



**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO  
FACULTAD DE CIENCIAS**

**CASO DE ESTUDIO DE CIENCIA DE DATOS EN LA POPULARIDAD DE LAS  
CANCIONES DE SPOTIFY**

**TESIS**

QUE PARA OBTENER EL TÍTULO DE  
**ACTUARÍA**

PRESENTA:  
**CRISTINA SÁNCHEZ MAYO**

DIRECTOR DE TESIS:  
**DR. FRANK PATRICK MURPHY HERNANDEZ**



CIUDAD UNIVERSITARIA, CD. MX.  
2024



**UNAM – Dirección General de Bibliotecas**

**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

Todo el material contenido en esta tesis está protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (Méjico).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.



## Agradecimientos

Quiero expresar mi más profundo agradecimiento a todas las personas que me han acompañado en este camino y han hecho posible la realización de esta tesis.

A mis padres, Juana Mayo y Miguel Sánchez, y a mis hermanos, Juan Sánchez y Alejandro Sánchez, por su amor incondicional, por creer en mí, y por darme la fuerza necesaria para superar los momentos difíciles. A pesar de las adversidades, siempre me han apoyado de la mejor manera, llenando mi vida de felicidad y motivación. Su compañía y apoyo emocional han sido esenciales para llegar hasta aquí.

También quiero agradecer de manera especial a mi director de tesis, el Dr. Frank Patrick Murphy. Su inigualable guía, paciencia y apoyo constante, así como su compromiso durante todo el tiempo que tomó este proyecto, fueron cruciales para su finalización. A lo largo de este proceso, él estuvo siempre presente, brindándome su experiencia y valiosos consejos en cada etapa, lo que ha sido fundamental para el desarrollo de este trabajo.

Finalmente, extiendo mi gratitud a mis sinodales y a los profesores de la H. Facultad de Ciencias, quienes me han proporcionado las bases del conocimiento necesarias para la elaboración de esta tesis, así como el entorno adecuado que hicieron posible llegar a este punto.



# Índice general

Introducción	7
Capítulo 1. Preliminares	11
1. A través de la música	11
2. Géneros más importantes en la historia	13
3. Digitalización de la música	14
4. Comercialización de la música en la actualidad	15
Capítulo 2. Modelos estadísticos	23
1. La base de datos Iris	23
2. Técnicas de aprendizaje (supervisado y no supervisado)	24
3. Análisis discriminante lineal (LDA)	25
4. Árboles de decisión con clasificación	38
5. Comparación de los modelos LDA y Árboles	44
6. Método de Bootstrap	52
Capítulo 3. Análisis Exploratorio	55
1. Limpieza de datos	55
2. Descripción de Variables	59
3. Relación de Variables	62
4. Análisis de variables por género	65
Capítulo 4. Aplicación Modelos LDA y árboles	85
1. Análisis discriminante lineal (LDA)	86
2. Árboles de decisión con clasificación	91
Capítulo 5. Informe ejecutivo	103
Capítulo 6. Conclusiones	113
Bibliografía	117



## Introducción

Actualmente la Ciencia de Datos tiene un impacto de gran relevancia, esto debido a que la ciencia de datos es un campo interdisciplinario que utiliza procesos y algoritmos para extraer valor de los datos. Asimismo, encuentra la manera de obtener el máximo provecho de los datos para arrojar resultados en distintas áreas. La ciencia de datos se ha vuelto de gran relevancia en el mundo y en México, de esta manera en la Facultad de Ciencias se puede cursar la carrera de Ciencia de Datos.

El objetivo de la tesis no es presentar la teoría estadística de algún modelo predictivo específico. El objetivo es presentar a la ciencia de datos por medio de un caso de estudio haciendo uso de las herramientas que se aprendieron a lo largo de la carrera de Actuaría como lo es: Estadística multivariada, Estadística no paramétrica, Programación en R y Base de Datos.

Por otro lado, este proyecto de tesis toma como punto de partida el concurso que hubo en 2009 por parte de Netflix (es un servicio de streaming por suscripción que les permite a sus miembros ver series y películas sin publicidades en un dispositivo con conexión a internet) en el cual se buscaba mejorar el algoritmo para predecir las preferencias de los usuarios en cuanto a películas. Con estos hechos precursores muchas otras plataformas de streaming han hecho estos concursos, puntualmente Spotify también ha aplicado algoritmos predictivos enfocados en los gustos de las personas.

Spotify es un servicio de música, podcasts y videos digitales en streaming que te da acceso a millones de canciones y otros contenidos de artistas de todo el mundo. Spotify tiene cierta competencia, por ende, hace públicos datos en específico, para este proyecto se programa la obtención de datos a través de la API, datos que son de utilidad para la realización de este proyecto.

La tesis cuenta con cinco capítulos, en el primero se encuentran los preliminares, en dicho capítulo se habla de los antecedentes de la música, se adentra a grandes rasgos en cómo fue la música a través del tiempo hasta la actualidad, también se hace

una descripción de los géneros musicales más importantes para llegar así a la parte en la que la música sufre un punto de inflexión al ser digitalizada y comercializada de una manera completamente diferente a como se venía haciendo. En gran parte que se digitalizara la música se considera que fue un logro para la sociedad ya que tenían acceso más libre a la música y con ello se logró comercializar de manera digital. Esto se ve en la actualidad, ya que existen muchas plataformas de streaming en las cuales hay millones de canciones que pueden ser consultadas a cambio de cierto pago. En este capítulo también se habla del crecimiento de usuarios que ha tenido Spotify y la importancia de este mismo.

En el capítulo dos: Modelos estadísticos, se hace una descripción de los modelos que se usan a lo largo de este proyecto. En primera instancia se habla de la base de datos Iris que será en donde se apliquen los modelos descritos. Iris es una base de datos la cual se presta para aplicar y ejemplificar con mayor claridad los modelos y los algoritmos. Posteriormente se plantean los dos modelos de aprendizaje supervisado:

- 1) Análisis discriminante lineal conocido como modelo LDA que consiste en hacer una clasificación en grupos dadas las similitudes que presenten las variables, para el caso de Iris sería las similitudes entre el ancho y alto de los sépalos y pétalos.
- 2) Modelo de árboles de clasificación que consiste en hacer una clasificación de las variables en función de sus atributos mediante reglas binarias, que se ejemplifican con la base de datos Iris.

En otra sección de este capítulo dos, se hace la comparación de estos dos modelos y se comparan los resultados obtenidos que se obtienen de aplicar estos modelos. Es importante mencionar que para hacer estas comparaciones se hace por medio de las matrices de confusión que dan un panorama general de asertividad al aplicar los modelos predictivos LDA y árboles, por ende, se puede hacer una comparación. Por último, en esta capítulo se abarca un método llamado Bootstrap, este método se usa debido a que se necesita completar la base de datos extraída por cada género, ya que si no es lo suficientemente robusta se puede completar de manera aleatoriamente con reemplazo dejando así una base de datos e información más consistente.

En el capítulo tres: Análisis exploratorio, primeramente se realiza una limpieza de datos, en la que se verifica que la información sea concisa, congruente, que los datos estén completos, que la información descargada este organizada de manera que

sea útil y que las variables estén homologadas, posteriormente se hace un análisis exploratorio de los datos, en donde el objetivo es explorar, resumir y entender de forma clara la naturaleza de los datos recolectados, también se obtiene información importante por medio de gráficos estadísticos, ponderaciones de las variables, correlaciones entre las variables, análisis de la utilidad de las variables y principales tendencias de los datos.

El capítulo cuatro: Aplicación de Modelos LDA y árboles. Se realiza la aplicación de los modelos estadísticos mencionados en el capítulo tres. Esta aplicación se realiza en la base de datos ya limpia que se tomó con la ayuda de la API de Spotify. La aplicación se efectúa por géneros musicales. En este capítulo se refuerza el resultado obtenido gracias a las matrices de confusión que muestran si cierto grupo aleatorio de canciones serán clasificadas por los modelos estadísticos como populares o no.

En el capítulo cinco se presenta un reporte ejecutivo que contiene de manera general los resultados y conclusiones de la investigación a modo de que sea comprensible para los ejecutivos, este proyecto se realizó con la intención de generar un bien social, es decir se plantea generar dinero con él, con el fin de saber si una canción puede ser popular o no, dejando mayores ganancias, por lo que este reporte ejecutivo está dirigido en términos de ganancias y muy resumido para ser usado por directivos que no necesariamente tienen estudios en una carrera a fin al área físicomatemático e ingenierías.



## Capítulo 1

# Preliminares

Este capítulo se muestran los hallazgos en la música, específicamente se da una descripción e historia acerca de cómo es que se clasificaron los géneros musicales, de dónde surge la necesidad de clasificar la música por género, como ha sido el proceso para llegar a la digitalización de la música, cómo es que se comercializa la música en la actualidad y por último se habla sobre Spotify una pieza de suma relevancia para esta investigación.

### 1. A través de la música

Es difícil saber cuándo exactamente empezó la música, pero los relatos populares cuentan que la música tuvo un origen divino y que su sonido representaba el mensaje de la naturaleza y del hombre. El ser humano sintió la necesidad de expresarse y de comunicarse. Buscaba cómo hacerlo: emitía ruidos, gritaba, gemía, imitaba, entre otros, necesitaba un lenguaje, entonces ocurrió hace aproximadamente 40 mil años, cuando el primer hombre (*Homo sapiens*) fue capaz de imitar los sonidos de la naturaleza, que eran diferentes a los que hacía cuando estructuraba su lenguaje. A este hombre se le conoce como *Homo musicus*. “*Homo musicus*, el hombre musical, que crea, interpreta y escucha música, es mayor que el *Homo sapiens*.. El hombre hacía una especie de música incluso cuando no sabía cómo medir las cosas o contarlas correctamente, y el concepto mismo de números todavía era un destello en su cerebro. Hizo música cuando no pudo encontrar la razón de los fenómenos naturales, la lluvia, el granizo y la sequía a su alrededor. Durante los muchos siglos de su desarrollo, la cultura europea mantuvo un .“centro musical” que jugó un papel activo en el proceso educativo... La música y el habla comparten una función común: la comunicación social.” <sup>1</sup>. Por ello, la música ha sido de gran relevancia a lo largo de la historia y que también comienza como una necesidad para la comunicación social.

#### ■ La música de Mesopotamia

La religión fue de suma importancia, puesto que los textos sagrados hablan de un tipo de solistas y del acompañamiento instrumental. En las esculturas

---

<sup>1</sup>Dina Kirnarskaya(2009) The Natural MusicianOn abilities, giftedness, and talent.

se observan cantantes e instrumentistas. Los músicos sumerios y babilonios se dividían en dos grupos: nar y gala. El grupo nar son los que cantaban las alabanzas de dioses o reyes, por otro lado, el grupo gala, los que cantaban lamentos.

- **La música de Egipto**

Sus temas son siempre religiosos: textos de himnos y salmos, estatuas de dioses músicos y retratos de músicos de los templos, de bailarines religiosos o de comidas sagradas que se acompañaban con música. Existen representaciones donde se ve a personas haciendo signos con las manos a cantantes e instrumentistas, como si fueran representaciones de notas.

- **La música en la antigua Grecia**

Es importante el impacto que tiene la antigua Grecia debido a que en este punto alcanzó un significado artístico parecido con el que contamos en la actualidad. La palabra música deriva de mousiké, que se aplicaba a toda expresión artística elevada. Así, la música pasó a ser un elemento de perfección, un instrumento para mejorar la conducta y el pensamiento de los seres humanos. Bajo estos períodos es importante resaltar que la música comenzaba con un componente religioso y por ende solo era para algunas personas que estuvieran privilegiadas bajo estas circunstancias. Se podría considerar que de cierta forma tenía algo divino. Cuando se llega a la antigua Grecia se aprecia desde otro ángulo y es el artístico.

- **Barroco**

El Barroco fue una época de arte exquisito, tanto en la pintura como en la música. Las primeras composiciones barrocas, en las que predominaba la música instrumental, las hicieron los maestros italianos, como: Corelli, Albinoni (renombrado por su adagio) y Vivaldi (famoso por su obra Las cuatro estaciones).

- **El romanticismo**

Lo que hay en mi corazón debe salir y por eso lo escribo, dijo Beethoven, siendo el primer músico romántico. El Romanticismo comprende casi todo el siglo XIX. En los siglos XX y XXI, la música se divide en dos etapas: la música moderna, que va de 1900 a 1945, y la música contemporánea, desde 1946 hasta la actualidad. Los músicos siempre buscan nuevos ritmos y sonidos.

## 2. Géneros más importantes en la historia

La música ha tenido a lo largo de la historia la necesidad de clasificarse, ya que la música siempre está buscando reinventarse por ello surge la necesidad de crear también nuevos géneros para ser clasificada. Cada década y sus acontecimientos sociales han determinado en gran parte los géneros musicales, dejando ver el gran impacto que tiene la música. A continuación, se muestra una breve compilación de los géneros más importantes a lo largo de la historia.

El jazz es la combinación de la tradicional música africana y de la europea y esta mezcla se da en Estados Unidos, debido a la llegada de los esclavos negros a principios del siglo XVII, los salmos de tradición africana dieron lugar a lo que se conoce como gospel (canto religioso). La música religiosa convivía con la profana: canciones de plantación, baladas y otras formas de expresión popular, tanto africanas como europeas.

A mediados de la década de 1950, surgió un nuevo tipo de música conocido como rock. Proviene de una danza afroamericana denominada rhythm and blues, que a su vez fusiona el blues, el jazz y el gospel. Los representantes del rhythm and blues fueron Little Richard, Chuck Berry.

El pop se originó en su forma moderna a mediados del decenio de 1950 en los Estados Unidos y el Reino Unido. La música pop define a grupos con un estilo bien definido: Suelen tener los coros y ganchos repetidos, las canciones de corta a mediana duración escritas en un formato básico además de los ritmos o tempos que pueden bailarse fácilmente. Gran parte de la música pop también toma prestados elementos de otros estilos, como el rock, el urbano, el dance, el latino y el country.

En las décadas de 1950 y 1960, la música pop abarcaba el rock and roll y los estilos orientados a la juventud. Los términos siguieron siendo aproximadamente sinónimos hasta finales del decenio de 1960, posteriormente se asoció con la música efímera y accesible, es decir la que resulta ser más comercial.

Para el año de 1960 surgió el género soul, con raíces en el gospel y relacionado a los afroamericanos. Los representantes fueron Ray Charles y Aretha Franklin. En 1964 empieza una nueva era del rock, con The Beatles, que en esa época fue una las bandas más famosas y que su legado ha perdurado con

el paso de los años, ya que actualmente sigue siendo muy conocida.

Para la década de 1970 predominaba el rock pero surge otro género que causa gran impacto y es la música disco, siendo Donna Summers una de sus principales exponentes. Por su parte el house, derivó de la música disco, a su vez, derivado del funk y el soul. Tuvo sus orígenes en las discotecas de afroestadounidenses y latinoamericanos, en donde consiguió la popularidad. Tiene ritmos característicos tales como: ruidos mecánicos y voces sintetizadas. Además, surgen tres géneros populares: El reggae en Jamaica, el punk en el Reino Unido y el hip-hop en Estados Unidos.

### 3. Digitalización de la música

Se puede observar que la música tiene un gran impacto en la sociedad por ende no debería de sorprender el hecho de que la música fuera una fuente de ingresos, por lo que la industria hoy constituye uno de los sectores más importantes en la economía.

Es importante mencionar que la sociedad se encuentra en una era digital y todo tiende hacia el lado de la digitalización, además se aprovecha los beneficios que ofrecen las tecnologías digitales ya que resulta ser que los costes son menores en almacenamiento, distribución y comercialización y sobre todo la facilidad con la que se puede llegar al otro lado del mundo por parte de la digitalización para así abrir más mercados digitales.

Es sabido que la música pasó por una gran evolución respecto a la forma en la que se escuchaba grabada, desde discos de acetatos, cassettes, discos CD y de forma digital, por ende, siempre hay un estancamiento para mejorar, en este caso la industria de la música sufrió al comenzar a tener pérdidas debido al desuso del CD y servicios Peer tu Peer (P2P) es decir, de colega a colega, y son aquellos programas que permiten a los usuarios de Internet conectarse entre sí y compartir archivos que están en sus ordenadores. En esta era donde todo puede ser digital hay quienes, si tomaron ventaja como principal ejemplo fue que, en 2001, por la compañía Apple, creadora de iTunes, una distribuidora de música online que convertiría el uso del MP3 de acceso público al pagar cierta cuota. Steve Jobs "la mejor forma de detener la piratería –la única forma, de hecho– consistía en ofrecer una alternativa más atractiva que aquellos absurdos servicios que estaban preparando las discográficas"

<sup>2</sup>. iTunes en un principio ofreció los catálogos de las cinco mejores disqueras (BMG, Sony, Warner, Universal y EMI) en 2011, entró a Latinoamérica. Steve Job fue de los principales precursores para que la música comenzaría a comercializarse digitalmente.

En la actualidad la música se paga mediante el pago por producto y la financiación indirecta por medio de la publicidad. Se tiene un acceso a la oferta musical a cambio de un pago. A continuación, se mencionan las principales formas de pago que hay para acceder a la música:

- Método de Pago por Descarga: Se le conoce como pago por descarga y consiste en el pago directo por la descarga de todo el álbum o por canción.
- Método por Suscripción: Es a lo que se le conoce como suscripción a algún Streaming que consiste en el pago de una cuota mensual para acceder al servicio que brinda la plataforma digital de música.

Es importante mencionar que todo se debe ir actualizando y ver que los servicios que se consumían presencialmente se comienzan a digitalizar a gran medida, el primer ejemplo es que ha sido muy mencionado y es el caso de blockbuster que es un servicio en dónde ibas y rentabas películas pero que gracias a la digitalización este negocio se ha ido casi a la quiebra y esto porque ahora existen muchas plataformas de Streaming en las cuales se puede tener acceso a las películas mucho más fáciles. De este modo pasa con la música la mayoría de los consumidores prefiere no comprar el CD de manera presencial, debido a que es muy sencillo acceder desde el celular y se puede hacer la consulta de esa canción hasta de manera gratuita en plataformas como Youtube, Spotify, iTunes entre otros.

#### 4. Comercialización de la música en la actualidad

Existió un periodo para la industria musical que se conoció como una época de oro, esto fue en el periodo de 1982 a 1999, debido a que se obtuvieron ganancias de los dividendos de la comercialización de álbumes y sencillos. En específico las disqueras resultaron tener mayores ganancias.

La aparición de los quemadores en el 2000 vino a crear un parteaguas en las ganancias de las disqueras. Ya que con la llegada de los quemadores: que son unos

---

<sup>2</sup>Arango Archila.(2015). LA INDUSTRIA DISCOGRÁFICA Y LOS CONSUMIDORES: ¿La música como bien comercial o gratuito?

aparatos integrados a los computadores, que su funcionalidad es grabar discos compactos, y que para su época era costoso, pero posibilitaba el acceso a tener los CDs. Los quemadores mantenían la calidad original de las canciones, lo cual creaba una opción muy viable para obtener CDs sin pagar el original. Tan viable era que el comercio ilegal de CDs no originales se triplicó entre los 2000 y 2001, sobre todo en países tercera mundistas, ya que la adquisición de discos conocidos como piratas era de lo más normal. Pero esto abrió paso completamente a como se comercializaría la música en la actualidad.

Otro factor importante fue internet y lo fácil que era acceder a incontables consultas. En 1999 internet hizo que las canciones fueran de acceso gratuito; con la aparición de servicios Peer tu Peer (P2P) es decir, de colega a colega, y son aquellos programas que permiten a los usuarios de Internet conectarse entre sí y compartir archivos que están en sus ordenadores. El mayo representante de estos servicios fue Napster que abrió por completo el pasó al intercambio de música. Ya no era necesario tener todo el disco, ahora solo se podía obtener la canción deseada y lo mejor es que era gratis. Se comenzó a descargar música en las universidades, en donde se tenía banda ancha. Era evidente que la industria estaba teniendo pérdidas notables. Hasta que hubo quienes se quejaron; tal es el ejemplo de Lars Ulrich, baterista de la famosa banda metálica, comenzó a manifestar su descontento en contra de Napster y tras procesos legales se logró que se cerrará Napster.

Las disqueras buscaron formas de intentar entrar en la era digital para no seguir teniendo pérdidas, intentaron hacer plataformas P2P por su cuenta, pero ninguna funcionaba y estaban acompañadas de grandes fracasos. El momento en el que hubo un real y funcional cambio fue hasta que Steve Job fue de los principales precursores para que la música comenzaría a comercializarse digitalmente. Apple al crear iTunes y hacer que se pagará para acceder a la música y que también contenía a las disqueras más importantes en sus catálogos y la calidad de la música era por mucho mejor comparada a la que se compartía gratuitamente resulto ser muy atractivo y comenzó a ganar terreno en el mercado. En iTunes se podía escuchar 30 segundos de cada canción, antes de comprarla, lo que de alguna manera resultaba conveniente para algunos usuarios que podían pagar estos servicios.

Aún con la existencia de las soluciones era evidente que no todos tenían los recursos para pagar estos servicios por lo que el aumento de la piratería llegó a niveles alarmantes, de suma preocupación para las compañías discográficas. En el año 2008 se presentó uno de los mayores años en los que se vendió más piratería y esto debido a que los teléfonos inteligentes iban cada vez teniendo más avances tales como; el infra rojo y el Bluetooth que permitía que los teléfonos inteligentes compartieran la

música descargada.

Por otro lado, bandas tales como; Nine Inch Nails o Radiohead subieron sus canciones a un sitio web permitiendo que los usuarios tuvieran acceso de manera gratuita sus discos, álbumes, sencillos etc.

Se sabe que siempre hay intereses comerciales y económicos de por medio y en Estados Unidos se creó la ley SOPA (Stop Online Piracy Act) y PIPA (Protect IP Act), en 2012 La Stop Online Piracy Act (Acta de Cese a la Piratería En Línea) también conocida como Ley SOPA con la finalidad expandir las capacidades de la ley estadounidense para combatir el tráfico de contenidos con derechos de autor y bienes falsificados a través de Internet, aunque esta ley fue muy sonada y para muchos de preocuparse, violaba por muchas razones colaterales la privacidad de los usuarios, Barack Obama como el congreso, rechazaron ambas propuestas por los términos tan polémicos como fueron redactadas.

Con esto es claro que las disqueras fueran cada vez mas en decadencia. Pero abrió otro método de comercialización que mostraría grandes ganancias en el presente; Spotify, YouTube y iTunes son de los principales negocios que obtienen ganancias considerables de esta forma de comercializar la música a través de medios digitales.

En la siguiente grafica ver Figura 1 Se muestra mediante una gráfica de pastel la importancia que tiene streaming y suscripción en cómo se comercializa actualmente la industria musical. Se habla que casi la mitad de los ingresos provienen de las plataformas de Streaming. Además, “Las descargas digitales de música registradas durante ese año supusieron alrededor de 6 % de los ingresos totales de esta industria, mientras que el streaming, incluyendo aquel financiado por publicidad, representó más del 62 % del total.”<sup>3</sup>, dato que da mucha relevancia porque se ve la importancia de como ponderan estas plataformas de Streaming en la actualidad, lo que lleva a que en el año 2020 el mayor consumo de música es por vía de las plataformas digitales de streaming. Es importante mencionar que a pesar de que se vive en una era digital aún el ingreso por compra física es del 20 % con respecto del total.

---

<sup>3</sup>Orús Abigail. (2021) Segmentos de la música grabada mundial por participación en los ingresos 2020

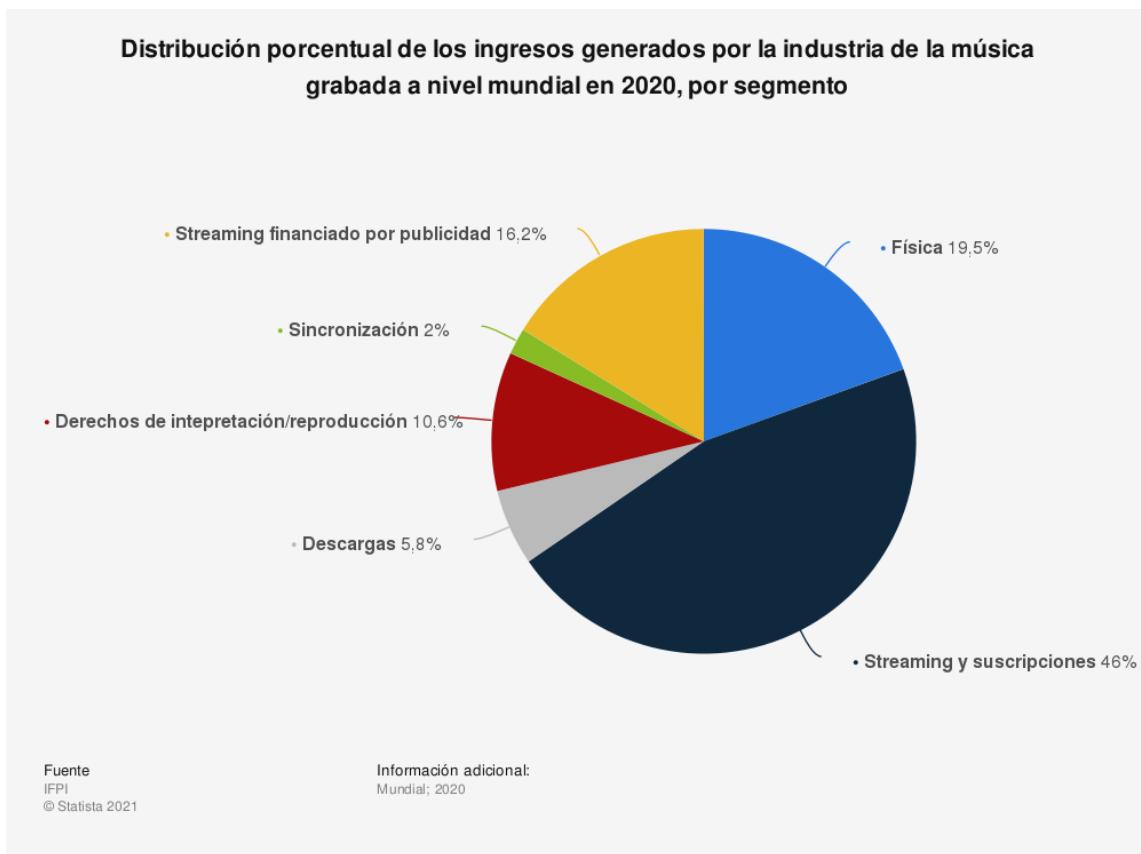


FIGURA 1. Distribución porcentual de los ingresos generados por la industria de la música grabada a nivel mundial en 2020, por segmento (Statista)

En dónde se entiende lo siguiente:

- Streaming financiado por publicidad: Es aquella plataforma la cual se tiene acceso a la música a cambio de consumir publicidad.
- Sincronización: Consiste en utilizar una obra musical sin modificarla para incluirla en otras obras como pueden ser obras audiovisuales tales como, películas, spots, anuncios publicitarios, documentales, cortos o algunas de las redes sociales como youtube, facebook entre otras.
- Derechos de interpretación / reproducción: Es decir el pago que se realiza a los derechos de autor a través de la interpretación o reproducción de la pireza musical.

- Descargas: Música descargada mediante las páginas oficiales de los artistas, es decir se puede pagar por descargar el disco o la canción. La obtención de éstas es por el medio digital.
- Física: Éste ingreso se refiere a los discos comprados de manera física que se adquiere en la mayoría de los casos por un CD.
- Streaming y suscripción: Plataforma a la cual se puede suscribir dando un pago para recibir a cambio acceso a bibliotecas digitales de música sin necesidad de escuchar publicidad.

**4.1. Spotify.** “Spotify es un servicio de música, podcasts y videos digitales en streaming que te da acceso a millones de canciones y otros contenidos de artistas de todo el mundo.

Las funciones básicas, como escuchar música, son totalmente gratis, pero también tienes la opción de mejorar tu cuenta con Spotify Premium. De cualquiera de las dos maneras, puedes:

- Elegir lo que quieres escuchar con Explorar y Buscar.
- Recibir recomendaciones en funciones personalizadas, como Descubrimiento semanal.
- Montar colecciones de música.
- Ver lo que escuchan amigos, artistas y famosos.
- Crear tus propias emisoras de radio.

Spotify está disponible en diversos dispositivos, como ordenador, teléfono, tablet, altavoces, televisores o coches, y puedes pasar fácilmente de uno a otro con Spotify Connect.”<sup>4</sup>.

Spotify es la plataforma de Streaming musical más importante a nivel mundial. Se lanzó por primera vez en el 2008 en Estocolmo, Suecia, pero con el tiempo ha ido creciendo para ser de gran relevancia dentro del mercado global. A México llegó en el año 2013. Es importante mencionar Spotify ya que es de las principales formas en los que en la actualidad la música se comercializa.

Spotify se ha convertido en la plataforma de streaming musical más importante, debido a que cuenta con millones de canciones, se puede escuchar música gratis, a cambio de pausas publicitarias y restricciones para saltar canciones. Sin embargo, si eres usuario premium tienes a tu disposición poder escuchar la música sin interrupción alguna y además se puede descargar para poder escucharla sin necesidad

---

<sup>4</sup><https://support.spotify.com/es/article/what-is-spotify/>

de estar conectado a una red de internet o usar datos móviles.

