



# *Final*

## *Gaussian Mixture Model (GMM)*

**Grupo 3:**

Tomas Marengo  
Santiago Rivas  
Franco De Simone  
Gastón Francois



# Índice



**Marco Teórico**



**Análisis de Parámetros  
y Métricas**



**Clustering**  
GMM vs K-Means



**Detección de anomalías**  
GMM vs Isolation Forest



**Análisis de Densidad**  
Iris & California Housing



**Procesamiento de audio**



**Conclusiones**

# GMM

# Marco Teórico

# Marco teórico

## ¿Qué es GMM?

Modelo probabilístico basado en la combinación de **múltiples distribuciones normales** (gaussianas) para representar la distribución subyacente de los datos.

Se utiliza para modelar conjuntos de datos donde se asume que cada punto pertenece a una de varias distribuciones gaussianas con cierta probabilidad.

$$P(x) = \sum_{i=1}^K \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

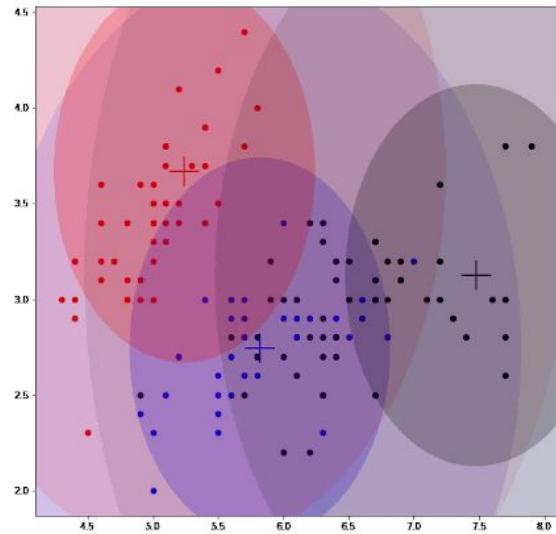
GMM es especialmente útil cuando los datos siguen una distribución multimodal (es decir, cuando hay varias agrupaciones naturales en los datos).

# Marco teórico

## ¿Cómo se usa GMM?

GMM se utiliza principalmente en problemas de **clustering, detección de anomalías y modelado de densidad**. Se puede aplicar en diversos campos, incluyendo:

- **Segmentación de datos:** identificación de subgrupos dentro de un conjunto de datos.
- **Reconocimiento de patrones:** clasificación de señales de audio o imágenes.
- **Análisis de densidad:** estimación de la distribución de datos en el espacio de características.
- **Detección de anomalías:** identificación de puntos de datos atípicos.



# Marco teórico

## ¿Cómo se entrena?

**Algoritmo EM:** método iterativo para encontrar los parámetros óptimos en modelos de mezcla de gaussianas cuando tenemos datos parcialmente observados o con incertidumbre en las asignaciones de cluster.

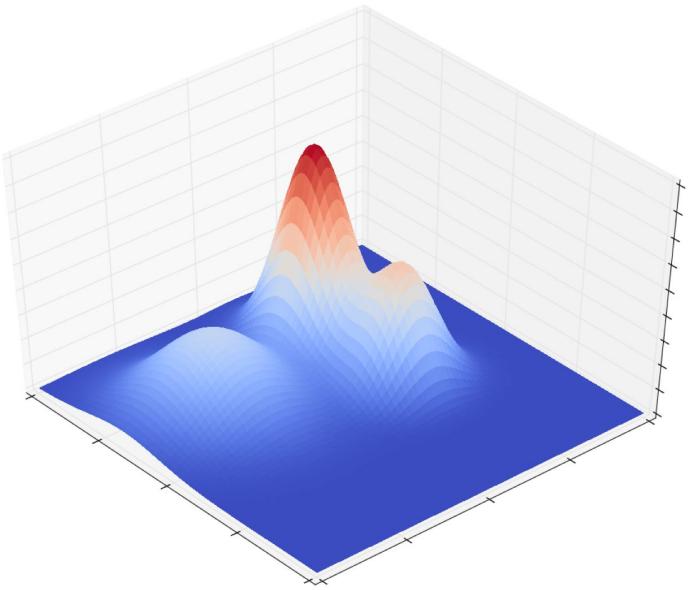
- **Expectation (E-Step):** Se calculan las probabilidades de pertenencia de cada punto a cada componente gaussiana, dadas las distribuciones actuales. Se usa la **regla de Bayes**.
- **Maximization (M-Step):** Se actualizan los parámetros del modelo () en base a las probabilidades calculadas en el paso E.

Se siguen los pasos hasta converger.

- No se garantiza alcanzar un óptimo global.
- Se suele inicializar varias veces con diferentes puntos de partida, y se elige la mejor según alguna métrica.

# Marco teórico

## Fortalezas y debilidades



### Fortalezas

- Capacidad para modelar distribuciones complejas
- Probabilístico
- Flexible en términos de estructura de datos

### Debilidades

- Sensibilidad a la inicialización - mínimos locales
- Sobreajuste
- Requiere la especificación de K
- Costo computacional alto

# GMM

# Análisis de parámetros

# Análisis de parámetros

## Parámetros y métricas

### Parámetros Importantes

- **n\_components:** Número de componentes gaussianas en la mezcla.
  - Un número bajo de componentes puede no capturar la verdadera estructura de los datos (*underfitting*).
  - Un número alto de componentes puede llevar a sobreajuste (*overfitting*).
  - Se elige generalmente con métricas como **AIC** y **BIC**, buscando un balance entre precisión y simplicidad.
- **covariance\_type:** Tipo de matriz de covarianza utilizada para modelar la dispersión de los datos en cada componente.
  - **full:** cada componente tiene su matriz de covarianza completa
  - **tied:** todos los componentes comparten la misma matriz de covarianza.
  - **diag:** se asume que las variables son independientes y cada componente tiene su propia matriz de covarianza diagonal.
  - **spherical:** cada componente tiene una matriz de covarianza isotrópica (una sola varianza para todas las dimensiones).

# Análisis de parámetros

## Parámetros y métricas

### Métricas

- **AIC:** evalúa qué tan bien un modelo se ajusta a los datos penalizando la complejidad. Valores **más bajos** indican un mejor equilibrio entre ajuste y simplicidad. Su fórmula toma:
  - $k$ : número de parámetros del modelo.
  - $L$  verosimilitud máxima del modelo.

$$AIC = 2k - 2 \ln(L)$$

- **BIC:** similar a AIC, pero con una penalización más fuerte para modelos complejos. Valores **más bajos** indican un mejor modelo. Su fórmula toma:
  - $k$ : número de parámetros del modelo.
  - $n$  número de datos del modelo.
  - $L$  verosimilitud máxima del modelo.

$$BIC = k \ln(n) - 2 \ln(L)$$

# Análisis de parámetros

## Parámetros y métricas

### Métricas

- **Silhouette Score:** separación entre clusters considerando la cohesión dentro de cada cluster y la distancia entre clusters diferentes. Valores **más altos** indican clusters bien definidos. Su fórmula toma:

- $a(i)$  es la distancia media entre el punto  $i$  y otros puntos del mismo cluster.
  - $b(i)$  es la distancia media entre el punto  $i$  y los puntos del cluster más cercano.

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

- **Calinski-Harabasz Index:** relación entre la varianza intra-cluster e inter-cluster. Valores **más altos** indican mejor clustering.

- $k$  es el número de clusters.
  - $n$  es el número total de puntos.
  - $B_k$  es la matriz de dispersión entre clusters.
  - $W_k$  es la matriz de dispersión dentro de los clusters.

$$CH = \frac{\text{Tr}(B_k)/(k-1)}{\text{Tr}(W_k)/(n-k)}$$

# Clustering

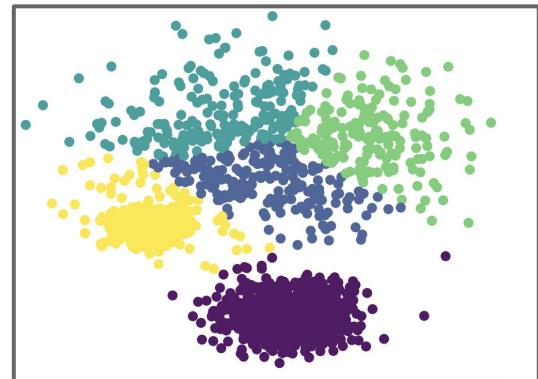
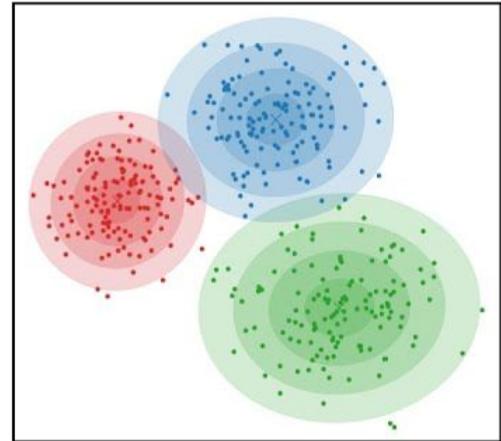
## GMM vs KMeans

# Clustering

## GMM vs KMeans

### Similitudes

- **Ambos son algoritmos de clustering:** Se utilizan para agrupar datos sin etiquetas en clusters.
- **Ambos requieren especificar el número de clusters:** En ambos métodos, es necesario definir **K** (en K-Means) o **n\_components** (en GMM) de antemano.
- **Utilizan iteraciones para converger:** Ambos se basan en un proceso iterativo hasta alcanzar una solución estable.

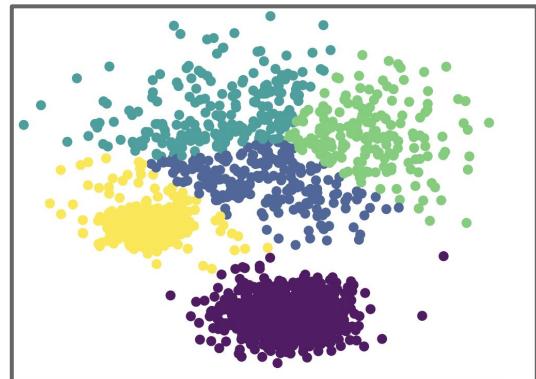
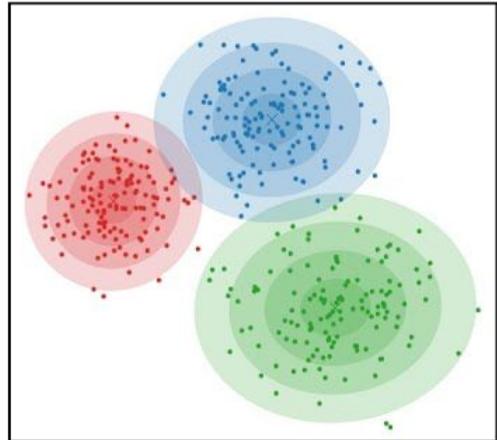


# Clustering

## GMM vs KMeans

### Diferencias

- **Tipo de modelo:** *hard clustering* vs *soft clustering*
- **Forma de los clusters:** Esféricos (radio fijo basado en distancia euclíadiana) para KMeans, mientras que GMM puede capturar formas elípticas y distribuciones más complejas
- **Método de asignación de puntos:** distancia mínima al centroide (KMeans) vs probabilidad de pertenencia basada en combinación de gaussianas
- **Sensibilidad a outliers:** GMM es menos sensible a outliers
- **Complejidad computacional:** GMM más costoso
- **Optimización:** mínimos cuadrados vs EM

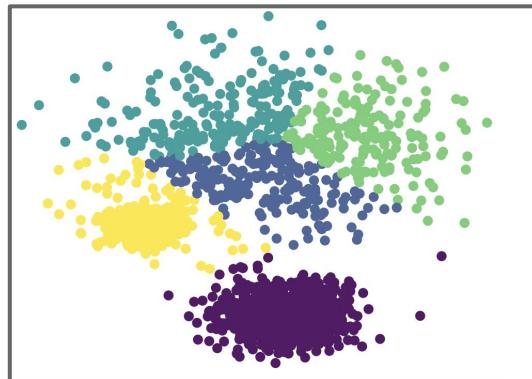
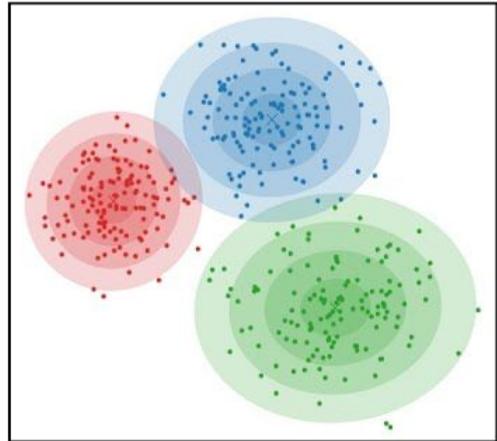


# Clustering

## GMM vs KMeans

### Usos

- **Usar K-Means** cuando los datos presentan clusters bien separados y aproximadamente esféricos. Es rápido y eficiente en grandes conjuntos de datos.
- **Usar GMM** cuando los clusters pueden tener formas elípticas o solapadas, o cuando se necesita una representación probabilística del clustering.

