Centro de Investigación en Matemáticas (CIMAT). Unidad Monterrey

Tarea 5

Victor Manuel Gómez Espinosa

27 de mayo de 2020

1. Problema 1

En este problema se implementó el modelo clásico perceptrón y se probó con un conjunto de datos sintéticos donde las clases son separables (Figura 1.1).

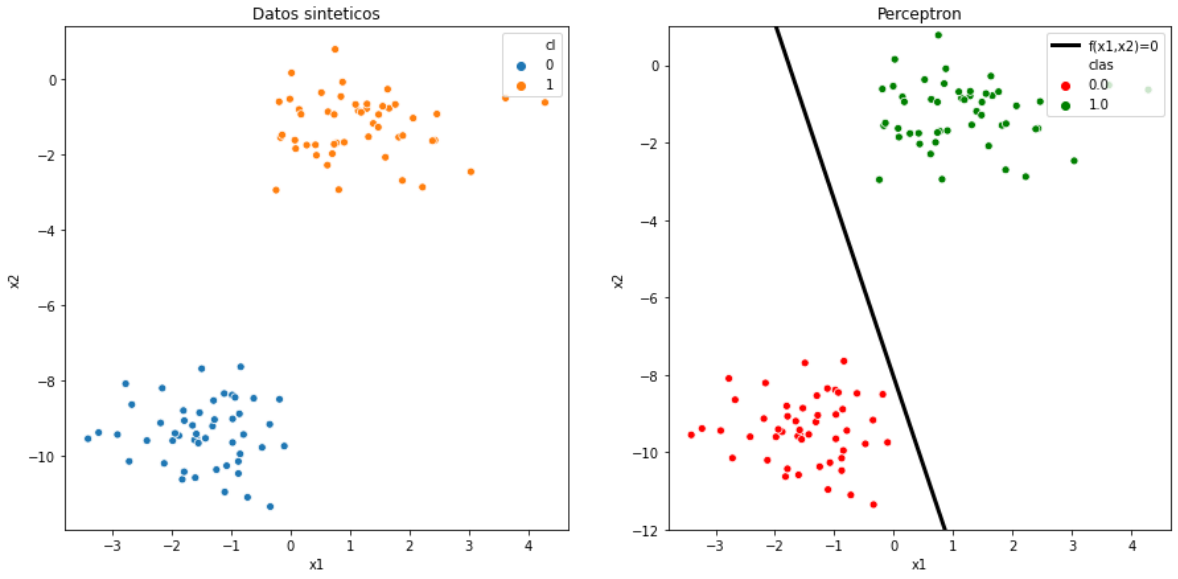

Figura 1.1: a) Datos sintéticos, b) Perceptrón aplicado a los datos sintéticos para clasificación, la línea negra representa el plano separador $f(X)=0$.

Posteriormente se utilizó el conjunto de datos *diabetes* que contiene 8 características médicas (embarazos, glucosa, presión, grosor de piel, insulina, edad, entre otras.) de 768 mujeres de la tribu de Indios Pima indicando si tiene diabetes (1) o no (0), se separó un conjunto de entrenamiento y uno de prueba del 70 y 30%. Observe la Figura 1.2 y observe que los datos se re etiquetaron (si 1, no -1) y además que hay más casos negativos que positivos para ambos conjuntos.

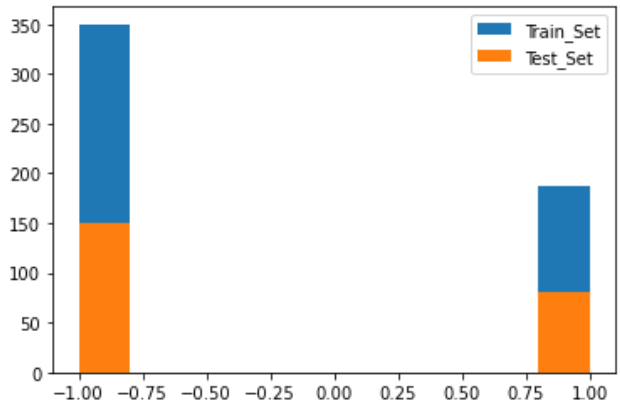

Figura 1.2: Conjuntos de datos de entrenamiento y de prueba.

