Universidad de Valladolid

ESCUELA DE INGENIERÍA INFORMÁTICA (SG)

Grado en Ingeniería Informática de Servicios y Aplicaciones

Estudio de la influencia de la edad en el proceso selectivo de acceso al Cuerpo Nacional de Policía

---

Alumno: Ángel Manuel Leal Herrero

Tutor: Juanjo José Álvarez Sánchez

“Lo bueno de la ciencia
es que es cierta independientemente
de si crees o no en ella”
- Neil deGrasse Tyson

Agradecimientos

Quiero expresar mi más sincero agradecimiento a mi tutor Juan José Álvarez Sánchez por su apoyo, orientación y comprensión durante la elaboración del presente proyecto. Su dedicación y pasión por la enseñanza han sido una fuente de inspiración para mí, y gracias a sus consejos y a su paciencia, que no ha sido poca, he podido superar los desafíos que se presentaron en el camino.

A mi querida familia, y en especial a mi compañera de vida, no tengo palabras suficientes para agradecerles su esfuerzo incondicional y su apoyo constante en todo este largo proceso, siendo mi pilar en cada paso, brindándome la motivación y la confianza que necesitaba para seguir adelante. Su confianza en mí ha sido un impulso fundamental para alcanzar mis metas (L+LATL).

Por último, quiero rendir homenaje a mi querido amigo Julián, aunque ya no esté físicamente con nosotros, su apoyo y aliento siempre estarán presentes. Su entusiasmo y su fe en mis capacidades me acompañaron en cada etapa de este proyecto. Estoy profundamente agradecido por los momentos compartidos y por todo lo que aprendí de él. Su memoria vivirá en cada logro que alcance.

Gracias a todos por ser parte de este viaje, sin ellos no habría sido posible haberlo completado.

Resumen

Debido a la eliminación en el 2011 del límite de edad para acceder al Cuerpo Nacional de Policía, se está produciendo un cambio en el perfil de los aspirantes que se presentan a las pruebas de acceso, notándose gran afluencia de aspirantes que rebasan el derogado límite de edad. El perfil de muchos de éstos nuevos aspirantes es muy diferente, dado que al tener una mayor edad ya tienen estructurada una vida que les conlleva unas obligaciones que los aspirantes más jóvenes no tienen, es decir, muchos están casados y con hijos, a su vez que tienen trabajos más o menos estables, motivos ambos que les obligan a cumplir unos horarios y no poder disponer de su tiempo como ellos quisieran para prepararse para el acceso.

Este nuevo cambio obliga a las academias que se dedican a preparar las pruebas de acceso a reestructurar la forma de impartir la formación, implantando cambios en su proceso formativo, dado que les conlleva desarrollar formas y modos más versátiles de poder llegar a estos aspirantes.

En este proyecto se ha utilizado la técnica de Clustering de Machine Learning, partiendo de los datos donados por una academia de formación, que ha permitido conocer si la edad del opositor influye de alguna manera, tanto positiva como negativamente, en el proceso de acceso al Cuerpo Nacional de Policía.

Abstract

Due to the elimination in 2011 of the age limit for access to the National Police Corps, there is a change in the profile of the applicants who take the entrance exams, with a large influx of applicants who exceed the repealed age limit. The profile of many of these new applicants is very different, since being older they already have a structured life that entails obligations that younger applicants do not have, many are married and have children, while they have more or less stable jobs, both reasons that force them to meet schedules and not be able to have their time as they would like to prepare for access.

This new change forces the academies that are dedicated to prepare the entrance exams to restructure the way of providing training, implementing changes in their training process, since it involves them to develop more versatile ways and means to reach these applicants.

In this project we have used the Machine Learning Clustering technique, based on data donated by a training academy, which has allowed us to know if the age of the candidate influences in any way, both positively and negatively, in the process of access to the National Police Force.

Índice general

Lista de figuras

Lista de tablas

Parte I
Memoria del Proyecto

Capítulo 1

Descripción del proyecto

1.1. Introducción

El presente TFG trata sobre el estudio de los resultados obtenidos por los distintos opositores al Cuerpo Nacional de Policía en función de su edad en España. Desde que la institución quitó el límite de edad propuesto de acceso en el año 2011, ha habido un cambio significativo en el perfil de los aspirantes. Dichos resultados han sido donados por una academia que se encarga de preparar todas las pruebas de acceso a dicha oposición. El proceso consta de 6 pruebas, siendo éstas las siguientes:

- **Pruebas físicas**: las pruebas físicas constan a su vez de tres pruebas, de un circuito de agilidad y velocidad, de un ejercicio para determinar la fuerza del tren superior del opositor diferenciado por sexo, consistente en un ejercicio de sujeción estática a una barra para las mujeres y dominadas para los hombres, y de una prueba de resistencia formada por la realización de 1 kilómetro corriendo. Siendo necesario para su aprobado la obtención entre las tres pruebas de 15 puntos como mínimo, no pudiendo obtener en ninguna cero puntos.

- **Examen de conocimientos:** examen tipo test sobre 45 temas. Siendo necesario para su aprobado la obtención de 5 puntos como mínimo.

- **Examen de ortografía:** examen de ortografía que va variando en función de la convocatoria, constando en éstas últimas en determinar si 100 palabras están bien o mal escritas. Siendo necesario para su aprobado superar la nota de corte establecida por el Tribunal Oposición una vez corregidos los exámenes.

- **Reconocimiento médico:** reconocimiento médico en el que se examina que el

- opositor no tenga ninguna de las causas de exclusión determinadas en las bases de la oposición.

- **Entrevista personal:** entrevista personal realizada por un Tribunal para determinar las cualidades del opositor en función de una prueba Biodata realizada previamente por el opositor, el cual es un cuestionario de información biográfica que pretende recabar datos, conocer su vida y trayectoria.

- **Test de psicotécnicos**: test psicotécnico con el objetivo de evaluar las habilidades y capacidades cognitivas de un candidato, incluida su capacidad de razonamiento, concentración, atención y velocidad de razonamiento de la información.

Para la realización del citado estudio, es necesario tener en cuenta los siguientes factores:

- **Comprender las características demográficas de los aspirantes a la Policía Nacional:** El proyecto busca investigar y comprender mejor la composición demográfica de los opositores que desean ingresar a la Policía Nacional, centrándose específicamente en su edad. Este análisis proporcionará una visión más clara de las tendencias y patrones en la demografía de los candidatos.

- **Identificar factores que influyen en la participación de diferentes grupos de edad:** El estudio pretende determinar si existen diferencias significativas en la participación de los aspirantes a la Policía Nacional según su grupo de edad. Al identificar estos factores, se podrán desarrollar estrategias específicas para promover la participación de grupos de edad que más interesen.

- **Optimizar los recursos y procesos de reclutamiento:** Mediante el análisis detallado de la citada base de datos de opositores, el proyecto tiene como objetivo identificar oportunidades para optimizar los recursos y procesos de captación de opositores. Al comprender las preferencias y características de diferentes grupos de edad, se podrán desarrollar estrategias de reclutamiento más eficaces y eficientes, asegurando que se atraiga a los mejores candidatos en todas las franjas de edad.

- **Fomentar la equidad y la igualdad de oportunidades:** La investigación sobre la base de datos de opositores en función de su edad tiene como objetivo fomentar la equidad y la igualdad de oportunidades en el proceso de selección para ingresar a la Policía Nacional. Al comprender cómo diferentes grupos de edad se ven representados en el proceso de oposición, se pueden tomar medidas para garantizar que todos los candidatos sean tratados de manera justa y tengan las mismas oportunidades de éxito.

- **Contribuir a la mejora de los futuros agentes de policía**: Al realizar un

análisis detallado de la base de datos de opositores a la Policía Nacional, el proyecto busca contribuir a la mejora de los diferentes medios y métodos ofrecidos por la academia para garantizar el éxito de sus alumnos.

Es por estos factores por los que a la academia en cuestión le interesa conocer si la edad de los opositores es una cuestión a tener en cuenta a la hora de determinar el proceso formativo de sus alumnos, dado que esto puede determinar la organización de las clases, los horarios de éstas, el método utilizado para la preparación de las pruebas físicas, la preparación de la entrevista personal, etc.

1.2. Marco histórico de la Policía Nacional

El primer antecesor de la Policía Nacional apareció en 1824, bajo Fernando VII, quien creó la Policía General del Reino por Real Cédula de 13 de enero.

Anteriormente, durante el llamado trienio constitucional (1820-1823), también se había proyectado un cuerpo de policía "destinado a proteger a las personas y los bienes y mantener el orden público", pero debido a las restricciones de la intervención francesas de los "Cien Mil hijos de San Luis y la restauración de la monarquía absoluta.

El 13 de enero de 1824 se considera la fecha de fundación de la Policía Española, que es clave para la modernidad. La Real Cédula prevé la doble clasificación de los agentes de policía. Por un lado, están los profesionales, integrados por el propio Superintendente General, el Secretario, el Tesorero, el Comisario de Cuartel, y y los Celadores de Barrio y de Puertas. Por otro lado, los semiprofesionales integrados por los llamados Alcaldes de Barrio.


824
REAL CEDULA
DE S. M.
A.H.N.
Y SEÑORES DEL CONSEJO,
Por la que se manda guardar y cumplir el Real decreto inserto compresivo de las reglas que han de observarse en el establecimiento de la Superintendencia general de la Policia del Reino, con lo demas que se expresa.
Año de 1824.

Figura 1.1: Cédula del 13 de enero de 1824.

En 1825, Fernando VII ordenó la formación de un regimiento de caballería como fuerza auxiliar al regimiento de caballería anterior. Obtuvo el título de "Royal Warden", otro claro precedente para un grupo de policías uniformados. Esta unidad, por orden del Comisionado, pretende extender su presencia a todas las provincias, pero solo por dos años.

A principios del reinado de Isabel II, en 1833, se crea un nuevo cuerpo uniformado, de breve existencia, la " Salvaguardas Reales", con la tarea de mantener el orden civil en Madrid y sus alrededores.

Continúa el establecimiento y disolución de instituciones, lo que ni siquiera prueba su eficacia. Entonces, en llegamos a 1844. Ese año, por Real Decreto de 26 de enero se crea el Ramo de Protección y Seguridad, restableciéndose los puestos vacantes de Comisario y Celador.

Curiosamente, en la casa del Celador se colocaba sobre su puerta un farol por la noche, iluminando un letrero con un anuncio de la "Celaduría de Protección y Seguridad".

Estos nuevos miembros se denominaron "Agentes" hasta que pasaron a llamarse "Salvaguardias de Madrid" en 1848. Ese mismo año se instauró el Gobierno Superior de Policía. Esta institución puede considerarse la primera Jefatura Superior de Policía, tal y como lo entendemos hoy en día.

El Cuerpo de Orden Público de Madrid se constituyó tras la revolución de "La Gloriosa" de octubre de 1868, que supuso el destronamiento y exilio de Isabel II. El 1 de junio de 1870, el gobierno provisional amplió su jurisdicción para incluir la totalidad de España después de restringirla inicialmente a la capital del Reino.

Si bien es claro que el Cuerpo de Orden Público no puede servir como policía judicial, cien de sus miembros fueron desmilitarizados y asignados a labores de investigación y prevención del delito. El Cuerpo de Orden Público es en esencia una organización militar.

Figura 1.2: Medalla que acreditaba la condición de Agente de la Autoridad.

Durante el breve gobierno de Amadeo I de Saboya se mantuvo la misma organización policial. El Gobierno y la Policía Judicial se reorganizaron nuevamente en 1873 después de la declaración de la República. El objetivo es crear un grupo que no esté involucrado en batallas políticas en curso. Las reformas posteriores estarán influidas por este apoliticismo, tanto durante la República como durante el reinado de Alfonso XIII.

Prueba de ello es el Real Decreto de 6 de noviembre de 1877, por el que se establece la existencia de dos servicios policiales, "Vigilancia" y "Seguridad", prestados por dos cuerpos distintos. Para convertirse en "la más fuerte garantía de la seguridad personal y el más poderoso auxiliar de la justicia", había que "huir siempre de lo que comúnmente se llama política, nunca ser un arma de partido, sirviendo sólo a los verdaderos intereses sociales".

El servicio de Vigilancia será prestado por un Cuerpo de empleados civiles, según este Real Decreto (Arto. 5) y el de seguridad, "por un Cuerpo organizado a imitación del Cuerpo Militar" (Arto. 6).

La Guardia Civil es la encargada de mantener la seguridad en los alrededores de la capital, según la misma cláusula. Hasta las reformas de 1886, el Cuerpo de Orden Público siguió funcionando en el resto del territorio nacional con su estructura anterior.

Un decreto real estableció la primera Dirección General de Seguridad y amplió la organización del Cuerpo de Vigilancia y Seguridad de Madrid al resto de España ese mismo año, 1886, que entró en vigencia después de la muerte del Rey Alfonso XII.

La Policía Gubernativa tenía una identidad nacional. Con algunas excepciones, como la creación de un cuerpo de Policía Judicial para Barcelona y Madrid únicamente, solo por la Orden Real del Ministerio de Gracia y Justicia el 19 de septiembre de 1896. De acuerdo con la "Ley para la Represión del Anarquismo", que se promulgó ese mismo año, su propósito era enjuiciar crímenes cometidos con los explosivos.

Figura 1.3: Medalla acreditativa de la Policía, Real Decreto de 30 de agosto de 1887.

Alfonso XIII, que firmó la "Ley Orgánica de la Policía del Gobierno" el 27 de febrero de 1908, fue el responsable de la reforma fundamental de la institución policial a principios del siglo XX.

La iniciativa y el arduo trabajo de Juan de la Cierva y Peñafiel, entonces Ministro del Interior, llevaron a la aprobación de esta importante Ley. De la Cierva trabaja para garantizar el respeto, la competencia y la estabilidad de los policías en sus puestos, poniendo fin a los casos anteriores en los que los agentes servían como nada más que las herramientas de los poderosos para sus propios fines.

El Gobernador Civil de cada provincia impartirá órdenes a los Cuerpos de Vigilancia y Seguridad, que serán los encargados de vigilar a todos los ciudadanos en virtud de esta Ley. Los Comisionados, Inspectores Jefes, Inspectores de Primera, Segunda y Tercera Clase, Agentes y Guardias integran el de Vigilancia. Adicionalmente, el de Seguridad está compuesta por clases, guardias, jefes y oficiales procedentes del Ejército.

Se restablece la Dirección General de Seguridad por Real Decreto de 27 de noviembre de 1912, asegurando su continuidad. Don Ramón Méndez Alans, Auditor de División del Cuerpo Jurídico Militar y jurista pertinente, fungirá como primer administrador del centro.

Méndez Alans, que fue el primer Jefe de Policía de Madrid en 1909, realiza una labor encomiable. Monta los laboratorios de fotografía y revelado de huellas dactilares, establece y organiza la fuerza policial en "brigadas". También establece los registros de la Dirección General de Seguridad. Con este grupo busca "la especialización de aptitudes", para lo que especifica normas específicas para los servicios de "Barrios", "Rondas" y "Ciclistas".

La Ley de Presupuestos de la Segunda República Española de 1932 acaba por cambiar la denominación de "Cuerpo de Vigilancia" por el de "Investigación y Vigilancia". Las llamadas "Secciones de Asalto" han ampliado su equipo de seguridad. Como eran tan conocidos, el Cuerpo pasó a llamarse de "Seguridad y Asalto".

Los nombres de los departamentos de policía (Investigación y Vigilancia y Seguridad y Asalto) siguieron siendo los mismos durante la guerra civil (1936-1939) en ambos lados. En la zona republicana se unificaron en el Cuerpo de Seguridad, dividiéndose en dos grupos, el “Uniformado” y el “Civil”. También se integran en este cuerpo las recién creadas Milicias de Vigilancia de la Retaguardia y la propia Guardia Civil, que por decreto de 30 de agosto de 1936 había cambiado su denominación por la de Guardia Nacional Republicana.

Tras el conflicto, el Estado español resultante organiza las fuerzas policiales de acuerdo con los principios de su régimen político.

Según la Ley del 8 de marzo de 1941, se crearon dos nuevos grupos policiales como resultado de dejar atrás la vieja organización liberal y democrática. Uno de ellos es un organismo civil denominado Cuerpo General de Policía, cuyas funciones incluyen "información, investigación y vigilancia". La Policía Armada y de Tráfico, otro organismo uniformado de carácter militar, tiene a su cargo "la vigilancia total y permanente, como

represión cuando sea necesario. Este último estaba integrado por antiguos empleados del "Cuerpo de Seguridad y Asalto" y del "Cuerpo de Vigilantes de Caminos", que fueron creados en 1933 para regular el tráfico en las carreteras y, hasta ese momento, dependientes del Ministerio de Obras Públicas. Este cuerpo perdió su autoridad sobre el tráfico interurbano en 1959, y su nombre fue cambiado al de Cuerpo de Policía Armada.

El Cuerpo Superior de Policía y el Cuerpo Nacional de Policía son dos nuevas corporaciones profesionales que la Ley de Policía de 4 de diciembre de 1978, "De la Policía", define como "Seguridad del Estado" tras el establecimiento del régimen de libertades civiles en España.

Los Cuerpos Generales de Policía y Cuerpos de Policía Armada funcionarán con estas nuevas designaciones hasta 1986. La citada ley creó el Cuerpo Nacional de Policía, que es uniformado, con el mandato de "defender el orden constitucional, proteger el libre ejercicio de los derechos y libertades, y garantizar la seguridad ciudadana".

El anuncio de las primeras vacantes para mujeres en el Cuerpo Superior de Policía en 1978 supuso un importante punto de inflexión para el futuro de la Policía española. Un año después se incorporaron las primeras 42 inspectoras de Policía; estas precursoras hicieron historia allanando el camino para que las siguieran muchas más mujeres. Cuando 53 mujeres ingresaron por primera vez a la Policía Nacional en 1984, se produjo el mismo fenómeno.

Figura 1.4: Emblemas Cuerpo de Policía Nacional y placa del Cuerpo Superior.

En 1986 entró en vigor una importante ley que establece la estructura actual del Cuerpo Nacional de Policía, proporcionando el modelo actual de las fuerzas policiales españolas.

Por la Ley Orgánica de "Fuerzas y Cuerpos de Seguridad", de 13 de marzo de 1986, se

establecen dos Cuerpos del Estado: la Guardia Civil y la Policía Nacional, que se formaron mediante la fusión del Cuerpo Superior de Policía y la Policía Nacional.

Además, otorga a estas instituciones el mandato de proteger el libre ejercicio de los derechos y libertades de los ciudadanos, garantizando su seguridad, todo ello bajo la dependencia del Gobierno Nacional. Este servicio permanente a la sociedad, cuyo sistema democrático defiende, se adhiere también a los principios defendidos por la Asamblea General de las Naciones Unidas y el Consejo de Europa.

Figura 1.5: Placa acreditativa del Cuerpo Nacional de Policía

La Policía Nacional ha sufrido una importante transformación desde 1986 hasta la actualidad, que la ha convertido en una de las instituciones más respetadas de la sociedad y un referente de la seguridad pública tanto a nivel nacional como internacional.

La Policía Nacional tiene el compromiso de velar por la seguridad de sus ciudadanos y, además, es capaz de hacer frente con eficacia a las graves amenazas al modelo de convivencia democrática que son de carácter global. Por ello, uno de sus objetivos estratégicos es combatir el crimen organizado, el terrorismo, el abuso de los más vulnerables, el cibercrimen y la trata de seres humanos.

Actualmente, la actividad policial se sustenta en una serie de pilares, que incluyen la formación de los miembros, la especialización de las unidades, la cooperación en todos los órdenes y niveles, la igualdad de oportunidades para todos los policías sin distinciones por razón de sexo y la modernización del servicio público a través de la transformación digital.

1.3. Objetivos y limitaciones

A la hora de la realización de este proyecto se han marcado los principales objetivos para su ejecución, los cuales se han diferenciado en objetivos de menor tamaño con el objetivo de objetivar el cumplimiento de estos:

1.3.1. Objetivos del proyecto

Obj-1 - Obtención de los datos producto de estudio y procesamiento de los mismos.

- Obj-1.1 - Estudio de las principales variables a estudiar.
- Obj-1.2 - Recopilación y procesamiento de datos producto del estudio a realizar.
- Obj-1.3 - Transformación y procesamiento de los datos a utilizar en siguiente estudio.
- Obj-1.4 - Análisis de todos los datos de los opositores.

Obj-2 - Investigación de los modelos y métodos de aprendizaje para la consecución de los objetivos perseguidos en este proyecto, así como elección de los mismos.

- Obj-2.1 - Estudio de métodos y modelos más apropiados para el logro de los resultados más adecuados.
- Obj-2.2 – Comparación entre la diversidad de modelos y métodos encontrados con el objetivo de determinar el más apropiado para el proyecto.
- Obj-2.3 - Modificación y mejora de modelos para la obtención de los resultados más concretos.
- Obj-2.4 - Elaboración de las medidas más convenientes utilizando los modelos creados sobre los datos iniciales.

Obj-3 - Aplicación del método de análisis y modelo a una determinada zona de test.

- Obj-3.1 - Recopilación y producción de los datasets necesitados para la zona de test.
- Obj-3.2 - Aplicación del método de aprendizaje seleccionados a los datos elegidos para la realización del test.

- Obj-3.3 - Aplicación a los datos de test del modelo de aprendizaje seleccionado con el objetivo de obtener de las mejores medidas.

1.3.2. Limitaciones

A continuación, se enumeran un listado de limitaciones encontradas durante el desarrollo del proyecto:

- **Obtención de la información a procesar**: obtener la información sobre los datos de opositores dada la privacidad de los mismos con el consiguiente permiso de los dueños de la academia.

- **Procesamiento de los datos objeto de este proyecto**: antes de realizar el procesamiento de los datos se han tenido que revisar comprobando que todos ellos se encontraban en el formato correcto y sin errores, dado que han sido introducidos por el personal de la academia.