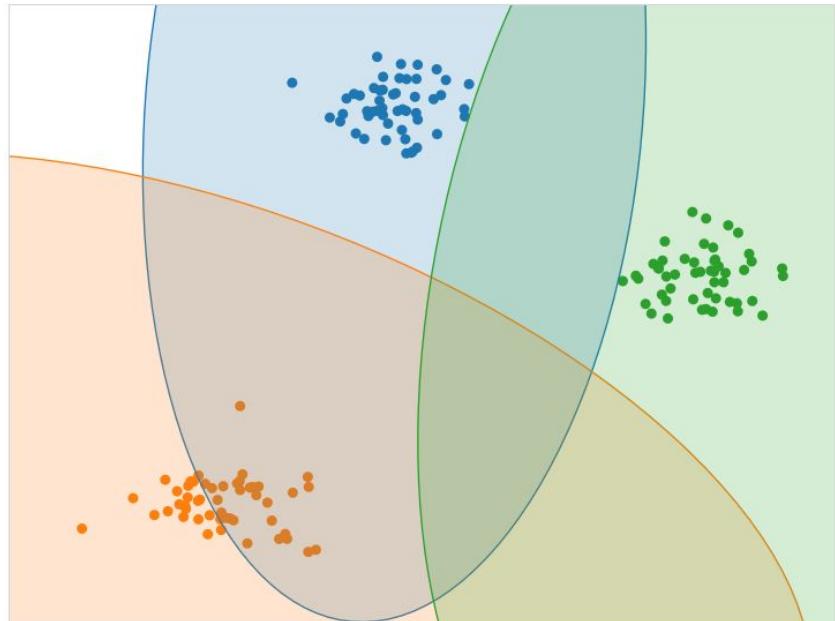


# Clustering

## Dataset

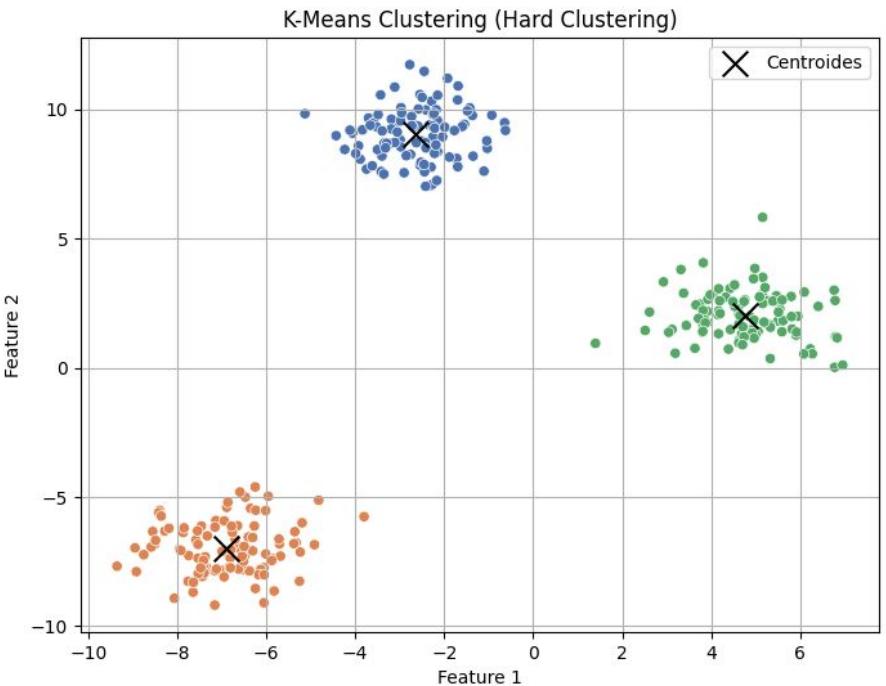
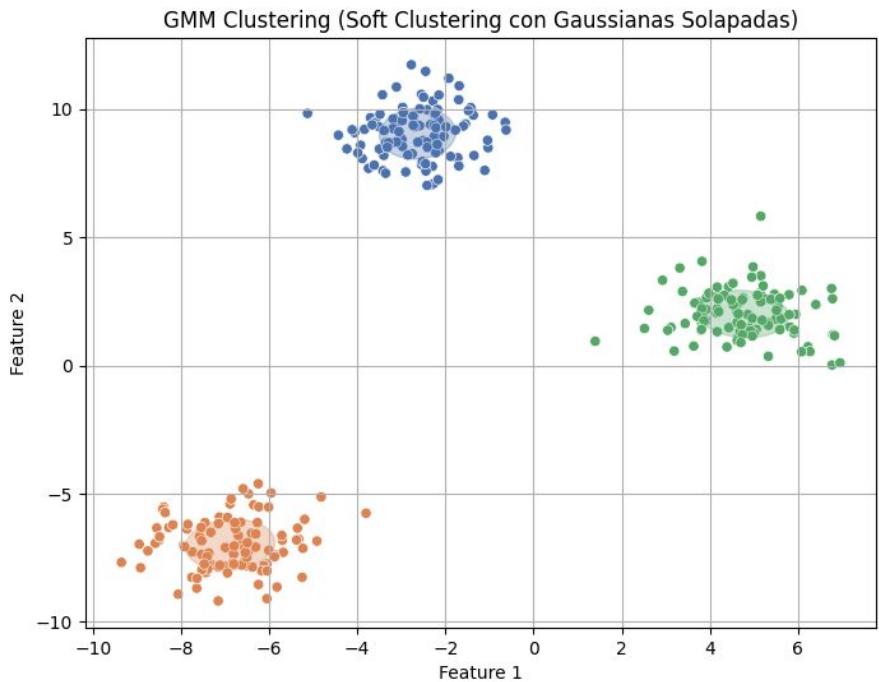
Se utiliza la librería **Scikit Learn** con la función **make\_blobs** que recibe:

- **n\_samples**: número de muestras.
- **centers**: cómo se van a distribuir esas muestras (un número).
- **n\_features**: características por muestra.
- **random\_state**: semilla (default 42).



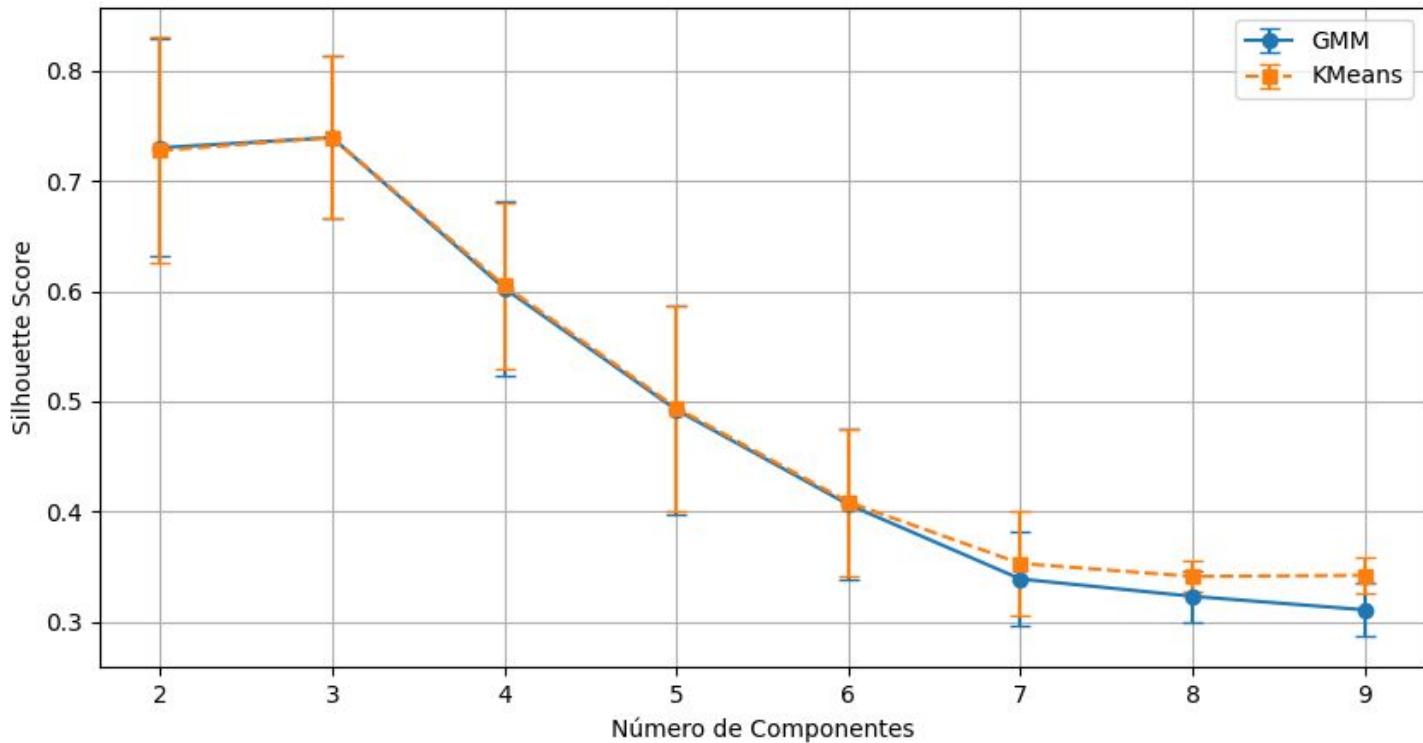
# Clustering

## GMM vs KMeans



# Clustering

## GMM vs KMeans: Silhouette Score



N\_samples: 300

N\_features: 2

Promedio 10 iteraciones

# Clustering

## Clusters Solapados

N\_components: 3

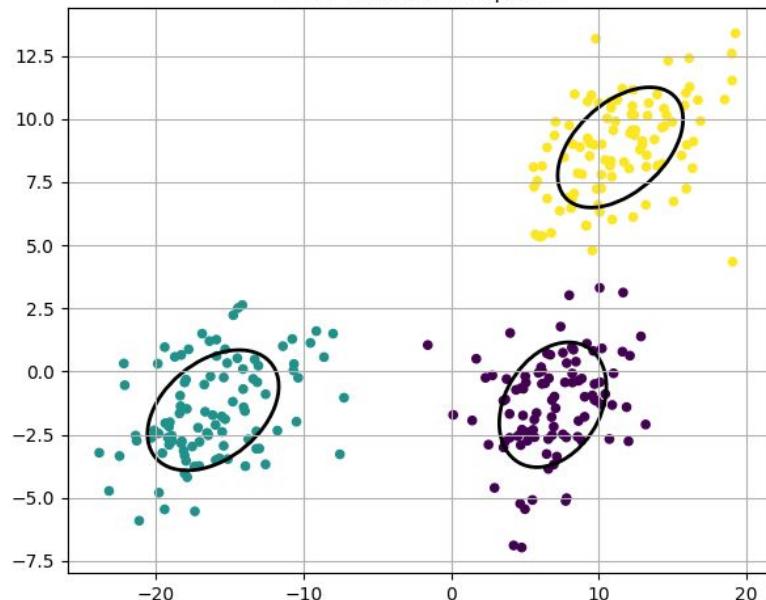
Random\_state: 42

N\_samples: 300

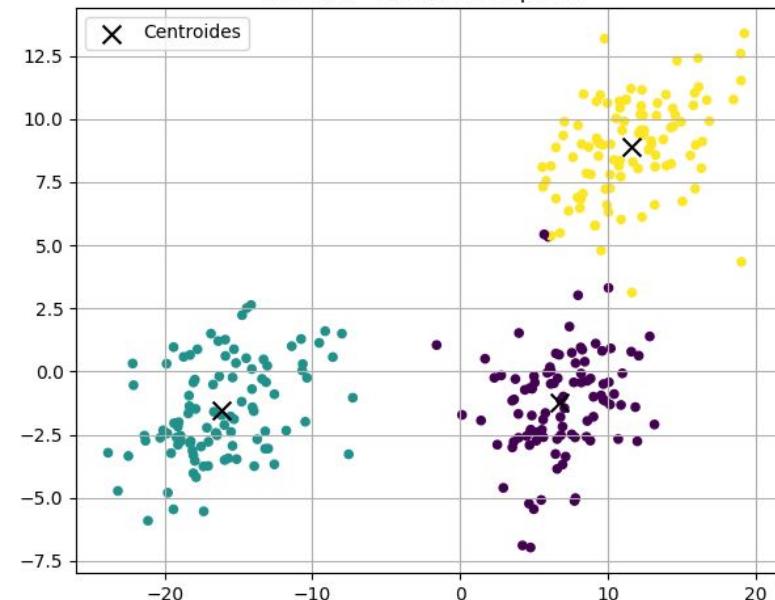
N\_features: 2

Covariance\_type: full

GMM - Clusters Solapados

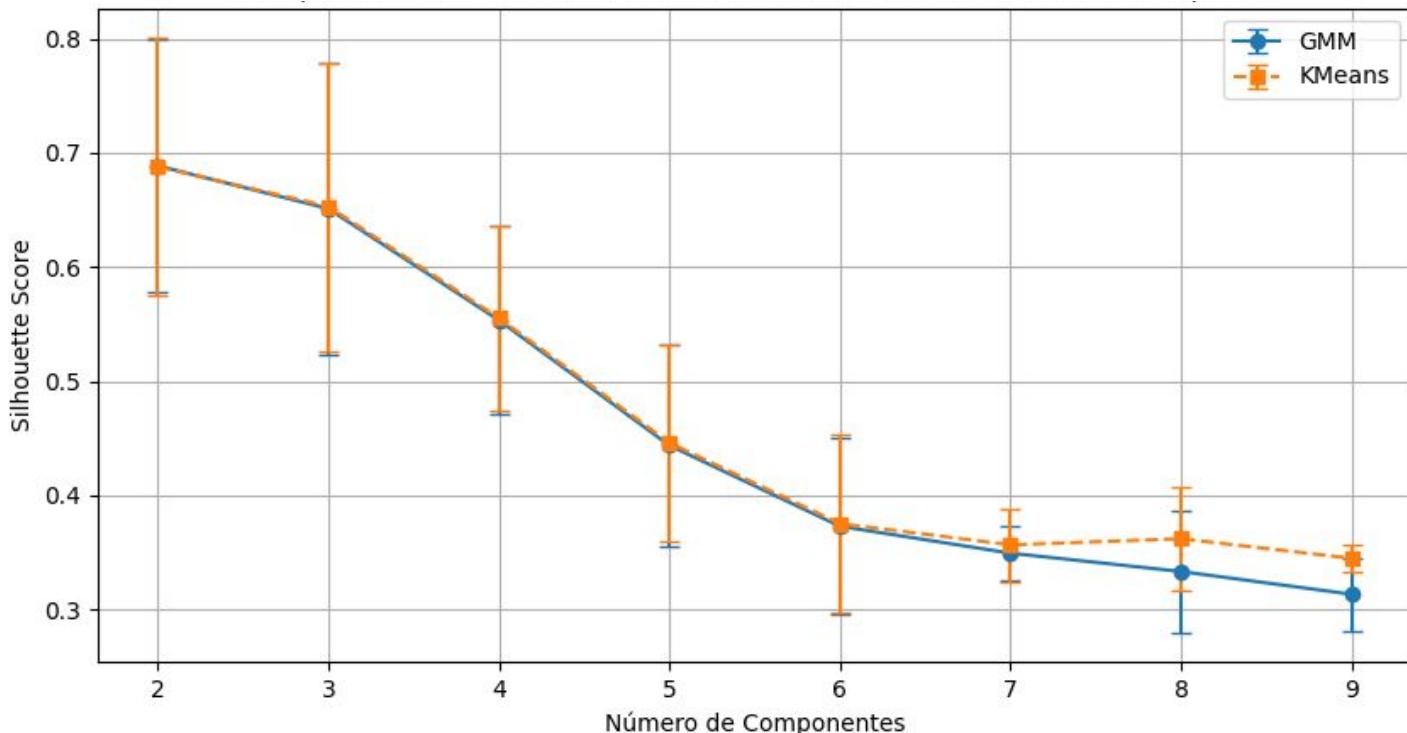


K-Means - Clusters Solapados



# Clustering

## Cluster Solapados: Silhouette Score



N\_components: 3

Random\_state: 42

N\_samples: 300

N\_features: 2

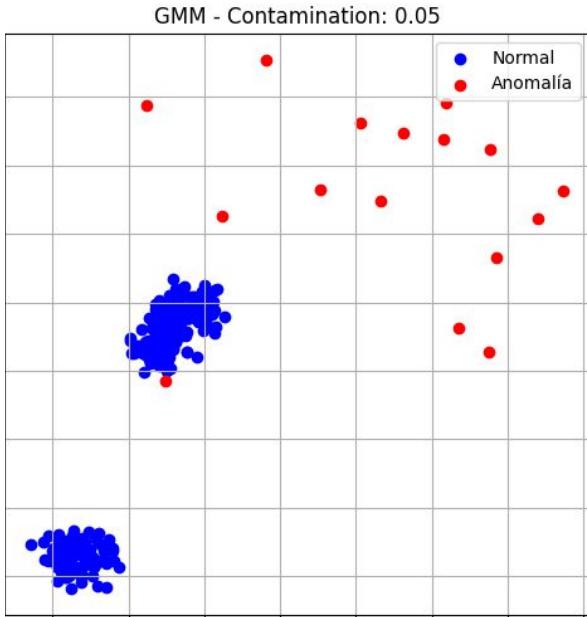
Covariance\_type: full

# GMM vs Isolation Forest

## Detección de anomalías

# Detección de anomalías

## GMM



GMM se usa para **detección de anomalías** para calcular la probabilidad de pertenencia de cada punto a la distribución aprendida.

- Si la probabilidad de un punto es **muy baja**, se considera una anomalía.
- Se define un umbral de **contaminación** que determina qué porcentaje de los datos será marcado como anómalo.
- Es útil en datos donde las anomalías se distinguen por su baja densidad relativa.

