



Universidad Autónoma de Querétaro  
Facultad de Contaduría y Administración  
Licenciatura en Actuaría

CLASIFICACIÓN SUPERVISADA DE SERIES DE TIEMPO MEDIANTE SU  
CARACTERIZACIÓN TEMPORAL A TRAVÉS DE MÉTODOS ACTUARIALES.

**TESIS CONJUNTA**

Que como parte de los requisitos para obtener el grado de  
Licenciado en Actuaría

**Presentan:**

Adriana Regina Huante Vargas y Samantha Aguilar Veyna

**Dirigido por:**

M en C. Wilfrido J. Paredes García

SINODALES

M en C. Wilfrido J. Paredes García  
Presidente

Firma

Dra. Denise Gómez Hernández  
Secretario

Firma

Dr. Eduardo Castaño Tostado  
Vocal

Firma

Dr. Felipe Abelardo Pérez Sosa  
Suplente

Firma

Dr. Samuel Estala Arias  
Suplente

Firma

M en I. Martín Vivanco Vargas  
Director de la Facultad

Dra. Ma. Guadalupe Flavia Loarca Piña  
Director de Investigación y Postgrado

Centro Universitario  
Querétaro, Qro.  
Agosto de 2020  
México

## Resumen

El propósito de esta tesis es determinar si son o no significativos los metadatos usados en la recomendación musical, puesto que esto limita la experiencia musical a géneros y artistas muy similares.

En este trabajo se presenta una propuesta de clasificación de canciones mediante el análisis matemático y estadístico de sus elementos temporales, es decir características propias de la música, generando así una alternativa a algunos clasificadores actuales que tienden a incluir en la caracterización de las canciones elementos ajenos a la música en sí, tales como año, artista, calificación, popularidad entre otras. Para ello se optó por utilizar el método de clasificación *K Nearest Neighbors*.

**Palabras clave:** Recomendación musical, Aprendizaje supervisado, K Nearest Neighbors, Clasificación

## Abstract

The aim of this thesis is to determine whether or not the usage of metadata in musical recommendation is significant, since this restricts the musical experience to very similar artists and genres.

This thesis presents a proposal for music classification through the mathematical and statistical analysis of the temporal elements of the songs, creating an alternative to some current classifiers that include other elements, such as year, artist, rating, among others. In order to create this classifier, the algorithm K-Nearest Neighbors (KNN) was used.

**Key words:** Music recommendation, supervised learning, K Nearest Neighbors, Classification

Dirección General de Bibliotecas UAQ

Esta tesis está dedicada a nuestras familias, especialmente a nuestros padres, quienes siempre nos han mostrado su apoyo, nos brindan su amor y nos motivan día con día. A nuestros amigos, que nos acompañaron a lo largo de este proceso y han estado en las buenas y en las malas.

## Agradecimientos

Agradecemos de una forma muy especial al M en C. Wilfrido J. Paredes García, nuestro director de tesis, por su tiempo, compromiso y por todo el apoyo brindado para que este trabajo fuera posible.

También queremos agradecer a nuestros sinodales, por tomarse el tiempo para revisar la tesis y brindarnos sus acertados comentarios que permitieron mejorar este trabajo.

Una mención muy especial al grupo 17 de actuaría y a todas las personas que nos ayudaron y nos brindaron su tiempo para escuchar canciones y clasificarlas. Gracias a ellos es que se obtuvo la base de datos de la cual partimos para realizar este trabajo.

Finalmente, queremos agradecer a la Universidad Autónoma de Querétaro y a la Facultad de Contaduría y Administración por brindarnos las herramientas y conocimientos necesarios para forjarnos profesionalmente.

---

## Índice general

<b>Resumen</b>	<b>I</b>
<b>Abstract</b>	<b>II</b>
<b>Dedicatoria</b>	<b>III</b>
<b>Agradecimientos</b>	<b>IV</b>
<b>1. Introducción</b>	<b>1</b>
1.1. Hipótesis y objetivos . . . . .	2
1.1.1. Hipótesis . . . . .	2
1.1.2. Objetivos . . . . .	3
<b>2. Marco teórico</b>	<b>4</b>
2.1. Recomendación musical . . . . .	4
2.1.1. Las canciones como series de tiempo . . . . .	5
2.2. Clasificación . . . . .	10
2.2.1. Aprendizaje supervisado . . . . .	11
2.2.2. Aprendizaje no supervisado . . . . .	13
<b>3. Metodología</b>	<b>15</b>
3.1. Selección aleatoria de una muestra de canciones . . . . .	15
3.2. Análisis estadístico . . . . .	16
3.2.1. Velocidad y aceleración . . . . .	17
3.2.2. Ciclos y periodos . . . . .	21
3.2.3. Varianza . . . . .	24
3.3. Desarrollo del algoritmo . . . . .	26
3.3.1. Evaluación del clasificador . . . . .	31
<b>4. Resultados</b>	<b>33</b>
4.1. Análisis de sensibilidad . . . . .	33
4.2. Análisis estadístico . . . . .	35
4.2.1. Velocidad y aceleración . . . . .	35
4.2.2. Ciclos y periodos . . . . .	36
4.2.3. Varianza . . . . .	37
4.3. Resultados del clasificador . . . . .	40
4.4. Evaluación del clasificador . . . . .	41
<b>5. Discusión y conclusiones</b>	<b>47</b>
5.1. Discusión . . . . .	47
5.2. Conclusiones . . . . .	47
<b>A. Algoritmo</b>	<b>49</b>

Dirección General de Bibliotecas UAQ

## Índice de tablas

3.1. Comparación de la velocidad . . . . .	18
3.2. Comparación de la aceleración . . . . .	20
3.3. Comparación de los ciclos . . . . .	22
3.4. Comparación de los periodos . . . . .	23
3.5. Comparación de las varianzas. . . . .	24
3.6. Representación de la matriz resultado . . . . .	29
3.7. Cálculo del número óptimo de canciones . . . . .	31
4.1. Etiquetas óptimas de clasificación “ <i>me gusta pero no guardaría la canción</i> ”	34
4.2. Valor $k$ óptimo . . . . .	35
4.3. Correlaciones. . . . .	36
4.4. Media de los ciclos y los periodos de la frecuencia. . . . .	37
4.5. Varianza. . . . .	38
4.6. Obtención del óptimo de canciones. . . . .	41
4.7. Resultados obtenidos comparando las primeras 50 canciones recomendadas y Spotify . . . . .	43
4.8. Resultados obtenidos comparando el número de canciones óptimo y Spotify	45

## Índice de figuras

2.1. Frecuencia de muestreo.Fuente: Mastering the mix . . . . .	6
2.2. Frecuencia de los primeros 30 segundos de una canción . . . . .	7
2.3. Periodograma de una canción . . . . .	10
2.4. Tipos de aprendizaje . . . . .	11
2.5. Ejemplo de KNN . . . . .	13
3.1. Comparación de la velocidad . . . . .	19
3.2. Comparación de la aceleración . . . . .	20
3.3. Comparación de los ciclos . . . . .	22
3.4. Comparación de los periodos . . . . .	23
3.5. Distancia a la canción original . . . . .	26
4.1. Varianza, media de ciclos y media de periodos . . . . .	39
4.2. Varianza, media de aceleración y media de velocidad . . . . .	40
4.3. Resultados de la prueba t comparando las primeras 50 canciones recomen- dadas y Spotify. . . . .	44
4.4. Resultados de la prueba t comparando el número de canciones óptimo y Spotify . . . . .	46
B.1. Conjunto de datos empleados y se puede consultar usando el siguiente DOI: 10.6084/m9.figshare.11933157. . . . .	52

Dirección General de Bibliotecas UAQ

# CAPÍTULO 1

## Introducción

La manera en la que las personas escuchan música ha ido evolucionando a la par con la tecnología. En un principio se escuchaba música en acetatos, para más adelante pasar a ser escuchada en Discos Compactos (CD's) y casetes, pero hoy eso se ha vuelto obsoleto y se han desarrollado aplicaciones en las cuales una persona puede tener acceso a millones de canciones de diferentes géneros.

Cuando las personas escuchaban música en acetato, CD's o casetes, estos eran pertenecientes a un género o artista específico, por lo que sus gustos estaban muy limitados a la música que consumían regularmente. Sin embargo, en la actualidad las aplicaciones para escuchar y descargar música son muy accesibles para aquellas personas que tienen acceso a internet y a los sitios o aplicaciones de música, por lo que se tiene un acceso casi ilimitado a cualquier tipo de música. Esto ha generado una diversificación en los gustos de los usuarios, lo que ha provocado que cada vez sea más difícil encontrar canciones que sean del agrado del usuario. La recomendación musical se ha vuelto fundamental, ya que ayuda a los usuarios a tratar con este problema.

En la actualidad existen algunas aplicaciones que recomiendan música basándose en la etiqueta de la canción, es decir, según el género, artista, año, etc., y otros, como menciona Germain and Chakareski (2013), en los que se recomiendan canciones según los gustos de usuarios similares. El problema del primer método de recomendación es que aún dentro de un mismo género existe demasiada diversidad, por lo que el hecho de que a una persona le guste una canción no necesariamente implica que le gusten otras canciones del mismo género. El problema del segundo método es que la recomendación se vuelve poco personalizada e ignora los gustos muy particulares del usuario, además de que es necesario que varios usuarios escuchen canciones para poder generar la recomendación.

Una forma de lidiar con las problemáticas de los recomendadores anteriormente expuestos es basarse en la señal de la canción, de esta forma se buscan similitudes entre

las canciones que son del agrado de una persona para ver si una nueva canción podría ser recomendada o no. Generalmente, en el análisis de la señal de la canción se utilizan métodos de *machine learning*, tales como redes neuronales, en los cuales se identifican características de las canciones de una misma categoría.

En esta tesis se presenta una alternativa a la recomendación musical usando únicamente elementos obtenidos de la señal de la canción. En esta alternativa se emplea una adaptación al método de clasificación *K Nearest Neighbors* para lograr una recomendación.

En el primer apartado del capítulo 2 se habla de los diferentes métodos de recomendación musical que existen actualmente y las desventajas que éstos presentan. También se hace referencia a las canciones como series de tiempo y se mencionan algunos de los elementos que caracterizan a las canciones. En el segundo apartado del capítulo 2 se da una introducción a lo que es el aprendizaje automático y se define el término de *clasificación*. Además, se describen algunos de los algoritmos comúnmente usados.

En el capítulo 3 se presenta a detalle la metodología empleada para: 1. Realizar el análisis matemático y estadístico de las canciones y determinar variables útiles para lograr una clasificación adecuada; 2. Construir el clasificador con las variables determinadas y 3. Evaluar los resultados obtenidos con el clasificador empleado y comparar la clasificación lograda con la clasificación obtenida usando una de las grandes plataformas de música.

Finalmente, en el capítulo 4 se muestran los resultados obtenidos del análisis empleado para la obtención de las variables características de las canciones, así como los resultados obtenidos al comparar el clasificador propuesto con el clasificador empleado por una de las grandes plataformas de música.

## 1.1 Hipótesis y objetivos

### 1.1.1 Hipótesis

Un clasificador de canciones basado únicamente en elementos de series de tiempo tiene un menor porcentaje de canciones mal clasificadas que un clasificador basado en

etiquetas.

### 1.1.2 *Objetivos*

Objetivo general

Encontrar una forma de clasificar las canciones que sea óptima, utilizando únicamente características temporales de cada una de las canciones.

Objetivos específicos

El siguiente trabajo posee los siguientes objetivos:

1. Identificar elementos en la serie de tiempo de la canción que permita su posterior agrupación.
2. Evaluar el método de clasificación supervisado usando los elementos del punto anterior como variables clasificadoras.
3. Comparar el método de clasificación obtenido en el punto anterior con un clasificador que usa las etiquetas de las canciones: género, país, idioma, duración, año.
4. Determinar las ventajas y desventajas del uso de las series de tiempo de las canciones.

## CAPÍTULO 2

### Marco teórico

#### 2.1 Recomendación musical

El principal objetivo de la recomendación de música es ayudar a los usuarios a descubrir nuevas canciones basándose en sus gustos y preferencias.

En la actualidad existen algunas aplicaciones que recomiendan música basándose en la etiqueta de la canción, es decir, según el género, artista, año, etc. El problema de este método de recomendación es que aún dentro de un mismo género existe demasiada diversidad, por lo que el hecho de que a una persona le guste una canción no necesariamente implica que le gusten otras canciones del mismo género, además de que es importante tener en mente que este sistema se basa en géneros musicales que el usuario ha escuchado con anterioridad, lo cual puede provocar que las recomendaciones sean muy similares a lo que el usuario escucha y no le permitan descubrir otros géneros que también podrían ser de su agrado (Zhang et al., 2012).

Algunas de las aplicaciones más representativas en la recomendación musical son Apple Genius, Pandora y Spotify. Apple Genius es una aplicación que toma una sola canción base y con ella crea una lista de reproducción con canciones similares de la biblioteca del usuario y otra bibliotecas de iTunes (Apple, 2019). Pandora utiliza un equipo de músicos entrenados que escuchan las canciones y asignan valores a aproximadamente 400 atributos de las canciones, usados para caracterizarlas. A partir de esta caracterización se basan en una sola canción para generar una radio con canciones similares (Pandora, 2019). Por otra parte, Spotify genera las recomendaciones vía EchoNest, la cual es una plataforma que caracteriza las canciones mediante atributos tales como el ritmo, tempo, etc. y metadatos tales como críticas y calificaciones (Spotify, 2019).

Algunos autores, como Germain and Chakareski (2013), han buscado mejorar es-

tos métodos de recomendación a través de la recomendación asistida por Facebook, donde se busca obtener datos de los artistas que el usuario sigue en Facebook y combinarlos con las canciones escuchadas por el usuario en alguna plataforma.

Otra forma en las que se puede recomendar música es basándose en las recomendaciones que hacen otros usuarios con gustos similares. Sin embargo, este método requiere de mucho más tiempo para empezar a hacer recomendaciones adecuadas y puede ser que las recomendaciones no sean las mejores, ya que no se consideran las preferencias y gustos particulares de cada usuario. Otro de los grandes problemas con esta técnica es que para que una canción sea recomendada debe haber sido escuchada por una cantidad significativa de usuarios, por lo que las recomendaciones dan preferencia a las canciones más populares (Kim et al., 2007).

