Análisis Predictivo de Éxito en la Industria Musical: Un Enfoque Basado en Redes de Colaboración y Atributos de Artistas en Spotify

Ana Conde Marrón

26 de mayo de 2024

1. Introducción

En este estudio se utiliza como objeto de estudio un conjunto de datos que contiene datos alrededor de 20.000 artistas cuyas canciones llegaron a las listas semanales de *Spotify* y alrededor de 136.000 artistas adicionales que colaboraron al menos una vez con uno de los artistas de las listas. Estos datos pueden representarse como un grafo pues contamos con los nodos (los artistas) y con las aristas (las colaboraciones). Además los nodos cuentan con atributos muy interesantes como sus géneros musicales, popularidad, seguidores... Uno de los que habría que destacar es el que muestra las listas de éxito en las que ha aparecido, pues se podría utilizar como variable objetivo a la hora de determinar si un artista puede considerarse que ha tenido éxito o no. O también la popularidad pues podríamos predecirla a partir de los atributos de este *dataset*, atributos que calcularemos en base a sus relaciones con otros nodos, y otros atributos derivados.

Del mismo modo que del atributo de las listas de éxitos se ha derivado el si ha pertenecido el artista o no a una lista de éxitos, se han derivado otros atributos como por ejemplo los potenciales oyentes a partir de las listas de éxito de los países a las que ha pertenecido. Los dos *datasets* que se utilizaron fueron el ya mencionado con los datos de spotify[1] y este con poblaciones mundiales[2]. Esto se ha podido llevar a cabo pues las listas en las que ha aparecido cada artista contienen las siglas del país al que pertenece la lista.

En este estudio trataremos la información que nos aportan los grafos mediante la extracción de características o construcción manual de *features* relacionales.

2. El conjunto de datos

En la sección anterior se dieron algunas pinceladas sobre los datos con los que se trabajará pero ahora nos adentraremos más en la naturaleza de sus datos. Cuenta con dos ficheros .csv. Uno contiene los nodos (nodes.csv) del grafo y los atributos de los mismos, estos son:

- *spotify_id*(cadena): aquí se almacena el identificador único del artista por el que se diferenciarán los nodos.
- *name*(cadena): el nombre del artista.
- *followers*(decimal): número de seguidores del artista.
- *popularity*(integer): número del 0 al 100 para indicar su popularidad, derivado de la API de *Spotify*.
- *genres*(lista): lista de géneros a los que pertenece. Tras el procesado de los datos se convertirá en el número de géneros a los que pertenece.
- *chart_hits*(lista): lista de listas de éxitos en las que ha estado. Lo procesaremos de dos maneras según el objetivo de predicción bien como 0 en caso de ser un artista que no ha pertenecido a una lista de éxitos o 1 si sí ha pertenecido. Por otro lado, también lo transformaremos en el número exacto de listas de éxitos a los que ha pertenecido cuando sea ese el objetivo de predicción.

En el cuadro 1 podemos ver dos ejemplos de nodos (se han acortado los identificadores para pode representar el cuadro de forma más clara).

spotify_id	name	followers	popularity	genres	chart_hits
48W...	Byklubben	1738.0	24	['nordic house', 'russelater']	['no (3)']
4lDi...	Kontra K	1999676.0	72	['christlicher rap']	['at (44)', 'de (1)']

Cuadro 1: Ejemplo entrada tabla de nodos

Y como vemos en el cuadro 2 así se representaría una arista que une los dos nodos mostrados en el cuadro 1.

id_0	id_1
48WvrUGoijadXXCsGocwM4	4lDiJcOJ2GLCK6p9q5BgfK

Cuadro 2: Ejemplo arista

Cuenta con un total de 156422 artistas como nodos y 300386 colaboraciones como aristas. No obstante antes de comenzar el estudio hubo que tratar los datos pues contenía nulos que tendrían que ser eliminados o sustituidos para las futuras predicciones.

3. Las colaboraciones

Como ya se indicó el hecho de que los datos vengan representados en forma de grafo nos va a aportar muchísima información sobre las relaciones entre sus nodos. A partir de las relaciones entre sus nodos podemos extraer las colaboraciones entre artistas, además de propiedades de los nodos que conforman el grafo, como sus centralidades. Esto es un forma de codificación del grafo, como se utilizarán Árboles de decisión para aprovechar su interpretabilidad tendremos que convertir los grafos en datos tabulares. Si se usasen *Graph Neural networks* se podría pasar como entrada el propio grafo, pero como se ha decidido escoger un modelo interpretativo aprovechando la interpretabilidad de los atributos del conjunto de datos nos limitaremos a extraer distintas medidas en l teoría de grafos.

Estas son las medidas que se tuvieron en cuenta:

- *Centralidad de grado*: es el cálculo del grado de cada uno de los nodos. Para calcularla, si tenemos la matriz de adyacencia de un grafo, donde cada $a_{i,j}$ asume el valor 1 su existe una arista (i, j)y el valor 0 cuando no existe. Entonces,

$$C_D(i) = \sum_j (a_{i,j}) = \sum_j (a_{i,j})$$

  considerando el vector $p = (C_E(1), C_E(2), ..., C_E(n))'$ con las centralidades de Eigen de todos los nodos de la red. [3]

- *Centralidad Eigen*: también llamada de Autovalor mide la influencia de un nodo en el grafo, y corresponde con el autovalor principal de la matriz de adyacencia del grafo. Si se define un grafo como $G = (V, E)$, donde V es el conjunto de nodos o vértices y E su conjunto de aristas. Entonces la centralidad de Eigen [4] para un nodo

$$C_E(j) = \sum_{i=1} n(a_{i,j}) C_E(i) = a_{1j} C_E(1) + a_{2j} C_E(2) + ... + anj C_E(n)$$

- *Coeficiente de clustering*: Probabilidad de que dos nodos vecinos a uno dado, sean vecinos entre sí. Se calcula de la siguiente forma,

$$C_i = \frac{E_i}{\frac{1}{2} k_i (k_i - 1)}$$

  donde E_i es el número de aristas que conectan entre sí los nodos adyacentes al nodo i.

- *Centralidad de cercanía:* es la suma o el promedio de las distancias de los caminos más cortos de un nodo a los demás del grafo [5]. Se calcula de la siguiente forma,

$$C(x) = \frac{N-1}{\sum_y (d(y,x))}$$

donde N es el número de nodos del grafo y $d(y,x))$ es la distancia del nodo y al nodo x.