N\_components: 3 Random\_state: 42 N\_samples: 300

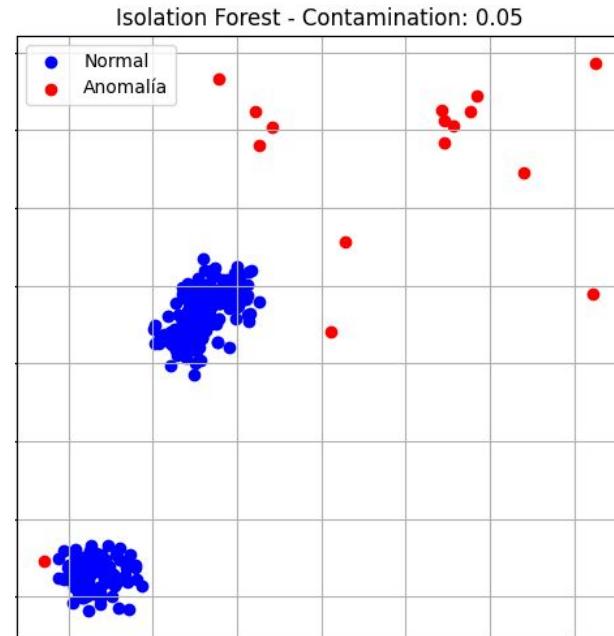
N\_features: 2 Covariance\_type: full

# Detección de anomalías

## Isolation Forest

Algoritmo basado en árboles de decisión que **aísla** anomalías.

- Construye múltiples árboles aleatorios donde los datos se dividen de manera recursiva.
- Los puntos que requieren **menos divisiones** para ser aislados son considerados anomalías.
- Funciona bien en **datasets de alta dimensión** y es más eficiente computacionalmente que GMM, además de ser menos sensible a la distribución de los datos



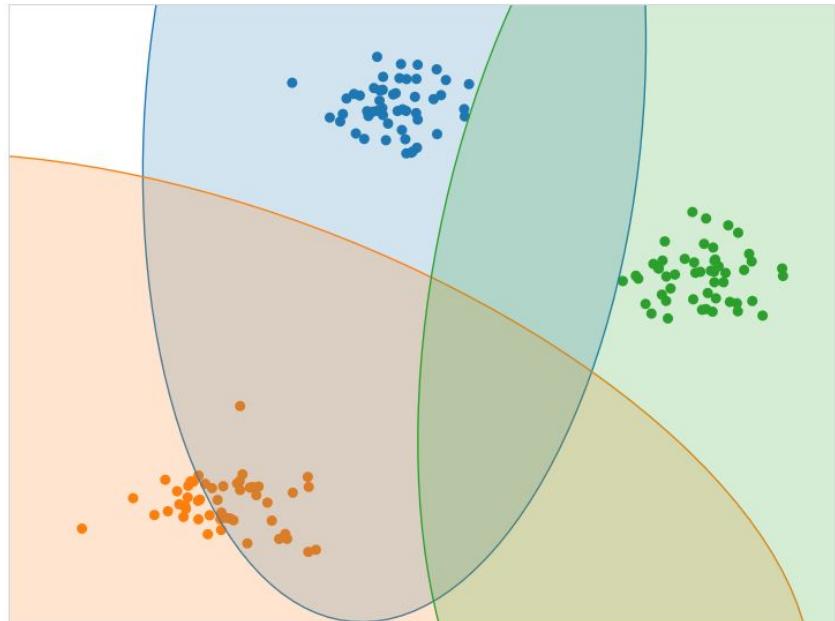
Random\_state: 42 N\_samples: 300  
N\_features: 2

# Clustering

## Dataset

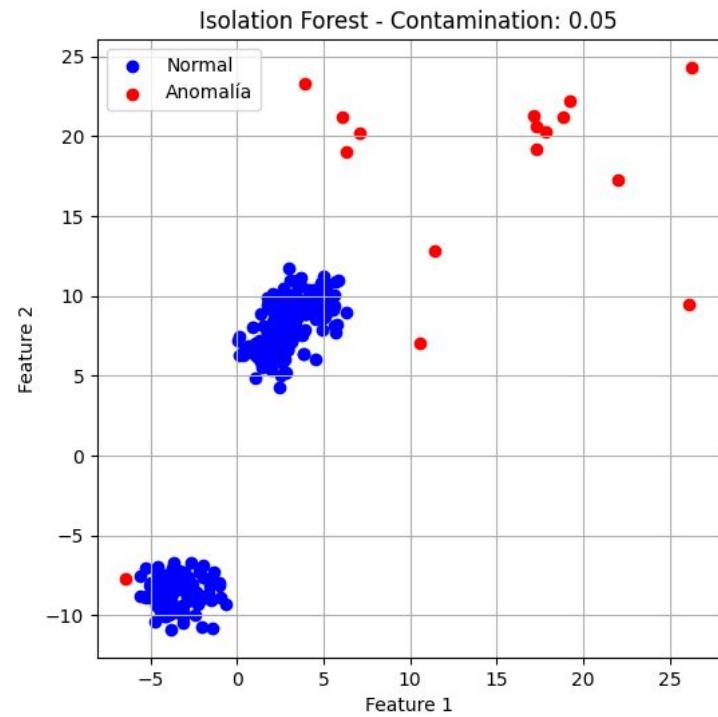
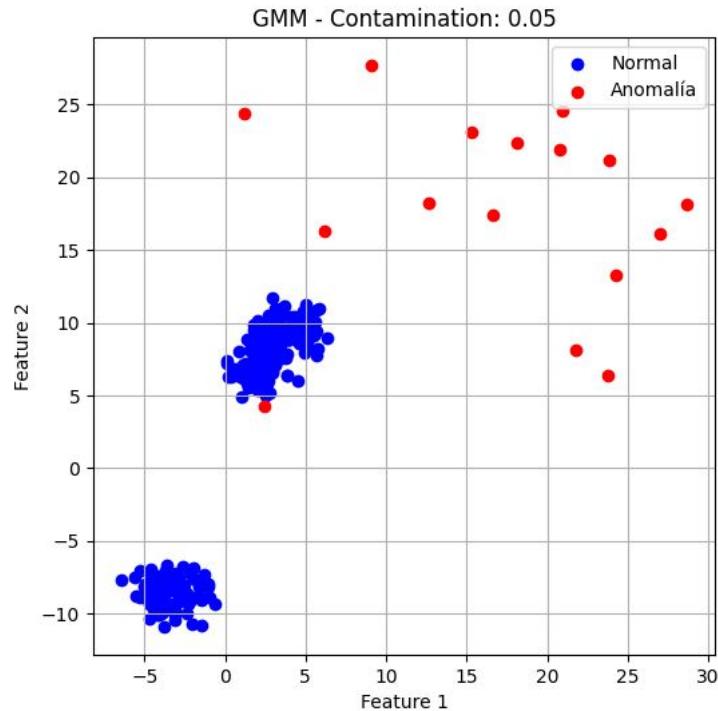
Se utiliza la librería **Scikit Learn** con la función **make\_blobs** que recibe:

- **n\_samples**: número de muestras.
- **centers**: cómo se van a distribuir esas muestras (un número).
- **n\_features**: características por muestra.
- **random\_state**: semilla (default 42).



# Detección de anomalías

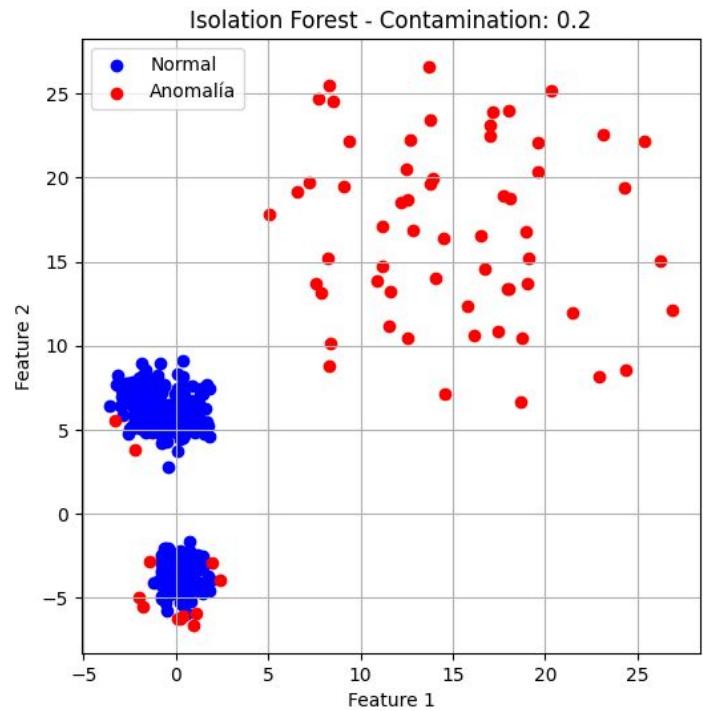
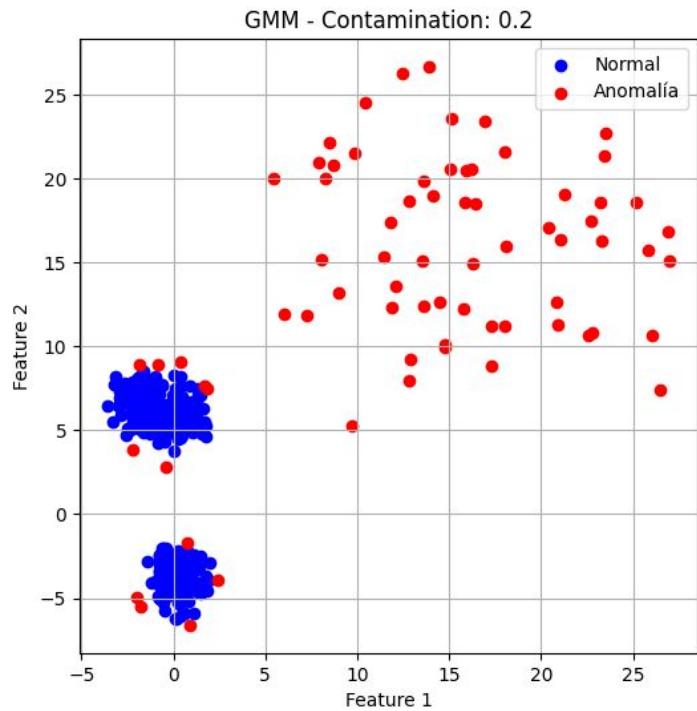
## GMM vs Isolation Forest: Contaminación Baja



# Detección de anomalías

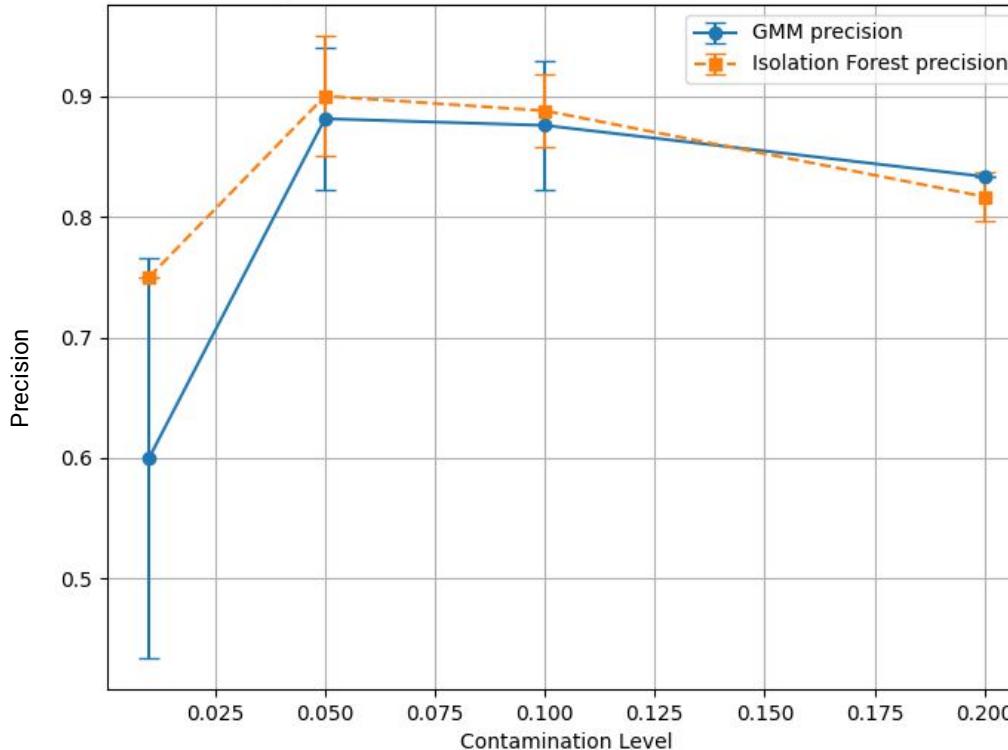
## GMM vs Isolation Forest: Contaminación Alta

N\_components: 3  
Random\_state: 42  
N\_samples: 300  
N\_features: 2  
Covariance\_type: full



# Detección de anomalías

## GMM vs Isolation Forest: Precisión



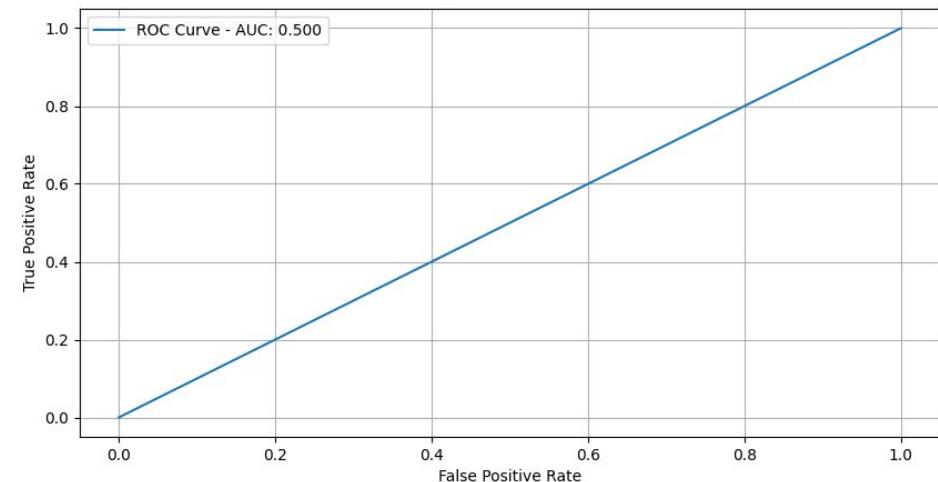
N\_samples: 300

N\_features: 2

# Detección de anomalías

## GMM vs Isolation Forest

GMM ROC

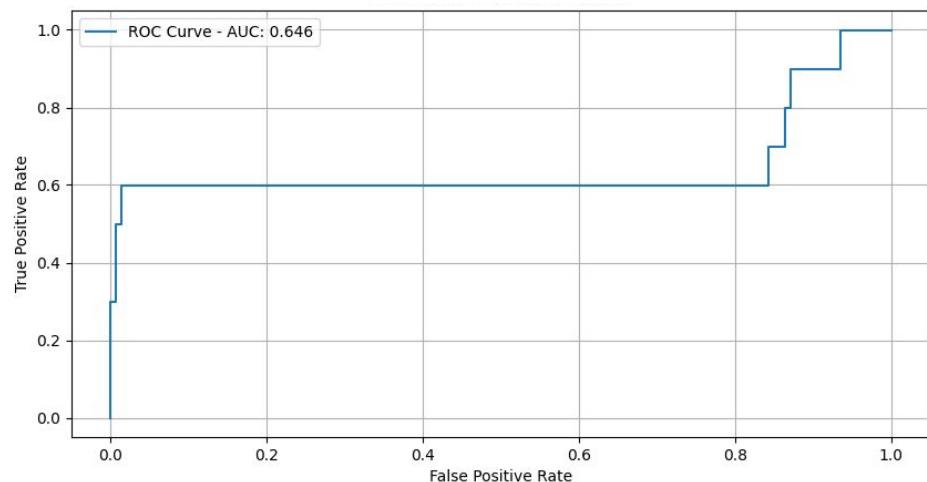


N\_components: 3    Random\_state: 42    N\_samples: 300

N\_features: 2

Covariance\_type: full

Isolation Forest ROC



Random\_state: 42    N\_samples: 300    N\_features: 2

# GMM

# Análisis de densidad

# Análisis de Densidad

## GMM

Un **análisis de densidad** permite estimar la distribución de probabilidad subyacente de los datos del dataset. Al ajustar un modelo GMM, podemos:

- Identificar regiones de alta y baja densidad, lo que ayuda a entender la estructura del dataset.
- Detectar posibles subgrupos dentro de cada grupo.
- Modelar la distribución de cada feature usando gaussianas.

