Análisis de datos médicos con aprendizaje no supervisado

Santiago Mora Cruz | A01369517
Gabriel Reynoso Escamilla | A01643561
Guillermo Villegas Morales | A01637169

Junio de 2024

La descripción de la base de datos seleccionada para este proyecto.

La base de datos seleccionada son 2126 medidas extraídas de cardiotocografías fetales, y fueron clasificados por Obstetras profesionales. Una cardiotocografía es un registro de la frecuencia cardiaca fetal y de las contracciones uterinas de la madre. Más información sobre la base de datos está disponible en la investigación:
Ayres de Campos et al. (2000) SisPorto 2.0 A Program for Automated Analysis of Cardiotocograms. J Matern Fetal Med 5:311-318
Los siguientes son los atributos de la base de datos, y lo que significan, como están descritos en el sitio de kaggle de la base de datos:

Atributo	Significado
baseline_value	Baseline Fetal Heart Rate (FHR)
accelerations	Number of accelerations per second
fetal_movement	Number of fetal movements per second
uterine_contractions	Number of uterine contractions per second
light_decelerations	Number of LDs per second
severe_decelerations	Number of SDs per second
prolongued_decelerations	Number of PDs per second
abnormal_short_term_variability	Percentage of time with abnormal short term variability
abnormal_short_term_variability	Percentage of time with abnormal short term variability
mean_value_of_short_term_variability	Mean value of short term variability

percentage_of_time_with_abnormal_long_term_variability	Percentage of time with abnormal long term variability
mean_value_of_long_term_variability	Mean value of long term variability
histogram_width	Width of the histogram made using all values from a record
histogram_min	Histogram minimum value
histogram_number_of_peaks	Number of peaks in the exam histogram
histogram_mode	Hist mode
histogram_mean	Hist mean
histogram_median	Hist Median
histogram_variance	Hist variance
histogram_tendency	Histogram trend
fetal_health	Fetal health: 1 - Normal 2 - Suspect 3 - Pathological

(https://www.kaggle.com/code/karnikakapoor/fetal-health-classification/input)

Los resultados de aplicar los tres métodos de agrupamiento.

Utilizamos tres métodos de agrupamiento: K-Means, Bisecting K-Means y Gaussian Mixture Model. El número óptimo de grupos para cada método fue determinado usando el Coeficiente *Silhouette* y el índice Davies-Bouldin para los tres casos el número fue dos (Fig. 1), por lo que revisando las distribuciones de cada variable por los diferentes grupos, podemos ver que los tres agruparon de manera muy similar alternando cuál es grupo 0 y grupo 1, pero la dicotomía para ser la misma. Para este conjunto de datos, contamos con una columna 'fetal_health' que es una variable categórica donde se describe la salud del feto según la votación de 3 expertos como normal, sospechosa y patológica (1,2 y 3 respectivamente). Decidimos omitir esta variable y utilizarla como label y después ver si el agrupamiento coincide con el diagnóstico de los expertos y nos diera dos grupos: sanos y no sanos. Como podemos ver en seguida con las distribuciones de los grupos, no parece que haya esta distinción ya que las distribuciones de estos en ambos grupos son similares. Pero dejando de lado la distribución de los labels, hay cuestiones más interesantes de los grupos que se discutirán en las conclusiones individuales.

Método: K-Means Clustering

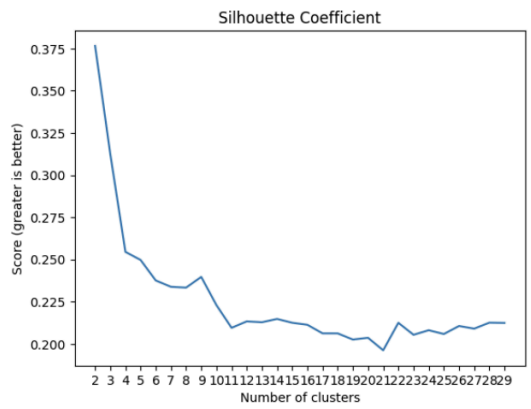
Grupo 0 (fetal_health)		Grupo 1 (fetal_health)	
1	829	1	826
2	229	2	111
3	65	3	66