- *Centralidad armónica*: es una centralidad derivada e la centralidad de cercanía, se calcula a partir de la inversa de la cercanía. Se calcula de la siguiente forma[6],

$$C_H(i) = \sum_{j \neq i} (\frac{n-1}{d(x,y)})$$

3.1. Análisis del grafo

En medidas como el *coeficiente de clustering* es interesante ver como se distribuyen las distintas comunidades del grafo y hacer un pequeño análisis para ver si podría darse el caso que nodos más cercanos a artistas con éxito hayan acabado apareciendo en las listas de éxitos de algún país. De esta forma si detectamos a pequeña escala un patrón podemos extrapolarlo a todo le grafo y por medio de estas propiedades alcanzar un mejor rendimiento en nuestro algoritmo predictivo. Por ejemplo, a mayor coeficiente de *clustering* significará que es más probable que tenga más colaboraciones pues la probabilidad de que esté conectado con un nodo cualquiera es alta. En la figura 1 vemos una muestra de artistas que han aparecido en listas de éxitos.

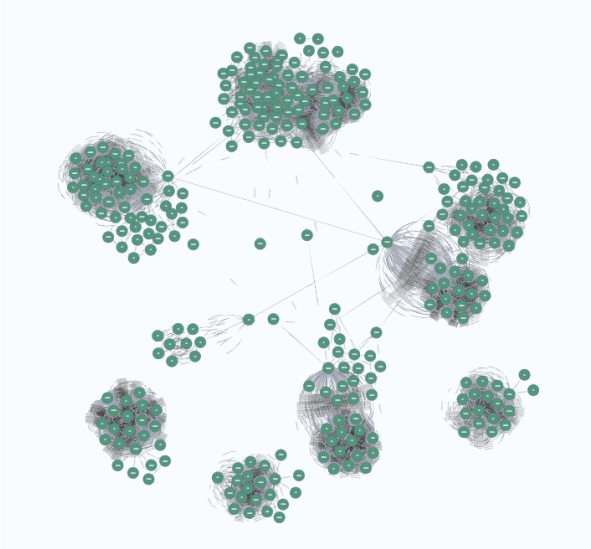

Figura 1: Representación nodos contenidos en listas éxitos

La figura 1 muestran tan sólo nodos que pertenecen a listas de éxitos con colaboraciones, se pueden identificar varias comunidades. Lo cual quiere decir que el hecho de pertenecer a una lista puede potenciarse si se encuentra en una comunidad de artistas donde se colabora.

Veamos ahora que ocurre para el caso contrario, es decir si no se tiene ninguna canción en a lista de éxitos.

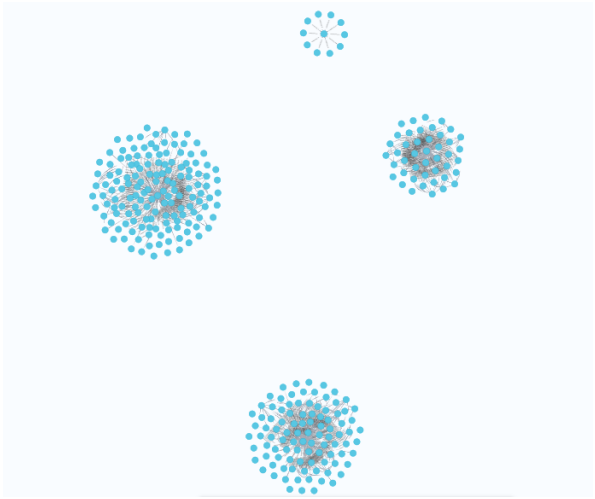

Figura 2: Representación nodos no contenidos en listas éxitos

Para este caso tal y como se muestra en la figura 2 vemos como siguen existiendo esas comunidades de artistas pero no existen relaciones entre estas comunidades lo que puede hacer que no llegue a tantos oyentes si no que se quede en pequeños nichos y no llegue a ser un éxito en las listas del país. Esto significa que habrá una alta modularidad en el grafo que conforma estos nodos.

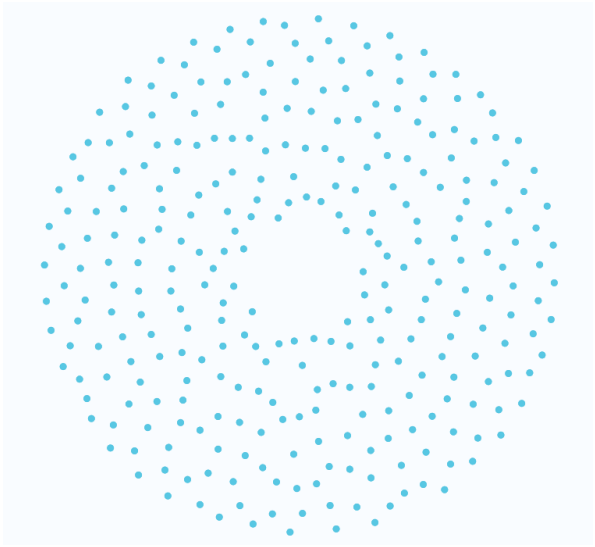

Figura 3: Representación nodos aislados contenidos en listas de éxitos

Algo que puede ser de interés es identificar artistas con éxitos que no presenten colaboraciones con ningún artista. Las colaboraciones hacen llegar a más oyentes pero también es importante identificar los casos aislados.

En la figura 3 vemos como hay artistas con canciones en listas de éxitos que no han colaborado con ningún artista, veamos si la mayoría tiene un grado alto de popularidad que pueda explicar la llegada a estas listas sin necesidad de colaboraciones.

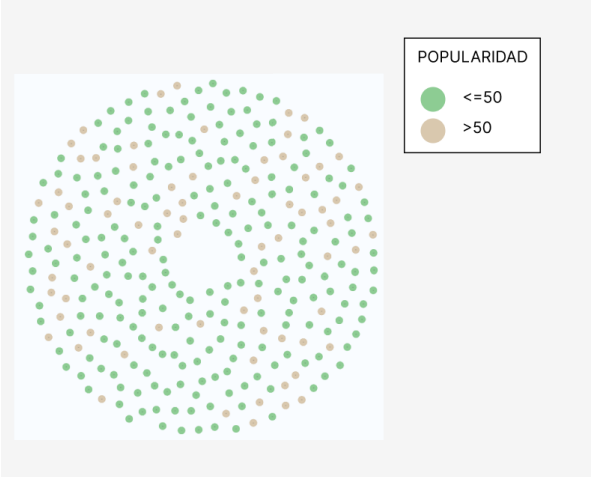

Figura 4: Representación artistas aislados diferenciados

En la figura 4 vemos en marrón aquellos artistas con canciones en listas de éxitos que tienen una popularidad mayor 50, el número de nodos que lo cumplen no es muy elevado.