- **Limitaciones en cuanto a la cantidad de información disponible:** las conclusiones obtenidas en este estudio pueden verse afectadas dado que el tamaño de la muestra viene limitado por el número de alumnos que están apuntados en la academia.

1.4. Estructura del proyecto

Capítulo 1.- Introducción: Este primer capítulo proporcionará una introducción general del tema principal del proyecto, así como una exposición de sus principales objetivos y las razones por las que se llevó a cabo.

Capítulo 2.- Metodología: Se describirá el tipo de proceso de desarrollo del proyecto y las herramientas utilizadas en el mismo.

Capítulo 3.- Gestión del proyecto: Dentro de este capítulo se expondrá la planificación seguida para el desarrollo del proyecto junto al presupuesto y balance obtenidos.

Capítulo 4.- Dominio del problema: En este capítulo se estudiará y analizará el entorno en el que se podrá dar uso a este proyecto, destacando todos los factores cruciales que debemos considerar antes de comenzar a construir nuestra herramienta.

Capítulo 5.- Obtención de datos: Se expondrán todos los aspectos de la búsqueda, adquisición y adaptación necesaria de los datos objeto del proyecto. Además, para la correcta comprensión del trabajo realizado, se dará una definición detallada de cada variable utilizada.

Capítulo 6.- Análisis de datos: Capítulo en el cual se realizará un estudio, junto con el análisis pertinente de los datos recopilados para la obtención de unas conclusiones que nos servirán de punto de partida.

Capítulo 7.- Métodos de aprendizaje: Clustering: En este capítulo de da una visión teórica y práctica del método de aprendizaje clustering, así como se aplica el mismo al conjunto de datos recopilados.

Capítulo 8.- Conclusiones y trabajo futuro: Siendo éste el Capítulo final, en el cual se muestran las conclusiones alcanzadas tras la realización del proyecto, así como las principales aportaciones realizadas junto con las posibles líneas de investigación futuras.

Capítulo 2

Metodología de Trabajo

En este capítulo se describe la metodología utilizada para desarrollar el proyecto. Se iniciará explicando el tipo y comportamiento de la metodología elegida y el por qué fue elegida.

2.1. Metodología SCORE

Se ha optado por la metodología SCORE, basada en SCRUM, dado que dicha metodología está enfocada a proyectos de investigación en el ámbito educativo, y el proyecto en cuestión que se está desarrollando es un proyecto de investigación.

SCRUM es una metodología ágil, ampliamente utilizada en la gestión de proyectos de desarrollo de software, siendo ideal para la realización del este proyecto dado que ofrece gran flexibilidad en la toma de decisiones que se realizan en el desarrollo del mismo, además de ofrecer gran versatilidad debido a lo imprevisible de los resultados obtenidos. Se centra en la entrega iterativa e incremental de productos de alta calidad, al tiempo que fomenta la colaboración y la flexibilidad.

2.1.1. Origen

Los profesores de la Universidad de Maryland Michael Hicks y Jeffrey S. Foster para evitar el dificultoso y lento desarrollo de los proyectos académicos de investigación que tenían

a su cargo, decidieron adaptar uno de los marcos existentes más efectivos (SCRUM) para la planificación y seguimiento de dichos proyectos, dado el incremento de trabajos universitarios de investigación que tenía que atender un tutor. Dicho volumen de trabajo hacía imposible el

correcto seguimiento y tutorización de cada uno de los proyectos. Esto dio pie al nacimiento del concepto de SCRUM o SCORE.

Con esta metodología lo que se pretende es proporcionar a los alumnos de su propia autonomía para desarrollar el proyecto, mientras que al mismo tiempo están recibiendo una supervisión adecuada, garantizando así la motivación de las investigaciones en curso, así como la garantía del mantenimiento de la calidad en las mismas.

2.1.2. Adaptación desde SCRUM

Como la metodología SCORE se creó partiendo de una adaptación de SCRUM, para poder comprender esta adaptación se hace necesario la profunda explicación de esta.

SCRUM es una metodología ágil ampliamente utilizada en la gestión de proyectos de desarrollo de software. Se centra en la entrega iterativa e incremental de productos de alta calidad, al tiempo que fomenta la colaboración y la flexibilidad.

A continuación, se muestra el principal flujo de trabajo para el desarrollo de software utilizado por la metodología SCRUM:

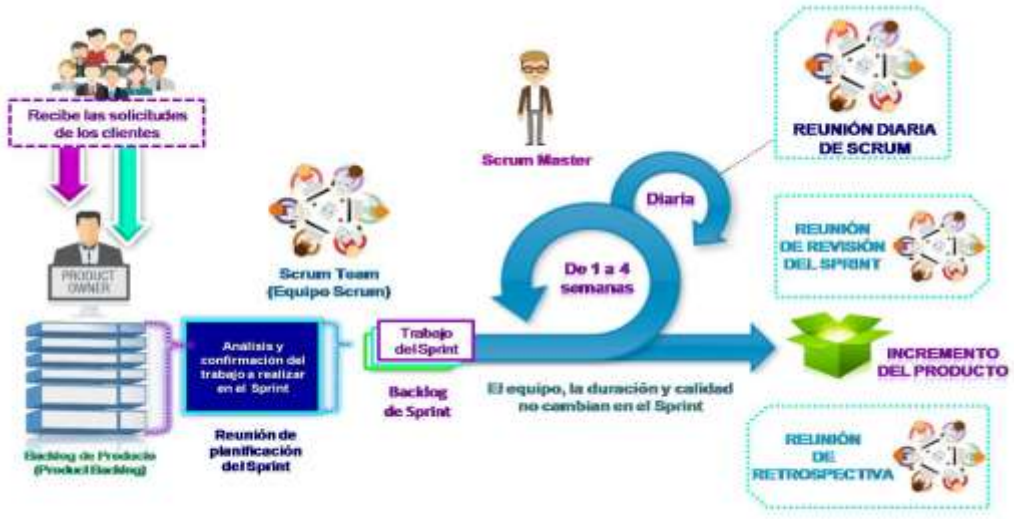


Figura 2.1: Flujo de trabajo SCRUM

Tal y como se puede observar en el esquema de la figura de la parte superior, el trabajo se divide entre varios equipos, siendo uno de ellos el denominado “scrum master”, el cual es

el encargado del seguimiento y el cumplimiento adecuado de los objetivos.

En Scrum, el trabajo se organiza en iteraciones llamadas "sprints" que generalmente tienen una duración de una a cuatro semanas. Cada sprint comienza con una reunión de planificación, llamada “sprint planning”, en la cual se seleccionan los elementos más importantes de una lista priorizada de requisitos o funcionalidades, denominada "product backlog", y se definen las tareas que se abordarán durante el sprint.

Durante el sprint, el equipo de desarrollo trabaja en las tareas definidas en colaboración con el "Scrum Master" y el "Product Owner", representante del cliente o responsable de la visión del producto. Se llevan a cabo reuniones diarias breves, no superiores a 15 minutos, llamadas "daily scrums", para sincronizar el trabajo y abordar cualquier problema o bloqueo, siendo el encargado de plantear soluciones para la resolución de los mismos el “scrum master”.

Al final de cada sprint, se lleva a cabo una revisión de éste y una retrospectiva del equipo. En la revisión del sprint, se muestra el trabajo realizado al "Product Owner" y a otros interesados, denominados “stakeholders”, para recibir retroalimentación. La retrospectiva del equipo es una oportunidad para reflexionar sobre el sprint y encontrar formas de mejorar el proceso de trabajo.

Al final de cada sprint se realizarán dos reuniones denominadas “sprint review” y “sprint retrospective”:

- **“Sprint review”**: Reunión realizada junto con los “**stakeholders”** vinculados en el proyecto, en la que se llevará a cabo la revisión del producto obtenido y su versión, así como la determinación de futuras adaptaciones.

- “**Sprint retrospective**”: Reunión realizada para la valoración del trabajo realizado durante el sprint, indicando a su vez elementos de mejora a tener en cuenta en el siguiente sprint.

En general, Scrum es una metodología ágil flexible que puede adaptarse a diferentes entornos y necesidades, por lo que puede adaptarse eficazmente al desarrollo de proyectos universitarios, ya que comparte muchos principios ágiles que promueven la colaboración, la flexibilidad y la entrega incremental de resultados. La explicación de cómo se puede adaptar Scrum a la metodología de desarrollo de proyectos universitarios comprende los siguientes pasos:

- **1. Formación del equipo**: Al igual que en Scrum, es importante formar un equipo multidisciplinario para el proyecto universitario.

- **2. Definir las tareas que se abordarán durante el sprint o "Product Backlog":** En el contexto universitario, el "Product Backlog" se refiere a la lista de requisitos, tareas o entregables que deben completarse para lograr los objetivos del proyecto. Estos elementos pueden incluir investigaciones, análisis, diseño, desarrollo de prototipos, redacción de informes, entre otros.

- **3. Establecer las iteraciones o "Sprints":** En lugar de utilizar los sprints de duración fija típicos de Scrum, es posible adaptarlos a la duración y estructura del desarrollo académico, pudiendo dividir el proyecto en fases o etapas más pequeñas que se correspondan con los plazos establecidos en el programa académico.

- **Reuniones de planificación de sprint o "sprint planning":** Antes de cada fase o etapa, se lleva a cabo una reunión de planificación donde el equipo y el profesor o tutor del proyecto seleccionan los elementos más importantes del "Product Backlog", definiendo las tareas que se abordarán durante esa fase, estableciendo metas y objetivos claros para cada etapa es esencial.

- **5. Reuniones diarias de seguimiento o "daily scrums":** Se realizan reuniones breves y diarias con el equipo para mantener la comunicación y el seguimiento del progreso. En estas reuniones, cada miembro puede compartir sus avances, identificar posibles obstáculos y coordinar las tareas pendientes. Estas reuniones serán adaptadas según las necesidades en el desarrollo del proyecto, convirtiéndose en "Status meetings" u "On-demand meetings":

    - **"Status meetings"**: Son reuniones similares a las "dailys scrums", siendo la diferencia con éstas que se realizarán en dos o tres días a la semana.

    - "**On-demand meetings**": Reuniones realizadas entre alumnos y profesores para la resolución de bloqueos y problemas técnicos surgidos bajo demanda.

- **6. Reuniones de revisión o "Status review" y reuniones retrospectivas o "Sprint retrospective":** Al final de cada fase o sprint, se lleva a cabo una revisión donde se presentan los resultados y se recibe retroalimentación del profesor o tutor. Además, se organizan retrospectivas del equipo para analizar el proceso de trabajo y encontrar áreas de mejora.

- **7. Adaptación continua:** A medida que avanza el proyecto, es importante adaptar la metodología según las necesidades y circunstancias que surjan, ajustando el "Product Backlog" y las tareas en función de los cambios en los objetivos o requisitos del proyecto.

Hay que tener en cuenta que adaptar SCRUM a un proyecto universitario implica flexibilidad y ajustes en función del contexto específico. La clave es fomentar la colaboración, la comunicación y la entrega incremental de resultados para lograr el éxito del proyecto universitario.

2.1.3. Reuniones de estado o Status meeting

Como hemos visto anteriormente, las reuniones diarias de seguimiento daily scrums de la metodología SCRUM, al ser adaptadas a la metodología SCORE pueden ser sustituidas por las reuniones de estado o status meeting y por las reuniones on-demand meeting. Esto es debido a que la cantidad de estudiantes tutorizados por un mismo tutor no hace posible que se realicen reuniones diarias o daily scrums, tal como se deberían de realizar en un entorno de desarrollo de software profesional.

Por este motivo, dichas reuniones serán acordadas entre el tutor y el alumno o alumnos que desarrollen el proyecto dependiendo de la disponibilidad de ambos, siendo recomendable un 2 o 3 reuniones por semana, para evitar dilaciones en el tiempo de desarrollo del proyecto, concretando los días fijos de las mismas y con una duración máxima de 15 minutos cada una.

En estas reuniones no se deberán abordar detalles técnicos del proyecto en cuestión, ya que para ese tipo de cuestiones esta adaptación de la metodología SCRUM a la metodología académica SCORE ya tiene previstas las reuniones bajo demanda o On-demand meetings.

En estas reuniones, el tutor del proyecto realizará una supervisión del trabajo realizado desde la última status meeting, así como marcará los nuevos objetivos a realizar hasta la siguiente reunión de estado. A su vez, todos los alumnos expondrán, además del trabajo realizado, los desafíos encontrados.

La participación de los tutores en estas reuniones sirve como una forma de motivación para los alumnos, ayudándoles a promover la aparición tanto de nuevas ideas como de curiosidades, siendo esto una experiencia muy enriquecedora para los estudiantes.

De esta interacción de los diferentes alumnos surgen resultados positivos, como la aportación de ideas de personas ajenas al proyecto y la exposición del trabajo realizado al resto de estudiantes, dando pie a la obtención de un feedback, pudiendo aportar éste por un lado la obtención nuevas ideas al desarrollo del proyecto, como el conocimiento de otro tipo de herramientas que podrían ser útiles para el desarrollo del proyecto en cuestión.

2.1.4. Reuniones técnicas bajo demanda o On-demand meetings

Como ya se ha comentado anteriormente, la adaptación de la metodología SCRUM a la metodología SCORE prevé las reuniones on-demand meeting para abordar toda la problemática técnica sobre resultados, modelos técnicas.

Estas reuniones on-demand meeting son una modificación de las reuniones sprint planning realizadas al principio de cada sprint, y en ellas se marcarán líneas de investigación para el desarrollo del proyecto.

Estas reuniones técnicas serán propuestas por los estudiantes dadas sus necesidades

técnicas, y se realizarán cuando sea necesario, no teniendo una programación fija en el tiempo, tal y como se lleva a cabo con las status meeting.

Para evitar que estas reuniones sean poco productivas a la hora de superar obstáculos en el desarrollo del proyecto encontrados por el alumno, estas reuniones solo serán propuestas cuando éste haya realizado previamente una investigación exhaustiva sobre la problemática encontrada.

El hecho de que estas reuniones se celebren solo cuando sea necesario, favorecen un desarrollo fluido y un aprovechamiento del tiempo mucho más eficaz permitiendo que el proceso de desarrollo del proyecto sea muy flexible y adaptable a las necesidades del alumno y del tutor.

2.1.5. Otros elementos de SCORE

La metodología SCORE, además de proporcionar como elementos clave las status meeting y las on-demad meeting, como hemos visto en los apartados anteriores, ofrece un conjunto adicional de componentes para mejorar el proceso de desarrollo de proyectos de investigación en el ámbito educativo:

- **Lugar para el desarrollo del trabajo**: Con el fin de dar calidad al proyecto en desarrollo, el tutor debe el acceso al alumno a herramientas adecuadas y proveerle de un canal de comunicación sostenible. Así mismo, debe facilitar acceso al estudiante a un lugar para el desarrollo del trabajo y la investigación dentro de las instalaciones universitarias que se están utilizando

- **Interacción social**: La metodología SCORE propone un marco de interacción social en el que tanto el tutor y el alumno tienen encuentros para fomentar y favorecer esta interacción entre ambos. Ejemplos de estos encuentros pueden ser la organización de comidas con todo el grupo de desarrollo para celebrar la consecución de objetivos, fomentar la interacción en los descansos con la toma de un refresco, etc.

- **Grupos de aprendizaje**: Para promover el intercambio de ideas y de nuevas líneas de investigación que mejoren los proyectos que se están desarrollando, la metodología SCORE anima al tutor a organizar grupos de aprendizaje entre los alumnos tutorizados.

2.1.6. Adaptación al proyecto

SCORE ofrece una forma de desarrollar el trabajo que se puede personalizar para el proyecto en cuestión que se está realizando. Sin embargo, ha sido necesaria la adaptación de la metodología SCORE dada la naturaleza del proyecto y la distancia con la ubicación

geográfica con la sede de la universidad, ubicación regular del tutor.

Por estos motivos, se optado por reemplazar las interacciones presenciales que propone la citada metodología, vistas en el apartado anterior, por interacciones en formato virtual a través de Microsoft Teams, correo electrónico a través de Gmail y WhatsApp.

Respecto a la organización en concepto de sprint que propone la metodología SCRUM, se ha optado por adaptarla a una organización del proyecto en bloques de trabajo, dada las características del presente proyecto, teniendo éstos una duración del proceso de desarrollo entre 1 y 3 semanas.

Estos bloques de trabajo contendrán a su vez reuniones bajo demanda, con el objetivo de determinar el estudio, la ejecución y la revisión de las tareas que compondrán dicho bloque. Dichas reuniones se determinarán como on-demand meetings, englobando las reuniones sprint meeting, sprint review y sprint retrospective determinadas en la metodología SCRUM.

El reemplazo de interacciones en las status meetings y en las on-demand meeting quedan determinados de la siguiente manera:

- **Status meetings:** Se reemplazan por mensajes intercambiados con el tutor responsable del proyecto, pasando a realizarse el envío de éstos sin ninguna asignación temporal fija, es decir, enviándose con una variabilidad semanal. A través de estos mensajes, se tratan dudas entre el tutor y el alumno con poca importancia técnica y capaces de ser resueltas en un espacio breve de tiempo. Por estos motivos, estas reuniones pasan a determinarse como status messages.

- **On-demand meetings**: Dada la gran flexibilidad de estas reuniones propuestas dentro de la metodología SCORE, se ha decidido mantenerlas, conservando su objetivo principal donde se abordan la resolución de problemática más técnica y compleja de resolver, así como la determinación de las próximas tareas a realizar. Por la misma problemática de la distancia entre las ubicaciones, la interacción determinada en la metodología SCORE se ve reemplazada por interacciones virtuales.

Podemos ver cómo esta adaptación de la metodología SCORE nos ofrece una forma de desarrollo flexible, adaptable y con una excelente comunicación dentro del entorno y la situación en la que se tuvo que realizar el proyecto.

Además, SCORE nos ha ayudado a completar con éxito la división por bloques del proyecto, lo que ha mejorado nuestra capacidad de gestión y nos ha dado una comprensión más clara de su alcance, lo que nos ha permitido cumplir con todas las metas establecidas.

Las diversas formas de comunicarse con el tutor a través de las herramientas y los medios vistos anteriormente son la fuente de todas las ventajas de SCORE.

2.2. Herramientas utilizadas

A continuación, se enumeran y detallan las herramientas utilizadas para el desarrollo y elaboración del presente proyecto.

- **WhatsApp:** Es una aplicación de mensajería instantánea para teléfonos inteligentes y otros dispositivos móviles. Fue lanzada en 2009 por dos ex empleados de Yahoo!, en 2014 fue adquirida por Facebook Inc. WhatsApp permite a los usuarios enviar mensajes de texto, realizar llamadas de voz y videollamadas, así como compartir imágenes, videos, documentos y ubicaciones de manera gratuita a través de la conexión a Internet.

- **Microsoft Teams:** Microsoft Teams es una plataforma de comunicación y colaboración desarrollada por Microsoft. Fue lanzada como parte de la suite de aplicaciones de productividad de Microsoft 365. La plataforma está diseñada para mejorar la colaboración y la comunicación en equipos y organizaciones, especialmente en entornos empresariales y educativos.

- **Microsoft Excel:** Potente herramienta de software creada por Microsoft, que es utiliza con frecuencia para realizar operaciones relacionadas con el procesamiento de datos, el análisis y la presentación de información numérica.

- **Google Colab:** Herramienta desarrollada por Google que proporciona un entorno de ejecución en la nube (cloud) permitiendo a los usuarios ejecutar y experimentar con código Python sin necesidad de configurar un entorno de desarrollo en su propio equipo. Esta herramienta utiliza Notebooks de Jupyter, permitiendo la utilización gratuita de recursos GPUs o TPUs de Google, asícomo las librerías Scikit-learn, Pytorch, TensorFlow, Keras y OpenCV.

- **Jupyter Notebooks:** Proyecto de desarrollo de software creado por Python, de código abierto. Siendo los principales lenguajes de programación compatibles Julia, Python y R .Jupyter. Ofrece un entorno de computación web interactivo para crear documentos y textos de tipo JSON versados con una lista ordenada de celdas de entrada y salida que pueden contener código, texto, gráficos matemáticos o texto rico con la extensión .jpynb.

2.3. Tecnologías utilizadas

A continuación, se citan los diferentes lenguajes de programación utilizados en el presente proyecto, así como las bibliotecas correspondientes:

- **Lenguaje de programación Python:** Python es un lenguaje de programación versátil que se utiliza en una amplia gama de disciplinas, desde desarrollo web y científico

hasta inteligencia artificial y automatización, debido a su legibilidad, facilidad de uso y a la gran cantidad de bibliotecas que proporciona, siendo las utilizadas en este proyecto las siguientes:

- **Mat-plotlib:** Biblioteca de visualización en Python que se utiliza para crear una amplia variedad de gráficos y visualizaciones de datos. Su objetivo principal es permitir la representación gráfica de datos de manera clara y efectiva.

- **Seaborn:** Biblioteca de visualización de datos en Python que se construye sobre Matplotlib. Se utiliza para crear visualizaciones estadísticas y gráficos más atractivos y estilizados con facilidad.

- **Pandas:** Biblioteca que proporciona estructuras de datos flexibles, como DataFrames y Series, que pueden almacenar y manipular datos tabulares en forma de tablas con filas y columnas. Esto es esencial para organizar y limpiar datos antes de su análisis.

- **Pickle:** Biblioteca que se utiliza para serializar objetos en Python, lo que facilita el almacenamiento, la comunicación y el intercambio de datos complejos entre diferentes partes de un programa, procesos o sistemas.

- **Numpy:** Biblioteca que proporciona estructuras de datos eficientes para representar y manipular matrices, siendo especialmente útil para tareas que involucran datos multidimensionales dado que dispone de una gran variedad de funciones matemáticas de alto nivel.

