Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: Data Analysis y Big Data

Análisis de la conexión entre trastornos mentales y la preferencia musical mediante algoritmos predictivos

---

Autor: Lucía Pérez Rego

Tutor: Rafael Luque Ocaña

Profesor: Albert Solé Ribalta

A Coruña, 11 de junio de 2024

Créditos/Copyright

Se especifica a continuación los créditos/copyright de este proyecto:

FICHA DEL TRABAJO FINAL

Título del trabajo:	Análisis de la conexión entre trastornos mentales y la preferencia musical mediante algoritmos predictivos
Nombre del autor:	Lucía Pérez Rego
Nombre del colaborador/a docente:	Rafael Luque Ocaña
Nombre del PRA:	Albert Solé Ribalta
Fecha de entrega (mm/aaaa):	06/2024
Titulación o programa:	Máster Universitario en Ciencia de Datos (Data Science)
Área del Trabajo Final:	Data Analysis y Big Data
Idioma del trabajo:	Español
Palabras clave	MongoDB, algoritmos predictivos, Power BI

Dedicatoria/Cita

A mi familia.

Resumen

La sociedad actual es cada vez más consciente de la importancia que tiene la salud mental. Por este motivo, se intenta comprender qué factores influyen en el bienestar de las personas que sufren algún tipo de trastorno. En este punto, la música aparece como un elemento cotidiano capaz de hacer sentir diferentes emociones y estados de ánimo. Surge por tanto la idea de estudiar la relación entre la salud mental y las preferencias musicales de las personas.

Para ello, se pretende almacenar datos recogidos a través de una encuesta con el objetivo de analizar la correlación entre dieciséis géneros musicales y cuatro trastornos mentales: ansiedad, depresión, insomnio y TOC. Además, se pretende realizar un estudio para ver si tras la afirmación de escuchar ciertos géneros musicales se puede predecir si ese individuo sufre algún trastorno mental. Finalmente, se recopilan los resultados extraídos de los estudios anteriores para poder reflejar visualmente los resultados y sacar las conclusiones pertinentes sobre la relación de estos dos ámbitos.

Por tanto, este proyecto busca comprender la relación entre la salud mental y la música, con el objetivo de proporcionar resultados que puedan resultar beneficiosos para profesionales de la salud y de la industria musical. De esta forma, se espera mejorar el bienestar de las personas que sufren algún tipo de trastorno mental.

Palabras clave: Salud mental, trastorno mental, música, género musical, MongoDB, algoritmos predictivos, Power BI.

Abstract

Society is increasingly aware of the importance of mental health. For this reason, researchers aim to understand which factors influence the wellbeing of people who suffer from some kind of mental disorder. At this point, music appears as an everyday element capable of making us feel different emotions and moods. Therefore, the idea of studying the relationship between mental health and people's musical preferences arises.

To do this, the aim is to store data collected through a survey in order to analyse the correlation between sixteen musical genres and four mental disorders: anxiety, depression, insomnia and OCD. In addition, the aim is to carry out a study to see if, the statement of listening to certain genres of music can predict whether the individual suffers from a mental disorder. Finally, the results extracted from the previous studies are compiled in order to visually reflect the results and draw the relevant conclusions about the relationship between these two areas.

Therefore, this project seeks to understand the relationship between mental health and music, with the aim of providing results that may be beneficial for health professionals and the music industry. In this way, it is hoped to improve the wellbeing of people who suffer from some kind of mental disorder.

Palabras clave: Mental health, mental disorder, music, music genre, MongoDB, predictive algorithms, Power BI.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

1. Contexto y motivación

La salud mental toma cada vez más importancia dentro de la sociedad actual. Dentro de este término se engloban los trastornos mentales, caracterizados por ser una alteración clínicamente significativa de la cognición, la regulación de las emociones o el comportamiento de un individuo [1]. En el año 2019, una de cada ocho personas en el mundo padecía un trastorno mental [2].

Por otra parte, según la Real Academia Española la música es el arte de combinar los sonidos de la voz humana o de los instrumentos, o de unos y otros a la vez, de suerte que produzcan deleite, conmoviendo la sensibilidad, ya sea alegre, ya tristemente [3]. Es decir, la música es capaz de provocar distintas emociones y estados de ánimo. Por ello, surge la necesidad de analizar y comprender la relación que existe entre la música y los trastornos mentales.

Este proyecto pretende establecer una relación entre la frecuencia con la que se escuchan diferentes géneros musicales con algunos trastornos mentales, concretamente, la ansiedad, la depresión, el insomnio y el TOC.

La idea de este proyecto surge debido a la estrecha relación que mantiene la autora con el mundo de la música, ya que es una actividad que ha estado unida a ella desde la infancia. Además, se une con uno de sus nuevos intereses, el área de la psicología. De esta manera, nace la idea de combinar la informática con estos dos ámbitos y dedicar este proyecto a estudiar un tema que resulta de su interés.

Por todo ello, se busca realizar un estudio que proporcione resultados que puedan ser beneficiosos para profesionales de la salud mental y de la industria musical. De esta manera, se

espera poder mejorar el bienestar de las personas que sufren algún tipo de trastorno mental.

2. Objetivos

2.1. Hipótesis

El objetivo principal de este proyecto es analizar si existe alguna relación entre los trastornos mentales y las preferencias musicales.

2.2. Objetivos parciales

Con el objetivo de alcanzar la hipótesis, es necesario almacenar la información y analizarla correctamente. Para ello, este proyecto se divide en tres partes diferenciadas:

- Almacenar en MongoDB los datos recogidos de un dataset y que se complementan con una encuesta realizada por la autora a sus contactos.
- Realizar la limpieza y preprocesado de los datos. Estudiar la relación existente mediante la aplicación de varios algoritmos predictivos.
- Crear un informe en Power BI para visualizar los resultados obtenidos en los estudios anteriores.

3. Sostenibilidad, diversidad y desafíos ético/sociales

La Universitat Oberta de Catalunya está públicamente comprometida con la CCEG y los ODS, los cuales se incluye en el programa del máster con la siguiente definición: "Actuar de manera honesta, ética, sostenible, socialmente responsable y respetuosa con los derechos humanos y la diversidad, tanto en la práctica académica como en la profesional, y diseñar soluciones para mejorar estas prácticas". La CCEG aborda tres grandes dimensiones, se comentan a continuación el impacto de cada una de ellas en el proyecto:

Sostenibilidad Este proyecto no tiene impacto en términos de sostenibilidad, ya que los recursos empleados para su realización son mínimos, ya que sólo es necesario el ordenador personal de la autora. En relación a términos legales, este proyecto cumple con la confidencialidad de los datos, ya que se tratan de manera totalmente anónima y bajo el permiso de los usuarios participantes.

Comportamiento ético y responsabilidad social Este proyecto tiene un impacto positivo en aspectos sociales, ya que su objetivo es analizar si existe relación entre las preferencias musicales de la sociedad y el sufrir algún trastorno mental. Para ello, es importante garantizar el comportamiento ético garantizando la confidencialidad de los datos y el consentimiento informado de todas las personas participantes.

Diversidad, género y derechos humanos En este caso puede tener un impacto positivo ya que los datos empleados pueden pertenecer a cualquier persona, de tal manera que las conclusiones extraídas no excluyen a ninguna parte de la población.

4. Enfoque y metodología

Con el objetivo de cumplir todos los objetivos marcados inicialmente, es importante establecer una metodología, lo cual permite estructurar la planificación y el desarrollo de las tareas del proyecto. La metodología del proyecto sigue los siguientes pasos:

- Determinar el tipo de encuesta a realizar. Para ello, es necesario estudiar los datos necesarios para el análisis y tener en cuenta el público al que se dirige.
- Implementar la encuesta diseñada para almacenar la información. En este paso es importante garantizar la confidencialidad de los datos y el consentimiento informado de los participantes.
- Almacenar los datos recogidos en la encuesta para su análisis posterior.
- Realizar un preprocesamiento de los datos mediante una limpieza de los mismos.
- Elaborar un análisis exploratorio de los datos para ver las relaciones entre los trastornos mentales y los géneros musicales escuchados.
- Implementar y evaluar algoritmos predictivos que sean capaces de predecir la presencia de trastornos mentales en función de los hábitos musicales de los participantes.
- Realizar un informe para la presentación visual de los resultados obtenidos.
- Detallar las conclusiones obtenidas a lo largo del desarrollo.

Con el objetivo de gestionar de manera adecuada el trabajo se emplea un tablero de Kanban, ya que permite visualizar de manera clara el flujo de trabajo a seguir [4]. Gracias a él, resulta sencillo identificar puntos de bloqueo y gestionar correctamente el tiempo de las tareas.

5. Planificación

La planificación del proyecto se detalla mediante un diagrama de Gantt. Para ello, se toman como referencia las fechas de las Pruebas de Evaluación Continua y se definen las principales tareas asociadas a ellas. Se puede ver la estimación y las posibles dependencias entre las tareas en el diagrama resultante mostrado en la figura 1.1. Cabe destacar que la planificación puede sufrir variaciones a lo largo del proyecto, ya que para cada una de las entregas se detallan las tareas a realizar, lo cual permite completar el diagrama con tareas más concretas. Además, la estimación de las tareas puede sufrir alguna desviación, especialmente en aquellas englobadas en la implementación del proyecto.

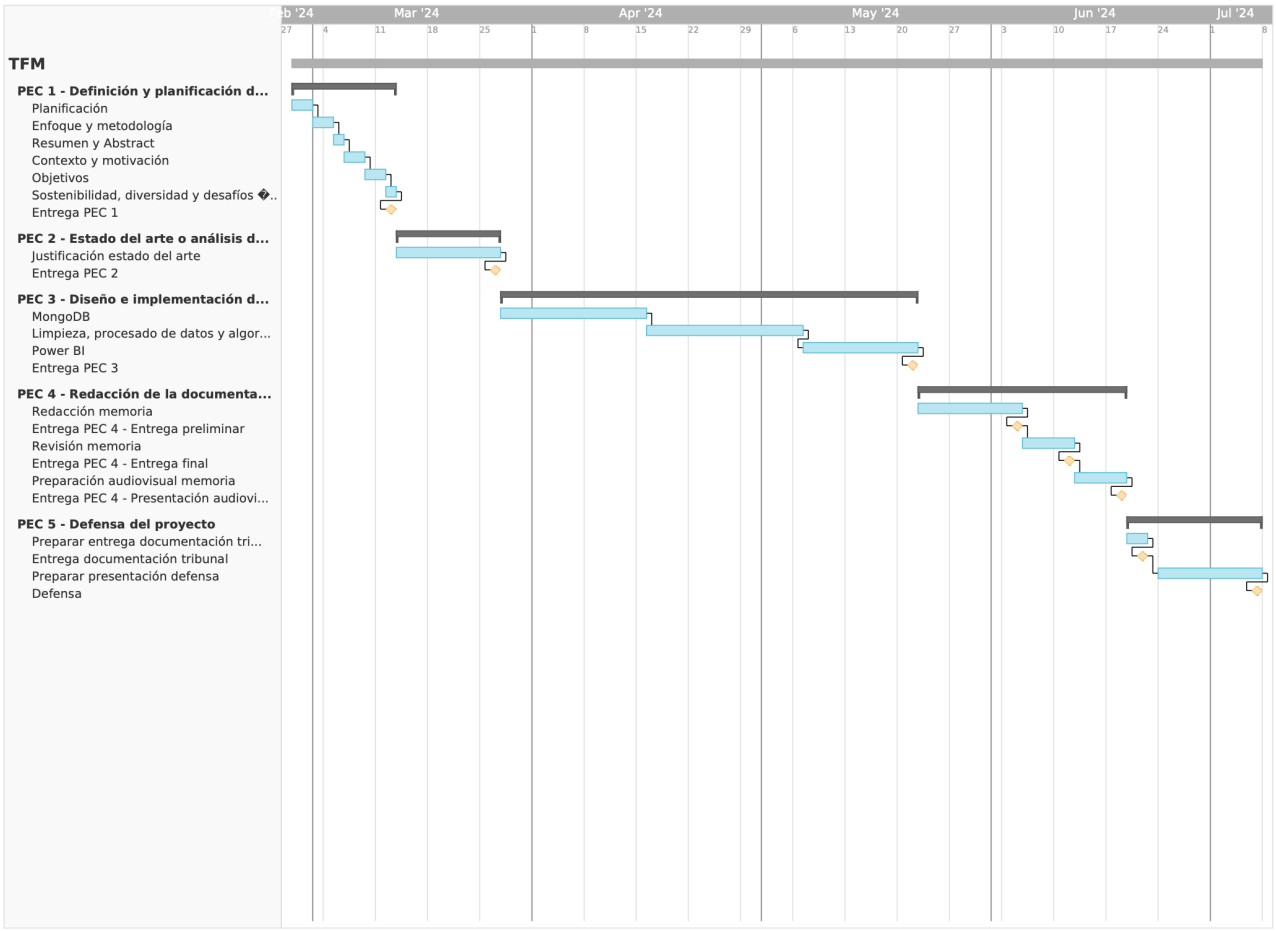

Figura 1.1: Diagrama de Gantt

5.1. Gestión de riesgos

La gestión de riesgos pretende identificar posibles dificultades que se pueden encontrar durante el desarrollo del proyecto y que afecten al cumplimiento de los objetivos. Estos son:

- Falta de participación en la encuesta. En este proyecto los datos se recogen de dos fuentes distintas: un conjunto de datos publicado que cuenta con una licencia que permite su uso y los datos recogidos a través de una encuesta. En caso de una baja participación en la encuesta y teniendo en cuenta que su única intención es añadir más muestras, se empleará únicamente el conjunto de datos, ya que es suficiente para el estudio que se va a realizar.
- Resultados no concluyentes. Existe la posibilidad de que una vez terminado el análisis no haya ninguna conclusión que nos indique si existe o no relación entre los trastornos mentales y los géneros musicales escuchados. En este caso, se detallarán los resultados y las conclusiones obtenidas, asumiendo que con los datos recogidos y el tipo de análisis realizado no es suficiente para llegar a una conclusión definitiva.