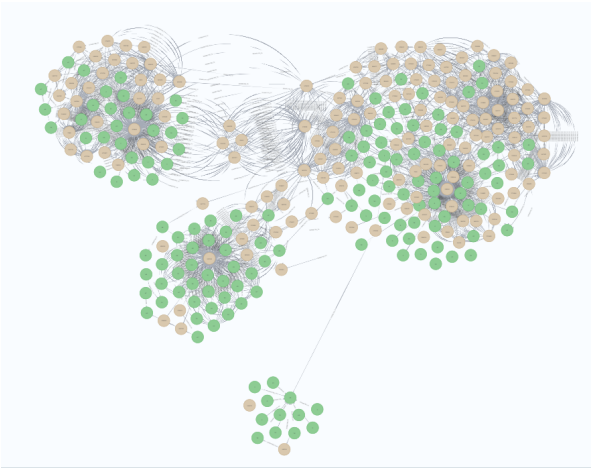

Figura 5: Representación nodos con éxito o con relación con éxito con popularidad

Lo que lleva a pensar que si los artistas representados en la figura 4 han alcanzado el éxito no ha sido debido a su popularidad.

En la figura 5 están representados los nodos que están en listas de éxito o han hecho colaboraciones con artistas exitosos. En este caso vemos como hay proporcionalmente más nodos con popularidad alta. Pero también como los nodos más populares tienden a estar más conectados. Esto es importante de cara a las predicciones pues podrá haber una relación entre el éxito y la popularidad.

A continuación, vamos a ver como se distribuyen los artistas en función de los géneros a los que pertenezcan.

En la figura 6 vemos como hay bastante variedad en lo que se refiere a números de géneros por artista(nodos). Hay 5 categorías a las que puede pertenecer. En la leyenda podemos ver el rango a que hace referencia cada color. Pero hay un patrón que se repite al menos para esta selección de nodos, los artistas exitosos tienen más de un género.

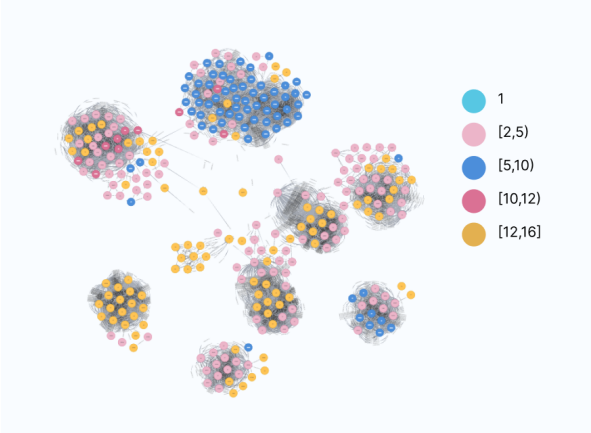

Figura 6: Representación nodos con éxito o con relación con éxito con géneros

Aclarar que todas las figuras son una muestra aleatoria pues debido al alto número de nodos y conexiones se hace imposible una representación total en *Neo4j*. El número de artistas con pertenecen a un único género es lo más multitudinario. Podemos sacar en conclusión

3.2. Correlación entre las variables

Antes de comenzar a ver los resultados de las predicciones en profundidad vamos a hacer un primer análisis de las variables. Para ello, veremos la correlación entre las variables. Para medir la correlación se van a utilizar los índices distintos en función de la naturaleza de los datos.

3.2.1. Índice de correlación de Pearson

El coeficiente de correlación de *Pearson* [10] mide la correlación entre dos variables continuas. Un valor igual a 0 significa que no hay relación entre las variables, cercano a 1 que la relación es alta y positiva y cercano a -1 que la relación es alta y negativa. Que la relación sea positiva significa que el aumento de una variable aumenta la otra y una relación negativa que el aumento de una supone la disminución de la otra y viceversa. En la figura 7 se ve que as variables que más relación tienen con que sea categorizado como éxito o no son: el número de géneros, la popularidad y el coeficiente de grado. Algo predecible es la alta correlación entre la cercanía y la centralidad harmónica pues son inversas la una de la otra. Y la popularidad y el coeficiente de grado también estarán correlacionadas. Por último la correlación más alta se encontrará entre la popularidad y el número de géneros al que pertenece, cosa que ya se pudo intuir en la figura 6.

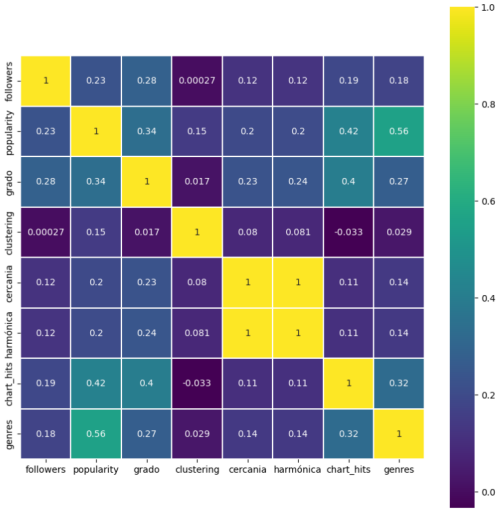

Figura 7: Índice de *Pearson*

Se ha incluido también la variable *chart_hits* pues pese a ser una variable de clasificación binaria, también se puede entender como una probabilidad. Además, como en parte de los siguientes experimentos se utilizará como variable objetivo puede aportar información sobre qué atributos que podrán ser de interés para la construcción de los modelos de predicción.

