Universidad Politécnica
de Madrid

Escuela Técnica Superior de
Ingenieros Informáticos

Grado en Ciencia de Datos e Inteligencia Artificial

Trabajo Fin de Grado

Análisis de Rendimiento en Equipos Profesionales mediante la Integración de Modelos de Aprendizaje Automático

Autor: Prisco García-Consuegra Martín

Tutor: Sergio Paraíso Medina

Madrid, 06/2024

Este Trabajo Fin de Grado se ha depositado en la ETSI Informáticos de la Universidad Politécnica de Madrid para su defensa.

Trabajo Fin de Grado
Grado en Ciencia de Datos e Inteligencia Artificial
Título: Análisis de Rendimiento en Equipos Profesionales mediante la Integración de Modelos de Aprendizaje Automático
06 / 2024

Autor: Prisco García-Consuegra Martín

Tutor:
Sergio Paraíso Medina
Departamento de Lenguajes y Sistemas Informáticos e Ingeniería del Software
ETSI Informáticos
Universidad Politécnica de Madrid

Resumen

En la actualidad, optimizar el desempeño de los equipos profesionales surge como una prioridad estratégica para la competitividad y el éxito organizacional. El desafío de fomentar la colaboración efectiva y alcanzar metas en entornos laborales dinámicos y complejos ha llevado a explorar la diversidad de roles y personalidades dentro de los equipos, reconociendo que estas interacciones pueden influir en la eficiencia general del equipo.

Este proyecto se embarcó en abordar este desafío mediante la investigación y el análisis del rendimiento de equipos en el ámbito laboral, prestando especial atención a las características de personalidad de sus miembros. Se recurrió al aprendizaje automático para aprovechar los datos disponibles sobre las dinámicas del equipo y las características individuales, integrando medidas de evaluación de personalidad con algoritmos avanzados. El objetivo primordial fue desarrollar un modelo que facilitara la formación de equipos eficientes y cohesionados, teniendo en cuenta las características de personalidad de sus integrantes.

El proyecto se dividió en diversas etapas, comenzando con un análisis exhaustivo del estado del arte de las tecnologías pertinentes. Se seleccionó el test *Big Five* como metodología principal debido a su amplia aceptación y su capacidad para proporcionar una comprensión profunda de las características individuales relevantes para el rendimiento laboral. Además, se llevó a cabo un análisis empírico del efecto de diferentes rasgos de personalidad en el rendimiento académico, proporcionando así una base sólida para la propuesta de optimización del rendimiento del equipo profesional.

Para el desarrollo del modelo, se utilizaron técnicas y librerías basadas en Python en un entorno de programación público externo, aprovechando su versatilidad y potencia computacional. Los datos se recopilaron a través de un test de personalidad interactivo en línea, utilizando marcadores de los cinco grandes factores de personalidad. El preprocesamiento de datos desempeñó un papel fundamental en la limpieza y preparación de los datos para el análisis posterior.

Se aplicaron técnicas de *clustering* para identificar patrones significativos en las características de personalidad de los participantes, segmentándolos en diferentes grupos. Se observaron diferencias significativas entre los *clusters* en términos de características de personalidad, lo que sugiere la importancia de considerar estas diferencias en la formación y gestión de equipos.

Este proyecto ha demostrado que la formación de equipos profesionales puede beneficiarse enormemente de la consideración de las características de personalidad individuales. Al utilizar un modelo de *clustering* basado en el cuestionario *Big Five*, se ha podido segmentar a los participantes en *clusters* que reflejan distintas combinaciones de características de personalidad. Estos resultados proporcionan una herramienta poderosa para la selección y gestión de equipos, asegurando un rendimiento óptimo y sostenible en las organizaciones.

Abstract

Nowadays, optimizing the performance of professional teams is emerging as a strategic priority for organizational competitiveness and success. The challenge of fostering effective collaboration and achieving goals in dynamic and complex work environments has led to exploring the diversity of roles and personalities within teams, recognizing that these interactions can influence overall team efficiency.

This project embarked on addressing this challenge by investigating and analyzing the performance of teams in the workplace, paying particular attention to the personality characteristics of their members. Machine learning was used to leverage available data on team dynamics and individual characteristics, integrating personality assessment measures with advanced algorithms. The primary objective was to develop a model that would facilitate the formation of efficient and cohesive teams, considering the personality characteristics of its members.

The project was divided into several stages, starting with a comprehensive state-of-the-art analysis of the relevant technologies. The Big Five test was selected as the main methodology due to its wide acceptance and its ability to provide an in-depth understanding of individual characteristics relevant to job performance. In addition, an empirical analysis of the effect of different personality traits on academic performance was conducted, thus providing a solid basis for the proposed optimization of professional team performance.

For the development of the model, Python-based techniques and libraries were used in an external public programming environment, taking advantage of its versatility and computational power. Data were collected through an online interactive personality test, using markers of the Big Five personality factors. Data preprocessing played a key role in cleaning and preparing the data for subsequent analysis.

Clustering techniques were applied to identify significant patterns in the personality characteristics of the participants, segmenting them into different groups. Significant differences were observed between clusters in terms of personality characteristics, suggesting the importance of considering these differences in team building and management.

This project has shown that the formation of professional teams can benefit greatly from the consideration of individual personality characteristics. By using a clustering model based on the Big Five questionnaire, it has been possible to segment participants into clusters reflecting different combinations of personality characteristics. These results provide a powerful tool for team selection and management, ensuring optimal and sustainable performance in organizations.

Tabla de Contenidos

Tabla de Ilustraciones

1 Introducción

En el panorama empresarial actual, la optimización del rendimiento de los equipos profesionales se ha convertido en un objetivo estratégico de máxima relevancia para la competitividad y el éxito de las organizaciones. La capacidad de los equipos para colaborar de manera efectiva y alcanzar sus objetivos en un entorno laboral cada vez más dinámico y complejo es crucial para mantener la ventaja competitiva y adaptarse a los cambios del mercado. Sin embargo, la formación y gestión de equipos eficientes presenta desafíos significativos, especialmente en lo que respecta a la diversidad de roles y personalidades que coexisten dentro de ellos.

El presente proyecto se propone abordar este desafío central mediante la investigación y análisis del rendimiento de los equipos en entornos laborales, poniendo especial atención en la variedad de roles y características de personalidad presentes en dichos equipos. Nos enfrentamos a un problema complejo: la interacción entre individuos con diferentes habilidades y estilos de trabajo puede generar tanto sinergias positivas como conflictos, lo que repercute directamente en la eficiencia y el rendimiento del equipo en su conjunto. La falta de herramientas adecuadas para evaluar y comprender estas dinámicas limita la capacidad de las organizaciones para optimizar su desempeño y aprovechar al máximo el potencial de sus equipos. Por lo tanto, es crucial desarrollar métodos avanzados que permitan analizar de manera integral las características individuales y su influencia en el rendimiento del equipo.

Para abordar este problema de manera efectiva, se utilizarán técnicas de aprendizaje automático que aprovechen los datos disponibles sobre las dinámicas de equipo y las características individuales de los miembros. Se explorarán diversas herramientas de evaluación de personalidad ampliamente utilizadas en el ámbito empresarial, como el cuestionario *Big Five*, para comprender mejor las interacciones interpersonales y su impacto en el rendimiento colectivo. La integración de estas medidas con algoritmos avanzados de aprendizaje automático permitirá la construcción de un modelo que facilite la formación de equipos eficientes y cohesionados.

De esta manera el objetivo principal de este proyecto es desarrollar un modelo de aprendizaje automático capaz de aportar valor en la formación de equipos profesionales, basado en las características individuales de los componentes. Este modelo permitirá a los encargados de selección de las empresas conformar grupos de trabajo teniendo en cuenta las características de personalidad de sus integrantes, facilitando así la creación de equipos equilibrados y eficientes.

Para alcanzar este objetivo, se plantean las siguientes metas o subobjetivos:

1. Análisis y estado del arte de tecnologías:
   En aras de satisfacer esta meta, se realizará una investigación de los diferentes tipos de equipos de trabajo para los que se puede implementar la propuesta, así como una exploración de los distintos métodos de evaluación de roles existentes con una selección de la metodología para

la propuesta de este proyecto. Además, se abordará un análisis de las diferentes técnicas de aprendizaje automático aplicables a la propuesta y una exploración de las distintas bases de datos accesibles. Por último, se investigarán estudios similares y metodología aplicada.

2. Estudio de necesidades y requisitos:
   Esta meta pretende comprender la situación previa al desarrollo del modelo, atendiendo especialmente a aquello referente a los datos y la estrategia de aprendizaje automático. De esta manera, para alcanzar esta meta se realizará una consideración y análisis de los datos seleccionados, así como de las tecnologías y conocimientos técnicos que se aplicarán en el desarrollo del modelo.

3. Desarrollo del modelo:
   En este apartado se realizará una administración y actualización constante del trabajo realizado. Se comenzará con preprocesado y limpieza de datos orientado a satisfacer el objetivo principal. Se dará paso a la creación del modelo de aprendizaje automático que buscamos, trabajando con diferentes técnicas y metodologías. Se realizará siempre que se dé la ocasión una depuración de mejoras, de cara a optimizar el modelo.

4. Resultados:
   Se realizará un análisis de resultados con el mejor modelo optimizado para la propuesta y se estudiaran las conclusiones obtenidas.

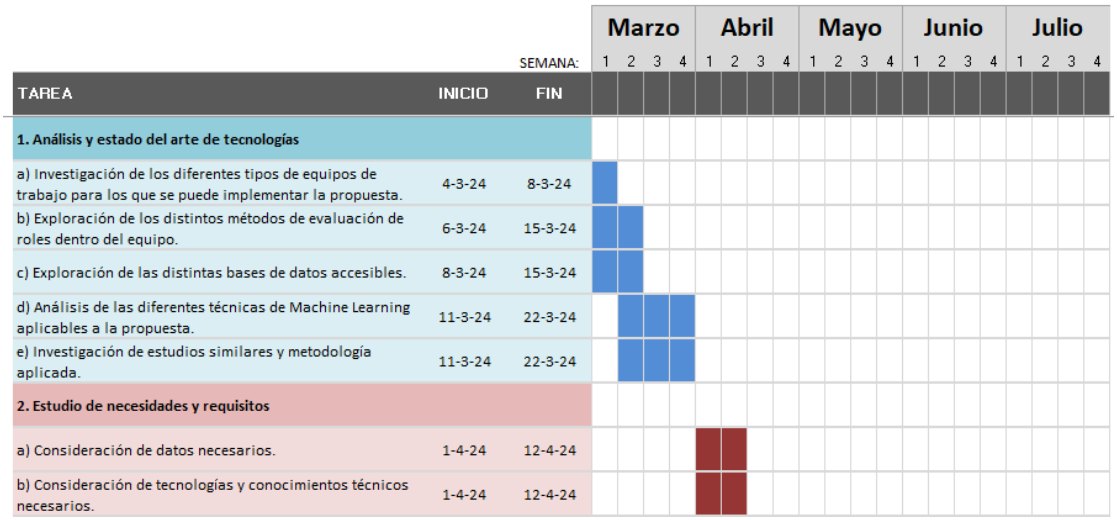

Figura 1: Diagrama de Gantt para los objetivos del proyecto

2 Estado del Arte

2.1 Investigación de los diferentes equipos de trabajo

En un entorno empresarial cada vez más dinámico y complejo, comprender cómo la interacción entre los distintos roles y personalidades influye en el desempeño colectivo de un equipo se ha convertido en una prioridad estratégica. Ejemplos concretos de la aplicación de esta propuesta pueden observarse en una amplia gama de sectores y contextos laborales.

En la industria tecnológica, se observa la prevalencia de equipos multidisciplinarios, como los equipos de desarrollo de software, donde ingenieros de software, diseñadores de UX/UI, especialistas en marketing y expertos en gestión de proyectos colaboran en el desarrollo de productos digitales. Estos equipos, por ejemplo, ilustran la variedad de roles y habilidades presentes en entornos laborales modernos y resaltan la necesidad de una comprensión profunda de cómo estas diferencias pueden influir en el rendimiento colectivo.

Además, el panorama laboral actual incluye la presencia creciente de equipos virtuales o remotos, como evidenciado en empresas que adoptan políticas de trabajo flexible o que operan a nivel internacional. Equipos de esta índole, como los departamentos de atención al cliente distribuidos geográficamente, ofrecen un caso de estudio interesante debido a los desafíos únicos que enfrentan en términos de comunicación, coordinación y cohesión de equipo.

A su vez, se identifican equipos de investigación científica, donde profesionales con especialidades diversas, desde biología hasta física, colaboran en proyectos de vanguardia, mostrando cómo la combinación de distintos perfiles contribuye a la generación de conocimiento y avances en sus respectivos campos.

En este sentido, resulta evidente que la aplicación de la propuesta de análisis y optimización del rendimiento de equipos no está limitada a un sector o tipo de equipo en particular. Por el contrario, la dinámica de equipo y la influencia de los roles en el rendimiento colectivo es de naturaleza general y abarca una amplia gama de contextos laborales. De este modo, este proyecto se centrará en explorar las diferentes personalidades existentes en un equipo de trabajo, con el objetivo de ofrecer una comprensión integral, aplicable en múltiples escenarios empresariales.

2.2 Exploración de los distintos métodos de evaluación

En esta sección se pretende hacer un reconocimiento general de los diferentes métodos utilizados a día de hoy en el entorno profesional para evaluar los roles existentes en un equipo de trabajo. Se realiza esta investigación como punto de contacto y vista general, no exhaustiva, de la situación actual, centrándose para cada método en su funcionalidad y usabilidad para la propuesta del proyecto.

2.2.1 Modelo de las cinco grandes dimensiones de la personalidad

Uno de los enfoques más prominentes en este dominio es el Modelo de los Cinco Grandes, ampliamente conocido como *Big Five*. Propuesto por los psicólogos Lewis Goldberg y Warren Norman en la década de 1960, la teoría de los cinco grandes rasgos de la personalidad, también denominada modelo de los Cinco Grandes, se fundamenta en la premisa de que existen cinco dimensiones fundamentales de la personalidad que explican las variaciones individuales en cognición, afecto y comportamiento.

Los Cinco Grandes rasgos comprenden la apertura a la experiencia, la conciencia, la extraversión, la amabilidad y el neuroticismo. Cada uno de estos factores se desglosa en múltiples subdimensiones que, de forma conjunta, delinean un perfil de personalidad único. A continuación, se detallan cada uno de estos factores.

- Apertura a la experiencia: Este atributo abarca la disposición hacia nuevas ideas, la creatividad y la imaginación. Individuos con altos niveles de apertura a la experiencia tienden a manifestar curiosidad intelectual, interés por el aprendizaje continuo y disposición para explorar nuevas actividades. En contraste, aquellos con bajos niveles de apertura a la experiencia prefieren la estabilidad y la rutina.

- Conciencia: Este rasgo refiere a la responsabilidad y la organización. Aquellos con altos niveles de conciencia son reconocidos por su fiabilidad, cumplimiento de obligaciones y mantenimiento de un entorno ordenado. Por el contrario, individuos con bajos niveles de conciencia pueden mostrar descuido y falta de confiabilidad.

- Extraversión: Este factor se relaciona con la sociabilidad y la energía. Personas con altos niveles de extraversión tienden a ser extrovertidas, amigables y disfrutar de la interacción social. Por el contrario, aquellos con bajos niveles de extraversión pueden mostrar timidez y preferir actividades solitarias.

- Amabilidad: Este atributo se refiere a la empatía y la cooperación. Individuos con altos niveles de amabilidad son caracterizados por su bondad,

compasión y disposición para colaborar. Por el contrario, aquellos con bajos niveles de amabilidad pueden mostrar menos empatía y tender hacia la competitividad.

- Neuroticismo: Este rasgo se refiere a la estabilidad emocional. Individuos con altos niveles de neuroticismo pueden experimentar emociones negativas con mayor intensidad y preocuparse en exceso. En contraste, aquellos con bajos niveles de neuroticismo suelen ser emocionalmente estables.

Una de las aplicaciones más frecuentes de la teoría de los Cinco Grandes radica en la evaluación de la personalidad en contextos laborales. Numerosas empresas emplean pruebas de personalidad basadas en este modelo para evaluar a candidatos durante procesos de selección. Dichas pruebas ofrecen información relevante sobre características de personalidad pertinentes para el ámbito laboral, lo que contribuye a la toma de decisiones informadas por parte de las organizaciones.

En el artículo [1], llevado a cabo por Barrick & Mount en 1991, revisa las distintas posiciones teóricas sobre las cinco grandes dimensiones de la personalidad, mostrando las semejanzas y diferencias entre las posturas teóricas se estudia el completo meta/análisis del poder predictivo de las cinco dimensiones de personalidad para el desempeño laboral. Distinguieron tres criterios de desempeño (habilidad laboral, habilidad de entrenamiento y datos personales) en cinco grupos (profesionales, políticos, gerentes, vendedores y trabajadores calificados y semi-calificados). Los autores son conscientes del hecho, a menudo demostrado empíricamente, que la personalidad es un predictor muy modesto del desempeño laboral. Ellos consideran a la taxonomía de las cinco grandes dimensiones como un esquema útil, puesto que organiza muchos rasgos de personalidad. Afirman que el consenso en número y contenido de los factores no son completos. No obstante, las cinco dimensiones y otras escalas son relativamente independientes de las capacidades cognitivas; por consiguiente, pueden contribuir a la predicción del desempeño laboral independientemente de las capacidades cognitivas.

2.2.2 DiSC

La metodología DiSC fue creada en 1928 por el psicólogo y escritor estadounidense William Marston. Presentó su método en el libro *Emotions of Normal People*, en el que explica cómo dos ejes producen cuatro factores que están conectados entre sí. Pese a que Marston no pretende encasillar a los individuos según una tipología específica, sí considera que los cuatro componentes individuales trabajan juntos para crear un todo. Mientras que una persona típica puede tener una o dos características dominantes, es la combinación de los cuatro componentes del DISC lo que define el

comportamiento de una persona. Los cuatro componentes del DISC crean un "maquillaje conductual" único para cada individuo.

DISC es un acrónimo de los cuatro principales perfiles de personalidad descritos en el modelo: (D)ominancia, (i)nfluencia, (S)teadiness (Firmeza) y (C)umplimiento o (C)onciencia.

- Las personas con personalidad **D** tienden a ser seguras de sí mismas y a hacer hincapié en la consecución de resultados.
- Las personas con personalidad **i** tienden a ser más abiertas y a hacer hincapié en las relaciones y en influir o persuadir a los demás.
- Las personas con personalidad **S** tienden a ser fiables y hacen hincapié en la cooperación y la sinceridad.
- Las personas con personalidad **C** tienden a hacer hincapié en la calidad, la precisión, la experiencia y la competencia.

A día de hoy DiSC es una herramienta de evaluación personal utilizada por más de un millón de personas cada año para ayudar a mejorar el trabajo en equipo, la comunicación y la productividad en el lugar de trabajo. Las organizaciones y los facilitadores utilizan estos perfiles como herramientas para ayudar a encender el cambio cultural, inspirando cambios de comportamiento duraderos que dan forma positiva a su fuerza de trabajo. Esta herramienta es considerada de gran utilidad en el entorno empresarial para aportar valor al entrenamiento, formación de equipos y liderazgo.

2.2.3 Los 9 roles de Belbin

Meredith Belbin, un destacado experto inglés en gestión de personas y gestión de equipos eficientes, ha dejado un legado significativo en el campo de los recursos humanos con su teoría de los roles de equipo, publicada en 1981. Esta teoría ha sido ampliamente reconocida y sigue siendo relevante en la actualidad como una metodología probada para maximizar el potencial tanto de individuos como de equipos en entornos laborales.

La teoría de Belbin identifica nueve roles distintivos, organizados en tres categorías principales: roles de acción, roles sociales y roles mentales.

1. Roles de acción:

    - El Finalizador: Se caracteriza por su dedicación, meticulosidad y enfoque en la finalización de tareas dentro de los plazos establecidos. Sin embargo, en momentos de estrés, puede mostrar tendencias perfeccionistas y una aversión a delegar responsabilidades.

- El Implementador: Destaca por su disciplina, fiabilidad y enfoque pragmático en la ejecución de tareas. En situaciones de presión, tiende a volverse inflexible y lento en la toma de decisiones.

- El Impulsor: Es dinámico, enérgico y orientado hacia la superación de obstáculos. Sin embargo, en momentos desfavorables, puede tender a la provocación y mostrar una menor tolerancia a la frustración.

2. Roles sociales:

    - El Coordinador: Se distingue por su madurez, capacidad para inspirar confianza y habilidad para facilitar la toma de decisiones grupales. En circunstancias adversas, puede mostrar pasividad y una expectativa excesiva de que otros asuman la iniciativa.

    - El Investigador de Recursos: Es extrovertido, entusiasta y hábil en el establecimiento de contactos. No obstante, puede perder interés fácilmente después del entusiasmo inicial.

    - El Cohesionador: Se caracteriza por su habilidad para fomentar la armonía y la cooperación dentro del equipo, así como por su diplomacia y empatía. En situaciones estresantes, puede mostrar indecisión y evitar el conflicto en exceso.

3. Roles mentales:

    - El Monitor Evaluador: Es estratégico, reflexivo y experto en evaluar diferentes escenarios. En momentos de tensión, puede carecer de energía y demostrar escepticismo, adoptando un enfoque demasiado crítico.

    - El Cerebro: Destaca por su imaginación, originalidad y habilidad para resolver problemas complejos. Sin embargo, puede enfrentar dificultades para comunicarse efectivamente y dirigir a otros en momentos de baja inspiración.

    - El Especialista: Se caracteriza por su especialización y dedicación en un área técnica específica. Su contribución se limita a un campo estrecho en el que posee un conocimiento profundo y habilidades técnicas destacadas.

Según la teoría de Belbin, todos los roles deben estar presentes en todos los equipos para funcionar como equipos de alto rendimiento. Sin embargo, es un factor interesante la siguiente idea expuesta en el artículo [2]: Los líderes necesitan tener en cuenta que los equipos de éxito necesitan una combinación adecuada de personas, de tal manera que los comportamientos asociados a los nueve Roles de Equipo estén representados. Esto no significa, necesariamente,

que cada equipo ha de contar con nueve personas como mínimo. Muchas personas se sienten cómodas desempeñando dos o tres roles de equipo (roles altos); pueden asumir otros tantos si les resulta necesario (roles medios); y el resto de roles prefieren no adoptarlos en absoluto (roles bajos).

2.2.4 Eneagrama

Al abordar el Eneagrama, es crucial distinguir entre el símbolo y los nueve tipos de personalidad que representa [3], ya que cada uno de estos elementos tiene su propio trasfondo histórico y desarrollo. De hecho, esta herramienta de autoconocimiento no surge de una única fuente; es más bien una síntesis compleja que acoge diversas tradiciones espirituales, religiosas y los avances contemporáneos en psicología.

En cuanto al origen del símbolo del círculo de nueve puntas, no se puede establecer con certeza su primera aparición. La creencia predominante sugiere que se originó en la región de Caldea en Babilonia (Mesopotamia) hace más de 2.500 años. Además, este diagrama se relaciona con figuras destacadas de la antigua Grecia, como Pitágoras (569-475 a.C.), Platón (427-347 a.C.) y Plotino (205-270 d.C.). Tanto los caldeos como los griegos consideraban el número nueve como "divino", utilizándolo para representar lo absoluto, lo infinito y la unidad primordial de donde emana y abarca todo. [4]

Volviendo a la tierra y utilizando esta sección únicamente como vista general, sin entrar en detalles, en la herramienta se distinguen 9 tipos diferentes de personalidad (eneatipos). Cada uno de ellos viene representado por un número del 1 al 9 y se relacionan entre sí por tríadas y alas (obsérvese Fig.2). Diferentes teorías sobre el eneagrama discuten sobre el poder de la relación entre los eneatipos. Por ejemplo, un individuo con eneatipo 2 predominante, tendrá influencia de las alas 1 y 3. A su vez, estará influido por las características de los eneatipos 4 y 8.

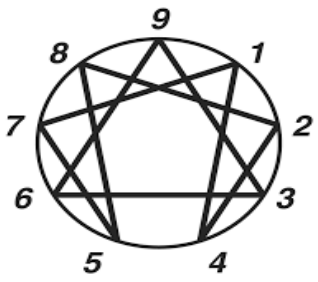

Figura 2: Eneagrama

Esta última herramienta coincidiendo con la mayoría de las diferenciaciones de personalidad y algunas de las metodologías que hemos comentado con anterioridad, pese a simplificar y racionalizar con nombres la complejidad de una personalidad, el conocimiento profundo de las habilidades, capacidades y aptitudes de un individuo emana del análisis de la predominancia y el equilibrio entre todos los roles y tipos que se estudien.

2.2.5 Selección de metodología

Se ha analizado un subconjunto de las herramientas existentes a día de hoy para aplicar una diferenciación de roles que aporten valor en la eficiencia de un equipo de trabajo. Finalmente, la metodología seleccionada para este proyecto es el test de las cinco grandes dimensiones de la personalidad o Big Five. Esta decisión se fundamenta en su amplia aceptación y su capacidad para proporcionar una comprensión profunda de las características individuales relevantes para el rendimiento en entornos laborales. La aplicación práctica de esta metodología en entornos empresariales respalda la perfecta alineación con el objetivo del proyecto de desarrollar un modelo de *clustering* para la formación de equipos, ya que proporciona una base sólida y confiable para la caracterización y agrupamiento de individuos en función de su personalidad.

Además, como factor importante a la hora de tomar la decisión, se ha tenido en cuenta la usabilidad de la metodología para el modelo de *clustering* que se pretende desarrollar. En aras de crear este modelo, la elección de las cinco grandes dimensiones se presenta como una opción idónea. Esto se debe a que estas dimensiones proporcionan una estructura clara y robusta para caracterizar la personalidad de cada individuo, lo que facilita el proceso de agrupamiento. Al utilizar esta metodología como base para el modelo de *clustering*, se puede capturar la complejidad de la personalidad de manera representativa, esto permite identificar patrones y relaciones significativas entre los miembros del equipo.

2.3 Exploración de las bases de datos accesibles

Al realizar esta investigación, se ha detectado que encontrar bases de datos de perfiles dentro de un equipo de trabajo real es una ardua tarea y realmente son casi inexistentes para el público, pues son datos confidenciales de las empresas y se requiere de derechos y permisos que el autor de este proyecto no posee. De esta manera, el enfoque del proyecto no puede ser en sí la variación del método de trabajo para garantizar la eficiencia de un equipo real.

Sin embargo, sí se ha tenido acceso a distintas bases de datos que recogen información sobre algunos de los métodos mencionados en este apartado. La mayoría de bases de datos abiertas disponibles parecen ser a priori insuficientes, ya sea por la cantidad de datos, por la calidad de los mismos o por su aplicabilidad en este proyecto.

Finalmente, dada la elección de trabajar con el modelo Big Five, se expone una tabla comparativa de la vista general de las bases de datos a las que se ha tenido acceso para poder comprender visualmente la decisión de escoger la base de datos final.

*BD	Instancias	Atributos	Ventajas	Desventajas
FigShare[1]	124	33	Cuestionario BigFive agrupado correctamente	Idioma: Japonés
			Alta calidad de datos por ser atributos enteros	Falta de definición de los atributos
			Nº atributos suficientes	Nº instancias bajo
Harvard Dataverse[2]	386	82	Alta cantidad de atributos	Relacionado a un tema diferente del objetivo de estudio
			Alta calidad de datos	Falta de definición de atributos
			Origen fiable y reciente (2022)	Conflictos con el acceso a la BD
Selfies[3]	30,935	6	Alta cantidad de Instancias	Atributos sólo indican de manera binaria la aparición o no de las 5 dimensiones
			Descripción adecuada de la BD	Falta de información relacionada con el cuestionario
			Origen reciente (2022)	Conflictos con el acceso a la BD
IPIP-FFM[4]	1,015,341	110	Alta cantidad de datos	No todos son enteros
			Alta calidad de datos a priori	Formato .csv desorganizado
			Descripción detallada de atributos	4 años sin actualizaciones

Como podemos comprobar, no son muchas las bases de datos a las que se tiene acceso relacionadas con este ámbito y que recojan datos de calidad y fiables para el estudio que se pretende llevar a cabo. Finalmente, la base de datos escogida es *IPIP-FFM,* esto se debe a que es la que más datos recoge, así como la descripción detallada de los mismos y su relación directa con el cuestionario Big Five.

2.4 Análisis de técnicas de aprendizaje automático

Al ser un proyecto centrado en ciencia de datos e inteligencia artificial, el análisis de las diferentes técnicas de aprendizaje automático aplicables a la propuesta es fundamental para el éxito del proyecto. Una vez identificados los datos necesarios, su calidad y relevancia para el análisis, se debe explorar el espectro de algoritmos disponibles que se adapten a las características de los datos y los objetivos del estudio.

[1]https://figshare.com/articles/dataset/Raw_data_of_survey_on_willingness_to_study_and_big_five/2814766
[2] https://dataverse.harvard.edu/file.xhtml?fileId=5600028&version=1.0&toolType=PREVIEW
[3] https://ieee-dataport.org/open-access/data-prediction-apparent-personality-traits-selfies-using-five-factor-model
[4] https://openpsychometrics.org/_rawdata/IPIP-FFM-data-8Nov2018.zip

En este contexto, las técnicas de aprendizaje supervisado pueden ser especialmente útiles. Estas técnicas permiten predecir o clasificar variables objetivo a partir de variables de entrada mediante la construcción de un modelo basado en ejemplos etiquetados. Algoritmos como regresión logística, árboles de decisión, máquinas de vectores de soporte (SVM) y métodos de ensamble como *Random Forest* y *Gradient Boosting* pueden ser aplicados para predecir el rendimiento del equipo en función de las características individuales de sus miembros.