# Análisis de Densidad

## Análisis de Dataset: Iris



Información sobre tres especies de flores del género *Iris*:

1. **Iris setosa**
2. **Iris versicolor**
3. **Iris virginica**

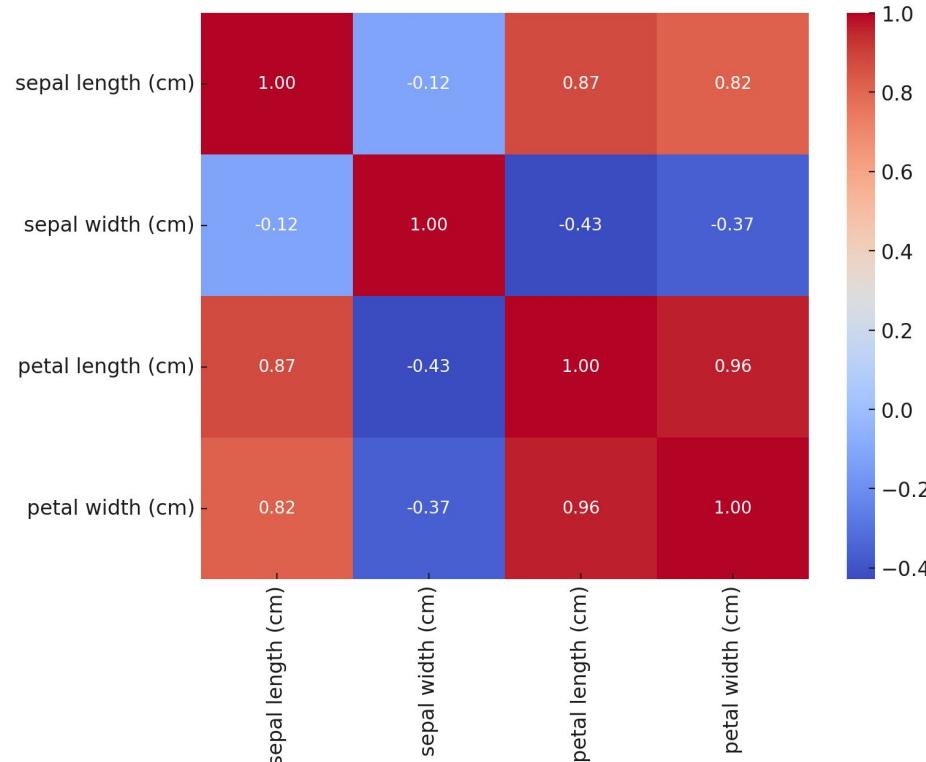
Cada observación en el dataset representa una flor y tiene **cuatro características**:

- **Largo del sépalo** (*sepal length*, en cm)
- **Ancho del sépalo** (*sepal width*, en cm)
- **Largo del pétalo** (*petal length*, en cm)
- **Ancho del pétalo** (*petal width*, en cm)

# Análisis de Densidad

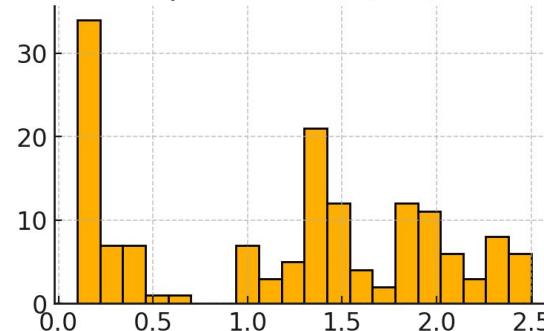
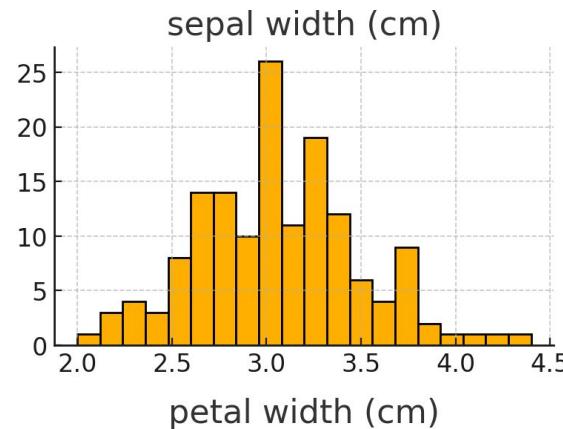
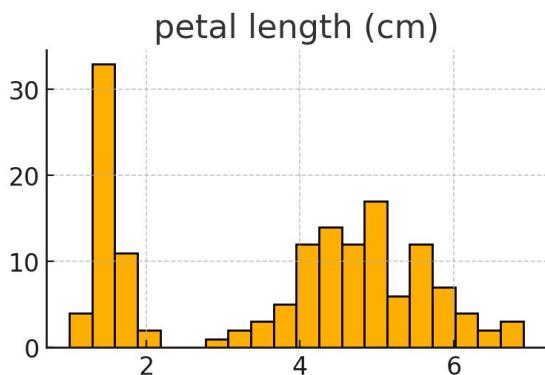
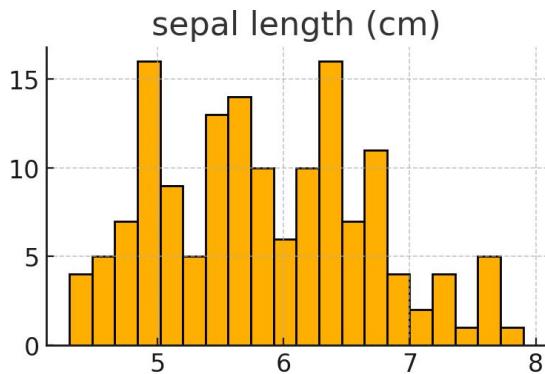
## Análisis de Dataset: Iris

Matriz de Correlación



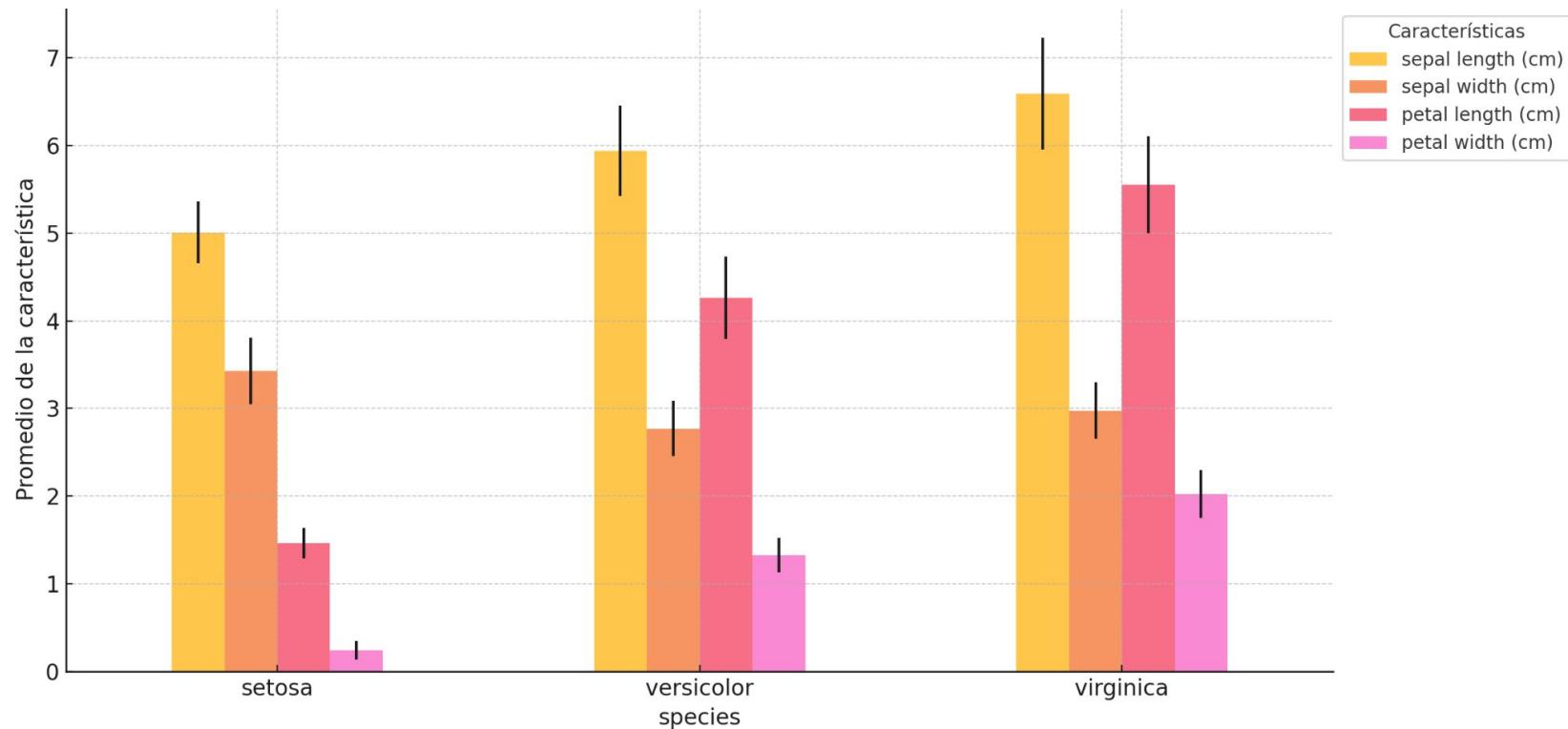
# Análisis de Densidad

## Análisis de Dataset: Iris



# Análisis de Densidad

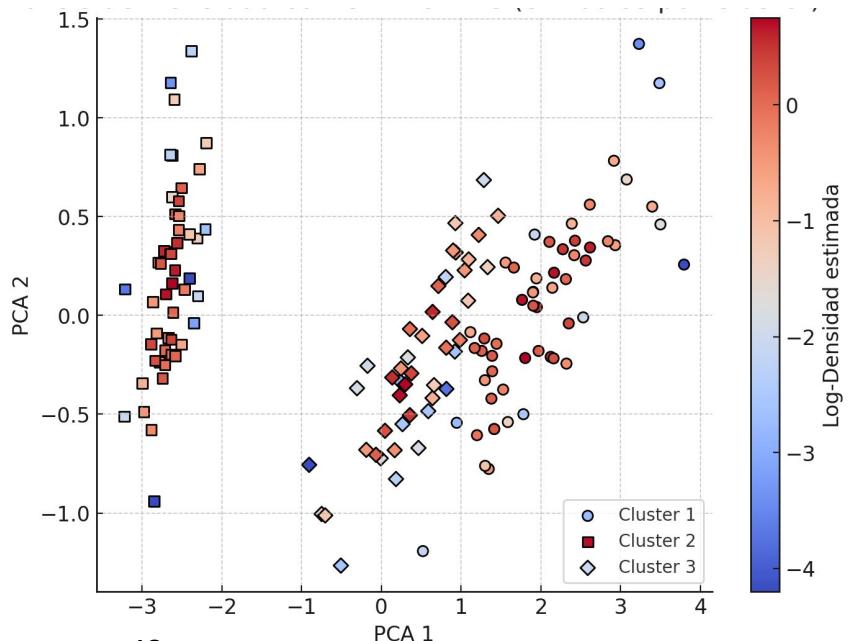
## Análisis de Dataset: Iris



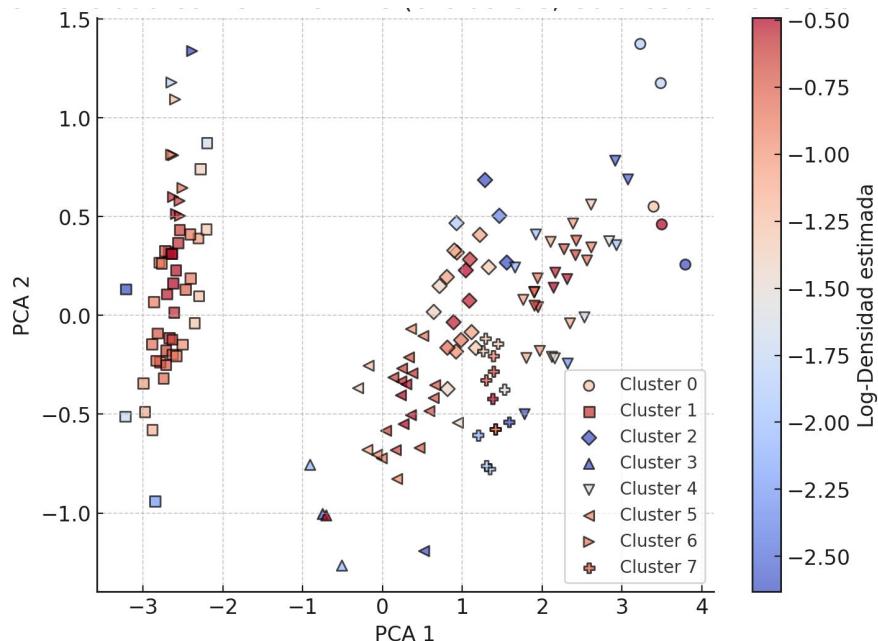
# Análisis de Densidad

GMM

3 Clusters



8 Clusters



# Análisis de Densidad

## Análisis de Dataset: California Housing

Proviene del censo de 1990 y contiene información sobre diferentes zonas residenciales en California.

