Clusterización de la población juvenil chilena utilizando un modelo de mezcla de gaussianas

Diego Ibáñez e Ignacio Vargas

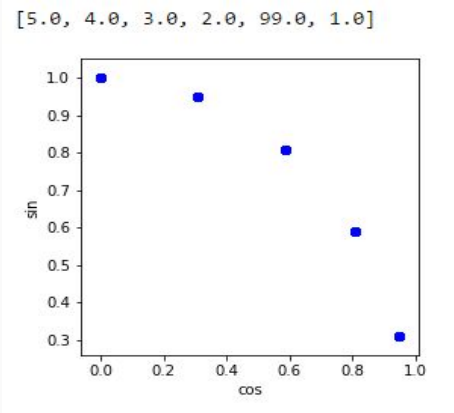
MA5203-1 Aprendizaje de Máquinas Probabilístico
Prof.: Felipe Tobar
Aux.: Alejandro Veragua - Alejandro Cuevas
8 de julio 2017

El problema y Metodología

Segmentación de la juventud para la focalización de políticas públicas usando datos de la Encuesta Nacional de la Juventud (INJUV)

Limpieza de los datos → Pre-procesamiento → Gaussian Mixture clustering → Análisis de varianzas y distribuciones

Variables categóricas ordinales



Variables categóricas nominales

One Hot Encoder

Normalización

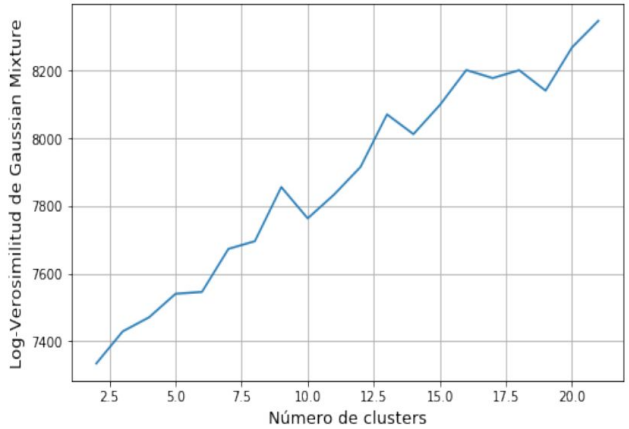
$$\| x \|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}.$$