Por otro lado, el aprendizaje semi-supervisado podría ser beneficioso si se dispone de una cantidad limitada de datos etiquetados y una gran cantidad de datos no etiquetados. Estas técnicas combinan principios de aprendizaje supervisado y no supervisado para mejorar el rendimiento del modelo utilizando tanto datos etiquetados como no etiquetados.

Sin embargo, en este caso, el proyecto estará enfocado a utilizar el aprendizaje no supervisado. Esta técnica permite explorar patrones y estructuras ocultas en los datos sin la necesidad de etiquetas previas. Algoritmos como *clustering* (agrupamiento), análisis de componentes principales (PCA) y redes neuronales *autoencoder* pueden ser utilizados para identificar grupos de perfiles de personalidad similares dentro del equipo o para reducir la dimensionalidad de los datos y extraer características relevantes.

Dada la naturaleza de los datos con los que se va a trabajar y el objetivo de estudio, la técnica de aprendizaje automático elegida es el *clustering*. Esta técnica es una opción adecuada para el objetivo de agrupar candidatos para un equipo en perfiles específicos debido a su capacidad para identificar automáticamente estructuras latentes en los datos sin la necesidad de etiquetas previas. En nuestro caso, al utilizar los resultados del test de las cinco grandes dimensiones de la personalidad como características de entrada, el *clustering* nos permitirá segmentar automáticamente a los candidatos en grupos homogéneos basados en sus perfiles de personalidad. Esto facilitará la formación de equipos equilibrados y complementarios, en función de los requisitos solicitados por el personal de selección.

Además, el *clustering* nos brinda la flexibilidad de explorar diferentes enfoques de agrupamiento, como *k-means*, DBSCAN o *hierarchical clustering*, dependiendo de la estructura y la distribución de los datos. Esto nos permite adaptar el proceso de agrupamiento a las características específicas de nuestra muestra de datos y a los requisitos del estudio. Por ejemplo, podemos ajustar el número de clústeres según la diversidad deseada en los equipos o explorar la similitud entre perfiles de personalidad mediante técnicas de *clustering* jerárquico.

Otra ventaja del *clustering* es su capacidad para manejar grandes volúmenes de datos de manera eficiente, lo que es crucial en nuestro proyecto, donde podemos tener una gran cantidad de candidatos y características de personalidad para analizar. Los algoritmos de *clustering* están diseñados para escalar bien con

conjuntos de datos de gran tamaño, lo que nos permite aplicar esta técnica de manera efectiva incluso en entornos empresariales con una gran cantidad de candidatos potenciales.

Dado que este proyecto apunta a construir un modelo que permita la asignación de individuos a roles específicos, para así poder seleccionar los perfiles más convenientes para el equipo de trabajo al que se aplique, esta técnica ayudará en gran medida a la formación de equipos efectivos y cohesionados.

2.5 Estudios similares y metodología aplicada

El artículo expuesto por la UPC de Madrid [5] presenta un análisis empírico del efecto de diferentes rasgos de personalidad en el rendimiento académico en asignaturas cuantitativas de Administración y Dirección de Empresas (ADE). Utiliza el modelo *Big Five* para medir los rasgos de personalidad y emplea modelos de regresión múltiple para explicar el rendimiento académico en varias asignaturas. Los datos fueron obtenidos mediante la aplicación del cuestionario *Big Five* a estudiantes de ADE en la Universidad Pontificia Comillas de Madrid. Los resultados destacan que la dimensión de *Conciencia*, con sus subdimensiones de *Escrupulosidad* y *Perseverancia*, según el estudio es la única relevante para explicar el rendimiento académico, sugiriendo que un mayor esfuerzo y búsqueda de excelencia conducen a un mejor rendimiento académico.

Para la propuesta de optimización del rendimiento de equipos profesionales, este estudio proporciona una base sólida al demostrar la influencia significativa de ciertos rasgos de personalidad en el rendimiento académico. Específicamente, el énfasis en la dimensión de *Conciencia* como un predictor clave del éxito académico puede extrapolarse a la dinámica laboral, sugiriendo que individuos con predominancia en esta dimensión pueden ser más propensos a contribuir positivamente al rendimiento del equipo en entornos laborales. Este hallazgo respalda la importancia de comprender y aprovechar las diferencias individuales en personalidad para mejorar la eficacia y la colaboración dentro de los equipos profesionales.

Además, sabemos que, para alcanzar el éxito de un equipo, es crucial la selección cuidadosa de miembros del equipo que puedan trabajar de manera efectiva en conjunto. Este estudio [6] analiza la relación entre los cinco grandes factores de personalidad y el rendimiento objetivo del equipo en equipos de diseño de productos. Se descubrió que los equipos exitosos mostraban niveles más altos de capacidad cognitiva general, *Extraversión* y *Amabilidad*, y niveles más bajos de *Neuroticismo*. Además, se observó que la heterogeneidad de la conciencia estaba relacionada negativamente con el rendimiento del producto en los equipos exitosos. Estos hallazgos tienen importantes implicaciones para la selección de equipos de diseño de productos y señalan áreas clave para futuras investigaciones.

3 Desarrollo

3.1 Entorno de desarrollo

3.1.1 Lenguaje de Programación

Python se ha consolidado como uno de los lenguajes de programación más utilizados en el campo de la inteligencia artificial y el aprendizaje automático debido a su versatilidad, eficiencia y amplio ecosistema de bibliotecas especializadas. Su sintaxis clara y legible facilita la implementación de algoritmos complejos, permitiendo así un enfoque directo en la lógica del modelo en lugar de detalles técnicos. Además, *Python* es un lenguaje de código abierto con una comunidad activa y colaborativa, lo que significa que constantemente se están desarrollando nuevas herramientas y mejoras que benefician a los proyectos de aprendizaje automático. Este lenguaje cuenta con bibliotecas como *TensorFlow*, *Keras*, *Scikit-learn* y *PyTorch*, que ofrecen herramientas poderosas y optimizadas para la construcción y entrenamiento de modelos de aprendizaje automático. Esta riqueza de recursos disponibles hace que sea la opción escogida para este proyecto, ofreciendo un entorno de desarrollo robusto y eficiente.

3.1.2 Google Colaboratory

Colab, como servicio alojado de *notebooks* de *Jupyter*, ha sido seleccionado para este proyecto debido a su accesibilidad y su conjunto de características especializadas. Al no requerir instalación, *Colab* ofrece una plataforma conveniente que elimina barreras de entrada y permite comenzar a trabajar de inmediato. Su acceso gratuito a recursos informáticos, como *GPUs* y *TPUs*, proporciona una infraestructura poderosa para ejecutar modelos de aprendizaje automático a gran escala y realizar análisis de datos complejos. También proporciona una interfaz muy visual por estar implementado con los *notebooks* mencionados. Esto hace que se pueda subdividir el trabajo en celdas y permite una mayor comprensión del trabajo realizado por la posibilidad de visualizar paso a paso el proceso de desarrollo de un modelo.

Esta combinación de facilidad de uso y potencia computacional hace que *Colab* sea una opción ideal para proyectos de ciencia de datos y aprendizaje automático, permitiendo a los investigadores centrarse en el desarrollo de modelos sin preocuparse por la infraestructura subyacente. En un estudio llevado a cabo por la Universidad Nacional de Chimborazo [7] se realizó un análisis del rendimiento de *Google Colaboratory* en el entrenamiento de una red neuronal convolucional para la clasificación de imágenes. Las pruebas se realizaron con cuatro *datasets*, se aplicó un enfoque cuantitativo, una investigación experimental y descriptiva. Se demostró que no existe una

diferencia significativa en el tiempo de entrenamiento de la red neuronal en *Google Colaboratory* y un computador personal, sin embargo, existe una menor pérdida y mayor precisión del modelo en la clasificación de imágenes.

3.2 Metadatos

Las cinco grandes dimensiones no están asociadas a ningún test en particular, sino que se han desarrollado diversas maneras de medirlas. El test que da lugar a la base de datos con la que vamos a trabajar utiliza los marcadores de los cinco grandes factores del *International Personality Item Pool*, desarrollado por Goldberg (1992) [8]. Estos datos se recogieron entre 2016 y 2018 a través de un test de personalidad interactivo en línea [9].

Los datos recogen las respuestas de los participantes según las afirmaciones que se presentaron en el test, cada una se valoró en una escala de cinco puntos utilizando botones de opción. La escala se etiquetó como *1=En desacuerdo, 3=Neutral, 5=De acuerdo*. El orden en la página era *EXT1*, *AGR1*, *CSN1*, *EST1*, *OPN1*, *EXT2*, etc. Estas son las afirmaciones que los participantes visualizaron al seleccionar una etiqueta:

EXT1	Soy el alma de la fiesta.	**AGR6**	Tengo un corazón blando.
EXT2	No hablo mucho.	**AGR7**	No me interesan los demás.
EXT3	Me siento cómodo con la gente.	**AGR8**	Dedico tiempo a los demás.
EXT4	Me mantengo en un segundo plano.	**AGR9**	Siento las emociones de los demás.
EXT5	Inicio conversaciones.	**AGR10**	Hago que la gente se sienta a gusto.
EXT6	Tengo poco que decir.	**CSN1**	Siempre estoy preparado.
EXT7	Hablo con mucha gente en las fiestas.	**CSN2**	Dejo mis pertenencias por ahí.
EXT8	No me gusta llamar la atención.	**CSN3**	Presto atención a los detalles.
EXT9	No me importa ser el centro de atención.	**CSN4**	Desordeno las cosas.
EXT10	Soy callado con los desconocidos.	**CSN5**	Hago las tareas enseguida.
EST1	Me estreso con facilidad.	**CSN6**	A menudo me olvido de poner las cosas en su sitio.
EST2	Estoy relajado/a la mayor parte del tiempo.	**CSN7**	Me gusta el orden.
EST3	Me preocupo por las cosas.	**CSN8**	Descuido mis obligaciones.
EST4	Rara vez me siento triste.	**CSN9**	Sigo un horario.
EST5	Me altero con facilidad.	**CSN10**	Soy exigente en mi trabajo.
EST6	Me altero con facilidad.	**OPN1**	Tengo un vocabulario rico.
EST7	Cambio mucho de humor.	**OPN2**	Tengo dificultades para comprender ideas abstractas.
EST8	Tengo frecuentes cambios de humor.	**OPN3**	Tengo una imaginación muy viva.
EST9	Me irrito con facilidad.	**OPN4**	No me interesan las ideas abstractas.
EST10	A menudo me siento triste.	**OPN5**	Tengo ideas excelentes.
AGR1	Siento poca preocupación por los demás.	**OPN6**	No tengo buena imaginación.
AGR2	Me intereso por la gente.	**OPN7**	Comprendo las cosas con rapidez.
AGR3	Insulto a la gente.	**OPN8**	Utilizo palabras difíciles.
AGR4	Simpatizo con los sentimientos de los demás.	**OPN9**	Dedico tiempo a reflexionar sobre las cosas.
AGR5	No me interesan los problemas de los demás.	**OPN10**	Tengo muchas ideas.

Las cinco dimensiones son etiquetadas de la siguiente manera:

- *EXT* - Afirmaciones para medir la *Extraversión*.
- *EST* - Afirmaciones para medir el *Neuroticismo*.
- *AGR* - Afirmaciones para medir la *Amabilidad*.
- *CSN* - Afirmaciones para medir la *Conciencia*.
- *OPN* - Afirmaciones para medir la *Apertura a la Experiencia*.

Los participantes fueron informados de que sus respuestas serían grabadas y utilizadas para la investigación al comienzo de la prueba, y se les pidió que confirmaran su consentimiento al final de la misma. Además, el tiempo empleado en cada pregunta se registró en milisegundos. Esta información se recoge en las variables terminadas en "*_E*". Se calculó tomando el tiempo en que se pulsó el botón de la pregunta menos el tiempo de la pulsación más reciente de otro botón.

Por último, tenemos 10 variables más recogidas en la base de datos, estas son:

dateload	Fecha y hora de inicio de la encuesta.
screenw	La anchura de la pantalla del usuario en píxeles.
screenh	La altura de la pantalla del usuario en píxeles.
introelapse	Tiempo en segundos transcurrido en la página de inicio.
testelapse	Tiempo en segundos empleado en la página con las preguntas de la encuesta.
endelpase *	Tiempo en segundos empleado en la página de finalización (en la que se pedía al usuario que indicara si había respondido correctamente y si sus respuestas podían almacenarse y utilizarse para la investigación).
IPC	Número de registros de la dirección IP del usuario en el conjunto de datos.
country	El país, determinado por la información técnica.
lat_appx_lots_of_err	Latitud aproximada del usuario determinada por información técnica.
long_appx_lots_of_err	Longitud aproximada del usuario determinada por información técnica.

** De nuevo: este conjunto de datos sólo incluye a los usuarios que respondieron "Sí" a esta pregunta; los usuarios eran libres de responder "No" y podían seguir viendo sus resultados de cualquier forma.*

Así pues, esta base de datos recoge, en 110 variables, información sobre las respuestas del test de personalidad de 1,015,341 participantes.

3.3 Preprocesado

El preprocesado de datos juega un papel fundamental en cualquier proyecto de ciencia de datos y aprendizaje automático. Este proceso comprende una serie de pasos diseñados para limpiar, transformar y preparar los datos antes de aplicar algoritmos de *machine learning*. Estas tareas nos permiten garantizar la calidad de los datos y maximizar la eficiencia del algoritmo. Su importancia radica en que los datos crudos suelen presentar irregularidades, ruido y falta de coherencia que pueden afectar negativamente el rendimiento de los modelos. Por lo tanto, el preprocesado se encarga de abordar estos desafíos con el fin de garantizar que los datos sean adecuados y representativos para el análisis posterior. En el contexto del *clustering*, un preprocesado eficaz es crucial para identificar patrones significativos en los datos y generar agrupamientos precisos y útiles para la toma de decisiones.

3.3.1 EDA: Análisis Exploratorio de Datos

Como contacto inicial con los datos, se realiza un análisis exploratorio de datos (EDA). Esta etapa inicial del proceso analítico tiene como propósito fundamental entender la estructura, distribución y características esenciales de los datos antes de la implementación del algoritmo de *machine learning*. Mediante técnicas de visualización, estadística descriptiva y detección de patrones preliminares, el EDA ofrece una visión comprensiva del conjunto de datos, facilitando la formulación de hipótesis, la orientación del preprocesado de datos y la selección de enfoques adecuados para el modelado subsiguiente. En nuestro caso, un análisis exploratorio exhaustivo resulta crítico para identificar interrelaciones entre variables, evaluar la pertinencia de los datos para el *clustering* y definir estrategias apropiadas para la segmentación de datos en grupos coherentes.

Tras realizar adecuadamente la importación de los datos y la creación del entorno de trabajo, comenzamos a explorar a fondo los datos. Como sabemos tenemos datos estructurados, esto facilita significativamente el preprocesado. Tenemos 1,015,341 participantes y 110 características. Vemos que no todas las columnas son del mismo tipo. Por supuesto, las 100 columnas correspondientes a las preguntas del test son de tipo `float64`. Sin embargo, en las otras 10, tenemos 4 de tipo `float64`, 2 de tipo `int64` y 4 de tipo `object` (podrían contener cualquier tipo de objeto de Python).

Comprobamos que las variables de tipo `object` son *'dateload', 'country', 'lat_appx_lots_of_err'* y *'long_appx_lots_of_err'* (esta información fue tomada técnicamente por el sistema). Sabemos que *'dateload'* guarda información sobre el momento en el que los participantes comenzaron el test y *'lat_appx_lots_of_err'* y *'long_appx_lots_of_err'* se refieren a la latitud expresadas

en grados y décimas de grado. Dado que sabemos que '`country`' es la columna que almacena información sobre el país, y observamos participantes de 223 países diferentes, vamos a comprobar el lugar desde el que se realizaron los cuestionarios.

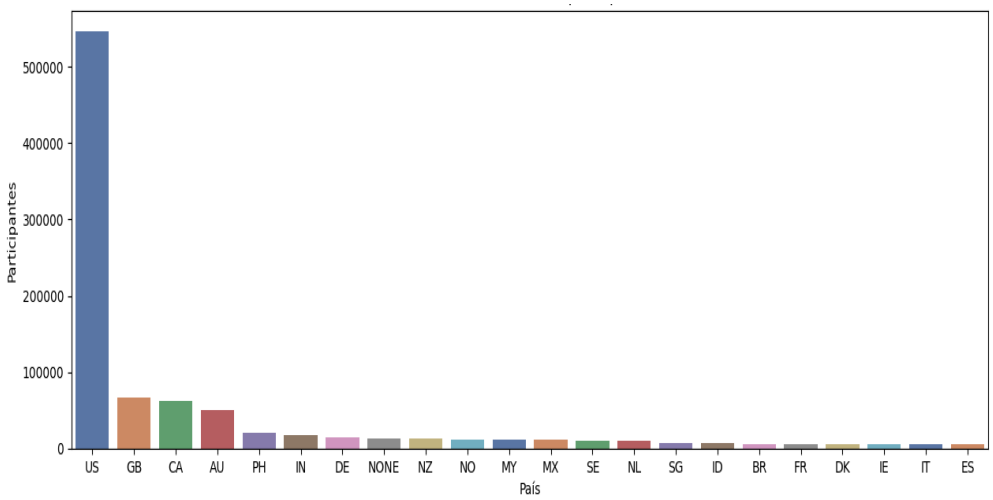

Figura 3: Países con más de 5000 participantes

Pese a que no tenemos información sobre la correspondencia exacta de las siglas y el país al que se refieren, podemos intuirlo. Sin embargo, lo que está claro es que el octavo lugar de la gráfica (obsérvese Fig.3) es "*NONE*", eso quiere decir que para una gran parte de participantes no se sabe el país. En concreto, hay 13,728 con información inexacta del país. Esto nos lleva a comprender que no todos los datos que tenemos son correctos o completos.

A continuación, tras explorar los valores nulos de nuestra base de datos, vemos que tenemos 186,358 en total. Además, todas las columnas asociadas a las respuestas tienen el mismo número de nulos, incluso las 50 que miden el tiempo de respuesta. Esto nos hace entender, a priori, que hay el mismo número de participantes con respuestas nulas y no algunos valores nulos en distintos participantes.

Como siguiente paso del análisis exploratorio, vamos a estudiar si todas las respuestas están en el rango esperado sabiendo que hay nulos, es decir, si los valores que tenemos en las primeras 50 columnas son un entero del 1 al 5 o un nulo. Comprobamos que esta hipótesis no se cumple, a diferencia de lo que esperábamos por la información expuesta en el codebook de la base de datos, hay respuestas evaluadas con la etiqueta 0. Sin embargo, sabemos que al responder en el test sólo hay 5 botones. Trataremos este asunto en el apartado correspondiente a la limpieza de datos.

Ahora vamos a explorar la variable '`IPC`', que se refiere al número de registros de la dirección IP del usuario en el conjunto de datos. De esta manera, los valores altos pueden deberse a redes compartidas (como universidades o compañías) o a envíos múltiples.

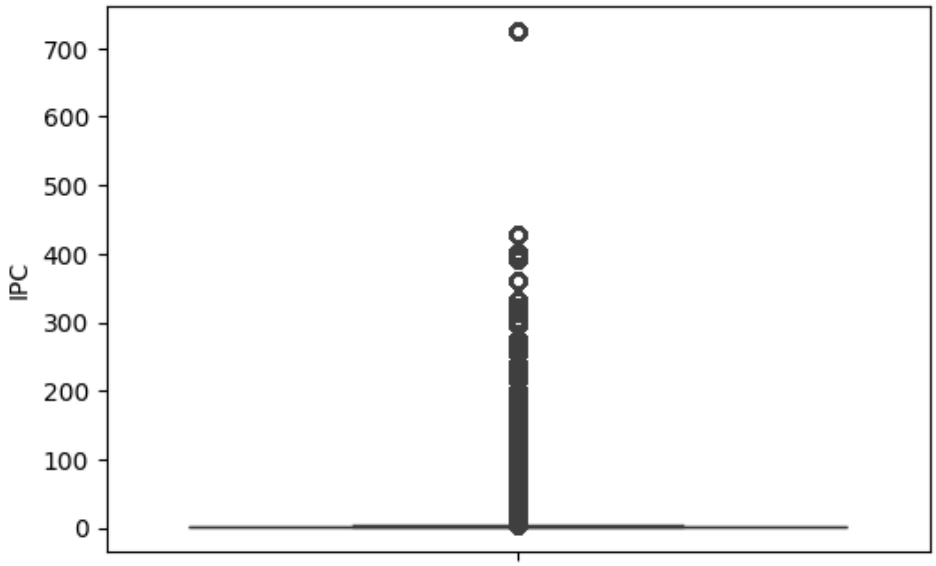

Figura 4: Boxplot para IPC

Vamos a realizar un análisis de la información que obtenemos de las estadísticas descriptivas de esta variable, pues proporcionan información sobre la distribución de los valores de la columna. La presencia de un gran número de usuarios con un mismo valor de 1 ha generado una distribución que se traduce en una gráfica de *boxplot* sin caja visible y todos los valores concentrados en ese mismo punto (obsérvese Fig.4). Esto se debe a que la mayoría de los datos están concentrados en este valor mínimo de 1, como se indica por el primer cuartil (25%) y la mediana (50%). Además, la presencia de valores extremadamente altos, como el máximo de 725, hace que el rango intercuartílico sea muy pequeño en comparación con el valor máximo, lo que resulta en la ausencia de una caja visible en el *boxplot*. Esta gráfica de *boxplot* con algunos valores extremadamente altos, que se representan como *outliers*, refleja la distribución sesgada de los datos, donde la gran mayoría de los valores están agrupados en el extremo inferior, siendo usuarios con un único registro de IP en la base de datos.

Finalmente, para terminar con el proceso de EDA, comprobamos que no hay participantes duplicados en nuestra base de datos. El hecho de contar con la variable '`dateload`' facilita mucho este proceso.

3.3.2 Limpieza de datos

3.3.2.1 Procesado de '*IPC*'

Ahora que ya hemos realizado un análisis de la situación en la que están los datos, comenzamos el proceso de limpieza para así dejarlos listos para el modelo.

Como primer paso, vamos a limpiar la base de datos atendiendo a la columna *'IPC'*. Hemos observado durante la exploración que la gran mayoría de datos corresponden con usuarios de 1 sólo registro. Aun así, hay muchos participantes que tienen más de uno, llegando incluso a 725. En este estudio nos queremos asegurar de tratar con individuos y no con organizaciones para un mejor resultado. Así pues, vamos a quedarnos sólo con aquellos participantes que tengan un solo registro en esta variable, evitando además el sesgo de usuarios con varios intentos del cuestionario.

3.3.2.2 Selección de características

Tras procesar la variable IPC hemos eliminado 318,496 participantes, quedando ahora 696,845 con registros únicos en la base de datos.

A continuación, vamos a eliminar aquellas características innecesarias:

- *'**dateload**'*: Esta información nos ha sido útil para estudiar los duplicados en la base de datos, es importante comprender que el estudio de duplicados debe realizarse como paso previo a la eliminación de columnas. Hacerlo de manera opuesta podría derivar en errores de consideración de duplicados falsos, en este caso se ve muy clara esta situación al usar como pivote la variable *'dateload'*. Sin embargo, no es información en absoluto relevante para el estudio. Es más, podría influenciar de manera errónea el rendimiento del modelo, pues sólo queremos estudiar la personalidad y eso no puede relacionarse con el momento en el que se realizó el test. Así pues, esta columna será eliminada de la base de datos.
- ***'screenw'*** y ***'screenh'***: Por el mismo motivo que la columna anterior, es decir, que son columnas irrelevantes y que podrían distorsionar los resultados, estas columna serán eliminadas.
- ***'IPC'***: Tras realizar el preprocesado correspondiente de esta columna, será eliminada por no aportar ninguna información adicional, pues ahora todos los valores son 1.
- *'**country**'*: Como pudimos comprobar durante la exploración de la base de datos, es un número muy relevante el de países con información inexacta. Por esta razón y por su ausencia de relación con el estudio que se va a realizar se eliminará esta columna de la base de datos.
  *'**lat_appx_lots_of_err**'* y *'**long_appx_lots_of_err**'*: La decisión de la eliminación de estas dos columnas se fundamenta en los conflictos que este tipo de informaciones han mostrado en múltiples ocasiones. Un caso interesante que muestra lo que puede suponer confiar en este tipo de informaciones es el que se expone en el artículo de Kashmir Hill [10]:

En el tranquilo y remoto pueblo de Potwin, Kansas, una parcela de 360 acres ha sido testigo de una década de caos tecnológico para sus habitantes. La granja, propiedad de la familia Vogelman por generaciones y ahora alquilada por Joyce Taylor, de 82 años, se ha convertido en el epicentro de un tormentoso relato digital. Desde acusaciones de delitos cibernéticos hasta visitas de agencias federales, ambulancias y vigilantes en línea, los residentes han sido acosados y tratados como delincuentes sin entender la razón de tal hostigamiento. Este calvario tiene su origen en un error de cartografía digital que sitúa incorrectamente la ubicación de las direcciones IP en la propiedad de Taylor, generando una confusión que ha transformado un apacible rincón rural en una pesadilla tecnológica.

El incidente revela las complejidades y consecuencias de la cartografía digital en la era moderna, destacando cómo una simple equivocación puede desencadenar un caos inimaginable en la vida de las personas. La empresa MaxMind, proveedora de servicios de geolocalización basados en direcciones IP, inadvertidamente seleccionó la granja de Taylor como el punto de referencia predeterminado para miles de direcciones IP no identificadas en los Estados Unidos. Este error ha expuesto las limitaciones y la falta de regulación en la infraestructura de Internet, donde empresas privadas como MaxMind asumen roles cruciales sin un control oficial. A medida que se revela el impacto devastador de esta falla en la vida cotidiana de los afectados, surge la urgencia de abordar las deficiencias en la cartografía digital y garantizar su precisión para evitar que más personas sean arrastradas a un torbellino de problemas injustificados.

Por la inexactitud general de este tipo de informaciones y, de nuevo, su irrelevancia en el estudio, se eliminarán estas dos columnas de la base de datos.

Tras deshacernos de estas características, tenemos 103 columnas. Como sabemos las 50 primeras son las etiquetas para las afirmaciones del cuestionario, las 50 siguientes el tiempo en milisegundos que se tardó en responder y las últimas 3 información técnica de tiempos en la pantalla de inicio (`'introelapse'`), del test (`'testelapse'`) y de finalización (`'endelapse'`).