Se utilizaron los datos de entrenamiento para ajustar 4 modelos de clasificación (perceptrón, regresión logística, LDA y QDA) y los datos de prueba para verificar dicho modelo. Para medir el desempeño de los modelos, se reporta para los conjuntos de entrenamiento y de prueba la precisión junto con sus respectivas matrices de confusión y para el conjunto de prueba las curvas ROC y su área bajo la curva (AUC Score).

Antes de continuar a los resultados se explicarán brevemente estas métricas empleadas.

Observe la Figura 1.3, que corresponde a una matriz de confusión para un caso de ejemplo, la precisión es la fracción que ayuda a cuantificar los positivos verdaderos del total de clasificados como positivos, es decir nos ayuda a medir que tan bueno es el modelo en predecir la categoría de interés, como por ejemplo que si tienen diabetes.

Ahora observe la Figura 1.4 b), corresponde al grafico de las curvas ROC donde el eje horizontal corresponde a la tasa de falsos positivos y el eje vertical a la tasa de verdaderos positivos (igual a la fracción Recall), que son las fracciones que se indican en la Figura 1.4 a). Las curvas ROC, dan para un modelo los posibles escenarios de matriz de confusiones.

Esta es otra forma de medir los modelos, y mientras mayor sea el área debajo de la curva mejor será el modelo (máximo 1, mínimo 0), por ejemplo una curva que pase por el punto

verde en 0,1, daría el caso perfecto de AUC=1, mientras una línea en 0 paralela al eje horizontal (puntos rojos) será la peor AUC=0 y por otro lado una línea que pase por el punto 1,1 (punto negro) tendrá tasas iguales, lo que significa que las proporciones de haber catalogado correctamente la categoría de interés es la misma proporción de clasificar incorrectamente las que no son de la categoría de interés.

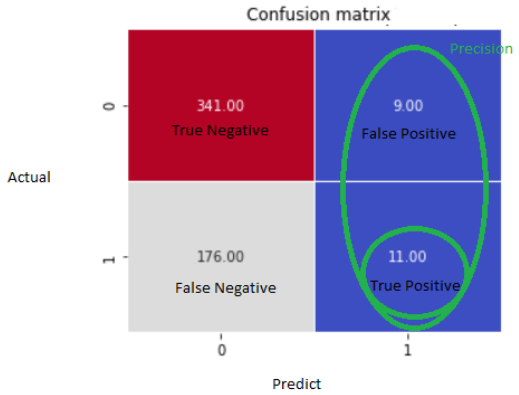

Figura 1.3: Matriz de confusión, Precisión.