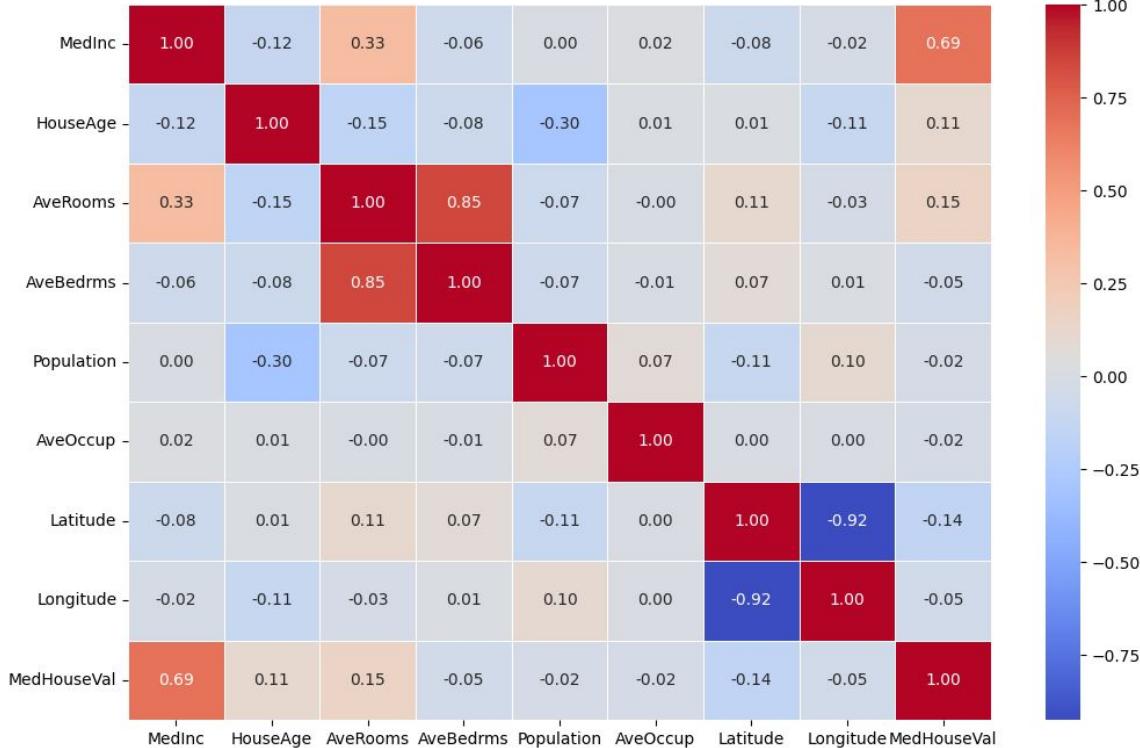
- **MedInc**: Ingreso medio en la zona (en decenas de miles de dólares).
- **HouseAge**: Edad promedio de las viviendas en la zona.
- **AveRooms**: Número promedio de habitaciones por vivienda.
- **AveBedrms**: Número promedio de dormitorios por vivienda.
- **Population**: Población total en la zona.
- **AveOccup**: Promedio de ocupantes por vivienda.
- **Latitude**: Latitud de la ubicación.
- **Longitude**: Longitud de la ubicación.
- **MedHouseVal (variable objetivo)**: Valor medio de las viviendas (en cientos de miles de dólares).



# Análisis de Densidad

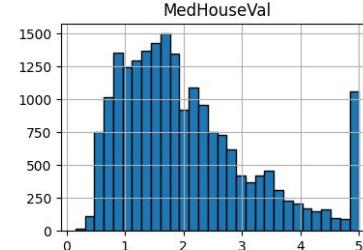
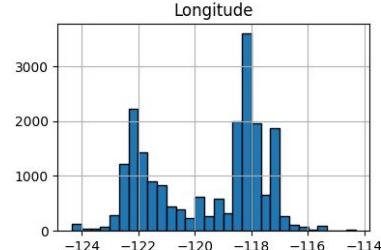
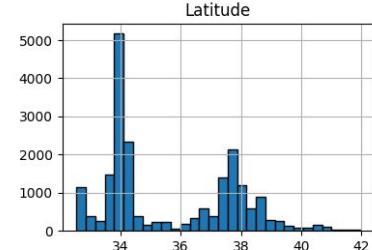
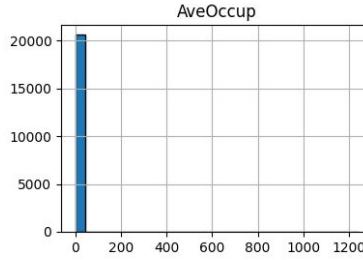
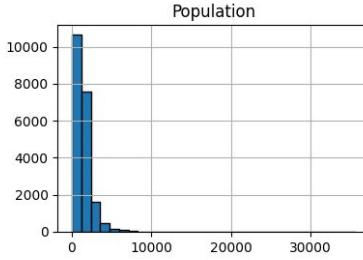
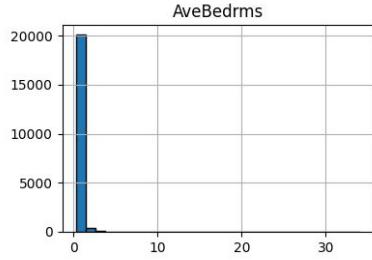
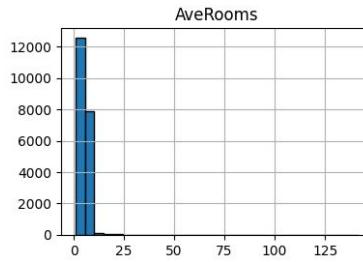
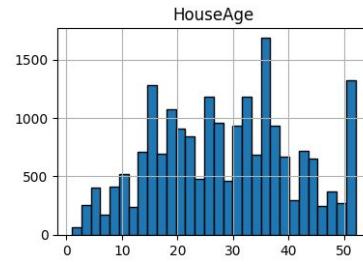
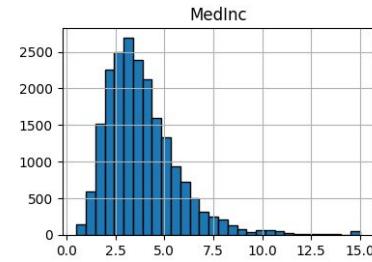
## Análisis de Dataset: California Housing