Queremos estudiar la relevancia de estas últimas 3 para el estudio que se va a realizar. Para ello vamos a trabajar con un nuevo conjunto de datos que combine las variables basándonos en descripciones similares. La lista de variables relacionadas con las afirmaciones en este nuevo conjunto será *'Social', 'Not_Social', 'Optimal_mood', 'Disturbed_mood', 'Positive_social_interactions', 'Negative_social_interactions', 'Organised', 'Unorganised', 'Thinker', 'Not_thinker'*, así como el tiempo dedicado a cada uno de estos tipos de preguntas para todos los conjuntos de preguntas. Ahora estudiamos la correlación entre las respuestas del test y los tiempos respectivos con las últimas 3 características del tiempo en las distintas pantallas:

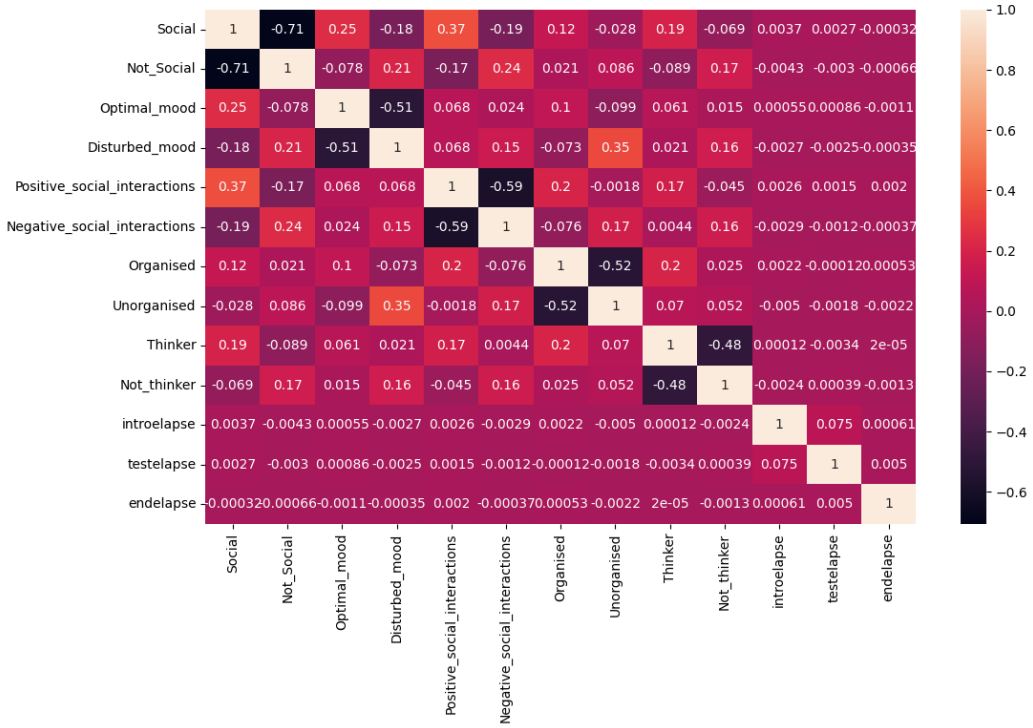

Figura 5: Matriz de correlación de respuestas agrupadas y variables independientes

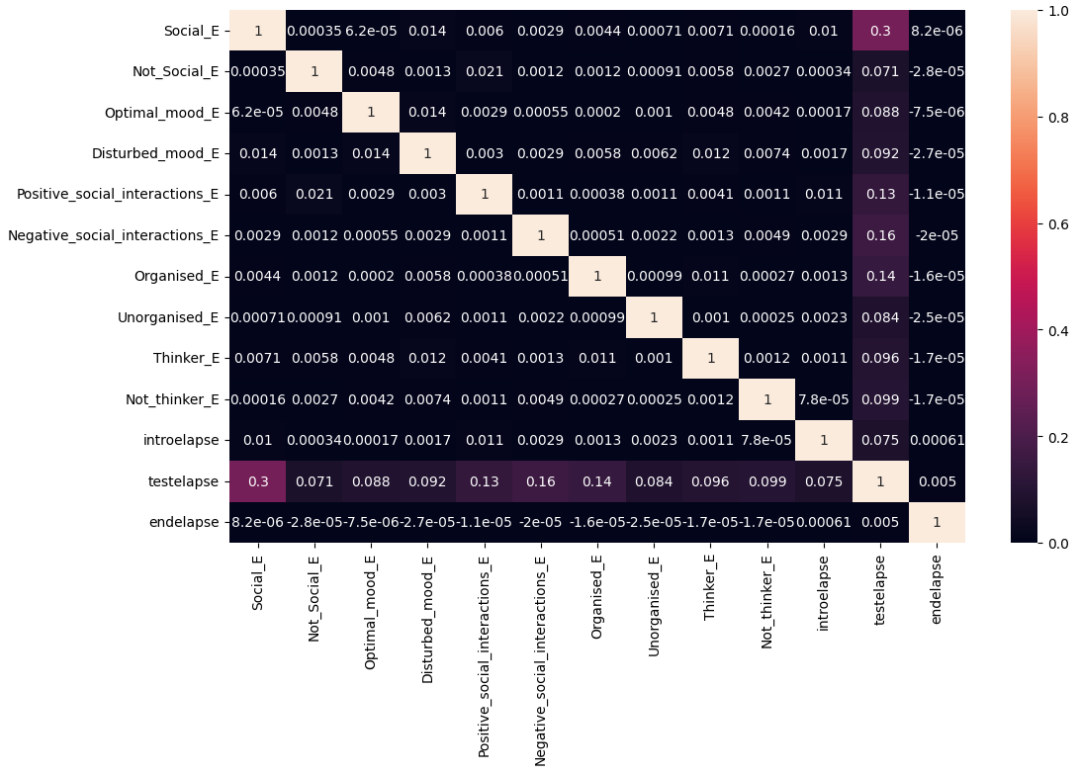

Figura 6: Matriz de correlación de tiempos de respuesta agrupados y variables independientes

Observamos que estas últimas 3 variables tienen una correlación casi nula para las respuestas del cuestionario (Fig.5). Obviamente se observa correlación de '`testelapse`' con todo el resto de variables de tiempo (Fig.6), pues es la suma total de todas ellas. Dada la ausencia de información que pueda aportarnos estas 3 variables y su mayor número de nulos que del resto de preguntas, vamos a eliminarlas, quedándonos sólo con las 100 variables correspondientes al cuestionario en sí.

Ahora todas las columnas son de tipo `float64`. Sabemos que algunos algoritmos de *clustering*, como *k-means*, trabajan mejor con variables continuas y requieren que todas las variables tengan la misma escala. En este caso, es apropiado tener todas las variables como números de punto flotante.

También se pretende verificar que realmente el tiempo de respuesta para las 50 afirmaciones es relevante en este estudio, así pues. Vamos a analizar la correlación de los tiempos con las respuestas, estudiando la matriz de correlación para el conjunto con respuestas agrupadas y para el conjunto original:

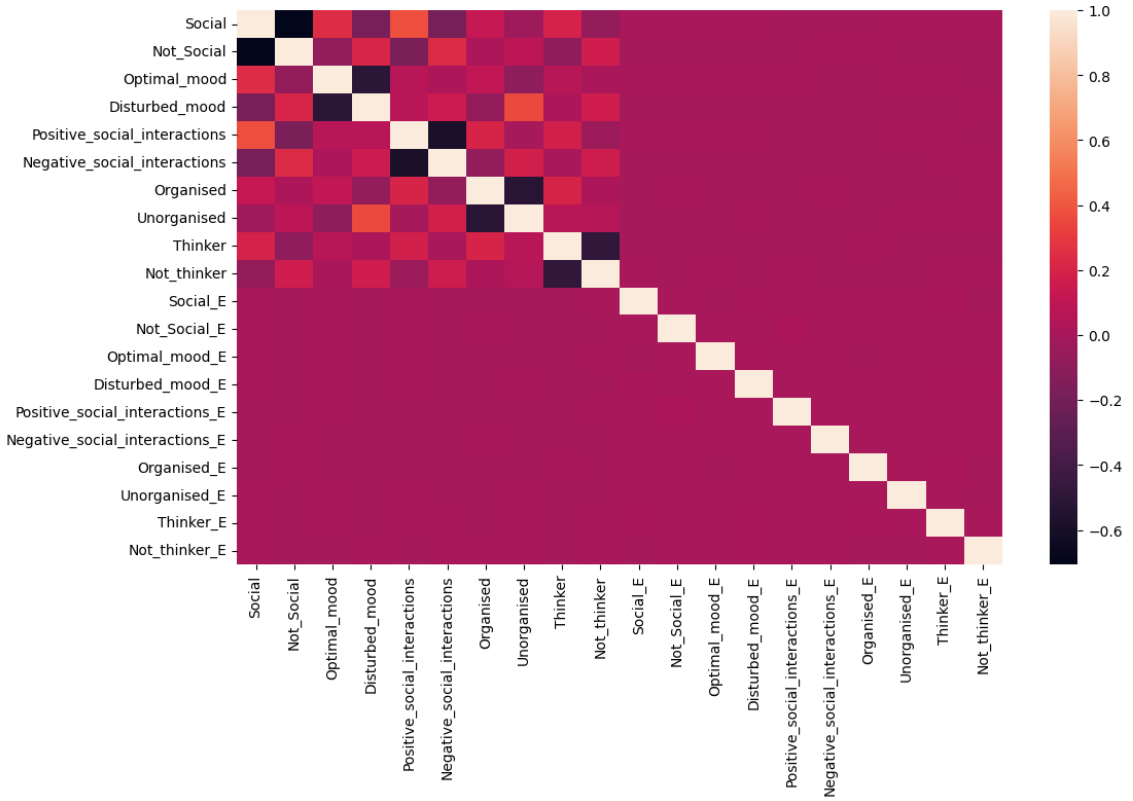

Figura 7: Matriz de correlación de respuestas agrupadas y sus tiempos asociados

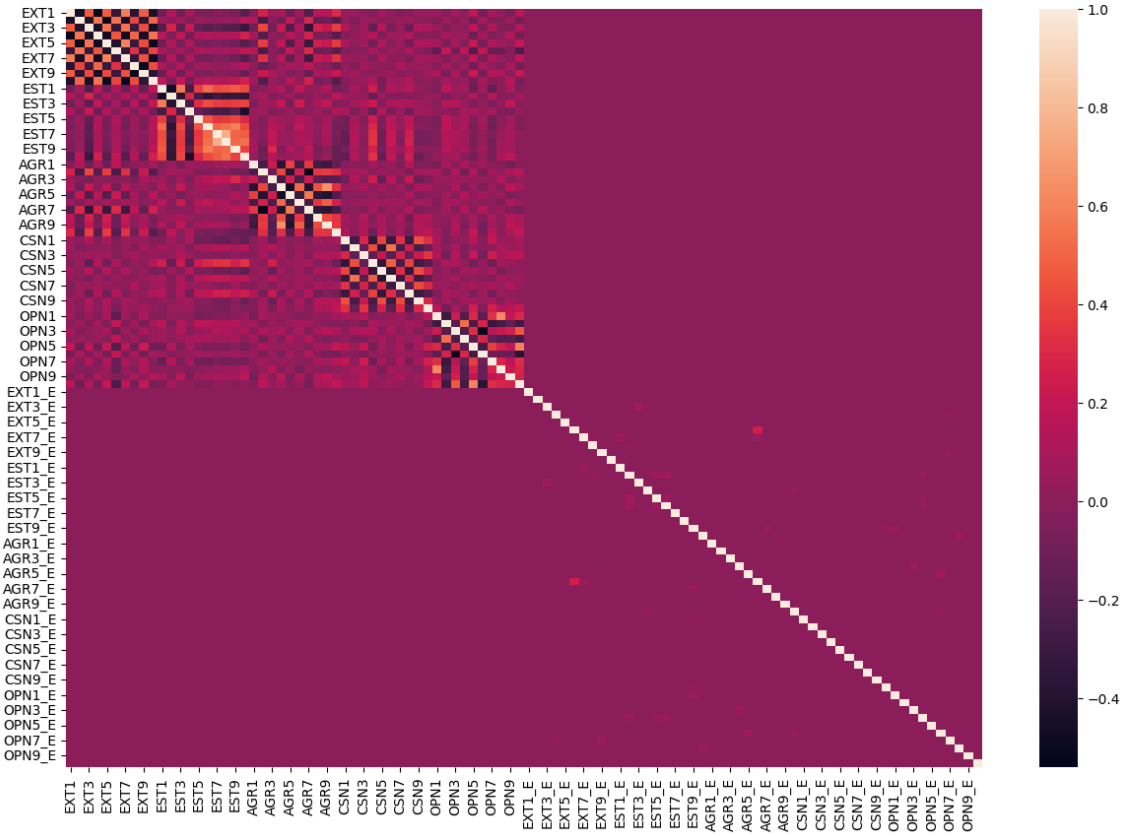

Figura 8: Matriz de correlación de respuestas y sus tiempos asociados

Se puede claramente comprobar que los tiempos de respuesta no están en absoluto correlacionados con las variables que nos interesan, ni si quieran lo están entre ellos (Fig.7 y Fig.8). Estas variables, por ende, no son relevantes para el modelo de *clustering* que busca agrupar las respuestas en función de sus similitudes o patrones en las variables relacionadas con las respuestas. Vamos a eliminar estas columnas. Finalmente, tras la selección de características, tenemos una base de datos de 696,845 participantes con 50 características, que en concreto son las respuestas de las 50 afirmaciones del cuestionario Big Five.

3.3.2.3 Tratamiento de valores nulos

A continuación, nos centramos en el tratado de nulos de la base de datos. Durante la exploración se comprobó que las columnas relacionadas con las respuestas del test tenían el mismo número de nulos, en concreto había 1,783 participantes nulos. Tras el procesado de '`IPC`', se han eliminado 642 participantes nulos. Dado a que el número de participantes nulos es insignificante (0.16% respecto a los datos totales), vamos a eliminar directamente los participantes con valores nulos.

Además, como vimos en la exploración, no se trata de valores nulos repartidos por la base de datos, si no de participantes completamente nulos. Queremos realizar un análisis de casos completos y no tendría mucho sentido imputar estos participantes con otros valores para las 50 variables.

3.3.2.4 Tratamiento de valores fuera de rango

Como última parte de la limpieza, se procesan los valores fuera de rango que observamos en la exploración. Se vio que había bastantes participantes con algún valor 0, concretamente en la base de datos en el momento de tratar este asunto hay 92,382 valores iguales a 0. Es un conjunto relevante, así que se descarta la opción de eliminar estos participantes de la base de datos. Al contrario que en el caso de los nulos, estaríamos perdiendo mucha información sin estos participantes. Volvemos a visualizar la correlación de nuestras variables para tomar una decisión apropiada y coherente con la situación:

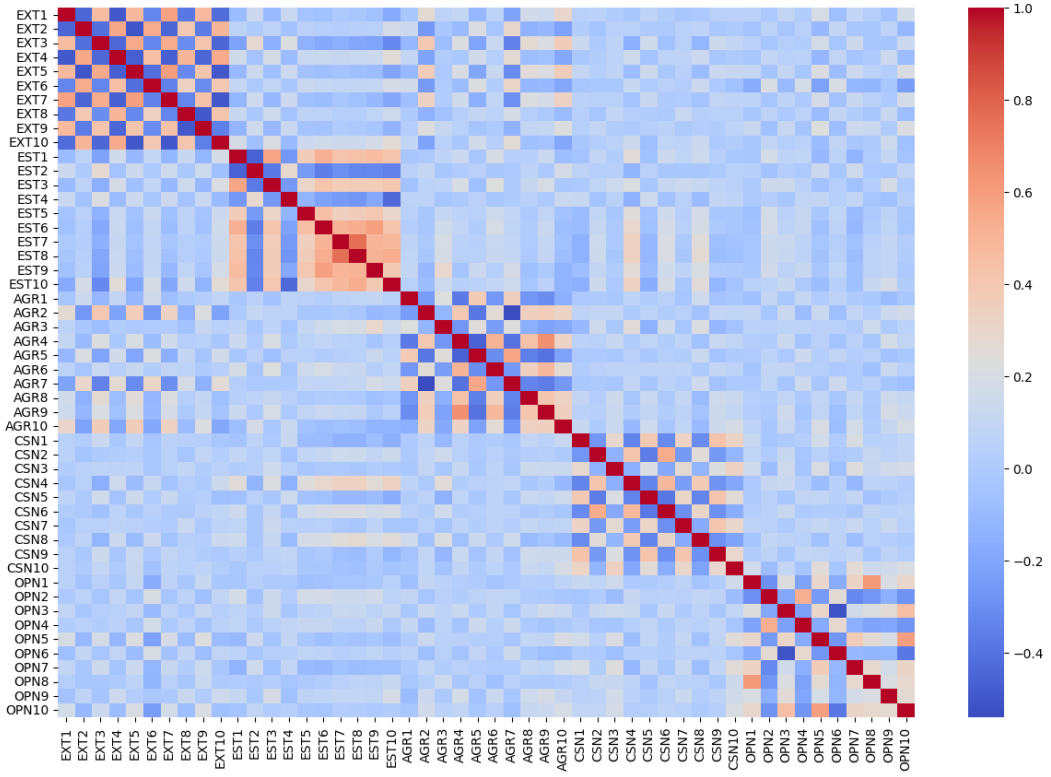

Figura 9: Matriz de correlación de respuestas

Efectivamente, vemos que nuestras variables están correlacionadas entre sí (Fig.9), generando 5 grupos de 10 características, lógicamente, atendiendo a las dimensiones de personalidad a las que hacen referencia.

Para tratar con el valor 0 de las respuestas, la librería *scikit-learn* proporciona opciones para imputar con valores faltantes o nulos, puede sernos útil en este caso. Un tipo de algoritmo de imputación es el método univariante, que imputa valores en la i-ésima dimensión de una característica utilizando sólo valores no faltantes en esa característica (p. ej. `SimpleImputer`). En cambio, los algoritmos de imputación multivariante utilizan todo el conjunto de dimensiones de características disponibles para estimar los valores que faltan (p. ej `IterativeImputer`).

Usar `IterativeImputer` puede ser una opción interesante ya que hay valores erróneos (fuera de rango) en varias columnas y estas columnas están correlacionadas entre sí. `IterativeImputer` es un enfoque más sofisticado que `SimpleImputer`, ya que puede estimar los valores que están fuera de rango basándose en el resto de las columnas del conjunto de datos. Lo hace de forma iterativa por turnos: en cada paso, una columna se designa como salida ***yy*** las otras columnas se tratan como entradas ***X***. Se ajusta un regresor en ***(X, y)*** para ***y*** conocido. A continuación, el regresor se utiliza para predecir los valores que faltan de ***y***. Esto se hace para cada característica de forma iterativa, y luego se repite para ***max_iter*** rondas de imputación. Se devuelven los resultados de la última ronda de imputación [11].

Sin embargo, este estimador sigue siendo experimental por ahora: los parámetros por defecto o los detalles de comportamiento podrían cambiar sin ningún ciclo de depreciación. Además, se han detectado distintos conflictos con este estimador en repetidas ocasiones para la versión actual, como se puede comprobar por ejemplo en el caso de la página [12]. Así pues, por el momento, pese que teóricamente parece ser una muy buena opción para este caso, no la vamos a utilizar esta vez.

Usaremos la clase `SimpleImputer`, que proporciona estrategias básicas para imputar valores perdidos o nulos. Estos valores se pueden imputar con un valor constante proporcionado, o utilizando los estadísticos (media, mediana o más frecuente) de cada columna en la que se encuentran los valores nulos. Esta clase también permite diferentes codificaciones de valores nulos. Ya sabemos que, en nuestro caso, estos valores nulos hacen referencia a las respuestas fuera de rango, es decir, los valores 0.

Concretamente, vamos a imputar los valores 0 por la moda de la característica (afirmación del cuestionario) en la que se encuentren. Esta decisión se debe a que necesitamos números enteros para crear un modelo adecuado y seguir un mismo formato para todas las respuestas, así pues, la media no es buena opción. Además, utilizando la moda influenciaremos en menor medida el sesgo que pueda estar generándose al imputar con un valor irreal para la respuesta del participante.

3.3.2.5 Normalización y escalado de características

La normalización y el escalado son técnicas comunes en el preprocesamiento de datos, utilizadas para ajustar las características de un conjunto de datos a una escala específica o distribución. Mientras que la normalización se refiere a la transformación de los datos para que estén en un rango específico, como 0 a 1, el escalado se refiere a la transformación de los datos para que tengan una media de 0 y una desviación estándar de 1, o alguna otra escala deseada.

En este caso particular, donde los datos ya están limitados al rango discreto de valores del 1 al 5, la normalización o el escalado no serían apropiados. Esto se debe a que estas técnicas generalmente se aplican cuando las características tienen diferentes escalas, lo que puede afectar negativamente el rendimiento de ciertos algoritmos de aprendizaje automático. Algunos algoritmos, como *K-means,* son sensibles a la escala y pueden producir resultados sesgados si las características no están en la misma escala.

Finalmente, sabiendo que la normalización y el escalado son técnicas valiosas en muchas situaciones de análisis de datos, en este caso particular no serían beneficiosos y podrían incluso distorsionar la información contenida en los datos. Es importante comprender cuándo aplicar estas técnicas y cuándo no, para evitar introducir sesgos innecesarios en el análisis de datos.

3.4 Visualización de datos

Ahora ya tenemos nuestra base de datos limpia y preparada para comenzar a trabajar. Vamos a visualizar la situación actual de los datos con diagramas de violín antes de entrar en el modelo de *clustering*.

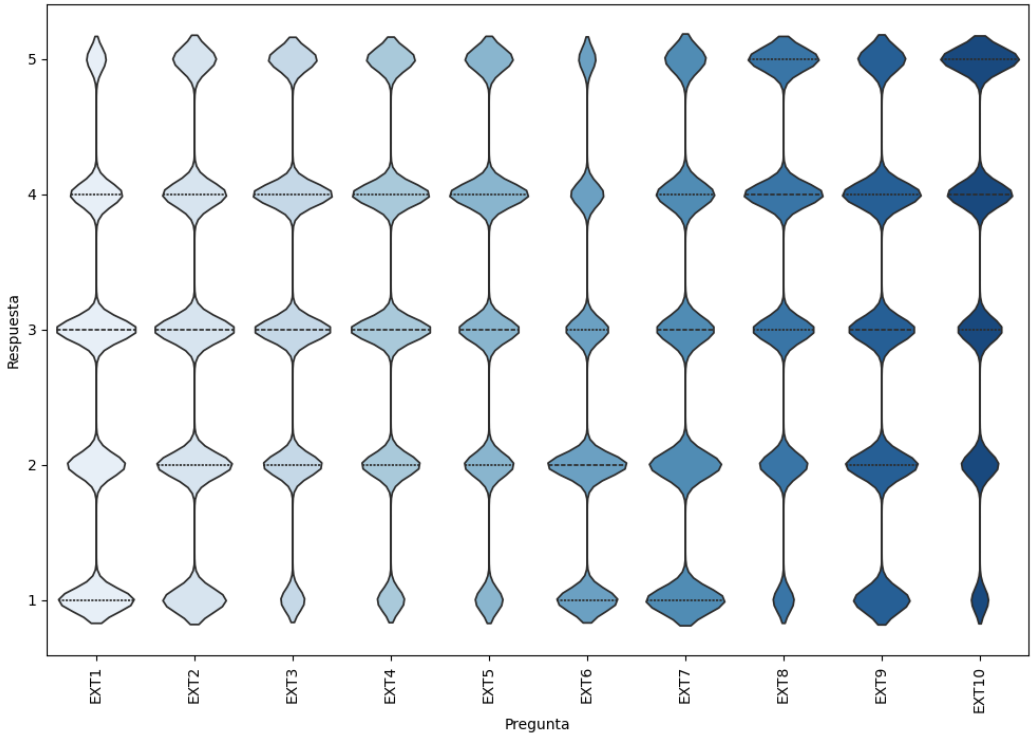

Figura 10: Distribución de respuestas para cada pregunta de Extraversión

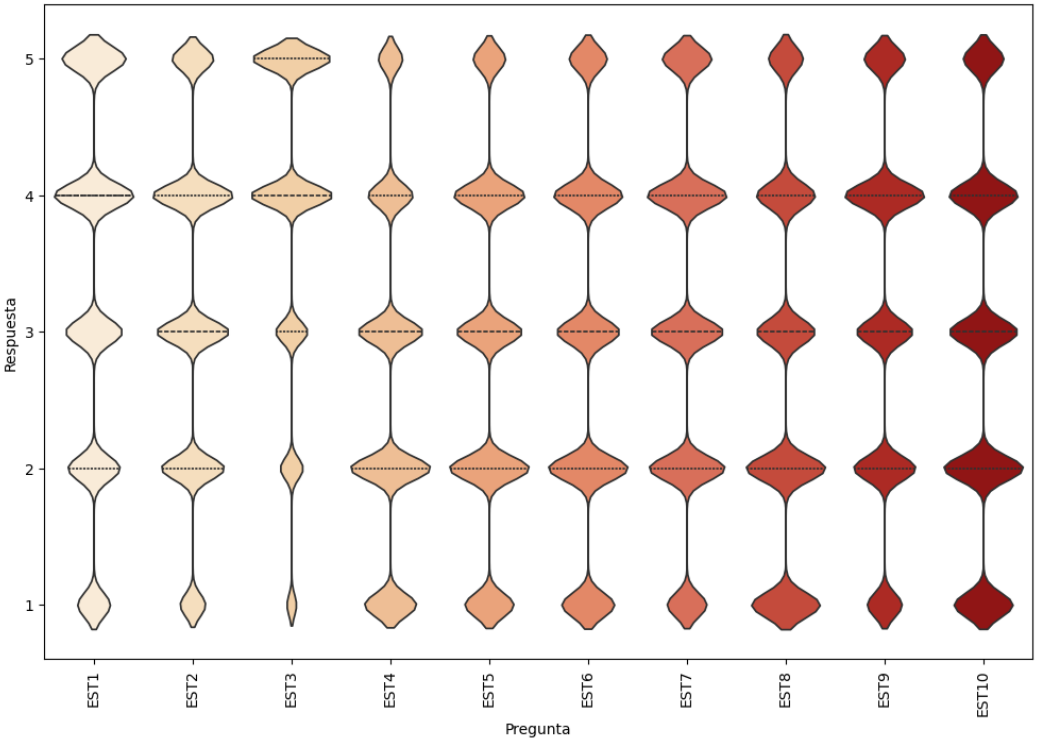

Figura 11: Distribución de respuestas para cada pregunta de Neuroticismo

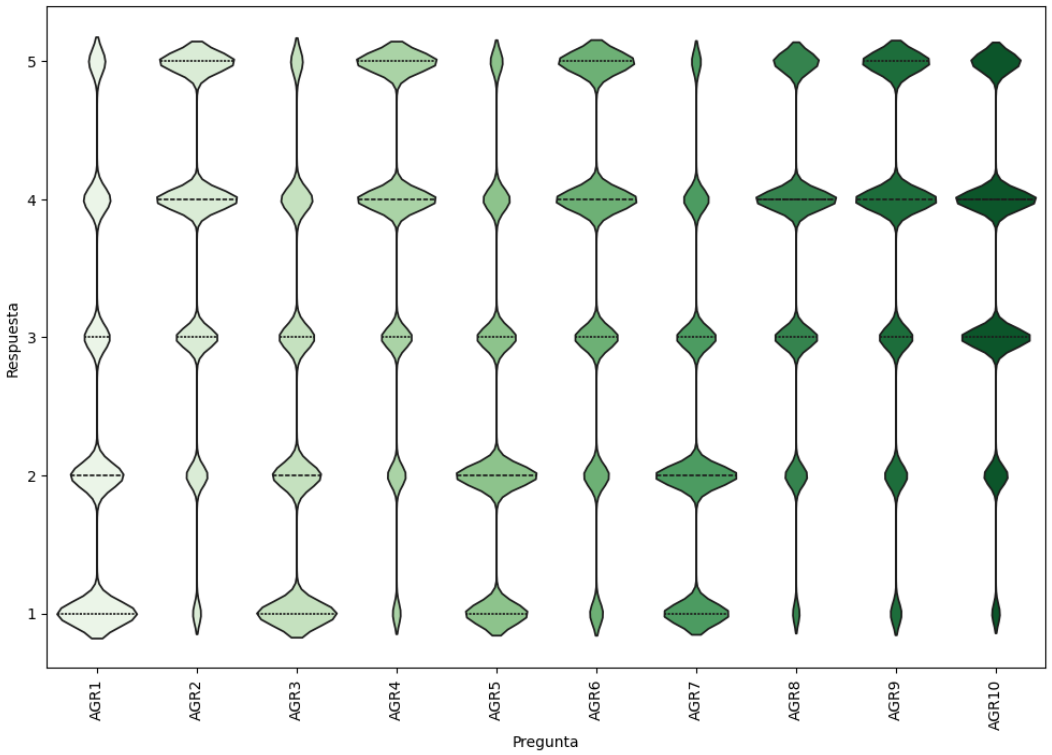

Figura 12: Distribución de respuestas para cada pregunta de Amabilidad

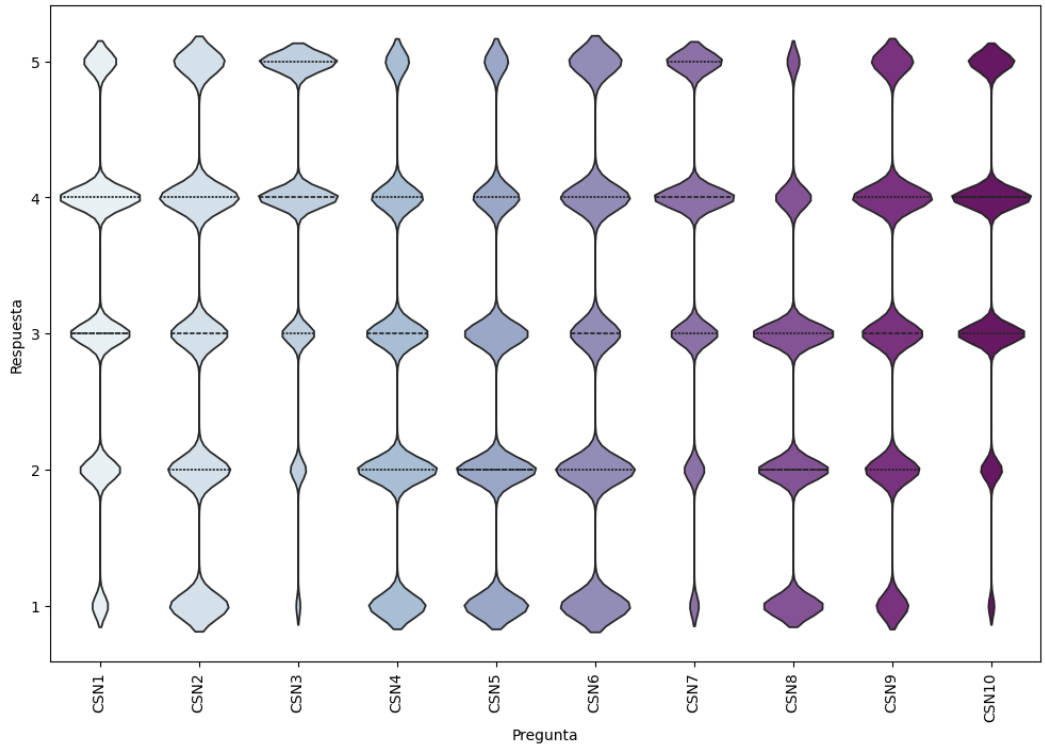

Figura 13: Distribución de respuestas para cada pregunta de Conciencia

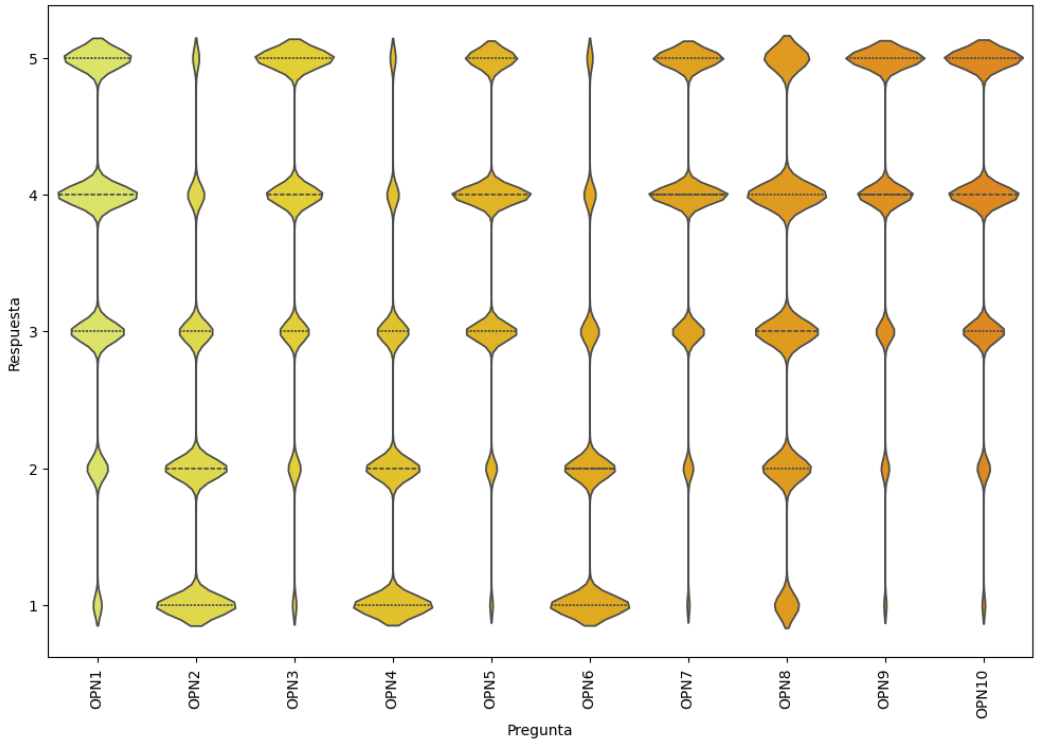

Figura 14: Distribución de respuestas para cada pregunta de Apertura a la experiencia

Estos diagramas de violín muestran la distribución de respuestas para cada grupo de preguntas asociado a cada dimensión de la personalidad.

