

ANÁLISIS DE FACTORES INFLUYENTES EN LA FELICIDAD PERSONAL

UN ENFOQUE DE APRENDIZAJE AUTOMÁTICO
UTILIZANDO DATOS MULTIDIMENSIONALES

ANÁLISIS DE FACTORES INFLUYENTES EN LA FELICIDAD PERSONAL

UN ENFOQUE DE APRENDIZAJE AUTOMÁTICO
UTILIZANDO DATOS MULTIDIMENSIONALES

Francesc Xavier REVERTÉ BARÓ



TRABAJO FINAL DE MÁSTER
Máster en Ciencia de Datos

Tutores:

María Beatriz LÓPEZ IBÁÑEZ

Núria FORNALÉS ORTEU

Septiembre 2023

Título:

**ANÁLISIS DE FACTORES INFLUYENTES
EN LA FELICIDAD PERSONAL
UN ENFOQUE DE APRENDIZAJE AUTOMÁTICO
UTILIZANDO DATOS MULTIDIMENSIONALES**

Copyright © Xavier Reverté Baró

Edición en papel:

ISBN: 978-84-291-6469-5

Propiedad de:

EDITORIAL REVERTÉ, S. A.

Loreto, 13-15, Local B

08029 Barcelona

Tel: (+34) 93 419 33 36

reverte@reverte.com

www.reverte.com

Reservados todos los derechos. La reproducción total o parcial de esta obra, por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos, queda rigurosamente prohibida sin la autorización escrita del titular del copyright, bajo las sanciones establecidas por las leyes.

Impreso en España · *Printed in Spain*

Depósito Legal: B 20756-2023

Impresión: Liberdigital

Resumen

El objetivo principal de este trabajo es identificar y comprender los factores que influyen en la felicidad personal de individuos en diferentes contextos y situaciones. Para lograrlo, empleamos dos conjuntos de datos, 'Encuesta' y 'Personal', y aplicamos una combinación de técnicas de análisis de datos y aprendizaje automático adaptadas a sus particularidades.

'Encuesta' consiste en un formulario diseñado para describir el perfil personal del encuestado, definir los tópicos que más influyen en su felicidad y revelar también el conjunto de actividades y hábitos más influyentes. Por otro lado, 'Personal' captura las actividades y hábitos diarios de diferentes participantes, incluyendo la variable objetivo: la evaluación subjetiva de felicidad percibida cada día.

Comenzando la metodología con el estudio 'Encuesta', aplicamos la técnica 'K-means' para clusterizar el conjunto de datos y determinar las actividades más influyentes en la felicidad de cada grupo mediante un análisis de diferencias significativas. También construimos y ensamblamos modelos de aprendizaje automático supervisado para predecir las etiquetas asignadas y extraer la importancia de las características en la predicción de grupos, identificando cuáles de las diferencias entre grupos son las más importantes para su caracterización.

En cuanto a 'Personal', construimos y ensamblamos múltiples modelos de aprendizaje automático supervisado para predecir el nivel de felicidad diario en función de las distintas actividades registradas. Esto nos permitió identificar las variables que tienen un mayor impacto en la felicidad de cada participante. Adicionalmente, cada participante respondió también la encuesta, permitiendo la determinación de los grupos a los que pertenecen, y de los supuestos factores que más influyen en su felicidad.

Entrelazando ambas aproximaciones, anticipamos las variables más influyentes de cada participante en 'Personal' en función de sus segmentaciones en 'Encuesta'. Esto proporciona una guía para enfocar la atención en las variables más relevantes de cada participante y, al mismo tiempo, confirmar o refutar generalidades previamente identificadas en 'Encuesta' mediante el estudio de 'Personal'.

Los **resultados** revelan discrepancias entre las expectativas basadas en la segmentación ('Encuesta') y las influencias reales según el estudio 'Personal'. Además, desvelan que los aspectos más influyentes en la felicidad varían entre los participantes, pero las actividades relacionadas con el cansancio, los factores externos (como el clima), y el tiempo libre, tienden a ser las más influyentes.

En resumen, este trabajo **aporta valor** al emplear conjuntos de datos diversos y una metodología adaptativa según sus particularidades. Contribuye al conocimiento del bienestar humano al ofrecer una visión más completa de cómo diferentes factores interactúan para influir en la felicidad individual. Además, aborda la brecha entre percepciones ('Encuesta') y comportamientos ('Personal'), lo que puede ser relevante en la investigación psicológica y sociológica.

Agradecimientos

Deseo expresar mi gratitud a todas las personas e instituciones que han contribuido a la realización de este trabajo. Sus apoyos, orientación y colaboraciones han dejado una huella imborrable en este proceso y en mi crecimiento personal y profesional.

Agradezco al conglomerado de la universidad por brindarme la oportunidad de sumergirme en el apasionante mundo de la ciencia de datos. Sus enseñanzas han reafirmado mi pasión por esta disciplina y han desafiado mis preconcepciones sobre el sistema educativo.

A mi tutora a lo largo del trabajo, María Beatriz López Ibáñez, cuyo equilibrado enfoque entre seguimiento y fomento de mi libertad creativa, así como sus enfoques, comentarios y aportaciones, fueron indispensables para el desarrollo de este trabajo.

Quiero extender mi agradecimiento a la Fundación Drissa por su apoyo y por las valiosas oportunidades brindadas. Expreso mi reconocimiento a la psicóloga Núria Fornalés Orteu, cuyas instrucciones y orientación han añadido una perspectiva enriquecedora a este proyecto.

Mi más sincero agradecimiento se extiende a mi familia, por su comprensión, apoyo incondicional, y por permitirme concentrarme en esta tarea, descuidando las tareas del hogar. Su paciencia y tolerancia durante las etapas finales del proyecto han sido de vital importancia.

Por último, pero quizás lo más importante, deseo expresar mi profundo agradecimiento a los verdaderos protagonistas de los datos: a todos los encuestados y, en particular, a todos los participantes en el estudio de datos personales. Sin su voluntaria participación y generosa contribución, este trabajo carecería de su base fundamental. Cada dato ha sido un pilar en la construcción de este estudio y en la búsqueda de conocimiento.

En resumen, este trabajo no habría sido posible sin el apoyo y las contribuciones invaluables de muchas personas. A todos ustedes, les ofrezco mi sincero agradecimiento por su influencia en esta travesía de aprendizaje y descubrimiento.

Charlie: 'Haz de este trabajo tu portafolio laboral'

Edogawa: '¡Yo ya he cumplido!'

Pato: 'Deberías registrar el trabajo'

Juju: 'Que pesado...'

Ajara: 'Tú puedes'

Dolphin: 'Cuenta conmigo para un posterior estudio'

Relookyou: 'Conócete a tí mismo'

Mack: 'Por fin has encontrado lo que te gusta'

Índice general

1. Introducción	1
2. Estado del arte	7
3. Preliminares	11
3.1 Dominio	11
3.2 Notación y terminología	13
3.3 Trasfondo metodológico	15
3.3.1 K-means	15
3.3.2 Análisis de diferencias significativas entre grupos	17
3.3.3 Modelos de aprendizaje automático supervisado	19
3.4 Datos	20
3.4.1 Encuesta	21
3.4.2 Personal	24
4. Planificación y Metodología	27
5. Contribución Metodológica	33
5.1 Encuesta	34
5.1.1 Transformación y adecuación del dataset	34
5.1.2 K-means	36
5.1.3 Análisis de significancia de diferencias entre grupos	38
5.1.4 Modelos de aprendizaje automático supervisado	40
5.1.5 Ensamblaje	41
5.2 Datos personales	43
5.2.1 Etiquetaje	45
5.2.2 Relookyou	45
5.2.3 Resto de datasets 'Personal'	55
6. Resultados	61
6.1 Encuesta	62
6.1.1 K-means	62
6.1.2 Análisis de significancia de diferencias entre grupos	64

6.1.3 Modelos de aprendizaje automático supervisado	75
6.1.4 Ensamblaje	84
6.2 Datos personales	91
6.2.1 Relookyou	91
6.2.2 Dolphin	95
6.2.3 Juju	97
6.2.4 Pato	100
6.2.5 Ajara	103
6.2.6 Charlie	106
6.3 Discusión de resultados	109
7. Conclusiones	111
7.1 Limitaciones	112
7.2 Trabajo futuro	113
7.3 Contribuciones	114
Apéndice A. Hoja de consentimiento	115
Apéndice B. Árboles de decisión – ‘Encuesta’	121
Apéndice C. Rendimiento ensamblajes - ‘Personal’	125
Apéndice D. Variables estudiadas - ‘Personal’	127
D.1 Relookyou	127
D.2 Dolphin	130
D.3 Juju	131
D.4 Pato	133
D.5 Ajara	134
D.6 Charlie	136
Resultados de “Mack” y “Edogawa”	139
Bibliografía	141

Índice de figuras

4.1. Esquema de CRISP-DM (Cross-Industry Standard Process for Data Mining)	27
4.2. Diagrama de Gantt retrospectivo	31
5.1. Representación gráfica del flujo de trabajo para el estudio de 'Encuesta'	58
5.2. Representación gráfica del flujo de trabajo para el estudio 'Personal' . . .	59
6.1. Test de comparación múltiple para cada par de grupos de las variables significativas - " <i>pProfile</i> " - 'Encuesta'.	65
6.2. Representación gráfica del promedio, según grupo, de las variables significativas pertenecientes al género ' <i>HomeCompany</i> ' - " <i>pProfile</i> " - 'Encuesta'.	67
6.3. Representación gráfica del promedio, según grupo, de las variables significativas pertenecientes al género ' <i>actualStatus</i> ' - " <i>pProfile</i> " - 'Encuesta'.	68
6.4. Representación gráfica del promedio, según grupo, de las variables significativas pertenecientes al género ' <i>aboutYou</i> ' - " <i>pProfile</i> " - 'Encuesta'.	68
6.5. Test de comparación múltiple para cada par de grupos de las variables significativas - " <i>topics</i> " - 'Encuesta'	69
6.6. Representación gráfica del promedio, según grupo, de las variables significativas - " <i>topics</i> " - 'Encuesta'.	70
6.7. Clasificación de grupos según Decission Tree - " <i>pProfile</i> " - 'Encuesta' . . .	78
B.1. Apéndice - Clasificación de grupos según Decission Tree - " <i>topics</i> " - 'Encuesta'	121
B.2. Apéndice - Clasificación de grupos según Decission Tree - " <i>subtopics</i> " - 'Encuesta'	122
B.3. Apéndice - Clasificación de grupos según Decission Tree - " <i>encuesta</i> " - 'Encuesta'	123
C.1. Apéndice - Representación gráfica del rendimiento de los ensamblajes y datasets estudiados - 'Personal'	125

Índice de cuadros

3.1. Tipos de satisfacción - 'R Veenhoven'	12
3.2. Cuadro resumen de la encuesta, con las preguntas de interés e información sobre sus respuestas - 'Encuesta'	23
3.3. Significado de los prefijos en las variables de los datasets 'Personal'	25
6.1. Cuadro resumen de los 3 mejores valores de 'K' según el método en cuestión y dataset estudiado - 'Encuesta'	63
6.2. Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - " <i>pProfile</i> " - solamente entradas significativas - 'Encuesta'	65
6.3. Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - " <i>pProfile</i> " - 'Encuesta'	66
6.4. Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - " <i>topics</i> " - 'Encuesta'	69
6.5. Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - " <i>topics</i> " - 'Encuesta'	70
6.6. Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - " <i>subtopics</i> " - solamente entradas significativas - 'Encuesta'	71
6.7. Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - " <i>subtopics</i> " - 'Encuesta'	72
6.8. Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - " <i>encuesta</i> " - solamente entradas significativas - 'Encuesta'	73
6.9. Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - " <i>encuesta</i> "	74
6.10. Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - " <i>pProfile</i> " - 'Encuesta'	76
6.11. Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - " <i>topics</i> " - 'Encuesta'	79
6.12. Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - " <i>subtopics</i> " - 'Encuesta'	81
6.13. Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - " <i>encuesta</i> " - 'Encuesta'	83
6.14. Ensamblaje final de la importancia de características - " <i>pProfile</i> " - 'Encuesta'	86

6.15. Ensamblaje final de la importancia de características - "topics" - 'Encuesta'	87
6.16. Ensamblaje final de la importancia de características - "subtopics" - 'Encuesta'	89
6.17. Ensamblaje final de la importancia de características - "encuesta" - 'Encuesta'	90
6.18. Importancia de características según modelos y ensamble y coeficientes lineales - "Relookyou" - 'Personal'	92
6.19. Importancia agrupada según tópicos - "Relookyou" - 'Personal'	93
6.20. Importancia de características según modelos y ensamble y coeficientes lineales - "Dolphin" - 'Personal'	95
6.21. Importancia agrupada según tópicos - "Dolphin" - 'Personal'	96
6.22. Importancia de características según modelos y ensamble y coeficientes lineales - "Juju" - 'Personal'	98
6.23. Importancia agrupada según tópicos - "Juju" - 'Personal'	99
6.24. Importancia de características según modelos y ensamble y coeficientes lineales - "Pato" - 'Personal'	101
6.25. Importancia agrupada según tópicos - "Pato" - 'Personal'	101
6.26. Importancia de características según modelos y ensamble y coeficientes lineales - "Ajara" - 'Personal'	104
6.27. Importancia agrupada según tópicos - "Ajara" - 'Personal'	104
6.28. Importancia de características según modelos y ensamble y coeficientes lineales - "Charlie" - 'Personal'	107
6.29. Importancia agrupada según tópicos - "Charlie" - 'Personal'	107
 D.1. Apéndice - Resumen de variables de "Relookyou" - 'Personal'	130
D.2. Apéndice - Resumen de variables de "Dolphin" - 'Personal'	131
D.3. Apéndice - Resumen de variables de "Juju" - 'Personal'	133
D.4. Apéndice - Resumen de variables de "Pato" - 'Personal'	134
D.5. Apéndice - Resumen de variables de "Ajara" - 'Personal'	136
D.6. Apéndice - Resumen de variables de "Charlie" - 'Personal'	137
 E.1. Apéndice - Importancia de características según modelos y ensamble y coeficientes lineales - "Mack" - 'Personal'	139
E.2. Apéndice - Importancia agrupada según tópicos - "Mack" - 'Personal'	140
E.3. Apéndice - Importancia de características según modelos y ensamble y coeficientes lineales - "Edogawa" - 'Personal'	140
E.4. Apéndice - Importancia agrupada según tópicos - "Edogawa" - 'Personal'	140

CAPÍTULO 1

Introducción

En este capítulo explicaremos las motivaciones y proporcionaremos un resumen más extenso de nuestro trabajo de investigación, explicando la literatura que ha servido como base para su desarrollo, los objetivos, la metodología empleada, los resultados y su discusión, así como sus principales contribuciones.

¿Comprendemos nuestra felicidad? ¿Sabemos qué la propicia? Quizás nos dejamos llevar demasiado por la facilidad y comodidad de los contenidos audiovisuales 'para ayudarte a desconectar', relegando opciones más enriquecedoras que promuevan el crecimiento personal, como la lectura de un libro, la práctica de deporte, el conocer gente nueva o sencillamente la reflexión. Esta 'zona de confort' en la que se halla nuestra sociedad contemporánea podría ser una espiral degenerativa y corrupta: trabajar, abstraerse mirando una serie y dormir. Las comodidades tecnológicas, tal vez el 'soma' del siglo XXI, pueden proporcionar recompensas inmediatas y anestesiar a quienes las consumen, pero ¿verdaderamente conducen a la felicidad? Reflexiones como las plasmadas en 'Un mundo feliz' de Aldous Huxley nos instan a considerar si la búsqueda de la felicidad debe ir más allá de la satisfacción superficial y alcanzar una profundidad más significativa [Huxley 1932].

La búsqueda de la felicidad y sus determinantes ha sido un tema recurrente a lo largo de la historia de la humanidad. En este mundo, en constante cambio, en el cual las dinámicas sociales y las tecnologías evolucionan de manera vertiginosa, se vuelve aún más plausible que los factores que moldean la felicidad también experimenten transformaciones. La **comprensión** de estos matices adquiere una importancia crucial, no solo para la satisfacción individual, sino también para abordar desafíos colectivos relacionados con la salud mental y el bienestar.

En este contexto de exploración y descubrimiento, el presente trabajo se dedica a examinar los **determinantes de la felicidad**, adoptando un enfoque multidimensional y analítico. Tratando de desentrañar la complejidad de cómo una interconexión de fac-

tores afecta la experiencia de la felicidad, y de contribuir al enriquecimiento de nuestro entendimiento sobre la búsqueda de la auténtica felicidad en el mundo actual.

Para lograr este cometido, empleamos dos conjuntos de datasets ('Encuesta' y 'Personal') y aplicamos una combinación de técnicas de análisis de datos y aprendizaje automático adaptadas a sus particularidades. '**Encuesta**', se trata de un formulario que busca describir el perfil personal del encuestado y definir el conjunto de actividades y hábitos que más influyen en su felicidad y bienestar. Estos hábitos y actividades (que denominamos 'subtópicos') se agrupan en 8 tópicos diferentes: hobbies, aspectos ajenos a la persona, deporte, alimentación, interacciones sociales, tiempo productivo, patrones de sueño y hábitos perjudiciales o vicios. '**Personal**', en cambio, captura las actividades y hábitos diarios de diferentes participantes, así como la variable objetivo, la evaluación subjetiva de felicidad percibida cada día. Del mismo modo que en el conjunto 'Encuesta', estas actividades se pueden agrupar en los mismos 8 tópicos.

La metodología comienza con el estudio de '**Encuesta**'. Una vez adecuado y procesado el dataset, éste se dividió en 3 fracciones: en función de su perfil personal ("*pProfile*"), de los tópicos que más influyen en su felicidad ("*topics*") y sus subtópicos vinculados ("*subtopics*") y se clusterizaron cada una de sus porciones y el conjunto completo ("*encuesta*"). Nuestra intención radica en segmentar los respectivos datasets mediante la técnica 'K-means', identificando las diferencias significativas entre grupos y construyendo y ensamblando diferentes modelos de aprendizaje automático supervisado que predigan las etiquetas asignadas. Se extraerá la importancia de características en la predicción de grupos de cada dataset, permitiéndonos identificar cuáles de las diferencias entre grupos son las más importantes.

Posteriormente, en el estudio de '**Personal**' trabajaremos con múltiples conjuntos de datos. El propósito es discernir las variables que mayor influencia tienen en la felicidad de cada participante. Se construyeron y ensamblaron diferentes modelos de aprendizaje automático supervisado predictores del nivel de felicidad diario, en función de las distintas actividades registradas, obteniendo la importancia de características en su predicción. Adicionalmente, cada participante respondió también la encuesta, permitiendo la determinación de los grupos a los que pertenecen, según sus diversas segmentaciones en 'Encuesta' ("*pProfile*", "*topics*", "*subtopics*" y "*encuesta*"), y de los supuestos factores que más influyen en su felicidad.

Entrelazando las dos aproximaciones, este procedimiento trata de anticipar las variables más influyentes de cada participante ('Personal') de acuerdo a sus segmentaciones ('Encuesta'), procurando una guía que ayude a focalizar correctamente la atención entre las variables de cada participante. Adicionalmente, podremos confirmar o refutar las generalidades previamente identificadas según el análisis de 'Encuesta'. Concretamente, para cada participante, se destacaron las actividades que, según su segmentación, se supone, deberían tener influencias extremas (elevadas o pequeñas en contraste al resto de grupos), a partir del análisis de diferencias entre grupos de 'Encuesta'. Éstas se contrastaron con la importancia de características obtenida en el ensamblaje de modelos predictivos del nivel de felicidad 'Personal'.

En resumen, este trabajo aborda la determinación de factores influyentes en la felicidad personal a través de dos estudios distintos, 'Encuesta' y 'Personal'. Utiliza técnicas de análisis estadístico de datos y aprendizaje automático para estudiar diferencias entre grupos, segmentar y predecir, con el fin de destacar la relevancia de características en la felicidad. La intersección de ambos enfoques busca comprender hasta qué punto una segmentación de la población puede proporcionar conocimiento sobre las variables que afectan a la felicidad en diferentes grupos.

En el marco de esta investigación, hemos explorado la **literatura** existente en busca de orientación y referencias que respalden nuestro enfoque. Aunque el campo de la minería de datos aplicada a la felicidad aún se encuentra en desarrollo, hemos identificado tres estudios que se alinean con nuestros objetivos.

Uno de estos estudios de referencia, titulado '**Hábitos de vida en una población escolar de Mataró (Barcelona) asociados al número diario de horas de televisión y al consumo de azúcares**' [Ruano Ruano 1997], se centró en aplicar técnicas de clusterización a datos obtenidos a través de encuestas a escolares. Su objetivo era identificar perfiles de estudiantes con hábitos perjudiciales. Aunque su enfoque se dirige a hábitos específicos relacionados con la salud en lugar de la felicidad, encontramos valiosa su metodología de análisis de clústeres y su enfoque estadístico para evaluar diferencias entre grupos. Estos aspectos han influido en la construcción de nuestra propia aproximación de clusterización en 'Encuesta'.

Otro estudio de relevancia, denominado '**An Empirical Study of Learning-Based Happiness Prediction Approaches**' [Kong 2021], se centra en la predicción de la felicidad utilizando diversas técnicas de aprendizaje automático, trabajando con datos de

encuestas de felicidad recopilados en la Encuesta Social General de China. A pesar de que ellos no utilizaron datos longitudinales, consideramos que su enfoque en la construcción y ensamblaje de modelos es muy relevante para nuestro caso de estudio en el apartado 'Personal'.

Por último, el artículo titulado '**Más felicidad para un mayor número de personas**' [Veenhoven 2013], se enfoca en la posibilidad de aumentar los niveles de felicidad en México, inspirándose en la teoría utilitarista de Jeremy Bentham. Aunque su enfoque es más descriptivo en comparación con nuestra investigación, nos ha proporcionado una comprensión más profunda de conceptos fundamentales relacionados con la felicidad y su medición.

Los **resultados** de nuestro caso de estudio muestran que, en general, las características previstas según la segmentación en la encuesta no se corresponden demasiado con los resultados obtenidos en el estudio 'Personal'. De modo que existen discrepancias entre las expectativas, entendidas como la influencia de las variables en la felicidad basadas en la segmentación, y las influencias reales de los participantes según su estudio personal.

Los **aspectos más influyentes** en la felicidad varían en función del participante, pero los tópicos y actividades relacionadas con el cansancio, los factores externos y el tiempo libre tienden a ser los más influyentes en la felicidad de los participantes.

Al tópico del **deporte** se le confirió una influencia, quizás, menor a la esperada. Esto puede deberse a la carencia de dispositivos que logren capturar las variables adecuadas entre los distintos participantes. En cambio, se le confirió mucha importancia al aspecto del **clima**, perteneciente al tópico de circunstancias externas, pudiendo ser por el motivo contrario: la inclusión de demasiadas variables capturando su información, respecto al resto de aspectos.

La relación entre las preferencias, actividades y la felicidad es compleja y **multifacética**. Las características que influyen son una combinación de factores relacionados con su estilo de vida, aspectos externos, interacciones sociales, tiempo libre y una lista interminable de otros elementos.

Aunque los resultados no logren identificar con precisión las características influyentes mediante la segmentación de la población, sientan las bases para futuros intentos.

El éxito en este planteamiento podría aportar una guía en la formulación de estrategias más efectivas para promover la felicidad y el bienestar en diferentes segmentos de la población. De modo que consideramos el diseño de la **metodología** empleada para la interpretación de los resultados de las dos aproximaciones de manera conjunta, ‘Encuesta’ y ‘Personal’, como la principal contribución del trabajo.

Este estudio contribuye al **conocimiento** del bienestar humano, permitiendo una visión más completa de cómo diversos factores interactúan para influir en la felicidad individual. Al profundizar, se abren oportunidades para intervenciones personalizadas, lo que podría mejorar la calidad de vida de las personas. De hecho, en finalizar el estudio, cada uno de los participantes en el conjunto de datos ‘Personal’ podrá disfrutar de los beneficios de conocer las actividades que más influyen en su felicidad.

El planteamiento llevado a cabo aborda la **brecha** entre percepciones ('Encuesta') y comportamientos ('Personal'), pudiendo devenir un fenómeno relevante en investigación psicológica y sociológica. Este trabajo ofrece una aproximación valiosa para entender cómo estas discrepancias afectan la felicidad.

Estado del arte

La felicidad es un tema que ha adquirido una creciente relevancia en la sociedad actual. La investigación sobre la felicidad personal, desde la perspectiva de aprendizaje automático, está experimentando un aumento en su interés. Aunque este enfoque se encuentra en desarrollo, la minería de datos en el ámbito de la felicidad aún no cuenta con una extensa literatura. En este capítulo, presentamos una revisión del estado del arte, donde destacamos tres estudios en particular, que han servido de referencia y orientación para nuestra investigación. Explicaremos por qué estos estudios son relevantes y cómo se distinguen de nuestro trabajo.

Uno de los estudios de referencia es '**Hábitos de vida en una población escolar de Mataró (Barcelona) asociados al número de veces diarias que ve televisión y al consumo de azúcares**' [Ruano Ruano 1997]. Este estudio aplicó técnicas de clusterización a datos recopilados a través de encuestas a casi 3000 escolares con el objetivo de identificar perfiles de estudiantes con hábitos nocivos. Sus resultados sugieren que la presencia de algunos hábitos perjudiciales, como el elevado número de horas de televisión y consumo de azúcares, facilita la presencia de otros hábitos no saludables en el mismo sujeto.

Sin embargo, es fundamental destacar que nuestro enfoque se centra en la felicidad personal, a diferencia de este estudio que se enfoca en hábitos más bien relacionados con la **salud**. Además, en términos de la diversidad de la población estudiada y de la amplitud de los hábitos analizados, nuestro enfoque de clusterización es más amplio y abarca una mayor variedad de factores.

A pesar de estas diferencias, encontramos valiosa la **metodología** de análisis de clústeres empleada en este estudio, que podría ser aplicada a otros grupos de encuestados para obtener perfiles similares. Nos enfocamos en su método de clusterización, así como en su enfoque estadístico para analizar las diferencias, que incluye el uso de una prueba de Chi-cuadrado con corrección de Yates. Estas técnicas nos inspiraron en la construcción de nuestro propio enfoque de clusterización.

Otro estudio relevante, titulado ‘**An Empirical Study of Learning Based Happiness Prediction Approaches**’ [Kong 2021]. En este estudio, los investigadores emplearon datos de encuestas de felicidad recopilados en la Encuesta Social General de China y aplicaron diversos enfoques de aprendizaje automático para predecir los niveles de felicidad de los encuestados.

Los resultados de este estudio destacaron la **efectividad** de los enfoques de aprendizaje automático para predecir la felicidad de los participantes, concluyendo que ‘Gradient Boosting’ resulta ser el más eficaz. Además, los resultados sugieren que factores como la equidad social, la salud mental, la situación económica y el estatus social desempeñan un papel significativo en la determinación de la felicidad.

A pesar de que no centramos nuestra atención en las conclusiones específicas de este estudio, encontramos inspiración en su **metodología**, particularmente en lo que respecta a la preparación de datos y la construcción de modelos. Nos llamó la atención la forma en que procesaron los datos antes de incluirlos en los modelos, empleando técnicas de estandarización y la conversión de variables binarias mediante Get-dummies. También nos interesó su enfoque en realizar un análisis de correlación como parte del proceso de selección de variables, así como la construcción y el ensamblaje de múltiples modelos para obtener predicciones más precisas. Estos aspectos del estudio, a pesar de que no utilizaron datos longitudinales, nos proporcionaron ideas valiosas para nuestro propio enfoque metodológico en el estudio ‘Personal’.

Por último, el artículo titulado ‘**Más felicidad para un mayor número de personas**’ [Veenhoven 2013] se inspira en la teoría utilitarista de Jeremy Bentham, que postula que el objetivo de la sociedad debe ser maximizar la felicidad de la mayoría de sus miembros. Este trabajo se enfoca en la posibilidad de aumentar los niveles de felicidad en México.

En este artículo, se realiza un análisis exhaustivo de la situación de México en términos de felicidad y bienestar. Además, se proponen una serie de medidas con el potencial de **mejorar la felicidad de la población**, como el fortalecimiento de la educación, la atención médica de calidad y la mejora de la calidad institucional del gobierno. Se destaca la importancia de medir la felicidad y el bienestar como parte integral de la toma de decisiones políticas y sociales. Se argumenta que esta medición puede ayudar a los gobiernos y las organizaciones a identificar áreas críticas que requieren mejoras y a evaluar la efectividad de las políticas y programas implementados.

En general, este artículo ofrece una visión interesante de cómo la teoría utilitarista puede aplicarse en la práctica para mejorar la felicidad y el bienestar de la población. Aunque su enfoque es más descriptivo en comparación con nuestro trabajo, nos ha proporcionado una comprensión más profunda de **conceptos fundamentales** relacionados con la felicidad y su medición.

En **resumen**, aunque la literatura sobre minería de datos y felicidad es aún incipiente, hemos identificado estudios relevantes que nos han proporcionado conocimientos valiosos y enfoques metodológicos que hemos adaptado para nuestro proyecto. En los capítulos siguientes, detallaremos cómo hemos aplicado estos conceptos y métodos en nuestro enfoque único de exploración de la felicidad personal. A pesar de las diferencias en los objetivos y alcances de estos estudios, todos comparten un interés común en comprender y mejorar la calidad de vida y el bienestar, contribuyendo así al crecimiento del campo de la ciencia de datos aplicada a la felicidad personal y social.

Preliminares

3.1 Dominio

La inclusión de esta sección es esencial para establecer cierta comprensión del concepto de 'felicidad' en el contexto de nuestro proyecto de investigación. Aquí definiremos lo que entendemos por felicidad, basándonos en diversas perspectivas teóricas y dimensiones clave. Esta definición proporcionará una base conceptual para nuestro estudio y permitirá una exploración más acertada de cómo los factores y actividades influyen en esta experiencia humana fundamental.

Existen **diversas perspectivas** sobre la felicidad. Algunos autores la ven como un logro personal, un triunfo primero sobre uno mismo y luego sobre el mundo que nos rodea [Margot 2007]. Otros la describen como una emoción de alegría y una sensación de tranquilidad y armonía que resulta de contar con ciertas condiciones básicas relacionadas con el bienestar personal, las relaciones sociales, la salud, el trabajo y el amor [Hernández Aburto 2017]).

En nuestro enfoque, nos basamos en la reflexión de [Veenhoven 1984], quien considera que la felicidad es un compuesto de dos dimensiones esenciales. La primera se refiere al grado en que las experiencias placenteras y positivas superan a las experiencias desagradables (felicidad como **balance emocional**). La segunda dimensión, que procederemos a explicar a continuación, se relaciona con el grado en que una persona se siente satisfecha con su autorrealización en los diferentes aspectos de la vida (felicidad como **satisfacción con la vida**).

Cuando hablamos de satisfacción subjetiva con la vida, como se propone en [Veenhoven 2013], sus conceptos asociados pueden clasificarse según las siguientes **dicotomías**: parte de la vida versus la vida como un todo, y deleite pasajero versus satisfacción duradera, que se presentan en el cuadro 3.1:

Satisfacción	Pasajera	Duradera
Aspectos de la vida	Placer	Satisfacción de dominio
La vida en general	Experiencia máxima	Satisfacción de vida

Cuadro 3.1: Tipos de satisfacción - 'R Veenhoven'

El cuadrante superior-izquierdo del cuadro 3.1, representa los '**placeres**' pasajeros de la vida. Como disfrutar de una taza de té, la satisfacción de completar una tarea o el placer de apreciar una obra de arte.

En el cuadrante superior derecho del cuadro 3.1 se encuentra la '**satisfacción de dominio**', que se refiere a una apreciación duradera de aspectos tales como la satisfacción laboral o con el matrimonio. Aunque ésta dependa de un flujo continuo de placeres, tiene cierta continuidad propia.

El cuadrante inferior derecho del cuadro 3.1 representa las '**experiencias máximas**', efímeras pero intensas, que combinan momentos de felicidad pasajera con una valoración positiva de la vida en general. Estas son a menudo vivencias culminantes en la existencia, la clase de felicidad de la que escriben los poetas.

Finalmente, en el cuadrante inferior derecho del cuadro 3.1 se encuentra la '**satisfacción con la vida**', representando la combinación de satisfacción perdurable con la vida en general. Este término se relaciona con el '**principio de mayor felicidad**' de Jeremy Bentham, que busca maximizar la satisfacción general en la vida.

Esta clasificación nos proporciona un **marco conceptual** para abordar la noción de felicidad en nuestro estudio. Siguiendo la perspectiva de Veenhoven, consideramos dos dimensiones esenciales de la felicidad: la felicidad como balance emocional y la felicidad como satisfacción con la vida.

En el estudio '**Personal**', exploramos la dimensión de la felicidad como **balance emocional**. Cada participante deberá hacer el balance al registrar diariamente el índice de felicidad, lo que nos permite explorar cómo varía el equilibrio emocional en relación con diversas actividades y circunstancias.

En cuanto a la dimensión de la **satisfacción con la vida**, en nuestro caso de estudio, consideramos todos los tipos de satisfacción comentados. En el estudio '**Encuesta**',

aparecen los 'placeres' cotidianos, como disfrutar de una taza de café, las 'satisfacciones de dominio', como la satisfacción con el trabajo, las 'experiencias máximas', como tener una cita romántica exitosa, y la 'satisfacción con la vida' en su conjunto, al evaluar la influencia de varios tópicos y actividades en la felicidad.

En el caso del estudio de '**Personal**', los dos tipos de satisfacción pasajera se evalúan cada día por los distintos participantes y, a pesar de que no se ha hecho, dada la naturaleza longitudinal de los datos también podríamos haber considerado analizar la felicidad del participante desde el punto de vista de satisfacciones duraderas, estudiando cómo fluctúa el índice de felicidad a lo largo del tiempo según las actividades registradas.

En conclusión, en este capítulo hemos definido cómo entendemos nosotros el concepto de felicidad en el contexto de este proyecto, establecido una base teórica para nuestro estudio.

3.2 Notación y terminología

Con el objetivo de facilitar la lectura y comprensión de este proyecto, se han considerado ciertas convenciones en cuanto a su formato y terminología. Esta sección establece la notación y terminología que se utilizarán a lo largo de este informe, buscando mantener la claridad y coherencia en la presentación de conceptos. La correcta interpretación de los términos empleados resulta esencial para una comprensión más precisa de la metodología, los resultados y las conclusiones expuestas.

Recordando que este estudio se divide en dos **secciones** según los conjuntos de datos analizados: 'Encuesta' y 'Personal'. A continuación, se describen ciertas particularidades de ambas secciones. No se abordará una explicación detallada en este momento, ya que ésta se presentará en las secciones correspondientes. La intención reside en que el lector pueda identificar las distintas partes de estudio de este proyecto.

El primer conjunto ('**Encuesta**') gira en torno al análisis de los resultados de una encuesta. Como se detalla en la subsección 5.1.1(Transformación y Adecuación del Dataset), este conjunto se subdivide en 3 segmentos ("*pProfile*", "*topics*" y "*subtopics*"), permitiendo su estudio desde 4 perspectivas diferentes: desde cada segmento individualmente y en su totalidad ("*encuesta*"). En las secciones pertinentes al estudio de la

'Encuesta', se hará referencia a 'géneros' en ciertas ocasiones. Como se detallará en la sección 5.1.1, se aplicó la técnica 'get dummies' al analizar este conjunto de datos. Los 'géneros' agrupan las nuevas columnas generadas mediante esta técnica, en función de sus columnas 'progenitoras'.

El segundo ('Personal'), avanzándonos a la sección 3.4.2, se basa en el estudio de un conjunto de datos recabados, diariamente, por ocho participantes. Los distintos datasets toman los siguientes nombres: "*Relookyout*", "*Edogawa*", "*Mack*", "*Dolphin*", "*Juju*", "*Pato*", "*Ajara*" y "*Charlie*". En este caso existe un aspecto importante en cuanto a la terminología que se explicará más adelante mediante el cuadro 3.3. Esta explica cómo se codificaron los prefijos de los nombres de las distintas variables para lograr una mayor agilidad en el manejo de datos, así como un mayor entendimiento en la interpretación de los resultados.

Dada la naturaleza del proyecto, en el cual se analizan múltiples **conjuntos de datos**, se hará mención constante al conjunto de datos relevante. Para asegurar que el lector identifique con claridad el conjunto en cuestión, estos se presentarán siempre entre comillas dobles y en cursiva, por ejemplo "*Relookyout*". Cuando se haga referencia a una variable específica dentro de alguno de los conjuntos de datos, se utilizará un formato similar: comillas simples y cursiva, por ejemplo '*FRE_sum*'.

Es importante mencionar que términos como **TFM**, estudio, proyecto y trabajo hacen referencia al mismo concepto: el Trabajo de Fin de Máster. Dependiendo del contexto, estos términos pueden referirse a partes específicas del trabajo en lugar de su totalidad.

A lo largo del informe, se han incluido fragmentos del código en **Python** utilizando [Python 023], cuando se ha considerado necesario. Estos extractos se reconocerán por estar enmarcados en recuadros, con un tamaño de fuente más pequeño y un formato distintivo.

Por último, mencionar que, con el fin de facilitar la lectura, se ha procurado destacar en **negrita** palabras o conceptos clave entre los distintos párrafos que comprenden el trabajo. De tal manera que el lector pueda identificar rápidamente el contexto del párrafo antes incluso de comenzar a leerlo.

3.3 Trasfondo metodológico

En esta sección, proporcionaremos un marco de referencia que servirá como base fundamental para comprender la metodología y las técnicas utilizadas en nuestro estudio. A medida que avanzamos en la exploración de la minería de datos aplicada a la felicidad, es crucial establecer un terreno común de conocimientos y conceptos clave, de modo que aquí abordaremos aspectos esenciales relacionados con las técnicas de análisis estadístico y los modelos de aprendizaje automático empleados.

3.3.1 K-means

El algoritmo K-means es un modelo de aprendizaje automático, no supervisado, que tiene como objetivo agrupar el conjunto de datos en función de su similitud, donde cada punto de datos se asigna al clúster cuyo centro (representado por la media aritmética y denominado centroide) está más cercano a él. Se trata de un proceso iterativo, en el que se asignan puntos al clúster más cercano en función de su distancia euclídea con respecto al resto de centroides, luego se inicializan de nuevo los centroides en diferentes posiciones, reiterando el proceso hasta que los centroides convergen o se alcanza un número predefinido de iteraciones. [Martínez Heras 023, Wagstaff 2001]

Entre los diferentes hiperparámetros del modelo, el más importante es el **número de clústeres** seleccionados 'K'. A continuación, describiremos los métodos que se usaron para determinar el valor de 'K' idóneo en nuestro caso de estudio:

- El método **Elbow**, en el algoritmo K-means, nos permite identificar el punto de inflexión en la curva de la suma de las distancias al cuadrado (inerzia) entre cada punto de datos y el centroide al que pertenece. Este punto de inflexión indica un equilibrio entre la capacidad de explicar la varianza de los datos y el número de clústeres [Tomar 023]. Si bien no es siempre determinante, nos proporcionó una estimación inicial del número óptimo de clústeres.

$$Inertia = \sum_{i=0}^N (x_i - c)^2 \quad (3.1)$$

Siendo la suma entre 0 y el número de clústeres (N) de las distancias al cuadrado de cada objeto del clúster (x) a su centroide (c) [Moya 023].

- Se ha utilizado método **Silhouette** para medir la calidad de un agrupamiento, basándonos en la comparación de la distancia promedio del punto (fila en el registro de datos) con su propio clúster y con el resto de clústeres [Bhardwaj 023, León Guzmán 023, Tomar 023]. A esta aproximación se le atribuyó una mayor importancia debido a su capacidad para medir la cohesión y separación de los clústeres.

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \quad (3.2)$$

Donde $a(x)$ mide la cohesión de x a todos los demás puntos en el mismo clúster; $b(x)$ mide la distancia promedio de x a todos los demás puntos en el clúster más cercano; el valor ' $s(x)$ ' puede variar entre -1 (muy mal agrupamiento) y 1 (muy bueno) [León Guzmán 023]. El coeficiente Silhouette para todo el agrupamiento es:

$$SC = \frac{1}{N} \sum_{i=1}^N s(x) \quad (3.3)$$

- Aunque en general tuvo un papel más secundario, el índice **Calinski-Harabasz** proporciona una medida de dispersión inter e intra-clústeres, lo que puede ser decisivo en ciertos casos [León Guzmán 023]. En el dataset "topics", esta medida fue especialmente relevante en la elección del valor de 'k'.

$$CH = \frac{SSB/(k-1)}{SSW/(n-k)} \quad (3.4)$$

Donde SSW es una medida interna para evaluar la cohesión de los clústeres generados: (siendo 'k' el número de clústeres, 'x' un punto del clúster ' C_i ' y ' m_i ' el centroide del clúster ' C_i ') [León Guzmán 023].

$$SSW = \sum_{i=1}^k \sum_{x \in c_i} dist^2(m_i, x) \quad (3.5)$$

Y SSB es una medida de separación utilizada para evaluar la distancia inter-clúster (separación): (siendo 'k' el número de clústeres, ' n_j ' el número de elementos en el clúster 'j', y ' x_{mean} ' el promedio del dataset) [León Guzmán 023].

$$SSB = \sum_{j=1}^k n_j dist^2(c_j - x_{mean}) \quad (3.6)$$

- El índice **Davies-Bouldin** mide la similitud promedio entre cada clúster y su clúster más similar [León Guzmán 023]. Éste, aunque no aportó información relevante en la elección del valor de 'k', se utilizó para evaluarlo junto con los demás métodos, como una medida adicional de la calidad de los clústeres.

$$DB = \frac{1}{k} \sum_{i=j, i \neq j}^k \max \left(\frac{o_i + o_j}{d(c_i, c_j)} \right) \quad (3.7)$$

Donde 'k' es el número de clústeres, o_i es la distancia promedio entre cada punto en el clúster i y el centroide del clúster, o_j es la distancia promedio entre cada punto del clúster j y el centroide del clúster, y $d(c_i, c_j)$ es la distancia entre los centroides de los 2 clústeres [León Guzmán 023]. Cuanto más pequeños los valores del índice, más compactos los clústeres, con los centros bien separados los unos de los otros.