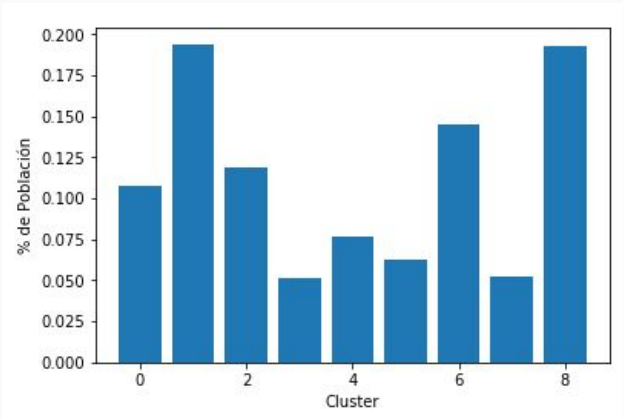
Distancia de Mahalanobis

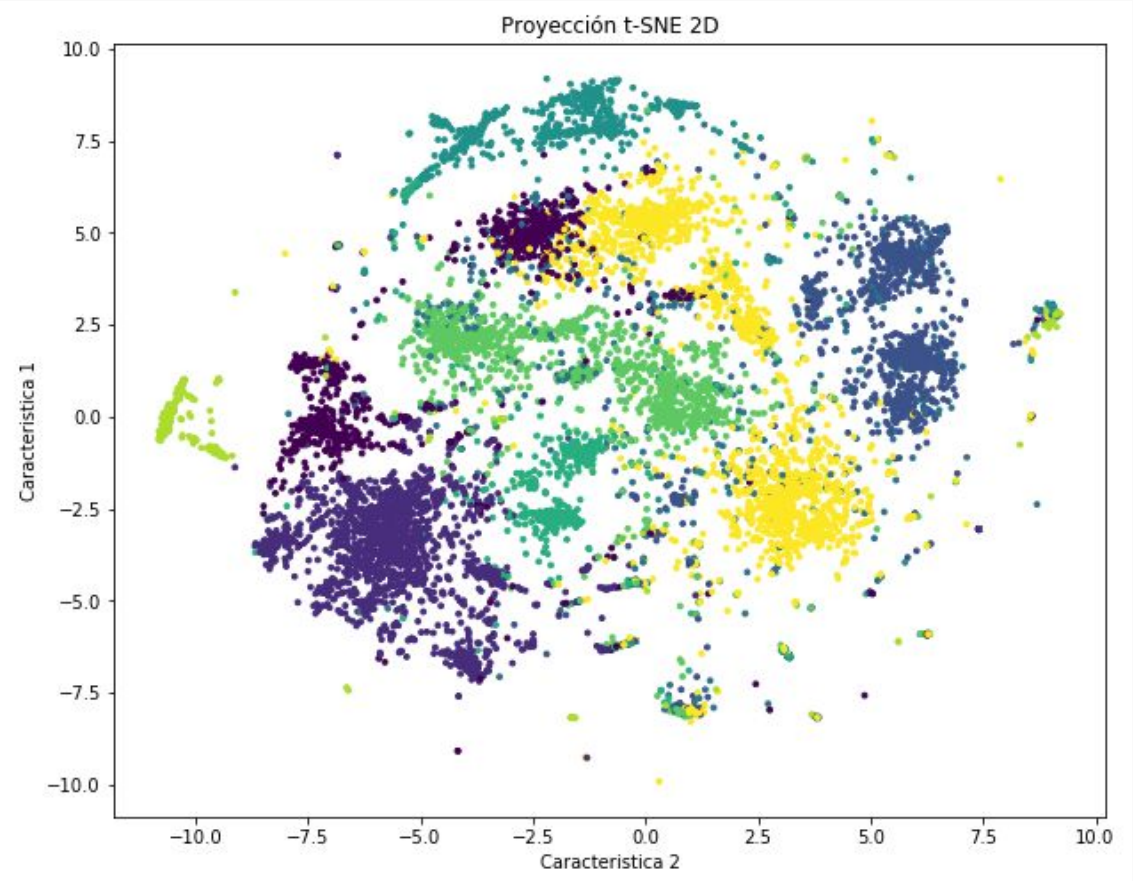
$$d(i,j) = \sqrt{(x-y)^T \sum^{-1} (x-y)}$$

Resultados: Clusterización

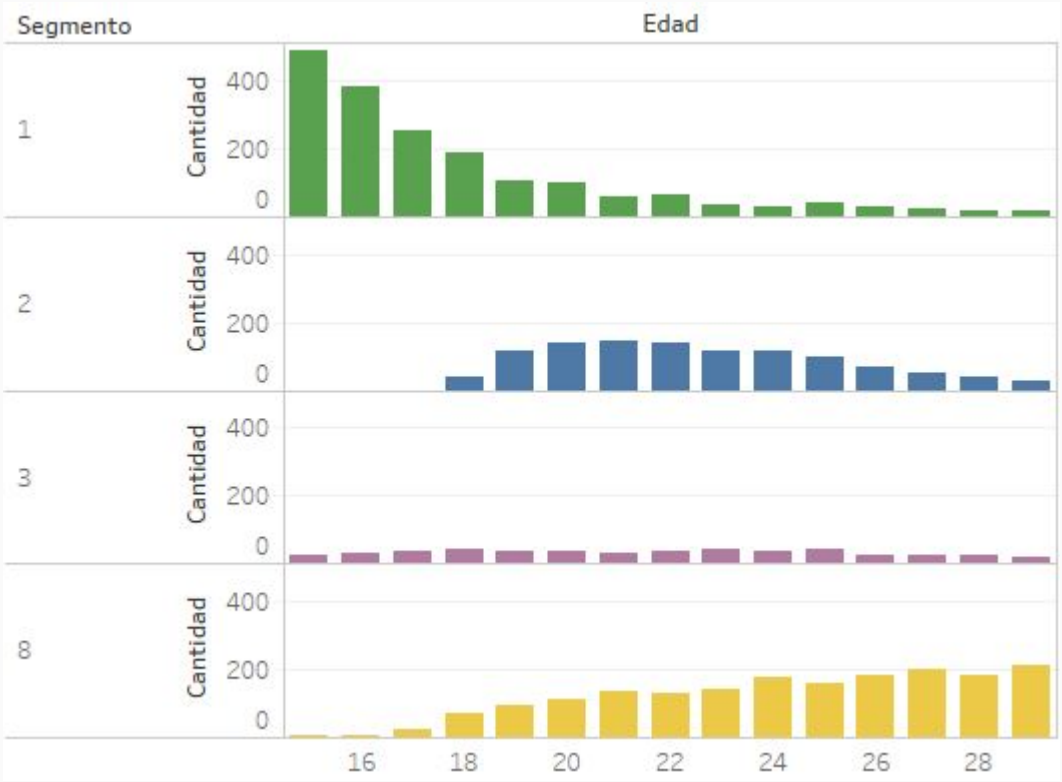
Fig. 1. Score con respecto al numńero de clusters