Sklearn: Biblioteca esencial para aquellos que desean aplicar técnicas de aprendizaje automático en Python. Con su amplia gama de algoritmos y herramientas, es una opción valiosa para crear, entrenar y evaluar modelos de aprendizaje automático en una variedad de aplicaciones, desde análisis de datos hasta problemas de predicción y clasificación, pudiendo encontrar algoritmos como Kmeans, DBS, Gradient Boosting.

Capítulo 3

Gestión del proyecto

Este capítulo describe la planificación y estimación del proyecto dentro del plazo sugerido para su finalización, que no debe exceder las 300 horas.

El proyecto ha sido desarrollado utilizando el marco SCORE, por lo que ha sido dividido en varios bloques de trabajo con diferentes duraciones, tal y como se ha explicado en el Capítulo 2.

A continuación, se abordará detalladamente la estimación del esfuerzo realizado para cada tarea que compone el proyecto y su planificación. Además, se mostrará el método de estimación del presupuesto económico del proyecto junto con un balance final en el que se incluyen todos los obstáculos encontrados durante el desarrollo, lo que muestra un cambio sobre la estimación realizada en un primer momento.

3.1. Estimación del esfuerzo

Estimar el tiempo y esfuerzo necesarios para completar las tareas que componen nuestro proyecto es uno de los pasos a dar antes de comenzar una planificación temporal. En este proyecto, se aplicó el marco SCORE, que utiliza las llamadas historias de usuario para una estimación precisa del esfuerzo basada en puntos de la historia. Los puntos de la historia pasarán a denominarse Puntos de Tarea (PT), los cuales darán un valor numérico a la cantidad de tiempo requerido para completar cada tarea, con el fin de comprender adecuadamente la adaptación que se realizó entre las tareas y las historias de los usuarios.

Para la asignación de estos PT, ha sido necesario consensuarlos entre el alumno y el tutor encargado de supervisar el proyecto.

A continuación, se muestra una tabla en la que se refleja la relación existente entre las distintas tareas que engloban el proyecto y los puntos de tareas asignados a cada una:

- **TAREA 0**.- Obtención y acondicionamiento de la información

Nombre tarea	**Puntos de tarea (PT)**
TAREA 0.1- BÚSQUEDA Y RECOPILACIÓN DE DATOS	2 PT
TAREA 0.2- ESTUDIO DE VARIABLES MÁS RELEVANTES	1 PT
TAREA 0.3- INSTALACIÓN Y APRENDIZAJE DE HERRAMIENTAS NECESARIAS	1 PT
TAREA 0.4 - OBTENCIÓN DATOS	1 PT
TAREA 0.5 - OBTENCIÓN DATOS MEDIDAS	1 PT
TAREA 0.6 - ELABORACIÓN DE DATASETS	5 PT
TOTAL	**11 PT**

Tabla 3.1: TAREA 0.- Obtención y acondicionamiento de la información

- **TAREA 1**.- Análisis y puesta en marcha de métodos de análisis

Nombre tarea	**Puntos de tarea (PT)**
TAREA 1.1 - BÚSQUEDA Y ELECCIÓN MÉTODO DE ANÁLISIS	3 PT
TAREA 1.2 – ELABORACIÓN DEL MÉTODO CLUSTERING	2 PT
TAREA 1.3 - ESTUDIO DE RESULTADOS OBTENIDOS CON MÉTODO CLUSTERING	3 PT
TOTAL	**8 PT**

Tabla 3.2: TAREA 1.- Análisis y puesta en marcha de métodos de análisis

- **TAREA 2**.- Obtención y procesamiento de datos

Nombre tarea	**Puntos de tarea (PT)**
TAREA 2.1 - BÚESQUEDA DE INFORMACIÓN DE LA ACADEMIA	1 PT
TAREA 2.2 - BÚESQUEDA DE INFORMACIÓN SOBRE EL PROCESO SELECTIVO	4 PT
TAREA 2.3-ESTUDIO Y PROCESAMIENTO DE DATASETS	3 PT
TOTAL	**8 PT**

Tabla 3.3: Tarea 2.- Obtención y procesamiento datos

- **TAREA 3**.- Elaboración de los modelos de aprendizaje

Nombre tarea	Puntos de tarea (PT)
TAREA 3.1 - ELABORACIÓN DE DATASETS	1 PT
TAREA 3.2 - ESTUDIO DE MEJORES MODELOS DE APRE	2 PT
TAREA 3.3 - ELABORACIÓN DE MODELOS	4 PT
TAREA 3.4 - ESTUDIO DE RESULTA DOS	2 PT
TAREA 3.5 - ELABORACIÓN DE MODELO ELEGIDO	5 PT
TOTAL	**14 PT**

Tabla 3.4: Tarea 3-Elaboración de los modelos de aprendizaje

- **TAREA 4**.- Evaluación de los datos

Nombre tarea	Puntos de tarea (PT)
TAREA 4.1 - ELABORACIÓN DE DATASETS ZONA DE EVALUACIÓN	3 PT
TAREA 4.2- OBTENCIÓN DATOS DE ZONA DE EVALUACIÓN	2 PT
TAREA 4.3 - APLICACIÓN MÉTODO CLUS- TERING	2 PT
TAREA 4.4 – APLICACIÓN MODELOS OBTENIDOS EN TAREA 4.1 Y TAREA 4.2	2 PT
TAREA 4.5 - ESTUDIO DE RESULTADOS	2 PT
TOTAL	**11 PT**

Tabla 3.5: Tarea 4.- Evaluación de los datos

- **TAREA 5**.- Elaboración de documentación

Nombre tarea	**Puntos de tarea (PT)**
TAREA 5.1 - DESCRIPCIÓN DEL PROYECTO	2 PT
TAREA 5.2 - METODOLOGÍA DE TRABAJO Y GESTIÓN	2 PT
TAREA 5.3 - DOMINIO DEL PROBLEMA	2 PT
TAREA 5.4 - OBTENCIÓN Y ESTUDIO DE DATOS	4 PT
TAREA 5.5 - MODELOS EMPLEADOS	2 PT
TAREA 5.6 - FASE DE EVALUACIÓN	2 PT
TAREA 5.7 - CONCLUSIÓN, REFERENCIAS Y BIBLIOGRAFÍA	2 PT
TAREA 5.8 - REVISIONES DE MEMORIA	2 PT
TOTAL	**18 PT**

Tabla 3.6: Tarea 5.- Elaboración de documentación

Finalmente, los Puntos de Tarea obtenidos (PT) en la estimación del esfuerzo de este proyecto suman un total de 70 PT.

3.2. Planificación temporal

Como podemos ver, al utilizar la metodología SCRUM ha sido necesario la adaptación del concepto sprint, determinado en esta metodología, y por ese motivo nuestro proyecto se ha dividido en bloques de trabajo, cada uno de los cuales tiene una duración fija de 2 semanas y está determinado a durar de 1 a 4 semanas. La combinación de estos componentes básicos y su desarrollo crea una imagen integral de toda la planificación realizada para completar el proyecto en las 300 horas, asignadas de acuerdo con la carga de créditos ETCS. A excepción de fines de semana, festivos y periodo vacacional de verano, comprendiendo éste el mes íntegro de agosto, dedicándose unas 3 horas diarias al proyecto.

A continuación, se muestra la duración aproximada de cada uno de estos bloques, que incluye únicamente los días de trabajo en el proyecto y las reuniones bajo demanda celebradas en cada uno de ellos:

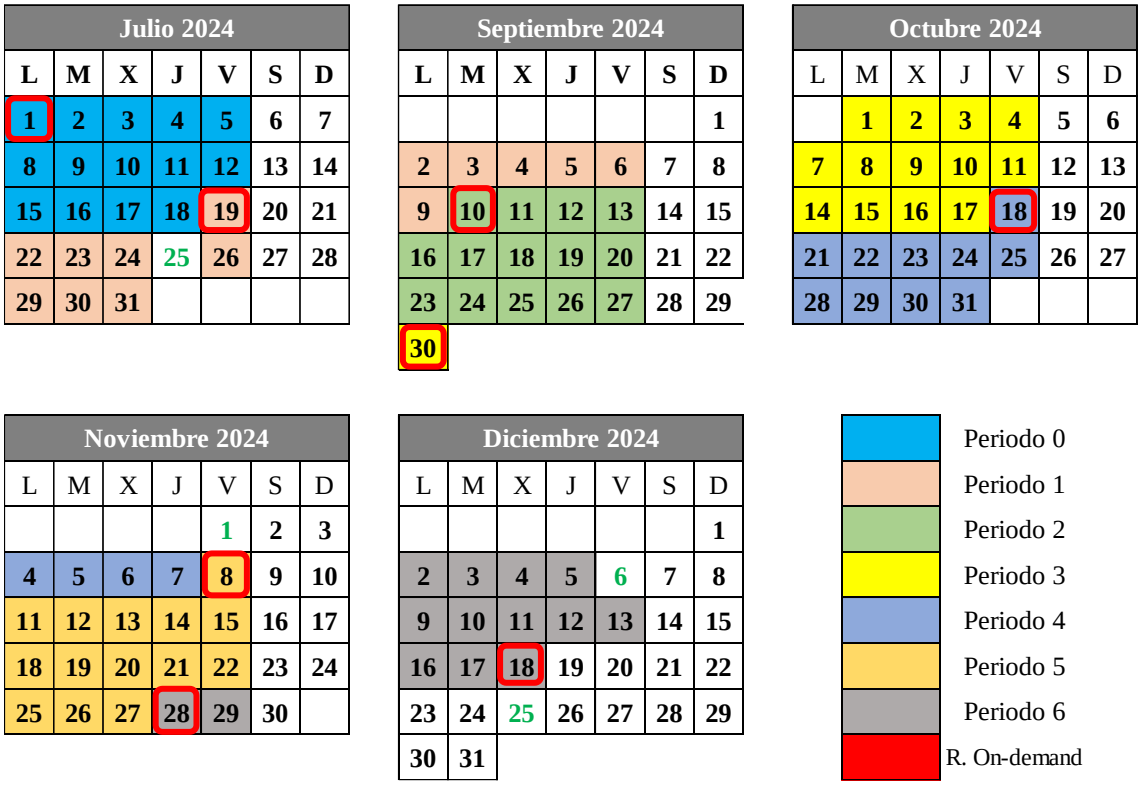


Figura 3.1: Calendario por periodos del proyecto

El proyecto se ha dividido en períodos de trabajo y, como se puede ver en la Figura 3.1, cada período de trabajo tiene la misma duración. Este es el resultado de la distribución de las tareas del proyecto en cada periodo, lo que llevó a igualar dichos periodos de trabajo, así como a una estimación de las reuniones bajo demanda, realizadas éstas para revisar y proponer tareas al inicio de estos periodos.

En las siguientes tablas se muestran las tareas realizadas en cada uno de los periodos:

Periodo	Número de días	Tareas	Total PT
PERIODO 0	14 días	0.1, 0.2, 0.3, 0.4, 0.5, 5,1, 5.2	10 PT
PERIODO 1	14 días	0.6, 1.1, 5.3	10 PT
PERIODO 2	14 días	1.2, 1.3, 2.1, 2.2	10 PT
PERIODO 3	14 días	2.3, 3.1, 3.2, 3.3	10 PT
PERIODO 4	14 días	3.4, 3.5, 4.1	10 PT
PERIODO 5	14 días	4.2, 4.3, 4.4, 4.5, 5.6	10 PT
PERIODO 6	14 días	5.4, 5.5, 5.7, 5.8	10 PT
TOTAL	**98 días**	**TOTAL**	**70 PT**

Tabla 3.7: Tareas repartidas por periodos

Se ha obtenido el tiempo estimado en realizar el proyecto teniendo en cuenta que cada día se trabaja una media de 3 horas, obteniendo un total de 98 días, así como su equivalente en cada punto de tarea teniendo en cuenta la cantidad de días que ha durado el mismo y el total del valor de los puntos de tarea.

DÍAS EJECUTANDO EL PROYECTO	**HORAS EQUIVALENTE**
98 días	300 horas
Total Puntos de Tarea (PT)	**Horas estimadas por Punto de Tarea (PT)**
70 PT	4.28 horas

Tabla 3.8: Cálculo del tiempo total

PERIODO	**NÚMERO DE PT**	**HORA / PT**	**HORAS EJECUTADAS**
PERIODO 0	10 PT	4,28 horas	42.8 horas
PERIODO 1	10 PT	4,28 horas	42.8 horas
PERIODO 2	10 PT	4,28 horas	42.8 horas
PERIODO 3	10 PT	4,28 horas	42.8 horas
PERIODO 4	10 PT	4,28 horas	42.8 horas
PERIODO 5	10 PT	4,28 horas	42.8 horas
PERIODO 6	10 PT	4,28 horas	42.8 horas
TOTAL	**70 PT**	**TOTAL**	**300 horas**

Tabla 3.9: Tiempo estimado de cada bloque

3.2.1. Metodología de gestión de proyecto Trello

Trello es una herramienta de gestión de proyectos que permite organizar tareas mediante tableros visuales. En estos tableros, se utilizan columnas y tarjetas para representar y organizar diferentes tareas, proyectos o flujos de trabajo. Es una plataforma muy popular debido a su simplicidad y flexibilidad.

Trello se basa en la metodología Kanban, siendo ésta una metodología de gestión visual del trabajo, que se utiliza para optimizar el flujo de tareas, basándose en principios de flujo continuo y mejora constante. Kanban emplea tableros visuales con columnas que representan diferentes estados de trabajo.

Inicialmente Trello nos ofrece un tablero con tres columnas, en función del estado en que se encuentra cada tarea, siendo éstas las siguientes:

- **LISTA DE TAREAS**: Columna en la que se asignan las tareas que no están iniciadas.
- **EN PROCESO**: En esta columna se incluyen las tareas que están comenzadas pero que no están terminadas al completo.
- **HECHO**: Esta es la columna en la que se incluyen las tareas que están totalmente completadas.

Para la realizar la parte de gestión de proyectos se ha suprimido la columna HECHO, creando a su vez varias columnas con el objetivo de conseguir una mejor organización de las tareas que engloba el proyecto.

Por un lado, se han creado una columna por cada periodo en el que se encuentra dividida la ejecución del mismo, incluyendo en cada columna las tareas que engloba dicho periodo. Por lo tanto, se han creado 7 columnas que comprenden los 7 periodos de que consta la ejecución del proyecto.

Y por último, se ha creado la columna BLOQUEDAS con el objetivo de ir incluyendo en dicha columna aquellas tareas que se encuentran temporalmente bloqueadas por algún motivo.

Para la mejor identificación del estado de las tareas con una simple mirada, se ha realizado una asignación de colores según el estado en el que se encuentre la tarea en cuestión. Dicha asignación de colores es la siguiente:

- **Color verde**: Para las tareas que no han sido iniciadas, es decir, que estén pendientes de que comience su elaboración.
- **Color naranja**: Para las tareas que estén ya en proceso de ejecución, es decir, que se han empezado ya a ejecutar, aunque haya sido una mínima parte de la misma.

- **Color rojo**: Para las tareas que se encuentran temporalmente bloqueadas, es decir, que durante su ejecución se ha detenido la misma por algún imprevisto.

- **Color azul**: Para las tareas ya realizadas, las cuales se encuentran asignadas a la columna con el periodo correspondiente en el que han sido ejecutadas.

A cada tarea, se le ha asignado una fecha de vencimiento de la misma con el objetivo de tener controlados los tiempos de ejecución del proyecto, además de ir adaptándola en función de las dilataciones temporales sufridas durante la elaboración de las mismas.

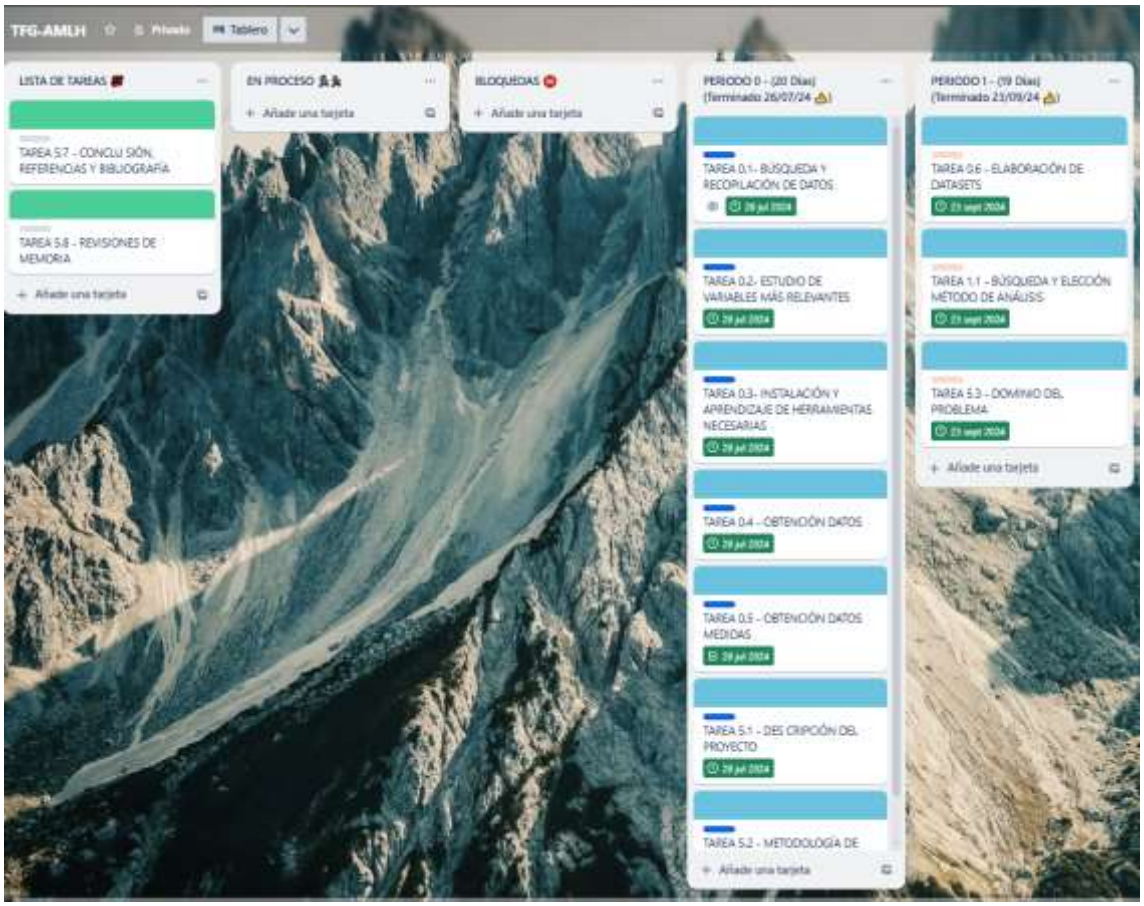


Figura 3.2: Tablero Trello-1 usado en el proyecto

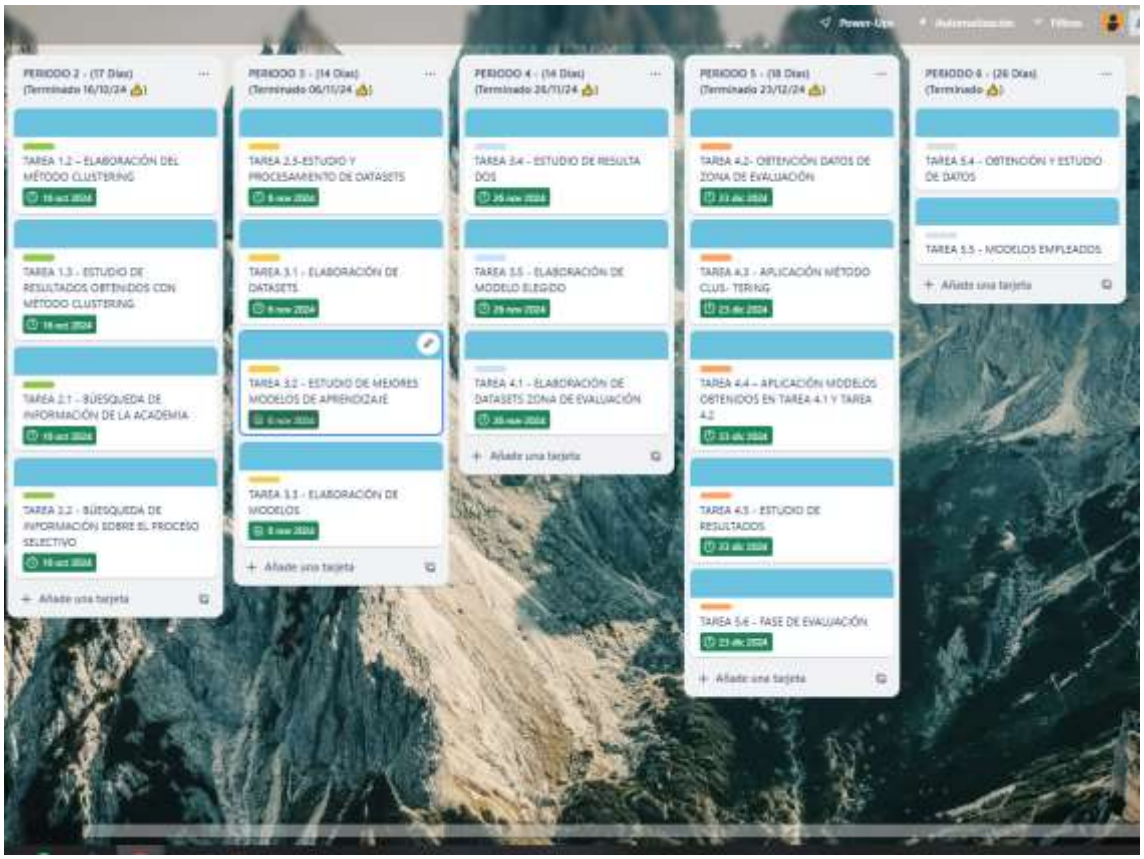


Figura 3.3: Tablero Trello-2 usado en el proyecto

3.3. Presupuesto inicial

Debido al desconocimiento de la variedad de tecnologías o herramientas necesarias para su desarrollo, estimar los costes que implicará un proyecto de investigación como este se vuelve difícil. Al desconocerse los resultados finales, la estimación tanto del tiempo como del esfuerzo se ha realizado de forma muy aproximada. Teniendo esto en cuenta, se ha realizado el desglose del coste del proyecto en cuestión, reflejándose los diferentes importes de los distintos componentes y basando nuestra estimación en estos principios.