3.3.2 Análisis de diferencias significativas entre grupos

En nuestra metodología, en el marco del estudio de 'Encuesta', para analizar las diferencias de las actividades y hábitos entre los grupos segmentados, utilizamos el test de **Chi cuadrado**. Se trata de una técnica estadística que evalúa si existe una conexión significativa entre dos variables categóricas dentro de un conjunto de datos [Waine 023].

Para ello, partiendo del cuadro de frecuencias observadas, que representa la distribución de las observaciones conjuntas de las dos variables categóricas, se construye el cuadro de **frecuencias esperadas**, según:

$$FE = (TF) * (TC) / (TT) \quad (3.8)$$

Donde FE = Frecuencia esperada, TF = Total en la fila, TC = Total en la columna, y TT = Total en el cuadro.

Entonces, suponiendo que las variables son independientes entre sí, comparando las frecuencias observadas y las esperadas, determina si las discrepancias entre las observaciones y las expectativas son lo bastante notables como para indicar una relevancia estadística. Si las diferencias son significativas, esto podría indicar una asociación o **relación entre las variables** categóricas que no se debe al azar. El test de Chi cuadrado

cuantifica esta diferencia y proporciona una medida de la significancia estadística [Wainne 023].

Cuando se utiliza el test de Chi cuadrado y se trabaja con tablas de contingencia de 2x2, con un tamaño de la muestra relativamente pequeño, se corre el riesgo de obtener **significancias sobreestimadas** y erróneas. Es decir, en muestras pequeñas, las diferencias observadas entre las categorías pueden no ser necesariamente indicativas de una verdadera asociación o relación entre las variables categóricas, especialmente cuando encontramos frecuencias bajas en las tablas de contingencia [how2stats 023].

La **corrección de Yates** es una técnica que se utiliza en este contexto. Ajusta los valores del test para mitigar la tendencia a producir resultados significativos cuando las diferencias son mínimas, reduciendo la magnitud de la estadística Chi cuadrado. La corrección de Yates es especialmente útil cuando se trabaja con muestras pequeñas o cuando se desean resultados más conservadores en el análisis de asociación entre variables categóricas [how2stats 023].

Como nos encontramos en un contexto de múltiples pruebas de hipótesis, es necesario aplicar algún tipo de corrección para reducir el riesgo de obtener falsos positivos. Nosotros aplicaremos primeramente la **corrección de Bonferroni**, una corrección conservadora que ajusta los valores p dividiéndolos por el número de pruebas realizadas [McDonald 023, Walsh 2004].

También usaremos la corrección de **Holm-Bonferroni**, una corrección secuencial que ordena de forma ascendente los p-valores obtenidos y los ajusta del siguiente modo:

$$a = 1 - (1 - p)^{(1/m)} \quad (3.9)$$

Donde 'a' es el p-valor corregido, 'p' el p-valor del índice evaluado y 'm' el número de pruebas restantes [Walsh 2004, Hervé 2010].

En última instancia, usaremos la corrección de **Benjamini-Hochberg**, también secuencial, que ajusta cada p-valor comparándolo con su respectivo valor crítico 'FDR', que se calcula del siguiente modo:

$$FDR = (i/n)Q \quad (3.10)$$

Donde 'FDR' es el valor máximo que puede tomar el p-valor para considerarse significativo, 'i' el rango (el índice que ocupa el p-valor a estudiar), 'n' el número total de tests y 'Q' la tasa de falso descubrimiento escogida (0.05) [McDonald 023, Walsh 2004].

3.3.3 Modelos de aprendizaje automático supervisado

En el marco de nuestra metodología de análisis de datos, hemos empleado diversos modelos de aprendizaje automático supervisado. A continuación, presentamos una descripción de Random Forest, Gradient Boosting, Extreme Gradient Boosting Classifier y Support Vector Machine. Cada uno de ellos ofrece enfoques únicos para la resolución de problemas y se seleccionan de acuerdo con las necesidades específicas de nuestro estudio.

El modelo **Random Forest** es un ensamblaje de múltiples árboles de decisión individuales. Cada árbol se entrena independientemente, con una muestra aleatoria de los datos de entrenamiento originales. En cada árbol individual, las observaciones se dividen en nodos, creando la estructura del árbol. Las predicciones de todos los árboles se combinan para obtener una predicción más robusta y precisa. Esta técnica es conocida por su capacidad para manejar datos de alta dimensionalidad y su resistencia al sobreajuste [Amat Rodrigo 2017].

Por otro lado, **Gradient Boosting** es otra técnica de ensamble que se puede utilizar con diversos algoritmos base, incluyendo árboles de decisión. En este enfoque, se ajustan secuencialmente múltiples 'weak learners' (modelos simples que predicen ligeramente mejor que el azar). Cada nuevo modelo se enfoca en aprender de los errores del anterior, mejorando iteración tras iteración. En contraste con Random Forest, las diferencias entre cada 'weak learner' no se deben al muestreo de conjuntos de datos diferentes, sino que se originan porque la importancia de las observaciones cambia en cada iteración. El Gradient Boosting es conocido por su capacidad para capturar relaciones complejas en los datos y su versatilidad y capacidad de generalización lo hacen valioso en una variedad de dominios [Amat Rodrigo 2017].

Extreme Gradient Boosting, es una variante optimizada del Gradient Boosting. Destaca por su eficiencia computacional y su capacidad para manejar grandes conjuntos de datos. Utiliza una función de 'penalización de la complejidad' para evitar el sobreajuste y mejorar la generalización del modelo [Espinosa-Zúñiga 2020].

Por último, el modelo **Support Vector Machine** es conocido por su capacidad para encontrar hiperplanos óptimos de separación en el espacio multidimensional de características. Su objetivo es maximizar la distancia entre las clases de datos, lo que lo convierte en una herramienta poderosa para la clasificación y la regresión. El SVM destaca, especialmente, cuando trabaja con conjuntos de datos de alta dimensionalidad y ha demostrado su utilidad en diversas áreas, desde la clasificación de textos hasta la detección de rostros [Berwick 023].

Estos modelos de aprendizaje automático supervisado se utilizaron en las dos aproximaciones del estudio. En ‘Encuesta’, considerados como modelos de clasificación (RF, GB, XGB y SVM), empleando la accuracy como métrica de rendimiento. Y en ‘Personal’, abordando la predicción numérica de los niveles de felicidad personal de los participantes, los consideramos como modelos de regresión (RFR, GBR, XGBR y SVR), usando el Error Cuadrático Medio (MSE) como medida de evaluación.

3.4 Datos

El objetivo central de este trabajo reside en la obtención de conclusiones sobre los factores determinantes de la felicidad personal. Para lograr este cometido, empleamos un conjunto diverso de datasets y aplicamos una combinación de técnicas de análisis de datos y aprendizaje automático adaptadas a las particularidades de cada uno de ellos.

En esta sección, presentamos una descripción detallada de los datos utilizados en el estudio, así como el proceso de obtención de los mismos. En particular, examinamos dos conjuntos de datos distintos.

El primero, que denominamos ‘**Encuesta**’, se trata de un formulario que busca describir el perfil personal del encuestado y definir el conjunto de actividades y hábitos que más influyen en su felicidad y bienestar. Este conjunto de actividades se agrupan 8 tópicos diferentes: hobbies, aspectos ajenos a la persona, deporte, alimentación, interacciones sociales, tiempo productivo, patrones de sueño y hábitos perjudiciales o vicios.

El otro, se trata de un conjunto de datos, que llamamos ‘**Personal**’, y captura las actividades y hábitos diarios de los participantes y su relación con la evaluación subjetiva de felicidad percibida cada día. Del mismo modo que en el conjunto ‘Encuesta’, estas actividades se pueden agrupar en los mismos 8 tópicos.

A lo largo de este apartado, describiremos las características y estructura de estos conjuntos de datos, así como los métodos empleados para su obtención.

3.4.1 Encuesta

La recopilación de estos datos se llevó a cabo mediante un formulario implementado en **Google Forms**, publicado el 1 de junio de 2022. En los días siguientes, se obtuvieron respuestas de 153 personas interesadas en participar en el estudio. El formulario fue diseñado con el propósito de investigar los factores influyentes en la felicidad personal y evaluar la viabilidad de desarrollar una aplicación móvil que pudiera analizar, de manera personalizada, los hábitos y actividades diarias que tienen un impacto significativo en la felicidad. De modo que, en realidad, el formulario es algo más 'complejo' de lo necesario para nuestro caso de estudio, pero aquí describiremos solamente la parte que atañe a nuestro proyecto.

El cuestionario se inicia con un conjunto de preguntas destinadas a recopilar información sobre el **perfil personal** de cada encuestado. Estas preguntas no se limitan solo a aspectos demográficos, como el género, edad, altura, situación laboral y estado civil, sino que también abarcan áreas relacionadas con actividades, hobbies y preferencias personales.

Posteriormente, se definen ocho categorías principales, los denominados **tópicos**, descritos en la introducción de esta sección 3.4, que se consideran relevantes para analizar su influencia en la felicidad personal. El objetivo es identificar y comprender cuáles de estos tópicos ejercen una mayor influencia en la felicidad de cada encuestado.

Para cada uno de estos tópicos, se plantean diversas actividades y hábitos específicos, que denominamos **subtópicos**, y podrían estar relacionados con la felicidad. A través de las respuestas proporcionadas por los encuestados, se busca identificar las actividades más influyentes dentro de cada tópico y comprender cómo afectan a la felicidad de cada individuo. Véase el cuadro 3.2 para advertir los distintos subtópicos, tópicos, y la caracterización del perfil personal.

Pregunta	Tipo de respuesta	Nº respuestas	Distintas respuestas
Sexo	Elección única	3	Hombre Mujer Otro
Edad	Elección única	6	<18 18-25 26-35 36-45 46-65 >65
¿Cuánto mides?	Elección única	6	<1.5m 1.5-1.6 1.61-1.7 1.71-1.8 1.81-1.9 >1.9
¿Cuánto pesas?	Elección única	6	<40kg 40-50 51-60 61-70 71-80 >80
Estado actual	Elección única	3	Estudio Trabajo Vivo del cuento
¿Con quién vives?	Elección única	5	Con mis padres Sólo En pareja Con amigos En familia
¡Háblame sobre ti! ¿Qué te representa?	Elección múltiple	22	Tengo o dispongo de vehículo de transporte Fumo Me gusta salir de fiesta Sofá y manta Activo en redes sociales Suelo pensar en mi felicidad, valorarla y pensar en cómo mejorarla No estoy satisfecho si no he aprovechado el día Procuro comer sano Practico deporte Muy Zen (naturaleza, Yoga, meditación...) Duermo a pierna suelta Antes playa que montaña Me estreso fácilmente Soy romántic@ Espontáneo Antes dulce que salado Películas y series Videojuegos Lectura Dependo económicamente de alguien Soy más de ciencias que de letras Mascotas
¿Cuáles de los siguientes tópicos influyen más en tu estado de ánimo, ya sea para bien o para mal? (selecciona 5)	Elección múltiple (5) [otro]	8+1	Horas productivas (estudio, trabajo, tareas...) Social (amigos, familia, citas...) Ocio/Hobbies (contenido audiovisual, lectura, juegos, cultural...) Deporte Sueño Alimentación Ajenos a la persona (clima, día de la semana...) Vicios (café, drogas, sexo...) Otra
Respecto al tópico del tiempo productivo. (Selecciona 4)	Elección múltiple (4) [otro]	6+1	El trabajo (tipo de trabajo, horas dedicadas, horario...) El estudio (tipo, horas dedicadas, horario...) El desarrollo mental: ajedrez, sudoku, juegos de memoria... La lectura didáctica El cumplimiento del programa diario (conseguir hacer lo que te has propuesto para hoy) Las tareas domésticas, recados... Otra
Respecto al tópico social. (Selecciona 3)	Elección múltiple (3) [otro]	5+1	Familia Amigos Citas Fiesta: discoteca, tomar algo, reunión con amigos... Redes sociales Otra
Respecto al tópico del ocio/hobbies. (Selecciona 4)	Elección múltiple (4) [otro]	8+1	Contenido audiovisual: TV, streaming, videojuegos... Lectura Relajamiento: meditar, tomar el sol, Yoga, no hacer nada... Culturales: cine, teatro, conciertos... Naturaleza / animales Viajes Hacer algo nuevo, salir de la zona de confort Música Otra
Respecto al tópico deportivo y de actividad. (Selecciona 2)	Elección múltiple (2) [otro]	4+1	Salir de casa, moverse un poco, pasear, desplazamientos andando... Tipo de deporte (quizás un partido de pádel te aporta más felicidad que ir al gimnasio) Intensidad y tiempo del deporte Deporte "pasivo": escaleras en vez de ascensor, apretar el abdomen durante el día... Otra
Respecto al tópico del sueño. (Selecciona 3)	Elección múltiple (3) [otro]	6+1	Calidad de sueño Cantidad de sueño Hora en la que te levantas Siesta Pantallas/dispositivos antes de dormir (TV, ordenador, Tablet...) Aspecto onírico (pesadillas, sueños plácidos, sueños lúidos...) Otra
Respecto al tópico de la alimentación. (Selecciona 2)	Elección múltiple (2) [otro]	4+1	Tipo de comidas de hoy (sana, rápida, casera, reparto, restaurante...) Que he comido hoy (pollo, ternera, legumbres, ensalada, verduras...) Dieta (sin carne, sin azúcar, refrescos...) Cantidad (ligero, copioso) Otra

Pregunta	Tipo de respuesta	Nº respuestas	Distintas respuestas
Respecto al tópico de los factores ajenos a tu persona. (Selecciona 2)	Elección múltiple (2) [otro]	4+1	Clima (soleado, lluvioso...) Día de la semana Recibir buenas o malas noticias Que gane tu equipo deportivo Otra
Respecto al tópico de los vicios. (Selecciona 2)	Elección múltiple (2) [otro]	5+1	Alcohol Tabaco Café Drogas Sexo Otra

[Reverté 022]

Cuadro 3.2: Cuadro resumen de la encuesta, con las preguntas de interés e información sobre sus respuestas - 'Encuesta'

Con el fin de tratar de encontrar el mejor compromiso entre máxima explicabilidad con el mínimo de información, se ha construido el cuadro 3.2, que explica de manera clara y concisa los determinantes de la encuesta. En ella se identifican como filas las distintas preguntas relevantes del formulario, y se destacan las siguientes columnas:

- La columna '**Pregunta**' contiene las distintas preguntas (relevantes) realizadas en la encuesta.
- La columna '**Tipo de Respuesta**' merece una atención especial. Ésta determina si las respuestas son de elección única (solamente se puede seleccionar una opción); o bien de elección múltiple (en cuyo caso se deberá seleccionar el número de respuestas indicado entre paréntesis). Algunas preguntas pueden no presentar un número determinado de respuestas a seleccionar. Si aparece la condición '[otro]', significa que los encuestados tienen la opción de invertir una de las respuestas en expresarse libremente y responder, con un texto breve, de forma distinta a las opciones indicadas.
- La columna '**Nº respuestas**' indica el número de respuestas entre las que podemos escoger. Observamos como aquellas preguntas con la condición '[otro]' se les añade un {+1}. Nos ha parecido útil incluir esta columna para que se pueda apreciar la dimensionalidad que tomará el conjunto de datos una vez adecuado el dataset.
- La columna '**Distintas respuestas**' muestra el espectro de posibles respuestas a seleccionar; observamos que algunas de sus celdas contienen, como última respuesta, 'Otra'.

Es importante destacar que la mayoría de las respuestas del formulario son de elección múltiple, lo que implica que, una vez importados los datos en formato dataframe,

ciertas columnas contendrán diferentes combinaciones de respuestas por cada fila. Esto se debe a la naturaleza **multidimensional** del estudio y permite obtener una visión más completa y detallada de las preferencias y hábitos de cada encuestado.

El proceso de adquisición de datos mediante el cuestionario permitirá analizar y comprender cómo los diferentes aspectos de la vida de cada individuo influyen en su felicidad y bienestar personal. Con estos datos, se realizará una **clusterización** para identificar grupos con características similares y asignar factores influyentes específicos a cada grupo, evitando generalizaciones simplistas y obteniendo conclusiones más precisas y personalizadas sobre los determinantes de la felicidad.

3.4.2 Personal

En este apartado, se detallan los datos utilizados en el análisis de la segunda parte del estudio, que aborda las actividades y hábitos personales de distintos participantes, y su influencia en la felicidad percibida, diariamente. Los distintos participantes se identifican con un alias, y **distinguimos**: Relookyou, Edogawa, Mack, Dolphin, Juju, Pato, Ajara y Charlie. Sus respectivos datasets tomarán esos mismos nombres.

Para procurar cierto grado de concisión en el trabajo, se ha decidido incluir solamente una **selección de análisis** entre los ocho realizados. Estos análisis han sido elegidos cuidadosamente para maximizar su contribución al estudio. Por el mismo motivo, se ha optado por explicar este apartado en términos generales, y añadir en los anexos del informe (7.3) cuadros descriptivos de las diferentes variables por cada dataset personal estudiado.

Los datos fueron recopilados a través de una aplicación móvil llamada '**Hábitos**' [Álison S 016], que permitió a los participantes registrar sus actividades diarias y hábitos en un formato estructurado y exportable en CSV. Esta aplicación permitió registrar datos binarios o numéricos, diariamente, de hasta 1000 variables definidas por el usuario; de modo que cada participante pudo tener la libertad de añadir cualquier variable que creyese influyente en su felicidad.

Mediante la aplicación, los participantes documentaron sus actividades y hábitos diarios en relación con las ocho áreas de interés predefinidas (tópicos). De este modo los respectivos análisis están en **sintonía** con el estudio de 'Encuesta'.

Es relevante destacar que, si bien todos los participantes registraron datos en relación con los mismos ocho tópicos, la naturaleza de las actividades y hábitos individuales varió entre los participantes. Esta diversidad en las variables capturadas para cada individuo implica la necesidad de abordar los datos de manera **individualizada**, considerando las particularidades de las distintas variables registradas.

Para garantizar la claridad y eficacia en el análisis, aplicamos una **codificación de prefijos** en los nombres de las variables según su temática y contexto. Cabe mencionar que estos prefijos no son mutuamente excluyentes, de modo que una variable puede llevar múltiples prefijos. Esta codificación se detalla en el cuadro 3.3.

Prefijo	Significado	Ejemplo
'XX_'	Variables binarias	'XX_FRE_shower': si el participante se ha duchado o no
'EVA_'	Variables de evaluación subjetiva	'EVA_happiness': evaluación de la felicidad
'TIR_'	Relacionadas con el aspecto del cansancio del participante	'EVA_TIR_tiredness': evaluación del cansancio
'FRE_'	Relacionadas con el tiempo libre	'FRE_VIC_audiovisual_content': tiempo de contenido audiovisual
'PRF_'	Relacionadas con el tiempo provechoso	'PRF_reading': tiempo dedicado a la lectura
'EAT_'	Relacionadas con la alimentación	'XX_EAT_fish': si el participante ha comido pescado
'ACT_'	Relacionadas con actividades sociales	'XX_ACT_family': actividad familiar
'VIC_'	Relacionadas con vicios o hábitos perjudiciales	'VIC_alcohol': número de copas tomadas
'ALI_'	Relacionadas con aspectos externos, ajenos a la persona	'ALI_tmed': temperatura media
'SPO_'	Relacionadas con el deporte	'SPO_sportTime': tiempo dedicado a practicar deporte

Cuadro 3.3: Significado de los prefijos en las variables de los datasets 'Personal'

Esta codificación de prefijos, cuidadosamente estructurada, **simplifica** la navegación y agiliza tanto el proceso de ingeniería de características como el análisis de variables, contribuyendo así a un proceso más efectivo y coherente.

Cabe mencionar que se procuró estandarizar, en la medida de lo posible, la medición de **variables subjetivas**. Escalando la medida en un rango comprensible y manejable, entre el 0 y 10, para todos los participantes, de la siguiente forma (pongamos como ejemplo la medida de felicidad): el 5 corresponde al estado normal para el participante y, conforme avanzamos en la numeración, la diferencia en el cambio de estado aumenta, siendo 6 'quizás un poco más feliz de lo normal', 7 'definitivamente más feliz', 8

‘mucho más feliz’. Lo mismo aplica en sentido opuesto y para el resto de variables de evaluación subjetiva.

En la mayoría de los casos, las variables relacionadas con los aspectos ajenos al participante ('ALI_') se basan principalmente en **datos climáticos**, como la precipitación y la temperatura. Estos datos climáticos han sido recopilados utilizando la API de AEMET [AEMET 023], obtenidos de una estación de Barcelona con el identificador '0201D', o de Madrid, Retiro con el identificador '3195', dependiendo de la ubicación del participante. En un caso particular donde un participante (Charlie) se encontraba en Milán, Italia, se descargó el histórico de datos climáticos desde la API 'Open-Meteo Historical Weather API' [Open-Meteo 023].

Es importante destacar que el conjunto de datos "*Relookyou*" es el más completo entre todos los participantes. Este conjunto no solo recopila datos de la aplicación 'Hábitos', sino que también incorpora una variedad de datos biométricos y de actividad personal, obtenidos de un dispositivo inteligente tipo smartwatch llamado 'Garmin Venu2' [Garmin 023]. Entre los datos adicionales, se incluyen mediciones diarias como la calidad y duración del sueño ('TIR_'), la estimación de la batería corporal (una medida inversa del cansancio físico, 'TIR_'), la frecuencia cardíaca media ('SPO_'), las calorías gastadas ('SPO_'), los minutos de intensidad deportiva ('SPO_'), actividades deportivas ('SPO_') y el número de pasos andados ('SPO_'). Por lo tanto, este conjunto de datos no solo es el que contiene el mayor número de registros, sino también el que considera un amplio espectro de variables. Debido a su complejidad, la metodología aplicada para este conjunto fue más elaborada, permitiendo un proceso más completo de ingeniería de características.

Para obtener detalles más precisos sobre los diversos datasets analizados, se incluyen cuadros descriptivos en los anexos 7.3, que proporcionan una visión detallada de cada variable en cada dataset según los distintos participantes.

Planificación y Metodología

A lo largo de este trabajo, el objetivo fundamental ha sido obtener conclusiones significativas sobre los factores determinantes de la felicidad personal. Para alcanzar este propósito, hemos seguido una metodología rigurosa y estructurada que abarca desde la comprensión inicial del área de estudio hasta la evaluación de los resultados obtenidos, pasando por la extracción y análisis de los datos, así como la aplicación de técnicas de aprendizaje automático. Siendo este un proceso iterativo de mejora continua, basado en la metodología **CRISP-DM** (Cross-Industry Standard Process for Data Mining), que nos ha proporcionado una estructura sólida para nuestro enfoque de investigación (véase la figura 4.1).

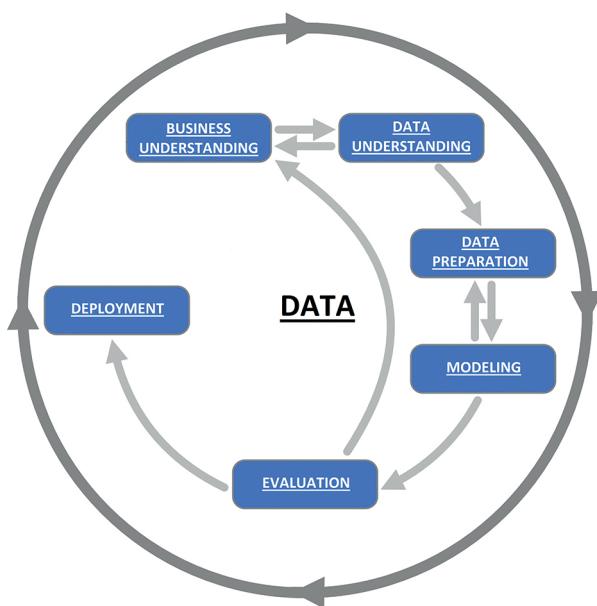


Figura 4.1: Esquema de CRISP-DM (Cross-Industry Standard Process for Data Mining)

El primer paso en el desarrollo de este trabajo, antes de que se aprobara la propuesta de TFM de manera extraoficial, fue la adquisición de los datos necesarios para la investigación. En un principio, nuestra intención era utilizar un conjunto de datos extenso y longitudinal que capturara información sobre las actividades diarias de una o varias personas, junto con su índice de felicidad. Sin embargo, no fue posible encontrar un conjunto de datos público que cumpliera con estas características, lo que nos llevó a tomar la decisión de **recopilar los datos personalmente**.

Dado que la **recopilación manual de datos** es un proceso que consume tiempo, y que nuestro calendario para la realización del TFM era limitado, optamos por abordar el problema del volumen de datos limitado mediante la inclusión de múltiples participantes en el estudio. Además, para enriquecer aún más nuestro trabajo, decidimos utilizar los resultados de una encuesta que proporcionaría información adicional para el análisis de los datos personales y aportaría amplitud al estudio.

Era necesario, pues, idear un método de obtención de datos. Con ese fin se utilizó la aplicación móvil '**Hábitos**', con la que los distintos participantes podrían registrar diariamente las distintas actividades y factores que, según su percepción, influían en su felicidad personal, así como el índice de felicidad percibida cada día. Esta metodología nos proporcionó una fuente rica y detallada de datos sobre las experiencias diarias de los participantes y sus niveles de felicidad.

Simultáneamente, realizamos una **búsqueda bibliográfica** de ciertos conceptos clave para nuestro marco de investigación, como la puesta en común de cómo debían evaluar la felicidad los distintos participantes (explicado en el apartado 3.4.2). También investigamos las mejores prácticas metodológicas en el campo de la minería de datos y el aprendizaje automático en nuestro contexto, lo que nos ayudó a establecer nuestro esquema metodológico.

Cabe mencionar que, a lo largo del trabajo, en ningún momento se ha usado una tecnología de “Organizational Prototyping”. La **planificación** y organización de tareas fue más bien rudimentaria. Este factor terminó siendo una gran carencia en el aspecto organizativo del trabajo, y nos arrepentimos de no haber tenido en cuenta dichas tecnologías. De otro modo, probablemente nos hubiésemos percatado de más de un error de planteamiento. Por ejemplo, seguramente, no se hubiese invertido tal cantidad de tiempo y esfuerzos en la parte del estudio relativa a ‘Encuesta’, que recibió una atención desproporcionadamente alta en relación a su trascendencia con el proyecto.

Nuestra tutora en el TFM, con muy buen criterio, nos instó a **planificar** y organizar las diferentes tareas previstas para tener una primera versión del TFM a finales de julio. Así que, de un modo quizás demasiado optimista, fijamos los plazos para las siguientes tareas (se trata de tareas prácticas, sin tener en cuenta el redactado de la memoria):

- La clusterización del conjunto de datos ‘Encuesta’ debía estar lista para el 4 de junio.
- Un primer análisis descriptivo de un conjunto de datos personales para el 25 de junio. La idea era empezar a trabajar con un conjunto, asentando la arquitectura de trabajo para el resto de datasets personales.
- El 9 de julio correspondía a la finalización del estudio de los datos personales para la detección de patrones de comportamiento. De nuevo, solamente para un conjunto de datos.
- Y el 23 de Julio la identificación de los factores más relevantes que influyen en el nivel de felicidad personal del participante.

Nada más lejos de la realidad. En día 24 de julio, con más de 230 horas de trabajo, recién finalizaba la primera de las tareas... Como ya se mencionó, se invirtieron demasiados esfuerzos en el estudio de ‘Encuesta’, y a pesar de eso seguía teniendo mucho margen de mejora, tareas que quedarán pendiente para el **trabajo futuro** (7.2).

Desde ese entonces, hasta el 6 de agosto, nos dedicamos a plasmar en la memoria los aspectos del trabajo relativos al estudio ‘Encuesta’. Nos encontrábamos en un momento en que, si bien la primera parte del trabajo estaba casi lista (solo faltaba la evaluación de los resultados), aún quedaba la, más laboriosa y **apremiante**, segunda parte. Conscientes de la situación en la que nos encontrábamos y de los motivos de la falta de tiempo, se dedicó todo el 7 de agosto solamente a un único objetivo: el de planificar correctamente el resto de tareas pendientes, que no eran pocas. Más adelante, en vista de cómo progresó esta segunda parte, recordamos el significado de la cita “Dame seis horas para cortar un árbol y pasaré las primeras cuatro afilando el hacha” [Lincoln sf].

Cabe mencionar que el aumento en la cantidad de horas trabajadas diarias, así como de su productividad, propiciados por el carácter cada vez más urgente e imperioso del

desarrollo del trabajo, también fueron factores clave en el **cumplimiento del programa**. Del mismo modo, el hecho de incluir en el estudio solamente una selección de entre todos los participantes, y que la mayor parte del código empleado en el estudio ‘Personal’ seguía la misma arquitectura que el desarrollado en ‘Encuesta’, también fueron determinantes. Concretamente la organización se basó en los siguientes puntos clave:

- Hasta el 13 de agosto, se debía completar el análisis de los conjuntos de datos de ‘Edogawa’ y ‘Juju’. Dado que se emplearon metodologías similares para analizar los distintos conjuntos de datos, se optó por comenzar con ‘Edogawa’, uno de los conjuntos de datos con menos características. Esto permitió diseñar el esquema de análisis, aplicable luego al resto de los conjuntos de datos, empezando con los más simples.
- El objetivo para el 20 de agosto era finalizar los análisis de los conjuntos de datos de ‘Dolphin’, ‘Charlie’, ‘Pato’ y ‘Ajara’. Además, se planificó concluir también con la redacción de la sección de la memoria correspondiente a la primera parte del estudio, es decir, el análisis de la encuesta.
- Para el 27 de agosto, se previó completar el análisis de los conjuntos de datos restantes, ‘Mack’ y ‘Relookyou’, que eran los más complejos. También se planeó finalizar la redacción de las secciones de descripción de datos y metodología de la segunda parte del estudio, los datos personales.
- La meta para el 3 de septiembre era finalizar la redacción completa de la memoria. Esto incluía los subapartados pendientes del capítulo de preliminares, así como los resultados personales y las conclusiones, entre otros.

Esta vez, aunque de manera muy ajustada, se pudo cumplir con éxito el programa establecido.

En la figura 4.2 se presenta una especie de diagrama de Gantt ‘retrospectivo’, el cual fue confeccionado una vez que el proyecto se encontraba en su fase final. Su objetivo principal es documentar el cronograma de las diversas tareas, más que el de organizar las mismas. En este gráfico, se pueden observar las tareas agrupadas en distintos colores, cada uno representando un contexto específico. Cabe señalar que, en lo que respecta al desarrollo de la encuesta, no fue viable distinguir sus diferentes etapas de manera clara; este proceso fue una iteración prolongada de mejora continua, lo que hace que las fases individuales no sean fácilmente discernibles.

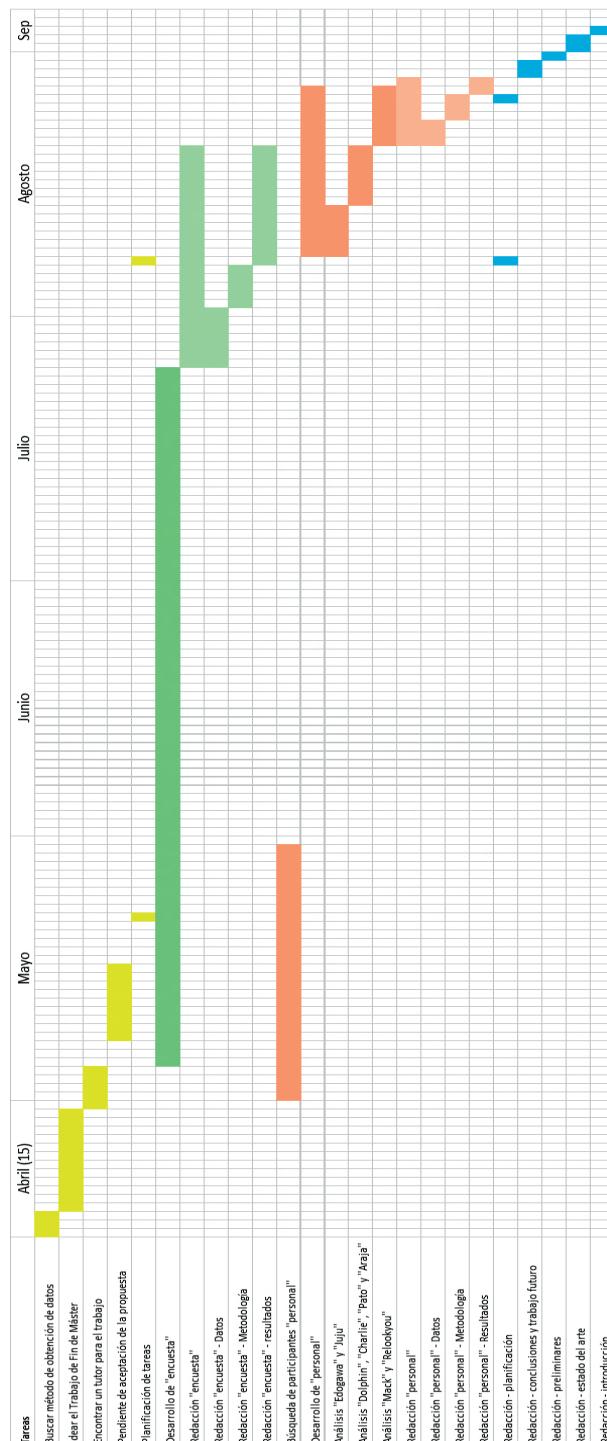


Figura 4.2: Diagrama de Gantt retrospectivo

Contribución Metodológica

La contribución metodológica se divide en **dos secciones** principales: 'Encuesta', que detalla la metodología aplicada en el análisis de los datos recopilados a través de la encuesta, y 'Personal', que se enfoca en el análisis de los datos concernientes a las actividades y hábitos diarios de los participantes, explorando cómo estos influyen en la felicidad personal.

Iniciamos con el primer conjunto de datos, '**Encuesta**'; que alberga las respuestas a un cuestionario diseñado para definir el perfil personal de los encuestados y los tópicos que más influyen en su felicidad, junto con el conjunto de actividades y hábitos asociados a cada tópico. Nuestra intención radica en segmentar este dataset mediante la técnica de K-means, permitiendo así no solo la identificación de las características distintivas de cada segmento, sino también la determinación de la importancia relativa de cada característica en la clasificación.

Posteriormente, trabajaremos con múltiples conjuntos de datos ('**Personal**'), que capturan las actividades diarias y los hábitos, así como la medición subjetiva de la felicidad experimentada, por ocho participantes distintos. El propósito es discernir las variables que mayor influencia tienen en la felicidad de cada participante. Estos, responderán también la encuesta, permitiendo la determinación de los grupos a los que pertenecen según su segmentación en 'Encuesta' y los supuestos factores que más influyen en su felicidad. Este procedimiento trata de procurar una guía que ayude a focalizar correctamente la atención entre las variables de cada dataset 'Personal'. Adicionalmente, este procedimiento permitirá confirmar o refutar las generalidades previamente identificadas según al análisis de 'Encuesta'.

En cada sección examinaremos las técnicas de análisis utilizadas y los desafíos específicos que emergieron durante el proceso, junto con las soluciones adoptadas para superarlos. A través de esta descripción exhaustiva, buscamos no solo brindar una comprensión clara de la metodología detrás de este estudio, sino también resaltar las

contribuciones distintivas que este trabajo puede aportar al campo de la investigación sobre la felicidad y el análisis de datos.

5.1 Encuesta

A continuación, se describirá en detalle cada paso del proceso metodológico empleado en el análisis conjunto '*Encuesta*'. Desde la adaptación y preparación del dataset para su análisis (5.1.1), hasta la aplicación de técnicas específicas, como K-means (5.1.2), para la clusterización de los datos y el análisis de significancia entre grupos (5.1.3). También, se describirán los modelos de aprendizaje automático supervisado utilizados (5.1.4) y se explicará cómo se llevó a cabo el ensamblado de resultados para obtener conclusiones más robustas (5.1.5). La figura 5.1, que se muestra al final del capítulo, ayudará a una mejor comprensión de la arquitectura seguida.

5.1.1 Transformación y adecuación del dataset

La adecuación y transformación del conjunto '*Encuesta*' es un paso fundamental para preparar los datos antes de su posterior análisis. En esta sección, describiremos los distintos procesos llevados a cabo para garantizar la calidad y validez de los datos, buscando un equilibrio entre explicar los procesos en orden cronológico y facilitar la comprensión del lector. Nos encontramos en el apartado azul de la figura 5.1; en ella se pueden apreciar algunos de los procesos que aquí se describen.

Como ya hemos comentado en el apartado 3.4.1, la finalidad del formulario era ligeramente diferente de la del estudio, de modo que se identificaron ciertas columnas que no aportaban información relevante para el caso de estudio. Estas columnas fueron descartadas para mantener el conjunto de datos más limpio y enfocado en los aspectos pertinentes para el análisis.

Gracias a las características del formulario, se aseguró que **no pudiera haber valores nulos** en las respuestas: el encuestado debía seleccionar siempre el número de respuestas indicado. Sin embargo, existía una única fila que contenía valores nulos: este registro se creó cuando el formulario aún se encontraba en una etapa temprana de desarrollo, con características ligeramente diferentes. Esta fila fue eliminada, asegurando la integridad de los datos.

Las columnas de elección múltiple, en las cuáles no se identifica una relación de orden o jerarquía, requerían de una adecuación para su posterior estudio. Se aplicó la técnica '**One-Hot Encoder**' para convertir cada posible valor de las columnas categóricas en una nueva columna binaria, facilitando su análisis y permitiendo su inclusión en modelos de aprendizaje automático. Se valoró la posibilidad de aplicar otras técnicas, como '**Target Encoder**', pero finalmente se asumió que el aumento de dimensionalidad del dataset en aplicar '**One-Hot**' era suficientemente manejable, así que se optó por esta técnica, tratando de conservar el máximo de información original [Brownlee 023]. En el apartado azul de la figura 5.1, apreciamos cómo, para cada variable, se indica entre paréntesis el número de columnas que abarca después de aplicar One-Hot Encoder.

Existen ciertas respuestas del formulario que debieron ser manejadas con especial atención, éstas se tratan de las **respuestas propuestas** por los encuestados que eligieron la opción “[otro]” (3.4.1). Éstas se catalogaron según su temática y contexto, agrupándolas, siempre que fuera posible, en categorías ya existentes [Chowdhury 023, Rençberoğlu 023]. Para las que su temática y contexto difería suficientemente respecto a las categorías existentes, si se podían formar grupos de 3 o más entre ellas, se consideraron significativas y, a pesar de su baja proporción en comparación al resto de respuestas establecidas por defecto, se incorporaron al análisis [Rençberoğlu 023]. Aquellas que no pudieron incorporarse en ninguna de las respuestas preestablecidas, ni agruparse entre ellas con un mínimo de 3 integrantes, fueron despreciadas.

Con el fin de buscar un análisis más detallado del dataset; ‘Encuesta’ se **dividió en tres porciones**, tratando que el estudio no sea solamente generalizado, a partir del dataset original, sino que se focalice también en cada uno de sus subconjuntos. Ya hemos hablado de cómo se estructuran las distintas preguntas del formulario en el subapartado 3.4.1, así que no nos debería sorprender que las porciones se delimiten del siguiente modo: la relativa a las respuestas del perfil personal, denominada “*pProfile*”; la porción que hace referencia a los ocho tópicos anteriormente mencionados, “tópicos”; y la porción específica para el análisis en profundidad de cada uno de estos tópicos, denominada “*subtópicos*”. Véase la figura 5.1 para visualizar gráficamente cómo se fragmenta el dataset; en cada subconjunto del dataset original se indica el número de columnas que contiene.

En la inicialización del algoritmo K-means para la clusterización de datos, se identificaron seis columnas [‘*sex_Man*’, ‘*sex_Woman*’, ‘*sex_Other*’, ‘*age*’, ‘*size*’, ‘*weight*’], pertenecientes a “*pProfile*”, que podrían actuar como **variables de confusión** (véase la figura

5.1). Durante las primeras iteraciones del análisis, observamos que el algoritmo tendía a otorgarles una influencia desproporcionada, afectando significativamente a la formación de grupos y a las conclusiones obtenidas; distorsionando así la identificación de los factores realmente relevantes para la felicidad. Se decidió descartar dichas columnas durante el proceso de clusterización, evitando que los resultados se vieran sesgados por atributos, digamos, menos pertinentes al estudio.

Estos procesos de transformación, adaptación y adecuación de datos, son fundamentales para garantizar la calidad y la validez de los resultados obtenidos a partir del análisis del dataset 'Encuesta'. Una vez realizados, el conjunto de datos se encuentra listo para ser analizado.

5.1.2 K-means

La aplicación del algoritmo K-means es un paso crucial en nuestro caso de estudio ya que las posteriores etapas dependerán enteramente de este proceso. Por esta razón, se han tomado precauciones para procurar que la elaboración de K-means sea lo más acertada posible. A continuación, se describen detalladamente las metodologías seguidas en la aplicación del algoritmo. Nos encontramos en el apartado verde de la figura 5.1; en ella se pueden visualizar gráficamente los procesos descritos en esta subsección.

El hiperparámetro 'k', que representa el número de clústeres, es fundamental en el algoritmo K-means. Se realizó una búsqueda exhaustiva del **valor óptimo de 'k'** considerando diferentes aproximaciones tales como los métodos Elbow y Silhouette y los índices Calinski-Harabasz y Davies-Bouldin (véase la sección 3.3.1). Para cada una de ellas, se contrastaron los resultados con 10 iteraciones y se aplicó la validación cruzada (cross-validation = 5) mediante k-Fold, para obtener métricas de evaluación y resultados más robustos.