Resultados: análisis de varianzas y de distribuciones



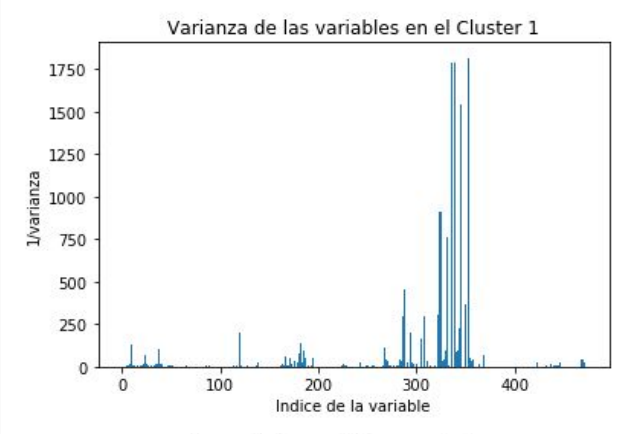
Conclusiones y trabajos futuros

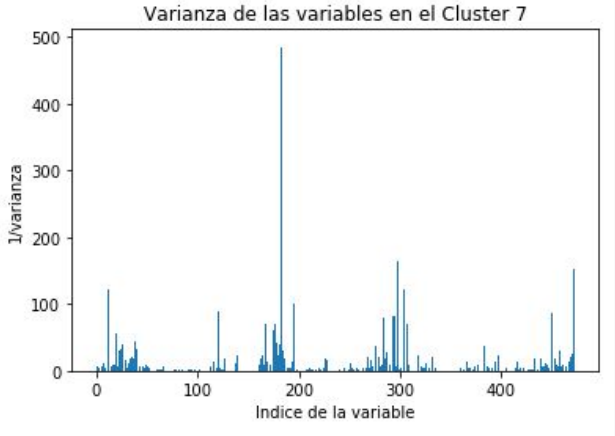
- Número de clusters en función de los resultados
- Comparar con clusterización en cascada
- Naturaleza de los datos → Interpretación “Manual”
- Agregar una selección de características → Usar variables que diferencian a los individuos

Ej: Actuales caracterizaciones:

Cluster 0: “*No ha probado drogas duras, No tiene crédito corfo, No percibe ingresos por inversiones, No usa métodos anticonceptivos poco convencionales*”

Cluster 1: “*Está estudiando, Iniciado sexualmente, Usa Facebook, Usa whatsapp, Usa preservativo, Tiene celular, Bebe alcohol*”





Referencias

- (Scikit-learn) Gaussian mixture models: http://scikit-learn.org/stable/modules/mixture.html
- 8va Encuesta Nacional de la Juventud, Instituto Nacional de la Juventud: http://www.injuv.gob.cl/portal/categoria/publicaciones/encuestas-de-juventud/