3.3.1. Software y Hardware

- **Software:** El coste del software es gratuito.

Herramienta	**Coste (€) / mes**	**Importe total (€)**
Microsoft Teams	0 €	0 €
Trello	0 €	0 €
Microsoft Excel – Word	0 €	0 €
Jupyter Notebooks	0 €	0 €
Google Colab	0 €	0 €
TOTAL	**0** €	**0** €

Tabla 3.10: Presupuesto componentes Software

- **Hardware:** Teniendo en cuenta que el tiempo de trabajo estimado en meses para le elaboración de este proyecto es de 5 meses.

Dispositivo	**Uso (%)**	**Coste (€) / mes**	**Importe total (€)**
PC portátil	50 %	150 €	375 €
Conexión a Internet	50 %	30 €	75 €
TOTAL		**180 €**	**450 €**

Tabla 3.11: Presupuesto componentes Hardware

3.3.2. Recursos humanos (RRHH)

Se utilizará como referencia para el cálculo de los costos de recursos humanos el salario promedio anual de un analista de datos Junior, puesto con un trabajo similar al realizado en este proyecto. El salario por hora se determinará dividiendo este salario anual por el número aproximado de horas trabajadas en un año.

- **Sueldo puesto de trabajo:**

Perfil del trabajador	**Sueldo bruto Anual (€)**	**Importe bruto por hora (€)**
Analista de datos Junior	25000 €	14.17 €

Tabla 3.12: Sueldo perfil del trabajador

- **Coste del proyecto:** Teniendo en cuenta que el tiempo estimado de trabajo de este proyecto son 300 horas.

Perfil del trabajador	**Coste por hora (€)**	**Importe total (€)**
Analista de datos Junior	14.17 €	4251 €

Tabla 3.13: Presupuesto RRHH

3.3.3. Presupuesto total

Una vez sumados todos los costos asociados a este proyecto, siendo estos hardware, software y recursos humanos, se determina el presupuesto total.

Concepto	**Importe (€)**
Software	0 €
Hardware	450 €
Recursos humanos	4251 €
TOTAL	**4701** €

Tabla 3.14: Presupuesto total del proyecto

3.4. Balance

Una vez finalizado el proyecto, se ha realizado una revisión del desarrollo, destacando los ajustes y diversos obstáculos que se habían descubierto en relación con la planificación propuesta inicialmente, viéndose alterada la longitud de los periodos de trabajo previstos en los que se divide el proyecto, obligando a ampliar la duración de alguno de ellos, hecho éste contemplado en la metodología SCORE.

3.4.1. Desviación de la planificación

En concreto, donde se ha producido una dilatación del tiempo inicialmente estimado ha sido en la elaboración de las subtareas (1.1, 1.2 y 1.3) que comprenden la Tarea 1.- Análisis y puesta en marcha de métodos de análisis, y de las subtareas (5.1, 5.2, 5.3, 5.4, 5.5, 5.6, 5.7 y 5.8) que comprenden la Tarea 5.- Elaboración de documentación. Al realizarse dichas subtareas en diferentes periodos, se han visto ampliados los tiempos estimados inicialmente de éstos, afectando por lo tanto a la duración completa estimada

del proyecto, tal y como se puede ver en la figura siguiente que muestra el calendario final del proyecto, así como en la tabla 3.15, que muestra la nueva asignación de PT y de días de trabajo en cada periodo.

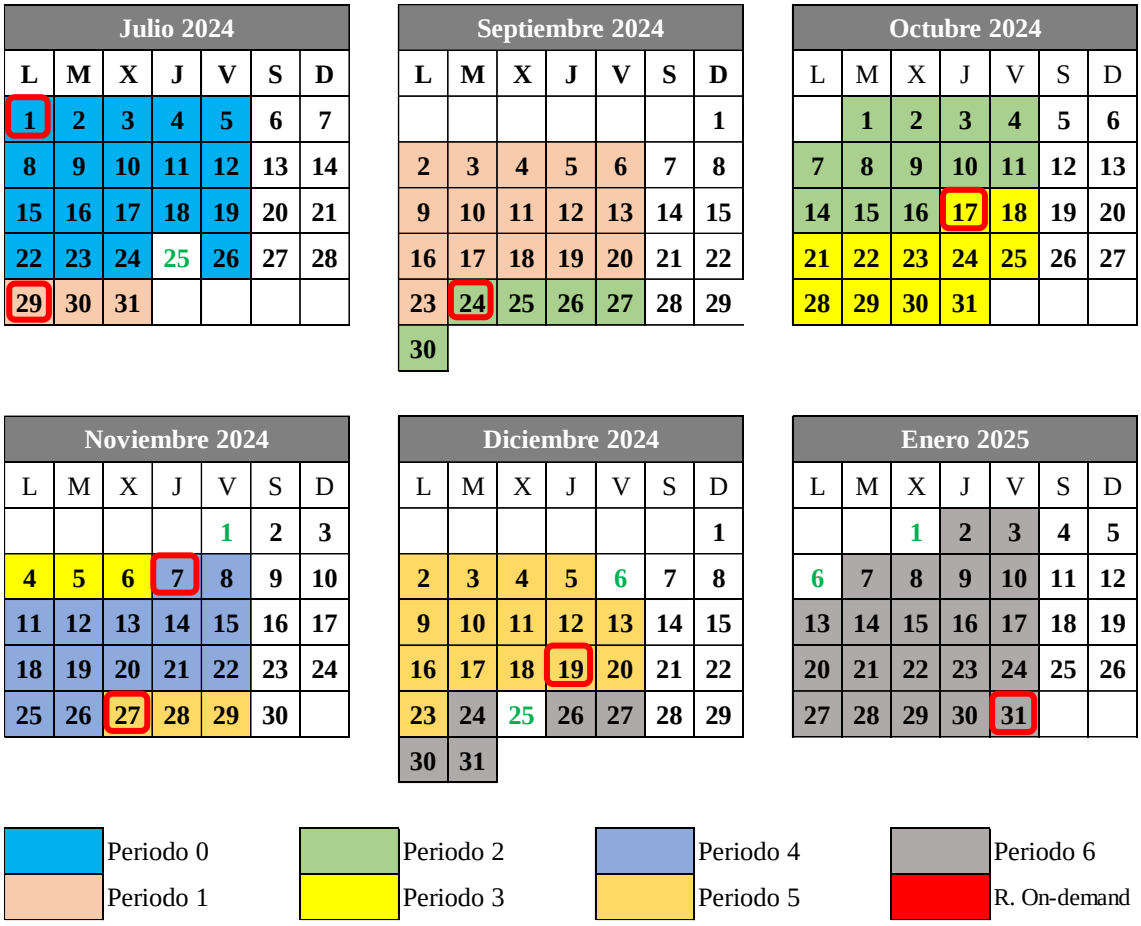


Figura 3.4: Planificación temporal balance

Periodo	Número de días	Tareas	Total PT
PERIODO 0	20 días	0.1, 0.2, 0.3, 0.4, 0.5, 5,1, 5.2	14 PT
PERIODO 1	19 días	0.6, 1.1, 5.3	13 PT
PERIODO 2	17 días	1.2, 1.3, 2.1, 2.2	12 PT
PERIODO 3	14 días	2.3, 3.1, 3.2, 3.3	10 PT
PERIODO 4	14 días	3.4, 3.5, 4.1	10 PT
PERIODO 5	18 días	4.2, 4.3, 4.4, 4.5, 5.6	12 PT
PERIODO 6	26 días	5.4, 5.5, 5.7, 5.8	19 PT
TOTAL	**128 d**	**TOTAL**	**90 PT**

Tabla 3.15: Balance tareas por periodos

3.4.2. Desviación del presupuesto

Debido a estos imprevistos, se ha ampliado el tiempo de ejecución del proyecto a 128 días, es decir, 30 días más de los que se estimaron inicialmente, implicando este hecho una ampliación a 384 horas, 84 horas más de la estimación inicial. Esta dilatación del tiempo inicial estimado con la consiguiente desviación de la planificación inicial se considera que se debe a la propia naturaleza de investigación del proyecto en cuestión.

Días totales empleados en proyecto	**Equivalencia en horas**
128 días	384 horas
Número total de Puntos de Tarea (PT)	**Horas aproximadas por punto de ta- reas (PT)**
90 PT	4.28 horas

Tabla 3.16: Cálculo del tiempo total

A continuación, se muestra la relación de horas invertidas en cada periodo teniendo en cuenta estas dilataciones en los tiempos de ejecución de cada periodo, correspondientes a 128 días totales de trabajo con 385,22 horas invertidas para la elaboración del presente proyecto.

Periodos	**Horas invertidas**
Periodo 0	59,92 horas
Periodo 1	55,64 horas
Periodo 2	51,36 horas
Periodo 3	42,81 horas
Periodo 4	42,81 horas
Periodo 5	51,36 horas
Periodo 6	81,32 horas
TOTAL	**385,22 horas**

Tabla 3.17: Tiempo final de cada bloque con las desviaciones temporales

- **Hardware:** Teniendo en cuenta las dilataciones en los tiempos estimados inicialmente del proyecto, ampliamos los 5 meses estimados inicialmente a 6 meses y medio, es decir, 30 días más de los que inicialmente estaban estimados.

Dispositivo	**Uso (%)**	**Coste (€) / mes**	**Importe total (€)**
PC portátil	50 %	150 €	488 €
Conexión a Internet	50 %	30 €	98 €
TOTAL		**180** €	**586** €

Tabla 3.18: Presupuesto componentes Hardware incluyendo dilataciones temporales.

- **Coste del proyecto:** Teniendo en cuenta la ampliación de horas producidas por las dilataciones temporales en ciertas tareas, ahora las horas invertidas en el proyecto han aumentado a 384, por lo que se ha producido una desviación en el coste del analista Junior.

Perfil del trabajador	**Coste por hora (€)**	**Importe total (€)**
Analista de datos Junior	14.17 €	5441.28 €

Tabla 3.19: Presupuesto RRHH teniendo en cuenta las dilataciones temporales

Una vez sumados todos los costos asociados a este proyecto habiendo tenido en cuenta las dilataciones temporales y su repercusión en cada una los conceptos a tener en cuenta para la obtención del presupuesto, siendo estos hardware, software y recursos humanos, se determina el presupuesto total.

Concepto	**Importe (€)**
Software	0 €
Hardware	586 €
Recursos humanos	5441.28 €
TOTAL	**6027,28** €

Tabla 3.20: Presupuesto total del proyecto teniendo en cuenta las dilataciones temporales

Capítulo 4

Dominio del problema

4.1. Estudio de la influencia de la edad de los aspirantes en el proceso selectivo

El objetivo de este proyecto es analizar si existe alguna tendencia clara en la edad de los aspirantes que han superado el proceso selectivo de acceso al Cuerpo nacional de Policía, siendo estos aspirantes alumnos de una academia que prepara el acceso a dicha oposición, con el fin de diseñar, en función de los resultados obtenidos, desde la publicidad para captar nuevos alumnos hasta la etapa de preparación de la oposición, dado que, incluso teniendo todos los aspirantes el mismo objetivo, que es superar el proceso selectivo, pueden existir variaciones notables en sus inquietudes, sus horarios, su método de preparación, etc.

4.2. Quien forma el Cuerpo Nacional de Policía en España

En España se encuentran varias Fuerzas y Cuerpos de Seguridad, entre ellos están los pertenecientes al Estado, denominados Fuerzas y Cuerpos de Seguridad del Estado. A su vez, estos últimos están formados por el Cuerpo Nacional de Policía y la Guardia Civil, teniendo competencias territoriales iguales, a excepción de la competencia de Tráfico que el Cuerpo Nacional de Policía no la ejerce. Dichos cuerpos pertenecen al Estado, motivo por el cual sus miembros son funcionarios del estado. En concreto, nuestro proyecto versa sobre el estudio del perfil del opositor que ha aprobado el acceso al Cuerpo Nacional de Policía., es decir del Proceso Selectivo del Cuerpo Nacional de Policía.

Las oposiciones para acceder al Cuerpo nacional de Policía constan de 6 pruebas independientes, todas ellas eliminatorias y sin validez para procesos selectivos de posteriores años. Dichas pruebas están recogidas en bases publicadas en cada convocatoria de plazas, siendo éstas las siguientes:

- **Aptitud física**: Constando a su vez de tres pruebas, siendo en un primer lugar de un circuito de agilidad y velocidad, en el cual se ordenará «listo»… «ya», y a esta señal, se deberá realizar el recorrido hasta completarlo en la forma que se indica en el siguiente gráfico, valorándose el tiempo invertido en segundos y décimas de segundo, empezando a contar éste desde la voz de «ya», hasta que el opositor toque el suelo con uno de los pies, habiendo superado con la totalidad del cuerpo la última valla del circuito.

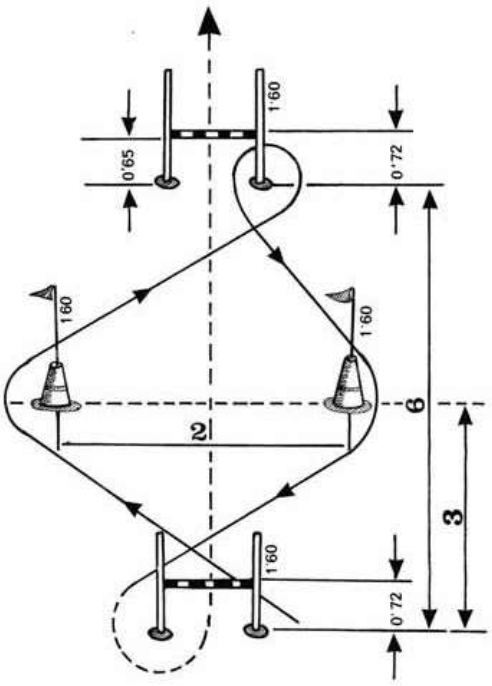


Figura 4.1: Ejecución de la prueba de circuito.

En un segundo lugar, se realizará una prueba para medir la fuerza resistencia de los principales músculos dorsales, flexores de los brazos y la cintura escapulo-humeral. Consistiendo en la realización de dominadas para los hombres, valorándose el mayor número de dominadas posibles hasta completar el máximo, y para las mujeres, suspensión en una barra horizontal, valorándose el mayor tiempo posible suspendida en la posición descrita. Como última prueba física, para medir la resistencia orgánica de los opositores, éstos tendrán que recorrer la distancia de 1.000 metros, valorándose el menor tiempo empleado en la

realización del recorrido. El conjunto de la puntuación de estas tres pruebas deberá superar como mínimo los 15 puntos, no pudiendo obtener una puntuación de 0 puntos en ninguna prueba.

- **Conocimientos**: Consistirá en la contestación por escrito, en un tiempo de cincuenta minutos, a un cuestionario de cien preguntas, con un enunciado y tres alternativas de respuestas, de las que solo una es verdadera, descontando dos acertadas por cada respuesta errónea. Dichas preguntas estarán relacionadas con un temario ofrecido en un anexo adjunto a las bases. Dicha prueba se calificará de 0 a 10 puntos, considerando que han superado la citada prueba los aspirantes que hayan superado la nota de corte prevista.

- **Ortografía**: Consistirá en la contestación por escrito a un cuestionario para evaluar la capacidad ortográfica del aspirante. Esta prueba se calificará de 0 a 10 puntos, siendo aptos los que superen la nota de corte establecida por el Tribunal del proceso selectivo.

- **Reconocimiento médico**: Esta prueba va dirigida a comprobar que no concurren en el aspirante ninguna de las causas de exclusión a que se el cuadro médico de exclusiones para el ingreso en el Cuerpo Nacional de Policía. Se aplicarán a los aspirantes las técnicas médicas de uso convencional que se estimen oportunas, incluida la analítica de sangre y orina o cualquier otra que se estimen convenientes para detectar el consumo de drogas tóxicas, estupefacientes o sustancias psicotrópicas. El resultado de dicha prueba será apto o no apto.

- **Entrevista personal**: Se realizará de carácter profesional y personal, y con la finalidad de comprobar la idoneidad de cada aspirante tomando como referencia criterios que tienen incidencia directa en las funciones policiales a desarrollar en un futuro, siendo realizada al menos por una persona integrante del tribunal calificador y con el asesoramiento de los especialistas que se consideren necesarios. Se evaluará atendiendo a los resultados y conclusiones obtenidos exclusivamente durante su transcurso, sin perjuicio de que se pueda realizar tanto un cuestionario de información biográfica como pruebas de exploración del perfil de personalidad del opositor, de modo que sirvan como información complementaria a la entrevista, apoyadas con un curriculum vitae y una vida laboral. Los criteriosa tener en cuenta por el tribunal serán: socialización, comunicación, orientación hacia las metas, características de la personalidad, observaciones clínicas y competencias profesionales, valorándose dicha prueba con apto o no apto.

- **Test psicotécnico**: Se realizarán uno o varios test dirigidos a determinar las aptitudes de inteligencia general del aspirante para el desempeño de la función policial. El resultado de dichos test establecerá el orden descendente de notas, declarándose aptos a un número de opositores igual al de plazas convocadas.

4.3. Motivos del aumento del número de opositores

Desde ya hace varios años se ha venido observando un aumento del número de opositores a las Fuerzas y Cuerpos de Seguridad. Dicho aumento se cree que es por la precariedad del mercado laboral existente, de hecho, a día de hoy muchos de los aspirantes poseen estudios superiores con los que podrían optar a dicho mercado, pero aun así, prefieren optar a ser funcionarios del estado y tener una estabilidad económica.

En el proceso selectivo de acceso a la Policía Nacional, hasta el año 2011 existía una limitación en la edad de los aspirantes, siendo el límite propuesto de 30 años de edad. Sin embargo, una sentencia del Tribunal Supremo de 21 de marzo de ese mismo año declaró nulo este límite de edad por considerarlo discriminatorio y no justificado.

4.4. Historia del proceso selectivo y antecedentes de estudios anteriores

Antes de la eliminación del límite de edad de 30 años en el 2011, la media de edad de los aspirantes rondaba los 25 años de edad, además, la gran mayoría de ellos carecía de estudios superiores. Se cree que este hecho se debía a que el mercado laboral era más atractivo, consecuencia directa de una situación económica del país más desahogada, por lo que los recién titulados optaban por la empresa privada en vez de tomar el camino del empleo público. En los últimos años, el número de opositores se ha incrementado notablemente, creyendo que ha podido ser fruto la crisis económica del 2008, endureciéndose el mercado laboral y haciendo que los jóvenes, incluso con estudios superiores, optasen por opositar. A este hecho, hay que sumarle la eliminación del límite de edad para acceder al Cuerpo Nacional de Policía, motivo que ha contribuido de manera notable tanto en el incremento del número de aspirantes como en el cambio de perfil de éstos.

4.4.1. Estudios previos sobre la influencia de la edad de los aspirantes

Hasta la fecha no hay constancia de que se haya realizado ningún estudio sobre la influencia de la edad de los aspirantes del proceso selectivo al Cuerpo Nacional de Policía.

4.5. Cambio de perfil del aspirante a raíz de la eliminación del límite de edad

La eliminación de este límite de edad por sentencia de 21 de marzo de 2001, llevó consigo, además de un nuevo aumento en el número de aspirantes, un cambio de perfil de

éstos. Muchos de los opositores que superan los 30 años, tienen limitaciones de horarios debido a ciertas obligaciones que, en muchos caos, van aparejadas con la edad, como pueden ser trabajos estables, familia con hijos, etc.

Dichas limitaciones se trasladan al ámbito académico para preparar la oposición, provocando que no puedan asistir a un horario de clases en horario convencional, por lo que es necesario diseñar una estructura de aprendizaje adaptada a esos perfiles, que no son pocos. Estas nuevas estructuras académicas van desde clases presenciales en horario intensivo de sábados y domingos hasta plataformas en las que la enseñanza se realiza totalmente online, con clases en directo e incluso grabadas para que sean accesibles en cualquier momento, así como contenido teórico y exámenes que la propia plataforma corrige haciendo comparativas con el resto de aspirantes de la academia con el fin de conocer tu posición respecto al resto.

Capítulo 5

Obtención y tratamiento de datos

5.1. Introducción

Hasta marzo del 2011 los aspirantes al Cuerpo Nacional de Policía no podían superar los 30 años de edad, pero a partir de esa fecha, en la que el Tribunal Supremo eliminaba esa restricción por considerarla discriminatoria y no acorde con el trabajo que desarrolla un policía, empezaron a sumarse al proceso selectivo muchos aspirantes que superaban el derogado límite de edad, cosa que hizo que se produjera un aumento considerable en el número de opositores, ayudado este además por la instabilidad del mercado laboral.

La inclusión en el proceso selectivo de aspirantes de mayor edad, trajo consigo una reestructuración del modelo tradicional de preparación de la oposición, dado que el aumento de edad lleva aparejada, en una gran mayoría de casos, responsabilidades ineludibles de los aspirantes, con las limitaciones que eso conlleva, sobre todo en el marco horario.

Estas limitaciones han llevado a las academias que preparan a los opositores a reestructurar su método de enseñanza con cursos completamente online, cursos intensivos, etc, adaptándose a las necesidades de los aspirantes.