En la siguiente gráfica ver Figura 2 se muestra la cantidad de usuarios activos de Spotify, hay 345 millones de usuarios activos, de los cuales 155 millones son usuarios premium, es decir pagan cierta cuota para obtener los servicios de Spotify, siendo así uno de los principales generadores de ingresos en el sector musical.

Por otro lado, Spotify resulta ser una gran fuente de información, y esto porque puede saber cuántos usuarios escuchan cuales canciones, álbumes, artistas, playlists. A través, de la base de datos de Spotify se puede saber cuáles son los artistas o canciones más populares del momento, siendo así una gran fuente de información para posibles estudios.

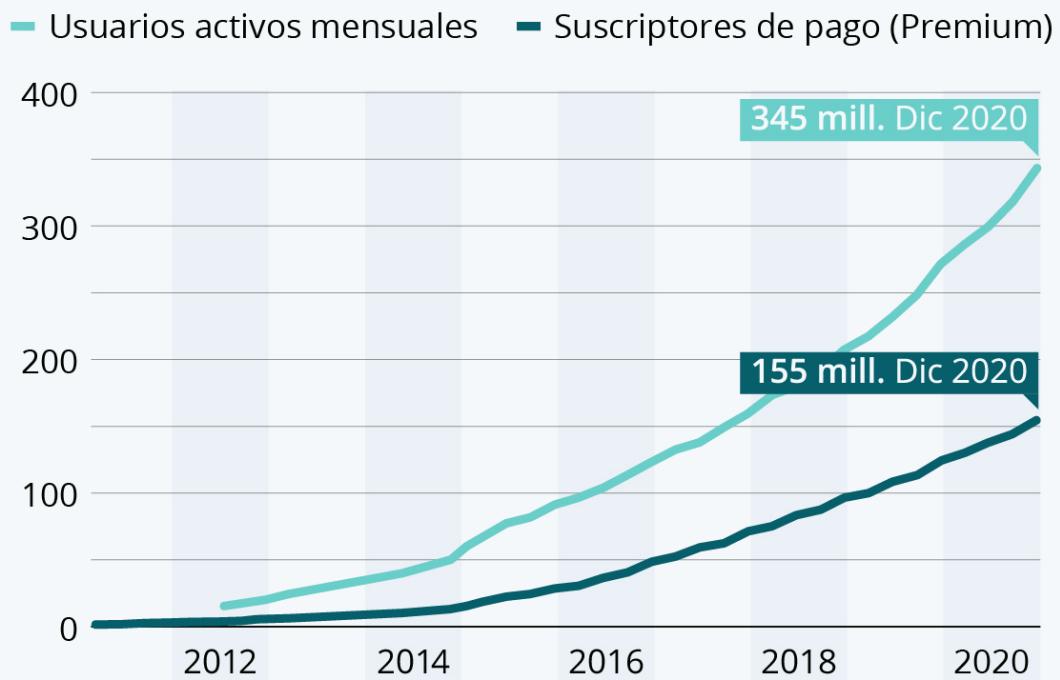
Este acceso a la información de los usuarios, Spotify también ha aprovechado para generar algoritmos en beneficio del usuario porque su algoritmo es característico por recomendar canciones que muy probablemente serán de gusto del usuario lo que cumple con la finalidad de que te quedarás ahí escuchando música en la plataforma porque básicamente mantiene al usuario satisfecho. Con esto surge algo que es importante para fines de esta investigación, y es como se manipula la música en la actualidad para que pueda comercializarse de manera más efectiva. Y esto porque los primeros segundos de la canción son de suma importancia porque si el usuario salta rápidamente la canción el algoritmo no te va a recomendar de ese estilo de canciones. Por lo que se mencionaba en un video de YouTube del canal Alvinsch con 1.4 millones de seguidores, el video cuenta con 905,870 vistas<sup>5</sup>, menciona que la música ahora tiene menor duración y no solo eso que el fenómeno de saltar las canciones apresuradamente hace que las bandas, músicos y creadores de música ahora lancen su coro al principio para así llamar la atención de los usuarios y se queden consumiendo esas canciones, o bien buscar al artista. Además, menciona que la estructura actual de la música más comercial es primero el coro seguido del verso, nuevamente coro y por último verso. Asimismo, que en promedio las canciones de ahora duran alrededor de 3.30 minutos, garantizando así llamar la atención del usuario y que sea fácil de escuchar dando como resultado que el usuario se quede consumiendo cada vez más.

---

<sup>5</sup>Alvinsch. (2020). Así te manipulan para que te guste su música

## Spotify sigue creciendo

Usuarios activos mensuales y suscriptores premium de Spotify en el mundo



Fuente: Spotify



**statista**

FIGURA 2. Cantidad de usuarios activos en Spotify. Mónica Mena Roa. (2021) Spotify alcanza los 155 millones de suscriptores de pago.



## Capítulo 2

# Modelos estadísticos

En este capítulo se muestran los distintos modelos estadísticos que se usan en este trabajo, para predecir la popularidad de una canción. La intención del capítulo es exponer la heurística y los algoritmos de estos modelos predictivos. Lo que se busca a través de estos modelos es un mecanismo que predice el comportamiento de la popularidad de una canción, utiliza las diversas variables de la base de datos como entrada y proporciona un valor predictivo como salida. En éste capítulo se observa como es que funciona el algoritmo de cada modelo, usando el lenguaje de programación R, para percibir cómo es que funciona el modelo computacionalmente, esto aplicada en la base de datos Iris.

### 1. La base de datos Iris

Para la descripción de los algoritmos usados para los modelos que se describen en este capítulo, se usa la base de datos Iris disponible en la librería {datasets} de R. Esta es una base de datos presentada por Ronald Fisher en su artículo publicado en 1936 , “*The use of multiple measurements in taxonomic problems*” en español (El uso de medidas múltiples en problemas taxonómicos). En el artículo se menciona que las especies de iris setosa e iris versicolor crecieron en la misma colonia y fueron medidas por el Dr. E. Anderson. Es importante mencionar que a la fecha esta es una de las bases de datos más utilizadas para la aplicación de modelos de clasificación, debido a su versatilidad para poder diferenciar con base en sus medidas a qué tipo de especie pertenece.

La base se encuentra compuesta por 50 observaciones de cada especie de iris; iris setosa, irirs versicolor e iris virginica, que tienen diferencias morfológicas significativas, estas diferencias se observan en Figura 1.



FIGURA 1. Especies de Iris

En total hay 150 observaciones, las variables o atributos que se miden de cada flor son las siguientes:

- El tipo de especie de la flor iris, como variable categórica, cuyas categorías son: Virginica, Versicolor y Setosa.
- El largo y el ancho del pétalo en centímetros como variables numéricas.
- El largo y el ancho del sépalo en centímetros como variables numéricas.

## 2. Técnicas de aprendizaje (supervisado y no supervisado)

En la ciencia de datos se cuenta principalmente con dos tipos de técnicas: el aprendizaje no supervisado y el aprendizaje supervisado, “El aprendizaje no supervisado encuentra patrones ocultos o estructuras intrínsecas en los datos. Se utiliza para extraer inferencias de conjuntos de datos que consisten en datos de entrada sin respuestas etiquetadas. La segmentación es la técnica de aprendizaje no supervisada más común. Se utiliza para el análisis exploratorio de datos para encontrar patrones ocultos o agrupaciones en los datos. Las aplicaciones para la agrupación incluyen análisis de secuencia de genes, investigación de mercado y reconocimiento de objetos... El aprendizaje supervisado entrena un modelo sobre datos de entrada y salida conocidos para que pueda predecir resultados futuros.”<sup>1</sup>

¿Cómo funciona el aprendizaje supervisado? Pérez César (2021) también señala que un algoritmo de aprendizaje supervisado toma un conjunto conocido de datos de entrada y respuestas conocidas a los datos (salida) y entrena un modelo para generar

---

<sup>1</sup>Pérez López, César. (2021) Machine learning. Técnicas de aprendizaje supervisado a través de R

predicciones razonables para la respuesta a nuevos datos. El aprendizaje supervisado utiliza técnicas de clasificación y regresión para desarrollar modelos predictivos, por ejemplo, si un correo electrónico es genuino o spam (clasificación) y cambios en la temperatura (regresión).

- Aprendizaje supervisado por clasificación: Predice respuestas categóricas, es decir se tienen distintas variables en cierto conjunto de datos, lo que se espera con estas variables (etiquetas) es que se pueda obtener una categoría con base en sus etiquetas. El algoritmo clasifica, es decir, da una etiqueta entre dos o más clases, esta etiqueta hace referencia al valor categórico que se asigna.
- Aprendizaje supervisado por regresión: Predice respuestas continuas, con base en los datos de entrenamiento el algoritmo predice a partir de un rango de valores.

### 3. Análisis discriminante lineal (LDA)

“El artículo de 1936 de Ronald A. Fisher, «The Use of Multiple Measurements in Taxonomic Problems» (Annals of Eugenics, 7: 179-188), es el comúnmente referenciado como primera aplicación de AD, hasta el punto de que el análisis discriminante lineal (ADL) a veces se denomina «análisis discriminante lineal de Fisher». No obstante, el AD que describe este artículo no cumple todos los supuestos que demanda la adecuada realización de ADL... el artículo sí introdujo el término discriminación y dio forma a la idea de combinación lineal de variables independientes para la diferenciación de grupos. Fisher lo aplicó al esclarecimiento de taxonomías tradicionales en el área de la biología y la antropología física. En su desarrollo influyeron propuestas anteriores de medidas de distancias entre grupos.”<sup>2</sup> Con el paso del tiempo se hicieron diferentes supuestos y adecuaciones al análisis discriminante lineal de Fisher, para poder llegar a lo que hoy se conoce como: El Análisis Discriminante Lineal (LDA) o *Linear Discriminant Analysis* que es una técnica de aprendizaje supervisado, que consiste en la descripción, predicción y clasificación.

Es importante mencionar que el análisis discriminante hace una selección de variables en donde se destaca el término de variable respuesta o dependiente, que se conoce como una variable de tipo categórica, por otro lado, las variables que darán la discriminación son las variables explicativas o predictoras.

---

<sup>2</sup>De Cea D' Ancona, María Ángeles. (2016) Cuadernos Metodológicos; 54 Análisis discriminante

El LDA se define como un método de clasificación supervisada que usa las variables cuantitativas para clasificar las observaciones.

Este método hace uso de uno de los teoremas más importantes y más usados en la probabilidad que es el teorema de Bayes. LDA hace una estimación de la probabilidad de que cada observación, dado un valor de los predictores, pertenezca a cada una de las clases de la variable, esto es, que  $P(G = k|X = x)$ . Por último, se destina la observación a la clase  $k$  para la probabilidad que se predice es mayor.

(Hastie Trevor *et al*, 2008, pp.106- 110). Mencionan y desarollan lo siguiente: la teoría de decisión para clasificación dice que se necesitan conocer las probabilidades a posteriori de clase  $P(G|X)$  para una clasificación óptima. Supongamos que  $f_k(x)$  es la densidad condicional de clase  $G = k$  para  $X$ , y  $\pi_k$  es la probabilidad previa de la clase  $k$ , con  $\sum_{k=1}^K \pi_k = 1$ . Una aplicación simple del Teorema de Bayes.

$$P(G = k|X = x) = \frac{P(X = x|G = k) \cdot P(G = k)}{P(X = x)}$$

Donde:

$P(G = k|X = x)$  es la probabilidad a posteriori de que la observación  $x$  pertenezca a la clase  $k$ ,

$P(X = x|G = k)$  es la probabilidad condicional de la observación  $x$  dada la clase  $k$ ,

$P(G = k)$  es la probabilidad previa de la clase  $k$ ,

$P(X = x)$  es la probabilidad marginal de la observación  $x$ .

Vemos que en términos de capacidad para clasificar, tener  $f_k(x)$  es casi equivalente a tener la cantidad  $P_r(G = k|X = x)$ .

Muchas técnicas se basan en modelos para las densidades de clase:

- El análisis discriminante lineal y cuadrático utiliza densidades gaussianas.
- Las mezclas más flexibles gaussianas permiten fronteras de decisión no lineales.
- Las estimaciones generales no paramétricas para cada densidad de clase permiten la mayor flexibilidad
- Los modelos de Bayes ingenuos son una variante del caso anterior y asumen que cada una de las densidades de clase son productos de densidades marginales; es decir, asumen que las entradas son condicionalmente independientes en cada clase.

Se supone que se modela cada densidad de clase como gaussiana multivariante:

$$f_k(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}$$

El análisis discriminante lineal (LDA) surge en el caso especial cuando se asume que las clases tienen una matriz de covarianza común  $\Sigma_k = \Sigma \forall k$ . Al comparar dos clases  $k$  y  $l$ , es suficiente observar la relación logarítmica, y vemos que se tiene una ecuación lineal en  $x$

$$\begin{aligned} \log \frac{\Pr(G = k | X = x)}{\Pr(G = l | X = x)} &= \log \frac{f_k(x)}{f_l(x)} + \log \frac{\pi_k}{\pi_l} \\ &= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l) + x^T \Sigma^{-1} (\mu_k - \mu_l) \end{aligned}$$

Las matrices de covarianza iguales hacen que los factores de normalización se cancelen, así como la parte cuadrática en los exponentes. Esta función lineal de *log-odds* implica que la frontera de decisión entre las clases  $k$  y  $l$ —el conjunto donde  $\Pr(G = k | X = x) = \Pr(G = l | X = x)$  es lineal en  $x$ ; en  $p$  dimensiones, un hiperplano. Esto es cierto para cualquier par de clases, por lo que todas las fronteras de decisión son lineales. Si dividimos  $\mathbb{R}^p$  en regiones que se clasifican como clase 1, clase 2, etc., estas regiones estarán separadas por hiperplanos. Vemos que las funciones discriminantes lineales:

$$\delta_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k$$

son una descripción equivalente de la regla de decisión, con  $G(x) = \text{argmax}_k \delta_k(x)$ .

Con más de dos clases, el LDA no es lo mismo que la regresión lineal de la matriz de indicadores de clase, y evita los problemas de enmascaramiento asociados con ese enfoque (Hastie et al., 1994).

Por otro lado, es importante mencionar que estas definiciones del modelo solo se usan como apoyo, ya que se usarán las herramientas de programación para hacer los cálculos pertinentes. El proceso del análisis discriminante puede resumirse en cuatro pasos, para fines de mejor entendimiento, a continuación se ejemplificará cada uno de los cuatro pasos con la ya mencionada base de datos Iris.

**3.1. Paso 1.** En este paso se lleva a cabo la división de la muestra, entre los conjuntos de entrenamiento (train) y prueba (test). ¿Por qué se debería dividir esta muestra?. Se supone que a mayor número de observaciones, mas precisas serán las estimaciones. Con esto se podría suponer que la mejor opción para determinar

los parámetros del modelo es usar todas las observaciones, pero esta no es la mejor opción, debido a que, si se usa todo el conjunto de datos sin hacer una división, para determinar los parámetros del modelo, se estaría haciendo un sobreajuste. De esta manera es importante hacer la división de la muestra. Empíricamente se ha mostrado que las mejores opciones para obtener los mejores resultados son; si se usa una división de los datos en la siguiente proporción: 20 % - 30 % de los datos para el conjunto de test y el restante 80 % - 70 % , respectivamente, para el conjunto de train. Esta proporción se ha usado de manera empírica en el ambiente de la ciencia de datos. La división dependerá del analista.

Para el entrenamiento de un modelo es necesario determinar parámetros, a partir de las observaciones que sean conocidas, esto es el conjunto de train, datos seleccionados al azar y el resto de la muestra que es elconjunto de test.

En este caso se decide usar la división 70 % 30 %. Se seleccionan aleatoriamente el 70 % de los renglones del conjunto de datos, con muestreo sin reemplazo, es decir, las unidades se extraen de la base de datos y en cualquier extracción que se haga, esta no se devuelve a la base de datos, cada unidad extraída de la base de datos no puede seleccionarse más de una vez.

A continuación, se muestran unas matrices de dispersión, estas herramientas se usan para visualizar la relación entre las variables. Representa pares de datos numéricos, con una variable en cada eje. Además, se busca visualizar si las especies están relacionadas entre sí, a través de su ancho y largo del pétalo y sépalo. Se definen las siguientes variables: *Sepal.Length*, como largo del sépalo, *Sepal.Width*, como ancho del sépalo, *Petal.Length*, como largo del pétalo y finalmente *Petal.Width* como ancho del pétalo.

Se busca hacer una previsualización de la base de datos Iris, Ver figura 2 en la cual se observa cómo es que las variables separan la muestra entre especies y como es que se comportan. Se puede analizar que las variables Petal.Length y Petal.Width tienen correlaciones positivas. También, Petal.Length y Petal.Width son las dos variables en donde se observa mejor la separación de clases, es decir, estas variables hacen una mejor distinción entre las especies; setosa, versicolor y virginica.

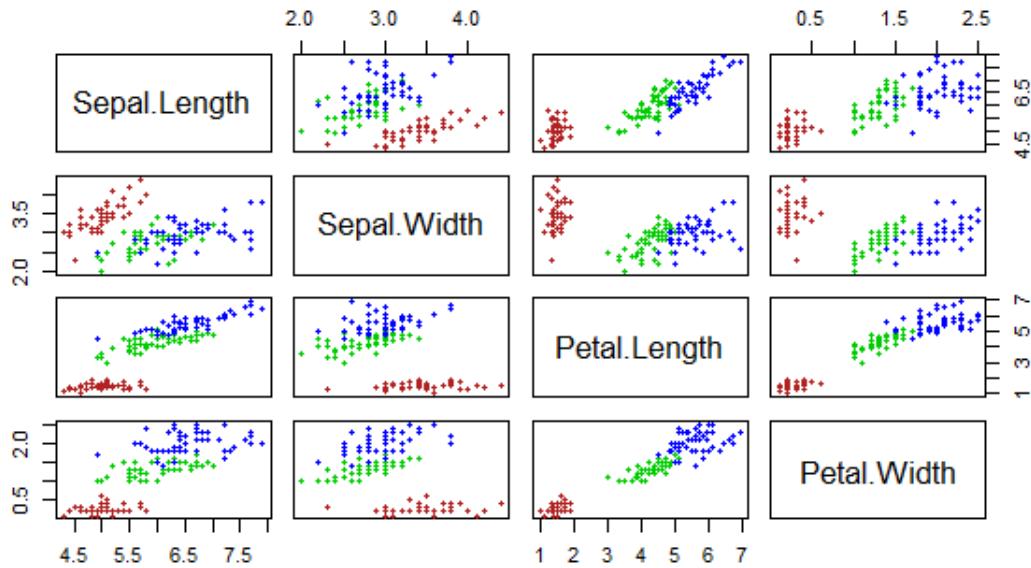


FIGURA 2. Gráficas de previsualización de los datos Iris. Setosa color rojo, versicolor color verde y virginica azul

Se muestran las siguientes matrices de dispersión, en donde el objetivo es observar que los datos de entrenamiento y prueba se comportan consistentemente con la base de datos de Iris. En ambos gráficos Ver figura 3 Ver figura 4 parece existir una consistencia con los datos de la base Iris, ya que la correlación es positiva con las variables Petal.Length y Petal.Width al igual que lo es para Iris, también las variables, Petal.Length y Petal.Width son las dos variables en donde se observa mejor la separación de clases, esto refleja que estas variables tienen mayor fuerza para separar entre especies.

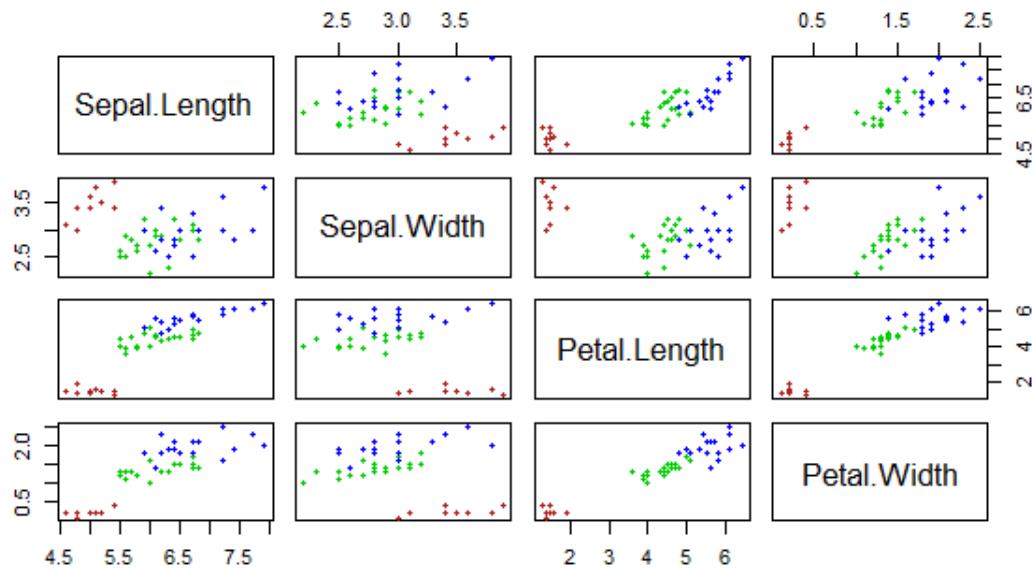


FIGURA 3. Gráficas de previsualización de los datos test. Setosa color rojo, versicolor color verde y virginica azul

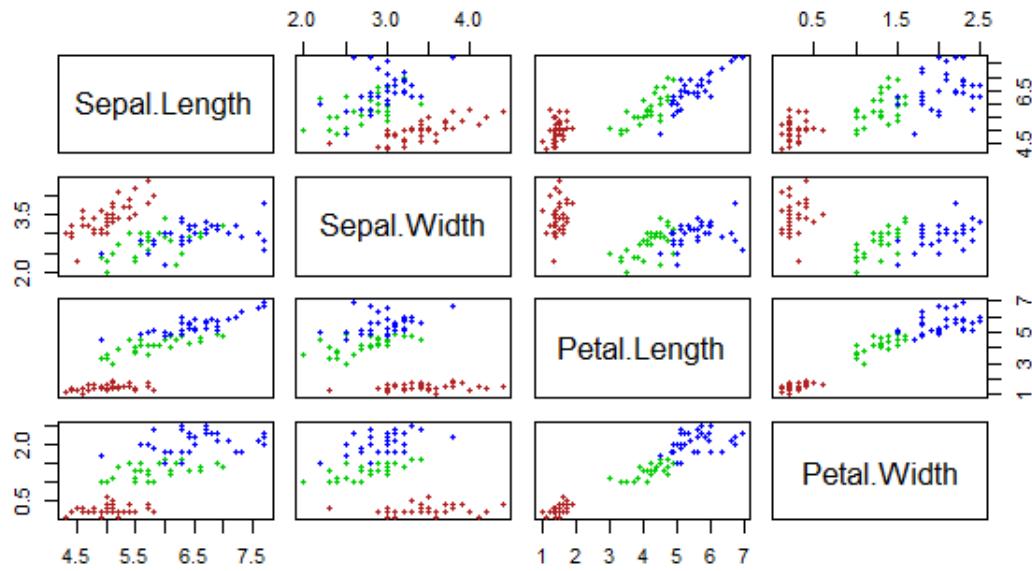


FIGURA 4. Gráficas de previsualización de los datos train. Setosa color rojo, versicolor color verde y virginica azul

**3.2. Paso 2.** En el paso 2 se prosigue con el cálculo de las probabilidades previas ( $\pi_k$ ): Es decir, esto es la proporción esperada de observaciones que pertenecen a cada grupo, de esta manera la probabilidad previa es la probabilidad de que una observación aleatoria pertenezca a la clase  $k$ . De este modo la probabilidad previa ( $\pi_k$ ) se hace la estimación como se sigue. La probabilidad de que una observación cualquiera sea de la clase  $k$  sea igual al número de observaciones de esa clase dividido por el total de observaciones que existen  $\pi_k = \frac{n_k}{N}$ . Con esta información, se tiene que para iris, la probabilidad previa de cada especie es, el número de observaciones de la especie, dividido por el número de observaciones totales que existen. Se tienen 3 especies (setosa, versicolor, virginica) y cada una de estas especies cuenta con 50 observaciones por lo que el número de observaciones totales es 150. Es decir, se hace la división de 50 entre 150 dando por resultado una probabilidad previa de igual a 0.33.  $\pi_{\text{setosa}} = \pi_{\text{versicolor}} = \pi_{\text{virginica}} = \frac{50}{150} = 0.33$

**3.3. Paso 3.** La metodología del LDA sugiere normalidad univariante. En éste gráfico (ver figura 5) se observa que la muestra se comporta de forma normal. Se graficó para cada variable de la base de datos Iris.

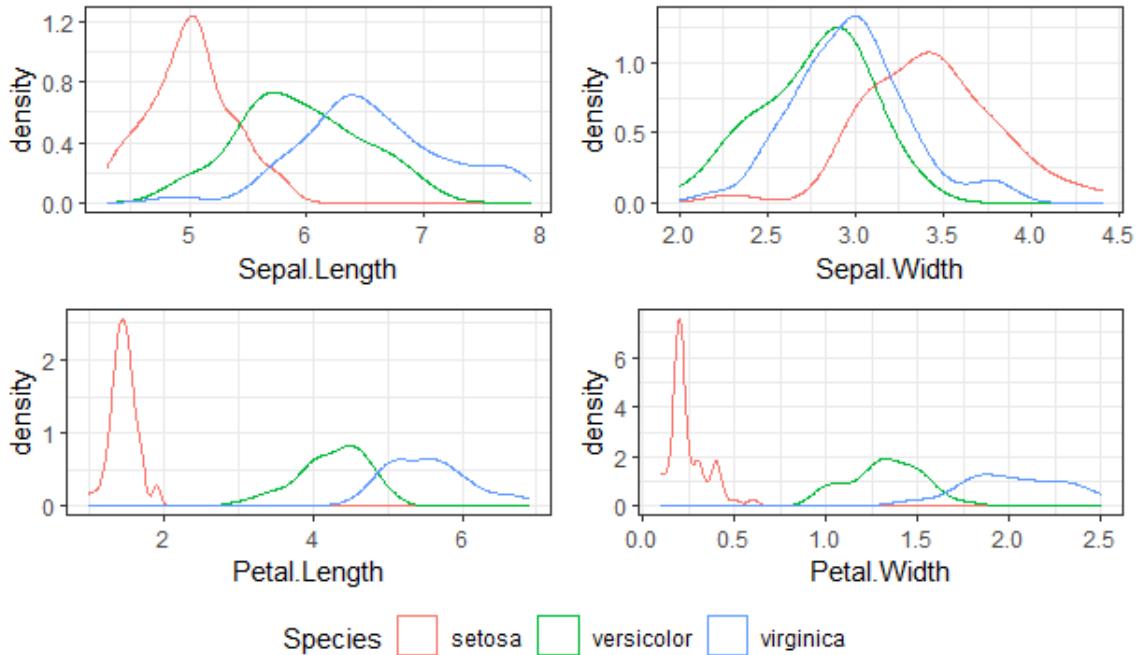


FIGURA 5. Gráficos de densidad

Ahora, se utiliza el test de normalidad de Shapiro-wilk para ver como se comporta cada variable para las tres especies en la base de datos Iris. Esto solo se usa como un recurso del lenguaje de programación interpretado llamado R, se usa la siguiente librería, se usan las siguientes librerías: library (knitr), library (dplyr) y library (reshape2) en la cuál se encuentra la función Shapiro.test que calcula el valor del test Shapiro-Wilk para cada grupo.

Se tiene que la variable P value\_Shapiro.test; representa el valor p obtenido del test. P se utiliza para evaluar la hipótesis nula de que los datos siguen una distribución normal. En dónde un valor pequeño ( $p\text{-value} < 0.05$ ) sugiere que los datos no siguen una distribución normal.

Species	Variable	P value_Shapiro.test
setosa	Sepal.Length	0.45951
setosa	Sepal.Width	0.27153
setosa	Petal.Length	0.05481
setosa	Petal.Width	0.00000
versicolor	Sepal.Length	0.46474
versicolor	Sepal.Width	0.33800
versicolor	Petal.Length	0.15848
versicolor	Petal.Width	0.02728
virginica	Sepal.Length	0.25831
virginica	Sepal.Width	0.18090
virginica	Petal.Length	0.10978
virginica	Petal.Width	0.08695

TABLA 1. Resultados del test de Shapiro-Wilk para la normalidad.

Como se muestra en la Tabla 1, los resultados indican para Sepal.Length y Sepal.Width en todas las especies, así como para Petal.Length en setosa y versicolor, los valores p son mayores que 0.05. En estos casos, no hay suficiente evidencia para rechazar la hipótesis nula de normalidad, y se puede considerar que los datos siguen una distribución normal.

Pero para el caso de Petal.Length en virginica y todas las variables en setosa y versicolor, los valores p son menores que 0.05. En estos casos, se puede rechazar la hipótesis nula y concluir que los datos no siguen una distribución normal.

Ahora para verificar si la base de datos iris sigue una distribución normal multivariante se usa Henze-Zirkler (hz). Esto solo se usa como un recurso del lenguaje de programación interpretado llamado R, se usa la siguiente librería: library (MVN) que con la función multivariateNormality se obtiene la información de la normalidad multivariante, el valor p y el estadístico de prueba.

Test	HZ	p value	MVN
Henze-Zirkler	2.336394	0	

TABLA 2. Resultados del test de normalidad multivariante de Henze-Zirkler.