Matriz de Correlación



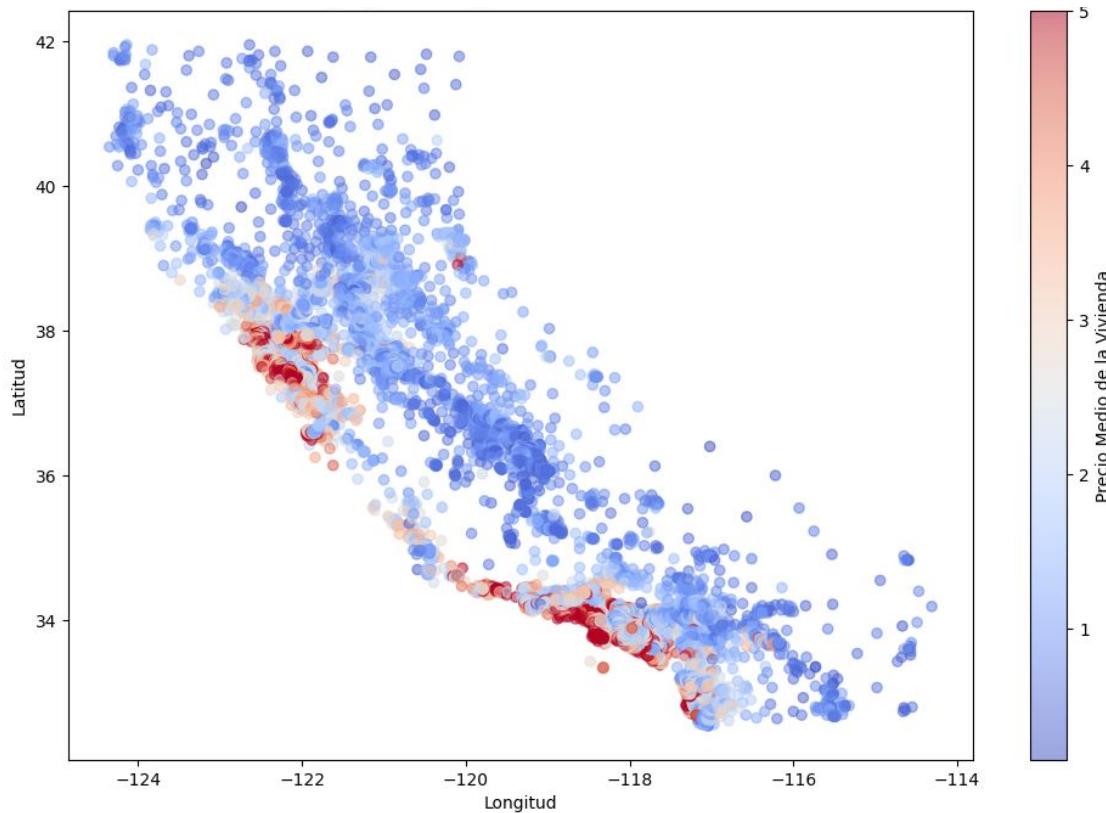
# Análisis de Densidad

## Análisis de Dataset: California Housing



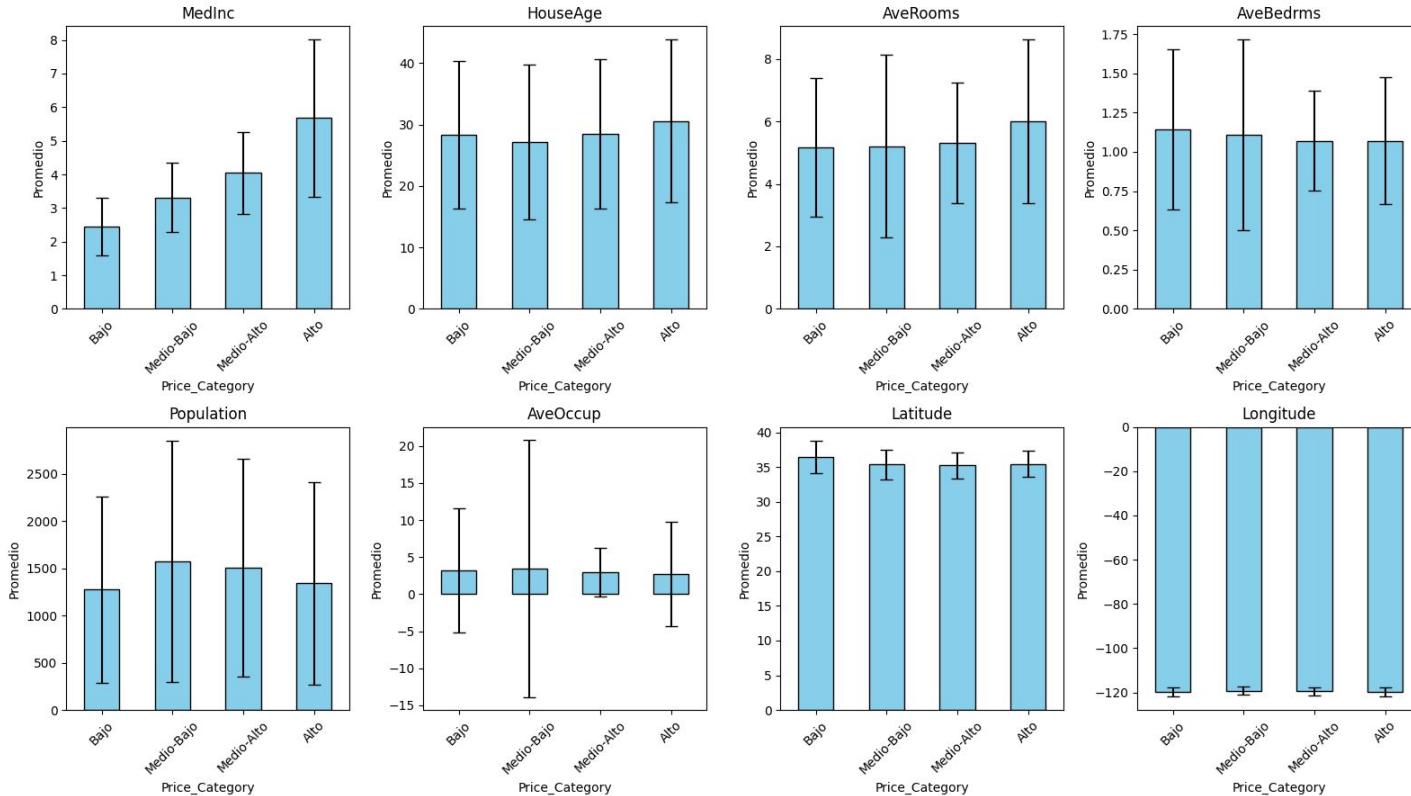
# Análisis de Densidad

## Análisis de Dataset: California Housing



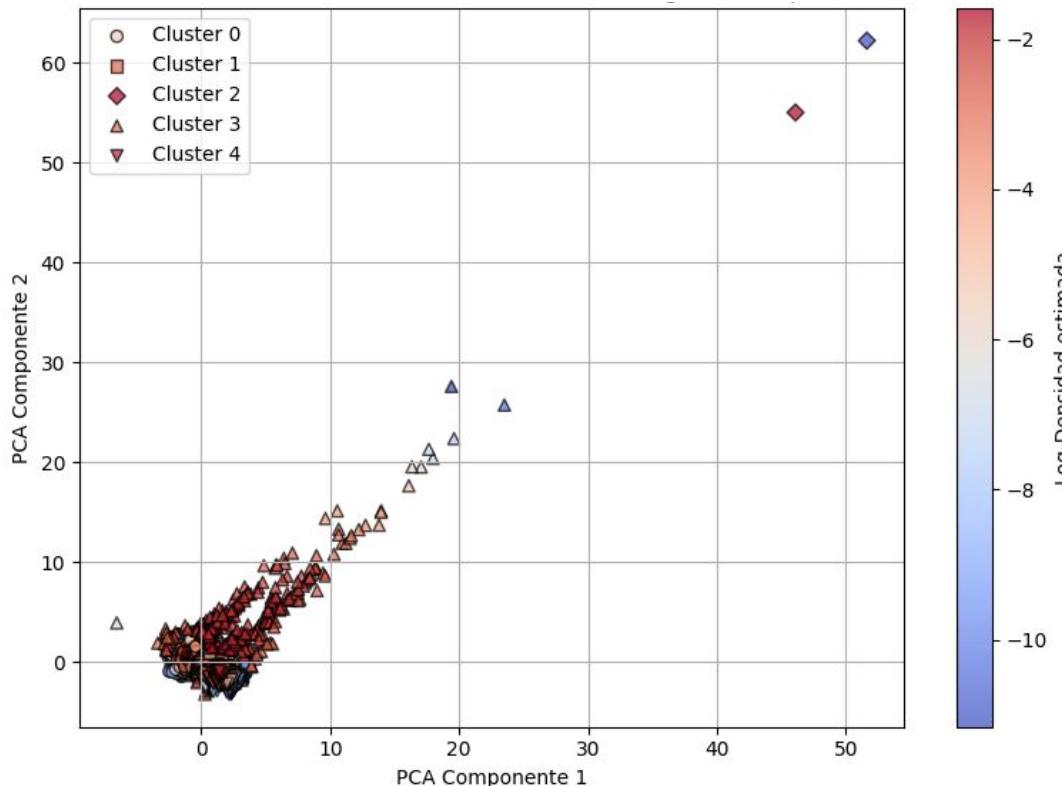
# Análisis de Densidad

## Análisis de Dataset: California Housing



# Análisis de Densidad

## GMM



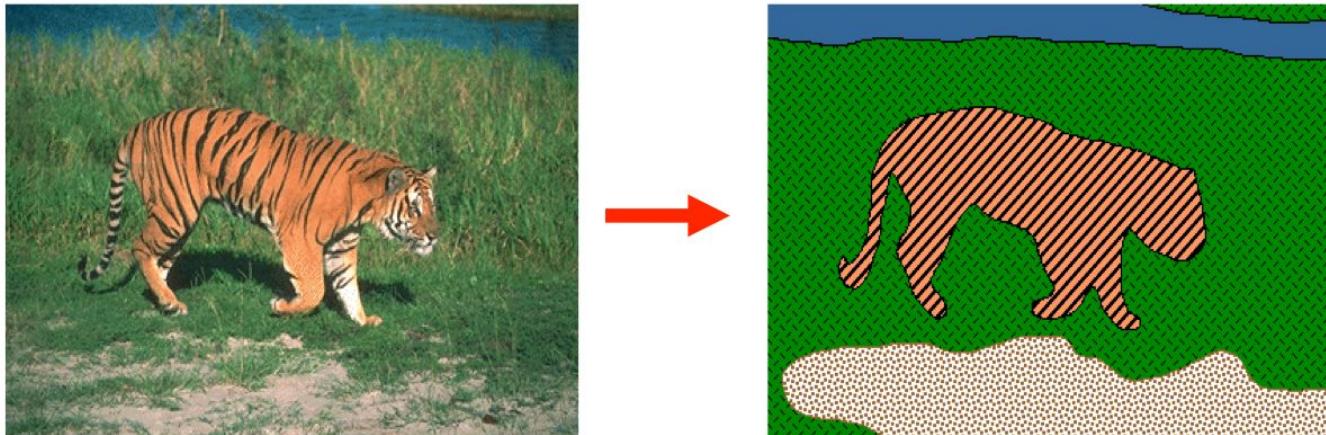
# GMM

# Segmentación de imágenes

# Segmentación de imágenes

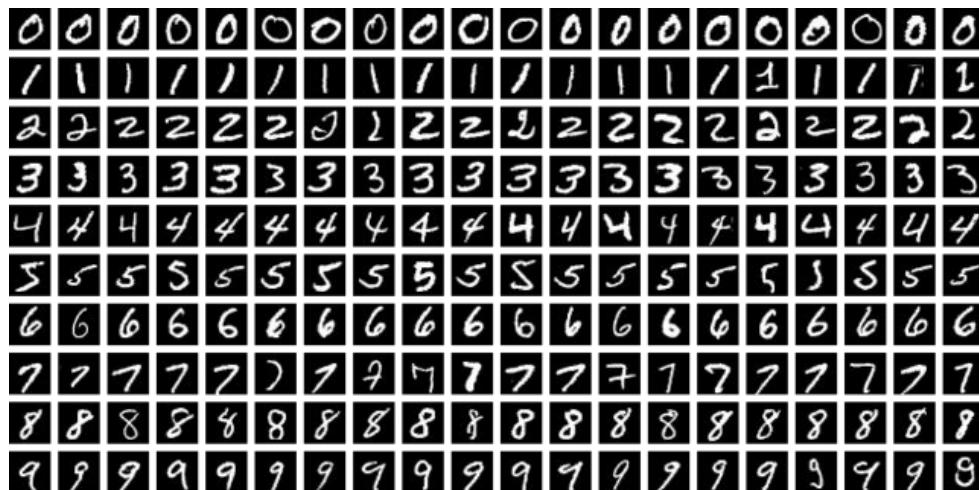
## Basada en Clustering

- Proceso de dividir una imagen en regiones o segmentos con características similares, como color, textura o intensidad.
- **Objetivo:** simplificar la representación de una imagen para facilitar su análisis o procesamiento posterior.



# Segmentación de imágenes

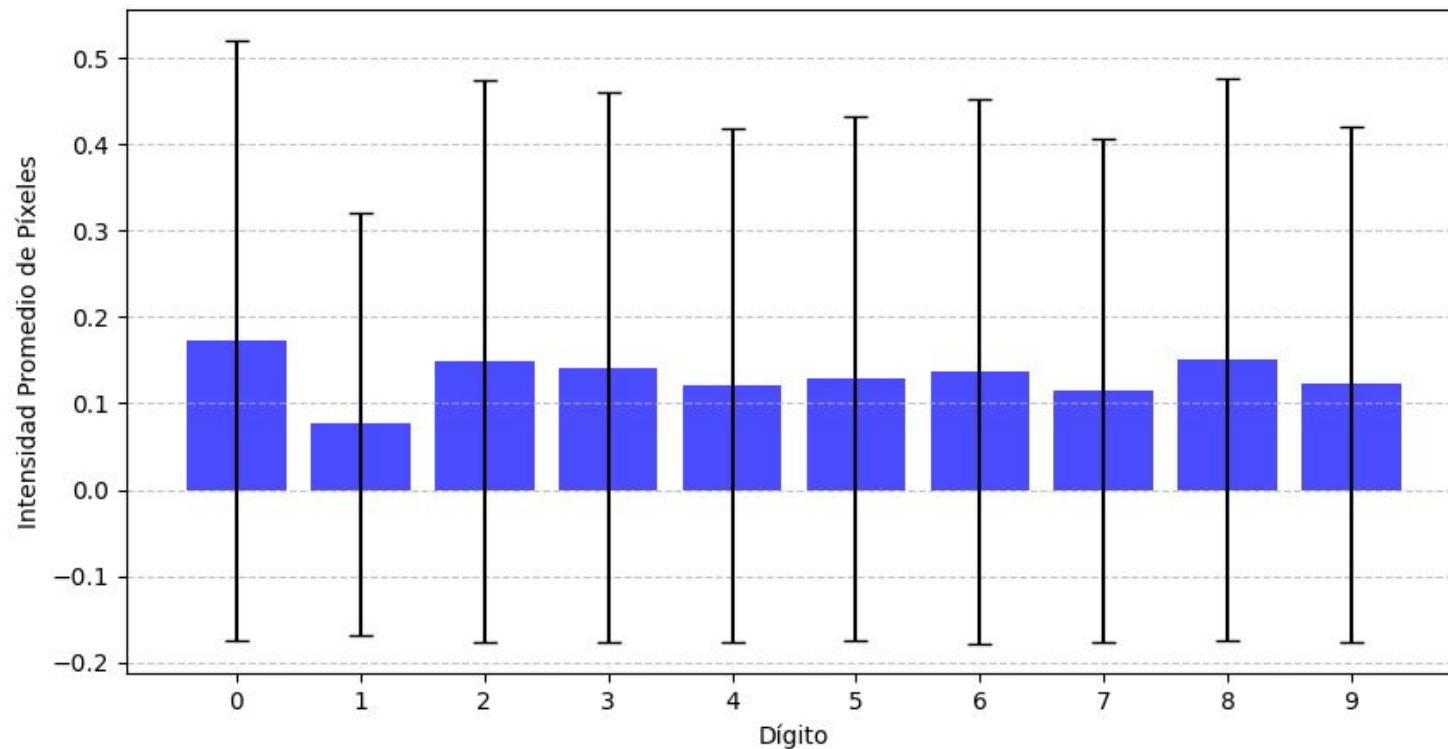
## Análisis del Dataset: MNIST



- Dígitos del 0 al 9 escritos a mano
- Imágenes en escala de grises de 28x28 píxeles
- Intensidades de píxeles que van de 0 (negro) a 255 (blanco)

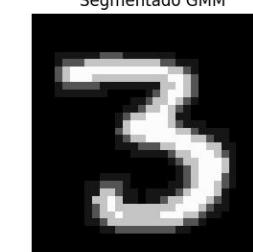
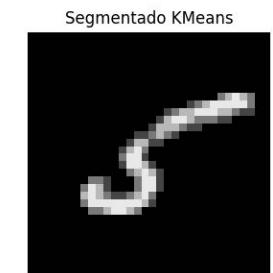
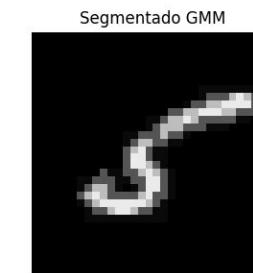
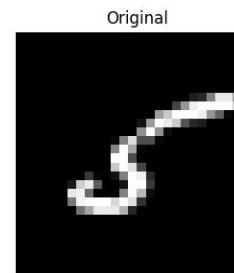
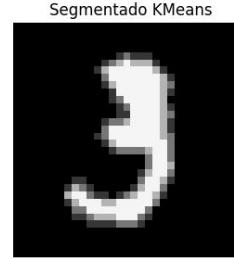
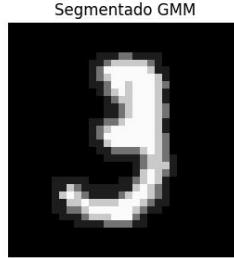
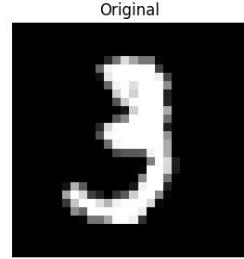
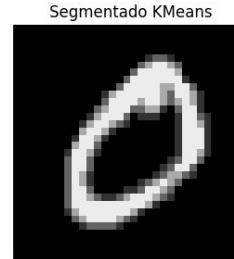
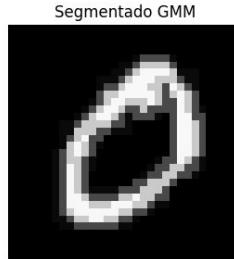
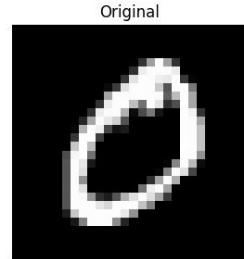
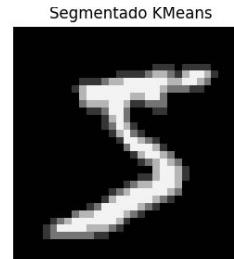
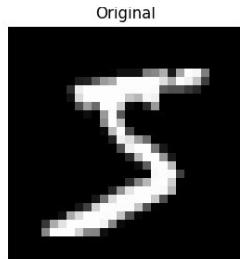
# Segmentación de imágenes

## Análisis del Dataset: MNIST



# Segmentación de imágenes

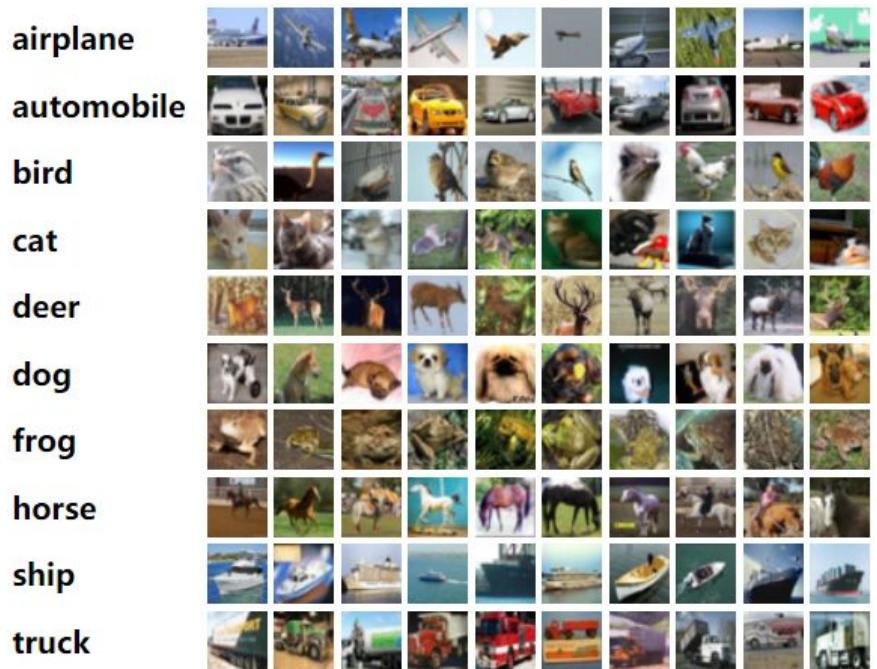
## MNIST



# Segmentación de imágenes

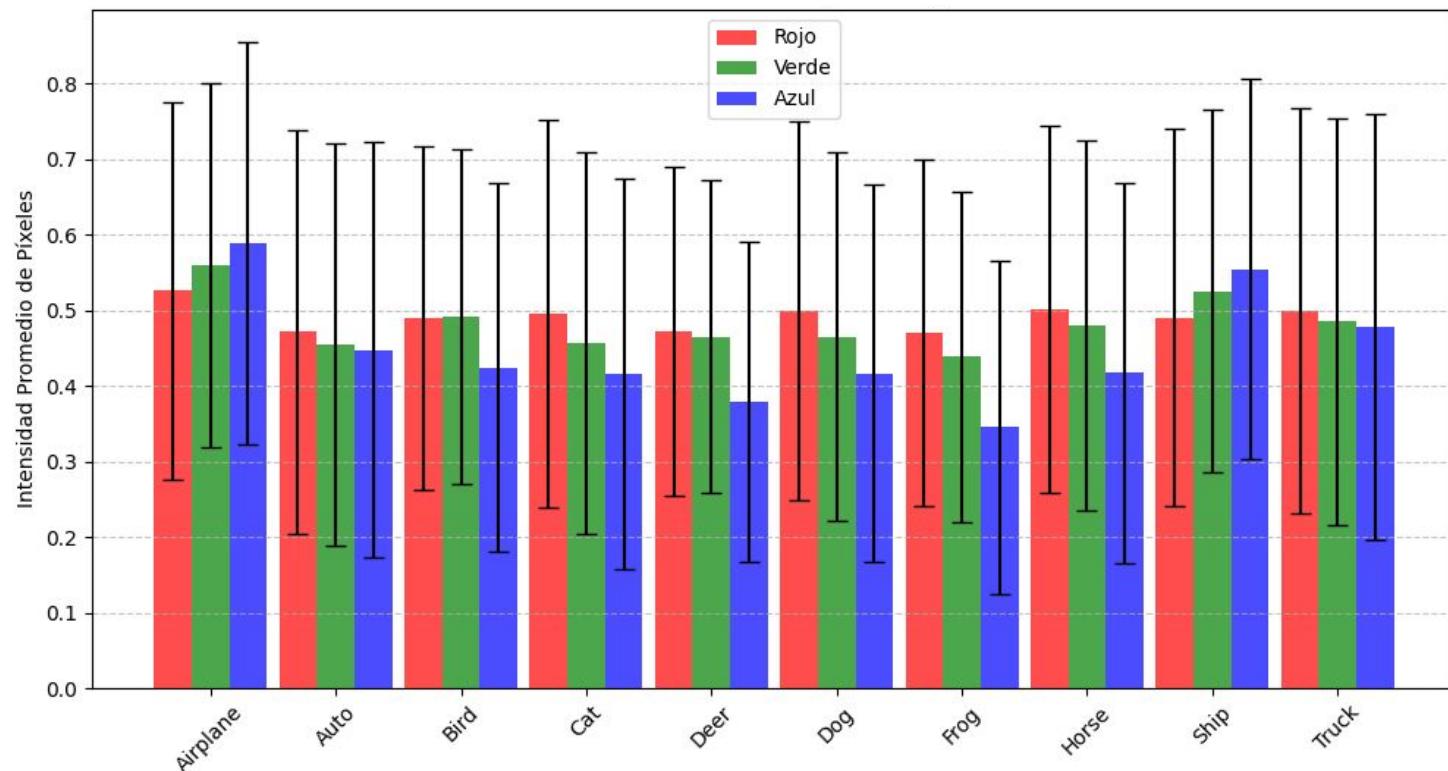
## Análisis del Dataset: CIFAR-10

- Imágenes de 10 clases diferentes: avión, automóvil, pájaro, gato, venado, perro, rana, caballo, barco y camión.
- Imágenes RGB de 32x32 píxeles



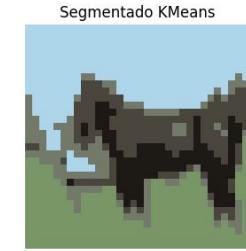
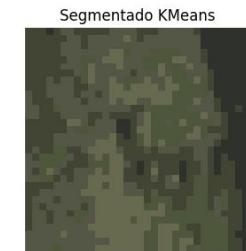
# Segmentación de imágenes

## Análisis del Dataset: CIFAR-10



# Segmentación de imágenes

## CIFAR-10



# GMM

# Procesamiento de audio

# Procesamiento de audios

## Objetivos y ventajas de GMM

GMM resulta útil para la clusterización de audio porque puede modelar distribuciones complejas en las señales, como los coeficientes cepstrales. Al tratar a los features (MFCC y Chromas) como un conjunto de variables aleatorias que siguen una distribución normal, GMM permite identificar patrones con precisión.