5.2. Caso de estudio

Este proyecto está realizado sobre los alumnos de una academia que prepara el acceso al Cuerpo Nacional de Policía que han superado la primera parte del proceso selectivo, que

consiste en superar las 6 pruebas de que consta esta primera fase. A estas 6 pruebas hay que sumarle una séptima prueba de Conocimiento de Idioma Extranjero, que solamente suma a la

nota final obtenida, es decir, en ningún caso es eliminatoria.

En concreto, estas 7 pruebas que componen esta primera fase del proceso selectivo son las siguientes:

- Aptitud Física.
- Conocimiento.
- Ortografía.
- Reconocimiento médico.
- Entrevista Personal.
- Test Psicotécnico.
- Conocimiento de Idioma extranjero.

Una vez que el aspirante haya superado esta primera etapa, tendrá que realizar durante un curso lectivo una formación ya más específica en la Academia del Cuerpo Nacional de Policía ubicada en la ciudad de Ávila, para posteriormente terminar el proceso selectivo con un año de prácticas trabajando en una plantilla real en una de las Comisarías de la Policía Nacional que se encuentran en todo el territorio nacional.

Superadas ya todas estas etapas, el aspirante por fin jurará el cargo como funcionario de carrera entregándosele la acreditación como tal, constando ésta de un carné profesional y de una placa emblema, y del destino a nivel nacional donde pasará a ejercer sus funciones como funcionario de carrera perteneciente a la Policía Nacional.

Figura 5.1: Carné profesional y placa emblema de la Policía Nacional.

El objetivo de este proyecto es realizar el estudio sobre la influencia que tiene la edad de los aspirantes en el proceso de preparación de la oposición, con el fin de determinar si existe alguna tendencia que relacione la edad con el resultado obtenido en cada prueba del proceso, así como con el resultado final del proceso, siendo éste el cómputo de todas y cada una de las pruebas que lo forman.

5.3. Fuentes de datos

Estos datos nos han permitido realizar el estudio de este proyecto en cuestión, los cuales han sido cedidos por una academia que prepara a los aspirantes para el proceso selectivo de acceso al Cuerpo Nacional de Policía del año 2022.

5.3.1. Metodología de recopilación

La academia en cuestión facilitó los datos de los alumnos en un único fichero en formato Excel llamado NOTAS PROMO 38-V01.xlsx, el cual contenía varias hojas con la afirmación obtenida por los opositores en cada una de las pruebas. Como se tratan de datos de personas concretas, éstos ya venían sin los datos personales de los opositores, es decir, sin nombre ni apellidos, ni número de D.N.I., solamente con un identificador de opositor que, en ningún caso, lleva a asociarlo con dichos datos personales, por lo que es imposible asociar cualquier dato de este archivo con una persona física.

En cada hoja del mencionado archivo, las columnas corresponden a las variables a tener en cuenta según la información facilitada por el archivo, y las filas a los valores de dichas variables de cada opositor. A continuación, se detalla el contenido de las hojas del archivo con sus cabeceras:

- **FÍSICAS-V0**: En esta hoja se reflejan las puntuaciones obtenidas en la primera prueba que se realiza en el proceso selectivo, prueba ésta de Aptitud Física, siendo aptas las notas a partir de 5,00 puntos. A continuación, se detallan las columnas que contiene el archivo:
    - **IdPer**: Número genérico que se le asigna a un opositor para su identificación, pero que en ningún caso permite conocer la identidad del mismo. Ejemplo: 2515.
    - **EDAD**: Edad del opositor en cuestión. Este campo viene defino en número de años. Ejemplo: 32.
    - **FISICAS**: Nota obtenida en la prueba de aptitud física. Dicha nota conlleva 2 decimales acorde a las puntuaciones que se pueden obtener. Ejemplo: 5,00.

- **GLOBAL FÍSICAS**: Comentario sobre la nota obtenida por el opositor o sobre si se ha presentado o no al examen. Ejemplo: APTO.

IdPer	EDAD	FÍSICAS	GLOBAL FÍSICAS
9032	24	5,670	APTO
8196	30	0,000	NO INSTANCIA
1508	29	5,000	APTO

Figura 5.2: Cabecera de hoja FÍSICAS-V0.

- **TEORÍA Y ORTO-V0**: Esta hoja recoge las notas obtenidas en las pruebas de Conocimiento y Ortografía. Seguidamente se detallan las columnas que contiene el archivo:

  - **IdPer**: Número genérico que se le asigna a un opositor para su identificación, pero que en ningún caso permite conocer la identidad del mismo. Ejemplo: 2515.

  - **EDAD**: Edad del opositor en cuestión. Este campo viene defino en número de años. Ejemplo: 32.

  - **TEORÍA**: Nota obtenida en la prueba de conocimientos, siendo aptas las notas a partir de 5,000 puntos. Dicha nota conlleva 3 decimales acorde a las puntuaciones que se pueden obtener. Ejemplo: 5,721.

  - **ORTO**: Nota obtenida en la prueba de ortografía, siendo aptas las notas a partir de la nota de corte proporcionada por el tribunal, la cual varía todos los años, siendo la nota de corte de este año en cuestión de 4,0 puntos. Dicha nota conlleva 1 decimal acorde a las puntuaciones que se pueden obtener. Ejemplo: 4,4.

  - **GLOBAL TEORÍA/ORTO**: Comentario sobre la nota obtenida por el opositor o sobre si se ha presentado o no al examen. Ejemplo: APTO.

IdPer	EDAD	TEORÍA	ORTO	GLOBAL TEORÍA/ ORTO
9032	24	1,597	0,800	NO APTO
1508	30	5,154	4,900	APTO
7624	29	5,000	4,100	APTO

Figura 5.3: Cabecera de hoja TEORÍA Y ORTO-V0.

- **ENTREVISTA Y MÉDICO-V0**: En esta hoja se reflejan si los opositores han superado las pruebas de Entrevista Personal y de Reconocimiento Médico. A continuación, se detallan las columnas que contiene el archivo:

    - **IdPer**: Número genérico que se le asigna a un opositor para su identificación, pero que en ningún caso permite conocer la identidad del mismo. Ejemplo: 2515.

    - **EDAD**: Edad del opositor en cuestión. Este campo viene defino en número de años. Ejemplo: 32.

    - **ENTREVISTA**: Resultado de la prueba de Entrevista Personal. En este campo solamente se especifica si a supera o no la prueba, es decir, no se indica ninguna nota numérica, solamente si apto o no. Ejemplo: APTO.

    - **MEDICO**: Igualmente que el campo anterior, en esta columna solamente si ha superado o no el reconocimiento Médico con un "APTO" o "NO APTO". Ejemplo: APTO.

    - **GLOBAL ENTREVISTA/MEDICO**: Columna en la que se refleja si el opositor es apto para pasar a la siguiente prueba. Para serlo, debe de haber supera las pruebas de Entrevista y Reconocimiento Médico. Ejemplo: APTO.

IdPer	EDAD	ENTREVISTA	MEDICO	GLOBAL ENTREVISTA/MEDICO
7624	32	APTO	APTO	APTO
4498	30	APTO	APTO	APTO
3605	27	APTO	APTO	APTO

Figura 5.4: Cabecera de hoja ENTREVISTA Y MÉDICO-V0.

- **PSICOS-V0**: Esta hoja muestran las notas obtenidas por los aspirantes en la prueba de Psicotécnicos, pasando dicha prueba los que hayan conseguido superar la nota de corte específica en el año concreto, variando esta última todos los años. Este año en concreto, la nota de corte fue de 4,0 puntos. Seguidamente se muestran las columnas que contiene el archivo:

    - **IdPer**: Número genérico que se le asigna a un opositor para su identificación, pero que en ningún caso permite conocer la identidad del mismo. Ejemplo: 2515.

    - **EDAD**: Edad del opositor en cuestión. Este campo viene defino en número de

años. Ejemplo: 32.

- **PSICOS**: Esta columna refleja la nota obtenida por el opositor en la prueba de psicotécnicos. Dicha nota conlleva 3 decimales acorde a las puntuaciones que se pueden obtener. Ejemplo: 5,808.

IdPer	Edad	PSICOS
7624	32	4,55020
4498	30	6,44185
3605	27	5,80028

Figura 5.5: Cabecera de hoja PSICOS-V0.

- **IDIOMA-V0**: La prueba de Idioma Extranjero es una prueba que solamente suma en el resultado del proceso selectivo, es decir, que en caso de obtener mala puntuación en ella no provocaría una bajada de la nota definitiva de esta parte del proceso. Se puede elegir entre dos idiomas, el inglés o el francés. A continuación, se detallan las columnas que contiene el archivo:

  - **IdPer**: Número genérico que se le asigna a un opositor para su identificación, pero que en ningún caso permite conocer la identidad del mismo. Ejemplo: 2515.

  - **EDAD**: Edad del opositor en cuestión. Este campo viene defino en número de años. Ejemplo: 32.

  - **IDIOMA**: Esta columna refleja la nota obtenida por el opositor en la prueba de Idioma Extranjero. El máximo de puntuación que se puede obtener son 2,0 puntos. Dicha nota conlleva 3 decimales acorde a las puntuaciones que se pueden obtener. Ejemplo: 0,675.

IdPer	Edad	IDIOMA
7624	32	0,000
4498	30	1,475
3605	27	0,875

Figura 5.6: Cabecera de hoja IDIOMA-V0.

- **APTOS OPOSICIÓN-V0**: En esta hoja se muestra la nota obtenida por los opositores que han superado todo el proceso selectivo y el escalafón que ocupan dentro del global de las plazas ofertadas por la Dirección General de la Policía. Seguidamente se muestran las columnas que contiene el archivo:

    - IdPer: Número genérico que se le asigna a un opositor para su identificación, pero que en ningún caso permite conocer la identidad del mismo. Ejemplo: 2515.

    - **EDAD**: Edad del opositor en cuestión. Este campo viene defino en número de años. Ejemplo: 32.

    - **FINAL**: Esta columna refleja un cómputo de la nota obtenida por el aspirante después de haber realizado todo el proceso selectivo. Este cómputo no concentra todas las pruebas realizadas, solamente contempla la suma de 3 pruebas, siendo éstas la de Aptitud Física, la de Conocimientos y la de Conocimiento de Idioma Extranjero. Dicha nota conlleva 3 decimales acorde a las puntuaciones que se pueden obtener. Ejemplo: 11,396.

    - **ESCA**: En esta columna se indica la posición en el escalafón en función de la nota obtenida por el opositor en las tres pruebas a las que se hacía mención en campo anterior, es decir, la de Aptitud Física, la de Conocimientos y la de Conocimiento de Idioma. Ejemplo: 2109.

IdPer	EDAD	FINAL	ESCALAFÓN
4498	30	17,258	2
3605	27	13,442	839
4935	27	12,451	1568

Figura 5.6: Cabecera de hoja APTOS OPOSICIÓN-V0.

5.4. Creación de datasets

Una vez estudiados los datos que facilitó la academia de oposiciones, se decidió separar cada hoja del archivo de Excel NOTAS-PROMO 38-V01.xlsx en archivos independientes de Excel para facilitar su manejabilidad. Por ello, se crearon los siguientes archivos Excel correspondientes a las pruebas que se evalúan en el proceso selectivo:

- NOTAS PROMO 38-Fisicas V01.xlsx

- NOTAS PROMO 38-Teoria V01.xlsx

- NOTAS PROMO 38-Orto V01.xlsx
- NOTAS PROMO 38-Teoria_Orto V01.xlsx
- NOTAS PROMO 38-Teoria_Medico V01.xlsx
- NOTAS PROMO 38-Teoria_Entrevista V01.xlsx
- NOTAS PROMO 38-Psicos V01.xlsx
- NOTAS PROMO 38- Idioma V01.xlsx
- NOTAS PROMO 38-Aptos Nota V01.xlsx
- NOTAS PROMO 38-Aptos Escalafon V01.xlsx

En cada archivo, las columnas corresponden a las variables a tener en cuenta según la información facilitada en el archivo, y las filas a los valores de dichas variables de cada opositor.

A continuación, se detalla cada fichero junto con sus variables:

- **NOTAS PROMO 38-Fisicas V01.xlsx**:

Variable		
Nombre	Tipo	Descripción
IdPer	INT	mero genérico que se le asigna a un opositor a su identificación
EDAD	INT	ad del opositor
FÍSICAS	FLOAT	ta obtenida en la prueba de Aptitud Física

Tabla 5.1: Cabecera fichero NOTAS PROMO 38-Fisicas V01.xlsx

- **NOTAS PROMO 38-Teoria V01.xlsx:**

Variable		
Nombre	Tipo	Descripción
IdPer	INT	úmero genérico que se le asigna a un positor para su identificación
EDAD	INT	dad del opositor
TEORIA	FLOAT	ota obtenida en la prueba de Conocimientos

Tabla 5.2: Cabecera fichero NOTAS PROMO 38-Teoria V01.xlsx

- **NOTAS PROMO 38-Orto V01.xlsx**:

Variable		
Nombre	Tipo	Descripción
IdPer	INT	mero genérico que se le asigna a un positor para su identificación
EDAD	INT	ad del opositor
ORTO	FLOAT	ta obtenida en la prueba de Ortografía

Tabla 5.3: Cabecera fichero NOTAS PROMO 38-Orto V01.xlsx

- **NOTAS PROMO 38-Teoria_Orto V01.xlsx**: La prueba de Conocimientos y de Ortografía se computan al mismo tiempo, siendo necesario para superarlas obtener una la puntuación igual o superior a la nota de corte. Por lo tanto, lo que refleja este fichero es si los opositores son aptos o no para pasar a la siguiente prueba, habiendo sido reemplazado el valor de "APTO" por un 1 y el valor de "NO APTO" por un 0.

Variable		
Nombre	Tipo	Descripción
IdPer	INT	úmero genérico que se le asigna a un positor para su identificación
EDAD	INT	dad del opositor
TEORTO	INT	esultado conjunto de ambas pruebas, omando el valor de 1 si se han superado mbas pruebas y de 0 en caso contrario.

Tabla 5.4: Cabecera fichero NOTAS PROMO 38-Teoria_Orto V01.xlsx

- **NOTAS PROMO 38-Medico V01.xlsx**: Lo que se refleja en este fichero es si los opositores son aptos o no para pasar a la siguiente prueba, es decir, los que hayan superado el reconocimiento médico serán valorados con un "APTO" y los que no con un "NO APTO", habiendo sido reemplazado el valor de "APTO" por un 1 y el valor de "NO APTO" por un 0.

Variable		
Nombre	Tipo	Descripción
IdPer	INT	mero genérico que se le asigna a un opositor a su identificación
EDAD	INT	ad del opositor
MEDICO	INT	sultado conjunto de ambas pruebas tomando el or de 1 si se han superado ambas pruebas y de n caso contrario

Tabla 5.5: Cabecera fichero NOTAS PROMO 38-Medico V01.xlsx

- **NOTAS PROMO 38-Entrevista V01.xlsx**: Lo que se refleja en este fichero es si los opositores han superado o no la prueba de Entrevista Personal para pasar a la siguiente prueba, es decir, valorando con "APTO" los que la hayan superado y con un "NO APTO" los que no, habiendo sido reemplazado el valor de "APTO" por un 1 y el valor de "NO APTO" por un 0.

Variable		
Nombre	Tipo	Descripción
IdPer	INT	Número genérico que se le asigna a un opositor para su identificación
EDAD	INT	Edad del opositor
ENTREVISTA	INT	Resultado conjunto de ambas pruebas tomando el valor de 1 si se han superado ambas pruebas y de 0 en caso contrario

Tabla 5.6: Cabecera fichero NOTAS PROMO 38-Entrevista V01.xlsx

- **NOTAS PROMO 38-Psicos V01.xlsx:**

Variable		
Nombre	Tipo	Descripción
IdPer	INT	Número genérico que se le asigna a un opositor para su identificación
EDAD	INT	Edad del opositor
PSICOS	FLOTA	Nota obtenida en la prueba de Psicotécnicos

Tabla 5.7: Cabecera fichero NOTAS PROMO 38-Psicos V01.xlsx

- **NOTAS PROMO 38- Idioma V01.xlsx:**

Variable		
Nombre	Tipo	Descripción
IdPer	INT	mero genérico que se le asigna a un ositor para su identificación
EDAD	INT	ad del opositor
IDIOMA	FLOTA	ta obtenida en la prueba de Idioma

Tabla 5.8: Cabecera fichero NOTAS PROMO 38-Idioma V01.xlsx

- **NOTAS PROMO 38-Aptos Nota V01.xlsx:**

Variable		
Nombre	Tipo	Descripción
IdPer	INT	mero genérico que se le asigna a un sitor para su identificación
EDAD	INT	d del opositor
NOTA	FLOTA	mputo de la nota obtenida en las pruebas Aptitud Física, la de Conocimientos y la Conocimiento de Idioma Extranjero

Tabla 5.9: Cabecera fichero NOTAS PROMO 38-Aptos Nota V01.xlsx

Capítulo 6

Análisis de datos

En este capítulo se aborda el análisis realizado a los datos obtenidos para el proyecto. Dicho análisis está realizado para cada uno de los ficheros de formato Excel que componen el dataset. A continuación, se mostrarán las gráficas obtenidas, así como las conclusiones a las que se han llegado después de cada análisis.

6.1. Herramientas de análisis empleadas

Para realizar el análisis de los datos que han sido facilitados por la academia de formación en cuestión, se han aplicado dos tipos de herramientas de la aplicación Excel, Gráfico de dispersión con Línea de Regresión Lineal y Coeficiente de Correlación, las cuales han ofrecido con claridad si existe o no alguna relación entre las puntuaciones obtenidas en cada prueba y la edad de los aspirantes.

6.1.1. Gráfico de Dispersión con Línea de Regresión Lineal

Un gráfico de dispersión, también conocido como diagrama de dispersión, es una representación gráfica que utiliza coordenadas cartesianas para mostrar valores de dos variables para un conjunto de datos. En este tipo de gráfico, cada punto representa un par de valores correspondientes a las dos variables, con una variable en el eje horizontal (eje X) y la otra en el eje vertical (eje Y). Este tipo de gráficos se utiliza principalmente para identificación de relaciones, para detección de patrones, para identificación de outliers, así como para el

análisis de regresión. En resumen, un gráfico de dispersión es una herramienta visual muy valiosa para analizar la relación entre dos variables y extraer conclusiones valiosas sobre su comportamiento conjunto.

Una línea de regresión lineal es una recta que se ajusta a un conjunto de datos en un gráfico de dispersión, de tal manera que minimiza la suma de los cuadrados de las distancias verticales de los puntos de datos a la línea, dándose a conocer a este proceso de ajuste como "mínimos cuadrados". El uso principal de una línea de regresión puede ser para realizar una predicción, o para un análisis de tendencias, o para identificación de outliers, así como para realizar una evaluación de la relación lineal, siendo el uso que se le da en el presente proyecto, permitiendo evaluar la fuerza y la dirección de la relación lineal entre dos variables, a lo que una pendiente positiva indica una relación positiva, mientras que una pendiente negativa indica una relación negativa.

Por lo tanto, la combinación de las dos maneras de análisis y representación de las variables proporciona una gran visualización que permite analizar y comprender la relación entre dos variables. De la visualización de la relación lineal tenemos que tener en cuenta los siguientes dos aspectos:

- Relación General: La línea de regresión muestra la tendencia general de los datos. Si los puntos de datos se alinean cerca de la línea de regresión, indica una fuerte relación lineal entre las variables.

- Pendiente: La pendiente de la línea de regresión indica la dirección y el grado de cambio de la variable dependiente (eje Y) con respecto a la variable independiente (eje X). Una pendiente positiva indica una relación positiva, mientras que una pendiente negativa indica una relación negativa.

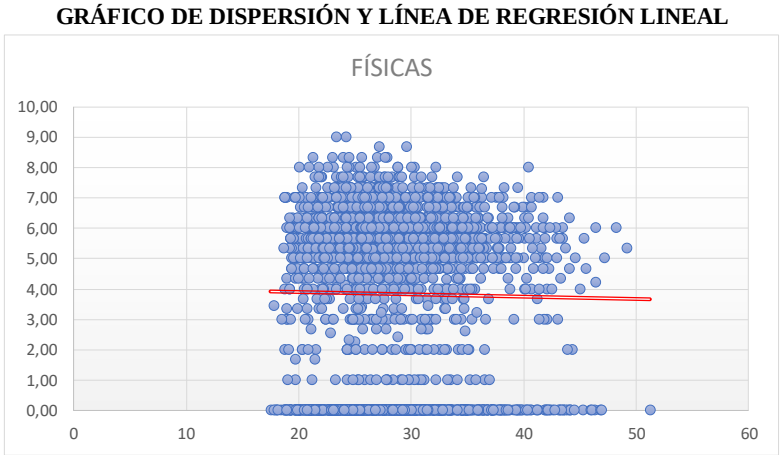


Figura 6.1: Ejemplo de las gráficas del gráfico de dispersión con una línea de regresión lineal.

6.1.2. Coeficiente de Correlación

El coeficiente de correlación es un valor numérico que mide la fuerza y la dirección de la relación lineal entre dos variables. Dicho coeficiente se utiliza en diversas disciplinas, como la estadística, economía, ciencias sociales, y en análisis de datos en general, para entender cómo dos variables están relacionadas. Dependiendo del tipo de correlación, se puede usar para determinar cómo una variable tiende a cambiar cuando la otra lo hace.

Existen los siguientes tipos de coeficiente de correlación:

- **Coeficiente de correlación de Pearson**: El cual se utiliza para medir la relación lineal entre dos variables.

- **Coeficiente de correlación de Spearman**: Este coeficiente de correlación se usa principalmente para medir la relación monótona (no necesariamente lineal) entre variables ordinales o no paramétricas.

