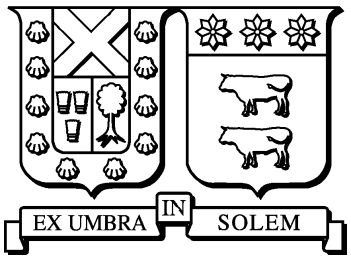
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INFORMÁTICA

SANTIAGO – CHILE



“TÍTULO DE LA MEMORIA”

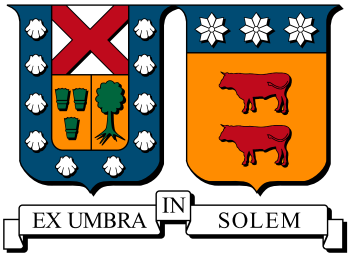
CARLOS ANDRÉS VARGAS POBLETE

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIUBOV DOMBROVSKAIA

ENERO 2018

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INFORMÁTICA
SANTIAGO – CHILE



"TÍTULO DE LA MEMORIA"

CARLOS ANDRÉS VARGAS POBLETE

MEMORIA DE TITULACIÓN PARA OPTAR AL TÍTULO DE INGENIERO CIVIL INFORMÁTICO

PROFESOR GUÍA: LIUBOV DOMBROVSKAIA
PROFESOR CORREFERENTE: PATRICIO RODRÍGUEZ

ENERO 2018

Agradecimientos

Resumen

Abstract

Índice de Contenidos

Índice de cuadros

Índice de figuras

Glosario

Introducción

Capítulo 1

Definición del Problema

Este trabajo realiza un estudio sobre algoritmos de clasificación no supervisados, especificamente X-Means, en bases de datos de establecimientos y estudiantes de la región metropolitana para entender cual es la estructura subyacente que estos poseen. Con esto se busca poder establecer una manera de clasificar los colegios y alumnos sin contar con una idea preconsevida, sino que permitir que los datos analizados entreguen dicha información.

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.

1.0.1. Datos a analizar

Las bases de datos de establecimientos y matrículas fueron facilitadas por el Centro de Investigación Avanzada en Educación (CIAE). Donde la primera fue complementada con la información disponible en la página web del Ministerio de Educación de Chile, MIME, mediante la técnica de *web scraping*.

La base de datos de establecimientos corresponde a los colegios de la región metropolitana que imparten enseñanza básica y media para niños y jóvenes, tanto humanísta científico como técnico profesional. Por otro lado las matrículas corresponden a estudiantes que se

encuentran inscritos en los colegios anteriormente descritos.

1.1. Identificación del Problema

1.2. Objetivos Generales

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.

Para esto se utilizará un algoritmo de clusterización sobre establecimientos y matrículas, los cuales se asociarán entre sí para determinar los atributos que los relacionan.

1.2.1. Objetivos Específicos

1. Generar clústers de establecimientos mediante el uso del algoritmo X-Means y comparar los resultados según el número de atributos escogidos para cada prueba.
2. Generar clústers de matrículas mediante el uso del algoritmo X-Means y comparar los resultados según el número de atributos escogidos para cada prueba.

Queremos entender cual es la estructura subyacente, tener forma de clasificar a los establecimientos y estudiantes de las escuelas de Chile, no tener ideas preconsevidas acerca de eso, que los datos entreguen la información

Capítulo 2

Estado del Arte

En el transcurso de los años han habido varios investigadores interesados en conocer los diferentes factores que influyen en los padres al momento de elegir un colegio en Chile.

En el 2002 Sapelli y Torche [8] 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 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 resultados que obtuvieron apuntan a que algunos de los factores más determinantes son el nivel de ingreso, la educación de los padres, la recepción de subsidios y la calidad del colegio. Además destacan que por ser los subsidios por colegios y no por alumno, es decir no son portables, genera que sea más difícil para las familias de menores recursos acceder a estos si deciden optar por un colegio donde el nivel de subsidio es menor. Otro punto importante

[1] 98,2 % de cobertura educacional para el nivel de enseñanza básica en el año 1996. Fuente: CASEN 1996.

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 [4] buscando resolver la interrogante de cómo los padres escogen el colegio para sus hijos usaron un modelo basado en el desarrollado por McFadden [6], junto con las especificaciones planteadas por Berry, Levinsohn y Pakes [1]. 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. A partir del modelo y los datos utilizados se obtuvo que existen dos variables que afectan más al momento de escoger un colegio, las cuales 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 [8].

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 [8], 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. Mediantes 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 [5], 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 [8]. 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 [2], 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.

Capítulo 3

Propuesta de Solución

3.1. Pre-procesamiento de los datos

Antes de ejecutar el algoritmo de clusterización se realiza un proceso ETL (extract, transform and load), el cual permite limpiar y estandarizar las bases de datos.