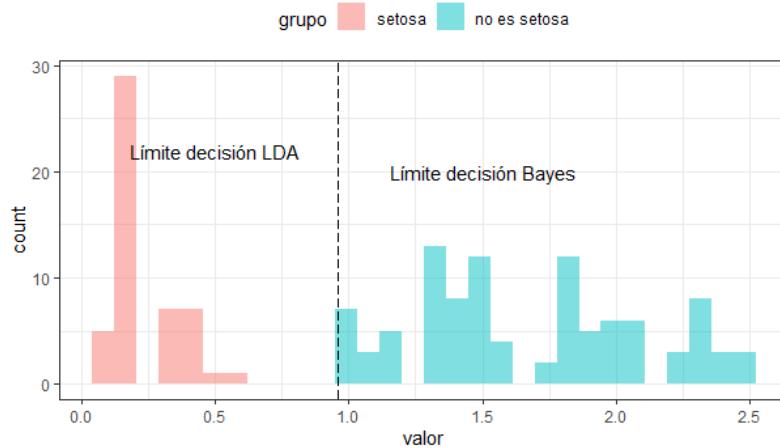
Los resultados de la prueba de normalidad multivariante, indican que si hay una p-value igual a cero generalmente significa que hay evidencia significativa para rechazar la hipótesis nula de que los datos siguen una distribución multivariante normal.

Como el resultado es 0, entonces no siguen una distribución multivariante normal a un nivel de significancia convencional ( $p\text{-value} < 0.05$ ). Por lo que se concluye que bajo el test de Henze-Zirkler, los datos de iris, no siguen una distribución multivariante normal.

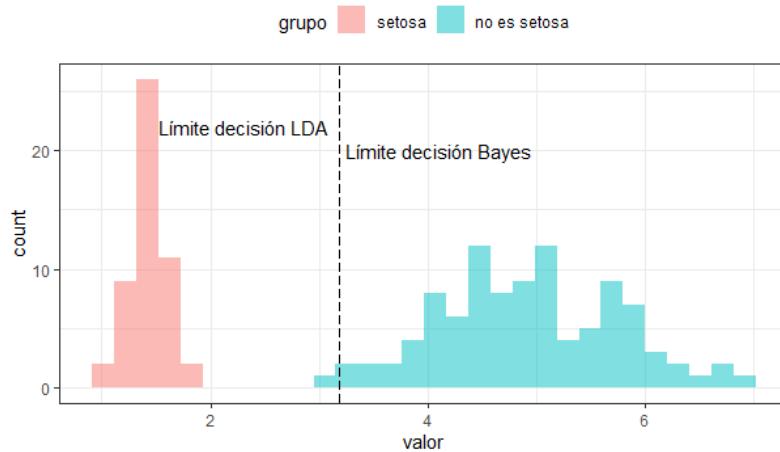
Se tiene que el LDA supone que las distribuciones condicionales de las variables predictoras, dadas las clases, son normales multivariantes. Esto significa que para cada clase, la combinación lineal de las variables predictoras sigue una distribución normal multivariante. Además, la normalidad multivariante es una suposición importante en LDA debido a que usa la información de covarianza entre las variables predictoras para poder construir las funciones discriminantes lineales.

Como se observa en ambas pruebas hay falta de normalidad pero se relaja un poco el supuesto suponiendo que la matriz de covarianza es constante y que el LDA puede obtener una buena precisión para la clasificación.

**3.4. Paso 4.** Aplicar la técnica del LDA, en dónde el resultado determina a qué grupo se asigna cada observación. En la gráfica (ver figura 2) es notorio que las especies versicolor y virginica tienen más parecido entre ellas, mientras que, la especie setosa se encuentra separada de estas dos. La muestra se puede clasificar mejor en dos grupos (setosa y no setosa), es decir, cada grupo clasifica de mejor manera si pertenece a la clase setosa o no, aunado a que, se observa que las variables Petal.Length y Petal.Width separan potencialmente bien las especies, por lo



(A) Separación de cada especie mediante la variable Petal.Width



(B) Separación de cada especie mediante la variable Petal.length

FIGURA 6. Visualización de la clasificación de dos grupos

que se usan estas dos variables para ilustrar la separación de clases a través del LDA.

Se muestran dos gráficas (ver figura 6). En dónde se aplica la separación de clases para ver cómo es que el LDA podría clasificar potencialmente. Por consiguiente, se decide que, para fines de entendimiento, la clasificación será para dos clases, es decir  $k = 2$ , con  $k_1 = \text{setosa}$  y  $k_2 = \text{no es de la especie setosa}$ .

Ahora, en la siguiente gráfica (ver figura 7), se hace uso del lenguaje de programación interpretado llamado R, se usa la librería (MASS) que contiene la función lda, con esta se hace una predicción de las clases, creando una columna llamada predicción\_lda y finalmente se crea un gráfico de dispersión en donde se puede apreciar las observaciones en función de las variables Petal.Length y Petal.Width, en dónde la predicción realizada por el modelo LDA se separa por color.

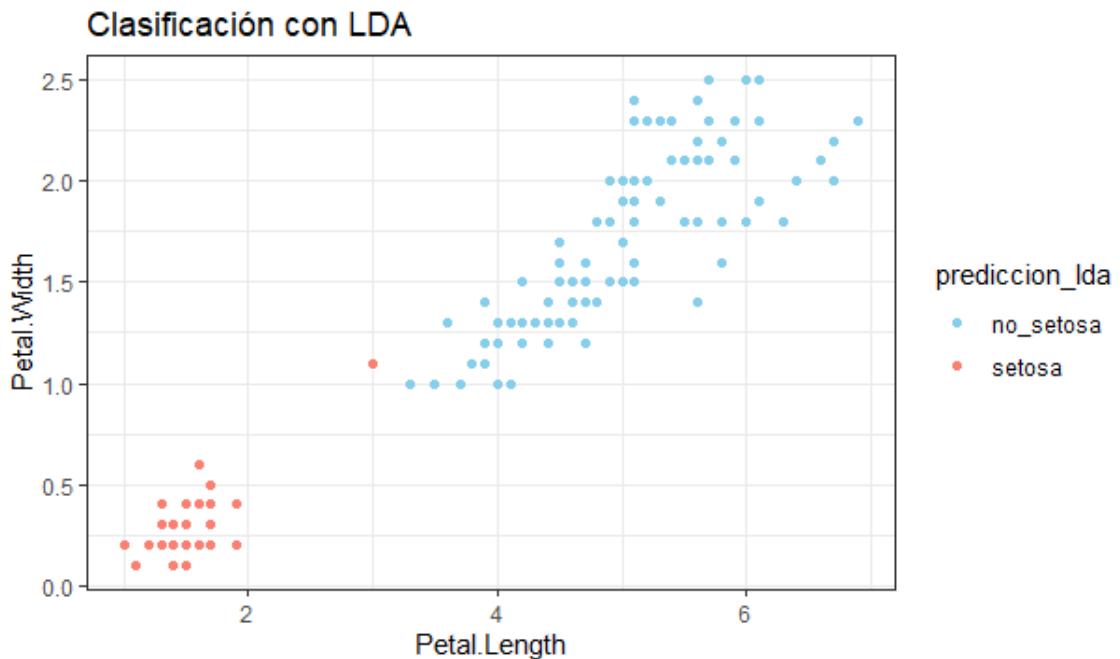


FIGURA 7. Clasificación con LDA para las variables Petal.Length y Petal.Width

Es importante, que gráficamente se observa como es que el algoritmo LDA, sí logra hacer una clasificación de la muestra, para ver a que especie pertenece, mediante dos variables, por ende, se logra hacer una clasificación adecuada.

También se puede ilustrar gráficamente que se puede suponer una distribución multivariada para dos variables predictoras Petal.Length y Petal.Width, es decir, se quiere saber cómo se distribuyen y cómo están relacionadas simultáneamente estas dos variables en las flores de Iris. Primeramente se calculan los parámetros estadísticos: Media, desviación estándar, la covarianza y coeficiente de correlación de las

variables Petal.Length y Petal.Width. Se define una función multivariante, que representa la densidad de probabilidad bivariante normal la función de densidad de probabilidad de una distribución normal bivariante, posteriormente se generan los valores de densidad para poder crear un gráfico en tercera dimensión en donde es posible ver la superficie de densidad multivariante en función de Petal.Width y Petal.Length. Todos estos cálculos, se hacen en el programa y solo se tiene el fin de observar graficamente la distribución.

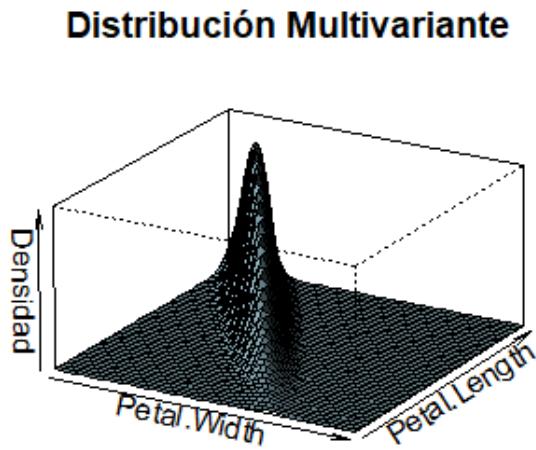


FIGURA 8. Extensión multivariante de para dos variables: Petal.Length y Petal.Width

En el siguiente gráfico (ver figura 9), se busca la visualización multivariante de las variables Petal.Length y Petal.Width para la clasificación de los grupos, setosa y no setosa. La gráfica en tercera dimensión, proporciona una representación visual de las densidades multivariantes para cada grupo. Primero, se separa la base de iris en los dos grupos, después, se calculan los parámetros estadísticos: Media, desviación estándar, la covarianza y coeficiente de correlación de las variables Petal.Length y Petal.Width. Se define una función multivariante que depende de la distribución bivariante normal para cada grupo, también se calculan los valores de densidad para crear un gráfico en tercera dimensión en donde se observa la clasificación de estos

dos grupos. Estos cálculos, se hacen en el programa y solo se tiene el fin de observar graficamente la distribución de la clasificación.

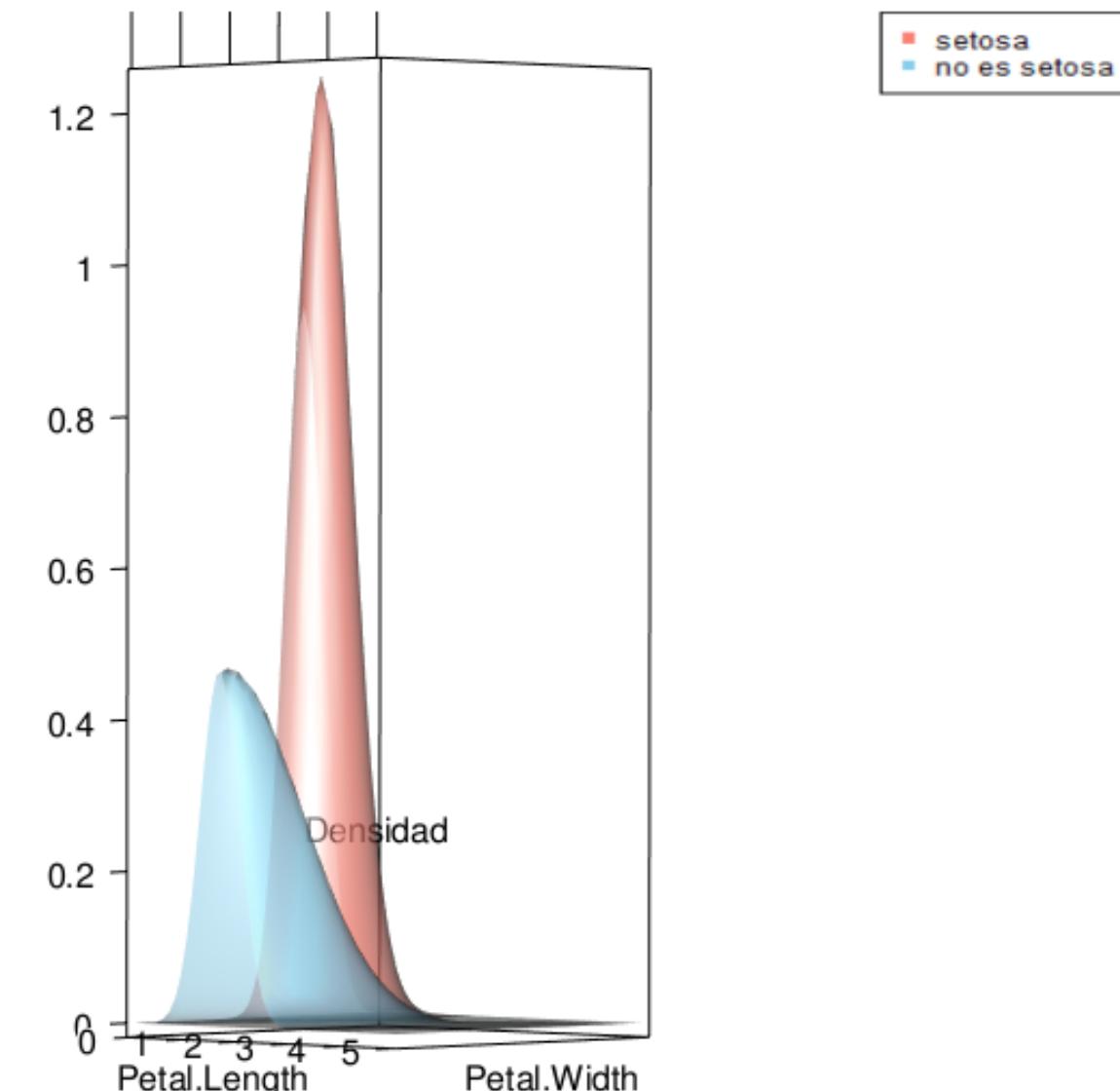


FIGURA 9. Representan multivariante de las variables Petal.Length y Petal.Width para la clasificación de los grupos, setosa y no setosa

#### 4. Árboles de decisión con clasificación

Los árboles de decisión son un método usado en distintas disciplinas como modelo de predicción. Estos son similares a diagramas de flujo, en los que se llega a puntos en los que se toman decisiones de acuerdo con cierta regla determinada.

En el campo del aprendizaje automático, hay distintas maneras de obtener árboles de decisión, la que se usa en esta ocasión es conocida como CART: Classification And Regression Trees. Esta es una técnica de aprendizaje supervisado. Donde se tiene una variable objetivo (dependiente) y la finalidad es obtener una función que me permita predecir, a partir de variables predictoras (independientes), el valor de la variable objetivo para casos desconocidos.

Como el nombre indica, CART es una técnica con la que se pueden obtener árboles de clasificación y de regresión. Se usa clasificación cuando la variable objetivo es discreta, mientras que se usa regresión cuando es continua. Se cuenta con una variable discreta, así que se decide usar la clasificación.

La implementación particular de CART que se usa es conocida como Recursive Partitioning and Regression Trees o RPART. De allí el nombre del paquete que usa del programa R.

De manera general, lo que hace este algoritmo es encontrar la variable independiente que mejor separa los datos en grupos, que corresponden con las categorías de la variable objetivo. Esta mejor separación es expresada con una regla. A cada regla corresponde un nodo.

Se propone el siguiente ejemplo en riesgo de crédito para mejor entendimiento. El riesgo de crédito es el riesgo de que no le paguen a quién le piden prestado. Es la posibilidad de sufrir una pérdida como consecuencia de un impago por parte de la contrapartida en una operación financiera. El riesgo de crédito tiene deudores y no deudores. Los deudores serían las personas que no pagan a quién les prestó, por otra parte, los no deudores pagan su deuda a la contraparte. Ahora supone un ejemplo de árboles con el riesgo de crédito, se propone una variable objetivo con dos niveles: deudor y no deudor. Se supone a una variable que mejor separa los datos y ésta es: ingreso mensual, como resultado se obtiene una regla donde el ingreso mensual es mayo a  $X$  pesos. Esto quiere decir que los datos para los que esta regla es verdadera tienen más probabilidad de pertenecer a un grupo, que al otro. En este ejemplo, se supone que, si la regla es verdadera, un caso tiene más probabilidad de formar parte

del grupo no deudor.

Una vez hecho esto, los datos son separados (particionados) en grupos a partir de la regla obtenida. Después, para cada uno de los grupos resultantes, se repite el mismo proceso. Se busca la variable que mejor separa los datos en grupos, se obtiene una regla, y se separan los datos. Se hace esto de manera recursiva hasta que es imposible obtener una mejor separación. Cuando esto ocurre, el algoritmo se detiene. Cuando un grupo no puede ser partido mejor, se le llama nodo terminal u hoja.

Una característica muy importante en este algoritmo es que una vez que alguna variable ha sido elegida para separar los datos, ya no es usada de nuevo en los grupos que ha creado. Se buscan variables distintas que mejoren la separación de los datos.

Además, después de una partición donde se han creado dos grupos: A y B. Es posible que para el grupo A, la variable que mejor separa estos datos sea diferente a la que mejor separa los datos en el grupo B. Una vez que los grupos se han separado, el algoritmo “no ve” lo que ocurre entre grupos, estos son independientes entre sí y las reglas que aplican para ellos no afectan en nada a los demás.

El resultado de todo el proceso anterior es una serie de bifurcaciones que tiene la apariencia de un árbol que va creciendo ramas, de allí el nombre del procedimiento.

La principal ventaja de este método es su interpretabilidad, pues da un conjunto de reglas a partir de las cuales se pueden tomar decisiones. Este es un algoritmo que no es demandante en poder de cómputo comparado con procedimientos más sofisticados y, a pesar de ello, que tiende a dar buenos resultados de predicción para muchos tipos de datos. Su principal desventaja es que tiene un tipo de clasificación “débil”, pues sus resultados pueden variar mucho dependiendo de la muestra de datos usados para entrenar un modelo. Aunado a que es fácil sobre ajustar los modelos, esto es, hacerlos excelentes para clasificar datos que conocemos, pero deficientes para datos no conocidos.

El algoritmo de árbol de decisiones consiste de 4 pasos los cuales se ejemplifica con la base de datos Iris:

- Paso 1. Determinar la medida de impureza de los datos.
- Paso 2. Ganancia de información de los datos.
- Paso 3. Elección variable de decisión para el nodo raíz.
- Paso 4. Construcción descendiente de los nodos.

**4.1. Paso 1. Determinar la medida de impureza de los datos.** El mejor atributo (variable predictora) es el que separa el conjunto de datos en diferentes clases, de manera más eficaz o es la característica que mejor divide el conjunto de datos. Existen distintas medidas que se conocen como medidas de impureza que son las siguientes:

$$\text{Entropía} = - \sum_{i=1}^n p_i \log_2 p_i$$

$$\text{Gini} = 1 - \sum_{i=0}^{c-1} p_i^2$$

$$\text{Error de clasificación} = 1 - \max p_i$$

En donde  $c$  es el número de clases y  $0 \log_2 0 = 0$  en el cálculo de la entropía.

Ahora, se calculan las medidas de impureza para la base de datos Iris que se utilizará para ajustar un árbol de decisión con las siguientes clases: setosa, versicolor y virginica. De este modo se tiene 3 tipos de especies con 50 registros cada una, por lo que la probabilidad para cada clase es la misma, entonces:  $p_1 = p_2 = p_3 = \frac{50}{150} = \frac{1}{3}$

Con el valor de esta probabilidad se prosigue a calcular el valor de la entropía, como se sigue:

$$\text{Entropía} = - \sum_{i=1}^3 \frac{1}{3} \log_2 \frac{1}{3} = -3 \left( \frac{1}{3} \log_2 \frac{1}{3} \right) = 1.5850$$

$$\text{Gini} = 1 - \sum_{i=1}^3 \left( \frac{1}{3} \right)^2 = 1 - 3 \left( \frac{1}{3} \right)^2 = \frac{2}{3} \approx 0.6667$$

$$\text{Error de clasificación} = 1 - \frac{2}{3} = \frac{1}{3} \approx 0.3333$$

**4.2. Paso 2. Ganancia de información.** La ganancia es un criterio que puede ser usado para determinar el mejor divisor:

$$\Delta = I(MI) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j)$$

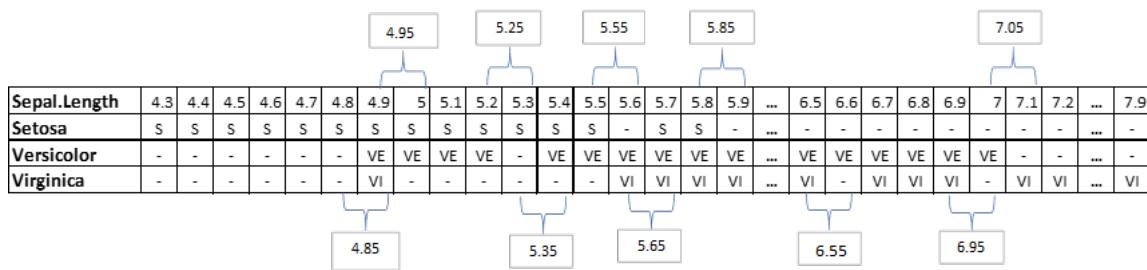
En dónde  $I(MI)$  representa la medida de impureza de un nodo dado (Entropía, Gini) y  $N$  es el total de los registros del padre nodo,  $k$  es el número de atributos y  $N(v_j)$

es el número de registros asociados con el nodo hijo  $v_j$ . Generalmente se elige una condición de prueba para maximizar la ganancia. Cuando la entropía se usa como medida de impureza en la ecuación, se le conoce como ganancia de información IG:

$$IG(S,A) = E(S) - \sum_{j=1}^k p(j)E(j)$$

La ganancia es la magnitud de la variación en la entropía, antes de la división. En dónde  $E(S)$  es la entropía del conjunto S,  $p(j)$  es la proporción del número de elementos dentro del número de elementos en el conjunto S y  $E(j)$  es la entropía del subconjunto j.

Para efectos de ejemplificar el algoritmo se toma a la entropía como medida de impureza para así poder determinar la ganancia de información. Antes de esto es necesario hacer un cálculo en los datos de Iris en los cuales se toman todos los datos de la base de datos y se ordenan de manera ascendente, posteriormente se examinan los puntos de corte de cada especie, es decir, se hace un promedio de los datos entre cada diferencia de clase, en otras palabras, cada que cambia la clase y por último se selecciona el punto de corte con el valor más alto obtenido. A continuación, se muestra un ejemplo para la variable Sepal Length de cómo se ilustra el cambio de clase y el promedio entre los datos de cada cambio:



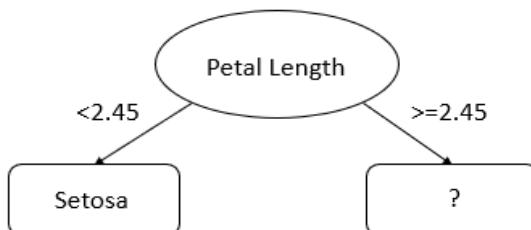
La variable Sepal.Length los valores que dividen estos datos están dados por los siguientes: 4.85, 4.95, 5.25, 5.35, 5.55, 5.65, 5.85, 6.55, 6.95 y 7.05.

Ahora se calcula la ganancia de información para Sepal Length con cada uno de los valores obtenidos anteriormente, de esta manera se tiene que la ganancia de información ( $Sepal.length = (> 4.85)$ ).

$$\begin{aligned}
 \text{IG}(S,A) &= E(S) - \sum_{j=1}^k p(j)E(j) \\
 &= 1.5850 - \left( \frac{16}{150} \left( -\frac{16}{16} \log_2 \frac{16}{16} - 0 - 0 \right) + \frac{134}{150} \left( -\frac{34}{134} \log_2 \right. \right. \\
 &\quad \cdots \frac{34}{134} - \frac{50}{134} \log_2 - \frac{50}{134} - \frac{50}{134} \log_2 - \frac{50}{134} \left. \right) 1.5634 \\
 &= 1.5850 - \frac{16}{150} 0 + \frac{134}{150} 1.5634 \\
 &= 1.5850 - 1.396 \\
 &= 0.1884
 \end{aligned}$$

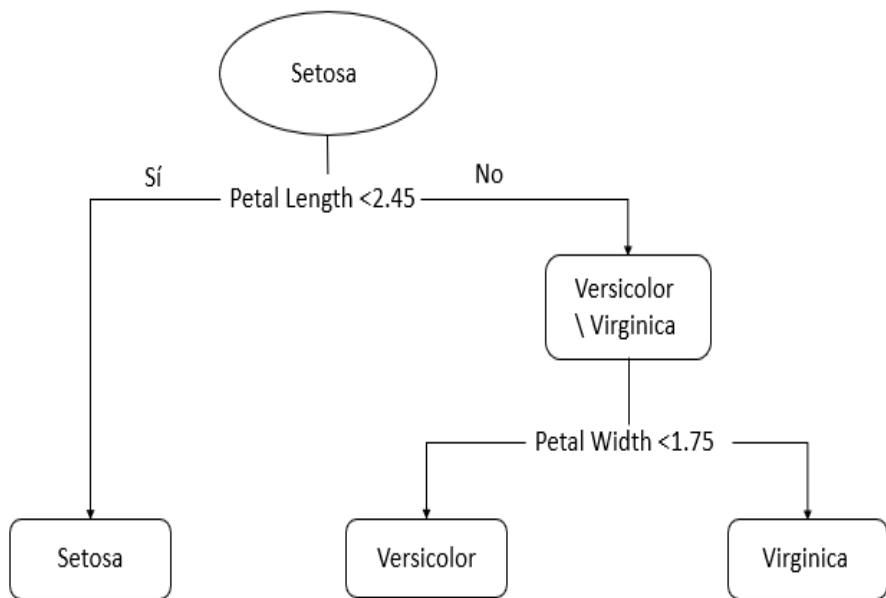
De este modo se obtuvo la ganancia de información para cada valor que divide las variables conforme cambia de especie. Se hizo para los valores 4.85, 4.95, 5.25, 5.35, 5.55, 5.65, 5.85, 6.55, 6.95 y 7.05. Ahora se tiene que hacer lo mismo para las variables faltantes; Sepal width, petal length y petal with.

**4.3. Paso 3. Elección variable de decisión para el nodo raíz.** Posteriormente al encontrar todos los valores correspondientes a la ganancia de información, se selecciona el que sea más grande que los demás ya que éste dará más información. Para las variables Petal.length y Petal.width se obtuvieron los valores más grandes. Para continuar con el árbol de decisiones con la información separada se toma el valor de la variable Petal.length: ganancia de información (Petal.length = (< 2.45)) = 0.9183.



Inicialmente, el total de los datos es de 150 después de encontrar a la variable Petal.length como nodo raíz, la muestra se divide en 2 subconjuntos en función de sus valores de clasificación, es decir, de la especie que sea menor o igual a 2.45 y mayor que 2.45. Ahora, el subconjunto que tenga valores menores o iguales a 2.45 es una clase pura que contiene la clase setosa, mientras que otro subconjunto que tiene valores mayores a 2.45 es una clase impura. Se muestra el nodo raíz graficado y que se observa que si es menor que 2.45 pertenece a la clase setosa.

**4.4. Paso 4. Construcción descendiente del nodo.** Ahora cuando existe un valor mayor o igual a 2.45 puede ser Versicolor o Virginica, para el siguiente nodo se tiene que la variable que presentó más ganancia de información fue Petal.Width, en esta variable se busca de forma recursiva un corte del cambio de clase, tal que al ser menor o igual ya no contemple a la clase setosa y que las dos clases restantes (Versicolor y Virginica) se pueda encontrar el divisor máximo. Después del proceso se verifica que existan valores menores o iguales a 1.75 . Se tiene que la mayoría de los datos se concentra en la clase versicolor. Esto representado en el diagrama se ve de la siguiente manera:



## 5. Comparación de los modelos LDA y Árboles

En esta sección, se busca el comparar los resultados obtenidos por medio de los modelos LDA y árboles de decisión, la comparación se hace mediante las matrices de confusión y verificando un error de entrenamiento.

### Matriz de Confusión

Una matriz de confusión es, una herramienta que permite observar el desempeño del algoritmo en uso. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. El objetivo de la matriz de confusión es que, facilita ver, si el sistema está clasificando de manera correcta las clases o que de alguna manera intuitiva que tanta confusión tiene el modelo, también ayudan a ver claramente los resultados obtenidos al aplicar los modelos en la base de datos Iris en el entorno de software libre y lenguaje de programación interpretado llamado R. Las entradas de la matriz de confusión son las siguientes:

Predicción \ Observación	Setosa	Versicolor	Virginica
Setosa	$VP_{Setosa}$	$FP_{Versicolor}$	$FP_{Virginica}$
Versicolor	$FP_{Setosa}$	$VP_{Versicolor}$	$FP_{Virginica}$
Virginica	$FP_{Setosa}$	$FP_{Versicolor}$	$VP_{Virginica}$

Donde:

$VP_{Setosa}$  : Verdaderos Positivos para Setosa

$FP_{Setosa}$  : Falsos Positivos para Setosa

$VP_{Versicolor}$  : Verdaderos Positivos para Versicolor

$FP_{Versicolor}$  : Falsos Positivos para Versicolor

$VP_{Virginica}$  : Verdaderos Positivos para Virginica

$FP_{Virginica}$  : Falsos Positivos para Virginica

Las entradas en la diagonal principal representan las clasificaciones correctas, mientras que las entradas fuera de la diagonal principal representan errores de clasificación.