Capítulo 2

Estado del arte

Es importante comprender el estado actual de la investigación de la relación entre la salud mental y las preferencias musicales. Este capítulo tiene como finalidad revisar el estado actual de la investigación, proporcionando una visión general de los objetivos alcanzados, los desafíos actuales y las posibles aplicaciones.

A través de la revisión de literatura existente, se entenderá la influencia que tiene la música en las emociones y se analizará desde el punto de vista psicológico, la relación que existe entre las preferencias musicales y los trastornos mentales. Además, se detallarán las posibles aplicaciones prácticas de este proyecto en diferentes áreas. Finalmente, se estudiarán las herramientas y tecnologías que permiten llevar a cabo el análisis de este proyecto. De esta manera, se establece un marco de trabajo sólido para orientar la investigación y proporcionar resultados dentro de este ámbito de conocimiento.

1. Influencia de la música en las emociones

La música tiene un papel importante dentro de la sociedad, siendo un medio de expresión y de identidad cultural. Se estima que las personas escuchan entre una y tres horas de música diariamente. Por ello, se comenzaron a realizar estudios para ver cómo afecta y si tiene algún tipo de influencia en las personas.

En un primer estudio [5], se observa que la música puede repercutir en las emociones a través de dos componentes: el *arousal* y la *valencia*. El *arousal* refleja la actividad que un estímulo genera sobre el sujeto que lo recibe, cuyo rango abarca desde la excitación a la calma. Por otra parte, la *valencia* evalúa la bondad o cualidad de un estímulo, con un rango que va desde lo

agradable a lo desagradable. En el ámbito musical se trabaja con el concepto de BPM o beats por minuto, que refleja la velocidad a la que va la música.

Teniendo en cuenta estos conceptos se realiza un estudio sobre la contribución de diversas características estructurales de la música [6], como por ejemplo el tempo, el timbre, el modo o la articulación rítmica sobre respuestas psicofisiológicas y se llevan a cabo evaluaciones subjetivas del estado emocional. Como conclusión, se obtuvo que existe una correlación entre los BPM y el *arousal*. Asimismo, también se observa una relación positiva entre la *valencia* y los juicios del *arousal* que en general correspondieron a las piezas con mayor BPM .

Además, la investigación analiza si los resultados obtenidos difieren entre personas introvertidas y extrovertidas. Se concluye que las personas extrovertidas prefieren un mayor nivel de excitación que los introvertidos, mientras que niveles menores de excitación ambiental serán más desagradables para ellos que para los introvertidos.

2. Preferencias musicales y trastornos mentales

Una vez comprendido lo anterior, surge la necesidad de hablar de los géneros musicales. Los géneros musicales son categorías que han nacen a través de una compleja interacción de culturas, artistas y del mercado para caracterizar similitudes entre músicos o composiciones y organizar colecciones de música [7]. No obstante, en ocasiones los límites entre los géneros pueden resultar confusos. Las composiciones o canciones asociadas a cada género musical suelen estar comprendidas en un rango de BPM. Las pulsaciones por minuto o BPM es la velocidad a la que se reproduce una composición musical. Por ejemplo, las canciones de rock suelen estar en el rango 120-140 BPM, mientras que hip-hop ronda los 70-100 BPM.

Existe una investigación cuyo objetivo es correlacionar los géneros musicales con los rasgos de personalidad de las personas que lo escuchan [8]. La conclusión obtenida es que sí que hay relación entre ambos aspectos. Un ejemplo de los resultados obtenidos es el caso del heavy metal, el cual se relaciona con asertividad, agresividad, indiferencia hacia los sentimientos de los demás, malhumor, pesimismo y una mayor probabilidad de actuar por impulso.

Se recalca que la música no es el factor causal si no que más bien la preferencia es indicativa de una vulnerabilidad emocional. Se necesita más investigación para determinar si las preferencias musicales de aquellas personas diagnosticadas difieren sustancialmente de la población en general.

3. Estudios similares

Los estudios anteriores estaban enfocados desde el punto de vista de la psicología y han ayudado a entender los conceptos que involucran este Trabajo de Fin de Máster y el tipo de resultados que se pueden llegar a esperar. En este punto es importante analizar los proyectos que engloban esta temática desde el prisma informático para ver que herramientas se están utilizando y en que punto se encuentran los estudios.

El primero de los estudios analizados se llama *Çlassification Algorithms based Mental Health Prediction using Data Mining"* [9]. Su objetivo es crear un sitio web en el que los usuarios introducen valores en un formulario para obtener información acerca de su salud mental. Para ello, se almacena un conjunto de datos proporcionado por una encuesta y que recoge información principalmente de personas que trabajan. Los datos son procesados para posteriormente aplicar los algoritmos Decision Tree y Random Forest para evaluar los datos y encontrar el algoritmo más preciso. A continuación, se crea un modelo basado en el mejor algoritmo que será empleado en el sitio web.

En segundo lugar, se ha analizado el estudio denominado *"Mental Healthcare Analysys using Power BI and Machine Learning"* [10]. Su objetivo principal es analizar diferentes factores como el sexo, la edad, el insomnio, la soledad, etc y ver cuáles afectan a una mala salud mental. Para ello, se ha difundido un formulario que contiene preguntas sobre el estilo de vida habitual de las personas. Posteriormente, se preprocesan los datos y se aplican algoritmos de aprendizaje automático para la clasificación y la predicción. Gracias a esto, se encuentran diferentes resultados que permiten determinar cuáles son los factores que más afectan a la salud mental de una persona y se puede sugerir qué tipo de cosas evitar para su buena salud mental. Para terminar, los resultados obtenidos son representados en Power BI, una herramienta de visualización que permite conectarse con facilidad a diferentes orígenes de datos y realizar visualizaciones de manera clara y sencilla [11].

El último artículo analizado es *Analyzing the Mental Health of Engineering Students using Classification and Regression* [12]. El objetivo es emplear minería de datos para comprender los factores que afectan a la salud mental de estudiantes de ingeniería. Para ello, realizan una encuesta que incluía preguntas sobre las posibles influencias académicas en la salud mental, como el año de estudio, el programa académico y la situación sentimental. A continuación, se aplican algoritmos de regresión lineal y clasificación para identificar cuáles de las influencias mencionadas tiene mayor efecto sobre cada aspecto de la salud mental. Los resultados de este

estudio sugieren una serie de recomendaciones para ayudar a mejorar la salud mental de los estudiantes universitarios de ingeniería.

Una vez comprendida la relación que existe entre la música y la salud mental desde el punto de vista psicológico, y tras haber analizado el estado actual en el que se encuentran los proyectos desde una perspectiva tecnológica, nace la idea de este proyecto. Siguiendo la línea de los estudios mencionados anteriormente, se parte de un conjunto de datos que proviene de una encuesta realizada a usuarios, complementada con una nueva encuesta análoga realizada por la autora del proyecto. Una vez almacenados y preprocesados los datos, se realizará un estudio utilizando diferentes algoritmos de predicción. Finalmente, se elaborará un informe en la herramienta Power BI para visualizar los resultados obtenidos y sacar las conclusiones derivadas de este proyecto.

4. Aplicación práctica

El estudio de la relación entre las preferencias musicales y los trastornos mentales abarca diferentes ámbitos, por lo que los resultados de este proyecto pueden tener un impacto significativo desde diferentes perspectivas.

En el ámbito de la medicina, puede contribuir en el diagnóstico de un paciente al estudiar sus hábitos de escucha, ya que puede dar información relevante sobre su estado emocional y mental permitiendo la detección de síntomas tempranos de trastornos mentales. Además, tras saber que la música afecta a las emociones, puede utilizarse de manera personalizada como un tratamiento. Esto se conoce como musicoterapia y es un proceso en el cual el terapeuta ayuda al paciente a fomentar su salud, utilizando experiencias musicales y las relaciones que se desarrollan a través de ellas [13]. De igual forma, se podría realizar una monitorización de los géneros escuchados por el paciente durante su periodo de terapia para proporcionar información sobre su progreso.

Por otro lado, tiene importancia en el ámbito educativo, los hallazgos obtenidos pueden ser utilizados para concienciar a los estudiantes de la importancia de la salud mental y del uso de la música como una herramienta de apoyo emocional.

Por último, tiene relevancia dentro de la industria musical. Los resultados pueden ser empleados por plataformas de streaming para mejorar la personalización de las recomendaciones musicales. Además, se podría crear música terapéutica orientada al bienestar de los oyentes, lo

que podría abrir nuevas oportunidades en el mercado y contribuir al desarrollo de la música como herramienta de salud.

Capítulo 3

Conjunto de datos

1. Descripción

El conjunto de datos utilizado en el desarrollo de este proyecto viene de dos orígenes diferentes. En primer lugar, se emplea como referencia un *dataset* de Kaggle que recoge datos mediante la realización de una encuesta sobre las preferencias musicales y algunos trastornos mentales: ansiedad, depresión, insomnio y TOC. Adicionalmente, se ha llevado a cabo una encuesta propia totalmente anónima a través de Google Forms que ha sido distribuida a los contactos cercanos de la autora con el objetivo de poder ampliar el conjunto de datos sobre el que se va a trabajar. A continuación se van a comentar todas las columnas que tiene el conjunto de datos utilizado.

Bloque 1

Los participantes responden a preguntas genéricas que tratan el trasfondo musical y los hábitos de escucha.

- Timestamp: Fecha y hora en que se envió el formulario.
- Age: Edad del participante.
- Primary streaming service: El principal servicio de transmisión que utiliza el encuestado.
- Hours per day: Número de horas que el encuestado escucha música al día.
- While working: Especifica si el participante escucha música mientras estudia o trabaja.
- Instrumentalist: Especifica si el participante toca algún instrumento.
- Composer: Especifica si el participante compone su propia música.

- Fav genre: Género favorito escuchado por el encuestado.
- Exploratory: Especifica si el encuestado escucha nuevos artistas o nuevos géneros.
- Foreign languages: Especifica si el encuestado escucha música en otros idiomas.
- BPM: Número de *beats* por minuto del género favorito. Determina la velocidad a la que se reproduce una canción. El *tempo* de las canciones puede tener una influencia directa en el estado anímico.
- Music effects: Especifica si el usuario considera que la música mejora, empeora o no tiene efecto en la salud mental.
- Permission: Sólo está en el conjunto de datos descargado de Kaggle. Indica si el usuario acepta las condiciones de la encuesta.

Bloque 2

Las personas que participan en la encuesta clasifican la frecuencia con la que escuchan 16 géneros musicales. Las posibles respuestas son: nunca, casi nunca, a veces o muy frecuentemente.

- Frequency [Classical].
- Frequency [Country].
- Frequency [EDM].
- Frequency [Folk].
- Frequency [Gospel].
- Frequency [Hip hop].
- Frequency [Jazz].
- Frequency [K pop].
- Frequency [Latin].
- Frequency [Lofi].
- Frequency [Metal].
- Frequency [Pop].
- Frequency [R&B].
- Frequency [Rap].
- Frequency [Rock].
- Frequency [Video game music].

Bloque 3

Los encuestados clasifican diferentes trastornos mentales en una escala del 0 al 10, donde 0 es no experimento este trastorno y 10, lo experimento regularmente.

- Anxiety.
- Depression.
- Insomnia.
- OCD.

2. Características

El primer conjunto de datos ha sido descargado de Kaggle y son el resultado de una encuesta realizada en Google. Esta encuesta, tal y como indica la autora del post de Kaggle, fue distribuida a través de foros de Reddit, servidores de Discord y redes sociales. Además, utilizaron carteles para anunciar el formulario en bibliotecas, parques y otros lugares públicos. La autora permitió que alguna de las preguntas fueran opcionales.

Este conjunto cuenta con 736 filas distribuidas en 33 columnas comentadas anteriormente. Dentro de las columnas, aparecen valores nulos en las siguientes:

- Age: 1.
- Primary service: 1.
- While working: 3.
- Instrumentalist: 4.
- Composer: 1.
- Foreign languages: 4.
- BPM: 107.
- Music effects: 8.

