TEC | Tecnológico de Costa Rica

Escuela de Ingeniería en Electrónica

Licenciatura en Ingeniería Electrónica

EL5852 Aprendizaje Automático

Tarea 6

Reducción de dimensiones, aglomeración y modelos generativos

Diego Bogarín Picado

diegobp11@gmail.com

2018160264

Oscar Andrés Rojas Fonseca

osaf2412@hotmail.com

2018102187

Cartago, Costa Rica

2 de mayo, 2023

Índice

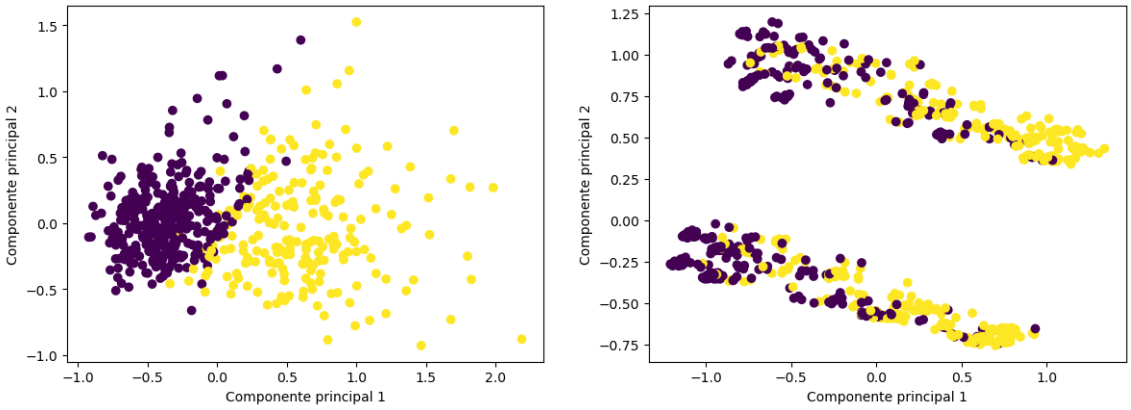
1. Reduccion de dimensiones

Como bien se indica, se debe proceder con reducción de dimensiones de la matriz de datos de 569×30 al realizar proyecciones sobre un plano, de manera que en dos dimensiones sea posible realizar el estudio de los mismos y diagnosticar de manera eficiente a cada muestra, esto considerando ambos conjuntos de datos, diagnóstico de cáncer y de problemas cardiacos. Para esto se realizó un proceso por medio de "Principal Component Analysis" (PCA), "Kernel Principal Component Analysis" (KPCA), se seleccionó como método extra el "Truncated SVM" (TSVM), "Lineal Discriminant Analysis" (LDA), "Uniform Manifold Approximation and Projection" (UMAP) y el "Pairwise Controlled Manifold Approximation" (PaCMAP) [1].

1.1. PCA

Este método proyecta las dimensiones linealmente sin escalarlas, además de centrarlas en el plano de dos dimensiones [2].

Así, aplicando el método con la librería sklearn.decomposition, se obtuvo el resultado mostrado en la Figura 1 para los dos conjuntos de datos.



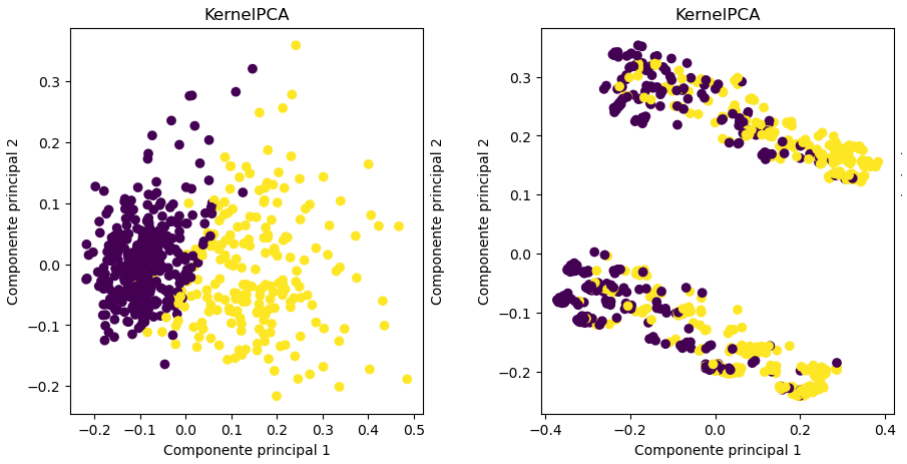
(a) Cáncer (b) Corazón

Figura 1: Visualización de proyección PCA.