4. Medidas de rendimiento

Antes de comenzar las predicciones se detalla como se medirá el rendimiento de los modelos de clasificación binaria. A partir de la matriz de confusión [9] 4:

	Clase predicha	
Clase correcta	**Positiva**	**Negativa**
Positiva	Verdaderos positivos (VP)	Falsos negativos (FN)
Negativa	Falsos positivos (FP)	Verdaderos negativos (VN)

Cuadro 3: Matriz de Confusión

Estas son las medidas que se utilizarán:

- Tasa de verdaderos positivos o sensibilidad o *recall*: es la proporción de ejemplos clasificados correctamente. Se calcula de la siguiente manera:

$$tpr = \frac{VP}{FP + VN}$$

- Tasa de verdaderos negativos o especificidad: proporción de ejemplos negativos clasificados correctamente. Se calcula de la siguiente manera:

$$tnr = \frac{VN}{FP + VN}$$

- Precisión: ejemplos realmente positivos clasificados como positivos. Se calculará así:

$$prec = \frac{VP}{VP + FP}$$

- Exactitud o *accuracy*: es el porcentaje de veces que el modelo predice correctamente y se calcula de la siguiente manera:

$$accuaracy = \frac{\text{predicciones correctas}}{\text{total de predicciones}}$$

- *F1 score*: es una medida que combina dos medidas expuestas anteriormente *recall* y precisión. Se calcula de la siguiente manera:

$$F1\ socre = \frac{2 \cdot \text{Precisión} \cdot \text{Recall}}{\text{Precisión} + \text{Recall}}$$

A la hora de medir el rendimiento en modelos regresivos utilizaremos las siguientes métricas:

- Coeficiente de determinación [7]:es la proporción de la variación en la variable dependiente que es predecible a partir de la variable o variables independientes. En otras palabras, mide cuan bien el modelo es capaz de explicar la variabilidad de la variable dependiente o variable objetivo. Como de bien predice el modelo. Se calcula así:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

  Donde:

  - R^2 es el coeficiente de determinación.
  - n es el número total de observaciones en tu conjunto de datos.
  - y_i son los valores observados.
  - $\hat{y}_i$ son los valores predichos por el modelo.
  - $\bar{y}$ es la media de los valores observados o etiqueta esperada.

- Error cuadrático medio (MSE) [8]: mide la media de los cuadrados de los errores, es decir, la diferencia cuadrada promedio entre los valores estimados y el valor real. Se calcula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

  Donde:

  - MSE es el Error Cuadrático Medio.
  - n es el número total de observaciones.
  - y_i son los valores reales observados.
  - $\hat{y}_i$ son los valores predichos por el modelo.

En casos en los que se trate de una tarea de clasificación pero no sea binaria se utilizará *accuracy*, precisión, *recall* y *F1-score*.

5. Primeras predicciones

5.1. Árboles de clasificación y regresión

Para las predicciones se utilizará como modelo un árboles de decisión. Un árbol de decisión es un algoritmo de aprendizaje supervisado que se utiliza para la clasificación y la regresión. La regresión es un método utilizado para el modelado predictivo, por lo que estos árboles se utilizan para clasificar datos o predecir valores futuros. La razón por la que se ha escogido este modelo es porque es interpretable, es decir, a partir de una representación de la estructura del modelo ya entrenado podemos comprender el porqué de la toma de sus decisiones.

Los árboles de decisión 8 están formados por:

- Nodos interiores: que será los atributos
- Arcos: posibles valores del nodos origen.
- Hojas: valor de clasificación. En el caso de árboles de regresión se consiguen valores en las hojas lo suficientemente para que sean aproximadamente constantes.

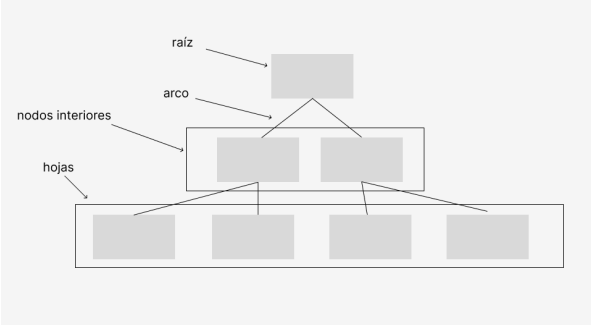

Figura 8: Ilustración árbol de decisión

Los árboles de clasificación son un tipo de árbol de decisión, el objetivo que se predice es la clase a la que el dato pertenece, por ejemplo clasificar un conjunto de imágenes en perro, gato y caballo. Podrán ser binarios si clasifican entre dos categorías o multiclase si clasifican entre más de 2 categorías.

Los árboles de regresión son otro tipo de árbol de decisión en el que en el que el objetivo a predecir es una variable continua.

5.2. Predicción de artistas exitosos

La primera tarea de predicción que se propuso fue la de predecir si un artista había pertenecido a alguna lista de éxitos por medio de sus atributos. Puesto que era una tarea de clasificación binaria se decidió llevarla a cabo con un árbol de decisión de clasificación.

Para ello habría que hacer algunas modificaciones a los datos, y fueron las siguientes:

- Si en la columna *name* hay un valor vacío le daremos el valor desconocido.
- Si en la columna *followers* hay un valor vacío le asignaremos el valor 0.
- Asignaremos a todos esos registros que tienen un valor no nulo en la columna **chart_hits** el valor 1 indicando que ha pertenecido a alguna lista de éxitos y el valor 0 el caso de que sea nulo.
- A la hora de categorizar los géneros, puesto que son listas de géneros, se han creado 6 categorías.
  - *0*: Ningún género.
  - *1*: Un género.
  - *2-5*: de 2 (incluido) a 5 géneros (no incluido) .
  - *5-10*: de 5 (incluido) a 10 géneros (no incluido) .
  - *10-12*: de 10 (incluido) a 12 géneros (no incluido) .
  - *12-16*: de 12 (incluido) a 16 géneros (incluido) .

Posteriormente transformaremos la columna *followers* a decimal y la columna *popularity* a entero. Ahora mismo el conjunto de datos está listo para ser categorizado. Una vez hecho se pasa a dividir el conjunto de datos en entrenamiento y test, utilizando los valores por defecto de la librería en este caso 75 % para entrenamiento y 25 % para validación. En esta primera prueba se consiguen los siguientes resultados:

Accuracy:	0,89296
Precisión:	0,64351
Recall:	0,36801
Especificidad:	0,97006
F1-Score:	0,46824

La exactitud o *accuacy* obtiene el segundo resultado más alto lo que quiere decir que en un gran porcentaje de las ocasiones predice correctamente tanto los ejemplos positivos como negativos.

Pasando a la precisión solo tiene en cuenta la correcta clasificación de ejemplos positivos entre los que se categorizan falsamente positivos y los que realmente lo son. Es más baja que la *accuracy* lo que nos lleva a suponer que el modelo tiende a clasificar incorrectamente instancias positivas.

Obtiene un *recall* alrededor del 30 por ciento, lo que quiere decir que el número de verdaderos positivos es bajo con respecto a los clasificados como falsos negativos y verdaderos negativos, por lo que el modelo estaría clasificando correctamente menos ejemplos de los que clasifica como negativos e incorrectamente como positivos. Pasando a la especificidad se obtiene el mejor resultado, lo cual es coherente con la afirmación anterior, el modelo tiende a clasificar mejor los verdaderos negativos antes que clasificar correctos positivos y falsos positivos.