Teniendo en cuenta que estos datos no van a afectar al estudio que se haga en los algoritmos predictivos, ya que tan sólo se utilizan las frecuencias de los géneros musicales como variables independientes y los trastornos como variables dependientes, se ha decidido no tratar los nulos y poder analizarlos en la herramienta de visualización Power BI.

Por otro lado, se han recogido datos desde una encuesta propia. Esta encuesta ha sido creada en Google Forms con las mismas preguntas que la encuesta comentada anteriormente a excepción de la columna "Permission", ya que va implícito al decidir enviar la encuesta. Todas las preguntas realizadas eran de respuesta obligatoria a excepción de los BPM, ya que se consideraba una pregunta con cierta complejidad. De manera aclaratoria, se añadió una nota en la que se explicaba el concepto y su repercusión en el estado anímico. Por otra parte, algunas respuestas como la edad o el género favorito eran de respuesta corta, aunque la mayoría de ellas tenían respuestas predefinidas con las mismas opciones que en la encuesta original.

La distribución de esta encuesta ha sido realizada a través de las redes sociales y de los contactos de la autora. Cabe destacar que en la configuración se ha especificado la opción de

no almacenar las direcciones de correo electrónico para garantizar el anonimato de los usuarios participantes. En total, se han obtenido 46 respuestas sin valores nulos en ninguna de las columnas.

De esta manera, el conjunto de datos final que se carga en la base de datos tiene un total de 782 filas y 32 columnas.

Capítulo 4

Implementación

Este capítulo está dedicado al desarrollo e implementación de este proyecto. Todo el código realizado se encuentra en el siguiente GitHub https://github.com/luciaprego/TFM.git. En primer lugar, en la sección 1 se comentan las características principales de MongoDB, las colecciones creadas y los *scripts* realizados. Posteriormente, en la sección 2 se realiza un análisis exploratorio de los datos para tener una visión general de sus principales características. A continuación, en la sección 3 se lleva a cabo la implementación de los algoritmos de predicción con la búsqueda de los mejores parámetros y el entrenamiento de los modelos. Finalmente, en la sección 4 se desarrollan dos *dashboards* para analizar la información vista a lo largo de todo el proyecto de una manera visualmente atractiva e intuitiva.

1. Mongo DB

Mongo DB es una base de datos no relacional y de código abierto [14]. Su modelo de datos está basado en documentos, lo que ofrece una gran escalabilidad y flexibilidad, y un modelo de consultas e indexación avanzado. Cada documento se almacena en una colección, un contenedor para almacenar documentos.

Para este proyecto se han instalado tres componentes principales de MongoDB:

- MongoDB Community Edition: Es una base de datos gratuita y de código abierto. Ofrece un modelo de datos de documentos flexible junto con soporte para consultas ad-hoc, indexación secundaria y agregaciones en tiempo real para proporcionar formas poderosas de acceder y analizar sus datos.
- MongoDB Shell: Una interfaz de línea de comandos basada en JavaScript que permite conectarse a MongoDB para trabajar con sus datos y configurar la base de datos.

- MongoDB Compass: Una interfaz gráfica que permite manipular de manera sencilla la base de datos proporcionando visualizaciones detalladas de esquemas y métricas de rendimiento en tiempo real, entre otras cosas.

Una vez realizada la instalación, desde MongoDB Compass se ha creado la base de datos que almacenará todos los datos del proyecto, denominada "mentalhealth". Dentro de la base de datos se han creado dos colecciones distintas, "mentalhealth" y "mentalhealth_duplicated". La primera de ellas recoge los datos obtenidos directamente de las encuestas, tanto la de Kaggle como la propia. Por otro lado, "mentalhealth_duplicated" se crea con la finalidad de aumentar el volumen de datos para el estudio que se hace posteriormente con los algoritmos predictivos. Esto se debe a que hay un número muy reducido número de datos que no son suficientes para poder realizar las predicciones.

Después de crear las colecciones, se realiza la carga de los datos mediante la importación de CSV. Cabe destacar que en ambas colecciones se realiza la misma carga inicial. Se ha decidido omitir de la carga el campo la columna "Permission" del dataset de Kaggle, ya que siempre tiene el mismo valor y no aporta información en el estudio. A continuación, se detallan los pasos posteriores realizados para cada una de las colecciones:

Colección "mentalhealth"

Se ha creado un *script* que realiza una actualización de los datos. Principalmente, las operaciones a realizar son:

- Composer, exploratory, foreign languages, instrumentalist y while working: Cambiar el valor "Sí" a "Yes".
- Frecuencias de todos los géneros musicales:
  - Cambiar "Muy frecuentemente" por "Very frequently".
  - Cambiar "A veces" por "Sometimes".
  - Cambiar "Casi nunca" por "Rarely".
  - Cambiar "Nunca" por "Never".
- Cambiar el formato del campo "Timestamp" para que tenga formato fecha.
- Añadir 4 campos nuevos asignando una categoría para cada uno de los trastornos. Esto se hace para seguir la estructura que se va a seguir en la otra colección y además estos campos resultarán útiles en Power BI para mostrar la información.

- Mayor o igual que 0 y menor que 4: Autopercepción baja.
- Mayor o igual que 4 y menor que 7: Autopercepción media.
- Mayor o igual que 7 y menor o igual que 10: Autopercepción alta.

Colección "mentalhealth_duplicated"

Para esta colección se han creado tres *scripts* diferentes. El primero de ellos, realiza una actualización inicial de los datos, aunque en este caso enfocado a las necesidades de los algoritmos de predicción, es decir, valores numéricos para los valores utilizados en el análisis, así como para las variables dependientes e independientes. Las variables dependientes son las 16 columnas con los géneros musicales y las independientes las 4 columnas con los trastornos mentales. Por ello, las actualizaciones a realizar son las siguientes:

- Script 1. Primera actualización.
  - Composer, exploratory, foreign languages, instrumentalist y while working: Cambiar el valor "Sí" o "Yes" a 1 y "No" a 0.
  - Frecuencias de todos los géneros musicales:
    - Cambiar "Muy frecuentemente" a 3.
    - Cambiar "A veces" a 2.
    - Cambiar "Casi nunca" a 1.
    - Cambiar "Nunca" 0.
  - Cambiar formato campo "Timestamp" para que tenga formato fecha.
- Script 2. Ampliación de los datos. Se realizan 300 copias de cada uno de los documentos originales. De cara a evitar datos duplicados que puedan generar problemas posteriormente en los algoritmos de predicción, se aplican variaciones para los campos que forman las variables dependientes e independientes del estudio, tal y como se observa en el *listing* 4.1. En el caso de las frecuencias de los géneros musicales, se aplica una variación de ± 1 respecto al valor original ya que su escala original comprende valores de 0 a 3, mientras que para los trastornos la escala está comprendida entre el 0 y el 10, por lo que se aplica una variación de ± 2. En el caso de que el nuevo valor esté fuera del rango de las escalas, se asigna el que había inicialmente.

```
async function variacionAleatoriaTrastornos(valor) {
  const variacion = Math.floor(Math.random() * 5) -2;

  let nuevoValor = valor + variacion;
```

```
5
6   if(nuevoValor < 0 || nuevoValor > 10) {
7     return valor } else { return nuevoValor }
8 }
9
10 async function variacionAleatoriaFrecuencias(valor) {
11   const variacion = Math.floor(Math.random() * 2) -1;
12
13   let nuevoValor = valor + variacion;
14
15   if(nuevoValor < 0 || nuevoValor > 3) {
16     return valor } else { return nuevoValor }
17 }
```

Listing 4.1: Script variaciones.

- Script 3. Segunda actualización. Una vez aplicadas las variaciones se lleva a cabo un escalado de los datos para cada uno de los trastornos que se corresponde con el nivel de sintomatologías realizado en la colección de "mentalhealth".
  - Mayor o igual que 0 y menor que 4: 0. Se corresponde con la autopercepción baja.
  - Mayor o igual que 4 y menor que 7: 1. Se corresponde con la autopercepción media.
  - Mayor o igual que 7 y menor o igual que 10: 2. Se corresponde con la autopercepción alta.

Finalmente, esta colección es exportada a CSV para su utilización posterior.

Para llevar a cabo estas actualizaciones de los datos se utiliza la función de MongoDB updateMany() que actualiza todos los documentos que coinciden con el filtro indicado para una colección. Tiene la siguiente sintaxis:

```
1 db.collection.updateMany(filter, update, options)
```

Listing 4.2: Sintaxis updateMany().

2. Análisis exploratorio

El análisis exploratorio sirve para analizar e investigar conjuntos de datos y resumir sus características principales, lo que permite descubrir patrones, detectar anomalías y estudiar las

relaciones entre las variables del conjunto. En primer lugar, es importante destacar que esta sección emplea la colección de datos "mentalhealth_duplicated". Los resultados obtenidos en esta parte del proyecto nos pueden revelar si con un conjunto de datos más grande se pueden obtener resultados interesantes. No obstante, es importante tener en cuenta que este proyecto es un Trabajo de Fin de Máster y su objetivo es demostrar el conocimiento de los métodos aplicados y no tanto conseguir unos resultados óptimos para conclusiones o aplicaciones futuras.

Una vez realizada la carga de datos y construido el *dataset* final, se procede a realizar un análisis exploratorio con el objetivo de observar las principales características del mismo. Para empezar, se realiza un estudio de los valores nulos que se encuentran en el conjunto de datos. Tal y como se adelantaba en el capítulo anterior, la encuesta descargada de Kaggle contiene valores nulos porque se permitía dejar preguntas sin responder. Estos valores no se van a tratar ya que no afectan a las variables que se van a utilizar en este estudio. Es importante destacar que en este conjunto de datos el número de nulos ha aumentado, tal y como se puede ver en el *listing* 4.3, ya que en la ampliación de los datos se han duplicado las filas y a estas columnas no se les ha aplicado ningún tipo de variación.

```
_id                                0
Timestamp                          0
Age                              501
Primary streaming service        501
Hours per day                      0
While working                   1503
Instrumentalist                 2004
Composer                         501
Fav genre                          0
Exploratory                        0
Foreign languages               2004
BPM                            58617
Frequency [Classical]              0
Frequency [Country]                0
Frequency [EDM]                    0
Frequency [Folk]                   0
Frequency [Gospel]                 0
Frequency [Hip hop]                0
Frequency [Jazz]                   0
Frequency [K pop]                  0
Frequency [Latin]                  0
Frequency [Lofi]                   0
Frequency [Metal]                  0
Frequency [Pop]                    0
```

```
Frequency [R&B]                    0
Frequency [Rap]                    0
Frequency [Rock]                   0
Frequency [Video game music]       0
Anxiety                            0
Depression                         0
Insomnia                           0
OCD                                0
Music effects                   4008
dtype: int64
```

Listing 4.3: Distribución de valores nulos en el conjunto de datos.

En primer lugar, se realiza un estudio de las preguntas más genéricas de la encuesta para poder analizar el perfil de las personas que han respondido las encuestas. Se obtiene la siguiente información:

- La media de edad de las personas que han respondido las encuestas es: 25.19.
- El número medio de horas al día que escuchan los participantes es: 3.58.
- El 78.75 % de los participantes escuchan música mientras trabajan o estudian.
- El 32.27 % de los participantes tocan algún instrumento.
- El 16.77 % de los participantes componen música.
- El 71.32 % de los participantes escuchan nuevos artistas.
- El 56.98 % de los participantes escuchan música en otro idioma.
- El 74.52 % de los participantes piensan que escuchar música afecta positivamente en la salud mental, mientras que el 22.28 % piensan que no tiene ningún efecto. Tan sólo el 2.18 % creen que tiene efectos negativos.

Además, la principal plataforma de streaming que utilizan los usuarios es Spotify, con un gran salto respecto a la siguiente plataforma en la lista YouTube Music, tal y como se puede observar en la figura 4.1.

Además, se ha realizado un estudio de la distribución que siguen los géneros musicales para cada una de sus opciones. En la figura 4.2 se pude ver que la opción de nunca tiene el mayor porcentaje es Gospel, mientras que el que tiene un mayor porcentaje escucha muy frecuente es el rock, tal y como avanzábamos con el gráfico comentado anteriormente.