1.2. KPCA

Se trata de un proceso basado en PCA que primero realiza un mapeo de los datos hacia un espacio no lineal, para luego aplicar el método PCA [3].

Al aplicar el método se obtuvo el resultado de la Figura 2.



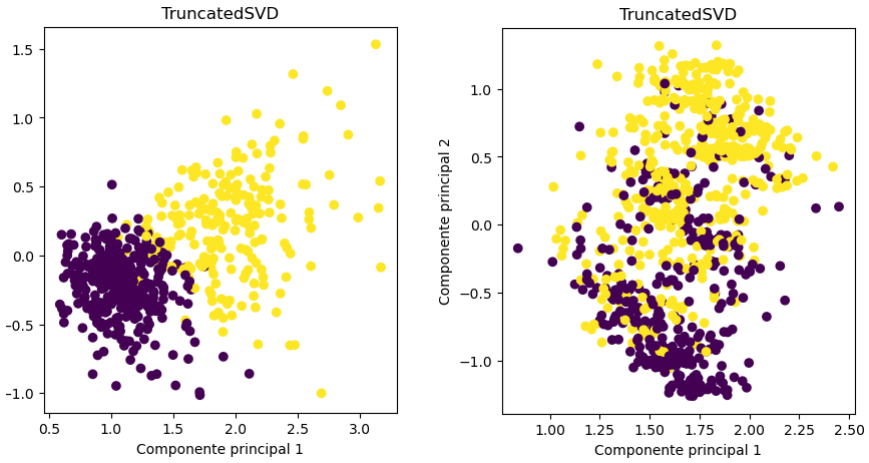
(a) Cáncer (b) Corazón

Figura 2: Visualización de proyección KPCA.

1.2.1. TSVM

Este proceso proviene del campo del álgebra lineal se utiliza para preparar un dato al generar una proyección de un conjunto de datos en menor cantidad de dimensiones y así usarlo en un modelo [4].

Aplicando el método a los datos originales, se obtuvo el resultado de la Figura 3



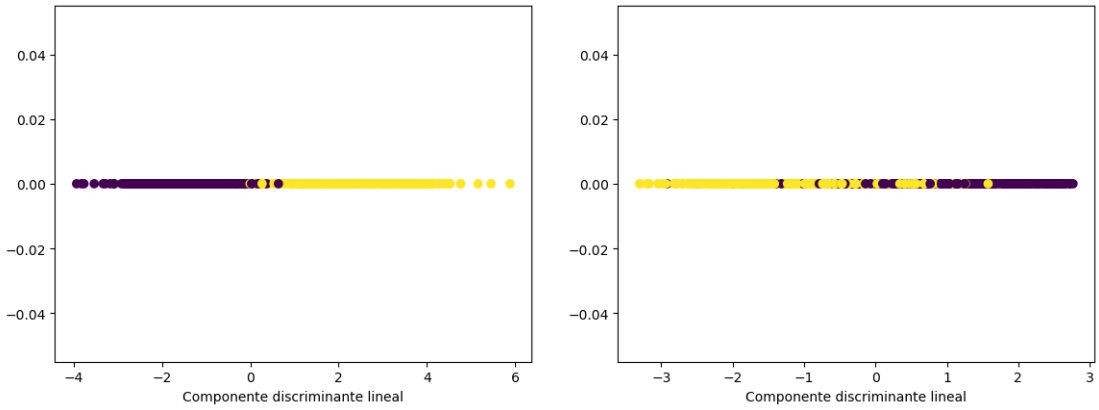
(a) Cáncer (b) Corazón

Figura 3: Visualización de proyección TSVM.

1.3. LDA

Proceso que se utiliza para encontrar una combinación lineal de características que caracterizan o separan dos o más clases de objetos o eventos que normalmente se emplea como preprocesamiento de datos para el aprendizaje [5].

Al emplear el método se obtuvo el comportamiento de la Figura 4.



