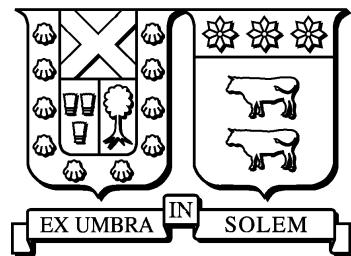


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



“PERFILAMIENTO DE ESTUDIANTES Y
ESCUELAS A TRAVÉS DE MODELOS DE
APRENDIZAJE AUTOMÁTICO PARA LA
ELECCIÓN DE ESCUELAS EN CHILE”

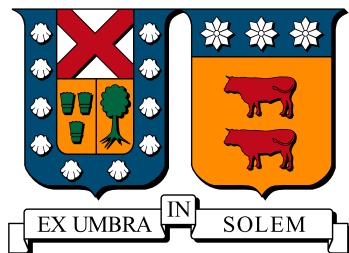
CARLOS ANDRÉS VARGAS POBLETE

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIOUBOV DOMBROVSKAIA

ABRIL 2018

**UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE**



**“PERFILAMIENTO DE ESTUDIANTES Y
ESCUELAS A TRAVÉS DE MODELOS DE
APRENDIZAJE AUTOMÁTICO PARA LA
ELECCIÓN DE ESCUELAS EN CHILE”**

CARLOS ANDRÉS VARGAS POBLETE
MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO
PROFESOR GUÍA: LIOUBOV DOMBROVSKAIA
PROFESOR CORREFERENTE: PATRICIO RODRÍGUEZ

ABRIL 2018

MATERIAL DE REFERENCIA, SU USO NO INVOLUCRA RESPONSABILIDAD DEL AUTOR O DE LA INSTITUCIÓN

Agradecimientos

Resumen

El análisis de datos es un proceso fundamental para obtener información y con esta poder tomar diversas decisiones. En el ámbito escolar los establecimientos y estudiantes se agrupan por su dependencia o por la dependencia del colegio al que asisten. Este estudio busca encontrar mediante un algoritmo de clasificación no supervisado los diferentes grupos de colegios y alumnos, determinando cuales son sus principales características. Se comparan los resultados de un algoritmo al ser ejecutado con tres grupos de variables distintas y luego al mejor resultado se le agregan variables de relación establecimiento - matrículas y se compara con la primera versión. Los resultados más importantes son que los establecimientos y matrículas encuentran un óptimo de 4 grupos de clasificación con el grupo de menor cantidad de variables, donde una de las más relevantes a la hora de separar los grupos es el nivel de copago existente.

Palabras Clave: Análisis de datos, Aprendizaje No Supervisado, X-Means.

Abstract

Keywords: Data Analysis, Unsupervised Learning, X-Means.

Índice de Contenidos

Agradecimientos	III
Resumen	IV
Abstract	V
Índice de Contenidos	VI
Lista de Tablas	VIII
Lista de Figuras	X
Glosario	XII
Introducción	1
1. Definición del Problema	3
2. Estado del Arte	5
2.1. Machine Learning	7
2.2. Clustering	9
2.2.1. K-Means	10
2.2.2. X-Means	11

2.3. CRISP-DM	15
3. Propuesta de Solución	18
3.1. Metodología de trabajo	18
4. Resultados y Análisis	24
4.1. Resultados obtenidos	24
4.2. Análisis de los resultados	30
4.2.1. Análisis cualitativo	30
4.2.2. Análisis geográfico	36
Conclusiones	42
A. Anexo I	43
Bibliografía	44

Índice de tablas

3.1. Atributos que componen la base de datos de los establecimientos.	19
3.2. Atributos que componen la base de datos de las matrículas.	20
3.3. Atributos de relación establecimiento-matrícula para la base de datos de los establecimientos.	21
3.4. Atributos de relación establecimiento-matrícula para la base de datos de las matrículas.	22
4.1. Clústers de establecimientos variando la cantidad de atributos (por grupos de importancia).	25
4.2. Comparativa de clústers de establecimientos por atributo.	25
4.3. Clústers de establecimientos (tabla 4.2) según su dependencia.	25
4.4. Comparativa de clústers de establecimientos por atributo, incluyendo los de relación establecimiento-matrícula.	26
4.5. Clústers de establecimientos (tabla 4.4) según su dependencia.	27
4.6. Detalle de la sobre edad en los clústers de establecimientos.	27
4.7. Clústers de matrículas variando la cantidad de atributos (por grupos de importancia).	28
4.8. Comparativa de clústers de matrículas por atributo.	28

4.9. Detalle de la sobre edad en los clústers de matrículas.	28
4.10. Comparativa de clústers de matrículas por atributo, incluyendo los de relación establecimiento-matrícula.	29
4.11. Detalle de la sobre edad en los clústers de matrículas, incluyendo atributos de relación.	29
A.1. Resumen	43

Índice de figuras

2.1.	Gráfico de 8 objetos. [8]	9
2.2.	Iteración X-Means.	14
2.3.	Diagrama de proceso que muestra la relación entre las diferentes fases de CRISP-DM.	16
4.1.	Promedios de atributos normalizados para establecimientos de la Región Metropolitana.	31
4.2.	Promedios de atributos normalizados de establecimientos con atributos relacionales (tabla 3.3) de la Región Metropolitana.	31
4.3.	Promedio de atributos normalizados de matrículas de la Región Metropolitana.	33
4.4.	Promedio de atributos normalizados de matrículas con atributos relacionales (tabla 3.4) de la Región Metropolitana.	33
4.5.	Composición de los clústers de establecimientos según los clústers de matrículas (sin variables de relación).	35
4.6.	Composición de los clústers de establecimientos según los clústers de matrículas (con variables de relación).	36
4.7.	Mapas de clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.	37

4.8. Mapas de calor de matrículas (con atributos relacionales) en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.	39
4.9. Mapas de calor de clústers de matrículas sobre mapa GSE de la Región Metropolitana.	40
4.10. Mapas de calor de clústers de matrículas (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.	41

Glosario

BIC: Bayesian Information Criterion (Criterio de información bayesiano).

CIAE: Centro de Investigación Avanzada en Educación.

CRISP-DM: Cross Industry Standard Process for Data Mining.

GSE: Grupo Socioeconómico.

IDE: Índice de Desarrollo de la Educación.

MICE: Multiple Imputation by Chained Equations (Imputación múltiple por ecuaciones encadenadas).

SEP: Subvención Escolar Preferencial.

Introducción

En Chile los establecimientos educacionales se encuentran categorizados según su dependencia administrativa en municipal, particular subvencionado, particular pagado y corporación de administración delegada, los cuales en la Región Metropolitana se distribuyen de la siguiente manera 23,8 %, 65,2 %, 10 % y 1 % respectivamente [13]. Por otro lado, los estudiantes matriculados en dichos colegios no poseen una categorización clara, y las clasificaciones más cercanas son por el tipo de colegio al cual asisten o por su nivel socioeconómico. La falta de perfiles de establecimientos y matrículas, basados en múltiples atributos, dificulta la tarea de llevar a cabo políticas públicas enfocadas a ciertos perfiles educaciones y estudiantiles.

Debido a esto es que con este estudio se busca encontrar, a partir de diversas variables, la estructura que tienen los establecimientos en la Región Metropolitana y los estudiantes que a ellos asisten, de manera de poder realizar una mejor clasificación.

Para esto se utilizó una metodología propia basada en CRISP-DM, en donde primero se tomaron diferentes fuentes de información, las cuales fueron corregidas y estandarizadas para poder trabajar de manera sencilla con ellas. A continuación de esto se escogieron variables de interés propias de cada establecimiento o estudiante, para poder estudiar su relevancia y también variables que las relacionan.

Luego, con el fin de no tener una categorización previa de los datos, se utilizó un algoritmo de aprendizaje no supervisado para poder deducir una clasificación directamente de ellos. Primero se realizó solo con las variables propias para así poder seleccionar las más relevantes. Después, el resultado obtenido se compara con los resultados que se obtienen al incluir

las variables de relación, tanto en el caso de los establecimientos como de las matrículas, para analizar si al añadir este tipo de información genera un enriquecimiento de la categorización obtenida anteriormente.

Finalmente, con las geolocalizaciones de establecimientos y estudiantes se generan mapas superpuestos al mapa GSE de la Región Metropolitana para determinar la relación existente entre los clústers generados y el sector socioeconómico en el cual están situados los colegios y en los lugares que residen los estudiantes.

En el primer capítulo se realiza un acercamiento al problema y los objetivos del estudio. Luego, en el segundo, se expone el estado del arte sobre el problema de elección de colegios y se presenta un breve marco teórico conceptual sobre el aprendizaje automático, el clustering y la metodología CRISP-DM. El tercer capítulo se enfoca en la propuesta de solución planteada para el problema en estudio, describiendo la metodología utilizada para el desarrollo del problema. Finalmente, en el cuarto capítulo, se presentan los resultados obtenidos y los análisis correspondientes, para luego presentar las conclusiones del estudio.

Capítulo 1

Definición del Problema

Uno de los principales problemas que surgen al momento de querer generar políticas públicas, es determinar el público objetivo sobre el cual estas se aplicarán, es decir, en qué segmento de establecimientos o estudiantes se enfocarán dichas políticas.

En la actualidad, como se señaló anteriormente, los colegios en Chile se clasifican principalmente por la dependencia administrativa, dejando fuera muchos criterios que podrían ser de gran interés. De manera similar, los alumnos que a ellos asisten no cuentan con una clasificación que describa de manera relevante los grupos y solo se dividen por el tipo de establecimiento al cual asisten o por el nivel de ingreso familiar.

Por esto es necesario identificar los diferentes perfiles de establecimientos y estudiantes presentes en Chile. Para esto se trabajó en conjunto con el Centro de Investigación Avanzada en Educación de la Universidad de Chile, el cual tiene dentro de su misión el dar soporte científico a la discusión y diseño de políticas públicas en el sector educación, para que las políticas nacionales, la gestión educativa local y la docencia en el aula estén basadas en la evidencia que genera la investigación.

El problema al cual se enfrenta el Estado al momento de realizar políticas públicas es conocer a cabalidad las características del conjunto o grupo a las cuales estas van dirigidas, lo cual es difícil de realizar con una clasificación que se basa en un solo atributo. Para realizar esta tarea sería de mayor utilidad contar con perfiles, tanto de establecimientos como de estudiantes, los

cuales mediante diferentes variables puedan definir de mejor manera un grupo determinado.

El objetivo principal es establecer las relaciones existentes entre perfiles de alumnos y grupos de establecimientos para que a partir de esto se puedan generar políticas públicas de acuerdo a la realidad escolar que se vive en Chile. Además, contar con perfiles establecidos directamente de los datos facilita realizar análisis en diferentes estudios complementarios.

Para llevar a cabo esto se utilizará un algoritmo de clusterización sobre establecimientos y matrículas, que en una primera instancia permitirá determinar los atributos más importantes. En segunda instancia se incorporan variables que los relacionan y así ver si estas tienen un impacto positivo o negativo en los grupos generados anteriormente.

La realización de esto cobra gran importancia, debido a que una mala caracterización del público objetivo impide que las políticas públicas tenga los efectos esperados. Además, el no tener un buen perfil de los grupos significa no conocer como estos están formados, lo que dificulta la creación de políticas que tenga un impacto positivo. Es decir, el problema afecta a la creación de las políticas y a las efectos que estas causan.

A la fecha no existen otros estudios que realicen un trabajo similar al planteado en el presente documento, ya que ha diferencia de este buscan establecer a que establecimiento debe asistir cada alumno según diferentes funciones de utilidad.

