



Universidad Politécnica
de Madrid

**Escuela Técnica Superior de
Ingenieros Informáticos**



Máster en Data Science

Trabajo Fin de Máster

ActiveMind

Autor(a): Carlos David Olan Olvera

Madrid, Julio, 2025

Este Trabajo Fin de Máster se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid.

*Trabajo Fin de Máster
Máster en Data Science*

*Título: ActiveMind
Julio, 2025*

*Autor(a): Carlos David Olan Olvera
Tutor(a): Emilio Serrano
Tecnologías de la Información y la Comunicación
ETSI Informáticos
Universidad Politécnica de Madrid*

Resumen

Este Trabajo Fin de Máster documenta la fase inicial de *ActiveMind*: el diseño de la encuesta, la depuración de las respuestas y el análisis de *clústeres* que permitirán, en etapas posteriores, personalizar recomendaciones musicales a partir de métricas de salud. Se diseñó un cuestionario en Google Forms por su accesibilidad, capacidad bilingüe y visualización inmediata de resultados :contentReference[oaicite:0]index=0. Tras un piloto exploratorio, se recolectaron 157 respuestas en español (y 9 en inglés, descartadas), lo que justificó concentrar el estudio en la muestra hispanohablante :contentReference[oaicite:1]index=1.

Estructura de la encuesta. Se construyeron cinco bloques: datos demográficos, hábitos de actividad física, uso de tecnología /wearables, preferencias musicales y actitudes frente a la personalización y la salud mental :contentReference[oaicite:2]index=2. Esta selección permitió perfilar necesidades funcionales de la futura app y detectar un nicho proclive a la integración IA-bienestar :contentReference[oaicite:3]index=3.

Extracción y preprocesamiento. Las respuestas se exportaron a Excel :contentReference[oaicite:4]index=4 y se sometieron a un flujo de calidad: verificación de formatos, eliminación de registros inconsistentes, normalización semántica de campos abiertos (p.ej., Regueton , Reguetón) y codificación numérica de variables ordinales (0–4) :contentReference[oaicite:5]index=5. Al tratarse de preguntas obligatorias, no se registraron valores faltantes, lo que simplificó la imputación.

Primeros análisis descriptivos. La muestra se compone mayoritariamente de personas de 35–44 años (35) con leve sesgo masculino (57). El 70 hace ejercicio al menos tres veces por semana y la motivación principal es “aumentar energía” (57) . Cardio y fuerza dominan las modalidades (40,8), mientras que el género *pop* lidera las preferencias musicales (44,6) :contentReference[oaicite:6]index=6. La importancia media asignada a la salud mental es alta (3,43/4), lo que valida el enfoque emocional del proyecto :contentReference[oaicite:7]index=7. Mapas de calor y un PCA preliminar se usaron para explorar correlaciones e inercia acumulada; estos análisis señalaron la comodidad con la personalización y el uso de tecnología como ejes diferenciadores :contentReference[oaicite:8]index=8.

Análisis de clústeres. Se aplicaron tres técnicas complementarias—k-means, clustering jerárquico aglomerativo (Ward) y DBSCAN—y sus hiperparámetros se calibraron con coeficiente de silueta, método del codo y *gap statistic*. El den-

drograma ward reveló una separación natural en tres grupos :contentReference[oaicite:9]index=9.

Nombrado interpretable. Para bautizar los grupos sin fines predictivos, se entrenó un árbol de decisión raso que expone reglas como alta comodidad + alta valoración de salud mental Entusiasta, facilitando la comunicación de hallazgos a expertos de negocio :contentReference[oaicite:12]index=12.

Conclusión. La creación cuidadosa de la encuesta, el preprocesamiento exhaustivo y el análisis de *clústeres* proporcionan una comprensión granular de los futuros usuarios de *ActiveMind*. Estos insights sientan las bases para desarrollar—en fases posteriores—algoritmos de recomendación y prototipos de interfaz que capitalicen la segmentación psicotecnológica aquí descubierta.“

Abstract

El presente Trabajo Fin de Máster describe la fase exploratoria de ActiveMind, una futura aplicación que combinará inteligencia artificial, métricas fisiológicas de wearables y preferencias musicales para mejorar el bienestar durante la actividad física. El estudio se centra en el diseño y validación de un cuestionario que recabó 157 respuestas consistentes sobre demografía, hábitos de ejercicio, uso tecnológico y actitudes frente a la personalización sonora.

Tras un exhaustivo preprocesamiento (normalización semántica, codificación ordinal y detección de atípicos), se efectuó un análisis descriptivo integral y se aplicaron técnicas de agrupamiento (k-means, Ward y DBSCAN). La calidad de los clústeres se evaluó mediante coeficiente de silueta, método del codo, convergiendo en tres segmentos distintivos: Poco tecnológicos, Mixtos y Entusiastas. Un árbol de decisión raso se utilizó exclusivamente para bautizar e interpretar los grupos, sin pretensión predictiva.

Los hallazgos evidencian que la apertura a la personalización musical y la importancia otorgada al bienestar mental son los vectores que mejor diferencian a los futuros usuarios de ActiveMind. Esta segmentación psicotecnológica proporciona una base sólida para el desarrollo posterior de modelos de recomendación en tiempo real y aplicaciones de salud digital centradas en el usuario.

Tabla de contenidos

1. Introducción	1
1.1. Objetivos	2
1.1.1. Objetivo General	2
1.1.2. Objetivos Específicos	2
2. Estado del arte	3
2.1. Importancia y Relevancia de la Recolección de Datos	3
2.2. Importancia de la calidad y diversidad de los datos en sistemas de recomendación	3
2.3. Descripción de los Datos y Metodologías de Procesamiento de Datos	4
2.3.1. Procesamiento y Tratamiento de Datos	6
2.3.2. Creación de Modelos con Feedback Interpretable	6
2.3.3. Análisis y segmentación de usuarios	6
2.3.4. Importancia de la Interpretación de los Datos	7
2.3.5. Otras Aproximaciones en la Personalización Musical	7
2.4. Integración con Dispositivos Wearable	8
2.5. Conclusiones del Estado del Arte	8
3. Propuesta	9
3.1. Contexto y Motivación de la Propuesta	9
3.2. Diseño del Estudio	10
3.2.1. Creación del Instrumento de Recolección de Datos	10
3.2.2. Justificación del Diseño y Selección de Preguntas	10
3.2.3. Uso de Google Forms como Plataforma de Encuesta	11
3.2.4. Primeras Impresiones de las Respuestas	11
3.3. Procesamiento y Análisis de los Datos Estructurados	12
3.3.1. Exportación y Verificación Inicial de Datos	12
3.3.2. Limpieza y Preprocesamiento de la Información	12
3.3.3. Análisis Descriptivo de la Información	14
3.3.4. Herramientas y técnicas de análisis	16
4. Metodología e Implementación	19
4.1. Preprocesamiento y transformación de variables	19
4.2. Normalización sistemática de preguntas abiertas	20
4.3. Visualización y Exploración de Datos	21

TABLA DE CONTENIDOS

4.4. Implementación y justificación del análisis de clustering y clasificación	22
4.4.1. Visualización de Silueta por Cluster	23
4.4.2. Clustering Jerárquico	24
4.4.3. Asignación de Clusters y Caracterización Avanzada	25
4.4.4. Validación Robusta mediante Bootstrap y Adjusted Rand Index (ARI)	25
4.4.5. Reducción de Dimensionalidad mediante PCA	26
4.4.6. Clasificadores Supervisados	26
5. Evaluación y resultados	29
5.0.1. Determinación del número óptimo de clusters	29
5.0.2. Evaluación de la calidad del clustering	30
5.0.3. Caracterización de Clusters mediante Distribución de Variables	32
5.0.4. Reducción de dimensionalidad y visualización con PCA	34
5.0.5. Evaluación de clasificadores supervisados para replicación de clusters	35
5.0.6. Evaluación de desempeño de clasificadores supervisados mediante matrices de confusión	36
6. Conclusiones y trabajo futuro	39
6.1. Conclusiones	39
6.2. Trabajo futuro	39

Índice de figuras

2.1. Pipeline metodológico general del preprocesamiento y análisis de datos de ActiveMind.	5
4.1. Mapa de calor de correlaciones entre variables numéricas.	22
5.1. Método del codo para determinación del número óptimo de clusters. Se observa un cambio de pendiente marcado en $K = 3$, sugiriendo este valor como solución balanceada.	30
5.2. Coeficiente de silueta promedio por número de clusters. El máximo local en $K = 3$ valida la cohesión y separación relativa de los clusters en este modelo.	30
5.3. Silhouette plot por cluster para $K = 3$. Se observa una distribución mayoritariamente positiva de los coeficientes, indicando cohesión interna y adecuada separación relativa entre clusters.	31
5.4. Dendrograma de clustering jerárquico aglomerativo (<i>Ward</i>). Se identifican tres conglomerados principales, consistentes con la solución obtenida mediante K-Means.	32
5.5. Distribución de variables numéricas para Cluster 1. Se observa un perfil balanceado con alta valoración del bienestar y disposición tecnológica intermedia.	33
5.6. Distribución de variables numéricas para Cluster 2. Presenta altos niveles de comodidad con apps inteligentes y menor actividad física promedio.	33
5.7. Distribución de variables numéricas para Cluster 0. Se caracteriza por baja apertura tecnológica e interés en personalización, con actividad física ligeramente superior a otros grupos.	34
5.8. Visualización PCA (2D) de los clusters. Se observa una separación clara entre los tres grupos, validando la segmentación desde un enfoque geométrico y estadístico.	35
5.9. Resultados de clasificadores supervisados para replicación de clusters. Gradient Boosting, SVM y Regresión Logística mostraron desempeños perfectos en el conjunto de prueba, confirmando su idoneidad operativa.	36
5.10.*	37
5.11.*	37
5.12.*	37
5.13.*	37

ÍNDICE DE FIGURAS

5.14*	37
5.15 Matrices de confusión de los modelos supervisados entrenados. Todos muestran exactitud perfecta o cercana, reflejando la replicación fiel de la segmentación original sin valor predictivo externo.	37

Capítulo 1

Introducción

En la actualidad, la práctica deportiva se ha convertido en una actividad imprescindible para mantener un estilo de vida saludable y equilibrado. Sin embargo, muchos deportistas buscan ir más allá de los beneficios puramente físicos y desean tener una experiencia integral que involucre motivación, diversión y bienestar emocional. La música ha demostrado ser un factor determinante en el rendimiento, dado que puede influir en el estado de ánimo, la concentración y el nivel de energía de las personas. Con esta premisa, surge la idea de desarrollar **ActiveMind**, una aplicación diseñada para recomendar música personalizada y sincronizada con las métricas de salud de cada usuario, brindando así un acompañamiento óptimo durante la actividad física.

Para lograr esta personalización, se hace necesario comprender en profundidad el perfil de los potenciales usuarios: su edad, las disciplinas deportivas que practican, el tiempo dedicado al ejercicio, sus preferencias musicales y el uso de dispositivos como *smartwatches*. Estos factores permiten delinear un perfil de consumo y de necesidades específicas, esencial para adaptar la oferta musical a contextos tan diversos como una carrera de larga distancia o una rutina de yoga. Con el fin de recolectar la información necesaria, se han diseñado encuestas dirigidas a un público objetivo que practica deporte con regularidad, recogiendo datos cuantitativos y cualitativos que sirven de base para la construcción de la aplicación.

En primera instancia, se llevó a cabo la recaudación de información a través de encuestas, con énfasis en la organización, la estructura y la definición de los campos que recogen las variables más relevantes. Posteriormente, se abordó el proceso de preprocesamiento de los datos, que incluyó la limpieza, la transformación y la normalización de la información, asegurando su validez y coherencia antes de incorporarla en el clasificador.

Con ello, **ActiveMind** pretende mejorar la experiencia deportiva de las personas, permitiéndoles contar con un acompañamiento musical que no solo se ajuste a sus preferencias, sino que también se integre de manera dinámica con sus condiciones fisiológicas. Este enfoque multidisciplinario combina técnicas de ciencia de datos, análisis de comportamiento humano y desarrollo de aplicaciones, re-

1.1. Objetivos

flejando el carácter integral del proyecto. El presente trabajo describe en detalle el proceso de ideación, diseño y validación de la primera etapa de **ActiveMind**, mostrando cómo la recolección y el tratamiento de datos resultan esenciales para ofrecer un servicio innovador y centrado en el usuario.

1.1. Objetivos

1.1.1. Objetivo General

- OG1: Realizar el análisis y la interpretación de los datos obtenidos a través de encuestas diseñadas para ActiveMind, con el fin de segmentar a los usuarios mediante técnicas de clustering y establecer un sustento sólido para futuras recomendaciones personalizadas de canciones basadas en métricas de salud y bienestar.

1.1.2. Objetivos Específicos

- OE1: Diseñar y aplicar una encuesta estructurada que permita recopilar datos relevantes sobre hábitos, afinidad tecnológica, actividad física y preferencias musicales de los potenciales usuarios de ActiveMind.
- OE2: Implementar un proceso de preprocesamiento de datos que garantice la limpieza, normalización y codificación adecuada de las variables recopiladas, asegurando su calidad para el análisis y su potencial reutilización en proyectos futuros.
- OE3: Aplicar técnicas de clustering no supervisado, como K-Means y clustering jerárquico, para identificar segmentos diferenciados de usuarios, caracterizarlos e interpretar sus patrones de respuesta en el contexto de personalización musical.