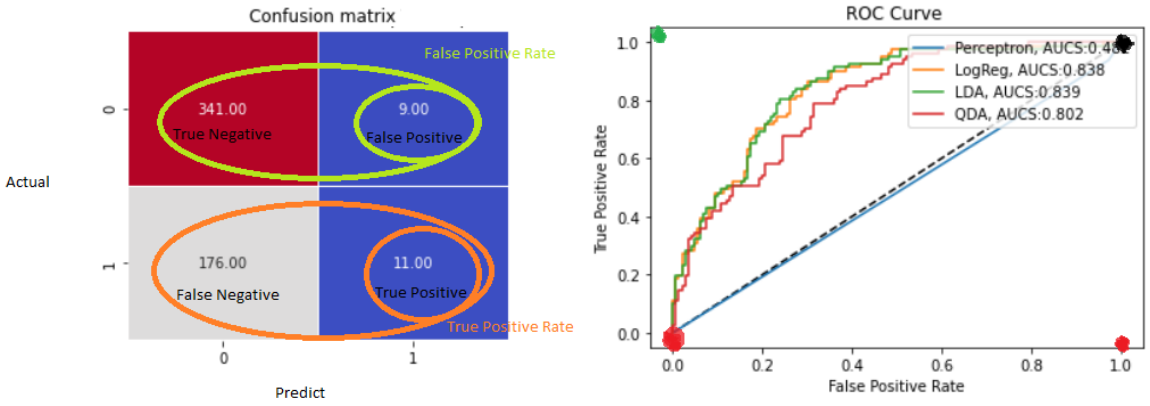

Figura 1.4: a) Matriz de confusión y tasas de falsos positivos y de verdaderos positivos, b) curvas ROC y área bajo las curvas.

En la Figura 1.5 puede observar los resultados de las matrices de confusión para cada modelo y conjunto de datos, mientras que en la tabla 1.1 puede observar la precisión para cada modelo para cada conjunto de datos.