Capítulo 2

Estado del Arte

A la fecha de realización de este estudio no existen trabajos que presenten como objetivo final la clasificación no supervisada de establecimientos educacionales y los alumnos matriculados en cada uno de ellos. Debido a esto se investigó sobre trabajos anteriores en el ámbito de la educación, en específico la elección de colegios, para así tener una guía de qué variables son relevantes de considerar.

En el 2002 Sapelli y Torche [12] estudiaron los diferentes determinantes que inciden en la elección del tipo de colegio que realizan los padres al momento de matricular a sus hijos en un determinado establecimiento educacional. En este suponen que una mayor educación de los hijos proveerá a los padres una probabilidad mayor de apoyo cuando estén en la vejez, por lo cual utilizan un modelo en donde se postula una función de utilidad para los padres. Dicha función depende del capital humano inicial de los hijos, el cual puede incrementarse con la educación, y del nivel de consumo presente. Para esto utilizaron diferentes fuentes de información, siendo las más relevantes la encuesta de caracterización socioeconómica (CASEN) de 1996 [4] y los resultados del SIMCE, en donde solo consideraron los datos referentes a la elección de establecimientos de enseñanza básica (niños entre 7 y 14 años), en donde el nivel de cobertura es cercano al 100 %¹ y se descarta la opción de no elegir un colegio. Los factores más determinantes son el nivel de ingreso, la educación de los padres, la recepción de subsidios y la calidad del colegio. Los subsidios por colegios y no por alumno,

¹98,2 % de cobertura educacional para el nivel de enseñanza básica en el año 1996. Fuente: CASEN 1996.

es decir no son portables, dificultan el acceso a los colegios con menores subsidios para las familias de menores recursos. Otro punto importante que destacan es la alta sensibilidad que los padres demuestran respecto a la calidad de los colegios, aún sin conocer los resultados SIMCE, actúan de tal forma que hace pensar que los conocieran.

En el año 2009 Gallego y Hernando [5] buscando resolver la interrogante de cómo los padres escogen el colegio para sus hijos mejorando un modelo desarrollado anteriormente. Para el estudio se consideraron diferentes variables, las cuales se pueden agrupar en dos grandes categorías: características del alumno y características del colegio. En ambos casos los datos son obtenidos del SIMCE del 2012 o calculados por los autores a partir de dichos datos para un universo de 70.000 alumnos de cuarto básico que asisten a 1.200 colegios. Las dos variables que afectan más la elección de un colegio, son el resultado del establecimiento en las pruebas y la distancia entre el hogar y el colegio, en donde la primera variable se repite respecto al estudio [12].

Dos años después, en el 2011, Daniel Gómez en conjunto con R. Chumacero y R. Paredes [3] realizan un estudio similar a los ya presentados, en donde consideran diversos factores que consideran los padres al escoger un determinado colegio. Dichos factores se pueden clasificar en características particulares de cada niño, las propias de cada establecimiento y las que asocian al niño con la escuela, como la distancia entre el hogar y el colegio. Para llevar a cabo esto establecieron una función similar a la presentada en [12], donde se mide la utilidad de que un niño asista a un determinado colegio y que depende de los tres grupos de factores mencionados. Al igual que en trabajos anteriores fueron considerados datos de la encuesta CASEN y del SIMCE, ambos correspondientes al año 2003. Mediante los estudios realizados llegaron a la conclusión de que de los factores analizados, la localización, el precio, la calidad y la potencial competencia de los establecimientos son determinantes al momento de realizar la elección, pero los más valorados por los padres son la calidad y la distancia.

Al año siguiente Gómez, Chumacero y Paredes [6], buscan determinar si el conocimiento de resultados de pruebas específicas (SIMCE) determina de manera importante la selección que realizan los padres sobre el colegio donde matricular a sus hijos. Para esto realizaron un estudio comparativo, tomando como base el estudio anterior y comparándolo con datos de 1996 (primer año donde se hicieron públicos los resultados del SIMCE, por lo cual no influyen en

la elección de colegios de ese año). Del estudio se obtuvo que aún sin conocer los resultados los padres actúan como si los conocieran escogiendo escuelas de mayor calidad, tal como se obtuvo en [12]. Además, cuando los resultados de las pruebas se hicieron públicos, este pasó a ser un factor aún más determinante al momento de tomar una decisión.

Finalmente, uno de los trabajos más recientes en torno a la selección de colegios fue realizado por Canales, Bellei y Orellana [1], donde a diferencia de los trabajos anteriormente señalados, este se enfoca en un sector social específico para determinar y comprender el sentido que tiene para los padres de clase media el elegir un colegio privado. Para este estudio utilizaron dos técnicas complementarias: grupo de discusión y entrevista focalizada, donde la primera apunta a conocer cuál es el valor o significado colectivo de la decisión y la segunda permite conocer como el sujeto entiende la decisión que esta tomando. Los resultados obtenidos son de un carácter preocupante, ya que la selección de colegios esta guiada por el interés del sector medio de distanciarse y diferenciarse de los más pobres, siendo esto una decisión netamente clasista. Además esta preocupa del lado de la educación, debido a que al parecer ni familias ni escuelas parecen orientadas a mejorar el nivel de educación.

Sumado a los trabajos anteriormente señalados es importante conocer algunos conceptos que serán clave para el estudio.

2.1. Machine Learning

El *Machine Learning* [7] o aprendizaje automático es una subrama de la inteligencia artificial que permite a un sistema aprender a través de los datos y no mediante una programación explícita. Este usa diferentes algoritmos que de manera iterativa aprenden de los datos, logrando mejorar la descripción de los datos y la predicción de resultados. Dichos algoritmos operan mediante la construcción de un modelo basado en conjuntos de datos de entrenamiento, que al ir iterando va consiguiendo modelos más precisos. Finalmente se obtiene un modelo de aprendizaje automático que es la salida generada al entrenar un algoritmo de aprendizaje con datos.

En otras palabras es una ciencia que permite el estudio del comportamiento o patrones presentes en diversos tipos de datos, para poder automatizar diversos procesos. Además esto implica un aprendizaje continuo que va mejorando con cada iteración haciéndolo más inteligente y capaz de resolver diferentes problemas en base a lo que va aprendiendo.

Su utilización en diversos ámbitos tiene variadas ventajas, esta permite mejorar la gestión organizacional, facilitar la toma de diferentes decisiones, automatizar y acelerar procesos, entre muchas otras. Pero como es de esperarse también presenta desventajas, las cuales vienen muy de la mano con las decisiones humanas, ya que estas decisiones afectan la resolución que toma el algoritmo en las tareas que se le asignan. Una mala decisión humana puede influir en malos resultados del algoritmo y un mal desempeño en las tareas asignadas.

Este aprendizaje se divide en dos: supervisado y no supervisado.

El aprendizaje supervisado consiste en intentar deducir a partir de datos de entrenamiento o ejemplo una función que clasifique datos sin una clasificación previa. En este caso los datos están compuestos por dos partes, por un lado están los diferentes atributos, ya sean numéricos o categóricos, y por otro lado una etiqueta que clasifica el dato. Entonces lo que se hace es determinar mediante los diferentes atributos el valor de la etiqueta, para luego poder predecir la clasificación de datos no categorizados. Un ejemplo de algoritmo de aprendizaje no supervisado es el de *K* vecinos mas cercanos.

El aprendizaje no supervisado, a diferencia del supervisado no posee un conocimiento a priori de una clasificación de los datos. Por lo tanto un modelo se ajusta al número de observaciones que contiene un data set. En este tipo de aprendizaje solo se cuenta con diferentes atributos, sobre los cuales se buscan semejanzas para poder clasificarlos y crear agrupaciones o clústers. Algunos ejemplos de algoritmos de aprendizaje no supervisado son K-Means y la extensión propuesta en [11], X-Means.

2.2. Clustering

El agrupamiento o *clustering* [8] es un procedimiento que consiste en buscar grupos en un conjunto de datos, estos se denominan clústers y encontrarlos es el objetivo principal del análisis de clústers. Lo que se busca con esto es básicamente formar grupos que contengan objetos similares entre sí y que sean lo más diferente posible con los objetos de otros grupos.

Esta tarea es una actividad importante del proceso de aprendizaje del ser humano, la cual comienza desde muy pequeños. Tareas como la distinción entre perros y gatos u hombres y mujeres, entre muchos otros, son claros ejemplos de esto. La clasificación siempre ha jugado un papel muy importante dentro de la ciencia y es por esto que existen innumerables aplicaciones, tales como la clasificación de compuesto en química, la identificación de diversas enfermedades en la medicina, la clasificación de estrellas en la astronomía, etc.

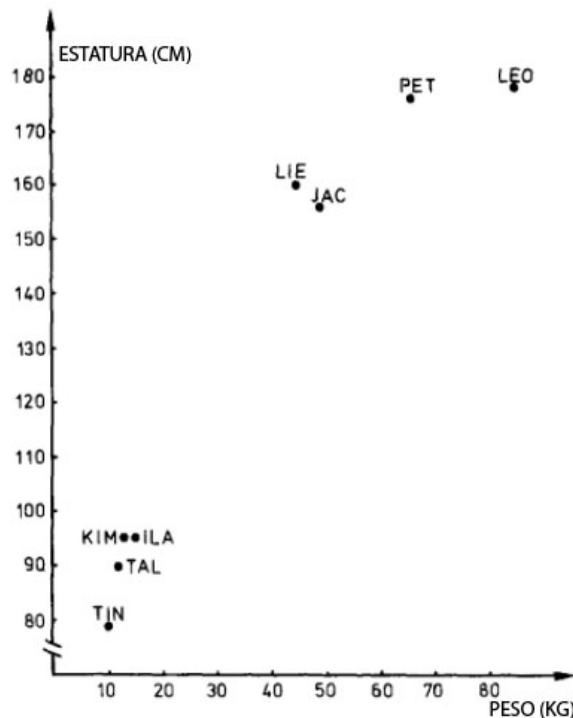


Figura 2.1: Gráfico de 8 objetos. [8]

En el pasado los agrupamientos se realizaban generalmente de manera subjetiva, basándose en el juicio y la percepción del observador. Esto es posible de realizar para casos simples

con 2 o 3 dimensiones, como el de la figura 2.1, donde se pueden reconocer a simple vista 2 grupos {ILA, KIM, TIN, TAL} y {JAC, LEO, LIE, PET}. Sin embargo, en la vida cotidiana es necesario clasificar objetos con una mayor dimensionalidad y realizar este procedimiento de manera más objetiva. Es por esto que se ha desarrollado el procedimiento automático de clasificación.

Con los años se han desarrollado múltiples algoritmos de clasificación, debido a que no existe una definición general de clúster y de hecho existen muchos tipos: clústers esféricos, clústers lineales, entre otros. Además, cada una de las distintas aplicaciones utilizan diferentes tipos de datos como variables discretas, variables continuas, similitudes y diferencias.

A continuación se presentan los algoritmos de *clustering* utilizados.

2.2.1. K-Means

Este es un método de agrupamiento que tiene por finalidad clasificar un grupo de datos en k grupos, basándose en la distancia que presenta cada dato respecto a un centroide. Este algoritmo fue propuesto por Lloyd en 1957, pero no fue publicado hasta el año 1982 [9].

Para realizar el agrupamiento el algoritmo realiza iteraciones, en donde lleva a cabo un seguimiento de los centroides de los grupos. Antes de la primera iteración se inicializan los centroides con valores aleatorios. En cada iteración se realizan las siguientes tareas. Primero, para cada punto de los datos se encuentra el centroide más cercano y se asocia a este. Despues se reestiman las ubicaciones de los centroides con el centro de masa de sus puntos asociados. Luego se repite el mismo proceso de asignar cada punto al centroide más cercano y reestimar la ubicación de los centroides hasta que estos permanezcan fijos en una iteración o se alcance un número determinado de iteraciones.