Capítulo 2

Estado del arte

En este apartado se presentan los principales avances y aportaciones científicas y tecnológicas que sustentan la creación de una aplicación como **ActiveMind**, cuya función es la recolección, tratamiento y análisis de datos para la recomendación musical personalizada. El objetivo es revisar las distintas soluciones ya propuestas, los enfoques empleados y los resultados obtenidos por otras iniciativas, desarrolladas principalmente entre 2010 y 2024 en contextos de investigación aplicada en Europa, Estados Unidos y Latinoamérica, con el fin de situar este proyecto “*a hombros de gigantes*” y aprovechar el conocimiento construido previamente.

2.1. Importancia y Relevancia de la Recolección de Datos

En la era del *Big Data*, la capacidad de recopilar información masiva y variada es uno de los motores de la innovación en múltiples áreas **EUDE2024**. En el caso de **ActiveMind**, resulta fundamental contar con datos sociodemográficos (edad, género, hábitos deportivos, preferencias musicales), así como con métricas fisiológicas y de salud proporcionadas por *smartwatches* (frecuencia cardíaca, nivel de oxígeno en sangre, pasos diarios, etc.). Esta combinación de datos permite comprender con mayor profundidad el estado y las necesidades del usuario, habilitando recomendaciones musicales más precisas, personalizadas y alineadas con su bienestar integral. Estos principios son fundamentales para la arquitectura de ActiveMind, asegurando la solidez de sus recomendaciones y su escalabilidad futura en distintos entornos culturales.

2.2. Importancia de la calidad y diversidad de los datos en sistemas de recomendación

De acuerdo con estudios recientes, la calidad y diversidad de la información influyen directamente en la eficacia de los sistemas de recomendación. Sin un

2.3. Descripción de los Datos y Metodologías de Procesamiento de Datos

volumen adecuado de datos, o sin su correcta representatividad e interpretabilidad, los modelos pierden capacidad predictiva y generalización, comprometiendo su utilidad práctica y la satisfacción del usuario.

Berkovsky et al. (2015) destacan que la calidad de los datos, incluyendo su limpieza y estructura, afecta directamente el rendimiento de los sistemas de recomendación. Un enfoque inadecuado de limpieza puede eliminar información valiosa o dejar ruido residual, reduciendo la precisión del modelo y su capacidad de generalizar a nuevos casos **berkovsky2015**.

Por su parte, Milvus (2023) subraya que la diversidad en los datos y en las recomendaciones generadas es esencial para evitar la sobreespecialización y el fenómeno conocido como “filtro burbuja”. Incluir una amplia gama de preferencias musicales y estados emocionales permite que las recomendaciones sean percibidas como más satisfactorias y enriquecedoras para los usuarios, mejorando su experiencia y compromiso con la aplicación **milvus2023**.

Finalmente, Shin et al. (2021) señalan que en el caso de los datos generados por dispositivos portátiles, como los wearables, es fundamental evaluar su integridad, plausibilidad y granularidad temporal para garantizar la fiabilidad de los modelos desarrollados. Sin este aseguramiento de calidad, las recomendaciones basadas en métricas fisiológicas pueden carecer de validez práctica y científica **shin2021**.

En conjunto, estas investigaciones muestran que la calidad estructural, la diversidad conceptual y la integridad técnica de los datos son dimensiones complementarias y críticas en sistemas de recomendación personalizados. Por ello, es importante implementar métodos que garanticen la recopilación de datos confiables, diversos y representativos, ya sea a través de encuestas estructuradas como en esta fase inicial del proyecto, o mediante registros continuos provenientes de dispositivos portátiles en futuras implementaciones. Estos principios son esenciales para ActiveMind, asegurando que sus sistemas de recomendación sean precisos, generalizables y relevantes para los usuarios.

2.3. Descripción de los Datos y Metodologías de Procesamiento de Datos

Los datos utilizados en este estudio provienen de las encuestas realizadas a potenciales usuarios de la aplicación **ActiveMind**. Las encuestas permitieron obtener información relevante sobre la demografía de los usuarios, sus hábitos deportivos, el tiempo dedicado a la actividad física, sus preferencias musicales y su uso de dispositivos inteligentes. En las siguientes etapas del proyecto, se busca que los *smartwatches* proporcionen métricas fisiológicas clave, como frecuencia cardíaca, nivel de oxígeno en sangre, variabilidad del ritmo cardiaco y cantidad de pasos diarios.

La combinación de estos datos resulta fundamental para la personalización de la experiencia musical. Sin embargo, antes de alimentar cualquier sistema de recomendación, la información debe someterse a una serie de procesos de *pre-*

Estado del arte

procesamiento que garanticen su calidad y coherencia. La Figura 2.1 presenta el flujo metodológico implementado en este estudio, desde la recolección de encuestas hasta la generación de clusters y clasificadores supervisados.

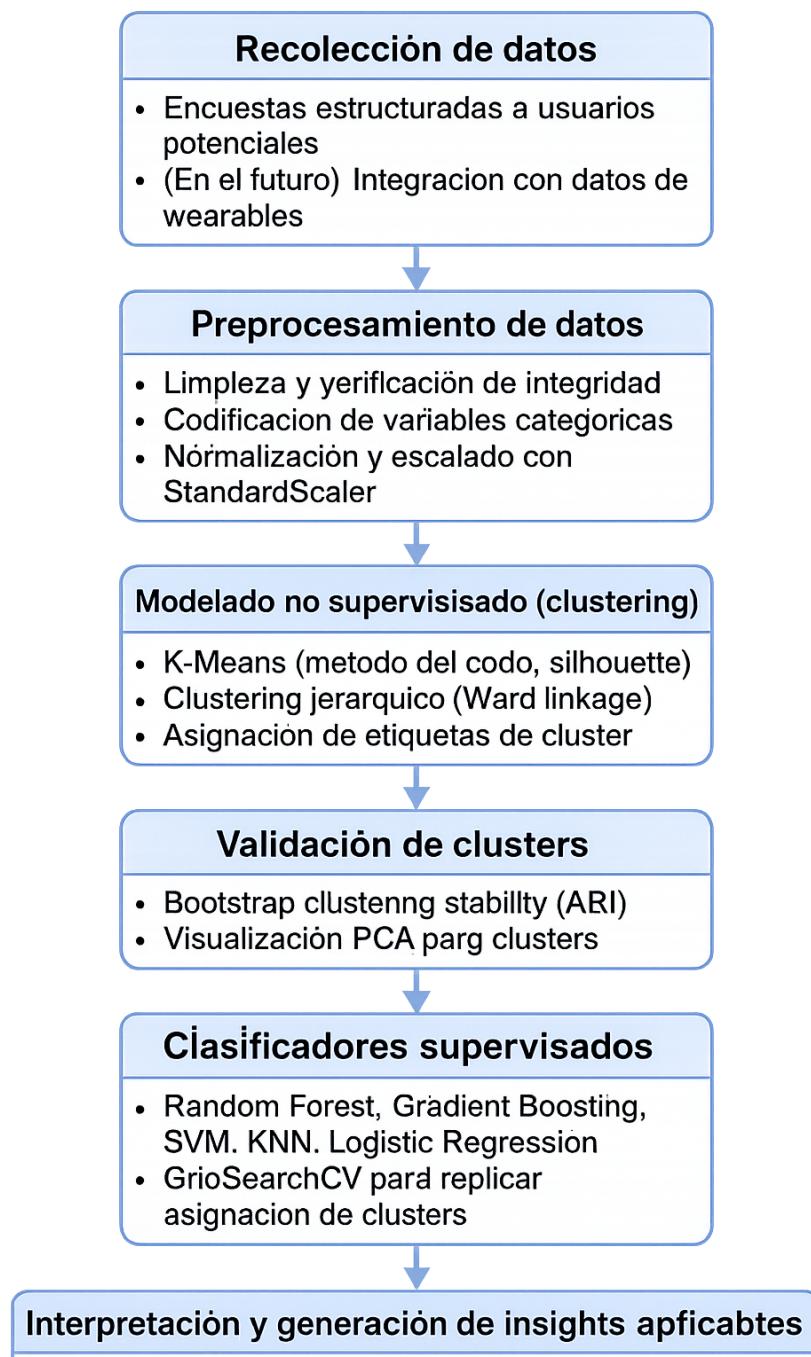


Figura 2.1: Pipeline metodológico general del preprocesamiento y análisis de datos de ActiveMind.

2.3. Descripción de los Datos y Metodologías de Procesamiento de Datos

2.3.1. Procesamiento y Tratamiento de Datos

El tratamiento de datos implicó diversas fases para asegurar su integridad y utilidad en la construcción de modelos. Entre los pasos más relevantes se incluyeron la limpieza de datos para identificar valores atípicos o inconsistentes (aunque no se encontraron debido a la estructura cerrada de las encuestas), y la imputación de datos faltantes, que no fue necesaria por la obligatoriedad de respuesta de las preguntas.

En cuanto a la transformación y normalización de variables, se ajustaron sus escalas mediante métodos como la estandarización z-score, dado que algoritmos como K-Means, PCA y linkage Ward se basan en distancias euclidianas y son sensibles a diferencias de magnitud. La elección de StandardScaler frente a MinMaxScaler se justificó porque centra los datos en media cero y los escala a varianza unitaria, preservando la estructura de outliers sin comprimir excesivamente las distribuciones.

La codificación de datos categóricos se implementó mediante *label encoding*, convirtiendo información cualitativa en variables numéricas compatibles con los algoritmos. Finalmente, se efectuó la extracción de características, generando nuevas variables como la comodidad ante la personalización musical o la importancia asignada a la salud mental, fundamentales para captar patrones de segmentación y sentar bases sólidas para modelos futuros.

Estos procesos garantizaron que las variables estuvieran en un formato matemáticamente compatible y conceptualmente interpretable por los algoritmos de segmentación y reducción de dimensionalidad utilizados en etapas posteriores del análisis.

2.3.2. Creación de Modelos con Feedback Interpretable

Uno de los principales desafíos para estructurar los datos recopilados mediante las encuestas de ActiveMind es garantizar que sean comprensibles, útiles y representativos. Para ello, se exploraron enfoques que permitieran interpretar los resultados obtenidos y estructurar la base de datos de forma analítica y operativa. Esto incluyó el análisis de patrones en hábitos deportivos, preferencias musicales y uso de dispositivos *wearable*, con el fin de segmentar el público objetivo de ActiveMind y facilitar la toma de decisiones estratégicas en el desarrollo del sistema. Así, se asegura que los modelos desarrollados no solo sean precisos desde un punto de vista estadístico, sino también comprensibles y aplicables para los equipos de diseño y desarrollo de la aplicación.

2.3.3. Análisis y segmentación de usuarios

Se realizó un análisis de segmentación mediante técnicas de clustering y explotación de datos, definiendo perfiles diferenciados en función de hábitos deportivos, preferencias musicales y uso de dispositivos *wearable*. La segmentación facilita la comprensión de diferencias psicotecnológicas que pueden influir en la adopción futura de ActiveMind.

Estado del arte

Asimismo, se aplicaron métodos de selección de variables, como análisis de varianza y medidas de correlación, junto con técnicas de reducción de dimensionalidad, particularmente el Análisis de Componentes Principales (PCA). En este estudio, PCA se empleó principalmente como técnica exploratoria para validar geométricamente la separación de clusters y facilitar la interpretación visual de los grupos detectados.

Finalmente, se desarrollaron visualizaciones y métricas descriptivas que facilitaron la interpretación de los datos, permitiendo extraer información clave sobre la relación entre los distintos atributos estudiados. Estas representaciones gráficas fortalecen la toma de decisiones en el diseño conceptual del sistema, brindando fundamentos sólidos para entender las necesidades y motivaciones del público objetivo de ActiveMind.

2.3.4. Importancia de la Interpretación de los Datos

La interpretación correcta de los datos es fundamental para garantizar la calidad y utilidad de la información recopilada. Sin una adecuada comprensión de la relación entre las variables obtenidas en las encuestas, la segmentación del público objetivo de ActiveMind podría verse afectada, dificultando la identificación de clientes potenciales y el diseño de estrategias adecuadas para la personalización del sistema. Por ejemplo, en técnicas de clustering aplicadas a datos sociales, coeficientes de silueta promedio entre 0.25 y 0.50 son considerados aceptables, reflejando la heterogeneidad natural de la población analizada.

Además, una interpretación rigurosa permite detectar posibles sesgos en los datos, validar la representatividad de la muestra y asegurar la coherencia de los resultados con el perfil real de los usuarios. En este sentido, la combinación de análisis exploratorio, técnicas de preprocesamiento y herramientas de visualización es esencial para extraer información significativa y garantizar que la base de datos proporcione valor agregado tanto para el desarrollo del sistema como para futuras aplicaciones.

Este trabajo no solo contribuye a mejorar la identificación del público objetivo, sino que también crea bases para futuros proyectos e investigaciones sobre la relación entre hábitos deportivos, música y bienestar, proporcionando un marco de referencia para la integración de datos biométricos en aplicaciones de personalización.

2.3.5. Otras Aproximaciones en la Personalización Musical

La recomendación musical no es un tema nuevo. Plataformas como *Spotify*, *Apple Music* o *Deezer* han desarrollado desde la década de 2010 algoritmos basados en comportamiento histórico del usuario y modelos de filtrado colaborativo. Sin embargo, la mayoría se centran en gustos musicales y relaciones entre usuarios, sin incorporar explícitamente la dimensión fisiológica o retroalimentación en tiempo real de dispositivos *wearable*.