- El eje x representa las preguntas del grupo.
- El eje y representa las respuestas dadas por los participantes.
- El ancho del violín indica la densidad de respuestas en diferentes valores.
- Las líneas discontinuas representan los cuartiles de los datos.
    - La línea inferior representa el primer cuartil (25º percentil)
    - La línea del medio representa la mediana (50º percentil)
    - La línea superior representa el tercer cuartil (75º percentil).

  Estas líneas proporcionan una indicación visual de la dispersión y la centralidad de los datos en cada respuesta a las preguntas.
- Las partes superiores e inferiores del violín representan el rango de respuestas, que como sabemos está entre 1 y 5.

Primero se observan lógicamente saltos discretos entre los 5 valores, se comprueba con las líneas que unen cada valor en el gráfico, esto quiere decir que no existen valores intermedios entre estos 5 para ninguna respuesta en los datos.

Atendiendo a la distribución en los grupos de preguntas de *Extraversión* y *Neuroticismo*, no se observa una clara relación entre la valoración de los participantes (ancho del violín) y los grupos que generamos según la descripción de las preguntas [Fig.10 y Fig.11]. Parece que no hay un patrón claro de asociación en base a descripción de preguntas para estos casos.

Sin embargo, para las otras 3 dimensiones de personalidad [Fig.12, Fig.13 y Fig.14], se observa claramente la relación entre las respuestas de los participantes y las características (preguntas) que se asociaron previamente para el estudio de correlación.

3.5 Aprendizaje automático

El *clustering*, es un técnica de aprendizaje automático no supervisado que consiste en agrupar los datos en conjuntos homogéneos llamados "*clusters*". El objetivo es que los puntos de datos dentro de un mismo *cluster* sean muy similares entre sí en términos de alguna medida de similitud o distancia, mientras que los puntos de datos de *clusters* diferentes sean distintos entre sí. A diferencia de los algoritmos de clasificación, donde se conocen las etiquetas de clase para entrenar al modelo, los algoritmos de *clustering* no tienen esa información y deben descubrir la estructura subyacente en los datos por sí mismos.

Hay varios tipos de algoritmos de *clustering*, que se pueden clasificar en función de cómo se realiza el proceso de agrupación de los datos. Algunas de las categorías comunes de *clustering* son las siguientes:

- ***Clustering* particional:** Estos algoritmos dividen el conjunto de datos en un número predefinido de *clusters*, donde cada punto de datos pertenece a uno de estos *clusters*. *K-means*, *K-medoids* y CLARA son ejemplos prominentes de algoritmos particional. Son eficientes y escalables, pero pueden ser sensibles a la inicialización de centroides y pueden no funcionar bien con *clusters* de formas y tamaños irregulares.

- **Algoritmos basados en densidad:** Estos algoritmos encuentran *clusters* identificando regiones de alta densidad en el espacio de características. Consideran como *clusters* áreas densamente pobladas, separadas por regiones de baja densidad. DBSCAN es un ejemplo prominente de algoritmo de *clustering* basado en densidad. Son robustos ante ruido y pueden detectar *clusters* de formas arbitrarias, pero pueden tener dificultades con *clusters* de densidades variables o con tamaños de *cluster* desconocidos.

- ***Clustering* jerárquico:** Estos algoritmos construyen una estructura de *clusters* en forma de árbol, donde los *clusters* pueden ser *subclusters* de otros *clusters* más grandes. Pueden ser *aglomerativos*, donde los *clusters* se forman fusionando *clusters* más pequeños, o divisivos, donde los *clusters* más grandes se dividen en *clusters* más pequeños. Estos algoritmos son útiles para visualizar la estructura de los datos y pueden identificar *clusters* a diferentes niveles de granularidad.

- **Algoritmos basados en modelos:** Estos algoritmos asumen que los datos fueron generados a partir de un modelo probabilístico y buscan ajustar este modelo a los datos para identificar *clusters*. Algunos ejemplos son el *Gaussian Mixture Model* (GMM) y el *Bayesian Gaussian Mixture Model* (BGMM). Son capaces de capturar estructuras de *cluster* más complejas y pueden manejar *clusters* de diferentes formas y tamaños, pero pueden ser computacionalmente intensivos y sensibles a la elección del modelo.

- **Algoritmos basados en subespacios:** Estos algoritmos encuentran *clusters* en subconjuntos de características o subespacios del espacio de características completo. Son útiles cuando los datos tienen características relevantes en diferentes subconjuntos de características. El *Subspace Clustering Algorithm* (SCA) y el *CLIQUE algorithm* son ejemplos de este enfoque. Pueden descubrir *clusters* en datos de alta dimensionalidad y son robustos ante datos ruidosos, pero pueden ser computacionalmente costosos y sensibles a la elección de los parámetros.

En nuestro caso, tenemos 696,845 instancias con 50 características en la misma escala. Son datos estructurados y regulares que no fueron generados a partir de modelos probabilístico. Además, no nos interesa trabajar con subespacios del conjunto pues queremos estudiar los resultados del test completo. De esta manera, dada la naturaleza de nuestros datos, vamos a trabajar con algoritmos de *clustering* particional, basado en densidades y jerárquico.

3.5.1 Clustering particional

El *clustering* particional es un proceso que descompone un conjunto de datos en un conjunto de *clusters* disjuntos. Aquí encontramos diferentes algoritmos que pueden utilizarse, la decisión de cuál utilizar, de nuevo, se fundamenta en la naturaleza de nuestros datos. Observamos brevemente las características de algoritmos de *clustering* particional:

- ***K-Means***: Es uno de los algoritmos de *clustering* más simples y utilizados Divide los datos en K *clusters*, donde K es un número predefinido. Comienza con K centroides aleatorios y asigna cada punto al *cluster* cuyo centroide está más cercano. Agrupa los datos tratando de separar las muestras en grupos de igual varianza, minimizando un criterio conocido como la inercia Luego recalcula los centroides como la media de los puntos en cada *cluster* y repite hasta que no haya cambios significativos en la asignación de puntos o se alcance un número máximo de iteraciones.

Consideraciones: Bueno para *datasets* grandes o pequeños con *clusters* de forma esférica o globular. Sensible a la inicialización de centroides y afectado por outliers.

- **MiniBatch *K-Means***: Variante del *K-means* diseñada para manejar grandes volúmenes de datos. En lugar de actualizar todos los puntos en cada iteración, utiliza *mini-batches* aleatorios de datos para actualizar los centroides de manera más eficiente.

  Consideraciones: Más rápido que *K-means* para grandes *datasets*, pero puede producir resultados ligeramente diferentes debido al uso de *mini-batches*. Adecuado cuando se tiene una gran cantidad de datos y se requiere eficiencia computacional.

- ***K-Medoids***: Similar a *K-means*, pero utiliza objetos reales de los datos como representantes de los *clusters* en lugar de centroides. Inicializa los *medoids* aleatoriamente y asigna puntos al *medoid* más cercano. Luego, actualiza los *medoids* eligiendo el punto que minimiza la distancia total a los demás puntos en su *cluster*.

  Consideraciones: Más robusto que *K-means* frente a outliers y ruido. Adecuado para *datasets* con *clusters* de formas no esféricas o no convexas.

- **CLARA**: Diseñado para *datasets* grandes que no caben en memoria. Divide el *dataset* en subconjuntos más pequeños, aplica *K-medoids* a cada subconjunto y luego combina los resultados para obtener una solución global.

  Consideraciones: Eficiente para grandes volúmenes de datos, pero puede ser costoso computacionalmente debido al número de subconjuntos necesarios. Adecuado para *datasets* extremadamente grandes o cuando la memoria es limitada.

Para elegir entre estos algoritmos, se considera el tamaño del *dataset*, la forma y distribución de los *clusters*, la presencia de *outliers* y la disponibilidad de recursos computacionales. *K-Means* y *MiniBatch K-Means* son buenos para *datasets* grandes con *clusters* globulares, mientras que K-medoids puede ser más robusto frente a *outliers* y *clusters* no esféricos. CLARA es útil para *datasets* muy grandes o cuando la memoria es limitada.

A priori, *K-Means* y *MiniBatch K-Means* (en caso de ser necesario) parecen las mejores opciones pues tenemos un conjunto de datos grande, pero no tanto como para utilizar CLARA, tenemos recursos suficientes para trabajar con él. Comenzaremos a trabajar con *K-Means* para estudiar la estructura interna de los *clusters* y verificar que es un algoritmo adecuado. Utilizaremos dos técnicas diferentes:

1. **PCA (Análisis de Componentes Principales)**:

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que se utiliza para transformar un conjunto de datos de alta dimensión en un conjunto de datos de menor dimensión, mientras se conserva la mayor cantidad posible de variabilidad presente en los datos originales. PCA funciona mediante la identificación de las direcciones (componentes principales) en las cuales los datos varían más. Estas direcciones son ortogonales entre sí y están ordenadas de tal manera que la primera componente principal captura la mayor variabilidad, la segunda componente principal captura la mayor variabilidad restante, y así sucesivamente. Aplicamos PCA a nuestro caso con 50 características para reducir las dimensiones a 2 componentes principales. Esta reducción permite visualizar los datos en un espacio bidimensional, facilitando la interpretación y el análisis visual de la forma de los *clusters* en los datos.

2. **t-SNE (t-Distributed Stochastic Neighbor Embedding)**:

t-SNE es una técnica de reducción de dimensionalidad no lineal que se utiliza principalmente para la visualización de datos de alta dimensión en un espacio de menor dimensión, típicamente 2 o 3 dimensiones. Se aplica comúnmente en el procesamiento de imágenes, la PNL, los datos genómicos y el reconocimiento de voz. t-SNE funciona optimizando la proximidad relativa de los puntos en el espacio de baja dimensión para que refleje la proximidad de los puntos en el espacio de alta dimensión. Esto se logra mediante la minimización de una divergencia de Kullback-Leibler entre las distribuciones de pares de puntos en los espacios de alta y baja dimensión.

La divergencia de Kullback-Leibler (KL) mide la diferencia entre dos distribuciones de probabilidad. En términos simples, mide cuánta información se pierde cuando se usa una distribución de probabilidad Q para aproximar otra distribución de probabilidad P. No es simétrica, lo que significa que la divergencia de KL de P a Q no es igual a la divergencia de KL de Q a P.

A diferencia de PCA, t-SNE es especialmente útil para capturar la estructura local, patrones complejos y las relaciones no lineales en los datos.

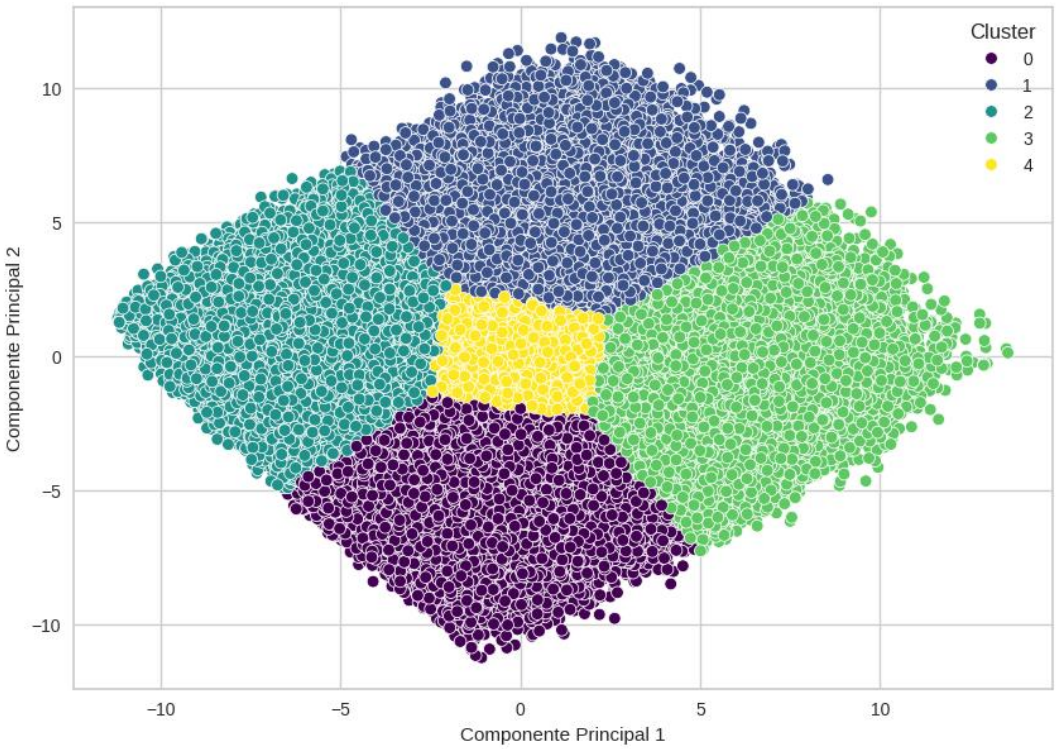

Figura 15: Visualización de clusters con PCA

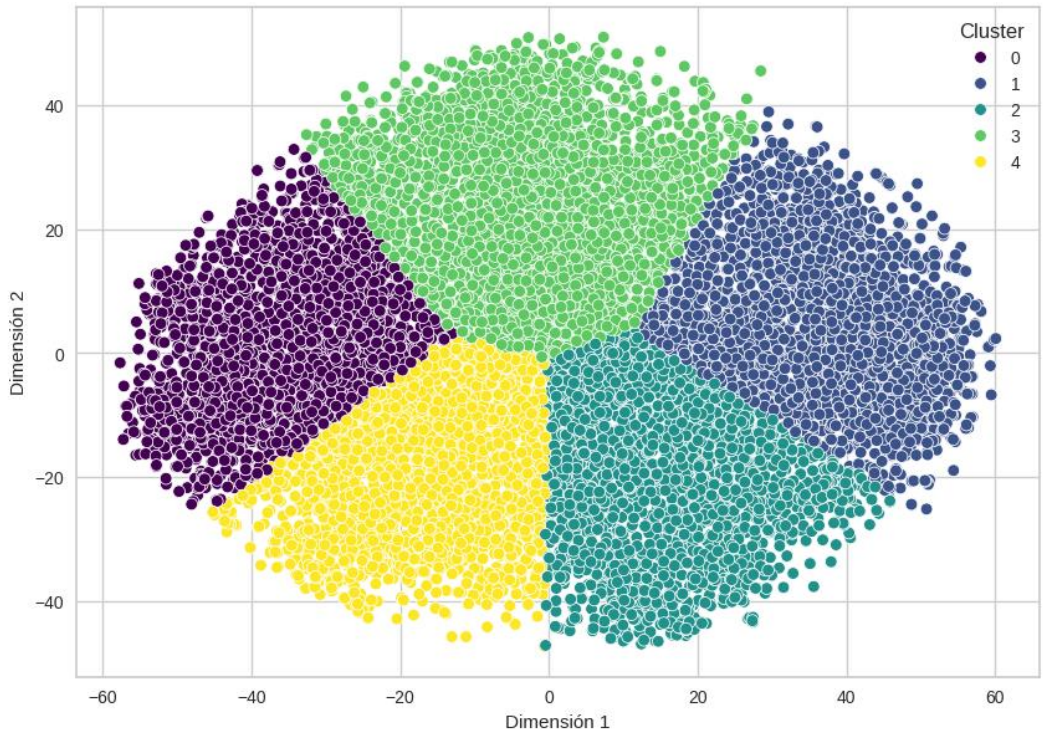

Figura 16: Visualización de clusters con t-SNE

Observamos *clusters* globulares (Fig.15 y Fig.16), es decir, grupos de puntos de datos que están distribuidos de manera similar a una esfera en el espacio de características. Los puntos de datos de los *clusters* están más concentrados cerca de un centroide y se dispersan uniformemente en todas las direcciones alrededor de ese centroide. De esta manera, confirmamos la elección de trabajar con *K-Means*.

El algoritmo alterna dos pasos y finaliza cuando la asignación de instancias a los *clusters* ya no cambia:

1. Elige un punto de datos al azar como el primer centroide.
2. Asigna cada punto de datos al centroide más cercano.
3. Establece cada centroide como la media de los puntos de datos que se le asignan.

De esta forma, *K-means* agrupa los datos tratando de separar las muestras en *k* grupos de igual varianza, minimizando un criterio conocido como la inercia o la suma de cuadrados dentro del grupo. Así pues, comenzamos el proceso estudiando la inercia para distintos valores de *k*.

`KElbowVisualizer` de la librería *Yellowbrick* es una herramienta muy útil para estudiar la inercia y determinar el número óptimo de *clusters* en el algoritmo de *clustering K-means*. Este visualizador ayuda a identificar el punto de inflexión (o "codo") en la gráfica de la inercia, donde se puede observar una disminución significativa en la inercia al agregar más *clusters*.

Estudiaremos esta gráfica para un subconjunto de nuestros datos y comprobar la optimalidad de *clusters* con eficiencia. Para un conjunto de datos de casi 700,000 instancias, una muestra de 50,000 es razonable y generalmente suficiente para capturar las características principales de la distribución y estructura de los datos.

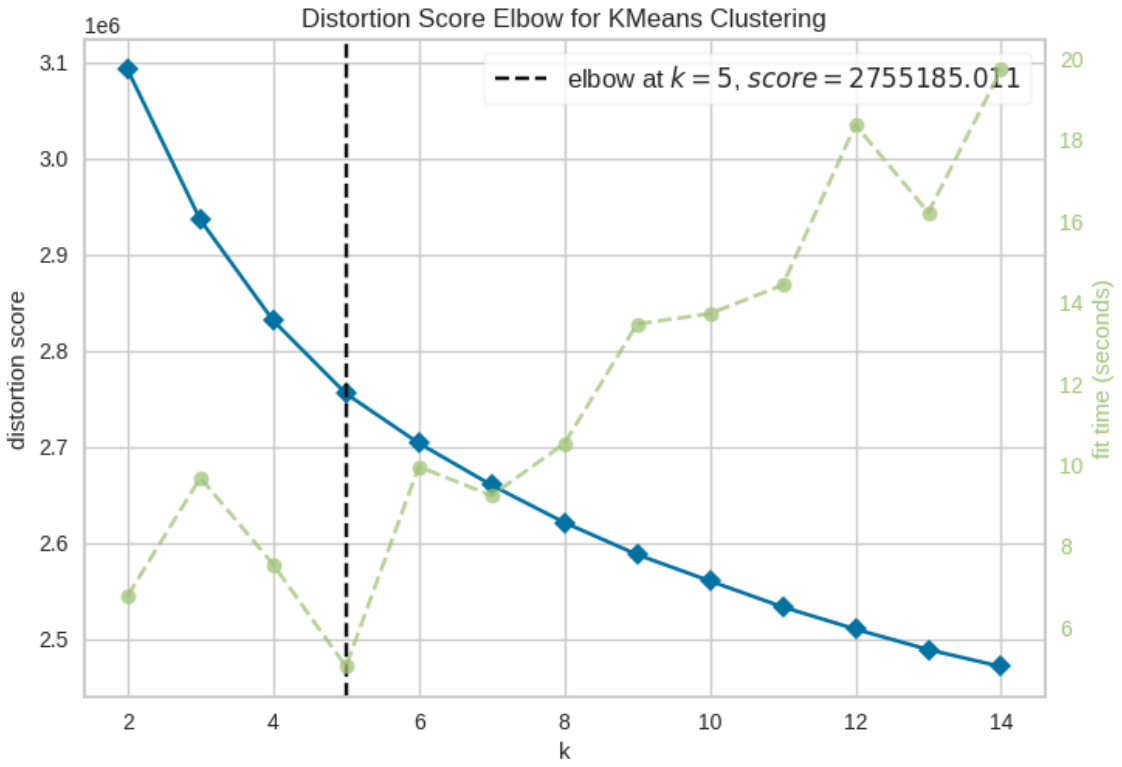

Figura 17: Inercia para distintos valores de k en K-Means

En la gráfica (Fig.17), observamos que la inercia disminuye significativamente hasta llegar a k = 5, después de lo cual la disminución se vuelve menos pronunciada. Esto sugiere que añadir *clusters* más allá de 5 no proporciona una mejora sustancial en la compactación de los datos dentro de los *clusters*. Por lo tanto, k = 5 se considera el número óptimo de *clusters* en este caso, ya que representa un buen equilibrio entre la estructura interna de los *clusters* y la simplicidad del modelo.

Esta técnica para elegir el mejor valor del número de conglomerados es bastante ambigua. Un enfoque más preciso, aunque más caro computacionalmente, es utilizar la puntuación de la silueta.

La puntuación se calcula promediando el coeficiente de silueta de cada muestra, calculado como la diferencia entre la distancia media intraclúster y la distancia media al clúster más cercano de cada muestra, normalizada por el valor máximo. Esto produce una puntuación entre 1 y -1, donde 1 corresponde a *clusters* muy densos y -1 a *clusters* completamente incorrectos.

El visualizador de siluetas muestra el coeficiente de silueta de cada muestra por *cluster*, visualizando qué *clusters* son densos y cuáles no. Esto es particularmente útil para determinar el desequilibrio de los *clusters*, o para seleccionar un valor de *k* comprobando varios visualizadores. Comprobamos esta visualización para valores de *k* que parecen relevantes además del supuesto valor óptimo 5.

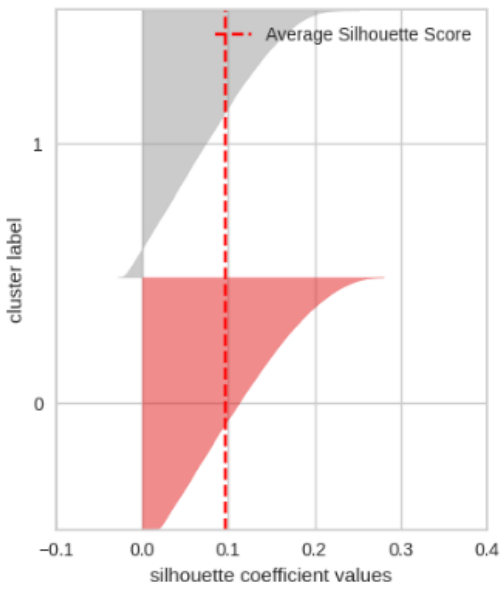

Figura 18: Coeficiente de silueta para k=2

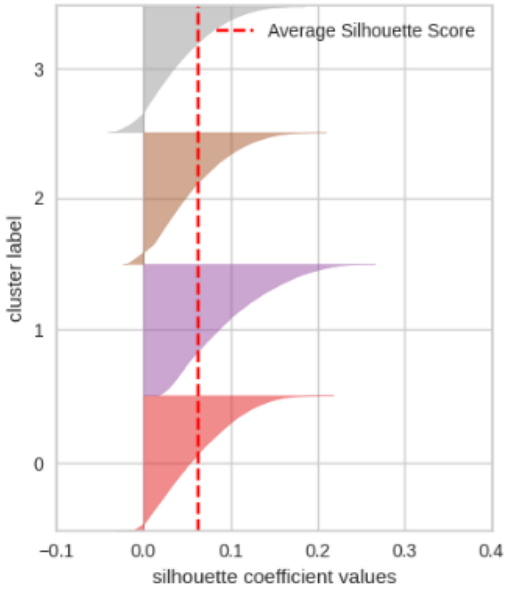

Figura 19: Coeficiente de silueta para k=4

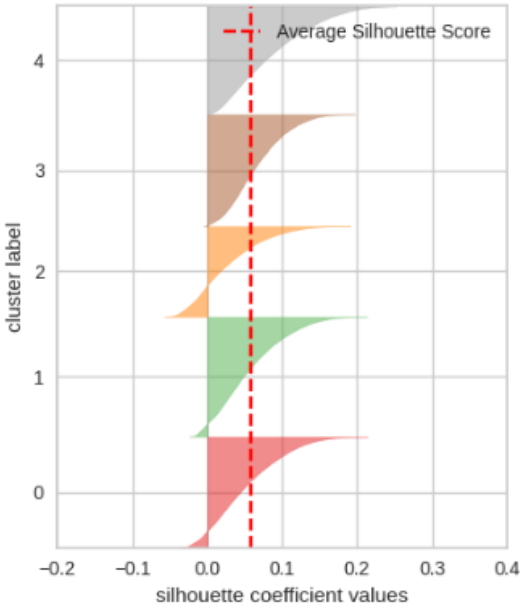

Figura 21: Coeficiente de silueta para k=5

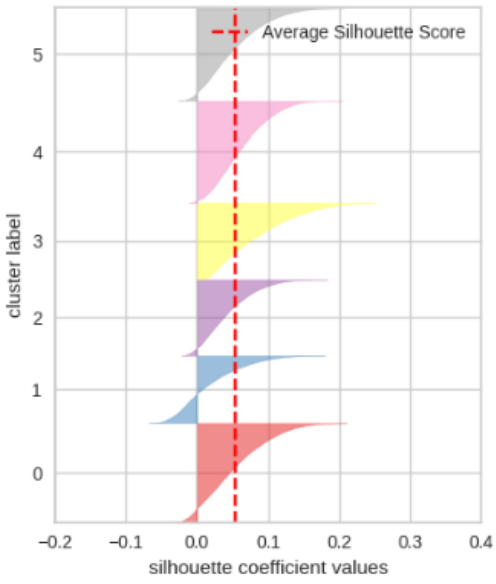

Figura 20: Coeficiente de silueta para k=6

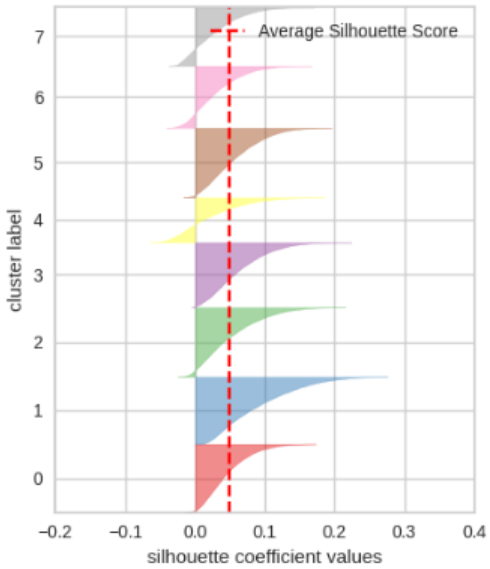

Figura 22: Coeficiente de silueta para k=8

En el gráfico, el ancho de cada barra representa el tamaño del *cluster*. En general, buscamos un gráfico donde la mayoría de las barras estén por encima del valor 0, lo que sugiere una buena calidad de agrupación. Además, queremos minimizar la variabilidad en el tamaño de las barras, lo que indica *clusters* de tamaños relativamente uniformes.

El valor k = 2 (Fig.18) parece mostrar buenos resultados, sin embargo, puede agrupar datos que tienen una estructura más compleja en solo dos grupos, lo que puede no ser suficiente para capturar todas las variaciones en los datos. Para k = 4 (Fig.19) también se pueden tener buenos resultados para el coeficiente de silueta, pero muestra instancias con valores negativos de esta métrica en todos los *clusters*, lo que indica que en todos los *clusters* hay alguna instancia que posiblemente esté mal agrupada.

Para k = 6 (Fig.21) y k = 8 (Fig.22), aparte de tener bastantes *clusters* con instancias negativas, observamos una varianza significativa entre el ancho de los *clusters*, esto implica que los *clusters* no son uniformes.

Así pues, pese a comprobar que para ningún valor k todas las instancias están perfectamente agrupadas, el valor k = 5 (Fig.20) muestra 2 *clusters* sin error, y tres con algunos participantes con un bajo pero negativo coeficiente de silueta. Además, muestra bastante uniformidad entre *clusters*.