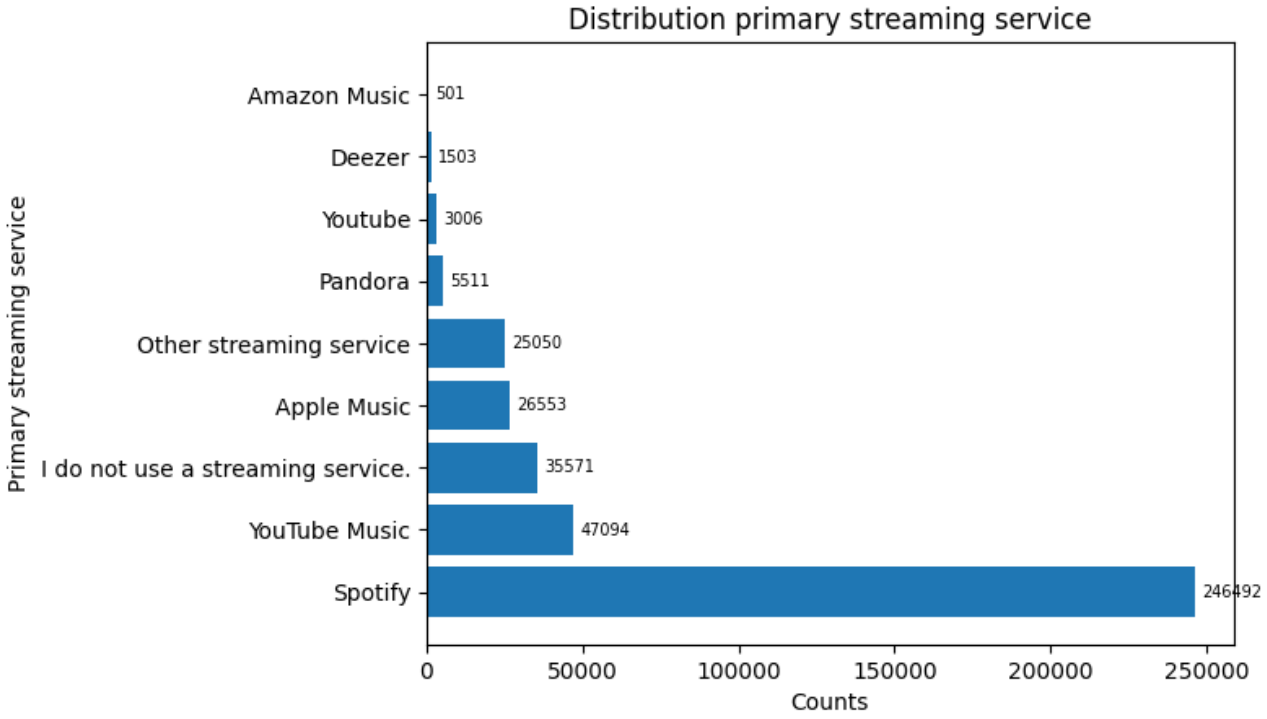

Figura 4.1: Número de personas que emplean cada plataforma de escucha.

De cara a los trastornos mentales, es necesario recordar que los valores respondidos en las encuestas son autoinformados y que además han sufrido las variaciones que se han aplicado para el aumento del volumen de datos. La escala se divide en autopercepción baja, autopercepción media y autopercepción alta. Teniendo en cuenta esta información, se puede ver en la figura 4.3 que la ansiedad es el trastorno que tiene un mayor número de casos de autopercepción alta, seguido por la depresión. De igual forma, la autopercepción baja, está más presente tanto en el insomnio como en el TOC.

A continuación, se ha llevado a cabo un estudio de la correlación existente entre diferentes variables para ver la relación lineal entre las variables. El coeficiente de correlación va desde -1 a 1, cuanto más se aproxima el valor a 0 menor es la relación lineal entre ambas variables. Por otro lado los valores negativos indican una correlación negativa, es decir, los valores de una variable tienen a incrementarse mientras que los valores de la otra variable descienden. De manera contraria, los valores positivos indican una correlación positiva, en la que los valores de ambas variables tienen a incrementarse juntos.

En primer lugar, se ha estudiado la correlación entre algunas de las características principales, tal y como se puede observar en la figura 4.4. Las correlaciones más significativas se encuen-

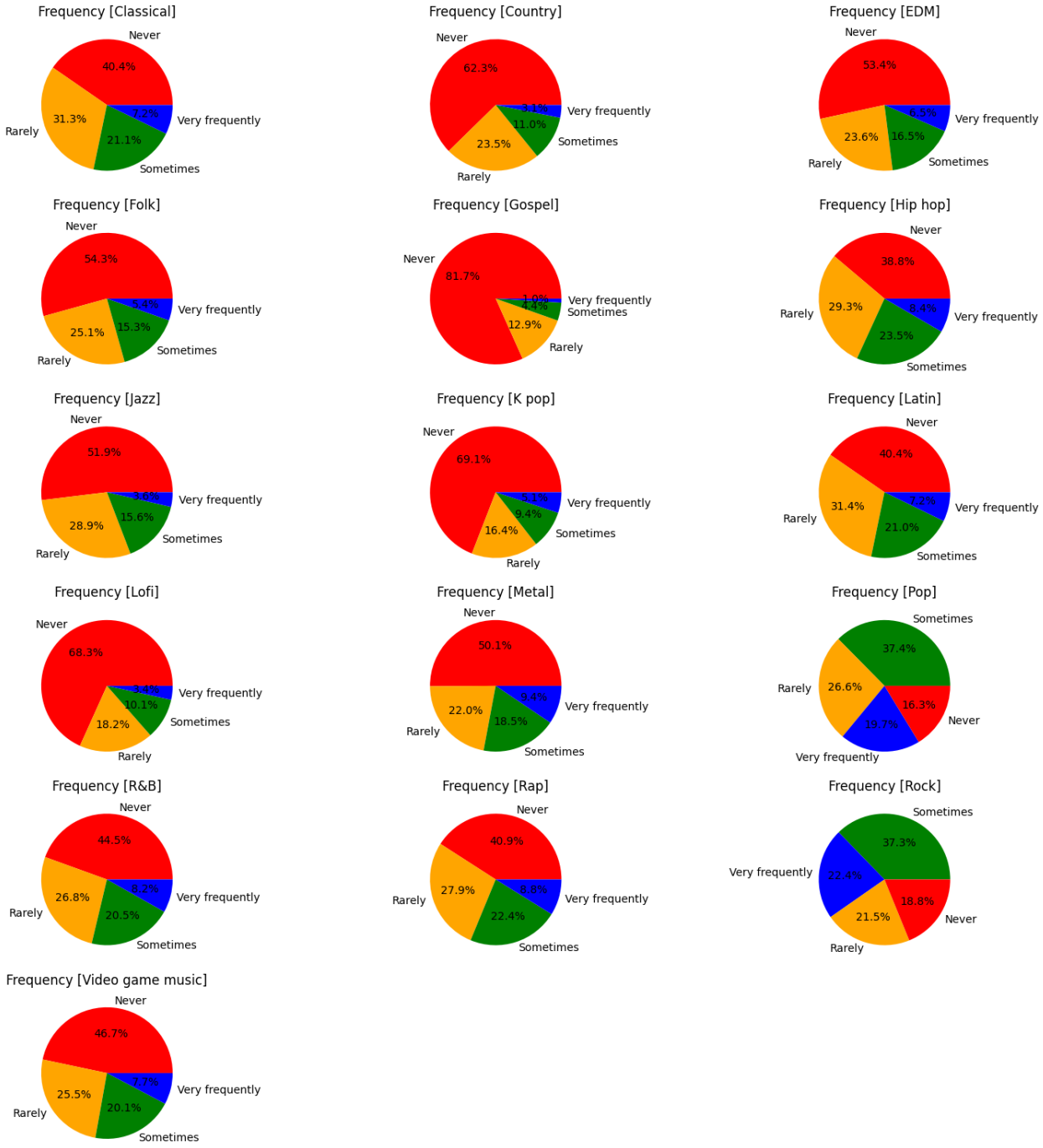

Figura 4.2: Porcentaje de frecuencia por género musical.

tran entre las personas que componen música y tocan un instrumento, junto con las horas de escucha diaria y las personas que escuchan música mientras trabajan. Por otro lado, existe una correlación negativa entre la edad y la escucha de nuevos géneros o artistas, de manera que a cuanta más edad menos se exploran nuevos estilos.

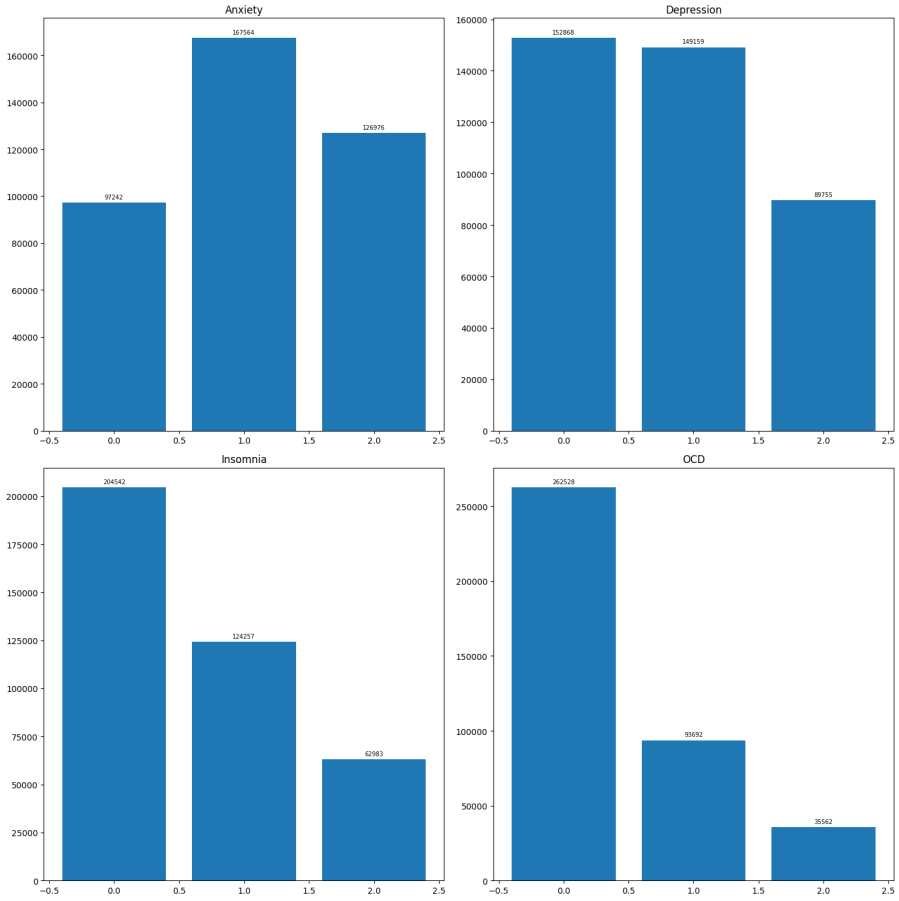

Figura 4.3: Número de casos por sintomatología en los trastornos mentales.

Además, se ha estudiado la correlación que existe entre los diferentes trastornos de salud mental. En la figura 4.5 se puede ver que la mayor relación se produce entre la ansiedad y la depresión, aunque en todos los casos la correlación entre las variables es positiva. De igual forma, la correlación menor se produce entre la depresión y el TOC.

Finalmente, se ha utilizado ha hecho un estudio entre la correlación entre los géneros musicales y los trastornos, tal y como se puede observar en la figura 4.6. En el caso de los géneros, las principales correlaciones positivas se encuentran entre la música clásica y la latina, seguido por la correlación entre el rap y el hip hop. Por otro lado, destaca la correlación negativa existente entre el rock y el K pop, así como entre el pop y el metal. En el caso del estudio entre géneros y trastornos se puede ver que tienen valores muy cercanos a 0 en la mayoría de los casos, por lo que existe una relación directa entre estas variables. El único valor a destacar es el valor entre el metal y el rock respecto a la depresión, así como del metal con el insomnio.

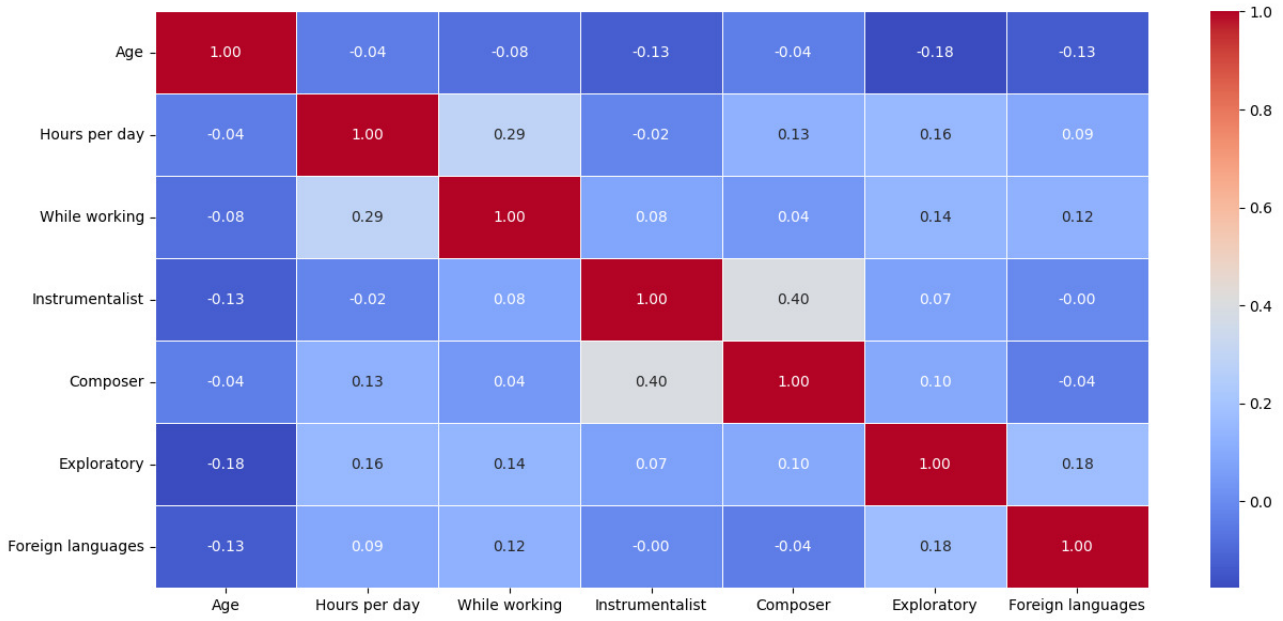

Figura 4.4: Correlaciones generales.