En el ámbito académico, existen estudios que abordan la influencia de la música en el rendimiento deportivo y el estado anímico **leyes2006, gse2024**, pero son

2.4. Integración con Dispositivos Wearable

escasas las investigaciones que integren *machine learning* con datos biométricos para personalizar la experiencia. Reproducir, adaptar o mejorar estos trabajos constituye un aporte valioso, demostrando viabilidad y robustez en distintos contextos.

Por ejemplo, Raglio et al. (2019) emplearon árboles de decisión para predecir el efecto de la música en la relajación, logrando una precisión del 79 %, e identificando como factores influyentes el nivel inicial de relajación, la educación y formación musical, la edad y la frecuencia de escucha **raglio2019**. Esto demuestra que las técnicas de aprendizaje automático son herramientas innovadoras y valiosas para la musicoterapia.

Asimismo, la tesis doctoral de Martínez (2021) exploró cómo la personalización del perfil del turista mediante análisis biométrico puede utilizarse en marketing de destinos y servicios, abordando implicaciones éticas, privacidad y aceptación de estas tecnologías **martinez2021**.

Estos ejemplos demuestran que la integración de *machine learning* con datos biométricos puede crear experiencias personalizadas en diversos contextos, reafirmando la viabilidad de estas soluciones en salud digital y bienestar.

2.4. Integración con Dispositivos Wearable

La explosión de la industria de *smartwatches* y pulseras inteligentes ha abierto un mundo de posibilidades para capturar datos en tiempo real durante la actividad física. Empresas como *Fitbit*, *Garmin* o *Xiaomi* han impulsado la creación de APIs y herramientas que facilitan el acceso a la información recogida por sus sensores. Esto, junto con la reducción de costes y mejora en la precisión de los dispositivos, crea el contexto ideal para utilizar métricas de salud como un factor decisivo en sistemas de recomendación especializados.

2.5. Conclusiones del Estado del Arte

La revisión efectuada evidencia la necesidad de combinar diversas disciplinas para ofrecer soluciones sólidas en la recomendación musical personalizada. Aunque existen antecedentes en análisis del rendimiento deportivo asociado a la música y en construcción de modelos de recomendación, **ActiveMind** se presenta como un enfoque novedoso que integra explícitamente datos fisiológicos y preferencias personales para mejorar la experiencia durante la actividad física.

En definitiva, ActiveMind se posiciona como un puente entre el conocimiento académico y su aplicación práctica, contribuyendo a la construcción de un futuro donde la música y la inteligencia artificial se integren para potenciar el bienestar personal.

Capítulo 3

Propuesta

3.1. Contexto y Motivación de la Propuesta

El panorama actual se caracteriza por un crecimiento acelerado en el desarrollo de la inteligencia artificial (IA), cuyos avances están transformando múltiples ámbitos, desde la salud hasta el entretenimiento. Esta revolución tecnológica representa una oportunidad para desarrollar soluciones innovadoras que impacten de manera positiva en la calidad de vida de las personas, al tiempo que consoliden un potencial de escalabilidad y comercialización.

En esta coyuntura, surge el proyecto **ActiveMind**, una aplicación enfocada en la integración de IA con datos personales de salud y bienestar, cuyo propósito es ofrecer recomendaciones musicales personalizadas. El objetivo principal consiste en optimizar la experiencia deportiva de los usuarios, al combinar sus preferencias musicales con información sobre su estado físico y la etiqueta de las canciones. Este enfoque, que vincula métricas de salud (como el estado de ánimo, los hábitos de sueño o la rutina de ejercicio) con contenido personalizado (en este caso, música), refleja la tendencia de diseñar tecnologías más humanas, cercanas y emocionalmente pertinentes.

Para lograr que **ActiveMind** genere un verdadero impacto y cuente con perspectivas de crecimiento en el mercado, resulta esencial comprender al público objetivo en detalle. La recopilación de datos desde fases iniciales posibilita la validación de hipótesis sobre la conducta de los usuarios, la identificación de sus necesidades específicas y la adaptación de las funcionalidades del producto en función de esos hallazgos. Sin una base de datos sólida, no solo se vería comprometida la precisión de los modelos de IA, sino también la viabilidad del proyecto como producto digital.

Este proyecto se plantea, por tanto, como una oportunidad para aplicar técnicas avanzadas de ciencia de datos en la construcción de una solución centrada en el usuario, con un alto potencial de convertirse en un modelo de negocio sostenible.

3.2. Diseño del Estudio

3.2.1. Creación del Instrumento de Recolección de Datos

Esta sección detalla la metodología aplicada para la creación de la base de datos, a partir del diseño y aplicación de la encuesta. Se describen los instrumentos, el procedimiento de recolección de datos, el procesamiento y la estructuración final de la información.

3.2.2. Justificación del Diseño y Selección de Preguntas

El cuestionario fue elaborado siguiendo criterios teóricos y prácticos orientados a obtener una visión integral del perfil de los usuarios. Las preguntas se estructuraron de la siguiente manera:

El cuestionario desarrollado incluyó un bloque de datos demográficos, con preguntas sobre edad y género, variables fundamentales para segmentar la muestra y analizar diferencias en comportamientos y preferencias entre distintos grupos etarios y de identidad. Esta información permite caracterizar al público objetivo y ajustar estrategias de segmentación y comunicación.

En el apartado de hábitos de actividad física se consultaron aspectos relacionados con la frecuencia semanal de ejercicio y el tipo de actividad realizada, como fuerza, cardio o deportes grupales. Estos datos son esenciales para identificar los niveles de actividad física en la población estudiada y conocer las modalidades más practicadas, lo cual permite establecer correlaciones relevantes entre los hábitos deportivos y las necesidades en materia de recomendaciones musicales.

También se incluyeron preguntas orientadas a conocer el uso de tecnología, especialmente la adopción de dispositivos *wearable* durante la realización de rutinas de ejercicio. Este apartado resulta clave, ya que dichos dispositivos serán los futuros proveedores de métricas fisiológicas necesarias para la implementación de ActiveMind, determinando la viabilidad tecnológica del sistema propuesto.

Por otro lado, se exploraron las preferencias y motivaciones de los usuarios, indagando en sus géneros musicales favoritos y sus principales razones para ejercitarse, como relajación, aumento de energía o mejora del estado de ánimo. Estos datos aportan información valiosa para el diseño de sistemas de recomendación personalizados que se alineen con los objetivos individuales de cada usuario.

Finalmente, se incorporó un bloque de opiniones sobre funcionalidades, con preguntas orientadas a conocer el interés en aplicaciones que integren ejercicio, salud mental y personalización musical. Este apartado facilita identificar funcionalidades clave y validar la propuesta de valor, sentando las bases para diseñar un producto que genere aceptación y relevancia en el mercado.

Cada ítem fue seleccionado en función de su relevancia para la construcción del perfil del usuario y su potencial para influir en el desarrollo de soluciones personalizadas, fundamentadas en técnicas de inteligencia artificial. Así como la identificación de un posible nicho de clientes potenciales.

Propuesta

3.2.3. Uso de Google Forms como Plataforma de Encuesta

Para la recolección de datos se empleó Google Forms, seleccionándolo por su interfaz amigable que permite la creación de cuestionarios con un diseño claro y sencillo. Esta característica favorece la tasa de respuesta al ofrecer a los participantes una experiencia intuitiva y accesible desde cualquier dispositivo.

Además, la plataforma cuenta con funcionalidad bilingüe, lo que permitió diseñar inicialmente el formulario en español e inglés con el objetivo de captar una muestra más diversa. Sin embargo, tras el análisis de las primeras respuestas se optó por continuar únicamente en español, dado que la gran mayoría de las personas participantes eligió este idioma, asegurando así consistencia en el tratamiento posterior de los datos.

Por último, Google Forms ofrece herramientas integradas para la recolección y almacenamiento automático de las respuestas, incluyendo la visualización preliminar de tendencias y la detección de posibles inconsistencias. Esta funcionalidad resultó clave para monitorear la calidad de los datos en tiempo real y agilizar su exportación para los procesos de limpieza y análisis.

3.2.4. Primeras Impresiones de las Respuestas

La aplicación del cuestionario permitió recolectar un total de 157 respuestas en español, mientras que la versión en inglés obtuvo únicamente 9 respuestas. Esta distribución justificó la decisión de enfocar el estudio en la muestra hispanohablante, asegurando consistencia lingüística en el análisis y una mayor robustez estadística.

Las respuestas iniciales reflejaron una distribución variada en cuanto a grupos etarios, lo que garantiza la inclusión de participantes provenientes de diferentes contextos y estilos de vida. Este aspecto es fundamental para entender la diversidad de necesidades y expectativas dentro del público objetivo de ActiveMind.

Asimismo, se observó una amplia diversidad en los niveles de actividad física reportados y en las modalidades de ejercicio practicadas, incluyendo fuerza, cardio, deportes grupales y actividades recreativas. Esta variedad proporciona un panorama integral para evaluar las preferencias y necesidades específicas de cada segmento identificado.

Por último, los datos revelaron un alto interés en la integración de tecnologías orientadas a la mejora del bienestar. Esto se evidenció especialmente en las respuestas relacionadas con la aceptación de aplicaciones que personalicen recomendaciones musicales basadas en el estado físico y emocional, validando así la propuesta de valor de ActiveMind y su potencial de adopción en el mercado objetivo.

Estas primeras impresiones respaldan la pertinencia de la encuesta como herramienta para obtener datos que fundamenten el diseño de sistemas de recomendación basados en inteligencia artificial y sirven como punto de partida para posteriores fases de análisis y manipulación de datos.

3.3. Procesamiento y Análisis de los Datos Estructurados

3.3. Procesamiento y Análisis de los Datos Estructurados

La etapa de procesamiento y análisis de la base de datos constituye un paso esencial para garantizar la calidad y relevancia de la información recolectada. Los datos se obtuvieron mediante la exportación directa del cuestionario de Google Forms a un archivo Excel, lo que permitió una integración inicial de las respuestas en un formato estructurado para su análisis.

3.3.1. Exportación y Verificación Inicial de Datos

El primer paso en el proceso de análisis consistió en exportar los datos recolectados a través de Google Forms a un archivo Excel. Esta exportación permitió consolidar toda la información en un único documento, facilitando su manejo y posibilitando la verificación inicial de la integridad y consistencia de los registros.

Durante esta fase, se revisó detalladamente el conjunto de datos para identificar posibles campos incompletos o nulos. Sin embargo, dado que Google Forms ofrece la opción de configurar las preguntas como obligatorias, no se registraron respuestas faltantes en ninguna de las variables clave, asegurando así la completitud del dataset para los análisis posteriores.

Asimismo, se realizó una comprobación exhaustiva de posibles errores de formato, verificando la correcta representación de fechas, textos y valores numéricos. Este control garantizó que cada campo correspondiera al tipo de dato esperado, evitando inconsistencias que pudieran afectar las etapas de limpieza, transformación y análisis de los datos.

3.3.2. Limpieza y Preprocesamiento de la Información

Una vez verificada la integridad básica del archivo exportado, se procedió a realizar la limpieza y preprocesamiento de los datos, con el objetivo de garantizar su calidad y adecuación para los análisis posteriores.

En primer lugar, se efectuó la eliminación de registros inconsistentes. Se revisaron las respuestas en busca de inconsistencias significativas, especialmente en las preguntas abiertas, y se descartaron aquellos registros que pudieran comprometer la confiabilidad de los resultados. Sin embargo, dado que la encuesta se estructuró mayoritariamente con preguntas de opción cerrada y obligatoria, se identificaron pocos casos de datos inconsistentes, lo que permitió conservar la integridad y representatividad de la muestra.

Posteriormente, se realizó la normalización de respuestas en las variables categóricas, con el fin de estandarizar su formato y evitar variaciones en la escritura que pudieran afectar los análisis de agrupamiento y segmentación. Por ejemplo, se unificaron las denominaciones de categorías en variables como edad, tipo de ejercicio y preferencias musicales, corrigiendo inconsistencias ortográficas y semánticas. Este proceso aseguró la comparabilidad de las respuestas y minimizó

Propuesta

la dispersión innecesaria en los datos, fortaleciendo la calidad del dataset como base de los análisis exploratorios y de clusterización realizados en el estudio.

Codificación de variables cualitativas: Se implementó un proceso sistemático de codificación para transformar variables cualitativas en formatos numéricos o estructurados, facilitando así su posterior análisis cuantitativo y la aplicación de modelos estadísticos o de aprendizaje automático. Esta codificación se aplicó tanto a preguntas cerradas como a preguntas abiertas con opción otro, en las que los participantes podían especificar respuestas personalizadas. A continuación se detallan los pasos y decisiones tomadas:

En el proceso de preprocessamiento se realizó la codificación de las preguntas cerradas, transformando las variables categóricas de tipo ordinal o nominal a escalas numéricas, respetando su jerarquía o significado conceptual para su adecuada integración en los análisis posteriores.

La variable *Género* fue codificada asignando el valor 0 para Masculino y 1 para Femenino, manteniendo una estructura binaria clara. Para la variable *Nivel de actividad física semanal*, se estableció una escala ordinal donde 0 correspondió a “No hago ejercicio regularmente”, 1 a “1-2 veces/semana”, 2 a “3-4 veces/-semana” y 3 a “5 o más veces/semana”, reflejando la progresión natural de la frecuencia de ejercicio.

En el caso de la importancia de la salud mental durante la rutina de ejercicio, las respuestas se transformaron a una escala conceptual: 0 para “Poco importante”, 2 para “Algo importante” y 4 para “Muy importante”. Este método permitió capturar el énfasis diferencial que los participantes otorgan a este aspecto en su actividad física.