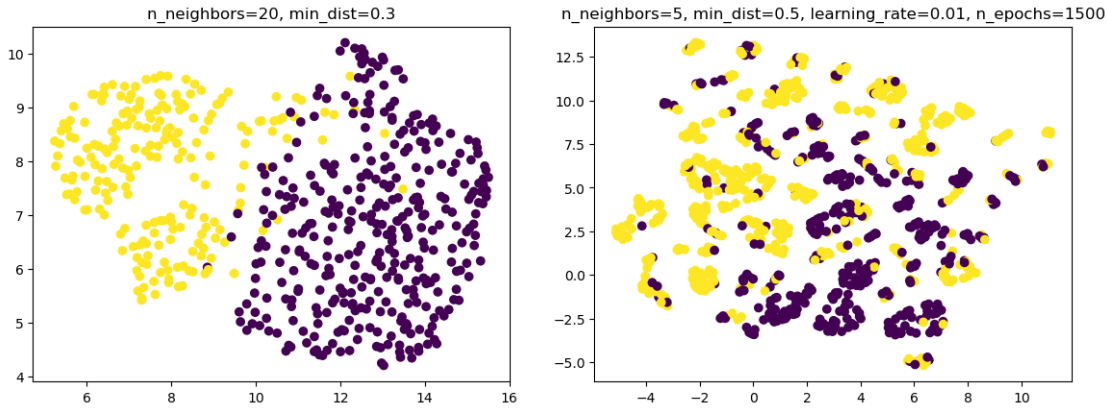
(a) Cáncer (b) Corazón

Figura 4: Visualización de proyección LDA.

1.4. UMAP

Se trata de una técnica de aprendizaje no supervisado que se basa en la topología y la geometría riemanniana para preservar la estructura local y global de los datos [6].

EL método aplicado se seleccionó con los parámetros *n_neighbors* de 20 y *min_dist* a 0.3, dado que las pruebas realizadas mostraron el comportamiento deseado de clasificación eficiente al mostrar aglomeraciones. El resultado se muestra en la Figura 5.



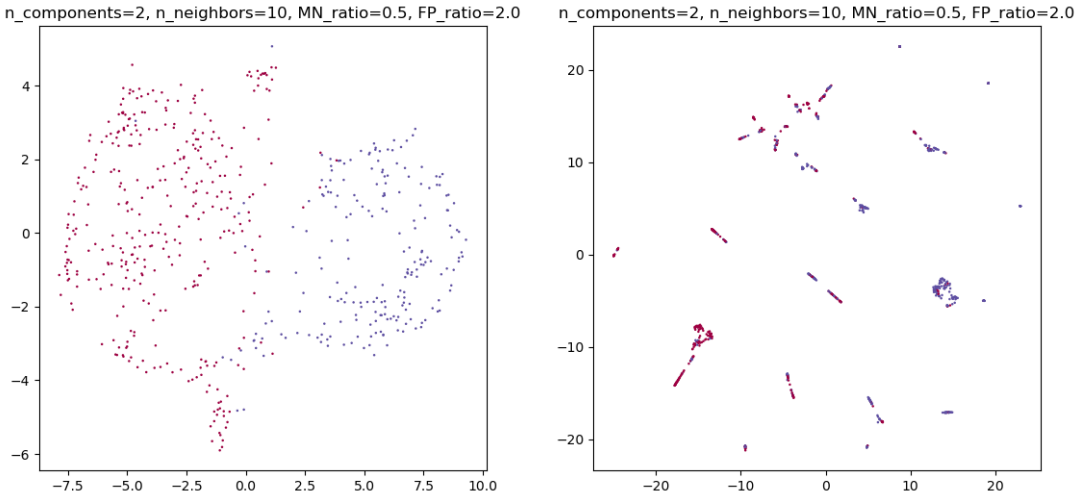
(a) Cáncer (b) Corazón

Figura 5: Visualización de proyección UMAP.

1.5. PaCMAP

Proceso que se basa en la optimización de la probabilidad de que los puntos en el espacio de alta dimensión se mapeen a los puntos correspondientes en el espacio de baja dimensión [7].

Los parámetros seleccionados para la aplicación del método fueron *n_components* con 2 (dimensiones), *n_neighbors* en 10 y por último *MN_ratio* y *FP_ratio* en 0.5 y 2 respectivamente (valores por defecto), para lo cual el resultado se muestra en la Figura 6.