### Error de entrenamiento

Como parte de esta interpretación, también se hace uso de una estimación del error en la clasificación de los modelos, conocido como error de entrenamiento, se usa con el nombre de, *training error*. Se define como el porcentaje de instancias en las que el modelo ha predicho incorrectamente la clase en comparación con las clases reales en el conjunto de datos de entrenamiento. Se muestra a continuación la fórmula para el training error:

$$\text{Training error} = \frac{\text{Número de predicciones incorrectas}}{\text{Número total de instancias}} \times 100$$

Por lo que, es el porcentaje de instancias en las que el modelo no ha predicho correctamente la clase real en el conjunto de entrenamiento.

**5.1. Análisis discriminante lineal (LDA).** El resultado de aplicar la función LDA en el entorno de software libre y lenguaje de programación interpretado llamado R. Da como resultado los pasos descritos anteriormente:

```
modelo_lda
```

```
Call:
```

```
lda(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
data = iris)
```

Prior probabilities of groups:

```
setosa versicolor virginica
0.3333333 0.3333333 0.3333333
```

Group means:

```
Sepal.Length Sepal.Width Petal.Length Petal.Width
```

```
setosa 5.006 3.428 1.462 0.246
```

```
versicolor 5.936 2.770 4.260 1.326
```

```
virginica 6.588 2.974 5.552 2.026
```

Coefficients of linear discriminants:

```
LD1 LD2
```

```
Sepal.Length 0.8293776 0.02410215
```

```
Sepal.Width 1.5344731 2.16452123
```

```
Petal.Length -2.2012117 -0.93192121
```

```
Petal.Width -2.8104603 2.83918785
```

Proportion of trace:

LD1 LD2

0.9912 0.0088

En donde, los coeficientes de los discriminantes lineales, LD1 y LD2 indican la contribución de las variables de la base de datos iris para la construcción de los discriminantes lineales. Se observa que LD1 la contribución mayor positiva proviene de Sepal.Length y Sepal.Width, mientras que Petal.Length y Petal.Width tienen contribuciones negativas. Finalmente, la mayor parte de la variabilidad se captura en LD1, lo que indica que una sola dimensión lineal puede ser suficiente para diferenciar las especies en este conjunto de datos. De este modo, el LDA cuenta con discriminantes lineales que separan correctamente las tres especies de iris en función de las variables Sepal.Length, Sepal.Width, Petal.Length, Petal.Width.

A continuación, se observa la matriz de confusión resultante del modelo LDA, al aplicarlo a la base de datos Iris. Toma los 150 datos. La función que calcula el LDA en el lenguaje de programación R. Se interpreta que de un total de 50 de la especie setosa está prediciendo que las 50 son setosa, para la especie versicolor también hay 50 datos de los cuales 48 los está prediciendo como versicolor y 2 clasificados en la especie virginica, por último, del total de 50 datos registrados como virginica predijo 1 como versicolor y el resto fue acertado para la especie virginica, de este modo se tiene un training error del 2 %. Por lo que se puede decir que el modelo LDA aplicado en la base de Iris funciona muy bien.

Observación	Predicción	setosa	versicolor	virginica
		setosa	versicolor	virginica
setosa	setosa	50	0	0
	versicolor	0	48	2
	virginica	0	1	49

Ahora, se muestra una matriz de confusión con los datos entrenados. Se ve la matriz de confusión para test y train. Datos que antes fueron divididos aleatoriamente bajo el criterio %70, %30 para poner a prueba el funcionamiento del modelo.

#### Matriz de confusión de los datos test

La matriz de confusión resultante del modelo LDA, al aplicarlo a los datos de la base test. Toma las 45 observaciones. La función que calcula el LDA en el lenguaje

de programación R. Se interpreta que de un total de 12 de la especie setosa está prediciendo que las 12 son setosa, para la especie versicolor hay 19 datos de los cuales 18 los está prediciendo como versicolor y 1 lo clasifica en la especie virginica. Por último, del total de 14 datos registrados como virginica predijo acertadamente las 14 observaciones para la especie virginica, de este modo se tiene un training error del 2 %. Por lo que se puede decir que el modelo LDA aplicado en los datos de entrenamiento funciona bien.

Observación \ Predicción	setosa	versicolor	virginica
setosa	12	0	0
versicolor	0	18	1
virginica	0	0	14

### Matriz de confusión de los datos train

La matriz de confusión que resultante de aplicar el modelo LDA a los datos train. Toma 105 observaciones aleatorias. Se interpreta que de un total de 38 observaciones de la especie setosa está prediciendo que las 38 son setosa, para la especie versicolor hay 31 datos de los cuales 31 los está prediciendo como versicolor, por último, del total de 36 datos registrados como virginica predijo 1 como versicolor y el resto fue acertado para la especie virginica, de este modo se tiene un training error del 2 %. Por lo que se puede decir que el modelo LDA aplicado a los datos train se obtiene un resultado acertado.

Observación \ Predicción	setosa	versicolor	virginica
setosa	38	0	0
versicolor	0	31	0
virginica	0	1	35

**5.2. Árboles de decisión por clasificación.** Se procede a aplicar el modelo de árboles de decisión para la base de datos Iris, se obtiene la siguiente el siguiente diagrama: Figura 10.

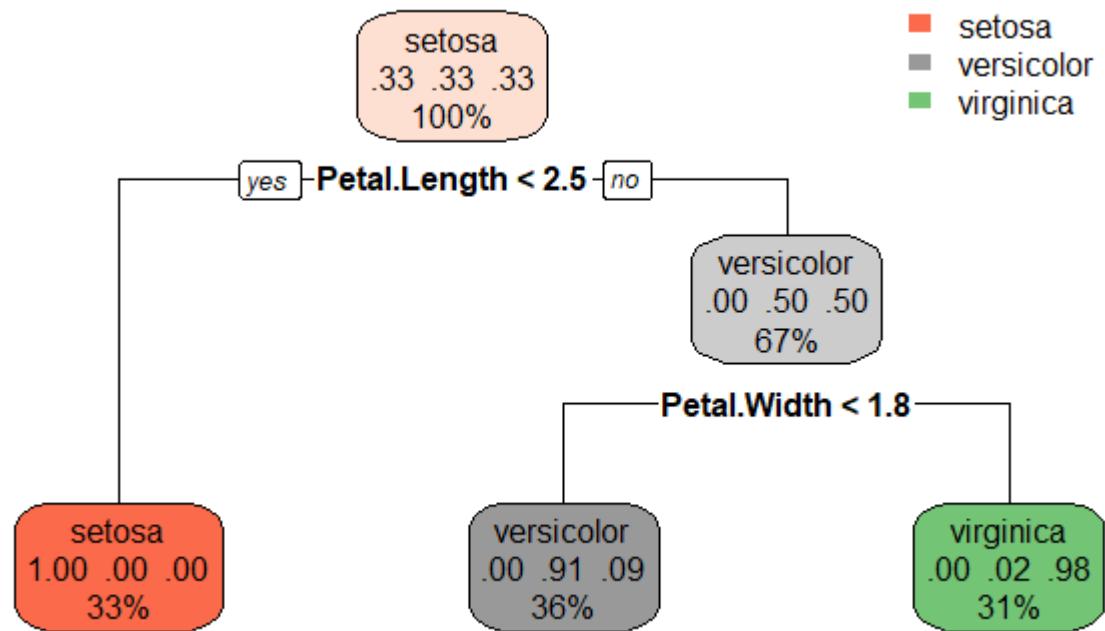


FIGURA 10. Diagrama de árbol

En dónde el nodo principal es separado por la variable Petal.length que en donde se observa el algoritmo separa la muestra con el valor 2.5 que es el valor del divisor dejando así que todos los que sean menores de este valor serán de la especie setosa dejando concentrada el 33 % de la información en la especie setosa. Todas se encuentran clasificadas aquí, por otro lado, el modelo toma el siguiente nodo que es Petal.width, en donde si es menor que 1.8 entonces la muestra indica que el 36 % de la información se encuentra concentrada en esta información, en la otra rama se encuentra la especie virginica con un 31 % de la información. Por lo que se puede apreciar que para la base de datos Iris resulta ser muy efectivo.

Al aplicar la función que calcula los árboles de decisión en el lenguaje de programación R . Se interpreta que de un total de la especie setosa de 50 está prediciendo que las 50 son setosa, para la especie Versicolor hay 54 datos de los cuales 49 los está prediciendo como Versicolor y por último del total de 46 datos registrados como

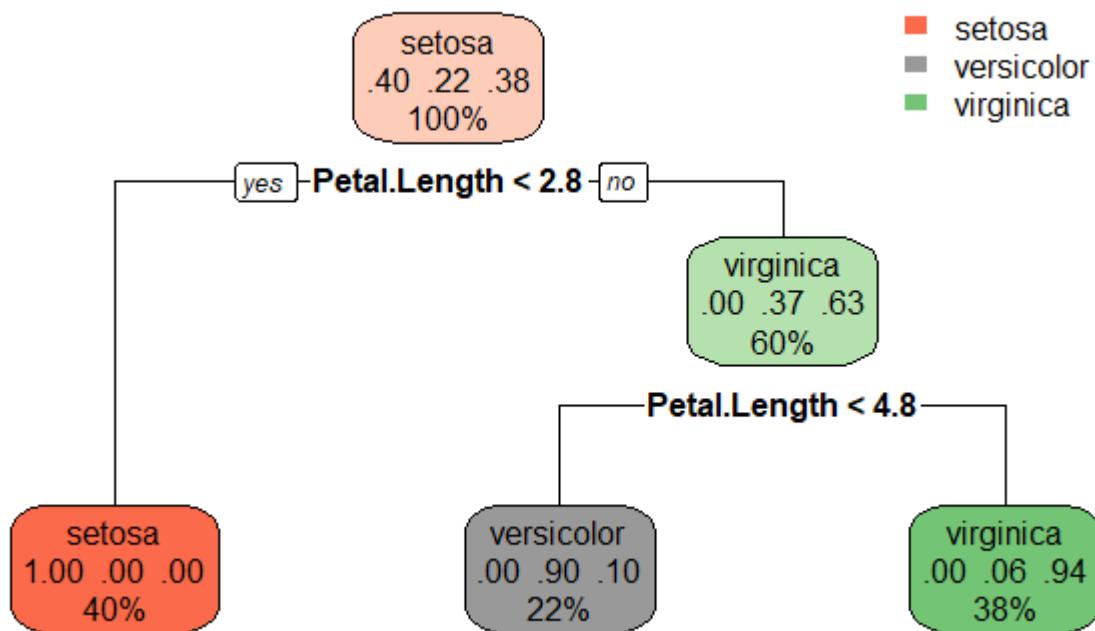
Virginica predijo 1 como versicolor y el resto fue acertado para la especie virginica, de este modo se tiene un training error del 2%. Por lo que se puede decir que el modelo árboles de decisión aplicado en la base de Iris funciona bien.

Observación \ Predicción			
	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	49	5
virginica	0	1	45

Se necesita entrenar el modelo y ver que funcione de forma en que se pueda obtener dos muestras aleatorias, en otras palabras, dividiremos la base de datos en train y test. Ahora se muestra una matriz de confusión con los datos entrenados. Se ve la matriz de confusión para test y train. Datos que antes fueron divididos aleatoriamente para poner a prueba el funcionamiento del modelo.

#### Matriz de confusión de los datos test

Al aplicar la función en el lenguaje de programación R se obtiene lo siguiente:

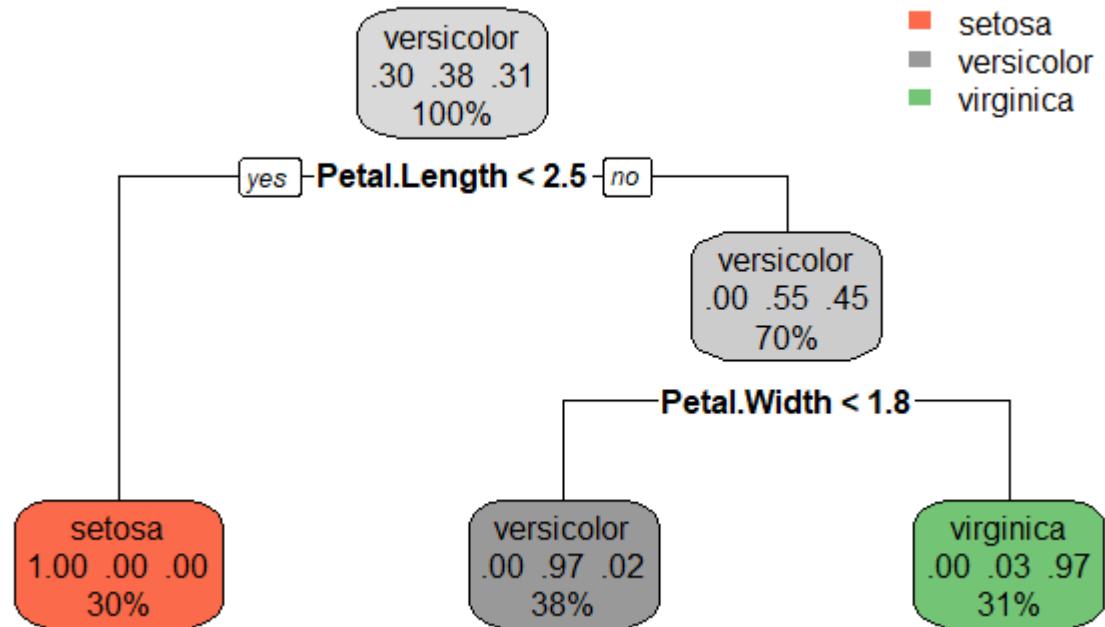


Al aplicar la función. Se interpreta que de un total de la especie setosa de 15 está prediciendo que las 15 son setosa, para la especie versicolor hay 13 datos los cuales está prediciendo como versicolor, por último, del total de 17 datos registrados como virginica predijo 1 como versicolor y el resto: 16 especies clasificadas como virginica, de este modo se tiene un training error del 2 %. Se concluye que el modelo funciona bien en los datos entrenados.

Con la siguiente matriz de confusión:

Observación \ Predicción			
	setosa	versicolor	virginica
setosa	15	0	0
versicolor	0	13	0
virginica	0	1	16

**Matriz de confusión de los datos train** Al aplicar la función del programa R se obtiene lo siguiente:



Al aplicar la función que calcula los árboles de decisión al lenguaje de programación R. Se interpreta que de un total de la especie setosa de 35 está prediciendo

que las 35 son setosa, para la especie versicolor hay 38 datos de los cuales 35 los está prediciendo como versicolor y 3 clasificados en la especie virginica, por último, del total de 32 datos registrados como virginica predijo 1 como versicolor y 31 especies clasificadas como virginica, así, se tiene un training error del 2 %. Se concluye que el modelo funciona bien en los datos entrenados.

Con la siguiente matriz de confusión:

Observación \ Predicción	setosa	versicolor	virginica
setosa	35	0	0
versicolor	0	35	3
virginica	0	1	31

## 6. Método de Bootstrap

Es sabido que puede existir el caso en donde la muestra de los datos sea tan pequeña, que hace difícil que el modelo se aplique de manera correcta, por ello se puede usar el método de Bootstrap, que ayuda a hacer más consistentes los datos y ayuda a agregar más volumen a la base de datos, generando datos aleatorios.

Se comienza por definir el método de Bootstrap: Bootstrap es un método de inferencia sobre una población que utiliza datos de muestra. Bradley Efron lo presentó en 1979. Bootstrap se basa en el muestreo con reemplazo de datos de muestra. Esta técnica se puede utilizar para estimar el error estándar de cualquier estadístico y obtener un intervalo de confianza.

A grandes rasgos, es un proceso en el que se toma una muestra de los datos existentes y se le aplica variabilidad para simular nuevas observaciones. La muestra aleatoria de los datos que se toma es con reemplazo; el muestreo con reemplazo es aquel en que un elemento puede ser seleccionado más de una vez en la muestra para ello se extrae un elemento de la base y se devuelve a la población, por lo que de esta forma se pueden hacer infinitas extracciones de la población aun siendo esta finita. La muestra Bootstrap es al menos del mismo tamaño que el conjunto de datos original. Algunas muestras estarán representadas varias veces, mientras que otras no se serán seleccionadas. Las muestras no seleccionadas suelen denominarse muestras “fuera de bolsa”. Para una iteración de bootstrap, se construye un modelo con las muestras seleccionadas. Las que están fuera de la bolsa y se utiliza para predecir estas mismas.

En la Figura 11 Se muestran doce elementos diferentes en el conjunto de entrenamiento, se representan con símbolos de esta manera se asignan a B subconjuntos. Cada subconjunto es del tamaño de la muestra original y puede contener algunos elementos del mismo tipo de símbolo. Las muestras que no son seleccionadas por el bootstrap se predicen y se utilizan para estimar el rendimiento del modelo, en el esquema se observa que están en la parte derecha en donde dice *Predict on*. En la parte izquierda del esquema *Build Model With* se muestra la construcción de como se puede ir tomando las muestras aleatoriamente.

En conclusión para la aplicación del método de Bootstrap se necesita una función que divida la muestra, para fines del lenguaje de programación R. Se tiene la función que se encarga de dividir muestras. A ésta función se le debe indicar el tamaño y además que debe ser un muestreo con remplazo. Para así poder generar datos aleatorios para agregarle más volumen a los datos.

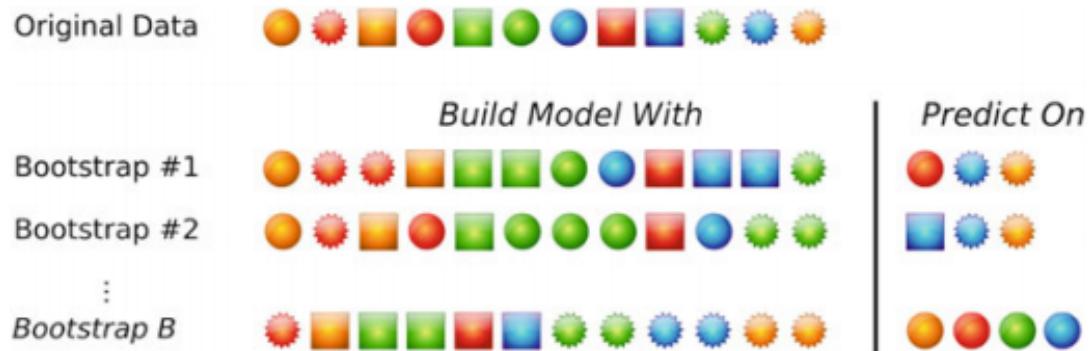


FIGURA 11. Esquema del remuestreo bootstrap. Obtenido de Applied Predictive Modeling. En dónde: *Original Data* como Datos Originales, *Build Model With* como Construir Modelo Con y *Predict On* como Realizar Predicciones En Muestras



## Capítulo 3

# Análisis Exploratorio

En este capítulo se busca hacer un análisis, es decir, dar una idea intuitiva de los datos mediante gráficos estadísticos. También se identifican características como; media, moda, mediana, valores máximos, valores mínimos, desviación estándar, varianza, coeficiente de asimetría entre otros. En éste capítulo se ve la limpieza de los datos, la descripción de las variables y las complementaciones a ésta misma y la relación que existe entre las variables.

### 1. Limpieza de datos

Los datos se obtuvieron con la ayuda de la Web API para desarrolladores de Spotify. En la sección *Audio Features & Analysis platform icon for web platform icon for ios platform icon for android*. Se exploran las características de audio y el análisis de audio a profundidad de las canciones. Con ayuda de la plataforma para desarrolladores se pueden leer las características cuantificadas de audio para las canciones en las que se puede conocer su capacidad de baile (danceability), energía (energy), valencia (valence). También se puede consultar casos más avanzados para mayor análisis sobre las canciones, tales como, los compases, los ritmos, los tonos entre otros. Las características se clasifican de la siguiente manera:

- Estado de ánimo: Danceability, Valence, Energy, Tempo.
- Propiedades: Loudness, Speechiness, Instrumentalness.
- Contexto: Liveness, Acousticness, segments, Tatums, Bars, Beats.

Las características de sonido que se usan en esta tesis son las siguientes:

- |          |                |               |
|----------|----------------|---------------|
| ▪ title  | ▪ danceability | ▪ speechiness |
| ▪ artist | ▪ loudness     |               |
| ▪ genres | ▪ liveness     |               |
| ▪ year   | ▪ valence      |               |
| ▪ tempo  | ▪ duration_ms  |               |
| ▪ energy | ▪ acousticness |               |

Estas características de las canciones se descargan de unas listas de reproducción llamadas playlist creadas por Spotify, en las cuales se encuentran las canciones más populares de cada década. Como ejemplo se supone la década de los años ochenta, la playlist utilizada sería *All Outs 80s* que es creada por Spotify. Para fines de esta investigación se descargan todas éstas playlist disponibles en Spotify con las canciones más populares de las décadas de los años de 1950 hasta el 2010. Posteriormente se unen estos archivos csv para crear una base de datos y poder trabajar de manera eficiente. A esta base donde se unieron los archivos, se le agregó una variable llamada; década, la cual nos indica el año de la década de a la cual pertenece.

De esta manera, se decidió agregar una variable más, la cual indica una reclasificación del género al que pertenece cada canción de la base de datos. Esta decisión se tomó ya que no se puede trabajar con la información proporcionada con la variable genres, que contemplaba 116 tipos de géneros distintos, los cuales dificultaban la forma de presentar los datos en forma de gráfica. Se decidió reclasificar con una variable genre; de la siguiente manera, por ejemplo la variable genres contenía glam rock, soft rock o dance rock, entonces se clasificó en la variable rock, otro ejemplo es si la variable genres era afropop, Brill building pop o europop se clasificó como pop, consiguiendo así una disminución de las variables en esta variable categórica. Se reclasificó de 116 géneros a 14 géneros.

Como referencia para la reclasificación de los géneros se tomó en cuenta los géneros musicales favoritos del mundo, Ver figura 1, de un estudio de la empresa Statista que es un proveedor líder de datos de mercado y consumidores. En donde muestran los 10 géneros más populares del mundo. Teniendo esta referencia se reclasificó el subgénero para agruparlos en los géneros, sin embargo, en la base original existen distintos géneros que no se pudieron agrupar en esta categoría de los 10 más populares, por lo que, se mantuvo el género que tenían originalmente, por ejemplo, el género Funk o la clasificación Adult standards.

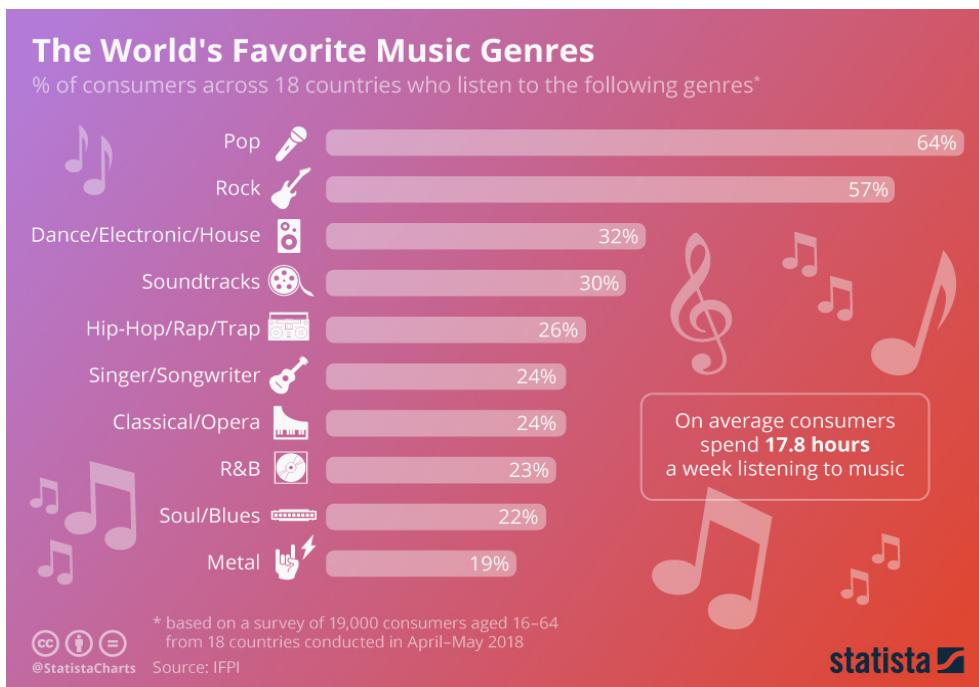


FIGURA 1. Gráfico de Statista proveedor de datos de mercado. *The World's Favorite Music Genres* como: Los géneros musicales favoritos en el mundo

A continuación, se muestra una gráfica con la reclasificación en la base de datos, reduciendo la agrupación por género. Ver figura 2.

En este diagrama ilustrativo se logra justificar porqué se decidió hacer una reagrupación del género. Como se observa en esta figura los nombres de los géneros son claros, debido a la disminución de la clasificación, por otro lado, se ve que los datos informativos son claros y nada saturados. Por ende, se puede obtener una interpretación útil. Viendo estos dos resultados obtenidos se puede concluir que la decisión de reclasificar el género fue de suma importancia y utilidad para continuar con el objetivo de éste capítulo, que es el análisis exploratorio de la base de datos.

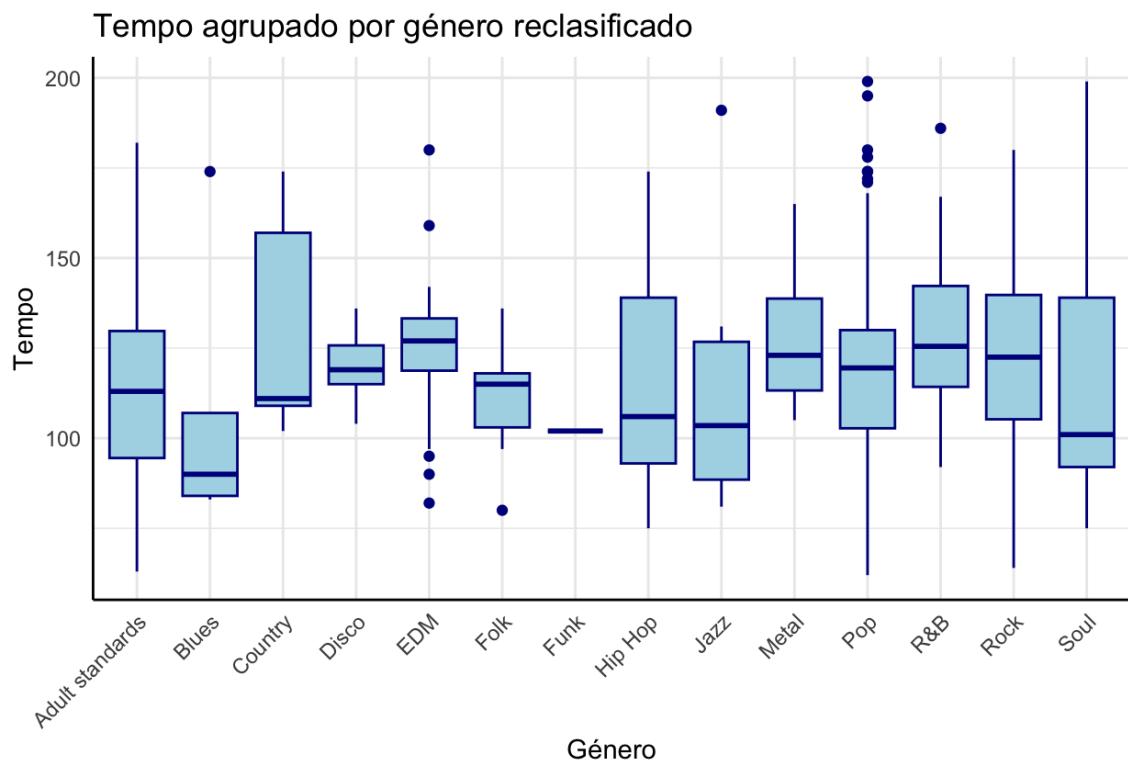


FIGURA 2. Diagrama de cajas con la reclasificación

## 2. Descripción de Variables

Los datos originalmente contaban con el nombre de las variables en inglés, por lo que, por mejor tratamiento de los datos se decidió renombrar los nombres de las variables al idioma del español. La descripción de las variables que se dará a continuación contiene las modificaciones que se mencionaron en la sección anterior, por consiguiente la base final que se trabaja cuenta con las siguientes variables:

- Década: Indica el Año de la década a la que pertenece la lista de reproducción de Spotify “All outs” desde los años 1950 hasta 2010, esta variable se agregó para darle mejor tratamiento a la base. Variable de tipo entero.
- Título: Indica el título de la canción, esto es el nombre de la canción. Es una variable de tipo string.
- Géneros: Una lista de los géneros utilizados para clasificar el álbum. Por ejemplo: ”Prog Rock”, ”Post-Grunge”. (Si aún no está clasificado, la matriz está vacía). Es de tipo
- Artista: Indica el cantante, la banda o el grupo, que interpretada la canción. Los artistas del álbum. Cada objeto artista incluye un enlace en href a información más detallada sobre el artista. Es una variable de tipo Array.
- Género: Variable que se agregó con la reclasificación del género para mejor tratamiento de la información. Esta variable se utiliza como una variable de clasificación de información. Los géneros se pueden ver como clases.
- Año: Ésta variable representa el año de lanzamiento o puede existir el caso de que sea el año de relanzamiento o de reediciones. Es una variable de tipo Array.
- Tempo: El tempo global estimado de una pista en pulsaciones por minuto (BPM). En la terminología musical, el tempo es la velocidad o el ritmo de una pieza determinada y se deriva directamente de la duración media de los tiempos. Variable de tipo Float.
- Energía: La energía es una medida de 0.0 a 1.0 y representa una medida perceptiva de intensidad y actividad. Por lo general, las pistas energéticas se sienten rápidas, ruidosas y con mucho ruido. Por ejemplo, el death metal tiene mucha energía, mientras que un preludio de Bach tiene una puntuación baja en la escala. Entre las características perceptivas que contribuyen a este

atributo se encuentran el rango dinámico, el volumen percibido, el timbre, la velocidad de aparición y la entropía general. Es de tipo Float.

