Análisis Usando Data Science Sobre la Participación en Elecciones Presidenciales en Chile

Andres Arenas R., Indy Navarro V. Juan Saavedra J.
andres.arenas.r@mail.pucv.cl , indy.navarro.v@mail.pucv.cl , juan.saavedra.j@mail.pucv.cl
Escuela de Ingeniería Industrial - Pontificia Universidad Católica de Valparaíso

Ingeniería Industrial
PONTIFICIA UNIVERSIDAD CATOLICA DE VALPARAISO

Introducción - Antecedentes

Esta investigación se enmarca en un análisis de las tendencias electorales de diversas zonas de Chile, considerando las elecciones presidenciales correspondiente a los años 2013 y 2017, tanto en primera como segunda vuelta.

- Este estudio surge por el interés de comprender la poca precisión de los instrumentos de medición en las últimas elecciones (*CADEM 2017*), donde el sistema de encuestas no fue capaz de predecir con precisión los resultados de la misma.
- Actualmente, la tendencia y el comportamiento electoral solamente se ve representado de manera oficial por medio de la realización de encuestas, como la *CADEM* o la realizada por el *CEP*, siendo las estadísticas oficiales que se presentan a través de los medios de comunicación.
- El foco de la investigación se centra en comparar la Región Metropolitana con el resto de las regiones, junto con encontrar características símiles en comunas como son la participación y tendencias políticas.

Metodología Aplicada

- **Obtención de datos**
  - Este estudio emplea principalmente el uso del datos pertenecientes a organismos públicos del gobierno de Chile, obtenido de organismos como lo son el *SERVEL* (Servicio electoral) y el *INE* (Instituto Nacional de Estadísticas). Los siguientes corresponden a algunos de los *dataset's* utilizados:
    1. Resultados electorales correspondientes a elecciones presidenciales de los periodos 2013-2017 (considera primera como segunda vuelta).
    2. Resultados demográficos regionales, obtenidos a través de los resultados del *Censo 2017*.
    3. Proyecciones de crecimiento demográficos de población regional, correspondientes al año 2014 (*INE*).

  Los datos mencionados anteriormente al pertenecer a organismos públicos son de libre acceso, con posibilidad de ser obtenidos desde la fuente principal, o bien desde terceros (*Datachile*)
- **Metodología empleada: K-Means**
  - Esta metodología permite realizar un agrupamiento de un set de datos que comparten ciertas características, considerando que, anteriormente no han sido etiquetados o definidos bajo algún criterio como se puede observar en la siguiente figura:

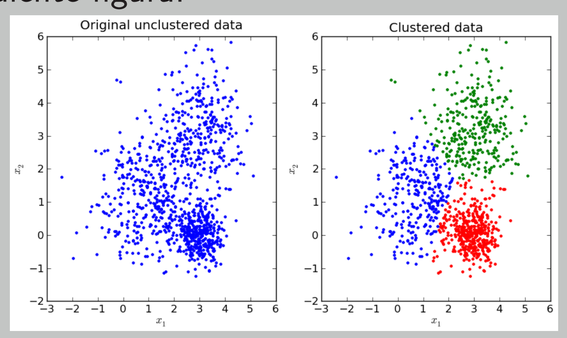

Figure 1: Representacion del algoritmo K-Means

  - A pesar que la metodología se encuentra principalmente basada en la minimización del cálculo de distancias entre dos puntos, este algoritmo por medio de su resultado de agrupación permite de igual forma generar análisis sobre *variables cualitativas*
  - Para la aplicación del método K-Means se utilizó el índice de *Calinski-Harabasz*, el cual está definido como la razón entre la dispersión interior de los clústers y la dispersión entre los clústers, como se indica a continuación:
  -
$$\frac{SS_b}{SS_w} \times \frac{(N-k)}{(k-1)} \tag{1}$$
  - Para la ecuación anterior los valores corresponden a:
  - SS_b : Varianza global entre clústers.
  - SS_W: Varianza general dentro del clúster.
  - N: número de observaciones. ; k: número de clústers.

Resultados de la metodología

- Para el desarrollo del análisis, se utilizan las votaciones de la segunda vuelta de las elecciones presidenciales del 2017. Para cada comuna se tienen 3 atributos en estudio: porcentaje votación Sebastián Piñera, porcentaje votación Alejandro Guiller y participación en las elecciones de la comuna.