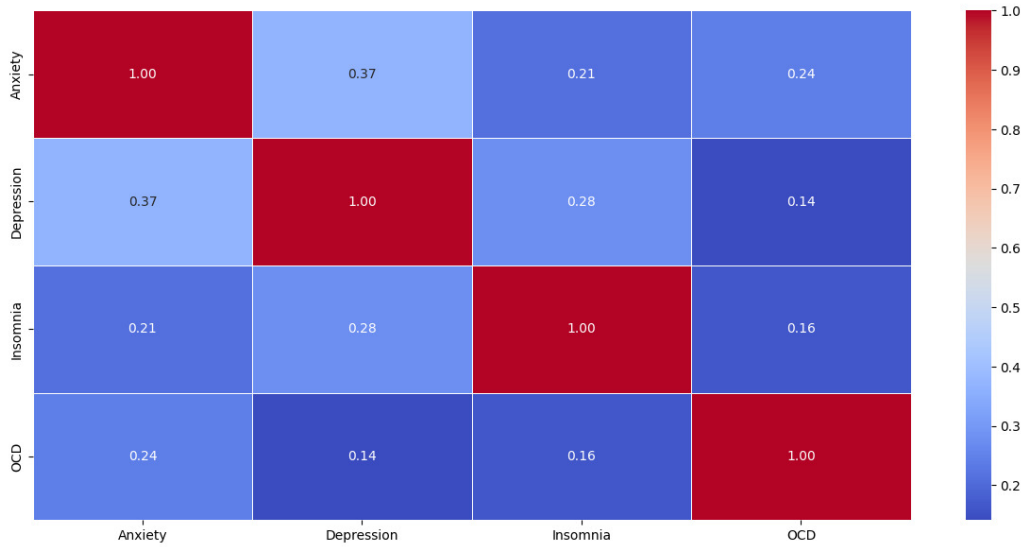

Figura 4.5: Correlaciones trastornos.

3. Algoritmos predictivos

Los algoritmos predictivos son herramientas que se utilizan para predecir resultados basándose en datos históricos. Para ello, se crean modelos con los que analizar patrones y relaciones dentro de un conjunto de datos y después utilizar esta información para predecir comportamientos futuros. En la actualidad, los modelos de predicción tienen aplicaciones en múltiples

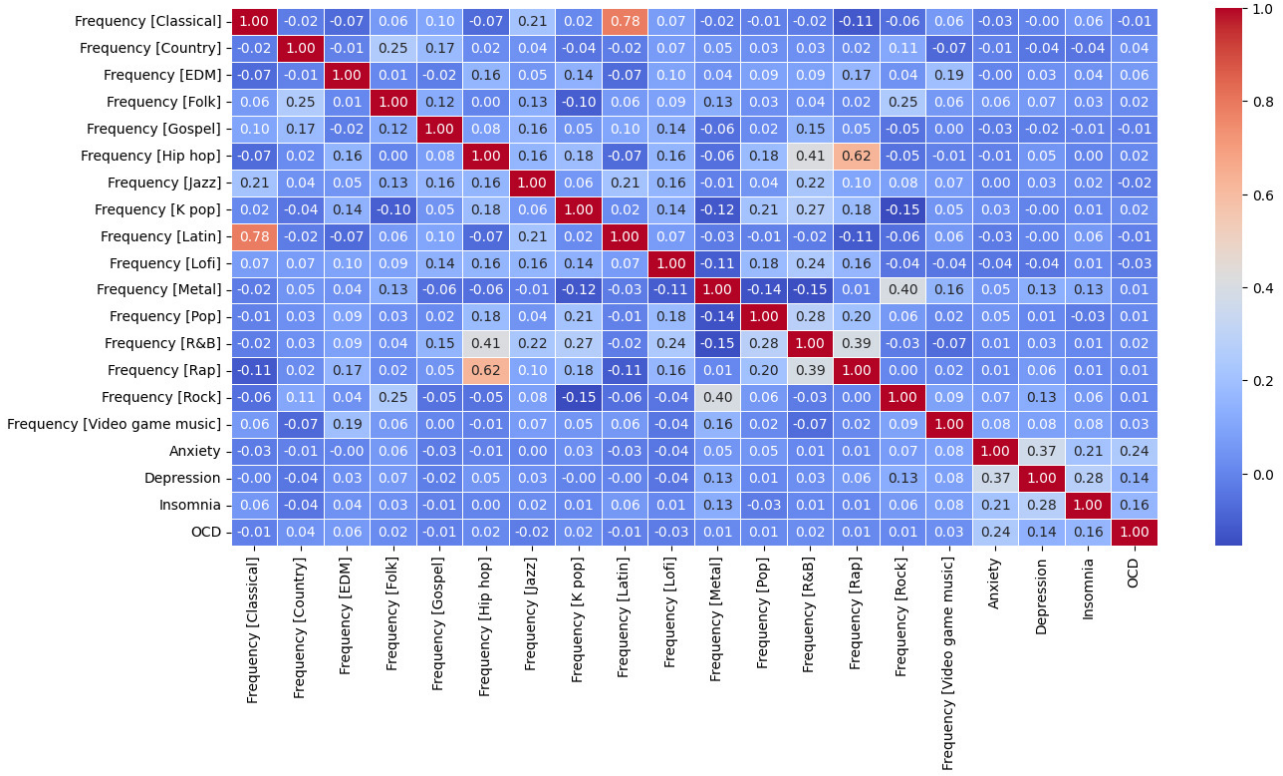

Figura 4.6: Correlaciones géneros musicales y trastornos.

campos, como las finanzas, el marketing o la salud.

En este proyecto el objetivo es predecir el valor de cada uno de los trastornos de salud mental mediante los hábitos de escucha de los usuarios. De cara a realizar una predicción más simplificada, se han escalado los valores de los trastornos en tres niveles, tal y como se comentó anteriormente en la sección 1 de este mismo capítulo. Debido a las características del *dataset* original, es necesario emplear algoritmos basados en el aprendizaje supervisado, caracterizados por el uso de conjuntos de datos etiquetados. De esta manera, el objetivo es aprender una función que sea capaz de mapear las entradas a las salidas y pueda predecir etiquetas en nuevos datos correctamente. Dentro de los modelos supervisados, se implementan modelos de clasificación, donde el objetivo es asignar una etiqueta a cada muestra de datos. Las salidas de estos algoritmos son categorías discretas o clases que pueden ser binarias o multiclase, tal y como ocurre en este proyecto.

Para comprender correctamente la implementación de los algoritmos de predicción, se comentan las herramientas, lenguajes y librerías utilizadas en la sección 3.1. Posteriormente, en la sección 3.2 se explican los modelos escogidos junto con sus parámetros. A continuación, en

la sección 3.3 se explican los pasos a seguir para realizar el entrenamiento de los modelos. Finalmente, en la sección 3.4 se detallan los resultados obtenidos en el entrenamiento.

3.1. Herramientas, lenguajes y librerías

El desarrollo de esta parte del proyecto se ha realizado en Google Colab[1], un servicio gratuito en la nube que permite crear y compartir cuadernos interactivos. La implementación se realizó en Python [2], y se utilizaron algunas librerías que se detallan a continuación:

- Pandas[3]: Es un paquete de Python que proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para trabajar con datos relacionales o etiquetados de manera sencilla e intuitiva.
- Matplotlib[4]: Es una biblioteca para crear archivos estáticos, animados y visualizaciones interactivas en Python.
- Numpy[5]: Es un paquete para matrices N-dimensionales, funciones matemáticas y cálculo numérico con Python.
- Seaborn[6]: Es una biblioteca de visualización de datos de Python basada en matplotlib. Proporciona una interfaz de alto nivel para dibujar gráficos estadísticos atractivos e informativos.
- Scikit-learn (sklearn)[7]: Es una librería de aprendizaje automático de código abierto para Python. Es una de las más utilizadas debido a su facilidad de uso y a la amplía variedad de funcionalidades que proporciona.

3.2. Modelos y búsqueda de parámetros

De cara a optimizar el rendimiento de los modelos, se realiza una búsqueda exhaustiva de un conjunto de parámetros para un modelo concreto mediante el uso de la función *GridSearchCV*[8] de *sklearn*. Esta búsqueda realiza validación cruzada, permitiendo evaluar los modelos mediante la división del conjunto de datos en varios subconjuntos para que el modelo entrene y evalúe en diferentes particiones.

[1] https://colab.research.google.com/
[2] https://www.python.org/
[3] https://pandas.pydata.org/
[4] https://matplotlib.org/
[5] https://numpy.org/
[6] https://seaborn.pydata.org/
[7] https://scikit-learn.org/stable/index.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

Cabe destacar que en el proceso de implementación de este proyecto se han probado diferentes modelos optando por aquellos que resultaban más familiares para la autora. Además, respecto a los que se describen a continuación, se han probado más hiperparámetros. Para cada uno de ellos, se realiza una descripción general y de los parámetros a utilizar. Además, se comentan los valores probados y los resultados obtenidos con la función *GridSearchCV*.

3.2.1. Regresión logística

La regresión logística estima la probabilidad de que ocurra un evento en función de un conjunto de datos de variables independientes. En este caso, se aplica la regresión logística multinomial en la que se pueden analizar problemas que tienen varios resultados posibles como es el caso de las tres autopercepciones que se distinguen en los trastornos mentales. El uso de este algoritmo destaca por su simplicidad respecto a otros métodos, la velocidad de procesar grandes volúmenes de datos y la flexibilidad para encontrar respuestas con dos o más resultados finitos. Este modelo devuelve valores entre 0 y 1 para la variable dependiente basándose en la siguiente ecuación:

$$f(x) = \frac{1}{1 + e^{-x}}$$

En este caso, existen múltiples variables independientes que afectan al valor de la variable dependiente. Para ello, las fórmulas de regresión asumen que existe una relación lineal entre las variables independientes del conjunto de datos.

Para la implementación de este algoritmo, se utiliza la función *LogisticRegression*[9] de *sklearn* con el parámetro C que indica el nivel de regularización que se aplica al modelo. Se evaluaron los valores 0.1, 1, 10. 100. En el cuadro 4.1 se muestra el mejor parámetro para cada trastorno.

Trastorno	C
Ansiedad	10
Depresión	0.1
Insomnio	0.1
TOC	0.1

Cuadro 4.1: Parámetros regresión logística.

[9] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

3.2.2. MLP

El perceptrón multicapa o Multilayer Perceptron (MLP) es una red neuronal que está compuesta por múltiples capas de neuronas interconectadas. En ella, las salidas de las neuronas de una capa se convierten en entradas de la siguiente. La primera capa se denomina capa de entrada y son las características o variables de las observaciones. Las capas intermedias se llaman capas ocultas y reciben las entradas que transforman utilizando una función de activación. La última capa es la de salida y produce las predicciones del modelo. El algoritmo se basa en la técnica de propagación hacia atrás, que ajusta los pesos de las conexiones de la red para minimizar el error de la predicción entre las salidas producidas por la red y las salidas deseadas. Se ha utilizado la función *MLPClassifier*[10] de *sklearn* con los siguientes parámetros:

- alpha: Controla la regularización L2 que se aplica a los pesos del modelo. Se evaluaron los valores 0.0001, 0.001 y 0.01.
- hidden_layer_sizes: Especifica la estructura de las capas ocultas de la red neuronal. Se evaluaron los valores (50,), (100,) y (50,50).

En el cuadro 4.2 se muestra el mejor parámetro para cada trastorno.

Trastorno	alpha	hidden_layer_sizes
Ansiedad	0.001	(50,)
Depresión	0.001	(50,)
Insomnio	0.001	(50,)
TOC	0.001	(50,)

Cuadro 4.2: Parámetros MLP.