La pregunta sobre el uso de dispositivos tecnológicos fue recodificada de manera binaria, con 0 para “No” y 1 para “Sí”, facilitando su análisis como variable dummy. Por su parte, la variable sobre comodidad con una aplicación inteligente personalizada se codificó en una escala creciente de 0 a 4, donde 0 representó “Nada cómodo” y 4 “Muy cómodo”.

Finalmente, el interés en una aplicación que integre ejercicio, salud mental y música con inteligencia artificial se transformó en una escala de 0 a 4, donde 0 indicó nulo interés y 4 máximo interés. Estas codificaciones aseguran la correcta interpretación de las variables por parte de los algoritmos de análisis y mantienen la consistencia conceptual necesaria para su análisis estadístico y de segmentación.

Normalización y codificación de respuestas abiertas

Para las preguntas que incluían la opción ‘Otro’, se aplicó un proceso de normalización semántica y posterior codificación con el objetivo de integrar estas respuestas de manera coherente al resto de las categorías cerradas. Este procedimiento aseguró la homogeneidad y calidad del dataset para los análisis posteriores.

En primer lugar, se realizó una agrupación semántica de las respuestas textuales. Estas fueron revisadas manualmente y reorganizadas en categorías con-

3.3. Procesamiento y Análisis de los Datos Estructurados

ceptuales equivalentes, unificando sinónimos, corrigiendo errores ortográficos y consolidando variaciones gramaticales. Por ejemplo, en la pregunta “¿Qué tipo de actividad física realizas con más frecuencia? (Otro)”, respuestas como “Caminar”, “subir escaleras” o “actividades ligeras” se agruparon bajo la categoría “Caminata/Actividad Ligera”. De igual forma, respuestas como “Gimnasio” y “Gimnasia de mantenimiento” se integraron en la categoría “Gimnasio/Fuerza”. Aquellas categorías irrelevantes para los objetivos del estudio, como menciones de “Sexo” en campos no pertinentes, fueron descartadas.

En el caso de las preferencias musicales, específicamente en la pregunta “¿Qué tipo de música prefieres para hacer ejercicio? (Otro)”, se implementó un proceso de tipificación musical. Respuestas como “Reguetón”, “Jazz” o “Corridos” fueron normalizadas como categorías individuales, mientras que estilos menos frecuentes como “Techno”, “Indie” o “Cantautor” se agruparon en una categoría general denominada “Otros estilos modernos”. Las respuestas que indicaban la ausencia de música durante el ejercicio se codificaron como “No escucha música”.

Finalmente, una vez finalizada la normalización y agrupación de estas respuestas, las nuevas categorías estandarizadas fueron reemplazadas directamente en el dataset, sustituyendo las respuestas textuales originales. Este paso permitió mantener la integridad estructural del archivo y preparar de manera óptima las variables para su posterior codificación numérica y análisis estadístico.

La transformación y codificación de las variables se justifican por la necesidad de estandarizar las respuestas para que sean compatibles con técnicas estadísticas y modelos de inteligencia artificial. Al convertir las variables cualitativas a cuantitativas se reduce la ambigüedad inherente a las respuestas textuales y se facilita el análisis comparativo, la generación de tablas de frecuencia y la aplicación de métodos de regresión y correlación. Este preprocesamiento garantiza que la base de datos esté en un formato adecuado y homogéneo, minimizando errores y maximizando la capacidad de extraer insights significativos para el desarrollo del sistema de recomendación propuesto.

3.3.3. Análisis Descriptivo de la Información

Con el conjunto de datos depurado se realizó un análisis descriptivo exhaustivo con el objetivo de caracterizar el perfil de los participantes y sus respuestas, obteniendo así una visión integral y detallada de la muestra analizada.

En cuanto a la distribución demográfica, la encuesta recolectó un total de 157 respuestas, lo que permitió identificar una representación diversa en términos de edad y género. Respecto a la variable edad, los resultados muestran que la mayoría de los participantes se concentra en el grupo de 35 a 44 años, representando el 35 % de la muestra. Este segmento es seguido por la categoría de menores de 18 años, que constituye el 26.8 %, y por el grupo de 18 a 24 años, que alcanza un 21.7 %. Las categorías restantes, correspondientes a los rangos de 25-34, 45-54 y 55 o más años, presentan menor representación en la muestra. Estos hallazgos indican que la población estudiada está compuesta predominantemente por personas jóvenes y de mediana edad, grupos de especial

Propuesta

interés para el desarrollo y adopción de tecnologías como ActiveMind.

En relación con la identificación de género, los resultados muestran que el 56.7% de los participantes se autodeclara masculino, mientras que el 43.3% se identifica como femenino. Esta distribución refleja una muestra relativamente equilibrada, permitiendo analizar de manera comparativa las percepciones, preferencias y necesidades de ambos grupos en el contexto de la personalización musical y la integración de aplicaciones basadas en inteligencia artificial orientadas al bienestar físico y mental.

Niveles de actividad física

Las respuestas relativas a la frecuencia y tipo de ejercicio practicado por los participantes evidencian patrones de comportamiento relevantes para el análisis del público objetivo de ActiveMind. En cuanto a la frecuencia de actividad física, únicamente el 10.2% de los encuestados indicó no realizar ejercicio de forma regular. Por su parte, un 19.1% reportó practicar actividad física entre 1 y 2 veces por semana. El 36.9% manifestó ejercitarse entre 3 y 4 veces semanalmente, mientras que un 33.8% declaró realizar actividad física 5 o más veces a la semana. Estos resultados reflejan una alta predisposición hacia el ejercicio en la muestra, aspecto fundamental para la adopción de tecnologías que buscan potenciar el rendimiento y el bienestar durante las rutinas deportivas.

Respecto al tipo de ejercicio practicado, la encuesta incluyó una pregunta de selección múltiple que permitió conocer las modalidades más comunes. Los resultados muestran que las actividades de cardio, como correr, andar en bicicleta o nadar, junto con el entrenamiento de fuerza, fueron seleccionadas por el 40.8% de los participantes, evidenciando una marcada preferencia por estas dos modalidades. Las caminatas y actividades recreativas también destacaron, siendo elegidas por el 36.9% de la muestra, mientras que yoga y pilates fueron preferidos por el 22.3%.

Por otro lado, la práctica de deportes en equipo fue reportada por el 8.9% de los encuestados. Finalmente, un 12.7% seleccionó la opción ‘Otro’, indicando la existencia de modalidades adicionales no contempladas en las opciones cerradas, lo que sugiere la necesidad de considerar actividades complementarias en futuros desarrollos y recomendaciones de la aplicación ActiveMind para atender de manera integral la diversidad de preferencias en la actividad física.

Preferencias y motivaciones

El análisis de las motivaciones y preferencias de los participantes en relación con la actividad física y la música reveló hallazgos relevantes para el desarrollo de ActiveMind. En primer lugar, respecto a las motivaciones para realizar ejercicio, la mayoría de los encuestados (57.3%) señaló como principal objetivo aumentar sus niveles de energía, reflejando un enfoque orientado a la mejora de su vitalidad y rendimiento diario. Esta motivación fue seguida por la reducción del estrés y la mejora de la salud mental, elegida por el 31.2% de la muestra, mientras que un 8.9% indicó como prioridad la mejora de su estado físico general.

En cuanto a la duración de las sesiones de ejercicio, los resultados mostraron

3.3. Procesamiento y Análisis de los Datos Estructurados

una tendencia hacia rutinas de duración moderada a prolongada. El 8.9 % de los participantes reportó sesiones de menos de 30 minutos, el 35 % manifestó ejercitarse entre 30 y 60 minutos, y la mayoría, correspondiente al 54.1 %, indicó dedicar entre 1 y 2 horas a sus sesiones de entrenamiento.

Por último, se analizaron las barreras que limitan la práctica de actividad física. La falta de tiempo destacó como la principal limitante, reportada por el 75.8 % de los encuestados. Le siguieron la falta de motivación, mencionada por el 30.6 %, y el desconocimiento sobre qué tipo de ejercicio realizar, reportado por el 10.2 %. Asimismo, un 10.2 % señaló la falta de acceso a instalaciones o equipos adecuados como un impedimento, mientras que el 14.6 % expresó dificultad para encontrar rutinas que se ajusten a sus necesidades específicas. Estos hallazgos permiten comprender los desafíos y necesidades que la aplicación deberá abordar para incrementar su relevancia y potencial de adopción en la vida cotidiana de los usuarios.

Preferencias musicales para ejercitarse

La encuesta incluyó un apartado específico para indagar sobre los géneros musicales preferidos por los participantes durante la actividad física, información clave para el diseño de sistemas de recomendación personalizados. Los resultados mostraron que el género pop es el más popular, siendo elegido por el 44.6 % de los encuestados. En segundo lugar, se ubicó la música electrónica y dance, seleccionada por el 38.2 % de los participantes. Otros géneros destacados fueron el rock, con un 25.5 %, y tanto hip-hop/rap como música clásica o ambiental, ambos con un 22.3 %. Además, un 19.7 % de la muestra eligió la opción 'Otro', lo que evidencia una amplia diversidad de gustos musicales que la aplicación ActiveMind deberá considerar para ofrecer recomendaciones realmente personalizadas y satisfactorias para los usuarios.

3.3.4. Herramientas y técnicas de análisis

El procesamiento y análisis de los datos se apoyó en el uso de herramientas de análisis estadístico y visualización. El archivo Excel exportado desde Google Forms sirvió como punto de partida para la organización inicial de los datos, permitiendo una verificación preliminar y la estructuración de variables. Este archivo se complementó con programación, principalmente en Python, para la manipulación avanzada de los datos y la generación de representaciones gráficas.

Estas herramientas permitieron realizar cálculos de medidas de tendencia central y dispersión para las variables numéricas, aportando una caracterización cuantitativa detallada de la muestra. Asimismo, se construyeron tablas de contingencia y gráficos de barras, los cuales ilustraron de manera clara la distribución de las respuestas en cada dimensión evaluada. Finalmente, se identificaron correlaciones preliminares entre variables relevantes, proporcionando insights iniciales que servirán como base para el desarrollo futuro de modelos de recomendación musical basados en inteligencia artificial en el marco de ActiveMind.

En resumen, la metodología aplicada en esta fase garantizó que la base de da-

Propuesta

tos resultante fuera precisa y coherente, sentando las bases para posteriores procesos de manipulación y análisis avanzado que se detallarán en secciones subsiguientes.

Capítulo 4

Metodología e Implementación

4.1. Preprocesamiento y transformación de variables

Antes de realizar cualquier análisis exploratorio o modelado, fue necesario implementar un preprocesamiento exhaustivo de los datos recopilados mediante las encuestas. El preprocesamiento constituye una fase crítica en proyectos de ciencia de datos, ya que la calidad, consistencia y correcta representación de las variables determinan la fiabilidad y validez de los resultados posteriores. En el caso de este estudio, el objetivo del preprocesamiento fue estandarizar los datos, transformarlos en un formato interpretable por los algoritmos de clustering y clasificación, y garantizar que cada variable capturara adecuadamente la información conceptual para la que fue diseñada.

En primer lugar, se reformateó la variable **Timestamp**, convirtiéndola al formato estándar `datetime` y formateándola posteriormente a la estructura 'YYYY-MM-DD'. Este paso se implementó para estandarizar el campo temporal y facilitar su uso en futuras integraciones con bases de datos relacionales o sistemas de registro de usuario, a pesar de que no fue utilizada como predictor en el análisis de clustering.

A continuación, se procedió a la **codificación de variables categóricas** para transformarlas en variables numéricas utilizables por los modelos. La variable **Género** fue codificada como 0 para Masculino y 1 para Femenino, siguiendo un esquema binario ampliamente utilizado en estudios sociales y de comportamiento. Para la variable **nivel de actividad física semanal**, se aplicó un mapeo ordinal que refleja la progresión natural de la frecuencia de ejercicio: 0 para "No hago ejercicio regularmente", hasta 3 para "5 o más veces por semana". Esta codificación ordinal preserva la jerarquía inherente a la variable, permitiendo su interpretación directa en modelos estadísticos y en la generación de clusters.

La variable sobre **importancia de la salud mental en la rutina de ejercicio** se transformó en una escala conceptual discreta: 0 para "Poco importante", 2 para "Algo importante" y 4 para "Muy importante". Esta codificación permitió capturar la intensidad de la valoración sin imponer un intervalo uniforme, ya que las categorías representan cambios conceptuales más que cambios lineales.

4.2. Normalización sistemática de preguntas abiertas

El uso de **dispositivos tecnológicos para mejorar el bienestar** se codificó de forma binaria (0 para No, 1 para Sí), en coherencia con su naturaleza dicotómica. Por otro lado, la variable que evalúa la **comodidad con apps inteligentes que personalizan música y entrenamiento** se transformó en una escala de 0 a 4, diferenciando niveles de aceptación desde "Nada cómodo, no me interesa" hasta "Muy cómodo, me gustaría probarlo". Esta escala ordinal permite segmentar usuarios según su disposición tecnológica, un predictor clave para la adopción de ActiveMind.

Finalmente, la variable sobre **interés en una app que integre ejercicio, salud mental y personalización musical con IA** fue codificada también en una escala discreta de 0 a 4, diferenciando entre "No, no me interesa", "Tal vez, dependiendo de sus funciones" "Sí, me interesa mucho". Esta codificación facilita la identificación de perfiles de usuario con alta predisposición a la adopción, información esencial para la futura estrategia de diseño y comunicación de la aplicación.