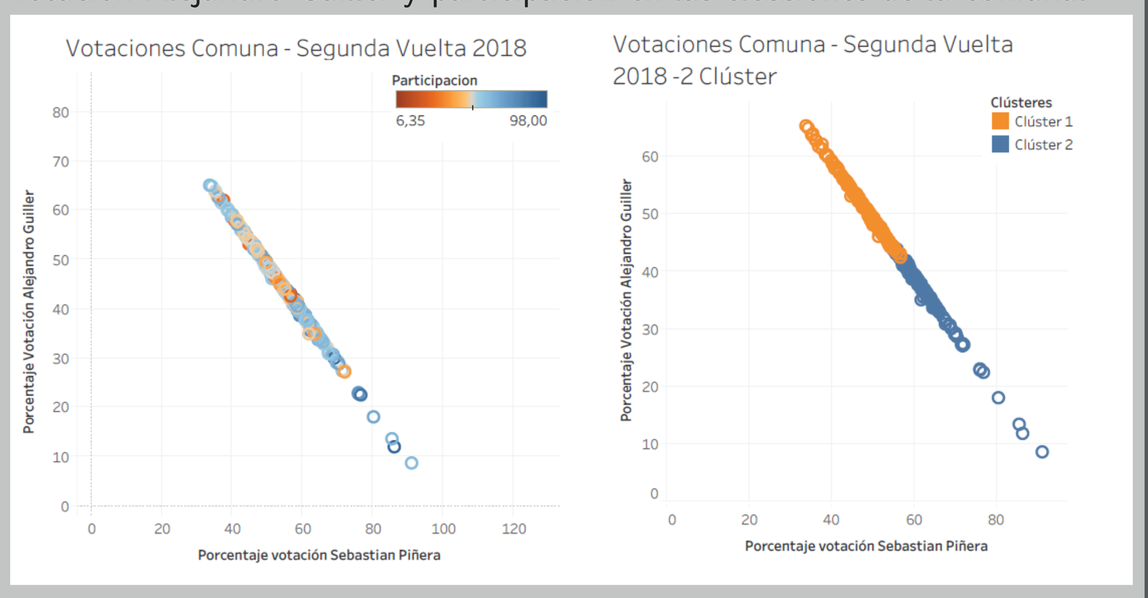

Figure 2: Comportamiento votaciones y clúster de elecciones

  - Un primer paso, fue graficar el comportamiento de las comunas, dejando en el eje horizontal los votos de Sebastián Piñera, y en el eje Vertical los votos de Alejandro Guiller, y cada punto corresponde a una comuna, con sus respectivos porcentajes de votación a cada candidato. En el extremo inferior derecho, se encuentra Colchane con un apoyo a Sebastián Piñera del **91.05%**, seguido de Vitacura con un **86, 37%**. En el otro extremo se encuentra Canela con un **65.11%** de apoyo a Alejandro Guiller, seguido de Petorca con un **64.82%**. Como tercera dimensión en análisis, se aade la participación de la comuna, que gráficamente se representa con el color del punto. Los puntos de color azul son comunas con una alta participación como Vitacura con un **93, 5%**. Y con rojo las comunas con baja participación como Navidad con **18.81%**.
  - En este grafico es posible ver que las comunas con mayor votación hacia Sebastián Piñera, tienden a tener una mayor participación de las elecciones. Y por el otro lado se tiene porcentajes de votaciones más divididos, pero también menor participación de la comuna.
  - Para poder cuantificar el resultado, estableciendo una métrica que considere las tres variables en estudio, se aplica k-means, que permite realizar agrupación de los datos. Al aplicar la metodología, se obtienen dos cl´ster, que, como aspectos generales, agrupa por un lado las comunas con un apoyo a Sebatián Piñera superior al **55%**, además de tener comunas con mayor participación.
  - Al analizar los resultados, se aprecia que las comunas con mayor apoyo al candidato Sebastián Piñera, son de las comunas con mayor participación, además son de las comunas con mejor situación socioeconómica, como Vitacura, Lo Barnechea y Las Condes.

Conclusiones y Trabajo futuro

- En un trabajo posterior se aplicará el mismo método para las elecciones presidenciales del año 2013 y 2017 primera y segunda vuelta. Se espera que el comportamiento de cada zona se sostenga en cuanto a nivel de participación y tendencia política.
- Se ha encontrado una relación respecto a las comunas con mayor porcentaje de participación y la preferencia a cierto candidato o tendencia política.
- Se considera la incorporación del método CHAID para la clasificación de diversas comunas y regiones.
- Se considera la incorporación de nuevas variables con la finalidad de profundizar los resultados obtenidos a partir del estudio.

Este proyecto está siendo desarrollado bajo la realización del curso de Data Science Aplicada, dictado por el PhD. Victor Leiva Sánchez en la Escuela de Ingeniería Civil Industrial (PUCV),