Procedemos pues a crear el modelo de *K-Means* para el conjunto completo con 5 *clusters*, que según hemos analizado, es el número de *clusters* que será más efectivo. Visualizamos durante el desarrollo en el *notebook* algunas de las predicciones del modelo y el número de participantes por *cluster* (Fig.23).

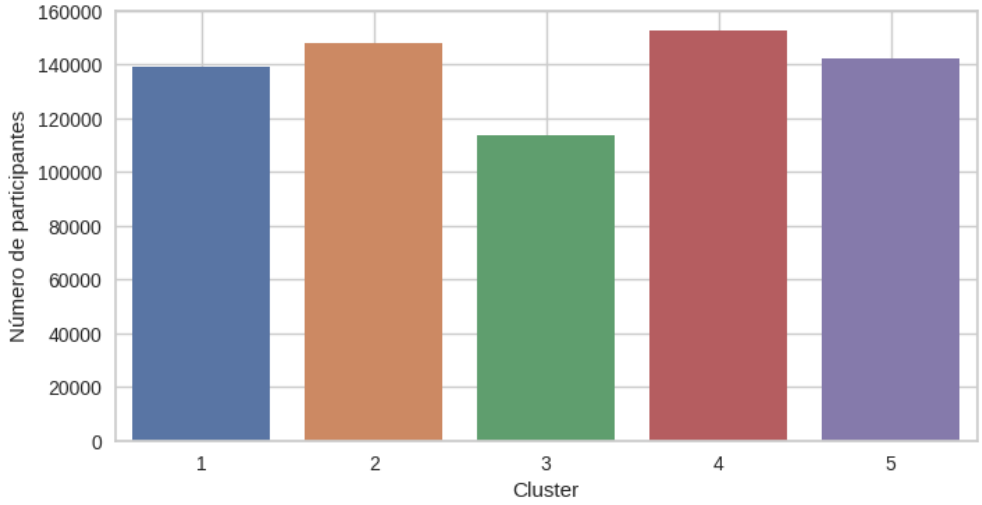

Figura 23: Número de participantes por cluster

Para seguir analizando los *clusters*, realizamos visualizaciones en 2 y 3 dimensiones las predicciones de los *clusters* con PCA.

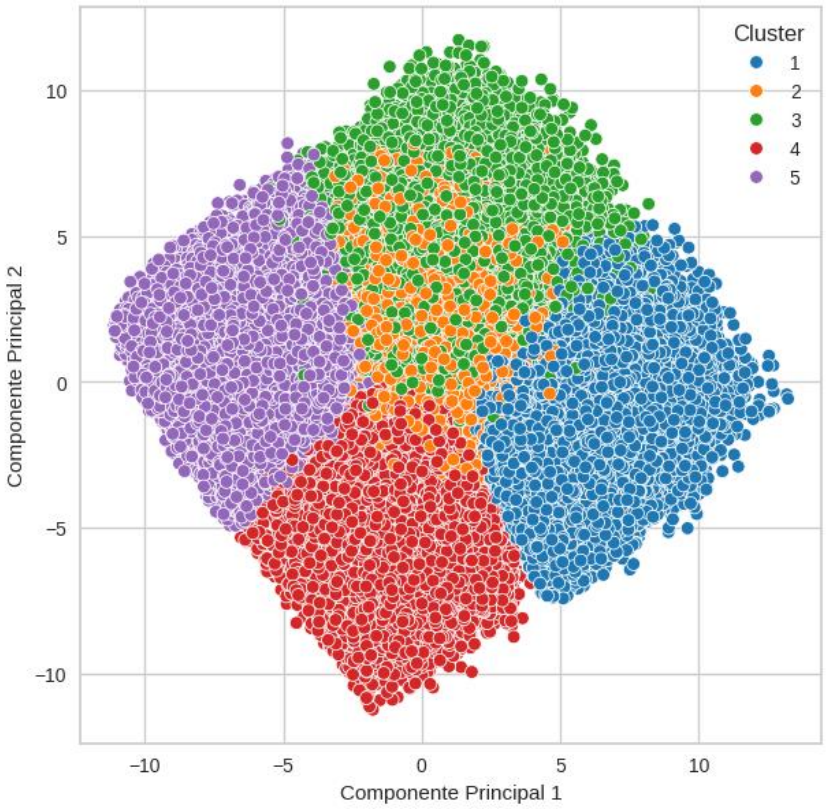

Figura 24: Clusters con PCA en 2 componentes principales

El gráfico de *scatterplot* (Fig.24) muestra la distribución de los cinco *clusters* en un espacio bidimensional después de la reducción de dimensionalidad con PCA.

Los cuatro *clusters* periféricos, ubicados en las esquinas del gráfico, muestran una agrupación densa y compacta, lo que sugiere una alta similitud entre las muestras dentro de cada *cluster*. Por otro lado, el *cluster* central (*cluster* 2) exhibe una dispersión más amplia y una falta de agrupación clara, lo que indica una mayor variabilidad y heterogeneidad en las muestras asignadas a este *cluster*.

Además, se observa una superposición significativa entre el *cluster* central y el *cluster* 3, lo que sugiere una posible ambigüedad en la asignación de muestras entre estos dos *clusters*.

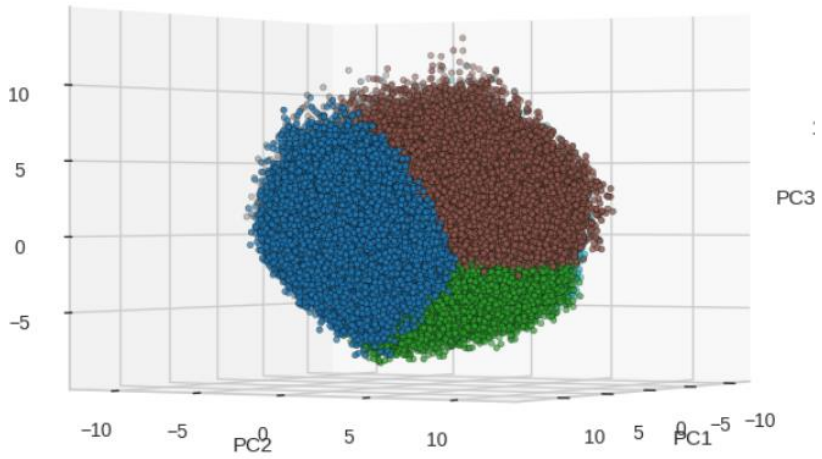

Figura 26: Clusters con PCA en 3 componentes principales (Vista 1)

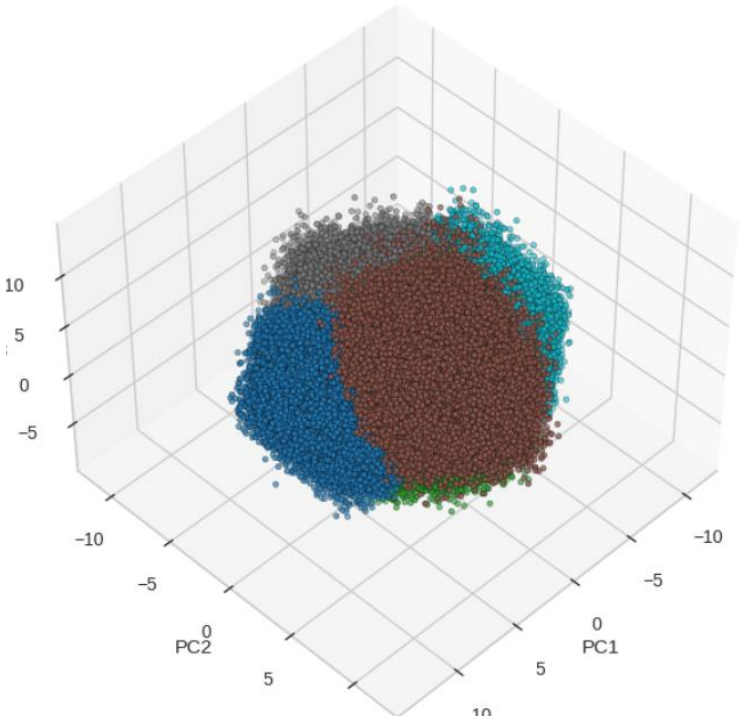

Figura 25: Clusters con PCA en 3 componentes principales (Vista 2)

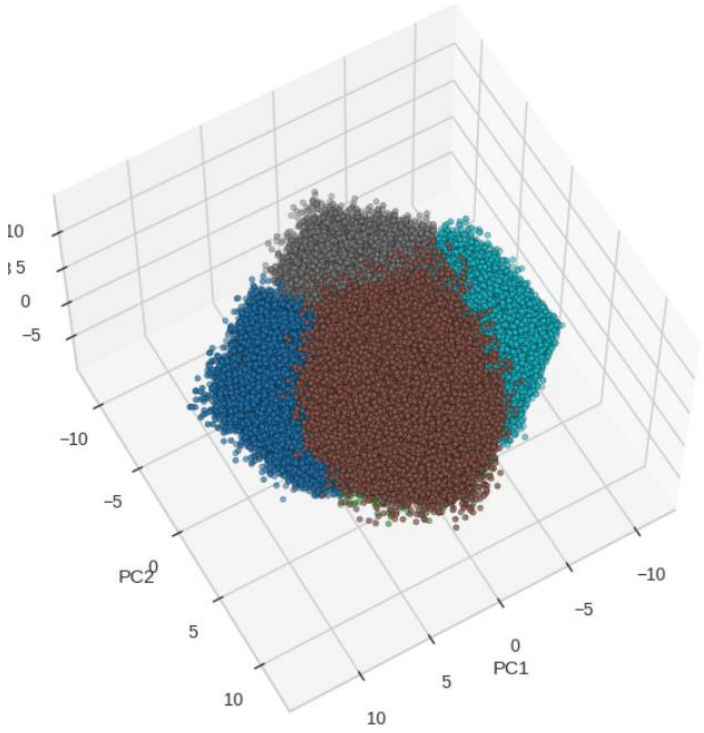

Figura 27: Clusters con PCA en 3 componentes principales (Vista 3)

Observando este tipo de visualizaciones, antes de hacer una interpretación, hacemos las siguientes anotaciones:

- El PCA reduce la dimensionalidad de los datos originales (con 50 características) a solo tres componentes principales. Esta reducción puede causar que las diferencias entre *clusters* se vean atenuadas, ya que la información de los datos se comprime en menos dimensiones. Es posible que la estructura original de los datos no se represente completamente en tres dimensiones.
- PCA tiende a agrupar datos que están correlacionados, lo que puede resultar en *clusters* que se ven más juntos en el espacio reducido.
- Las primeras tres componentes principales seleccionados por PCA capturan la mayor parte de la varianza de los datos, pero no toda. Si los componentes principales no capturan suficiente varianza, los *clusters* pueden parecer más juntos de lo que realmente están en el espacio original de mayor dimensión.

En cuanto a la interpretación, todos los *clusters* aparecen juntos y densos, lo que sugiere que las diferencias entre estos *clusters* no son suficientemente grandes en las tres dimensiones principales seleccionadas. Esto puede indicar que los *clusters* son similares en términos de las principales características que están siendo capturadas por las primeras tres componentes principales.

Además, el solapamiento de instancias entre *clusters* indica que hay algunos puntos de datos que no se distinguen claramente entre los *clusters* en el espacio reducido. Esto puede deberse a la pérdida de información durante la reducción de dimensionalidad o a la naturaleza intrínseca de los datos.

Así pues, pese a que es útil realizar una visualización de nuestros *clusters* en 2 y 3 dimensiones, no vamos a basar ciegamente nuestra interpretación en estas gráficas.

Se ha analizado el algoritmo *K-means* y su aplicabilidad el caso de nuestros datos, se ha estudiado el valor óptimo del número de *clusters* para el algoritmo y se ha estudiad con visualizaciones el significado de estas agrupaciones. Ahora entraremos a estudiar otros enfoques de *clustering* y volveremos a *K-Means* cuando sea necesario.

3.5.2 DBSCAN

DBSCAN, o *Density-Based Spatial Clustering of Applications with Noise*, es un algoritmo de *clustering* ampliamente utilizado en *machine learning*. Su funcionamiento se basa en la idea de identificar regiones densas de puntos en el espacio de características, considerando puntos que están dentro de un radio específico como parte del mismo *cluster*. De esta manera, DBSCAN es capaz de encontrar *clusters* de forma eficiente sin requerir que el número de *clusters* sea especificado de antemano, y es especialmente útil para identificar *clusters* de forma arbitraria y con densidades variables.

El algoritmo asigna cada punto de datos a uno de tres tipos: núcleo, borde o punto de ruido. Un punto es considerado como núcleo si hay un número mínimo de puntos de su vecindario (definido por los parámetros del algoritmo: `epsilon` y `minPts`) dentro, lo que indica una región densa. Los puntos que están dentro del vecindario de un punto núcleo, pero no son núcleos se clasifican como puntos de borde, mientras que los puntos que no están dentro del vecindario de ningún punto núcleo se consideran ruido.

Una de las principales ventajas de DBSCAN es su capacidad para identificar *clusters* de forma automática, lo que lo hace útil en situaciones donde no se conoce el número de *clusters* a priori o cuando los *clusters* tienen formas y densidades irregulares. Sin embargo, DBSCAN también tiene sus limitaciones. Por ejemplo, su rendimiento puede verse afectado por la elección de los parámetros, como el radio de búsqueda y el número mínimo de puntos requeridos para formar un *cluster*. Además, DBSCAN puede tener dificultades para identificar *clusters* de diferentes densidades en conjuntos de datos de alta dimensionalidad.

Tras probar el algoritmo *K-Means* en nuestros datos, hemos confirmado que los *clusters* son globulares y que no hay presencia significativa de *outliers*. *K-Means* es un algoritmo adecuado para este tipo de situación, ya que es eficiente en términos de tiempo de ejecución y puede manejar conjuntos de datos de gran tamaño. Además, *K-Means* es fácil de implementar e interpretar, lo que lo convierte en una opción atractiva para nuestro proyecto.

Dado que nuestros datos se ajustan a las características que son mejor abordadas por *K-Means* y no se observan las condiciones ideales para aprovechar las ventajas de DBSCAN, se ha decidido no utilizar DBSCAN en este proyecto. En lugar de eso, nos centraremos en explorar otra técnica de *clustering* que pueda ser más adecuada para futuros análisis, concretamente el *clustering* jerárquico.

3.5.3 Clustering jerárquico

Esta técnica construye *clusters* midiendo las disimilitudes entre los datos. Puede utilizarse con cualquier tipo de datos para visualizar e interpretar la relación entre puntos de datos individuales.

En este caso, utilizaremos la agrupación jerárquica para agrupar los puntos de datos y visualizar los *clusters* mediante diagramas de árbol.

Este tipo de *clustering* requiere:

- Especificar el `linkage` (vínculo), una medida de disimilitud entre grupos.
- Elegir `K` número de *clusters*, esto no condiciona la salida del algoritmo y puede hacerse después de que haya terminado.

Existen distintos tipos de vínculos, pero es importante saber que, si todos los grupos son compactos y están bien separados, los diferentes vínculos producen resultados similares. Para tener un primer contacto, trabajamos con un vínculo de tipo `complete` y utilizando como métrica de distancia la de tipo `minkowski`. Mostramos los resultados con *`dendrogram()`* para una muestra representativa de nuestro conjunto de datos.

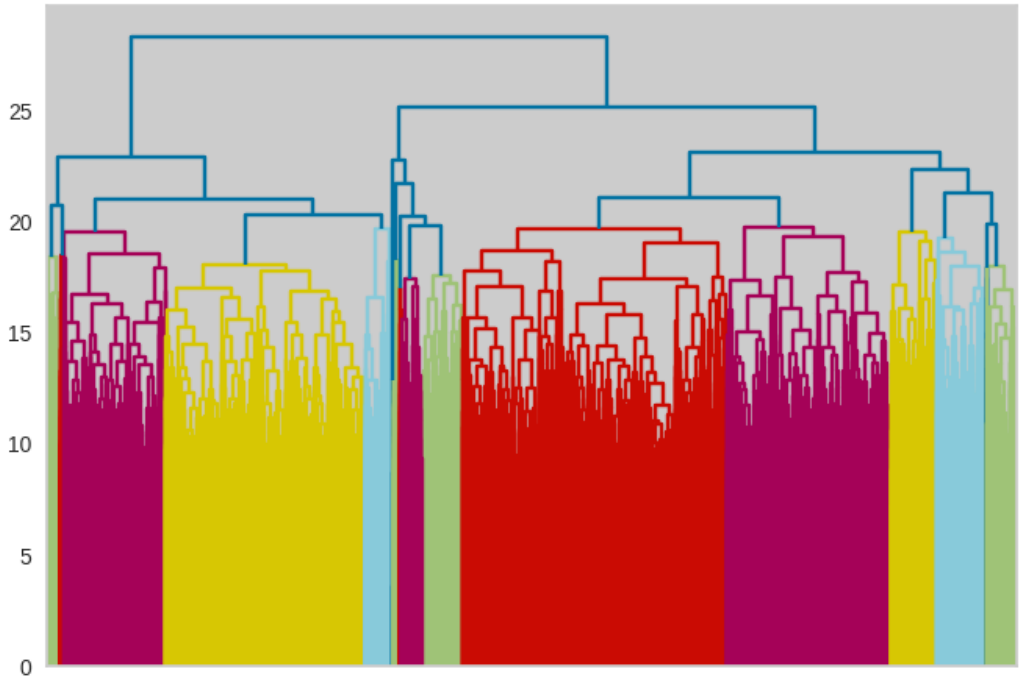

Figura 28: Diagrama de árbol para una muestra de 10000 participantes

Podemos comprobar que pese a estar usando una muestra del conjunto, un diagrama de árbol (Fig.28) en este caso no es la manera óptima de visualizar los datos y *clusters* pues tenemos un conjunto de instancias muy grande.

Cuando el Algoritmo de *Clustering* Jerárquico (HCA) comienza a enlazar los puntos y a encontrar *clusters*, puede primero dividir los puntos en 2 grandes grupos, y luego dividir cada uno de esos dos grupos en otros 2 más pequeños, teniendo 4 grupos en total, que es el enfoque divisivo y descendente.

Alternativamente, puede hacer lo contrario: puede mirar todos los puntos de datos, encontrar 2 puntos que están más cerca el uno del otro, vincularlos, y luego encontrar otros puntos que son los más cercanos a esos puntos vinculados y seguir construyendo los 2 grupos de abajo hacia arriba. Este es el enfoque *aglomerativo* que desarrollaremos.

Al tener un conjunto de datos grande, el enfoque *aglomerativo* puede ser más eficiente computacionalmente, ya que solo se necesita calcular las distancias entre los puntos una vez. Este enfoque permite una exploración gradual de la estructura de los *clusters*, comenzando con cada participante como un *cluster* individual y luego fusionando gradualmente los *clusters* más similares. Esto facilita la comprensión de la estructura general de los datos y proporciona una jerarquía completa de *clusters* que puede ser útil para analizar diferentes niveles de detalle en los patrones de respuesta.

Los pasos del algoritmo de Agrupación Jerárquica *Aglomerativa* (AHC) son:

1. **Inicialización:** Comienza asignando cada punto de datos a su propio *cluster*, es decir, cada punto es un *cluster* inicialmente.

2. **Cálculo de la matriz de distancias:** Calcula la matriz de distancias entre todos los pares de puntos de datos en función de alguna medida de distancia.

3. **Búsqueda de los pares más cercanos:** Encuentra los dos *clusters* más cercanos entre sí en función de la distancia entre ellos. Esto se puede hacer gracias a la matriz de distancias calculada en el paso anterior.

4. **Fusión de *clusters*:** Fusiona los dos *clusters* más cercanos en un nuevo *cluster*. La distancia entre los clústeres fusionados puede ser calculada de varias maneras, dependiendo del criterio de fusión seleccionado (por ejemplo, `'ward'`, `'complete'`, `'average'`, `'single'`).

5. **Actualizar la matriz de distancias:** Actualiza la matriz de distancias para reflejar la distancia entre el nuevo *cluster* y los demás *clusters*. Esto implica recalcular las distancias entre el nuevo *cluster* y todos los demás *clusters*, lo cual puede hacerse de manera eficiente si se utiliza una estructura de datos adecuada.

6. **Repetición:** Repite los pasos 3 a 5 hasta que todos los puntos de datos estén en un solo *cluster* o hasta que se alcance el número deseado de *clusteres*.

Probamos a trabajar con la función `AgglomerativeClustering()` para nuestro conjunto de datos. Resulta que, a pesar de sus ventajas, el *clustering* jerárquico *aglomerativo* puede ser computacionalmente intensivo y consumir grandes cantidades de memoria, especialmente en conjuntos de datos grandes. En este caso, al intentar aplicar el *clustering* jerárquico *aglomerativo* con diversas métricas de distancia y ajustando los parámetros correspondientes, nos hemos enfrentado con una limitación de recursos computacionales. Este enfoque consume una cantidad significativa de memoria, lo que resulta en una incapacidad para completar el desarrollo del modelo debido a la falta de recursos disponibles. Dado este problema de recursos, se descarta esta opción como enfoque práctico en nuestro caso.

4 Resultados

Una vez se ha trabajado con las distintas técnicas de *clustering* aplicables al proyecto, entrando a fondo a cada una de ellas y detectando ventajas, desventajas y usabilidad real en este caso, se ha comprobado que la decisión más acertada para el modelo final es utilizar *K-Means* para cinco *clusters*. Así pues, una vez el modelo está listo, procedemos a analizar los resultados, de manera que quede una vinculación clara entre la cuestión que se aborda y la solución implementada.

Para analizar los resultados del clustering y relacionar los clusters con las dimensiones de personalidad, realizaremos varias visualizaciones y análisis.

4.1 Visualización de dimensiones de personalidad

Podemos visualizar las puntuaciones de las dimensiones de personalidad en *scatter plots* para diferentes *clusters*. Comprobaremos así si hay patrones específicos en cómo las dimensiones se distribuyen entre los *clusters*.

Para realizar esta tarea, calculamos la media a lo largo del eje de las columnas para cada fila. Esto significa que, en lugar de representar cada punto de datos individual, que dejaría ver una simple maya de valores discretos en todos los valores posibles de las 2 dimensiones, representamos el punto promedio de las características relativas a cada dimensión para cada participante.