Se seleccionan atributos numéricos y categóricos para los establecimientos y se generan otros a partir de la información extraída del MIME, debido a que mucha de esta no se encuentra estandarizada. De la misma forma se seleccionan atributos numéricos y categóricos para las matrículas, y a partir de variables no seleccionadas se calculan nuevas como es el caso de la sobre edad. Además se incorpora de manera relacional con el establecimiento al que asisten su nivel de copago y la distancia a la cual viven del colegio. De igual manera, para los establecimientos se agregan atributos relacionales, como lo es la distancia a la que viven sus estudiantes en su percentil 75 y el índice de desarrollo de la educación (IDE) por rango.

Luego de esto se transforman las variables categóricas o nominales a numéricas para poder utilizarlas en el algoritmo de clusterización junto al resto de las variables numéricas.

Se imputaron los datos faltantes dentro de cada uno de los atributos previamente escogidos mediante el algoritmo MICE (*(Multiple imputation by chained equations*) para eliminar todos los valores nulos.

3.2. X-Means

X-Means es un algoritmo de agrupamiento, extendido de K-Means, propuesto por Pelleg y Moore [7] en el cual se busca dar solución a los principales problemas de K-Means. Estos son: baja escalabilidad computacional, requiere el ingreso de un número determinado de clústers y es sensible a mínimos locales.

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

3.2.1. Pseudocódigo de X-Means[?]

Algoritmo 1 X-Means

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

Require: Función 2Means(S) retorna 2 clústers

```
1:  Clustering ← 2Means(S)
2:  Mejor_Puntuacion ← −∞
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
```

Cuadro 3.1: Atributos seleccionados y generados para la base de datos de establecimientos.

Atributo	**Descripción**
area_metropolitana_rbd	Pertenencia al area metropolitana.
cod_depe	Código de dependencia del establecimiento.
gen_rbd	Género del establecimiento.
mat_total	Matrícula total de alumnos.
prom_alu_cur	Promedio de alumnos por curso.
pago_mat	Nivel de pago de matrícula.
pago_men	Nivel de pago de mensualidad.
becas_disp	Becas disponibles en el establecimiento.
convenio_sep	Posee convenio de subvención escolar preferencial (SEP).
deportivo	Nivel deportivo del establecimiento.
req_papeles	Requisitos de papeles para postular.
req_pruebas	Requisitos de prueba para postular.
req_entrevista	Requisitos de entrevista para postular.
req_pago	Requisitos de pago para postular.
req_otros	Requisitos de cualquier tipo que no clasifique en las categorías anteriores.
enf_academico	Enfoque acádemico.
enf_valorico	Enfoque valórico.
enf_laboral	Enfoque laboral.
enf_otros	Enfoque de otro tipo que no clasifique en las categorías anteriores.
apoyo_tutorias	Ofrece ayuda a los alumnos mediante tutorías.
apoyo_especialistas	Ofrece ayuda a los alumnos mediante especialistas.
apoyo_otros	Ofrece ayuda a los alumnos de cualquier otra forma que no clasifique en la categorías anteriores.
s_basica	Establecimiento de enseñanza básica.
s_media	Establecimiento de enseñanza media.
completa	Establecimiento de enseñanza completa.
IDE_rango	Índice de desarrollo de la educación para todos por rango.
dist_percentil_75	Distancia del percentil 75 de los alumnos que asisten al establecimiento.

Cuadro 3.2: Atributos seleccionados y generados para la base de datos de matrículas.

Atributo	**Descripción**
area_metropolitana_alu	Pertenencia al área metropolitana.
gen_alu	Género del establecimiento.
area_metropolitana_alu	
cod_sec	Código del sector económico.
cod_espe	Código de especialidad.
cod_rama	Código de rama.
grado_sep	Corresponde a un nivel SEP.
beneficiario_sep	Indicador del alumno beneficiario de la SEP.
criterio_sep	Criterio por el cual se considera prioritario.
sobre_edad	Diferencia entra la edad actual y la esperada para el nivel.
dist_actual	Sitancia del alumno a su establecimiento actual
pago_mat	Nivel de pago de matrícula.
pago_men	Nivel de pago de mensualidad.

Cuadro 3.3: 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

Capítulo 4

Implementación

Conclusiones

Bibliografía

[1] Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995.

[2] 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. *Revista Estudios Pedagógicos*.

[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] Francisco A Gallego and Andrés Hernando. School choice in chile: Looking at the demand side. *Pontificia Universidad Catolica de Chile Documento de Trabajo*, (356), 2010.

[5] Daniel Gómez, Rómulo A Chumacero, and Ricardo D Paredes. School choice and information. *Estudios de economía*, 39:143 – 157, 12 2012.

[6] Daniel McFadden. Conditional logit analysis of qualitative choice. pages 105–142, 1974.

[7] 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.

[8] 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.