En el algoritmo 1 se muestra el pseudocódigo del algoritmo de agrupamiento descrito.

Las principales ventajas que presenta este algoritmo son: facilidad para ser implementado, simple y general. Por otro lado, sus desventajas son: se debe especificar un k , muy sensible a la inicialización de los centroides y su baja escalabilidad. Una mala elección del número de

Algoritmo 1 K-Means [10]

Require: $K \geq 1$

```
1: for all  $i \in [1..K]$  do
2:    $c_i \leftarrow$  ObtenerPuntoAleatorio()
3: repeat
4:   Set all  $C_i$  equal to  $\emptyset$ 
5:   for all item  $\in S$  do
6:      $j \leftarrow \operatorname{argmin}_i(\|c_i - \text{item}\|)$ 
7:      $C_j = C_j \cup \{\text{item}\}$ 
8:   for all  $i \in [1..K]$  do
9:     old\_ $c_i = c_i$ 
10:     $c_i = \frac{1}{|C_i|} \sum_{j \in C_i} j$ 
11: until  $\forall_i, \text{old\_}c_i = c_i$ 
12: return all  $C_i$ 
```

clústers puede repercutir en la obtención de malos resultados, al igual que la inicialización aleatoria de los centroides. Además, mientras más grande es el número de clústers, mayor es la probabilidad de incurrir en mínimos locales.

2.2.2. X-Means

X-Means es un algoritmo de agrupamiento, extendido de K-Means, propuesto por Pelleg y Moore [11] el cual busca mejorar la baja escalabilidad computacional, uso de un número determinado de clústers y sensibilidad a mínimos locales.

El principal problema que viene a solucionar este método es el de ingresar con anticipación el número deseado de clústers. A diferencia de K-Means, X-Means recibe un límite inferior y uno superior, dentro del cual el algoritmo es capaz de determinar cual es el número de centroides correcto basándose en una heurística.

El algoritmo 2 generado por Montresor y Guerrieri [10] muestra el funcionamiento de X-Means, el cual está basado en un K-Means reiterativo con $K = 2$. Lo que realiza este método es dividir en dos el data set inicial, para luego ir dividiendo en dos cada clúster que se va generando y detenerse cuando el número de clústers es mayor al límite superior.

En palabras sencillas el algoritmo realiza las siguientes operaciones:

1. Ejecuta K-Means ($K = 2$) en el conjunto completo de datos, tomando dos centroides a partir de un vector aleatorio que pasa por el centro de masa del conjunto original y a una distancia proporcional al tamaño de la región total.
2. Si los clústers "hijos" tienen un desempeño mejor según el criterio de información bayesiano (BIC) que el clúster original, estos se conservan y lo reemplazan.
3. Si no existe una mejor representación del clúster original escoge una fracción constante de los clústers y los reemplaza por sus dos "hijos".
4. El algoritmo se detiene cuando el número de clústers es mayor al límite superior entregado al algoritmo.

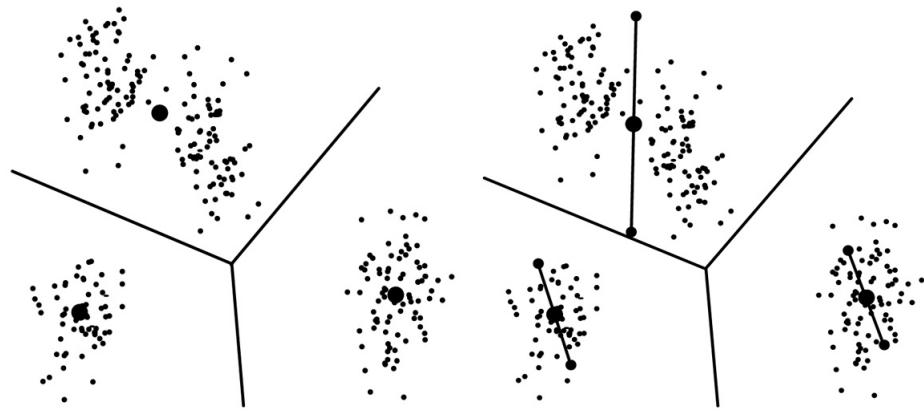
Lo anteriormente descrito se puede apreciar de forma gráfica en la figura 2.2, donde se encuentra una representación para una iteración del algoritmo, con un conjunto inicial de 3 clústers.

Algoritmo 2 X-Means (simplificado) [10]

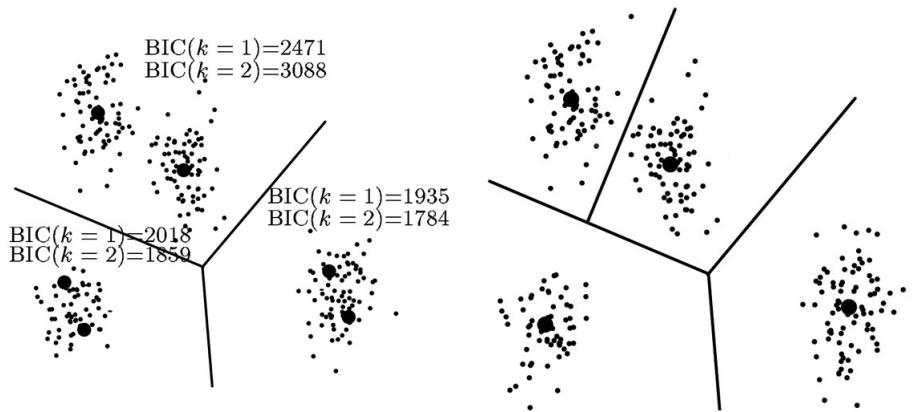
Require: Set de datos S, número máximo de cluster MAX

Require: Función 2Means(S) retorna 2 clústers

```
1: Clustering ← 2Means(S)
2: Mejor_Puntuacion ←  $-\infty$ 
3: while |Clustering| < MAX do
4:     Nuevo_Clustering ← {}
5:     for all Cl ∈ Clustering do
6:         Cl2 ← 2Means(Cl)
7:         if Medida(Cl) > Medida(Cl2) then
8:             Nuevo_Clustering ← Nuevo_Clustering ∪ {Cl}
9:         else
10:            Nuevo_Clustering ← Nuevo_Clustering ∪ Cl2
11:        Clustering ← Nuevo_Clustering
12:        if Measure(Clustering) > Mejor_Puntuacion then
13:            Mejor_Puntuacion ← Medida(Clustering)
14:            Mejor_Clustering ← Clustering
15: return Mejor_Clustering
```



(a) El resultado de ejecutar K-Means con 3 centroides.
(b) Cada centroide original se divide en dos hijos.



(c) El resultado después de que todos los 2- Means paralelos hayan terminado.
(d) Los centroides sobrevivientes después de todas las pruebas de puntuación.

Figura 2.2: Iteración X-Means.

En la figura 2.2a se muestra una solución de K-Means con 3 centroides y los límites de cada región. En la 2.2b se dividen los centroides en dos nodos hijos, trazando un vector aleatorio a una distancia proporcional al tamaño de la región. Luego de esto se ejecuta un K-Means con $k = 2$ de manera local, esto debido a que los hijos solo se distribuyen los puntos de los padres y no se relacionan con el resto. El resultado al finalizar los K-Means se muestra en la figura 2.2c, donde cada hijo tiene sus puntos asociados.

Una vez finalizado lo anterior es momento de evaluar si el padre o los hijos representan mejor

la distribución de cada región. Según este criterio se eliminará el padre o su descendencia. Si la representación original de una región es la adecuada, entonces sobrevivirá el padre y se matará a sus hijos. Por otro lado, las regiones que no estén bien representadas por los centroides recibirán mayor atención, aumentando en ellas el número de centroides. En la figura 2.2d se muestra el resultado final al llevar a cabo la prueba de puntuación (BIC) en los 3 pares de hijos.

Una de las principales ventajas que posee este algoritmo radica en que es más escalable debido a que con cada iteración se reduce más el número de datos en los cuales K-Means se debe ejecutar, lo que hace que sea más fácil utilizarlo en conjuntos de datos de mayor tamaño. Otra ventaja es que se sabe que con K-Means de pocos clústers es menos probable incurrir en mínimos locales en comparación a uno realizado con muchos clústers. Por lo tanto, el hecho de que X-Means utilice un K-Means con $K = 2$ favorece a que no se atasque en mínimos locales.

Además este método permite realizar una visualización del tipo árbol, la que crea una estructura jerárquica de los clústers.

El porqué se utilizará este algoritmo es debido a su sencillez y facilidad de implementación y ejecución. Además, otro punto importante por el cual se optó por este algoritmo es porque el problema que se busca resolver cuenta con una gran cantidad de datos, lo que no es una dificultad debido a la escalabilidad que este posee gracias a que con cada iteración el conjunto de puntos sobre el cual se aplica K-Means ($k = 2$) se ve reducido. Por último, pero no menos importante, se escogió este algoritmo debido a que no necesita conocer de antemano un número de clústers, sino que es calculado automáticamente.

2.3. CRISP-DM

CRISP-DM [14] (*Cross Industry Standard Process for Data Mining*) es una de las metodologías que más se utiliza para el desarrollo de proyectos de minería de datos. Esta proporciona una visión general del ciclo de vida de un proyecto de minería de datos. Además contiene las fases de un proyecto, sus respectivas tareas y sus resultados.

Este ciclo está compuesto por 6 fases que se aprecian en la figura 2.3. Dentro del diagrama las flechas solo indican las dependencias más importantes y frecuentes entre las fases, en donde la secuencia no es estricta. En la figura 2.3 el círculo exterior representa la naturaleza cíclica de la minería de datos. Este proceso no termina una vez se implemente la solución, debido a que todo lo aprendido durante el proyecto puede desencadenar nuevas preguntas de negocio, en donde dichos procesos se beneficiarán de experiencias anteriores.

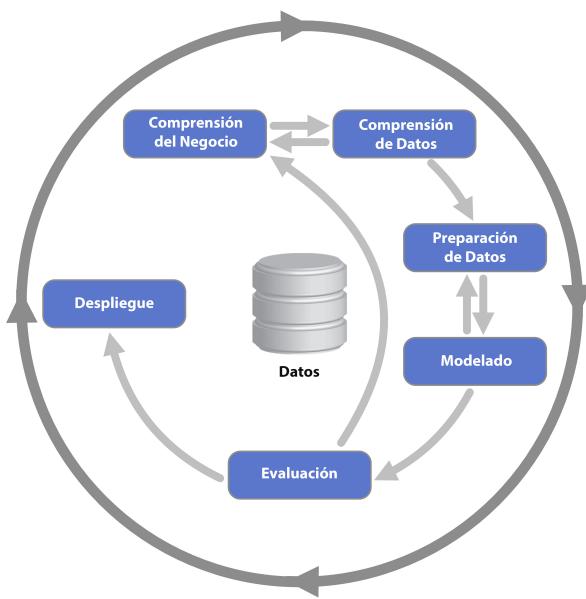


Figura 2.3: Diagrama de proceso que muestra la relación entre las diferentes fases de CRISP-DM.

1. **Comprensión del negocio:** Esta fase se centra en comprender los objetivos y requisitos del proyecto desde una perspectiva comercial, para luego convertirlo en una definición del problema de minería de datos. Además, crear un plan preliminar diseñado para alcanzar los objetivos.
2. **Comprensión de datos:** Esta fase comienza con una recopilación inicial de datos y actividades para familiarizarse con los datos, identificar problemas de calidad de datos, descubrir primeras ideas sobre los datos o detectar subconjuntos interesantes para formar una hipótesis de información oculta.
3. **Preparación de datos:** Esta cubre todas las actividades para construir el conjunto de datos final a partir de los datos brutos iniciales. Dichas tareas incluyen la selección

y transformación de tablas, registros y atributos, construcción de nuevos atributos y limpieza de datos para las herramientas de modelado. Las tareas de preparación de datos puede que se realicen varias veces y sin un orden específico.