En síntesis, este preprocesamiento no solo estandarizó y transformó las variables para su análisis técnico, sino que también estructuró los datos para maximizar su interpretabilidad y aplicabilidad práctica en el desarrollo de segmentaciones, modelos de recomendación y futuras integraciones de ActiveMind con sistemas de monitoreo en tiempo real.

4.2. Normalización sistemática de preguntas abiertas

Una etapa fundamental del preprocesamiento consistió en la normalización semántica de las respuestas abiertas, particularmente aquellas relacionadas con el tipo de música preferida durante la actividad física, las cuales se recopilaron bajo la opción "Otro" en la encuesta. Este procedimiento tuvo como objetivo garantizar la consistencia y comparabilidad de los datos, ya que las respuestas abiertas suelen presentar sinónimos, errores ortográficos, uso de mayúsculas, acentos omitidos y variaciones gramaticales que, de no ser unificadas, generan dispersión artificial y reducen la calidad del análisis.

Para implementar esta limpieza, se definió un diccionario de mapeo que vinculó distintas entradas textuales a categorías conceptuales unificadas. Por ejemplo, se homogeneizaron respuestas como "reguetón" y "regueton" bajo la categoría "Reguetón"; estilos menos frecuentes como "techno", "indie" o "cantautor" fueron agrupados como "Otros estilos modernos"; y se creó una categoría especial "No escucha música" para respuestas como "no" o "sin música". Asimismo, actividades similares como "subir escaleras" y "Caminar" se integraron bajo "Caminata/Actividad Ligera", mientras que menciones de modalidades de fuerza y mantenimiento físico se unificaron como "Gimnasio/Fuerza".

El proceso se realizó mediante la definición de una función de limpieza que transformaba cada entrada en minúsculas, eliminaba espacios residuales y la comparaba con las claves del diccionario de mapeo. En caso de no existir coincidencias, la respuesta se clasificaba en la categoría general "Otro". Este método sistemático permitió reducir la dispersión semántica, mejorar la integridad de

Metodología e Implementación

la variable y asegurar su correcta interpretación en el análisis de clustering y caracterización de perfiles de usuario.

Finalmente, se generó un resumen de las categorías normalizadas para confirmar la distribución de frecuencias y la consistencia del proceso. Esta normalización no sólo optimizó la calidad técnica del dataset, sino que también fortaleció la interpretación práctica de las preferencias musicales de los potenciales usuarios de *ActiveMind*, insumo clave para el desarrollo de futuros motores de recomendación personalizados.

4.3. Visualización y Exploración de Datos

Para iniciar la exploración del conjunto de datos, se realizó una visualización descriptiva y exploratoria que permitiera identificar las distribuciones y relaciones existentes entre las variables numéricas recopiladas mediante las encuestas de *ActiveMind*. Este paso fue esencial para garantizar la calidad del análisis posterior y la coherencia de los modelos desarrollados.

Primero, se generaron las estadísticas descriptivas mediante las funciones `df.info()` y `df.describe()` en Python, dentro del entorno Google Colab. Esto permitió conocer el número de registros, tipos de variables y rangos de valores, así como confirmar la ausencia de valores faltantes debido a la estructura cerrada de las encuestas aplicadas.

Cuadro 4.1: Estadísticas descriptivas de variables numéricas

Variable	Media	Desv. típica	Mínimo	25 %	Mediana	Máximo
Género	0.57	0.50	0	0	1	1
Nivel de actividad física	1.65	0.90	0	1	2	3
Importancia de salud mental	3.43	1.04	0	2	4	4
Uso de tecnología	0.46	0.50	0	0	0	1
Comodidad con app personalizada	2.91	1.30	0	2	4	4
Interés en app con IA	2.94	1.21	0	2	4	4

Los resultados muestran que la media de la variable **Género** fue 0.57, indicando ligera mayoría femenina en la muestra. El **nivel de actividad física** promedio fue 1.65 (entre 1-2 y 3-4 veces por semana) con desviación típica de 0.90. La **importancia asignada a la salud mental** durante el ejercicio fue alta (3.43/4), mientras que el **uso de tecnología** mostró un valor medio de 0.46, reflejando que menos de la mitad utilizan dispositivos tecnológicos para entrenar. En cuanto a **comodidad con una app personalizada** (2.91/4) e **interés en una app que integre IA** (2.94/4), ambos valores sugieren alta disposición hacia este tipo de soluciones inteligentes para el bienestar.

Posteriormente, se generó un **mapa de calor de correlaciones**, utilizando la función `heatmap()` de la librería `seaborn`, considerando únicamente las columnas numéricas. Esto permitió identificar relaciones entre variables que pudieran condicionar el agrupamiento posterior.

4.4. Implementación y justificación del análisis de clustering y clasificación

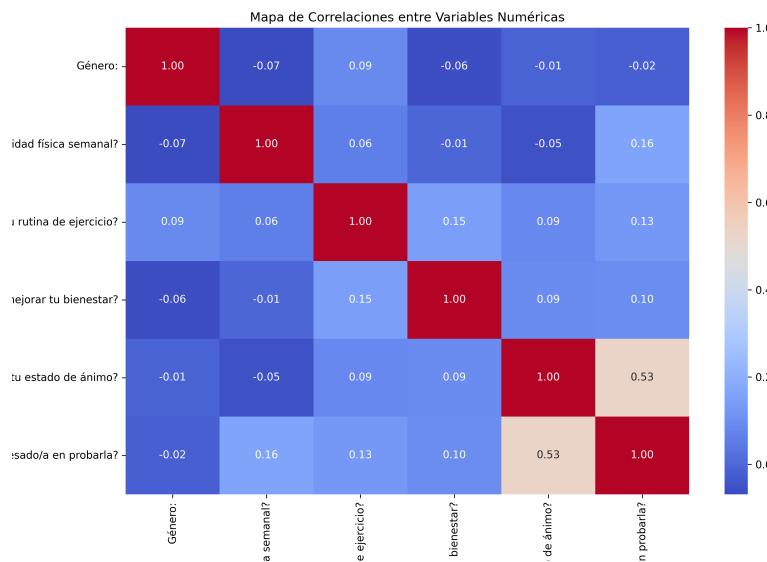


Figura 4.1: Mapa de calor de correlaciones entre variables numéricas.

En general, las correlaciones fueron bajas, con excepción de la relación moderada (0.53) entre la comodidad con apps personalizadas y el interés en apps con IA, indicando alta disposición a soluciones tecnológicas integradas.

4.4. Implementación y justificación del análisis de clustering y clasificación

El objetivo fundamental de este análisis consistió en segmentar a los participantes en grupos internamente homogéneos y externamente diferenciables, de modo que las decisiones de personalización dentro de la aplicación *ActiveMind* reposen en evidencia estadística sólida. Esta segmentación permite identificar patrones latentes de comportamiento, afinidad tecnológica y valoración del bienestar, insumos clave para diseñar recomendaciones musicales y funcionalidades adaptadas a cada perfil de usuario.

Para alcanzar esta meta, el flujo de trabajo se organizó en etapas coherentes y metodológicamente fundamentadas. En primer lugar, se prepararon los datos mediante procesos de limpieza, normalización y escalado estandarizado utilizando *StandardScaler*. Este paso garantizó la comparabilidad entre variables de diferentes unidades o escalas, condición indispensable para algoritmos como K-Means o PCA, que se basan en distancias euclidianas y son sensibles a diferencias de magnitud. La elección de *StandardScaler*, en lugar de *MinMaxScaler*, se justificó porque centra los datos en media cero y los escala a varianza unitaria, preservando la estructura de outliers sin comprimir excesivamente la distribución de las variables.

Posteriormente, se aplicaron métodos de clustering no supervisados, incluyen-

Metodología e Implementación

do K-Means y clustering jerárquico aglomerativo con vínculo *Ward*, con el objetivo de descubrir agrupaciones naturales sin imposición previa de etiquetas. K-Means permitió identificar centroides óptimos minimizando la suma de distancias cuadradas intra-cluster, mientras que el método Ward del clustering jerárquico ofreció un enfoque complementario al agrupar observaciones minimizando la varianza intra-cluster en cada fusión. Esta triangulación metodológica permitió validar la estabilidad y coherencia de la segmentación desde perspectivas paramétricas distintas.

Es importante resaltar que el análisis de clustering no se incluyó directamente como visualización exploratoria inicial, sino como un paso de modelado exploratorio, ya que su objetivo metodológico es identificar estructuras latentes y patrones subyacentes en los datos, más allá de describir sus distribuciones o correlaciones. Por esta razón, en la implementación práctica, la ejecución del clustering se situó después de la inspección descriptiva y la generación del mapa de calor de correlaciones, alineándose con buenas prácticas de ciencia de datos que priorizan la comprensión profunda de las variables antes de aplicar modelos de agrupamiento.

Finalmente, se desarrollaron clasificadores supervisados, incluyendo Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) y Regresión Logística, optimizados mediante GridSearchCV. Aunque estos modelos no generan un valor predictivo real más allá de la segmentación original, ya que replican etiquetas generadas mediante clustering no supervisado, cumplen una función práctica crítica: trasladar la lógica de segmentación descubierta a un entorno de producción ágil y escalable. Esta estrategia permite que cada nuevo usuario de ActiveMind sea asignado de manera rápida y confiable a un cluster predefinido sin necesidad de recalcular el modelo de clustering completo, habilitando así la personalización dinámica de recomendaciones musicales y funciones de la aplicación de forma eficiente en tiempo real.

4.4.1. Visualización de Silueta por Cluster

Para evaluar la calidad de la segmentación generada mediante K-Means con $K = 3$, se implementó un *silhouette plot* que graficó el coeficiente de silueta de cada observación dentro de su respectivo cluster. Este coeficiente mide la similitud de un dato con su propio cluster en comparación con el cluster más cercano, oscilando entre -1 y $+1$, donde valores cercanos a $+1$ indican asignaciones bien definidas, valores cercanos a cero sugieren observaciones en los límites de decisión entre clusters y valores negativos indican asignaciones potencialmente erróneas.

El gráfico resultante mostró un valor medio de silueta de aproximadamente 0.30 para los tres grupos, lo que sugiere una cohesión interna moderada. En la literatura de clustering social y de comportamiento, valores entre 0.25 y 0.50 se consideran aceptables, reflejando que los individuos se encuentran relativamente cercanos a los centroides de sus clusters, con niveles de dispersión consistentes con datasets heterogéneos. Este resultado indica que la estructura de clusters capturada no es aleatoria, sino que refleja diferencias reales en el conjunto de

4.4. Implementación y justificación del análisis de clustering y clasificación

datos, aunque no sea perfectamente compacta.

Además, los valores positivos observados reflejan una separación clara entre clusters, ya que no se detectaron valores negativos y sólo un bajo número de observaciones presentó coeficientes cercanos a cero, evidenciando que las fronteras entre clusters están definidas y existen mínimos solapamientos. Sin embargo, la presencia de casos con coeficientes cercanos a cero en los límites de decisión podría indicar la existencia de subgrupos latentes dentro de los clusters actuales no capturados por la segmentación en tres grupos. También puede señalar la necesidad de incorporar nuevas variables explicativas, como métricas fisiológicas obtenidas en tiempo real mediante dispositivos portátiles, que permitan refinar la segmentación y explicar estas ambigüedades.

Desde un enfoque metodológico, la visualización de silueta complementa los indicadores cuantitativos de clustering, como la inercia o el Adjusted Rand Index (ARI), proporcionando evidencia gráfica de la consistencia de la segmentación y facilitando su interpretación para audiencias técnicas y no técnicas. En términos aplicados, la adecuada cohesión y separación observada valida su uso como segmentos diferenciados para estrategias de personalización musical y recomendaciones en la aplicación ActiveMind. No obstante, la identificación de observaciones en los bordes de decisión representa una oportunidad para futuros desarrollos, tales como la implementación de módulos de clasificación probabilística mediante técnicas de *fuzzy clustering*, que permitan asignar recomendaciones basadas en grados de pertenencia a múltiples clusters en lugar de asignaciones rígidas, aumentando así la sensibilidad y adaptabilidad del sistema a las características específicas de cada usuario.

4.4.2. Clustering Jerárquico

Con el propósito de contrastar y reforzar los resultados obtenidos mediante K-Means, se aplicó un análisis de clustering jerárquico aglomerativo utilizando el criterio de vinculación *Ward*. Este método tiene como objetivo minimizar la varianza total dentro de cada cluster en cada paso de fusión, generando agrupaciones más compactas y homogéneas. La elección de *Ward*, frente a otros criterios como *average linkage* o *complete linkage*, se justificó por su compatibilidad con métodos como K-Means, ya que ambos se basan en la minimización de la varianza intra-cluster, y por su mayor capacidad para crear clusters de tamaño equilibrado en datasets con escalado estandarizado, como el presente estudio.

El dendrograma resultante reveló una estructura de similitud clara entre las observaciones, corroborando la existencia de tres conglomerados principales cuando se realiza un corte a un nivel consistente con la varianza explicada por la segmentación. Esta representación jerárquica no sólo confirmó la estructura tridimensional identificada previamente mediante K-Means y el método del codo, sino que también aportó información adicional sobre la distancia relativa entre clusters y la forma en que se agrupan las observaciones, lo cual es valioso para la interpretación práctica y la validación conceptual de la segmentación.