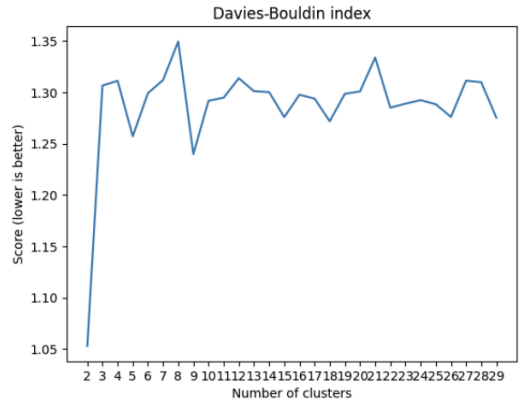
Método: Bisecting K-Means

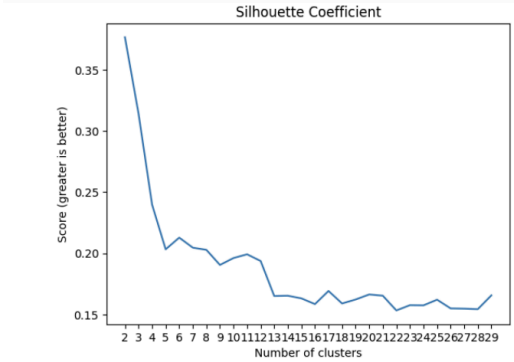
Grupo 0 (fetal_health)		Grupo 1 (fetal_health)	
1	826	1	829
2	111	2	229
3	66	3	65

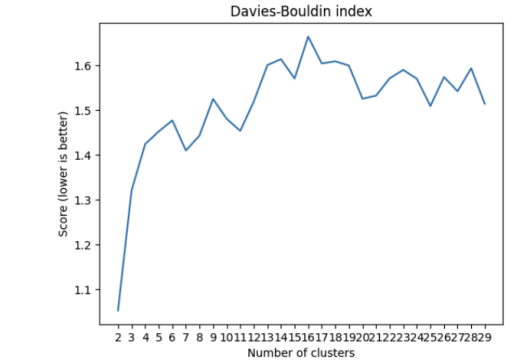
Método: Gaussian Mixture Model

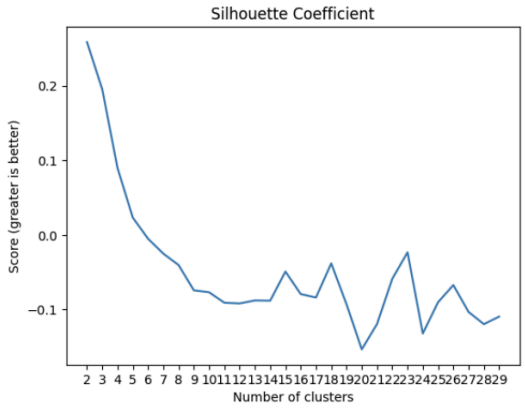
Grupo 0 (fetal_health)		Grupo 1 (fetal_health)	
1	673	1	982
2	110	2	265
3	30	3	66











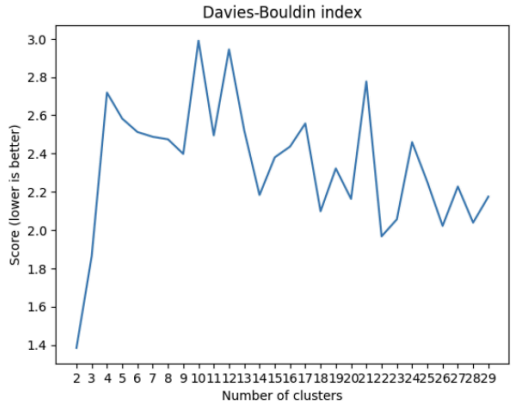
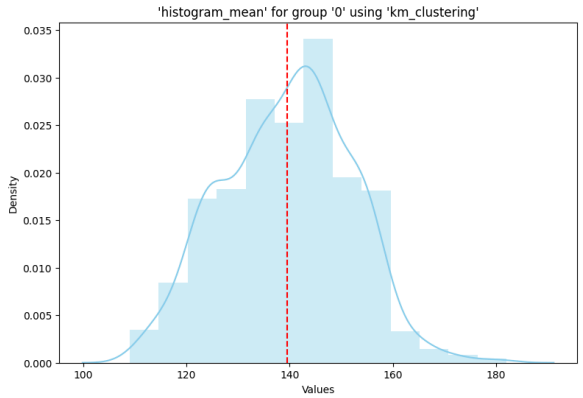
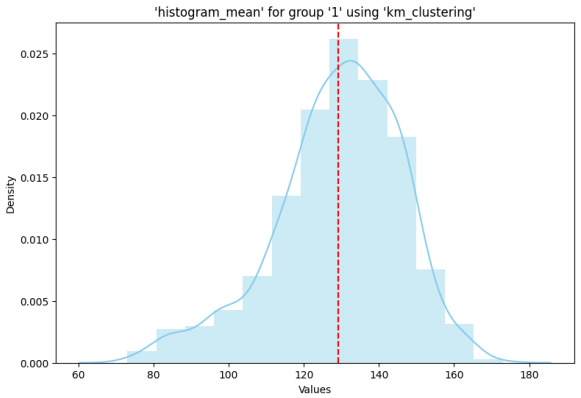