Por último *F1-score*, con el segundo peor resultado, para una mejor interpretación vamos a calcular el número de ejemplos del total de los datos de validación son ejemplos positivos y cuáles son ejemplos negativos. El número de positivos es igual a 4903 y en número de negativos es igual a 34203. El número de ejemplos negativos el del orden de 7 veces más grande que el de positivos, lo cual explica los resultados obtenidos. La medida más representativa será *accuracy*.

Para sólo haber tenido en cuenta datos propios del conjunto de datos está bien, aunque habría que tener en cuenta que se trata de una tarea de clasificación binaria en la que un modelo que predijese aleatoriamente obtendría un 50 % de acierto. Es en este momento cuando se buscan formas de mejorar las predicciones.

Esta es una representación del modelo generado, aprovechando la interpretabilidad de los árboles de decisión veamos que atributos y en qué medida se han tenido en cuenta para la clasificación de las instancias.

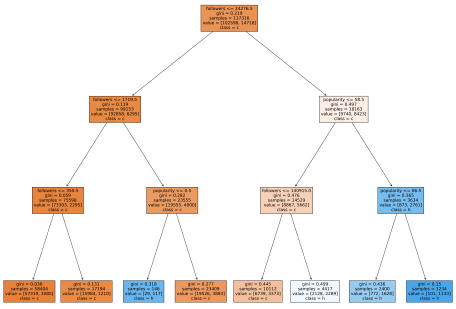

Figura 9: Representación del primer árbol.

En la figura 9 se ve como lo primero por lo que se diferencian dos conjuntos es el número de seguidores, después pasa a ser la popularidad y los seguidores de nuevo. En definitiva, las decisiones se toman a partir del éxito que confiere un mayor número de seguidores o ser considerado popular. Un mayor número de seguidores significa un mayor número de oyentes y por tanto una mayor probabilidad de que lo que esté produciendo ese artista sean grandes éxito. Por tanto atributos que potencien la popularidad y lo seguidores también estarán directamente relacionados con el éxito del artista. Esto ya se podía ver en la figura 7 en el índice de *Pearson* entre esas variables y la objetivo.

El hecho de que los datos estén siendo representados en grafos nos aporta la información sobre las relaciones entre sus nodos lo cuál está muy relacionado con aquellas propiedades que potenciarán el éxito.

Puesto que ya se habían calculado las nuevas medidas derivadas de las colaboraciones como son: el coeficiente de *clutering*, la centralidad de grado... se incluirán en los datos con los que se entrenará el modelo y se validará, tras haber visto una representación en forma de grafo de los artistas exitosos, se puede afirmar que hay ciertos patrones. Es decir, artistas que se encuentran en comunidades colaboradoras o *clusters* tienden a tener éxito y un mayor grado de popularidad entre otras características.

Tras incluir esas tres medidas en el *dataset* se consigue un rendimiento mucho mayor:

Accuracy:	0,94492
Precisión:	0,89352
Recall:	0,64473
Especificidad:	0,98878
F1-Score:	0,74901

Los resultados son mejores que cuando no se tenían en cuenta los atributos derivados de las colaboraciones, veamos cuanto ha cambiado el número de falsos negativos, verdaderos positivos... con respecto a la primera predicción.

Modelo de Predicción	**Primer Modelo**	**Segundo Modelo**
Número de Verdaderos Negativos	33854	33826
Número de Falsos Positivos	1074	382
Número de Falsos Negativos	3077	1789
Número de Verdaderos Positivos	3073	3109

Del primer modelo al segundo hay una mejora en la predicción de falsos positivos y falsos negativos, es por esto que aumenta su *accuracy* pues más instancias se clasifican correctamente. La precisión también mejora pues el número de falsos positivos disminuye drásticamente. En general el modelo mejora considerablemente en no predecir incorrectamente más que en predecir más instancias correctamente. Representamos el árbol aprovechando la interpretabilidad de este modelo para comprobar como estás nuevas columnas han modelado las decisiones del modelo.

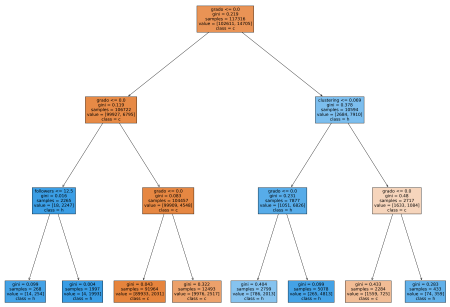

Figura 10: Representación segundo árbol

Como se indica en su representación 10 las centralidades que se han tenido en cuenta son la de grado y el coeficiente de *clustering* y en una de sus ramas los seguidores. Una mayor centralidad de grado significará que tiene un mayor número de conexiones que la media de los nodos del grafo, este mayor número de conexiones se traduce como un mayor número de colaboraciones. A más colaboraciones sus canciones aparecerán en los perfiles de más artistas. De este modo, sus canciones no sólo serán escuchadas por sus seguidores si no también por los seguidores de los artistas con los que colaboran y su música tendrá un mayor alcance, llevando al aumento de seguidores y la popularidad.

Pasando al coeficiente de *clustering*, aumenta si la nodos a los que está conectado un nodo están altamente conectados entre sí. Un alto coeficiente de *clustering* deriva a que se cree una comunidad de nodos altamente conectados entre sí. De esta forma los oyentes de un artista tienen más probabilidades de conocer no sólo los artistas con los que ha colaborado su artista favorito si no también las colaboraciones de sus colaboraciones. Esto de nuevo potencia los seguidores y la popularidad.

6. Introducción de nuevos atributos

6.1. Población

Otro aspecto muy a tener en cuenta es el número de seguidores y esto lo relacionaremos con un nuevo atributo que añadiremos que son los oyentes potenciales.

Si un artista aparece en una lista de éxitos de un país será probable que sea descubierto por más oyentes, pero además en función de la población del propio país podrá haber haber más posibilidades que que lo escuchen más personas o menos. Es por ello que en este punto se buscó un conjunto de datos que recogiese de cada país su abreviatura, pues las listas están nominadas siguiendo el siguiente patrón `"siglas_país (número)"`, por ejemplo, para estados unidos sería `üs (1)"`.

Después de añadir esa columna al conjunto de datos se consigue un rendimiento del 100 % en el modelo y el árbol de decisión quedaría de la siguiente manera.