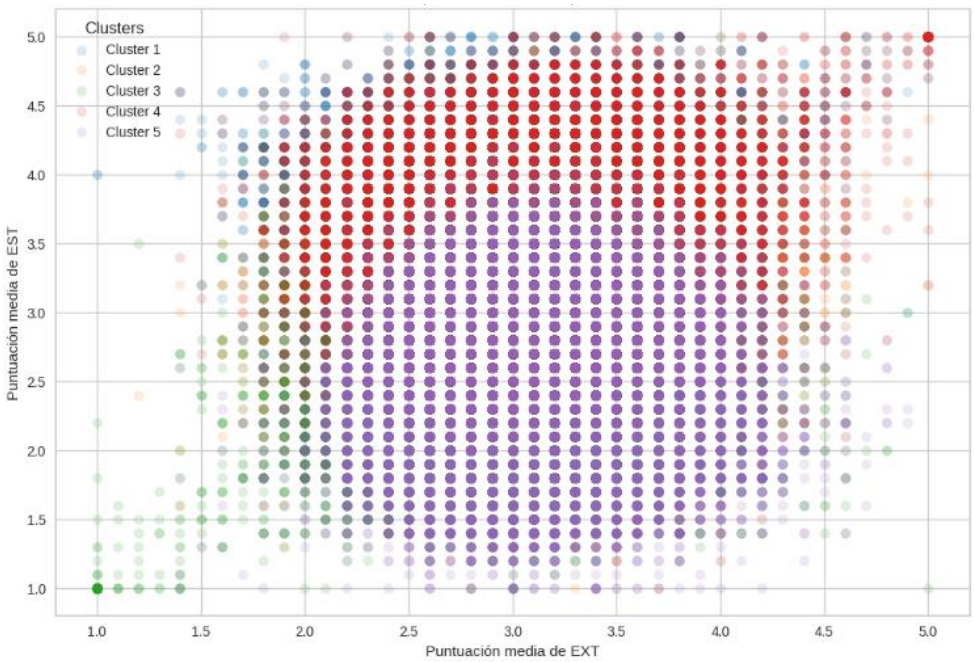

Figura 29: Scatter Plot de EXT vs EST por clusters

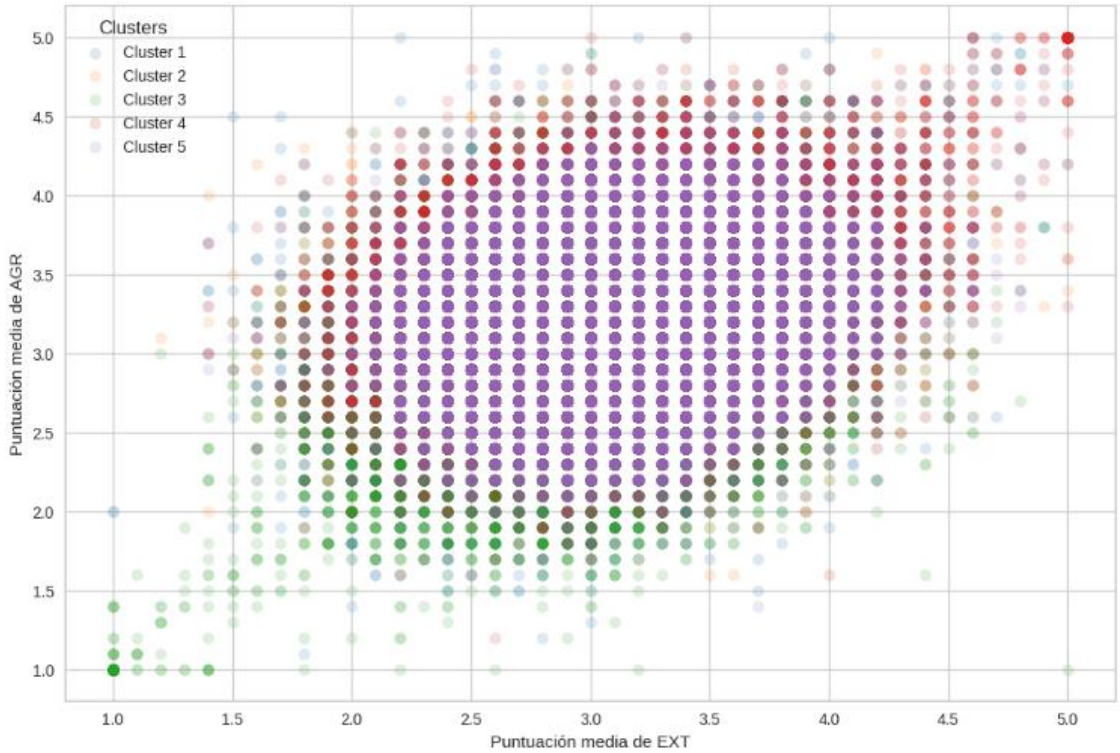

Figura 30: Scatter Plot de EXT vs AGR por clusters

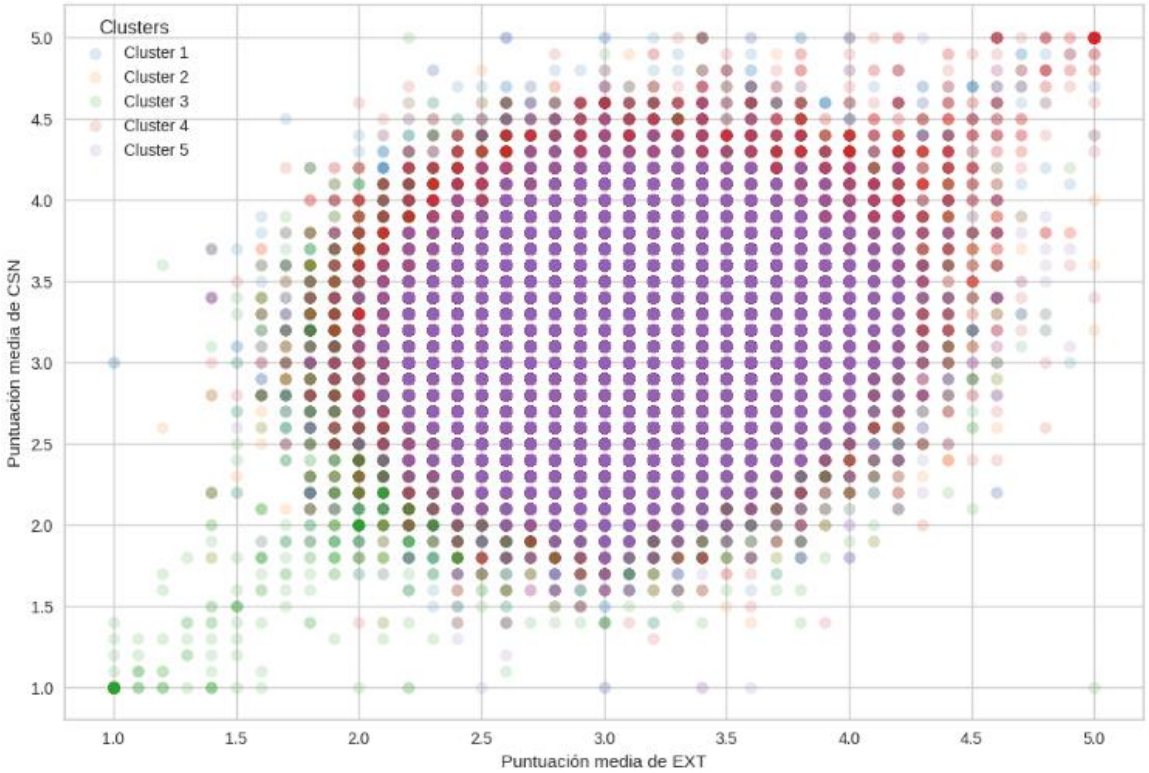

Figura 31: Scatter Plot de EXT vs CSN por clusters

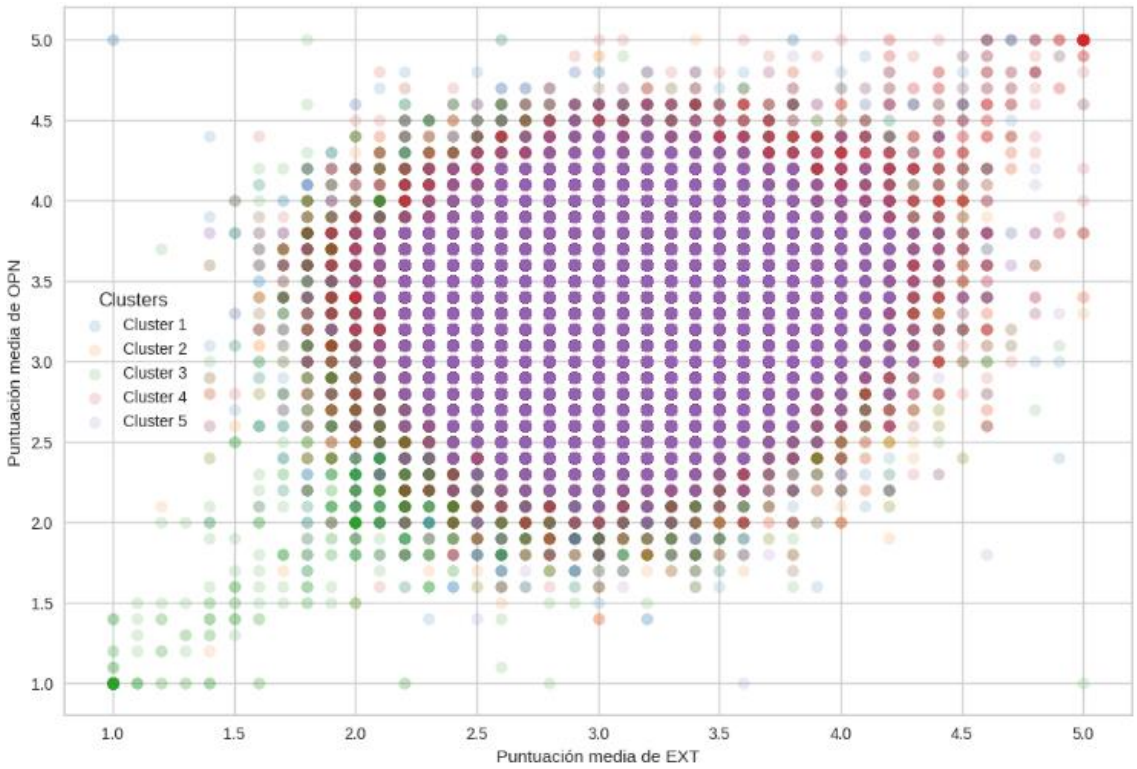

Figura 32: Scatter Plot de EXT vs OPN por clusters

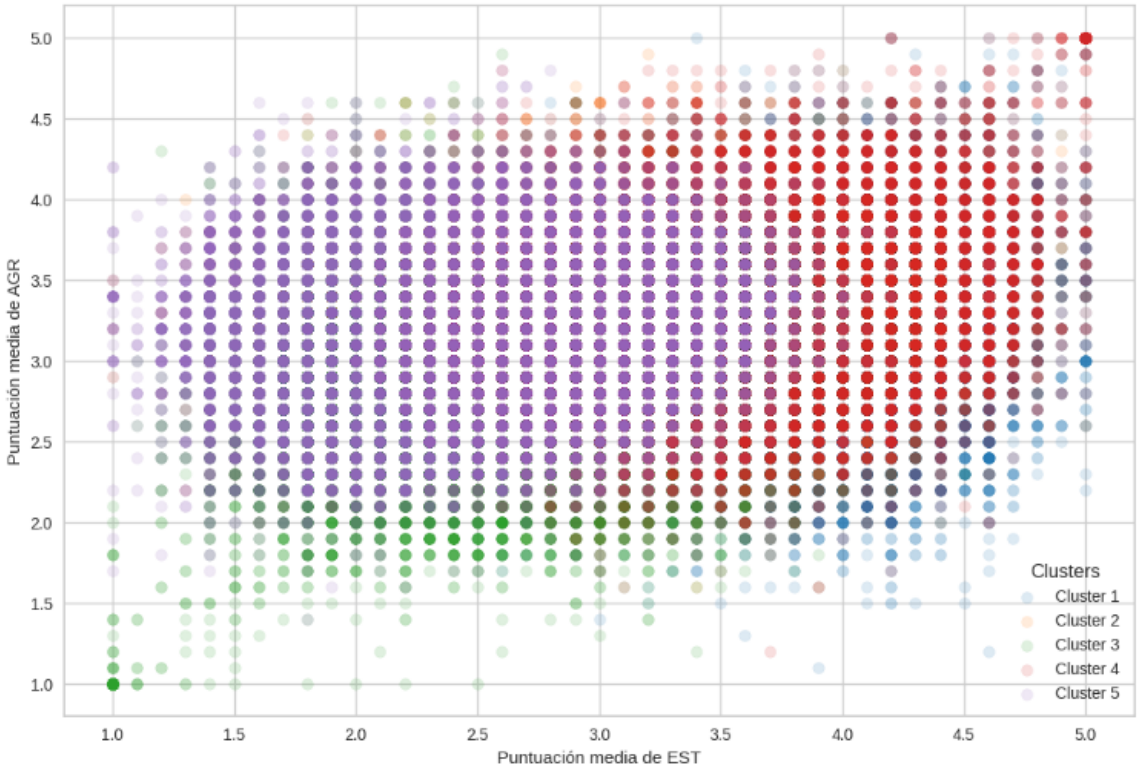

Figura 33: Scatter Plot de EST vs AGR por clusters

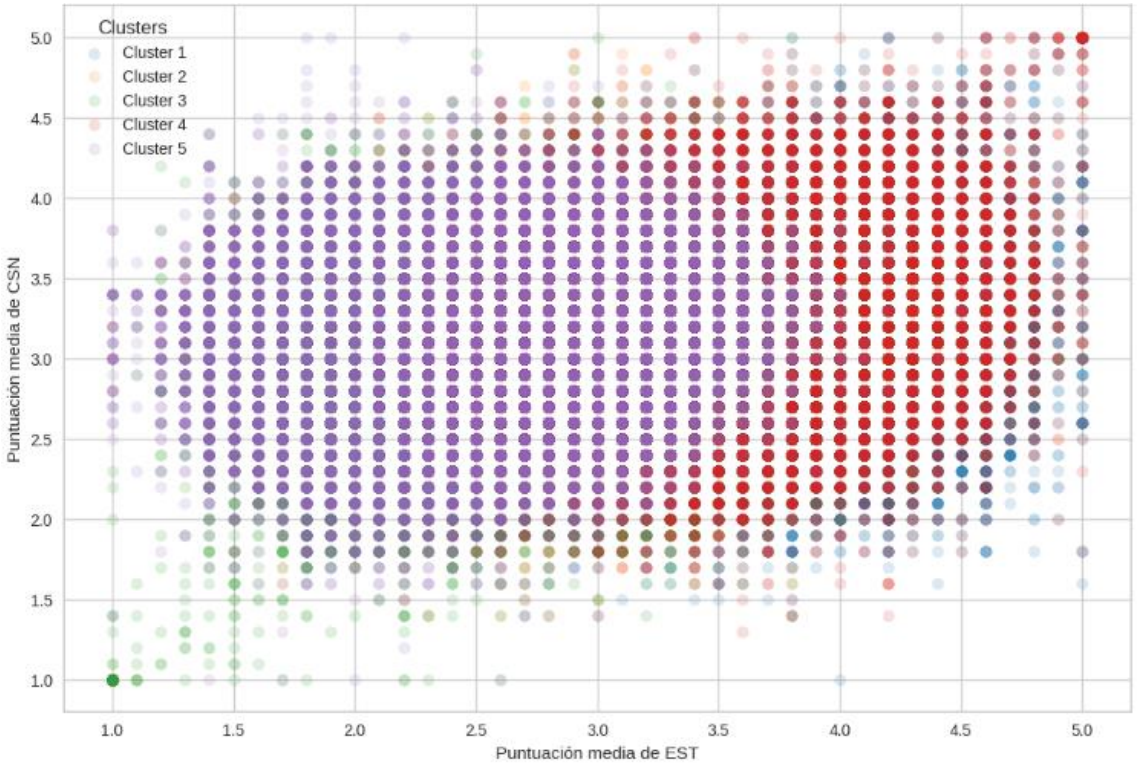

Figura 34: Scatter Plot de EST vs CSN por clusters

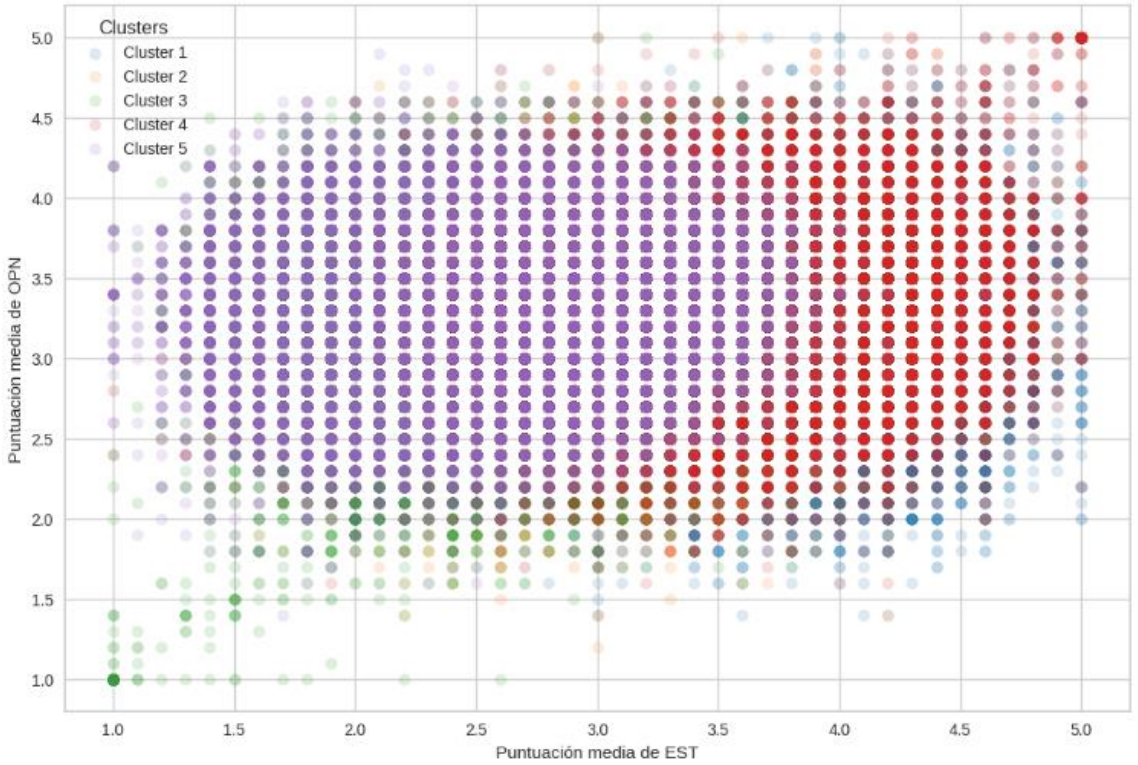

Figura 35: Scatter Plot de EST vs OPN por clusters

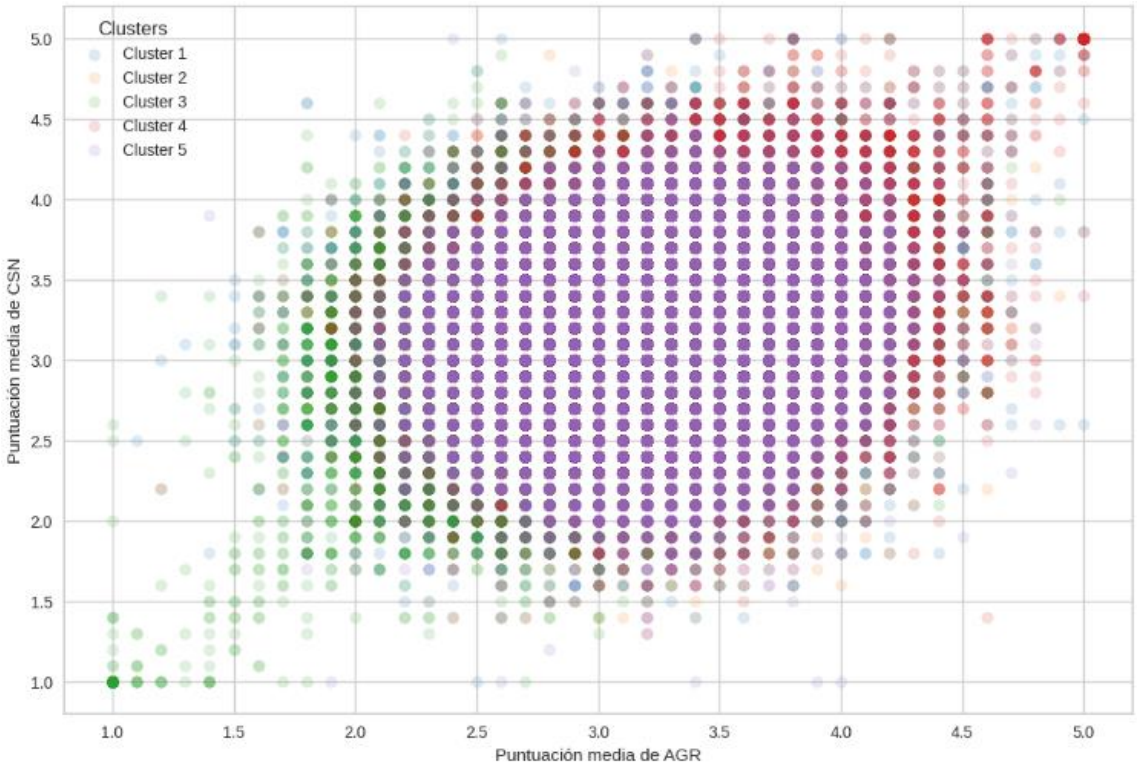

Figura 36: Scatter Plot de AGR vs CSN por clusters

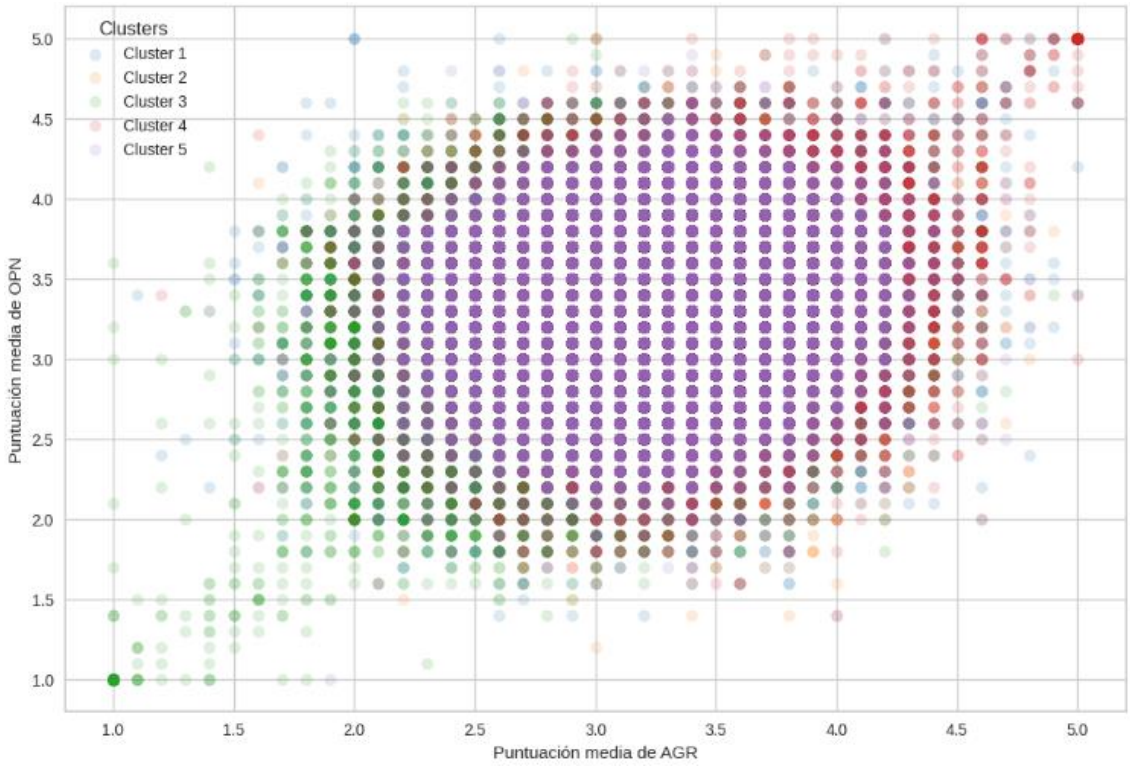

Figura 37: Scatter Plot de AGR vs OPN por clusters

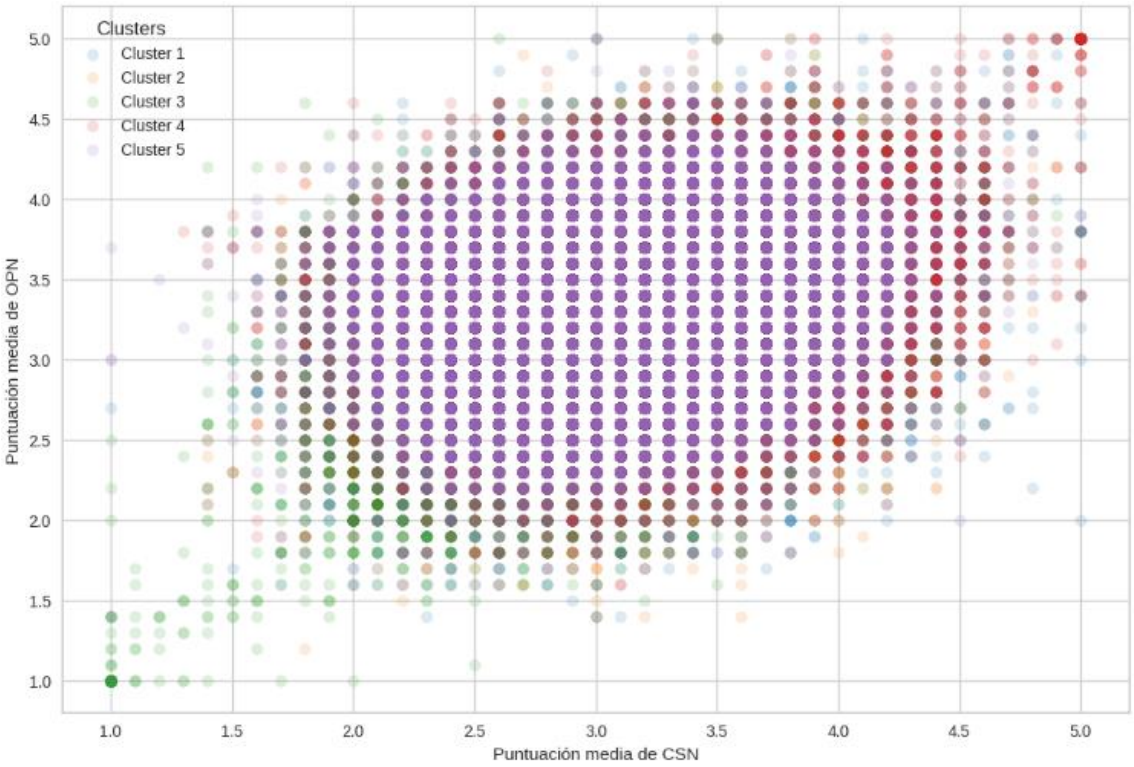

Figura 38: Scatter Plot de CSN vs OPN por clusters

Las gráficas previas muestran la relación entre las dimensiones de personalidad representadas en los ejes X e Y, utilizando los *clusters* generados por el modelo para visualizar cómo se agrupan los individuos en función de estas dimensiones. Cada punto en la gráfica representa la media de las dimensiones analizadas para un sólo participante. Los colores de los puntos representan los diferentes *clusters*, lo que permite identificar visualmente la distribución y separación de los *clusters* en el espacio de las dimensiones de personalidad. Además, la transparencia de los puntos, permite visualizar la densidad de puntos en áreas superpuestas, destacando áreas con una mayor concentración de individuos. Los puntos en la gráfica que presentan colores más fuertes indican una mayor densidad de individuos en esas áreas específicas del espacio de las dimensiones de personalidad.

A continuación, para seguir estudiando las dimensiones, representaremos en cada gráfico sólo un *cluster*, de manera que se evite superposición de *clusters* y se pueda ver con mayor claridad dónde están los puntos de datos asociados a cada participante.

