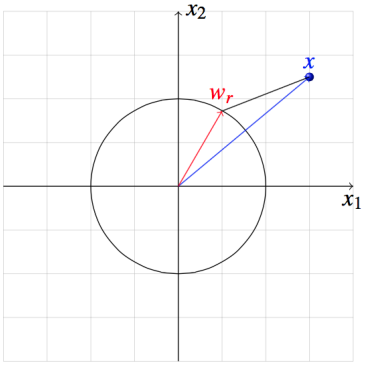


Figura 16: La actualización de los pesos de acuerdo con la regla.

Mapas auto-organizados Los mapas auto-organizados, con sigla SOM del inglés, son redes no supervisadas similares a las redes competitivas, con la diferencia de que no sólo se actualiza la neurona ganadora w_r, sino que se actualizan todos los pesos correspondientes a toda una vecindad de w_r.

Teniendo en cuenta los 5 grupos definidos, el mapa SOM utilizado es una matriz de tamaño 5×1.

Para la base de datos, se preparó una función en MATLAB, como se muestra a continuación.

```
function[A, tr]=formasf(a,r)
at=a';
p=at(2:10,1:623);
A=zeros(623,r);
for i=1:r;
net = selforgmap([5,1]);
net.trainParam.epochs = 2000;
net=configure(net,p);
net=setwb(net,rand(5,9));
net = train(net,p);
s = sim(net,p);
sc = vec2ind(s)';
A(:,i) = sc;
end
```

La función depende de dos parámetros:

- La matriz que contiene los datos: a.
- El número de réplicas que se requieren: r.

Cada iteración del ciclo `For` define una nueva red SOM, se entrena por 2000 épocas, simula los datos y las resultados se almacenan en una de las columnas de la matriz A.

La instrucción: `net=setwb(net,rand(5,9));` establece nuevos pesos aleatorios a la red *net*. Los resultados de cada réplica se almacenan en la matriz A, mediante el comando: `A(:,i) = sc;`.

3.5. Análisis de formas fuertes con redes SOM

Las cinco clases obtenidas para los 623 colegios de la base de datos, mediante una red SOM específica da lugar a los siguientes grupos:

Obviamente cada red SOM determina grupos diferentes, no obstante se trata de capturar las formas fuertes, para lo cual se construye una red diferente para cada una de las 5, 10, 20 y 30 réplicas. Los grupos obtenidos en cada uno de estos casos se muestran en las tablas de las figuras 17 a 20.

Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
7	7	68	67	186	186	222	2	140	140
			1				219		
							1		

Figura 17: Tabla de resultados con 5 réplicas.

Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
7	7	68	67	186	184	221	1	141	140
			1		1		1		1
							1		
					1		218		

Figura 18: Tabla de resultados con 10 réplicas.

Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
8	7	67	1	188	1	220	1	140	140
	1		66		1		217		
					184		1		
					1		1		
					1				

Figura 19: Tabla de resultados con 20 réplicas.

Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final	Inicial	Final
7	7	68	1	188	1	219	1	141	140
			66		1		217		
			1		182		1		
					1				
					1				
					1				
					1				

Figura 20: Tabla de resultados con 30 réplicas.

La columna inicial corresponde a la clasificación preliminar. Luego se hacen 5, 10, 20 y 30 réplicas y se buscan los individuos que en todas las réplicas quedan en el mismo grupo. Por ejemplo, en la segunda columna de la tabla de la figura 19 se observa: Inicial 67; final 1, 66. Esto significa que de después de 20 réplicas 66 individuos de los 67 iniciales quedaron siempre en el mismo grupo y sólo 1 ha quedado en en grupo diferente.

De acuerdo con las tablas anteriores, puede verse que las formas fuertes están presentes en las diferentes aplicaciones del método SOM. Aunque es evidente que a mayor número de réplicas, hay una mayor atomización de los grupos, puede decirse, en términos generales que las redes SOM en general son robustas, ya que conservan los diferentes grupos de la clasificación.

A partir de los resultados proporcionados del método de Redes Neuronales no Supervisadas en los 4 ciclos y con pesos aleatorios, se obtiene que presenta un comportamiento con variabilidad minima en el tamaño de los grupo entre ciclos, comparando con los grupo iniciales, se mantiene la misma distribución. De igual manera teniendo en cuenta la distribución poco homogenea observada desde la primera replica, se evidencia que no hay cambios después de aplicar varias de estas, teniendo en cuenta que se aplicaron diferentes pesos a los individuos. Cabe anotar que no todos los grupo tuvieron el mismo comportamiento, siendo asi que dos de los grupo con mayor densidad presentaron variaciones en una porción minima de individuos, pero que se debe mirar con cuidado, dado que son individuos que no se puedieron ajustar al grupo. otro punto a favor dentro de este método es la fácil identificación de en que grupo quedó agrupado cada individuo en cada replica, dado que no hay variación en el numero del grupo.

4. Comparación

Inicialmente vamos a comparar los tamaños de las clases obtenidas con los difrentes métodos y distancias. En el caso de k-means y redes neuronales, se reporta el tamaño de los grupos de la primera réplica del ciclo de 5, dado que el comportamiento a nivel de tamaños es muy similar en cada método.