4. Modelado: Fase de selección y aplicación de diversas técnicas de modelado, además de la calibración de sus parámetros para obtener óptimos resultados. Existen varias técnicas para el mismo tipo de problemas de minería de datos. Algunas de ellas requieren formatos de datos específicos.
5. Evaluación: En esta etapa se ha creado uno o varios modelos que parecen ser de alta calidad. Antes de realizar el despliegue final es importante evaluar más a fondo el modelo y revisar los pasos ejecutados para construirlo.
6. Despliegue: En fase se deberá definir como implementar los resultados. Dicha implementación dependerá de los requisitos, esta puede ser tan simple como generar un informe o tan compleja como implementar un proceso repetible de minería de datos.

Capítulo 3

Propuesta de Solución

Este trabajo presenta un estudio tanto de establecimientos educacionales como de sus alumnos, para poder determinar la estructura que estos presentan y poder generar un modelo de selección de colegios basado en los grupos que se generen a partir de los conjuntos de datos iniciales.

3.1. Metodología de trabajo

Con la finalidad de llevar a cabo los objetivos planteados para este estudio se optó por utilizar una metodología propia basada en la ya existe CRSIP-DM.

El trabajo se desarrolla de la siguiente forma:

- **Comprendión de datos:** esta etapa consiste básicamente en la colección de datos desde diferentes fuentes. En este caso fueron utilizadas las bases de datos de establecimientos y matrículas de la Región Metropolitana facilitadas por el CIAE. Además se complementó la base de datos de los colegios con la información pública disponible en la página Web del Ministerio de Educación de Chile (MIME [2]), los cuales fueron obtenidos mediante la técnica de *Web scraping*.

- **Preparación de datos:** en esta fase, a partir de los datos obtenidos anteriormente, se seleccionaron los atributos más relevantes para el estudio y se realizó un proceso de limpieza y estandarización de los datos. Además, se agregaron algunos atributos derivados o calculados de los datos extraídos originalmente.

Para cada atributo se calculó su nivel de estabilidad, la cual mide cuán estable o constante es el atributo. La estabilidad se calcula mediante la división del número de filas no nulas del valor más frecuente y el número total de filas de datos no nulos.

Lo anteriormente descrito se resume en las tablas 3.1 y 3.2, las cuales presentan los atributos correspondientes a los establecimientos y matrículas, respectivamente, ordenados de menor a mayor según su estabilidad.

Tabla 3.1: Atributos que componen la base de datos de los establecimientos.

Atributo	Descripción	Estabilidad	Fuente
mat_total	Matrícula total de alumnos.	0,44 %	MIME
prom_alu.cur	Promedio de alumnos por curso.	5,22 %	MIME
nivel_deportivo	Nivel deportivo del establecimiento.	48,36 %	MIME
nivel_ensenanza	Nivel de enseñanza que ofrece el establecimiento.	48,4 %	Derivado de datos CIAE.
cod_depe	Código de dependencia del establecimiento.	53,82 %	CIAE
enf_valorico	Enfoque valórico.	54,79 %	MIME
pago_men	Nivel de pago de mensualidad.	60,64 %	MIME
req_entrevista	Requisitos de entrevista para postular.	62,81 %	MIME
becas_disp	Becas disponibles en el establecimiento.	64,07 %	MIME
convenio_sep	Posee convenio de subvención escolar preferencial (SEP).	71,91 %	MIME
pago_mat	Nivel de pago de matrícula.	77,66 %	MIME
enf_academico	Enfoque académico.	84,62 %	MIME
area_metropolitana_rbd	Pertenencia al área metropolitana.	85,74 %	CIAE
continúa ...			

Tabla 3.1: Atributos que componen la base de datos de los establecimientos. (continuación)

apoyo_tutorias	Ofrece ayuda a los alumnos mediante tutorías.	86,65 %	MIME
req_papeles	Requisitos de papeles para postular.	87,57 %	MIME
req_pago	Requisitos de pago para postular.	89,12 %	MIME
apoyo_especialistas	Ofrece ayuda a los alumnos mediante especialistas.	90,18 %	MIME
gen_rbd	Género del establecimiento.	94,54 %	CIAE
req_otros	Requisitos de cualquier tipo que no clasifique en las otras categorías.	95,02 %	MIME
enf_laboral	Enfoque laboral.	95,21 %	MIME
req_prueba	Requisitos de prueba para postular.	95,94 %	MIME
enf_otros	Enfoque de otro tipo que no clasifique en las otras categorías.	96,32 %	MIME
apoyo_otros	Ofrece ayuda a los alumnos de cualquier otra forma que no clasifique en las otras categorías.	98,79 %	MIME

Tabla 3.2: Atributos que componen la base de datos de las matrículas.

Atributo	Descripción	Estabilidad	Fuente
IPAE	Indicador de barrio crítico.	0,6 %	CIAE
edad_alu	Edad del alumno.	8,79	CIAE
gen_alu	Género del alumno.	9,04 %	CIAE
criterio_sep	Criterio por el cual se considera prioritario.	9,23 %	CIAE
beneficiario_sep	Indicador del alumno beneficiario de la SEP.	50,73 %	CIAE
gse_decil_nac	Grupo Socio Económico del alumno.	64,04 %	CIAE
continúa ...			

Tabla 3.2: Atributos que componen la base de datos de las matrículas. (continuación)

area_metropolitana_alu	Pertenencia al área metropolitana.	90,48 %	CIAE
cod_sec	Código del sector económico.	94,81 %	CIAE
cod_espe	Código de especialidad.	94,81 %	CIAE
cod_rama	Código de rama.	94,81 %	CIAE
grado_sep	Corresponde a un nivel SEP.	100 %	CIAE

Además de los atributos propios de cada establecimiento y matrículas se añadieron variables que relacionan a ambas entidades, las variables de relación establecimiento-matrícula. En el caso de los establecimientos se agregó la sobre edad promedio, el GSE que predomina en el establecimiento, la desviación de la moda GSE, los porcentajes de hombres y mujeres, la distancia máxima a la que viven el 75 % de los alumnos de cada colegio y el índice de desarrollo de la educación (IDE) por rangos. Por otro lado, para las matrículas se añadió la distancia a la que vive el alumno del colegio, el nivel de sobre edad¹, la dependencia del colegio y el nivel de copago (matrícula y mensualidad). Estos se pueden ver en las tablas 3.3 y 3.4, respectivamente.

Tabla 3.3: Atributos de relación establecimiento-matrícula para la base de datos de los establecimientos.

Atributo	Descripción	Fuente
sobre_edad_prom	Sobre edad promedio de los alumnos que estudian en el establecimiento.	Calculado con datos CIAE.
desviacion_moda	Desviación que presenta el grupo GSE con respecto a la moda.	Calculado con datos CIAE.
porc_femenino	Porcentaje de mujeres del establecimiento.	Calculado con datos CIAE.
porc_masculino	Porcentaje de hombres del establecimiento.	Calculado con datos CIAE.
dist_percentil_75	Distancia del percentil 75 de los alumnos que asisten al establecimiento.	Calculado con datos CIAE.
continúa ...		

¹Diferencia entre la edad actual y la edad esperada para el curso en el que se encuentra.

Tabla 3.3: Atributos de relación establecimiento-matrícula para la base de datos de los establecimientos. (continuación)

IDE_rango	Índice de Desarrollo de la Educación para Todos por rango.	Calculado con datos CIAE.
gse_moda_nac	Grupo Socio Económico que más se repite en el establecimiento.	Calculado con datos CIAE

Tabla 3.4: Atributos de relación establecimiento-matrícula para la base de datos de las matrículas.

Atributo	Descripción	Fuente
dist_actual	Distancia del alumno a su establecimiento actual.	CIAE
pago_men	Nivel de pago de mensualidad.	MIME
cod_depe	Dependencia del colegio al que asiste.	CIAE
sobre_edad	Diferencia entre la edad actual y la esperada para el nivel.	Calculado con datos CIAE.
pago_mat	Nivel de pago de matrícula.	MIME

Adicionalmente se imputaron los campos de datos faltantes mediante el algoritmo MICE (*Multiple Imputation by Chained Equations*) para eliminar los valores nulos.

- **Modelado:** en esta etapa se escoge el algoritmo de agrupamiento X-Means, debido a que uno de los objetivos del estudio es poder encontrar sin una idea previa la estructura que poseen los establecimientos y los alumnos, el cual es ejecutado sobre el software RapidMiner. En el programa se transforman las variables categóricas a numéricas y se realiza una normalización. Además se ajustan los diferentes parámetros del algoritmo, incluyendo el número de variables que se utiliza (según la estabilidad de las variables propias).
- **Análisis y resultados:** se realiza un análisis de los resultados obtenidos de la fase de modelado, los cuales deben ser capaces de satisfacer los objetivos planteados para el estudio.

- **Conclusiones:** finalmente se deben realizar las conclusiones respectivas del trabajo efectuado y validar el cumplimiento de los objetivos que fueron planteados para el trabajo.

Capítulo 4

Resultados y Análisis

Una vez realizados los experimentos con el algoritmo de agrupamiento X-Means, se presentan los resultados en diferentes figuras y tablas. Primero se encuentran los resultados conseguidos al ejecutar el algoritmo con los atributos de las tablas 3.1 y 3.2. Luego se presentan los resultados al comparar el mejor resultado de la etapa anterior con el resultado que se obtiene al agregarle las variables de relación establecimiento-matrícula.

4.1. Resultados obtenidos

La tabla 4.1 muestra los resultados al ejecutar X-Means sobre la base de datos de establecimientos en 3 diferentes ocasiones, diferenciándose por la cantidad de atributos que se utilizan. Una con todos los atributos, una con los de importancia alta y media, y finalmente una solo con los atributos de importancia alta.

Las siglas de los clústers de establecimientos sin considerar los atributos de relación establecimientos-matrícula (tablas 4.1, 4.2 y 4.3) corresponden a:

- BCBM: Bajo Costo Baja Matricula (menor a 390 alumnos).
- BCAM: Bajo Costo Alta Matricula (sobre 700 alumnos).

- MCAM: Medio Costo Alta Matrícula (sobre 700 alumnos).
- ACMM: Alto Costo Media Matrícula (entre 391 y 700 alumnos).

Tabla 4.1: Clústers de establecimientos variando la cantidad de atributos (por grupos de importancia).

Atributos	BCBM	BCAM	MCAM	ACMM
Todos	959	826	247	36
Alta + Media	959	540	311	258
Alta	977	524	308	259

La tabla 4.2 muestra los resultados obtenidos al ejecutar X-Means en los atributos de la base de datos de establecimientos (tabla 3.1) con importancia alta.

Tabla 4.2: Comparativa de clústers de establecimientos por atributo.

	BCBM	BCAM	MCAM	ACMM
Dependencia	P.Subvencionado / Municipal	P. Subvencionado / Municipal / Corp. Adm. Deleagada	P. Subvencionado	P. Pagado
Educación	Básica / Media	Media / Completa	Media / Completa	Completa
Matrícula	Gratuita / Menor a \$25.000	Gratuita / Menor a \$25.000	Menor a \$10.000	Mayor a \$50.000
Mensualidad	Gratuita / Menor a \$25.000	Gratuita / Menor a \$25.000	25,000–100.000	Mayor a \$50.000
Convenio SEP	Si	Si	No	No
Prom. matrículas	383	781	905	686
Prom. alumnos por curso	26	30	32	21
Prom. de becas	19	59	91	8

Tabla 4.3: Clústers de establecimientos (tabla 4.2) según su dependencia.