Una de las problemáticas con las que se encuentra la recomendación musical es que en ocasiones se recomiendan canciones que sí son del agrado del usuario, sin embargo, no es lo que desea escuchar en ese momento, ya que el contexto en el que se encuentra el usuario influye la percepción de la música que se tiene en ese momento (Kaitila, 2017).

### **2.1.1 *Las canciones como series de tiempo***

El sonido comienza con la vibración de un objeto, que a su vez crea patrones de compresión y rarefacción de las moléculas de aire. El sonido generado es una señal análoga continua que, para que pueda ser capturado por una computadora, necesita registrar valores discretos a intervalos de tiempo discretos (Patel and Gopi, 2015). Este proceso de registrar valores de tiempo discreto se denomina muestreo debido a que se particiona la onda continua y se toman únicamente ciertos valores de ésta. Debemos preguntarnos qué tan rápido se debe muestrear una señal continua para preservar su contenido de información. Esta rapidez a la que se muestrea la señal es conocida como frecuencia de muestreo, podemos muestrear una señal continua a cualquier frecuencia de muestreo y obtendremos una serie de valores discretos, por lo que es importante que estos valores sean representativos de la señal original (figura 2.1). Una frecuencia de muestreo de audio común para música es 44,100 muestras por segundo, esto quiere decir que las canciones

con esta frecuencia de muestreo cuentan con 44,100 datos por segundo. La unidad para la frecuencia de muestreo es Hertz (Hz), por lo que 44,100 muestras por segundo son 44,100 hertz o 44.1 kilohertz (kHz).

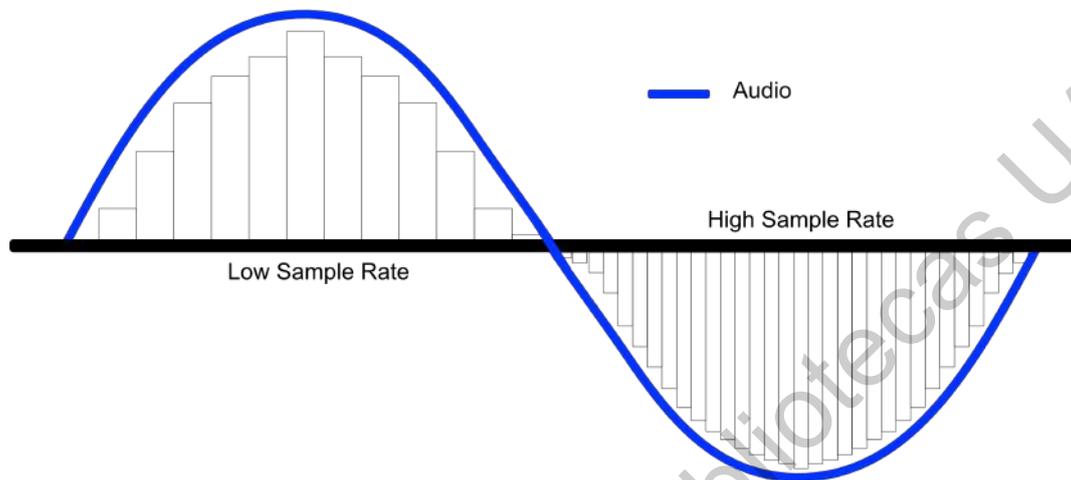


Figura 2.1: Frecuencia de muestreo. Fuente: *Mastering the mix*

Musicalmente, una canción puede ser caracterizada por una combinación de algunos de sus elementos como lo son el ritmo, tempo, tono, timbre o su frecuencia. En ese sentido se entiende por *ritmo* a la proporción guardada entre los acentos, pausas y repeticiones de diversa duración en una composición musical. El *tempo* es el grado de celeridad en la ejecución de una composición musical; el *tono* es una cualidad de los sonidos, dependiente de su frecuencia, que permite ordenarlos de grave a agudos; mientras que el *timbre* es el conjunto de marcas diferenciadoras que individualizan el sonido de cada nota musical, y la *frecuencia* es el número de oscilaciones completas que se producen en una unidad de tiempo (Colomer Blasco, 2016).

Por otro lado, si se considera una canción como una serie de tiempo, entonces una primera forma de intentar visualizar tales características es por medio del análisis descriptivo de la serie de tiempo (Beran, 2003). Dentro del análisis descriptivo de la música como serie de tiempo se tiene el gráfico de tiempo-frecuencia (figura 2.2).

En particular, la frecuencia resulta relevante de analizar, ya que esta cualidad del sonido nos ayuda a identificar cuántas veces se repite el patrón de vibración de una

onda sonora en un segundo. Mientras menor sea la frecuencia, la longitud de la onda será mayor y los sonidos serán más graves, en cambio, si la frecuencia es mayor, la longitud de la onda será menor y los sonidos serán más agudos.

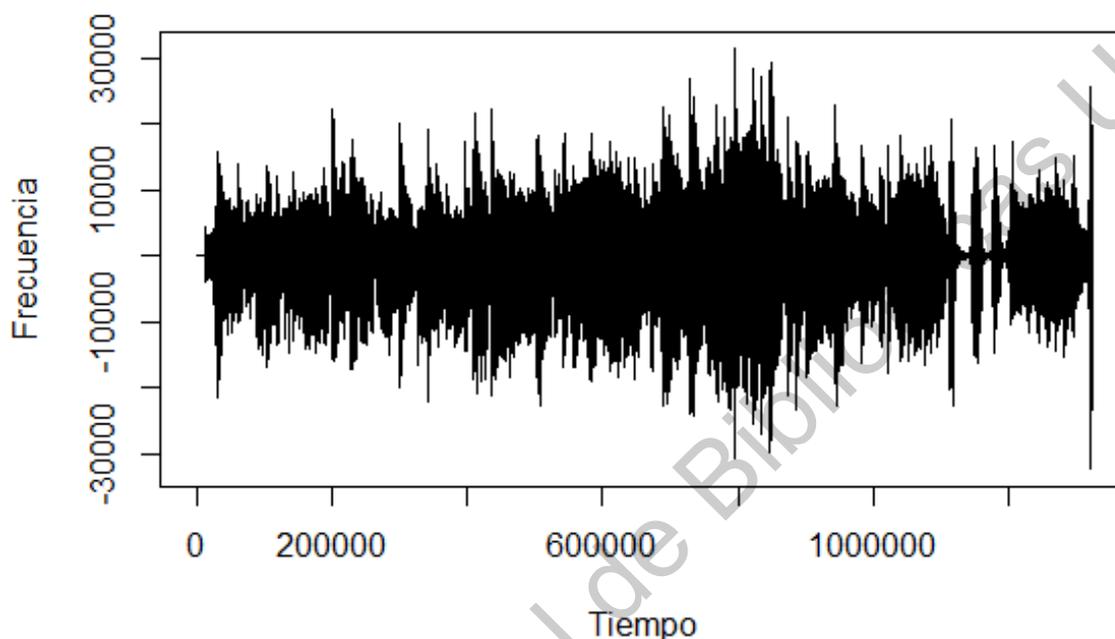


Figura 2.2: Frecuencia de los primeros 30 segundos de una canción

Otra característica de la música que es importante analizar es la *amplitud*, ya que refleja la intensidad o volumen del sonido, de tal forma que a medida que la amplitud aumenta, el volumen aumenta. En general, la amplitud se define como el valor máximo de una cantidad variable respecto de su valor medio o de base.

### Análisis espectral

El análisis espectral tiene por objetivo descomponer una serie de tiempo estacionaria en una suma de componentes senoidales periódicos con diversas frecuencias y amplitudes, así como determinar en que proporción está presente cada frecuencia en la serie.

Al análisis espectral también se le llama análisis en el dominio de la frecuencia ya

que las frecuencias más significativas sirven para explicar ciclos, tales como ciclos económicos, estacionalidad, o características estadísticas de un proceso (Montenegro et al., 2009).

### Transformada de Fourier

La Transformada de Fourier es una herramienta utilizada para obtener la información frecuencial de una determinada función. Esta transformada descompone una serie de tiempo en una suma finita de funciones de senos y cosenos y permite que una función sea expresada en el dominio del tiempo o de la frecuencia (Cabrejo and Sánchez, 2006). Esta serie es conocida como la serie de Fourier y se define como:

$$x(t) = \frac{a_0}{2} + \sum_{j=1}^{\infty} \left( a_j \cos \frac{2\pi jt}{n} + b_j \sin \frac{2\pi jt}{n} \right) \quad (2.1)$$

Donde  $x(t)$  es una serie de tiempo,  $a_j$  y  $b_j$  son los coeficientes de Fourier y  $n$  es la longitud de la serie.

De igual forma, la serie de tiempo se puede representar en el dominio de la frecuencia, donde el fenómeno de interés puede ser descrito por su amplitud en función de la frecuencia  $X(w)$ . De esta forma  $x(t)$  y  $X(w)$  son dos representaciones de la misma función y la conexión entre estas se encuentra dada mediante la transformada de Fourier (Cabrejo and Sánchez, 2006).

La Transformada de Fourier de  $x(t)$  se define como

$$X(w) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi wt} dt \quad (2.2)$$

Donde  $w$  es la frecuencia,  $t$  el tiempo e  $i = \sqrt{-1}$

### Periodograma

Uno de los factores que se busca al hacer un análisis espectral a una serie de tiempo es identificar las frecuencias que tienen mayor importancia. El periodograma es una herramienta que nos ayuda a hacer esto, ya que grafica una medida de la importancia de posibles valores de la frecuencia que puedan explicar el patrón de oscilación en la

serie analizada (of Science, 2018). Si los datos contienen algunos componentes periódicos fuertes, los valores del periodograma correspondientes a esas frecuencias (o cerca de esas frecuencias) serán grandes. Por otro lado, los valores del periodograma serán pequeños para los componentes periódicos que no están presentes en los datos (Shumway and Stoffer, 2015).

El periodograma se calculó de la siguiente manera:

$$Per = \frac{(fft(x - \bar{x}))^2}{n} \quad (2.3)$$

donde

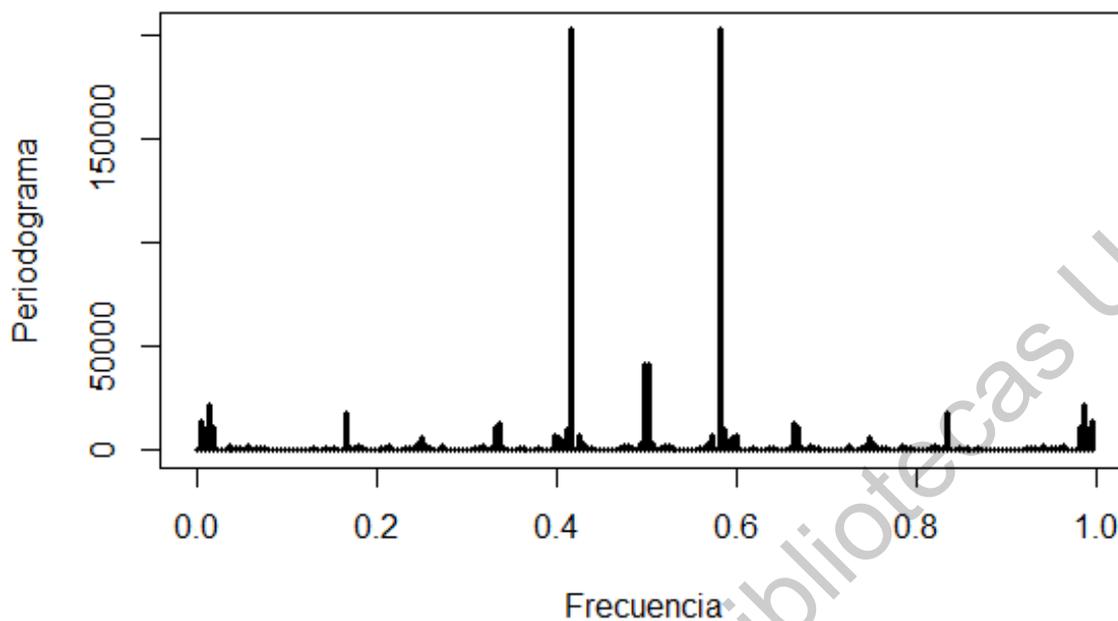
$fft$  es la transformada rápida de Fourier

$x$  son los valores de la frecuencia o amplitud de la canción

$\bar{x}$  es la media del vector  $x$

$n$  es el número de elementos del vector  $x$ .

La figura 2.3 muestra el periodograma para una canción cuya duración es de 223 segundos y en la que el periodo es de 1 segundo, es decir, se cuenta con los valores de la frecuencia por segundo, por lo que en total se tienen 223 registros. En ella se pueden observar algunos valores que resaltan más que otros, las dos barras más grandes corresponden a los registros 94 y 131, cuyas frecuencias son  $94/223 = 0.4215247$  y  $131/223 = 0.5874439$  ciclos por segundo respectivamente. Note que estos dos ciclos son los componentes periódicos más destacados de la canción.



*Figura 2.3: Periodograma de una canción*

## 2.2 Clasificación

El aprendizaje automático es un subcampo de la inteligencia artificial, usado para explorar la estructura de los datos y ajustar modelos que puedan ser entendidos y utilizados por los usuarios. Gangadhar Shobha (2018) afirma que el aprendizaje automático responde a la pregunta de cómo construir un programa de computadora utilizando datos históricos para resolver algún problema y automáticamente mejorar la eficiencia del programa. Lo que facilita a las computadoras construir modelos partiendo de una muestra de datos disponible, y automatizar el proceso de la toma de decisiones, basándose en las entradas de los datos y experiencia. Estas técnicas identifican patrones en los datos y proporcionan herramientas para la minería de datos.

El aprendizaje automático se divide en aprendizaje supervisado, no supervisado, semi-supervisado y reforzado. Éstos a su vez se dividen en diferentes métodos. Figura 2.4.