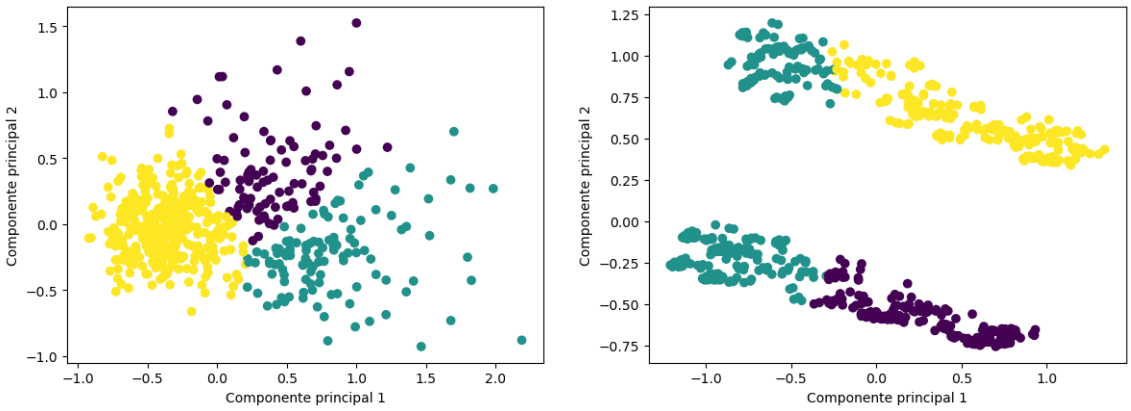


(a) Cáncer (b) Corazón

Figura 6: Visualización de proyección PaCMAP.

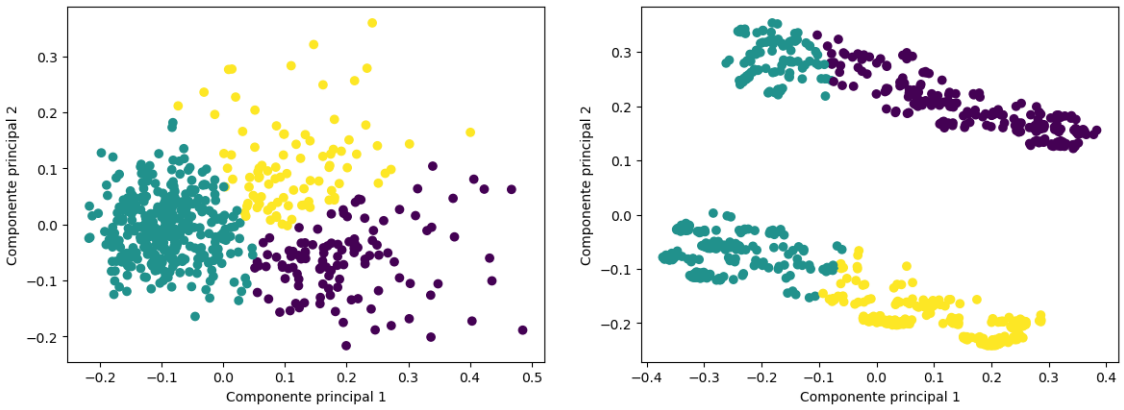
2. Aglomeraciones

Para la parte de aglomeraciones se utilizaron las reducciones de datos anteriores y se acomodaron los datos por acumulaciones, para esto se escogió un $K = 3$ acumulaciones para cada método, observable en la mayoría de los casos, así, se aplicaron los métodos anteriores ajustando la cantidad de clases y sus resultados en las Figuras 7, 8, 9, 10, 11 y 16.