El método Elbow, Si bien no es siempre fue determinante, nos proporcionó una estimación inicial del número óptimo de clústeres. Al método Silhouette se le atribuyó una mayor importancia debido a su capacidad para medir la cohesión y separación de los clústeres. Aunque en general tuvo un papel más secundario, el índice Calinski-Harabasz tuvo especial relevancia en el dataset "topics". En cuanto al índice Davies-Bouldin, aunque no aportó información relevante en la elección del valor de 'k', se utilizó para evaluarlo junto con los demás métodos, como una medida adicional de la calidad de los clústeres.

La valoración conjunta de los diferentes métodos, basada en la elección del valor 'k' más votado, combinada con un **criterio propio de evaluación**, permitió la elección del valor óptimo de 'k' para cada uno de los cuatro datasets que formarán parte del estudio. El proceso de selección de 'k' fue riguroso y se buscó garantizar que el número de clústeres fuera adecuado para un análisis significativo de los factores influyentes en la felicidad personal.

Una de las dificultades encontradas en el proceso de aplicación del algoritmo K-means fue la elección del valor óptimo de 'k' para cada dataset. Si bien las diferentes aproximaciones sugerían valores distintos, se tomó la decisión de escoger un valor de '**k=3 para todos los datasets**'. Esta elección se basó en la búsqueda de un término medio que evitara una clusterización 'dualizada' con dos clústeres, y permitiera una representación más equilibrada de los grupos, a pesar de que algunas aproximaciones indicaban 'k=2 como valor óptimo. En [Wagstaff 2001] se destaca la importancia de tener en cuenta el propio criterio y los conocimientos previos sobre el dominio o dataset en la aplicación del K-means.

Una vez seleccionado el valor óptimo de 'k', se llevó a cabo una búsqueda de la **mejor combinación de hiperparámetros** para el algoritmo, mediante 'Grid Search', con un 'cross validation' consistente en 15 folds. En este proceso se invirtieron grandes esfuerzos, realizando sutiles variaciones en las diferentes opciones posibles para cada hiperparámetro, hasta comprobar que la métrica de rendimiento no mejoraba significativamente.

Finalmente, se ejecutó el algoritmo K-means con el 'k' idóneo seleccionado según las distintas aproximaciones mencionadas, y con la mejor combinación del resto de hiperparámetros encontrada mediante GridSearch. Se asignaron **etiquetas** a cada registro, correspondientemente al clúster al que pertenecen, en copias de los respectivos datasets "*encuesta_*", "*pProfile_*", "*topics_*" y "*subtopics_*". Estas etiquetas servirán para clasificar a los encuestados y sus respuestas en clústeres específicos, permitiéndonos distinguir los determinantes de la felicidad propios de cada grupo.

Los esfuerzos invertidos en la elección del valor de 'k' y la búsqueda de la mejor combinación de hiperparámetros, prometen una mayor precisión y relevancia de los resultados obtenidos a través del algoritmo K-means. Un enfoque metodológico riguroso en esta etapa es esencial para asegurar que los patrones identificados sean fiables y representativos.

5.1.3 Análisis de significancia de diferencias entre grupos

Una vez etiquetado el dataset en cuestión, podemos proseguir con el análisis de significancia de diferencias entre grupos. Los procesos realizados que aquí se describen se han tratado de explicar de modo entendedor, pero revisar el trasfondo metodológico 3.3.2 y visualizar los procesos en el apartado amarillo de la figura 5.1 puede ayudar a una mejor comprensión.

Puesto que el número de grupos formados al realizar el K-means es siempre 3, independientemente del dataset clusterizado, y que todas las variables estudiadas son binarias; todas las tablas de contingencia formadas para analizar las diferencias entre grupos serán de **3x2**.

En el contexto de tablas de contingencia de estas características, el test de **Chi cuadrado** es comúnmente usado [Waine 023]. Fijamos un nivel de significancia en 0.05 para determinar si existen diferencias significativas entre los grupos formados por la clusterización, para cada una de las variables presentes en el dataset analizado (ya sea "*pProfile_*", "*topics_*", "*subtopics_*" o "*encuesta_*").

Durante el análisis, se comprobó que ciertas tablas de contingencia generadas para su evaluación, presentaban **frecuencias bajas**, de modo que se creyó conveniente usar algún tipo de corrección para prevenir errores del tipo I.

La **corrección de Yates** se desarrolló originalmente para ajustar el valor del estadístico de contraste cuando se trabaja con tablas de contingencia de 2x2, que son muy sensibles a frecuencias bajas [how2stats 023]. A pesar de que nuestras tablas de contingencia son de 3x2, teniendo en cuenta que la dimensión de la tabla no aumenta demasiado (y por ende tampoco los grados de libertad de Chi2), y la presencia relativamente elevada de tablas de contingencia con bajas frecuencias (en ocasiones, nulas), hemos considerado acertada aplicar la corrección mencionada para reducir la posibilidad de obtener resultados significativos sobreestimados.

Como nos encontramos en un contexto de múltiples pruebas de hipótesis, es necesario aplicar algún tipo de corrección para reducir el riesgo de obtener falsos positivos. Nosotros aplicaremos primeramente la **corrección de Bonferroni**. Una vez identificadas las variables que presentan diferencias, haremos tests de comparación múltiple para cada par de grupos de las variables significativas, con el fin de encontrar cuáles son los

grupos que difieren para cada variable que presenta diferencias entre grupos. También visualizaremos gráficamente los resultados para encontrar nuestras primeras conclusiones en cuanto a las características propias de cada grupo.

Luego, para asegurarnos de obtener resultados más robustos, también aplicamos las siguientes correcciones: **Holm-Bonferroni**, y **Benjamini-Hochberg**. En esencia, calculamos los p-valores ajustados y buscamos el **promedio** entre las tres aproximaciones, según las tres correcciones; luego creamos un dataset que recoja los resultados, al que llamamos "`{ }analysis`" (donde `{ }` se sustituye por "`pProfile_`", "`topics_`", "`subtopics_`" o "`encuesta_`", según el dataset estudiado). Esta información nos permitirá tener una visión más precisa de las diferencias significativas encontradas entre los grupos.

De modo que la exploración de diferencias significativas entre grupos, para tener una primera idea, simple pero efectiva, de cuáles son las características propias de cada grupo, se ha hecho teniendo en cuenta solamente las correcciones de Yates para frecuencias bajas y de Bonferroni para tests de múltiples pruebas. En cambio, los resultados que se usarán posteriormente en complementación con los obtenidos a partir de los diferentes modelos de aprendizaje automático supervisado (véase los siguientes apartados 5.1.4 y 5.1.5), deberán ser lo más **robustos posible**, así que serán el promedio de los obtenidos en las 3 aproximaciones; a estos resultados los denominamos "`{ }analysis`". Quizás sea necesario ver de nuevo la figura 5.1 para esclarecer cualquier duda.

El análisis de significancia entre grupos nos proporcionará una comprensión más profunda de los factores y atributos que distinguen a cada uno de los clústeres identificados. Al examinar detenidamente las características específicas que varían significativamente entre los grupos, podemos descubrir preferencias o atributos que son relevantes para **definir las particularidades de cada segmento**. Este análisis es esencial para identificar patrones de comportamiento y determinar qué variables influyen, supuestamente, en la felicidad personal según cada grupo.

Este proceso de análisis de diferencias significativas entre grupos difiere de los siguientes apartados en el **enfoque y objetivo**. Mientras que en los próximos apartados (5.1.4) nos sumergiremos en la creación y evaluación de modelos de aprendizaje automático supervisado para comprender la importancia relativa de las características en la determinación de los grupos, aquí nos concentraremos en la exploración directa de las divergencias observadas en las variables entre los clústeres identificados. El análisis de

diferencias significativas entre grupos nos permite identificar elementos intrínsecos de los conjuntos de datos que pueden contribuir a la formación de grupos bien definidos, antes de adentrarnos en la interpretación que proporcionarán los modelos predictivos.

5.1.4 Modelos de aprendizaje automático supervisado

Para poder encontrar la importancia relativa de las diferentes características en cuanto a la determinación de si un registro debe pertenecer a un grupo u otro, aplicamos diferentes modelos supervisados de aprendizaje automático, incluyendo Random Forest (RF), Decision Tree (DT), Gradient Boosting (GB), Extreme Gradient Boosting (XGB) y Support Vector Machine (SVM). Cada modelo se seleccionó teniendo en cuenta sus características y su capacidad para abordar el objetivo del estudio (véase 3.3.3). Puesto que todos los modelos siguen la misma estructura, este apartado lo explicaremos en términos generales. Los procesos descritos en aquí pertenecen a la parte roja de la figura 5.1.

En primer lugar, se utilizó GridSearchCV para encontrar la **mejor combinación de hiperparámetros** para cada modelo. Los hiperparámetros relevantes en cada caso fueron ajustados cuidadosamente, lo que permitió obtener un mejor rendimiento y resultados más precisos. Posteriormente, se construyeron los modelos utilizando los hiperparámetros seleccionados.

Estos modelos se plantearon desde una **óptica predictora**, es decir, no se utilizaron simplemente para encontrar la importancia de características en la totalidad del dataset en cuestión, sino que se concibieron como modelos predictivos. Realizando, para cada uno de ellos, diferentes iteraciones, ensamblando las distintas importancias de características obtenidas, y ponderándolas respectivamente según la accuracy de sus predicciones. De este modo tratamos de obtener una mayor solidez en las medidas de importancia de características.

En cuanto al método de división de datos entre entrenamiento y validación, se usó KFold (con 'k'=5). Los resultados y **métricas de rendimiento finales**, que nos servirán para ponderar la importancia de características en el ensamble (véase apartado 5.1.5), se calcularon como el promedio del accuracy para cada división e iteración.

El número de divisiones en KFold se mantuvo constante en 5 para todos los modelos, mientras que el número de iteraciones varió dependiendo del **tiempo de computación**

y la complejidad de éste. Por ejemplo, para el modelo SVM, con un costo computacional menor, se permitió un número máximo de iteraciones de hasta 2000. En general, el número de iteraciones nunca fue menor a 10 y se procuró que el tiempo de computación de los diferentes modelos estuviera entre 30 y 60 segundos.

Es importante mencionar que el modelo **Decision Tree**, aunque se construye de manera similar a los demás, no se tendrá en cuenta para el ensamblaje de modelos (que se explicará en la siguiente subsección 5.1.5). Se utilizará principalmente para visualizar gráficamente cómo se caracterizan los diferentes grupos y, de forma intuitiva pero menos precisa, para comprender la importancia relativa de las características (véase figura 5.1).

La suma de la importancia de características atribuidas a cada variable resultará siempre en 1, para cada modelo, con una excepción. En cuanto al modelo **Support Vector Machine**, la importancia de características se presenta en una escala diferente debido a la naturaleza de su proceso de clasificación. Éste extrae la importancia de las características calculando el hiperplano que mejor separa la distribución multidimensional de los datos y su distancia respecto los puntos de datos [Berwick 023]. De modo que se requiere de un proceso previo de escalado antes de extraer los resultados de este modelo.

Extraemos la importancia de características de todos los modelos, excepto del Decision Tree, para entender la relevancia de las distintas características en la determinación de los grupos. A los datasets resultantes de cada modelo se les asignaron los nombres “{ }RF”, “{ }GB”, “{ }XGB” y “{ }SVM” para Random Forest, Gradient Boosting, Extreme Gradient Boosting y Support Vector Machine, respectivamente; de nuevo, substituyendo ‘{}’ por el nombre de la porción de dataset estudiado.

5.1.5 Ensamblaje

El ensamblaje de modelos es un proceso crucial para obtener resultados más sólidos y precisos en el análisis de la felicidad personal. Este proceso se divide en dos etapas principales (véase la parte morada de la figura 5.1).

La primera etapa se trata del ensamblaje de los resultados del **conjunto de modelos**. Aquí se recolectan los conjuntos de datos generados por cada modelo de aprendizaje automático supervisado, es decir, “{ }RF”, “{ }GB”, “{ }XGB” y “{ }SVM”. Estos datasets

contienen la importancia de características para la determinación de los grupos, obtenida por cada modelo. Cada uno de estos datasets representa una fuente de información valiosa sobre la relevancia de las diferentes características.

La idea general es ensamblar los resultados de acorde a una **ponderación** según su métrica de rendimiento (el promedio de accuracy de todas sus CV e iteraciones), obteniendo un nuevo dataset al que llamamos "`{ }importance`", que reúne la importancia de características de los diferentes modelos. Por esto era necesario asegurarse que todos ellos estuvieran en la misma escala.

Se calcularon los **factores de peso** para cada modelo a partir de la precisión obtenida por cada uno ellos, del siguiente modo:

```
1 weights = [accuracy / sum(accuracies) for accuracy in accuracies]
```

Donde 'accuracies' es una lista que contiene los promedios de las accuracies de los diferentes modelos. De esta forma, cada modelo contribuye en proporción a su precisión, asegurando que los modelos más precisos tengan un mayor impacto en el resultado final. La suma de todos los weight factors será igual a 1.

La segunda etapa trata de combinar el ensamblaje recientemente obtenido ("`{ }importance`") con los resultados del **análisis de significancia de diferencias entre grupos** ("`{ }analysis`"). Antes de eso, es preciso un proceso de adecuación del dataset "`{ }analysis`". Primero tratamos de transformar la 'probabilidad de que no hayan diferencias entre grupos', entiéndase como el p-valor en este contexto de estudio, a algo parecido a una medida de la importancia relativa de cada característica, sencillamente computando ' $1 - p\text{-value}$ '. Luego será necesario dividir cada valor por la suma de todo el conjunto, asegurando así que se encuentre en la misma escala que los resultados de las importancias de los modelos de aprendizaje automático supervisado, cuya suma de importancia de características resulta en uno.

Para **ensamblar** las dos fuentes de información, de modo parecido al anterior proceso, se calculará el factor que ponderará el peso que deberá tener la importancia ensamblada de los modelos "`{ }importance`" respecto la de "`{ }analysis`". Concretamente se usa:

```
1 weight_factor = sum([accuracy / sum(accuracies) * accuracy for accuracy in accuracies])
```

El peso de la importancia ensamblada se calcula como la suma de los distintos pesos para cada modelo del ensamble. Estos pesos, como vemos, se calculan de un modo parecido al anterior, pero se multiplican nuevamente por su precisión, lo que da un enfoque ponderado donde los modelos más precisos tienen un mayor impacto en el weight factor total.

Asumiendo que los modelos tengan un buen rendimiento, se espera que el weight factor sea cercano a su máximo, 1; de modo que se le atribuya mayor importancia a los resultados del ensamblaje de modelos que al del análisis de diferencias significativas entre grupos. El resultado obtenido se almacena en el mismo dataset "`{ }importance`".

Se ha otorgado **mayor importancia al ensamblaje de modelos** debido a su naturaleza multifacética en la metodología empleada. Al combinar los resultados de diversos modelos de aprendizaje automático supervisado, obtenemos una visión más completa y robusta de la importancia relativa de las características en la determinación de los grupos. Cada modelo ofrece perspectivas únicas sobre las relaciones entre las variables y la felicidad personal, explorando cómo diferentes características se combinan para influir en ella, lo que nos permite abordar el problema desde distintos enfoques. En contraste, aunque el análisis de diferencias entre grupos proporciona información valiosa, se enfoca en identificar diferencias significativas entre los clústeres sin considerar las complejas interacciones entre las características.

5.2 Datos personales

En esta sección, nos adentraremos en la explicación metodológica seguida para el análisis de los datos 'Personal', centrándonos en las actividades y hábitos diarios de los participantes y su impacto en la percepción individual de la felicidad. A medida que avanzamos en esta parte del estudio, exploraremos cómo las elecciones y acciones cotidianas tienen el poder de moldear la experiencia subjetiva de la felicidad.

Este análisis adopta un enfoque individualizado, donde cada participante se convierte en una entidad única de estudio. Al contrastar las actividades y hábitos diarios con las evaluaciones subjetivas de felicidad, buscamos resaltar las variables más influyentes en la felicidad personal. Concretamente, el **objetivo** de esta parte del estudio trata de determinar la importancia de características en la predicción de la evaluación subjetiva de la felicidad de cada participante.

Esta sección explicará cómo los datasets ‘Personal’ se integran con los resultados del estudio de la encuesta y describirá el enfoque metodológico adoptado para perseguir el objetivo comentado. Desde cómo se procesaron y adecuaron los datos personales para su estudio, hasta cómo se obtuvieron los resultados de importancia de características; pasando por el diseño de ingeniería de características, la selección de características, la construcción de diversos modelos y el ensamblaje de sus resultados.

En consonancia con lo mencionado en la introducción del capítulo 5, es relevante recordar que estos participantes también completarán la **encuesta**, lo que nos permitirá categorizarlos en función de su pertenencia a los diversos grupos identificados en el estudio de ‘Encuesta’. Recordemos que no se realiza una sola categorización, sino que, al haber segmentado el dataset en tres partes distintas, se obtienen cuatro agrupaciones (según la porción del perfil personal, de los tópicos, de los subtópicos y según el conjunto completo del dataset). Este enfoque garantiza que se preste especial atención a las variables que, se presume, tienen una mayor influencia, y también posibilita la contrastación de los resultados obtenidos en ‘Personal’ con las generalidades de ‘Encuesta’.

Es importante recalcar que, como se explicó en la subsección 3.4.2 , con el fin de mantener la concisión del trabajo, se ha optado por incluir solo una **selección de análisis** entre los ocho realizados. Estos han sido elegidos cuidadosamente para maximizar su contribución al estudio, seleccionando aquellos participantes que, tras su participación en la encuesta, hayan sido asignados a grupos diferentes. Los protagonistas de esta selección incluyen a Relookyou, Dolphin, Juju, Pato, Ajara y Charlie.

En las siguientes subsecciones detallaremos primeramente el procedimiento empleado para categorizar a los participantes en sus respectivos grupos según las respuestas proporcionadas en la encuesta. Seguidamente explicaremos la metodología seguida para el estudio del dataset “Relookyou”, el más completo de todos. Por último, describiremos de forma general la metodología seguida para el resto de datasets, destacando sus particularidades.

Puesto que la metodología es muy parecida entre todos los participantes, se ha decidido añadir una sola figura (5.2), para describir gráficamente el flujo de trabajo seguido para “*Relookyou*”.

5.2.1 Etiquetaje

Dentro del contexto del estudio ‘Encuesta’, se siguió una metodología que permite predecir las etiquetas de **nuevos registros** sin tener que repetir el proceso completo de clustering (K-means). Continuando en el marco de ‘Personal’, este enfoque permitió asignar etiquetas a los nuevos datos de ‘Encuesta’ registrados para cada participante (‘Personal’). Concretamente se utilizaron los centroides previamente calculados con los datos originales, evitando la necesidad de reajustar el modelo K-means. Nos encontramos en la parte verdeazulada de la esquina inferior derecha de la figura 5.2.

Después de entrenar el modelo K-means con los datos de ‘Encuesta’, se obtuvieron los **centroides** de cada clúster. Estos centroides representan los puntos centrales de agrupación en el espacio de características. Una vez obtenidos los centroides, se calculó la **distancia euclíadiana** entre cada nuevo registro (uno para cada participante) y los centroides preexistentes. Esta distancia permitió determinar cuál de los centroides existentes estaba más cerca del nuevo registro, prediciendo así la etiqueta del participante en función de la menor distancia euclíadiana [Martínez Heras 023].

Este proceso de etiquetaje se replicó para cada porción del dataset en el que se aplicó una clusterización (“*pProfile*”, “*topics*”, “*subtopics*” y “*encuesta*”), generando **cuatro etiquetas** distintas para cada participante, según el subconjunto del dataset clusterizado.

La metodología presentada permitió predecir eficientemente las etiquetas de nuevos registros sin necesidad de repetir el proceso completo de clustering ni recurrir a modelos predictivos adicionales. Esto amplió la aplicabilidad de los resultados del modelo K-means a datos novedosos, permitiendo la asignación de etiquetas a los participantes de manera **ágil y eficaz**.

5.2.2 Relookyout

A continuación, procederemos a detallar la metodología del conjunto de datos “*Relookyou*”, que se trata del **más completo** de entre todos los participantes. De este modo, luego, podremos abordar de manera más concisa las metodologías para el resto de conjuntos de datos, resaltando las particularidades de cada uno de ellos.

En cuanto a este participante, después de responder la encuesta y ser etiquetado en los respectivos grupos de acuerdo a su segmentación en función de las distintas por-

ciones del dataset, encontramos que pertenece a los grupos [1, 0, 0, 0] para “*pProfile*”, “*topics*”, “*subtopics*” y “*encuesta*”, respectivamente. En la sección de evaluación de resultados (6.2.1), se profundizará en el significado de esta agrupación. Cabe mencionar que “*Mack*” y “*Edogawa*” también compartieron esta misma combinación de agrupaciones, por lo que sus estudios **no se detallarán** en este informe.

5.2.2.1 Transformación y adecuación del dataset

Ubicándonos en la parte azul de la figura 5.2, primero se **importaron los datos** desde la aplicación ‘Hábitos’ [Álison S 016]; codificando, desde un buen inicio, los prefijos de los nombres de las distintas variables, de acuerdo a el cuadro 3.3.

Luego (en parte verde central de la figura 5.2 hasta que se indique lo contrario), se creó, a partir de la fecha, la columna ‘*ALI_day_of_week*’, indicando el **día de la semana** en formato string. Esta solamente nos servirá para ayudar en el proceso de relleno de valores faltantes; después de cumplir con esta función, se transformará a valor numérico y se le conferirá una codificación cíclica, buscando un mejor rendimiento en los modelos.

Para abordar los **valores nulos**, se implementó un enfoque diferenciado. Se empleó el promedio del día de la semana para completar los valores faltantes, considerando además dos períodos distintos: antes y después del 24 de julio de 2023, fecha en que inició el período de vacaciones del participante. En algunos casos, como en la variable ‘*EVA_happiness_16*’, se pudo tomar en consideración información de otras columnas para mejorar el proceso de llenado, auxiliándonos en los valores de la medida de felicidad de la tarde para arreglar los valores nulos de la felicidad por la mañana. Recordemos que en los anexos se encuentran las descripciones de las distintas variables para cada participante (D.1). A continuación se muestra el proceso:

```

1 #fill (5) 'EVA_happiness_16' NaN with mean of the day of the week (*0.5)
  + values of EVA_happiness_22 (*0.5) (for rows below 2023-07-24)
2 mask_minus_07_24 = habits['date'] < '2023-07-24',
3 mean_day_EVA_happiness_16 = habits[mask_minus_07_24].groupby(
    day_of_week)['EVA_happiness_16'].mean()
4 habits.loc[mask_minus_07_24, 'EVA_happiness_16'] = habits.loc[
    mask_minus_07_24, 'EVA_happiness_16'].fillna(habits.loc[
    mask_minus_07_24, 'day_of_week'].map(mean_day_EVA_happiness_16) *
    0.5 + habits.loc[mask_minus_07_24, 'EVA_happiness_22'] * 0.5)

```

La imputación de valores nulos se limitó a un conjunto selecto de variables, específicamente aquellas que debían ser **constantes** y siempre presentes en cada registro. Este conjunto incluyó características como ['TIR_time_get_up', 'EVA_TIR_sleep_quality', 'EVA_happiness_16', 'EVA_happiness_22', 'EVA_TIR_tiredness_16', 'EVA_TIR_tiredness_22']. Una vez completados estos valores, los demás valores nulos se sustituyeron por 0.

Por último, de nuevo en la parte azul 5.2, se **integraron los datos** biométricos del smartwatch 'Garmin Venu2' y los datos climáticos obtenidos de la API de 'AEMET', en función de la fecha [Garmin 023] [AEMET 023]. Estos datos, que generalmente carecían de valores nulos, se combinaron con los datos de 'Hábitos', también procurando una temprana y coherente codificación de prefijos, facilitando el posterior manejo de datos.

5.2.2.2 Ingeniería de características

Encontrándonos en la parte naranja de la figura 5.2, se generó la columna 'EVA_happiness' como el **promedio** de las evaluaciones subjetivas de felicidad por la mañana y la tarde ('EVA_happiness_16' y 'EVA_happiness_22', respectivamente). Para evitar problemas de multicolinealidad y optimizar el rendimiento del modelo, se eliminaron estas últimas características justo antes de la implementación de los modelos. El mismo enfoque se aplicó para 'EVA_TIR_tiredness_16' y 'EVA_TIR_tiredness_22'.

Luego se calculó el **tiempo productivo total**, sumando las características con prefijo 'PRF_', con ciertas consideraciones adicionales. Se excluyó 'PRF_TFMDrissa' debido a que, de otro modo, se tendría en cuenta dos veces, en 'PRF_TFMDrissa' y 'PRF_study_of_happiness' (véase el anexo D.1 para encontrar una descripción de cada variable). Se ajustaron las variables binarias relativas al tiempo productivo multiplicándolas por valores numéricos adecuados para representar las horas invertidas en cada actividad. Se ajustaron las horas de transporte 'PRF_sum' partiendo de la suposición de que más de la mitad del tiempo de transporte estaba destinado a actividades provechosas. A continuación podemos ver en detalle cómo se construyó 'PRF_sum':

```
1 #sum of all PRF columns
2 habits_prf = habits.filter(regex='^PRF', axis=1)
3 habits_prf = habits_prf.drop(['PRF_TFMDrissa'], axis=1)
4 habits['PRF_sum'] = habits_prf.sum(axis=1)
5 habits['PRF_sum'] = habits['PRF_sum']+habits['XX_PRF_internship']*6
6 habits['PRF_sum'] = habits['PRF_sum']+habits['XX_PRF_chess']*1/3
7 habits['PRF_sum'] = habits['PRF_sum']-habits['PRF_TFMDrissa']
8 habits['PRF_sum'] = habits['PRF_sum']+habits['PRF_FRE_transport']*0.6
```

De un modo parecido se sumaron también las horas relativas al **tiempo libre** en la nueva variable '*FRE_sum*', confiriéndoles el tiempo adecuado a las distintas variables binarias.

En cuanto a la **alimentación**, se desarrolló una metodología para estimar un índice de su nivel de salud o calidad, en función de los distintos alimentos ingeridos. Las variables que determinan los alimentos ingeridos son todas binarias, de modo que se encuentran en la misma escala. El procedimiento consistió en multiplicar cada alimento por un índice determinado y, posteriormente, sumarlos todos, obteniendo el índice diario de salud/calidad en la alimentación. Estos índices para cada alimento se seleccionaron partiendo de cierta búsqueda bibliográfica [OMS 023] [Nestlé 023], pero premió sobre todo el propio criterio: teniendo en cuenta que el participante conoce sus hábitos alimenticios, se creyó lo más apropiado. Por ejemplo, comer legumbres debería tener un indicativo elevado, acorde con sus beneficios, pero quizás, teniendo en cuenta que el participante casi nunca come legumbres, entonces se le debería asignar un indicativo todavía más elevado. Así es como se confeccionó el cálculo del índice de sanidad en la alimentación:

```

1 #state index
2 EAT_index={'XX_EAT_processed':-0.5, 'XX_EAT_fried':-0.4, 'XX_EAT_meat',
   :-0.3, 'XX_EAT_carbohydrates':0.1, 'XX_EAT_dairy':0.3, ,
   'XX_EAT_legumes':0.6, 'XX_EAT_fish':0.7, 'XX_EAT_fruit':0.8, ,
   'XX_EAT_vegetables':1}
3 #build EAT_healthy column
4 habits['EAT_healthy'] = habits['XX_EAT_VIC_meat'] * EAT_index['
   XX_EAT_meat'] + \
5   habits['XX_EAT_fish'] * EAT_index['XX_EAT_fish'] + \
6   habits['XX_EAT_carbohydrates'] * EAT_index['XX_EAT_carbohydrates'] + \
   \
7   habits['XX_EAT_legumes'] * EAT_index['XX_EAT_legumes'] + \
8   habits['XX_EAT_vegetables'] * EAT_index['XX_EAT_vegetables'] + \
9   habits['XX_EAT_fruit'] * EAT_index['XX_EAT_fruit'] + \
10  habits['XX_EAT_dairy'] * EAT_index['XX_EAT_dairy'] + \
11  habits['XX_EAT_fried'] * EAT_index['XX_EAT_fried'] + \
12  habits['XX_EAT_VIC_processed'] * EAT_index['XX_EAT_processed']

```

Como observamos, a algunas variables se les confiere un valor negativo, penalizando el índice, como [`textit'XX_EAT_VIC_processed':-0.5, 'XX_EAT_fried':-0.4, 'XX_EAT_VIC_meat':-0.3`], mientras que el resto se consideraron beneficiosas [`'XX_EAT_carbohydrates':0.1, 'XX_EAT_dairy':0.3, 'XX_EAT_legumes':0.6, 'XX_EAT_fish':0.7, 'XX_EAT_fruit':0.8, 'XX_EAT_vegetables':1`].

De un modo parecido, se creó una variable que resumiera el nivel de **vicios** consumidos diariamente. En este caso, primero fue necesario escalar el conjunto de variables para que su rango de valores se encuentre entre el 0 y 1. Luego asignamos un índice a cada tipo de vicio; de nuevo, con cierta búsqueda bibliográfica para encaminar correctamente el procedimiento [Francia 023] [Montagud Rubí 023], pero premiando sobre todo el propio criterio, en función de la frecuencia y perjuicios personales de cada vicio. Así es como se calculó el índice de vicios diarios consumidos 'VIC', observamos que los que más penalizan son el consumo de cannabis, tabaco, contenidos audiovisuales y alcohol:

```

1 #standarize to values 0-1 habits['VIC_cigars']
2 habits['VIC_cigars_std'] = (habits['VIC_cigars'] - habits['VIC_cigars'].min()) / (habits['VIC_cigars'].max() - habits['VIC_cigars'].min())
3 #standarize to values 0-1 habits['FRE_VIC_audiovisual_content']
4 habits['FRE_VIC_audiovisual_content_std'] = (habits['FRE_VIC_audiovisual_content'] - habits['FRE_VIC_audiovisual_content'].min()) / (habits['FRE_VIC_audiovisual_content'].max() - habits['FRE_VIC_audiovisual_content'].min())
5 #standarize to values 0-1 habits['VIC_alcohol']
6 habits['VIC_alcohol_std'] = (habits['VIC_alcohol'] - habits['VIC_alcohol'].min()) / (habits['VIC_alcohol'].max() - habits['VIC_alcohol'].min())
7 #standarize to values 0-1 habits['VIC_coffee']
8 habits['VIC_coffee_std'] = (habits['VIC_coffee'] - habits['VIC_coffee'].min()) / (habits['VIC_coffee'].max() - habits['VIC_coffee'].min())
9
10 #state VIC index
11 VIC_index={'XX_VIC_weed':2,'VIC_cigars':1.9,'FRE_VIC_audiovisual_content':1.8,'VIC_alcohol':1.7,'XX_FRE_anime':1.6,'XX_FRE_self_touching':1.5,'XX_EAT_meat':1.2,'XX_EAT_processed':1.1,'VIC_coffee':1}
12 #build VIC column
13 habits['VIC'] = habits['XX_VIC_weed'] * VIC_index['XX_VIC_weed'] + \
14     habits['VIC_cigars_std'] * VIC_index['VIC_cigars'] + \
15     habits['FRE_VIC_audiovisual_content_std'] * VIC_index['FRE_VIC_audiovisual_content'] + \
16     habits['VIC_alcohol_std'] * VIC_index['VIC_alcohol'] + \
17     habits['XX_FRE_VIC_anime'] * VIC_index['XX_FRE_VIC_anime'] + \
18     habits['XX_VIC_self_pleasure'] * VIC_index['XX_VIC_self_pleasure'] + \
19     habits['XX_EAT_VIC_meat'] * VIC_index['XX_EAT_VIC_meat'] + \
20     habits['XX_EAT_VIC_processed'] * VIC_index['XX_EAT_VIC_processed'] + \
21     habits['VIC_coffee_std'] * VIC_index['VIC_coffee']
22 #drop std columns
23 habits.drop(habits.filter(regex='std', axis=1).columns, axis=1, inplace=True)

```

Luego se creó la columna 'XX_ACT' como conjunción de las distintas columnas en las que se hubiese hecho algún tipo de **actividad social**. Se creó la columna 'ALI_week',

como una medida del paso del tiempo desde que se inició el estudio, en semanas. Se modificó la columna ‘*ALI_day_of_week*’ para conferirle una codificación cíclica, obteniendo ‘*ALI_day_of_week_sin*’ y ‘*ALI_day_of_week_cos*’; del mismo modo se obtuvieron las columnas ‘*ALI_month_sin*’ y ‘*ALI_month_cos*’, relativas al período mensual.

Luego se creó la columna ‘*SPO_sport*’, indicando la **actividad deportiva**. Esta comprendía diferentes valores registrados desde el dispositivo ‘GARMIN’: 1 para las actividades de caminata, 2 de gimnasio, 3 de cardio y 4 de running. Hubo días en los que se hizo más de una actividad; en esos casos se sumaron los valores indicativos. Adicionalmente, para los días en que no se registró ninguna actividad, se le confirió un valor de 0.5 si se superaba el percentil 75 de las siguientes variables: ‘*SPO_active_calories*’ (550), ‘*SPO_intens_minut_value*’ (19) y ‘*SPO_steps*’ (6565). También se asignó un valor de 2 (gimnasio) los días en que el número de dominadas superaba las 60.

A continuación, se trató de combinar las distintas características relativas al **cansancio** (‘*TIR_*’). Para ello fue necesario un proceso de ingeniería de características algo más complejo. Primero se creó la columna ‘*TIR_sleep_score_mix*’, combinando ‘*EVA_TIR_sleep_quality*’ y ‘*TIR_sleep_score*’ (reescalando antes esta última para que sus valores comprendieran entre el 0 y 10 y ponderándolas del siguiente modo:

```
1 habits['TIR_sleep_score_mix'] = habits['TIR_sleep_score']/10*0.6+ habits
   ['EVA_TIR_sleep_quality']*0.4
```

Luego se creó una nueva columna ‘*inverted_bat*’ (que posteriormente sería eliminada), para adecuar los valores de ‘*TIR_body_bat_mean*’, como una medida de cansancio entre el 2.5 y 8.5, y combinarla con la evaluación subjetiva del cansancio ‘*EVA_TIR_tiredness*’, que comprende el mismo rango de valores, del siguiente modo:

```
1 # invert 'TIR_body_bat_mean'
2 habits['inverted_bat'] = 1 / (habits['TIR_body_bat_mean']/10)
3 #scale 'TIR_body_bat_mean' between 2.5 and 8.5
4 min_scaled_bat = habits['inverted_bat'].min()
5 max_scaled_bat = habits['inverted_bat'].max()
6 scaled_range_min = 2.5
7 scaled_range_max = 8.5
8 habits['inverted_bat'] = scaled_range_min + (scaled_range_max -
   scaled_range_min) * (habits['inverted_bat'] - min_scaled_bat) / (
   max_scaled_bat - min_scaled_bat)
9 # build 'EVA_TIR_tiredness_mix' column
10 habits['EVA_TIR_tiredness_mix'] = habits['inverted_bat']*0.6+ habits[,
   EVA_TIR_tiredness]*0.4
```

A continuación, se escaló la columna 'TIR_sleep_time' entre el 1 y el 9 para que no tome valores demasiados extremos (las distintas características de evaluación subjetiva tampoco son extremas):

```

1 #define the scale range for 'TIR_sleep_time'
2 new_min = 1
3 new_max = 9
4 #make the transformation
5 current_min = habits['TIR_sleep_time'].min()
6 current_max = habits['TIR_sleep_time'].max()
7 habits['TIR_sleep_time_scaled'] = ((habits['TIR_sleep_time'] -
8     current_min) / (current_max - current_min)) * (new_max - new_min) +
9     new_min

```

Y finalmente se construyó la columna combinatoria de las diversas características:

```

1 #state TIR index
2 TIR_index={'TIR_sleep_score_mix':0.8,'EVA_TIR_tiredness_mix':-0.6,'
3   TIR_sleep_time':0.2,'TIR_stress':-0.4}
4 #build TIR column
5 habits['TIR'] = habits['TIR_sleep_score_mix'] * TIR_index['
6   TIR_sleep_score_mix'] + \
7     habits['EVA_TIR_tiredness_mix'] * TIR_index['EVA_TIR_tiredness_mix'] +
8     \
9     habits['TIR_sleep_time_scaled'] * TIR_index['TIR_sleep_time'] + \
10    habits['TIR_stress']/10 * TIR_index['TIR_stress']
11 #drop columns
12 habits.drop(['inverted_bat', 'TIR_sleep_time_scaled'], axis=1, inplace=
13             True)

```

Observamos como la medición combinada de la puntuación de sueño se pondera muy positivamente, en conjunción con la medida del tiempo de sueño escalado, no tan ponderada; en cambio la medición combinada del cansancio se le confiere una ponderación muy negativa, en conjunción con el nivel de estrés, ponderada no tan negativamente.

Finalmente, las variables numéricas que no eran binarias, no tenían carácter evaluativo ni correspondían a características cíclicas ni a la fecha, se **estandarizaron**. Para ello, se aplicó una transformación estándar a estas variables, utilizando el método de StandardScaler. La selección de variables a estandarizar se basó en patrones de prefijos predefinidos para agilizar el proceso.

En resumen, el proceso de ingeniería de características resultó en la creación de una variedad de variables que capturan diferentes aspectos de los hábitos, actividades y características del participante. Estas nuevas características proporcionarán una base más sólida para el análisis posterior y la construcción de modelos predictivos.

5.2.2.3 Correlación y selección de características

Después de realizar la ingeniería de características, se procedió a evaluar la relación entre las variables para identificar posibles redundancias y mejorar el rendimiento de los modelos subsiguientes (parte amarilla de la figura 5.2). Se creó una **matriz de correlación** que permitió visualizar las asociaciones entre las variables y facilitó la selección adecuada de características.

Se generó un mapa de calor (**heatmap**) que visualiza de manera gráfica las correlaciones existentes entre las variables. En el heatmap, se destacaron aquellas correlaciones cuyos valores superaban el umbral de (+/-)0.25. Este enfoque permitió identificar rápidamente que algunas de las variables 'originales' presentaban una alta correlación con las nuevas características generadas a través del proceso de ingeniería.

Dado que se esperaba que la explicabilidad de las características altamente correlacionadas ya hubieran sido capturadas por las nuevas variables generadas durante el proceso de ingeniería de características, se planteó la posibilidad de eliminar algunas de las características originales para simplificar el modelo sin perder significado. Para evaluar el impacto de la eliminación de estas características, se optó por utilizar un modelo de **regresión lineal** de manera sencilla y preliminar. La idea es sencilla como primera aproximación: se calculó la suma del valor absoluto de los coeficientes del modelo entre todas las características, si esta suma aumentaba en eliminar una variable, ésta se eliminó y se procedió a buscar otras variables para eliminar.

```

1 #split the data into features (X) and target variable (y)
2 X = habits.drop(columns=['EVA_happiness'])
3 y = habits["EVA_happiness"]
4 #Build a linear regression model
5 linear_model = LinearRegression()
6 #train the model
7 linear_model.fit(X, y)
8 #obtain the feature importance of the models
9 linear_feature_importance = linear_model.coef_
10 # linear_feature_importance to absolute value
11 linear_feature_importance_abs = abs(linear_feature_importance)
12 #build a datafram with the absolute feature importance
13 linear_feature_importance_abs = pd.DataFrame({
14     "Feature": X.columns,
15     "Linear_Importance": linear_feature_importance_abs,
16 })
17 #sort by importance
18 linear_feature_importance_abs.sort_values(by="Linear_Importance",
    ascending=False, inplace=True)

```

```
19 # obtain the cumulative sum of the absolute importance  
20 print(linear_feature_importance_abs[‘Linear_Importance’].sum())
```

El proceso de selección de características se llevó a cabo con el objetivo de simplificar el modelo y evitar la redundancia en las variables. Una vez eliminadas las características seleccionadas, los datos quedaron preparados para ser integrados en los diferentes modelos de análisis.

5.2.2.4 Modelos predictivos y ensamblaje

Encontrámonos en la parte roja de la figura 5.2, para determinar la importancia de las características en la predicción de la variable objetivo ‘EVA_happiness’ se implementaron diversos modelos. En este caso, atendiendo que la variable EVA_Happiness es numérica, se escogieron modelos de **regression**, concretamente RandomForestRegressor (RFR), GradientBoostingRegressor (GBR), XGBRegressor (XGBR) y Support Vector Regression (SVR). Esta elección permitió aprovechar la metodología previamente utilizada en el análisis de ‘Encuesta’, adaptándola a la naturaleza continua de la variable objetivo.

Para ilustrar, consideremos la metodología aplicada al modelo **RFR**. Inicialmente, se utilizó GridSearchCV para identificar la combinación óptima de hiperparámetros, evaluando diferentes combinaciones en función del error cuadrado medio (MSE). Luego, se construyó y ajustó el modelo utilizando la configuración que demostró un mejor rendimiento, optimizando los hiperparámetros seleccionados en el mismo modelo.

El enfoque principal radica en cuantificar la **importancia de las características** en las predicciones del modelo. Para lograrlo, de modo parecido al del estudio ‘Encuesta’, tratando de asegurar la robustez de los resultados, se estableció una validación cruzada con un valor de 3 folds y múltiples iteraciones. La cantidad de iteraciones se ajustó cuidadosamente para garantizar una ejecución en un tiempo razonable. En cada iteración, el modelo se evaluó utilizando el MSE como métrica de rendimiento.

Dentro de cada iteración, se calculó la importancia de las características en función de su contribución a la predicción. Posteriormente, se **combinaron estas importancias** de características de todas las iteraciones para obtener una evaluación ensamblada. La ponderación se basó en el MSE promedio obtenido en cada iteración, asegurando así que las características se consideraran en función de su impacto en la precisión del modelo.