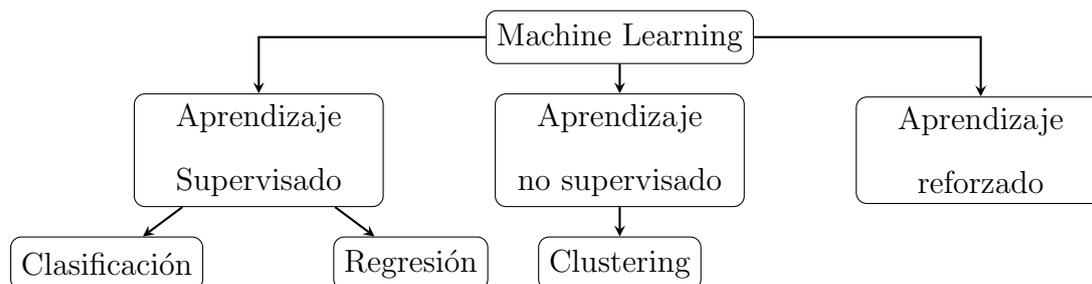


Figura 2.4: Tipos de aprendizaje

### 2.2.1 Aprendizaje supervisado

En el aprendizaje supervisado se tienen dos conjuntos de datos etiquetados. Estos datos son generalmente llamados datos de entrenamiento y datos de prueba. Los datos de entrenamiento, o predictores, contienen valores en la variable que se busca predecir, de esta forma se puede construir un clasificador con la información proporcionada por cada usuario. Los datos de prueba servirán, como su nombre lo dice, para probar qué tan bien funciona el clasificador. Sea  $\mathcal{X}$  el dominio de las características de los datos y sea  $\mathcal{Y}$  el dominio de las etiquetas. El aprendizaje supervisado entrena una función  $f : \mathcal{X} \mapsto \mathcal{Y}$  en alguna familia de funciones  $\mathcal{F}$ , con el objetivo que  $f(x)$  prediga la etiqueta real ( $y$ ) en datos futuros ( $x$ ). Esta función es conocida como clasificador y relaciona las características con las etiquetas o respuestas.

Este tipo de aprendizaje usa algoritmos de clasificación y técnicas de regresión para crear modelos predictivos. Mientras que la clasificación se utiliza para predecir variables categóricas o cualitativas, la regresión se utiliza para predecir respuestas cuantitativas. En la clasificación, para cada observación del conjunto de entrenamiento existe una respuesta asociada y se desea ajustar un modelo que relacione la respuesta a los predictores con el fin de predecir acertadamente la respuesta de futuras observaciones. En cambio, la regresión lineal modela la relación entre dos variables al ajustar una ecuación lineal a los datos observados.

Los algoritmos de aprendizaje supervisado incluyen regresión lineal, regresión logística, redes neuronales, árboles de decisión, y K-nearest neighbors, entre otros.

## K Nearest Neighbors

El algoritmo de *KNN* se basa en la idea de que las cosas similares están cerca entre ellas. *K- vecinos más próximos*, *KNN* por sus siglas en inglés, es un método de clasificación mediante el cual, para un patrón desconocido  $x_j$ , se asigna una etiqueta de clase según los *K- más próximos* patrones, en el espacio de datos. Esto se realiza al calcular la probabilidad condicional para la clase  $j$  como una fracción de los puntos en  $N_0$  cuya etiqueta de clase sea igual a  $j$ , siendo  $N_0$  los  $k$  puntos más cercanos a  $x_j$ . Finalmente, se clasifica al patrón desconocido  $k_j$  con la etiqueta de clase que tiene la probabilidad más alta (James et al., 2013).

$$Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j) \quad (2.4)$$

Para lograr lo anteriormente expuesto, es necesario definir una medida de distancia en el espacio de patrones. La manera más común de hacer esto es usando la distancia Euclidiana como se muestra en la ecuación 2.5 (Schott, 2019).

$$d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (2.5)$$

La figura 2.5 muestra un ejemplo de cómo funciona el algoritmo de *KNN*. Se pueden ver dos grupos de 7 observaciones cada uno, el objetivo es predecir la cruz negra. Supongamos  $K = 3$ . *KNN* primero identificará las tres observaciones más cercanas a la cruz, es decir, las tres observaciones cuya distancia con respecto a la cruz sea la más pequeña. La vecindad de puntos seleccionados se encuentran dentro del círculo, y puede observarse que hay dos cuadrados y un triángulo, por lo que la probabilidad de que la cruz pertenezca a los cuadrados es  $2/3$  y la probabilidad de que la cruz pertenezca a los triángulos es  $1/3$ . Por lo tanto, *KNN* predice que la cruz negra pertenece al grupo de los cuadrados.

La elección de  $K$  determina la localidad del clasificador *KNN* y juega un papel



## Clustering

*Clustering* se refiere a un conjunto muy amplio de técnicas para encontrar subgrupos o *clusters* en un conjunto de datos. Cuando agrupamos las observaciones del conjunto de datos, buscamos particionarlos en distintos grupos, de tal forma que las observaciones dentro de cada grupo sean similares entre ellas, mientras que las observaciones en diferentes grupos sean distintas de ellas. (James et al., 2013). Para poder realizar el agrupamiento primero es necesario definir qué significa que las observaciones sean similares o diferentes.

Existen muchos métodos de *clustering* o agrupamiento, uno de los más utilizados es el de K-medias. En este método se busca particionar las observaciones en un número previamente especificado de *clusters*,  $K$ , de tal forma que estos clusters no se traslapen y que cada observación pertenezca a no más de un cluster.

El algoritmo de K-means clustering funciona de la siguiente manera:

1. Se asigna aleatoriamente un número de 1 a  $K$  a cada una de las observaciones. Esto servirá como clusters iniciales.
2. Para cada uno de los  $K$  clusters se calcula el centroide, el cual es la media de los valores dentro de cada cluster.
3. Se asigna cada una de las observaciones al cluster cuyo centroide es el más cercano.

Los puntos 2 y 3 se repiten hasta que el algoritmo converge. Según James et al. (2013), se puede decir que un buen cluster es aquel donde la suma de las Distancias Euclidianas cuadradas de los elementos dentro del cluster, dividido entre el número total de observaciones dentro del cluster, es la más pequeña posible. Es decir, se busca

$$\min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\} \quad (2.6)$$

donde

$C_1, \dots, C_K$  son conjuntos que contienen los índices de las observaciones de cada cluster.

## CAPÍTULO 3

### Metodología

#### 3.1 Selección aleatoria de una muestra de canciones

Para este trabajo se optó por utilizar la plataforma Spotify como punto de comparación al clasificador debido a que es una de las plataformas más utilizadas para reproducir canciones. Por esta razón, se tomó una muestra de 30 personas a quienes se les preguntó una canción que fuera de su agrado y se generó *la radio* de la canción elegida. La *radio* de Spotify crea una colección de 50 canciones basada en el artista, álbum, playlist o, en este caso, canción que el usuario elija (Spotify, 2019). A partir de esta radio se tomó una muestra aleatoria de 15 canciones, las cuales fueron escuchadas por la persona que proporcionó la *canción base*.

A cada una de las personas se les pidió que clasificaran las canciones en tres categorías: “*me gusta y guardaría la canción*”, “*me gusta pero no guardaría la canción*” y “*no me gusta*”. De esta forma sabríamos si el clasificador de Spotify hizo un buen trabajo y nos ayudaría a encontrar patrones para la clasificación de las canciones.

Una vez que se recolectaron los datos obtenidos para cada persona, se conformó una base de datos inicial de 480 observaciones, con 30 características cada una (en adelante *matriz M*), donde las columnas son los sujetos, los renglones son las canciones y el elemento  $i, j$  de la matriz  $M$  está formado por una etiqueta según si la canción le gustó o no le gustó al sujeto.

$$M[i, j] = \begin{cases} 1 & \text{Le gusta y la guardaría} \\ p & \text{Le gusta y no la guardaría} \\ 0 & \text{No le gusta} \end{cases} \quad (3.1)$$

A la clasificación “Le gusta y la guardaría” se le otorgó el valor de 1, ya que quiere

decir que la recomendación hecha por Spotify fue buena. Por otro lado, a la clasificación "No le gusta" se le otorgó el valor de 0. Para la clasificación "Le gusta y no la guardaría", se decidió hacer un análisis de sensibilidad para detectar el valor  $p$  que mejor representaba a esta etiqueta.

Cabe destacar que, debido a que sólo se pidió a cada persona que escucharan 15 canciones adicionales a la *canción base* que proporcionaron, para cada sujeto en la matriz solo se van a contar con 16 renglones con el indicador de clasificación, los demás renglones se mostrarán como NA debido a que no corresponden a las canciones que escucharon. En el apéndice B se puede encontrar las etiquetas proporcionadas por cada uno de los individuos.

### 3.2 Análisis estadístico

Con las canciones previamente clasificadas, se buscaron atributos de las canciones que las caracterizaran de tal forma que resultara en una separación evidente de las canciones que les gustaron o no a los individuos, y así crear un vector con los atributos de cada canción.

Inicialmente se realizaron pruebas con diversos atributos tales como media, varianza, correlación, velocidad y aceleración. Estas medidas fueron tomadas tanto de la frecuencia como de la amplitud de las canciones.

Es importante destacar que estas pruebas se realizaron únicamente para los primeros treinta segundos de cada canción. Se registró el tiempo que tardaban los individuos estudiados en decidir si una canción era de su agrado, lo suficiente para seguir escuchándola, o el momento en que decidían cambiar de canción y, a pesar de que el tiempo resultante fue superior, se decidió usar los primeros 30 segundos dado que hay estudios que indican que este tiempo basta para decidir si a una persona le gusta o no una canción, entre ellos están North and Hargreaves (1995) y the National Association for Music Education MENC (1984).

Por otra parte, la frecuencia de muestreo de las canciones analizadas es de 44,100 muestras por segundo, por lo que, aún analizando los primeros 30 segundos, se contaba

con muchos datos que consumían mucha memoria de la computadora. Por esta razón se decidió tomar el promedio de las 44,100 muestras por segundo para los valores de la frecuencia o la amplitud, así, al final se contaba con 30 valores que representaban los primeros 30 segundos de la frecuencia o la amplitud de la canción. A este procedimiento se le llamará suavizamiento.

### 3.2.1 *Velocidad y aceleración*

El tempo se refiere a la velocidad a la que se ejecuta una pieza musical. Sin embargo, dado que no se cuenta con esta información de las canciones, se optó por calcular la velocidad de cada una de ellas mediante la primer diferencia de la serie de tiempo de acuerdo a la ecuación 3.2, así, se obtendría una aproximación del tempo.

$$y_t = x_t - x_{t-1} \quad (3.2)$$

Donde  $x_t$  son los valores de la frecuencia o la amplitud de la canción y  $y_t$  son los valores calculados pa la velocidad.

Por otra parte, la aceleración permite analizar qué tan rápidos son los cambios de ritmo en la canción, es decir, las variaciones de velocidad. Esta variable se obtuvo mediante la diferenciación de la serie de tiempo de la velocidad, de acuerdo a la ecuación 3.3, donde  $z_t$  representa los valores obtenidos para la aceleración.

$$z_t = y_t - y_{t-1} \quad (3.3)$$

Una vez obtenidas estas dos variables, se realizó un análisis de correlación de la velocidad y la aceleración de la frecuencia de las canciones etiquetadas por las personas. El objetivo de realizar este análisis era ver si existía relación alguna entre las canciones que contaban con la misma etiqueta, de tal forma que se mostrara una correlación alta entre las canciones con etiqueta igual y una correlación baja o nula entre las canciones con etiqueta diferente.

La figura 3.1 muestra las gráficas de la velocidad de la frecuencia y la amplitud

de las canciones que obtuvieron el valor más bajo y el más alto de esta variable para los primeros 30 segundos. Puede verse que la canción con mayor velocidad promedio (localizada en la parte inferior de las figuras 3.1a y 3.1b muestra más variaciones a lo largo de la mayor parte de los primeros 30 segundos. Las tablas 3.1a y 3.1b muestran el género al que pertenecen estas canciones, y puede observarse que tanto para la frecuencia como para la amplitud, las canciones con mayor velocidad pertenecen al género *Hard Rock*.