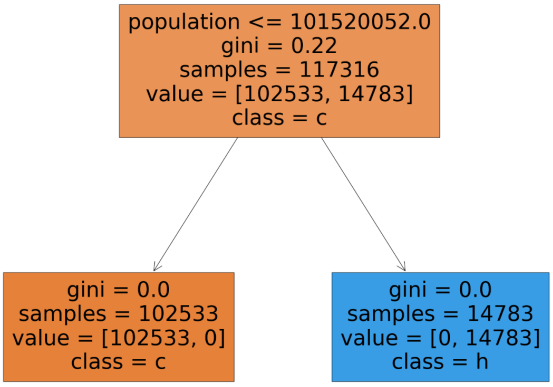

Figura 11: Representación tercer árbol

En la figura 11 vemos como únicamente teniendo en cuenta el atributo recién añadido se puede conseguir un modelo que acierte siempre. No obstante, no es algo que sorprenda pues estamos usando una propiedad derivada del atributo que queremos predecir para predecirlo. como a mayor número de listas es más probable que sea mayor la población pues es coherente que se rija sobre eso el si un artista pertenece a una lista de éxitos o no.

Si lo que se calculase en vez de ser si pertenece a una lista de éxitos o no fuese cualquiera de los otros atributos de os que no deriva, puede que se valiese de los otros atributos del *dataset*. Por tanto estas predicciones no tienen ningún valor, no obstante más adelante se utilizarán los oyentes potenciales para otras predicciones cambiando la variable objetivo.

7. Cambio del objetivo de la predicción

En las secciones anteriores se puso como objetivo de predicción el si un artista pertenecería a una lista de éxitos o no. En este caso se ha querido calcular el número exacto de listas a las que pertenecería. Para ello se han tratado los datos para que en vez de contar con con una lista de las listas a las que ha pertenecido, obtener el número de listas a las que perteneció. Además se siguieron incluyendo las medidas relativas a las relaciones entre sus nodos como son: el coeficiente de *clustering*, el autovalor... pues ya se comprobó que conferían una mejora en los resultados pues las colaboraciones entre artistas son un aliciente para el éxito.

En este momento se plantea la siguiente problemática, ¿sería mejor para esta tarea utilizar un modelo de regresión? Calcular el número de listas a las que podría pertenecer un artista no es una tarea que se pueda considerar de clasificación pues no hay un número exacto de categorías estipuladas. Es cierto que si con los datos con los que contamos construyésemos un modelo de clasificación como es un árbol de decisión de clasificación, acabaría cogiendo como categorías el número de listas presentes en el conjunto de entrenamiento, lo que podría derivar en un sobreajuste.

En primer lugar comenzamos por el árbol de decisión de clasificación. Estos serían los resultados obtenidos.

Accuracy:	0,91503
Precisión:	0,87859
Recall:	0,91503
F1-Score:	0,89643

Tiene una alta exactitud y el resto de métricas también alcanzan valores bastante altos a diferencia de los que se objetivo en la tarea de predicción binaria. No obstante, al comparar la variedad de valores reales con los predichos se observa que el modelo acaba limitando el número de posibles clases a dos.

$$\begin{aligned}\text{reales} = \{&0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19,\\ &20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37,\\ &38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 49, 50, 51, 52, 53, 54, 55, 56,\\ &57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71\}\end{aligned}$$

$$\text{predichos} = \{0, 1\}$$

Debido a que en la mayoría de casos los artistas tienen 0 o 1 género el modelo ha simplificado entre estos dos valores en vez de intentar abarcar todos los casos. La precisión no es baja debido a que los artistas que son mal clasificados son mucho menos que los que pueden aspirar a una buena clasificación pues su valor objetivo está dentro de la hipótesis del modelo.

Por lo que este buen resultado es una falsa esperanza, llegado el momento de calcular predicciones para un conjunto de datos nuevos y se diese que la mayoría de artistas tuviesen más de un género, el rendimiento sería muy pobre. Esto es un claro ejemplo de sobreajuste.

En la figura 12 vemos como los atributos a tener en cuenta son la popularidad, el grado y el coeficiente de *clustering*. Muy similar a los atributos tenidos en cuenta en 10.

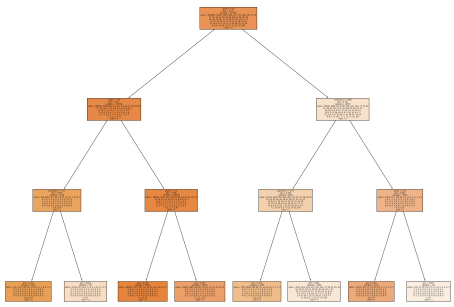

Figura 12: Representación cuarto árbol

Posteriormente se pasa a probar con un árbol de decisión de regresión, debería esperarse una mejora en los resultados a obtener, pero se obtiene un rendimiento muy bajo. Estos serían los resultados:

Coeficiente de determinación:	− 0,08085075050777091
Error cuadrático medio:	13,717514431129166

Se obtiene un bajo coeficiente de determinación lo que significa que el modelo no explica bien la variabilidad de la variable objetivo. Además teniendo en cuenta que para los datos de validación van desde 0 a 71 un error cuadrático medio bastante alto.

Se añade la representación gráfica de las predicciones y los valores reales 13.

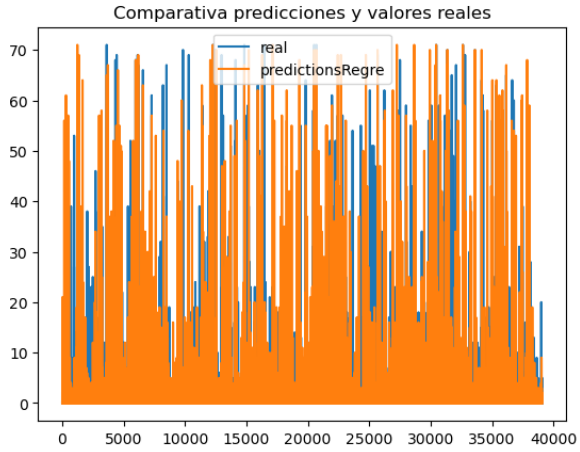

Figura 13: Representación comparativa predicicones árbol de regresión

No se ha incluido una ilustración del modelo para intentar interpretarlo pues está formado por mucha profundidad, y por ende hojas. Por tanto los resultados predichos se alejan mucho de los valores reales.

Se prueba otro algoritmo de regresión en este caso un *ensemble*.

7.1. Ensemble

Esta clase de aprendizaje utiliza múltiples algoritmos de aprendizaje mejorando así el rendimiento del modelo frente a los que tan sólo utilizan un único algoritmos de aprendizaje.