La ponderación de las importancias de características se efectuó en función del **inverso del error cuadrado medio** promedio obtenido en cada iteración. La elección de invertir el valor MSE se basa en que, a menor MSE, se logra una mejor calidad de predicción. Al invertirlo, se resalta la importancia de las características que contribuyen a reducir el error. Para asegurar la coherencia de la ponderación, los valores invertidos del MSE se escalaron de manera que la suma de los pesos resultara en 1. Esta estrategia garantiza que las características sean valoradas en proporción a su capacidad para mejorar la precisión general del modelo en el contexto de ensamblaje.

Este proceso se repitió para los **diversos modelos** seleccionados, asegurando la coherencia en la evaluación de la importancia de las características a través de los diferentes enfoques de modelado. Igual que en el caso del estudio de la encuesta, es preciso mencionar que fue necesario un proceso de escalado de la importancia de características del modelo SVM, asegurando que su suma resulte en uno.

Finalmente se concluyó con el proceso de **ensamblado** (parte morada de la figura 5.2). Se extrajo el MSE promedio de cada modelo. Luego, siguiendo una metodología parecida al ensamblaje individual de los modelos, se invirtieron los valores promedios MSE de cada modelo y se normalizaron para que la suma de todos ellos resultara en uno. De este modo se logró ensamblar las importancias relativas de características de cada modelo, ponderándolas de acuerdo a la medida, invertida y escalada, del MSE promedio de cada modelo.

Adicionalmente, al dataset de las importancias de características según los distintos modelos y su ensamblado, que se extrajo para estudiar los resultados, se le añadió también una columna indicando los coeficientes resultantes del modelo de **regresión lineal** usado para la selección de características. Esta adición permite comprender no solo la magnitud de la importancia de cada característica, sino también la dirección de su influencia en la felicidad, identificando si impacta positiva o negativamente.

Por último, sirviéndonos de la codificación de prefijos (véase el cuadro 3.3), construimos un dataset aglomerando las importancias atribuidas según los **ocho tópicos** de interés, utilizando los resultados del ensamblaje de modelos. Este dataset contiene la suma y promedio de importancias según los distintos tópicos.

En resumen, en este estudio se empleó una metodología sistemática para construir, evaluar y ensamblar modelos predictivos que permitieran identificar, de manera preci-

sa y confiable, las variables más relevantes en la predicción del nivel de felicidad. Los resultados obtenidos mediante este enfoque, riguroso y sistemático, proporcionan información valiosa para comprender cómo diferentes características pueden influir en la felicidad de una persona.

5.2.3 Resto de datasets 'Personal'

Como mencionamos previamente, dado que ya hemos detallado el estudio más complejo en la descripción anterior, y que la metodología es **muy similar** para los demás datasets (con excepción de los datos extraídos del smartwatch), en esta sección nos centraremos en describir, en términos generales, la metodología seguida para el resto de datasets, resaltando algunas particularidades específicas de cada uno de los otros participantes.

Cabe mencionar que el resto de datasets descritos en este informe se **segmentaron** en los siguientes grupos: "*Dolphin*": [0, 2, 0, 0], "*Juju*": [1, 2, 0, 0], "*Pato*": [0, 2, 2, 2], "*Ajara*": [0, 2, 1, 0] y "*Charlie*": [1, 0, 1, 1]; según las porciones "*pProfile*", "*topics*", "*subtopics*" y "*encuesta*", respectivamente.

Del mismo modo que en estudio de "*Relookyou*", en cuanto a la **transformación y adecuación** de los datasets, se importaron los datos desde la aplicación 'Hábitos' codificando los prefijos de los nombres de las distintas variables. Se creó la columna '*ALI_day_of_week*', para llenar, en función del promedio según el día de la semana, los valores nulos de aquellas variables de carácter constante, y se substituyeron el resto de valores faltantes por cero. Por último, se integraron los datos climáticos obtenidos de las APIs de 'AEMET' o 'Open-Meteo Historical Weather', dependiendo de la ubicación del participante.

Se debe destacar que algunos conjuntos de datos, como "*Juju*" y "*Ajara*", registraron las horas en **formato de tiempo** (hh:mm), en lugar de horas como unidad, de modo que se requirió un proceso de adecuación para ciertas variables, del siguiente modo:

```
1 #select columns to convert
2 columns_to_convert = ['TIR_time_get_up', 'TIR_sleep_time', 'PRF_study',
3   'PRF_reading', 'FRE_VIC_audiovisual_content', 'FRE_music', 'FRE_VIC_SocialNetworks']
4 #as string
5 habits[columns_to_convert] = habits[columns_to_convert].astype(str)
6 #define function to convert "h.mm" format to hours
```

```

6 def convert_hours(time_str):
7     parts = time_str.split('.')
8     hours = int(parts[0])
9     minutes_str = parts[1]
10    if len(minutes_str) == 1:
11        minutes = int(minutes_str) * 10
12    else:
13        minutes = int(minutes_str)
14    return hours + minutes / 60
15 #apply function to columns
16 for col in columns_to_convert:
17     habits[col] = habits[col].apply(convert_hours)
18 habits[columns_to_convert] = habits[columns_to_convert].round(2)

```

Durante el proceso de ingeniería de características, se generó la columna '*EVA_happiness*' como el **promedio** de las evaluaciones subjetivas de felicidad realizadas por la mañana y la tarde. Paralelamente, se obtuvo '*EVA_TIR_tiredness*'. A excepción del dataset "*Charlie*", que solo contenía una evaluación diaria para estas métricas.

Luego se calculó el **tiempo total productivo** ('*PRF_sum*') y el **tiempo total libre** ('*FRE_sum*'). Es importante mencionar que en el caso de "*Charlie*", debido a la carencia de variables en la categoría de tiempo libre, se optó por calcular '*FRE_sum*' como la diferencia entre 24 horas y la suma total de tiempo invertido en diversas actividades.

En el ámbito de la **alimentación**, se aplicó una metodología similar para calcular el índice de calidad nutritiva, con excepción de dos datasets: "*Dolphin*", que carecía de las variables requeridas para dicho cálculo, aunque presentaba una variable binaria '*XX_EAT_Healthy*' para indicar si se había comido de forma saludable; y "*Charlie*", que a pesar de tener las variables necesarias, tenía un alto número de valores faltantes, por lo que no se tuvo en cuenta ninguna clase de variable indicativa de la calidad de alimentación.

De manera análoga, se creó la variable '*VIC*' para resumir el consumo diario de vicios, adaptándola a las particularidades de cada participante en función de sus hábitos. Es relevante mencionar que "*Dolphin*" no registró ninguna variable categorizada como vicio, excepto '*FRE_VIC_audiovisual_content*', lo que imposibilitó calcular '*VIC*' en este caso.

Posteriormente, se creó la columna '*XX_ACT*' como conjunción de las distintas columnas en las que se hubiese hecho algún tipo de **actividad social**. Asimismo, se creó la columna '*ALI_week*', midiendo **tiempo** transcurrido, en semanas, desde el inicio del

estudio. Se implementó una codificación cíclica en '*ALI_day_of_week*' para obtener las columnas '*ALI_day_of_week_sin*' y '*ALI_day_of_week_cos*'; del mismo modo se obtuvieron las columnas '*ALI_month_sin*' y '*ALI_month_cos*', relativas al período mensual.

En cuanto al resumen del **deporte**, dado que no se disponía de datos biométricos ni de actividad deportiva, se consideró simplemente el tiempo dedicado a actividades deportivas por cada participante.

Finalmente, las variables numéricas que no eran binarias, no reflejaban evaluaciones ni pertenecían a características cíclicas o fechas, se sometieron a un proceso de **estandarización** utilizando el método StandardScaler.

Todos los procesos posteriores, desde la selección de características utilizando la métrica de suma del valor absoluto de los coeficientes del modelo de regresión lineal, hasta la construcción de los diferentes modelos y su ensamblado, se llevaron a cabo de manera **idéntica** al caso de 'Relookyou'.

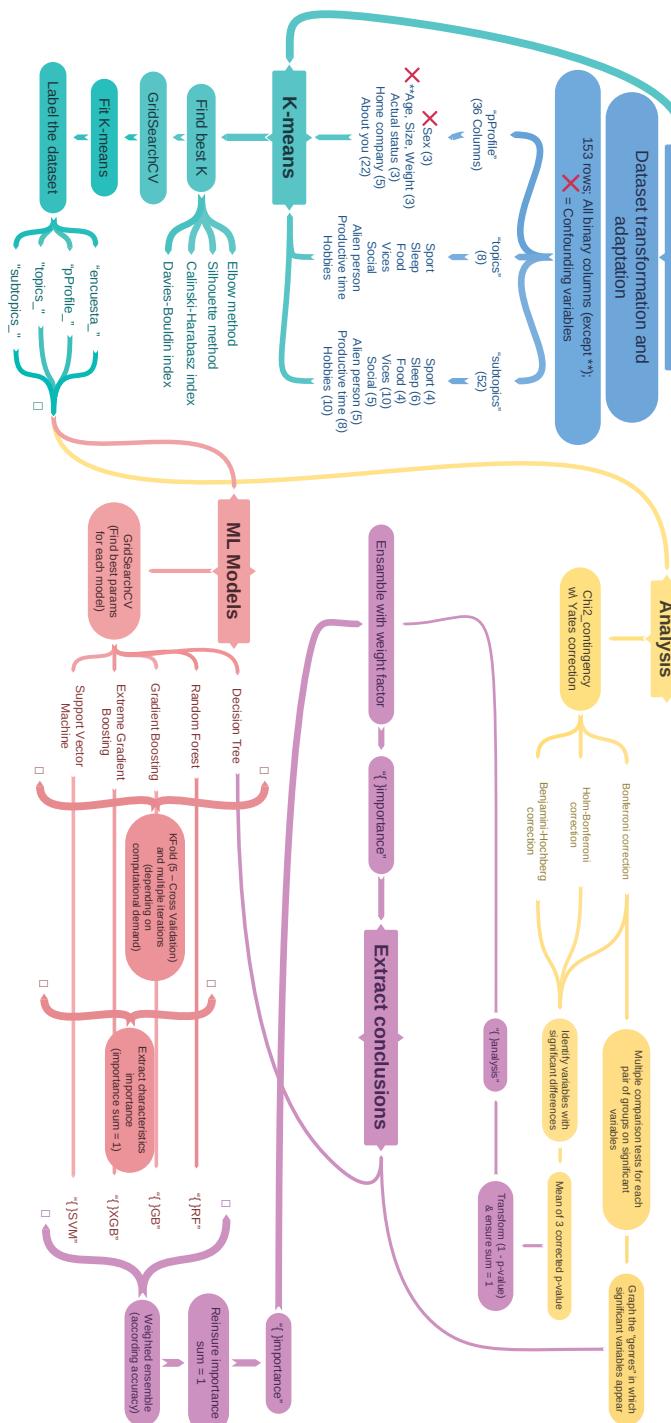


Figura 5.1: Representación gráfica del flujo de trabajo para el estudio de 'Encuesta'

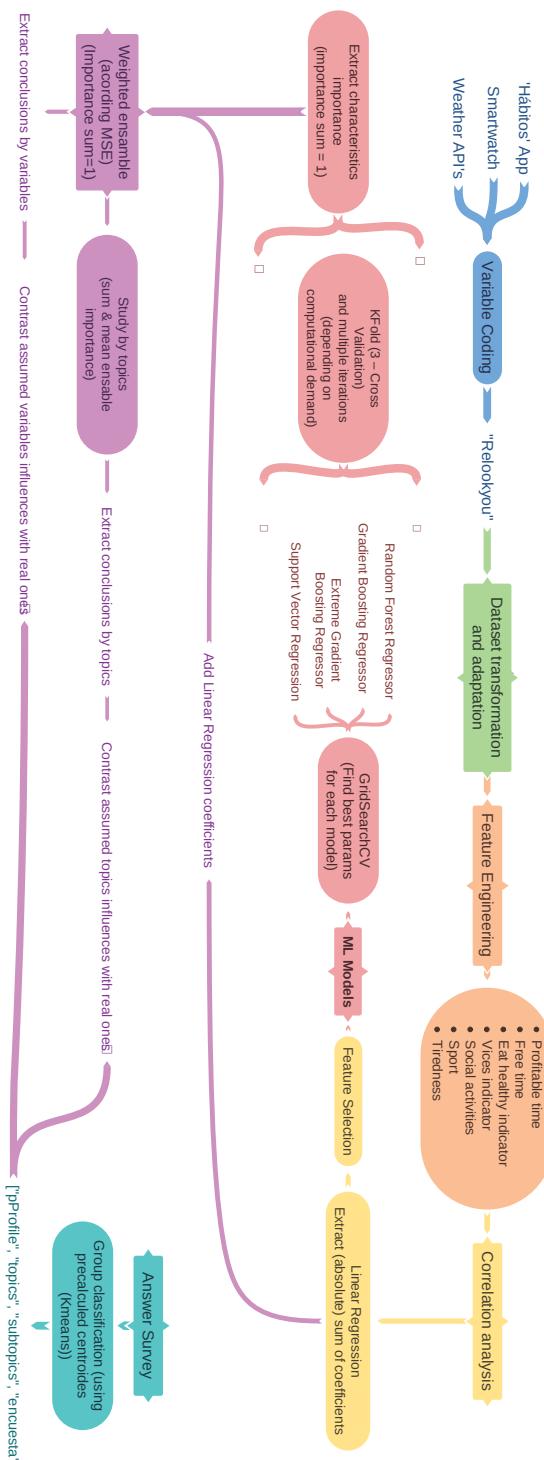


Figura 5.2: Representación gráfica del flujo de trabajo para el estudio 'Personal'

CAPÍTULO 6

Resultados

La intención de comprender los factores que contribuyen a la felicidad y cómo se relacionan con las preferencias, actividades diarias y percepciones individuales ha sido el motor impulsor de este estudio. En este capítulo, presentamos los resultados que han surgido a través de la ejecución de las diferentes etapas metodológicas de dos estudios complementarios, que abordan esta cuestión desde perspectivas diferentes pero convergentes: el estudio 'Encuesta' y el estudio 'Personal'.

La primera sección, '**Encuesta**', detalla los hallazgos obtenidos mediante el análisis de cada uno de los datasets ("*pProfile*", "*topics*", "*subtopics*" y "*Encuesta*"). Nos enfocamos en justificar el número de clústeres elegidos en función de los resultados de diversos métodos de selección de 'k', y en el análisis detallado de las diferencias significativas entre grupos. También exploramos la importancia de características en la determinación de los grupos y expondremos los resultados del proceso de ensamblaje final, que combina los resultados de la combinación de modelos y las diferencias significativas observadas.

La segunda sección, '**Personal**', profundiza en los resultados del estudio personalizado llevado a cabo en aquellos participantes que se segmentaron de manera distinta al estudiar sus respuestas en la encuesta. Aquí presentamos los resultados derivados de la aplicación de las metodologías, previamente definidas (5.2), para analizar cómo diversas características se relacionan con el nivel de felicidad reportado. Extraemos las importancias de las características, obtenidas de los distintos modelos construidos, para identificar las variables más influyentes en la predicción de la felicidad de cada participante. También discutiremos cómo estas características se relacionan con las preferencias y percepciones identificadas en el estudio de 'Encuesta', de acuerdo a la segmentación particular de cada participante.

En la tercera y última sección de este capítulo, se encuentra la **discusión de resultados**, en la que profundizaremos en los resultados obtenidos a través de nuestras dos

aproximaciones, 'Encuesta' y 'Personal', y exploraremos cómo se entrelazan para extraer nuestras conclusiones y lograr un mayor entendimiento de los factores determinantes de la felicidad personal.

6.1 Encuesta

En esta sección se detallan los resultados de acuerdo a los principales pasos del método explicado en la sección 5.1.

6.1.1 K-means

En esta sección, presentamos los resultados obtenidos al aplicar el algoritmo K-means en los diferentes conjuntos de datos: "*pProfile*", "*topics*", "*subtopics*" y "*encuesta*". La elección del número óptimo de clústeres, representado por el valor '*k*', desempeña un papel fundamental en el proceso de clusterización y en la interpretación de los resultados.

Para determinar el **valor óptimo de '*k*'**, empleamos varios métodos ampliamente reconocidos en la literatura: el Método del Codo (Elbow), el Método de Silueta (Silhouette), el Índice de Davies-Bouldin y el Índice Calinski-Harabasz. Cada método nos proporciona una perspectiva diferente sobre la estructura subyacente de los datos y nos ayuda a seleccionar un valor adecuado de '*k*'.

A continuación, presentamos en el cuadro 6.1 los tres valores más relevantes de '*k*' para cada método en cada uno de los conjuntos de datos:

	Method	First Best ' <i>k</i> '	Second Best ' <i>k</i> '	Third Best ' <i>k</i> '
Perfil	Elbow	2; Mean_inert=635	3; Mean_inert=590	4; Mean_inert=560
	Silhouette	2; Mean_coef=0.081	3; Mean_coef=0.066	4; Mean_coef=0.061
	Davies-Bouldin	10; Mean_index=2.35	9; Mean_index=2.45	8; Mean_index=2.5
	Calinski-Harabasz	2; Mean_index=9.8	3; Mean_index=8.8	4; Mean_index=8
Tópicos	Elbow	2; Mean_inert=155	3; Mean_inert=130	4; Mean_inert=115
	Silhouette	10; Mean_coef=0.5	9; Mean_coef=0.46	8; Mean_coef=0.43
	Davies-Bouldin	10; Mean_index=1.05	9; Mean_index=1.15	8; Mean_index=1.22
	Calinski-Harabasz	2; Mean_index=28.2	10; Mean_index=26.4	3; Mean_index=25.7
Subtópicos	Elbow	2; Mean_inert=980	3; Mean_inert=940	4; Mean_inert=910
	Silhouette	2; Mean_coef=0.0485	3; Mean_coef=0.0455	4; Mean_coef=0.0425
	Davies-Bouldin	10; Mean_index=2.7	9; Mean_index=2.75	8; Mean_index=2.85
	Calinski-Harabasz	2; Mean_index=7.3	3; Mean_index=6.3	4; Mean_index=5.6
General	Elbow	2; Mean_inert=1810	3; Mean_inert=1755	4; Mean_inert=1710
	Silhouette	2; Mean_coef=0.048	3; Mean_coef=0.037	4; Mean_coef=0.03
	Davies-Bouldin	10; Mean_index=2.35	9; Mean_index=3.25	8; Mean_index=3.35

Method	First Best 'k'	Second Best 'k'	Third Best 'k'
Calinski-Harabasz	2; Mean_index=6.3	3; Mean_index=5.35	4; Mean_index=4.5

Cuadro 6.1: Cuadro resumen de los 3 mejores valores de 'K' según el método en cuestión y dataset estudiado - 'Encuesta'

Estos resultados proporcionan una sólida guía para determinar el número óptimo de clústeres en cada conjunto de datos. Sin embargo, es importante recordar la relevancia del criterio basado en el conocimiento del dominio y del dataset. Para una comprensión más profunda de las diferentes métricas proporcionadas, consulta nuevamente el subapartado 5.1.2.

En nuestro caso el índice de **Davies-Bouldin**, que mide la similitud promedio entre cada clúster y su clúster más similar, siempre son seleccionados como óptimos los valores máximos propuestos en la determinación del 'k' óptimo. Dado que estos valores difieren notablemente de los obtenidos por los otros métodos y con el objetivo de evitar la formación de demasiados grupos que puedan complicar la comprensión y la extracción de información significativa, se ha considerado conferirle menos importancia a este método.

Por lo tanto, al observar los resultados de los otros métodos, podría ser justificable elegir 'k'=2 para todos los datasets. No obstante, como se mencionó en el subapartado 5.1.2, para evitar una clusterización dualizada, en la que no hubiera término medio en los respectivos grupos, y fomentar una mayor diferenciación, se ha optado por el segundo valor óptimo en cada dataset, asignando un valor de 'k'=3 en cada caso.

En el caso del dataset "**topics**", la decisión final no fue tan sencilla. Como se comentó anteriormente (5.1.2), el método Calinski-Harabasz tuvo un peso significativo en la elección de 'k'. Aunque excepcionalmente se podría haber elegido un 'k'=10 para este dataset, finalmente se optó por 'k'=3, ya que consideramos que 'k'=10 generaría demasiados grupos y podrían dificultar la comprensión y la interpretación del estudio.

Estos valores de 'k' reflejan adecuadamente la diversidad y la complejidad de los datos, facilitando una clusterización significativa y una extracción de información más clara en el contexto de este estudio sobre la felicidad personal. En el siguiente apartado, profundizaremos en la interpretación de los clústeres analizando cuáles son las características que presentan diferencias significativas entre ellos.

6.1.2 Análisis de significancia de diferencias entre grupos

En esta sección, nos adentramos en un análisis detallado de las diferencias significativas de las diferentes características entre los grupos formados mediante el algoritmo K-means en cada uno de los conjuntos de datos: "*pProfile*", "*topics*", "*subtopics*" y "*encuesta*". A medida que hemos identificado y definido los clústeres en la sección anterior, surge la necesidad de entender las características distintivas que separan a estos grupos. Este proceso complementa la perspectiva que ofrecerán los modelos de aprendizaje automático supervisado que exploraremos posteriormente (6.1.3).

Es relevante subrayar que, si bien los conjuntos de datos individuales, como "*pProfile*", "*topics*" y "*subtopics*", se abordan de forma independiente, forman partes interrelacionadas dentro del conjunto de datos más amplio "*encuesta*". A pesar de esta interconexión, los grupos identificados mediante la aplicación del algoritmo K-means en cada uno de estos conjuntos de datos no tienen **ninguna relación** o vinculación entre sí. Cada conjunto de datos representa un enfoque único y autónomo para analizar la totalidad de los datos de la encuesta desde diversos enfoques.

6.1.2.1 Perfil Personal

En este apartado, presentamos los resultados del análisis de diferencias significativas entre los grupos identificados por el algoritmo K-means en el conjunto de datos "*pProfile*". Hemos aplicado las correcciones, para pruebas múltiples a los p-valores obtenidos para cada variable, de Bonferroni, Holm-Bonferroni y Benjamini-Hochberg. La quinta columna muestra el p-valor promedio de las tres correcciones, lo que nos proporciona una medida más robusta de la significancia.

En el cuadro 6.2 se pueden observar los **p-valores ajustados** para las diferentes variables en el dataset "*pProfile*", ordenada según el p-valor promedio, en orden ascendente, de modo que las variables más significativas se encuentran en el inicio del cuadro:

Variable	Bonf.	H-Bonf.	B-Hoch.	Mean
homeCompany_family	4,08E-30	4,08E-30	4,08E-30	4,08E-30
homeCompany_couple	1,26E-18	1,22E-18	6,29E-19	1,03E-18
actualStatus_student	6,28E-14	5,86E-14	2,09E-14	4,74E-14
homeCompany_parents	1,91E-10	1,72E-10	4,77E-11	1,37E-10
actualStatus_work	5,44E-10	4,71E-10	1,09E-10	3,75E-10
aboutYou_dependent	3,91E-07	3,26E-07	6,52E-08	2,61E-07
homeCompany_friends	8,27E-05	6,61E-05	1,18E-05	5,35E-05

Variable	Bonf.	H-Bonf.	B-Hoch.	Mean
aboutYou_videogames	1,83E-03	1,40E-03	2,28E-04	1,15E-03
aboutYou_romantic	1,89E-02	1,39E-02	2,10E-03	1,16E-02
homeCompany_alone	2,38E-02	1,66E-02	2,38E-03	1,43E-02

Cuadro 6.2: Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - "*pProfile*" - solamente entradas significativas - 'Encuesta'

Los p-valores ajustados nos indican la significancia estadística de las diferencias entre grupos para cada variable en el conjunto de datos. Los valores más bajos sugieren que las diferencias observadas son menos probables de ocurrir por casualidad y pueden proporcionar información sobre las características que distinguen a los grupos generados por el algoritmo K-means.

Como se ha mencionado en el subapartado 5.1.3, también se han realizado test de comparación múltiple **para cada par de grupos** de las variables significativas, para identificar así cuáles son los grupos que difieren para cada variable significativa. La figura 6.1 recoge los resultados de este análisis (recordamos, llevados a cabo con las correcciones de Yates y Bonferroni solamente).

	Variable	Group X	Group Y	p-value
0	actualStatus_student	0	1	9.448141e-11
1	actualStatus_student	1	2	2.336188e-08
2	actualStatus_work	0	1	1.364698e-09
3	actualStatus_work	1	2	2.204256e-07
4	homeCompany_friends	0	1	8.178743e-04
5	homeCompany_friends	1	2	2.747637e-03
6	homeCompany_alone	0	1	1.179067e-02
7	homeCompany_alone	0	2	3.933555e-04
8	homeCompany_parents	0	1	3.284264e-10
9	homeCompany_parents	0	2	3.787623e-02
10	homeCompany_parents	1	2	2.087827e-05
11	homeCompany_couple	0	2	2.282895e-16
12	homeCompany_couple	1	2	9.700492e-09
13	homeCompany_family	0	1	3.783943e-19
14	homeCompany_family	0	2	1.074825e-27
15	aboutYou_romantic	0	1	1.092728e-02
16	aboutYou_romantic	1	2	3.444654e-04
17	aboutYou_videogames	0	1	9.968685e-05
18	aboutYou_videogames	1	2	1.067940e-02
19	aboutYou_dependent	0	1	2.419584e-05
20	aboutYou_dependent	1	2	3.772898e-06

Figura 6.1: Test de comparación múltiple para cada par de grupos de las variables significativas - "*pProfile*" - 'Encuesta'.

El cuadro 6.3 recoge la información de la figura 6.1 de un modo **más simple** y entendedor de manera que, a pesar que no muestre el nivel de significancia, la simplicidad del cuadro permite una evaluación rápida y acertada de los resultados. En ella se observan 4 columnas, una para las distintas características con diferencias significativas entre grupos, y una para cada grupo. El contenido del cuadro trata de expresar la proporción, según cada grupo, de cada variable significativa, al mismo tiempo que indica cuáles son los grupos que difieren según el test de comparación múltiple para cada par de grupos de las variables significativas. El modo de interpretación es el siguiente: si dos grupos presentan el mismo valor para una característica, significa que no difieren significativamente; si el valor es diferente, sí que difieren, y la magnitud del valor expresa la proporción de la característica. Por ejemplo, en cuanto a la característica '*homeCompany_parents*', el grupo '0' presenta significativamente menos encuestados con esa característica que el grupo '2'; éste, del mismo modo, presenta significativamente menos encuestados con esa característica que el grupo '1'.

Caract/Grupo ("pProfile")	0	1	2
actualStatus_student	1	3	1
actualStatus_work	3	2	3
homeCompany_friends	1	2	1
homeCompany_alone	1	2	2
homeCompany_parents	1	3	2
homeCompany_couple	1	1	3
homeCompany_family	3	1	1
aboutYou_romantic	2	3	2
aboutYou_videogames	1	2	1
aboutYou_dependent	1	2	1

Cuadro 6.3: Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - "pProfile" - 'Encuesta'

Una vez añadidos los distintos resultados, y explicada la manera de interpretarlos, podemos proseguir con su discusión. Para no cargar demasiado el texto, el resto del apartado no se usará demasiado 'significativo' y sus derivados, cada vez que se hable de mayor o menor que, deberá entenderse como significativamente mayor o menor.

Observamos en el cuadro 6.2 como muchas de las variables pertenecientes al 'género' '**homeCompany**' se posicionan al inicio en cuanto al test de diferencias, de modo que existen grandes diferencias entre grupos en cuanto al tipo de convivencia del encuestado (en familia, en pareja, con los padres, etc.). Observando la figura 6.2 y contrastándola con la figura 6.1 y el cuadro 6.3, podemos extraer las distintas conclusiones:

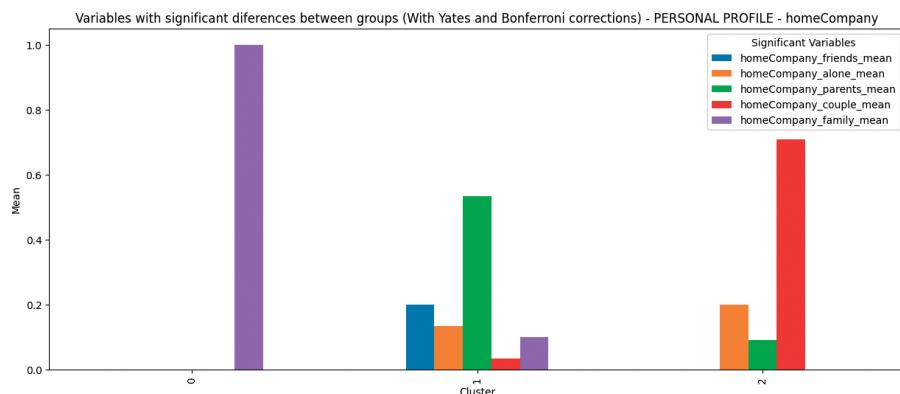


Figura 6.2: Representación gráfica del promedio, según grupo, de las variables significativas pertenecientes al género 'HomeCompany' - "pProfile" - 'Encuesta'.

- El **grupo '0'** solamente **vive en familia**, característica que es significativamente mayor al resto de grupos; de modo que suele tener siempre menos encuestados que viven de cualquier otro modo, que el resto de grupos.
- El **grupo '2'** solamente vive con: padres (significativamente menor al grupo '1'), solo, o en **pareja** (mayor al resto de grupos).
- El **grupo '1'** es el que tiene un espectro de convivencia más amplio, abarcando todas las posibilidades. Su convivencia **con padres** es significativamente mayor al resto y es el único que vive, a veces, **con amigos**.

Las respuestas demográficas en cuanto a si el encuestado trabaja o estudia, también difieren en gran medida según el grupo. En la figura 6.3 identificamos que los **grupos '0' y '2'** se caracterizan por **trabajar**, y '1' más bien se trata de **estudiantes**. Según la figura 6.1 y el cuadro 6.3 confirmamos que no existen diferencias significativas entre los grupos '0' y '2' para el género 'actualStatus', pero si entre el '1' y el resto.

En cuanto al género '**aboutYou**' (6.4), que procura describir al encuestado desde el punto de vista de los hábitos y características personales; de entre todas sus variables, las únicas significativas son el romanticismo, videojuegos y la dependencia económica. En este caso solamente podemos decir que el **grupo '1'** se caracteriza por presentar un mayor número de encuestados **románticos**, que juegan a **videojuegos** y/o con **dependencia económica**, que el resto de grupos.

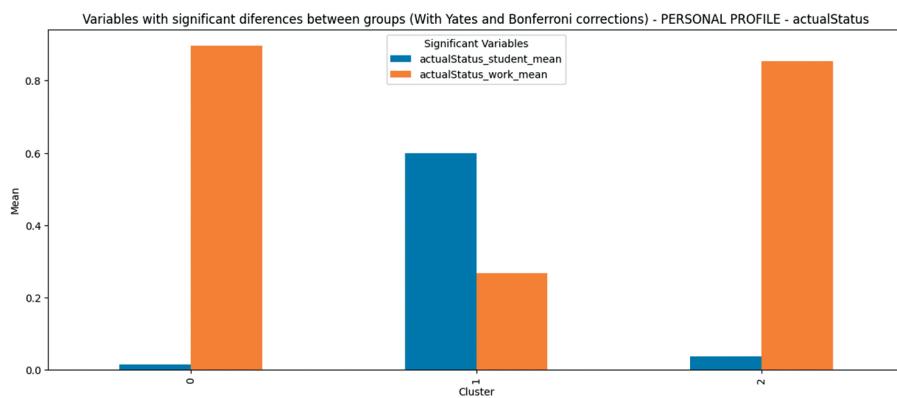


Figura 6.3: Representación gráfica del promedio, según grupo, de las variables significativas pertenecientes al género 'actualStatus' - "pProfile" - 'Encuesta'.

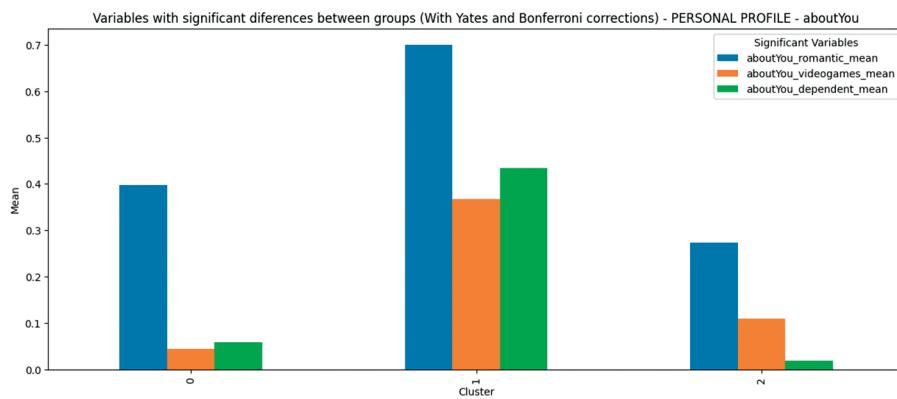


Figura 6.4: Representación gráfica del promedio, según grupo, de las variables significativas pertenecientes al género 'aboutYou' - "pProfile" - 'Encuesta'.

6.1.2.2 Tópicos

En este apartado, presentamos los resultados del análisis de diferencias significativas entre los grupos identificados mediante el algoritmo K-means en el conjunto de datos "topics". Cabe mencionar que, puesto que los análisis del resto de datasets presentan la misma metodología que la del anterior dataset "pProfile", en los siguientes apartados procuraremos evitar redundancias y nos enfocaremos en resaltar los puntos clave.

Al igual que en el análisis anterior, en el cuadro 6.4, se presentan los **p-valores ajustados** para las diferentes variables y correcciones en el dataset "topics", ordenados según el p-valor promedio en orden ascendente.

Variable	Bonf.	H-Bonf.	B-Hoch.	Mean
topics_hobbies	4,78E-33	4,78E-33	4,78E-33	4,78E-33
topics_alienPerson	8,30E-22	7,26E-22	4,15E-22	6,57E-22
topics_sport	4,58E-07	3,43E-07	1,53E-07	3,18E-07
topics_food	4,51E-05	2,82E-05	1,13E-05	2,82E-05
topics_social	4,77E-01	2,38E-01	8,99E-02	2,68E-01
topics_productiveTime	5,40E-01	2,38E-01	8,99E-02	2,89E-01
topics_sleep	1,00E+00	6,85E-01	3,92E-01	6,92E-01
topics_vices	1,00E+00	6,85E-01	6,61E-01	7,82E-01

Cuadro 6.4: Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - "topics" - 'Encuesta'

Del mismo modo que en el anterior dataset de estudio, los resultados de pruebas de comparación múltiple para cada par de grupos de las variables significativas se plasman en la figura 6.5.

	Variable	Group X	Group Y	p-value
0	topics_sport	0	1	1.608598e-04
1	topics_sport	0	2	2.990309e-02
2	topics_sport	1	2	7.584504e-08
3	topics_food	0	1	5.186255e-04
4	topics_food	1	2	6.854116e-06
5	topics_alienPerson	0	1	3.117404e-23
6	topics_alienPerson	0	2	4.216117e-10
7	topics_alienPerson	1	2	3.230481e-06
8	topics_hobbies	0	2	1.747216e-25
9	topics_hobbies	1	2	1.368528e-19

Figura 6.5: Test de comparación múltiple para cada par de grupos de las variables significativas - "topics" - 'Encuesta'

El cuadro 6.5 ofrece una **visión intuitiva** de la proporción de cada variable significativa en relación con cada grupo, así como una indicación de cuáles grupos presentan diferencias significativas según el test de comparación múltiple para cada par de grupos.

Caract/Grupo ("topics")	0	1	2
topics_sport	2	1	3
topics_food	2	1	2

Caract/Grupo ("topics")	0	1	2
topics_alienPerson	1	3	2
topics_hobbies	3	3	1

Cuadro 6.5: Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - "topics" - 'Encuesta'

La figura 6.6 es una representación gráfica del valor promedio, según cada grupo, de las variables significativas pertenecientes al 'género' 'topics'. Basándonos en la interpretación conjunta de los resultados, se pueden extraer conclusiones clave para cada grupo.

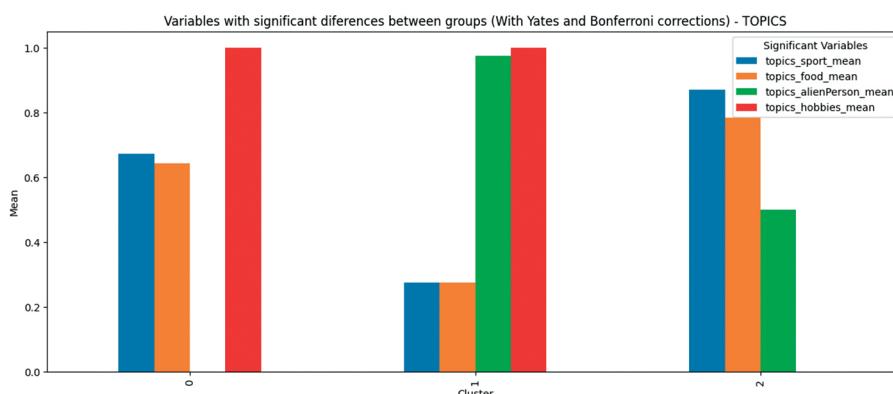


Figura 6.6: Representación gráfica del promedio, según grupo, de las variables significativas - "topics" - 'Encuesta'.

El **grupo '0'** es el que cree que menos influyen en la felicidad las circunstancias ajenas a la persona (incontrolables, como el tiempo) de entre todos los grupos; le da más importancia al deporte que el grupo '1', pero menos que el '2'; le importa más el tópico de la alimentación que el grupo '1'; y piensa que el tópico de 'hobbies' tiene una influencia en el bienestar significativamente mayor que lo que piensa el grupo '2'. En otras palabras: teniendo en cuenta solamente las variables con diferencias significativas de entre los diferentes tópicos, y comparando este grupo respecto a los demás, ellos creen que el **deporte** puede influir en la felicidad, y creen que sin duda lo hace la **alimentación** y los **hobbies**, pero no los aspectos ajenos a la persona.

El **grupo '1'**: atribuye a los hobbies una influencia en la felicidad significativamente mayor a la que le atribuye el grupo '2'; es el que menos importancia le da al aspecto deportivo y de alimentación, pero el que más a aspectos ajenos a la persona. De modo

que las personas de este grupo piensan que el deporte y la alimentación no son determinantes de la felicidad, pero sí los **aspectos ajenos a la persona** y los **hobbies**.

El **grupo '2'**: es el que mayor importancia le atribuye al deporte, y el que menos a los hobbies; le asigna a la alimentación una mayor importancia de la que le concede el grupo '**1**'; y le otorga a los aspectos ajenos a la persona, mayor importancia de lo que lo hace el grupo '**0**', pero menor al '**1**'. Entonces, creen que los **aspectos ajenos a la persona** pueden influir en la felicidad, y creen fervientemente que el **deporte** y la **alimentación** sí que lo hacen, contrariamente a los hobbies.

6.1.2.3 Subtopics

En este apartado, presentamos los resultados del análisis de diferencias significativas entre los grupos identificados a través del algoritmo K-means en el conjunto de datos "*subtopics*". De nuevo, debido a la similitud metodológica con los análisis anteriores, nos centraremos en resaltar los aspectos clave, evitando redundancias.

Al igual que en los análisis previos, el cuadro 6.6 presenta los **p-valores ajustados** para las variables y correcciones correspondientes en el dataset en cuestión. Estos p-valores, ordenados según el p-valor promedio en orden ascendente, resaltan las diferencias significativas entre los grupos.

Variable	Bonf.	H-Bonf.	B-Hoch.	Mean
sport_leavingHome	1,84E-13	1,84E-13	1,84E-13	1,84E-13
sport_type	8,78E-09	8,61E-09	3,48E-09	6,96E-09
sport_passive	1,04E-08	1,00E-08	3,48E-09	7,98E-09
sport_intensity	2,06E-06	1,94E-06	5,15E-07	1,50E-06
social_party	9,17E-06	8,47E-06	1,83E-06	6,49E-06
food_watHaveIEaten	3,09E-05	2,79E-05	5,15E-06	2,13E-05
hobbies_outConfort	5,84E-04	5,16E-04	7,82E-05	3,93E-04
vices_coffee	6,25E-04	5,41E-04	7,82E-05	4,15E-04
social_networks	1,35E-03	1,14E-03	1,50E-04	8,80E-04
hobbies_culture	8,67E-03	7,17E-03	8,67E-04	5,57E-03
sleep_nap	1,26E-02	1,02E-02	1,10E-03	7,96E-03
vices_sex	1,32E-02	1,04E-02	1,10E-03	8,27E-03
alienPerson_sportsTeam	1,48E-02	1,14E-02	1,14E-03	9,10E-03
sleep_timeGetUp	2,99E-02	2,25E-02	2,14E-03	1,82E-02
alienPerson_weekDay	4,09E-02	2,99E-02	2,73E-03	2,45E-02
alienPerson_goodBadNews	6,20E-02	4,41E-02	3,88E-03	3,67E-02

Cuadro 6.6: Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - "*subtopics*" - solamente entradas significativas - 'Encuesta'

El cuadro 6.7 proporciona una **visión intuitiva** de la proporción de cada variable significativa en relación con cada grupo, junto con una indicación de cuáles grupos presentan diferencias significativas según el test de comparación múltiple para cada par de grupos.

Caract/Grupo ("subtopics")	0	1	2
social_networks	1	1	2
social_party	3	3	2
hobbies_culture	3	3	2
hobbies_outConfort	2	1	3
sport_leavingHome	1	3	3
sport_intensity	3	1	2
sport_passive	1	2	1
sport_type	3	1	3
sleep_timeGetUp	3	2	2
sleep_nap	1	1	2
food_watHaveIEaten	2	2	1
alienPerson_weekDay	3	2	3
alienPerson_sportsTeam	2	1	1
vices_sex	2	1	1
vices_coffee	2	3	2

Cuadro 6.7: Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - "subtopics" - 'Encuesta'

A partir de la interpretación de los distintas cuadros y figuras, podemos destacar algunas conclusiones clave. Hasta ahora no se ha creído necesario, pero ahora quizás es recomendable releerse de nuevo el cuadro 3.2 para interpretar correctamente el significado de cada variable del dataset "subtopics".