<b>Canción</b>	<b>Género</b>	<b>Canción</b>	<b>Género</b>
Don't Kill My High	Electrónica	Me Hubieras Dicho	Pop
Lights Out	Hard Rock	Lights Out	Hard Rock

(a) *Frecuencia* (b) *Amplitud*

Tabla 3.1: Comparación de la velocidad

Dirección General de Bibliotecas UAQ

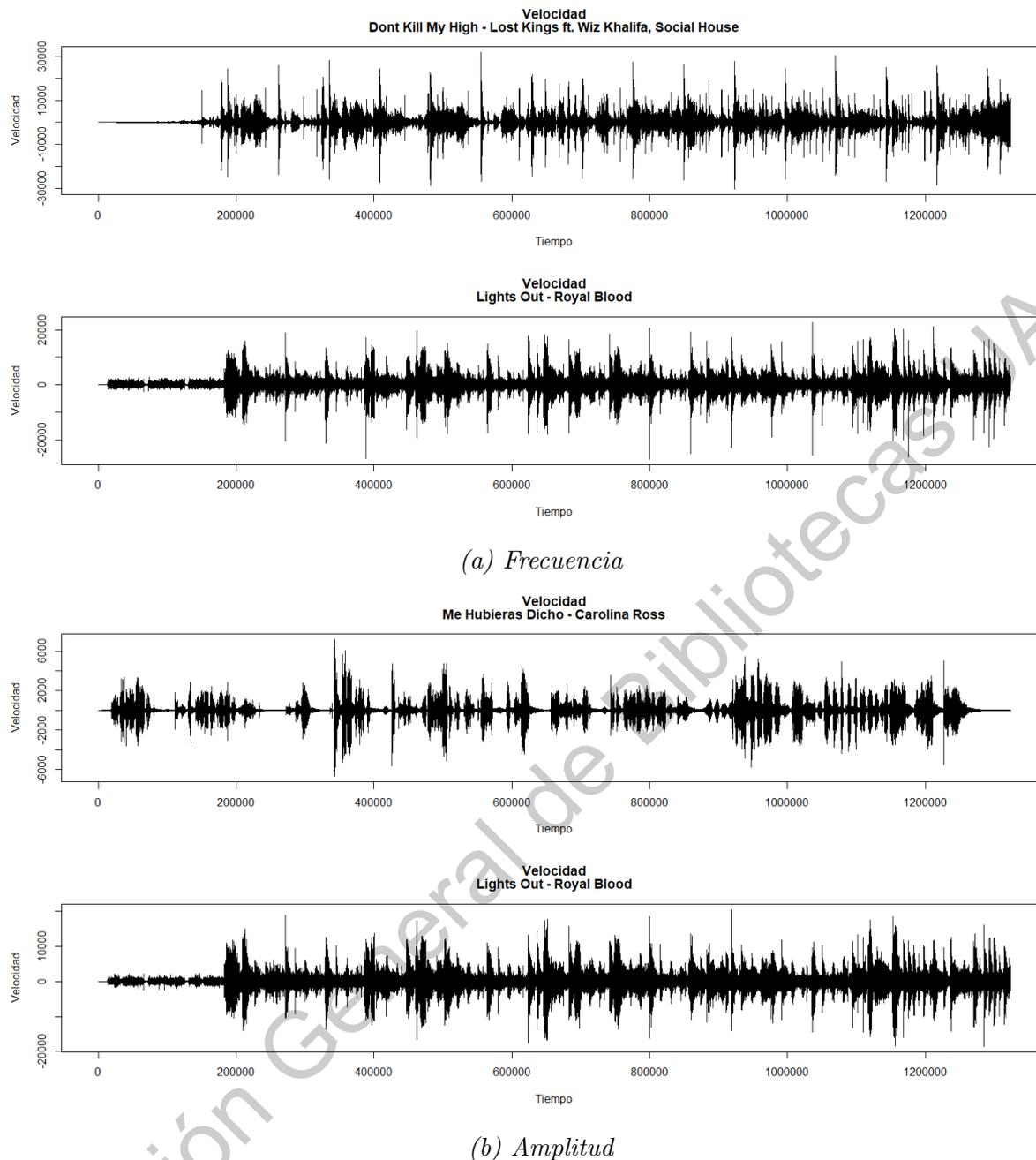


Figura 3.1: Comparación de la velocidad

Respecto a la aceleración, la figura 3.2 muestra las gráficas para la frecuencia y la amplitud de las canciones que obtuvieron el valor más bajo y el más alto de esta variable en los primeros 30 segundos. Puede verse que la canción con mayor amplitud muestra más variaciones a lo largo de la mayor parte de los primeros 30 segundos. Las tablas 3.2a y 3.2b muestran el género al que pertenecen estas canciones, y puede observarse que las canciones con mayor velocidad pertenecen al género *Cumbia* o *Indie alternativo*.

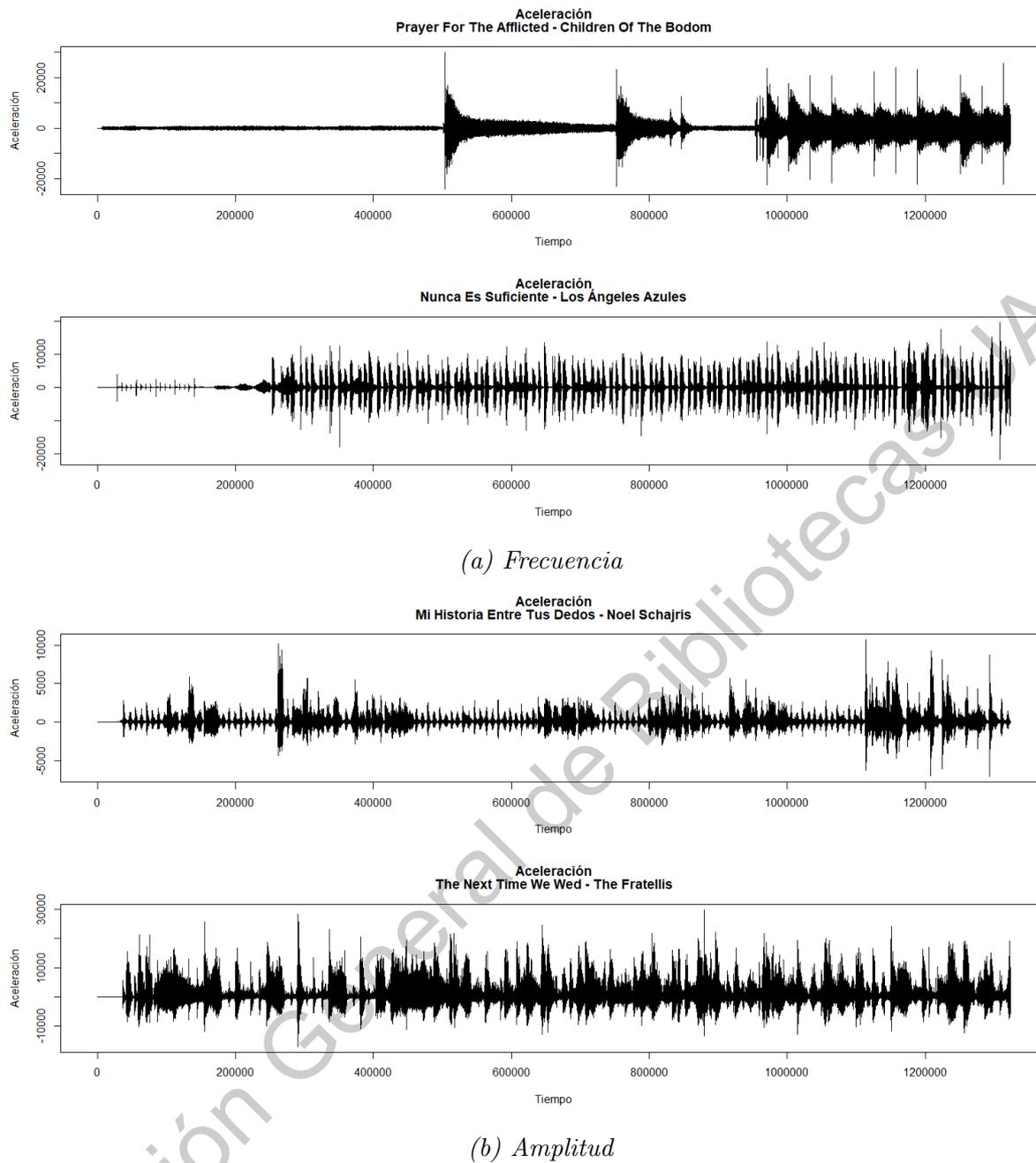


Figura 3.2: Comparación de la aceleración

Canción	Género	Canción	Género
Prayer For The Afflicted	Black/Death Metal	Mi Historia Entre Tus Dedos	Pop
Nunca Es Suficiente	Cumbia	The Next Time We Wed	Alternative/Indie, Rock

(a) Frecuencia

(b) Amplitud

Tabla 3.2: Comparación de la aceleración

### 3.2.2 Ciclos y periodos

Los ciclos en música se refieren a una vibración completa (Apel, 2003), siendo su unidad base los Hertz (un ciclo por segundo).

Para calcular los ciclos se utilizó el periodograma, ya que éste nos muestra las frecuencias más importantes. Se tomaron las frecuencias resultantes del periodograma que excedían el tercer cuartil debido a que estos son los componentes periódicos más destacados de la canción. Estos valores se dividieron entre 30 porque la longitud de cada canción es de 30 segundos. De esta manera, se obtuvo un vector que nos indica la frecuencia de las frecuencias más importantes.

$$ciclos = \frac{Per \geq Per_{Q_3}}{30} \quad (3.4)$$

Donde  $Per$  es el periodograma calculado de acuerdo a la ecuación 2.3 y  $Per_{Q_3}$  corresponde al tercer cuartil de los periodos.

Para calcular los periodos simplemente se obtuvo la inversa de los ciclos, de esta forma se obtendría el tiempo que tardan en repetirse las frecuencias obtenidas en el periodograma.

$$periodos = \frac{1}{ciclos} \quad (3.5)$$

Para ver si alguna de estas variables era relevante, y dado que cada canción contaba con un número diferente de ciclos y periodos con valores que variaban, se decidió calcular el promedio de los valores de estas variables para cada canción y ordenarlos de menor a mayor con la finalidad de ver si se lograba algún agrupamiento entre las canciones con la misma etiqueta.

En la tabla 3.3 se muestran las canciones que obtuvieron el mínimo (en la parte superior) y el máximo (en la parte inferior) del promedio de los ciclos tanto para la frecuencia como para la amplitud. En la tabla 3.3b puede verse que ambas canciones pertenecen al género *Metal*, lo cual es importante resaltar ya que, aunque ambas canciones pertenecen al mismo género, son muy diferentes y un clasificador que se base en el género de la canción no haría distinción entre éstas.

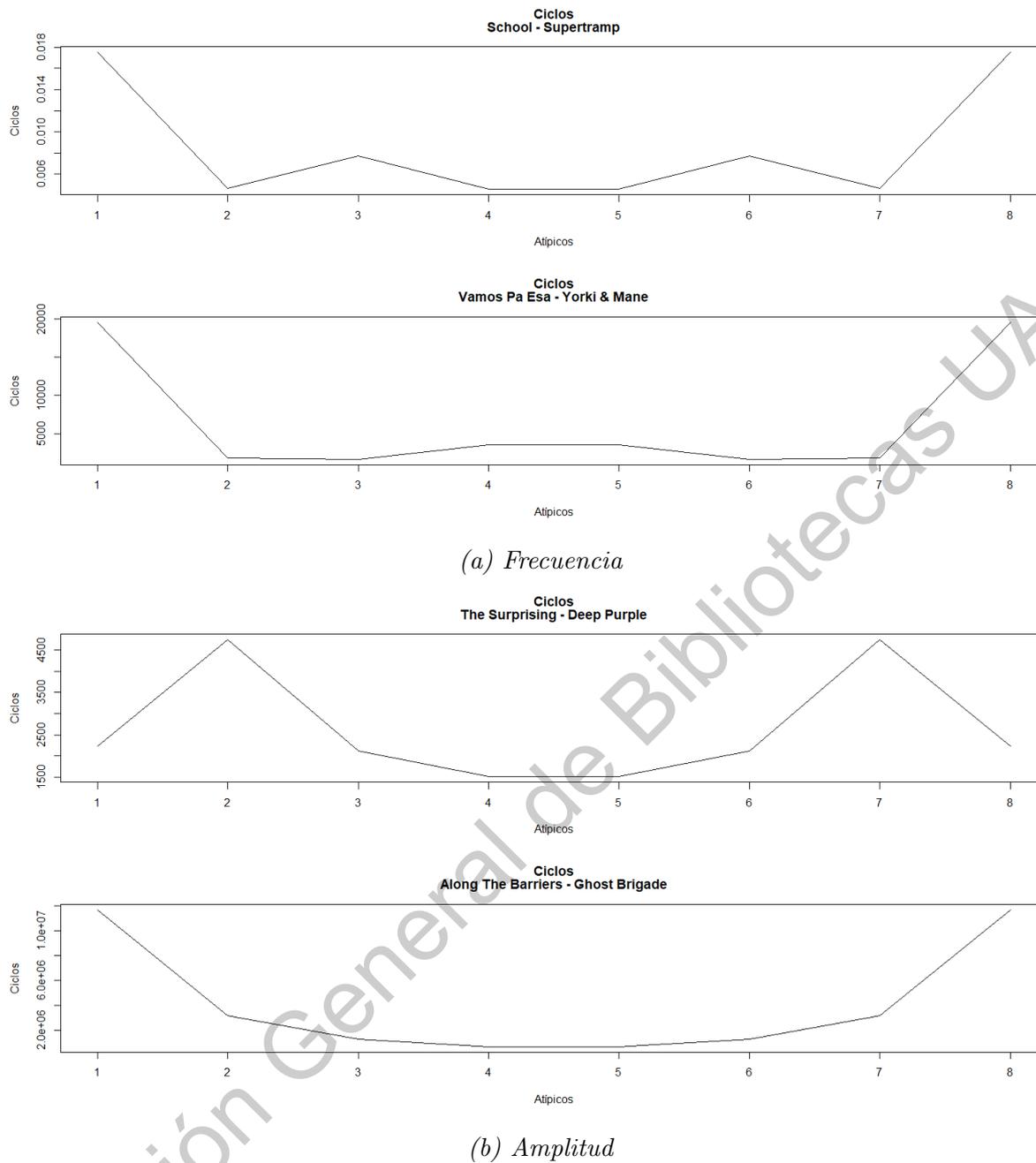


Figura 3.3: Comparación de los ciclos

Canción	Género	Canción	Género
School	Progressive/Art Rock	The Surprising	Metal, Rock
Vamos Pa Esa	Reggaeton	Along The Barriers	Progressive metal, Metal

(a) Frecuencia

(b) Amplitud

Tabla 3.3: Comparación de los ciclos



### 3.2.3 Varianza

La varianza, como medida de dispersión, es útil para identificar la variabilidad en la frecuencia y amplitud de cada canción. Con esta medida se buscó clasificar a las canciones de acuerdo a su variabilidad, para ello, se calculó la varianza para la frecuencia y la amplitud de las canciones y los resultados se ordenaron de menor a mayor para ver cuáles canciones contaban con una mayor varianza y cuáles con una menor varianza. Con esto, se observó que se generaban agrupaciones de las canciones según su etiqueta, por lo que se decidió que la varianza era un buen atributo para la caracterización de las canciones.

La tabla 3.5a muestra el valor de la varianza más bajo (en la parte superior) y el más alto (en la parte inferior) para la frecuencia de las canciones. De igual forma, la tabla 3.5b muestra los valores de la varianza más altos y bajos para la amplitud de las canciones. En ambas tablas puede observarse que la canción con mayor varianza pertenece al género *Dance / Electronic*.

Canción	Varianza	Género
Move on - Mike Ponser	239353.8	Alternative/Indie
Who's the boss - Loud Luxury	144657091.4	Dance/Electronic
<i>(a) Frecuencia</i>		
Canción	Varianza	Género
The Surprising - Deep Purple	149846.7	Metal, Rock
Who's the boss - Loud Luxury	144657091.4	Dance/Electronic
<i>(b) Amplitud</i>		

Tabla 3.5: Comparación de las varianzas.