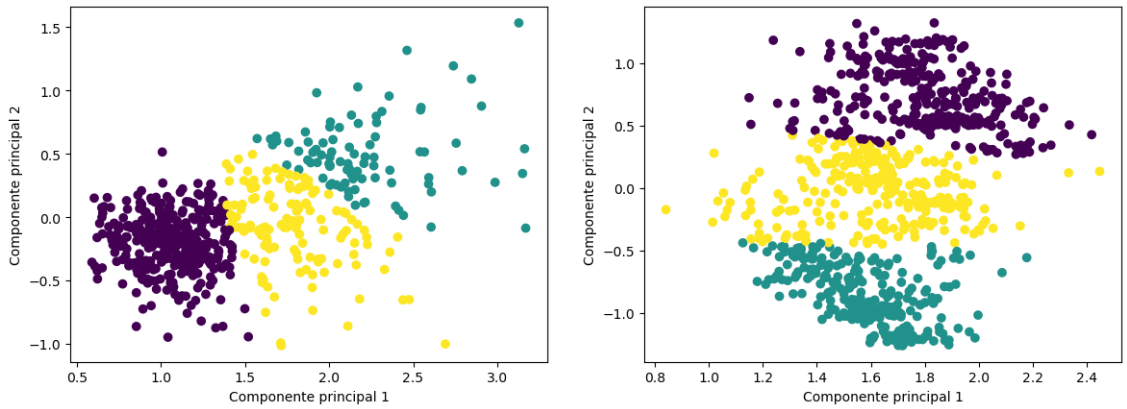
(a) Cáncer (b) Corazón

Figura 7: Visualización de proyección PCA con k=3.



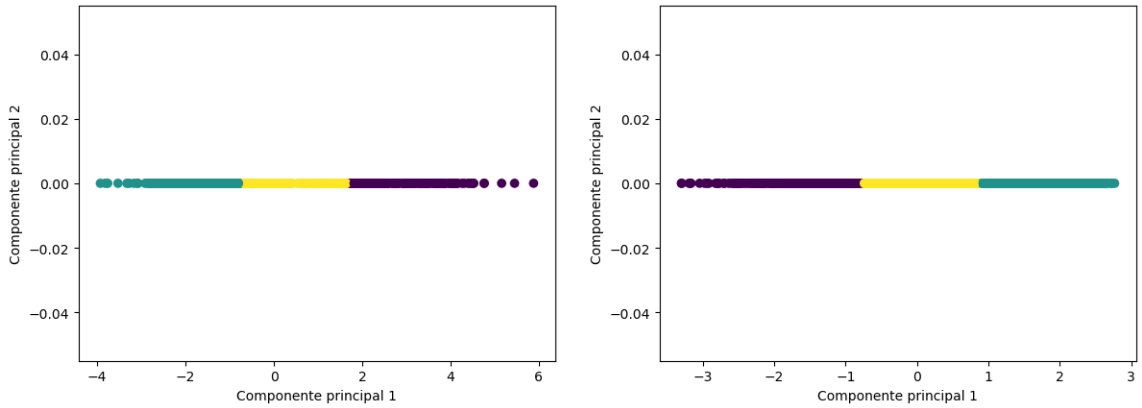
(a) Cáncer (b) Corazón

Figura 8: Visualización de proyección KPCA con k=3.



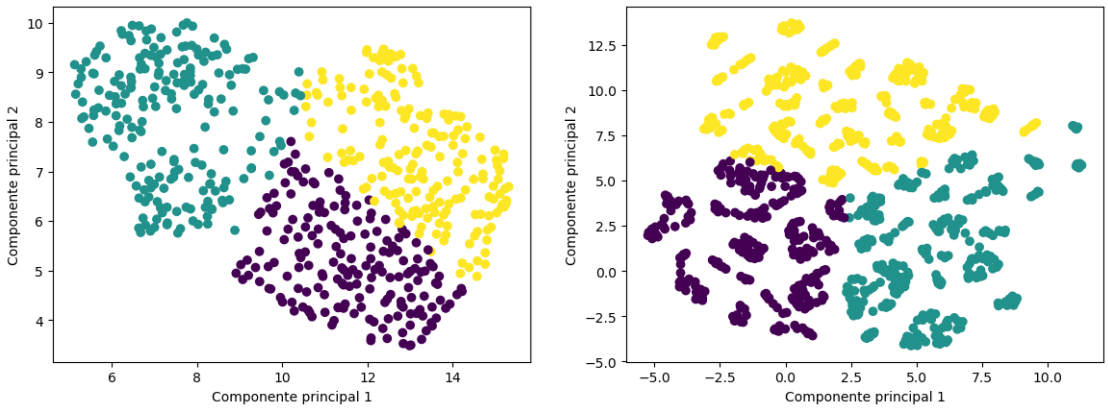
(a) Cáncer (b) Corazón

Figura 9: Visualización de proyección TSVM con k=3.