Los subtópicos a los que el **grupo '0'** les confiere una mayor importancia que el resto de grupos son: las fiestas, las actividades culturales, el día de la semana, la **hora a la que uno se levanta**, la **intensidad deportiva** y el tipo de deporte, que gane su **equipo deportivo** favorito, y el **sexo**.

En cuanto al **grupo '1'**, se caracteriza por concederle una importancia mayor respecto al resto de grupos a los subtópicos: las fiestas, las actividades culturales, el **deporte pasivo**, el salir de casa como actividad deportiva, y el tomar **café**.

Por último, el **grupo '2'** le brinda una mayor importancia a las redes sociales, el salir de la **zona de confort**, el tipo de deporte y el salir de casa como actividad deportiva, hacer una **siesta** y el día de la semana.

6.1.2.4 Encuesta

En este apartado, exponemos los resultados del análisis de diferencias significativas entre los grupos identificados a través del algoritmo K-means en el conjunto de datos “encuesta”. Siguiendo el enfoque anterior, nos centraremos en resaltar los aspectos clave.

El cuadro 6.8 presenta los **p-valores ajustados** para las variables y correcciones en el dataset en cuestión. Los p-valores están ordenados según el p-valor promedio en orden ascendente, resaltando las diferencias significativas entre los grupos.

Variable	Bonf.	H-Bonf.	B-Hoch.	Mean
topics_sport	2,20E-21	2,20E-21	2,20E-21	2,20E-21
aboutYou_sport	2,97E-18	2,93E-18	1,48E-18	2,46E-18
sport_passive	1,35E-15	1,32E-15	4,51E-16	1,04E-15
aboutYou_profitableDay	5,74E-09	5,54E-09	1,43E-09	4,24E-09
sport_type	6,17E-08	5,89E-08	1,23E-08	4,43E-08
aboutYou_couch	5,83E-06	5,51E-06	9,72E-07	4,10E-06
sport_intensity	1,81E-04	1,69E-04	2,59E-05	1,25E-04
vices_coffee	5,27E-04	4,86E-04	6,59E-05	3,60E-04
aboutYou_stress	1,03E-03	9,42E-04	1,15E-04	6,97E-04
vices_sex	4,48E-03	4,03E-03	4,48E-04	2,99E-03
topics_alienPerson	1,08E-02	9,59E-03	9,81E-04	7,12E-03
aboutYou_party	3,75E-02	3,29E-02	3,12E-03	2,45E-02

Cuadro 6.8: Cuadro resumen de diferencias significativas según las 3 correcciones y su promedio - “encuesta” - solamente entradas significativas - ‘Encuesta’

El cuadro 6.9 brinda una **visión intuitiva** de la proporción de cada variable significativa en relación con cada grupo, junto con una indicación de cuáles son los grupos que difieren significativamente según el test de comparación múltiple para cada par de grupos.

Caract/Grupo (“encuesta”)	0	1	2
aboutYou_party	2	1	1
aboutYou_couch	2	2	1
aboutYou_profitableDay	3	1	1
aboutYou_sport	3	1	3
aboutYou_stress	2	2	1
topics_sport	3	1	3
topics_alienPerson	1	2	1
sport_intensity	2	1	2

Caract/Grupo ("encuesta")	0	1	2
sport_passive	2	3	1
sport_type	3	2	3
vices_sex	3	2	2
vices_coffee	2	3	3

Cuadro 6.9: Representación intuitiva de las diferencias significativas y proporciones de cada variable respecto al resto de grupos - "encuesta"

Tengamos en cuenta que la extensión de este dataset comprende cada uno de los anteriores, evaluando tanto el perfil personal, como los tópicos y subtópicos indicados, así que puede ser necesaria de nuevo otra lectura del cuadro 3.2. La interpretación de los cuadros que en este apartado en el que nos encontramos, revela algunas observaciones clave.

En cuanto a perfil personal propio del **grupo '0'** se distingue por necesitar **sentirse productivos** y querer aprovechar el día, gustarles la '**fiesta**' y el deporte como algo necesario. Este grupo le confiere, respecto al grupo '1', una gran importancia al tópico del deporte. En cuanto a los subtópicos, le brindan gran importancia al tipo de deporte y al **sexo**.

Respecto al perfil personal característico del **grupo '1'**, no se distingue demasiado del resto de grupos, solamente se puede decir que no les gusta la fiesta, en cierto sentido no les importa no aprovechar el día y se caracterizan por **no ser deportistas**. Ellos piensan que los factores **ajenos a la persona**, como el tiempo, son factores clave en la influencia de la felicidad. También sienten que hacer **deporte pasivo** y el café pueden influir en gran medida.

Por último, el perfil personal de los integrantes del **grupo '2'** suele ser el siguiente: no les gusta la fiesta, pero **no son sedentarios** (en el sentido de gustarles quedarse confortablemente en casa); no les importa no aprovechar el día y **no se estresan** fácilmente, pero son deportistas. Las variables a las que les asignan una mayor importancia en función de la influencia que ejercen en su bienestar son el tomar café, el tópico del deporte y con él el subtópico de la distinción entre el tipo de deporte practicado (no es lo mismo caminar que correr).

6.1.3 Modelos de aprendizaje automático supervisado

En este subapartado, mostramos los resultados obtenidos mediante la aplicación de diversos modelos de aprendizaje automático supervisado. Nuestro objetivo principal radica en desentrañar la importancia inherente de las características al clasificar a los individuos en los distintos grupos identificados, según los conjuntos de datos "*pProfile*", "*topics*", "*subtopics*" y "*encuesta*". Los modelos seleccionados para este propósito abarcan una variedad de enfoques, incluyendo Random Forest (RF), Gradient Boosting (GB), Extreme Gradient Boosting (XGB) y Support Vector Machine (SVM). Adicionalmente, empleamos el algoritmo Decision Tree para visualizar de manera gráfica el proceso de clasificación de los individuos en los distintos grupos.

El análisis de la importancia de características proporciona información valiosa sobre los factores que más influyen en la pertenencia de un individuo a un grupo en particular. Al comprender qué características son más relevantes para cada grupo, podemos obtener una visión más profunda de los factores que pueden estar contribuyendo, 'significativamente', a las diferencias observadas entre grupos.

Las siguientes secciones presentan, de manera entendedora, los resultados derivados de los diversos modelos aplicados a los cuatro conjuntos de datos. Destacamos aquellas características que emergen como influyentes en la clasificación de individuos en grupos específicos. Adicionalmente, proporcionamos representaciones gráficas de los árboles de decisión, que brindan una visualización intuitiva del proceso de clasificación.

6.1.3.1 Perfil Personal

En esta sección, exponemos los resultados derivados de la aplicación de diversos modelos de aprendizaje automático supervisado al conjunto de datos "*pProfile*". Como se explicó en la subsección 5.1.4, cada uno de estos modelos se empleó con el propósito de evaluar la importancia de las distintas características en la predicción de la pertenencia a los diferentes grupos identificados en el conjunto de datos.

Para obtener una mayor **precisión y robustez**, recordamos que estos modelos no solo se usan para análisis interpretativos, sino que también son algoritmos predictores que trabajan en predecir la pertenencia de individuos específicos a los distintos grupos. Este proceso implica la utilización de folds de validación cruzada y múltiples iteraciones para obtener medidas de precisión, representadas por la exactitud (accuracy), en cada fold e iteración. Estas medidas de precisión se utilizan posteriormente para ponderar la

importancia de características asignada por cada modelo, obteniendo así la importancia de características final del modelo. Las distintas importancias finales, una por cada modelo, se agregan en un proceso de ensamblado, ponderándolas según la exactitud promedio de cada modelo.

El cuadro 6.10 presenta los resultados obtenidos mediante los distintos modelos de aprendizaje automático supervisado aplicados al dataset "*pProfile*". Cada fila representa una característica, identificada según la primera columna. Las últimas cuatro columnas muestran la importancia resultante de cada modelo específico (la etiqueta de la columna indica el modelo aplicado y su exactitud promedio). La columna '*Ensamble*' combina las importancias de las diferentes características, ponderadas por la exactitud promedio de los modelos. El cuadro está ordenada de mayor a menor importancia según la columna de '*Ensamble*'.

Characteristics	RF_95	GB_95	SVM_97	XGB_96	Ensamble
homeCompany_family	0,282	0,295	0,159	0,201	0,234
homeCompany_couple	0,136	0,186	0,001	0,115	0,109
actualStatus_student	0,065	0,061	0,099	0,131	0,089
homeCompany_parents	0,066	0,073	0,097	0,086	0,081
actualStatus_work	0,066	0,052	0,071	0,036	0,056
homeCompany_alone	0,026	0,044	0,028	0,078	0,044
aboutYou_dependent	0,034	0,028	0,025	0,071	0,039
aboutYou_romantic	0,032	0,027	0,072	0,020	0,038
aboutYou_videogames	0,020	0,016	0,018	0,035	0,022
aboutYou_stress	0,016	0,013	0,044	0,011	0,021
aboutYou_read	0,025	0,020	0,020	0,017	0,021
aboutYou_audiovisual	0,014	0,010	0,046	0,011	0,020
aboutYou_couch	0,023	0,019	0,018	0,016	0,019
homeCompany_friends	0,015	0,024	0,033	0,000	0,018
aboutYou_socialNet	0,013	0,012	0,028	0,015	0,017
aboutYou_sport	0,018	0,010	0,026	0,012	0,017
actualStatus_free	0,007	0,008	0,028	0,022	0,016
aboutYou_spontaneous	0,013	0,008	0,029	0,010	0,015
aboutYou_profitableDay	0,017	0,011	0,020	0,013	0,015
aboutYou_smoke	0,012	0,007	0,014	0,020	0,013
aboutYou_sleepWell	0,012	0,011	0,015	0,015	0,013
aboutYou_pets	0,013	0,007	0,024	0,006	0,012
aboutYou_thinkHapp	0,012	0,010	0,019	0,008	0,012
aboutYou_car	0,012	0,009	0,013	0,012	0,011

Cuadro 6.10: Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - "*pProfile*" -'Encuesta'

Los resultados revelan que ciertas características tienen una mayor influencia en la clasificación de los individuos en grupos específicos. Por ejemplo, la característica '*homeCompany_family*' obtiene una importancia ponderada del 0.2336 en el ensamblado, lo que sugiere que desempeña un papel significativo en la predicción de la pertenencia a grupos. Cada modelo también asigna diferentes niveles de importancia a estas características, lo que puede indicar su contribución relativa en función de la metodología del modelo.

Observamos cómo las 5 características más influyentes **coinciden** con las 5 características cuyas diferencias entre grupos son más significativas (véase el cuadro 6.2). El resto de características, en general, también coinciden con el orden de significancia otorgado en el cuadro 6.2, de modo que se puede intuir que los dos métodos de obtención de las importancias relativas de las distintas variables, son similares a la hora de clasificar a los encuestados en los distintos grupos, aunque sus enfoques sean diferentes.

Ahora es todavía más seguro confirmar que el **tipo de convivencia**, entendido como las distintas personas con quien podemos convivir (familia, pareja, solo, etc.), es de gran importancia a la hora de determinar los grupos en "*pProfile*". Del mismo modo que la distinción entre si eres un **estudiante** o estás **trabajando**, y si el encuestado se considera **romántico**, **juega videojuegos** o es **dependiente económico**. Parece que, a pesar de haber eliminado las variables de confusión (entre ellas, la edad del encuestado), el algoritmo sigue tratando de clasificar el conjunto de datos "*pProfile*" según la **edad**.

Es importante destacar que esta evaluación de la importancia de características proporciona una perspectiva valiosa sobre los factores que influyen en la clasificación de los individuos en los grupos identificados previamente, enriqueciendo nuestra comprensión de las características clave que definen la pertenencia a los distintos grupos.

A continuación, explicaremos como se debe interpretar el **árbol de decisión** resultante en la predicción de las etiquetas del conjunto de datos "*pProfile*", que se muestra en la figura 6.7. Cabe mencionar que en el cuerpo de la memoria solamente se incluirá éste árbol de decisión, se podrá encontrar el resto en los anexos 7.3.

En el primer nodo del árbol, la característica '*homeCompany_family*' se utiliza como criterio para dividir los datos. Si su valor para un registro dado es menor o igual a 0.196, el árbol seguirá por la rama izquierda, de lo contrario lo hará por la derecha.

En el mismo nodo observamos también un valor de ‘entropía’, que se trata de una medida de la incertidumbre en la clasificación de los datos en este nodo; a mayores valores de entropía, mayor impureza de los datos en el nodo. Al tratarse del primer nodo, es previsible que tome un valor elevado (1.585).

Encontramos también un valor de ‘samples’, indicando cuántos ejemplos de entrenamiento llegaron a este nodo. Relacionándose con el concepto de entropía, encontramos ‘value’, que muestra cómo se distribuyen las clases en este nodo. Los valores [40.667, 40.667, 40.667] indican que un tercio de los ejemplos pertenecen a cada una de las tres clases (0, 1 y 2), respectivamente.

Finalmente, ‘class’, asigna los ejemplos a una clase particular. En este caso, los ejemplos que llegan a este nodo se asignan a la clase 2 (hasta que se reasignen con mayor precisión en los siguientes nodos). El árbol completo consta de múltiples nodos y reglas que se aplican secuencialmente para clasificar los datos de manera efectiva.

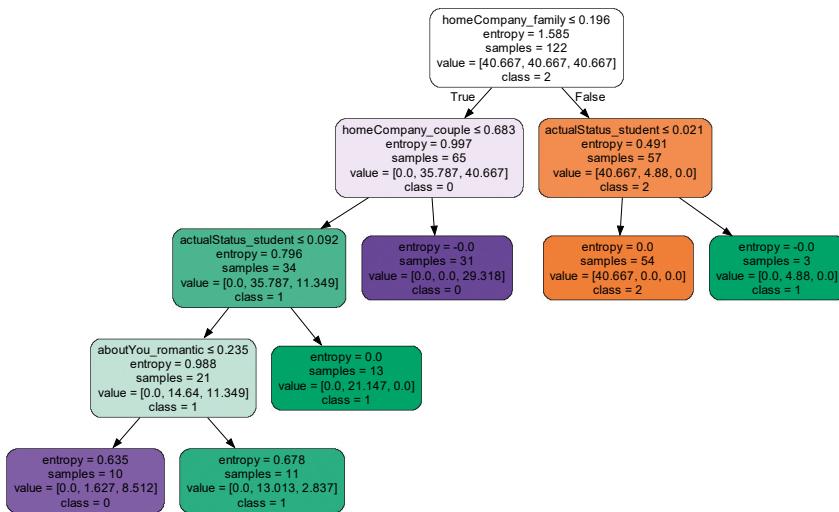


Figura 6.7: Clasificación de grupos según Decission Tree - ”*pProfile*” - ’Encuesta’

A través de esta metodología, hemos explorado cómo los modelos de aprendizaje automático supervisado pueden revelar las características más influyentes en la clasificación de individuos en grupos específicos dentro del dataset ”*pProfile*”. En los siguien-

tes apartados de esta sección, continuamos este análisis de resultados para el resto de conjunto de datos "*topics*", "*subtopics*" y "*encuesta*", de un modo menos detallado, para evitar redundancias.

6.1.3.2 Topics

En este apartado, presentamos los resultados obtenidos al aplicar diversos modelos de aprendizaje automático supervisado al conjunto de datos "*topics*". Siguiendo la metodología explicada en el apartado anterior, los modelos evaluaron la importancia de las características en la clasificación de individuos en los grupos previamente identificados. Cada modelo predictivo fue empleado con folds de validación cruzada y distintas iteraciones para obtener medidas de precisión con las que ponderar las importancias de características.

El cuadro 6.11 resume los resultados de los modelos de aprendizaje automático supervisado aplicados al dataset "*topics*". Cada fila representa una característica, y las columnas muestran la importancia resultante de cada modelo específico, así como el ensamblado ponderado por la exactitud promedio de cada modelo (indicada en la etiqueta de cada columna). Observamos que ciertas características tienen una influencia más destacada en la clasificación de individuos en grupos específicos.

Characteristics	RF_99	GB_99	SVM_99	XGB_99	Ensamble
topics_hobbies	0,4333	0,6342	0	0,5084	0,394
topics_alienPerson	0,2855	0,3419	0,4425	0,4557	0,3814
topics_food	0,081	0,0166	0,0984	0,0109	0,0517
topics_sport	0,0896	0,0018	0,0968	0,0115	0,0499
topics_productiveTime	0,0353	0,0043	0,0954	0,0055	0,0351
topics_vices	0,0224	0,0012	0,0917	0,008	0,0308
topics_sleep	0,0343	0	0,0868	0	0,0303
topics_social	0,0185	0	0,0884	0	0,0267

Cuadro 6.11: Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - "*topics*" - 'Encuesta'

Se destaca que solamente las dos primeras variables '*topics_hobbies*' y '*topics_alienPerson*' recogen una importancia de características de más del 77 %. Repasando el cuadro de análisis de significancia de diferencias entre grupos (6.4), no nos sorprende que a las mismas variables se les confiera un p-value mucho menor que al resto.

Esta coincidencia refuerza la idea de que los **hobbies y factores externos** a la persona son elementos determinantes en la clasificación entre los grupos de individuos en el dataset *"topics"*. La alta importancia asignada a *'topics_hobbies'* indica que las preferencias y actividades de ocio influyen en la agrupación de individuos. Por otro lado, *'topics_alienPerson'* está relacionado con las actitudes y perspectivas hacia elementos externos, lo que también desempeña un papel relevante en la agrupación.

Es relevante destacar una observación intrigante en los resultados. Mientras que los modelos de Random Forest, Gradient Boosting, Extreme Gradient Boosting y la media ponderada del ensamblado coinciden en otorgar la mayor importancia a la característica *'topics_hobbies'*, el modelo de Support Vector Machine **difiere**, asignándole una importancia de cero. Esta discrepancia puede surgir debido a la naturaleza del algoritmo SVM y su enfoque en la búsqueda de un hiperplano óptimo de separación.

El **SVM** se basa en la identificación del hiperplano que maximiza el margen entre las clases, considerando solo los puntos de datos más cercanos a dicho margen (vectores de soporte). Es posible que *'topics_hobbies'* presente una distribución o separación en los datos que no sea adecuada para la forma en que SVM construye su margen de separación. En contraste, los otros modelos, que no están limitados por esta construcción de margen, pueden percibir de manera más destacada la influencia de *'topics_hobbies'* en la agrupación de individuos [Berwick 023].

6.1.3.3 Subtopics

En este apartado, presentamos los resultados obtenidos al aplicar diversos modelos de aprendizaje automático supervisado al conjunto de datos *"subtopics"*. Siguiendo la metodología descrita previamente, cada modelo evaluó la importancia de las características en la clasificación de individuos en los grupos identificados.

El cuadro 6.12 muestra los resultados de los modelos de aprendizaje automático supervisado aplicados al dataset *subtopics*". Las filas representan características y las columnas exhiben la importancia resultante de cada modelo, así como el ensamblado ponderado por la exactitud promedio de los modelos (indicada en la etiqueta de cada columna).

Characteristics	RF_88	GB_90	SVM_89	XGB_90	Ensamble
sport_leavingHome	0,090	0,114	0,074	0,123	0,101
sport_passive	0,063	0,078	0,030	0,058	0,057
sport_type	0,068	0,061	0,046	0,038	0,053
sport_intensity	0,053	0,058	0,048	0,051	0,052
hobbies_outConfort	0,042	0,054	0,025	0,033	0,039
food_watHavelEaten	0,049	0,065	0,012	0,024	0,038
social_party	0,052	0,051	0,008	0,030	0,035
vices_coffee	0,036	0,031	0,032	0,030	0,032
social_networks	0,025	0,030	0,008	0,064	0,032
vices_sex	0,030	0,029	0,040	0,023	0,031
alienPerson_weekDay	0,029	0,027	0,043	0,022	0,030
sleep_timeGetUp	0,027	0,024	0,045	0,014	0,028
productiveTime_compliance	0,008	0,016	0,043	0,043	0,027
productiveTime_housework	0,021	0,013	0,028	0,044	0,026
alienPerson_goodBadNews	0,019	0,016	0,035	0,033	0,026
productiveTime_study	0,029	0,024	0,023	0,021	0,024
hobbies_culture	0,031	0,024	0,014	0,013	0,020
sleep_nap	0,029	0,025	0,008	0,019	0,020
hobbies_relax	0,020	0,019	0,031	0,007	0,019
alienPerson_sportsTeam	0,012	0,014	0,006	0,045	0,019
food_quantity	0,020	0,017	0,023	0,011	0,018
alienPerson_climate	0,019	0,014	0,015	0,020	0,017
sleepScreens	0,018	0,014	0,021	0,014	0,017
hobbies_music	0,022	0,013	0,020	0,008	0,016
social_couple	0,014	0,011	0,014	0,023	0,015
sleep_quantity	0,015	0,011	0,025	0,011	0,015
vices_tobacco	0,015	0,014	0,015	0,018	0,015
hobbies_reading	0,018	0,009	0,029	0,004	0,015
vices_drugs	0,004	0,009	0,009	0,036	0,015
productiveTime_mental	0,015	0,009	0,028	0,006	0,014
hobbies_nature	0,013	0,006	0,029	0,003	0,012
food_diet	0,014	0,010	0,007	0,018	0,012
food_type	0,008	0,006	0,011	0,020	0,011
vices_alcohol	0,017	0,012	0,007	0,007	0,011
productiveTime_work	0,006	0,005	0,013	0,017	0,010

Cuadro 6.12: Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - "subtopics" - 'Encuesta'

Observando el cuadro se identifica claramente que los subtópicos relacionados con el **deporte** son elementos esenciales en la clasificación de los grupos de individuos en el dataset, sugiriendo que las preferencias y comportamientos deportivos desempeñan un papel crucial en la agrupación.

Sin embargo, nos pareció interesante que los subtópicos relacionados con el tiempo libre (aquellos que empiezan con '*hobbies_*') generalmente tienen importancias más bajas, con la excepción de '*hobbies_outConfort*' es decir, 'salir de la zona de confort'.

En cambio, los tópicos relacionados con el **tiempo productivo**, como el sentirse satisfecho, las faenas del hogar y los estudios, se posicionan como variables de importancia relativamente grande.

En consonancia con el análisis de significancia de diferencias entre grupos (véase el cuadro 6.6), notamos que las características destacadas (las posicionadas más arriba en el cuadro y, por ende, de mayor importancia) tienen p-values bajos, lo que **refuerza** su importancia en la clasificación.

Sin embargo, es cierto que se observan pequeñas **discrepancias** comparando los dos cuadros. Existen variables que se les brinda una importancia mayor a la que uno se pudiera imaginar observando sus p-values, relativamente elevados, como '*sport_passive*', '*hobbies_outConfort*', '*alienPerson_weekDay*'; y a la inversa, como '*sport_type*', '*social_party*', '*hobbies_culture*'.

Estos desajustes, en realidad, confirman que cada método de análisis es único, y aporta información desde **distintos enfoques**, otorgando valor añadido al que será el ensamblaje final de importancia de características, cuyos resultados se mostrarán en la subsección 6.1.4.

6.1.3.4 Encuesta

En este apartado, presentamos los resultados obtenidos al aplicar varios modelos de aprendizaje automático supervisado al conjunto de datos "*encuesta*". Siguiendo la metodología explicada anteriormente, cada modelo evaluó la importancia de las características en la clasificación de individuos en los grupos identificados.

El cuadro 6.13 resume los resultados de los modelos de aprendizaje automático supervisado aplicados al dataset "*encuesta*". Cada fila representa una característica, mientras que las columnas muestran la importancia resultante de cada modelo específico, así como el ensamblado ponderado por la exactitud promedio de los modelos (indicada en la etiqueta de cada columna).

Characteristics	RF_89	GB_90	SVM_89	XGB_87	Ensamble
topics_sport	0,113	0,104	0,057	0,111	0,096
aboutYou_sport	0,131	0,063	0,044	0,069	0,077
aboutYou_profitableDay	0,073	0,051	0,044	0,055	0,056
sport_passive	0,090	0,063	0,022	0,035	0,053
aboutYou_couch	0,032	0,036	0,010	0,027	0,026
vices_coffee	0,030	0,026	0,019	0,026	0,025
vices_sex	0,023	0,033	0,019	0,021	0,024
sport_type	0,033	0,045	0,009	0,008	0,024
aboutYou_stress	0,021	0,020	0,019	0,021	0,020
topics_alienPerson	0,020	0,010	0,031	0,015	0,019
aboutYou_party	0,016	0,025	0,013	0,013	0,017
sport_intensity	0,022	0,015	0,017	0,011	0,016
aboutYou_thinkHapp	0,016	0,013	0,021	0,015	0,016
hobbies_reading	0,012	0,009	0,024	0,019	0,016
homeCompany_parents	0,009	0,012	0,016	0,020	0,014
hobbies_outConfort	0,010	0,008	0,028	0,010	0,014
aboutYou_videogames	0,006	0,010	0,016	0,024	0,014
aboutYou_mountain	0,011	0,017	0,011	0,015	0,013
sleep_nap	0,014	0,015	0,009	0,015	0,013
hobbies_nature	0,013	0,019	0,007	0,013	0,013
productiveTime_reading	0,009	0,008	0,007	0,026	0,012
sleep_timeGetUp	0,011	0,008	0,021	0,009	0,012
aboutYou_sciences	0,012	0,010	0,012	0,012	0,012
homeCompany_family	0,013	0,011	0,008	0,013	0,011
productiveTime_housework	0,009	0,009	0,011	0,016	0,011
aboutYou_healthyFood	0,010	0,013	0,012	0,008	0,011

Cuadro 6.13: Importancias de características de cada modelo y su ensamblaje (ponderado según su precisión) - "encuesta" - 'Encuesta'

Notamos que en este caso las distintas características presentan niveles de importancia relativa muy similares. Esto podría deberse a la inclusión de numerosas características en el análisis, lo que podría contribuir a que las diferencias de importancia sean menos pronunciadas. Recordemos que este dataset combina aglomerados "*pProfile*", "*topics*", y "*subtopics*", lo que implica que la clasificación entre grupos tiene en cuenta **todas las características** estudiadas.

Se destaca que las características relacionadas con el **deporte** siguen encabezando el cuadro, independientemente de la porción del dataset de la que provengan. Esto abarca desde la descripción personal del encuestado ('*aboutYou_sport*') hasta el tópico ('*topics_sport*'), y también los subtópicos ('*sport_passive*', '*sport_type*', y '*sport_intensity*'). Esta consistencia resalta la gran importancia del ámbito deportivo en la clasificación de los grupos.

Observamos que otras características del perfil personal también poseen cierta importancia, como la necesidad de sentirse **productivo** (*'aboutYou_profitableDay'*), la tendencia a tomarse un **descanso ocasional** (*'aboutYou_couch'*), la facilidad para experimentar **estrés** (*'aboutYou_stress'*) y el interés por la fiesta (*'aboutYou_party'*).

Adicionalmente, se ha conferido considerable importancia al tópico de **factores ajenos** a la persona (*'topics_alienPerson'*), así como a ciertas características específicas de la porción del dataset *"subtopics"*, como los 'vicios' del café y el sexo.

Al contrastar estos resultados con los análisis de significancia de diferencias entre grupos (consultar el cuadro 6.8), se puede apreciar que las características con importancias más altas también presentan p-values relativamente bajos. En realidad, todas las características que hemos discutido hasta ahora encabezan el cuadro y son las únicas que muestran diferencias significativas. Esto **fortalece** la idea de que estas características influyen de manera significativa en la clasificación de los individuos en el dataset *"encuesta"*.

6.1.4 Ensamblaje

En esta sección, exponemos los resultados del proceso de ensamblado final para cada uno de los datasets bajo análisis: *"pProfile"*, *"topics"*, *"subtopics"* y *"encuesta"*. A lo largo de esta investigación, hemos empleado dos enfoques distintos para evaluar la importancia de las características en la clasificación de individuos en grupos. Uno de ellos involucra el análisis de diferencias significativas entre grupos, mientras que el otro abarca el ensamblado de modelos de aprendizaje automático supervisado.

En esta etapa, convergimos estos **dos enfoques** con el objetivo de obtener una perspectiva más enriquecedora sobre las características más influyentes en la clasificación de grupos. Combinar estas metodologías ofrece la oportunidad de aprovechar las fortalezas de ambas aproximaciones.

El análisis de **diferencias significativas** entre grupos destaca características que muestran disparidades sustanciales entre los grupos y, por ende, podrían desempeñar un rol crucial en la clasificación. A demás, esta metodología se revelará especialmente útil en la segunda fase de nuestro estudio, la relativa al análisis de datos diarios de los hábitos de los 8 distintos participantes (véase el apartado 5.2).

En contraste, el enfoque de **ensamblado de modelos** de aprendizaje automático supervisado brinda una perspectiva más amplia y versátil para evaluar la importancia de las características. A diferencia de la mera identificación de diferencias entre grupos, este enfoque considera la capacidad predictiva de cada característica. Evalúa cómo cada atributo contribuye a mejorar la precisión general de la clasificación, en un contexto más amplio al tener en cuenta varios modelos, lo que se traduce en una convergencia de diferentes perspectivas. Esta convergencia proporciona una visión sólida y robusta de las características clave que definen cada grupo.

Mediante la **fusión** de estas dos perspectivas, aspiramos a obtener un nivel más sólido de confianza en la identificación de las características verdaderamente influyentes. Ya hemos comenzado a observar cómo algunas características pueden destacar tanto por sus diferencias entre grupos como por su contribución a la precisión del modelo de clasificación. Esta estrategia combinada enriquece nuestra comprensión de la relevancia de las características.

A continuación, presentamos los resultados del ensamblado total para cada dataset, destacando las características que emergen como especialmente significativas al considerar ambos enfoques de manera conjunta.

6.1.4.1 Perfil Personal

En este apartado presentamos los resultados del proceso de ensamblado final para el conjunto de datos "*pProfile*". Como mencionamos anteriormente, fusionamos dos enfoques distintos para evaluar la importancia de las características en la clasificación de los grupos. Estos enfoques involucran el análisis de diferencias significativas entre grupos y el ensamblado de modelos de aprendizaje automático supervisado.

El cuadro 6.14 presenta los resultados del ensamble total para el dataset "*pProfile*". En este cuadro, cada fila representa una característica específica, y las columnas reflejan diferentes aspectos de la evaluación de la importancia de la característica en la clasificación de los grupos. Las columnas incluyen '*models*' y '*differences*', que son las importancias asignadas según el ensamblaje de modelos y según el análisis de significancia de diferencias entre grupos, respectivamente. '*ensamble*', que es el resultado del ensamblado de las dos aproximaciones, y la columna '*p_value*', que indica el valor promedio de significancia de diferencias entre grupos. Los datos están ordenados descendientemente según la columna '*ensamble*'.

Characteristics	Models	Differences	Ensamble	p_value
homeCompany_family	0,234	0,060	0,226	4,08E-30
homeCompany_couple	0,109	0,060	0,107	1,03E-18
actualStatus_student	0,089	0,060	0,088	4,74E-14
homeCompany_parents	0,081	0,060	0,080	1,37E-10
actualStatus_work	0,056	0,060	0,056	3,75E-10
homeCompany_alone	0,044	0,060	0,045	1,43E-02
aboutYou_dependent	0,039	0,060	0,040	2,61E-07
aboutYou_romantic	0,038	0,060	0,039	1,16E-02
aboutYou_videogames	0,022	0,060	0,024	1,15E-03
aboutYou_read	0,021	0,054	0,022	1,00E-01
aboutYou_stress	0,021	0,028	0,021	5,44E-01
aboutYou_couch	0,019	0,056	0,020	6,89E-02
aboutYou_audiovisual	0,020	0,014	0,020	7,69E-01
homeCompany_friends	0,018	0,060	0,020	5,35E-05
aboutYou_socialNet	0,017	0,052	0,019	1,32E-01
aboutYou_sport	0,016	0,017	0,017	7,13E-01
aboutYou_profitableDay	0,015	0,028	0,016	5,41E-01
actualStatus_free	0,016	0,003	0,016	9,49E-01
aboutYou_spontaneous	0,015	0,014	0,015	7,69E-01
aboutYou_sleepWell	0,013	0,015	0,013	7,45E-01
aboutYou_smoke	0,013	0,012	0,013	8,04E-01
aboutYou_pets	0,012	0,017	0,013	7,14E-01
aboutYou_thinkHapp	0,012	0,017	0,012	7,13E-01
aboutYou_car	0,011	0,017	0,012	7,14E-01

Cuadro 6.14: Ensamblaje final de la importancia de características - "pProfile" - 'Encuesta'

Recordamos que el ensamblaje de las dos aproximaciones se ha ponderado de la siguiente manera:

```
1 weight_factor = sum([accuracy / sum(accuracies) * accuracy for accuracy
in accuracies])
```

De esta manera, el ensamblaje final se calcula como:

```
1 'importance'* weight_factor + 'importance-p-value' * (1 - weight_factor)
```

Asumiendo que los modelos tienen un buen rendimiento, el weight factor será cercano a su máximo, es decir, 1, asignando una mayor importancia a los resultados del ensamblaje de modelos que al análisis de diferencias significativas entre grupos. Concretamente el weight factor resultante en este caso es 0.956.

Los resultados obtenidos no nos sorprenden, ya que **concuerdan** con los hallazgos previamente identificados mediante los enfoques del análisis de diferencias y ensamblaje de modelos. Es evidente que las características relacionadas con el atributo

'home_Company' encabezan el cuadro, destacando la relevancia que tiene el tipo de convivencia de los encuestados en la clasificación de grupos en el conjunto de datos "pProfile". La categorización en grupos también se ve significativamente influenciada por las respuestas demográficas relacionadas con el estado laboral y educativo de los encuestados. Asimismo, ciertos aspectos del estilo de vida, como el romanticismo, la dependencia económica y el hábito de jugar videojuegos, siguen desempeñando un papel destacado en la clasificación de los individuos en grupos específicos.

6.1.4.2 Topics

En esta sección, presentamos los resultados del proceso de ensamblado final para el conjunto de datos "topics". Como se mencionó previamente, combinamos dos enfoques distintos para evaluar la importancia de las características en la clasificación de los grupos: el análisis de diferencias significativas entre grupos y el ensamblado de modelos de aprendizaje automático supervisado.

El cuadro 6.15 presenta los resultados del ensamblado total para el dataset "topics". Cada fila representa una característica específica, mientras que las columnas reflejan diferentes aspectos de la evaluación de la importancia de cada característica en la clasificación de grupos. Las columnas incluyen 'models' y 'differences', que son las importancias asignadas según el ensamblado de modelos y según el análisis de significancia de diferencias entre grupos, respectivamente. 'ensamble' muestra el resultado de combinar ambas aproximaciones (con un weight factor igual a 0.993), y la columna 'p_value' indica el valor promedio de significancia de diferencias entre grupos. Los datos están ordenados de manera descendente según la columna 'ensamble'.

Characteristics	Models	Differences	Ensamble	p_value
topics_hobbies	0,394	0,168	0,392	4,78E-33
topics_alienPerson	0,381	0,168	0,380	6,57E-22
topics_food	0,052	0,168	0,053	2,82E-05
topics_sport	0,050	0,168	0,051	3,18E-07
topics_productiveTime	0,035	0,119	0,036	2,89E-01
topics_vices	0,031	0,037	0,031	7,82E-01
topics_sleep	0,030	0,052	0,030	6,92E-01
topics_social	0,027	0,123	0,027	2,68E-01

Cuadro 6.15: Ensamblaje final de la importancia de características - "topics" - 'Encuesta'

En contraposición al análisis de significancia, se observa que el proceso de ensamblado otorga una mayor importancia al tópico de la alimentación en comparación con el tema deportivo. Sin embargo, en términos generales, los resultados obtenidos muestran **similitudes** sustanciales entre los distintos enfoques. Resulta interesante destacar que los tópicos relacionados con el tiempo libre y los aspectos ajenos a la persona se mantienen como las características más influyentes, sobresaliendo con claridad en este proceso de ensamblado.

En resumen, el proceso de ensamblado total ha reforzado y validado los hallazgos obtenidos previamente mediante las diferentes aproximaciones de análisis. La combinación de estas perspectivas enriquece nuestra comprensión de las características clave que definen la clasificación en grupos dentro del conjunto de datos "topics".

6.1.4.3 Subtopics

En este apartado, exponemos los resultados del proceso de ensamblado final para el conjunto de datos "subtopics". Siguiendo la metodología previamente descrita, hemos combinado dos enfoques diferentes para evaluar la importancia de las características en la clasificación de los grupos: el análisis de diferencias significativas entre grupos y el ensamblado de modelos de aprendizaje automático supervisado.

El cuadro 6.16 muestra los resultados del ensamblado total para el dataset "subtopics". En cada fila, se representa una característica específica, mientras que las columnas reflejan varios aspectos de la evaluación de la importancia de cada característica en la clasificación de los grupos. Las columnas incluyen 'models' y 'differences', que son las importancias asignadas mediante el ensamblado de modelos y el análisis de significancia de diferencias entre grupos, respectivamente. La columna 'ensamble' refleja el resultado de combinar ambas aproximaciones (con un weight factor igual a 0.895), y la columna 'p_value' indica el valor promedio de significancia de diferencias entre grupos. Los datos se encuentran ordenados en orden descendente según la columna 'ensamble'.

Characteristics	Models	Differences	Ensamble	p_value
sport_leavingHome	0,101	0,038	0,094	1,84E-13
sport_passive	0,057	0,038	0,055	7,98E-09
sport_type	0,053	0,038	0,052	6,96E-09
sport_intensity	0,052	0,038	0,051	1,50E-06
hobbies_outConfort	0,039	0,038	0,039	3,93E-04
food_watHaveEaten	0,038	0,038	0,038	2,13E-05

Characteristics	Models	Differences	Ensamble	p_value
social_party	0,035	0,038	0,036	6,49E-06
vices_coffee	0,032	0,038	0,033	4,15E-04
social_networks	0,032	0,038	0,032	8,80E-04
vices_sex	0,031	0,038	0,031	8,27E-03
alienPerson_weekDay	0,030	0,037	0,031	2,45E-02
sleep_timeGetUp	0,027	0,038	0,029	1,82E-02
productiveTime_housework	0,026	0,036	0,027	6,36E-02
alienPerson_goodBadNews	0,026	0,037	0,027	3,67E-02
productiveTime_compliance	0,027	0,011	0,026	7,13E-01
productiveTime_study	0,024	0,036	0,025	6,16E-02
hobbies_culture	0,020	0,038	0,022	5,57E-03
sleep_nap	0,020	0,038	0,022	7,96E-03
alienPerson_sportsTeam	0,019	0,038	0,021	9,10E-03
hobbies_relax	0,019	0,032	0,020	1,70E-01
sleep_screens	0,017	0,031	0,018	1,98E-01
food_quantity	0,018	0,023	0,018	4,13E-01
alienPerson_climate	0,017	0,023	0,018	3,96E-01
hobbies_music	0,016	0,025	0,017	3,51E-01
hobbies_reading	0,015	0,020	0,015	4,73E-01
sleep_quantity	0,015	0,011	0,015	7,13E-01
social_couple	0,015	0,010	0,015	7,45E-01
vices_tobacco	0,015	0,009	0,015	7,69E-01
vices_drugs	0,015	0,010	0,014	7,32E-01
productiveTime_mental	0,014	0,011	0,014	7,25E-01
food_diet	0,012	0,011	0,012	7,23E-01
hobbies_nature	0,012	0,007	0,012	8,18E-01
food_type	0,011	0,006	0,011	8,40E-01
vices_alcohol	0,011	0,010	0,010	7,39E-01
productiveTime_work	0,010	0,005	0,010	8,69E-01

Cuadro 6.16: Ensamblaje final de la importancia de características - "subtopics" - 'Encuesta'

Al comparar el análisis de significancia y el proceso de ensamblado, se destaca que el ensamblado otorga mayor importancia a 'hobbies_outConfort' en comparación con el análisis de significancia. Sin embargo, en términos generales, los resultados obtenidos a través de ambas metodologías presentan **similitudes** notables. Estos resultados reafirman la relevancia de ciertos aspectos en la agrupación de individuos, y destacan el papel de los temas relacionados con el deporte, las actividades recreativas y la alimentación en la clasificación de grupos en el conjunto de datos "subtopics".

6.1.4.4 Encuesta

En esta sección, presentamos los resultados del proceso de ensamblado final para el conjunto de datos "encuesta". Como ya se ha mencionado, hemos agregado dos enfo-

ques distintos para evaluar la importancia de las características en la clasificación de los grupos: el análisis de diferencias significativas entre grupos y el ensamblado de modelos de aprendizaje automático supervisado.

El cuadro 6.17 exhibe los resultados del ensamblado total para el dataset "encuesta". En cada fila, se representa una característica específica, mientras que las columnas reflejan diversas facetas de la evaluación de la importancia de cada característica en la clasificación de los grupos. Las columnas incluyen 'models' y 'differences', que son las importancias asignadas mediante el ensamblado de modelos y el análisis de significancia de diferencias entre grupos, respectivamente. La columna 'ensamble' muestra el resultado de combinar ambas aproximaciones (con un weight factor igual a 0.888), y la columna 'p_value' indica el valor promedio de significancia de diferencias entre grupos. Los datos se encuentran ordenados en orden descendente según la columna 'ensamble'.