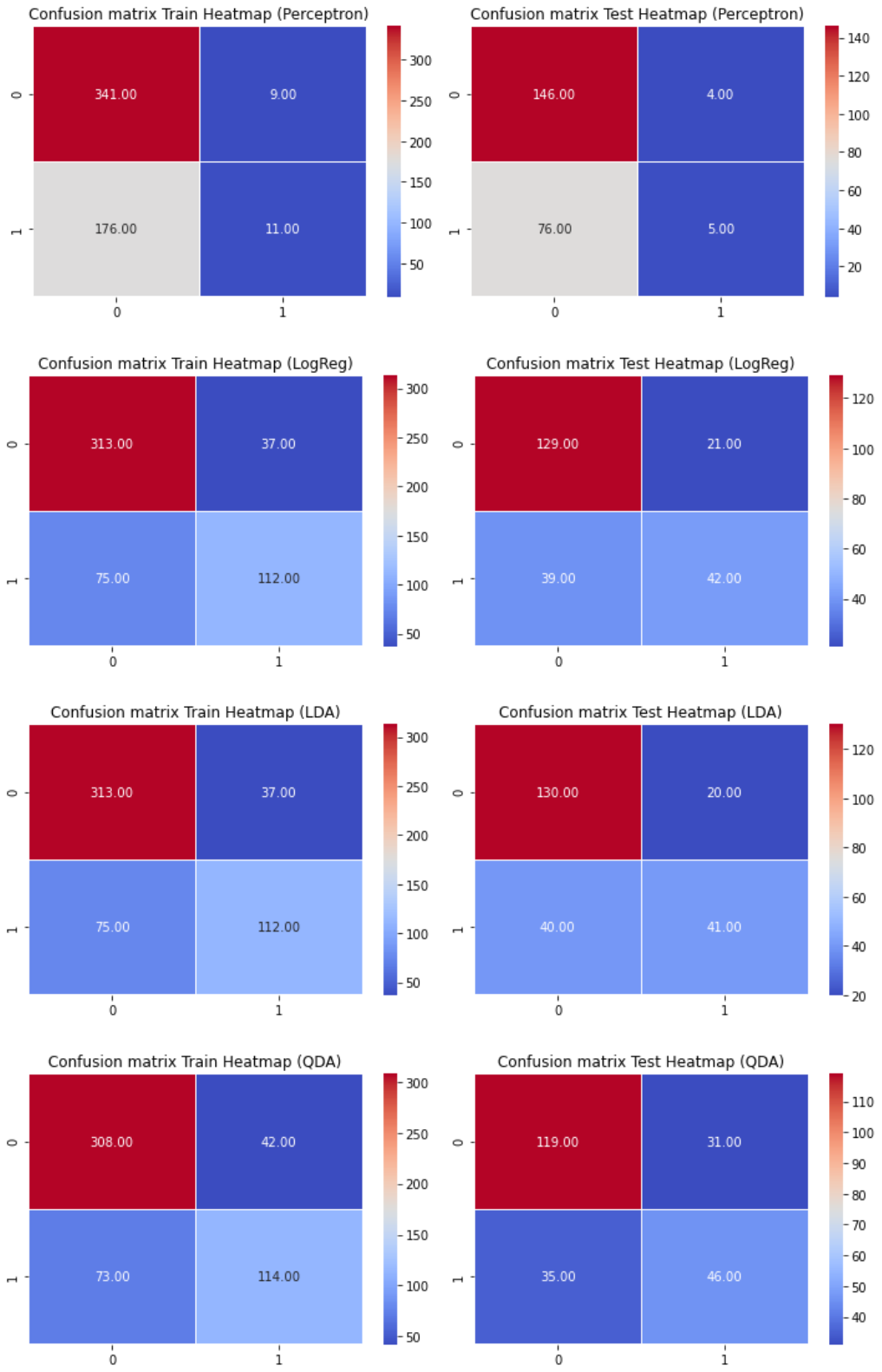

Figura 1.5: Matrices de confusiones para cada modelo (renglones) por conjunto de datos (columnas).

MODEL/SET	TRAIN_SET	TEST_SET
PERCEPTRON	0.55	0.56
LINEAR REGRESION	0.75	0.67
LDA	0.75	0.67
QDA	0.73	0.6

Tabla 1.1: precisión para cada modelo por cada conjunto de datos.

Observe la Tabla 1.1 y note que para ambos conjuntos de datos bajo el criterio de la precisión los modelos LDA y de regresión lineal son los mejores modelos, esto se corrobora también bajo el criterio de las curvas ROC y el área bajo la curva (AUCS) ya que justamente son estos dos los que tienen una mayor área (prácticamente la misma) y el peor el modelo perceptrón (lo cual tiene sentido, ya que indica que las categorías no son claramente separables, entonces el algoritmo no converge a una solución).

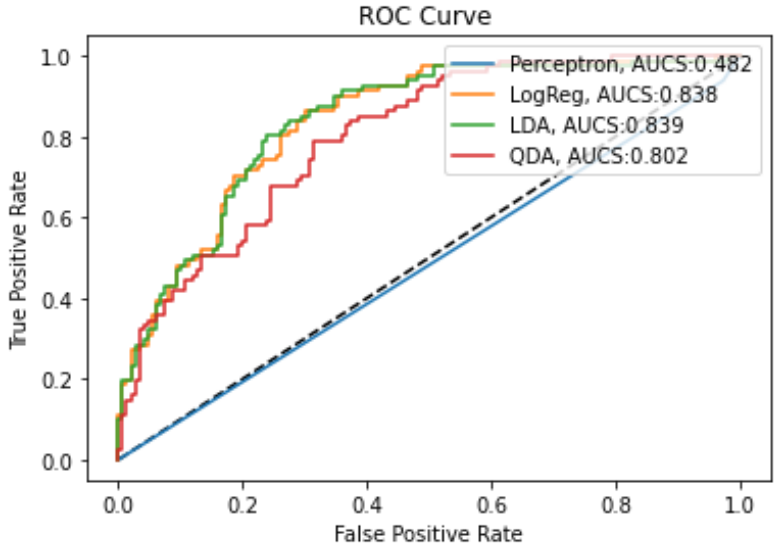

Figura 1.6: Curvas ROC y áreas bajo la curva (AUCS) para cada modelo para el conjunto de prueba.

2. Problema 2

Para este problema se utilizaron los datos del archivo my_all_tracks_2019.csv que corresponde a un extracto de una base de datos muy extensa (FMA), los datos describen diferentes características de los audios (3 bloques por así decirlo), información correspondiente a la canción, características del audio y características de la señal, para 13129 canciones (Figura 2.1).