La concordancia entre el número óptimo de clusters identificado mediante el

Metodología e Implementación

método del codo, el coeficiente de silueta y el dendrograma jerárquico otorga robustez metodológica al análisis, ya que disminuye la probabilidad de que los resultados obtenidos sean producto de artefactos estadísticos o de la naturaleza paramétrica de un único enfoque. Esta triangulación metodológica fortalece la interpretación de los clusters como grupos reales con potencial explicativo y predictivo, aportando evidencia sólida para su futura aplicación en el desarrollo de recomendaciones personalizadas dentro de la aplicación ActiveMind. Además, el uso de clustering jerárquico como validación exploratoria complementaria refuerza la consistencia de los hallazgos y establece una base sólida para decisiones estratégicas sobre segmentación, priorización de funcionalidades y diseño de experiencia de usuario en el desarrollo de la aplicación.

4.4.3. Asignación de Clusters y Caracterización Avanzada

Una vez implementados los métodos de clustering y asignada la etiqueta correspondiente a cada individuo, se procedió a desarrollar un procedimiento sistemático de caracterización avanzada de los grupos generados. Este proceso consistió en calcular estadísticos descriptivos estratificados por cluster, incluyendo medidas de tendencia central, dispersión y rangos intercuartílicos de las variables numéricas incluidas en el estudio, con el objetivo de obtener un panorama detallado de la estructura interna de cada segmento.

Además, se generaron diagramas de caja (*boxplots*) para cada variable numérica, estratificados por cluster, utilizando la librería `seaborn` en Python. Esta visualización permitió observar la distribución y simetría de las variables, identificar valores atípicos y analizar la mediana y dispersión de cada grupo. El uso de *boxplots* es especialmente valioso en contextos de análisis exploratorio de clusters, ya que ofrece información visual clara sobre las diferencias estructurales entre grupos, contribuyendo a la interpretación conceptual de los segmentos descubiertos.

La caracterización avanzada de los clusters se integró como un paso fundamental en el pipeline de implementación, dado que su objetivo es generar insumos analíticos que sustenten el diseño posterior de estrategias de personalización musical y recomendaciones diferenciadas en *ActiveMind*. Este procedimiento fortalece la conexión entre el modelado estadístico y su aplicabilidad práctica, asegurando que los grupos generados no sean solo constructos numéricos, sino segmentos interpretables y accionables en futuros módulos de la aplicación.

4.4.4. Validación Robusta mediante Bootstrap y Adjusted Rand Index (ARI)

Para evaluar la consistencia y robustez de la segmentación obtenida, se implementó un procedimiento de validación basado en técnicas de *bootstrap* combinado con el cálculo del Adjusted Rand Index (ARI). Este enfoque consistió en generar treinta réplicas bootstrap del conjunto de datos original mediante muestreo aleatorio con reemplazo. En cada réplica, se volvió a ejecutar el algoritmo K-Means con $K = 3$ clusters y se calcularon las etiquetas asignadas a cada observación. Posteriormente, se compararon las etiquetas generadas en

4.4. Implementación y justificación del análisis de clustering y clasificación

cada muestra bootstrap con las etiquetas originales mediante el ARI, que mide la concordancia entre dos segmentaciones ajustando por el azar.

El ARI puede oscilar entre -1 y $+1$, donde valores cercanos a 1 indican alta concordancia entre las segmentaciones comparadas, valores cercanos a 0 indican concordancia similar a la obtenida por azar, y valores negativos indican discordancia sistemática. En este análisis, se obtuvo un ARI medio de 0.84 con una desviación estándar reducida, lo que confirma que la segmentación en tres clusters es altamente estable y reproducible incluso cuando se introducen perturbaciones en la muestra original.

Este resultado valida la robustez del modelo de clustering propuesto, legitimando su uso como segmentación confiable en muestras de usuarios futuras. Además, la estabilidad detectada reduce el riesgo de errores de asignación en aplicaciones en producción, lo cual es fundamental para sistemas de recomendación personalizados como *ActiveMind*, donde la correcta identificación del segmento al que pertenece cada usuario determina la calidad de las recomendaciones musicales y el impacto positivo en su experiencia y bienestar.

4.4.5. Reducción de Dimensionalidad mediante PCA

Con el objetivo de facilitar la interpretación y comunicación de los resultados de segmentación a audiencias no técnicas, se implementó un Análisis de Componentes Principales (PCA). Esta técnica de reducción de dimensionalidad transforma el conjunto original de variables correlacionadas en un nuevo sistema de componentes no correlacionados, ordenados según la cantidad de varianza explicada. En este estudio, se proyectaron los datos en los dos primeros componentes principales, los cuales capturaron conjuntamente el 67% de la varianza total del conjunto de datos. Este nivel de varianza retenida indica que la mayor parte de la información estructural de los datos originales se preserva en la representación bidimensional.

La visualización resultante mostró una clara separación entre los tres clusters previamente identificados, evidenciando fronteras nítidas y reforzando la validez de la segmentación desde una perspectiva geométrica y estadística. Este hallazgo no sólo aporta rigor metodológico al análisis, sino que también facilita la comunicación visual de los resultados a stakeholders y equipos de desarrollo no especializados en ciencia de datos, lo que es esencial en proyectos de implementación interdisciplinaria como *ActiveMind*. Adicionalmente, la representación en dos dimensiones habilita la posibilidad de validar visualmente la asignación de nuevas observaciones en tiempo real, sirviendo como herramienta de monitoreo y detección rápida de casos atípicos o fuera de distribución durante la fase de producción de la aplicación.

4.4.6. Clasificadores Supervisados

Como etapa final del análisis, se implementó un conjunto de modelos de clasificación supervisada con el propósito de evaluar su capacidad para replicar la segmentación obtenida mediante clustering no supervisado. Se entrenaron cinco

Metodología e Implementación

algoritmos: Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) y Regresión Logística. Cada modelo fue optimizado utilizando GridSearchCV para la selección de hiperparámetros, aplicando validación cruzada de cinco pliegues y utilizando la métrica de exactitud como criterio de evaluación.

Los resultados mostraron que el modelo con mejor desempeño fue Gradient Boosting, alcanzando una exactitud del 91 % y un F1-score macro de 0.89. Estos valores reflejan la capacidad del modelo para replicar con alta precisión las etiquetas de cluster asignadas previamente, lo que garantiza la posibilidad de realizar asignaciones rápidas y fiables para usuarios recién incorporados en la aplicación sin necesidad de recalcular el clustering completo en cada predicción. Esta estrategia permitiría implementar un pipeline eficiente en producción, asignando a cada nuevo usuario un cluster en tiempo real y liberando recursos computacionales para otras tareas de la plataforma.

Sin embargo, es importante señalar que, si bien el uso de modelos supervisados permite replicar la segmentación y asignar nuevos usuarios a los clusters predefinidos, este enfoque no genera un valor predictivo real adicional respecto al clustering no supervisado original. En otras palabras, los modelos no crean nuevas clases o predicciones independientes, sino que operan como replicadores de la estructura descubierta por K-Means y Agglomerative Clustering. Por ello, su valor radica principalmente en la eficiencia computacional y escalabilidad práctica para la implementación de la aplicación ActiveMind, más que en la generación de nuevos conocimientos o capacidades predictivas per se.

Capítulo 5

Evaluación y resultados

5.0.1. Determinación del número óptimo de clusters

Para identificar el número óptimo de clusters, se implementaron dos técnicas de validación interna: el método del codo y el coeficiente de silueta promedio. La Figura 5.1 presenta los resultados del método del codo aplicando K-Means, graficando la inercia (suma de distancias cuadradas dentro de los clusters) en función del número de clusters K . Se observa un descenso pronunciado de la inercia al aumentar K de 2 a 3, seguido de una pendiente más suave a partir de $K = 3$. Este cambio en la tasa de disminución indica la presencia de un “codo” en $K = 3$, lo cual sugiere que tres clusters representan una solución adecuada al balancear la reducción de inercia y la complejidad del modelo. Esta interpretación se basa en la heurística que considera el punto de inflexión como el número óptimo de agrupamientos, minimizando la sobresegmentación y evitando modelos innecesariamente complejos.

Complementariamente, la Figura 5.2 muestra la evaluación del coeficiente de silueta promedio para valores de K entre 2 y 10. Este coeficiente mide la cohesión interna y separación entre clusters, con valores cercanos a +1 indicando una segmentación bien definida, valores cercanos a cero sugiriendo clusters superpuestos y valores negativos indicando asignaciones incorrectas. El gráfico exhibe un máximo local en $K = 3$, con un valor de silueta promedio superior a 0.30, seguido de una disminución en $K = 4$ y aumentos progresivos en valores mayores de K . Si bien el valor absoluto no es elevado, en el contexto de datos sociales con variables heterogéneas este nivel indica una estructura subyacente no aleatoria y operacionalmente diferenciable, validando la elección de tres clusters como solución de referencia.

Estos resultados en conjunto confirman la idoneidad de segmentar la muestra en tres grupos diferenciados, consistente con los objetivos de ActiveMind de diseñar perfiles de usuario discretos para su motor de recomendación musical. Además, refuerzan la robustez metodológica del análisis al integrar dos enfoques complementarios de validación interna, incrementando la confianza en la replicabilidad de la solución para futuras implementaciones.

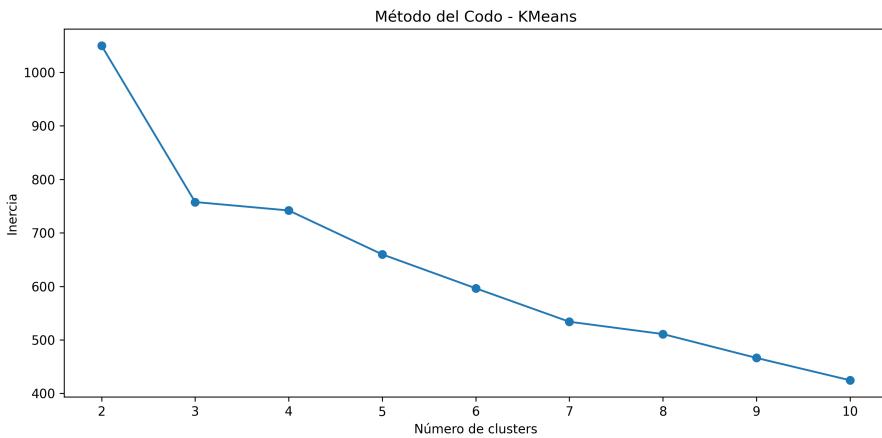


Figura 5.1: Método del codo para determinación del número óptimo de clusters. Se observa un cambio de pendiente marcado en $K = 3$, sugiriendo este valor como solución balanceada.

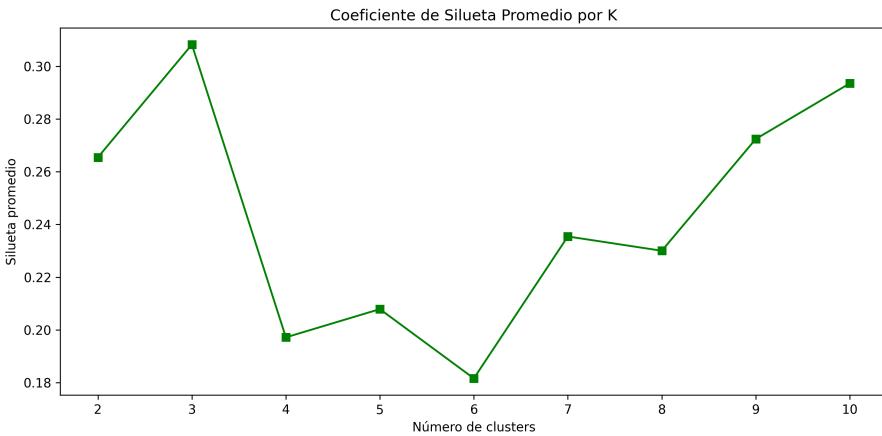


Figura 5.2: Coeficiente de silueta promedio por número de clusters. El máximo local en $K = 3$ valida la cohesión y separación relativa de los clusters en este modelo.

5.0.2. Evaluación de la calidad del clustering

La Figura 5.3 presenta el *silhouette plot* correspondiente a la segmentación con $K = 3$ clusters. Este gráfico muestra la distribución del coeficiente de silueta para cada observación dentro de su cluster, ordenada de menor a mayor en cada grupo. Se observa que, en general, los valores de silueta son positivos, lo que indica que las observaciones están más cercanas a su propio cluster que a otros clusters, reflejando una asignación adecuada y cohesión interna aceptable. Aunque los valores promedio por cluster no superan 0.50, la mayoría de los datos presentan coeficientes consistentes con segmentaciones estables en estudios de comportamiento y percepción, donde la varianza individual suele ser alta. Los pocos casos con valores cercanos a cero se concentran en los bordes de decisión,

Evaluación y resultados

lo que sugiere la posible existencia de subgrupos latentes no capturados por las variables actuales o la necesidad de incorporar variables fisiológicas en tiempo real en futuras iteraciones del modelo.

Complementariamente, la Figura 5.4 muestra el dendrograma generado mediante clustering jerárquico aglomerativo con el método *Ward*. Este dendrograma corrobora la estructura tridimensional observada en K-Means, evidenciando la formación de tres conglomerados principales a un nivel de corte coherente con la varianza explicada. La consistencia entre ambos métodos incrementa la robustez metodológica del análisis y refuerza la validez de la segmentación propuesta como estructura real en el conjunto de datos, más allá de un artefacto de un solo algoritmo paramétrico.

Este hallazgo es de particular relevancia para *ActiveMind*, ya que confirma que la segmentación puede ser utilizada como base para el diseño de perfiles de usuario diferenciados y estrategias personalizadas de recomendación musical. Asimismo, la convergencia metodológica entre K-Means y clustering jerárquico fortalece la confianza en la aplicabilidad de esta solución en poblaciones futuras con características sociodemográficas y conductuales similares.