Characteristics	Models	Differences	Ensamble	p_value
topics_sport	0,096	0,028	0,088	2,20E-21
aboutYou_sport	0,077	0,028	0,071	2,46E-18
aboutYou_profitableDay	0,056	0,028	0,053	4,24E-09
sport_passive	0,052	0,028	0,050	1,04E-15
aboutYou_couch	0,026	0,028	0,026	4,10E-06
vices_coffee	0,025	0,028	0,025	3,60E-04
vices_sex	0,024	0,028	0,025	2,99E-03
sport_type	0,024	0,028	0,024	4,43E-08
aboutYou_stress	0,020	0,028	0,021	6,97E-04
topics_alienPerson	0,019	0,028	0,020	7,12E-03
aboutYou_party	0,017	0,027	0,018	2,45E-02
sport_intensity	0,016	0,028	0,017	1,25E-04
aboutYou_thinkHapp	0,016	0,024	0,017	1,28E-01
hobbies_reading	0,016	0,024	0,017	1,28E-01
homeCompany_parents	0,014	0,027	0,015	5,04E-02
aboutYou_videogames	0,014	0,023	0,015	1,89E-01
sleep_nap	0,013	0,023	0,014	1,89E-01
hobbies_nature	0,013	0,022	0,014	2,23E-01
hobbies_outConfort	0,014	0,009	0,013	6,88E-01
sleep_timeGetUp	0,012	0,020	0,013	2,86E-01
aboutYou_mountain	0,013	0,010	0,013	6,33E-01
aboutYou_healthyFood	0,011	0,026	0,012	8,19E-02
aboutYou_sciences	0,012	0,015	0,012	4,60E-01
homeCompany_family	0,011	0,015	0,012	4,62E-01
productiveTime_reading	0,012	0,006	0,012	7,80E-01
alienPerson_goodBadNews	0,010	0,022	0,011	2,18E-01
productiveTime_housework	0,011	0,009	0,011	6,88E-01
sport_leavingHome	0,009	0,023	0,011	1,74E-01

Cuadro 6.17: Ensamblaje final de la importancia de características - "encuesta" - 'Encuesta'

Al realizar una comparación entre el análisis de significancia y el proceso de ensamblado, se resalta que algunas características obtienen una menor importancia a través del ensamblado. Por ejemplo, los temas '*sport_passive*', '*sport_type*', '*aboutYou_stress*' y '*sport_intensity*' muestran una menor relevancia en el proceso de ensamblado. Sin embargo, en términos generales, los resultados obtenidos a través de ambas metodologías presentan notables **similitudes**. Estos hallazgos refuerzan la importancia de ciertos aspectos en la clasificación de individuos en grupos, especialmente subrayando la relevancia de temas relacionados con el deporte, el deseo de aprovechar el día al máximo o descansar cómodamente en el sofá, y ciertos hábitos cotidianos como el consumo de café y la actividad sexual.

6.2 Datos personales

En esta sección se detallan los resultados de acuerdo a los principales pasos del método explicado en la sección 5.2.

6.2.1 Relookyou

En esta sección, presentamos los resultados obtenidos del análisis del conjunto de datos '*Personal*' correspondiente al participante "*Relookyou*". Los resultados se presentan a través de dos cuadros distintos.

El cuadro 6.18 registra la importancia de las características en la predicción del nivel de felicidad según varios modelos, y su ensamblaje ponderado en función del MSE promedio de cada uno de ellos. Este cuadro refleja la influencia de cada variable estudiada en la felicidad del participante. Los resultados también toman en consideración los coeficientes de regresión lineal para determinar si una variable influye positiva o negativamente en la felicidad. En el cuadro, solo se incluyen las variables que tienen una importancia mayor a 0.01 en el ensamble, y el nombre de las columnas indican el modelo usado y el factor de peso para su ensamblaje (según su MSE promedio).

Feature	RFR_0.223	GBR_0.255	XGBR_0.249	SVR_0.272	Ensemble	Linear_coef
FRE_sum	0.054	0.051	0.034	0.047	0.046	-4.410
TIR_body_bat_20h	0.057	0.052	0.041	0.036	0.046	-0.135
FRE_music	0.042	0.045	0.037	0.034	0.039	0.270
SPO_sport	0.036	0.033	0.037	0.044	0.038	0.644
EVA_TIR_tiredness_mix	0.038	0.047	0.026	0.038	0.037	-0.252

Feature	RFR_0.223	GBR_0.255	XGBR_0.249	SVR_0.272	Ensemble	Linear_coef
XX_ACT	0.035	0.028	0.047	0.032	0.036	0.334
SPO_steps	0.046	0.041	0.034	0.020	0.035	-0.208
SPO_active_calories	0.052	0.043	0.033	0.013	0.035	-0.370
VIC_cigars	0.033	0.033	0.030	0.042	0.034	-3.316
TIR_stress	0.032	0.031	0.028	0.039	0.033	0.451
VIC_alcohol	0.028	0.028	0.033	0.028	0.029	-2.457
TIR_sleep_time	0.024	0.028	0.020	0.033	0.027	-0.463
PRF_sum	0.025	0.023	0.018	0.025	0.023	5.450
VIC_coffee	0.014	0.015	0.024	0.029	0.021	-1.707
ALI_tmed	0.026	0.027	0.019	0.005	0.019	14.976
PRF_reading	0.010	0.013	0.017	0.032	0.019	-0.476
ALI_day_of_week_cos	0.013	0.016	0.022	0.023	0.019	-0.086
ALI_tmax	0.023	0.025	0.019	0.005	0.018	-15.029
TIR	0.022	0.026	0.020	0.003	0.017	-0.017
SPO_intens_minut_value	0.021	0.022	0.018	0.010	0.017	0.161
SPO_steps_goal	0.019	0.019	0.014	0.017	0.017	0.188
EAT_healthy	0.016	0.019	0.014	0.016	0.016	-3.568
PRF_study_of_happiness	0.018	0.021	0.018	0.008	0.016	-4.868
TIR_body_bat_mean	0.020	0.023	0.016	0.006	0.016	0.012
ALI_trange	0.019	0.020	0.016	0.010	0.016	2.387
TIR_body_bat_8h	0.017	0.018	0.016	0.013	0.016	-0.438
TIR_body_bat_4h	0.016	0.018	0.015	0.013	0.015	0.417
TIR_body_bat_12h	0.019	0.018	0.017	0.009	0.015	-0.103
VIC	0.020	0.018	0.014	0.009	0.015	18.761
TIR_sleep_score_mix	0.019	0.021	0.017	0.004	0.015	0.083
ALI_week	0.018	0.016	0.018	0.007	0.015	0.310
SPO_card_freq_rest	0.016	0.013	0.017	0.012	0.014	-0.150
FRE_TIR_meditate	0.007	0.007	0.013	0.029	0.014	0.467
TIR_body_bat_16h	0.016	0.018	0.015	0.006	0.013	0.290
ALI_day_of_week_sin	0.009	0.010	0.012	0.020	0.013	0.105
EVA_TIR_sleep_quality	0.010	0.011	0.015	0.015	0.013	0.102
TIR_sleep_score	0.014	0.014	0.013	0.007	0.012	0.086
TIR_body_bat_0h	0.012	0.013	0.013	0.009	0.012	0.114
XX_FRE_shower	0.005	0.008	0.016	0.016	0.011	0.741
XX_EAT_fruit	0.004	0.004	0.012	0.021	0.011	2.968
XX_FRE_VIC_anime	0.004	0.006	0.015	0.015	0.010	-14.279

Cuadro 6.18: Importancia de características según modelos y ensamble y coeficientes lineales - "Relookyoo" - 'Personal'

En el cuadro 6.19, se agrupan las diferentes importancias en según los 8 tópicos correspondientes, en función de su suma y su promedio, y se ordena según este último. Esto permite describir la importancia relativa de los mismos tópicos estudiados en 'Encuesta' a través de los datos del participante.

Prefix	sum_imp	mean_imp
'SPO_': Deporte	0.1558	0.0260
'FRE_': Tiempo libre	0.1177	0.0235
'TIR_': Cansancio	0.2883	0.0206
'VIC_': Vicios o hábitos perjudiciales	0.1346	0.0150
'ALI_': Aspectos externos, ajenos	0.1232	0.0137
'PRF_': Tiempo provechoso	0.0753	0.0108
'ACT_': Actividades sociales	0.0450	0.0090
'EAT_': Alimentación	0.0600	0.0075

Cuadro 6.19: Importancia agrupada según tópicos - "Relookyou" - 'Personal'

A continuación, explicaremos las características que se anticiparían para el participante según su segmentación en la encuesta, al tiempo que contrastaremos estos resultados con los hallazgos del estudio 'Personal'. Esta descripción de características 'previstas' según la segmentación no será abordada en detalle ni con rigor, para una comprensión más precisa y específica del significado de la pertenencia del participante a uno u otro grupo, remitimos al apartado 6.1.2. Dado que al participante se le asignaron las etiquetas de grupo [1, 0, 0, 0] según "*pProfile*", "*topics*", "*subtopics*", y "*encuesta*", respectivamente, discutimos:

En "*pProfile*", el grupo 1 abarca una amplia gama de situaciones de convivencia, aunque tiende a vivir con sus padres. Es posible que este participante sea un estudiante con interés en los videojuegos, dependa económicamente de alguien y se considere romántico. Sin embargo, la segmentación por perfil personal puede proporcionar información limitada en términos de relación con los resultados del estudio 'Personal'. Solo es posible identificar si ciertos aspectos de la descripción personal (como los videojuegos) se reflejan en las actividades diarias del participante, evaluando la calidad del algoritmo de clusterización realizado. En este estudio de "Relookyou", no se refleja ninguna de las actividades mencionadas (consultar el cuadro D.1).

En "*topics*", el grupo 0 considera que las circunstancias externas a la persona no influyen en su felicidad, a diferencia de la alimentación y los hobbies. Además, el deporte tiene una influencia notable.

Al analizar el cuadro 6.19, observamos que el deporte es un tópico con una influencia mayor de lo que sugiere la segmentación en "*topics*". Los otros tópicos también difieren: los aspectos externos a la persona tienen una influencia notoria, mientras que la alimentación tiene menos importancia. Los hobbies sí están correspondientemente asignados en el cuadro, de acuerdo con la previsión en la segmentación del participante. En

resumen, las importancias de características aglomeradas en los tópicos correspondientes, es decir, la importancia de los tópicos, no coincide con lo que podríamos anticipar según la segmentación del participante en “topics”.

En “**subtopics**”, el grupo 0 cree que los subtópicos que más influyen en su felicidad son: fiestas, actividades culturales, el día de la semana, la hora de levantarse, la intensidad deportiva, el tipo de deporte, que su equipo deportivo favorito gane y el sexo.

Teniendo en cuenta que en el cuadro 6.18 solamente se muestran los dos tercios de mayor importancia de entre todas las características, las fiestas ‘XX_ACT’, el día de la semana ‘ALI_day_of_week_cos’, la intensidad deportiva ‘SPO_intens_minut_value’, y el tipo de deporte ‘SPO_sport’ presentan, ciertamente, una notable influencia en la felicidad de “Relookyout”. Las actividades culturales y la hora de levantarse presentan una influencia menor, y ‘que su equipo deportivo favorito gane’ y el sexo ni siquiera se contemplaron (es razonable pensar que, en caso contrario, no tendrían una influencia significativa).

En “**encuesta**”, en cuanto al perfil personal del grupo 0, se distingue por la necesidad de sentirse productivos, gustarles ’la fiesta’ y el deporte. El tópico del deporte tiene una gran influencia en su felicidad, al igual que los subtópicos tipo de deporte y sexo.

Estudiando simultáneamente los dos cuadros 6.18 y 6.19, podemos contrastar la veracidad de estas asunciones con los resultados del participante. En esencia, esta contrastación podría entenderse como un resumen de las anteriores. Confirmamos nuevamente que a las actividades de ’fiesta’ y deportivas se les confiere una notable influencia, y acordamos que el tiempo productivo ‘PRF_sum’ tiene una influencia no despreciable, con una relación directamente proporcional a la felicidad (observemos el carácter positivo del signo en el coeficiente de regresión lineal). El tópico del deporte se alinea con la importancia esperada según su segmentación (véase el cuadro 6.19, del mismo modo que el subtópico del tipo de deporte ‘SPO_sport’).

En **conclusión**, los tópicos del deporte, el tiempo libre y el cansancio son los más influyentes en la felicidad de Relookyout. En cuanto a las actividades, independientemente de si se encuadran dentro de los mismos tópicos o no, destacan: el tiempo total dedicado a actividades de entretenimiento, el aspecto del cansancio y las actividades deportivas, las tres con una relación inversa con la felicidad percibida. En el aspecto de su segmentación en la encuesta, el participante cumple con las asunciones hechas según “subtopics” y “encuesta”, pero no según “pProfile” ni “topics”.

6.2.2 Dolphin

En esta sección, presentamos los resultados derivados del análisis del dataset personal correspondiente al participante “*Dolphin*”. Los resultados se presentan en dos cuadros distintos.

El cuadro 6.20 registra la importancia de las características en la predicción del nivel de felicidad según varios modelos, y su ensamblaje (el nombre de las columnas indican el modelo usado y el factor de peso para su ensamblaje, según su MSE promedio). También incluye los coeficientes de regresión lineal para determinar si una variable influye positiva o negativamente en la felicidad. El cuadro solamente contiene las variables con una importancia mayor a 0.01 en el ensamble.

Feature	RFR_0.247	GBR_0.272	XGBR_0.259	SVR_0.222	Ensemble	Linear_coef
ALI_trange	0.253	0.491	0.344	0.131	0.314	13.104
SPO_sportTime	0.116	0.175	0.101	0.087	0.122	3.666
XX_EAT_healthy	0.077	0.088	0.073	0.063	0.076	0.889
ALI_prec	0.042	0.028	0.065	0.036	0.043	0.771
XX_ACT_Family	0.039	0.021	0.052	0.057	0.041	3.577
EVA_tiredness	0.061	0.025	0.029	0.041	0.039	0.146
ALI_tmax	0.049	0.038	0.030	0.024	0.035	-59.654
XX_ACT	0.020	0.023	0.045	0.045	0.033	0.185
TIR_sleep_time	0.050	0.004	0.000	0.072	0.029	-0.059
ALI_tmmed	0.037	0.025	0.033	0.017	0.028	56.239
FRE_sum	0.020	0.008	0.056	0.025	0.027	-4.137
PRF_sum	0.039	0.014	0.033	0.021	0.027	-1.752
ALI_week	0.034	0.011	0.032	0.028	0.026	1.988
ALI_day_of_week_cos	0.030	0.016	0.021	0.030	0.024	-2.044
XX_EAT_abundant	0.021	0.008	0.036	0.028	0.023	-0.040
SPO_Alberto	0.028	0.004	0.002	0.046	0.019	-3.440
ALI_day_of_week_sin	0.028	0.011	0.018	0.017	0.018	0.446
SPO_Yoga	0.000	0.000	0.000	0.076	0.017	1.197
ALI_month_cos	0.014	0.004	0.015	0.012	0.011	4.079
SPO_Running	0.002	0.000	0.000	0.048	0.011	0.703
PRF_Work	0.009	0.002	0.013	0.018	0.010	3.070

Cuadro 6.20: Importancia de características según modelos y ensamble y coeficientes lineales - “*Dolphin*” - ‘Personal’

En el cuadro 6.21, se agrupan las diferentes importancias según los 8 tópicos correspondientes, en función de su suma y su promedio, ordenada según este último, describiendo la importancia relativa de los mismos tópicos estudiados en ‘Encuesta’ a través de los datos del participante.

Prefix	sum_imp	mean_imp
'ALI_': Aspectos externos, ajenos	0.5082	0.0565
'SPO_': Deporte	0.1686	0.0421
'EAT_': Alimentación	0.1079	0.0360
'TIR_': Cansancio	0.0678	0.0339
'FRE_': Tiempo libre	0.0272	0.0272
'ACT_': Actividades sociales	0.0788	0.0263
'PRF_': Tiempo provechoso	0.0367	0.0183
'VIC_': Vicios o hábitos perjudiciales	0.0049	0.0049

Cuadro 6.21: Importancia agrupada según tópicos - "Dolphin" - 'Personal'

A continuación, explicaremos de manera genérica las características que se anticipan para el participante según su segmentación en la encuesta (véase el apartado (6.1.2) para una explicación más rigurosa), y su contrate según el estudio 'Personal'. Dado que al participante se le asignaron las etiquetas de grupo [0, 2, 0, 0] según "*pProfile*", "*topics*", "*subtopics*", y "*encuesta*", respectivamente, el perfil del participante debería caracterizarse de la siguiente manera:

En "*pProfile*", el Grupo 0 vive exclusivamente en un entorno familiar. Es posible que este participante tenga empleo y no dependa económicamente de nadie, no esté interesado en los videojuegos y no se considere tan romántico como los del Grupo 1. En este caso, la ausencia de interés por los videojuegos se refleja en el análisis de 'Dolphin', ya que no se registran actividades relacionadas con este aspecto (ver el cuadro D.2).

En "*topics*", el grupo 2 es el que mayor importancia le atribuye al deporte, y el que menos a los hobbies. También atribuye mucha importancia a la alimentación, y los aspectos ajenos a la persona tienen una influencia notable.

Observamos que, en este caso, las importancias de características aglomeradas en los correspondientes tópicos sí parecen reafirmar su anticipación según la segmentación del participante en "*topics*". Los aspectos ajenos, el deporte y la alimentación se posicionan encabezando el cuadro, con una gran importancia de características promedio. Cabe destacar, por eso, que al tiempo libre se le confiere una importancia mayor a la que cabría esperar según la segmentación del participante (véase el cuadro 6.21).

En "*subtopics*", el grupo 0 considera que los subtópicos que más influyen en su felicidad son: fiestas, actividades culturales, el día de la semana, la hora de levantarse, la intensidad deportiva, el tipo de deporte, que su equipo deportivo favorito gane y el sexo.

La previsión de la influencia en las actividades no se corresponde con las importancias brindadas en el estudio de “*Dolphin*”, a excepción del deporte. En este caso, de nuevo, apenas se muestran dos terceras partes de las características; aun teniendo esto en cuenta, quizás no podamos confirmar que el día de la semana ‘*ALI_day_of_week_cos*’ tenga demasiada importancia. La intensidad deportiva y el tipo de deporte no se contemplaron, pero sí que es cierto que el tiempo de deporte ‘*SPO_sportTime*’ juega un papel de primordial importancia. La ‘fiesta’ tiene un papel negligible, y las actividades culturales, la hora de levantarse, ‘que su equipo deportivo favorito’ y el sexo, ni siquiera se tuvieron en cuenta (véase el cuadro 6.20).

En “*encuesta*”, en cuanto al perfil personal del grupo 0, se distingue por la necesidad de sentirse productivos, gustarles las fiestas y el deporte. El tópico del deporte tiene una gran influencia en su felicidad, al igual que los subtópicos tipo de deporte y sexo.

En este apartado, la contrastación entre resultados requiere del estudio simultáneo de los dos cuadros 6.20 y 6.21. En el aspecto del perfil personal, si bien es cierto que el tiempo productivo ‘*PRF_sum*’ tiene una influencia no despreciable, este presenta una relación inversamente proporcional a la felicidad; en cuanto al deporte, se confirma que se trata de una característica presente e influyente. El tópico del deporte se alinea con la importancia esperada según su segmentación, contrariamente a los subtópicos descritos.

En **conclusión**, los tópicos de aspectos externos, el deporte y la alimentación son los más influyentes en la felicidad de *Dolphin*. En cuanto a las actividades, se comprendan dentro de los mismos tópicos o no, destacamos: las condiciones climáticas (rango de temperaturas, precipitación y temperatura máxima; directa, directa e inversamente relacionadas respectivamente); el tiempo dedicado al deporte y la alimentación sana. En el aspecto de su segmentación en la encuesta, el participante cumple con las asunciones hechas según “*pProfile*”, “*topics*” y “*encuesta*”, pero no para “*subtopics*”.

6.2.3 Juju

En esta sección, presentamos los resultados derivados del análisis del dataset personal correspondiente al participante “*Juju*”. Los resultados se presentan en dos cuadros distintos:

El cuadro 6.22 registra la importancia de las características en la predicción del nivel de felicidad según varios modelos, y su ensamblaje (el nombre de las columnas indican el modelo usado y el factor de peso para su ensamblaje, según su MSE promedio). También incluye los coeficientes de regresión lineal para determinar si una variable influye positiva o negativamente en la felicidad. El cuadro solamente contiene las variables con una importancia mayor a 0.01 en el ensamble.

Feature	RFR_0.253	GBR_0.25	XGBR_0.258	SVR_0.239	Ensemble	Linear_coef
EVA_tiredness	0.205	0.178	0.067	0.095	0.137	-0.388
VIC	0.112	0.112	0.051	0.055	0.083	-0.101
XX_ACT_Social	0.044	0.084	0.056	0.043	0.057	0.213
XX_ACT	0.044	0.083	0.054	0.043	0.056	0.213
EAT_healthy	0.058	0.066	0.041	0.052	0.054	-1.354
PRF_sum	0.065	0.052	0.046	0.020	0.046	47.520
ALI_day_of_week_sin	0.053	0.056	0.046	0.022	0.045	-0.211
FRE_sum	0.065	0.049	0.044	0.016	0.044	0.693
TIR_sleep_time	0.028	0.017	0.026	0.069	0.034	0.069
ALI_trange	0.046	0.032	0.036	0.023	0.034	1.677
TIR_time_get_up	0.030	0.024	0.040	0.038	0.033	-0.223
PRF_reading	0.033	0.026	0.029	0.029	0.029	-20.047
ALI_tmed	0.020	0.033	0.027	0.029	0.027	10.511
ALI_tmax	0.022	0.030	0.028	0.029	0.027	-10.245
PRF_study	0.011	0.015	0.035	0.039	0.025	-34.277
ALI_day_of_week_cos	0.032	0.018	0.023	0.020	0.023	0.027
FRE_VIC_SocialNetworks	0.021	0.022	0.029	0.021	0.023	-0.539
FRE_VIC_audiovisual_content	0.017	0.017	0.026	0.022	0.020	-0.466
ALI_week	0.013	0.017	0.026	0.024	0.020	0.454
XX_EAT_eating_out	0.010	0.012	0.024	0.031	0.019	0.229
VIC_coffee	0.010	0.011	0.020	0.029	0.017	0.197
ALI_month_sin	0.006	0.011	0.030	0.016	0.016	1.489
XX_EAT_legumes	0.005	0.002	0.025	0.022	0.013	0.983
ALI_prec	0.005	0.003	0.024	0.012	0.011	0.101
XX_EAT_carbohydrates	0.005	0.004	0.017	0.016	0.011	0.152

Cuadro 6.22: Importancia de características según modelos y ensamble y coeficientes lineales - "Juju" - 'Personal'

En el cuadro 6.23, se agrupan las diferentes importancias según los 8 tópicos correspondientes, en función de su suma y su promedio, ordenada según este último, describiendo la importancia relativa de los mismos tópicos estudiados en 'Encuesta' a través de los datos del participante.

Prefix	sum_imp	mean_imp
'TIR_': Cansancio	0.2039	0.0680
'ACT_': Actividades sociales	0.1220	0.0407
'FRE_': Tiempo libre	0.0537	0.0268
'VIC_': Vicios o hábitos perjudiciales	0.1589	0.0265
'ALI_': Aspectos externos, ajenos	0.2127	0.0236
'PRF_': Tiempo provechoso	0.1213	0.0202
'EAT_': Alimentación	0.1230	0.0176
'SPO_': Deporte	0.0045	0.0023

Cuadro 6.23: Importancia agrupada según tópicos - "Juju" - 'Personal'

A continuación, explicaremos de manera genérica las características que se anticipan para el participante según su segmentación en la encuesta (véase el apartado 6.1.2 para una explicación más rigurosa), y su contrate según el estudio 'Personal'. Dado que al participante se le asignaron las etiquetas de grupo [1, 2, 0, 0]según "*pProfile*", "*topics*", "*subtopics*", y "*encuesta*", respectivamente, el perfil del participante debería caracterizarse de la siguiente manera:

En "*pProfile*", el grupo 1 presenta el rango más amplio de convivencia, aunque tiene a vivir con sus padres. Probablemente se trate de un estudiante, con intereses en los videojuegos, dependa económicamente de alguien y se considere romántico. Ninguna de las actividades mencionadas se refleja en las actividades diarias registradas por el participante (véase el cuadro D.3).

En "*topics*", el grupo 2 es el que mayor importancia le atribuye al deporte, y el que menos a los hobbies. También atribuye mucha importancia a la alimentación, y los aspectos ajenos a la persona tienen una influencia notable.

En este caso, las importancias de características aglomeradas no se corresponderse en absoluto con la previsión según su segmentación. El deporte juega un papel negligible en la predicción de la felicidad, mientras que el tiempo libre tiene una influencia muy notoria. La alimentación se posiciona casi en la base del cuadro, indicando su baja influencia, y los aspectos ajenos tienen, quizás, una importancia menor a la esperada (véase el cuadro 6.23).

En "*subtopics*", el grupo 0 cree que los subtópicos que más influyen en su felicidad son: fiestas, actividades culturales, el día de la semana, la hora de levantarse, la intensidad deportiva, el tipo de deporte, que su equipo deportivo favorito gane y el sexo.

La variable que captura las actividades con la fiesta “XX_ACT_Party” tiene una importancia negligible; sin embargo, otras características relacionadas “XX_ACT_Social” y “XX_ACT”, tienen una gran importancia. El día de la semana tiene una gran influencia según ‘ALI_day_of_week_sin’, al igual que la hora de levantarse ‘TIR_time_get_up’. Hasta aquí se cumple con las influencias esperadas según su segmentación, pero las variables que capturan información relativa al deporte tienen una influencia negligible, y las actividades culturales, ‘que su equipo deportivo favorito gane’ y el sexo, ni siquiera se contemplaron (véase el cuadro 6.22).

En “*encuesta*”, en cuanto al perfil personal del grupo 0, se distingue por la necesidad de sentirse productivos, gustarles las fiestas y el deporte. El tópico del deporte tiene una gran influencia en su felicidad, al igual que los subtópicos tipo de deporte y sexo.

En este apartado, la contrastación entre resultados requiere del estudio simultáneo de los dos cuadros 6.22 y 6.23. En el aspecto del perfil personal no se reflejan sus supuestas preferencias en los datos estudiados: si bien es cierto que el aspecto productivo es influyente, este se relaciona inversamente con la felicidad. En cuanto a los tópicos y subtópicos descritos, tampoco coinciden con los resultados de “Juju”.

En **conclusión**, los tópicos del cansancio, las actividades sociales y el tiempo libre son los más influyentes en la felicidad de Juju. En cuanto a las actividades, se comprendan dentro de los mismos tópicos o no, destacamos: el cansancio y el índice de vicios consumidos, inversamente proporcionales, y las actividades sociales. En el aspecto de su segmentación en la encuesta, el participante cumple con las asunciones hechas según “subtopics” pero no para “*pProfile*”, “*topics*” ni “*encuesta*”.

6.2.4 Pato

En esta sección, presentamos los resultados derivados del análisis del dataset personal correspondiente al participante “Pato”. Los resultados se presentan en dos cuadros distintos:

El cuadro 6.24 registra la importancia de las características en la predicción del nivel de felicidad según varios modelos, y su ensamblaje (el nombre de las columnas indican el modelo usado y el factor de peso para su ensamblaje, según su MSE promedio). También incluye los coeficientes de regresión lineal para determinar si una variable influye positiva o negativamente en la felicidad. El cuadro solamente contiene las variables con una importancia mayor a 0.01 en el ensamble.

Feature	RFR_0.255	GBR_0.254	XGBR_0.236	SVR_0.255	Ensemble	Linear_coef
EVA_stress	0.579	0.556	0.152	0.157	0.365	-0.347
ALI_day_of_week_cos	0.173	0.233	0.103	0.089	0.150	-0.476
ALI_week	0.035	0.034	0.055	0.053	0.044	-0.036
ALI_tmax	0.037	0.054	0.054	0.016	0.040	17.672
EAT_healthy	0.022	0.020	0.043	0.034	0.029	-0.569
ALI_tmmed	0.019	0.019	0.052	0.015	0.026	-18.062
XX_ACT_Family	0.010	0.011	0.053	0.030	0.025	-0.092
ALI_day_of_week_sin	0.017	0.011	0.054	0.018	0.024	0.041
ALI_month_sin	0.005	0.007	0.046	0.029	0.022	-0.543
ALI_trange	0.012	0.003	0.046	0.019	0.019	-3.812
XX_EAT_eating_out	0.011	0.006	0.029	0.033	0.019	-0.137
FRE_sum	0.006	0.002	0.029	0.032	0.017	1.203
VIC_alcohol	0.005	0.002	0.028	0.034	0.017	0.879
PRF_sum	0.008	0.004	0.035	0.017	0.016	-0.304
EVA_tiredness	0.010	0.011	0.025	0.016	0.015	-0.015
VIC	0.009	0.002	0.035	0.014	0.015	-2.728
TIR_time_get_up	0.005	0.004	0.019	0.030	0.014	0.028
XX_EAT_VIC_meat	0.004	0.003	0.014	0.027	0.012	1.995
TIR_sleep_time	0.005	0.008	0.015	0.020	0.012	-0.036
FRE_TIR_siesta	0.006	0.003	0.016	0.021	0.011	-0.661
ALI_prec	0.003	0.000	0.011	0.030	0.011	0.065
EVA_TIR_sleep_quality	0.004	0.002	0.003	0.033	0.011	-0.125

Cuadro 6.24: Importancia de características según modelos y ensamble y coeficientes lineales - "Pato" - 'Personal'

En el cuadro 6.25, se agrupan las diferentes importancias según los 8 tópicos correspondientes, en función de su suma y su promedio, ordenada según este último, describiendo la importancia relativa de los mismos tópicos estudiados en 'Encuesta' a través de los datos del participante.

Prefix	sum_imp	mean_imp
'ALI_': Aspectos externos, ajenos	0.7120	0.0647
'FRE_': Tiempo libre	0.0285	0.0142
'TIR_': Cansancio	0.0519	0.0130
'ACT_': Actividades sociales	0.0375	0.0125
'EAT_': Alimentación	0.0677	0.0113
'VIC_': Vicios o hábitos perjudiciales	0.0649	0.0108
'PRF_': Tiempo provechoso	0.0204	0.0102
'SPO_': Deporte	0.0172	0.0057

Cuadro 6.25: Importancia agrupada según tópicos - "Pato" - 'Personal'

A continuación, explicaremos de manera genérica las características que se anticipan para el participante según su segmentación en la encuesta (véase el apartado 6.1.2

para una explicación más rigurosa), y su contraste según el estudio ‘Personal’. Dado que al participante se le asignaron las etiquetas de grupo [0, 2, 2, 2] según “*pProfile*”, “*topics*”, “*subtopics*”, y “*encuesta*”, respectivamente, el perfil del participante debería caracterizarse de la siguiente manera:

En “*pProfile*”, el grupo 0 solamente vive en familia. Probablemente trabaje y no tenga intereses en los videojuegos, no dependa económicamente de alguien y no se considere tan romántico como los del grupo 1. De nuevo, el no tener intereses por los videojuegos se ve reflejado en el análisis de “*Pato*”, por no tener registrada ninguna actividad relacionada con este aspecto (véase el cuadro D.4).

En “*topics*”, el grupo 2 es el que mayor importancia le atribuye al deporte, y el que menos a los hobbies. También atribuye mucha importancia a la alimentación, y los aspectos ajenos a la persona tienen una influencia notable.

En este caso, las importancias de características aglomeradas no parecen corresponderse demasiado con la previsión según su segmentación. Si bien es cierto que los aspectos ajenos presentan una influencia muy notoria (quizás demasiado), el deporte presenta una influencia negligible, al contrario que el tiempo libre, que se le confiere la segunda mayor de las importancias (véase el cuadro 6.25).

En “*subtopics*”, el grupo 2 le brinda una mayor importancia a las redes sociales, el salir de la zona de confort, el tipo de deporte y el salir de casa como actividad deportiva, el hacer una siesta y el día de la semana.

A pesar que el uso de redes sociales, y el salir de la zona de confort no fueron contempladas por el participante, y que las características relativas al deporte comprenden una importancia negligible; el día de la semana y la actividad de siesta, se alinean con las previsiones (véase el cuadro 6.24).

En “*encuesta*”, el perfil personal del grupo ‘2’ suele ser el siguiente: no les gusta la fiesta, pero no son sedentarios, no les importa no aprovechar el día y no se estresan fácilmente, pero son deportistas. Las variables a las que les asignan una mayor importancia son el tópico del deporte y los subtópicos de tomar café y la distinción entre el tipo de deporte practicado.

En este apartado, la contrastación entre resultados requiere del estudio simultáneo de los dos cuadros 6.24 y 6.25. En el aspecto del perfil personal no se reflejan sus supuestas preferencias en los datos estudiados: no se contemplaron las actividades de fiesta y sedentarismo, pero el tiempo de provecho tiene una notoria influencia, al contrario que el deporte. Tampoco coincide en cuanto a tópicos ni subtópicos.

En **conclusión**, los tópicos de circunstancias externas, tiempo libre y cansancio son los más influyentes en la felicidad de Pato. En cuanto a las actividades, se comprendan dentro de los mismos tópicos o no, destacamos: el nivel de estrés, circunstancias ajenas como el día de la semana, el paso del tiempo y la temperatura (máxima y media) y el comer saludable (todas inversas exceptuando la temperatura máxima). En el aspecto de su segmentación en la encuesta, el participante cumple con las asunciones hechas según “*pProfile*” pero no para “*topics*”, “*subtopics*” ni “*encuesta*”.

6.2.5 Ajara

En esta sección, presentamos los resultados derivados del análisis del dataset personal correspondiente al participante “Ajara”. Los resultados se presentan en dos cuadros distintos:

El cuadro 6.26 registra la importancia de las características en la predicción del nivel de felicidad según varios modelos, y su ensamblaje (el nombre de las columnas indican el modelo usado y el factor de peso para su ensamblaje, según su MSE promedio). También incluye los coeficientes de regresión lineal para determinar si una variable influye positiva o negativamente en la felicidad. El cuadro solamente contiene las variables con una importancia mayor a 0.01 en el ensamble.

Feature	RFR_0.253	GBR_0.255	XGBR_0.248	SVR_0.243	Ensemble	Linear_coef
EVA_tiredness	0.084	0.099	0.078	0.065	0.082	-0.116
VIC	0.092	0.092	0.077	0.044	0.077	0.691
EVA_sleep_quality	0.068	0.063	0.072	0.089	0.073	0.040
TIR_sleep_time	0.072	0.072	0.075	0.037	0.064	0.175
ALI_tmmed	0.084	0.076	0.063	0.028	0.063	-6.431
ALI_tmax	0.076	0.069	0.046	0.030	0.056	6.656
FRE_VIC_SocialNetworks	0.043	0.033	0.060	0.073	0.052	-0.465
PRF_sum	0.048	0.054	0.057	0.022	0.045	11.227
TIR_time_get_up	0.049	0.052	0.048	0.027	0.044	-0.250
ALI_week	0.041	0.038	0.037	0.031	0.037	-0.155

Feature	RFR_0.253	GBR_0.255	XGBR_0.248	SVR_0.243	Ensemble	Linear_coef
FRE_sum	0.039	0.040	0.034	0.031	0.036	0.282
PRF_houseWork	0.028	0.028	0.028	0.048	0.033	-2.632
ALI_trange	0.025	0.031	0.030	0.018	0.026	-1.561
EAT_healthy	0.027	0.030	0.025	0.015	0.024	-1.213
XX_ACT_granddaughter	0.014	0.015	0.034	0.033	0.024	0.499
FRE_VIC_phoneGames	0.019	0.016	0.030	0.026	0.023	-0.621
ALI_prec	0.022	0.024	0.022	0.023	0.023	0.071
EVA_OTH_personalCare	0.018	0.019	0.012	0.033	0.020	0.067
ALI_day_of_week_sin	0.019	0.021	0.019	0.018	0.019	-0.068
FRE_VIC_audiovisual_content	0.013	0.012	0.017	0.021	0.016	-0.515
ALI_month_sin	0.017	0.011	0.020	0.014	0.015	0.166
XX_EAT_eating_out	0.009	0.009	0.012	0.024	0.013	-0.619
XX_EAT_legumes	0.011	0.012	0.009	0.017	0.012	1.504
PRF_Work	0.004	0.004	0.009	0.028	0.011	-11.328
XX_EAT_VIC_meat	0.007	0.009	0.007	0.022	0.011	-0.577
XX_ACT_Family	0.009	0.009	0.013	0.012	0.011	-0.007
ALI_day_of_week_cos	0.010	0.011	0.008	0.014	0.011	-0.015

Cuadro 6.26: Importancia de características según modelos y ensamble y coeficientes lineales - "Ajara" - 'Personal'

En el cuadro 6.27, se agrupan las diferentes importancias según los 8 tópicos correspondientes, en función de su suma y su promedio, ordenada según este último, describiendo la importancia relativa de los mismos tópicos estudiados en 'Encuesta' a través de los datos del participante.

Prefix	sum_imp	mean_imp
'TIR_': Cansancio	0.2631	0.0658
'VIC_': Vicios o hábitos perjudiciales	0.1836	0.0306
'PRF_': Tiempo provechoso	0.0894	0.0298
'ALI_': Aspectos externos, ajenos	0.2634	0.0239
'FRE_': Tiempo libre	0.0625	0.0208
'ACT_': Actividades sociales	0.0513	0.0103
'EAT_': Alimentación	0.0793	0.0099
'SPO_': Deporte	0.0074	0.0074

Cuadro 6.27: Importancia agrupada según tópicos - "Ajara" - 'Personal'

A continuación, explicaremos de manera genérica las características que se anticipan para el participante según su segmentación en la encuesta (véase el apartado 6.1.2 para una explicación más rigurosa), y su contrate según el estudio 'Personal'. Dado que al participante se le asignaron las etiquetas de grupo [0, 2, 1, 0] según "pProfile", "topics", "subtopics", y "encuesta", respectivamente, el perfil del participante debería caracterizarse de la siguiente manera:

En “*pProfile*”, el grupo 0 solamente vive en familia. Probablemente trabaje y no tenga intereses en los videojuegos, no dependa económicamente de alguien y no se considere tan romántico como los del grupo 1. De nuevo, el no tener intereses por los videojuegos se ve reflejado en el análisis de “Ajara”, por no tener registrada ninguna actividad relacionada con este aspecto (véase el cuadro D.5).

En “*topics*”, el grupo 2 es el que mayor importancia le atribuye al deporte, y el que menos a los hobbies. También atribuye mucha importancia a la alimentación, y los aspectos ajenos a la persona tienen una influencia notable.

El deporte, seguidamente de la alimentación, son los tópicos que menor importancia reciben, contrariamente a lo esperado. Al tiempo libre se le asigna una notable influencia con lo que, a pesar que los aspectos ajenos tienen una influencia adecuada, las importancias de características aglomeradas no parecen corresponderse con la previsión según su segmentación (véase el cuadro 6.27).

En “*subtopics*”, el grupo 1 se caracteriza por concederle una importancia mayor a las fiestas, las actividades culturales, el deporte pasivo, el salir de casa como actividad deportiva, y el tomar café. En este caso no podemos contrastar los resultados, ninguna de las actividades comentadas fue contemplada por el participante (véase el cuadro 6.26).

En “*encuesta*”, en cuanto al perfil personal del grupo 0, se distingue por la necesidad de sentirse productivos, gustarles las fiestas y el deporte. El tópico del deporte tiene una gran influencia en su felicidad, al igual que los subtópicos tipo de deporte y sexo.

En este apartado, la contrastación entre resultados requiere del estudio simultáneo de los dos cuadros 6.26 y 6.27. En el aspecto del perfil personal no se reflejan sus supuestas preferencias en los datos estudiados: no se contemplaron actividades relativas a la fiesta, y la importancia del deporte es negligible; sin embargo, el tiempo de provecho es ciertamente influyente. En cuanto a tópicos y subtópicos, los resultados tampoco son coincidentes.

En **conclusión**, los tópicos de cansancio, hábitos perjudiciales y tiempo provechoso son los más influyentes en la felicidad de Ajara. En cuanto a las actividades, se comprendan dentro de los mismos tópicos o no, destacamos: la evaluación del cansancio

(inversa), el índice de vicios consumidos, la calidad y tiempo de sueño, y circunstancias externas como la temperatura. En el aspecto de su segmentación en la encuesta, el participante cumple con las asunciones hechas según “*pProfile*” pero no para “*topics*”, “*subtopics*” ni “*encuesta*”.

6.2.6 Charlie

En esta sección, presentamos los resultados derivados del análisis del dataset personal correspondiente al participante “*Charlie*”. Los resultados se presentan en dos cuadros distintos:

El cuadro 6.28 registra la importancia de las características en la predicción del nivel de felicidad según varios modelos, y su ensamblaje (el nombre de las columnas indican el modelo usado y el factor de peso para su ensamblaje, según su MSE promedio). También incluye los coeficientes de regresión lineal para determinar si una variable influye positiva o negativamente en la felicidad. El cuadro solamente contiene las variables con una importancia mayor a 0.01 en el ensamble.