3.2.3. k-Nearest Neighbors

Este algoritmo se presenta en problemas de regresión para datos con etiquetas continuas y en problemas de clasificación para datos con etiquetas discretas, tal y como ocurre en este caso. Su objetivo es buscar los k puntos más cercanos a un punto concreto y devolver el valor que se utiliza con más frecuencia. Este modelo sigue los siguientes pasos:

1. Calcula la distancia de todos los elementos con nuestro punto.
2. Ordena las distancias de menor a mayor.

[10] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

3. Selecciona los k elementos más cercanos.

4. Obtiene la predicción del algoritmo.

Existen diferentes métricas para el cálculo de la distancia que se pueden aplicar en este modelo. A continuación, se comentan las opciones valoradas en este proyecto:

- Distancia euclidiana (p=2): Mide una línea recta entre el punto de consulta y el otro punto que se está midiendo.
- Distancia de Manhattan (p=1): Es la suma de las distancias absolutas entre las coordenadas correspondientes de dos puntos.
- Distancia de Minkowski: Es la forma generalizada de las métricas de distancia euclidiana y de Manhattan. El parámetro p, permite la creación de otras métricas de distancia. La distancia euclidiana se representa con p=2 y la de Manhattan con p=1.

Se ha utilizado la función *KNeighborsClassifier*[11] de *sklearn* con el parámetro *n_neighbors* que indica el número de vecinos que se usan de forma predeterminada para las consultas. Se evaluaron los valores 2,3,5,7,8 y 10. Además, se ha utilizado la opción predeterminada del parámetro *metric* que utiliza la distancia de Minkowski con p=2.

En el cuadro 4.3 se muestra el mejor parámetro para cada trastorno.

Trastorno	n_neighbors
Ansiedad	10
Depresión	2
Insomnio	2
TOC	2

Cuadro 4.3: Parámetros k-Nearest Neighbors.

3.2.4. Árbol de decisión

Un árbol de decisión es una estructura jerárquica que se usa para tomar decisiones y predecir resultados. De esta forma, el árbol comienza en un nodo raíz que no tiene ramas entrantes. Las ramas salientes de este nodo generan los nodos internos, los cuales en función de las características disponibles realizan evaluaciones para formar subconjuntos que se indican mediante nodos hojas. Los nodos hojas representan todos los resultados posibles dentro del conjunto de datos [15]. Este modelo emplea la estrategia divide y vencerás realizando una búsqueda para identificar los puntos de división óptimos.

[11] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Los árboles de decisión son fáciles de interpretar y flexibles, ya que permiten realizar tareas de clasificación y regresión. Por la contra, son propensos al sobreajuste y no generalizan bien en datos nuevos. En este proyecto se utiliza la clase *DecisionTreeClassifier*[12] de *sklearn* con los siguientes parámetros:

- *max_depth*: La profundidad máxima del árbol. Si el valor es None, los nodos se expanden hasta que todas las hojas sean puras o hasta que todas las hojas contengan menos muestras que *min_samples_split*. Se evaluaron los valores None, 5, 10, 15 y 20.

- *min_samples_split*: Número mínimo de muestras necesario para dividir un nodo internos. Se evaluaron los valores 2,3,5 y 10.

En el cuadro 4.4 se muestra el mejor parámetro para cada trastorno.

Trastorno	max_depth	min_samples_split
Ansiedad	5	2
Depresión	5	2
Insomnio	5	2
TOC	5	3

Cuadro 4.4: Parámetros árbol de decisión.

3.2.5. Random Forest

Random Forest es una extensión de los árboles de decisión en los que combina múltiples árboles para alcanzar un único resultado. Gracias a este esquema y a la no correlación entre los árboles existentes, se produce una mejora en el sobreajuste de los datos, lo que implica que el modelo generaliza mejor en nuevos datos no vistos durante el entrenamiento. Se utiliza la función *RandomForestClassifier*[13] de *sklearn* y se estudian los siguientes parámetros:

- *max_depth*: La profundidad máxima del árbol. En caso de que el valor sea None, ocurre lo mismo que en los árboles de decisión. Se evaluaron los valores None, 10 y 20.

- *n_estimators*: El número de árboles. Se evaluaron los valores 100, 200 y 300.

En el cuadro 4.5 se muestra el mejor parámetro para cada trastorno.

[12] https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html

[13] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Trastorno	max_depth	n_estimators
Ansiedad	10	200
Depresión	10	100
Insomnio	10	300
TOC	10	200

Cuadro 4.5: Parámetros Random Forest.

3.3. Entrenamiento

Una vez encontrados los mejores parámetros, se realiza el entrenamiento y validación de los datos mediante el uso de *K-Fold*. Esta técnica de validación cruzada divide el conjunto de datos en k subconjuntos diferentes. Luego, evalúa k veces estos subconjuntos obteniendo coeficientes de rendimiento en cada una de las iteraciones. Finalmente, se calcula el promedio de las iteraciones individuales para obtener el rendimiento general del modelo. Este método permite realizar una evaluación más consistente y confiable del modelo, ya que cada uno de los subconjuntos se utiliza tanto para entrenar como para el test.

Por lo tanto, se realiza el entrenamiento para cada una de las 4 variables dependientes aplicando los algoritmos comentados anteriormente con sus parámetros óptimos. El modelo obtendrá resultados tanto del conjunto de datos de entrenamiento como del conjunto de prueba, lo que permite evaluar ambos conjuntos y extraer conclusiones. Una vez completado el entrenamiento del modelo, se aplican diversas métricas para evaluar su funcionamiento. Estas métricas están integradas en la biblioteca *sklearn*, y son las siguientes:

- *Accuracy*[14]: Mide la exactitud general del modelo. Se calcula como el número de predicciones correctas dividido por el número total de predicciones realizadas. Su rango de valores se encuentra entre 0 y 1. Se calcula de la siguiente forma:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

- *Recall*[15]: Indica la proporción de ejemplos positivos que están identificados correctamente por el modelo entre todos los positivos reales. Su rango de valores se encuentra entre 0 y 1. Se calcula de la siguiente forma:

$$Recall = \frac{\text{VP}}{\text{VP} + \text{FN}}$$

[14] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html
[15] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.recall_score.html

Donde VP es el número verdaderos positivos y FN el número de falsos negativos.

- *F1 score*[16]: Combina las métricas de Precision y Recall para dar un único resultado. Su rango se encuentra entre 0 y 1. Se calcula de la siguiente forma:

$$F1 = \frac{2 * \text{VP}}{2 * \text{VP} + \text{FP} + \text{FN}}$$

Donde VP es el número de verdaderos positivos, FP el número de falsos positivos y FN el número de falsos negativos.

3.4. Resultados

A continuación, se detallan los valores de las métricas obtenidos para cada uno de los trastornos, según los modelos utilizados. A partir de ellos, se extraen conclusiones sobre el modelo que mejor funciona en cada caso.

3.4.1. Ansiedad

Los valores obtenidos se pueden observar en el cuadro 4.6. Tal y como se puede ver, *k-Nearest Neighbors* es el mejor modelo, obtiene para el conjunto de entrenamiento 0.730 en *accuracy*, 0.734 en *recall* y 0.736 en *F1 score*. En el conjunto de test, obtiene unos resultados ligeramente inferiores, 0.669, 0.675 y 0.676 para *accuracy*, *recall* y *F1 score*, respectivamente. Por otra parte, *Random Forest* y *MLP* obtienen unos valores inferiores en todas sus métricas. En el caso del conjunto de entrenamiento, los valores se encuentran entre 0.55 y 0.61 y en test entre 0.54 y 0.6. Los modelos que presentan unos resultados más bajos son los árboles de decisión y regresión logística en la que ninguno de los conjuntos tiene valores en sus métricas superiores a 0.451.

	Entrenamiento			Test		
Modelo	Accuracy	Recall	F1 score	Accuracy	Recall	F1 score
Logistic Regression	0.427	0.344	0.250	0.427	0.344	0.249
MLP	0.561	0.552	0.559	0.558	0.549	0.556
k-Nearest Neighbors	**0.730**	**0.734**	**0.736**	**0.669**	**0.675**	**0.676**
Decision Tree	0.451	0.393	0.387	0.451	0.392	0.368
Random Forest	0.605	0.562	0.577	0.596	0.553	0.567

Cuadro 4.6: Resultados ansiedad.

[16] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

3.4.2. Depresión

Los valores obtenidos se pueden observar en el cuadro 4.7. En este trastorno, el mejor modelo es *k-Nearest Neighbors*. En el conjunto de entrenamiento obtiene 0.788 en *accuracy*, 0.755 en *recall* y 0.769 en *F1 score*. En el conjunto de test, los resultados son 0.647, 0.612 y 0.616 para *accuracy*, *recall* y *F1 score*, respectivamente.*Random Forest* y *MLP* obtienen valores más bajos en todas sus métricas. En el caso del conjunto de entrenamiento, los valores se encuentran entre 0.56 y 0.63 y en test entre 0.55 y 0.621. Los resultados más bajos se consiguen en los árboles de decisión y regresión logística en la que no hay valores superiores a 0.47 en ninguna de sus métricas.

	Entrenamiento			Test		
Modelo	Accuracy	Recall	F1 score	Accuracy	Recall	F1 score
Logistic Regression	0.440	0.381	0.339	0.439	0.382	0.339
MLP	0.589	0.560	0.565	0.586	0.557	0.561
K Neighbors	**0.788**	**0.755**	**0.769**	**0.647**	**0.612**	**0.616**
Decision Tree	0.468	0.426	0.417	0.468	0.426	0.416
Random Forest	0.630	0.578	0.578	0.621	0.569	0.567

Cuadro 4.7: Resultados depresión.

3.4.3. Insomnio

Los valores obtenidos se pueden observar en el cuadro 4.8. Se puede ver que en este caso el mejor modelo también es *k-Nearest Neighbors*. Obtiene para el conjunto de entrenamiento 0.816 en *accuracy*, 0.737 en *recall* y 0.769 en *F1 score*. En el conjunto de test, obtiene unos resultados ligeramente inferiores, 0.698, 0.605 y 0.625 para *accuracy*, *recall* y *F1 score*. En el caso de *Random Forest* y *MLP*, se obtienen unos valores inferiores en todas sus métricas. En el conjunto de entrenamiento, los valores se encuentran entre 0.46 y 0.62 y en test entre 0.45 y 0.62. Los modelos con resultados más bajos son los árboles de decisión y regresión logística en la que ninguno de los valores supera el 0.532.

	Entrenamiento			Test		
Modelo	Accuracy	Recall	F1 score	Accuracy	Recall	F1 score
Logistic Regression	0.523	0.341	0.260	0.523	0.341	0.260
MLP	0.620	0.536	0.545	0.617	0.532	0.540
k-Nearest Neighbors	**0.816**	**0.737**	**0.769**	**0.698**	**0.605**	**0.625**
Decision Tree	0.532	0.357	0.292	0.531	0.356	0.291
Random Forest	0.618	0.468	0.466	0.613	0.461	0.458

Cuadro 4.8: Resultados insomnio.

3.4.4. TOC

Los valores obtenidos se pueden observar en el cuadro 4.9. Tal y como se puede ver, *k-Nearest Neighbors* vuelve a ser el mejor modelo. Obtiene para el conjunto de entrenamiento 0.853 en *accuracy*, 0.710 en *recall* y 0.762 en *F1 score*. En el conjunto de test, obtiene unos resultados ligeramente inferiores, 0.766, 0.586 y 0.621 para *accuracy*, *recall* y *F1 score*. Por otra parte, *MLP* tiene unos valores inferiores en todas sus métricas, con un 0.556 de *F1 score* en entrenamiento y 0.551 en test. En este caso, *Random Forest* alcanza valores más bajos en *F1 score*, siendo 0.391 en entrenamiento y 0.383 en test. Finalmente, los árboles de decisión y la regresión logística obtienen los valores más bajos con un *F1 score* que no supera el 0.3 en ninguno de los conjuntos.

	Entrenamiento			Test		
Modelo	Accuracy	Recall	F1 score	Accuracy	Recall	F1 score
Logistic Regression	0.670	0.333	0.267	0.670	0.333	0.267
MLP	0.725	0.527	0.556	0.722	0.524	0.551
k-Nearest Neighbors	**0.853**	**0.710**	**0.762**	**0.766**	**0.586**	**0.621**
Decision Tree	0.674	0.343	0.291	0.674	0.343	0.292
Random Forest	0.700	0.398	0.391	0.698	0.394	0.383

Cuadro 4.9: Resultados TOC.

3.5. Conclusiones

Una vez realizado el estudio completo de los algoritmos predictivos se puede afirmar que el modelo que ha obtenido mejores resultados en todos los trastornos es *k-Nearest Neighbors*, tal y como se muestra en el cuadro 4.10, en la que se recoge el mejor modelo de cada trastorno. El trastorno para el que obtiene unos resultados mejores a nivel de *F1 score* es la ansiedad. No obstante, los valores obtenidos son similares en todos los trastornos, por lo que podemos decir que este algoritmo consigue predecir la variable dependiente y confirmar que mediante los géneros musicales escuchados se puede predecir el nivel que sufre cada encuestado en estos trastornos.

Trastorno	Modelo	Accuracy	Recall	F1 score
Ansiedad	k-Nearest Neighbors	0.669	0.675	0.676
Depresión	k-Nearest Neighbors	0.647	0.612	0.616
Insomnio	k-Nearest Neighbors	0.698	0.605	0.625
TOC	k-Nearest Neighbors	0.766	0.586	0.621

Cuadro 4.10: Resultados test.