Clúster	Municipal	P. Subvencionado	P. Pagado	Corp. Admin. Del.
BCBM	485	491	1	0
BCAM	173	318	0	33
MCAM	3	303	2	0
ACMM	0	1	258	0

En la tabla 4.4 se encuentran los resultados que se obtuvieron al ejecutar el algoritmo de agrupamiento en la base de datos conformada por los atributos de importancia alta de la tabla 3.1 y los atributos de relación establecimiento-matrícula de la tabla 3.3.

Las siglas utilizadas para los clústers de establecimientos considerando los atributos de relación establecimientos-matrícula (tablas 4.4, 4.5 y 4.6) corresponden a:

- BCBMCD: Bajo Costo Baja Matricula (menor a 390 alumnos) Corta Distancia (menor a 4 km).
- BCAMMD: Bajo Costo Alta Matricula (sobre 700 alumnos) Media Distancia (entre 4 y 8 km).
- MCAMMD: Medio Costo Alta Matricula (sobre 700 alumnos) Media Distancia (entre 4 y 8 km).
- ACMMLD: Alto Costo Media Matrícula (entre 391 y 700 alumnos) Larga Distancia (sobre 8 km).

Tabla 4.4: Comparativa de clústers de establecimientos por atributo, incluyendo los de relación establecimiento-matrícula.

	BCBMPD	BCAMMD	MCAMMD	ACMMLD
Dependencia	P. Subvencionado / Municipal	P. Subvencionados / Municipales / Corp. Adm. Delegada	P. Subvencionado	P. Pagado
Educación	Básica	Media / Completa	Media / Completa	Completa
Matrícula	Gratuita	Gratuita / Menor a \$25.000	Menor a \$10.000	Mayor a \$50.000
Mensualidad	Gratuita / Menor a \$25.000	Gratuita / Menor a \$25.000	25,000–100.000	Mayor a \$50.000
Convenio SEP	Si	Si	No	No
Prom. matrículas	383	781	895	686
Prom. alumnos por curso	26	30	32	21
Prom. de becas	19	58	91	8
IDE	0,5 - 1	-0,5 - 0	0 - 1,5	1 - 1,5
Distancia				
Establecimiento - Hogar	2,960 km	5,204 km	5,207 km	9,761
Sobre edad promedio	0,524	0,488	0,297	0,488

Tabla 4.5: Clústers de establecimientos (tabla 4.4) según su dependencia.

Clúster	Municipal	P. Subvencionado	P. Pagado	Corp. Admin. Del.
BCBMPD	485	489	1	0
BCAMMD	173	307	0	33
MCAMMD	3	316	2	0
ACMMLD	0	1	258	0

Tabla 4.6: Detalle de la sobre edad en los clústers de establecimientos.

	% del total	% 1 año¹	% 2 años¹¹	% 3 años¹	% 4 años¹
BCBMPD	31	74,8	18,9	5,2	1,1
BCAMMD	32,9	75,1	20,5	3,9	0,5
MCAMMD	24	85,3	13	1,6	0,1
ACMMLD	38,1	94,9	4,6	0,4	0,1

Las siguientes tablas, y de manera similar a lo descrito anteriormente para los establecimientos, muestran los resultados obtenidos usando los datos de las matrículas. En la tabla 4.7 se muestran los resultados con los datos de la tabla 3.2 en 3 diferentes versiones, según la importancia del atributo. La primera con todos los atributos, luego solo con los de alta y media, y finalmente solo los de alta importancia.

Las siglas utilizadas para los clústers de matrículas sin atributos de relación (tablas 4.7, 4.8 y 4.9) corresponden a:

- MCB: Mujeres Con Beneficio.
- HCB: Hombres Con Beneficio.
- MSB: Mujeres Sin Beneficio.
- HSB: Hombres Sin Beneficio.

¹Porcentajes en base al total de alumnos con sobre edad mayor o igual a 1.

Tabla 4.7: Clústers de matrículas variando la cantidad de atributos (por grupos de importancia).

Atributos	MCB	HCB	MSB	HSB
Todos	551932	442048	40803	13585
Alta + Media	551932	442048	34088	20300
Alta	286089	300007	230486	231786

En las tablas 4.8 y 4.10 se muestran primero los resultados obtenidos al utilizar los atributos de alta importancia de la tabla 3.2 y luego los que se obtienen al incluir los atributos de relación de la tabla 3.4.

Tabla 4.8: Comparativa de clústers de matrículas por atributo.

	MCB	HCB	MSB	HSB
Género	Femenino	Masculino	Femenino	Masculino
Beneficiario SEP	Si	Si	No	No
Criterio SEP	Pertenece a Chile solidario / Puntaje de ficha de protección social	Pertenece a Chile solidario / Puntaje de ficha de protección social		
Sobre edad	0,364	0,488	0,284	0,365

Tabla 4.9: Detalle de la sobre edad en los clústers de matrículas.

	% del total	% 1 año ²	% 2 años ²	% 3 años ²	% 4 años ²
MCB	28,5	76,9	18,5	4	0,6
HCB	36,4	72,5	21,5	5,1	0,9
MSB	25,7	90,2	8,7	1	0,1
HSB	32	87,3	11,1	1,5	0,1

Las siglas utilizadas para los clústers de matrículas, incluyendo los atributos de relación, (tablas 4.10 y 4.11) corresponden a:

²Porcentajes en base al total de matrículas con sobre edad mayor o igual a 1.

- MCBBC: Mujeres Con Beneficio Bajo Costo.
- HCBBC: Hombres Con Beneficio Bajo Costo.
- MiSBMC: Mixto Sin Beneficio Medio Costo.
- MiSBAC: Mixto Sin Beneficio Alto Costo.

Tabla 4.10: Comparativa de clústers de matrículas por atributo, incluyendo los de relación establecimiento-matrícula.

	MCBBC	HCBBC	MiSBMC	MiSBAC
Género	Femenino	Masculino	Femenino / Masculino	Femenino / Masculino
Beneficiario SEP	Si	Si	No	No
Criterio SEP	Pertenece a Chile solidario / Puntaje de ficha de protección social	Pertenece a Chile solidario / Puntaje de ficha de protección social		
Sobre edad	0,370	0,492	0,270	0,401
Matrícula que paga	Gratuita	Gratuita	Gratuita	Mayor a \$100.000
Mensualidad que paga	Gratuita	Gratuita	10,000–100.000	Mayor a \$100.000
Distancia Establecimiento - Hogar	4,711 km	4,939 km	5,912 km	4,911 km

Tabla 4.11: Detalle de la sobre edad en los clústers de matrículas, incluyendo atributos de relación.

	% del total	% 1 año³	% 2 años³	% 3 años³	% 4 años³
MCBBC	28,9	76,7	18,7	3,9	0,7
HCBBC	36,7	72,4	21,6	5,1	0,9
MiSBMC	23,4	85,9	12,4	1,6	0,1
MiSBAC	38,1	94,9	4,6	0,4	0,1

³Porcentajes en base al total de matrículas con sobre edad mayor o igual a 1.

4.2. Análisis de los resultados

4.2.1. Análisis cualitativo

En esta sección se analizan los diferentes resultados obtenidos y presentados anteriormente, comenzando con el análisis para los establecimientos y seguido por el de las matrículas.

Lo primero que se aprecia en la tabla 4.1, es que para las 3 versiones el clúster BCBM agrupa casi un 50 % del total de los colegios, en donde además en la segunda y tercera versión el resto de los clústers presentan una cardinalidad similar. Esto demuestra que en este caso incluir variables consideradas de importancia media y baja no generan un gran impacto en el resultado final.

Teniendo en cuenta lo anterior, y de que por tratarse de un aprendizaje no supervisado es difícil establecer una medida de eficiencia, se consideró la tercera versión para el resto del estudio. Es decir, se tomará la versión en la cual solo se utilizaron los atributos clasificados como de alta importancia. Además se debe considerar que al ejecutar el algoritmo con menos variables el tiempo de ejecución es menor.

A partir de los resultados obtenidos para los establecimientos y de las figuras generadas con estos (figuras 4.1 y 4.2), podemos ver que al realizar una comparación entre ambas ejecuciones del algoritmo (sin y con atributos de relación) no se generan grandes cambios en los clústers, pero si aumenta el nivel de información y detalle de cada uno de ellos, incorporando nuevas características distintivas. Es por esto que el análisis se centrará en dichos resultados.

Los primeros dos clústers son colegios gratuitos o baratos con un IDE bajo que se diferencian entre ellos principalmente por el nivel de educación que imparten, en el primero predominan los de enseñanza básica y en el segundo establecimientos que imparten educación media o completa. El tercer y cuarto clúster se diferencia de los otros dos por tener un nivel de copago e IDE superiores, donde en el tercero se tienen valores medios y en el cuarto valores elevados. Por otro lado, al analizar la distancia que separa a los establecimientos de sus estudiantes, se aprecia notoriamente que el desplazamiento crece al aumentar el nivel del copago. Es decir, las familias que más pagan están dispuestas a desplazarse distancias mayores para llegar al

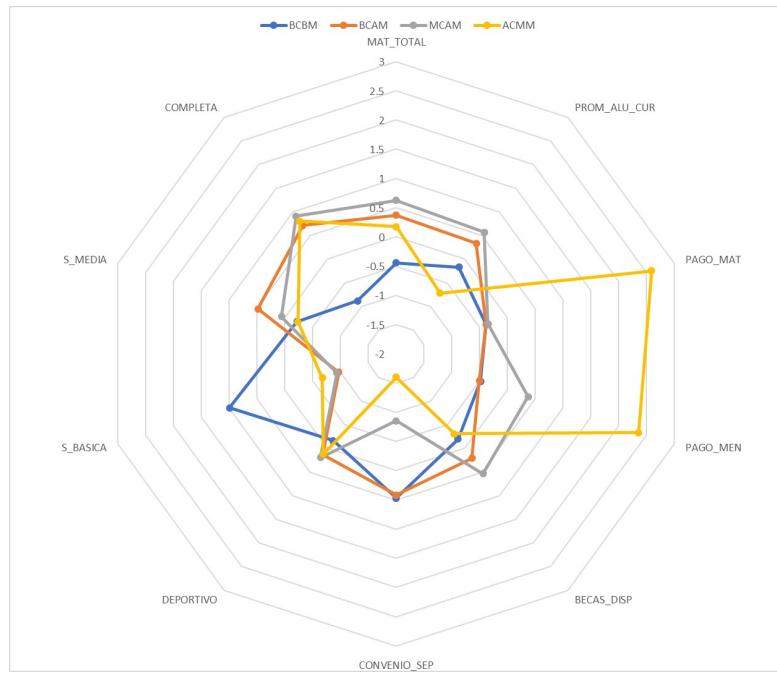


Figura 4.1: Promedios de atributos normalizados para establecimientos de la Región Metropolitana.