Feature	RFR_0.269	GBR_0.258	XGBR_0.264	SVR_0.209	Ensemble	Linear_coef
EVA_stress	0.078	0.083	0.045	0.136	0.083	-0.323
XX_ACT_Couple	0.073	0.083	0.108	0.052	0.080	2.439
XX_ACT	0.068	0.066	0.106	0.074	0.079	-0.794
ALI_tmax	0.092	0.088	0.040	0.062	0.071	-1.792
ALI_tmed	0.095	0.086	0.041	0.060	0.071	0.531
VIC	0.058	0.059	0.046	0.030	0.049	1.376
VIC_coffee	0.047	0.051	0.044	0.047	0.047	-0.167
ALI_prec	0.050	0.054	0.033	0.035	0.043	0.213
ALI_week	0.058	0.053	0.040	0.016	0.043	1.381
ALI_trange	0.052	0.051	0.028	0.030	0.041	0.309
EVA_TIR_tiredness	0.042	0.046	0.026	0.048	0.040	-0.180
TIR_sleep_time	0.047	0.053	0.029	0.021	0.039	0.123
PRF_sum	0.034	0.037	0.031	0.026	0.032	1.027
FRE_sum	0.031	0.031	0.021	0.033	0.029	0.847
ALI_day_of_week_sin	0.026	0.025	0.022	0.044	0.029	0.014
XX_VIC_alcohol	0.014	0.016	0.050	0.016	0.024	-1.659
ALI_month_sin	0.021	0.018	0.036	0.018	0.024	0.574
PRF_uni	0.010	0.010	0.019	0.044	0.020	-0.654
SPO	0.015	0.013	0.021	0.030	0.019	0.177
XX_VIC	0.012	0.012	0.031	0.019	0.018	-2.703
XX_EAT_eating_out	0.015	0.011	0.025	0.016	0.017	0.666
PRF_personalProjects	0.010	0.013	0.021	0.022	0.016	-0.628
XX_ACT_Friends	0.007	0.007	0.026	0.025	0.016	0.953
XX_ACT_Family	0.006	0.004	0.023	0.032	0.015	0.732

Feature	RFR_0.269	GBR_0.258	XGBR_0.264	SVR_0.209	Ensemble	Linear_coef
ALI_month_cos	0.011	0.010	0.032	0.005	0.015	-3.585
XX_EAT_Breakfast	0.007	0.006	0.017	0.030	0.014	1.039
XX_VIC_smoke	0.009	0.006	0.022	0.012	0.012	-0.984
ALI_day_of_week_cos	0.010	0.009	0.014	0.016	0.012	-0.035

Cuadro 6.28: Importancia de características según modelos y ensamble y coeficientes lineales - "*Charlie*" - 'Personal'

En el cuadro 6.29, se agrupan las diferentes importancias según los 8 tópicos correspondientes, en función de su suma y su promedio, ordenada según este último, describiendo la importancia relativa de los mismos tópicos estudiados en 'Encuesta' a través de los datos del participante.

Prefix	sum_imp	mean_imp
'ACT_': Actividades sociales	0.1906	0.0477
'ALI_': Aspectos externos, ajenos	0.4316	0.0432
'TIR_': Cansancio	0.0785	0.0393
'VIC_': Vicios o hábitos perjudiciales	0.1520	0.0304
'FRE_': Tiempo libre	0.0289	0.0289
'PRF_': Tiempo provechoso	0.0685	0.0228
'SPO_': Deporte	0.0191	0.0191
'EAT_': Alimentación	0.0307	0.0154

Cuadro 6.29: Importancia agrupada según tópicos - "*Charlie*" - 'Personal'

A continuación, explicaremos de manera genérica las características que se anticipan para el participante según su segmentación en la encuesta (véase el apartado 6.1.2 para una explicación más rigurosa), y su contrate según el estudio 'Personal'. Dado que al participante se le asignaron las etiquetas de grupo [1, 0, 1, 1] según "pProfile", "topics", "subtopics" y "encuesta", respectivamente, el perfil del participante debería caracterizarse de la siguiente manera:

En "*pProfile*", el grupo 1 presenta el rango más amplio de convivencia, aunque tiende a vivir con sus padres. Probablemente se trate de un estudiante, con intereses en los videojuegos, dependa económicamente de alguien y se considere romántico. En este caso, sí se identifica en "*Charlie*" una variable contabilizando las horas dedicadas a los videojuegos (véase el cuadro D.6).

En "*topics*", el grupo 0 considera que las circunstancias externas a la persona no influyen en su felicidad, a diferencia de la alimentación y los hobbies. Además, el deporte tiene una influencia notable.

En este caso, las importancias de características aglomeradas no podrían estar más desvinculadas con los resultados de la segmentación. Las actividades sociales y las circunstancias externas encabezan el cuadro, y la alimentación y el deporte se encuentran en su base, indicando su baja influencia. Al tiempo libre también se le atribuye una menor importancia (véase el cuadro 6.29).

En “*subtopics*”, el grupo 1 se caracteriza por concederle una importancia mayor a las fiestas, las actividades culturales, el deporte pasivo, el salir de casa como actividad deportiva, y el tomar café.

Para el caso de “*Charlie*”, observamos cierta alineación entre las previsiones y sus resultados: a pesar que se le confirió una importancia negligible a la fiesta, y que las actividades culturales no se tuvieron en cuenta, las actividades deportivas tuvieron cierta importancia, y el tomar café todavía más (véase el cuadro 6.28).

En “*encuesta*”, el perfil personal del grupo ‘1’, se caracteriza por no gustares la fiesta, no les importa no aprovechar el día y no son deportistas. Piensan que los factores ajenos a la persona son factores clave, y sienten que hacer deporte pasivo y el café pueden influir en gran medida.

En este apartado, la contrastación entre resultados requiere del estudio simultáneo de los dos cuadros 6.28 y 6.29. En el aspecto del perfil personal parece que se reflejan sus supuestas preferencias en los datos estudiados: a pesar de que el tiempo productivo sí presenta una importancia relativamente elevada, parece que la ni la fiesta ni el deporte influyen demasiado en su bienestar. En cuanto a tópicos y subtópicos, los resultados coinciden exceptuando el deporte pasivo, que no se tuvo en cuenta.

En **conclusión**, los tópicos de actividades sociales, aspectos externos y cansancio son los más influyentes en la felicidad de Charlie. En cuanto a las actividades, se comprendan dentro de los mismos tópicos o no, destacamos: el nivel de estrés (inverso), las actividades sociales, sobre todo con la pareja, los aspectos externos y los vicios consumidos. En el aspecto de su segmentación en la encuesta, el participante cumple con las asunciones hechas según “*pProfile*”, “*subtopics*” y “*encuesta*” pero no para “*topics*”.

6.3 Discusión de resultados

Existen **discrepancias** entre las expectativas de la influencia de las variables en la felicidad, basadas en la segmentación, y las influencias reales de los participantes según su estudio personal. Los resultados muestran que, en general, las características previstas según la segmentación en la encuesta no se corresponden demasiado con los resultados obtenidos en el estudio 'Personal', especialmente para Juju, Pato y Ajara. Adicionalmente en cuanto a la segmentación de la encuesta según el perfil personal, las preferencias que se indican no se reflejan en las actividades registradas en el estudio 'Personal'.

Los resultados revelan que ciertos aspectos tienen una influencia más significativa en la predicción de la felicidad que otros. Estos varían en función del participante, pero los tópicos y actividades relacionadas con el **cansancio**, los **factores externos** y el **tiempo libre** tienden a ser los más influyentes en la felicidad de los participantes. Por otro lado, elementos como el deporte y la alimentación, en ocasiones, tienen una influencia menor de la esperada.

En lo que respecta concretamente al **deporte**, si bien es cierto que, en general, se le ha atribuido una influencia menor a la esperada entre los distintos participantes, quizás se deba a la carencia de dispositivos que logren capturar las variables adecuadas entre los distintos participantes (a excepción de Relookyou), y realmente éste tenga una influencia mayor a la contemplada.

En lo que respecta al **clima**, quizás se tuvieron en cuenta demasiadas variables relacionadas con la meteorología en comparación a otros factores, brindándole a los aspectos ajenos a la persona una importancia desproporcionada. O bien quizás debiese haberse tomado en cuenta el clima como un tópico en sí mismo, separándolo de los factores externos.

La relación entre las preferencias, actividades y la felicidad es **compleja y multifacética**. Las características que influyen en la felicidad de los participantes no se limitan a un solo aspecto, sino que son una combinación de diversos factores relacionados con su estilo de vida, aspectos externos, interacciones sociales, tiempo libre y una lista interminable de otros elementos.

En resumen, los resultados muestran que las relaciones entre las preferencias declaradas en la encuesta y las actividades reales en el estudio 'Personal' no siempre son

directas ni consistentes. A pesar de ello, se encuentran indicios de que ciertas áreas, como el cansancio y los factores externos, son de gran importancia para predecir la felicidad de los participantes. Sin embargo, la relación entre estas áreas y la felicidad es intrincada y suele variar entre individuos, destacando la complejidad de la naturaleza humana y la ineeficacia de una segmentación de la población, como la que en este estudio planteamos, para predecir los factores relevantes en la felicidad.

CAPÍTULO 7

Conclusiones

En este trabajo se aborda la identificación de los factores que pueden estar relacionados con la felicidad de las personas. Con este objetivo se ha desarrollado una metodología que analiza datos de personas encuestadas sobre diferentes tópicos influyentes en la felicidad, y datos personales longitudinales que recogen además valores subjetivos de felicidad de un conjunto de individuos.

La metodología permitió **segmentar correctamente** a los encuestados según sus perfiles personales y preferencias declaradas, así como la caracterización de los grupos formados según sus hábitos y actividades, entendiendo cómo diferentes segmentos persiguen la felicidad.

Se **analizaron efectivamente** los datos longitudinales personales, determinando la influencia relativa de las variables y factores en la felicidad de los participantes.

Sin embargo, atendiendo que existen **discrepancias** entre las expectativas de la influencia de las variables en la felicidad, basadas en la segmentación, y las influencias reales de los participantes según su estudio personal, sospechamos que nuestra segmentación no es apropiada para conjeturar las características influyentes de un individuo.

Aún con todo, consideramos exitoso el **esquema metodológico** diseñado para vincular ambas aproximaciones y evaluar la precisión de la segmentación al predecir la influencia de las actividades en un individuo.

Cabe mencionar una notable desviación respecto la **planificación** original: debido tanto al margen de tiempo limitado, como a la gran extensión que estaba tomando el informe, se optó por incluir solamente una selección entre todos los participantes de ‘Personal’. Esta elección se respalda con que los participantes no seleccionados “Mack” y “Edogawa”, se segmentaron de la misma forma que “Relookyoo”, tratándose este último del conjunto de datos más completo.

Se debe hacer otro inciso para aclarar que se tomaron medidas para asegurar el no transgredir aspectos **éticos** ni **legales** al tratar con esta tipología de datos tan personales. Concretamente, se obtuvo el consentimiento informado de cada participante. En la hoja de consentimiento se explicaron los detalles relevantes para que el participante pudiese decidir, informadamente, si participar o no en el estudio. En los apéndices del documento se puede encontrar una muestra de ésta.

En conclusión, a pesar de ciertas limitaciones que se mencionarán a continuación, hemos logrado una mayor comprensión de los factores que influyen en la felicidad personal, y conseguimos también diseñar un esquema metodológico apropiado para el análisis de nuestro marco contextual.

7.1 Limitaciones

Se debe reconocer que este estudio no está exento de ciertas limitaciones que podrían haber influido en los resultados y en su interpretación. Estas limitaciones son fundamentales para comprender la extensión y la aplicabilidad de las conclusiones presentadas. A continuación, se detallan algunas de las limitaciones clave que deben considerarse.

El **tamaño de la muestra** utilizada en este estudio, si bien proporciona información valiosa, puede no ser representativo de la población en su conjunto. Además, la muestra puede haber carecido de diversidad en términos de diversos factores (demográficos, socioeconómicos, culturales, etc.). Esta falta de representatividad puede limitar la generalización de los hallazgos a grupos más amplios y diversos.

A pesar de los esfuerzos para estandarizar la medición de variables subjetivas como la felicidad, es importante reconocer que la naturaleza de estas mediciones puede llegar a ser muy compleja. La escala utilizada para evaluar la felicidad podría haber sido **interpretada de manera diferente** por los participantes, introduciendo cierta subjetividad en los resultados.

Aunque se realizó un esfuerzo para recopilar datos precisos y confiables, existe la posibilidad de que algunos participantes no hayan proporcionado **información del todo precisa** en sus registros de actividades. Esta falta de precisión podría haber afectado la coherencia y la confiabilidad de los resultados. En retrospectiva, la imputación de

valores nulos podría haberse abordado mediante algoritmos de predicción en lugar de depender únicamente del promedio basado en el día de la semana y las fechas oportunas.

El estudio se centró en un conjunto específico de variables relacionadas con las preferencias y actividades de los participantes. Sin embargo, sin duda, existen **muchos otros factores**, que podrían desempeñar un papel significativo en la predicción de la felicidad y que no fueron considerados en este análisis.

Las preferencias y actividades de los individuos pueden cambiar con el tiempo. El estudio se basa en un **momento específico** en la vida de los participantes, no permitiendo capturar la dinámica evolutiva de las preferencias a lo largo del tiempo.

Por último, y quizás más importante, se reconoce la falta de **representaciones gráficas** en los resultados. Una visualización simple, efectiva, novedosa e impactante de los resultados es esencial para comunicar claramente los hallazgos e impresionar al lector.

Estas limitaciones resaltan la complejidad en la investigación en torno a la felicidad, así como nuestras propias áreas de mejora en el ámbito de la ciencia de datos. No obstante, a pesar de estas limitaciones, los resultados siguen aportando información de gran valor sobre la interrelación entre las actividades, las preferencias y la felicidad.

7.2 Trabajo futuro

El presente estudio ha proporcionado valiosas perspectivas sobre la compleja interrelación entre las preferencias individuales, las actividades reales y la felicidad experimentada por los participantes. Sin embargo, se ha hecho evidente que el trabajo presenta ciertas carencias. En el trabajo futuro se tratarán de mitigar estas limitaciones ya descritas y, adicionalmente, se explorarán nuevas áreas de mejora y enriquecimiento en la investigación.

En un esfuerzo por mejorar la precisión y comprensión de los resultados, se propone una **mayor exploración de modelos**. Esto incluye la evaluación de una mayor variedad de hiperparámetros de los modelos y métodos de procesamiento de datos más adecuados. Se buscará entre una mayor variedad de enfoques para mejorar la identificación de relaciones entre las preferencias, las actividades y la felicidad, considerando posibles

interacciones y correlaciones más sutiles que pueden haber pasado desapercibidas en el presente estudio.

Para identificar patrones más sólidos relacionados con la felicidad, se pretende aplicar técnicas de **clusterización a los datos personales**. Mediante esta técnica, se agruparán los registros del participante con características y actividades similares, permitiendo una comprensión más profunda de cómo ciertas combinaciones de actividades pueden influir en la felicidad.

En lugar de simplemente eliminar las **variables confusoras** en el estudio de 'Encuesta', se buscará abordarlas de manera más completa y sofisticada. Esto podría implicar la aplicación de técnicas estadísticas avanzadas para controlar el efecto de estas variables en las relaciones observadas entre preferencias, actividades y felicidad.

7.3 Contribuciones

Aunque los resultados no logren identificar con precisión las características influyentes mediante la segmentación de la población, sientan las bases para futuros intentos. El éxito en este planteamiento podría aportar una guía en la formulación de estrategias más efectivas para promover la felicidad y el bienestar en diferentes segmentos de la población. De modo que consideramos el diseño de la **metodología** empleada para la interpretación de los resultados de las dos aproximaciones de manera conjunta, 'Encuesta' y 'Personal', como la principal contribución del trabajo.

Este estudio contribuye al **conocimiento** del bienestar humano, permitiendo una visión más completa de cómo diversos factores interactúan para influir en la felicidad individual. Al profundizar, se abren oportunidades para intervenciones personalizadas, lo que podría mejorar la calidad de vida de las personas. De hecho, al finalizar el estudio, cada uno de los participantes en el conjunto de datos 'Personal' podrá disfrutar de los beneficios de conocer las actividades que más influyen en su felicidad.

El planteamiento llevado a cabo aborda la **brecha** entre percepciones ('Encuesta') y comportamientos ('Personal'), pudiendo devenir un fenómeno relevante en investigación psicológica y sociológica. Este trabajo ofrece una aproximación valiosa para entender cómo estas discrepancias afectan la felicidad.

Hoja de consentimiento

A continuación, se incluye una muestra de la hoja de consentimiento informado, entregada y firmada por los diferentes participantes que se incluyeron en el estudio.

HOJA DE CONSENTIMIENTO INFORMADO

Título del Estudio: 'Análisis de Factores Influyentes en la Felicidad Personal: Un Enfoque de Aprendizaje Automático Utilizando Datos Multidimensionales'.

Institución y/o departamento responsable: Universitat de Girona, Escuela Politécnica Superior.

Población de estudio: 8 participantes.

Datos de contacto del investigador: Francesc Xavier Reverté Baró,
correo: 00000@00.com, teléfono: 00000000.

El presente informe tiene como objetivo el proporcionar toda la información necesaria para que usted pueda decidir, libre y voluntariamente, si quiere participar en este estudio. Por favor, lea atentamente el informe y pregunte cualquier duda al respecto dirigiéndose a la información de contacto proporcionada.

PROPÓSITO DEL ESTUDIO

El propósito de este estudio es analizar y comprender cómo diversos aspectos de la vida cotidiana influyen en la felicidad y el bienestar personal. Los resultados de este estudio pueden contribuir a una mayor comprensión de los determinantes de la felicidad y el bienestar de las personas.

PROCEDIMIENTO

En su momento se formuló una encuesta con el objetivo de recabar información del perfil personal de los encuestados e identificar aquellos factores que más influencian en su

felicidad. Usando diferentes técnicas de aprendizaje automático se ha podido agrupar a los participantes según sus similitudes e identificar los factores determinantes de la felicidad de cada grupo.

Una vez en este punto del proceso, empieza la segunda parte del estudio, la que se beneficiará de su participación, si así lo desea.

Será necesario que los participantes registren diariamente, durante más de 30 días, datos sobre las diferentes actividades que hace a lo largo del día, así como una valoración, del 0 al 10, de la felicidad sentida. La elección de las actividades que se deberán registrar dependerá del criterio del participante; éste deberá tratar de tener en cuenta aquellas que más puedan influenciar en su felicidad personal. Sí que se pedirá que se incluyan los siguientes tópicos: deporte (p. ej., horas de deporte), sueño (p. ej., horas de sueño, siesta), alimentación (p. ej., comer carne, pescado, comer sano), vicios (p. ej., cafés tomados, cigarrillos), social (p. ej., si hoy he tenido alguna actividad social), ajenos a la persona (p. ej., buenas o malas noticias), tiempo productivo (p. ej., horas de trabajo) y tiempo libre (p. ej., horas de televisión).

Para la comodidad del participante y del investigador, se requiere que los datos se recojan en una aplicación móvil que permita su exportación en formato csv, como la aplicación “Hábitos”.

También se le pedirá a cada participante que responda la encuesta (de la primera parte del estudio), así poder determinar el grupo al que pertenece y los supuestos factores que más influyen en su felicidad. Estos supuestos ayudarán a enfocar con mayor precisión el estudio de datos del participante en cuestión.

De este modo se podrá averiguar cuáles son realmente las variables más influyentes en el bienestar de cada participante, y contrastarlos con los estimados en la primera parte del estudio (según el análisis de la encuesta).

RIESGOS E INCOMODIDADES

No se anticipan riesgos graves o molestias significativas asociadas con su participación en este estudio. La información recopilada se manejará con confidencialidad y se utilizará exclusivamente para fines de investigación. Si existe una incomodidad, ésta se

trata de las molestias generadas por el hecho de requerir de un compromiso diario a la hora de registrar los datos. Se estima que cada participante deberá invertir unos 5 minutos cada día para registrarlos, por lo que realmente es necesaria un notable grado constancia y compromiso.

BENEFICIOS

Se contemplan 3 distintos beneficios por parte del participante:

El primero es tomar constancia de las actividades y hábitos diarios. A lo largo del proceso de captación de datos, es esperable que el participante se dé cuenta de algunos excesos o carencias en algunas de sus actividades. ¿Quizás no estoy comiendo sano? ¿Puede que haga poco deporte? ¿Estoy bebiendo demasiado?

La segunda es que, una vez finalizado el estudio, el investigador se compromete a compartirlo; de modo que el participante se beneficiará de un mayor autoconocimiento de su persona, de tal manera que, no solo podrá averiguar cuáles de sus actividades influyen más en su felicidad, sino que también encontrará las distintas correlaciones entre las diferentes variables. Por ejemplo -Parece que los días que veo a la familia tengo un estilo de vida más saludable-.

Los beneficios de este estudio incluyen contribuir al conocimiento científico sobre los factores que influyen en la felicidad y el bienestar personal. Su colaboración puede ayudar a mejorar la comprensión general de estos temas.

CONFIDENCIALIDAD

Los datos recopilados durante el estudio serán tratados con la máxima confidencialidad. Se tomarán medidas para proteger su privacidad y su identidad. Cualquier dato recopilado se utilizará exclusivamente para fines de investigación. Los resultados del estudio solo se presentarán en forma agregada y no se identificará a ningún individuo específico en las publicaciones o informes resultantes.

Aseguramos la confidencialidad y protección de los datos, según lo establecido en la Ley Orgánica 3/2018, de 5 de diciembre, de protección de datos personales y garantía de derechos digitales, y en el Reglamento (UE) 2016/679 del Parlamento Europeo y del

Consejo, de 7 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos, y por el cual se deroga la Directiva 95/46/CE (Reglamento general de protección de datos).

DERECHO A TENER MÁS INFORMACIÓN SOBRE EL ESTUDIO

Si en cualquier momento desea obtener más información sobre el estudio, siempre que quiera, a lo largo del registro, puede ponerse en contacto con el investigador responsable, Francesc Xavier Reverté Baró, a través de su dirección de correo electrónico (00000@00.com) o su número de teléfono (000000000). Estará disponible para responder cualquier pregunta o preocupación que pueda tener.

RECHAZO O ABANDONO DE LA PARTICIPACIÓN

La participación en este estudio es voluntaria y su decisión de participar o no, no tendrá consecuencias negativas para usted. Si decide participar, puede retirarse del estudio en cualquier momento sin necesidad de dar explicaciones. Del mismo modo, el investigador puede retirar su participación del estudio si no cumple con los requisitos o si, por alguna razón, se interrumpe el estudio.

FIRMA

Al firmar a continuación, afirmo que se me ha explicado el propósito y los procedimientos de la presente investigación, así como los posibles riesgos, beneficios y mis derechos como participante. He tenido la oportunidad de hacer preguntas y todas mis dudas han sido respondidas satisfactoriamente. Entiendo que mi participación es voluntaria y que puedo retirarme en cualquier momento sin repercusiones. Mi firma a continuación expresa mi deseo de participar libre y voluntariamente en este estudio.

Participante

Fecha

El abajo firmante declara haber explicado el propósito de la investigación, los procedimientos utilizados en el estudio, identificando aquellos que tienen un propósito meramente de investigación, los posibles riesgos e incomodidades que puedan originarse

y que ha respondido lo mejor que ha podido a las preguntas que se le han formulado respecto al estudio.

Investigador

Fecha

Árboles de decisión – ‘Encuesta’

A continuación, mostraremos los resultados de los diferentes modelos (Decision Tree) según la porción del dataset ‘Encuesta’ estudiado. El relativo a “*pProfile*” se encuentra en la figura 6.7)

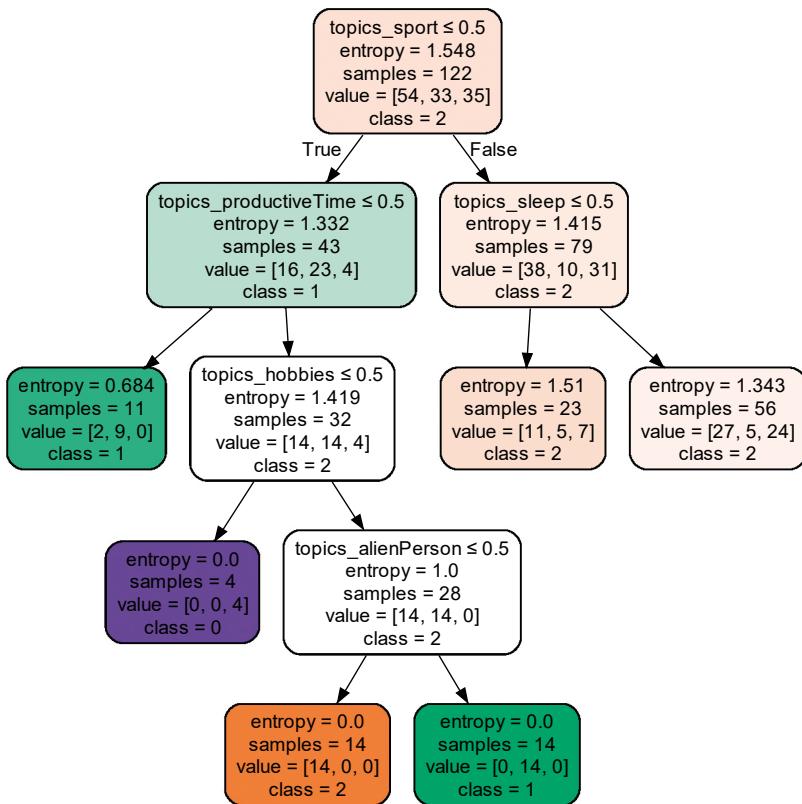


Figura B.1: Apéndice - Clasificación de grupos según Decission Tree - "topics" - 'Encuesta'

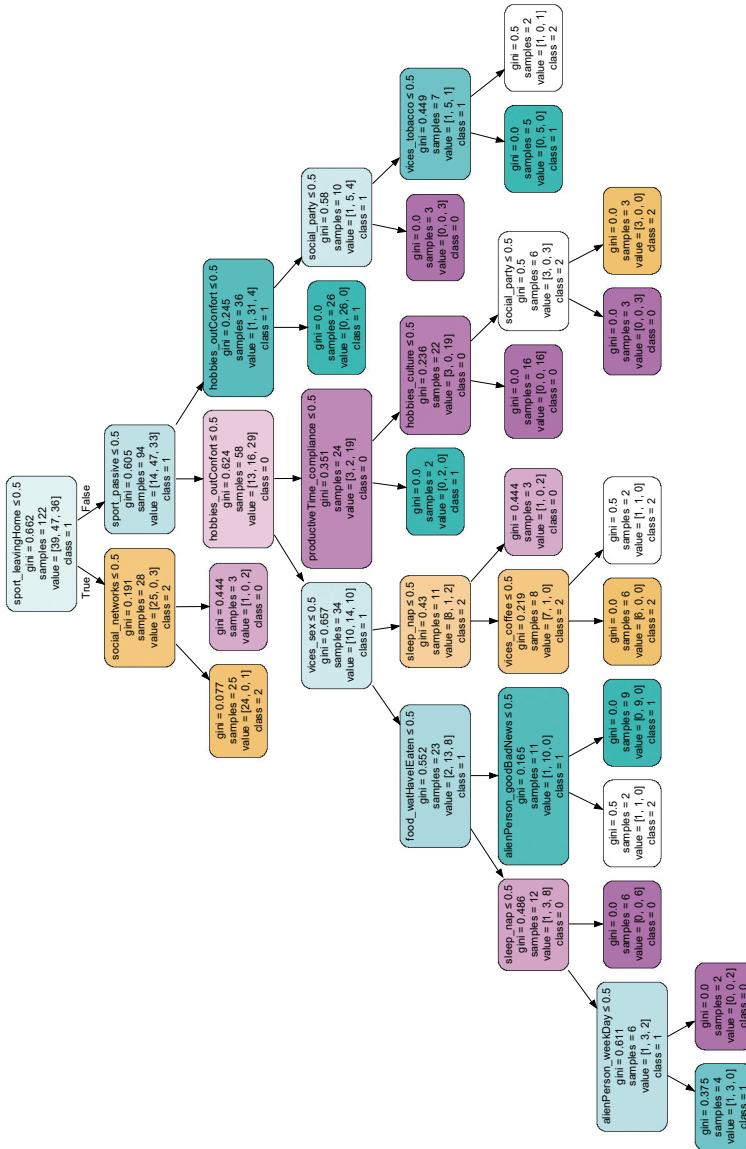


Figura B.2: Apéndice - Clasificación de grupos según Decission Tree - "subtopics" - 'Encuesta'

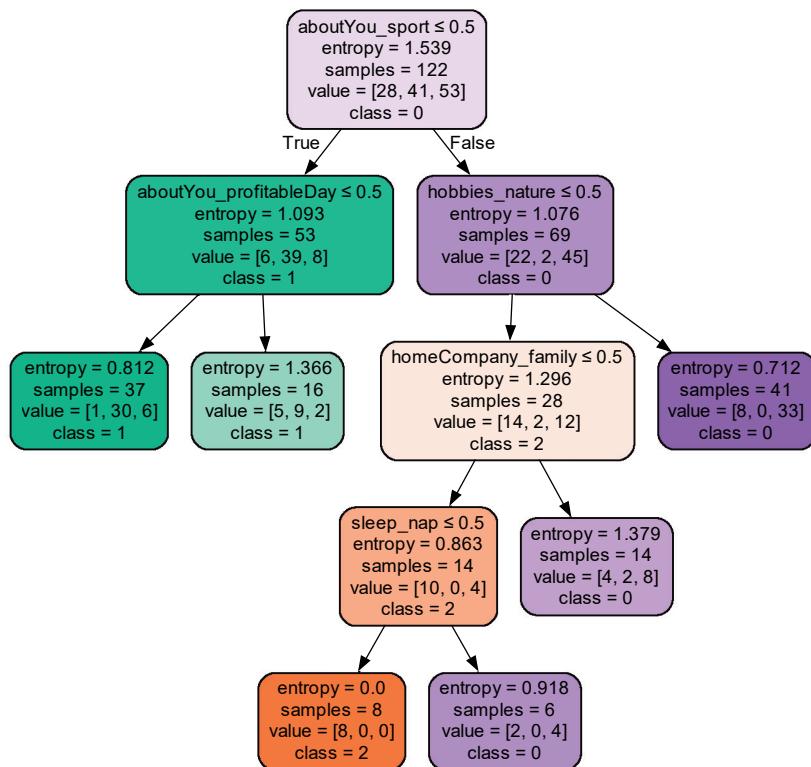


Figura B.3: Apéndice - Clasificación de grupos según Decission Tree - "encuesta" - 'Encuesta'

Rendimiento ensamblajes - 'Personal'

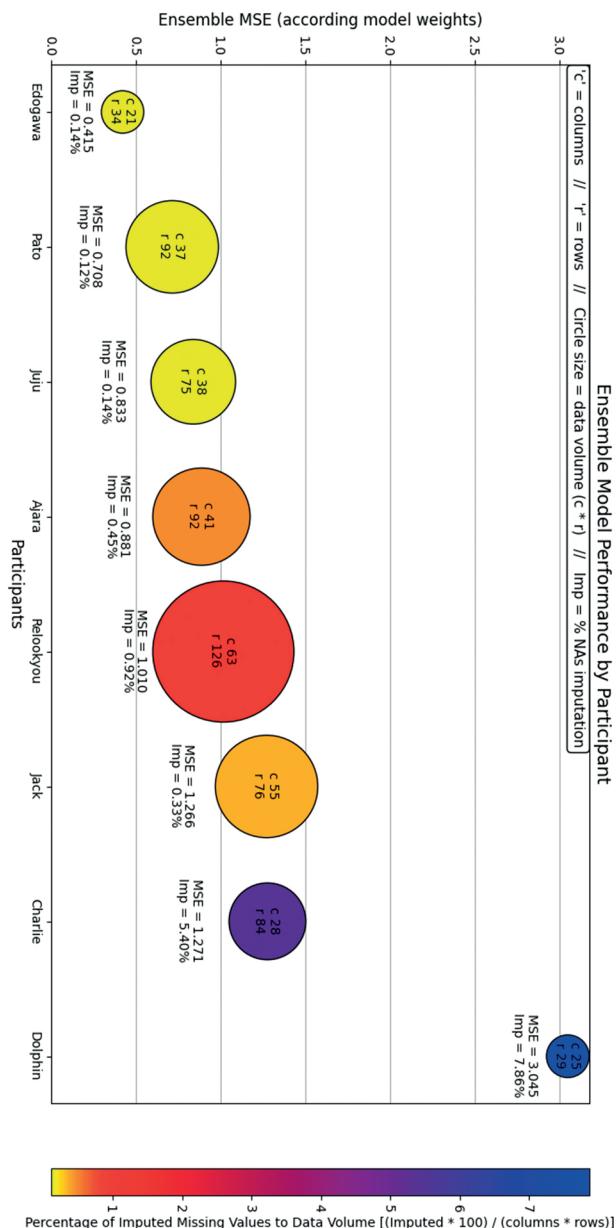


Figura C.1: Apéndice - Representación gráfica del rendimiento de los ensamblajes y datasets estudiados - 'Personal'

Variables estudiadas - ‘Personal’

A continuación, mostraremos los cuadros resumen de las variables registradas para los distintos participantes que se incluyeron en el estudio. Cabe destacar diversas consideraciones:

En los cuadros encontramos distintas columnas: ‘variable’, en la que se indican las distintas características; ‘Mean’, ‘Min’, ‘Max’ y ‘std’, que describen el resumen estadístico de las variables; ‘Type’, indicando el tipo de objeto de pandas y ‘Descripción’, que explica la característica, indicando primero su procedencia (HABITS provenientes de la aplicación ‘Hábitos’, GARMIN procedentes del smartwatch ‘Garmin Venu2’, y FE derivadas de la ingeniería de características).

Después de una lectura del cuadro 3.3, indicando el significado de la codificación de prefijos, la descripción de la mayoría de variables ya se puede entender sencillamente leyendo su nombre. Existen algunas variables, que se muestran en los cuadros, pero que posteriormente fueron eliminadas según la selección de características, como el caso de ‘TIR_time_get_up’ de “Dolphin”.

D.1 Relookyou

Variable	Mean	Min	Max	std	Type	Descripción
date	2023-06-20	2023-04-20	2023-08-20	35	date-time	HÁBITOS - Fecha
TIR_time_get_up	10.1	6.0	16.0	2.3	float64	HÁBITOS - Hora de levantarse
EVA_TIR_sleep_quality	5.6	2.0	9.0	1.9	float64	HÁBITOS - Evaluación subjetiva (0-10) de la calidad de sueño
EVA_happiness_16	5.8	2.0	9.0	1.5	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la mañana
EVA_TIR_tiredness_16	5.9	2.0	9.0	1.5	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la mañana
FRE_siesta	0.1	0.0	1.0	0.1	float64	HÁBITOS - Horas de siesta
EVA_happiness_22	5.6	2.0	9.0	1.7	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la tarde

Variable	Mean	Min	Max	std	Type	Descripción
EVA_TIR_tiredness_22	5.1	2.0	9.0	1.7	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la tarde
FRE_TIR_meditate	0.1	0.0	1.0	0.2	float64	HÁBITOS - Horas de meditación
XX_FRE_shower	0.6	0.0	1.0	0.5	int64	HÁBITOS - Ducha
PRF_FRE_transport	1.2	0.0	5.0	1.2	float64	HÁBITOS - Horas de transporte
SPO_pull_ups	13.3	0.0	111.0	24.5	int64	HÁBITOS - Número de dominadas
PRF_study_of_happiness	3.2	0.0	12.5	3.5	float64	HÁBITOS - Horas dedicadas a este estudio
PRF_data_science_study	0.1	0.0	7.0	0.8	float64	HÁBITOS - Horas dedicadas a la ciencia de datos (al margen del estudio)
XX_PRF_internship	0.3	0.0	1.0	0.5	int64	HÁBITOS - Prácticas de empresa
PRF_reading	0.2	0.0	2.5	0.5	float64	HÁBITOS - Horas de lectura
XX_PRF_FRE_chess	0.3	0.0	1.0	0.4	int64	HÁBITOS - Jugar a ajedrez
PRF_other_didactics	0.2	0.0	7.0	0.9	float64	HÁBITOS - Horas dedicadas a otros didácticos
XX_EAT_eating_out	0.4	0.0	1.0	0.5	int64	HÁBITOS - Comer fuera
XX_EAT_VIC_meat	0.8	0.0	1.0	0.4	int64	HÁBITOS - Comer carne
XX_EAT_fish	0.3	0.0	1.0	0.4	int64	HÁBITOS - Comer pescado
XX_EAT_carbohydrates	0.8	0.0	1.0	0.4	int64	HÁBITOS - Comer carbohidratos
XX_EAT_legumes	0.2	0.0	1.0	0.4	int64	HÁBITOS - Comer legumbres
XX_EAT_vegetables	0.6	0.0	1.0	0.5	int64	HÁBITOS - Comer vegetales
XX_EAT_fruit	0.2	0.0	1.0	0.4	int64	HÁBITOS - Comer fruta
XX_EAT_dairy	0.5	0.0	1.0	0.5	int64	HÁBITOS - Comer/beber lácteos
XX_EAT_fried	0.4	0.0	1.0	0.5	int64	HÁBITOS - Comer fritos
XX_EAT_VIC_processed	0.4	0.0	1.0	0.5	int64	HÁBITOS - Comer procesados
XX_ACT_activities	0.4	0.0	1.0	0.5	int64	HÁBITOS - Actividad social
XX_ACT_Kurt	0.1	0.0	1.0	0.2	int64	HÁBITOS - Actividad social con Kurt
XX_ACT_Edogawa	0.1	0.0	1.0	0.2	int64	HÁBITOS - Actividad social con Edogawa
XX_ACT_Mack	0.1	0.0	1.0	0.3	int64	HÁBITOS - Actividad social con Mack
XX_ACT_Charlie	0.0	0.0	1.0	0.2	int64	HÁBITOS - Actividad social con Charlie
XX_ACT_other_friends	0.1	0.0	1.0	0.3	int64	HÁBITOS - Actividad social con otros amigos
FRE_VIC_audiovisual_content	3.5	0.0	11.0	3.0	float64	HÁBITOS - Horas de contenido audiovisual
XX_FRE_VIC_anime	0.3	0.0	1.0	0.4	int64	HÁBITOS - Mirar anime
FRE_music	2.5	0.0	10.0	2.3	float64	HÁBITOS - Horas de música
XX_VIC_self_pleasure	0.6	0.0	1.0	0.5	int64	HÁBITOS - Onanismo
VIC_cigars	3.7	0.0	13.0	3.0	int64	HÁBITOS - Cigarrillos fumados
VIC_coffee	1.3	0.0	4.0	0.8	int64	HÁBITOS - Cafés tomados
VIC_alcohol	1.8	0.0	12.0	2.3	float64	HÁBITOS - Alcohol tomado (en cervezas como unidad)
XX_VIC_weed	0.2	0.0	1.0	0.4	int64	HÁBITOS - Fumar cannabis
PRF_TFMDrissa	0.2	0.0	6.0	1.1	int64	HÁBITOS - Horas de prácticas de empresa dedicadas, en realidad, a 'PRF_study_of_happiness'
EVA_happiness	5.7	2.5	8.0	1.4	float64	FE - Promedio de las evaluaciones subjetivas de la felicidad

Variable	Mean	Min	Max	std	Type	Descripción
EVA_TIR_tiredness	5.5	2.5	8.5	1.3	float64	FE - Promedio de las evaluaciones subjetivas del cansancio
PRF_sum	6.4	0.0	14.2	4.0	float64	FE - Suma de horas de tiempo productivo
FRE_sum	4.9	0.5	11.5	3.0	float64	FE - Suma de horas de tiempo libre
EAT_healthy	0.7	-0.8	3.0	0.9	float64	FE - Indicador de la sanidad en la alimentación
VIC	4.8	0.8	10.6	2.2	float64	FE - Indicador de los vicios consumidos
XX_ACT	0.5	0.0	1.0	0.5	int64	FE - Actividad social (1 si hay alguna de cualquier tipo)
ALI_week	9.8	1.0	18.0	5.1	int64	FE - Número de la semana desde el inicio del estudio
TIR_body_bat_0h	13.3	5.0	83.0	10.7	float64	GARMIN - Batería corporal a las 00h
TIR_body_bat_4h	67.1	5.0	100.0	29.6	float64	GARMIN - Batería corporal a las 04h
TIR_body_bat_8h	71.9	5.0	100.0	30.6	float64	GARMIN - Batería corporal a las 08h
TIR_body_bat_12h	55.0	14.0	100.0	18.7	float64	GARMIN - Batería corporal a las 12h
TIR_body_bat_16h	50.2	6.0	100.0	15.3	float64	GARMIN - Batería corporal a las 16h
TIR_body_bat_20h	17.8	5.0	84.0	12.5	float64	GARMIN - Batería corporal a las 20h
TIR_body_bat_mean	45.9	8.8	68.7	13.0	float64	FE - Promedio diario de Batería corporal
TIR_sleep_time	7.7	4.1	11.5	1.3	float64	GARMIN - Horas de sueño
TIR_sleep_score	81.7	25.0	100.0	13.9	float64	GARMIN - Puntuación de sueño (0-100)
TIR_stress	38.0	21.0	63.0	7.7	float64	GARMIN - Nivel de estrés (0-100)
SPO_card_freq_rest	44.4	38.0	57.0	3.2	float64	GARMIN - Frequencia cardíaca en reposo
SPO_active_calories	419.9	12.0	1848.0	384.7	float64	GARMIN - Calorías activas
SPO_intens_minut_value	25.5	0.0	235.0	43.7	float64	GARMIN - Minutos de intensidad deportiva
SPO_steps	5815.0	74.0	15746.0	3703.5	float64	GARMIN - Pasos andados
SPO_steps_goal	5862.0	3740.0	8430.0	1143.6	float64	GARMIN - Objetivo diario de pasos
ALI_tmed	22.5	13.8	28.6	4.0	float64	AEMET - temperatura media (°C) (Barcelona)
ALI_prec	1.2	0.0	35.9	3.7	float64	AEMET - precipitación (mm) (Barcelona)
ALI_tmax	25.2	17.2	32.1	4.0	float64	AEMET - temperatura máxima (°C) (Barcelona)
ALI_trange	5.3	2.5	10.3	1.3	float64	FE - Rango de temperaturas (°C) (Barcelona)
ALI_month_sin	-0.1	-0.9	0.9	0.6	float64	FE - Codificación cíclica del mes (sinus)
ALI_month_cos	-0.8	-1.0	-0.5	0.2	float64	FE - Codificación cíclica del mes (cosinus)
ALI_day_of_week_sin	0.0	-1.0	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (sinus)
ALI_day_of_week_cos	0.0	-0.9	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (cosinus)
SPO_sport	0.9	0.0	7.0	1.5	float64	FE - Actividad deportiva (1 caminar, 2 gimnasio, 3 cardio, 4 correr, 0.5)