Un ejemplos de *ensemble* serie el modelo *Random Forest*, que consiste en, la combinación del resultado de varios árboles de decisión en un único resultado. De ahí que reciba el nombre de bosque. Ver figura 14

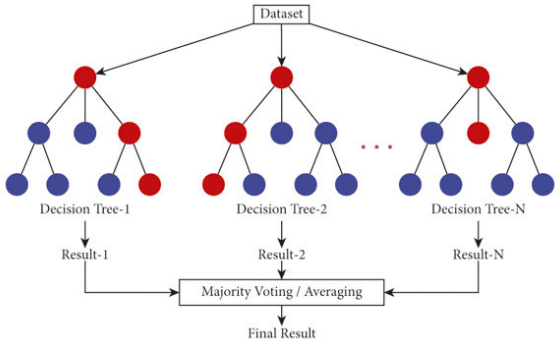

Figura 14: Ilustración explicativa modelo *Random Forest*

7.2. Pruebas con *Random Forest*

Viendo el bajo rendimiento del árbol de regresión, se prueba con el modelo *Random Forest*. Al igual que pasaba con los árboles de decisión, podemos diferenciar entre modelos de regresión y modelos de clasificación. Se probará con los dos tipos para ver si el bajo rendimiento provenía de la simplicidad

de los modelos de árbol de decisión o porque realmente con los datos de los que se dispone y la tarea que se pretende hacer no es posible con un modelo regresivo.

Después de hacer las predicciones se consiguen un mejor rendimiento para el modelo de clasificación que el regresivo:

Coeficiente de determinación:	0,4447286478480238
Error cuadrático medio:	7,739443818942585

Es cierto que la tasa de acierto ronda casi el 50 por ciento para el modelo de regresión de *Random Forest*, un porcentaje mucho mayor frente a lo que se consiguió con el Árbol de regresión. En lo que se refiere al error cuadrático medio es un valor bastante alto teniendo en cuenta que los valores dentro de la variable objetivo van de 0 a 100.

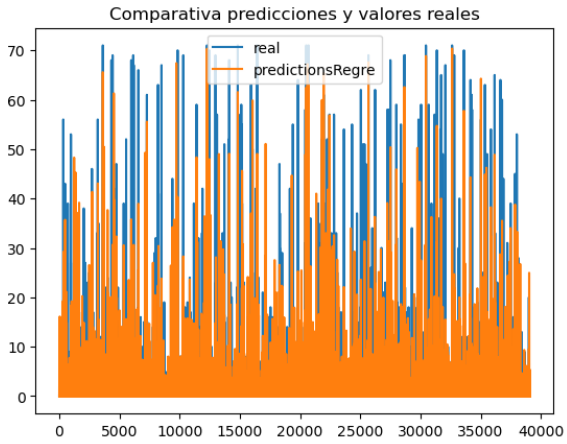

Figura 15: Representación comparativa predicciones árbol de regresión*Random Forest*

Este modelo puede parecer que tiene una menor tasa de acierto por su representación, pero lo que realmente ha sucedido es que este modelo no sobrestima y el otro sí.

Por último veamos los resultados de *Random Forest Classifier*, es decir el modelo de clasificación de *Random Forest*.

Accuracy:	0,91482
Precisión:	0,89135
Recall:	0,91482
F1-Score:	0,90243

Los resultados son casi idénticos a los obtenidos con el árbol de clasificación.

8. Predicción de la popularidad a partir de nuevos atributos

Se vio interesante aprovechar el cálculo de los oyentes potenciales extraídos a partir de las listas a las que pertenece cada artista. Todo el preprocesado de los datos ya había tenido lugar en experimentos anteriores. Tan sólo habría que hacer las predicciones.

Debido a la naturaleza de la variable *popularity* se puede entender de dos formas: como una tarea de regresión, pudiendo ir el resultado desde cero hasta 100, o como una tarea de clasificación multiclase. Se probarán las distintas alternativas y si discutirán los resultados.

8.1. Búsqueda de hiperparámetros

Son técnicas que nos permiten obtener en base a una matriz de hiperparámentros aquella combinación que conferirá el mejor rendimiento al modelo. En este estudio trataremos con dos técnicas distintas: búsqueda en rejilla (*Grid Search*) y búsqueda aleatoria (*Random Search*) .

8.1.1. Grid Search

En esta técnica primero se definirá una matriz de hiperparámetros donde se incluyen todos los valores de los hiperparámetros que querremos usar para construir el modelo.

Esta técnica hace una búsqueda exhaustiva sobre todas las posibles combinaciones de valores de esa matriz de parámetros. Es por esto que será computacionalmente costoso.

8.1.2. Random Search

Esta técnica a diferencia de la búsqueda en rejilla no busca sobre todas las posibles combinaciones de la matriz de parámetros. Va seleccionando combinaciones aleatorias de de hiperparámetros.

Según como funciona podemos deducir que como cualquier algoritmo con aleatoriedad será menos costoso computacionalmente.

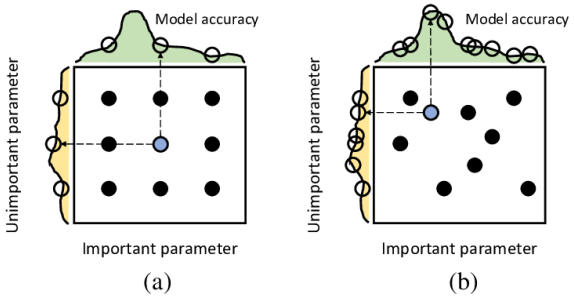

Figura 16: Ilustración comparativa *Grid Search* y *Random Search*

8.2. Árbol de decisión de regresión

En primer lugar se predice usando un árbol de decisión de regresión. Se quiso utilizar técnicas de búsqueda de hiperparámetros para escoger la mejor profundidad del árbol dentro de los valores dados. La razón de esto es debido a que si no se le restringía la profundidad del árbol acababa haciendo árboles muy profundo que tendían a al sobreajuste. Además se añadirán nuevos valores a hiperparámetros como el número mínimo de instancias por hoja para evitar que las divisiones se hagan demasiado específicas.

En primer lugar, usándose la técnica de la búsqueda en rejilla se pasa como parámetros las siguientes profundidades de árbol: $[3, 4, 5, 6, 7]$, el mejor rendimiento se consigue para la profundidad 7 y estos son los resultados obtenidos sobre el conjunto de validación:

Coeficiente de determinación:	0,7469
Error cuadrático medio:	85,0349

Puesto que el mejor rendimiento se obtiene para la profundidad 7 podría caber la posibilidad e que un aumento en la profundidad siga confiriendo mejores resultados, por ello vamos utilizar la búsqueda aleatoria para aportar más valores al parámetro profundidad y comprobar si una mayor complejidad sigue confiriendo mejores resultados.