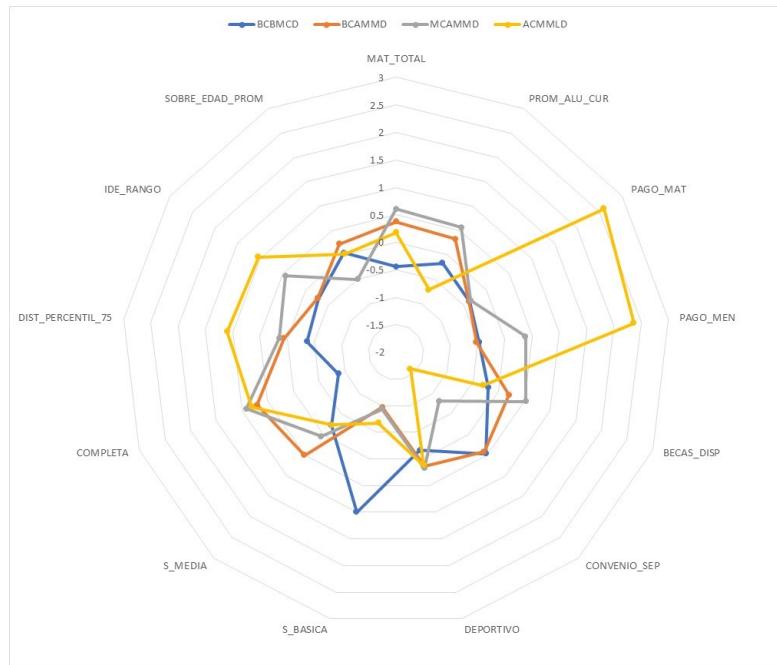


Figura 4.2: Promedios de atributos normalizados de establecimientos con atributos relacionales (tabla 3.3) de la Región Metropolitana.

establecimiento en comparación a familias que optan por colegios gratuitos o de bajo costo. Finalmente otro punto interesante de analizar es la sobre edad, que en el caso de los colegios más caros se concentra con un 95 % en un año de sobre edad. En el resto de los clústers el valor fluctúa entre un 75 % y 85 %, y el resto de distribuye de 2 a 4 años de sobre edad.

Para el caso de las matrículas lo primero que se debe analizar es la tabla 4.7, en donde se aprecian los resultados de las 3 ejecuciones con los diferentes atributos, agrupados por su nivel de importancia. Se puede ver que las primeras dos ejecuciones generan clústers de cardinalidad muy similares, diferenciándose claramente con el tercer resultado, el cual presenta clústers de tamaños similares. La diferencia se en que los clústers de la última ejecución poseen atributos que describen de mejor manera los grupos, los cuales se van perdiendo al ir añadiendo atributos adicionales.

Además de lo ya mencionado, es importante tener en cuenta que agregar más atributos a una base de datos de más de 1.000.000 de registros, provocará que el tiempo requerido para su ejecución aumente significativamente. Por lo tanto, se analizarán más a fondo los resultados de la tercera ejecución y se utilizará para comparar con los resultados que se obtienen al agregar los atributos de relación.

Lo siguiente a analizar son las diferentes características de los clústers obtenidos para las matrículas y compararlos con los que se obtienen cuando se incluyen los atributos de relación. Para facilitar la comparación entre clústers se generaron las figuras 4.3 y 4.4.

En la primera instancia, se aprecian 2 atributos que generan la mayor categorización dentro de los clústers y que permiten etiquetar al grupo. Estos atributos, el género y el ser o no beneficiario SEP, son del tipo binarios y al ser combinados generan los 4 clústers obtenidos. Es decir, al grupo que pertenezca un alumno se basa principalmente en si es hombre o mujer y si tiene o no la subvención escolar preferencial.

Otro aspecto importante que se aprecia en los clústers generados es que en los grupos de mujeres el nivel de sobre edad es menor a su símil masculino. Esto se puede ver con mayor precisión en la tabla 4.9, en donde los porcentajes de matrículas con sobre edad mayor o igual a un año es mayor en los clústers de hombres. Al analizar los porcentajes por años de sobre edad, estos tienen una distribución similar, aunque los porcentajes del género femenino

tienden a ser mayores para 1 año y menores para 2 o más años.



Figura 4.3: Promedio de atributos normalizados de matrículas de la Región Metropolitana.

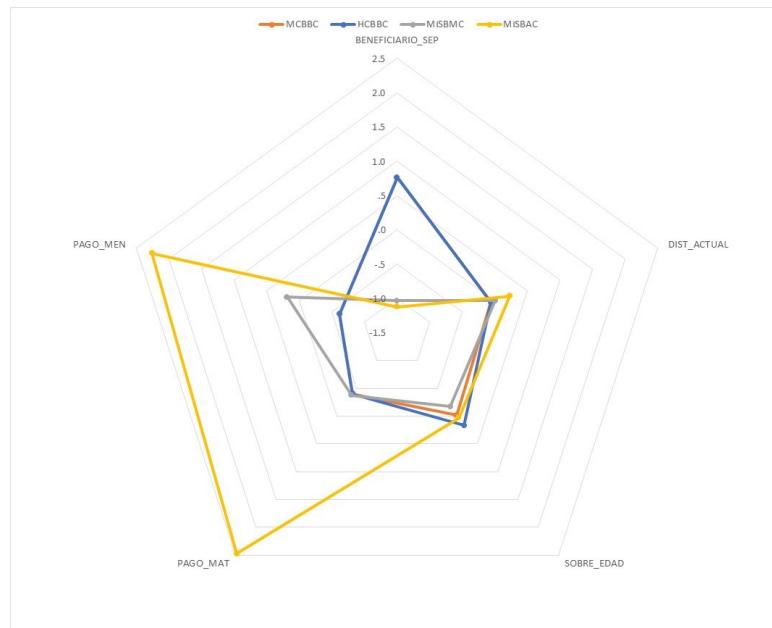


Figura 4.4: Promedio de atributos normalizados de matrículas con atributos relacionales (tabla 3.4) de la Región Metropolitana.

Al momento de incluir los atributos de relación (nivel de copago y distancia al establecimiento), los atributos más relevantes siguen siendo el género y el ser o no beneficiario del SEP, sumándose a estos el nivel de copago. A pesar de esto se debe destacar que la importancia del género no es la misma que en la situación anterior, debido a que en este caso 2 de los 4 clústers no son excluyentes en este atributo, grupos en los que cobra mas importancia el nivel de copago.

Los primeros dos grupos presentan características similares a los de la primera prueba, incorporando como atributo principal el nivel de copago, que para ambos es gratuito. Además para estos grupos se mantiene constante el hecho de que las mujeres presentan un nivel de sobre edad menor al de los varones. Como se mencionó anteriormente, el tercer y cuarto clúster son no excluyentes por género, por lo cual en estos dos pasa a ser más importante el copago. En el tercer clúster el nivel de copago es un costo bajo o medio y en el cuarto es mucho más alto.

En estos clústers la distancia no es un factor tan determinante para clasificar un alumno en uno u otro grupo, debido a que tienen valores muy similares, los cuales tienen una máxima diferencia de 1 kilómetro.

Es difícil realizar una comparación entre los resultados obtenidos sin incluir los atributos de relación establecimiento-matrícula y los que se obtienen al incluirlos, pero se puede apreciar que en ambos casos el algoritmo clasificó todas las matrículas en 4 clústers. También se puede señalar que el género es importante en este estudio, independiente de que en uno de los resultados tenga menor importancia.

Además del análisis individual de los clústers de establecimientos y matrículas, se estudia la relación que tienen entre ellos. En las figuras 4.5 y 4.6 se aprecia como están compuestos los clústers de colegios según los perfiles de sus alumnos, tanto en el caso de los clústers obtenidos sin los atributos de relación y con los atributos.

En la figura 4.5 se muestra como están compuestos los clústers de establecimientos, según los clústers de matrículas sin considerar los atributos de relación. En este se puede apreciar que los clústers se encuentran balanceados en género, es decir, no predomina un género en un tipo de colegio. Por otro lado, en los clústers BCBMCD y BCAMMD predominan los

estudiantes que son beneficiarios por sobre los que no lo son, lo que es totalmente opuesto en los grupos MCAMMD y ACMMLD, donde predominan los alumnos sin beneficio. Según las características de los colegios y de los alumnos, se observa que al aumentar el nivel de copago disminuye la cantidad de alumnos con beneficios y aumenta la cantidad de los que no tienen.

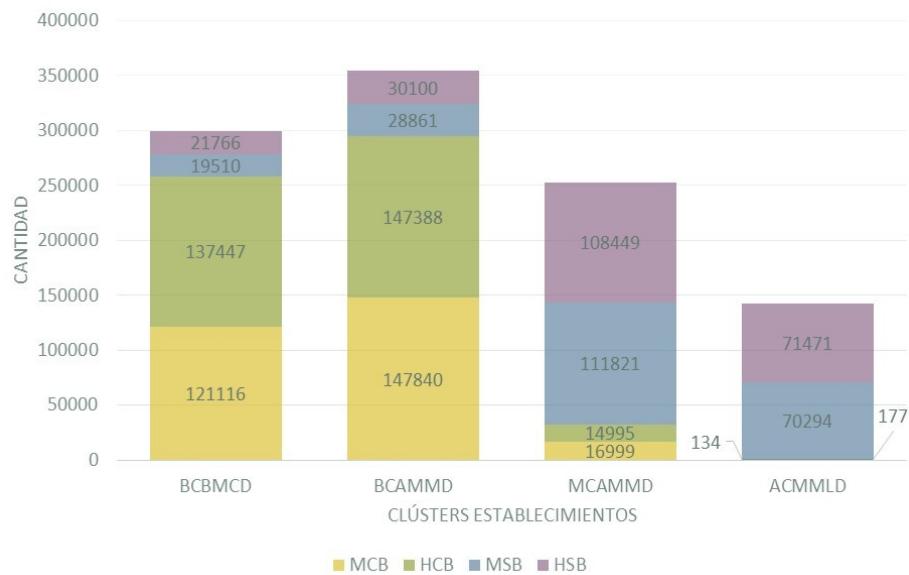


Figura 4.5: Composición de los clústers de establecimientos según los clústers de matrículas (sin variables de relación).

Al momento de incluir los atributos de relación para las matrículas, los grupos que se forman varían, lo que también hace variar la composición de los clústers de colegios. Los primeros dos clústers de establecimientos están formados principalmente por los grupos de alumnos MCBBC y HCBBC, es decir, por alumnos que pagan poco y tienen beneficios. El tercer clúster está compuesto principalmente por el grupo MiSBMC, que son alumnos (hombres y mujeres) que tienen un nivel de copago medio y no poseen beneficio. Por último, el clúster ACMMLD está formado en su mayoría por casi todo el clúster MiSBAC, el cual tiene alumnos de ambos géneros con un nivel de copago elevado. Lo anteriormente descrito se puede apreciar en la figura 4.6.

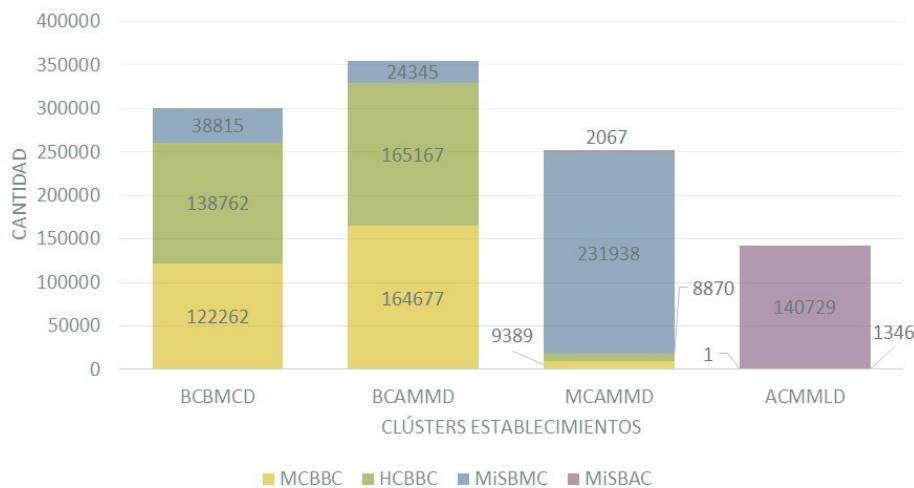


Figura 4.6: Composición de los clústers de establecimientos según los clústers de matrículas (con variables de relación).