4. Power BI

Power BI es una colección de servicios de software, aplicaciones y conectores que funcionan conjuntamente para convertir orígenes de datos sin relación entre sí en información coherente, interactiva y atractiva visualmente [11]. Power BI cuenta con tres elementos diseñados para crear, compartir y usar información de manera eficaz. Estos elementos son:

1. Aplicación de escritorio de Windows denominado Power BI Desktop.
2. Servicio de software como servicio (SaaS) en línea llamado servicio Power BI.
3. Aplicaciones para Power BI Mobiles para dispositivos Windows, iOS y Android.

Para empezar a usar Power BI, el flujo habitual comienza con la conexión de datos en Power BI Desktop y la creación de un informe. Posteriormente, ese informe se puede publicar en el servicio y compartir para que los usuarios puedan verlo e interactuar con él.

Para este proyecto se desarrollan dos Power BI diferentes. El primero de ellos se basa en la colección "mentalhealth" y pretende mostrar de manera visual la información que se ha recogido a través de las encuestas. Por otro lado, se desarrolla un Power BI que refleja los resultados obtenidos en los algoritmos de predicción.

4.1. Power BI encuestas

En este primer archivo, es necesario que el origen datos que se utiliza en el modelo sea MongoDB. Este origen no está disponible en Power BI, por lo que es necesario realizar una configuración adicional [16]. En primer lugar, es necesario instalar el conector BI de MongoDB[17] y un driver de MongoDB[18]. A continuación, es necesario consultar el puerto que se está utilizando en el archivo "mongosqld" del conector BI, tal y como se puede observar en la figura 4.7. Finalmente, se añade la conexión de ODBC de 64 bits con la configuración mostrada en la figura 4.8 con el puerto visto anteriormente. Una vez completada la configuración anterior, se puede añadir como origen de datos en Power BI el ODBC y cargar las colecciones necesarias.

Este archivo consta de tres páginas. La primera de ellas muestra información general recogida en las encuestas, tal y como se puede observar en la figura 4.9. En el margen izquierdo de la página, se muestran filtros sobre los campos generales de la encuesta. A su derecha se encuentra el porcentaje de cumplimiento de cada uno de esos campos. En los gráficos, se puede observar

[17] https://www.mongodb.com/docs/bi-connector/current/
[18] https://github.com/mongodb/mongo-bi-connector-odbc-driver/releases/

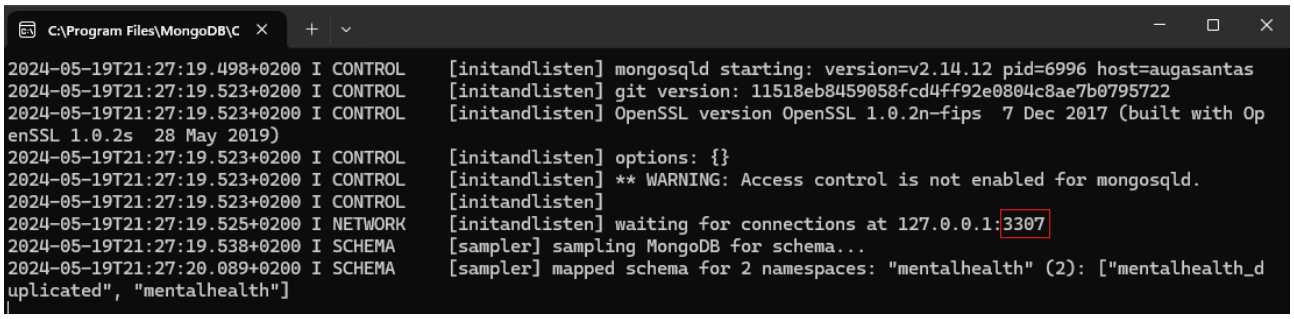

Figura 4.7: Conector BI MongoDB. Puerto.

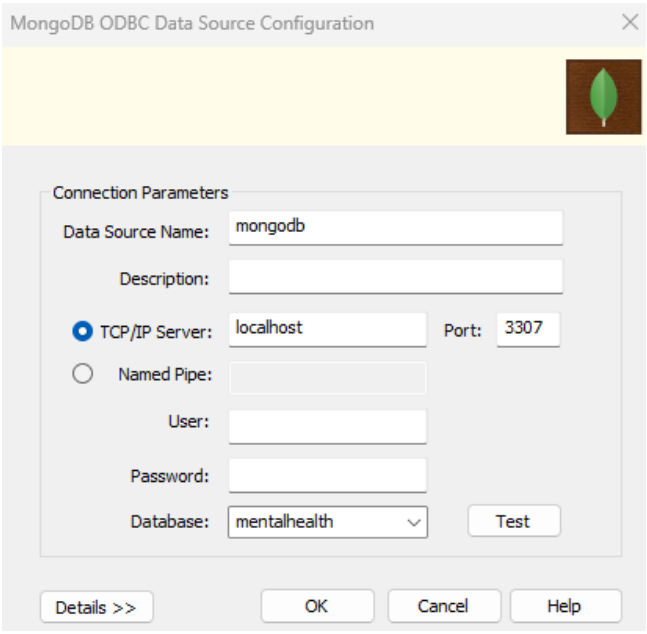

Figura 4.8: MongoDB conexión ODBC.

la distribución de valores tanto por edad como por número de horas diarias de escucha. En la parte derecha de la pestaña, se pueden ver dos tarjetas en la parte superior que incluye la media de estos valores, viendo que la edad media de los participantes es 25.19 y la media diaria de escucha son 3.58 horas. Se muestran también dos gráficos con las principales plataformas de escucha en la que destaca con gran diferencia Spotify y la percepción de los usuarios de como afecta la música en la salud mental, obteniendo mayoritariamente como respuesta la mejora de

la misma.

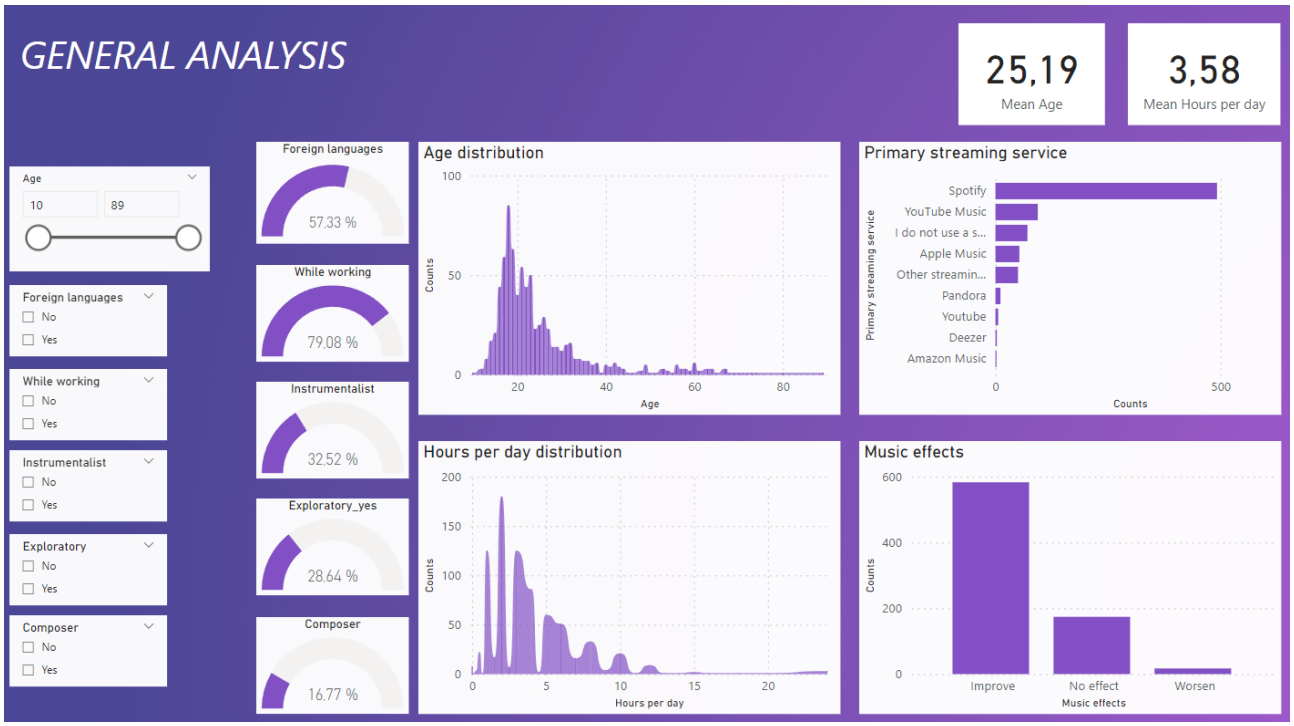

Figura 4.9: Power BI encuestas. Página “Genres and mental disorders”.

La segunda página, tal y como se observa en la figura 4.10, presenta en unas tarjetas los valores medios de cada uno de los trastornos de salud mental. Se puede observar que el trastorno con una mayor media en la autopercepción de los usuarios es la ansiedad con un 5.74 y la menor el TOC con 2.59. Además, en la primera mitad superior de la página, se muestra la información relativa a los géneros musicales y a sus frecuencias de escucha, permitiendo al usuario filtrar en el margen izquierdo los valores que desee para poder realizar comparaciones entre los géneros. La parte inferior está dedicada a los trastornos mentales, viendo también su distribución por edad y por el nivel de autopercepción de los usuarios. De igual forma, se presentan los filtros en la parte izquierda que permiten seleccionar los trastornos y el grado de autopercepción.

Finalmente, la última página mostrada en la figura 4.11, pretende mostrar la relación que existe entre la frecuencia de escucha de los géneros musicales y los trastornos mentales. Para ello se han incluido dos filtros, uno que permite mostrar los géneros deseados por el usuario y otro que incluye las frecuencias de escucha. De esta manera, para cada uno los trastornos se muestra la autopercepción de los participantes teniendo en cuenta los géneros y frecuencias seleccionadas en los filtros. Así, se puede saber que para las personas que nunca escuchan

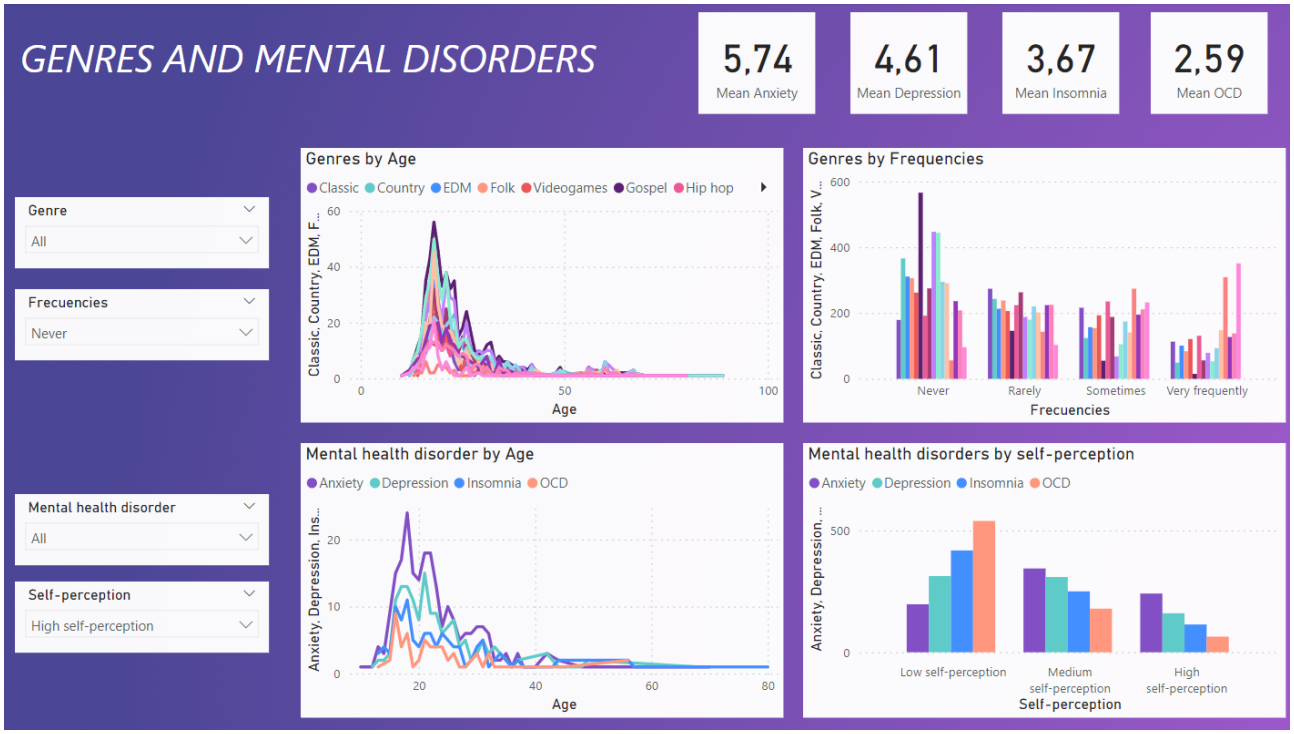

Figura 4.10: Power BI encuestas. Página “General analysis”.