- **Coeficiente de correlación de Kendall**: Este coeficiente mide la concordancia entre dos variables ordinales.

El coeficiente más comúnmente utilizado es el coeficiente de correlación de Pearson, el cual mide la correlación lineal entre dos variables continuas, por lo que es el que se va a utilizar en el presente proyecto.

La interpretación de este coeficiente teniendo en cuenta el valor de "r" es la siguiente:

- **Correlación positiva ($0 < r \leq 1$)**: A medida que una variable aumenta, la otra también lo hace. Por ejemplo, más horas de estudio suelen estar asociadas con mejores notas.

- **Correlación negativa ($-1 \leq r < 0$)**: A medida que una variable aumenta, la otra disminuye. Por ejemplo, menos horas de estudio pueden estar asociadas con peores resultados académicos.

- **No correlación ($r \approx 0$)**: No hay una relación lineal clara entre las variables. Las variaciones en una no permiten predecir las variaciones en la otra.

También se pude tener en cuenta a la hora de realizar la interpretación de los valores la siguiente escala, en la cual nos da una idea más aproximada del grado de correlación entre ambas variables:

- **De 0.7 a 1 o de -0.7 a -1**: Correlación fuerte, positiva o negativa respectivamente.

- **De 0.4 a 0.7 o de -0.4 a -0.7**: Correlación moderada, positiva o negativa respectivamente.

- **De 0 a 0.4 o de -0.4 a 0**: Correlación débil, positiva o negativa respectivamente.

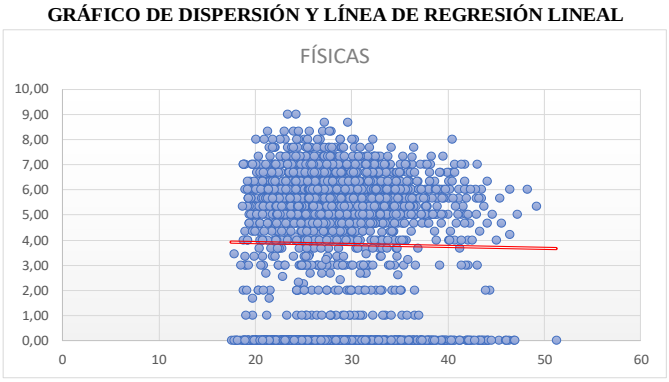


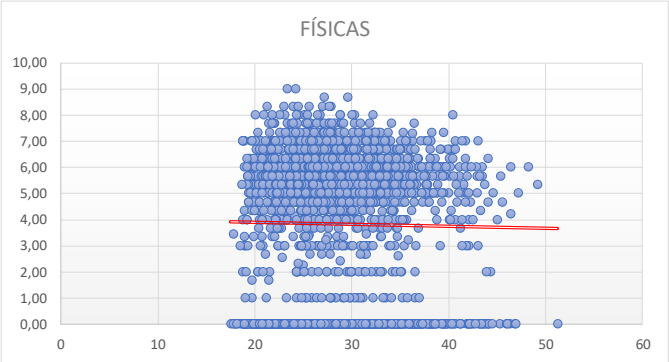
Figura 6.2: Ejemplo del gráfico del coeficiente de correlación.

6.2. Aplicación de herramientas para el análisis de datos

A continuación, se aplicarán estos dos métodos de Excel para realizar en ánalisis a todos los ficheros de Excel, antes de ser transformados al formato .csv, que forman nuestro dataset, mostrando a continuación el resultado que se puede apreciar en las siguientes imágenes en función de cada archivo csv que forman el dataset del presente proyecto:

- **Fichero NOTAS PROMO 38-Fisicas V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



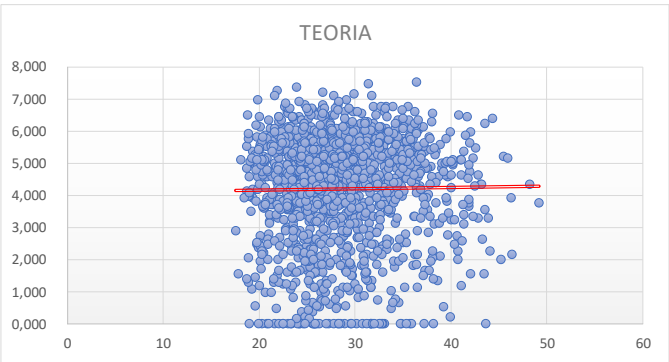
COEFICIENTE DE CORRELACIÓN
-0,013985158

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.3: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Físicas.

- **Fichero NOTAS PROMO 38-Teoria V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



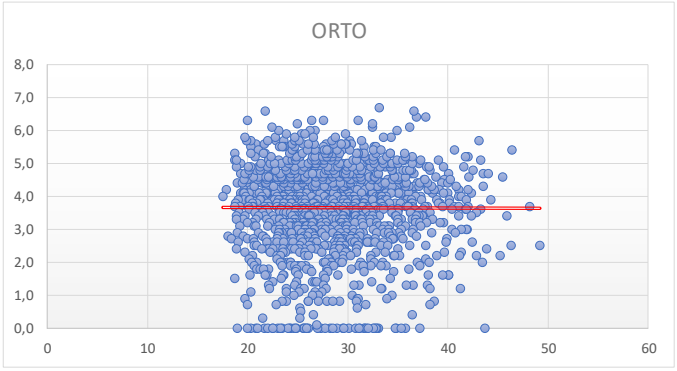
COEFICIENTE DE CORRELACIÓN
0,013800619

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.4: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Teoría.

- **Fichero NOTAS PROMO 38-Orto V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



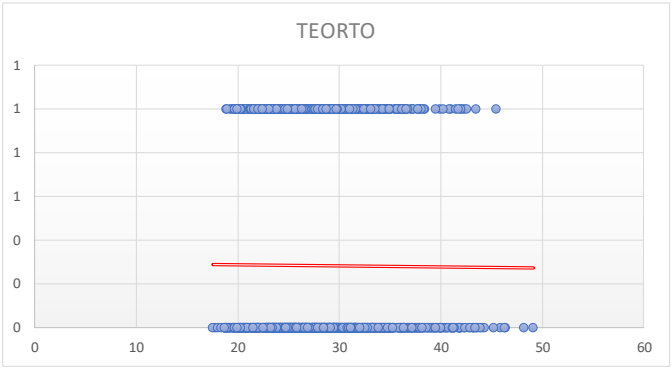
COEFICIENTE DE CORRELACIÓN
-0,002528304

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.5: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Ortografía.

- **Fichero NOTAS PROMO 38-Teoria_OrtoV01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



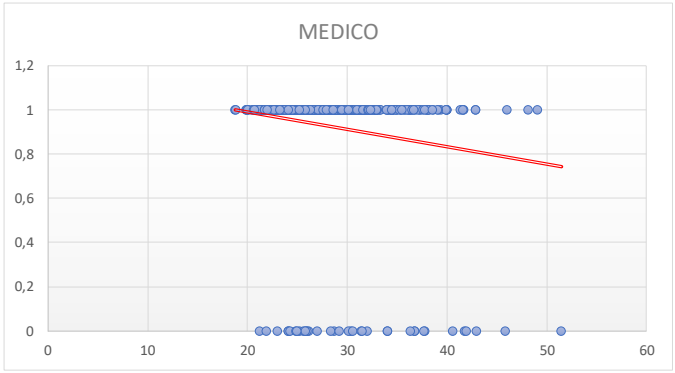
COEFICIENTE DE CORRELACIÓN
-0,005614741

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.6: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Teoría y Ortografía.

- **Fichero NOTAS PROMO 38-Medico V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



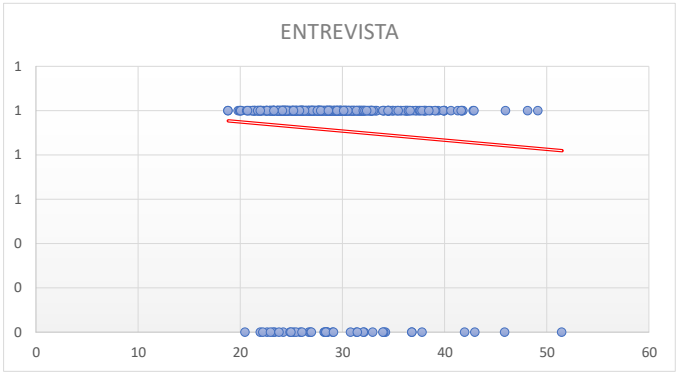
COEFICIENTE DE CORRELACIÓN
-0,155917052

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.7: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Médico.

- **Fichero NOTAS PROMO 38-Entrevista V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



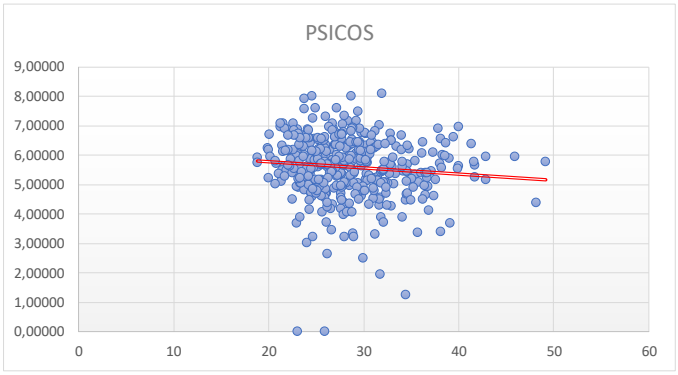
COEFICIENTE DE CORRELACIÓN
-0,077363347

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.8: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Entrevista.

- **Fichero NOTAS PROMO 38-Psicos V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



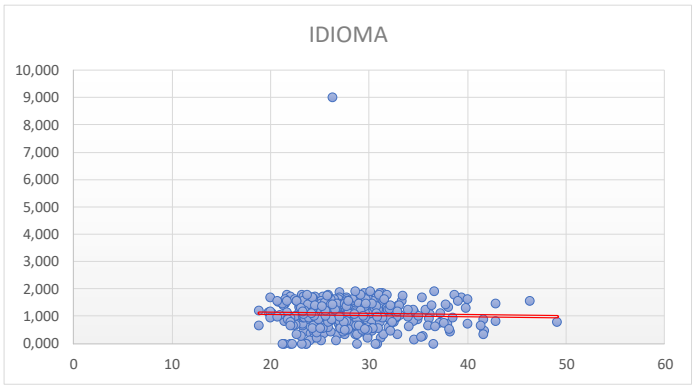
COEFICIENTE DE CORRELACIÓN
-0,102411416

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.9: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Psicotécnicos.

- **Fichero NOTAS PROMO 38-Idioma V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



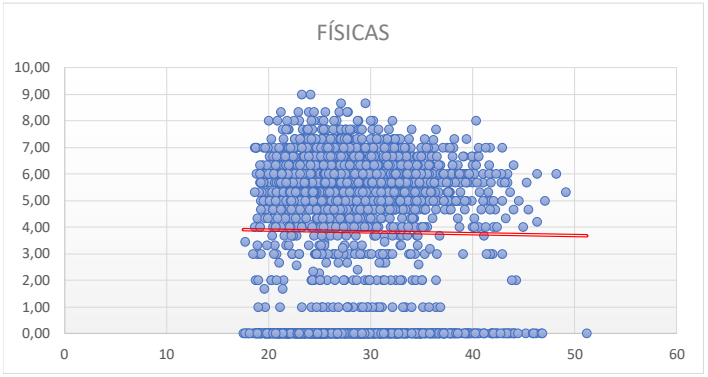
COEFICIENTE DE CORRELACIÓN
-0,030410018

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.10: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Idioma.

- **Fichero NOTAS PROMO 38-Aptos Nota V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



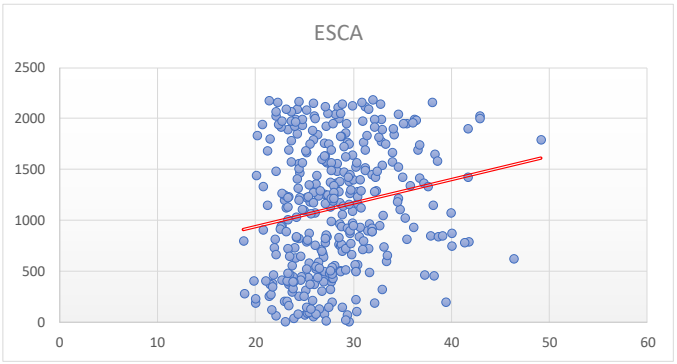
COEFICIENTE DE CORRELACIÓN
-0,013985158

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.11: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Aptos Nota.

- **Fichero NOTAS PROMO 38-Aptos Escalafon V01-Análisis:**

GRÁFICO DE DISPERSIÓN Y LÍNEA DE REGRESIÓN LINEAL



COEFICIENTE DE CORRELACIÓN
0,185454625

Interpretación: El valor resultante oscilará entre -1 y 1.
De 0.7 a 1 o de -0.7 a -1: Correlación fuerte, positiva o negativa
De 0.4 a 0.7 o de -0.4 a -0.7: Correlación moderada, positiva o negativa
Cercano a 0: No hay correlación significativa entre las variables.

Figura 6.12: Análisis del gráfico de dispersión y del coeficiente de correlación del archivo Aptos Esca.

Una vez analizadas todas las gráficas de cada uno de los archivos csv que componen el dataset, se llega a la conclusión de que no existe prácticamente relación entre la edad del opositor y el resultado de cada una de las pruebas, dado que en el gráfico de dispersión junto con la línea de regresión lineal se puede apreciar que no existe agrupamiento cerca de dicha línea, además de que ésta se encuentra prácticamente en posición horizontal en la mayoría de los casos, indicando estos dos aspectos que no existe relación entre la variables analizadas. Del mismo modo, se puede apreciar que el coeficiente de correlación se encuentra muy cercano a 0 en todos los casos, motivo éste que indica que no existe una correlación significativa entre ambas variables.

Por lo tanto, la conclusión del análisis realizado con estos dos métodos de Excel es que no hay relación significativa entre la edad del aspirante y cada una de las notas obtenidas por éste en las diferentes pruebas realizadas en esta parte del proceso selectivo.

Capítulo 7

Métodos de aprendizaje: Clustering

7.1. Introducción

Para la realización de este proyecto se ha realizado la aplicación de diferentes métodos y modelos a los datos del mismo, siendo este uno de los pasos más importantes en la ejecución del mismo. Este trabajo nos ha llevado a la elección de las técnicas más indicadas para la consecución de nuestro objetivo.

Los métodos utilizados en el trabajo de investigación no fueron especificados desde un inicio, por lo que para lograr resultados en el campo de la inteligencia artificial fue necesario buscar y analizar las formas de trabajar con los datos disponibles, según los objetivos planificados.

7.2. Clustering

El clustering es un método de aprendizaje automático no supervisado que juega un papel crucial en el análisis de datos. Se utiliza para identificar estructuras o grupos intrínsecos en un conjunto de datos sin etiquetas, donde las instancias dentro de cada grupo son más similares entre sí que con las de otros grupos. Esta técnica es fundamental en diversas aplicaciones, desde la segmentación del mercado y la organización social hasta el análisis de secuencias genéticas y la detección de patrones en el estudio de datos de salud, como en el proyecto en cuestión. Su importancia radica en la capacidad de revelar patrones ocultos y estructuras desconocidas en grandes volúmenes de datos, lo que permite a los investigadores y analistas tomar decisiones informadas y descubrir conocimientos sin preconcepciones explícitas.

7.2.1. Fundamentos del Clustering

El clustering, también conocido como agrupamiento o segmentación, es una técnica de aprendizaje no supervisado en la que se busca dividir un conjunto de datos en grupos o clusters, de manera que los elementos dentro de un mismo grupo sean más similares entre sí que con los elementos de otros grupos. El objetivo principal del clustering es descubrir patrones y estructuras ocultas en los datos, sin necesidad de etiquetas previas.

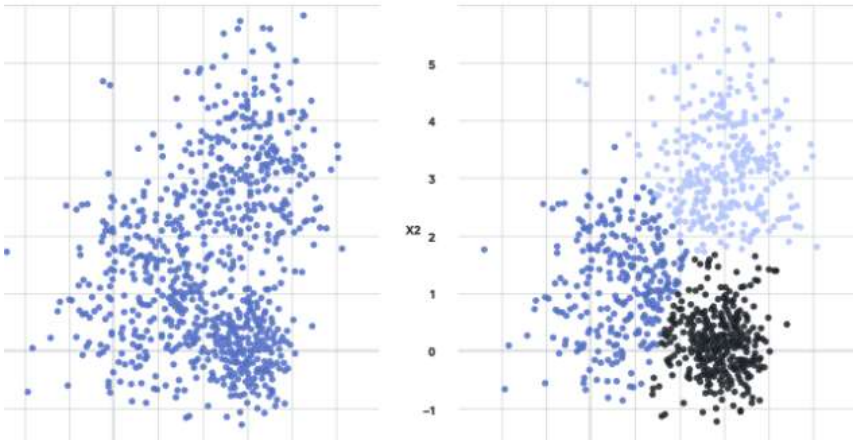


Figura 7.1: Ejemplo agrupación de clustering.

El clustering es una técnica de análisis de datos que se encarga de la organización de un conjunto de objetos en subgrupos, o clusters, en función de su similitud. En el corazón de esta técnica está el concepto de distancia o similitud entre los datos: los objetos que son más similares se agrupan en el mismo cluster, mientras que los que son diferentes quedan en clusters separados.

Esta metodología es esencial en la inteligencia artificial para descubrir estructuras escondidas sin la necesidad de etiquetas previas, lo que facilita el descubrimiento de relaciones que no se ven a primera vista entre los datos. Por ejemplo, es muy útil para el clustering para identificar patrones de síntomas que caracterizan a subgrupos, lo que puede ser crucial para diagnósticos personalizados o para entender mejor la progresión de ciertas tendencias.

Similaridad o distancia

La efectividad del clustering depende en gran medida de la definición de "similitud" utilizada. Las medidas de similitud más comunes incluyen la distancia entre dos vectores dimensionales. Esta distancia se puede realizar de muchas maneras, siendo las más comunes las siguientes:

- **Distancia Euclídea**: Es la más común y se utiliza para calcular la distancia más corta entre dos puntos en un espacio euclidiano. Se define como la raíz cuadrada de la suma de las diferencias cuadradas entre las coordenadas de los puntos. Esta medida es intuitiva y efectiva en muchos escenarios, especialmente en espacios de baja dimensión.

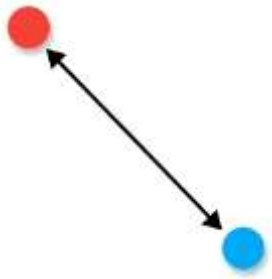


Figura 7.2: Distancia Euclídea.

- **Distancia de Manhattan**: Calcula la suma de las diferencias absolutas de las coordenadas de los puntos. Es útil en entornos de cuadrícula, como la planificación urbana, donde el movimiento solo puede ser horizontal o vertical, pero no diagonal.

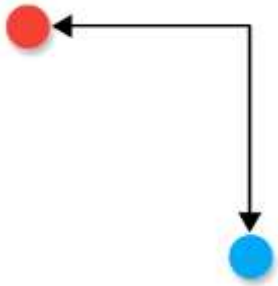


Figura 7.3: Distancia Manhattan.

- **Distancia de Minkowski**: Es una generalización de las distancias Euclídea y Manhattan. Dependiendo del valor del parámetro p (potencia), la distancia de Minkowski puede representar la distancia Euclídea (p=2) o la distancia de Manhattan (p=1). Permite una mayor flexibilidad y se puede adaptar a diferentes tipos de datos y estructuras.

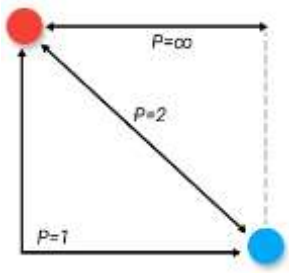


Figura 7.4: Distancia de Minkowski.

- **Distancia de Chebyshev**: Esta distancia es el máximo de las diferencias absolutas entre las coordenadas de los puntos. Es útil en aplicaciones donde el movimiento puede ser en cualquier dirección, pero solo un eje a la vez.

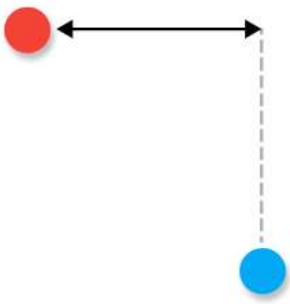


Figura 7.5: Distancia de Chebyshev.

Función objetivo

El clustering busca optimizar una función objetivo que cuantifique la cohesión intra-cluster y la separación inter-cluster. Esto significa que los elementos dentro de un cluster deben ser similares entre sí y diferentes de los elementos de otros clusters.

Algoritmos de clustering

Hay varios algoritmos de clustering, cada uno con sus propias suposiciones y enfoques. Algunos de los algoritmos más comunes son k-means, K-medoids, clustering jerárquico, DBSCAN y algoritmos basados en densidad, entre otros.

Inicialización

Muchos algoritmos de clustering requieren una inicialización, es decir, una selección inicial de los centroides o de los clusters. La calidad de la inicialización puede afectar significativamente los resultados finales.

Preprocesamiento de datos

El preprocesamiento de datos es crucial en el clustering. Esto puede incluir la normalización de características, la eliminación de valores atípicos, la reducción de dimensionalidad, etc.

Robustez y escalabilidad

Los algoritmos de clustering deben ser robustos frente a ruido en los datos y ser capaces de manejar conjuntos de datos grandes.