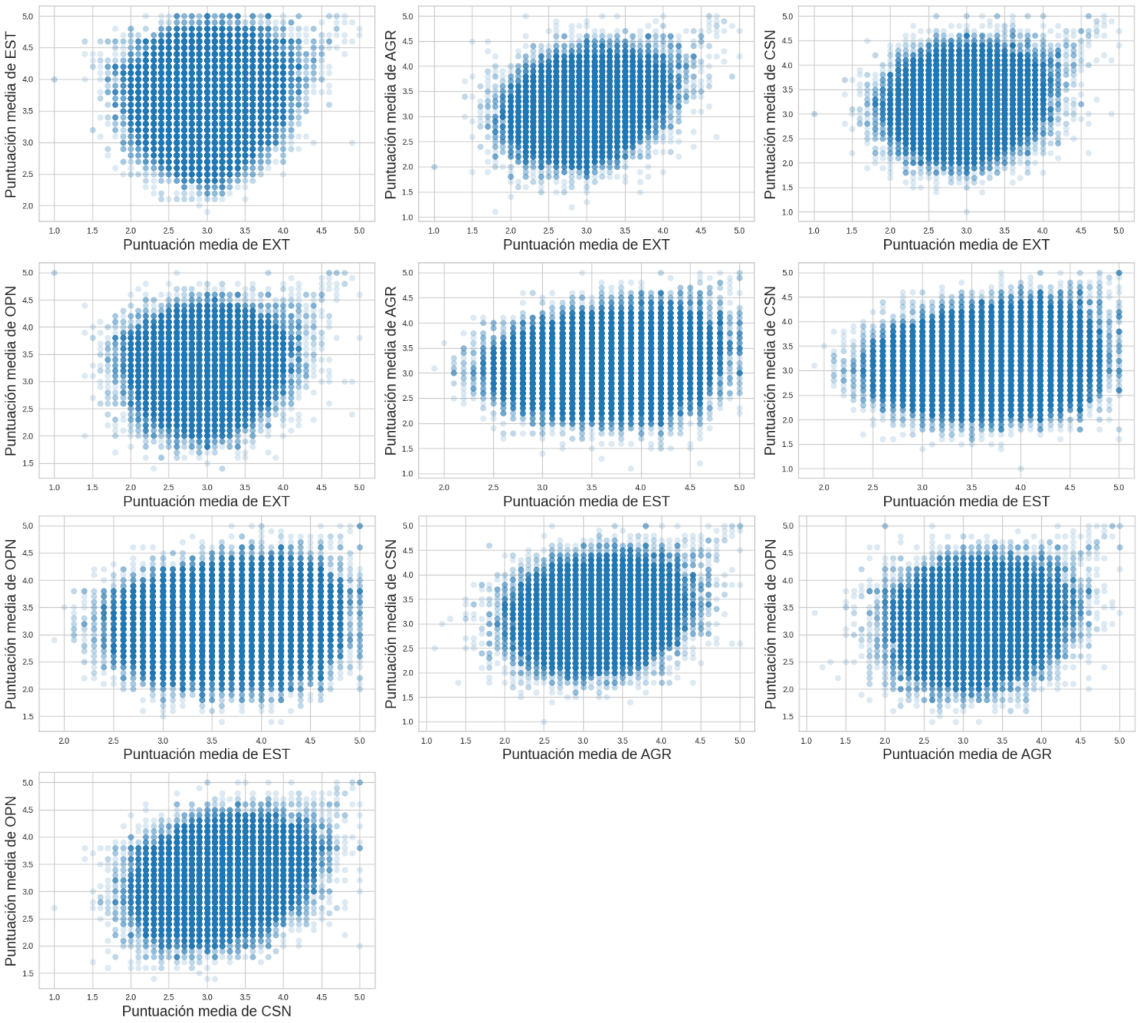

Figura 39: Scatter Plots por pares de dimensiones del Cluster 1

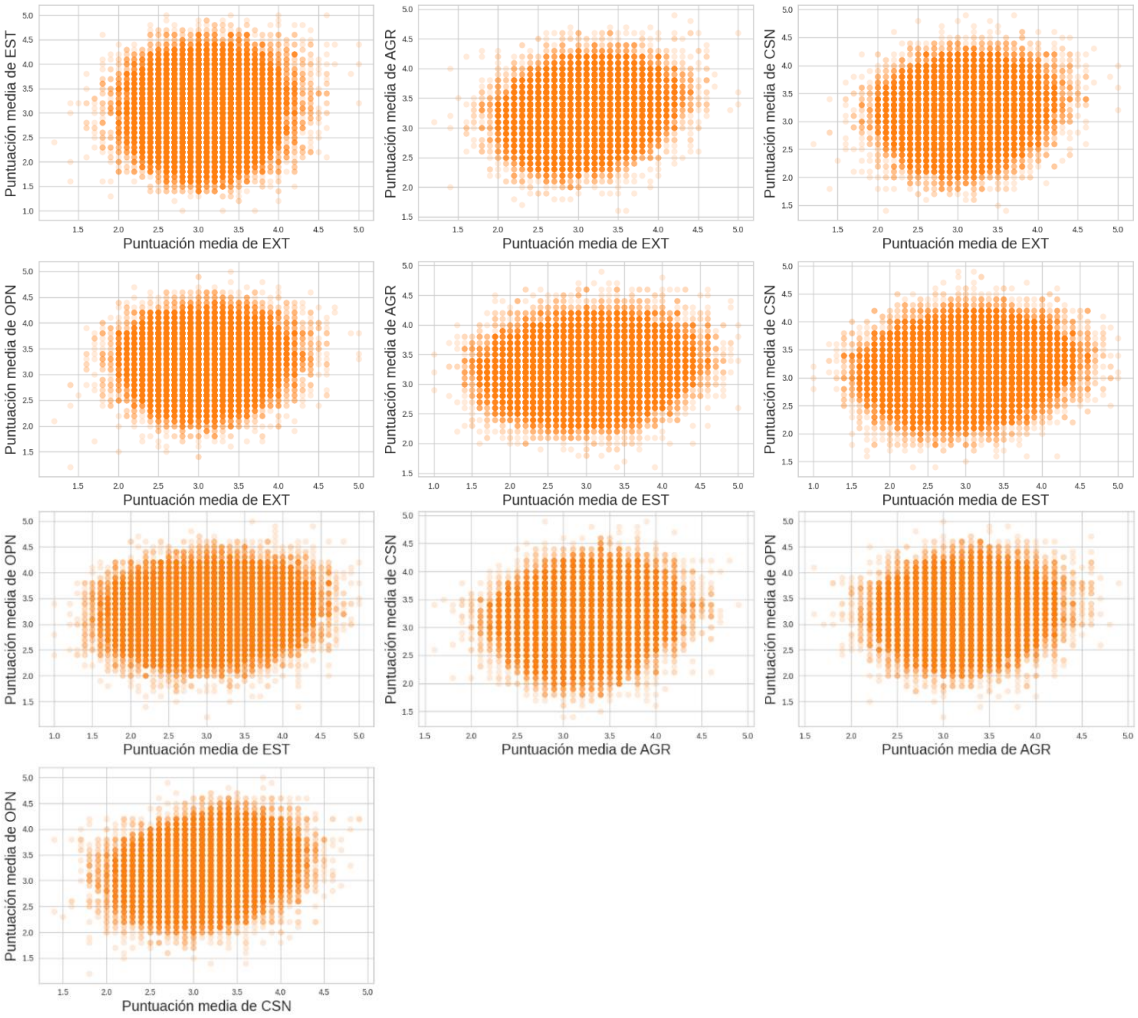

Figura 40: Scatter Plots por pares de dimensiones del Cluster 2

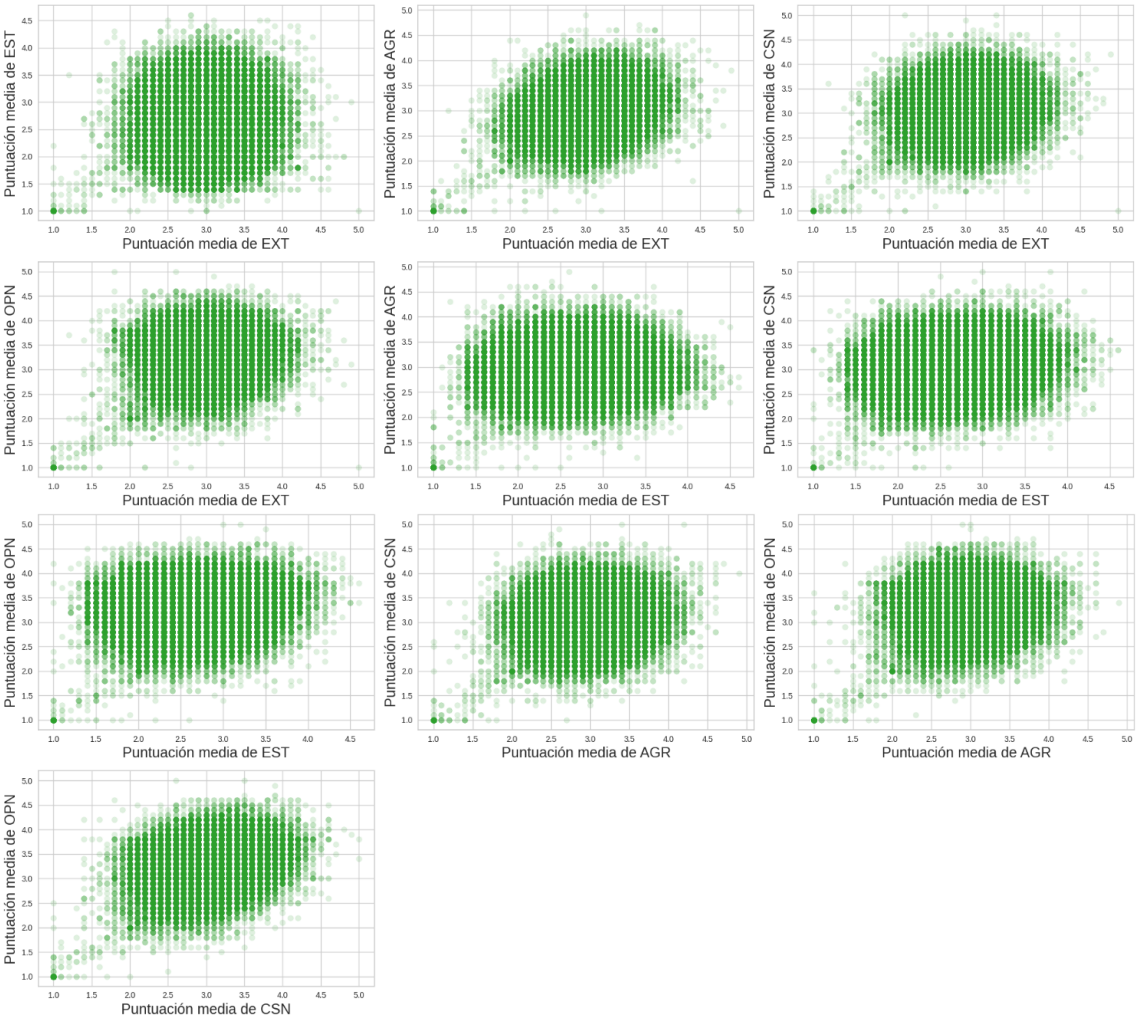

Figura 41: Scatter Plots por pares de dimensiones del Cluster 3

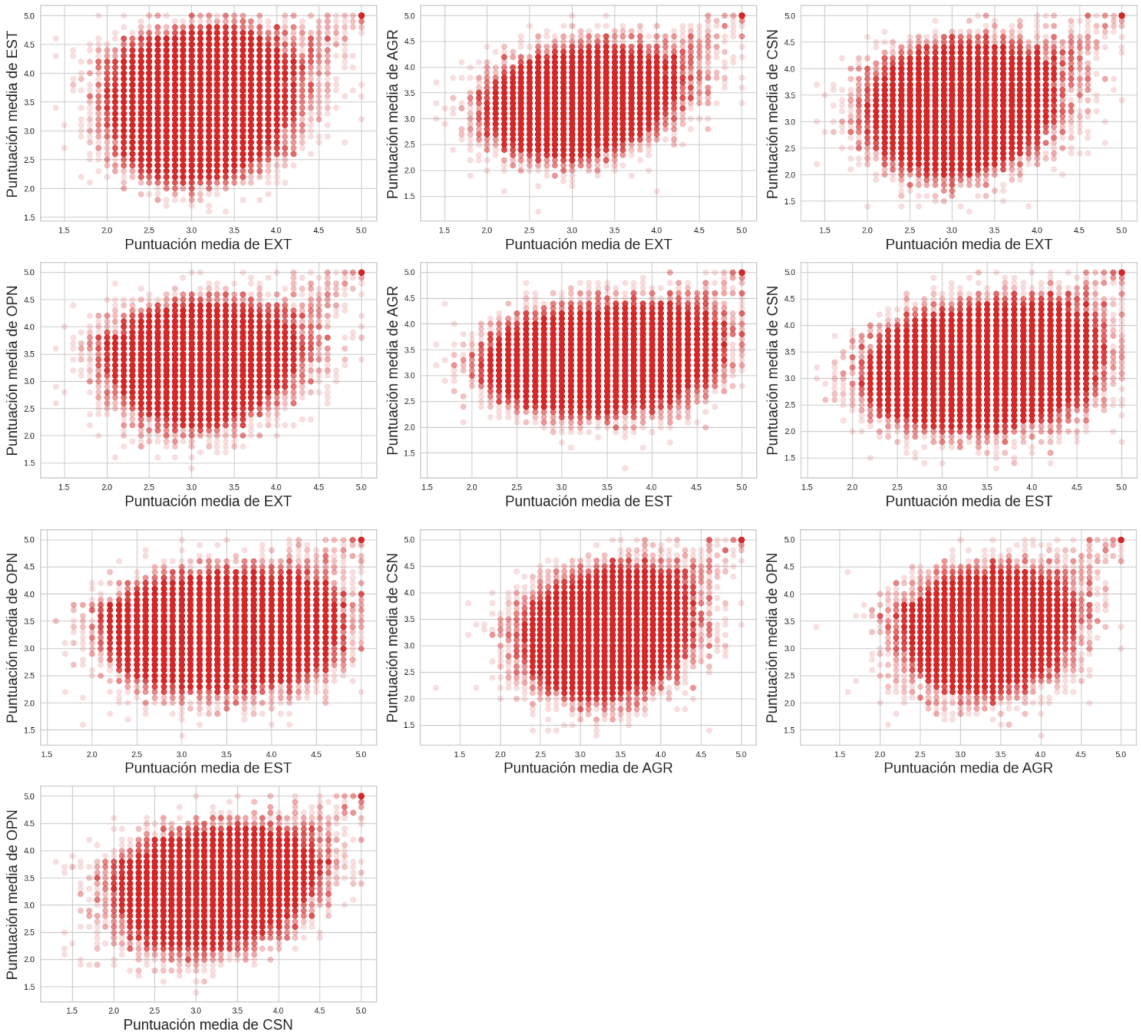

Figura 42: Scatter Plots por pares de dimensiones del Cluster 4

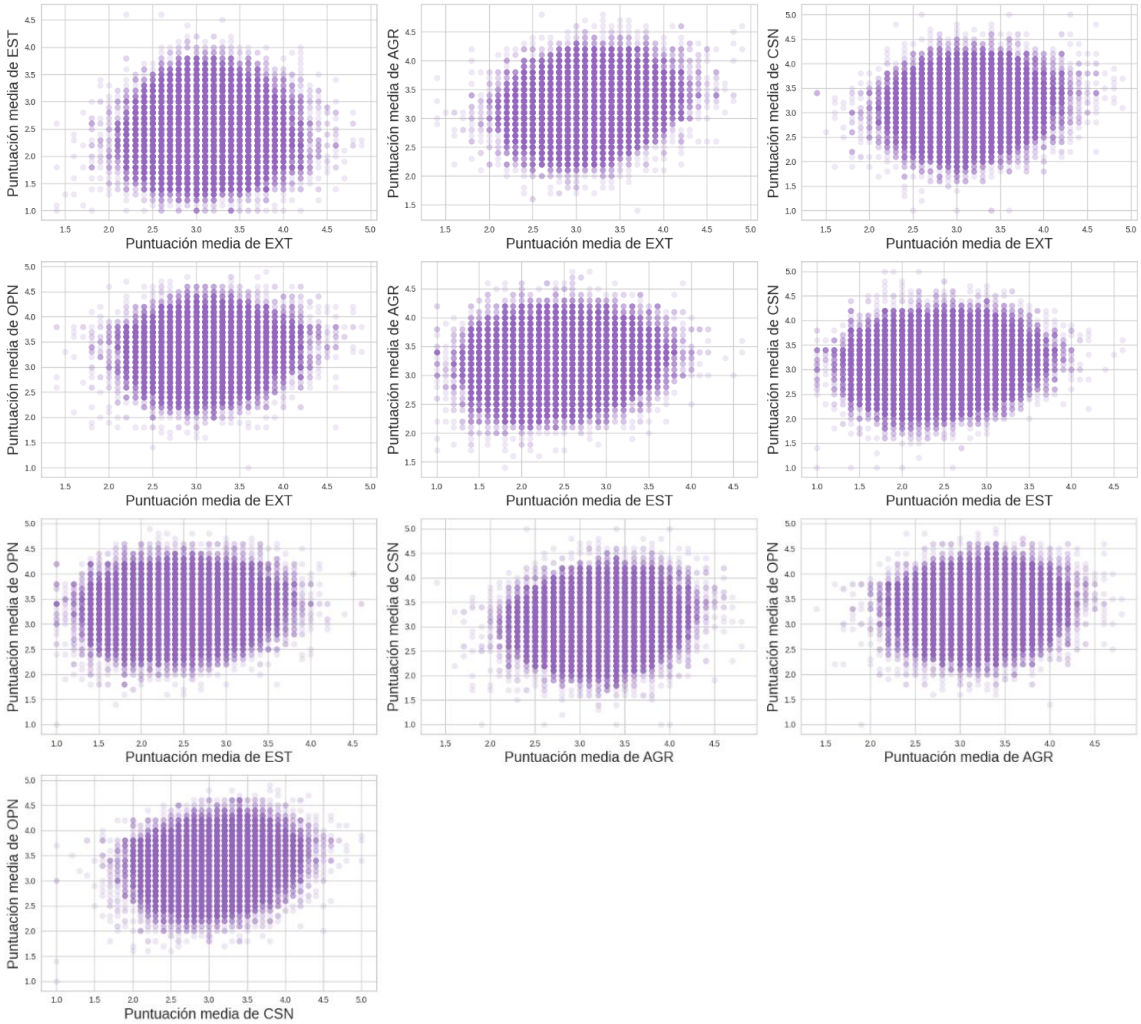

Figura 43: Scatter Plots por pares de dimensiones del Cluster 5

Analizamos los resultados atendiendo a cada *cluster* de manera individual:

***Cluster* 1 (Fig.39):**

- Hay 2 participantes con el valor máximo en la dimensión de *Extraversión* y sólo uno con el valor mínimo. Se observa que la gran mayoría de datos tienen una dispersión equilibrada en torno al valor central 3. Sin embargo, hay una mayor tendencia a valores altos.

- Es destacable en este *cluster* la correlación general para altos valores de *Extraversión*, es decir, aquellos participantes de este *cluster* que tienen puntuaciones altas en *Extraversión*, también las tienen en el resto de dimensiones.

- La gran mayoría de las medias de *Neuroticismo* son mayores que 2, excepto un participante. Además, hay un subconjunto destacable con valor máximo 5 en esta dimensión. De manera general, este cluster tiene valores altos de *Neuroticismo*.

- Aunque para *Amabilidad* hay valores distribuidos por toda la gráfica (con gran mayoría equilibrada en torno al 3), para *Conciencia* y *Apertura*, se observan subconjuntos destacables de participantes con altos valores en estas dos dimensiones.

***Cluster* 2 (Fig.40):**

- Este *cluster* muestra claramente una neutralidad en todas las dimensiones. Esto se observa por su forma globular equilibrada en el centro de las gráficas para todos los pares de dimensiones.

- Es destacable que pese a algunas excepciones (1, 2 o 3 participantes), ningún valor es máximo o mínimo en ninguna dimensión.

***Cluster* 3 (Fig.41):**

- En este *cluster* se observa una clara nube de puntos con colores más suaves en torno a la gran forma densa central. Esta nube de puntos exterior va distorsionándose en forma de flecha hacia la zona inferior izquierda de la gráfica para todos los pares, convergiendo en un punto muy denso (color muy fuerte) para el valor mínimo (1,1). Esto indica que este *cluster* posee un gran subconjunto de participantes con valores mínimos en las dimensiones.

- Además, pese a alguna excepción, no hay valores máximos en las gráficas.

***Cluster* 4 (Fig.42):**

- Se podría decir que este *cluster* muestra la distribución opuesta del anterior. Se observa claramente una convergencia hacia el valor máximo (5,5) en todas las dimensiones.

- Además, esta vez la distribución en forma de flecha es mucho más densa, es decir hay más valores fuera del glóbulo central y no hay ningún valor mínimo para ninguna dimensión.

- Para varias dimensiones se observan subconjuntos destacables con valor máximo 5, destacando con claridad *Neuroticismo*.

***Cluster* 5 (Fig.43):**

- Este último *cluster*, pese a mostrar neutralidad como el segundo *cluster* por su forma globular equilibrada y su posición céntrica en la mayoría de las gráficas, se destaca que muestra valores más altos de *Amabilidad* y *Conciencia* que el segundo *cluster* y valores menores de *Neuroticismo*.

- De hecho, este *cluster*, es el que de manera general muestra valores menores de *Neuroticismo*.

4.2 Visualización de clusters agrupados

Puesto que durante la visualización del modelo se comprobó que la agrupación que se hizo en base a la descripción de las respuestas era correcta, se trabajará también con esa agrupación. Esto se debe a que al trabajar directamente con las dimensiones asociadas a cada grupo de 10 columnas estamos trabajando inevitablemente con preguntas evaluadas de manera inversa por ser negativas.

Existe una técnica utilizada en el análisis de cuestionarios de personalidad donde ciertas preguntas están formuladas de manera negativa llamada *reverse scoring*. Esto significa que una alta puntuación en estas preguntas indica una característica opuesta a lo que se mide en preguntas formuladas positivamente. Por ejemplo, si una afirmación dice "Normalmente inicio conversaciones" y otra dice "No suelo hablar con desconocidos", las respuestas deberían interpretarse inversamente. Para garantizar que todas las preguntas en una dimensión evalúen consistentemente la misma característica, las respuestas negativas se recodifican, transformando las puntuaciones bajas en altas y viceversa.

Pese a que no se aplicará esta técnica por no haber certeza de que es aplicable al caso, se considera una buena práctica estudiar cada dimensión por agrupación de manera subjetiva, dividiendo las preguntas en subconjuntos positivos y negativos, esto permite una interpretación más precisa y coherente de los resultados. Al hacerlo, se eliminan posibles confusiones que puedan surgir de tener respuestas evaluadas en direcciones opuestas dentro de la misma dimensión y se amplía el campo de visión desde el que analizar los resultados y mejora la interpretación de los mismos.

Vamos a agrupar como hicimos anteriormente en el preprocesado, esta vez, sin tener en cuenta los tiempos de las preguntas pues ya no contamos con ellos. Visualizamos la información referente a estas agrupaciones una vez tenemos los participantes con *clusters* asociados.

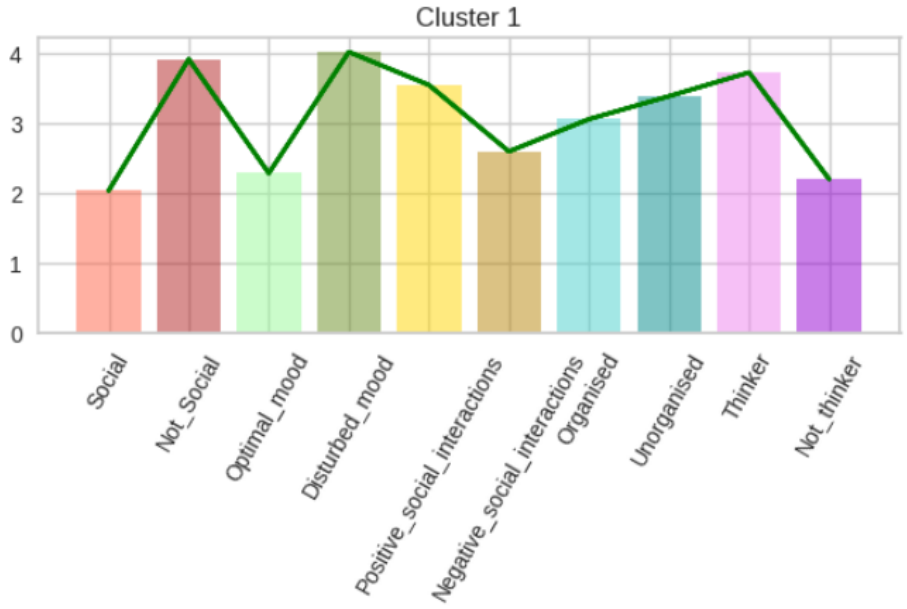

Figura 44: Cluster 1 agrupado según descripciones

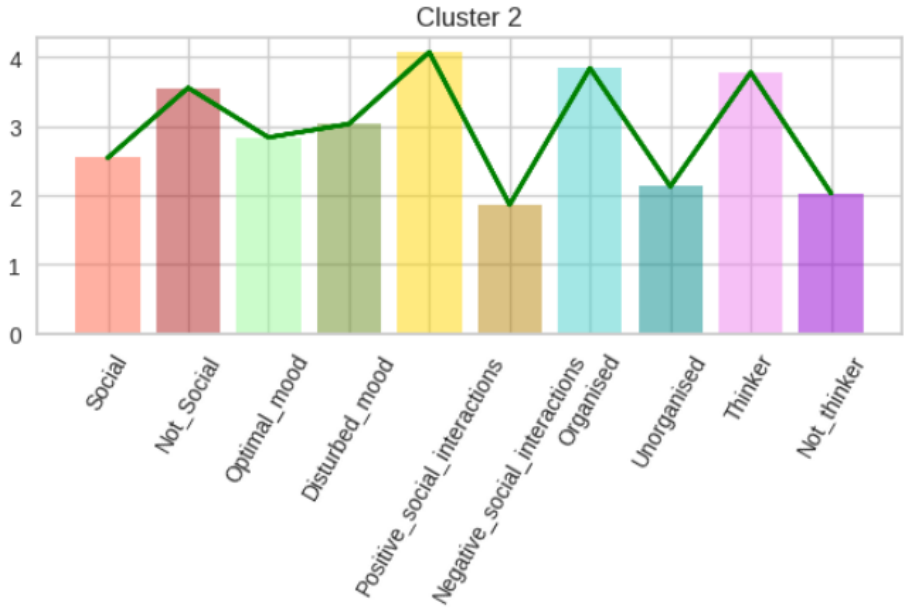

Figura 45: Cluster 2 agrupado según descripciones

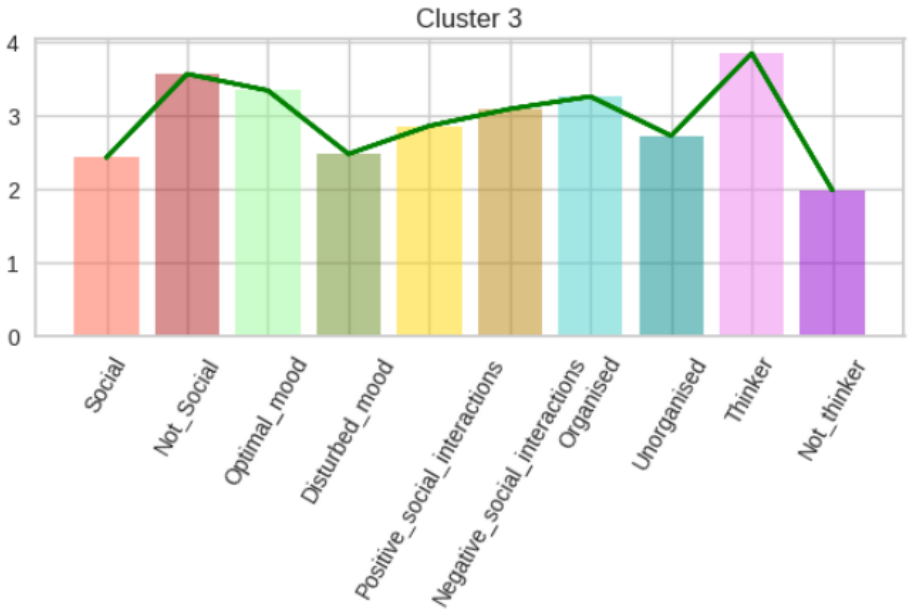

Figura 46: Cluster 3 agrupado según descripciones

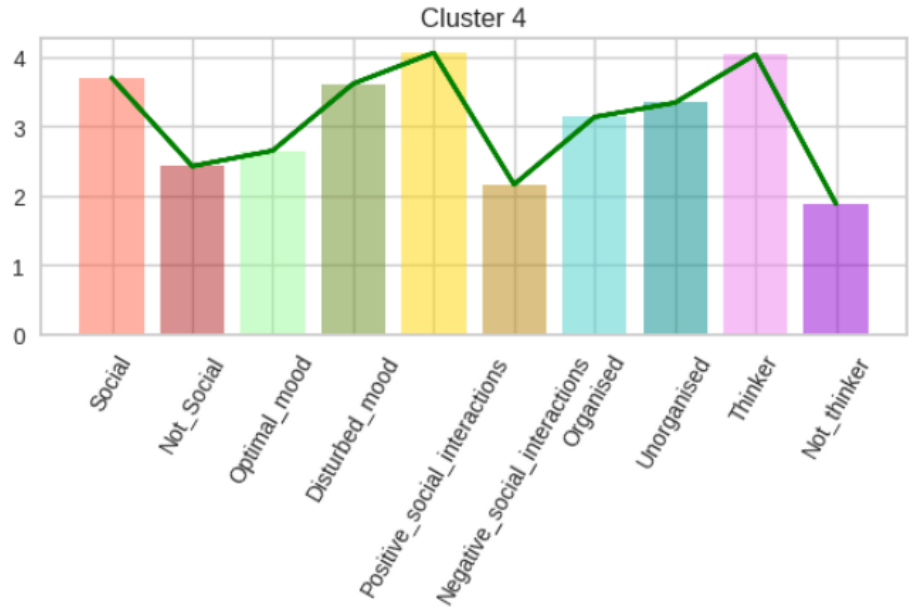

Figura 47: Cluster 4 agrupado según descripciones

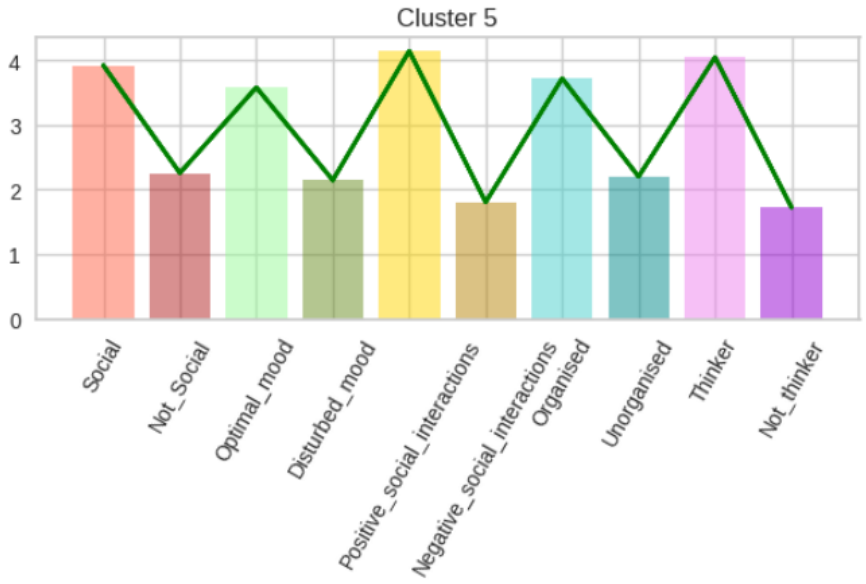

Figura 48: Cluster 5 agrupado según descripciones

Atendiendo a las gráficas de los *clusters* se pueden comprobar las siguientes observaciones:

- Parece haber una relación adecuada entre cada par de agrupaciones, como vimos durante la visualización posterior a la limpieza de datos, lo que da seguridad ante las interpretaciones de las gráficas.

- En el primer *cluster* (Fig.44) se observa una mayor puntuación en el ámbito "*Not_Social*", es decir, preguntas negativas de *Extraversión* y en el de "*Disturbed_mood*" de *Neuroticismo*.

- En el segundo *cluster* (Fig.45) se observa claramente una diferencia destacable en los pares referentes a la interacción social, organización y pensamientos. Eso quiere decir que en esta agrupación hay bastantes participantes con altos valores en estos 3 ámbitos.

- El tercer *cluster* (Fig.46), que tiene menos participantes que el resto, muestra una distribución bastante lineal, en el par de *Apertura a la experiencia* se puede observar más diferencia.

- En el cuarto *cluster* (Fig.47) se puede destacar que junto con el primero, tiene alto valor de "*Disturbed_mood*", que habla del *Neuroticisimo*. A diferencia del primero aquí también hay valores altos en términos de *Extraversión*.

- El último *cluster* (Fig.48) es el más diferencias muestra en todos los pares, de hecho, para todos los pares menos el referente a la dimensión de *Conciencia*, es el *cluster* que muestra valores más altos de manera general.

5 Conclusiones y Líneas Futuras

Tras analizar los resultados, se puede afirmar que se ha logrado identificar patrones significativos en las características de personalidad de los participantes mediante el uso de técnicas de aprendizaje automático. La creación de un modelo de *clustering* basado en las respuestas del cuestionario *Big Five* ha permitido segmentar a casi 700,000 individuos en cinco *clusters* distintos.

Analizando las observaciones de los *clusters*, el *Cluster* 1 destaca por su tendencia a altos valores en *Extroversión*, *Conciencia* y *Apertura a la Experiencia*, con una notable correlación positiva entre estas dimensiones. La presencia de altos valores de *Neuroticismo* sugiere una predisposición a la inestabilidad emocional, aunque también indica una capacidad para altos niveles de interacción social y adaptabilidad. Este *cluster*, por tanto, muestra individuos que pueden ser muy efectivos en roles que requieren comunicación y creatividad, pero podrían beneficiarse de apoyo adicional para manejar el estrés.

El *Cluster* 2 se caracteriza por una neutralidad destacada en todas las dimensiones, reflejando personalidades equilibradas sin extremos significativos. Esto sugiere una estabilidad y consistencia en el comportamiento de estos individuos, ideales para roles que requieren fiabilidad y previsibilidad. Esta característica de equilibrio y moderación puede ser muy útil en contextos laborales que necesitan una base sólida y estable de desempeño, que además no requieran de altos niveles de adaptación.

El *Cluster* 3 revela una clara tendencia hacia valores bajos en todas las dimensiones, indicando individuos con menor proactividad y adaptabilidad. Este perfil es coherente con personalidades que prefieren entornos de baja estimulación y menor interacción social, alineándose con descripciones de personalidades introvertidas y menos abiertas a nuevas experiencias. Puede desempeñar bien tareas que requieren concentración y autonomía

En contraste al anterior, el *Cluster* 4 muestra la convergencia hacia valores máximos en todas las dimensiones, sugiriendo una combinación de alta extroversión, apertura, y consciencia, junto con una alta reactividad emocional. Estos individuos podrían ser altamente efectivos en roles dinámicos, estimulantes y de liderazgo, aunque la tendencia a altos valores de *Neuroticismo* puede requerir estrategias de gestión emocional para mantener el rendimiento.

Finalmente, el *Cluster* 5, aunque similar en neutralidad al *Cluster* 2, se distingue por mayores puntuaciones en *Amabilidad* y *Conciencia* y menores en *Neuroticismo*. Esto sugiere una tendencia hacia la cooperación y la autoorganización con menor susceptibilidad a la ansiedad, ideal para trabajos en equipo y tareas que requieren precisión y estabilidad emocional.

En conclusión, este proyecto ha demostrado que la formación de equipos profesionales puede beneficiarse enormemente de la consideración de las características de personalidad individuales. Al utilizar un modelo de clustering basado en el cuestionario *Big Five*, se ha podido segmentar a los participantes en clusters que reflejan distintas combinaciones de características de personalidad. Los *clusters* identificados muestran que no existen dimensiones aisladas en los individuos, sino que todas las dimensiones del modelo *Big Five* están presentes en mayor o menor medida. Los resultados obtenidos pueden ser de gran utilidad para las dinámicas interpersonales, que son cruciales para el desempeño del equipo, tal y como se planteó en la introducción.

Estas observaciones son una herramienta poderosa para la selección y gestión de equipos, proporcionando a las organizaciones una ventaja estratégica en la optimización de su capital humano. Al considerar que todas las dimensiones de personalidad están presentes en cada individuo, se pone de manifiesto que no se trata de encontrar individuos con características extremas en una dimensión específica, sino de equilibrar las diversas dimensiones dentro del equipo. Este equilibrio es fundamental para crear equipos cohesionados y eficientes. Al proporcionar una herramienta basada en datos para evaluar y formar equipos, las organizaciones pueden tomar decisiones más informadas que consideren tanto las fortalezas individuales como el balance general de características de personalidad en el equipo, asegurando así un rendimiento óptimo y sostenible.

A continuación, se exponen distintas líneas futuras, ordenadas en orden de prioridad, que podrían dar espacio a una mayor comprensión de los resultados y optimización de equipos profesionales.