Además, Se realizaron gráficas con la varianza, la media de los ciclos y la media de los periodos para ver si estas tres variables podían ser consideradas como atributos característicos de las canciones que permitieran separarlas según su etiqueta.

Finalmente se determinaron diez características que resultaban significativas para

la caracterización de las canciones, cinco correspondientes a la frecuencia y estas mismas fueron determinadas para la amplitud. Los elementos que conforman el vector característico son: Varianza, media de velocidad, media de la aceleración, medida de los ciclos y de los periodos.

Adicionalmente, se calculó la distancia a la canción original con los distintos elementos del vector característico de acuerdo a la ecuación 2.5, donde  $q$  corresponde a la variables del vector característico de la canción  $i$  y  $p$  corresponde a la variables del vector característico de la canción  $j$ . En la figura 3.5 se muestran las gráficas de distancia para 5 individuos seleccionados aleatoriamente. Cabe destacar que las gráficas presentadas son de individuos seleccionados aleatoriamente para no sesgar los resultados mostrados, sin embargo, se presentaron casos donde la separación de las canciones que le gustaron y las que no le gustaron era muy evidente. Con estos usuarios seleccionados aleatoriamente se nota una tendencia de las canciones que no le gustaron a estar más alejadas ya que en cuatro de los cinco individuos mostrados, algunas de las canciones etiquetadas con 0 muestran una distancia mayor a las canciones etiquetadas con 1.

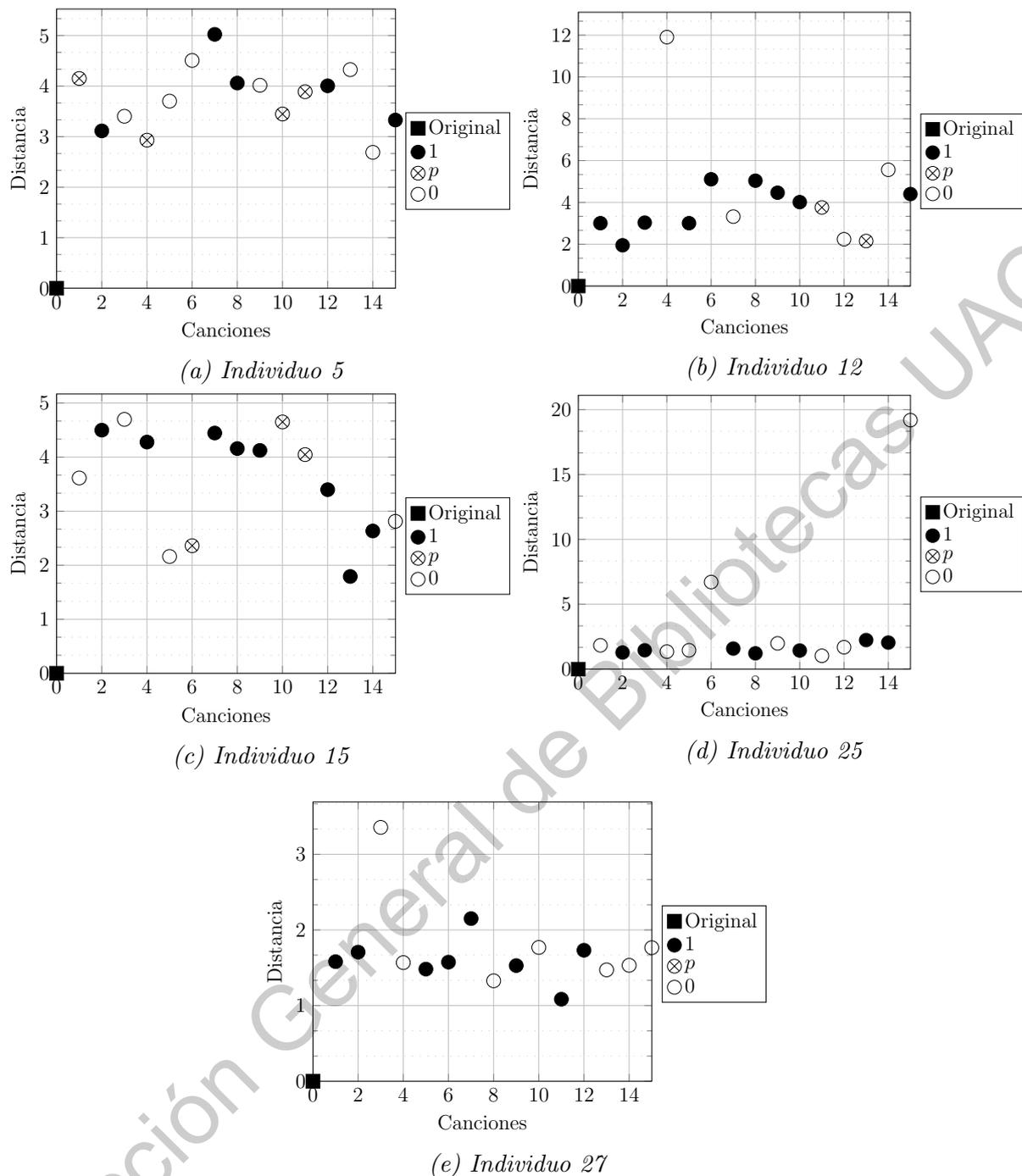


Figura 3.5: Distancia a la canción original

### 3.3 Desarrollo del algoritmo

El algoritmo se desarrolló en el software estadístico RStudio, el cual utiliza lenguaje de programación R.

Para estimar la función que permita clasificar las canciones, se empleó un modi-

ficación del algoritmo KNN que permitiera ponderar las características. Este algoritmo puede encontrarse en el apéndice A y toma como entrada la matriz mencionada en el apéndice B.

Debido a la diferencia de escala de las variables de los vectores característicos, se decidió estandarizarlos de acuerdo a la ecuación 3.6 de tal forma que todas las variables tuvieran el mismo peso. En esta ecuación  $y_i$  representa el nuevo valor estandarizado,  $x_i$  es el valor original obtenido para la variable a estandarizar, y  $\mu$  y  $\sigma$  representan la media y la desviación estándar de esta misma variable.

$$y_i = \frac{x_i - \mu}{\sigma} \quad (3.6)$$

Estas variables estandarizadas se utilizaron para crear una matriz de distancias Euclidianas, la cual es una matriz simétrica de  $480 \times 480$  con ceros en la diagonal y cada elemento de la matriz fuera de la diagonal se encuentra definido por la ecuación 2.5, donde  $q$  corresponde a la variables del vector característico de la canción  $i$  y  $p$  corresponde a la variables del vector característico de la canción  $j$ .

Para realizar la clasificación se creó una función que se tiene que correr para cada uno de los 30 individuos, tiene como entrada cuatro argumentos:

- Valor de  $k$ , es decir, cuántas canciones serán tomados en cuenta para realizar la clasificación. Para determinar este valor se realizó un análisis de sensibilidad variando los valores de  $k$  desde 10 hasta 20 para así encontrar el valor que proporcionara la mejor recomendación.
- Matriz de características, la cual consta de 11 columnas y 480 filas, donde las filas corresponden a las canciones analizadas, la primer columna corresponde al identificador único de cada elemento y las 10 columnas restantes a los elementos que caracterizan a cada canción.
- Vector de etiquetas, el cual tiene longitud 480 y contiene las canciones previamente clasificadas por el individuo. Puede tomar los valores de 0,0.6 ó 1, las entradas

correspondiente a las canciones sin clasificación se encuentran vacías.

- Número mínimo de canciones previamente etiquetadas necesarias para realizar la clasificación. Esto quiere decir que al clasificar un nuevo elemento, dentro de las  $k$  canciones más próximas a esta, debe haber un mínimo de canciones etiquetadas para poder asignarle un valor al elemento no clasificado, de lo contrario se queda sin predicción, a este parámetro se le asignó el valor 3.

Los pasos que se siguen dentro de la función son los siguientes:

1. Se calcula la matriz de distancias omitiendo la primera columna, pues ésta, como se había mencionado antes, corresponde a los indicadores de las canciones. La matriz resultante es una matriz cuadrada de  $480 \times 480$ , simétrica y con ceros en la diagonal.
2. Se determina cuáles son las canciones que no se encuentran clasificadas, y también se identifican sus posiciones en el vector de etiquetas.
3. Se concatena la matriz de distancias con el vector de etiquetas.
4. Se crea un vector llamado generaciones, que posteriormente será usado para almacenar un valor que indique la confianza de cada canción clasificada. A las canciones clasificadas originalmente por cada individuo se le asigna el valor 1.
5. Se inicia un ciclo "for" que realizará la clasificación mediante KNN, únicamente se realizarán 6 iteraciones puesto que iteraciones posteriores no tendrían un resultado confiable. Como se había mencionado anteriormente el vector de generaciones nos indicará el número de iteración en que se calculó cierto valor, para esto se elevará 0.9 a la potencia correspondiente a la generación, por ejemplo la segunda generación corresponderá a 0.81 ( $0.9^2$ ). Esto nos dará un indicador de qué tan confiable es cada recomendación arrojada por el algoritmo, es por esto que la primera generación tiene asociado el valor 1, ya que es completamente confiable puesto que proviene de la fuente original. El valor 0.9 fue seleccionado puesto que es un valor cercano a 1 y hará que la confianza no disminuya tan rápidamente. Las canciones que no sean clasificadas no tendrán ningún valor en este campo.

6. Se evalúa si hay alguna canción sin clasificar, de no ser así se rompe el ciclo, pues esto significaría que ya todas las canciones están clasificadas y no es necesario correr más iteraciones.
7. Se generará un segundo ciclo for, para realizar la clasificación para cada una de las canciones que no está etiquetada.
8. Las canciones son ordenadas de menor a mayor respecto a su distancia a la canción sin etiqueta. Las primeras  $k$  canciones son seleccionadas y se evalúa si cumple con el número mínimo de canciones previamente etiquetadas. Si lo cumple, se le asigna en la probabilidad de la canción sin etiqueta la media de las probabilidades de las primeras  $k$  canciones, de lo contrario, se queda vacío.
9. El vector de las canciones sin etiqueta es actualizado y se realiza la siguiente iteración.

Después de correr el algoritmo KNN para cada una de las 30 personas, se tiene como resultado una matriz con las etiquetas de las canciones como fila y las personas en la columna, es decir, la entrada  $i, j$  de la matriz corresponde a la etiqueta asignada a la canción  $i$  para la persona  $j$ . Cabe destacar que no necesariamente se tiene una etiqueta asignada para cada canción, esto se puede deber a que el número de iteraciones para clasificar algunas canciones es muy alto, y como se había mencionado anteriormente a mayor iteraciones menor confianza, también se puede deber a que las canciones están muy alejadas una de otra por lo que no se cumple el porcentaje mínimo necesario de canciones previamente etiquetadas.

Canciones	Persona 1	Persona 2	...	Persona 30
Canción 1	$eti_{q1,1}$	$eti_{q1,2}$	...	$eti_{q1,30}$
Canción 2	$eti_{q2,1}$	$eti_{q2,2}$	...	$eti_{q2,30}$
⋮	⋮	⋮	...	⋮
Canción 480	$eti_{q480,1}$	$eti_{q480,2}$	...	$eti_{q480,30}$

Tabla 3.6: Representación de la matriz resultado

Es importante tener en cuenta que las etiquetas asignadas con KNN no son recomendaciones hechas con base en la canción original. Es por esto que para dar preferencia en la recomendación a las canciones similares a la original se multiplicó la etiqueta asignada por la inversa de la distancia entre la canción  $i$  y la canción original para la persona  $j$ . De tal forma que los valores altos indican que la canción tiene una alta probabilidad de gustarle y se encuentra cercana a la canción original.

Una vez teniendo esta multiplicación, se realizaron dos pruebas para determinar los parámetros óptimos: La primer prueba consistió en filtrar solo las canciones que mediante KNN tuvieran un valor mayor a 0.6, ya que como 0.5 es el valor medio, se tiene la misma probabilidad de gustarle o no, sin embargo con un valor de 0.6 es más probable que le guste. Una vez realizado este corte se multiplicó por la inversa de la distancia y se ordenó de mayor a menor. Se siguió este proceso para aumentar las probabilidades de que las canciones recomendadas le gusten ya que sin realizar este corte se puede dar el caso que las canciones sean recomendadas únicamente por estar cercanas a la canción original a pesar de que la probabilidad de gustarle sea muy baja. La segunda prueba fue usar todos los valores de KNN, sin filtros, multiplicar por el inverso de la distancia a la canción original y ordenar de mayor a menor.

Es necesario resaltar que la radio de una canción en Spotify consiste en una lista de reproducción formada por 50 canciones. Ya que las canciones que se tomaron como muestra de Spotify provienen de la radio de la canción seleccionada por cada persona, se consideró que para tener una comparación equiparable de las recomendaciones hechas por Spotify con las presentadas en esta tesis también se debe escoger una aleatoria de las primeras 50 recomendaciones. Sin embargo también se consideró que es relevante obtener los resultados.

Para determinar el número óptimo de canciones a recomendar se calculó para cada número de canción recomendada el promedio de las etiquetas asignadas por las personas hasta ese número de canción como se muestra en la tabla 3.7.

Número de Canciones	K=10	...	K = 20	Personas
1	$media_{1,k_{10}}$	...	$media_{1,k_{20}}$	$Personas_1$
2	$media_{2,k_{10}}$	...	$media_{2,k_{20}}$	$Personas_2$
3	$media_{3,k_{10}}$	...	$media_{3,k_{20}}$	$Personas_3$
⋮	⋮	⋮	⋮	⋮
168	$media_{168,k_{10}}$	...	$media_{168,k_{20}}$	$Personas_{168}$

Tabla 3.7: Cálculo del número óptimo de canciones

Por ejemplo, la tercera fila corresponde al promedio de las etiquetas asignadas por las personas desde la canción 1 hasta la 3 desde K=10 hasta K=20. Dado que no todas las canciones recomendadas fueron catalogadas por las personas, un punto importante es también saber cuántas personas de las 30 llegan hasta la recomendación  $i$  y ese es el número asignado en la quinta columna. Esto se realizó para ambas pruebas con filtro de 0.6 y sin filtro. El número máximo de canción recomendada fue 168 por lo que es el máximo valor de tabla.