- Bailable: La bailableabilidad describe la idoneidad de una pista para el baile basándose en una combinación de elementos musicales como el tempo, la estabilidad del ritmo, la fuerza del compás y la regularidad general. Un valor de 0.0 es el menos bailable y 1.0 el más bailable. Tipo Float.
- Volúmen: La sonoridad general de una pista en decibelios (dB). Los valores de sonoridad se promedian en toda la pista y son útiles para comparar la sonoridad relativa de las pistas. La sonoridad es la cualidad de un sonido que es el principal correlato psicológico de la fuerza física (amplitud). Los valores suelen oscilar entre -60 y 0 db. Tipo Float.
- En vivo: Detecta la presencia de público en la grabación. Los valores más altos de liveness representan una mayor probabilidad de que la pista haya sido interpretada en directo. Un valor superior a 0.8 proporciona una fuerte probabilidad de que la pista sea en directo. Tipo Float.
- Positividad: Lo que en inglés se conoce como "Valence" Una medida de 0.0 a 1.0 que describe la positividad musical que transmite una pista. Las pistas con alta valencia suenan más positivas (por ejemplo, felices, alegres, eufóricas), mientras que las pistas con baja valencia suenan más negativas (por ejemplo, tristes, deprimidas, enfadadas). Es de tipo Float.
- Duración: Duración de la canción que está medida en segundos.
- Acústico : Una medida de confianza de 0.0 a 1.0 sobre si la pista es acústica. 1.0 representa una alta confianza en que la pista es acústica. Es de tipo Float.
- Habla: detecta la presencia de palabras habladas en una pista. Cuanto más exclusivamente hablada sea la grabación (por ejemplo, un programa de entrevistas, un audiolibro o una poesía), más se acercará a 1.0 el valor del atributo. Los valores superiores a 0.66 describen pistas que probablemente estén compuestas exclusivamente por palabras habladas. Los valores entre 0.33 y 0.66 describen pistas que pueden contener tanto música como voz, ya sea en secciones o en capas, incluyendo casos como la música rap. Los valores inferiores a 0.33 representan probablemente música y otras pistas no habladas. Es de tipo Float.

- Popularidad: La popularidad del álbum. El valor estará entre 0 y 100, siendo 100 el más popular. La popularidad se calcula a partir de la popularidad de las pistas individuales del álbum. La variable es de tipo Integer.

### 3. Relación de Variables

Una de las técnicas del análisis exploratorio es buscar e identificar la relación que existe entre las variables. Una manera de poder llegar a este resultado de forma ilustrativa es con ayuda de un diagrama de correlaciones, que en este caso resulta ser el siguiente: (Ver figura 3)

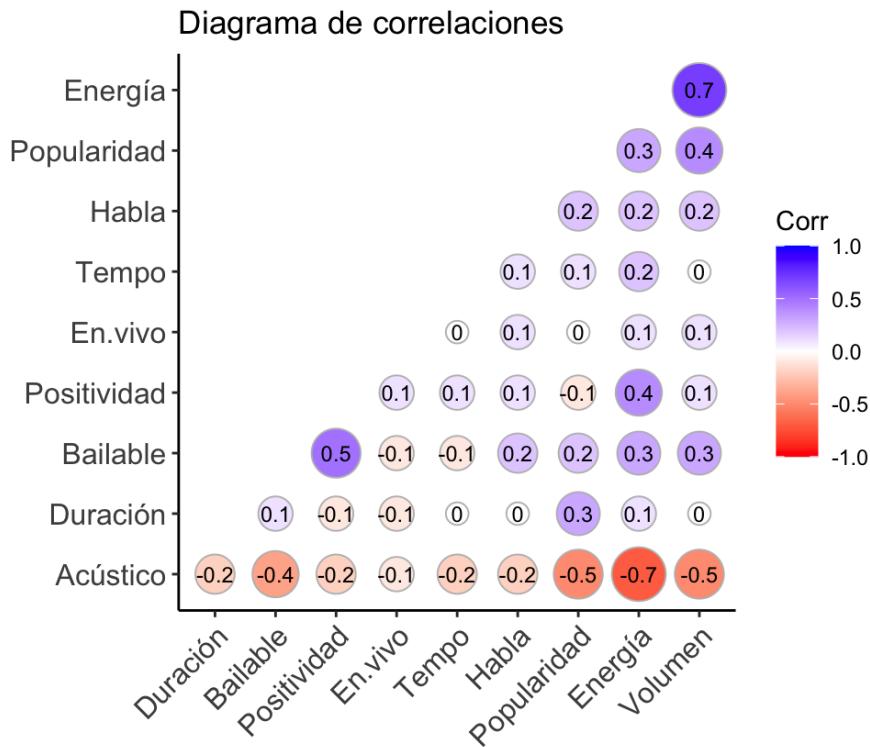


FIGURA 3. Diagrama de correlaciones entre las variables que contienen cada canción

Se observa que se puede dividir la correlación entre las variables de la siguiente manera; la primera es ver las correlaciones que son muy pequeñas casi cero, es decir que no dependen una de la otra, la segunda son las correlaciones negativas, que indican que si una variable incrementa la otra disminuye y por último es ver las correlaciones positivas, que definen que ambas variables disminuyen o decrementan.

Se comienza con el análisis de las correlaciones casi cero. La variable En vivo, donde En vivo determina si la canción puede estar grabada en vivo. Se observa en el

diagrama de correlaciones que ésta variable tiene correlación casi cero con la mayoría de las variables y esto hace sentido, la correlación de En vivo con Tempo es cero ya que no depende del tempo si la canción está grabada en vivo. La correlación entre la variable En vivo y duración es -0.1 indica que la duración no depende de si la canción es grabada en vivo. La correlación con la variable Bailable es -0.1 que de la misma manera no depende de que la canción sea grabada en vivo.

Ahora, se observa la correlación negativa que existe entre las variables Energía y Acústico. Se puede decir que esta observación es cierta ya que las canciones acústicas no suelen ser energéticas. Existe una correlación negativa de -0.4 entre las variable Acústico y Bailable, ésta correlación representa que si una canción es acústica no es bailable y viceversa. Cabe agregar que de las variables de más importancia para fines de éste análisis es la variable Popularidad, que nos dice que tan popular es una canción. En el diagrama de correlaciones se ve que se correlaciona negativamente con lo acústico, es decir que si la canción es acústica la canción no será popular, por otro lado, la correlación negativa que existe entre la variable positividad, que es la que indica el impacto positivo del estado del ánimo, está correlación negativa dice que las canciones populares son tristes. Y esto es un fenómeno que es cierto, ya que en la actualidad y a lo largo del tiempo las canciones tristes suelen ser populares, ”¿Por qué nos atrae la música que nos triste? La tristeza no es una emoción deseable en absoluto (para la mayoría de las personas), entonces, ¿por qué la buscamos a menudo? ¿Y qué tipo de placer extraño y doloroso obtenemos de él?”<sup>1</sup>. A éste efecto se le conoce como la paradoja de la tristeza, que en éste mismo artículo menciona un estudio que hicieron los investigadores de la Universidad de Limerick, que arrojó en resumen los siguientes resultados respecto a las personas adultas que decidieron escuchar música triste; se identificó que la música triste les sirvió para experimentar afecto; es decir procesar sus emociones. También a hacer un análisis cognitivo, que tiene que ver con la empatía ya que la tristeza es un sentimiento en común que tenemos los humanos y de cierta forma se podría entender que no se tiene el sentimiento de estar solo, otro aspecto fue la distracción, ya que al escuchar música triste se encontró que se olvidan de los pensamientos tristes o abrumadores. Con esto se podría dar un poco respuesta a él ¿por qué de sentirnos atraídos por la música triste? y ver que en el impacto de la popularidad es de suma importancia.

De las correlaciones positivas que se pueden rescatar son la de las variables Energía y Volumen, que se relacionan positivamente, y tiene sentido que una canción que sea muy energética tenga un alto volumen. Para las variables Bailable y

---

<sup>1</sup>Danielle Fong. (2018). The Sadness Paradox – Why do we enjoy listening to music that makes us sad?.

Positividad se ve una correlación alta, lo que se interpreta es que si la canción es bailable resulta tener un mayor impacto positivo en el estado de ánimo que causa la canción. Las siguientes correlaciones que se observan son de las variables Popularidad y Volumen, se interpreta que la popularidad incrementa si la sonoridad incrementa o en su defecto decrementa para ambas, este es un fenómeno cierto que se ve en la actualidad, debido a que las canciones con mayor sonoridad resultan ser populares. Otra correlación positiva que se identifica con las variables Energía y Popularidad, por lo que se puede decir que si una canción es muy energética es popular. La variable Bailable, que nos dice que tan bailable es una canción también se relaciona positivamente con la variable Popularidad, y de cierta manera se puede decir que esa observación hace sentido; si la canción es muy bailable o tiene ritmo resulta que es popular.

#### 4. Análisis de variables por género

En ésta sección se ve la descripción de las variables con ayuda de los géneros: Adult standards, Blues, Country, Disco, EDM, Folk, Funk, Hip Hop, Jazz, Metal, Pop, R&B, Rock, Soul, con técnicas gráficas que son utilizadas en el análisis exploratorio, tales como, gráficos de barras, diagramas de cajas y diagramas de correlaciones, con el fin de obtener información que sirva de interpretación de los datos.

Se comienza con un gráfico de barras, para mostrar la proporción que le corresponde a cada categoría, en este caso es el género.

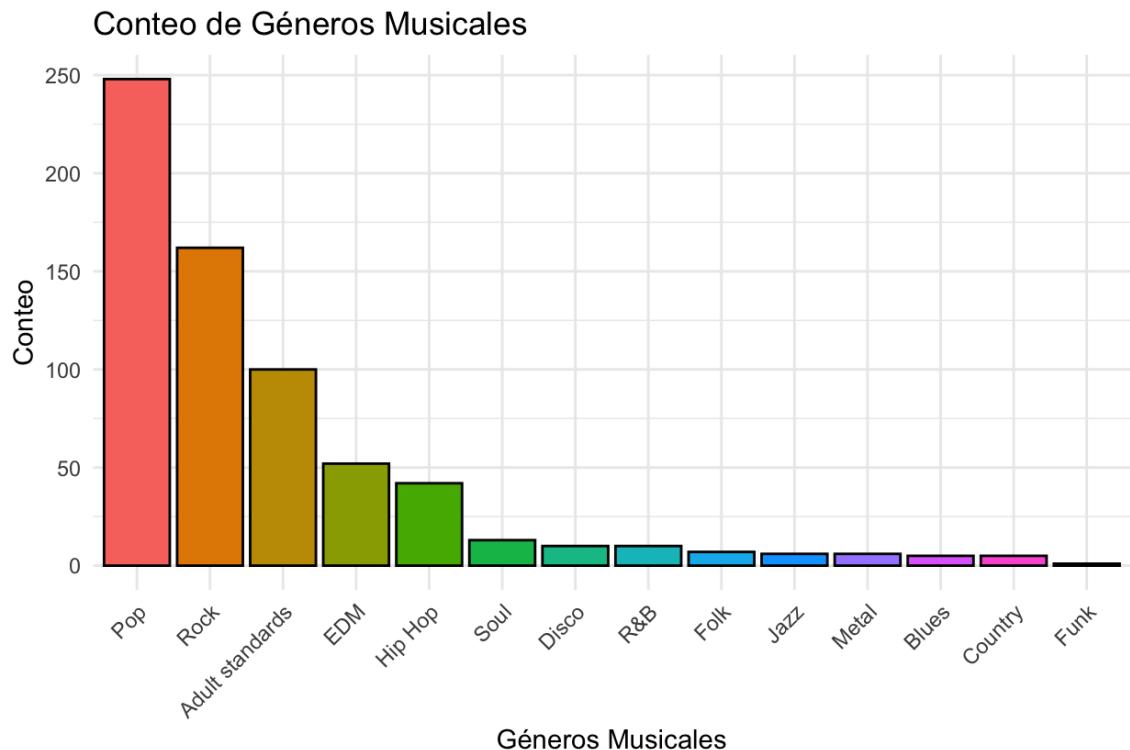


FIGURA 4. Gráfica de proporción por Género

Se observa que el género Pop es la muestra más grande en cuestión de proporción, seguido del Rock y Adult Standards (Canciones dirigidas principalmente para personas mayores de 50 años). Por lo que se tiene que estos tres géneros son los que más predominan en la base. Ahora, el género Funk se ve muy pequeño gráfica, debido a

que la muestra solo tiene una canción para este género.

Otra de las técnicas gráficas del análisis exploratorio es elaborar diagramas de caja que indican la descripción de características importantes; tales como la dispersión y simetría. Se representan los cuartiles, los valores mínimos y máximos de los datos, ayuda a proporcionar una visión rápida de la distribución y posibles datos atípicos. Se puede evaluar la dispersión de los datos y la presencia de valores extremos. Se tiene que se puede evaluar la simetría y posición de la caja esta proporciona información sobre la distribución de los datos. Si la caja está centrada y simétrica puede ser que los datos estén distribuidos de manera uniforme. Si se observa que la caja está desplazada hacia arriba o hacia abajo puede ser una asímetría. Se tiene un diagrama de caja para cada una de las variables de la base de datos clasificado por género. Se muestran a continuación los diagramas de cajas:

Figura 5. El primer diagrama para analizar es para la variable Acústico en la cual se observa que está clasificada por género. Para los géneros Country y R&B se observa que se puede determinar que en su mayoría las canciones pertenecen al género acústico ya que la mayoría de los datos se quedan dentro de la caja.

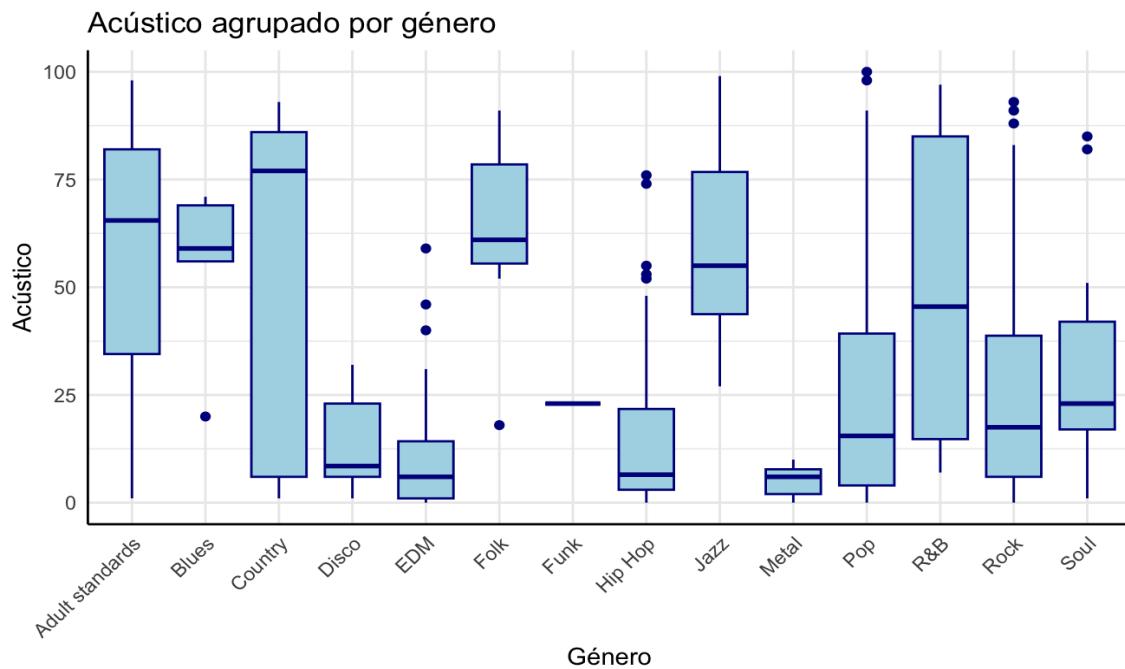


FIGURA 5. Acústico por género

Figura 6 . Para el género Country los datos se comportan típicamente ya que todos los datos están ahí, por otro lado para el género de la música pop se tienen datos atípicos. Con ésta información no se puede decir que si una canción es bailable esta pueda determinar el género.

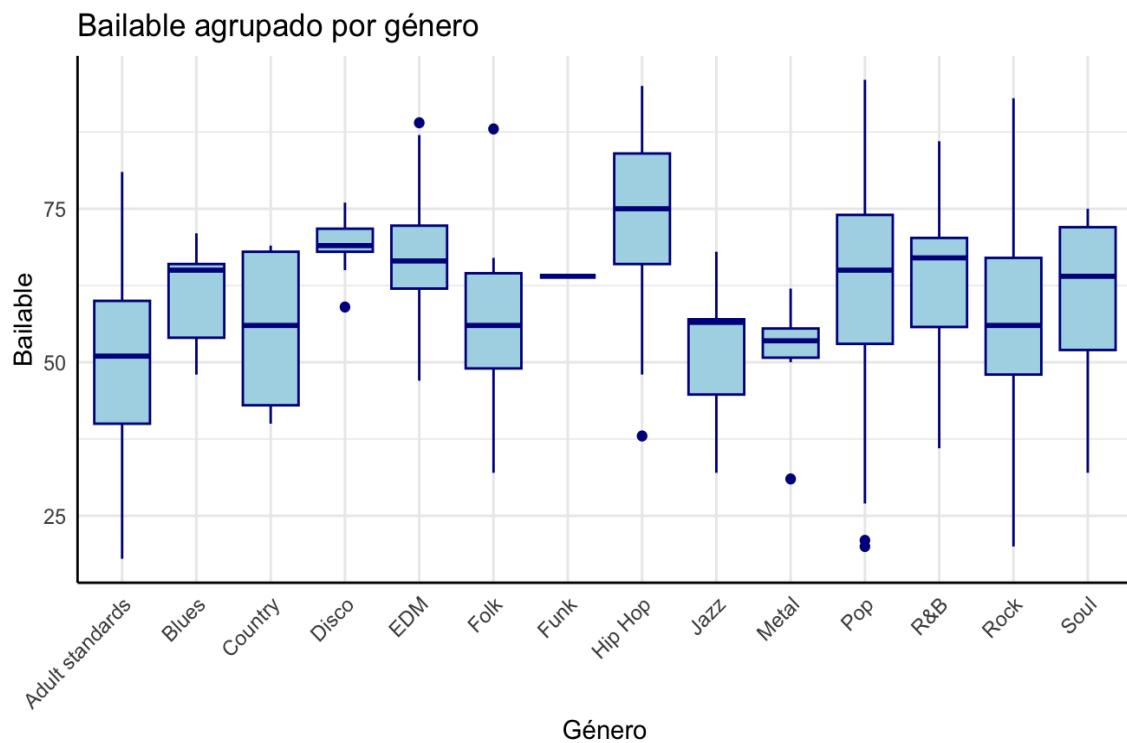


FIGURA 6. Bailable por género

Figura 7 . Se observa que la duración de las canciones no determina el género.

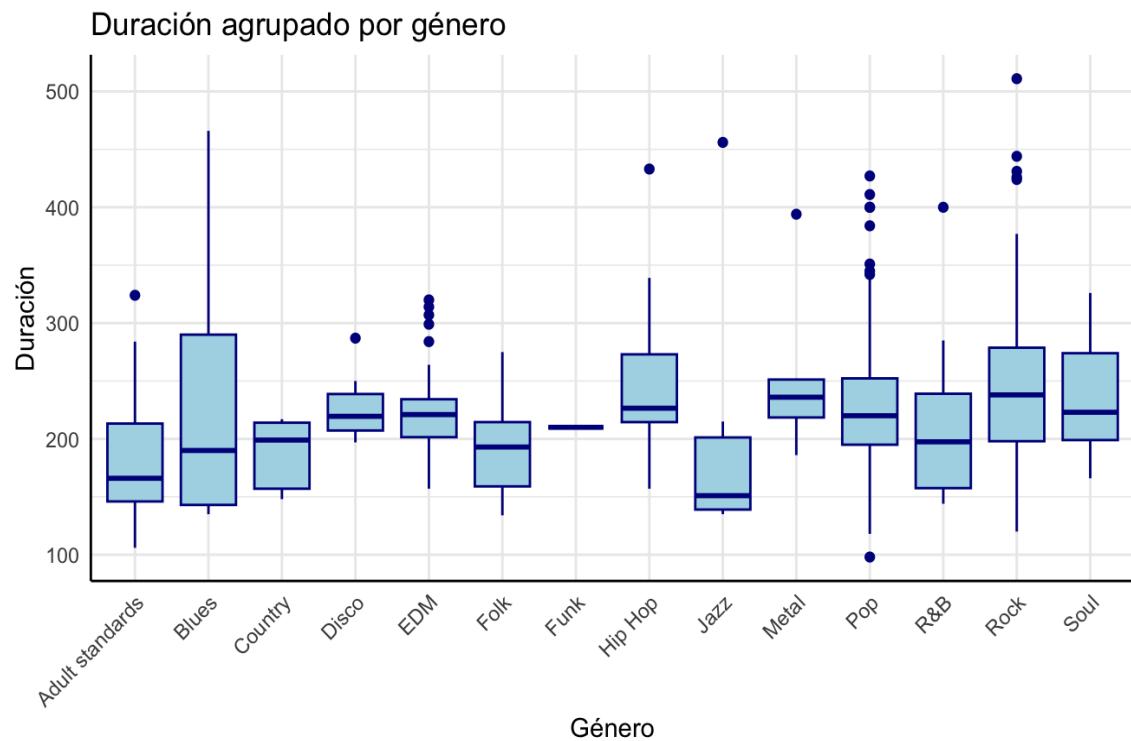


FIGURA 7. Duración por género

Figura 8 . Se ilustra que la energía esta relacionada con el género. Se puede decir que se puede hacer una separación de dos grupos, el primero contiene Adult Standards, Blues, Country, Folk y Jazz y para el segundo grupo se ve el resto de los géneros. Se aprecia que el Metal queda hasta arriba de la escala comportándose de cierta manera típicamente. Por lo que la energía puede ayudar a clasificar.

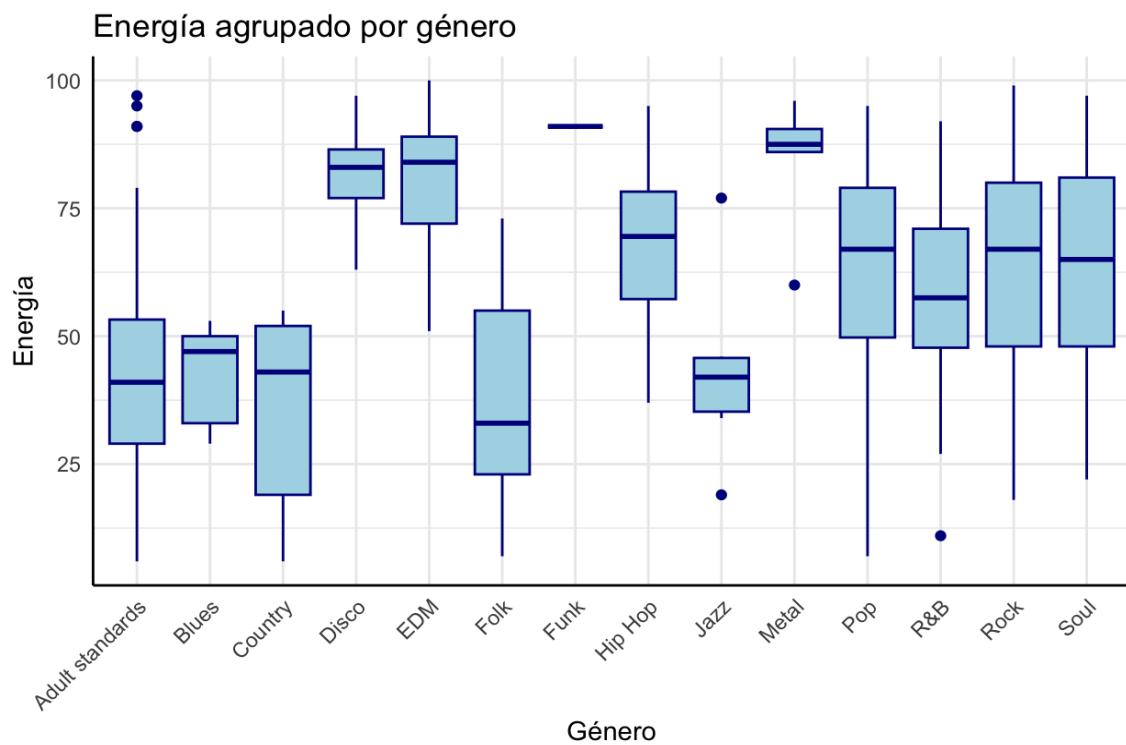


FIGURA 8. Energía por género

Figura 9 . Para la variable que nos determina que tan hablada es una canción, se percibe que para el género del Hip Hop los datos abarcan la mayoría de la escala, lo que da a entender que este género es el que contiene más palabras en sus canciones y esto es cierto en la actualidad. Para el género del Pop existen muchos datos atípicos entonces el habla no determina el género pop. Para los géneros Blues, Country, Jazz, Metal se observa un comportamiento típico de los datos, además de que la concentración de los datos se encuentra en la parte más baja de la escala, es decir estos géneros no tienen muchas palabras en sus canciones. Se puede concluir que si la variable habla tiene un valor mayor a 10 es Hip Hop, para los demás géneros no aplica.

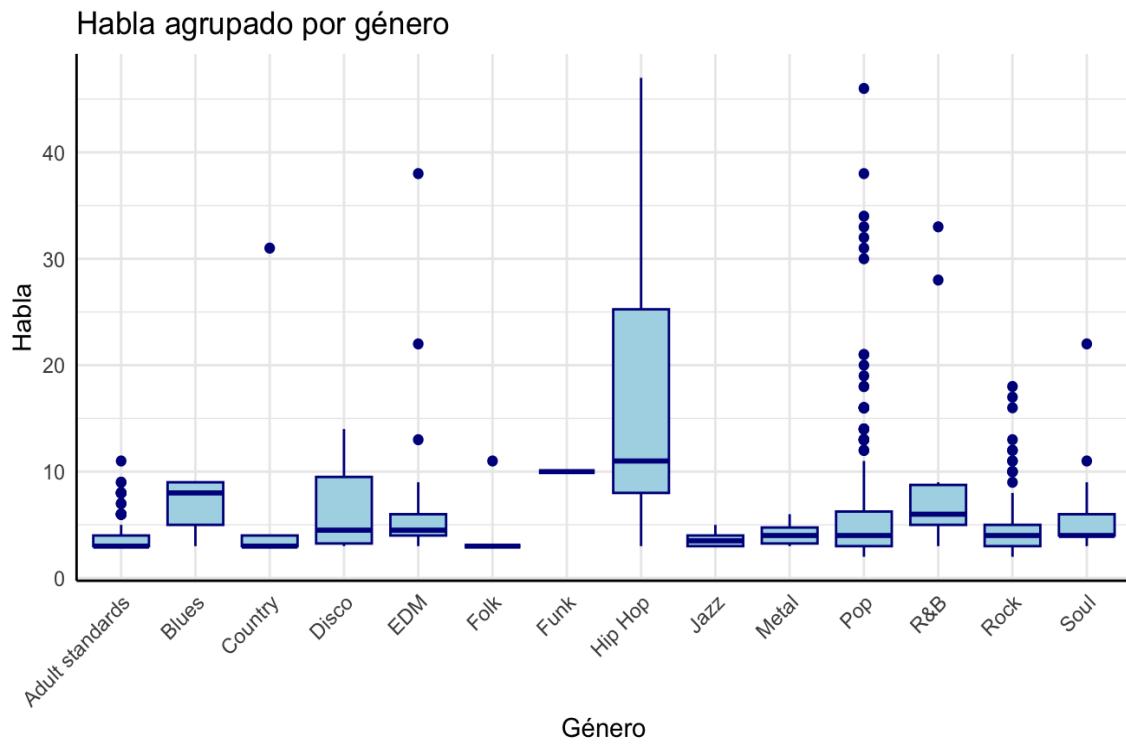


FIGURA 9. Habla por género

Figura 10 . En éste diagrama de la variable En vivo que determina si una canción es grabada en vivo, las cajas tienen concentración en la mayoría de los géneros, exceptuando el género Funk. Se sabe que las canciones en vivo por lo regular son covers o conciertos lo que indica que cualquier género puede grabarse en vivo.