7.2.2. Métodos de Clustering

El clustering puede ser abordado a través de varios métodos, cada uno adaptado a diferentes tipos de datos y resultados deseados. Estos métodos se pueden clasificar en tres categorías principales:

- **Clustering por Particiones**: Este método divide el conjunto de datos en distintos grupos, y cada partición forma un cluster. El objetivo es realizar la división de tal manera que los datos dentro de cada cluster sean tan similares como sea posible, mientras que los datos de diferentes clusters sean lo más distintos posible. Algoritmos como K-means y K-medoids son ejemplos clásicos de clustering por particiones, donde cada cluster está representado por un centroide o un medoid.

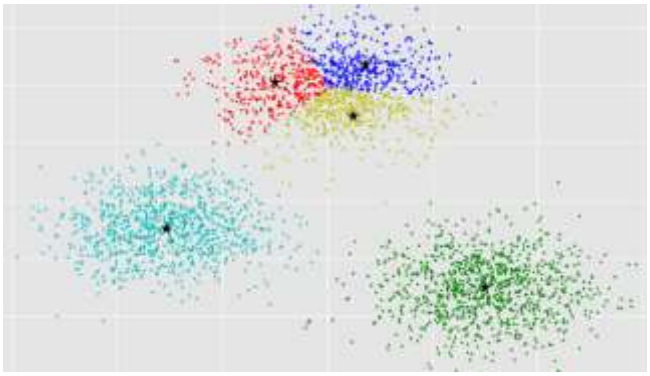


Figura 7.6: Ejemplo de clustering por particiones.

- **Clustering Jerárquico**: A diferencia del clustering por particiones, el método jerárquico organiza los datos en una estructura de árbol que representa las relaciones de jerarquía entre los datos. Puede ser aglomerativo, iniciando con cada dato como un cluster individual y fusionando clusters en pasos sucesivos, o divisivo, comenzando con un solo cluster que se divide progresivamente. Este enfoque es útil cuando se desea entender la relación de subgrupos dentro de los clusters más grandes.

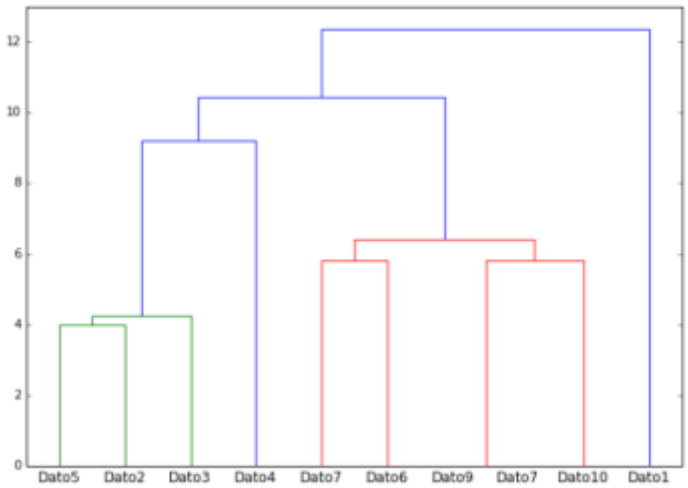


Figura 7.7: Ejemplo de clustering jerárquico.

- **Métodos Basados en Densidad**: Estos métodos, como DBSCAN, identifican clusters como áreas de alta densidad separadas por áreas de baja densidad. Son especialmente útiles cuando los clusters son irregulares o intercalados, y cuando los datos contienen ruido y outliers.

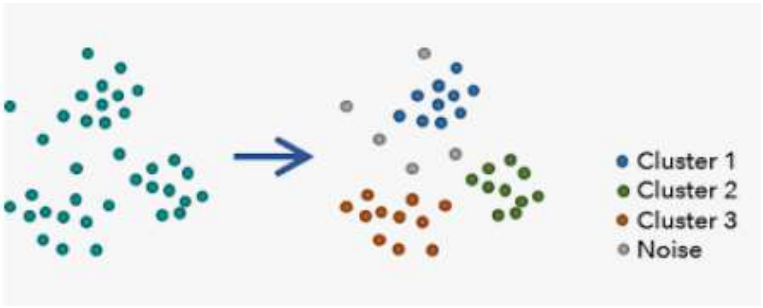


Figura 7.8: Ejemplo de clustering basado en densidad.

Cada uno de estos métodos tiene sus fortalezas y limitaciones, y la elección entre ellos depende del conjunto de datos específico y del contexto del problema.

7.2.3. Elección del Método de Clustering en el Proyecto

Para este proyecto específico, se optó por el método de clustering por particiones debido a una serie de razones estratégicas y prácticas. Primero, la naturaleza de los datos utilizados sugiere una distribución donde la formación de grupos naturales es clara, lo que favorece la aplicación de métodos por particiones como K-means y K-medoids. Estos algoritmos son efectivos en identificar centroides o medoids que actúan como representantes de los clusters, facilitando la interpretación y análisis subsiguiente.

Además, la experiencia previa con estos métodos proporcionó una base sólida para su implementación y ajuste. La simplicidad relativa del enfoque por particiones también permite una ejecución más directa y una menor complejidad computacional comparada con otros métodos como el clustering jerárquico o basado en densidad.

7.3. Clustering por particiones

El clustering por particiones es un enfoque para agrupar un conjunto de datos en subconjuntos, o "particiones", donde cada punto de datos pertenece a exactamente uno de los grupos. Este enfoque se basa en dividir el conjunto de datos en un número predeterminado de clusters o grupos, sin superposición entre ellos.

Se comienza con el proceso de Inicialización, donde se selecciona el número de clusters (k) que deseas crear. La posibilidad de utilizar el método Elbow (Método del codo) para determinar el número óptimo de clusters, es otra ventaja muy importante de los métodos por particiones. Esta técnica permite una evaluación más intuitiva de la homogeneidad dentro de los clusters y facilita la elección de un número de clusters que equilibra la precisión del modelo

con la generalización.

Posteriormente, se realiza la asignación de puntos a clusters. En esta fase se asignan aleatoriamente los puntos de datos a los k clusters iniciales. En muchos algoritmos de clustering por particiones, se utiliza una medida de distancia para determinar qué puntos de datos son más similares entre sí y, por lo tanto, deben pertenecer al mismo cluster.

Una vez realizada la asignación de puntos a cada cluster, se ejecuta la actualización de centroides, recalculando el centroide de cada cluster. Los centroides son los puntos "promedio" de los datos asignados a un cluster específico. La forma en que se calcula el centroide depende del algoritmo de clustering específico que se esté utilizando. En este proyecto se han estudiado los algoritmos K-Means y K-Medoids, eligiendo el segundo por su robustez ante la presencia de outliers, siendo éstos los datos que son bastantes diferentes al conjunto de datos, dado que si utilizamos el algoritmo K-Means pueden salir distorsionados los centroides obtenidos.

Cuando se tienen actualizados los centroides, se vuelven a asignar los puntos de datos a los clusters en función de su proximidad al centroide más cercano. Esto puede implicar recalcular las distancias entre los puntos de datos y los centroides actualizados y asignar cada punto al cluster con el centroide más cercano.

La actualización de centroides y la reasignación de los datos a los clusters en función de los nuevos centroides se repite iterativamente hasta que se cumpla algún criterio de detención, como puede ser que la convergencia de los centroides ya no cambie significativamente entre iteraciones, o cuando se cumpla un número máximo de iteraciones.

7.3.1. Método Elbow

La determinación del número K de clusters a utilizar es una tarea delicada, dado que si dicha determinación se realiza erróneamente puede verse alterada la agrupación de datos, siendo esta poco precisa, afectando esto al resultado obtenido.

Existen varias técnicas para realizar este cálculo, no siendo ninguna de ellas la técnica óptima que permita determinar de manera exacta el número de clusters a utilizar. Entre las citadas técnicas, tenemos el método Método de la silueta (Silhouette Method), Método de la densidad (Density-Based Method), Método de la validación interna (Internal Validation Method), Método del codo (Elbow Method), etc, que nos dan una aproximación del número de clusters a utilizar.

Para el desarrollo del presente proyecto, se ha elegido el Método del codo (Elbow Method) por la experiencia de uso con él. Este Método busca encontrar el punto en el que se observa un cambio brusco en la pendiente de la gráfica de la suma de las distancias intra-cluster versus el número de clusters. Este punto indica el número óptimo de clusters donde se obtiene el mejor equilibrio entre la cohesión intra-cluster y la simplicidad del modelo.

El Método del Codo se lleva a cabo siguiendo estos pasos:

1. Ejecutar el algoritmo de clustering (por ejemplo, K-Means) en el conjunto de datos para diferentes valores de k, donde k es el número de clusters que se desea probar.

2. Calcular la suma de las distancias cuadradas intra-cluster para cada valor de k. Esto implica calcular la distancia de cada punto al centroide de su cluster y sumar estas distancias para todos los puntos en el cluster.

3. Realizar la gráfica de la suma de las distancias cuadradas intra-cluster en función de los diferentes valores de k.

4. Identificar el "codo" en el gráfico. El codo es el punto en el gráfico donde la tasa de disminución de la suma de las distancias intra-cluster comienza a disminuir drásticamente. Visualmente, este punto suele ser donde la curva comienza a aplanarse significativamente.

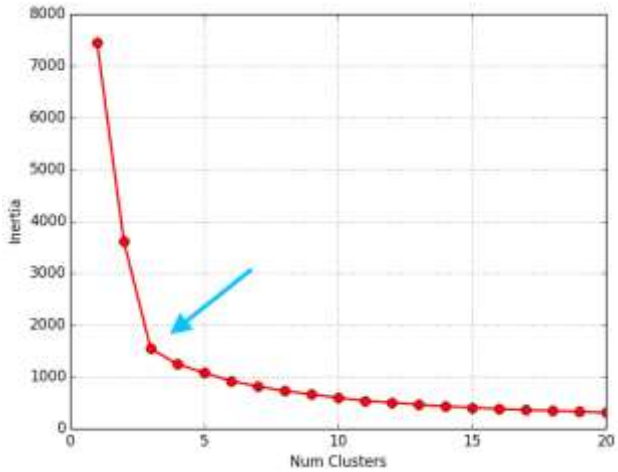


Figura 7.9: Ejemplo de gráfica del Método del Codo.

Como se puede ver en la figura, el punto de inflexión sería el número de clusters a tener en cuenta para el conjunto de datos.

7.3.2. K-means

K-means es un algoritmo de optimización que busca minimizar la suma de las distancias cuadradas entre los puntos de datos y el centroide más cercano, asignando cada punto al

centroide más próximo y luego recalculando los centroides basados en los puntos asignados. Este proceso se repite iterativamente hasta que se cumple un criterio de convergencia, como puede ser un número máximo de iteraciones o una variación mínima en los centroides entre iteraciones sucesivas.

Es particularmente efectivo cuando se trata de grandes volúmenes de datos debido a su simplicidad y rapidez computacional. Sin embargo, uno de sus inconvenientes es que funciona mejor con clusters de forma esférica y similar tamaño. Además, el resultado final puede verse afectado por la elección inicial de centroides, lo que a veces requiere múltiples ejecuciones con diferentes inicializaciones para obtener un resultado óptimo.

Determinación del número k de clusters

La determinación del Número K de clusters se realizará con cualquiera de las técnicas desarrolladas para ello, siendo aplicada en el presente proyecto el Método del Codo, tal y como se he explicado en el apartado anterior.

Determinación de centroides

Teniendo en cuenta el número k de clusters obtenido anteriormente, se determinará un centroide Ci para cada uno de estos grupos, eligiendo los valores que minimicen la función objetivo:

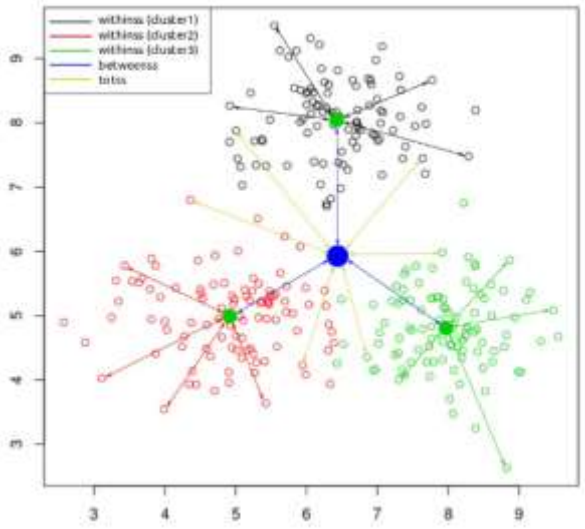


Figura 7.10: Ejemplo de determinación de centroides para cada cluster.

Proceso de aprendizaje

En esta etapa, cada punto de datos se asigna a uno de los cluster cuyo centroide es el más cercano. Esto se basa en alguna medida de distancia vistas anteriormente. Después de asignar todos los puntos a los clusters, los centroides se recalculan tomando la media de todos los puntos asignados a ese cluster. Estos pasos se repiten iterativamente hasta que los centroides no cambien significativamente entre iteraciones o hasta que se alcance un criterio de detención predefinido, como un número máximo de iteraciones.

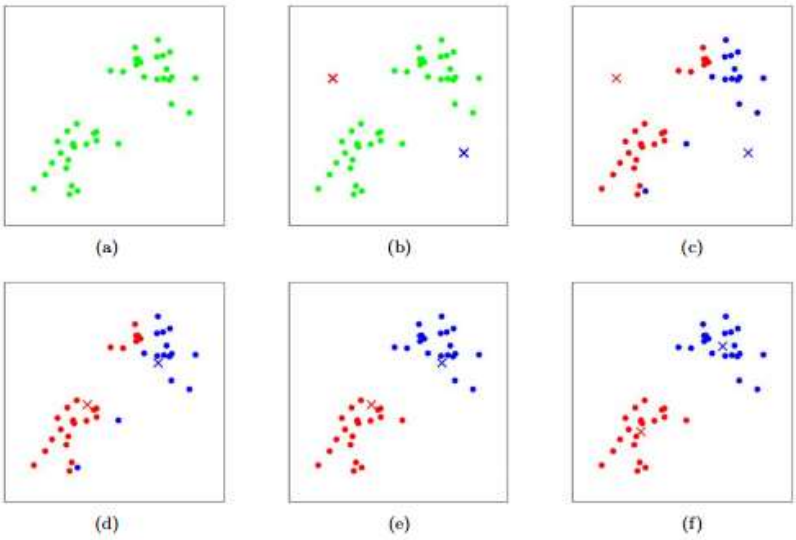


Figura 7.11: Ejemplo de proceso de iteraciones en K-means.

7.3.3. K-medoids

A diferencia de K-means, K-medoids selecciona puntos de datos reales como representantes centrales, o medoids. Esta característica hace que el algoritmo sea más robusto frente a los outliers, ya que un outlier es menos probable que sea elegido como un medoid en comparación con afectar la posición de un centroide en K-means.

Determinación del número k de clusters

La determinación del Número K de clusters se realizará de igual manera que se ha realizado para el algoritmo K-means, es decir, siendo aplicada la técnica del Método del Codo, tal y como se he explicado anteriormente.

Proceso de aprendizaje

El algoritmo más conocido para implementar K-medoids es el Partitioning Around Medoids (PAM). PAM comienza con un conjunto inicial de medoids y luego itera, tratando de mejorar el modelo, intercambiando medoids con no-medoids si mejora el coste total, que se mide como la suma de las distancias entre cada punto y el medoid más cercano. Este proceso se repite hasta que no se puede hacer ninguna mejora.

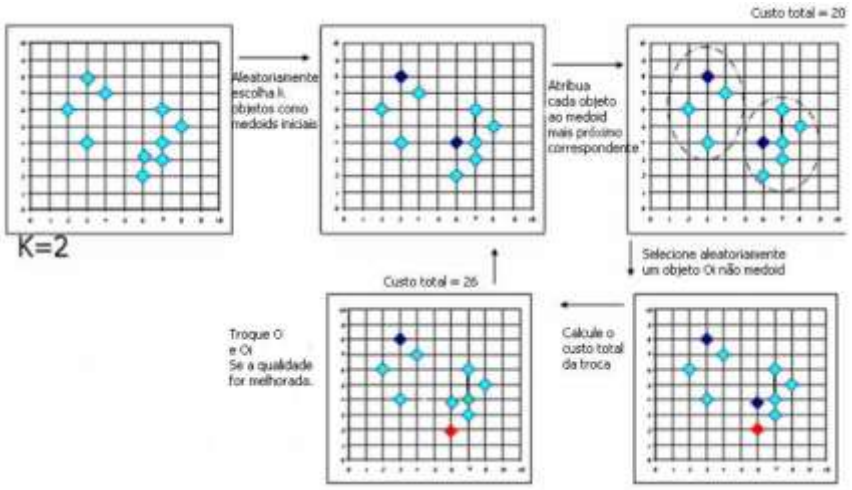


Figura 7.12: Ejemplo de proceso de iteraciones en K-medoids.

7.3.4. Desventajas de K-means y K-medoids

Ambos algoritmos tienen desventajas. La sensibilidad a la elección de los centroides o medoids iniciales puede llevar a resultados nada óptimos en K-means y K-medoids. Además, ambos asumen que la especificación previa del número de clusters, lo cual no siempre es evidente en conjuntos de datos reales.

También, ambos algoritmos tienen dificultades con clusters de formas no esféricas y tamaños variables. Los métodos son propensos a encontrar clusters locales óptimos en lugar de globales, lo que significa que pueden necesitar varias ejecuciones con diferentes inicializaciones para obtener una solución satisfactoria.

En resumen, la elección entre K-means y K-medoids depende de la naturaleza de los datos y los requisitos específicos del análisis. Para datos con outliers y cuando se desea una representación más realista de los centros de los clusters, K-medoids puede ser la mejor opción. Para conjuntos de datos grandes y bien comportados donde la velocidad es una prioridad, K-means puede ser más adecuado.

7.4. Construcción de los algoritmos

El objetivo de aplicar las distintas técnicas de clustering usadas en el presente proyecto es conocer si la edad de los opositores condiciona de alguna manera cada una de las pruebas que se realizan durante el proceso de oposición, por ello, se realizará la construcción e implementación de las distintas técnicas de clustering a aplicar en el proyecto.

A través de la plataforma on-line Colab de Google, se ha realizado la construcción en lenguaje Python de los distintos métodos. Para el tratamiento, análisis de datos y aprendizaje automático se han hecho uso de las bibliotecas matplotlib, sklearn y pandas.

7.4.1. Proceso de Normalización

El proceso de normalización es importante para mejorar el rendimiento de ciertos algoritmos de aprendizaje automático, ayudando a evitar que las características de los datos a tratar con escalas muy diferentes dominen la influencia en el modelo. La normalización de características también puede ser útil para visualizaciones de datos, donde tener todas las características en la misma escala facilita la interpretación.

Por este motivo, se hará uso de la función MinMaxScaler(). Dicha función es utilizada en análisis y aprendizaje automático de datos para realizar la normalización de los mismos. Su objetivo es transformar los datos para que estén dentro de un rango específico, en nuestro caso entre 0 y 1, siendo aplicado a cada uno de los ficheros con formato csv que forman nuestro data dataset.

La interpretación matemática de la función MinMaxScaler() es la siguiente, la cual funciona calculando y aplicando la siguiente transformación a cada característica:

Donde:

- **Xscaled**: Es el valor escalado.
- **X**: es el valor original de la característica.
- **Xmin**: Es el valor mínimo de la característica en el conjunto de datos.
- **Xmax**: Es el valor máximo de la característica en el conjunto de datos.

Este método asegura que el valor mínimo se transforme a 0 y el valor máximo se transforme a 1, mientras que los otros valores se escalan proporcionalmente entre estos dos extremos.

En Python, la función MinMaxScaler() es proporcionada por la biblioteca scikit-learn y se utiliza comúnmente en tareas de preprocesamiento de datos.

Este proceso de normalización de datos, se realiza con cada uno de los ficheros en formato csv que hemos preparado para su análisis,

```
scaler = MinMaxScaler()
normalizado = scaler.fit_transform(data.iloc[:, [1, 2]])

print(normalizado)
```

Figura 7.13: Código de normalización de datos.

Tal y como se puede observar en la figura 7.8, en la variable scaler se almacena la aplicación MinMaxScaler(), creando posteriormente un dataset nuevo llamado "normalizado", donde se almacenan los datos normalizados entre 0 y 1.

7.4.2. Método Elbow

Como ya se ha explicado anteriormente, para determinar el número de k clusters recomendado se realiza mediante el método elbow con el cáculo y aplicación de la inercia. Se utiliza un rango de 1 a 9 clusters para la obtención de dicho número k. Los resultados obtenidos en cada iteración del bucle for correspondientes a los tipos de clustering utilizados en el presente proyecto, Kmeans y Kmedoids, se guarda en la variable wcss, la cual es utilizada posteriormente para la representación gráfica.

A continuación, se muestra el código para el modelo K-means:

```
wcss = []
for i in range(1, 10):
    kmeans = KMeans(n_clusters=i, init='k-means++', max_iter=300, n_init=10, random_state=0)
    kmeans.fit(normalizado)
    wcss.append(kmeans.inertia_)

# Creación de una gráfica utilizando Matplotlib para visualizar el método elbow
plt.plot(range(1, 10), wcss)
plt.title('Método del codo - K-means')
plt.xlabel('Número de clusters')
plt.ylabel('Inertia')
plt.show()
```

Figura 7.14: Código determinación k clusters con K-means.

Seguidamente, se muestra el código para el modelo K-medoids:

```
wcss = []
for i in range(1, 10):
    kmedoids = KMedoids(n_clusters=i, init='k-medoids++', max_iter=300, random_state=0)
    kmedoids.fit(normalizado)
    wcss.append(kmedoids.inertia_)

# Creación de una gráfica utilizando Matplotlib para visualizar el método elbow
plt.plot(range(1, 10), wcss)
plt.title('Método del codo - K-medoids')
plt.xlabel('Número de clusters')
plt.ylabel('Inertia')
plt.show()
```

Figura 7.15: Código determinación k clusters con K-medoids.