Si se comparan ambas figuras podemos notar a simple vista que la composición de los clústers de establecimientos varían al incorporar variables de relación en las matrículas, esto debido a que los grupos en esta segunda clusterización están fuertemente influenciados por el nivel del copago que tienen los alumnos.

4.2.2. Análisis geográfico

Una vez realizado el análisis cualitativo de los clústers, y con los datos geográficos disponibles, se analiza la distribución geográfica y socioeconómica en el área metropolitana de los clústers de establecimientos, de matrículas y de las matrículas en los clústers de establecimientos.

En la figura 4.7 se muestra la distribución de los establecimientos de cada clúster en la Región Metropolitana, incluyendo los grupos socioeconómicos. En el mapa de la figura 4.7a se aprecia que los establecimientos del clúster BCBMPD se encuentran distribuidos de manera uniforme sobre gran parte del área metropolitana, exceptuando la zona oriente. Estos se concentran principalmente en las zonas de color naranja, las cuales corresponden a los GSE de los deciles 3 al 7. Los establecimientos del clúster BCAMMD (figura 4.7b) se distribuyen de

la misma forma descrita anteriormente, con la diferencia de que su cardinalidad es menor, por lo cual los colegios se encuentran a una mayor distancia entre ellos.

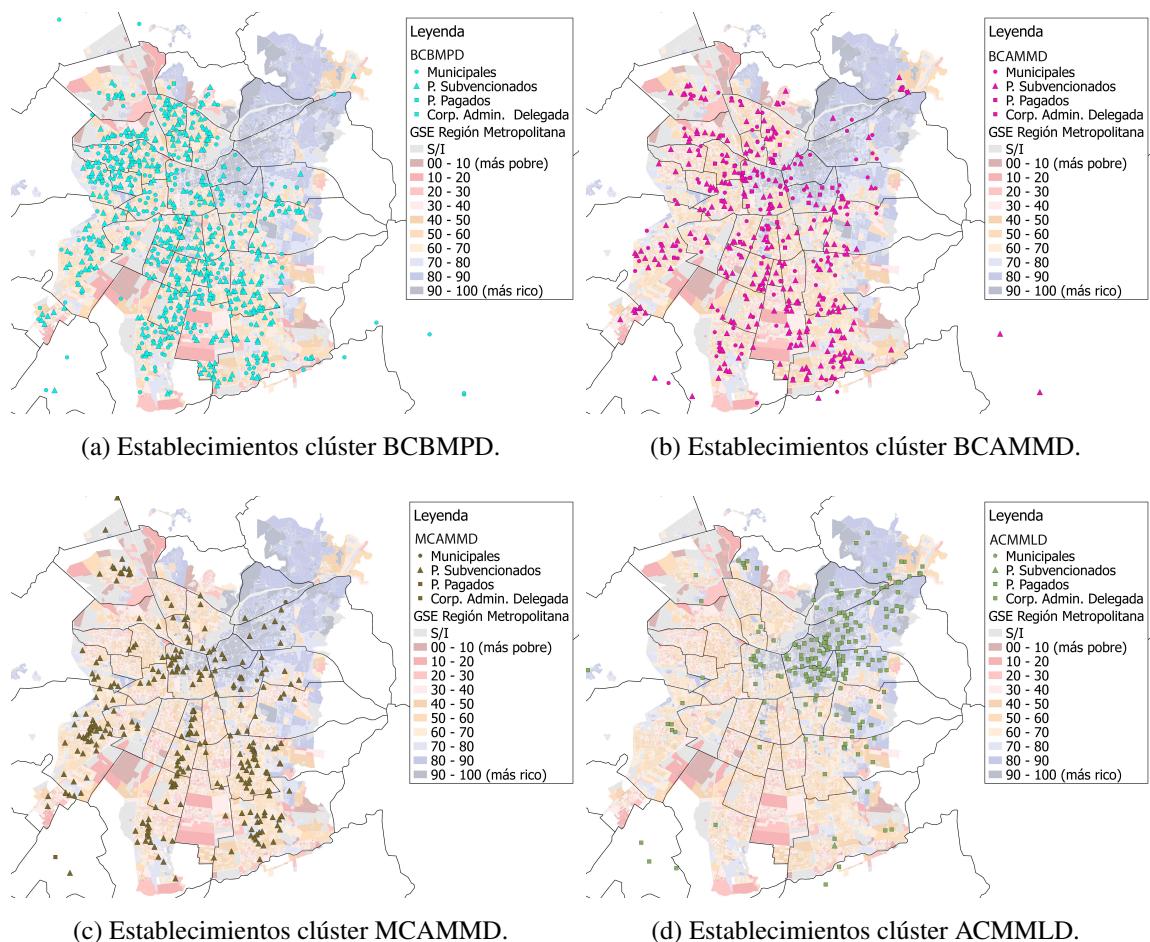


Figura 4.7: Mapas de clústers de establecimientos (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.

En la figura 4.7c los colegios del grupo MCAMMD se encuentran ubicados en las zonas anaranjadas, pertenecientes a los deciles 4 al 7. Pero a diferencia de los anteriores, estos se encuentran distribuidos en grupos y no se extienden de manera uniforma por la capital. Por último, el clúster ACMMLD (figura 4.7d) se ubica en la zona oriente de la Región Metropolitana, en donde se encuentran los GSE de mayor ingreso per cápita (deciles del 8 al 10). Esto concuerda con que los establecimientos de este clúster son los que tienen las matrículas y mensualidades más elevadas de la capital.

De manera general se puede apreciar que los establecimientos se encuentran principalmente en el área metropolitana de la capital (zonas coloreadas según su GSE). Además, estos se ubican principalmente en las zonas de colores naranjos y azules, correspondientes a los deciles 3 al 10, dejando a los sectores más pobres (deciles 1 y 2) con una presencia casi nula. Esto, de una u otra manera, refleja la importancia que tiene el grupo socioeconómico sobre la ubicación de un establecimiento y los niveles de costo que tienen.

En los mapas de la figura 4.8, y utilizando la geolocalización de las matrículas, se muestra la distribución de los alumnos que asisten a los establecimientos de los diferentes clústers. Es decir, en cada mapa se muestran los alumnos pertenecientes a los colegios que conforman los clústers BCBMPD, BCAMMD, MCAMMD y ACMMLD. En el mapa 4.8a los alumnos se distribuyen por casi toda el área metropolitana, exceptuando el sector oriente y concentrándose en sectores de la zona sur y poniente. Los estudiantes del clúster BCAMMD (figura 4.8b) se encuentran en toda el área metropolitana, con menor proporción y densidad en el sector oriente y presentando una mayor densidad en el sur de la capital.

En 4.8c el alumnado se encuentra por toda la capital, concentrándose de manera más densa en la periferia del sector poniente y seguida por dos sectores del sur de la capital. Por último, en el cuarto clúster (figura 4.8d) todos sus estudiantes se sitúan en el sector oriente, concentrándose en el sector nororiente.

A partir de los mapas de la figura 4.8 y lo anteriormente descrito, se puede apreciar que en el caso de los primeros tres clústers de establecimientos, sus alumnos predominantemente viven en zonas pertenecientes a los deciles del 4 al 7, a diferencia de lo que ocurre con los del último clúster. Los alumnos de colegios pertenecientes al clúster ACMMLD residen casi

en su totalidad en sectores de los deciles 8 al 10, es decir, en los sectores donde viven las familias con mayor ingreso per cápita.

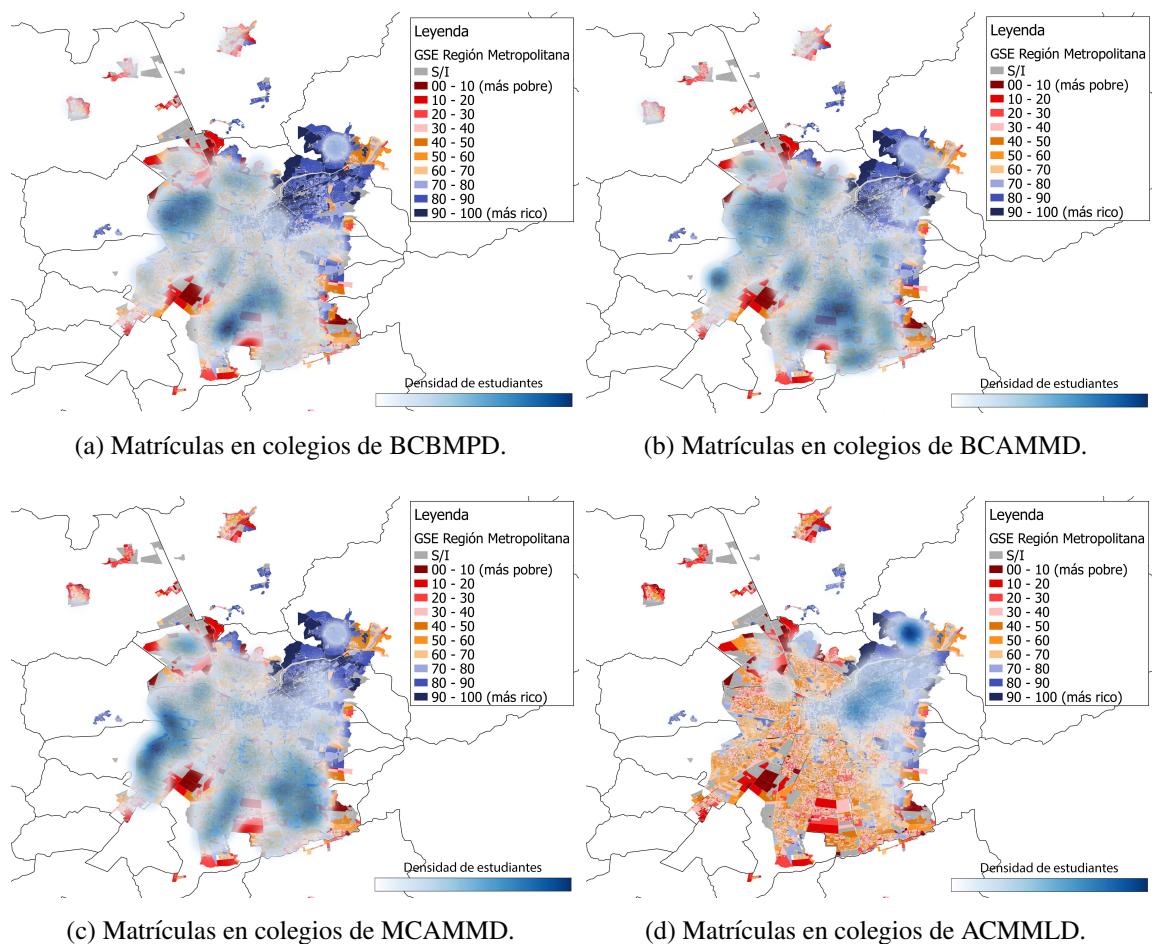


Figura 4.8: Mapas de calor de matrículas (con atributos relacionales) en clústers de establecimientos sobre mapa GSE de la Región Metropolitana.

Al momento de analizar la distribución de los clúster de matrículas sin considerar los atributos de relación, podemos ver en la figura 4.9 que los clúster MCB y HCB se distribuyen de manera similar. Los estudiantes de dichos clústers están dispersos por casi toda el área metropolitana, disminuyendo su cantidad en el sector oriente y concentrándose en el sector sur y poniente. Por otro lado los clústers MSB y HSB se distribuyen de igual manera, ocupando toda el área metropolitana. Estos presentan varias zonas de alta densidad de alumnos, destacando entre ellas la zona norte del sector oriente de la capital.