### 3.3.1 Evaluación del clasificador

Para comparar el clasificador propuesto en este trabajo con el clasificador de Spotify se realizó una prueba t de dos muestras dependientes. La prueba t de muestras dependientes se usa para comparar la media de dos muestras en las que las observaciones de una muestra se pueden emparejar con las observaciones de la otra, por ejemplo, cuando se quiere hacer comparaciones de antes y después, o cuando se quieren comparar dos métodos o tratamientos diferentes aplicados a las mismas personas. En este trabajo, las muestras son dependientes porque la evaluación de los dos métodos de clasificación se realizó a las mismas personas.

Para poder determinar si las observaciones de las muestras X y Y han sufrido una diferencia significativa, se calcula para cada uno de ellos la diferencia entre los dos métodos evaluados  $d_i = x_i - y_i$ , donde  $i = 1, 2, \dots, n$  y  $n$  es el número de observaciones dentro de cada muestra. El análisis asume que las observaciones siguen una distribución normal con medias  $\mu_x$  y  $\mu_y$ . Así, la variable aleatoria  $D = X - Y$  también sigue una distribución normal con media  $\mu$  y varianza  $\sigma_d^2$  (Hsu and Lachenbruch, 2007).

El estadístico t de student se calcula como

$$t = \frac{\bar{d}}{(s_d/\sqrt{n})} \quad (3.7)$$

donde  $\bar{d} = \sum_{i=1}^n d_i$

Bajo la hipótesis nula de que la diferencia de las medias de las muestras es cero, podemos comparar el estadístico t calculado con el valor crítico t obtenido de una distribución t de Student con  $n - 1$  grados de libertad.

Cuando realizamos un contraste de hipótesis siempre partimos de que la hipótesis nula,  $H_0$ , es cierta y nos preguntamos hasta qué punto los resultados obtenidos no nos permiten rechazar nuestra hipótesis, es decir, si partíamos de la hipótesis de que la diferencia de las medias era nula, qué tan alejado está el valor de  $\bar{d}$  que obtuvimos, de cero. Debemos preguntarnos qué tan improbable es el resultado obtenido cuando la hipótesis nula es cierta, es decir, queremos conocer la probabilidad de que se presente nuestro resultado obtenido cuando  $H_0$  es cierta. Esta probabilidad recibe el nombre de p valor. Mientras más pequeño sea el p valor, más pequeña es la probabilidad de que se presente el resultado obtenido cuando asumimos cierta la hipótesis nula y, por lo tanto, se rechaza esta hipótesis.

Los tests de hipótesis pueden ser de una cola o de dos colas. Si la hipótesis alterna es del tipo "diferente de", entonces es una prueba de dos colas, en cambio, si la hipótesis alterna implica "mayor que" o "menor que", entonces es una prueba de una cola, ya sea derecha o izquierda, dependiendo de lo que se quiera probar con esta hipótesis.

Es necesario definir un nivel de confianza que nos indique en qué límites se encuentra la verdadera diferencia de las medias, generalmente se usa al 95% y para pruebas de una cola se calcula de la siguiente manera

$$\bar{d} \pm t^* \frac{s_d}{\sqrt{n}} \quad (3.8)$$

donde  $t^*$  es el valor al 5% de la distribución t en  $n - 1$  grados de libertad.

## CAPÍTULO 4

### Resultados

#### 4.1 Análisis de sensibilidad

Para determinar los valores óptimos de la etiqueta asignada a las canciones en la categoría de “*me gusta pero no guardaría la canción*” y del parámetro  $k$ , se realizaron pruebas de sensibilidad.

Para establecer la etiqueta óptima de las canciones clasificadas como “*me gusta pero no guardaría la canción*”, se comenzó por definir una malla de valores para  $k$  donde;

$$k \in \{10, 11, 12, \dots, 19, 20\}$$

Para este primer análisis únicamente fueron consideradas las personas que tenían por lo menos una canción en esa categoría, puesto que los resultados no se verían alterados para las demás personas.

Una vez establecida la malla de valores para  $k$ , se generó una malla de de posibles valores  $p$  definidos en el siguiente conjunto:

$$p \in \{0.5, 0.55, 0.6, \dots, 0.9, 0.95\}$$

Partiendo de estas dos mallas de valores, se corrió el algoritmo KNN para todas las combinaciones de valores  $k$  y  $p$ , y se calculó la diferencia absoluta de las probabilidades asignadas por el algoritmo y las etiquetas asignadas por los individuos, así, mientras más pequeña sea la diferencia, más parecidas son las probabilidades asignadas a la etiqueta real. Se calculó la media de estas diferencias para cada valor de  $p$  y de  $k$  para precisar la combinación en la cual se encontraba la menor diferencia. En la tabla 4.1 se puede ver la etiqueta óptima para cada valor de  $k$ .

Valor $k$	$p$
10	0.5
11	0.5
12	0.55
13	0.55
14	0.55
15	0.6
16	0.6
17	0.65
18	0.65
19	0.65
20	0.65

*Tabla 4.1: Etiquetas óptimas de clasificación “me gusta pero no guardaría la canción”*

Posteriormente se corrió nuevamente el algoritmo KNN, esta vez para los 30 individuos, con los 11 valores de  $k$  y cada uno con la etiqueta  $p$  correspondiente, y de igual forma se calculó la media de las diferencias absolutas de las probabilidades estimadas y las etiquetas asignadas por los individuos.

En la tabla 4.2 se puede ver que el valor óptimo se alcanza en  $k = 15$  con una media de 0.408. De la tabla 4.1 obtenemos que la etiqueta óptima para  $k = 15$  es  $p = 0.6$ .

Valor $k$	Media Diferencias
10	0.425
11	0.426
12	0.432
13	0.436
14	0.424
15	0.408
16	0.413
17	0.420
18	0.420
19	0.420
20	0.420

Tabla 4.2: Valor  $k$  óptimo

## 4.2 Análisis estadístico

### 4.2.1 Velocidad y aceleración

Para los valores de la velocidad y la aceleración de las canciones, se realizó un análisis de correlación con la finalidad de probar si existía alguna relación entre las canciones que contaban con la misma etiqueta, sin embargo, no se cumplió el objetivo buscado ya que no se encontró relación alguna entre las canciones, por lo que se optó por no utilizar la correlación como atributo. Las tablas 4.3a y 4.3b muestran a manera de ejemplo la correlación de la velocidad y la aceleración de la frecuencia para 6 canciones de una persona. Se señalan con tres asteriscos (\*\*\*) las canciones etiquetadas con 1, con dos asteriscos (\*\*) las canciones etiquetadas con 0.6 y con un asterisco (\*) las canciones etiquetadas con 0.

	Alt For Meg **	Celebrating Life **	Quiero Que Sepas ***	Chocobongo ***	Te Mando Flores *	Dime Cuando Volverás *
Alt For Meg **	1	-0.00103	0.001041	-0.00141	-0.00011	-0.00109
Celebrating Life **	-0.00103	1	6.93E-05	0.0009	0.000876	0.000514
Quiero Que Sepas ***	0.001041	6.93E-05	1	-0.00049	0.00194	-0.00102
Chocobongo ***	-0.00141	0.0009	-0.00049	1	-0.00265	-0.00096
Te Mando Flores *	-0.00011	0.000876	0.00194	-0.00265	1	0.003044
Dime Cuando Volverás *	-0.00109	0.000514	-0.00102	-0.00096	0.003044	1

(a) *Velocidad*

	Alt For Meg **	Celebrating Life **	Quiero Que Sepas ***	Chocobongo ***	Te Mando Flores *	Dime Cuando Volverás *
Alt For Meg **	1	-0.00049	0.001865	-0.0008	0.000536	-0.0001
Celebrating Life **	-0.00049	1	1.13E-05	0.000932	0.000422	0.00333
Quiero Que Sepas ***	0.001865	1.13E-05	1	-0.00128	0.000587	-0.0022
Chocobongo ***	-0.0008	0.000932	-0.00128	1	-0.00264	-0.00289
Te Mando Flores *	0.000536	0.000422	0.000587	-0.00264	1	0.00319
Dime Cuando Volverás *	-0.0001	0.00333	-0.0022	-0.00289	0.00319	1

(b) *Aceleración*Tabla 4.3: *Correlaciones.*

#### 4.2.2 Ciclos y periodos

Se calculó el promedio de los valores de los ciclos y los periodos para cada canción y se ordenaron de menor a mayor con la finalidad de ver si se lograba algún agrupamiento entre las canciones con la misma etiqueta. La tabla 4.4a muestra a manera de ejemplo los resultados obtenidos ordenando la media de los ciclos de la frecuencia para las canciones de cinco personas elegidas aleatoriamente. Pueden verse pequeñas agrupaciones de las canciones etiquetadas con 1 (indicadas con \*\*\*), que, en general, permanecen cerca y se encuentran separadas por una canción. También pueden verse agrupaciones de las canciones etiquetadas con 0 (indicadas con \*). La tabla 4.4b muestra los resultados obtenidos ordenando la media de los periodos de la frecuencia, y, de igual manera que con los ciclos, se observan pequeñas agrupaciones de las canciones. Ambas tablas muestran que la media de los ciclos y de los periodos son buenos indicadores para generar una primera separación de las canciones, es por esta razón que se incluyeron estas variables en el vector característico.

Individuo 5	Individuo 12	Individuo 15	Individuo 25	Individuo 27	Individuo 5	Individuo 12	Individuo 15	Individuo 25	Individuo 27
0.1855 ***	0.0340 *	1.1478 ***	0.0087 *	0.2016 ***	<b>0.0219</b>	0.0089 ***	0.0180 *	0.0104 ***	0.0221 *
0.6650 *	0.2281 ***	1.4090 ***	0.0859 *	0.2467 ***	0.0779 **	0.0203 ***	0.0262 ***	0.0896 *	0.0354 ***
0.7095 **	1.4211 ***	2.1171 **	0.2110 *	0.2783 *	0.0784 ***	<b>0.0292</b>	0.0287 *	0.1088***	0.1279 *
4.5603 ***	1.7542 ***	4.8501 ***	0.2338 ***	0.3604 ***	0.0971 **	0.0376 ***	0.0294 **	0.1357 ***	0.1407 *
6.5150 **	3.7796	16.0364 ***	0.2472 *	0.3696 ***	0.0991 *	0.0387 *	0.0328 *	0.3481 ***	0.1893 *
6.8569 *	4.3209 ***	18.7065 *	2.1772 *	1.1961 *	0.1064 *	0.0635 **	<b>0.0347</b>	<b>0.3600</b>	0.2518 *
7.3415 *	6.2108 **	19.6228 ***	2.2094 *	1.2501 ***	0.1088 *	0.0811 ***	0.0422 ***	0.4149 *	<b>0.3168</b>
8.7854 ***	10.4535 *	21.7873 **	2.6510 *	1.4434 ***	0.1252 ***	0.0859 *	0.0557 **	0.5267 *	0.6401 ***
8.8827 ***	12.5263 *	30.6128 ***	2.7324 *	1.7318 ***	0.1430 *	0.1018 *	0.0582 ***	0.5731 *	0.7843 ***
10.5147 *	18.0911 ***	31.7784 *	2.8956 *	<b>3.4944</b>	0.1563 ***	0.1826 **	0.0584 ***	0.6148 *	0.8821 ***
10.9028 *	18.5785 **	<b>33.2359</b>	3.2128 ***	4.0937 *	0.1699 **	0.2465 ***	0.0597 *	0.7848 *	0.9058 *
13.2102 **	27.3946 *	38.6328 **	<b>6.2084</b>	6.3592 *	0.1727 *	0.3556 ***	0.0716 ***	4.2041 *	2.9673 ***
14.8252 **	30.2862 ***	39.1101 *	9.1988 ***	7.8718 *	0.2362 ***	0.5898 ***	0.2151 ***	4.6287 ***	3.1456 ***
17.2360 *	<b>39.2568</b>	46.6146 ***	9.2409 ***	9.1226 *	1.4260 **	0.7683 ***	0.4829 **	4.8938 *	4.1576 ***
20.7472 ***	53.6256	69.0152 *	12.4697 *	37.7588 ***	1.6656 *	4.5079 ***	0.7625 ***	13.9777 *	4.4115 *
<b>49.9171</b>	143.5406 ***	69.2588 ***	159.8901 ***	52.5769 *	5.4552 ***	34.0361 *	1.1329 ***	119.1951 *	5.0582 ***

(a) *Media de ciclos*(b) *Media de periodos*Tabla 4.4: *Media de los ciclos y los periodos de la frecuencia.*

### 4.2.3 Varianza

El objetivo de calcular la varianza de la frecuencia y la amplitud de las canciones y ordenarla de menor a mayor, fue ver si esta variable era útil para clasificar las canciones. La tabla 4.5 muestra los resultados de la varianza de la frecuencia y la amplitud, respectivamente, para las canciones clasificadas de 5 personas elegidas aleatoriamente. En verde se muestran las canciones etiquetadas con 1, en naranja las canciones etiquetadas con 0.75, en rojo las canciones etiquetadas con 0 y en azul la canción original de la cual se generó la lista.

Pueden verse pequeñas agrupaciones para las canciones de una misma persona, por ejemplo, para el individuo 15, la mayoría de las canciones etiquetadas con 1 muestran una menor varianza de la frecuencia, mientras que para el individuo 25 las canciones que muestran una menor varianza de la frecuencia son las etiquetadas con 0.