1. **Trabajo con expertos**

   La prioridad más alta en este esfuerzo de mejorar los resultados del proyecto debe ser la colaboración con expertos en psicología, especialmente en el cuestionario *Big Five*, y en liderazgo. Trabajar junto a un psicólogo especializado permitirá un análisis profundo sobre la implementación del *reverse scoring*, que es la técnica de invertir las puntuaciones de ciertas preguntas para evitar respuestas sesgadas y asegurar que los puntajes reflejan con precisión la variabilidad de las dimensiones de personalidad. Así se podría evaluar si la aplicación de *reverse scoring* en este caso específico mejora la validez del modelo y ofrece recomendaciones sobre cómo ajustar la metodología para obtener resultados más precisos.

   Simultáneamente, la colaboración con un especialista en liderazgo será crucial para asociar los diferentes *clusters* identificados a roles específicos dentro de diversos equipos. Este perfil puede aportar una perspectiva práctica sobre cómo las distintas combinaciones de rasgos de personalidad pueden influir en el desempeño de roles específicos, ayudando a definir estrategias para formar equipos equilibrados y eficientes. Su conocimiento permitirá crear una guía detallada para la asignación de individuos a roles concretos, optimizando así la sinergia entre las habilidades y personalidades de los miembros del equipo.

2. Imputación Multivariante para valores fuera de rango

En segundo lugar, una línea futura de gran relevancia es la implementación del `IterativeImputer` para tratar valores fuera de rango en el dataset. El uso de esta técnica es particularmente interesante debido a la naturaleza correlacionada de las columnas en el cuestionario Big Five. A diferencia del `SimpleImputer` que se ha usado, puede estimar valores erróneos basándose en la información de las demás columnas del conjunto de datos, proporcionando una imputación más precisa y coherente.

Aunque actualmente `IterativeImputer` sigue siendo experimental y algunos de sus parámetros por defecto o comportamientos podrían cambiar en futuras versiones, su potencial para mejorar la calidad del dataset justifica su consideración en trabajos futuros. Implementar esta técnica podría dar lugar a una gestión más sofisticada de los valores fuera de rango, reduciendo el sesgo y mejorando la fiabilidad de los resultados obtenidos del modelo de *clustering*.

3. *Clustering* jerárquico

En tercer y último lugar, se plantea la exploración del *clustering* jerárquico como una línea futura prometedora para este proyecto. Aunque en el presente trabajo se optó por el algoritmo *K-means* debido a problemas de recursos y memoria, el *clustering* jerárquico ofrece ventajas significativas que pueden enriquecer el análisis de los datos en estudios futuros.

Este enfoque no requiere especificar el número de *clusters* de antemano, lo que lo convierte en una técnica flexible y adecuada para explorar la estructura subyacente de los datos. Además, el *clustering* jerárquico permite identificar subgrupos y relaciones jerárquicas dentro de los *clusters*, lo cual podría ofrecer *insights* más profundos sobre la formación de equipos y las dinámicas interpersonales. A pesar de los desafíos de recursos que se han presentado, las mejoras continuas en capacidad computacional y técnicas de optimización podrían hacer viable la aplicación de *clustering* jerárquico en conjuntos de datos grandes como el utilizado en este proyecto.

6 Análisis de Impacto

El análisis de impacto de este proyecto se centra en evaluar los resultados obtenidos durante la realización del Trabajo de Fin de Grado (TFG) en diversos contextos: personal, empresarial, social, económico, medioambiental y cultural. Este análisis destaca tanto los beneficios esperados como los posibles efectos perjudiciales, además de relacionar los hallazgos con los Objetivos de Desarrollo Sostenible (ODS) de la Agenda 2030.

En el contexto personal, los resultados del proyecto pueden ofrecer a los individuos una comprensión más profunda de sus propias personalidades y cómo estas influyen en sus interacciones y en su desempeño en equipos. Esto puede llevar a un mayor autoconocimiento y desarrollo personal, fomentando habilidades de comunicación y colaboración. Sin embargo, también existe el riesgo de que las personas se encasillen en ciertos roles o perfiles, lo que podría limitar su crecimiento personal y profesional.

Desde una perspectiva empresarial, la aplicación de este modelo de análisis de personalidad puede transformar la forma en que se seleccionan y gestionan los equipos de trabajo. Al identificar las combinaciones óptimas de rasgos de personalidad, las empresas pueden formar equipos más equilibrados y eficientes, mejorando así la productividad y el ambiente laboral. No obstante, es importante manejar con cuidado la implementación para evitar la discriminación o el uso indebido de la información personal.

En el ámbito social, este proyecto puede contribuir a una mayor inclusión y diversidad en los equipos, al promover una cultura de respeto y apreciación por las diferencias individuales. El conocimiento de las dinámicas de personalidad puede ayudar a reducir conflictos y mejorar la cohesión social. Sin embargo, una mala interpretación de los datos podría reforzar estereotipos y prejuicios, exacerbando las divisiones sociales.

Económicamente, el impacto positivo se refleja en la potencial mejora de la eficiencia y la innovación dentro de las organizaciones, lo que podría traducirse en mayores beneficios y crecimiento económico. Sin embargo, el costo de implementar y mantener un sistema óptimo de análisis de personalidad, de la mano de expertos en inteligencia artificial, así como de psicólogos y perfiles de liderazgo, puede ser significativo, especialmente para las pequeñas y medianas empresas, lo que podría limitar su acceso a estos beneficios.

En cuanto al impacto medioambiental, aunque este proyecto no tiene una relación directa con cuestiones ambientales, una mayor eficiencia y productividad en las empresas puede conducir a un uso más sostenible de los recursos.

Culturalmente, el proyecto puede fomentar una mayor apreciación de la diversidad de personalidades y estilos de trabajo, enriqueciendo las culturas organizacionales y promoviendo una mentalidad más inclusiva y colaborativa. Sin embargo, es esencial asegurar que los métodos utilizados respeten las diferencias culturales y no impongan normas de personalidad específicas que podrían no ser apropiadas en todos los contextos.

Al relacionar los resultados del proyecto con los Objetivos de Desarrollo Sostenible (ODS), se observa que este trabajo puede contribuir significativamente a varios de ellos. Por ejemplo, el ODS 3 (Salud y Bienestar) se ve beneficiado al promover ambientes de trabajo más saludables y equilibrados. El ODS 4 (Educación de Calidad) puede beneficiarse mediante programas de formación que incluyan la comprensión de la personalidad y las dinámicas de equipo. El ODS 5 (Igualdad de Género) se refuerza al garantizar que las evaluaciones de personalidad se utilicen para promover la igualdad de oportunidades en el lugar de trabajo.

En conclusión, los resultados de este proyecto tienen el potencial de generar impactos significativos y positivos en diversos contextos, siempre y cuando se implementen con cuidado y responsabilidad. La comprensión y la aplicación de los análisis de personalidad pueden mejorar la eficiencia, la inclusión y el bienestar en múltiples niveles, contribuyendo así al logro de un desarrollo sostenible y equitativo.

7 Bibliografía

[1] Jan ter Laak, "Las cinco grandes dimensiones de la personalidad", *Revista de Psicología de la PUCP*, Vol. XIV, N° 2, 1996

[2] Inanna Catalá M., "Teoría de Belbin: roles en los equipos de trabajo", *Universidad Politécnica de Valencia*, Valencia, 2022

[3] Borja Vilaseca, "El origen y la historia del Eneagrama", 2020. Disponible en: https://borjavilaseca.com/el-origen-y-la-historia-del-eneagrama/ [Accedido el 3/30/2024]

[4] F. Fernández Christlieb, "¿De dónde demonios salió el Eneagrama?", México, 2017

[5] C. M. I. Zorita, R. R. Palomo, A. R. Vieites, E. F. Florit, Eugenia, "Factores de personalidad (Big Five) y rendimiento académico en asignaturas cuantitativas de ADE", *Departamento de Métodos Cuantitativos Universidad Pontificia Comillas de Madrid*, XIX Jornadas ASEPUMA, 2011

[6] Susan L. Kichuk, Willi H. Wiesner, "The big five personality factors and team performance: implications for selecting successful product design teams", *Journal of Engineering and Technology Management*, Volume XIV, Issues 3–4, Pages 195-221, 1997

[7] E. M. B. Poveda, M. P. López, A. E. Congacha, E. E. Cajamarca, C. Morales, "Google Colaboratory como alternativa para el procesamiento de una red neuronal convolucional", *Revista Espacios*, Vol. XLI, N° 7, 2020

[8] Lewis Goldberg, "Big-Five Factor Markers", 2018. Disponible en: https://ipip.ori.org/newBigFive5broadKey.htm [Accedido el 05/5/2024]

[9] Open-Source Psychometrics Project, "Big Five Personality Test", 2019. Disponible en: https://openpsychometrics.org/tests/IPIP-BFFM/ [Accedido el 05/5/2024]

[10] Kashmir Hill, "How an internet mapping glitch turned a random Kansas farm into a digital hell", 10, abril 2016. Disponible en: https://www.splinter.com/how-an-internet-mapping-glitch-turned-a-random-kansas-f-1793856052 [Accedido el 10/5/2024]

[11] *Scikit-Learn*, "Imputation of missing values". Disponible en: https://scikit-learn.org/stable/modules/impute.html [Accedido el 10/5/2024]

[12] amueller, "IterativeImputer not converging (at all)". Disponible en: https://github.com/scikit-learn/scikit-learn/issues/14338 [Accedido el 10/5/2024]

[13] A. Géron, "Hands-On *Machine Learning* with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.", *O'Reilly Media*, Inc., 3° edition, 2022.

8 Anexos

8.1 Muestra del conjunto de datos original

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	...	dateload	screenw	screenh	introelapse	testelapse	endelapse	IPC	country	lat_appx_lots_of_err	long_appx_lots_of_err
1	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	...	2016-03-03 02:01:01	768.0	1024.0	9.0	234.0	6	1	GB	51.5448	0.1991
2	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2016-03-03 02:01:20	1360.0	768.0	12.0	179.0	11	1	MY	3.1698	101.706
3	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	...	2016-03-03 02:01:56	1366.0	768.0	3.0	186.0	7	1	GB	54.9119	-1.3833
4	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	...	2016-03-03 02:02:02	1920.0	1200.0	186.0	219.0	7	1	GB	51.75	-1.25
5	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	...	2016-03-03 02:02:57	1366.0	768.0	8.0	315.0	17	2	KE	1.0	38.0
6	3.0	3.0	4.0	2.0	4.0	2.0	2.0	3.0	3.0	4.0	...	2016-03-03 02:03:12	1600.0	1000.0	4.0	196.0	3	1	SE	59.3333	18.05
7	4.0	3.0	4.0	3.0	3.0	3.0	5.0	3.0	4.0	3.0	...	2016-03-03 02:05:00	360.0	640.0	36.0	179.0	10	1	US	30.3322	-81.6556
8	3.0	1.0	5.0	2.0	5.0	2.0	5.0	2.0	3.0	2.0	...	2016-03-03 02:05:08	1440.0	900.0	15.0	210.0	17	1	MY	2.9927	101.7909
9	2.0	2.0	3.0	3.0	4.0	2.0	2.0	2.0	4.0	4.0	...	2016-03-03 02:05:27	2560.0	1440.0	2.0	181.0	4	1	GB	53.423	-2.2166
10	1.0	5.0	3.0	5.0	2.0	3.0	2.0	4.0	5.0	4.0	...	2016-03-03 02:08:06	1600.0	900.0	6.0	261.0	13	1	FI	60.1708	24.9375
11	3.0	3.0	2.0	3.0	3.0	2.0	4.0	3.0	3.0	5.0	...	2016-03-03 02:08:17	1440.0	900.0	6.0	110.0	7	1	UA	50.4333	30.5167
12	3.0	1.0	5.0	3.0	5.0	1.0	5.0	5.0	5.0	3.0	...	2016-03-03 02:08:52	1280.0	720.0	10.0	172.0	8	1	PH	14.5833	120.9667
13	4.0	1.0	5.0	4.0	5.0	1.0	4.0	1.0	5.0	2.0	...	2016-03-03 02:10:44	320.0	480.0	128.0	459.0	8	1	FR	48.8539	2.604
14	1.0	5.0	1.0	5.0	1.0	5.0	1.0	5.0	1.0	5.0	...	2016-03-03 02:10:50	1920.0	1080.0	2.0	129.0	5	1	GB	53.8	-1.5833
15	1.0	5.0	2.0	5.0	1.0	4.0	1.0	2.0	2.0	5.0	...	2016-03-03 02:11:06	1920.0	1080.0	6.0	120.0	8	1	AU	-37.9333	145.2333
16	2.0	1.0	3.0	4.0	4.0	3.0	5.0	3.0	3.0	5.0	...	2016-03-03 02:11:53	1366.0	768.0	9.0	641.0	19	1	IN	20.0	77.0
17	1.0	4.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	4.0	...	2016-03-03 02:14:11	1920.0	1080.0	6.0	169.0	11	1	CA	47.4596	-69.7547
18	4.0	1.0	5.0	2.0	4.0	2.0	3.0	2.0	4.0	2.0	...	2016-03-03 02:14:27	1536.0	864.0	229.0	388.0	26	6	NL	51.9225	4.4792
19	4.0	2.0	5.0	3.0	4.0	4.0	5.0	2.0	5.0	2.0	...	2016-03-03 02:16:17	1280.0	768.0	15.0	592.0	39	1	ZA	-29.0	24.0
20	5.0	1.0	5.0	2.0	5.0	1.0	5.0	3.0	5.0	4.0	...	2016-03-03 02:17:05	1829.0	1029.0	5.0	130.0	10	1	HK	22.3167	114.2167
21	3.0	3.0	2.0	3.0	4.0	3.0	1.0	5.0	1.0	2.0	...	2016-03-03 02:19:29	1280.0	800.0	14.0	193.0	10	1	GB	51.5	-0.13
22	3.0	2.0	2.0	4.0	4.0	4.0	5.0	3.0	1.0	3.0	...	2016-03-03 02:22:02	1366.0	768.0	8.0	305.0	17	1	US	34.1073	-118.3719
23	1.0	4.0	3.0	4.0	2.0	3.0	2.0	5.0	2.0	5.0	...	2016-03-03 02:22:12	1280.0	1024.0	14.0	161.0	22	1	GB	51.4629	-2.5589
24	3.0	4.0	4.0	3.0	3.0	4.0	2.0	2.0	0.0	5.0	...	2016-03-03 02:22:24	320.0	568.0	31.0	435.0	19	1	BR	-23.5475	-46.6361
25	1.0	5.0	1.0	4.0	1.0	5.0	1.0	5.0	1.0	5.0	...	2016-03-03 02:22:56	320.0	568.0	11.0	211.0	34	1	AU	-27.0	133.0
26	2.0	2.0	3.0	5.0	5.0	3.0	5.0	5.0	4.0	5.0	...	2016-03-03 02:23:08	360.0	640.0	22.0	220.0	16	1	CA	42.3997	-82.1996
27	4.0	1.0	5.0	3.0	5.0	1.0	5.0	5.0	5.0	1.0	...	2016-03-03 02:23:23	1600.0	1200.0	24.0	940.0	8	3	GB	51.5	-0.13
28	1.0	3.0	1.0	5.0	1.0	4.0	1.0	5.0	1.0	5.0	...	2016-03-03 02:26:37	1366.0	768.0	4.0	189.0	5	1	AU	-31.6448	152.7946
29	2.0	2.0	3.0	5.0	3.0	3.0	1.0	4.0	1.0	3.0	...	2016-03-03 02:27:49	1920.0	1080.0	275.0	99.0	4	1	GB	51.5	-0.13
30	2.0	4.0	4.0	2.0	3.0	3.0	3.0	3.0	3.0	3.0	...	2016-03-03 02:28:06	2560.0	1440.0	7.0	101.0	8	1	GB	51.5833	-3.2167
31	2.0	4.0	5.0	4.0	4.0	1.0	5.0	4.0	4.0	3.0	...	2016-03-03 02:30:46	1366.0	768.0	5.0	586.0	14	1	SE	58.3013	14.2878
32	2.0	4.0	2.0	4.0	3.0	4.0	2.0	5.0	2.0	4.0	...	2016-03-03 02:31:31	360.0	640.0	7.0	234.0	6	1	CH	46.0083	8.9495
33	3.0	2.0	4.0	2.0	2.0	3.0	3.0	2.0	4.0	4.0	...	2016-03-03 02:33:42	768.0	1024.0	33.0	305.0	8	2	FR	44.737	2.1233
34	3.0	4.0	3.0	2.0	5.0	4.0	2.0	4.0	4.0	4.0	...	2016-03-03 02:33:52	1920.0	1080.0	2.0	299.0	18	1	FR	48.8833	2.2667
35	2.0	2.0	0.0	2.0	5.0	1.0	2.0	3.0	3.0	4.0	...	2016-03-03 02:35:21	1536.0	864.0	6.0	243.0	10	1	GB	52.75	-3.8833
36	1.0	4.0	4.0	5.0	4.0	3.0	3.0	5.0	1.0	3.0	...	2016-03-03 02:38:28	1440.0	900.0	33.0	170.0	158	1	GB	51.5142	-0.0931
37	5.0	1.0	4.0	1.0	5.0	2.0	5.0	2.0	5.0	3.0	...	2016-03-03 02:39:16	1024.0	768.0	4.0	156.0	14	1	TH	13.75	100.4667
38	3.0	2.0	3.0	4.0	2.0	2.0	3.0	4.0	3.0	4.0	...	2016-03-03 02:39:37	1920.0	1200.0	3.0	113.0	8	1	GB	52.2	0.1167
39	2.0	3.0	3.0	4.0	2.0	4.0	1.0	3.0	3.0	4.0	...	2016-03-03 02:40:14	1920.0	1080.0	15.0	196.0	9	1	GB	51.5	-0.13
40	5.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	2.0	...	2016-03-03 02:42:18	1440.0	900.0	4.0	224.0	12	3	GB	51.5	-0.13
41	2.0	3.0	3.0	4.0	3.0	3.0	1.0	4.0	2.0	3.0	...	2016-03-03 02:42:39	1366.0	768.0	25.0	163.0	8	1	IT	42.8333	12.8333
42	3.0	4.0	3.0	4.0	4.0	2.0	2.0	2.0	4.0	4.0	...	2016-03-03 02:43:01	1440.0	900.0	474.0	147.0	5	5	AU	-27.0	133.0
43	3.0	3.0	5.0	2.0	5.0	1.0	4.0	2.0	5.0	1.0	...	2016-03-03 02:43:12	768.0	1024.0	170.0	547.0	14	2	FR	44.737	2.1233
44	2.0	1.0	4.0	1.0	5.0	2.0	2.0	5.0	1.0	3.0	...	2016-03-03 02:43:54	1024.0	768.0	13.0	224.0	15	1	IN	20.0	77.0
45	2.0	4.0	5.0	3.0	5.0	1.0	3.0	5.0	5.0	5.0	...	2016-03-03 02:46:07	1366.0	768.0	1254.0	1396.0	53	1	ES	41.3994	2.1757
46	3.0	1.0	4.0	3.0	4.0	2.0	3.0	2.0	4.0	5.0	...	2016-03-03 02:46:28	1438.0	808.0	90.0	208.0	13	3	IN	20.0	77.0
47	4.0	2.0	4.0	2.0	3.0	0.0	2.0	2.0	4.0	3.0	...	2016-03-03 02:53:26	360.0	640.0	648.0	295.0	17	1	FR	43.6043	1.4437
48	3.0	3.0	4.0	3.0	4.0	0.0	4.0	3.0	2.0	3.0	...	2016-03-03 02:54:06	1366.0	768.0	8.0	223.0	11	1	IN	20.0	77.0
49	2.0	4.0	4.0	2.0	3.0	2.0	2.0	4.0	1.0	5.0	...	2016-03-03 02:54:14	1366.0	768.0	29.0	246.0	8	1	FR	48.86	2.35
50	3.0	5.0	4.0	3.0	3.0	4.0	2.0	2.0	2.0	4.0	...	2016-03-03 02:55:00	1600.0	900.0	3.0	285.0	31	2	AE	24.0	54.0
51	4.0	1.0	5.0	1.0	5.0	2.0	5.0	1.0	5.0	1.0	...	2016-03-03 02:57:21	2560.0	1440.0	108.0	161.0	15	1	HR	45.1667	15.5
52	1.0	2.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2016-03-03 02:59:55	1366.0	768.0	24.0	179.0	14	1	US	36.2082	-86.879
53	1.0	4.0	1.0	3.0	2.0	3.0	2.0	3.0	4.0	4.0	...	2016-03-03 03:00:23	1280.0	800.0	16.0	157.0	6	1	GR	37.9667	23.7167
54	1.0	4.0	5.0	4.0	4.0	1.0	2.0	1.0	1.0	2.0	...	2016-03-03 03:01:38	1680.0	1050.0	2.0	409.0	10	1	IE	53.3478	-6.2597
55	2.0	4.0	4.0	4.0	4.0	2.0	2.0	4.0	4.0	4.0	...	2016-03-03 03:02:23	1280.0	720.0	22.0	246.0	18	1	IN	28.6667	77.2167
56	5.0	1.0	5.0	2.0	4.0	1.0	4.0	3.0	4.0	2.0	...	2016-03-03 03:03:51	1280.0	1024.0	3.0	141.0	4	1	GB	51.5166	-0.192
57	2.0	4.0	3.0	5.0	2.0	4.0	3.0	3.0	4.0	5.0	...	2016-03-03 03:06:44	1366.0	768.0	2.0	142.0	5	1	BR	-12.25	-38.95
58	3.0	2.0	5.0	1.0	3.0	2.0	5.0	3.0	2.0	4.0	...	2016-03-03 03:07:45	1600.0	1200.0	5.0	293.0	13	10	GB	53.45	-2.7333
59	4.0	5.0	4.0	4.0	4.0	1.0	2.0	2.0	0.0	3.0	...	2016-03-03 03:07:49	1366.0	768.0	18.0	192.0	10	1	IN	20.0	77.0
60	3.0	3.0	4.0	3.0	2.0	2.0	3.0	5.0	4.0	5.0	...	2016-03-03 03:08:02	1440.0	900.0	5.0	220.0	13	10	GB	53.45	-2.7333

8.2 Informe de originalidad

Prisco

INFORME DE ORIGINALIDAD

13%	13%	3%	%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	oa.upm.es Fuente de Internet	1 %
2	qu4nt.github.io Fuente de Internet	1 %
3	kipdf.com Fuente de Internet	1 %
4	riunet.upv.es Fuente de Internet	1 %
5	canalinnova.com Fuente de Internet	1 %
6	www.coursehero.com Fuente de Internet	1 %
7	docplayer.es Fuente de Internet	1 %
8	www.researchgate.net Fuente de Internet	1 %
9	hdl.handle.net Fuente de Internet	<1 %

10	borjavilaseca.com Fuente de Internet	<1 %
11	repositorio.puce.edu.ec Fuente de Internet	<1 %
12	ttisuccessinsights.do Fuente de Internet	<1 %
13	cybertesis.unmsm.edu.pe Fuente de Internet	<1 %
14	repositorio.ufsc.br Fuente de Internet	<1 %
15	repositorio-aberto.up.pt Fuente de Internet	<1 %
16	qdoc.tips Fuente de Internet	<1 %
17	comomedir.club Fuente de Internet	<1 %
18	sciendo.com Fuente de Internet	<1 %
19	repositorio.ug.edu.ec Fuente de Internet	<1 %
20	Eduardo Rafael Jáuregui Romero. "Machine Learning model based on the Light Gradient Boosting Machine to predict the probability of default in customers of the Credit Card	<1 %

	portfolio", Revista de investigación de Sistemas e Informática, 2023 Publicación	
21	api.research-repository.uwa.edu.au Fuente de Internet	<1 %
22	itegam-jetia.org Fuente de Internet	<1 %
23	es.scribd.com Fuente de Internet	<1 %
24	uvadoc.uva.es Fuente de Internet	<1 %
25	aplicaciones.mec.es Fuente de Internet	<1 %
26	eprints.uanl.mx Fuente de Internet	<1 %
27	www.ual.es Fuente de Internet	<1 %
28	core.ac.uk Fuente de Internet	<1 %
29	repositorio.utc.edu.ec Fuente de Internet	<1 %
30	ieee-dataport.org Fuente de Internet	<1 %
31	www.cienciadedatos.net Fuente de Internet	<1 %

32	Lijie Liu, Fan Li. " The vision of ideological culture in contemporary Chinese education: preserving society and state ( ) ", Culture and Education, 2023 Publicación	<1 %
33	www.cs.us.es Fuente de Internet	<1 %
34	dev.to Fuente de Internet	<1 %
35	cdn.cronista.com Fuente de Internet	<1 %
36	revistalajunta.jdccpp.org.pe Fuente de Internet	<1 %
37	slidetodoc.com Fuente de Internet	<1 %
38	zaragozaciudad.net Fuente de Internet	<1 %
39	alertachiapas.com Fuente de Internet	<1 %
40	repositorio.pucesa.edu.ec Fuente de Internet	<1 %
41	worldwidescience.org Fuente de Internet	<1 %
42	www.ebizlatam.com Fuente de Internet	<1 %

43	revistas.um.es Fuente de Internet	<1 %
44	www.eeoc.gov Fuente de Internet	<1 %
45	"Inter-American Yearbook on Human Rights / Anuario Interamericano de Derechos Humanos, Volume 25 (2009)", Brill, 2013 Publicación	<1 %
46	Jaume Gómez Caturla. "Desarrollo y caracterización de polímeros de alto rendimiento medioambiental derivados de residuos agroindustriales y aditivos de origen renovable", Universitat Politecnica de Valencia, 2024 Publicación	<1 %
47	aprenderly.com Fuente de Internet	<1 %
48	fastercapital.com Fuente de Internet	<1 %
49	recreadigital.jalisco.gob.mx Fuente de Internet	<1 %
50	upcommons.upc.edu Fuente de Internet	<1 %
51	www.revistacomunicar.com Fuente de Internet	<1 %

52	www.ruben-dario.net Fuente de Internet	<1 %
53	www.scielo.org.mx Fuente de Internet	<1 %
54	diarioelplanetablog.wordpress.com Fuente de Internet	<1 %
55	ephsheir.uhsp.edu.ua Fuente de Internet	<1 %
56	gobiernoabierto.navarra.es Fuente de Internet	<1 %
57	pingpdf.com Fuente de Internet	<1 %
58	prezi.com Fuente de Internet	<1 %
59	repository.udistrital.edu.co Fuente de Internet	<1 %
60	www.masquemunicipios.com Fuente de Internet	<1 %
61	www.mdpi.com Fuente de Internet	<1 %
62	www.news-medical.net Fuente de Internet	<1 %
63	www.omaze.com Fuente de Internet	<1 %

64	1library.co Fuente de Internet	<1 %
65	Luis Mora, Ricardo Lugo, Carlos Moreno, Jhon Edgar Amaya. "Parameters optimization of PID controllers using metaheuristics with physical implementation", 2016 35th International Conference of the Chilean Computer Science Society (SCCC), 2016 Publicación	<1 %
66	albertovillalobos1.wordpress.com Fuente de Internet	<1 %
67	arxiv.org Fuente de Internet	<1 %
68	dspace.ucuenca.edu.ec Fuente de Internet	<1 %
69	eprints.ucm.es Fuente de Internet	<1 %
70	es.coursera.org Fuente de Internet	<1 %
71	es.slideshare.net Fuente de Internet	<1 %
72	eur-lex.europa.eu Fuente de Internet	<1 %
73	geeks.linuxuanl.org Fuente de Internet	<1 %

74	latin.saxomarkets.com Fuente de Internet	<1 %
75	moam.info Fuente de Internet	<1 %
76	rstudio-pubs-static.s3.amazonaws.com Fuente de Internet	<1 %
77	runebook.dev Fuente de Internet	<1 %
78	sql-server-performance.com Fuente de Internet	<1 %
79	vdocumento.com Fuente de Internet	<1 %
80	www.cacic2016.unsl.edu.ar Fuente de Internet	<1 %
81	www.fae.usach.cl Fuente de Internet	<1 %
82	www.gestionar-facil.com Fuente de Internet	<1 %
83	www.kerwa.ucr.ac.cr Fuente de Internet	<1 %
84	www.mific.gob.ni Fuente de Internet	<1 %
85	www.sinembargo.mx Fuente de Internet	<1 %

86	www.theibfr.com Fuente de Internet	<1 %
87	www.uhu.es Fuente de Internet	<1 %
88	Laura De Dominicis, Giuseppe Arbia, Henri L.F. De Groot. "Concentration of Manufacturing and Service Sector Activities in Italy: Accounting for Spatial Dependence and Firm Size Distribution", Regional Studies, 2013 Publicación	<1 %
89	M.J. Suarez-Cabal. "Coverage Measurement for SQL Queries", IEEE Latin America Transactions, 3/2005 Publicación	<1 %
90	Carmen Florido, Juan-Luis Jiménez, Yaiza Navarro. "Students' continuity norms in the university and exam calendar: do they affect university academic performance? / Normas de permanencia y calendario de exámenes: ¿afectan al rendimiento académico universitario?", Cultura y Educación, 2019 Publicación	<1 %
91	documentop.com Fuente de Internet	<1 %

Este documento esta firmado por

	**Firmante**	CN=tfgm.fi.upm.es, OU=CCFI, O=ETS Ingenieros Informaticos - UPM, C=ES
	Fecha/Hora	Sun Jun 02 09:15:07 CEST 2024
	Emisor del Certificado	EMAILADDRESS=camanager@etsiinf.upm.es, CN=CA ETS Ingenieros Informaticos, O=ETS Ingenieros Informaticos - UPM, C=ES
	Numero de Serie	561
	Metodo	urn:adobe.com:Adobe.PPKLite:adbe.pkcs7.sha1 (Adobe Signature)