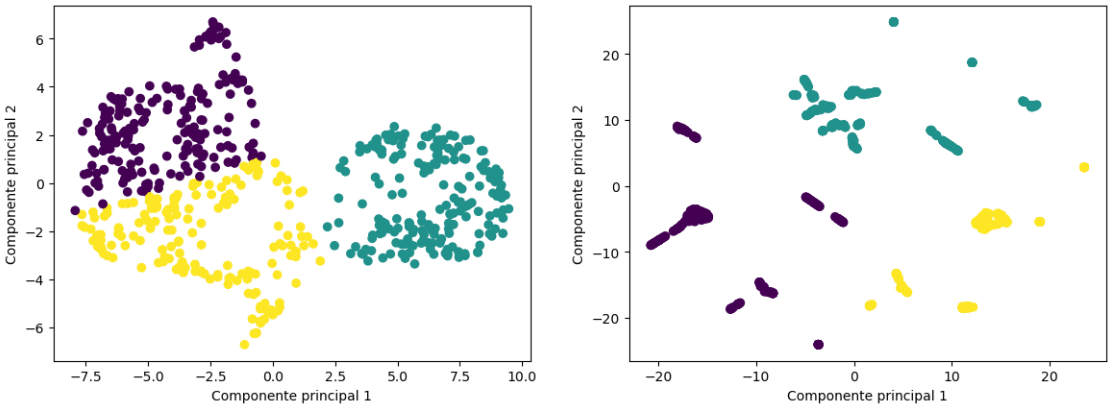
(a) Cáncer (b) Corazón

Figura 10: Visualización de proyección LDA con k=3.



(a) Cáncer (b) Corazón

Figura 11: Visualización de proyección UMAP con k=3.



(a) Cáncer (b) Corazón

Figura 12: Visualización de proyección PaCMAP con k=3.

3. Gaussian Mixture

Se construyo un clasificador generativo, utilizando una mezcla de gaussianas (GaussianMixture) para representar la distribución de densidad de cada una de las clases p(x—y). Donde se probaron diferentes parámetros basados en [8] y se obtuvieron los siguientes resultados del BIC y los gráficos.

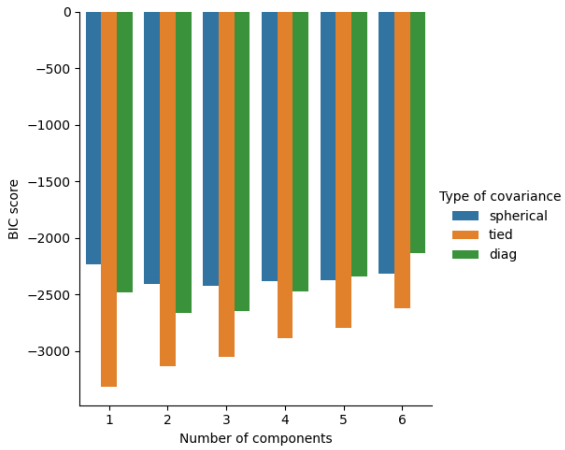
	Number of components	Type of covariance	BIC score
18	1	full	-6040.017307
6	1	tied	-6040.017307
7	2	tied	-5893.085927
8	3	tied	-5772.264686
9	4	tied	-5648.048280

(a) Cáncer

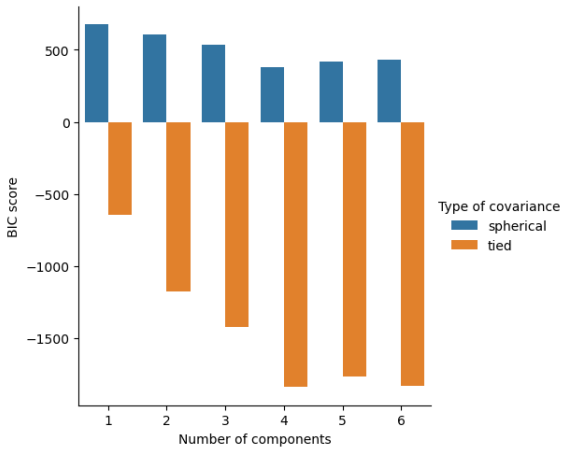
	Number of components	Type of covariance	BIC score
9	4	tied	-1841.440307
11	6	tied	-1829.747024
10	5	tied	-1768.803535
8	3	tied	-1424.741801
7	2	tied	-1180.098162

(b) Corazón

Figura 13: Resultados BIC para el benigno y saludable.