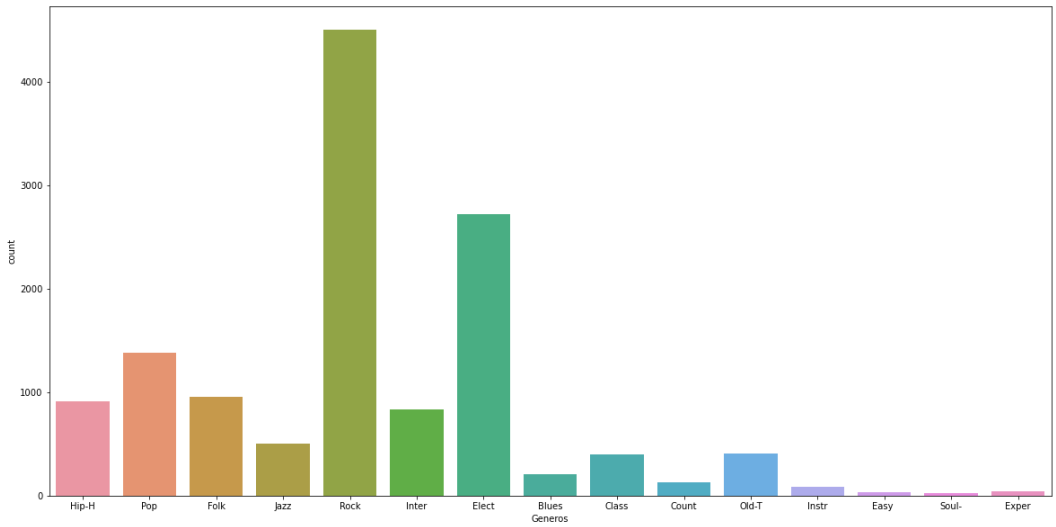

Figura 2.1: Distribución de los géneros musicales en la base de datos del archivo my_all_tracks_2019.csv.

El procedimiento seguido fue el siguiente, primero se cargaron los datos, se dividieron en los 3 bloques anteriormente mencionados y se seleccionaron variables para analizar. Del primer bloque se seleccionaron variables que reflejen alguna característica de la canción (duración e interés) y el tipo de género. Del segundo bloque se tomaron principalmente la variable tempo (que es la que mostraba entre todas, la mayor variabilidad y quizá la que podría aportar más información), pero también se utilizaron las variables correspondientes a energía y bailabilidad para hacer un análisis exploratorio por género que más adelante describiré. Por último, del tercer bloque únicamente se tomaron las variables que mostraban variabilidad y finalmente, todas las variables seleccionadas se estandarizaron.

Después de elegir las variables, primero con la selección del bloque 1 y 2 se hizo un análisis exploratorio y descriptivo. Primero se tomaron las variables (v) de bailabilidad y energía (Figura 2.2) que son variables continuas, índices numéricos entre 0-1. Se tomó el primer y tercer cuartil de cada uno para definir límites y crear nuevas variables categóricas donde 0 indica baja $(v \leq Q_1)$, 1 media $(Q_1 < v \leq Q_3)$, y 2 alta $(v > Q_3)$ para cada variable.

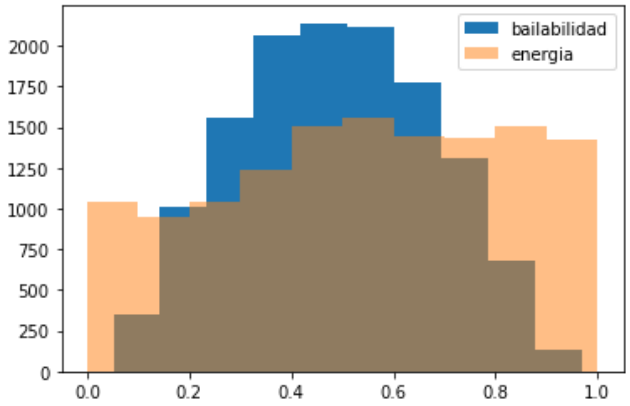

Figura 2.2: Histograma de las variables de bailabilidad y energía.