Además, en las imágenes 4.9 se aprecian los grupos socioeconómicos en los cuales los clústers se encuentran, los primeros dos están presentes en los deciles del 4 al 7 principalmente y los otros dos en los deciles del 4 al 10, donde su *peak* está en los deciles del 8 al 10.

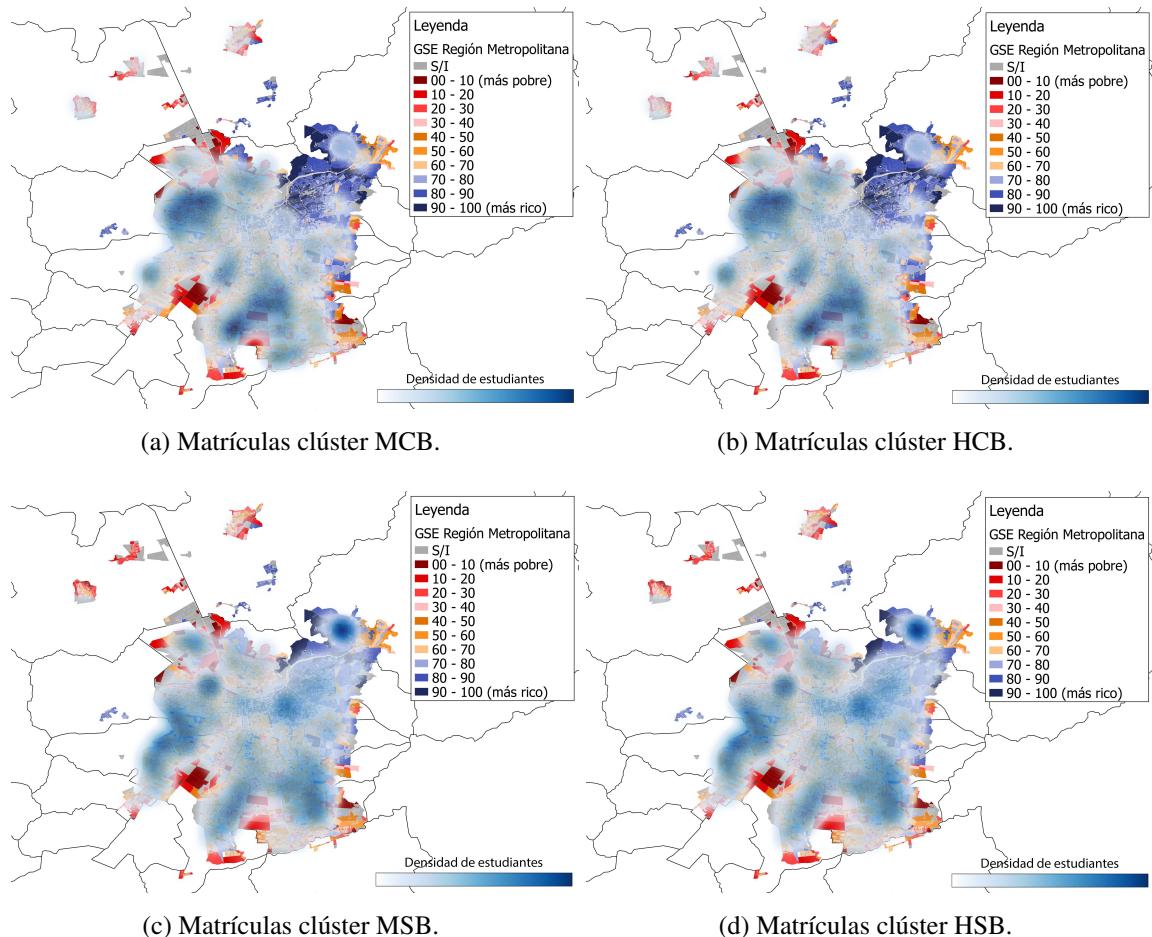


Figura 4.9: Mapas de calor de clústers de matrículas sobre mapa GSE de la Región Metropolitana.

Al analizar las imágenes 4.10a y 4.10b podemos observar una gran similitud, en donde los alumnos se distribuyen por casi toda el área metropolitana, pero disminuye notoriamente en el sector oriente. En la figura 4.10c encontramos alumnos en casi toda la capital, pero se concentran en 3 sectores, 2 en la zona sur y 1 en la zona poniente. Por último, en 4.10d, casi todos los alumnos residen en el sector oriente de la capital, siendo la periferia de este donde

existe una mayor concentración.

A diferencia del análisis sin variables de relación, las primeras 3 figuras muestran que los alumnos viven en sectores socioeconómicos que van del decil 2 al decil 8, es decir, comprenden grupos muy variados. Caso aparte es la última imagen, donde los estudiantes residen en sectores pertenecientes a los deciles del 8 al 10.

Un punto importante a destacar en ambos análisis de clústers de matrículas es que en general los alumnos no residen en sectores pertenecientes a lo deciles 1 y 2, es decir, a los sectores de menor ingreso.

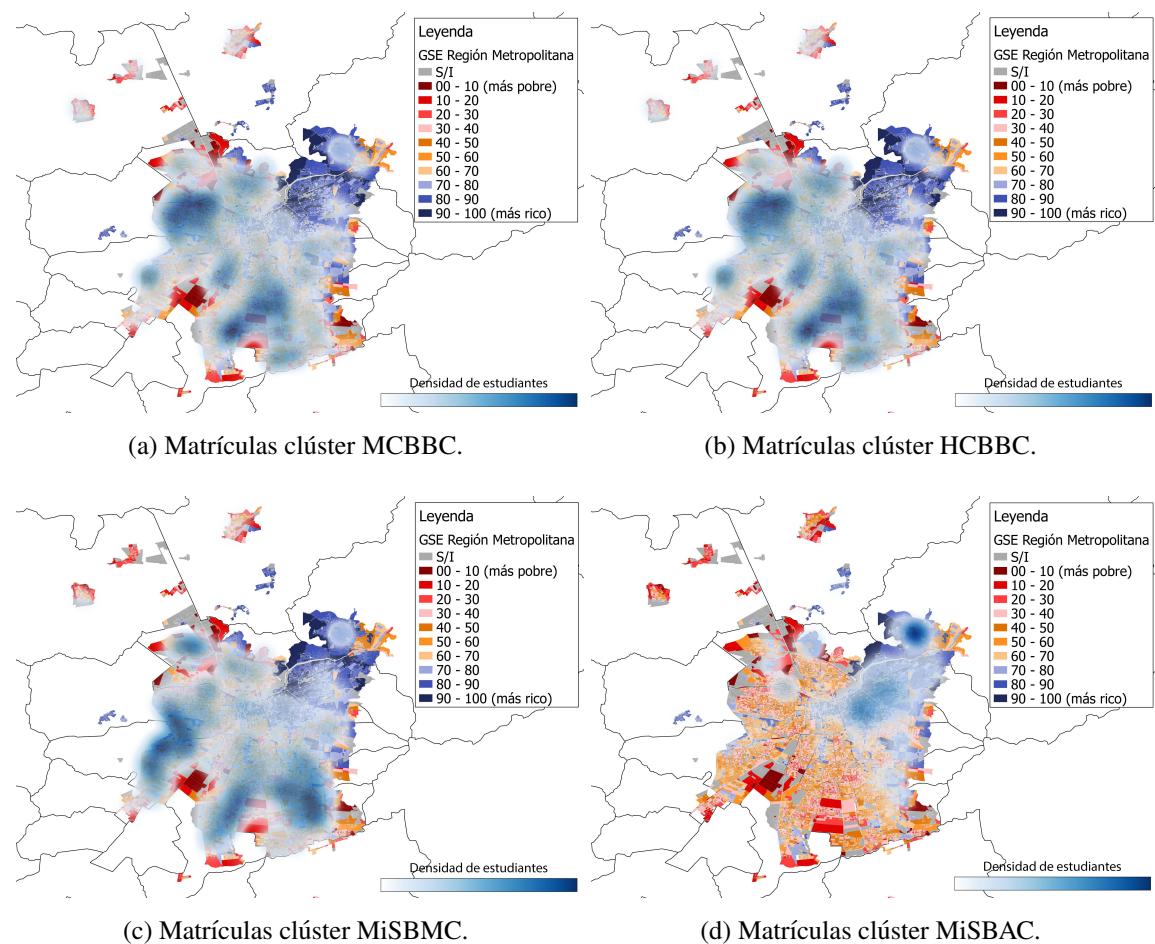


Figura 4.10: Mapas de calor de clústers de matrículas (con atributos relacionales) sobre mapa GSE de la Región Metropolitana.

Conclusiones

En este capítulo se presentan las conclusiones de cada etapa del estudio desarrollado. Al final de este se encuentran las conclusiones finales de la investigación, posibles extensiones y el cumplimiento de los objetivos propuestos en la definición del problema.

Anexos A

Anexo I

Tabla A.1: Resumen

Año	CIAE	MIME	% de coincidencia
2013	2110	2040	96,68
2014	2095	2049	97,8
2015	2088	2057	98,52
2016	2068	2061	99,66

Bibliografía

- [1] Manuel Canales, Cristián Bellei, and Víctor Orellana. ¿Por qué elegir una escuela particular subvencionada? sectores medios emergentes y elección de escuela en un sistema de mercado. *Estudios Pedagógicos (Valdivia)*, 42:89 – 109, 2016.
- [2] Ministerio Chile. MIME - Ministerio de Educación de Chile. <http://www.mime.mineduc.cl/mvc/mime/portada>. Accedido: 7 Oct. 2017.
- [3] Rómulo A. Chumacero, Daniel Gómez, and Ricardo D. Paredes. I would walk 500 miles (if it paid): Vouchers and school choice in Chile. *Economics of Education Review*, 30(5):1103 – 1114, 2011. Special Issue on Education and Health.
- [4] Ministerio de Desarrollo Social. Observatorio Social - Ministerio de Desarrollo Social - Gobierno de Chile. http://observatorio.ministeriodesarrollosocial.gob.cl/casen/basededatos_historico.php. Accedido: 22 Feb. 2018.
- [5] Francisco A. Gallego and Andrés Hernando. School choice in Chile: Looking at the demand side. <http://economia.uc.cl/publicacion/school-choice-in-chile-looking-at-the-demand-side/>, 09 2009.
- [6] Daniel Gómez, Rómulo A. Chumacero, and Ricardo D. Paredes. School choice and information. *Estudios de economía*, 39:143 – 157, 12 2012.
- [7] Judith Hurwitz and Daniel Kirsch. *Machine Learning for dummies*. John Wiley Sons, Inc, 2018.
- [8] Leonard Kaufman and Peter Rousseeuw. *Finding Groups in Data: An Introduction To Cluster Analysis*. 01 1990.
- [9] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, March 1982.
- [10] Alberto Montresor and Alessio Guerrieri. Decentralized clustering with estimation of the number of clusters. 2010.

- [11] Dau Pelleg and Andrew Moore. X-means: Extending k-means with efficient estimation of the number of clusters. In *In Proceedings of the 17th International Conf. on Machine Learning*, pages 727–734. Morgan Kaufmann, 2000.
- [12] Claudio Sapelli and Arístides Torche. Subsidios al alumno o a la escuela: efectos sobre la elección de colegios públicos. *Cuadernos de economía*, 39:175 – 202, 08 2002.
- [13] C. Soto, X. Saavedra, I. Larraguibel, and F. Flores. Estadísticas de la educación 2016. page 15, 2017.
- [14] Rüdiger Wirth. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*, pages 29–39, 2000.