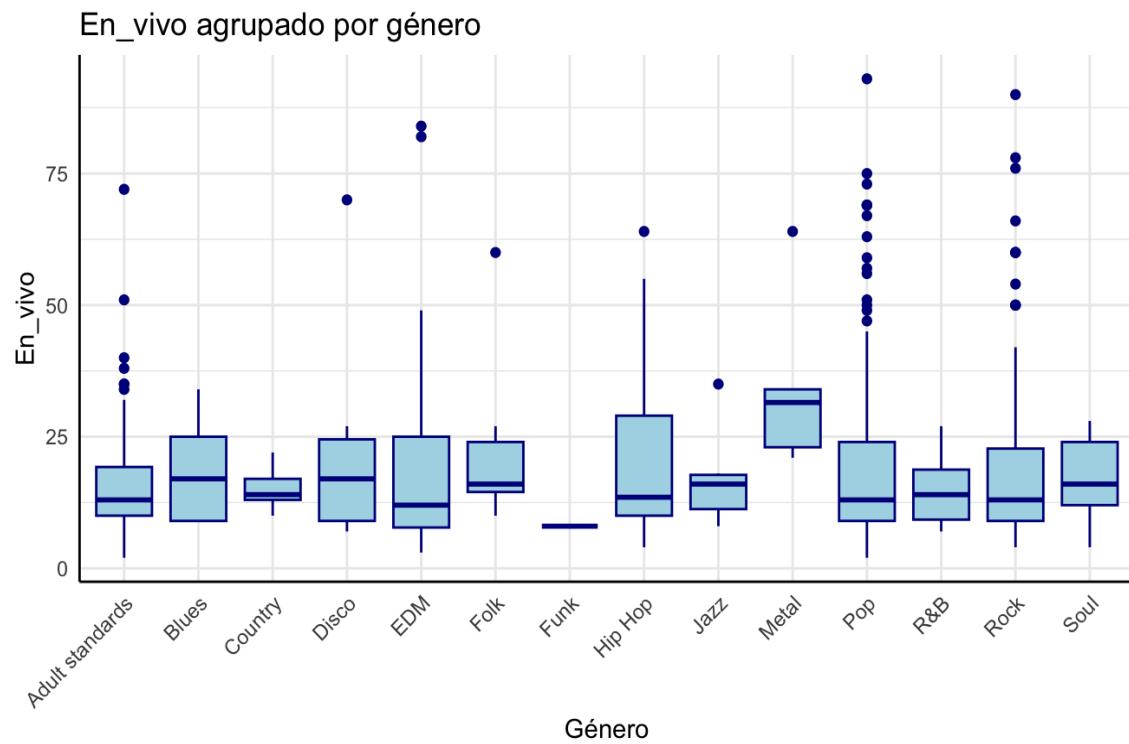


FIGURA 10. En vivo por género

Figura 11. Se percibe que para el género Blues aunque existen datos atípicos se puede concluir que la concentración de los mismos es típica, de cierta manera también para el Folk. En general para todos los géneros se ve que si existe impacto en la positividad. “Según los científicos de la Universidad de McGill en Montreal, Canadá, ésta es la primera vez que se comprueba que este compuesto químico, llamado dopamina, está vinculado a la música. Se sabe que la dopamina se incrementa en respuesta a otros estímulos o actividades de recompensa...”<sup>2</sup>. Por lo que tiene sentido que todos los géneros causen un impacto positivo en el estado de ánimo. mo.

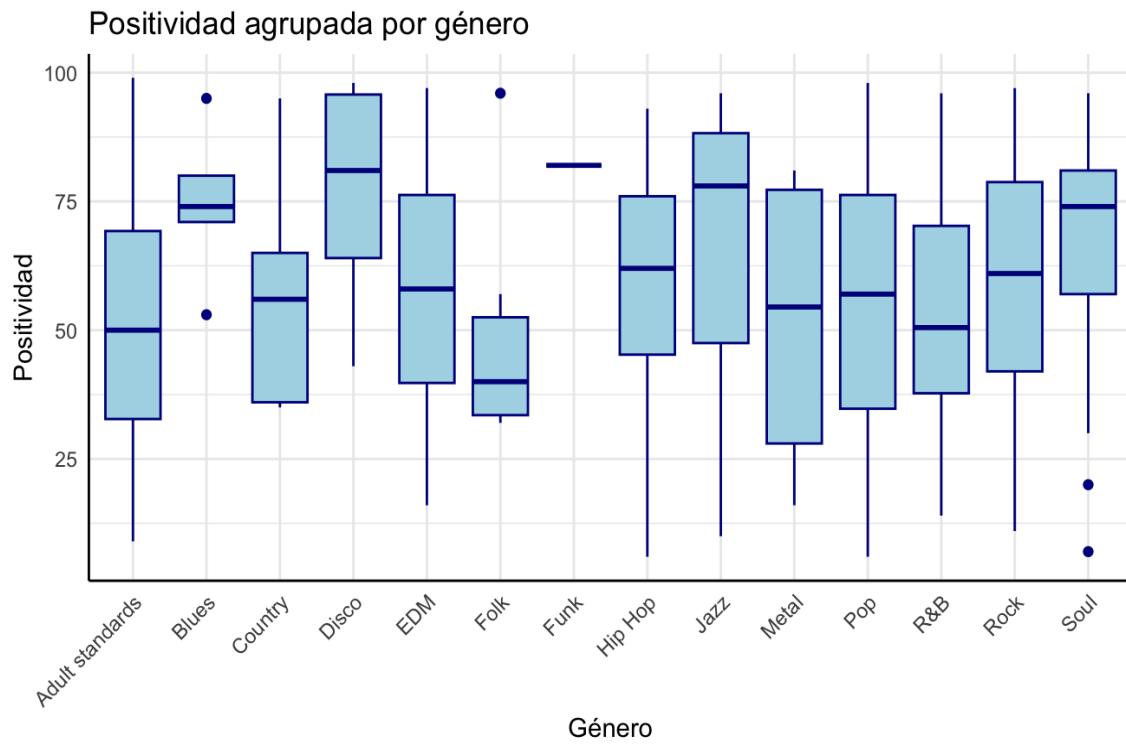


FIGURA 11. Positividad por género

<sup>2</sup>BBC Ciencia (2011). Por qué la música nos hace sentir bien. Recuperado de [https://www.bbc.com/mundo/noticias/2011/01/110113\\_musica\\_cerebro\\_animo\\_](https://www.bbc.com/mundo/noticias/2011/01/110113_musica_cerebro_animo_).

Figura 12 Para éste diagrama de caja se observa que el tempo si está fuertemente relacionado con el género de una canción. Se observa que si existen datos atípicos; para el género del pop se tienen algunos datos atípicos lo que se interpreta que estas canciones populares del género pop no se tienen que comportar igual respecto al tempo, por otro lado el género EDM (Electronic Dance Music) también tiene datos atípicos, donde se ve que el tempo no se comporta de manera típica respecto al tempo. Para los géneros jazz y blues los datos se comportan de manera estándar, es decir son más típicos y los datos se concentran aquí.

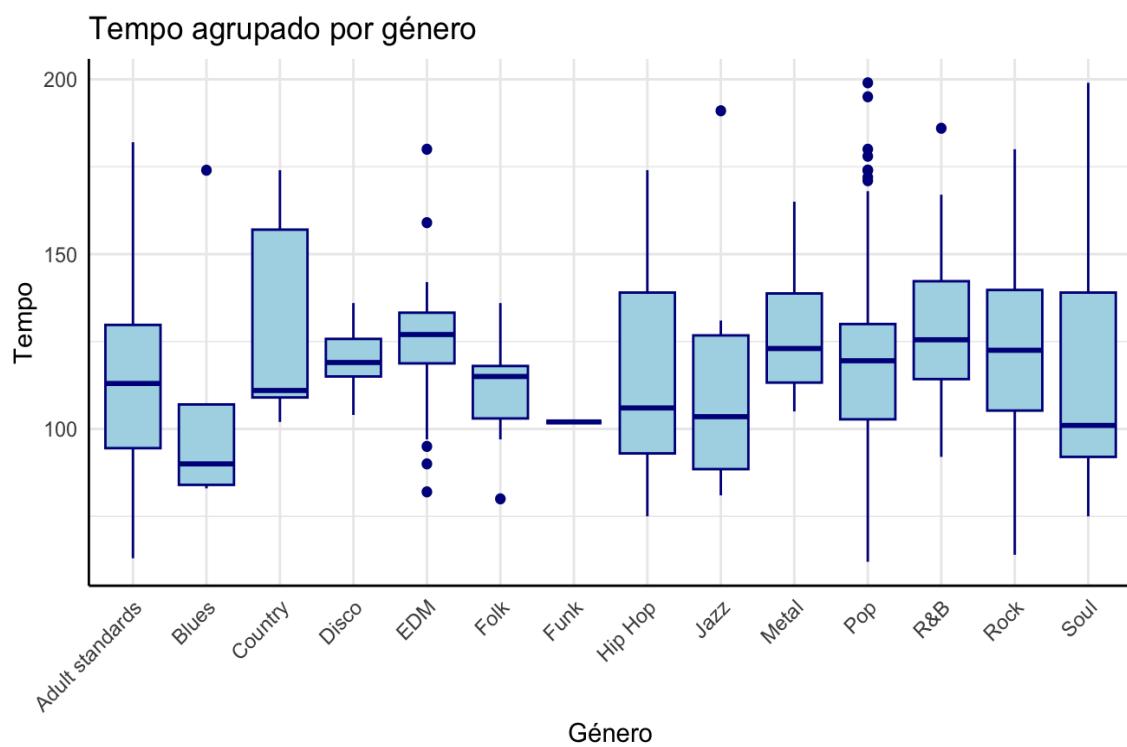


FIGURA 12. Tempo por género

Figura 13. La variable Popularidad respecto al género pop abarca casi toda la escala, es decir, que si quieres que una canción sea popular se debe apostar al género pop.

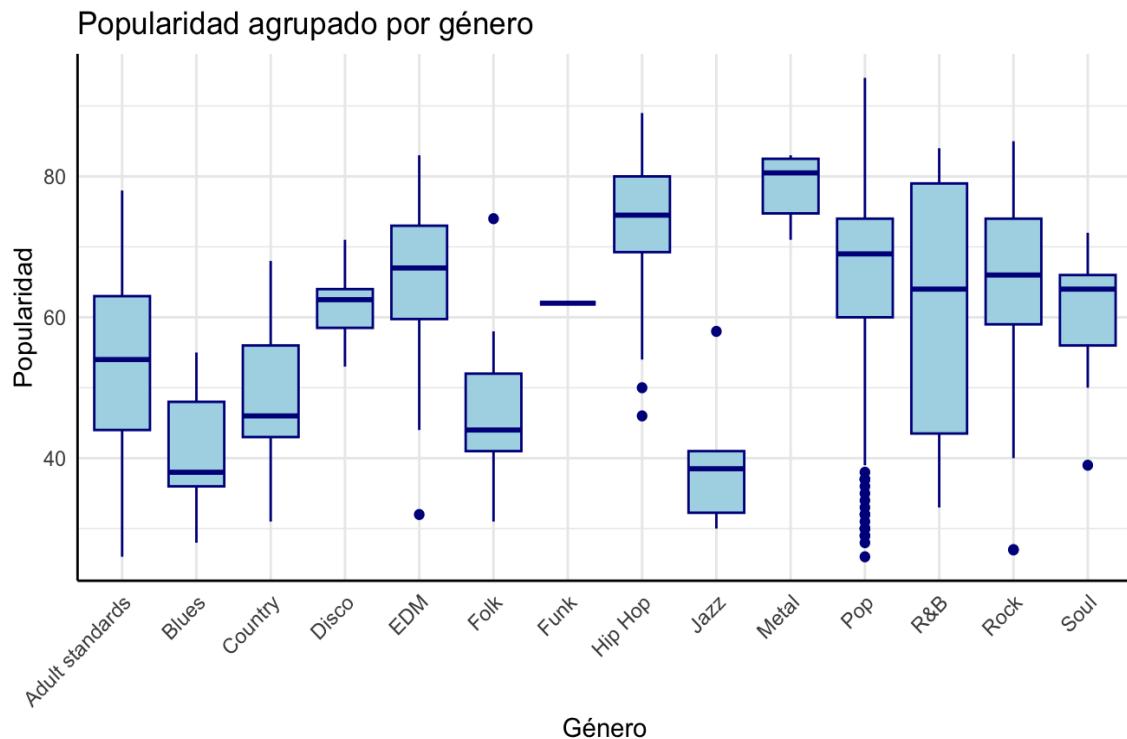


FIGURA 13. Popularidad por género

Por otro lado se observa que el metal está posicionado en la parte más alta de la gráfica, lo que indica mayor popularidad, como se observa en la figura (Ver figura 4) la proporción no es tan grande, de hecho en el género metal se tienen 6 datos, de los cuales son los artistas Bon Jovi, Linkin Park y Nickelback. Pero se observa un dato importante ya que Nickelback ha sido una de las bandas con más desprecio a lo largo de su carrera musical, lo que indica que la mala publicidad es publicidad y esta publicidad se traduce en popularidad. “De hecho, la memeificación de Nickelback en Internet ha hecho que ser fanático de la banda sea casi vergonzoso. Pero con shows agotados en el Madison Square Garden, un puñado de singles y álbumes número uno,

y ser uno de los grupos más vendidos de la década de 2000, no ha afectado su éxito.”<sup>3</sup>.

Existe una estrategia conocida de marketing y es que la mala publicidad, también es publicidad y en este caso para la banda Nickelback pudo ser usada a su favor, ya que se consiguió mucha popularidad, debido a que en los usuarios se puede causar curiosidad de consumir su música, para responder al ¿por qué esta banda es tan criticada? pero a fin de cuanta se cumple el objetivo; que es el consumir su producto, en este caso consumir su música.

Con estos diagramas de caja se puede observar que el género Funk no da mucha información de utilidad, debido a que en la base de datos solo existe un dato y con este no se puede hacer una interpretación acertada. En este diagrama (Ver Figura 4) se ve que la proporción que representa el género Funk es casi nula, por lo que se ha reclasificado en el género del Soul ya que en los diagramas de cajas se encontraba para la mayoría de las variables contenida en esta. El porqué del Funk se ha ido perdiendo a lo largo del tiempo es algo que se vive en la actualidad, debido a que su auge fue hace muchos años atrás. “Los acordes de séptima menor fueron introducidos a través del funk, del soul y la música disco en los años 70... Eso no causó una revolución, pero estos acordes no estaban antes y desde entonces no han desaparecido.”<sup>4</sup>.

---

<sup>3</sup>Radio CBC (2020). THow Nickelback became the internet’s most-hated band. Recuperado de <http://tinyurl.com/ycrxwmvk>

<sup>4</sup>BBC Mundo (2015). Las 3 grandes revoluciones de la música pop, según la ciencia. Recuperado de <http://tinyurl.com/ywdkcas7>

Como última parte de la sección se presentan los diagramas de correlación para cada género:

Figura 14 . Se observa que para la variable Popularidad no hay alguna relación fuerte para el género Soul. Se puede decir que el que una canción sea Bailable, tenga impacto en la Positividad, la Duración, el Tempo y el Habla no está relacionados con que sea popular, es específico para este género.

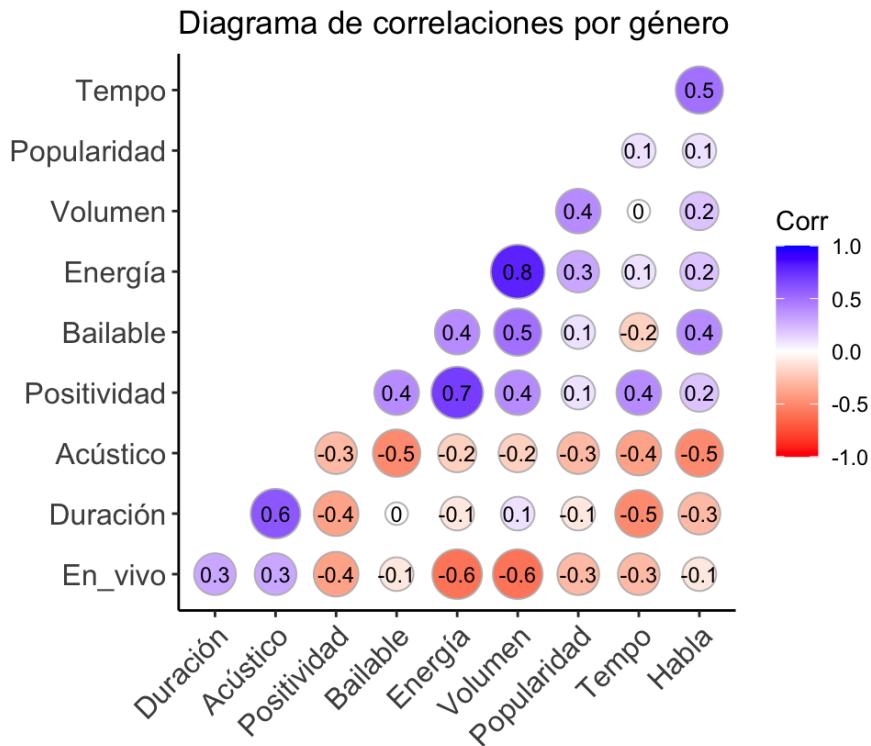


FIGURA 14. Correlaciones de las variables del género Soul

Figura 15 . Para el género Rock se percibe que la variable Popularidad no está relacionada significativamente con ninguna de las que hay.

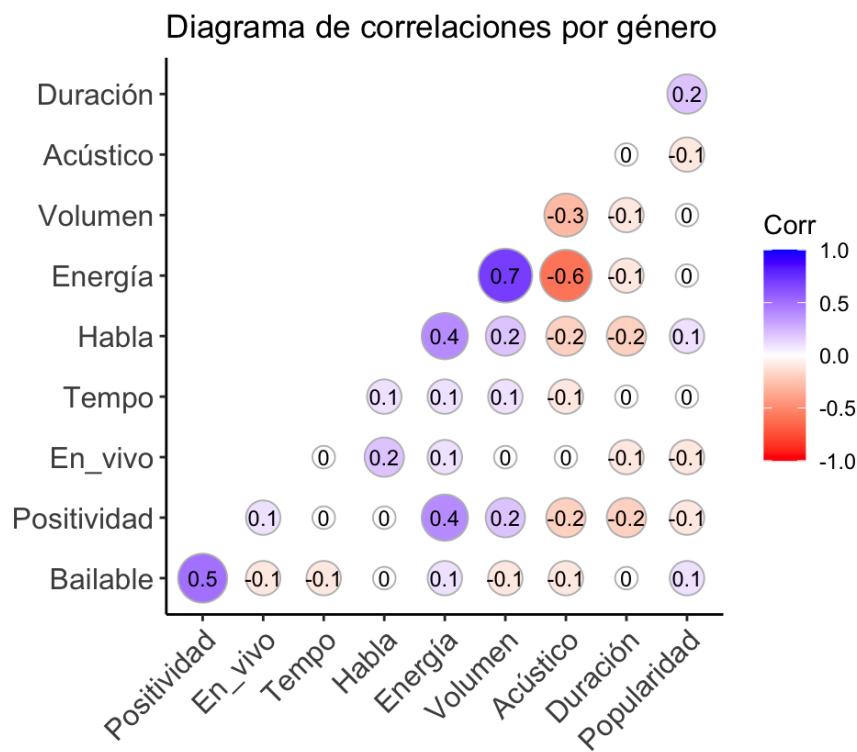


FIGURA 15. Correlaciones de las variables del género Rock

Figura 16 . Para el género Pop se tiene la correlación más fuerte con la variable Acústico, esta relación es negativa lo que nos indica que si es menos acústica será más popular, resultado que es cierto en la actualidad, debido a que la música más popular no es acústica en la mayoría de los casos. Otra de las más fuertes es la de Volumen esta relación es positiva lo que indica que si el volumen incrementa la popularidad también.

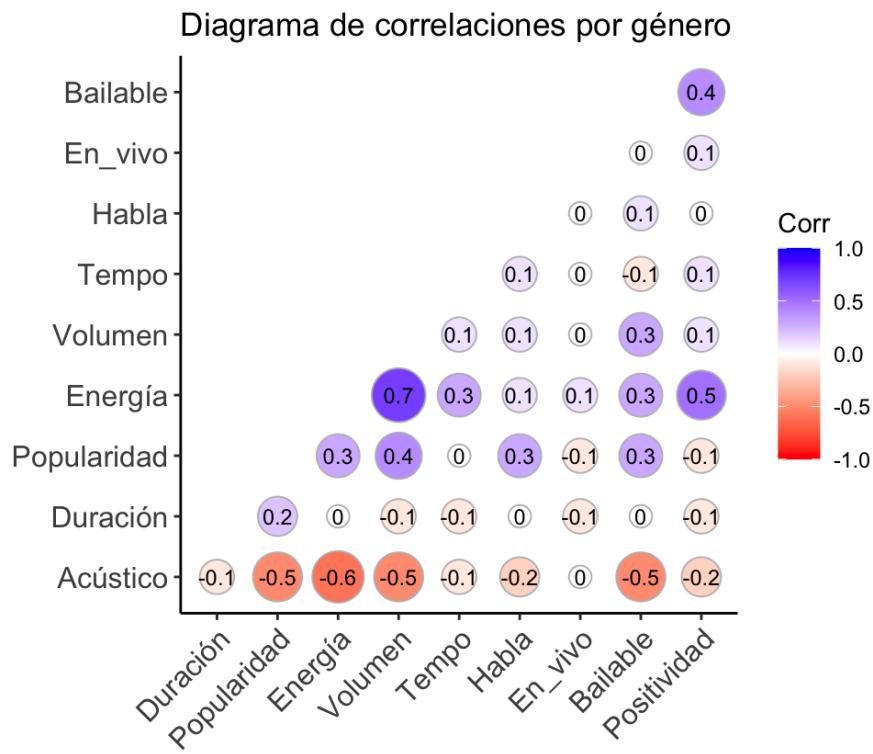


FIGURA 16. Correlaciones de las variables del género Pop

Figura 17 . Para el género Metal se tienen correlaciones más fuertes; la variable Popularidad está correlacionada negativamente con el Tempo lo que dice que si el tempo disminuye la popularidad incrementa al igual que la variable Duración que indica que si dura menos es más popular. La variable Popularidad se relaciona positivamente con la Energía y el Habla, es decir que, si la energía incrementa la popularidad también, lo mismo para la cantidad de palabras en la canción.

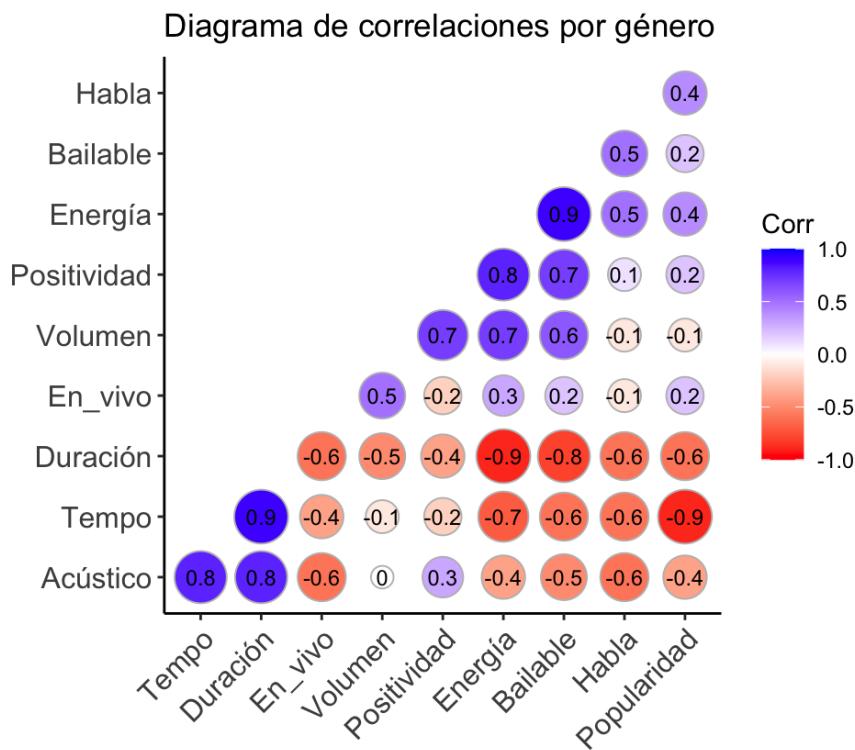


FIGURA 17. Correlaciones de las variables del género Metal

Figura 18 . En el género Jazz la popularidad se relaciona negativamente con que la variable En vivo es decir con canciones que estén en vivo, se podría decir que si no es en vivo es más popular, al igual que para las variables del Volumen, Bailable, Habla y con el impacto en la Positividad. Por otro lado, la variable Popularidad está relacionada positivamente con el Tempo, lo que indica que si el tempo incrementa la popularidad también.

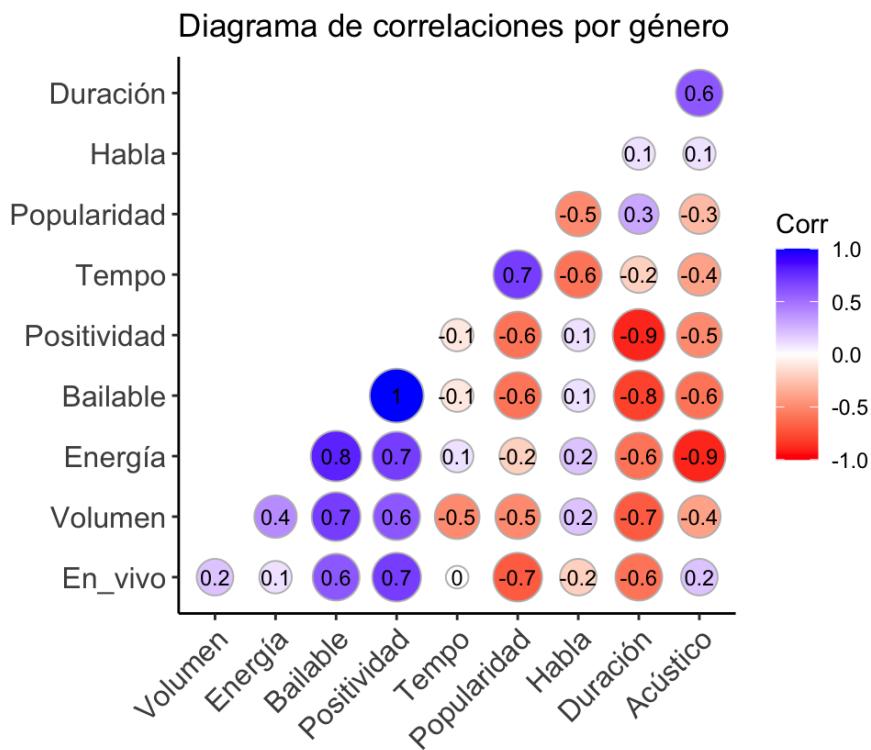


FIGURA 18. Correlaciones de las variables del género Jazz

Figura 19 . Existe una correlación negativa con la variable Popularidad y la Duración, lo que nos indica que en las canciones del género Hip Hop que si duran menos entonces serán más populares, otras relaciones negativas son con las variables Habla y Energía; indica que si contiene menos palabras es más popular y si es menos energética también es más popular.

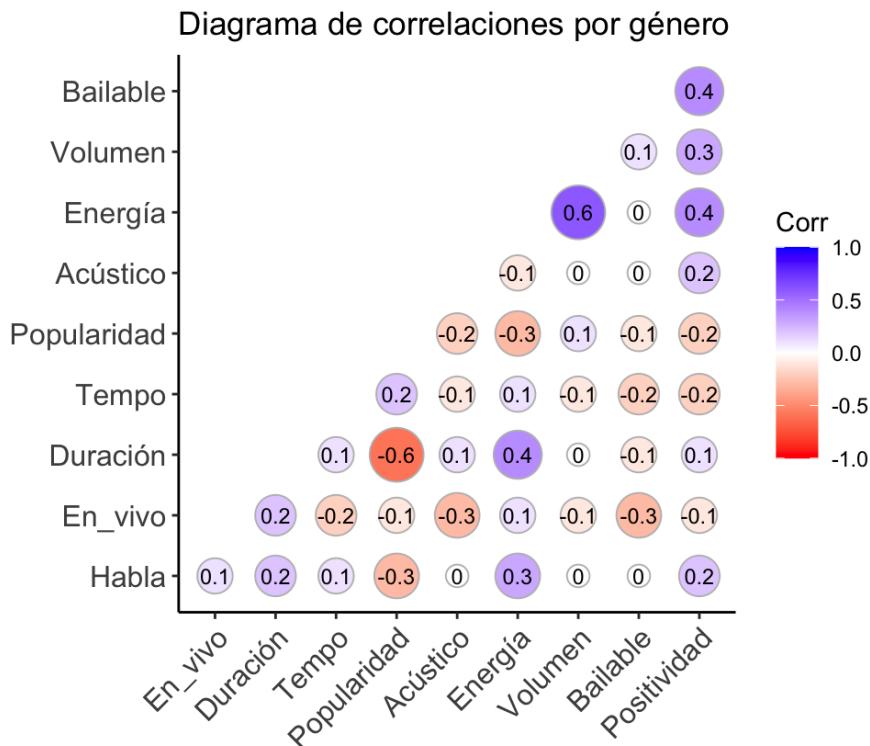


FIGURA 19. Correlaciones de las variables del género Hip Hop

Figura 20 . Se puede identificar que para la variable Popularidad existe una relación negativa con la variable Bailable, es decir que si la canción es menos bailable entonces es más popular, al igual que con las variables, Energía, Duración y Positividad. Aunado a esto, existe una relación positiva de la variable Popularidad con la variable Volumen, lo que indica que si el volumen incrementa la popularidad también.