Posteriormente se utilizaron estas nuevas variables para realizar un análisis visual por genero para tratar de ver si había características que compartían entre algunos géneros para tratar de agruparlos y crear una nueva variable categórica con menos categorías, ya que el proceso de identificar los 15 géneros por separado no resulta muy intuitivo.

Observe la Figura 2.3, donde se agrupan los géneros contra alguna variable numérica y categórica (bailabilidad y energía), observe la figura de la derecha, donde en el eje vertical tenemos la variable continua bailabilidad y como variable categórica tenemos la energía, se puede notar que hay géneros donde predomina el azul, es decir que son principalmente de baja energía y que a su vez muestran baja distribución en cuanto a bailabilidad, hay otros que predomina el verde (alta energía) y también su distribución en bailabilidad es alta y por otro lado hay otros que simplemente lucen muy dispersos en cualquiera de las características y se observa el mismo patrón en la figura del lado izquierdo.

Por lo tanto se creó la nueva variable categórica Generos2, en la cual la categoría 0 (Blues, Instr, Easy, Soul, Exper)incluye los géneros que se notan muy dispersos, en la categoría 1 (Class, Old-T, Folk, Jazz) los que son de baja energía o bailabilidad y en la categoría 2 los que son de alta energía o bailabilidad (Hip-Hop, Pop, Rock, Inter, Elect).

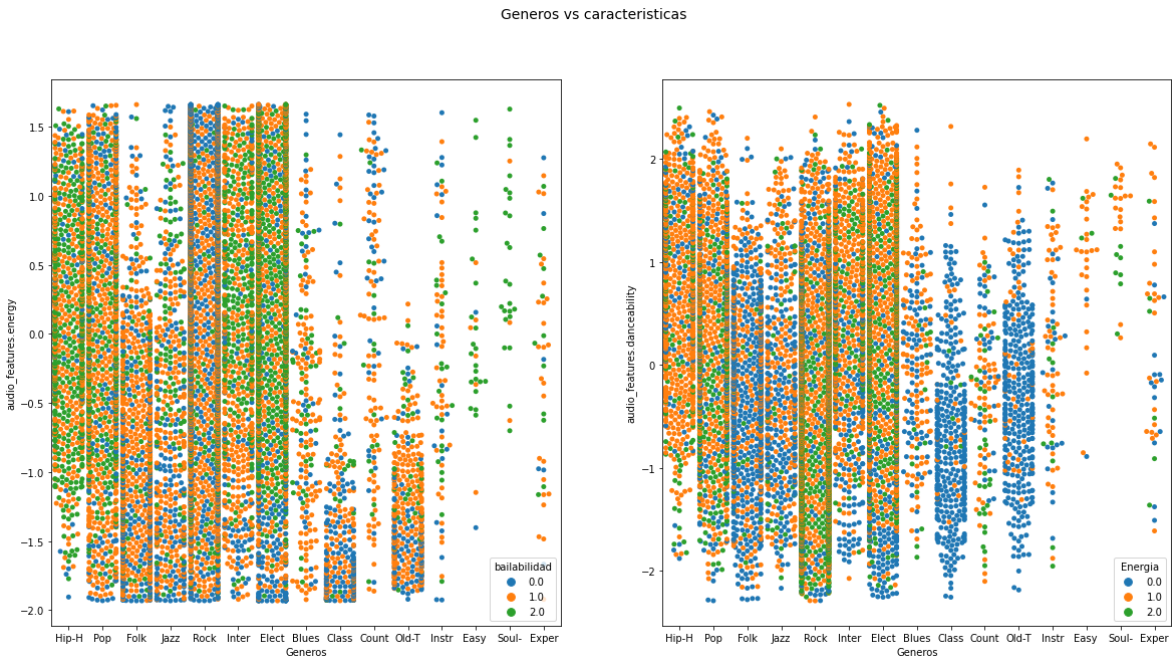

Figura 2.3: Géneros contra características. a) variable continua: Energía, variable categórica: bailabilidad. b) variable continua: bailabilidad, variable categórica: energía.