Clasificación Jerárquica	Clases	K means - dist. euclidea	K means - dist. City block	Redes
179	Cluster 1	188	152	188
141	Cluster 2	140	169	141
226	Cluster 3	219	121	219
12	Cluster 4	7	70	7
65	Cluster 5	68	111	68

Tabla 8: Reúmen del tamaño de las clases según el método

Podemos ver que ni el k-means con diferente distancia, ni las redes, clasifican con clases de tamaños igual a la clasificación por el método jerárquico.

Las redes neuronales y el k-means con distancia euclídea consiguen la misma clasificación. Considerando el tamaño de las clases podría pensarse que es similar a la jerárquica, sin embargo al hacer la matriz de confusión se concluye que es una clasificación totalmente diferente.

		Euclídea					
		1	2	3	4	5	Total
		n	n	n	n	n	n
redes	1		140			1	141
	2	1		218			219
	3	187				1	188
	4		1	1		66	68
	5				7		7
Total		188	141	219	7	68	623

		Jerárquica					
		1	2	3	4	5	Total
		n	n	n	n	n	n
redes	1	50	54	29	5	3	141
	2	54	82	49	29	5	219
	3	55	69	38	22	4	188
	4	20	19	23	6	0	68
	5	0	2	2	3	0	7
Total		179	226	141	65	12	623

		Jerárquica					
		1	2	3	4	5	Total
		n	n	n	n	n	n
euclidea	1	54	71	38	21	4	188
	2	51	53	29	5	3	141
	3	53	83	49	29	5	219
	4	0	2	2	3	0	7
	5	21	17	23	7	0	68
Total		179	226	141	65	12	623

Tabla 9: Matrices de confusión

Analizando las formas fuertes generadas con k-means con distancia euclídea y las redes, se observa que las redes son muy estables y se enfocan a conseguir los mejores 5 grupos, mientras que el k-means encuentra nuevos subgropos.

Formas fuertes con distancia Euclidea				
Réplicas	5	10	20	30
Cluster 1	117, 45	101, 45	103, 46	101, 46
Cluster 2	140	128	137	128
Cluster 3	159, 52	130, 52	140, 53	130, 53
Cluster 4	7	7	7	7
Cluster 5	63	55	55	55

Formas fuertes con Redes neuronales				
Réplicas	5	10	20	30
Cluster 1	186	184	184	182
Cluster 2	140	140	140	140
Cluster 3	219	218	217	217
Cluster 4	7	7	7	7
Cluster 5	67	67	66	66

Tabla 10: Formas fuertes según los métodos

Conclusiones

A nivel social encontramos 5 clases de colegios que se diferencian por el nivel de escolaridad de los padres y que a su vez está relacionada con el nivel de ingresos.

Respecto a las clasificaciones encontramos que el algoritmo que define las redes neuronales equivale al k-means con la distancia euclídea, sin embargo, las formas fuertes para las redes neuronales no surten el mismo efecto que para las nubes dinámicas.

En este contexto la distancia del city block no es eficiente para conseguir formas fuertes dado que sus ciclos deben ser grandes para que las capte.

Referencias

[1] Hartigan, J. A. and Wong, M. A., *A K-means Clustering Algorithm*, Applied Statistics, 28:100-108, (1979)

[2] Le Donné N. et Rocher T., *Une meilleure mesure du contexte socio-éducatif des élves et des écoles. Construction d'un indice deposition sociale á partir des professons des parents*, Éducation & Formations, n 79, p. 103-116, (2010).

[3] Lebart, L., Piron, M. and Morineau, A., *Statisitique exploratoire multidimensionnelle. Visualisation et inférence en fouilles de donnes*, 4 ed, pg 256, (2006), Dunod, Paris.

[4] Lebart, L., Morineau, A., Lambert, T. & Pleuvret, P., SPAD. Systeme Pour L' Analyse des Donees, (1999),Paris.
*http://www.coheris.com/produits/analytics/logiciel-data-mining/

[5] Oberti, M., Préteceille O., Rivire C.,*Les effets de l'assouplissement de la carte scolaire dans la banlieue parisienne*, Rapport de la recherche réalisée pour la HALDE,pg 22-23, (2012).

[6] Rubia, F., *La segregación escolar en nuestro sistema educativo*. Revista digital fórum aragón. Volumen 10. Recuperado de http://feae.eu/doc/ara/revista_digital_forum_aragon_10.pdf (2013)

[7] Hector Gomez A.., *Clasificación de imagenes multiespectrales lansat TM por medio de redes neuronales no supervisadas.* Universidad Nacional de estudios a distancia (2006)

[8] Secretaría de Educación del Distrito. Alcaldía mayor de Bogotá. Recuperado de http://www.educacionbogota.edu.co/es/

[9] Secretaría de Educación del Distrito. Alcaldía mayor de Bogotá. (2014). Clima escolar y victimización, en Bogotá 2013. Encuesta de convivencia escolar.

[10] Documento realizado por estudiantes de Departamento de Organización y Estructura de la Información, Escuela Universitaria de Informática, Universidad Politécnica de Madrid. Ctra. De Valencia, Km. 7, 28031, Madrid, Espantildea.

[11] Publicación tesis de grado de Hernán Grosser, Facultad de ingeniería, Universidad de Buenos Aires.

[12] Publicacion Percepcion de las ciencias y tecnologías en Colombia, Observatorio Colombiano de Ciencia y Tecnologia.