Dado que la varianza parecía ser un buen indicador para separar las canciones, se realizaron gráficas con los ciclos, periodos, velocidad y aceleración para ver cómo se comportaba en conjunto con otras variables. Este análisis se realizó tanto para la frecuencia como para la amplitud. La figura 4.1 muestra los resultados obtenidos para 5 personas elegidas aleatoriamente, en ella se comparan la varianza, los ciclos y los periodos

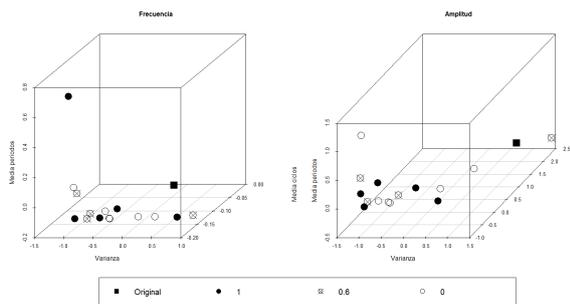
Individuo 5	Individuo 12	Individuo 15	Individuo 25	Individuo 27	Individuo 5	Individuo 12	Individuo 15	Individuo 25	Individuo 27
5567341***	977767***	3268397**	554241*	4251548*	4020034***	499483*	0.00167278**	391726*	4045416***
6465729***	1019117*	7472774***	1746930*	5809548***	4844779*	845635***	0.0023824***	1729276*	4279922*
7851113*	3006840***	9097276***	4727735*	6061569***	4955120.44**	3023825***	0.0024340*	3774531*	5071786***
8749523**	6499563***	14696690***	5886872*	6139785*	5111542***	5278395***	0.0028451***	5478404*	5464112*
9268568**	12644823**	18398903***	6997819*	7349887***	6872347.41**	6223188.92**	0.00362055**	6412147*	5805819***
11429073**	19880262***	27388027***	7858471*	7432874***	11332562.8**	11288510*	0.0036430***	6452875*	6764509***
14778009*	20646036*	27945161*	<b>8182114</b>	12353633***	11921816*	13132078***	0.00423274*	6902172*	8566290***
16424402***	23928820***	38061879*	9210163*	14337765***	11933164*	16519347***	0.00451231**	<b>7209055</b>	11387345***
18760619***	31803085***	51646020***	13281244*	<b>16790451</b>	12252776***	20221184*	0.0062135***	11529365*	<b>13018096</b>
21920381*	40661271*	54618739*	13447074***	19919254*	12849733*	21326025***	0.0067985***	12811434*	14462370*
21952387*	<b>41705868</b>	59344187*	13942153*	27872460***	17136056***	31654934.9**	0.0069523***	13052107***	17829982*
<b>31500392</b>	63264889**	59973381***	24198775***	28455104*	26781460*	<b>33895951.9</b>	0.0070056*	15509957***	22256382*
34086514*	69004288***	60694323**	27259620*	33434662***	31725821***	47588393*	0.0080489***	16576097*	25258163***
41995981*	73637067*	72501116***	28291076***	36155217*	32290201*	57027889***	0.0102198***	21736458***	26020332***
53671280***	78487908***	73290526***	31012457***	40586307*	<b>38127828</b>	61262222***	0.0110506*	22860223***	26434914*
59179851**	92780183***	<b>74144405</b>	41868143***	50590303*	51294205.3**	75599880***	<b>0.01204003</b>	26080582***	30314857*

(a) Frecuencia

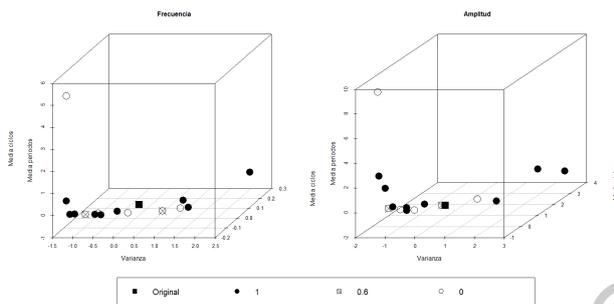
(b) Amplitud

Tabla 4.5: Varianza.

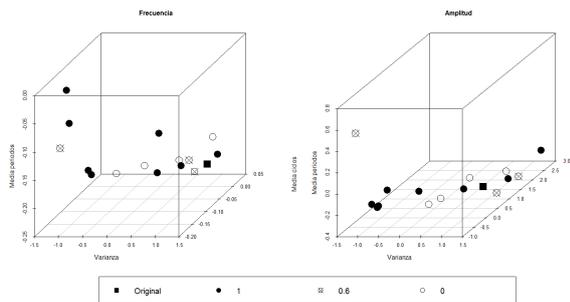
y, en general, se puede ver una buena agrupación de las canciones, ya que en la mayoría de los casos, las canciones que tienen la misma etiqueta se encuentran en la misma zona de la gráfica. Por otro lado, en la figura 4.2 se compara la varianza con la aceleración y la velocidad. De igual manera se puede observar una buena agrupación de las canciones tanto para la frecuencia como para la amplitud.



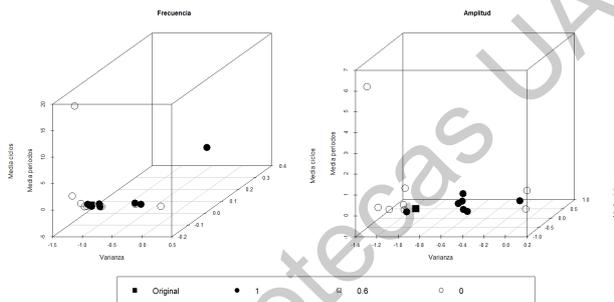
(a) Individuo 5



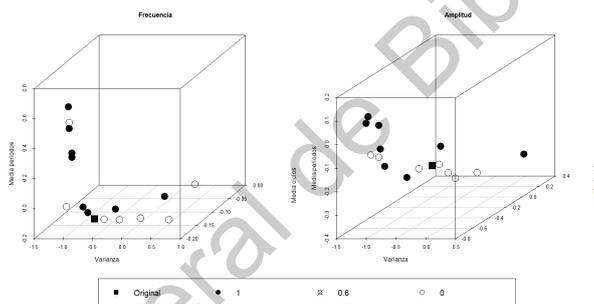
(b) Individuo 12



(c) Individuo 15



(d) Individuo 25



(e) Individuo 27

Figura 4.1: Varianza, media de ciclos y media de periodos

Dirección General de Bibliotecas UAQ

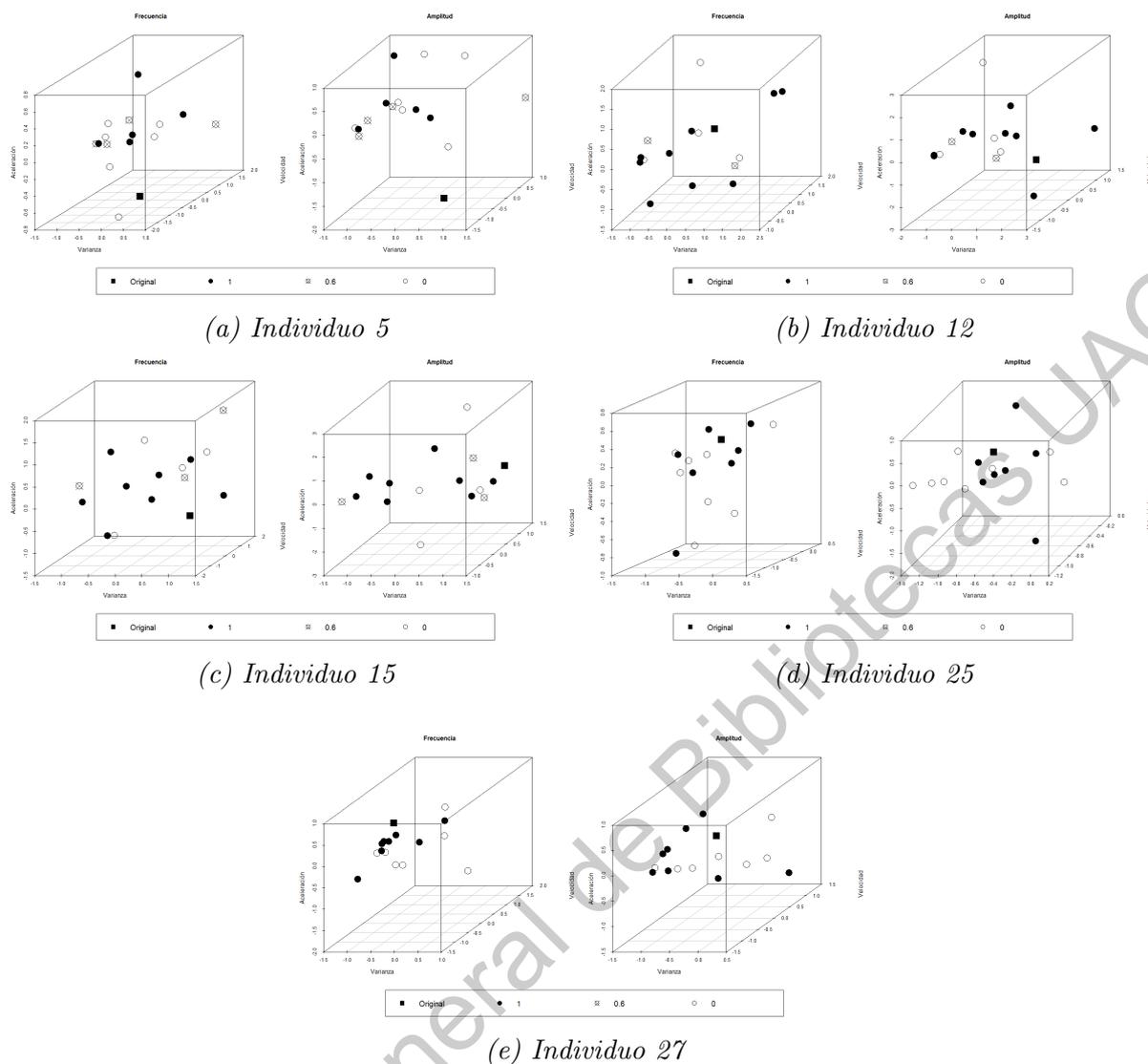


Figura 4.2: Varianza, media de aceleración y media de velocidad

### 4.3 Resultados del clasificador

Como se mencionó en la sección 3.3, la radio de una canción en Spotify consiste en una lista de reproducción formada por 50 canciones, por lo que, para tener una consideración equiparable de las recomendaciones hechas por Spotify con las presentadas en esta tesis, se eligió aleatoriamente una de las primeras 50 canciones recomendadas.

Además, se decidió comparar la recomendación de Spotify con la mejor recomendación obtenida utilizando el método propuesto en el presente trabajo. Para ello, se calculó el número óptimo de canciones a recomendar de acuerdo a lo presentado en la

sección 3.3. La tabla 4.6a muestra los resultados obtenidos sin corte de probabilidad y la tabla 4.6b muestra los resultados obtenidos de las canciones que mediante el algoritmo empleado tuvieran una probabilidad de agrado mayor a 0.6.

Canciones	Valor k	Media	Número de Personas
Óptimo (24)	15	0.5431081	23
50	15	0.511726416	23

(a) Sin corte

Canciones	Valor k	Media	Número de Personas
Óptimo (14)	15	0.5674603	10
50	15	0.3494949	3

(b) Con corte en 0.6

Tabla 4.6: Obtención del óptimo de canciones.

En la tabla 4.6a puede verse que la probabilidad más alta sin hacer ningún corte en la probabilidad se alcanza al recomendar 24 canciones, sin embargo, esta probabilidad se encuentra por debajo de la que se alcanza al recomendar 14 canciones realizando un corte en 0.6, como se muestra en la tabla 4.6b. Por esta razón, se decidió que el número de canciones óptimo a recomendar es 14, tomando en cuenta  $K = 15$  y con corte de probabilidad en 0.6.

#### 4.4 Evaluación del clasificador

Una vez elegidas las variables a utilizar para la clasificación de las canciones, se aplicó el algoritmo mencionado en la sección 3.2

La hipótesis de este trabajo es que un clasificador de canciones basado únicamente en elementos de series de tiempo tiene un menor porcentaje de canciones mal clasificadas que un clasificador basado en etiquetas, es decir, un clasificador de canciones que se base en elementos temporales de las canciones puede funcionar de mejor manera que un clasificador basado en etiquetas.

Para probar si hubo una mejora en la recomendación, se realizaron dos pruebas t de dos muestras. La primera prueba t compara los resultados obtenidos de la recomendación realizada por spotify contra los resultados obtenidos del recomendador propuesto tomando en cuenta las primeras 50 canciones recomendadas. La segunda prueba t compara los resultados obtenidos de la recomendación realizada por Spotify contra los resultados obtenidos del recomendador propuesto tomando en cuenta el número de canciones óptimo a recomendar.

Las pruebas t utilizadas fueron para muestras dependientes, ya que se probaron los dos clasificadores a las mismas personas, y dado que se busca probar que hay una mejora con el clasificador propuesto, las hipótesis que se contrastaron en la prueba fueron las siguientes:

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 > \mu_2$

Donde  $\mu_1$  es la media de la etiqueta de las canciones recomendadas con la metodología propuesta y  $\mu_2$  es la media de la etiqueta de las canciones recomendadas por Spotify.

Para realizar la primer comparación, para cada persona se tomó aleatoriamente una de las primeras 50 canciones recomendadas que tuvieran una probabilidad de agrado mayor o igual a 0.6 y esta canción se dio a escuchar a la persona correspondiente para etiquetarla con 1 si le gustó la canción y la guardó, con 0.6 si le gustó la canción pero no la guardó y con un 0 si no le gustó. También se tomó aleatoriamente una de las quince canciones tomadas de la radio de spotify que ya estaban previamente etiquedadas. Con esta nueva base de datos se calcularon las diferencias entre ambas muestras, así como la media y desviación estándar de cada una de las muestras y de las diferencias. La tabla 4.7 muestra los resultados obtenidos.

	Clasificador propuesto	Spotify	Diferencia
Media ( $\mu$ )	0.468965517	0.466666667	-0.013333333
Desviación estándar (s)	0.404530988	0.464856918	0.511747064

Tabla 4.7: Resultados obtenidos comparando las primeras 50 canciones recomendadas y Spotify

Para probar si existe alguna mejora, se obtiene el estadístico t como se describe en la ecuación 3.7, siendo  $\bar{d} = -0.013333333$  la media de las diferencias entre las dos muestras,  $s_d = 0.511747064$  la desviación estándar de las diferencias, y  $n = 30$  el tamaño de las muestras. Esto da como resultado un estadístico t de -0.1427.