Después se aplicó a las variables seleccionadas del bloque 1 y 2 técnicas de reducción de dimensión (PCA y kernel PCA con diferentes tipos de kernel y parámetros) y la que mostró mejores resultados fue Kernel PCA con kernel gaussiano con parámetro gamma=2. Observe la Figura 2.4 y note en la figura del lado izquierdo como en su gran mayoría se nota la separación de las 3 categorías de la variable anteriormente creada Géneros2 o todos los géneros en la figura del lado derecho.

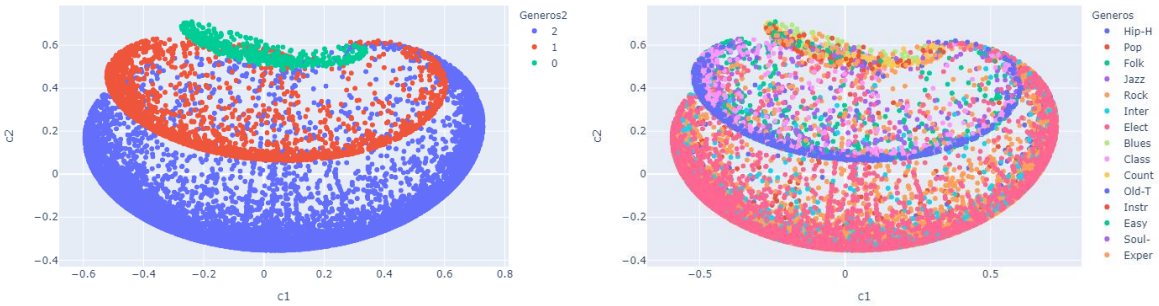

Figura 2.4: Kernel PCA, kernel Gaussiano (gamma=2). a) por categoría Generos2, b) por todos los géneros.

A los resultados obtenidos previamente se aplicaron técnicas de clustering y como podrá notar en la Figura 2.5, los resultados no concuerdan muy bien con lo que se encontró previamente.

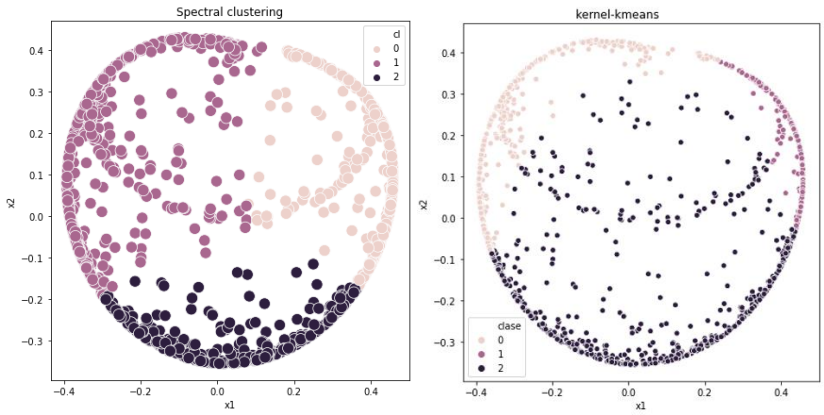

Figura 2.5: Clustering (3 clústers) a los resultados obtenidos por Kernel PCA. a) clustering espectral, b) clustering polinomial de grado 2.

Posteriormente se tomaron las variables del bloque de características de la señal (X) y las variables categóricas anteriormente creadas de Bailabilidad y Energía (de forma separada) como variables respuesta (y) y se seleccionaron conjuntos de entrenamiento y prueba para cada una en proporción 70 – 30% respectivamente (Figura 2.6).