Variable	Mean	Min	Max	std	Type	Descripción
TIR_sleep_score_mix	7.1	3.1	9.3	1.4	float64	FE - Combinación de TIR_sleep_score y EVA_TIR_sleep_quality
EVA_TIR_tiredness_mix	4.1	2.6	7.9	0.8	float64	FE - Combinación de TIR_body_bat_mean y EVA_TIR_tiredness
TIR	2.7	-3.6	5.8	1.8	float64	FE - Combinación de TIR_sleep_score_mix, EVA_TIR_tiredness_mix, TIR_sleep_time y TIR_stress

Cuadro D.1: Apéndice - Resumen de variables de "Relookyou" - 'Personal'

D.2 Dolphin

Variable	Mean	Min	Max	std	Type	Descripción
date	2023-06-17	2023-05-05	2023-08-08	28	date-time	HÁBITOS - Fecha
TIR_time_get_up	0.2	0.0	6.0	1.1	float64	HÁBITOS - Hora de levantarse
TIR_sleep_time	6.8	4.0	7.0	0.6	float64	HÁBITOS - Horas de sueño
SPO_sportTime	0.6	0.0	3.0	0.9	float64	HÁBITOS - Horas de deporte
XX_ACT_FRE_Max	0.4	0.0	1.0	0.5	int32	HÁBITOS - Jugar con Max
PRF_Work	2.9	0.0	6.0	3.1	int32	HÁBITOS - Trabajo
FRE_VIC_audiovisual_content	0.4	0.0	5.0	1.3	float64	HÁBITOS - Horas de contenido audiovisual
SPO_Yoga	0.0	0.0	0.5	0.1	float64	HÁBITOS - Horas de deporte (Yoga)
ACT_FRE_Pool	0.1	0.0	1.5	0.3	float64	HÁBITOS - Horas de piscina (relajación)
SPO_Running	0.0	0.0	0.5	0.1	float64	HÁBITOS - Horas de deporte (correr)
SPO_Alberto	0.2	0.0	1.0	0.4	int32	HÁBITOS - Deporte (gimnasio)
XX_ACT_Family	0.4	0.0	1.0	0.5	int32	HÁBITOS - Actividad social en familia
XX_EAT_eating_out	0.3	0.0	1.0	0.5	int32	HÁBITOS - Comer fuera
PRF_reading	0.2	0.0	0.5	0.2	float64	HÁBITOS - Horas de lectura
EVA_happiness_16	5.6	2.0	10.0	2.7	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la mañana
EVA_happiness_22	5.5	2.0	9.0	2.1	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la tarde
EVA_TIR_tiredness_16	5.2	2.0	9.0	2.0	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la mañana
EVA_TIR_tiredness_22	5.9	4.0	10.0	1.4	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la tarde
XX_EAT_healthy	0.3	0.0	1.0	0.5	int32	HÁBITOS - Comida sana
XX_EAT_abundant	0.4	0.0	1.0	0.5	int32	HÁBITOS - Comida abundante
XX_ACT_Friends	0.3	0.0	1.0	0.5	int32	HÁBITOS - Actividad social con amigos
ALI_tmmed	22.4	17.1	27.4	2.8	float64	AEMET - temperatura media (°C) (Barcelona)
ALI_prec	0.8	0.0	10.8	2.1	float64	AEMET - precipitación (mm) (Barcelona)

Variable	Mean	Min	Max	std	Type	Descripción
ALI_tmax	25.1	19.5	31.7	2.9	float64	AEMET - temperatura máxima (°C) (Barcelona)
ALI_trange	5.5	3.1	8.6	1.3	float64	FE - Rango de temperaturas (°C) (Barcelona)
EVA_happiness	5.6	2.0	8.9	2.1	float64	FE - Promedio de las evaluaciones subjetivas de la felicidad
EVA_tiredness	5.6	3.0	9.5	1.5	float64	FE - Promedio de las evaluaciones subjetivas del cansancio
FRE_sum	0.9	0.0	5.0	1.3	float64	FE - Suma de horas de tiempo libre
PRF_sum	3.1	0.0	6.5	3.1	float64	FE - Suma de horas de tiempo productivo
XX_ACT	0.7	0.0	1.0	0.5	int32	FE - Actividad social (1 si hay alguna de cualquier tipo)
ALI_week	7.4	1.0	15.0	4.0	int64	FE - Número de la semana desde el inicio del estudio
ALI_month_sin	-0.1	-0.9	0.5	0.4	float64	FE - Codificación cíclica del mes (sinus)
ALI_month_cos	-0.9	-1.0	-0.5	0.2	float64	FE - Codificación cíclica del mes (cosinus)
ALI_day_of_week_sin	0.2	-1.0	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (sinus)
ALI_day_of_week_cos	-0.2	-0.9	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (cosinus)

Cuadro D.2: Apéndice - Resumen de variables de "Dolphin" - 'Personal'

D.3 Juju

Variable	Mean	Min	Max	std	Type	Descripción
date	2023-06-14	2023-05-06	2023-08-05	25	date-time	HÁBITOS - Fecha
EVA_happiness_16	5.2	3.0	7.0	1.2	int32	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la mañana
EVA_happiness_22	5.3	3.0	7.0	1.2	int32	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la tarde
EVA_TIR_tiredness_16	5.3	3.0	8.0	1.3	int32	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la mañana
EVA_TIR_tiredness_22	5.2	3.0	8.0	1.4	int32	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la tarde
TIR_time_get_up	8.7	5.0	12.7	1.4	float64	HÁBITOS - Hora de levantarse
TIR_sleep_time	7.7	2.0	20.6	2.0	float64	HÁBITOS - Horas de sueño
XX_FRE_shower	0.7	0.0	1.0	0.5	int32	HÁBITOS - Ducha
PRF_uni	0.6	0.0	3.0	1.2	int32	HÁBITOS - Horas de universidad
PRF_exams	0.1	0.0	2.5	0.6	float64	HÁBITOS - Examen de universidad
PRF_study	1.2	0.0	8.0	1.9	float64	HÁBITOS - Horas de estudio
PRF_reading	0.9	0.0	5.0	1.1	float64	HÁBITOS - Horas de lectura
PRF_duolingo	0.0	0.0	0.0	0.0	float64	HÁBITOS - Horas de estudio de idiomas
PRF_driving	0.2	0.0	0.8	0.3	float64	HÁBITOS - Horas de prácticas de coche

Variable	Mean	Min	Max	std	Type	Descripción
XX_EAT_eating_out	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer fuera
XX_EAT_VIC_meat	0.7	0.0	1.0	0.4	int32	HÁBITOS - Comer carne
XX_EAT_fish	0.1	0.0	1.0	0.4	int32	HÁBITOS - Comer pescado
XX_EAT_carbohydrates	0.7	0.0	1.0	0.5	int32	HÁBITOS - Comer carbohidratos
XX_EAT_legumes	0.2	0.0	1.0	0.4	int32	HÁBITOS - Comer legumbres
XX_EAT_vegetables	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer vegetales
XX_EAT_fruit	0.4	0.0	1.0	0.5	int32	HÁBITOS - Comer fruta
XX_EAT_fried	0.2	0.0	1.0	0.4	int32	HÁBITOS - Comer fritos
XX_EAT_VIC_processed	0.2	0.0	1.0	0.4	int32	HÁBITOS - Comer procesados
XX_ACT_Social	0.5	0.0	1.0	0.5	int32	HÁBITOS - Actividad social
XX_ACT_Party	0.1	0.0	1.0	0.3	int32	HÁBITOS - Actividad social de fiesta
FRE_VIC_audiovisual_content	1.1	0.0	4.0	1.2	float64	HÁBITOS - Horas de contenido audiovisual
FRE_music	1.9	0.0	6.0	1.5	float64	HÁBITOS - Horas de música
FRE_VIC_SocialNetworks	1.9	0.0	4.0	1.0	float64	HÁBITOS - Horas de redes sociales
VIC_alcohol	0.5	0.0	7.0	1.4	float64	HÁBITOS - Alcohol tomado (en cervezas como unidad)
VIC_coffee	0.6	0.0	3.0	0.7	float64	HÁBITOS - Cafés tomados
XX_SPO_tennis	0.1	0.0	1.0	0.3	int32	HÁBITOS - Deporte (tenis)
XX_SPO_Other	0.1	0.0	1.0	0.2	int32	HÁBITOS - Deporte (otros)
XX_ACT_shopping	0.2	0.0	1.0	0.4	int32	HÁBITOS - Actividad social (salir de compras)
ALI_tmed	22.3	13.8	27.4	3.5	float64	AEMET - temperatura media (°C) (Barcelona)
ALI_prec	1.0	0.0	35.9	4.4	float64	AEMET - precipitación (mm) (Barcelona)
ALI_tmax	25.0	18.7	31.7	3.4	float64	AEMET - temperatura máxima (°C) (Barcelona)
ALI_trange	5.4	3.1	10.3	1.3	float64	FE - Rango de temperaturas (°C) (Barcelona)
EVA_happiness	5.2	3.0	7.0	1.0	float64	FE - Promedio de las evaluaciones subjetivas de la felicidad
EVA_tiredness	5.2	3.0	7.0	1.0	float64	FE - Promedio de las evaluaciones subjetivas del cansancio
FRE_sum	3.1	0.0	7.1	1.6	float64	FE - Suma de horas de tiempo libre
PRF_sum	3.0	0.0	11.5	2.6	float64	FE - Suma de horas de tiempo productivo
EAT_healthy	0.8	-1.1	2.8	0.9	float64	FE - Indicador de la sanidad en la alimentación
VIC	2.7	0.4	5.1	1.0	float64	FE - Indicador de los vicios consumidos
XX_ACT	0.5	0.0	1.0	0.5	int32	FE - Actividad social (1 si hay alguna de cualquier tipo)
ALI_week	6.9	1.0	14.0	3.6	int64	FE - Número de la semana desde el inicio del estudio
ALI_month_sin	0.0	-0.9	0.5	0.4	float64	FE - Codificación cíclica del mes (sinus)
ALI_month_cos	-0.9	-1.0	-0.5	0.1	float64	FE - Codificación cíclica del mes (cosinus)
ALI_day_of_week_sin	0.0	-1.0	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (sinus)

Variable	Mean	Min	Max	std	Type	Descripción
ALI_day_of_week_cos	0.0	-0.9	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (cosinus)

Cuadro D.3: Apéndice - Resumen de variables de "Juju" - 'Personal'

D.4 Pato

Variable	Mean	Min	Max	std	Type	Descripción
date	2023-07-01	2023-05-17	2023-08-16	27	date-time	HÁBITOS - Fecha
TIR_time_get_up	6.3	4.0	11.0	0.9	float64	HÁBITOS - Hora de levantarse
TIR_sleep_time	6.0	4.0	9.0	1.0	float64	HÁBITOS - Horas de sueño
EVA_TIR_sleep_quality	5.3	3.0	7.0	1.3	float64	HÁBITOS - Evaluación subjetiva (0-10) de la calidad de sueño
EVA_happiness_16	5.6	3.0	8.0	1.2	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la mañana
EVA_TIR_tiredness_16	5.1	2.2	8.0	1.4	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la mañana
FRE_TIR_siesta	0.7	0.0	2.0	0.5	float64	HÁBITOS - Horas de siesta
EVA_happiness_22	5.7	3.0	7.0	1.0	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la tarde
EVA_TIR_tiredness_22	5.1	2.0	8.0	1.5	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la tarde
XX_FRE_shower	0.8	0.0	1.0	0.4	int32	HÁBITOS - Ducha
PRF_loreto	3.1	0.0	6.5	3.3	float64	HÁBITOS - Horas de trabajo
PRF_home	3.0	0.0	6.0	1.3	float64	HÁBITOS - Horas de teletrabajo
SPO_tennis	0.4	0.0	1.0	0.5	int32	HÁBITOS - Horas de deporte (tenis)
SPO_swimming	0.1	0.0	0.8	0.3	float64	HÁBITOS - Horas de deporte (natación)
SPO_other	0.5	0.0	3.0	0.7	float64	HÁBITOS - Horas de deporte (otros)
XX_EAT_eating_out	0.6	0.0	1.0	0.5	int32	HÁBITOS - Comer fuera
XX_EAT_VIC_meat	0.7	0.0	1.0	0.5	int32	HÁBITOS - Comer carne
XX_EAT_fish	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer pescado
XX_EAT_legumes	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer legumbres
XX_EAT_vegetables	0.8	0.0	1.0	0.4	int32	HÁBITOS - Comer vegetales
XX_EAT_fried	0.2	0.0	1.0	0.4	int32	HÁBITOS - Comer fritos
XX_EAT_VIC_processed	0.3	0.0	1.0	0.5	int32	HÁBITOS - Comer procesados
XX_ACT_Friends	0.6	0.0	1.0	0.5	int32	HÁBITOS - Actividad social con amigos
XX_ACT_Family	0.3	0.0	1.0	0.5	int32	HÁBITOS - Actividad social en familia
XX_ACT_otherSocial	0.7	0.0	1.0	0.5	int32	HÁBITOS - Otras actividades sociales
FRE_VIC_audiovisual_content	1.2	0.0	4.0	0.9	float64	HÁBITOS - Horas de contenido audiovisual
VIC_alcohol	3.7	1.0	7.0	1.2	float64	HÁBITOS - Alcohol tomado (en cervezas como unidad)
VIC_cigars	29.6	10.0	35.0	4.5	float64	HÁBITOS - Cigarrillos fumados
VIC_coffee	2.0	0.0	3.0	0.6	float64	HÁBITOS - Cafés tomados
XX_EAT_carbohydrates	0.7	0.0	1.0	0.5	int32	HÁBITOS - Comer carbohidratos

Variable	Mean	Min	Max	std	Type	Descripción
XX_ALI_profitableInvestments	0.5	0.0	1.0	0.5	int32	HÁBITOS - Inversiones provechosas
EVA_stress	4.4	2.0	7.0	1.3	float64	HÁBITOS - Evaluación subjetiva (0-10) del estrés
ALI_tmed	23.8	16.8	27.5	2.9	float64	AEMET - temperatura media (°C) (Barcelona)
ALI_prec	0.4	0.0	7.4	1.1	float64	AEMET - precipitación(mm) (Barcelona)
ALI_tmax	26.5	19.5	31.7	2.9	float64	AEMET - temperatura máxima (°C) (Barcelona)
ALI_trange	5.3	2.5	8.7	1.2	float64	FE - Rango de temperaturas (°C) (Barcelona)
EVA_happiness	5.7	3.5	7.5	0.9	float64	FE - Promedio de las evaluaciones subjetivas de la felicidad
EVA_tiredness	5.1	2.1	7.5	1.2	float64	FE - Promedio de las evaluaciones subjetivas del cansancio
FRE_sum	2.1	0.3	5.8	1.0	float64	FE - Suma de horas de tiempo libre
PRF_sum	6.1	0.0	11.5	3.6	float64	FE - Suma de horas de tiempo productivo
SPO_sportTime	1.0	0.0	3.0	0.7	float64	FE - Horas de deporte
EAT_healthy	1.1	-0.6	2.4	0.7	float64	FE - Indicador de la sanidad en la alimentación
VIC	4.7	2.3	7.6	1.2	float64	FE - Indicador de los vicios consumidos
XX_ACT	0.9	0.0	1.0	0.4	int32	FE - Actividad social (1 si hay alguna de cualquier tipo)
ALI_week	7.4	1.0	14.0	3.8	int64	FE - Número de la semana desde el inicio del estudio
ALI_month_sin	-0.2	-0.9	0.5	0.4	float64	FE - Codificación cíclica del mes (sinus)
ALI_month_cos	-0.8	-1.0	-0.5	0.2	float64	FE - Codificación cíclica del mes (cosinus)
ALI_day_of_week_sin	0.0	-1.0	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (sinus)
ALI_day_of_week_cos	0.0	-0.9	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (cosinus)

Cuadro D.4: Apéndice - Resumen de variables de "Pato" - 'Personal'

D.5 Ajara

Variable	Mean	Min	Max	std	Type	Descripción
date	2023-06-29	2023-05-15	2023-08-14	27	date-time	HÁBITOS - Fecha
TIR_time_get_up	7.6	5.0	10.0	1.0	float64	HÁBITOS - Hora de levantarse
TIR_sleep_time	6.7	3.0	10.0	1.3	float64	HÁBITOS - Horas de sueño
EVA_sleep_quality	5.7	1.0	9.0	2.2	float64	HÁBITOS - Evaluación subjetiva (0-10) de la calidad de sueño
EVA_happiness_16	7.1	4.0	9.0	1.1	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la mañana

Variable	Mean	Min	Max	std	Type	Descripción
EVA_TIR_tiredness_16	5.1	0.0	10.0	2.4	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la mañana
FRE_siesta	0.0	0.0	0.3	0.1	float64	HÁBITOS - Horas de siesta
EVA_happiness_22	7.3	2.0	9.0	1.3	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad por la tarde
EVA_TIR_tiredness_22	5.9	0.0	10.0	2.3	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio por la tarde
FRE_shower	0.3	0.0	0.5	0.2	float64	HÁBITOS - Ducha
XX_OTH_houseClean	0.4	0.0	1.0	0.5	int32	HÁBITOS - Tener limpio el hogar
EVA_OTH_personalCare	6.5	1.0	9.0	1.7	float64	HÁBITOS - Evaluación subjetiva (0-10) del cuidado personal
PRF_Work	1.8	0.0	9.0	3.6	int32	HÁBITOS - Trabajo
PRF_reading	0.0	0.0	0.3	0.0	float64	HÁBITOS - Horas de lectura
PRF_houseWork	0.8	0.0	4.0	0.9	float64	HÁBITOS - Tareas del hogar
SPO_sportTime	0.1	0.0	1.0	0.4	int32	HÁBITOS - Horas de deporte
XX_EAT_eating_out	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer fuera
XX_EAT_fish	0.3	0.0	1.0	0.5	int32	HÁBITOS - Comer pescado
XX_EAT_VIC_meat	0.6	0.0	1.0	0.5	int32	HÁBITOS - Comer carne
XX_EAT_eggs	0.2	0.0	1.0	0.4	int32	HÁBITOS - Comer huevos
XX_EAT_carbohydrates	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer carbohidratos
XX_EAT_legumes	0.2	0.0	1.0	0.4	int32	HÁBITOS - Comer legumbres
XX_EAT_vegetables	0.5	0.0	1.0	0.5	int32	HÁBITOS - Comer vegetales
XX_EAT_fruit	0.4	0.0	1.0	0.5	int32	HÁBITOS - Comer fruta
XX_EAT_fried	0.1	0.0	1.0	0.4	int32	HÁBITOS - Comer fritos
XX_EAT_VIC_processed	0.4	0.0	1.0	0.5	int32	HÁBITOS - Comer procesados
XX_ACT_shopping	0.3	0.0	1.0	0.5	int32	HÁBITOS - Actividad social (compras)
XX_ACT_mother	0.3	0.0	1.0	0.5	int32	HÁBITOS - Actividad social con la madre
XX_ACT_Friends	0.3	0.0	1.0	0.5	int32	HÁBITOS - Actividad social con amigos
XX_ACT_granddaughter	0.5	0.0	1.0	0.5	int32	HÁBITOS - Actividad social con nieta
XX_ACT_Family	0.2	0.0	1.0	0.4	int32	HÁBITOS - Actividad social en familia
XX_OTH_discussion	0.1	0.0	1.0	0.2	int32	HÁBITOS - Discusión con alguien
FRE_VIC_audiovisual_content	0.7	0.0	5.0	1.1	float64	HÁBITOS - Horas de contenido audiovisual
FRE_VIC_phoneGames	0.8	0.0	3.0	0.7	float64	HÁBITOS - Horas de juegos de móvil
FRE_VIC_SocialNetworks	0.1	0.0	1.0	0.2	float64	HÁBITOS - Horas de redes sociales
VIC_alcohol	0.4	0.0	3.0	0.6	float64	HÁBITOS - Alcohol tomado (en cervezas como unidad)
ALI_tmed	23.7	16.8	27.5	3.0	float64	AEMET - temperatura media (°C) (Barcelona)
ALI_prec	0.4	0.0	7.4	1.1	float64	AEMET - precipitación (mm) (Barcelona)
ALI_tmax	26.3	19.5	31.7	3.0	float64	AEMET - temperatura máxima (°C) (Barcelona)
ALI_trange	5.3	2.5	8.7	1.3	float64	FE - Rango de temperaturas (°C) (Barcelona)
EVA_happiness	7.2	4.5	9.0	1.0	float64	FE - Promedio de las evaluaciones subjetivas de la felicidad

Variable	Mean	Min	Max	std	Type	Descripción
EVA_tiredness	5.5	1.5	10.0	1.9	float64	FE - Promedio de las evaluaciones subjetivas del cansancio
FRE_sum	2.0	0.0	7.5	1.6	float64	FE - Suma de horas de tiempo libre
PRF_sum	2.6	0.0	11.0	3.5	float64	FE - Suma de horas de tiempo productivo
EAT_healthy	0.8	-0.5	3.0	0.7	float64	FE - Indicador de la sanidad en la alimentación
VIC	1.1	0.0	3.9	0.8	float64	FE - Indicador de los vicios consumidos
XX_ACT	0.8	0.0	1.0	0.4	int32	FE - Actividad social (1 si hay alguna de cualquier tipo)
ALI_week	7.1	1.0	14.0	3.8	int64	FE - Número de la semana desde el inicio del estudio
ALI_month_sin	-0.2	-0.9	0.5	0.5	float64	FE - Codificación cíclica del mes (sinus)
ALI_month_cos	-0.9	-1.0	-0.5	0.2	float64	FE - Codificación cíclica del mes (cosinus)
ALI_day_of_week_sin	0.0	-1.0	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (sinus)
ALI_day_of_week_cos	0.0	-0.9	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (cosinus)

Cuadro D.5: Apéndice - Resumen de variables de "Ajara" - 'Personal'

D.6 Charlie

Variable	Mean	Min	Max	std	Type	Descripción
date	2023-06-23	2023-05-13	2023-08-04	24	date-time	HÁBITOS - Fecha
TIR_sleep_time	6.0	1.0	10.0	1.7	float64	HÁBITOS - Horas de sueño
EVA_happiness	5.9	3.0	9.0	1.4	float64	HÁBITOS - Evaluación subjetiva (0-10) de la felicidad
EVA_TIR_tiredness	6.7	4.0	10.0	1.2	float64	HÁBITOS - Evaluación subjetiva (0-10) del cansancio
SPO	0.7	0.0	2.3	0.8	float64	HÁBITOS - Horas de deporte
PRF_uni	1.6	0.0	7.0	2.2	int32	HÁBITOS - Universidad
PRF_reading	0.3	0.0	1.5	0.6	float64	HÁBITOS - Horas de lectura
PRF_personalProjects	0.7	0.0	2.5	1.1	float64	HÁBITOS - Horas dedicadas a proyectos personales
XX_EAT_eating_out	0.4	0.0	1.0	0.5	int32	HÁBITOS - Comer fuera
XX_ACT_Family	0.1	0.0	1.0	0.3	int32	HÁBITOS - Actividad social en familia
XX_ACT_Couple	0.1	0.0	1.0	0.3	int32	HÁBITOS - Actividad social en pareja
XX_ACT_Friends	0.1	0.0	1.0	0.3	int32	HÁBITOS - Actividad social con amigos
XX_VIC_smoke	0.2	0.0	1.0	0.4	int32	HÁBITOS - Fumar tabaco
XX_VIC	0.2	0.0	1.0	0.4	int32	HÁBITOS - Fumar cannabis
XX_VIC_alcohol	0.1	0.0	1.0	0.3	int32	HÁBITOS - Alcohol tomado (en cervezas como unidad)
VIC_coffee	1.3	0.0	4.0	1.3	float64	HÁBITOS - Cafés tomados

Variable	Mean	Min	Max	std	Type	Descripción
EVA_stress	6.5	3.0	10.0	1.2	float64	HÁBITOS - Evaluación subjetiva (0-10) del estrés
FRE_VIC_videogames	1.4	0.0	7.5	3.0	float64	HÁBITOS - Horas de videojuegos
XX_EAT_Breakfast	0.3	0.0	1.0	0.5	int32	HÁBITOS - Desayunar
ALI_tmax	25.5	15.3	36.3	5.4	float64	OPEN-METEO - temperatura máxima (°C) (Italia)
ALI_tmed	20.4	11.4	29.9	4.8	float64	OPEN-METEO - temperatura media (°C) (Italia)
ALI_prec	2.9	0.0	30.1	5.3	float64	OPEN-METEO - precipitación (mm) (Italia)
ALI_trange	10.4	4.9	15.2	2.4	float64	FE - Rango de temperaturas (°C) (Italia)
PRF_sum	2.6	0.0	8.5	2.3	float64	FE - Suma de horas de tiempo productivo
VIC	1.6	0.0	7.7	2.1	float64	FE - Indicador de los vicios consumidos
XX_ACT	0.3	0.0	1.0	0.4	int32	FE - Actividad social (1 si hay alguna de cualquier tipo)
FRE_sum	8.6	0.3	15.0	3.3	float64	FE - Suma de horas de tiempo libre
ALI_week	7.2	1.0	13.0	3.5	int64	FE - Número de la semana desde el inicio del estudio
ALI_month_sin	-0.1	-0.9	0.5	0.4	float64	FE - Codificación cíclica del mes (sinus)
ALI_month_cos	-0.9	-1.0	-0.5	0.1	float64	FE - Codificación cíclica del mes (cosinus)
ALI_day_of_week_sin	0.0	-1.0	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (sinus)
ALI_day_of_week_cos	0.0	-0.9	1.0	0.7	float64	FE - Codificación cíclica del día de la semana (cosinus)

Cuadro D.6: Apéndice - Resumen de variables de "Charlie" - 'Personal'

Resultados de “Mack” y “Edogawa”

A continuación, mostraremos los cuadros que recogen los resultados de los participantes que no fueron seleccionados: “Mack” y “Edogawa”.

Feature	RFR_0.247	GBR_0.272	XGBR_0.259	SVR_0.222	Ensemble	Linear_coef
XX_ALI_lonely	0.041	0.049	0.071	0.037	0.049	-1.014
ALI_trange	0.055	0.066	0.030	0.037	0.046	-8.581
ALI_day_of_week_sin	0.046	0.051	0.044	0.037	0.044	-0.025
ALI_tmax	0.048	0.049	0.022	0.046	0.041	34.071
ALI_tmmed	0.048	0.050	0.021	0.044	0.041	-27.392
PRF_FRE_transport	0.044	0.043	0.018	0.055	0.040	4.536
FRE_sum	0.059	0.055	0.025	0.019	0.039	-11.721
TIR_time_get_up	0.048	0.048	0.027	0.032	0.039	0.372
EAT_healthy	0.044	0.041	0.028	0.019	0.033	3.114
PRF_sum	0.049	0.044	0.022	0.018	0.033	-1.018
EVA_tiredness	0.032	0.039	0.019	0.039	0.032	-0.139
FRE_VIC_audiovisual_content	0.039	0.039	0.025	0.021	0.031	-1.625
ALI_week	0.037	0.034	0.030	0.019	0.030	-0.395
VIC	0.035	0.043	0.018	0.023	0.030	157.359
VIC_alcohol	0.021	0.028	0.014	0.051	0.028	-8.119
ALI_imposing	0.020	0.015	0.035	0.024	0.024	-0.056
EVA_TIR_sleep_quality	0.021	0.022	0.020	0.029	0.023	0.084
ACT_bravery	0.024	0.023	0.015	0.027	0.022	-0.111
VIC_cigars	0.025	0.025	0.015	0.022	0.022	-154.890
PRF_study	0.027	0.020	0.014	0.019	0.020	0.261
XX_PRF_ACT_psychologist	0.013	0.001	0.052	0.010	0.020	-0.893
PRF_class	0.011	0.007	0.028	0.031	0.020	0.091
ALI_month_sin	0.019	0.019	0.030	0.008	0.019	-1.525
XX_EAT_eating_out	0.013	0.022	0.015	0.025	0.019	1.271
ALI_month_cos	0.017	0.017	0.025	0.004	0.016	-2.335
XX_ALI_cry	0.009	0.005	0.035	0.012	0.015	0.017
ALI_day_of_week_cos	0.013	0.016	0.012	0.014	0.014	-0.089
PRF_singing	0.006	0.008	0.011	0.024	0.012	0.304
VIC_coffee	0.009	0.011	0.010	0.017	0.012	-5.085
XX_EAT_VIC_meat	0.006	0.006	0.020	0.012	0.011	-23.962
XX_ACT_Roberto	0.006	0.010	0.016	0.012	0.011	-0.095
FRE_VIC_selfPleasure	0.005	0.006	0.015	0.016	0.010	-14.910
XX_OTH_wellDressed	0.009	0.009	0.014	0.009	0.010	0.089
XX_OTH_paintedNails	0.008	0.006	0.008	0.017	0.010	0.709
XX_EAT_fried	0.007	0.009	0.015	0.008	0.010	1.147

Cuadro E.1: Apéndice - Importancia de características según modelos y ensamble y coeficientes lineales - “Mack” - ‘Personal’

Prefix	sum_imp	mean_imp
'ALI_': Aspectos externos, ajenos	0.3481	0.0290
'TIR_': Cansancio	0.0962	0.0241
'VIC_': Vicios o hábitos perjudiciales	0.1458	0.0182
'PRF_': Tiempo provechoso	0.1538	0.0171
'FRE_': Tiempo libre	0.0719	0.0144
'EAT_': Alimentación	0.0830	0.0138
'ACT_': Actividades sociales	0.1013	0.0092

Cuadro E.2: Apéndice - Importancia agrupada según tópicos - "Mack" - 'Personal'

Feature	RFR 0.247	GBR 0.272	XGBR 0.259	SVR 0.222	Ensemble	Linear_coef
ALI_tmax	0.143	0.159	0.083	0.089	0.117	-14.119
ALI_tmed	0.114	0.106	0.067	0.056	0.085	15.388
ALI_trange	0.092	0.074	0.054	0.098	0.080	5.009
EVA_TIR_tiredness	0.048	0.089	0.095	0.081	0.078	-0.071
ALI_day_of_week_sin	0.074	0.072	0.087	0.051	0.071	-0.707
FRE_sum	0.074	0.087	0.053	0.054	0.066	-0.071
VIC	0.084	0.069	0.054	0.044	0.062	0.826
TIR_sleep_time	0.027	0.064	0.059	0.093	0.061	0.179
ALI_week	0.062	0.045	0.056	0.039	0.050	-0.219
ALI_day_of_week_cos	0.050	0.039	0.037	0.071	0.050	-0.256
PRF_Work	0.020	0.029	0.041	0.087	0.045	-0.306
XX_SPO_sportTime	0.031	0.027	0.062	0.050	0.043	0.272
VIC_alcohol	0.033	0.041	0.044	0.039	0.039	-0.378
XX_EAT_healthy	0.030	0.019	0.039	0.042	0.033	1.265
ALI_prec	0.041	0.025	0.034	0.027	0.032	0.246
XX_VIC_cigars	0.014	0.013	0.034	0.015	0.019	-0.939
ALI_month_sin	0.018	0.009	0.030	0.013	0.018	-0.078
XX_ACT	0.016	0.011	0.015	0.024	0.017	0.628
ALI_month_cos	0.019	0.010	0.028	0.003	0.015	-0.021
XX_ACT_Family	0.010	0.010	0.012	0.021	0.014	-0.942

Cuadro E.3: Apéndice - Importancia de características según modelos y ensamble y coeficientes lineales - "Edogawa" - 'Personal'

Prefix	sum_imp	mean_imp
'TIR_': Cansancio	0.1389	0.0694
'FRE_': Tiempo libre	0.0659	0.0659
'ALI_': Aspectos externos, ajenos	0.5172	0.0575
'PRF_': Tiempo provechoso	0.0452	0.0452
'SPO_': Deporte	0.0432	0.0432
'VIC_': Vicios o hábitos perjudiciales	0.1205	0.0402
'EAT_': Alimentación	0.0330	0.0330
'ACT_': Actividades sociales	0.0362	0.0121

Cuadro E.4: Apéndice - Importancia agrupada según tópicos - "Edogawa" - 'Personal'

Bibliografía

- [AEMET 023] AEMET. *API de Información Meteorológica de la Agencia Estatal de Meteorología*, (2023). API, Available at https://www.aemet.es/es/datos_abiertos/AEMET_Data_API. (Cited on pages 26 and 47.)
- [Amat Rodrigo 2017] Joaquín Amat Rodrigo. *Árboles de decisión, random forest, gradient boosting y C5.0*. cienciadedatos.net, 2017. Available at https://cienciadedatos.net/documentos/33_arboles_decision_random_forest_gradient_boosting_c50#Random_Forest. (Cited on page 19.)
- [Berwick 023] B Berwick and Idiot Village. *An Idiot's guide to Support vector machines (SVMs)*, 2003 (Accessed: August 2023). Available at <https://web.mit.edu/6.034/wwwbob/svm.pdf>. (Cited on pages 20, 41 and 80.)
- [Bhardwaj 023] Ashutosh Bhardwaj. *Silhouette Coefficient*, May 2020 (Accessed: August 2023). Available at <https://towardsdatascience.com/silhouette-coefficient-validating-clustering-techniques-e976bb81d10c>. (Cited on page 16.)
- [Brownlee 023] Jason Brownlee. *Ordinal and One-Hot Encodings for Categorical Data*, June 2020 (Accessed: July 2023). Available at <https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/>. (Cited on page 35.)
- [Chowdhury 023] MZI Chowdhury and TC Turin. *Variable selection strategies and its importance in clinical prediction modelling*, February 2020 (Accessed: July 2023). Available at <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7032893/>. (Cited on page 35.)
- [Espinosa-Zúñiga 2020] Javier Jesús Espinosa-Zúñiga. *Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito*. Ingeniería Investigación y Tecnología, 2020. Available at <https://www.revistaingenieria.unam.mx/numeros/2020/v21n3-02.pdf>. (Cited on page 19.)
- [Francia 023] Gianluca Francia. *¿Qué tipos de vicios existen?*, May 2021 (Accessed: July 2023). Available at <https://www.psicologia-online.com/que-tipos-de-vicios-existen-5748.html>. (Cited on page 49.)

- [Garmin 023] Garmin. *Garmin Venu2 [Dispositivo de seguimiento de actividad]*, 2023). (Cited on pages 26 and 47.)
- [Hernández Aburto 2017] Karen Hernández Aburto, Marcia Muñoz Rioseco and Emilio Moyano-Díaz. *Concepto de Felicidad en Adultos de Sectores Populares*. SciELO - Scientific Electronic Library Online, 2017. Available at <https://www.scielo.br/j/paideia/a/dTXF4pcTmMgPXz7dmZgxYJb/?lang=es#>. (Not cited.)
- [Hervé 2010] Abdi Hervé. *Holm's Sequential Bonferroni Procedure*. In Neil Salkind (Ed.), *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage., 2010. <https://personal.utdallas.edu/~herve/abdi-Holm2010-pretty.pdf>. (Cited on page 18.)
- [how2stats 023] how2stats. *Yates' Correction*, (Accessed: July 2023). Available at <http://www.how2stats.net/2011/09/yates-correction.html>. (Cited on pages 18 and 38.)
- [Huxley 1932] Aldous Huxley. *Un mundo feliz*. Debolsillo, 1932. (Cited on page 1.)
- [Kong 2021] M Kong, L Li, R Wu and X Tao. *An Empirical Study of Learning Based Happiness Prediction Approaches*. Human-Centric Intelligent Systems, 2021. Available at <https://www.atlantis-press.com/journals/hcis/125958421>. (Cited on pages 3 and 8.)
- [León Guzmán 023] Elizabeth León Guzmán. *Métricas para la validación de Clustering*, July 2019 (Accessed: July 2023). Available at https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf. (Cited on pages 16 and 17.)
- [Lincoln sf] Abraham Lincoln. *Dame seis horas para cortar un árbol y pasaré las primeras cuatro afilando el hacha*, s.f. (Cited on page 29.)
- [Álison S 016] Xavier Álison S. *Loop - Analizador de Hábitos*, (2016). Aplicación móvil, Available at <https://play.google.com/store/apps/details?id=org.isoron.uhabits&pli=1>. (Cited on pages 24 and 46.)
- [Margot 2007] Jean-Paul Margot. *La felicidad*. Praxis Filosófica, Universidad del Valle, 2007. Available at <https://www.redalyc.org/pdf/2090/209014642004.pdf>. (Cited on page 11.)

- [Martínez Heras 023] Jose Martínez Heras. *Clustering (Agrupamiento), K-Means con ejemplos en python*, September 2020 (Accessed: July 2023). Available at <https://www.iartificial.net/clustering-agrupamiento-kmeans-ejemplos-en-python/#Resumen>. (Cited on pages 15 and 45.)
- [McDonald 023] John H McDonald. *Multiple Comparisons*, April 2022 (Accessed: August 2023). Available at [https://stats.libretexts.org/Bookshelves/Applied_Statistics/Biological_Statistics_\(McDonald\)/06%3A_Multiple_Tests/6.01%3A_Multiple_Comparisons](https://stats.libretexts.org/Bookshelves/Applied_Statistics/Biological_Statistics_(McDonald)/06%3A_Multiple_Tests/6.01%3A_Multiple_Comparisons). (Cited on pages 18 and 19.)
- [Montagud Rubí 023] Nahum Montagud Rubí. *Los 14 tipos de vicios y sus características*, July 2020 (Accessed: July 2023). Available at <https://psicologiyamente.com/psicología/tipos-de-vicios>. (Cited on page 49.)
- [Moya 023] Ricardo Moya. *Selección del número óptimo de Clusters*, September 2016 (Accessed: August 2023). Available at <https://jarroba.com/seleccion-del-numero-optimo-clusters/>. (Cited on page 15.)
- [Nestlé 023] Nestlé. *6 grupos de alimentos que debes conocer para una dieta saludable*, (Accessed: July 2023). Available at <https://www.nestle-contigo.co/elige-tu-medida/6-grupos-de-alimentos-que-debes-conocer-para-una-dieta-saludable>. (Cited on page 48.)
- [OMS 023] OMS. *Dieta sana*. Organización Mundial de la Salud, (Accessed: July 2023). Available at https://www.who.int/es/health-topics/healthy-diet#tab=tab_1. (Cited on page 48.)
- [Open-Meteo 023] Open-Meteo. *Open-Meteo Historical Weather API*, (2023). API, Available at <https://open-meteo.com/>. (Cited on page 26.)
- [Python 023] Python. *Python Software Foundation: Lenguaje de programación Python [versión '3.11.3']*, 2023). Available at <https://www.python.org/>. (Cited on page 14.)
- [Rençberoğlu 023] Emre Rençberoğlu. *Fundamental Techniques of Feature Engineering for Machine Learning*, April 2019 (Accessed: July 2023). Available at <https://towardsdatascience.com/feature-engineering-for-machine-learning-3a5e293a5114>. (Cited on page 35.)

- [Reverté 022] Francesc Xavier Reverté. *Estudio de la influenciabilidad de actividades en el estado de ánimo [Google Forms]*, (Published: June 2022). Available at https://docs.google.com/forms/d/e/1FAIpQLSz58izClGK79RUXy2tE_W9ZyYszyav-x2GrdL722BIh0xBlw/viewform?usp=sf_link. (Cited on page 23.)
- [Ruano Ruano 1997] I Ruano Ruano and M Serra Pujol. *HABITOS DE VIDA EN UNA POBLACION ESCOLAR DE MATARÓ (BARCELONA) ASOCIADOS AL NÚMERO DE VECES DIARIAS QUE VE TELEVISIÓN Y AL CONSUMO DE AZÚCARES*. Instituto de investigación Epidemiológica y Clínica (IREC). PASS, 1997. Available at https://scielo.isciii.es/scielo.php?pid=S1135-57271997000500007&script=sci_arttext. (Cited on pages 3 and 7.)
- [Tomar 023] Anmol Tomar. *Stop Using Elbow Method in K-Means Clustering*, August 2023 (Accessed: August 2023). Available at <https://builtin.com/data-science/elbow-method>. (Cited on pages 15 and 16.)
- [Veenhoven 1984] R Veenhoven. *Conditions of Happiness*. Erasmus Universiteit Rotterdam, 1984. Available at <https://personal.eur.nl/veenhoven/Pub1980s/84a-full.pdf>. (Cited on page 11.)
- [Veenhoven 2013] R Veenhoven. *MÁS FELICIDAD PARA UN MAYOR NÚMERO DE PERSONAS ¿Es posible esto en México?* Erasmus Universiteit Rotterdam, 2013. Available at <https://personal.eur.nl/veenhoven/Pub2010s/2013m-fulls.pdf>. (Cited on pages 4, 8 and 11.)
- [Wagstaff 2001] Kiri Wagstaff, Claire Cardie, Seth Rogers and Stefan Schroedl. *Constrained K-means Clustering with Background Knowledge*. Department of Computer Science, Cornell University and DaimlerChrysler Research and Technology Center, 2001. Available at https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf. (Cited on pages 15 and 37.)
- [Waine 023] W Waine. *Tests for Two or More Independent Samples, Discrete Outcome*, September 2016 (Accessed: July 2023). Available at https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_HypothesisTesting-ChiSquare/BS704_HypothesisTesting-ChiSquare3.html. (Cited on pages 17, 18 and 38.)
- [Walsh 2004] Bruce Walsh. *Multiple Comparisons: Bonferroni Corrections and False Discovery Rates*. Lecture Notes for EEB 581, 2004. Available at https://physiology.med.cornell.edu/people/banfelder/qbio/resources_2008/1.5_Bonferroni_FDR.pdf. (Cited on pages 18 and 19.)