Figura 1: Gráficas utilizadas para determinar el número óptimo de clusters para los métodos *K-Means* (Superior izquierdo), *Bisecting K-Means* (Superior derecho) y *Gaussian Mixture Model* (Inferior izquierdo)

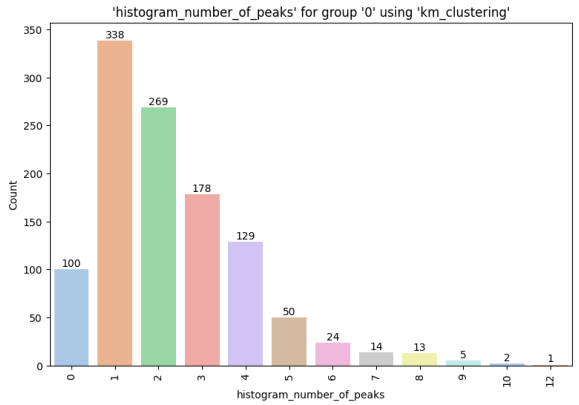
Conclusión personal

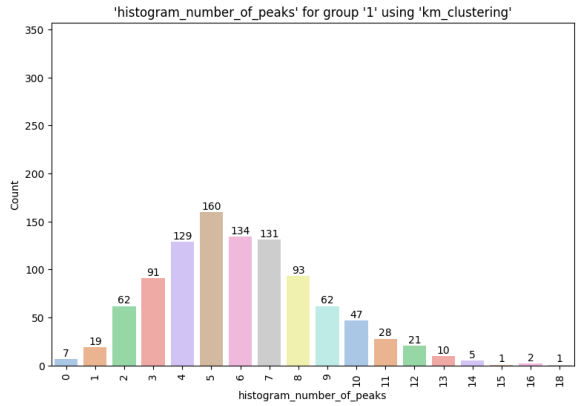
Como se menciona en el reporte, hay cuestiones interesantes más allá de la cantidad de cada label que hay en cada grupo. Por ejemplo, en el siguiente par de histogramas utilizando K-Means Clustering, podemos observar que las med de los histogramas de las cardiotocografías tienen una media menor en el grupo 1, así como una distribución más uniforme.





Además, podemos observar en los siguientes dos histogramas que los picos en las cardiotocografías son más comunes en mayores cantidades en el grupo 1, y estos picos son importantes ya que pueden proporcionar información valiosa para poder determinar si un feto tiene hipoxia, y son uno de los aspectos más importantes a analizar de una cardiotocografía.





A pesar de que no los agrupó simplemente como sanos o no sanos, nos dió información interesante sobre los grupos y, posiblemente, cual es el grupo más propenso a tener hipoxia a pesar de que tiene fetos que en ambos grupos hay fetos clasificados por expertos como saludables, sospechosos y patológicos.

Video tutorial:

El enlace de mi video tutorial, en el que hablo sobre K-Means, es el siguiente: https://youtu.be/W1iNDsFdFoU?si=3icB4pfgbgkNSRid

Además de esto, los videos que tomé como inspiración para el mío fueron los siguientes:
https://youtu.be/R2e3Ls9H_fc?si=VUWaezGb1Hv4Y47-
https://youtu.be/_aWzGGNrcic?si=L8ddCpxkeZOtwnUh
https://youtu.be/4b5d3muPQmA?si=cJuDlm8ZJ8-QLhnM

La razón por la que escogí estos videos es porque me parece que explican K-Means, que es el mismo del que yo hice mi video, de manera clara y dinámica. En el primer video, la explicación es breve y concisa, y logra explicar de manera clara el modelo en menos de 4 minutos. el segundo es un poco más técnico y te explica con ecuaciones matemáticas y ejemplos gráficos cómo es que funciona. Finalmente, el tercer video lo explica de una manera más dinámica y divertida, y es la principal inspiración de mi video tutorial. A pesar de que me inspiré de los tres, el tercero fue el que se me hizo la forma más coloquial y más entendible para cualquier persona de explicar el algoritmo de agrupamiento y la selección de la K más adecuada.