Esta sería la matriz de hiperparámetros:

max_depth	min_samples_leaf
8	200
9	100
10	300
11	400
12	-
13	-
14	-

La configuración con mejor rendimiento es con profundidad 9 y 100 instancias mínimas por hoja. Estos serían los resultados:

Coeficiente de determinación: 0,7459
Error cuadrático medio: 85,3510

El modelo aprende bastante bien la variabilidad de la variable objetivo, no obstante su error cuadrático medio es bastante alto. Esto está relacionado con que la variabilidad entre los posibles valores de la variable objetivo es bastante grande pues va de 0 a 100. Por tanto, significaría que de media la variabilidad entre lo que predice y lo real sería igual a $\sqrt{85{,}35.....}$ lo que es alrededor de 9,23..., que puede verse como alto pues se equivocaría en 10 puntos.

En vez de representar en una ilustración del árbol hemos extraído del mismo la importancia de los atributos obtenidos en la clasificación y estos son los resultados obtenidos. El hecho de que se utilice el número de seguidores como atributo para dividir los conjuntos es significativo pero coherente. En la figura 7 se muestra como la relación entre los seguidores y la popularidad es más alta que con cualquier otro atributo. Siguen siendo valores muy bajos, por lo que la correlación no es directa pero no deja de ser mayor que con el resto de propiedades. Además de los seguidores también se basa en el coeficiente de grado, que con vemos también tiene una correlación por encima de la media.

Importancias de los atributos:

followers: 0,9598
chart_hits: 0,0009
genres: 0,0020
grado: 0,0319
eigen: 0,0002
clustering: 0,0023
cercania: 0,0009
harmónica: 0,0008
population: 0,0012

Los atributos que gana más importancia serán los seguidores en primer lugar y posteriormente pertenecer a una lista de éxitos, los géneros, la centralidad de grado,la centralidad *Eigen*, el coeficiente de *clustering*, la centralidad de cercanía y la población. que la centralidad de *Eigen* o autovalor sea considerado es coherente pues a mayor influencia de un nodo se puede considerar que es más popular. No obstante, de todas las centralidades a la que se le da más importancia a la centralidad de grado, por tanto, podemos suponer que a más colaboraciones con artistas se considerará más popular. Algo que sorprende es el alto peso que tienen los seguidores cuando en la figura7 se ve como que el Coeficiente de grado tiene una mayor correlación con la popularidad.

8.3. Árbol de decisión de clasificación

A continuación se prueba con un árbol de clasificación, se esperan peores resultados pues a mayor número de clases es más probable que el modelo sea propenso al sobreajuste. El número de ejemplos de las distintas alternativas será mucho más bajo al tener que dividir el total de datos en 100 categorías distintas. Teniendo esto en cuenta, se muestra los resultados obtenidos:

Accuracy:	0,31210
Precisión:	0,76548
Recall:	0,31210
F1-Score:	0,44266

El mejor resultado lo tiene para la precisión. Pero el rendimiento del modelo de regresión sigue siendo mejor.Debido a la gran dimensionalidad del árbol no vamos a poder mostrar su estructura en una figura, pero podemos ver la importancia de los atributos. Da máxima importancia al atributo seguidores por lo que podemos deducir que basa sus decisiones únicamente en este atributo.

9. Conclusiones

Tras el desarrollo de este estudio se ha visto la morfología de los grafos que forman los nodos que cumplen cierta característica. Al representar los nodos contenidos en listas de éxitos se crearán comunidades exitosas. Los nodos no contenidos en listas de éxitos crean comunidades con baja comunicación entre nodos de otras comunidades, por lo que podemos presuponer que tendrán una alta modularidad. En relación al estudio de la popularidad y el éxito se ve una alta proporción de artistas con popularidad por encima de la media cuando se trata de artistas exitosos. Por último, para nodos con éxitos o colaboradores de artistas con éxito pertenecen a más de 1 género.

Tras hacer un estudio sobre la correlación entre variables mediante el índice de *Pearson*, vemos una alta correlación del éxito entre la popularidad y el coeficiente de grados. Y una alta correlación entre la popularidad y el número de géneros al que pertenecerá.

El uso de las medidas derivadas de las relaciones de los nodos nos permite obtener información inherente al grafo que pueda ser pasada como parámetro a un árbol de decisión. Son bien conocidas las *Graph Neural network* pero en este estudio se ha querido primar la interpretabilidad del modelo frente a un mejor rendimiento, además los atributos del *dataset* eran muy descriptivos y con un significado fácil de deducir por lo que se aprovechó esta oportunidad. para usar Árboles de decisión.

Es importante adaptar el como medir el rendimiento al tipo de tarea ya sea clasificación o regresión. Es por esto que se ha establecido desde el inicio como se mediría el rendimiento de cada modelo. Además se ha acompañado con gráficas que permitiesen ver realmente la magnitud del error y se ha discutido sobre los errores obtenidos. Hay medidas como el error cuadrático medio a las que un contexto le aporta todo el significado y permite saber si realmente podría ser un error despreciable en función de las magnitudes de los atributos o no.

Además la búsqueda de hiperparámetros tiene un peso bastante importante en la obtención de modelos con buen rendimiento. Existen distintas técnicas que permiten explorar las distintas combinaciones de sus hiperparámetros, y siempre serán más eficientes que probar individualmente en distintos modelos.

Las primeras predicciones tienen buenos resultados pero han añadir los atributos derivados de la construcción manual de *features* relacionales mejora aproximadamente un 5 % en *accuracy* y significativamente la precisión. Su mejora se basa en el número de instancias que predice incorrectamente, pues baja.

Tras el intento de mejora del rendimiento por medio de incluir los oyentes potenciales puesto que los oyentes potenciales derivan de la variable objetivo el rendimiento será del 100 %

Se predice a continuación el número de listas de éxito a las que pertenecerá un artista, pese a que es un claro ejemplo de regresión se prueba a hacer las predicciones con un árbol de clasificación. Los resultados obtenidos con el árbol de clasificación son significativamente mejores con respecto a los obtenidos con el árbol de regresión, debido a un sobreajuste. Pues limita el número de posibles clases a dos y como las instancias en en el conjunto de validación que se salen de ese espacio de hipótesis son pocas su rendimiento no es tan malo como se esperaría.

Por último se aprovecha la nueva variable de oyentes potenciales para predecir la popularidad.