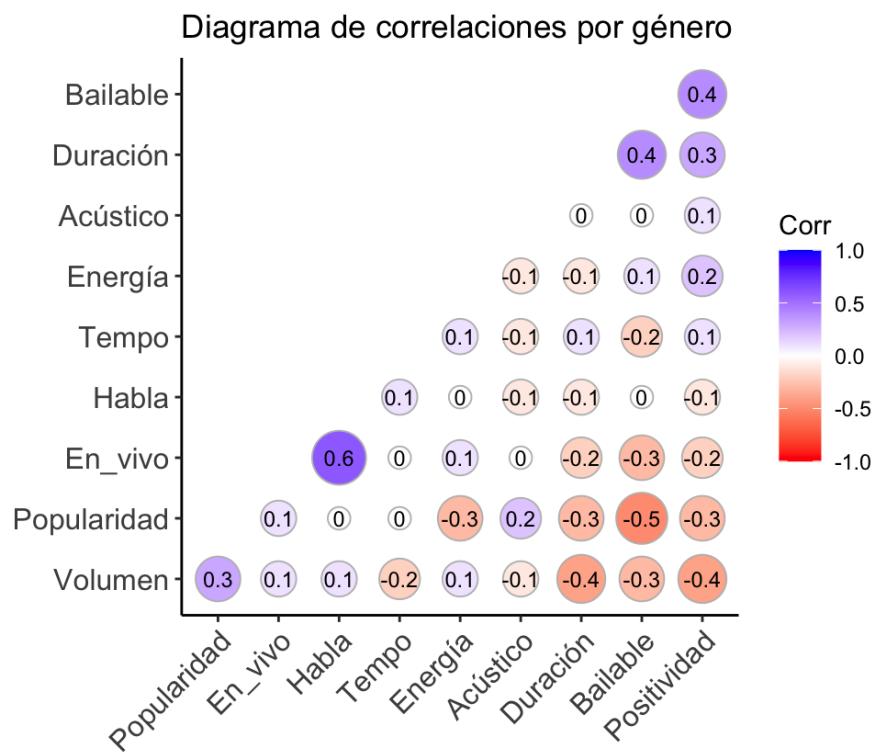


FIGURA 20. Correlaciones de las variables del género EDM

Figura 21 . Para el género disco se aprecia que la variable Popularidad está relacionada positivamente con la Energía, la Duración, el Volumen y el Habla, indica que si alguna de estas variables incrementa la popularidad también. Por otro lado, para la música disco se interpreta que no está relacionada la popularidad con que la canción sea grabada en vivo, sea bailable o esté determinada por la variable Tempo. Existe una correlación negativa de la variable Popularidad con Acústico, lo que indica que si la canción no es acústica es más popular.

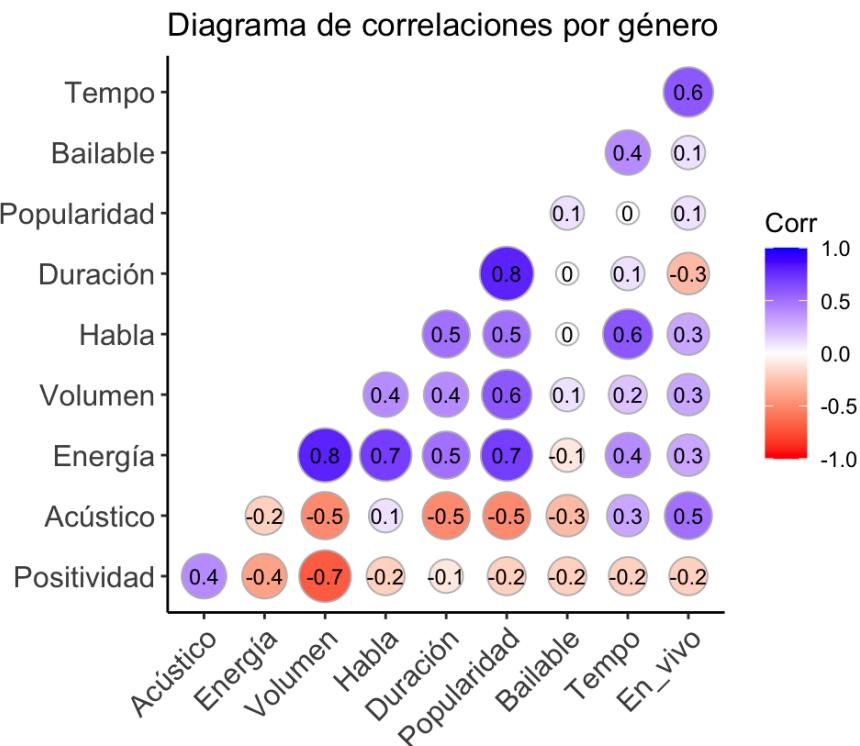


FIGURA 21. Correlaciones de las variables del género Disco

Figura 22 . En el género R&B la variable Popularidad está fuertemente relacionada negativamente con la variable Acústico, es decir que si la canción no es acústica entonces es popular. Seguidamente hay relaciones positivas respecto a la variable popularidad con las variables; Bailable, Habla y Volumen, lo que indica que si alguna de estas incrementa la popularidad también.

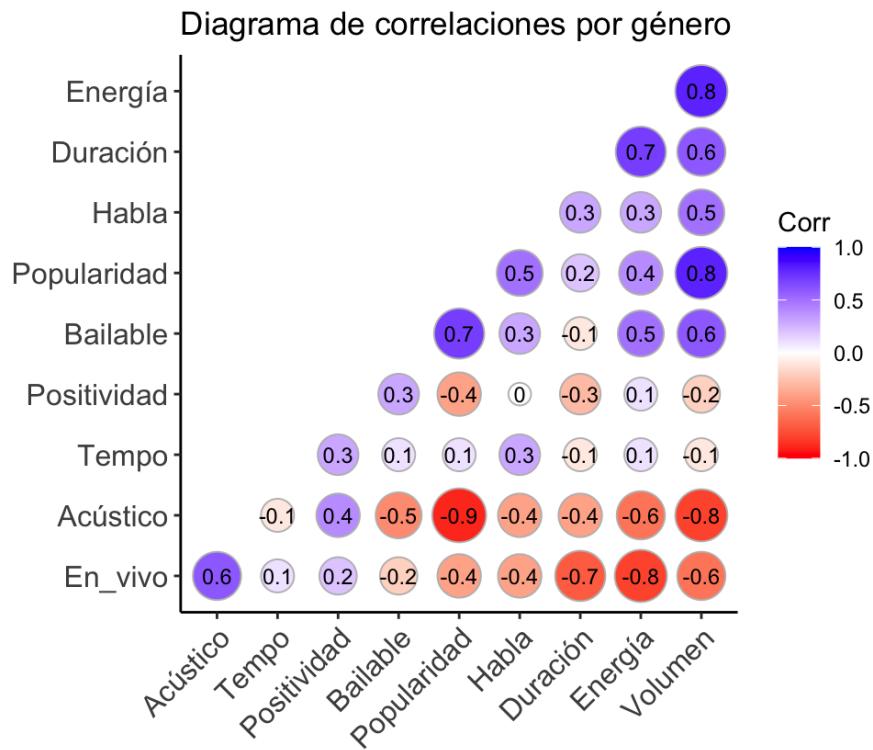


FIGURA 22. Correlaciones de las variables del género R&B

## Capítulo 4

### Aplicación Modelos LDA y árboles

En éste capítulo se pretende mostrar las aplicaciones de los distintos modelos estadísticos; a través, del recurso del lenguaje de programación interpretado llamado R, que se usan para predecir la popularidad de una canción. Estos algoritmos que sirven para predecir el comportamiento de la popularidad de una canción. Utiliza las diversas variables de la base de datos como entrada y proporciona un valor predictivo como salida, para este caso determina si la canción es popular o no. Debido a que se empleó aprendizaje supervisado: Es una técnica para deducir alguna relación mediante datos de entrenamiento, los cuales se dividen como test y train para calibrar el modelo.

Para evaluar el desempeño de ambos modelos se usa una matriz de confusión. Una matriz de confusión es una herramienta que permite observar el desempeño del algoritmo en uso. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. El principal punto de la matriz de confusión es que, facilita ver si el sistema está clasificando de manera correcta las clases o que de alguna manera intuitiva que tanta confusión tiene el modelo.

Se hace una aplicación de los modelos para cada género de la base de datos con la que se cuenta. Recordamos las variables que estos modelos usan y son las siguientes; Tempo, Energía, Bailable, Volumen, En vivo, Positividad, Duración, Acústico, Habla y Popularidad. Además de contar con la variable Popularidad1, ésta variable es la variable de clasificación en dónde se indica si es popular o no con ayuda de ciertos parámetros estadísticos de decisión que se toman de la variable Popularidad.

Los géneros con los que la base de datos cuenta son los siguientes; Adult standards, Blues, Country, Disco, EDM, Folk, Funk, Hip Hop, Jazz, Metal, Pop, R&B, Rock, Soul. Debido a que no para todos los géneros se cuenta con información suficiente. La aplicación de los modelos solo se hará para los géneros que contienen la suficiente información para ser útiles, ya que de otra manera se pudiese dar un resultado erróneo acerca de la clasificación. Seguido de esto los géneros que tiene

suficiente información resulta que son los mismos que se encuentran dentro de los 5 géneros favoritos en el mundo, esto se vio en el capítulo 3 Ver figura 1.

### 1. Análisis discriminante lineal (LDA)

Para la aplicación del modelo LDA se usa una librería en el lenguaje de programación R, ésta se llama library(MASS), en la cuál se encuentra una función lda() con ésta se obtiene el cálculo de la función discriminante. Posteriormente se hace una evaluación de los errores de clasificación a través de una matriz de confusión.

A continuación, se muestran las predicciones obtenidas de aplicar el modelo de LDA a los datos de entrenamiento con train; se entrenó el modelo con el 70 % de información y con test usamos datos que aún no han pasado por el modelo, es decir el 30 % restante.

- Pop

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 37 No populares está prediciendo que 25 de ellos no son populares y 12 populares y de los datos reales de popular se tienen 38 de los cuales está prediciendo que 7 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 25.33 %.

Observación \ Predicción			Total
	No popular	Popular	
No popular	25	12	37
Popular	7	31	38

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 86 No populares está prediciendo que 59 de ellos no son populares y 27 populares y de los datos reales de popular se tienen 87 de los cuales está prediciendo que 22 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 28.32 %.

Predicción Observación	No popular	Popular	Total
No popular	59	27	86
Popular	22	65	87

#### ■ Rock

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 14 No populares está prediciendo que 4 de ellos no son populares y 10 populares y de los datos reales de popular se tienen 35 de los cuales está prediciendo que 3 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 26.53 %.

Predicción Observación	No popular	Popular	Total
No popular	4	10	14
Popular	3	32	35

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 23 No populares está prediciendo que 3 de ellos no son populares y 20 populares y de los datos reales de popular se tienen 90 de los cuales está prediciendo que 1 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 18.58 %.

Predicción Observación	No popular	Popular	Total
No popular	3	1	20
Popular	1	98	90

■ **Adult Standards**

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 10 No populares está prediciendo que 6 de ellos no son populares y 4 populares y de los datos reales de popular se tienen 20 de los cuales está prediciendo que 1 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 16.67 %.

Observación \ Predicción			
	No popular	Popular	Total
No popular	6	4	10
Popular	1	19	20

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 33 No populares está prediciendo que 24 de ellos no son populares y 9 populares y de los datos reales de popular se tienen 37 de los cuales está prediciendo que 10 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 27.14 %.

Observación \ Predicción			
	No popular	Popular	Total
No popular	24	9	33
Popular	10	27	37

- **Electronic dance music (EDM)**

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cuál se tiene que de 7 No populares está prediciendo que 6 de ellos no son populares y 1 populares y de los datos reales de popular se tienen 9 de los cuales está prediciendo que 1 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 12.5 %.

Predicción \ Observación	No popular	Popular	Total
No popular	6	1	7
Popular	1	8	9

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cuál se tiene que de 9 No populares está prediciendo que 5 de ellos no son populares y 4 populares y de los datos reales de popular se tienen 27 de los cuales está prediciendo que 1 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 13.88 %.

Predicción \ Observación	No popular	Popular	Total
No popular	5	4	9
Popular	1	26	27

■ **Hip Hop**

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 7 No populares está prediciendo que 7 de ellos no son populares. De los datos reales de popular se tienen 6 de los cuales está prediciendo que 1 de ellos no son populares y 5 populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 7.7%.

Predicción \ Observación	No popular	Popular	Total
No popular	7	0	7
Popular	1	5	6

**Train** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 5 No populares está prediciendo que 4 de ellos no son populares y 1 popular. De los datos reales de popular se tienen 24 de los cuales está prediciendo que 2 de ellos no son populares y 22 populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 10.3%.

Predicción \ Observación	No popular	Popular	Total
No popular	4	1	5
Popular	2	22	24

## 2. Árboles de decisión con clasificación

Para la aplicación del modelo de árboles de decisión con clasificación se usan las siguientes librerías que se encuentran en la herramienta de programación R, library(tidyverse), library(rpart), library(rpart.plot) y library(caret), en la cual se encuentra una función principal que es rpart(), la cual da como resultado el árbol de decisión, posteriormente se hace una gráfica de cómo es que se el diagrama del árbol de decisión para cada género y finalmente se hace una evaluación de los errores de clasificación a través de una matriz de confusión.

A continuación, se muestran las predicciones obtenidas de aplicar el modelo de árboles de decisión con clasificación a los datos de entrenamiento con train; se entrenó el modelo con el 70 % de información y con test usamos datos que aún no han pasado por el modelo, es decir el 30 % restante.

- **Pop**

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 24 No populares está prediciendo que 21 de ellos no son populares y 3 populares y de los datos reales de popular se tienen 50 de los cuales está prediciendo que 11 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 18.92 %.

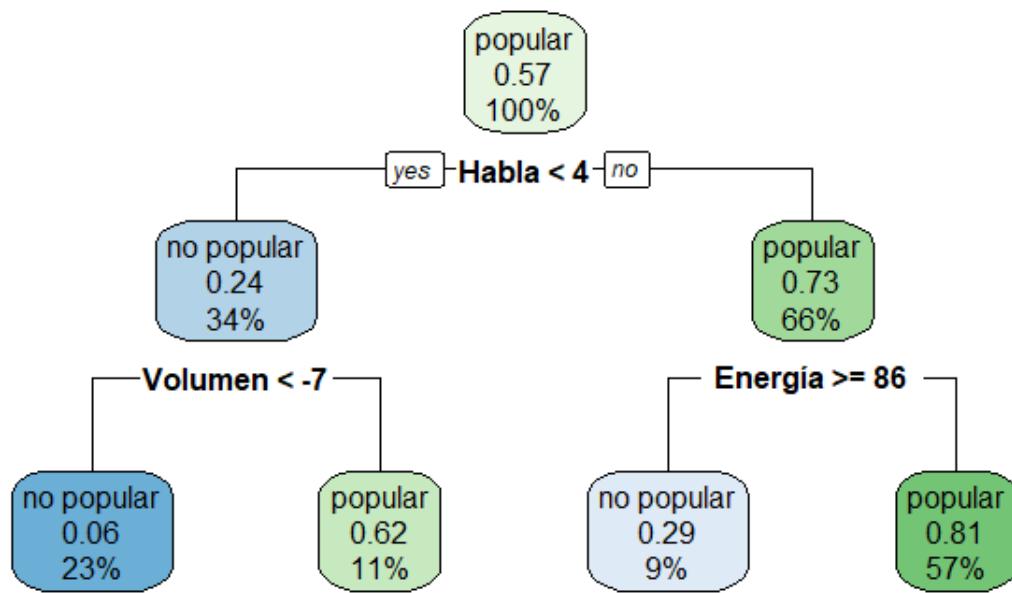


FIGURA 1. Diagrama de árbol género pop

Observación \ Predicción			
	No popular	Popular	Total
No popular	21	3	24
Popular	11	39	50

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 99 No populares está prediciendo que 78 de ellos no son populares y 21 populares y de los datos reales de popular se tienen 75 de los cuales está prediciendo que 13 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 19.54 %.

Observación \ Predicción			
	No popular	Popular	Total
No popular	78	21	99
Popular	13	62	75

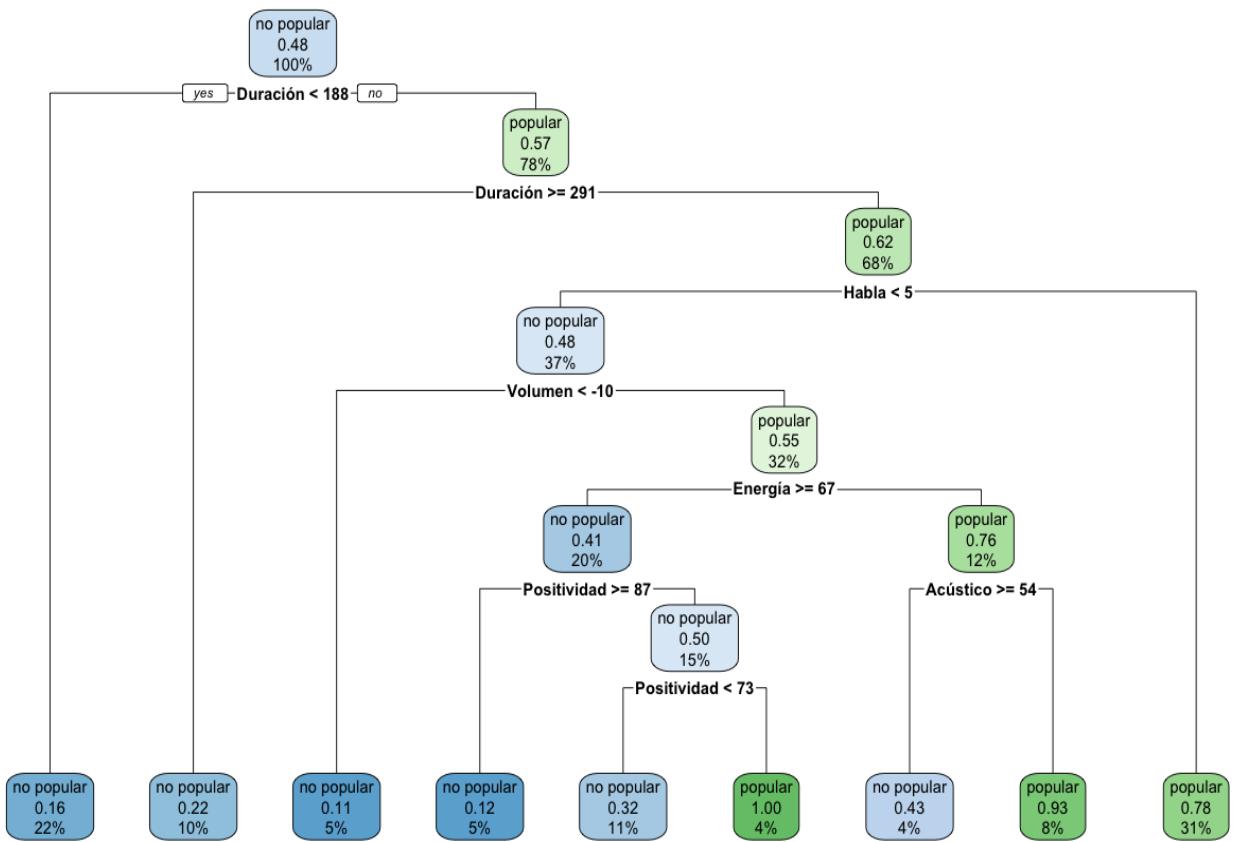


FIGURA 2. Diagrama de árbol género pop

- Rock

**Test** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 12 No populares está prediciendo que 8 de ellos no son populares y 4 populares. De los datos reales de popular se tienen 61 de los cuales está prediciendo que 53 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 16.44 %.

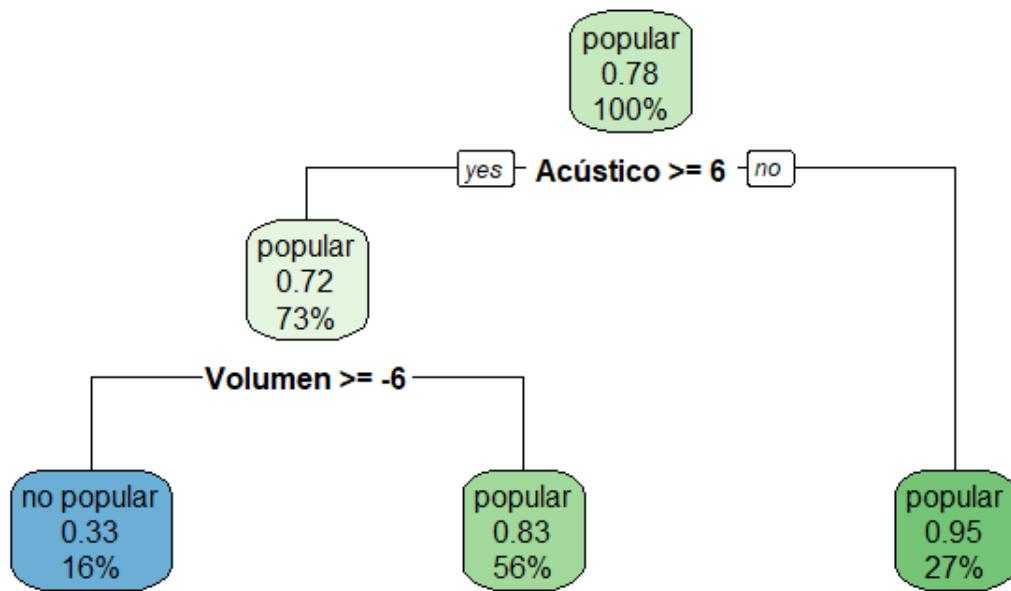


FIGURA 3. Diagrama de árbol género Rock

Observación \ Predicción	No popular	Popular	Total
No popular	8	4	12
Popular	8	53	61

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 29 No populares está prediciendo que 23 de ellos no son populares y 6 populares y de los datos reales de popular se tienen 141 de los cuales está prediciendo que 24 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 17.65 %.

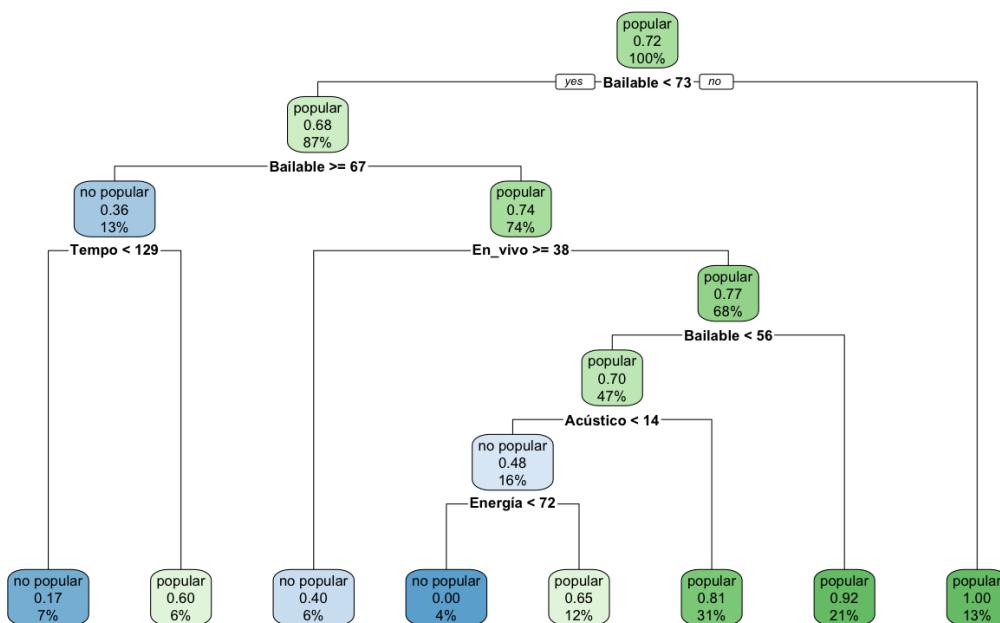


FIGURA 4. Diagrama de árbol género Rock

Observación \ Predicción	No popular	Popular	Total
No popular	23	6	29
Popular	24	117	141

### ■ Adult Standars

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de un total de 16 No populares está prediciendo que 10 de ellos no son populares y 6 populares y de los datos reales de popular se tienen 14 de los cuales está prediciendo que 1 de ellos no es popular. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 23.33 %.

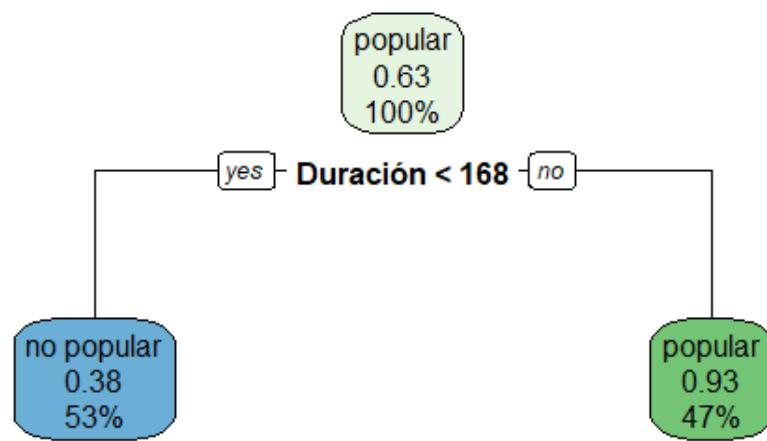


FIGURA 5. Diagrama de árbol género Adult Standars

Observación \ Predicción	No popular	Popular	Total
No popular	10	6	16
Popular	1	13	14

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 31 No populares está prediciendo que 25 de ellos no son populares y 6 populares y de los datos reales de popular se tienen 39 de los cuales está prediciendo que 7 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 18.57%.

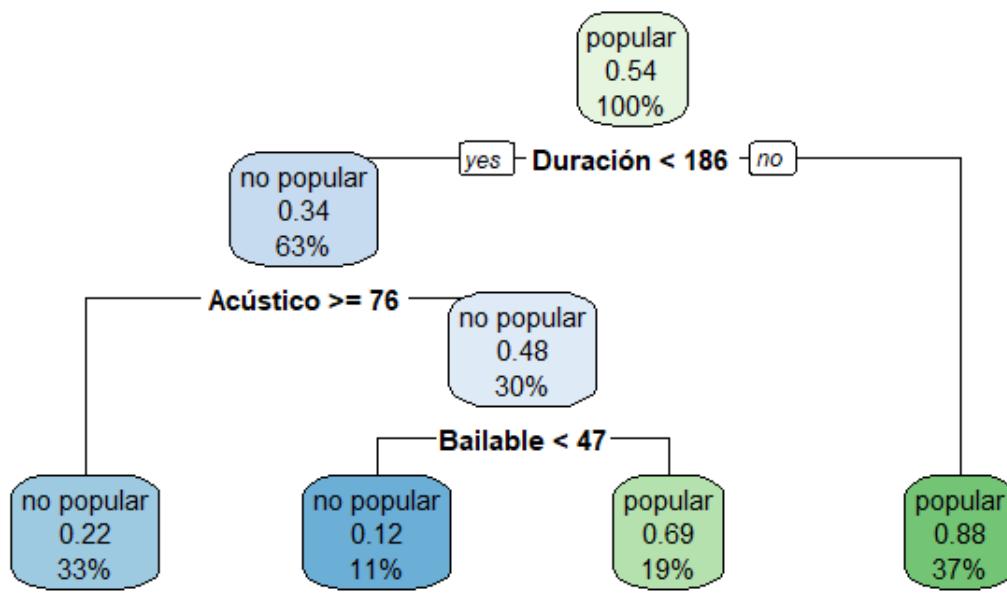


FIGURA 6. Diagrama de árbol género Adult Standars

Observación \ Predicción	No popular	Popular	Total
Observación			
No popular	25	6	31
Popular	7	37	39

- **Electronic dance music (EDM)**

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 7 No populares está prediciendo que 4 de ellos no son populares y 3 populares y de los datos reales de popular se tienen 17 de los cuales está prediciendo que 1 de ellos no es popular. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 16.67 %.

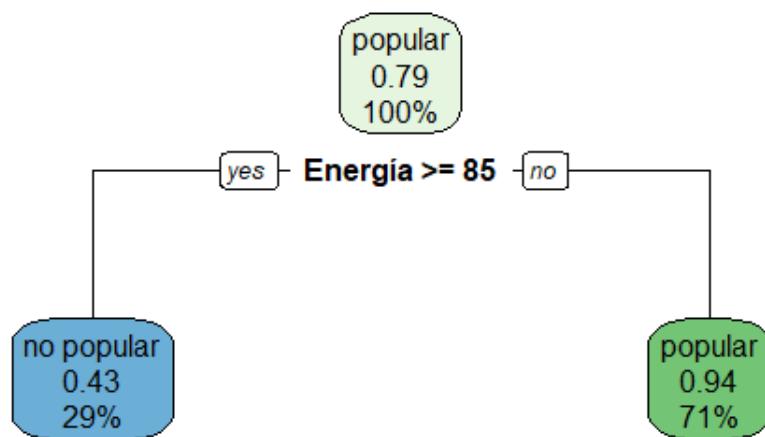


FIGURA 7. Diagrama de árbol género EDM

Observación \ Predicción			Total
	No popular	Popular	
No popular	4	3	7
Popular	1	16	17

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 27 No populares está prediciendo que 19 de ellos no son populares y 8 populares y de los datos reales de popular se tienen 27 de los cuales está prediciendo que 3 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 20.37 %.