Gospel, la autopercepción más común es la media en la ansiedad. En cambio, en la depresión la percepción está más igualada entre los valores bajo y medio para las personas que escuchan Gospel con esta frecuencia.

4.2. Power BI resultados algoritmos

Este *dashboard* pretende mostrar de manera visual los resultados obtenidos en los algoritmos predictivos. Para ello, se ha completado un CSV que recoge los valores de las métricas para cada uno de los modelos en los distintos trastornos y que se utiliza como origen de datos en el nuevo archivo. La información se visualiza en una única página y pretende que el usuario pueda identificar de manera sencilla que modelo ha obtenido un rendimiento mayor en los algoritmos. Se puede observar en la figura 4.12 el informe resultante.

En la parte superior se pueden observar unas tarjetas que recogen el valor medio de las métricas de todos los trastornos y algoritmos en los dos conjuntos de datos utilizados, prueba y test. De esta manera, se puede ver que en términos generales no existe discrepancia significativa entre ambos conjuntos, lo que indica que el entrenamiento se ha realizado correctamente y sin sobreajustes. No obstante, inmediatamente debajo de estas tarjetas tenemos dos nuevas

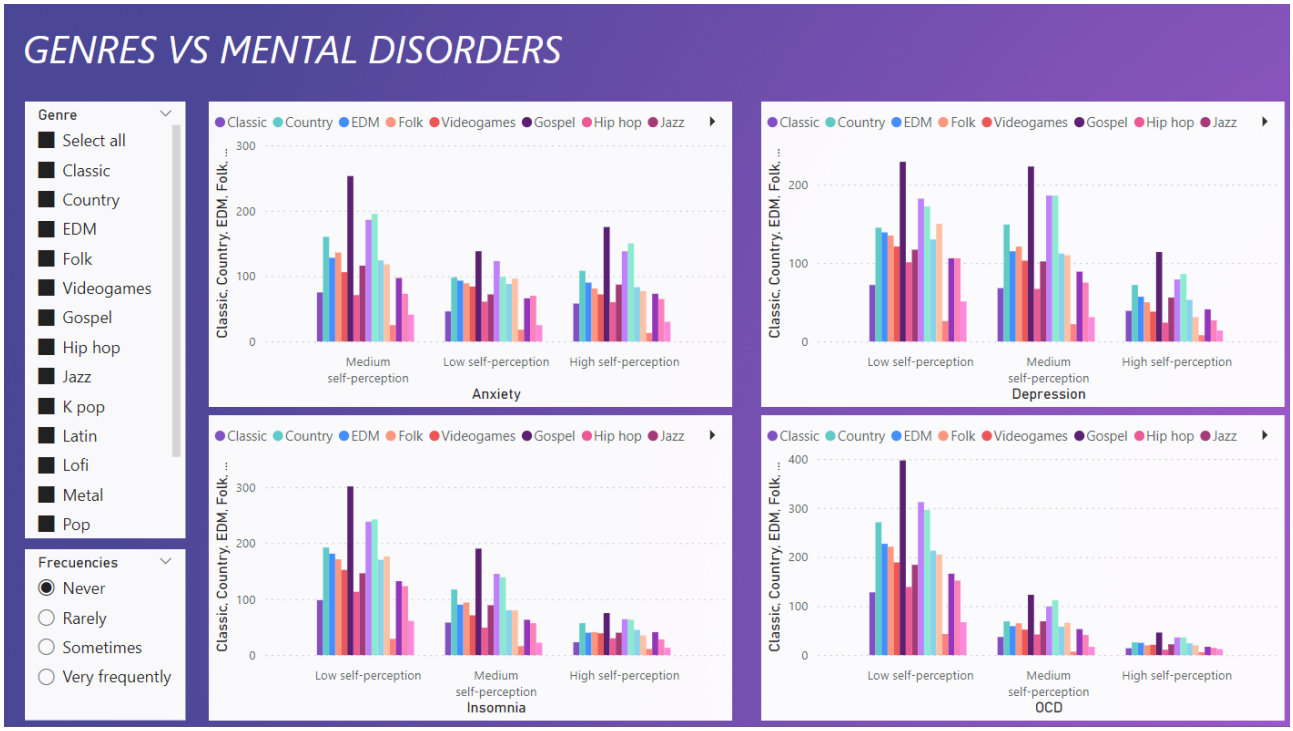

Figura 4.11: Power BI encuestas. Página "Genres vs mental disorders".

secciones. En primer lugar, en el margen izquierdo existe un filtro de selección única que permite al usuario escoger el trastorno que quiere analizar. A la derecha del filtro, existen tres gráficos, uno para cada métrica que permiten comparar para cada modelo el resultado obtenido en entrenamiento y en test, lo que nos permite ver si realmente existe diferencia entre estos conjuntos y se pudo dar un sobreajuste en el conjunto de datos de entrenamiento.

En la parte inferior del *dashboard*, se encuentran unas gráficas similares a las que se comentaron anteriormente, pero en este caso el filtro nos deja seleccionar un modelo concreto. Las gráficas muestran los diferentes de manera simultánea, lo que permite comparar qué trastorno tuvo mejores resultados, tanto en entrenamiento como en test en un modelo determinado.

Figura 4.12: Power BI resultados algoritmos.

Capítulo 5

Conclusiones y líneas futuras

En este último capítulo, se van a comentar en la sección 1 las conclusiones obtenidas tras el desarrollo del proyecto. Posteriormente, en la sección 2 se detallan posibles líneas de trabajo futuro que no forman parte del alcance de este Trabajo de Fin de Máster y que pueden resultar interesantes para proyectos posteriores.

1. Conclusiones

Este proyecto tiene como finalidad estudiar la relación entre las preferencias musicales de los usuarios y algunos trastornos de salud mental. De cara a sacar conclusiones sobre el desarrollo realizado, resulta útil revisar los objetivos marcados inicialmente para ver el grado de cumplimiento de cada uno de ellos y los resultados obtenidos.

En primer lugar, se cumple con el almacenamiento en MongoDB de los datos recogidos en la encuesta de la autora del proyecto junto con el conjunto de datos descargado. Estos datos se almacenan en una base de datos formada por dos colecciones, una con el volumen de datos original de las encuestas y otra con una ampliación del volumen da datos, ya que el volumen recogido no fue suficiente y se decidió realizar una ampliación de los datos realizando variaciones en las variables utilizadas para la predicción de los datos. Además, se lleva a cabo una limpieza de los datos, así como un análisis exploratorio inicial para ver la distribución de los mismos.

Por otro lado, se realiza un estudio con diferentes algoritmos de predicción sobre el conjunto de datos ampliado. Para ello, se hace una búsqueda de los mejores parámetros a aplicar en cada modelo y se entrena el modelo aplicando el método *K-Fold*. Los resultados obtenidos son similares en todos los trastornos mentales, obteniendo los mejores resultados según las métricas valoradas, *accuracy*, *recall* y *F1 score*, en el modelo *k-Nearest Neighbors*.

Finalmente, se desarrollan dos informes en Power BI. El primero de ellos, refleja la información recogida en las encuestas de manera visual, permitiendo sacar conclusiones realistas sobre los datos almacenados. Adicionalmente, se desarrolla otro informe que muestra los valores de las métricas obtenidas en el entrenamiento de los modelos.

Por tanto, se puede decir que se ha cumplido con todos los objetivos marcados al inicio del proyecto, realizando un estudio completo de la relación entre las preferencias musicales y algunos trastornos mentales.

2. Trabajo futuro

A lo largo del desarrollo de este proyecto, han surgido numerosas ideas para su ampliación, las cuales resultan interesantes y permiten tener un conocimiento más detallado de la relación que existe entre las preferencias musicales y los trastornos mentales. En esta sección se presentan algunas propuestas para el trabajo futuro, enfocándose en la ampliación de los valores y el volumen de datos recogidos en las encuestas, la mejora de los algoritmos de predicción, la actualización de datos y la colaboración con profesionales de los sectores implicados.

En primer lugar, es importante la ampliación del conjunto de datos mediante la realización de la encuesta a un mayor número de usuarios de diferentes perfiles. Para ello, la encuesta sería difundida mediante la redes sociales, en colegios, universidades o conservatorios. Podría enviarse por correo electrónico a diferentes listas de usuarios relacionadas con la temática del estudio. Además, podría compartirse en organizaciones de salud mental y en tiendas de música.

En la encuesta sería interesante añadir nuevos campos, como el sexo y el lugar de origen, que aportarían gran valor a los datos y permitirían obtener similitudes y diferencias entre los perfiles de los participantes. Por otro lado, también sería útil que los usuarios que participaron en la encuesta la repitieran varias veces a lo largo del tiempo para ver si los cambios en los hábitos de escucha influyen en la autopercepción de los trastornos.

De cara a la aplicación de los algoritmos de predicción, se podría estudiar el uso de otras técnicas o modelos que ofrezcan ventajas en un conjunto de datos de gran volumen. Además, también resultaría interesante realizar la validación de datos con un conjunto que utilice diagnósticos clínicos y no basados en la autopercepción de los usuarios.

Por otro lado, si la encuesta se difunde en lugares en los que esté presente de manera permanente o se hace llegar en diferentes periodos, podría ser interesante que la carga de los datos se

realizara de manera programada para cargar los nuevos registros. Esta actualización programada se establecería también en Power BI, de manera que se podría ver una evolución a lo largo del tiempo tanto de las preferencias musicales como de los trastornos y la relación que existe entre ambos.

Por último, los resultados obtenidos podrían ser compartidos con las plataformas de *streaming* para promover la salud mental de los usuarios. Además, sería interesante que este estudio pudiera ser distribuido a profesionales de la salud mental para que puedan validar los resultados obtenidos y, al mismo tiempo, vez trabajar con estos datos de manera útil durante las terapias.

Bibliografía

[1] D. Telles-Correia, S. Saraiva, and J. Gonçalves, "Mental disorder—the need for an accurate definition," *Frontiers in Psychiatry*, vol. 9, 2018. [Online]. Available: https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsyt.2018.00064

[2] O. M. de la Salud, "Trastornos mentales," 2022. [Online]. Available: https://www.who.int/es/news-room/fact-sheets/detail/mental-disorders

[3] R. A. Española, "Diccionario de la lengua española," 11 de junio de 2024. [Online]. Available: https://www.rae.es/

[4] K. University, "Kanban University Home — Kanban University — kanban.university," 11 de junio de 2024. [Online]. Available: https://kanban.university/kanban-guide/

[5] E. V. Zamora, I. Introzzi, M. del Valle, and M. Richard's, "Marco teórico del efecto de interferencia en contextos neutrales y emocionales," *Escritos de Psicología (Internet)*, vol. 13, pp. 23 – 33, 06 2020. [Online]. Available: http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1989-38092020000100003&nrm=iso

[6] J. A. C. Meneses and M. M. Díaz, "Influencia del tempo de la música en las emociones," *Revista Colombiana de Psicología*, vol. 19, no. 1, pp. 37–44, 2010.

[7] G. y. M. D. Scaringella, N. y Zoia, "Clasificación automática de géneros del contenido musical: una encuesta," *Revista de procesamiento de señales IEEE*, vol. 23, no. 2, 2006.

[8] F. Baker and W. Bor, "Can music preference indicate mental health status in young people?" *Australasian Psychiatry*, vol. 16, no. 4, pp. 284–288, 2008, pMID: 18608148. [Online]. Available: https://doi.org/10.1080/10398560701879589

[9] V. Laijawala, A. Aachaliya, H. Jatta, and V. Pinjarkar, "Classification algorithms based mental health prediction using data mining," in *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1174–1178.

[10] J. Virani, N. Daredi, A. Bhanushali, M. Shukla, and P. Shah, "Mental healthcare analysis using power bi machine learning," in *2023 4th International Conference on Signal Processing and Communication (ICSPC)*, 2023, pp. 73–76.

[11] M. 2024, "¿qué es power bi?" 11 de junio de 2024. [Online]. Available: https://learn.microsoft.com/es-es/power-bi/fundamentals/power-bi-overview

[12] M. Deziel, D. Olawo, L. Truchon, and L. Golab, "Analyzing the mental health of engineering students using classification and regression," in *Educational Data Mining 2013*, 2013.

[13] K. E. Bruscia, *Musicoterapia*. Editorial Pax México, 2007.

[14] MongoDB, "¿qué es mongodb?" 11 de junio de 2024. [Online]. Available: https://www.mongodb.com/es/company/what-is-mongodb

[15] B. Charbuty and A. Abdulazeez, "Classification based on decision tree algorithm for machine learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20 – 28, Mar. 2021. [Online]. Available: https://www.jastt.org/index.php/jasttpath/article/view/65

[16] MongoDB, "Connect from microsoft power bi desktop," 11 de junio de 2024. [Online]. Available: https://www.mongodb.com/docs/bi-connector/current/connect/powerbi/