El objetivo de este análisis es utilizar GMM para la clusterización de audios en función de features extraídos y su comparación y vínculo con las emociones de estos.

- Mediante clusterización separar audios en emociones positivas y negativas.
- Mediante clusterización separar audios en audios felices, enojados o calmos.

# Análisis del dataset

## Etiquetas y datos del dataset

Conjunto de 1.440 audios de actores en formato: AA-BB-CC-DD-EE-FF-GG.wav

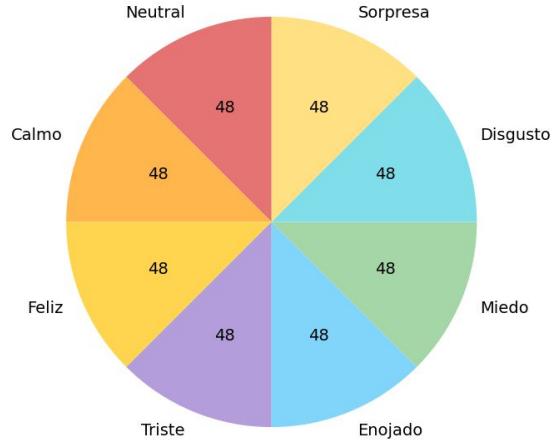
- **AA (Modo):** 01 = *audio y video completos*, 02 = *solo video*, 03 = *solo audio*
- **BB (Canal):** 01 = *habla*, 02 = *canto*
- **CC (Emoción):** 01=Neutral, 02=Calmado, 03=Feliz, 04=Triste, 05=Enojado, 06=Miedo, 07=Disgusto, 08=Sorpresa
- **DD (Intensidad):** 01 = *normal*, 02 = *fuerte*
- **EE (Frase):** 01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door".
- **FF (Repetición):** 01 = *primera repetición*, 02 = *segunda repetición*
- **GG (Actor):** 01 a 24 número del actor

Nosotros tomaremos un subset con: **03-01-CC-02-02-FF-GG.wav**

# Análisis del dataset

## Muestra de audios

El subset de **03-01-CC-02-02-FF-GG** se compone de 384 audios compuestos por:



Feliz



Enojado



Calmo



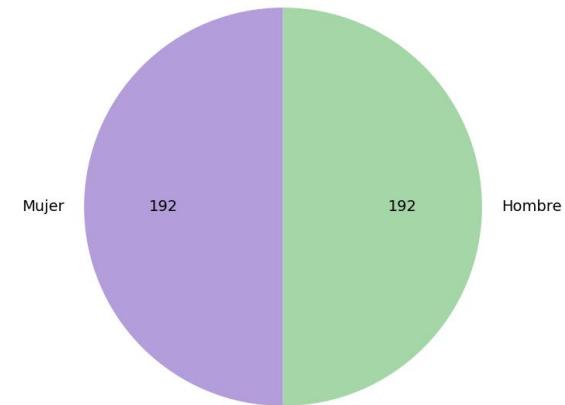
Sorprendido



Hombre



Mujer

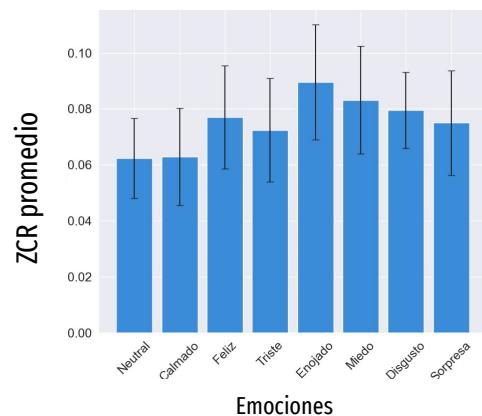
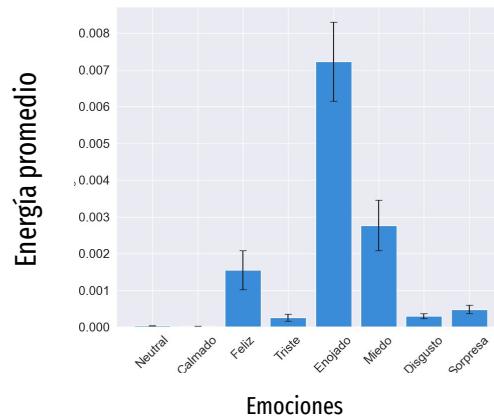
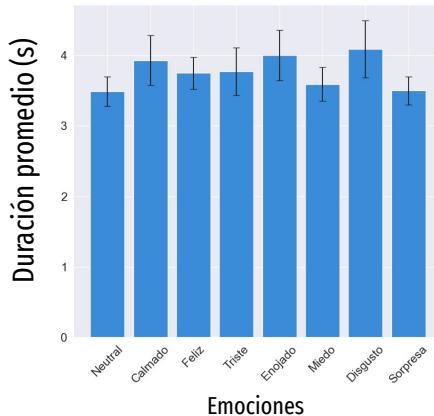


# Análisis del dataset

## Features relevantes a utilizar

De cada audio extraemos features para luego ingresarlos al modelo GMM:

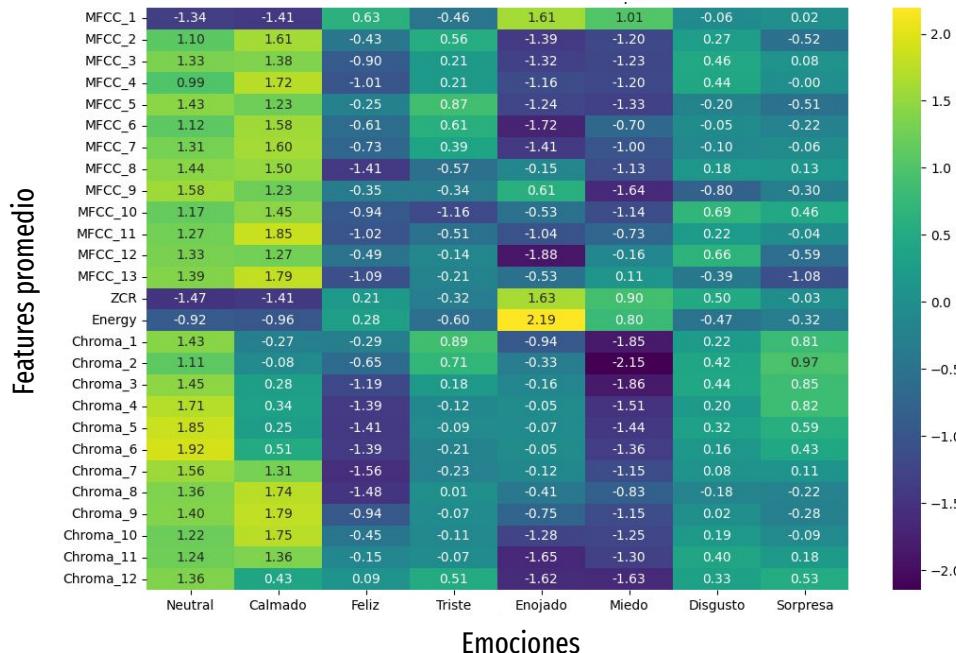
- Energía: Suma de los cuadrados de las amplitudes.
- ZCR: Número de veces que la señal cambia de signo, indicando la agudeza del sonido.
- Chroma(12): Energía de cada una de las 12 notas musicales (tonalidad) en la señal.
- MFCC(13): Características espectrales que describen la forma del espectro de la señal



# Análisis del dataset

## Etiquetas y datos del dataset

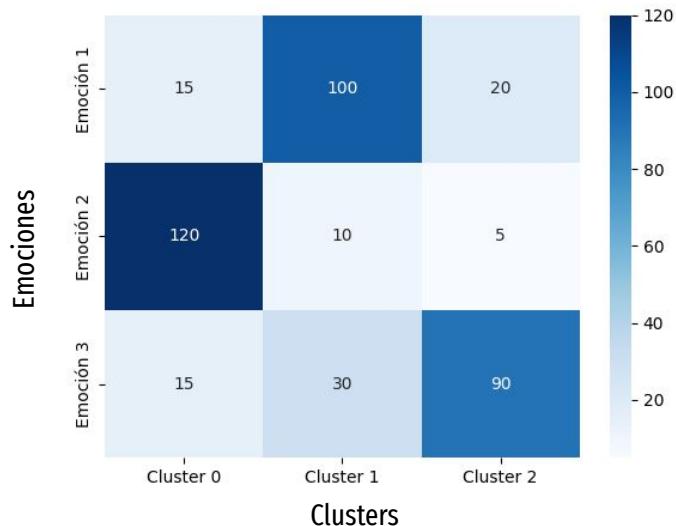
Con un análisis general de las features a extraer y calculamos el promedio obtenemos:



# Clusterización de emociones

## Métrica a utilizar

Nuestro objetivo es determinar cómo se están clusterizando las emociones, por lo que nuestra métrica tiene como objetivo asignarle una emoción a un cluster y calcular el accuracy de todo el conjunto. Siempre la asignación será la más favorable.



Asignamos emoción a su cluster más representativo

- Emoción 2 - Cluster 0
- Emoción 1 - Cluster 1
- Emoción 3 - Cluster 2

$$\text{Accuracy} = (120 + 100 + 90) / (135 \times 3) = 0.765$$

# **Clusterización de emociones: Positivas y Negativas**

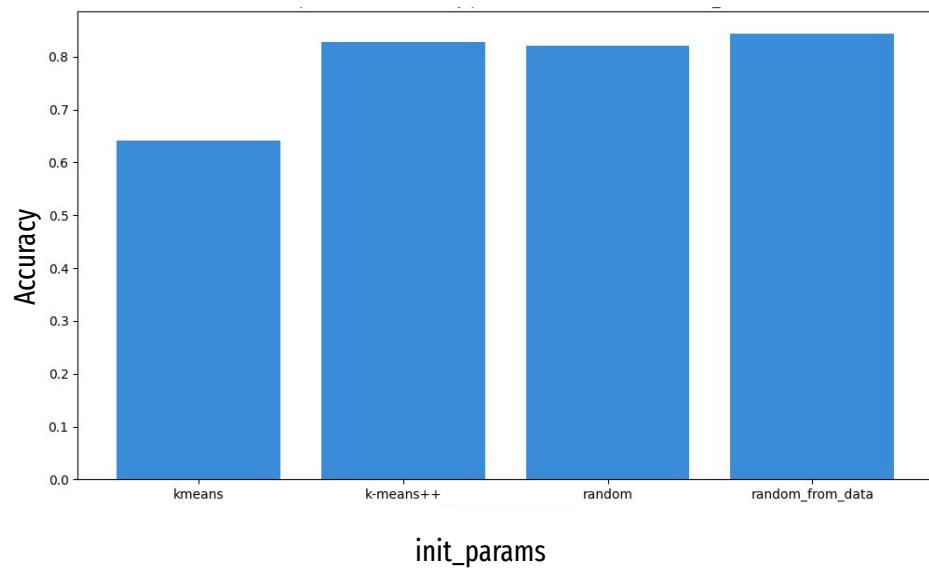
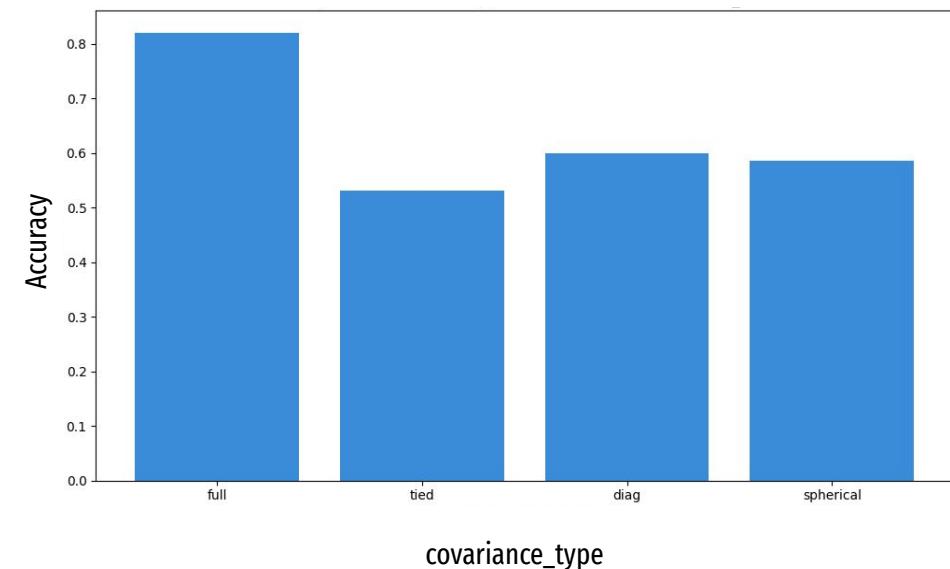
## **Procedimiento**

El proceso para clusterización de audios con emociones positivas y negativas es:

- Lectura de todos los audios del subset elegido.
- Extracción de 27 features (MFCCs, Chromas, ZCR, Energía).
- Estandarización de cada feature.
- Variación de parámetros según métrica definida con 2 clusters:
  - Positivas: Neutral - Calmo - Feliz - Sorpresa.
  - Negativas: Triste - Enojado - Miedo - Disgusto.
- Obtención de mejores parámetros y exposición de resultados.