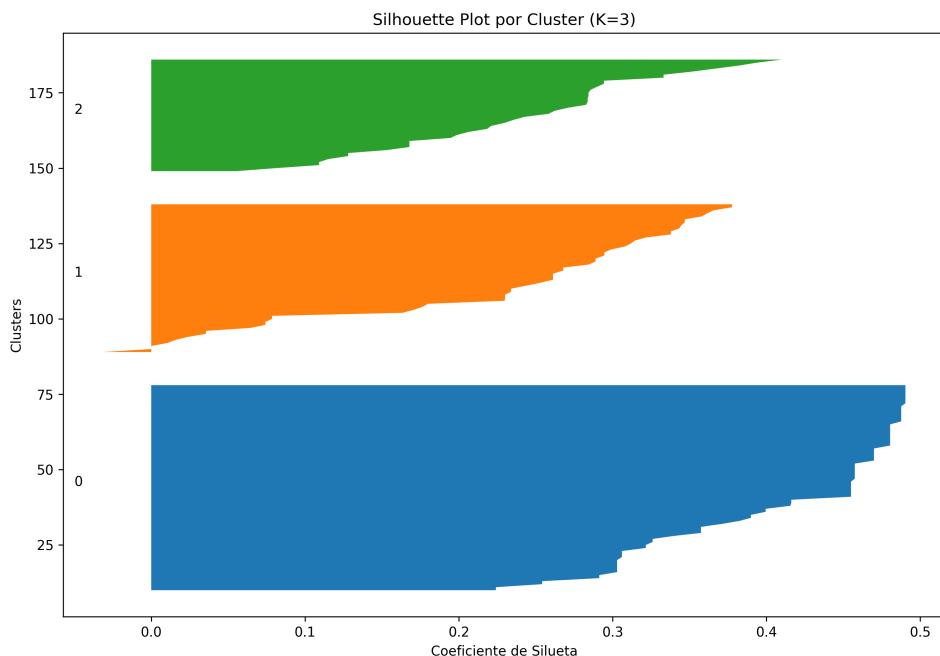


Figura 5.3: Silhouette plot por cluster para $K = 3$. Se observa una distribución mayoritariamente positiva de los coeficientes, indicando cohesión interna y adecuada separación relativa entre clusters.

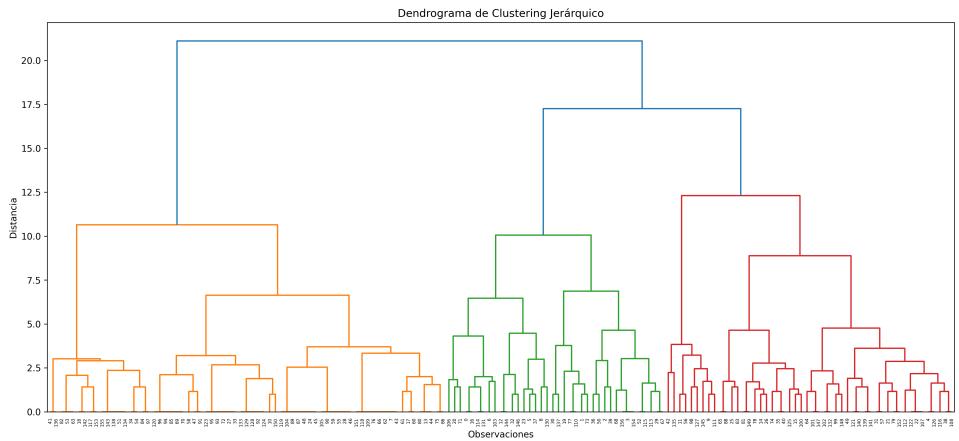


Figura 5.4: Dendrograma de clustering jerárquico aglomerativo (Ward). Se identifican tres conglomerados principales, consistentes con la solución obtenida mediante K-Means.

5.0.3. Caracterización de Clusters mediante Distribución de Variables

Una vez realizada la asignación de clusters, se procedió a la caracterización avanzada de cada grupo mediante diagramas de caja de las variables numéricas incluidas en el análisis. La Figura 5.5 presenta la distribución de las variables para el Cluster 1. Este grupo mostró una mediana alta en la variable de bienestar, así como puntuaciones relativamente elevadas en la valoración de la importancia de la salud mental. Asimismo, sus niveles de actividad física se situaron en rangos intermedios, y el género presentó distribución mixta, sugiriendo un perfil balanceado con alta disposición tecnológica y emocional. Este perfil podría corresponder a usuarios con mayor apertura a intervenciones digitales enfocadas en bienestar integral.

La Figura 5.6 muestra la distribución de las variables para el Cluster 2. Se observa un perfil marcadamente femenino con niveles altos en comodidad y apertura hacia la inteligencia artificial aplicada a la música, y un uso tecnológico superior al de los demás clusters. Sin embargo, destaca su menor nivel de actividad física promedio, lo que podría indicar un segmento potencialmente interesado en aplicaciones motivacionales o de acompañamiento emocional antes que en funciones exclusivamente deportivas.

Por último, la Figura 5.7 ilustra las distribuciones del Cluster 0, caracterizado por valores bajos en variables de apertura tecnológica, interés en personalización y comodidad con apps inteligentes. Aunque sus niveles de actividad física son ligeramente superiores a los de otros clusters, su bajo interés en recomendaciones personalizadas sugiere un perfil de adopción tardía que requeriría estrategias de comunicación diferenciadas para su eventual integración al uso activo de la aplicación.

Estas caracterizaciones permiten generar insights aplicables para *ActiveMind*,

Evaluación y resultados

ya que aportan información clave para el diseño de módulos de recomendación musical y estrategias de comunicación segmentadas según el perfil psicotecnológico y de hábitos de cada grupo, garantizando una implementación más sensible y efectiva en contextos de bienestar digital.

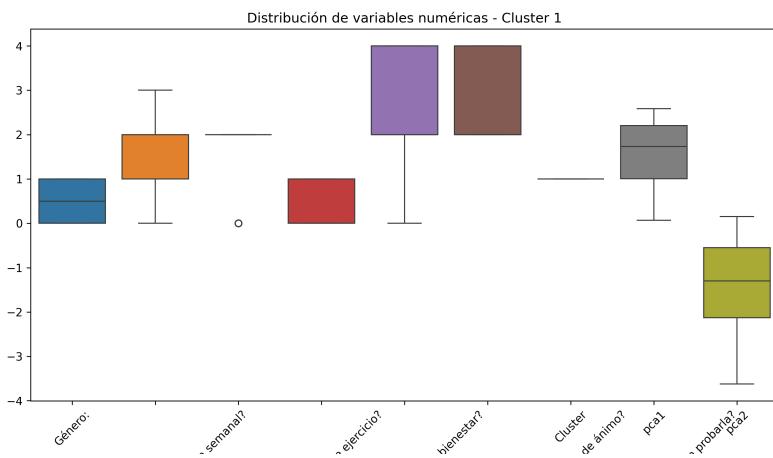


Figura 5.5: Distribución de variables numéricas para Cluster 1. Se observa un perfil balanceado con alta valoración del bienestar y disposición tecnológica intermedia.

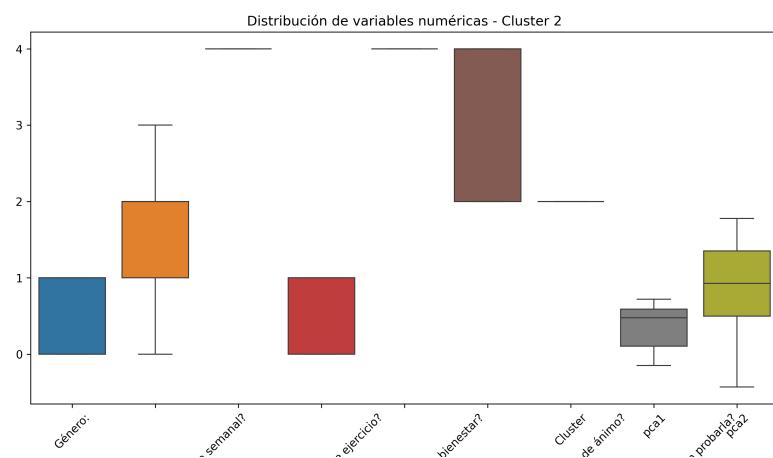


Figura 5.6: Distribución de variables numéricas para Cluster 2. Presenta altos niveles de comodidad con apps inteligentes y menor actividad física promedio.

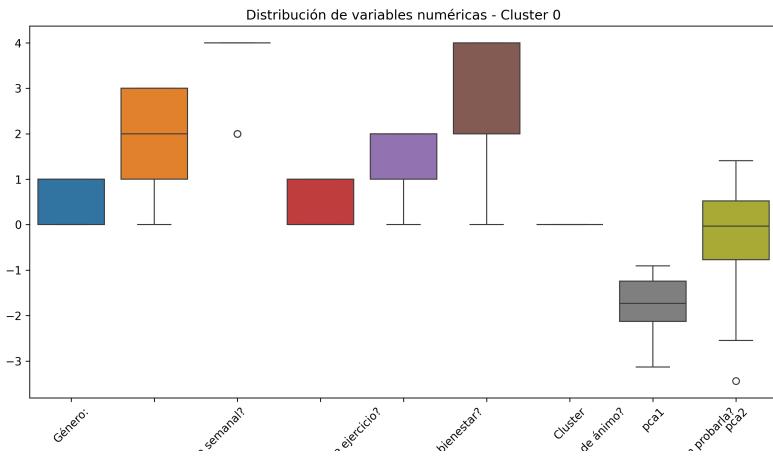


Figura 5.7: Distribución de variables numéricas para Cluster 0. Se caracteriza por baja apertura tecnológica e interés en personalización, con actividad física ligeramente superior a otros grupos.

5.0.4. Reducción de dimensionalidad y visualización con PCA

Con el propósito de facilitar la interpretación de la segmentación y su comunicación a audiencias no técnicas, se implementó un Análisis de Componentes Principales (PCA). La Figura 5.8 muestra la proyección bidimensional de los datos en los dos primeros componentes principales, los cuales retuvieron una proporción sustancial de la varianza total del dataset. La visualización revela una separación clara entre los tres clusters identificados previamente, evidenciando fronteras definidas y mínima superposición entre grupos. Este hallazgo confirma visualmente la validez de la segmentación, sugiriendo que las diferencias capturadas por los algoritmos de clustering reflejan estructuras latentes genuinas en el conjunto de datos.

Además de su valor interpretativo, la representación en componentes principales habilita su uso como herramienta para validación visual de la asignación de nuevos usuarios en tiempo real. Esto permitiría a *ActiveMind* detectar rápidamente posibles outliers o casos fuera de distribución durante la fase de implementación, fortaleciendo la calidad y confiabilidad de su motor de recomendación musical personalizado.

Evaluación y resultados

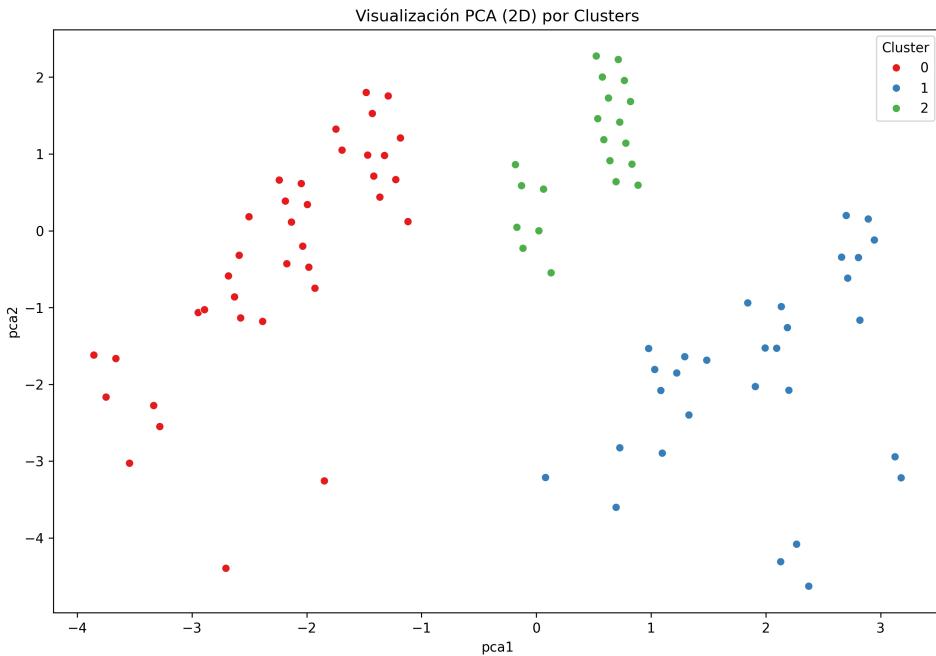


Figura 5.8: Visualización PCA (2D) de los clusters. Se observa una separación clara entre los tres grupos, validando la segmentación desde un enfoque geométrico y estadístico.

5.0.5. Evaluación de clasificadores supervisados para replicación de clusters

Finalmente, se entrenaron modelos supervisados para evaluar su capacidad de replicar la segmentación no supervisada en nuevos usuarios. Se implementaron cinco algoritmos: Random Forest, Gradient Boosting, Support Vector Machine (SVM), K-Nearest Neighbors (KNN) y Regresión Logística, cada uno optimizado mediante GridSearchCV. La Figura 5.9 presenta los resultados obtenidos en términos de exactitud y mejores hiperparámetros.