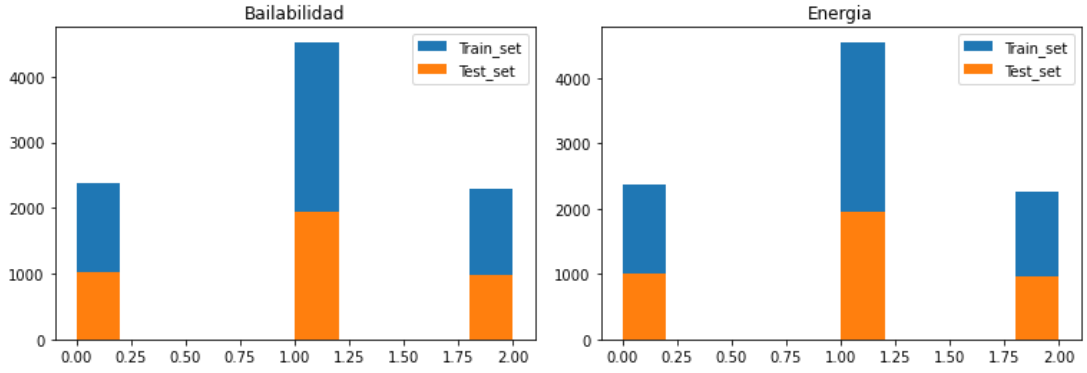

Figura 2.6: Conjuntos de datos de entrenamiento y de prueba para Bailabilidad y Energía.

Con estos conjuntos de forma separada se ajustaron clasificadores LDA, QDA, Multilogit y Redes neuronales y se compararon sus desempeños de forma general para los conjuntos de prueba y entrenamiento con la métrica de la exactitud (o accuracy), en la Tabla 2.1 podrá observar los resultados y notará que en ambos el mejor modelo resultó ser el de redes neuronales.

Para el modelo de bailabilidad se utilizaron 2 capas ocultas con 20 neuronas cada una, con la función de activación ReLU y regularización de 6, para el modelo de energía se utilizaron también dos capas, pero con 10 neuronas cada una, de igual forma se utilizó la función de activación ReLU y parámetro de regularización de 10.

	BAILABILIDAD		**ENERGÍA**	
	Train	Test	Train	Test
LDA	0.63	0.61	0.77	0.75
QDA	0.59	0.47	0.71	0.61
MULTILOGIT	0.64	0.51	0.79	0.76
REDES NEUR	**0.65**	**0.62**	**0.78**	**0.77**

Tabla 2.1: Accuracy por tipo de modelo para cada conjunto de cada variable respuesta.

Finalmente, del conjunto de datos del archivo my_all_tracks_No_genre_2019.csv se seleccionaron de forma aleatoria 5 canciones, las cuales se escucharon y se evaluaron cualitativamente y se les otorgó una categoría de Bailabilidad y Energía y posteriormente se utilizaron los modelos de redes neuronales anteriormente ajustados para predecir las categorías de estas variables y se compararon los resultados.

Observe la Tabla 2.2 y observe que de forma general los resultados son bastante buenos, por un lado, la Energía tiene muy buenos resultados puesto que sólo en una no concuerda con la cualitativa, aunque por otro lado la Bailabilidad sólo en 2 los resultados coincidieron, pero sin embargo no están tan alejados del valor cualitativo, aunque también hay que considerar que cualitativamente para otra persona que asigne valores estos podrían ser diferentes.

	CUALITATIVA		**CUANTITATIVA**	
TRACK	Bailabilidad	Energía	Bailabilidad	Energía
095631	1	2	1	2
084238	1	2	2	2
039060	0	0	1	0
047627	0	0	1	1
089887	1	0	1	0

Tabla 2.2: Resultados cualitativos y cuantitativos para 5 canciones de prueba.