(a) Cáncer



(b) Corazón

Figura 14: Gráficos BIC para el benigno y saludable.

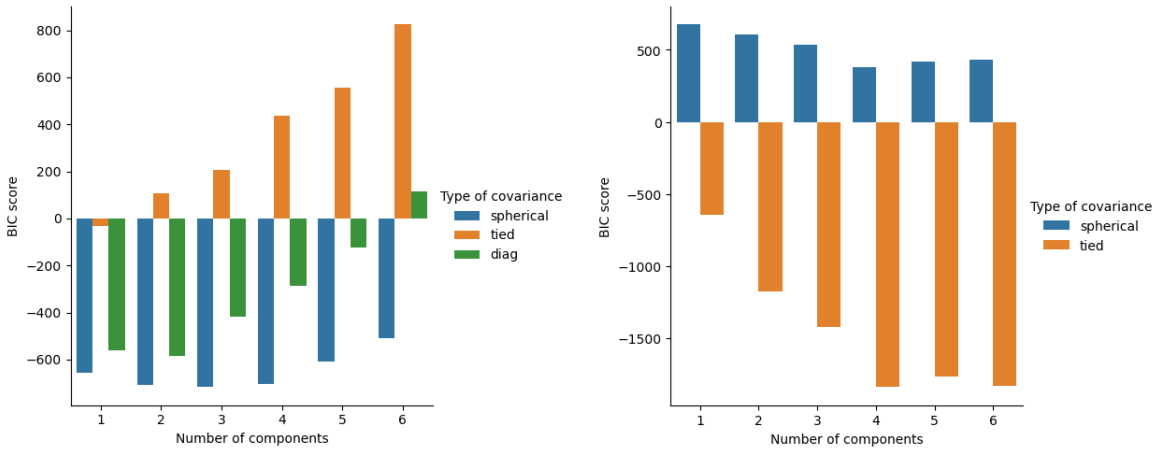
	Number of components	Type of covariance	BIC score
6	1	tied	-3312.690780
7	2	tied	-3129.973825
8	3	tied	-3050.016005
9	4	tied	-2884.775620
10	5	tied	-2798.705533

(a) Cáncer

	Number of components	Type of covariance	BIC score
9	4	tied	-1841.440307
11	6	tied	-1829.747024
10	5	tied	-1768.803535
8	3	tied	-1424.741801
7	2	tied	-1180.098162

(b) Corazón

Figura 15: Resultados BIC para el maligno e insuficiencia.



(a) Cáncer (b) Corazón

Figura 16: Gráficos BIC para el maligno e insuficiencia.

Bibliografía

[1] P. Alvarado-Moya. "Tarea 6: Reducción de dimensiones, aglomeración y modelos generativos". Aprendizaje Automático IS2023 ITCR. 2023.

[2] Area Tutorial. "Scikit-Learn: Reducción de dimensiones con PCA". [En línea]. Disponible en: https://areatutorial.com/scikit-learn-reduccion-de-dimensiones-con-pca/

[3] Complex Systems AI. "Técnicas de Reducción Dimensional". [En línea]. Disponible en: https://complex-systems-ai.com/es/analisis-de-datos/tecnicas-de-reduccion-dimensional/.

[4] Machine Learning Mastery. "Singular Value Decomposition for Dimensionality Reduction in Python". [En línea]. Disponible en: https://machinelearningmastery.com/singular-value-decomposition-for-dimensionality-reduction-in-python/.

[5] KnowledgeHut. "Linear Discriminant Analysis for Machine Learning". [En línea]. Disponible en: https://www.knowledgehut.com/blog/data-science/linear-discriminant-analysis-for-machine-learning.

[6] Canadian Centre for Cyber Security. "Uniform Manifold Approximation and Projection (UMAP)". [En línea]. Disponible en: https://www.cse-cst.gc.ca/en/culture-and-community/research/uniform-manifold-approximation-and-projection-umap.

[7] Y. Wang, H. Huang, C. Rudin and Y. Shaposhnik. "PaCMAP". [En línea]. Disponible en: https://pypi.org/project/pacmap/.

[8] "Gaussian mixture model selection", scikit-learn. [En línea]. Disponible en:https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_selection.html. [Consultado: 17-may-2023].