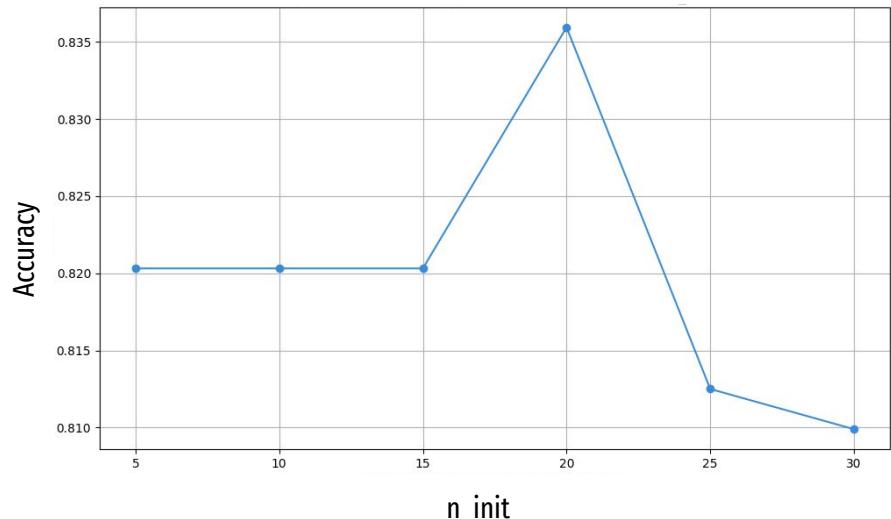
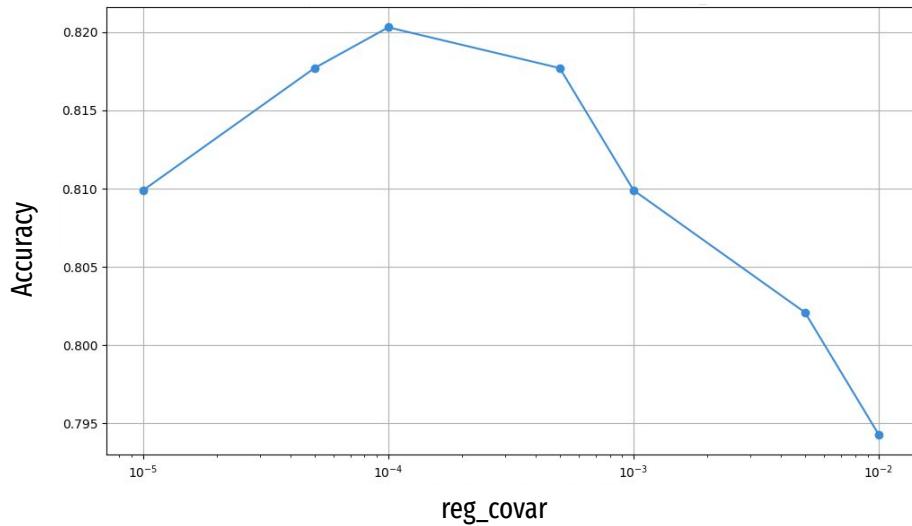
# Clusterización de emociones

## Variación de parámetros



# Clusterización de emociones

## Variación de parámetros



# Clusterización de emociones

## Resultados

Utilizando los parámetros obtenidos del análisis obtenemos la matriz correspondiente

### **Mejores parámetros:**

init\_params: random\_from\_data

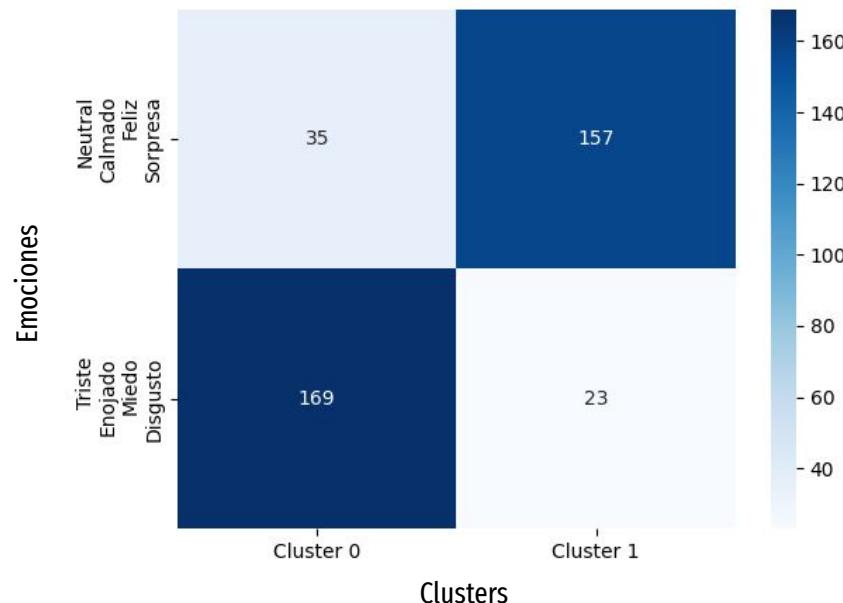
reg\_covar: 1e-4

covariance\_types: full

n\_init: 20

n\_components: 2

**Accuracy: 0.849**



# Clusterización de emociones

## Resultados

Si desagrupamos las categorías obtenemos la matriz siguiente por emoción

### **Mejores parámetros:**

init\_params: random\_from\_data

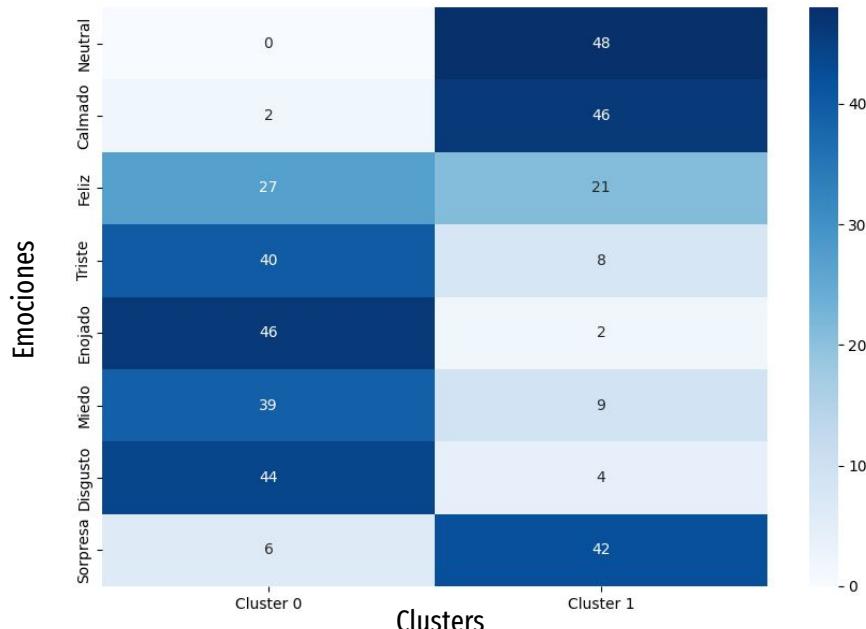
reg\_covar: 1e-4

covariance\_types: full

n\_init: 20

n\_components: 2

**Accuracy: 0.849**



# **Clusterización de emociones: Feliz - Calmo - Enojado**

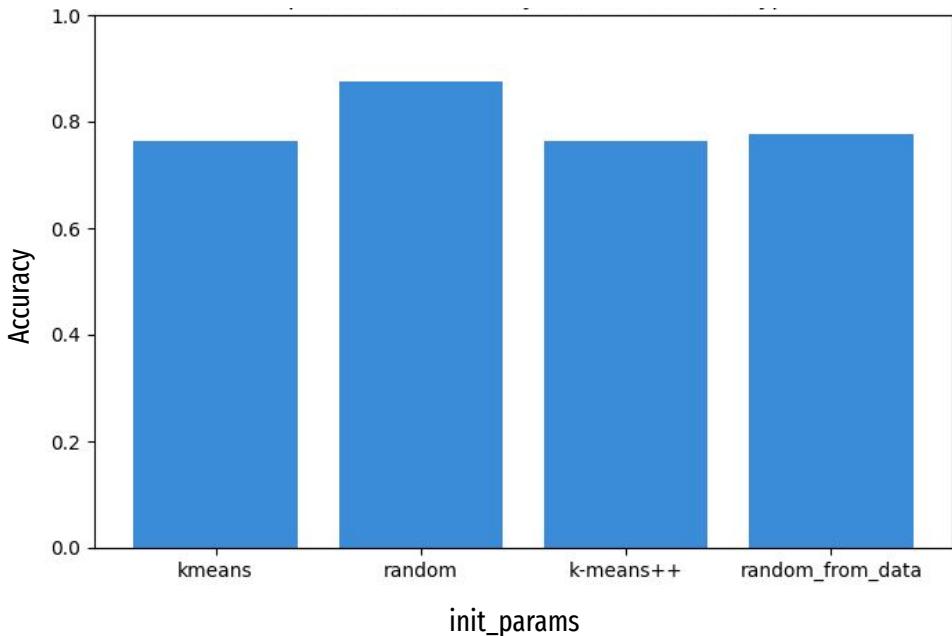
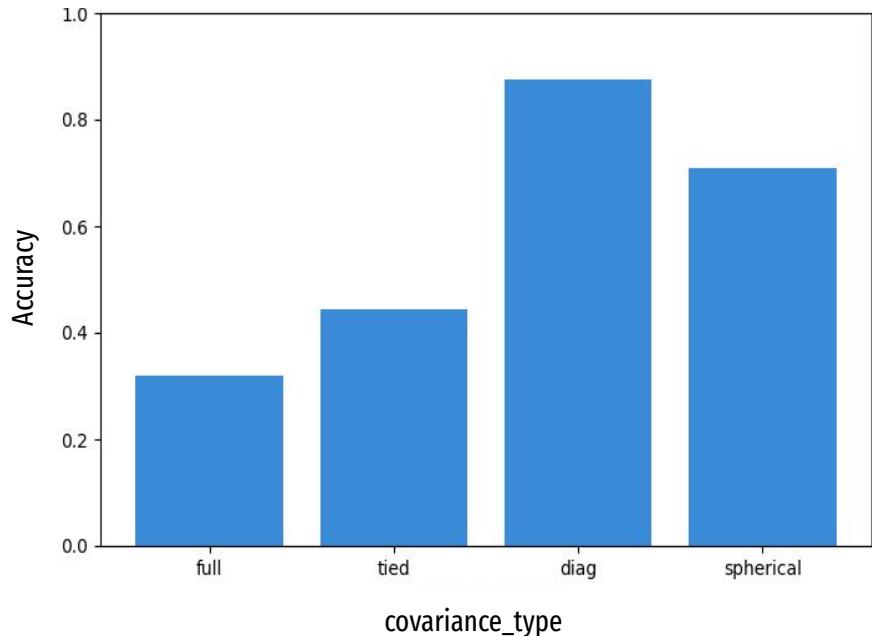
## **Procedimiento**

El proceso para clusterización de audios con emociones positivas y negativas es:

- Lectura de solo los audios del subset elegido filtrados por
  - Solo audio de hombres
  - Solo audios con emociones: Feliz - Calmo - Enojado.
- Extracción de 27 features (MFCCs, Chromas, ZCR, Energía)
- Estandarización de cada feature
- Variación de parámetros según métrica definida con 3 clusters: Feliz - Calmo - Enojado
- Obtención de mejores parámetros y exposición de resultados

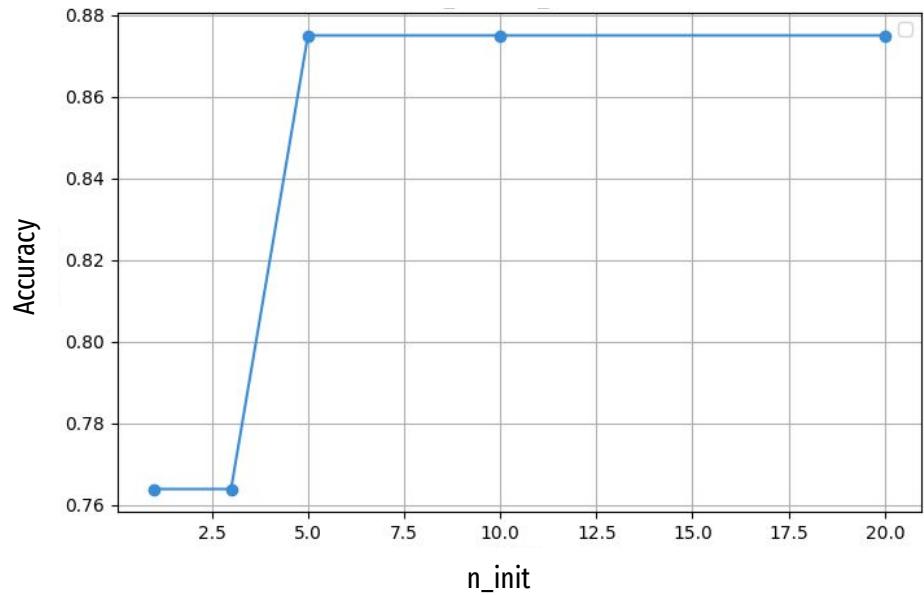
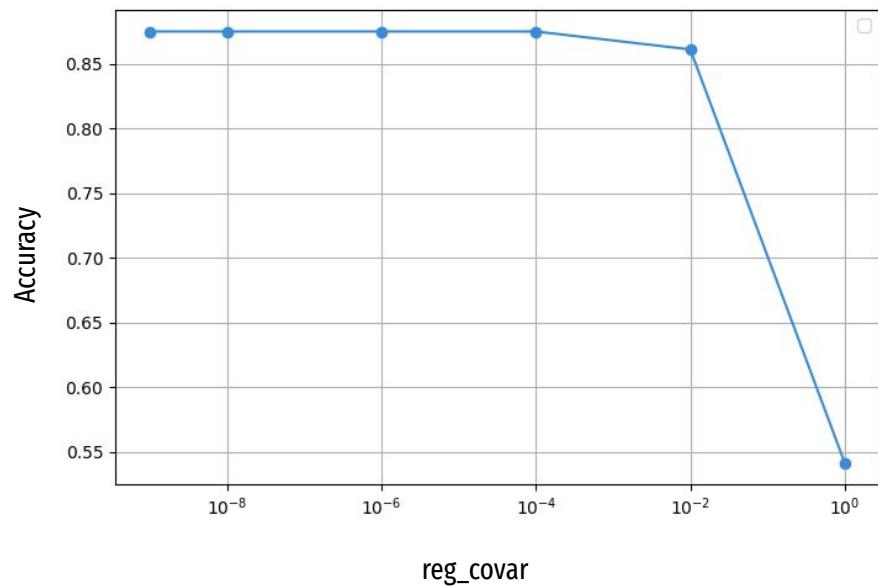
# Clusterización de emociones

## Variación de parámetros



# Clusterización de emociones

## Variación de parámetros



# Clusterización de emociones

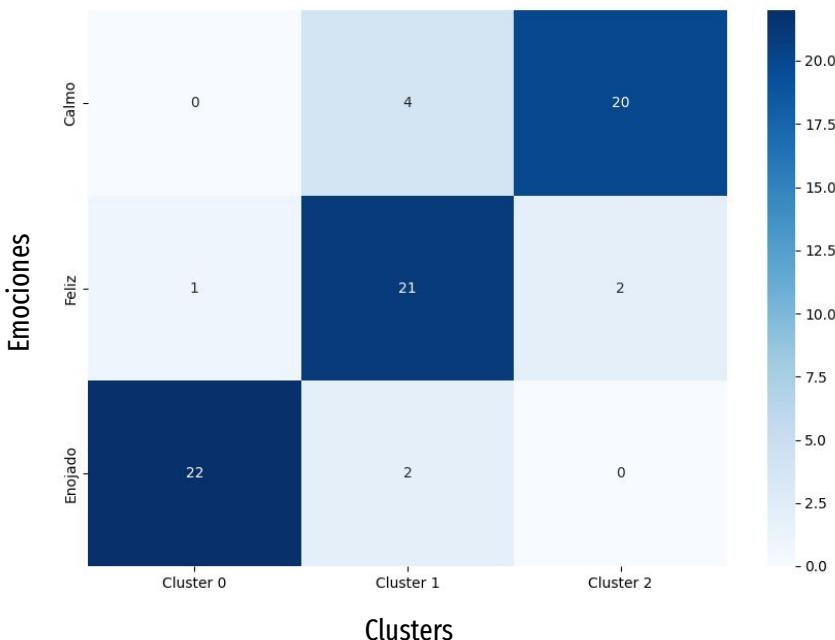
## Resultados

Utilizando los parámetros obtenidos del análisis obtenemos la matriz correspondiente

### **Mejores parámetros:**

init\_params: random  
reg\_covar: 1e-8  
covariance\_types: diag  
n\_init: 10  
n\_components: 3

**Accuracy: 0.875**



# Conclusiones

# Conclusiones

- KMeans es levemente mejor para clusterizar que GMM.
- Isolation Forest es mejor para contaminaciones bajas, y GMM para contaminaciones altas.
- El análisis de densidad puede servir para rápidamente analizar el dataset, si heterogéneo u homogéneo.
- KMeans y GMM tienen buenos resultados en segmentación de imágenes, pero en general KMeans es un poco mejor.
- En procesamiento de audios, covariace\_types e init\_param son los parámetros más relevantes.
- GMM resulta apto para la clusterización de audios en función de las emociones.