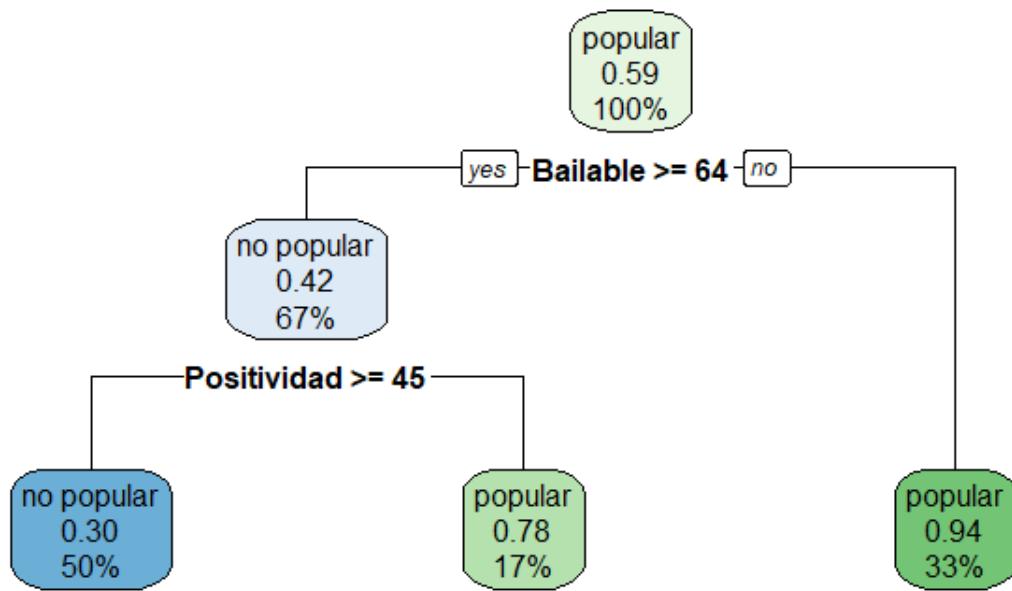


FIGURA 8. Diagrama de árbol género EDM

Observación \ Predicción	No popular	Popular	Total
Observación			
No popular	19	8	27
Popular	3	24	27

### ■ Hip Hop

**Test** Resultado de aplicar el modelo a la muestra de test se obtiene la siguiente matriz de confusión, en la cual se tiene que de 12 No populares está prediciendo que 8 de ellos no son populares y 4 populares y de los datos reales de popular se tienen 14 de los cuales está prediciendo que ninguno de ellos no es popular. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 15.38 %.

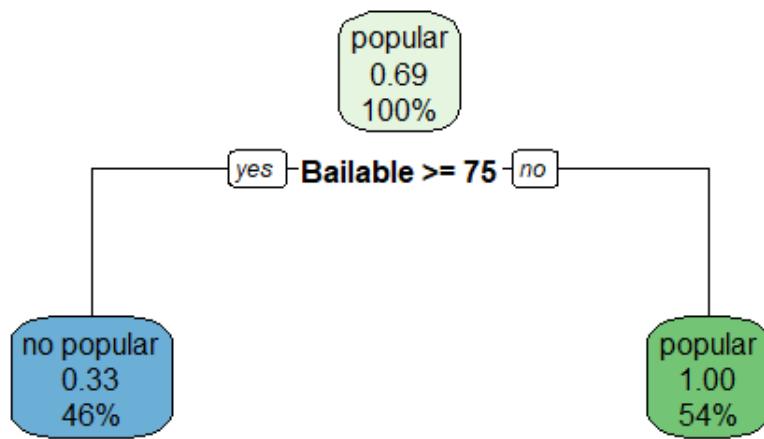


FIGURA 9. Diagrama de árbol género Hip hop

Observación \ Predicción	No popular	Popular	Total
No popular	8	4	12
Popular	0	14	14

**Train** Resultado de aplicar el modelo a la muestra de train se obtiene la siguiente matriz de confusión, en la cual se tiene que de 30 No populares está prediciendo que 17 de ellos no son populares y 3 populares y de los datos reales de popular se tienen 38 de los cuales está prediciendo que 2 de ellos no son populares. Se aplica una métrica en el lenguaje de programación R, para medir el error y se obtiene que es del 8.62 %.

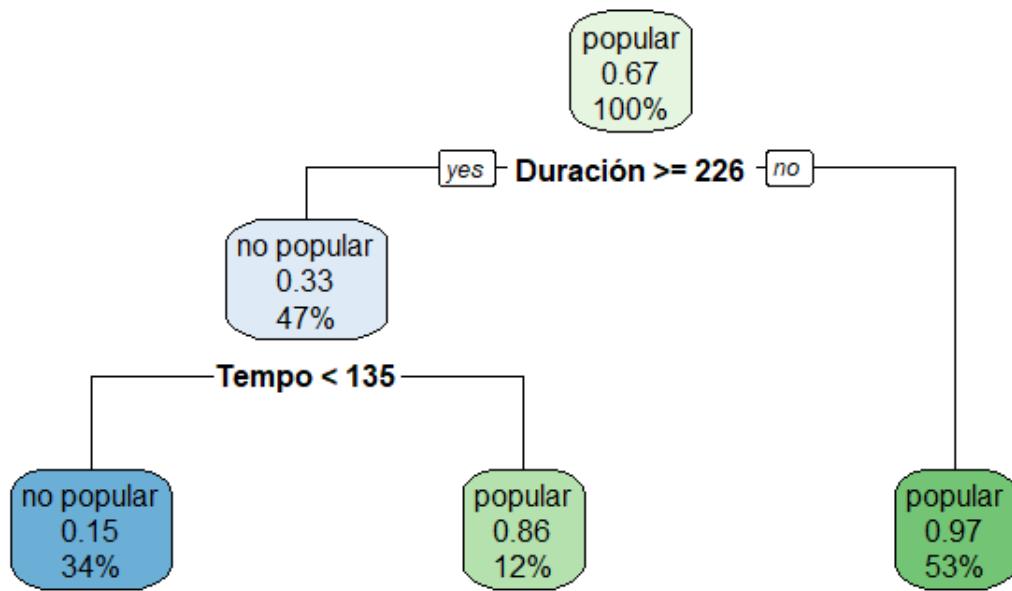


FIGURA 10. Diagrama de árbol género Hip Hop

Observación \ Predicción	No popular	Popular	Total
Observación			
No popular	17	3	30
Popular	2	36	38



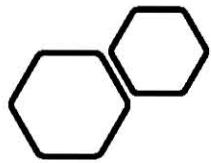
## Capítulo 5

### **Informe ejecutivo**

Un informe ejecutivo tiene el objetivo de mostrar un resumen del proyecto a las personas que pueden decidir sobre su financiamiento, los directivos, es decir, es una presentación de un proyecto comercial. Este informe debe ser conciso y claro para informar sobre los principales aspectos de este proyecto por cual no profundiza en tecnicismos.

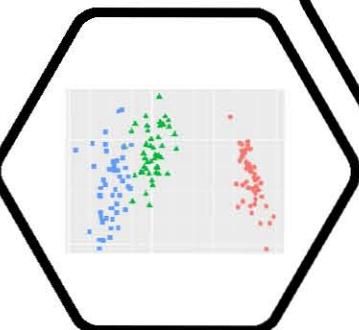
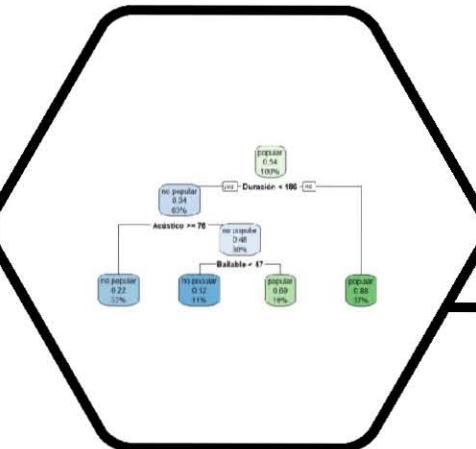
Se busca que la presentación, este enfocada más a resultados que a metodología. Este informe resulta ser de gran relevancia porque permite a los representantes tomar decisiones que puedan ser beneficiosas para la empresa. Por lo que se presenta de manera resumida los resultados obtenidos a través de todo lo que se trabajó tanto como en la exploración de datos, investigación de los modelos y aplicación de los modelos en la base de datos extraída con la API de Spotify.

A continuación, se muestra el informe ejecutivo de este proyecto de investigación a manera de presentación, en la cual se describe la problemática de este proyecto, se presenta los modelos aplicados a grandes rasgos y también se presentan los porcentajes de asertividad de los modelos y sobre todo las variables que resultan ser de mayor relevancia al determinar si una canción puede ser popular o no, culminando con información porcentual de los ingresos en la industria musical.



## INFORME EJECUTIVO

# Caso de estudio de Ciencia de Datos en la popularidad de las canciones de Spotify



Conocer la popularidad de una canción tiene un gran impacto para las disqueras, cantantes y productores. Debido a que invierten grandes cantidades de dinero en publicidad para la promoción de una canción, con el fin de que sea popular, si una canción se vuelve popular resultará tener mejores ganancias.

Es por ello que se trabaja en este caso de estudio de Ciencia de Datos en el cual se pretende buscar la popularidad de las canciones de la plataforma musical más popular del mundo; Spotify que cuenta con 345 millones de usuarios activos. Spotify tiene una API de la cuál se extrajo una muestra de datos. Se le aplicaron modelos estadísticos para predecir si una canción es popular o no.





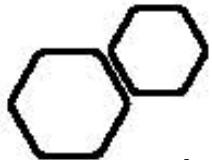
## Variables de la muestra

- **Tempo:** Hace referencia a la velocidad con la que debe ejecutarse una pieza musical.
- **Energía :** Medida perceptiva de intensidad y actividad. Por lo general se sienten rápidas y ruidosas.
- **Bailable:** Capacidad de baile cuanto mayor sea el valor, más fácil será bailar esta canción.
- **Volumen:** Cuanto mayor sea el valor, más fuerte será la canción es decir mide la sonoridad.

## Variables de la muestra

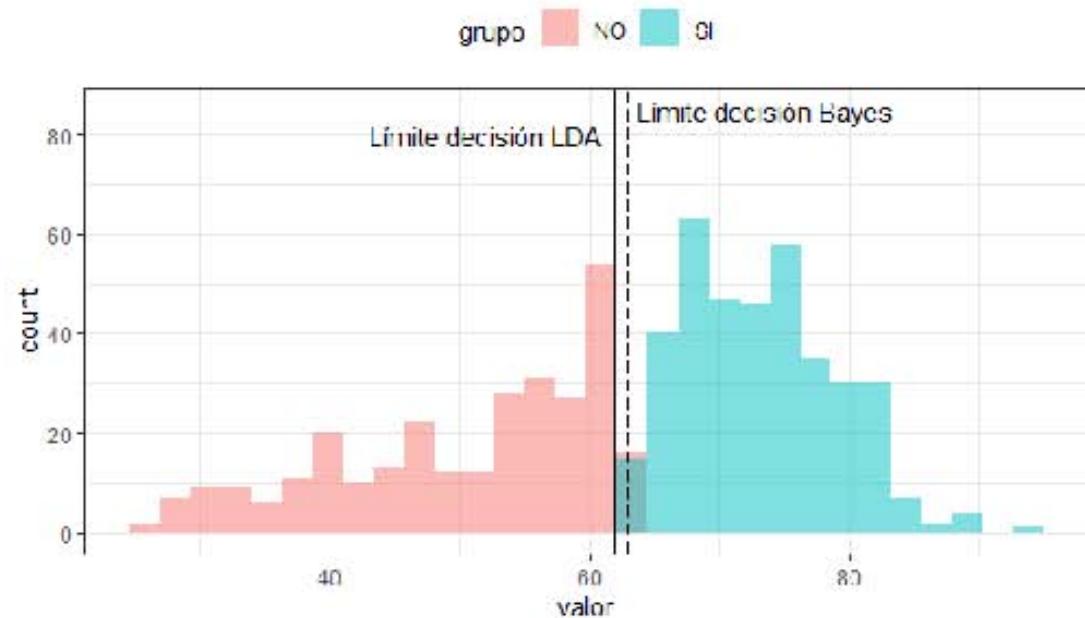
- Positividad : Será el estado de ánimo que transmite la canción.
- Duración : La duración de la canción.
- Acústico : Cuanto mayor sea el valor, más acústica será la canción.
- Habla : Cuanto mayor sea el valor, más palabras hablada contiene la canción.
- Popularidad : Cuanto mayor sea el valor, más popular será la canción.
- En vivo: Determina grabación en vivo





## Análisis discriminante lineal (LDA)

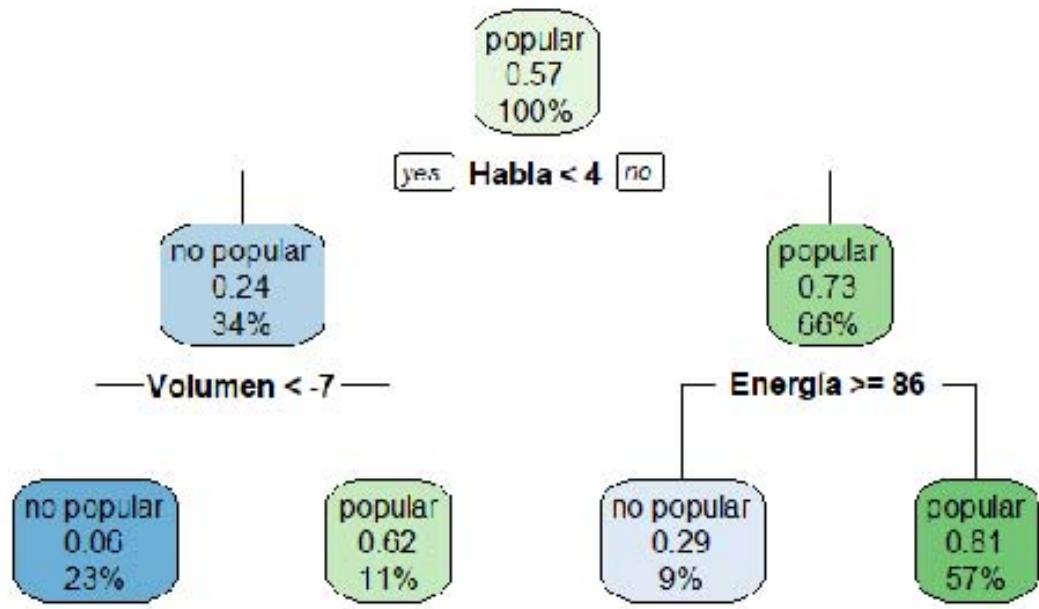
- Clasificar casos en grupos dadas las similitudes entre las variables.
- A través del teorema de Bayes, estima la probabilidad de que una observación, pertenezca a la clasificación popular o no popular.
- Particularmente aquí se observa que con una variable separa lo rosa en No popular y lo azul en Si popular





## Árboles de decisión con clasificación

- Selecciona el mejor atributo utilizando una medida de selección.
- Éste atributo se convierte en nodo de decisión y divide el conjunto en subconjuntos.
- Comienza la construcción del árbol repitiendo este proceso recursivamente, hasta que pase lo siguiente:
  - Las variables pertenecen al mismo valor de atributo.
  - No quedan más atributos.
  - No hay más casos.



# Variables de mayor impacto

- Algunas variables resultan ser más relevantes al determinar que canción podría ser popular. Se muestra con una palomita la variable que requiere cada género.
- Es importante mencionar que en cada variable resultó un rango entre cuanto deben estar esos valores para determinar si es popular o no.

	POP	Rock	Adult Standars	EDM	Hip Hop
Tempo					✓
Energía	✓			✓	
Bailable		✓	✓	✓	✓
Volumen	✓	✓			
Positividad				✓	
Duración			✓		✓
Aúctisco		✓	✓		
Habla	✓				

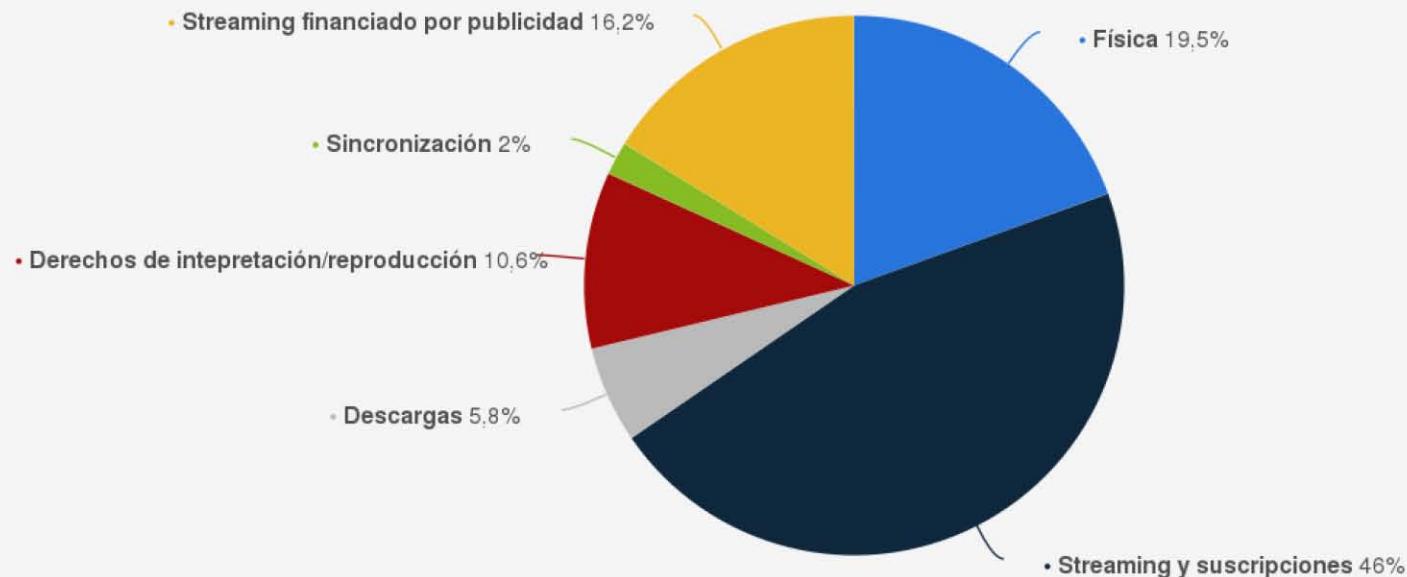
# Asertividad de los modelos

- Se observa que independientemente del género, los dos modelos tienen un poder predictivo del 82%.
- Si se tiene un disco con 10 canciones, al aplicar el modelo en la construcción de las canciones se sabe que al menos 8 de ellas serán populares.
- Si se eligieran las canciones de un disco usando alguno de los dos modelos propuestos y este disco tuviera 10 canciones entonces se esperaría que 8 de 10 canciones en promedio tuvieran éxito.

	LDA	Árboles
POP	73.2%	80.8%
Rock	77.5%	83.0%
Adult Standards	81.0%	79.0%
EDM	87.0%	81.5%
Hip Hop	91.0%	88.0%
Promedio	81.9%	82.5%

Es importante ver que el **46%** de los ingresos los generan las plataformas de streaming. Lo que da una gran oportunidad en el mercado al problema de determinar si una canción puede ser popular o no.

Distribución porcentual de los ingresos generados por la industria de la música grabada a nivel mundial en 2020, por segmento



Fuente  
IFPI  
© Statista 2021

Información adicional:  
Mundial; 2020

## Capítulo 6

# Conclusiones

El objetivo fundamental de ésta tesis era presentar a la ciencia de datos por medio de un caso de estudio haciendo uso de las herramientas adquiridas en la carreara de Actuaría, como lo es, la estadística multivariada, la estadística no paramétrica, programación en R y el conocimiento de la materia basa de datos. En esta tesis no se pretendía presentar la teoría de algún modelo estadístico predictivo en específico, sino presentarlo a modo de aplicación que resultara óptimo, esto quiere decir que el tiempo en el cual se ejecuta el algoritmo fuera rápido y, además, la asertividad fuera mayor al 80 %.

El caso de estudios que se presentó en esta tesis es un caso de estudio de la ciencia de datos para conocer la popularidad de las canciones de Spotify. Se pretendía conocer la popularidad de una canción debido a que tiene un gran impacto económico para las disqueras, cantantes y productores. Ya que invierten grandes cantidades de dinero en publicidad para la promoción de una canción, con el fin de que sea popular. Puesto que si una canción se vuelve popular resultará ser más conocida y con esto tener mejores ganancias mediante la popularidad de dicha canción.

Es por ello por lo que se trabaja en este caso de estudio de Ciencia de Datos, ya que la principal retribución es que se adquiere un beneficio económico. Se buscó la popularidad de las canciones de la plataforma musical más popular del mundo; Spotify que cuenta con 345 millones de usuarios activos. Spotify tiene una API de la cual se extrajo una muestra de datos. Se le aplicaron modelos estadísticos LDA y árboles de clasificación para predecir si una canción es popular o no, obteniendo los siguientes resultados:

Para el modelo LDA primeramente clasificó casos en grupos dadas las similitudes entre las variables. A través del teorema de Bayes, estimó la probabilidad de que una observación, perteneciera a la clasificación popular o no popular. Esta clasificación se hizo por género y se obtuvo un porcentaje de asertividad del 81.9 % que si se redondea queda en 82 %.

En el modelo de árboles se seleccionó el mejor atributo utilizando una medida de selección. Éste atributo se convierte en nodo de decisión y divide el conjunto en subconjuntos. Comienza la construcción del árbol repitiendo este proceso recursivamente, hasta que ocurrió lo siguiente: Las variables pertenecen al mismo valor de atributo, no quedan más atributos o no hay más casos. Al igual que el modelo LDA la clasificación se hizo por género obteniendo una assertividad del 82.5 % que redondeado es del 82 %.

Se observó que dio como resultado que independientemente del género, los dos modelos tienen un poder predictivo del 82 %. Lo que resulta muy beneficioso porque es lo que se buscaba en el planteamiento de la tesis.

De este poder predictivo se puede asumir lo siguiente: si se tiene un disco con 10 canciones, al aplicar el modelo en la construcción de las canciones se podría decir que al menos 8 de las canciones del disco serán populares. Lo que da como resultado un beneficio económico para las personas que se dedican a comercializar en la industria musical, en vista de que destinan gran presupuesto a publicidad ya sea de un artista, disco o canción. Y el tener al menos 8 canciones que pueden resultar populares puede ser de gran ayuda. Y esto es si el modelo se aplica a la construcción de las canciones.

Se puede aplicar el modelo a la construcción de las canciones gracias a que con los correlogramas se observó cómo la energía de una canción, su volumen, que tan bailable es o que tan positiva es, pueden hacer que sea una canción popular.

Con ayuda del algoritmo de árboles de clasificación se puede concluir que algunas variables causan mayor impacto en la popularidad, este impacto se pudo observar para cada género musical puesto que se aplicó de esta manera. Las variables que tienen mayor impacto para el género Pop son: energía que para ser popular debería de ser una canción enérgica, el volumen que debería de ser un poco más alto y por último cuantas palabras en promedio debería tener la canción. Para el género Rock las variables que pueden determinar si una canción es popular o no son las siguientes: que tan bailable es, esto es que entre menos bailable sea la canción más probabilidad tendrá de ser popular, también el volumen es una variable importante, debido a que sobre las demás canciones de otros géneros si es importante el volumen un poco más alto en comparación de las demás y por último la variable acústico que indica que entre menos acústico resultara ser más popular. Para el género que se clasificó como Adult Standards que básicamente era la música más antigua que tiene así clasificada Spotify. Las variables para este género fueron: que tan bailable es la canción y no debía de ser tan bailable comparada con otros géneros, la duración de la canción

que podría ser mayor a 3 minutos y por último que tan acústica es la canción que tampoco deber ser acústica. Ahora para el género EDM (Electronic dance music) las variables que pueden determinar si la canción es popular o no son las siguientes: Energía que debe tener en específico un rango para poder ser popular, bailable esto es que no tenga tanta capacidad de baile y positividad tampoco debe tener tanta positividad. Por último, en el género Hip Hop el tempo es de suma importancia, al igual que la capacidad de baile que debe ser muy pequeña para que tenga mas probabilidad de ser popular y por último la duración es una variable importante para este género.

Sin embargo, el modelo no solo se puede usar en la construcción de las canciones, sino que lo ideal sería aplicarlo a un disco ya hecho. Por ejemplo, si se eligieran las canciones de un disco usando alguno de los dos modelos propuestos y este disco tuviera 10 canciones entonces se esperaría que 8 de 10 canciones en promedio tuvieran éxito.

Lo que da un muy buen resultado de los modelos propuestos en la tesis debido a que se cumple su función planteada y que es la económica. La industria musical podría asumir que muchas de sus canciones pueden resultar populares después de aplicar estos modelos propuestos. Es importante mencionar que la popularidad de las canciones también depende de otros factores, como cuanta publicidad tiene el artista, si el artista es reconocido o no, si pertenece a una banda famosa es un ex integrante de alguna banda conocida, es alguien influyente conocido en el medio, entre otros factores. No obstante, esta tesis solo plante como hacerlo mediante un caso de estudio con los modelos propuestos (LDA y árboles de clasificación) asumiendo que solo se miden las características de las canciones con un poder predictivo para ambos modelos del 82 % y no se cuantifican otro tipo de factores.

También se toma en cuenta que como ambos modelos tienen el mismo poder predictivo se pueden usar cualquiera de los propuestos debido a que la información requerida es la misma y el tiempo de ejecución no representa ninguna ventaja uno sobre el otro. Por lo que se puede concluir que cualquiera de los dos resulta ser muy útil.



## Bibliografía

1. Abdi, H. (2007). Discriminant correspondence analysis.
2. Alvinsch. (2020). Así te manipulan para que te guste su música. En <http://tinyurl.com/2tkdthnj>
3. Aliyari Ghassabeh. (2014). Fast incremental LDA feature extraction.
4. Arango Archila. (2015). LA INDUSTRIA DISCOGRÁFICA Y LOS CONSUMIDORES: ¿La música como bien comercial o gratuito?
5. Barnett. (1987). Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Monthly Weather Review*.
6. Brownlee J. (2018). A Gentle Introduction to the Bootstrap Method. En <http://tinyurl.com/2z54z5b9>
7. Celine Ven. (2008). Decision trees for hierarchical multi-label classification.
8. Danielle Fong. (2018). The Sadness Paradox -Why do we enjoy listening to music that makes us sad?
9. De Cea D'Ancona, María Ángeles. (2016) Cuadernos Metodológicos; 54 Análisis discriminante.
10. Deng, H.; Runger, G.; Tuv, E. (2011). Bias of importance measures for multi-valued attributes and solutions.
11. Dina Kirnarskaya. (2009) The Natural Musician On abilities, giftedness, and talent.
12. Farley, B.G.; W.A. Clark. (1954). Simulation of Self-Organizing Systems by Digital Computer.
13. Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems.
14. Hastie, H.; Tibshirani R.; Friedman J. (2008) The Elements of Statistical Learning.
15. Hayashi, Chikio (1998). Studies in Classification, Data Analysis, and Knowledge Organization.
16. K. Karimi and H.J. Hamilton (2011), Generation and Interpretation of Temporal Decision Rules.

17. Kamiński, B.; Jakubczyk, M.; Szufel, P. (2017). A framework for sensitivity analysis of decision trees. *Central European Journal of Operations Research*. 26 (1): 135–159.
18. Kriegel, H. P.; Kröger, P.; Schubert, E.; Zimek, A. (2008). A General Framework for Increasing the Robustness of PCA-Based Correlation Clustering Algorithms. *Scientific and Statistical Database Management*.
19. Kuhn M, Johnson K. (2016). Applied Predictive Modeling.
20. McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience.
21. Mónica Mena Roa. (2021) Spotify alcanza los 155 millones de suscriptores de pago. En <http://tinyurl.com/ypmyrprw>
22. Ochoa, Ana María (2003). Músicas locales en tiempos de globalización.
23. Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*.
24. Quinlan, J. R. (1987). Simplifying decision trees. *International Journal of Man-Machine Studies*.
25. Radio CBC (2020). How Nickelback became the internet's most-hated band. En <http://tinyurl.com/ycrwxmvk>
26. Shin T. (2020). What is Bootstrap Sampling in Machine Learning and Why is it Important?. En <http://tinyurl.com/47m9jvhy>
27. Tan, Pang-Ning, Steinbach, Michael. (2005) Introduction to Data Mining.
28. Tukey, John W. (1962). The Future of Data Analysis.
29. Utgoff, P. E. (1989). Incremental induction of decision trees.
30. Wagner, Harvey M. (1975). Principles of Operations Research: With Applications to Managerial Decisions.
31. Wetcher-Hendricks. (2011). Analyzing Quantitative Data: An Introduction for Social Researchers.