Para obtener el valor crítico de t, se usó un nivel de significancia de 5 % para una distribución t de 29 grados de libertad, lo cual da como resultado un valor crítico de t de 1.6991. La figura 4.3 ilustra los resultados obtenidos.

Además, se utilizó el software estadístico R para obtener el p valor, el cual es de 0.079, y se calculó el intervalo de confianza al 95 % de acuerdo a la ecuación 3.8, obteniendo un límite inferior de -0.1720858 y un límite superior infinito. Como el p valor es mayor al nivel de significancia ( $0.556 > 0.05$ ), no se tienen evidencias suficientes para rechazar la hipótesis nula y por lo tanto, se puede concluir que no hay una mejora significativa en el recomendador propuesto.

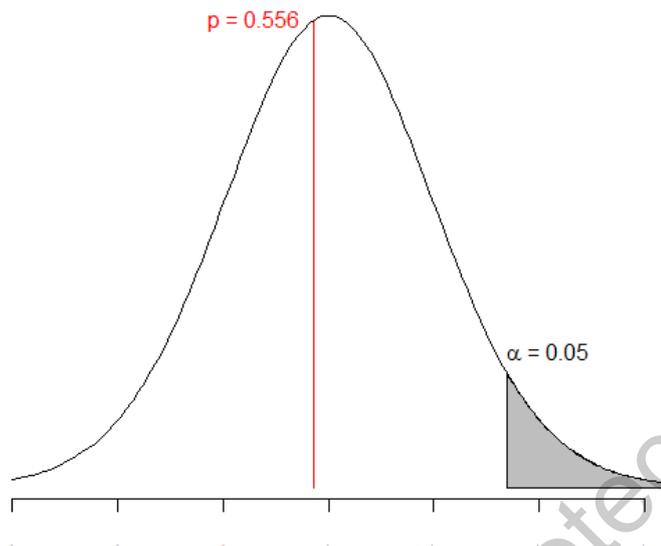


Figura 4.3: Resultados de la prueba  $t$  comparando las primeras 50 canciones recomendadas y Spotify.

Para realizar la segunda comparación, el número de canciones recomendadas óptimo para obtener una mayor cantidad de canciones que sean del agrado de la persona es 14, por lo que para cada persona se tomó aleatoriamente una de las primeras 14 canciones recomendadas que tuvieran una probabilidad de agrado mayor o igual a 0.6 y esta canción se dio a escuchar a la persona correspondiente para etiquetarla con 1 si le gustó la canción y la guardó, con 0.6 si le gustó la canción pero no la guardó y con un 0 si no le gustó. También se tomó aleatoriamente una de las quince canciones tomadas de la radio de spotify que ya estaban previamente etiquedadas. Con esta nueva base de datos se calcularon las diferencias entre ambas muestras, así como la media y desviación estándar de cada una de las muestras y de las diferencias. La tabla 4.8 muestra los resultados obtenidos.

	<b>Clasificador propuesto</b>	<b>Spotify</b>	<b>Diferencia</b>
Media ( $\mu$ )	0.506666667	0.466666667	0.04
Desviación estándar (s)	0.3990794	0.464856918	0.652633921

Tabla 4.8: Resultados obtenidos comparando el número de canciones óptimo y Spotify

Para probar si existe alguna mejora, se obtiene el estadístico t como se describe en la ecuación 3.7, siendo  $\bar{d} = 0.04$  la media de las diferencias entre las dos muestras,  $s_d = 0.652633921$  la desviación estándar de las diferencias, y  $n = 30$  el tamaño de las muestras. Esto da como resultado un estadístico t de 0.3357.

Dado que los grados de libertad y el nivel de significancia es el mismo que el de la primer prueba t, sabemos que el valor crítico de t es de 1.6991. La figura 4.4 ilustra los resultados obtenidos.

Además, se utilizó el software estadístico R para obtener el p valor, el cual es de 0.369757, y se calculó el intervalo de confianza al 95 % de acuerdo a la ecuación 3.8, obteniendo un límite inferior de -0.162458 y un límite superior infinito. Como el p valor es mayor que el nivel de significancia ( $0.369757 < 0.05$ ), no se tienen evidencias suficientes para rechazar la hipótesis nula y por lo tanto, se puede concluir que no hay una mejora significativa en el recomendador propuesto.

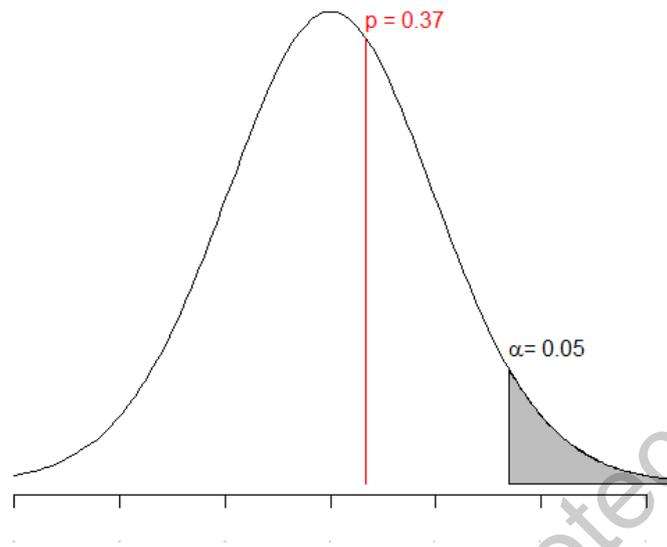


Figura 4.4: Resultados de la prueba  $t$  comparando el número de canciones óptimo y Spotify

## CAPÍTULO 5

### Discusión y conclusiones

#### 5.1 Discusión

En el desarrollo del algoritmo se hicieron muchos supuestos sobre algunos de los valores usados, los cuales representan una gran oportunidad de mejora para la clasificación propuesta. Por cuestiones de tiempo y limitaciones en capacidad de cómputo no se exploraron todos los escenarios posibles que, en futuras investigaciones se pudieran llegar a considerar. Entre estos escenarios se encuentran:

1. La cantidad de datos utilizados. En el presente trabajo se usaron los primeros 30 segundos suavizados de la canción. Contemplar la canción completa y sin promediar las muestras por segundo podría suponer una mejora en la clasificación ya que se contaría con la información completa de la canción y se podrían obtener más datos relevantes para la clasificación.
2. El peso de las variables usadas para la clasificación. Se asumió que los 10 elementos usados para caracterizar las canciones afectaban en la misma medida la recomendación, lo cual es un supuesto razonable y el más simple. Explorar todas las posibles combinaciones de pesos implicaba una investigación que iba más allá de los objetivos planteados, por lo cual no se realizó el análisis.

#### 5.2 Conclusiones

La recomendación musical es un factor que desde hace tiempo está tomando importancia para los sitios o aplicaciones de música. Por ello, se han buscado diferentes formas de mejorar este servicio y así brindar una mejor recomendación para los usuarios.

En el presente trabajo se muestra una alternativa a la recomendación musical, la cual se basa en el análisis matemático y estadístico de los elementos temporales de la música, para ello se optó por realizar una modificación al algoritmo *K Nearest Neighbors* y se realizaron dos pruebas para comparar los resultados obtenidos con el método de recomendación propuesto y la recomendación realizada por Spotify.

Los resultados obtenidos en este trabajo no muestran evidencia de mejora en la clasificación respecto al algoritmo de Spotify, sin embargo, puede verse que no hay una diferencia significativa en la recomendación obtenida a pesar de que se logró una media más alta con el clasificador propuesto.

Cabe destacar que el algoritmo empleado es muy simple y las canciones a recomendar están limitadas a 480, mientras que spotify cuenta con un número de canciones más amplio a recomendar, el cual asciende a 50 millones (Spotify, 2020). Por esta razón se considera que se pueden construir clasificadores más sofisticados que empleen los elementos temporales expuestos en el presente trabajo para generar mejores recomendaciones, y permitiendo mayor diversidad en estas, abriendo muchas posibilidades para perfeccionar los algoritmos de recomendación musical en la actualidad.

## APÉNDICE A

### Algoritmo

```

knn<-function(valor_k,matriz_caracteristicas,vector_etiquetas,min
_etiquetado){

#Genera matriz de distancias
matriz_distancias<- as.matrix(dist(matriz_caracteristicas
[,,-1],diag = TRUE,upper=TRUE))
nombres<-matriz_caracteristicas[,1]

#Determina cuales y en que óposicin se encuentran las
canciones que úan no tienen etiqueta y las guarda en el
vector sin_etiqueta
canciones<-length(vector_etiquetas)
sin_etiqueta<-vector(mode="numeric")
sin_etiqueta<-which(is.na(vector_etiquetas))

#Concatena matriz de distancias con vector de etiquetas
matriz_distancias<-cbind(matriz_distancias,vector_etiquetas)

#Asigna el valor 1 para la primera generacion
generaciones<-vector(length = canciones,mode="numeric")
for(i in 1:canciones){
  if(is.na(vector_etiquetas[i])){
    generaciones[i]=NA
  }else{
    generaciones[i]=1}
}

```

```

}

#Genera KNN para todas las canciones sin etiqueta
matriz_completa<-cbind(matriz_distancias,as.vector(
  generaciones))
probabilidad<-matriz_completa[,canciones+1]
generacion<-matriz_completa[,canciones+2]
for (r in 1:6){
  gen<-(0.9)^r

  if(length(sin_etiqueta)>0){

    for(i in 1:length(sin_etiqueta)){
      orden_distancias<-matriz_completa[order(matriz_completa[,
        sin_etiqueta[i]]),]
      k_primeras<-orden_distancias[1:valor_k+1,canciones+1]
      if(length(na.omit(k_primeras))>=min_etiquetado){
        probabilidad[sin_etiqueta[i]]<-mean(na.omit(k_primeras))
        generaciones[sin_etiqueta[i]]<-gen
      }
    }

  }
matriz_completa[,canciones+1]<-probabilidad
matriz_completa[,canciones+2]<-generaciones

sin_etiqueta<-which(is.na(matriz_completa[,canciones+1]))
}else{
  break
}
}

```

```
resultado<-cbind(as.vector(nombres),matriz_completa[,
  canciones+1]*matriz_completa[,canciones+2])
return(resultado)
}
```

Dirección General de Bibliotecas UAQ

## APÉNDICE B

### Valores obtenidos de la frecuencia y amplitud

Los valores mostrados en el presente anexo son los obtenidos una vez realizada la estandarización de acuerdo a la ecuación 3.6. Además, se muestra la clasificación que otorgaron las personas a la lista de canciones que escucharon originalmente, la cual fue obtenida a partir de la radio de Spotify.



*Figura B.1: Conjunto de datos empleados y se puede consultar usando el siguiente DOI: 10.6084/m9.figshare.11933157.*

## Bibliografía

- Apel, W. (2003). *The Harvard dictionary of music*. Harvard University Press.
- Apple (2019). Genius. <https://www.apple.com/legal/internet-services/itunes/us/genius.html>. [En línea; acceso 01-Abril-2019].
- Beran, J. (2003). *Statistics in musicology*. CRC Press.
- Cabrejo, O. J. G. and Sánchez, G. M. (2006). Sobre la utilización del análisis de fourier, análisis espectral singular y redes neuronales artificiales en estratigrafía. parte 1: Teoría y caso sintético. *Geología Colombiana*, 31:105–120.
- Colomer Blasco, L. (2016). Acústica musical.
- Gangadhar Shobha, S. R. (2018). Machine learning. In Rao, C. R. and Gudivada, V. N., editors, *Handbook of Statistics, Volume 38: Computational Analysis and Understanding of Natural Languages*, chapter 8, pages 197–228. Elsevier, Oxford.
- Germain, A. and Chakareski, J. (2013). Spotify me: Facebook-assisted automatic playlist generation. In *Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on*, pages 025–028. IEEE.
- Hsu, H. and Lachenbruch, P. A. (2007). Paired t test. *Wiley encyclopedia of clinical trials*, pages 1–3.
- James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.
- Kaitila, J. (2017). A content-based music recommender system. Master's thesis.
- Kim, D., Kim, K.-s., Park, K.-H., Lee, J.-H., and Lee, K. M. (2007). A music recommendation system with a dynamic k-means clustering algorithm. In *Machine Learning and*

*Applications, 2007. ICMLA 2007. Sixth International Conference on*, pages 399–403. IEEE.

Montenegro, Á. et al. (2009). Análisis espectral. Technical report, Universidad Javeriana-Bogotá.

North, A. and Hargreaves, D. (1995). Subjective complexity, familiarity, and liking for popular music. *psychomusicology: A journal of research in music cognition*, 14 (1-2), 77–93.

of Science, E. C. (2018). Lesson 6: The teriodogram. In *Course Stat 510 - Applied tyme series analysis*. Cambridge MA. .

Pandora (2019). Recommendations and personalization on pandora. [https://help.pandora.com/s/article/000001078?language=en\\_US](https://help.pandora.com/s/article/000001078?language=en_US). [En línea; acceso 01-Abril-2019].

Patel, J. K. and Gopi, E. (2015). Musical notes identification using digital signal processing. *Procedia Computer Science*, 57:876–884.

Schott, M. (2019). K-nearest neighbors (knn) algorithm for machine learning. <https://medium.com/capital-one-tech/k-nearest-neighbors-knn-algorithm-for-machine-learning-e883219c8f26>. [En línea; acceso 31-October-2019].

Shumway, R. H. and Stoffer, D. S. (2015). *Time series analysis and its applications: with R examples*. Springer.

Spotify (2019). Spotify radio. [https://support.spotify.com/es/using\\_spotify/features/spotify-radio/](https://support.spotify.com/es/using_spotify/features/spotify-radio/). [En línea; acceso 01-Abril-2019].

Spotify (2020). Company info. <https://newsroom.spotify.com/company-info/>. [En línea; acceso 04-Marzo-2020].

the National Association for Music Education MENC (1984). *Update: applications of research in music education*. Sage.

Zhang, Y. C., Séaghdha, D. Ó., Quercia, D., and Jambor, T. (2012). Auralist: introducing serendipity into music recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 13–22. ACM.

Zhu, X. and Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130.

Dirección General de Bibliotecas UAQ