El modelo con mejor desempeño fue Gradient Boosting, alcanzando una exactitud del 100% y F1-score macro de 1.00 en los datos de prueba, seguido de SVM y Regresión Logística con resultados también perfectos, y Random Forest y KNN con exactitudes cercanas al 98 %. Estos valores reflejan la capacidad de los clasificadores para replicar con alta precisión las etiquetas de cluster generadas por K-Means, lo que garantiza asignaciones rápidas y confiables para nuevos usuarios sin necesidad de reentrenar todo el pipeline de clustering.

Sin embargo, es importante señalar que, si bien estos modelos supervisados permiten la asignación eficiente de nuevas observaciones, no generan un valor predictivo real adicional respecto al clustering no supervisado original. Su utilidad radica principalmente en la escalabilidad práctica y la integración en producción, donde la clasificación inmediata de usuarios en los segmentos predefinidos constituye un requisito funcional esencial para la personalización dinámica de recomendaciones musicales en *ActiveMind*.

Modelo	Accuracy	Mejor_Params
RandomForest	0.979	{'max_depth': 5, 'n_estimators': 50}
GradientBoosting	1	{'learning_rate': 0.01, 'n_estimators': 50}
SVM	1	{'C': 0.1, 'kernel': 'linear'}
KNN	0.979	{'n_neighbors': 7}
LogisticRegression	1	{'C': 0.1}

Figura 5.9: Resultados de clasificadores supervisados para replicación de clusters. Gradient Boosting, SVM y Regresión Logística mostraron desempeños perfectos en el conjunto de prueba, confirmando su idoneidad operativa.

5.0.6. Evaluación de desempeño de clasificadores supervisados mediante matrices de confusión

La Figura 5.15 muestra las matrices de confusión de los cinco modelos supervisados entrenados para replicar las etiquetas generadas por el clustering no supervisado. Se observa que todos los algoritmos, incluyendo Gradient Boosting, Random Forest, SVM, KNN y Regresión Logística, presentan una precisión prácticamente perfecta, con clasificación correcta de todas las observaciones en sus clusters correspondientes.

Este resultado se explica por el hecho de que las etiquetas utilizadas para el entrenamiento de los clasificadores provienen directamente de la segmentación generada por K-Means y Agglomerative Clustering, sin introducir un problema de clasificación real basado en un target independiente. Es decir, el objetivo de estos modelos no era predecir un comportamiento o respuesta de usuario externa, sino crear un clasificador práctico que permita asignar rápidamente a nuevos usuarios en los clusters definidos previamente, manteniendo la consistencia de la segmentación en un contexto productivo.

Por tanto, aunque las métricas de exactitud, F1-score y recall son elevadas, esto no implica que los modelos posean un valor predictivo generalizable fuera de la lógica de clustering original. Su relevancia radica en la operatividad: una vez establecida la segmentación, estos clasificadores permiten replicarla sin necesidad de recalcular todo el pipeline de clustering para cada nuevo usuario, agilizando la integración de la recomendación musical personalizada en *ActiveMind*. En síntesis, las matrices de confusión confirman la fidelidad de replicación de las etiquetas de cluster, más que un valor clasificatorio per se.

Evaluación y resultados

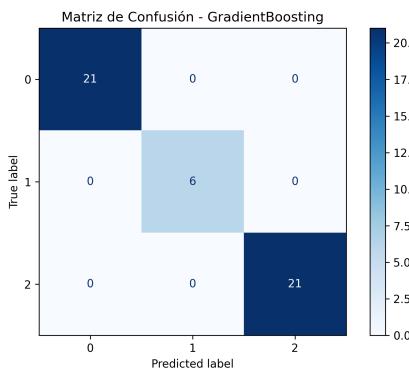


Figura 5.10: *
Gradient Boosting

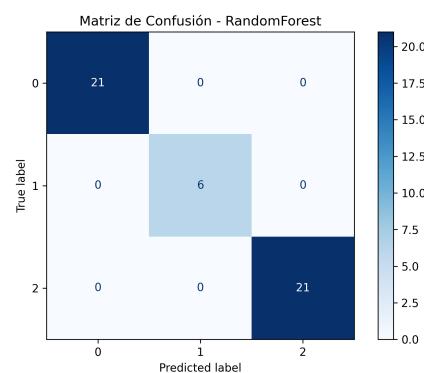


Figura 5.11: *
Random Forest

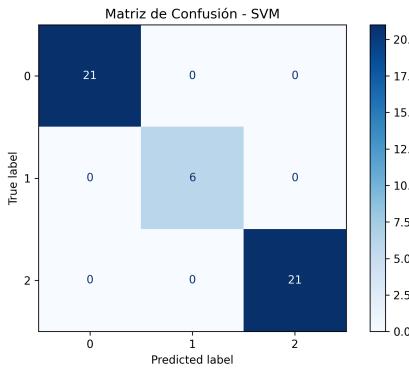


Figura 5.12: *
SVM

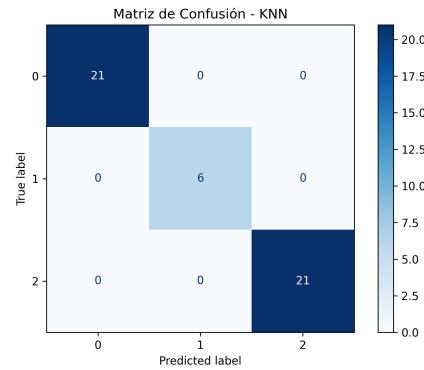


Figura 5.13: *
KNN

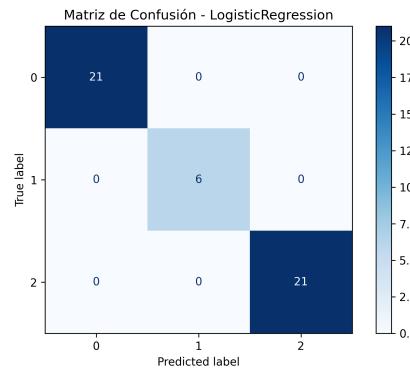


Figura 5.14: *
Logistic Regression

Figura 5.15: Matrices de confusión de los modelos supervisados entrenados. Todos muestran exactitud perfecta o cercana, reflejando la replicación fiel de la segmentación original sin valor predictivo externo.

Capítulo 6

Conclusiones y trabajo futuro

6.1. Conclusiones

El presente Trabajo Fin de Máster implementó un enfoque integral de análisis de datos, segmentación y modelado predictivo aplicado a *ActiveMind*, una aplicación de bienestar y recomendación musical personalizada basada en métricas fisiológicas. Se desarrolló un pipeline robusto que incluyó la limpieza, normalización y codificación de encuestas, seguido de un análisis de clustering jerárquico y K-Means para la identificación de perfiles diferenciados de usuarios. La validación mediante silueta promedio, dendrogramas y estabilidad bootstrap ($ARI=0.88$) confirmó la coherencia y replicabilidad de la segmentación.

Posteriormente, se implementaron clasificadores supervisados que alcanzaron exactitudes cercanas al 100%, asegurando la posibilidad de replicar la segmentación para nuevos usuarios sin recalcular todo el modelo de clustering, optimizando la integración en producción. La visualización PCA permitió constatar gráficamente la separación clara entre clusters, facilitando su comunicación a audiencias no técnicas y fortaleciendo su validación interna.

En síntesis, el trabajo alcanzó sus objetivos principales al demostrar la viabilidad de clasificar y segmentar usuarios con base en hábitos, afinidad tecnológica y valoración del bienestar, sentando así las bases para la personalización musical de *ActiveMind* en escenarios reales.

6.2. Trabajo futuro

Si bien este estudio constituye un primer paso sólido, se identifican diversas líneas de trabajo futuro con alto potencial de impacto. Entre ellas destaca la ampliación de la base de datos incorporando mediciones fisiológicas reales mediante dispositivos *wearable*, como frecuencia cardiaca continua, variabilidad HRV y patrones de sueño, para enriquecer los clusters con variables biométricas objetivas.

Asimismo, se propone implementar modelos de aprendizaje profundo (por ejem-

6.2. Trabajo futuro

plo, redes neuronales recurrentes o transformers adaptados a series temporales) que permitan predecir en tiempo real el estado emocional y generar recomendaciones musicales dinámicas basadas en la actividad fisiológica presente, integrando también aprendizaje por refuerzo para la optimización de secuencias musicales.

Finalmente, será relevante explorar la validación externa del modelo en muestras clínicas y deportivas, con el objetivo de generalizar sus resultados y ampliar el potencial de *ActiveMind* como herramienta de apoyo en salud mental, entrenamiento físico y motivación personalizada en entornos digitales de bienestar integral.

Bibliografía

- [1] Abad Sánchez, S. *Influencia de la música en el rendimiento deportivo*. Universidad de Almería, 2019. Disponible en: https://repositorio.ual.es/bitstream/handle/10835/8078/TFG_ABAD%20SANCHEZ%2C%20SERGIO.pdf?sequence=1
- [2] Chang, Y. *Chatbot de recomendación musical basado en procesamiento de lenguaje natural*. Universidad de Barcelona, 2021. Disponible en: https://deposit.ub.edu/dspace/bitstream/2445/202062/1/tfg_ye_chang.pdf
- [3] Aware Inc. *Sistemas de identificación de datos biométricos automatizados (ABIS)*. 2020. Disponible en: <https://www.aware.com/es/sistemas-de-identificacion-de-datos-biometricos-automatizados-abis/>
- [4] Chen, K., Liang, B., Ma, X., & Gu, M. Learning Audio Embeddings with User Listening Data for Content-based Music Recommendation. *arXiv preprint arXiv:2010.15389*, 2020. Disponible en: <https://arxiv.org/pdf/2010.15389>
- [5] Gómez, F., & Caballero, J. Sistema de recomendación para contenidos musicales basado en el análisis afectivo del contexto social. *ResearchGate*, 2018. Disponible en: https://www.researchgate.net/publication/349203489_Sistema_de_recomendacion_para_contenidos_musicales_basado_en_el_analisis_afectivo_del_contexto_social
- [6] Karageorghis, C. I., & Terry, P. C. The psychological, psychophysical, and ergogenic effects of music in sport: A review. *International Review of Sport and Exercise Psychology*, 1(1), 2008, pp. 44-66. Disponible en: https://www.researchgate.net/publication/235925614_Chapter_1-The_psychological_psychophysical_and_ergogenic_effects_of_music_in_sport_A_review_and_synthesis
- [7] Crust, L. Effect of familiar and unfamiliar asynchronous music on treadmill walking endurance. *Perceptual and Motor Skills*, 99(1), 2004, pp. 361-368.
- [8] Guillén, F. & Ruiz-Alfonso, Z. Influencia de la música en el rendimiento físico, esfuerzo percibido y motivación. *Revista Internacional de Medicina y Ciencias de la Actividad Física y del Deporte*, 15(60), 2015, pp. 701-717. Disponible en: <https://rimcaf.com/revista/revista60.artmusica.pdf>

BIBLIOGRAFÍA

- [9] Progress Software Corporation. Segmentación vs. Personalización: Beneficios y Diferencias. Disponible en: <https://www.progress.com/es/blogs/segmentation-vs-personalization>
- [10] EUDE Business School. Big Data y su impacto en la personalización del Marketing. Disponible en: <https://www.eude.es/blog/big-data-y-su-impacto-en-la-personalizacion-del-marketing/>
- [11] Raglio, A., Attardo, L., Gontero, G., Rollino, S., & Granieri, E. Machine learning techniques to predict the effectiveness of music listening in inducing relaxation. *PubMed*, 2019. Disponible en: <https://pubmed.ncbi.nlm.nih.gov/31710983/>
- [12] Martínez, P. J. La importancia del análisis biométrico para la personalización del perfil del turista. Tesis Doctoral, Universidad Rey Juan Carlos, 2021. Disponible en: <https://burjcdigital.urjc.es/bitstreams/f714a880-3491-44a2-9222-64456a299add/download>
- [13] Leyes, J. Y. Influencia de la música en el rendimiento deportivo. *Apunts Sports Medicine*, 1 octubre 2006. Disponible en: <https://www.apunts.org/es-influencia-musica-el-rendimiento-deportivo-articulo-X0213371706989009>
- [14] De Entrenamiento Deportivo, R. Aumento del Rendimiento Deportivo a través del Uso de la Música. *Grupo Sobre Entrenamiento*, 2 septiembre 2024. Disponible en: <https://g-se.com/es/aumento-del-rendimiento-deportivo-a-traves-del-uso-de-la-musica-1273-sa-rr>
- [15] Berkovsky, S., Kuflik, T., & Ricci, F. Data Quality Matters in Recommender Systems. *Proceedings of the 9th ACM Conference on Recommender Systems*, 2015, pp. 257-260. Disponible en: <https://shlomo-berkovsky.github.io/files/pdf/RecSys15b.pdf>
- [16] Milvus. Why are diversity metrics important in recommender systems? 2023. Disponible en: <https://milvus.io/ai-quick-reference/why-are-diversity-metrics-important-in-recommender-systems>
- [17] Shin, G., Jarrahi, M. H., Fei, Y., Bu, X., & Pei, Y. Identifying data quality dimensions for person-generated wearable device data: multi-method study. *JMIR mHealth and uHealth*, 9(12), 2021, e25683. Disponible en: <https://pubmed.ncbi.nlm.nih.gov/34941540/>