Se aplican a cada uno de los ficheros en formato csv que forman nuestro dataset ambos métodos, obteniéndose, tal y como se puede apreciar en el las siguientes figuras, que en ambos modelos de clustering se determina como número ideal de cluster 3, apreciándose de forma más clara con el modelo K-means.

- **Fichero archivoFisicas.csv:**

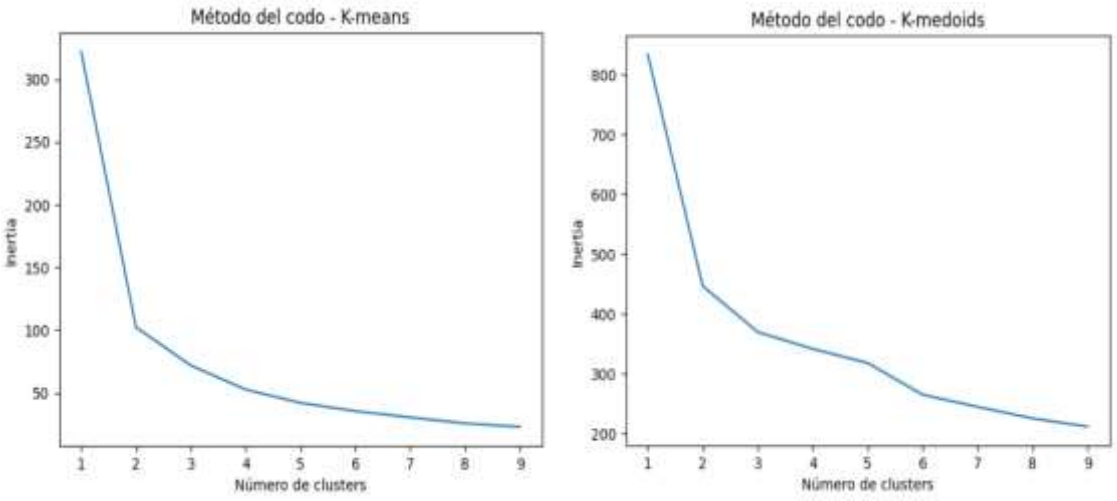


Figura 7.16: Código determinación k clusters con K-means y K-medoids para el archivo de Pruebas Físicas.

- **Fichero archivoTeoria.csv:**

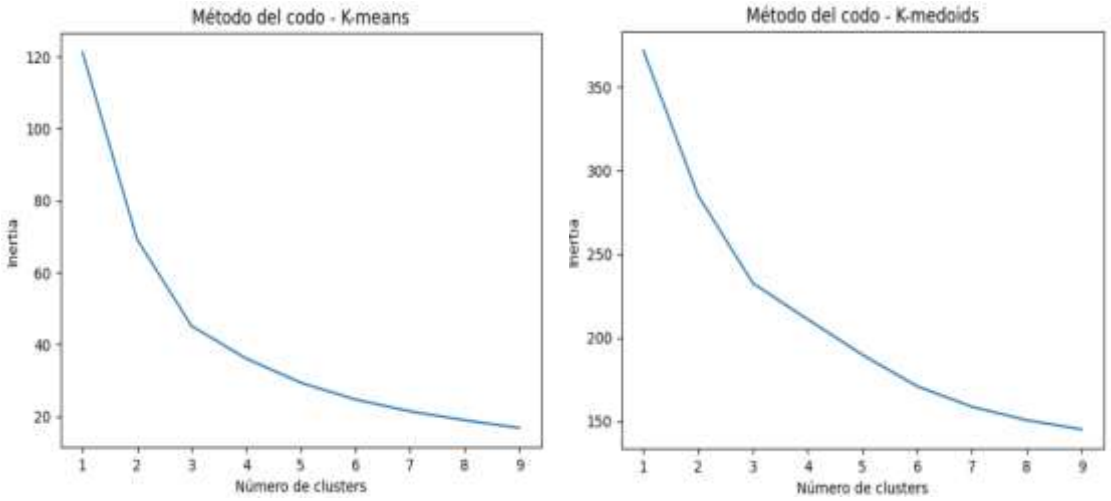


Figura 7.17: Código determinación k clusters con K-means y K-medoids para el archivo de Prueba de Teoría.

- **Fichero archivoOrto.csv:**

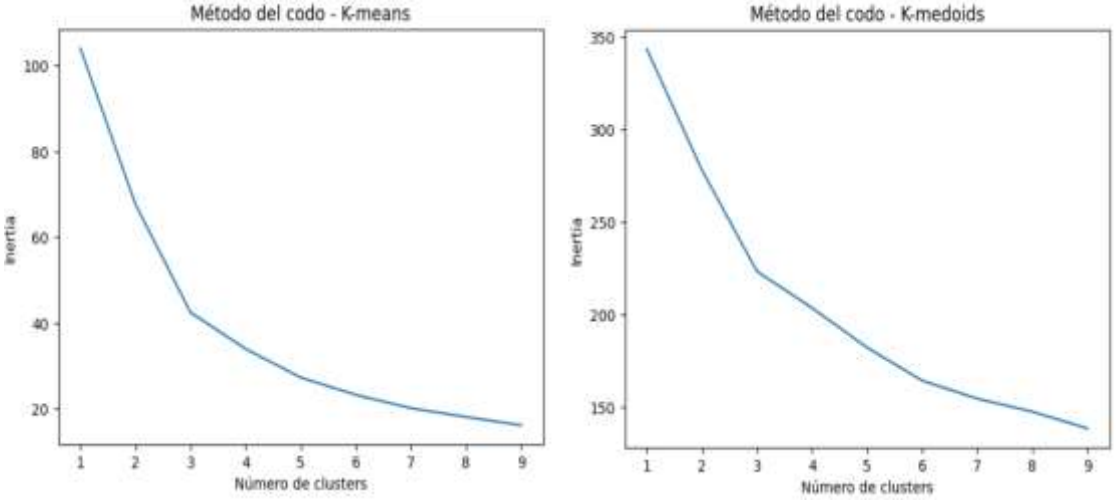


Figura 7.18: Código determinación k clusters con K-means y K-medoids para el archivo de Prueba de Ortografía.

- **Fichero archivoTeoriaOrto.csv:**

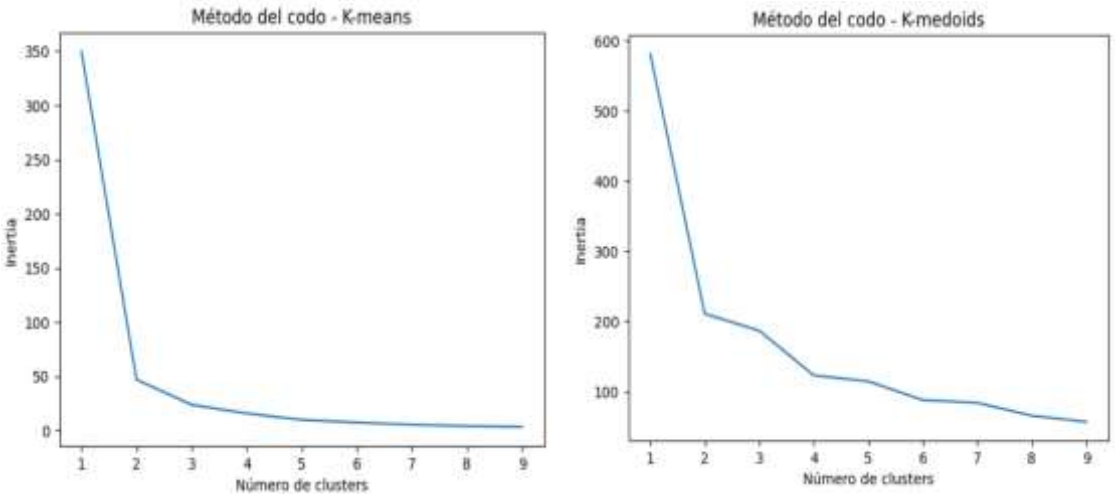


Figura 7.19: Código determinación k clusters con K-means y K-medoids para el archivo de Pruebas de Teoría y Ortografía conjunta.

- **Fichero archivoMedico.csv:**

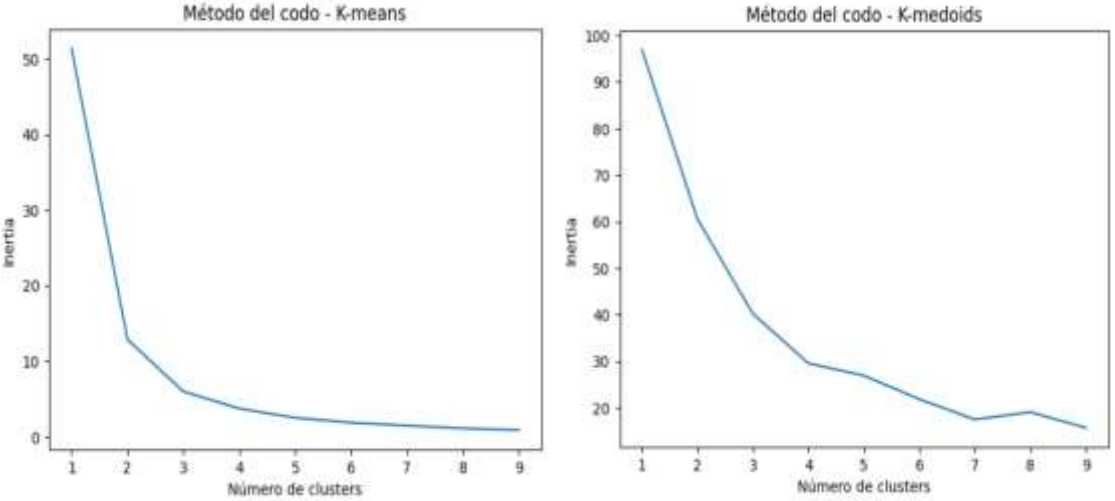


Figura 7.20: Código determinación k clusters con K-means y K-medoids para el archivo de Prueba Reconocimiento Médico.

- **Fichero archivoEntrevista.csv:**

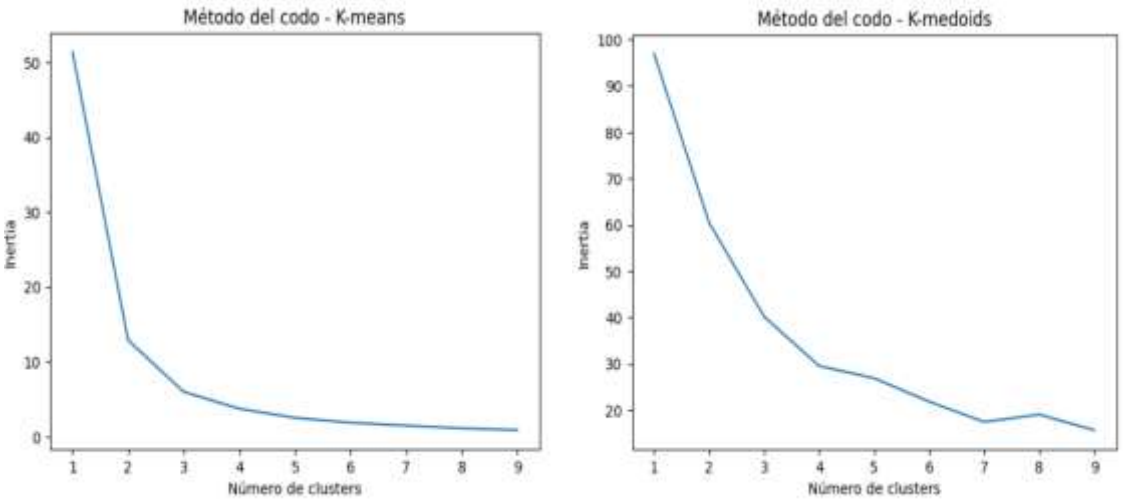


Figura 7.21: Código determinación k clusters con K-means y K-medoids para el archivo de Prueba Entrevista.

- **Fichero archivoPsico.csv:**

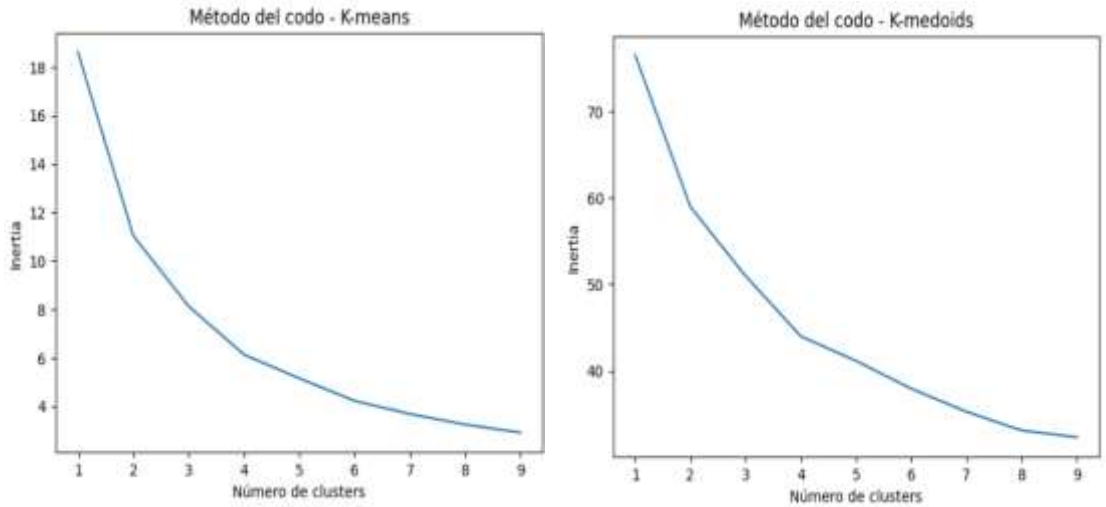


Figura 7.22: Código determinación k clusters con K-means y K-medoids para el archivo de Prueba Psicotécnicos.

- **Fichero archivoIdioma.csv:**

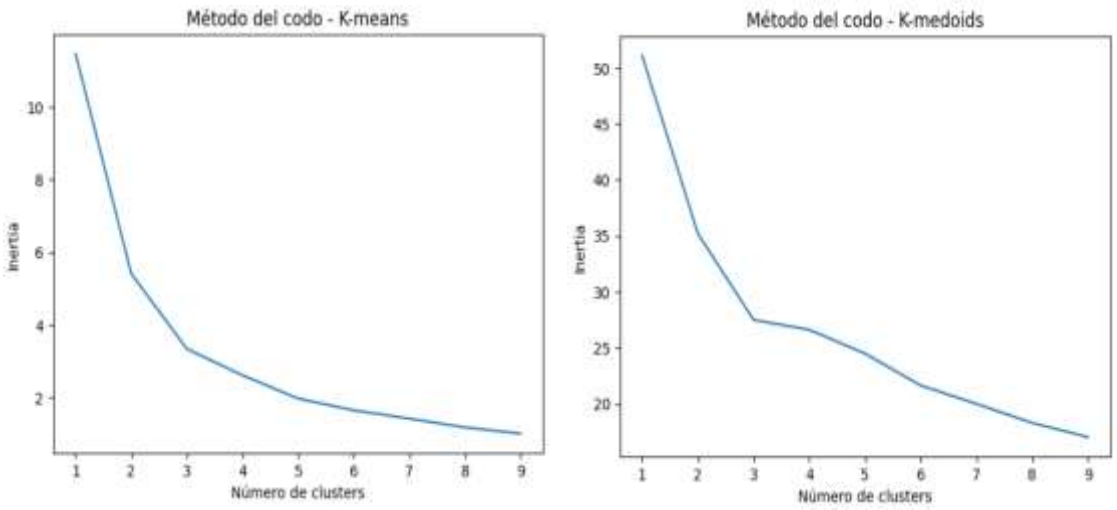


Figura 7.23: Código determinación k clusters con K-means y K-medoids para el archivo de Prueba Idioma.

- **Fichero archivoAptosNota.csv:**

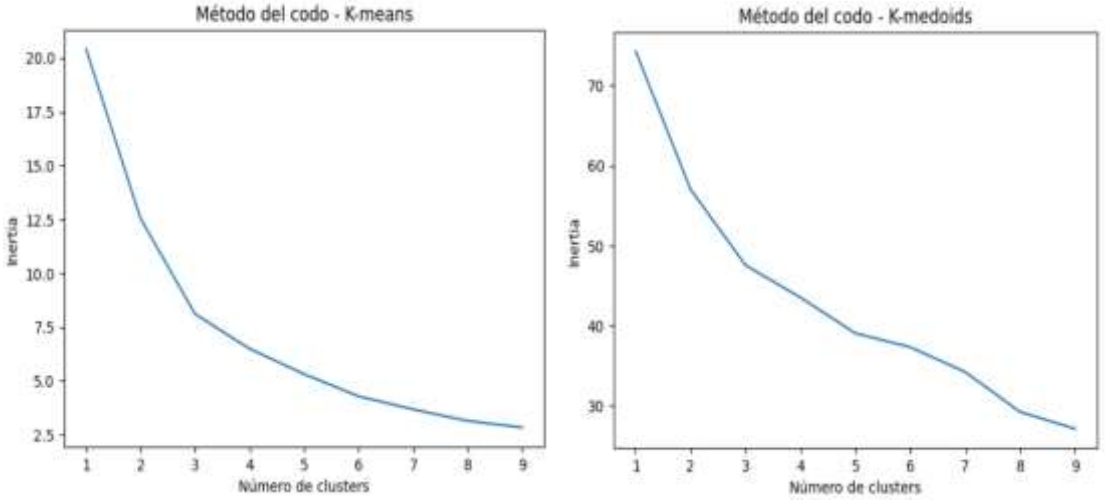


Figura 7.24: Código determinación k clusters con K-means y K-medoids para el archivo de Aptos Nota.

- **Fichero archivoAptosEsca.csv:**

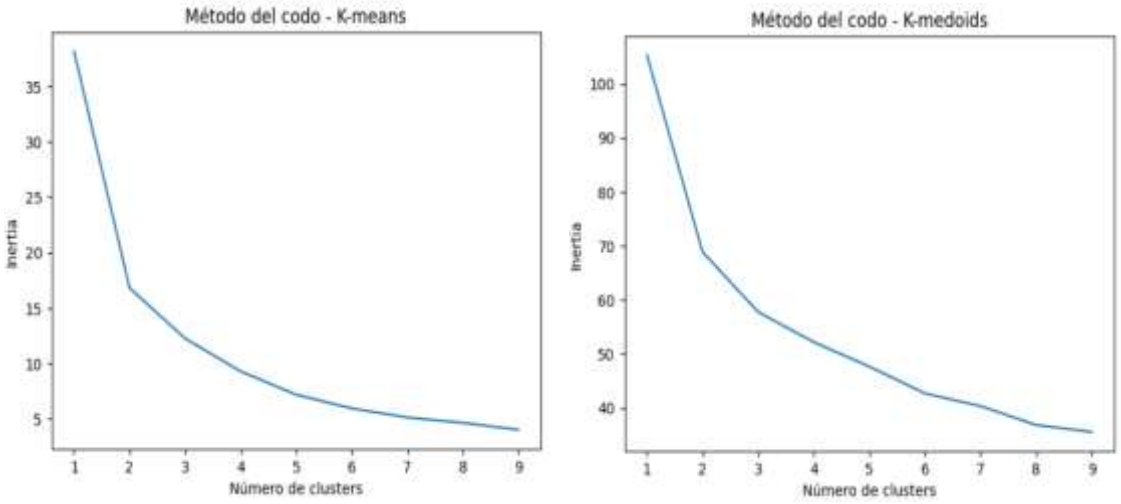


Figura 7.25: Código determinación k clusters con K-means y K-medoids para el archivo de Escalafón.

7.4.3. Centroides en K-means

Se ha tomado la decisión de utilizar el algoritmo K-means porque, además de que el número k de clusters se aprecia mucho mejor con este algoritmo, los datos no tienen muchos outliers, siendo las variables continuas y numéricas, y asumiendo que los clusters son aproximadamente esféricos

Seguidamente se aplica el método K-means al conjunto de datos con el objetivo de mostrar las coordenadas dadas a cada centroide, realizando este proceso en cada uno de los ficheros csv que forman nuestro dataset.

```
kmeans = KMeans(n_clusters=3, init='k-means++', max_iter=300, n_init=10, random_state=0).fit(normalizado)
centroids = kmeans.cluster_centers_
print(centroids)

#Este código realiza clustering utilizando el algoritmo K-Means en los datos normalizados
#ubicándolos en el espacio
```

```
[[0.4799689  0.61213913]
 [0.31190032 0.01651987]
 [0.21141077 0.6353154 ]]
```

Figura 7.26: Código determinación de la ubicación espacial de los centroides en K-means.

7.4.4. Asignación de datos a cada centroide en K-means

Una vez obtenida la ubicación en el espacio de los centroides, mediante un gráfico de dispersión, un gráfico de dispersión por colores y mediante la función pairplot se puede visualizar la asignación realizada de cada dato del dataset a cada centroide.

A continuación, se detalla el código de cada uno de los tres métodos utilizados para dicha visualización:

- **Código del gráfico de dispersión de las muestras y sus centroides**: Un gráfico de dispersión "scatter plot" es una representación gráfica que utiliza coordenadas cartesianas para mostrar los valores de dos variables para un conjunto de datos. Este tipo de gráfico ayuda a identificar si existe una relación lineal, curvilínea u otra entre las variables, siendo una herramienta clave en el análisis de datos, permitiendo visualizar patrones o tendencias en los datos que podrían no ser obvios en tablas de datos. En el contexto de K-means nos facilita la visualización de cómo los puntos de datos se agrupan en diferentes clusters. Cada punto en el gráfico representa un par de valores. Si los puntos tienden a subir hacia la derecha, hay una relación positiva, es decir, cuando una variable aumenta la otra también. Si los puntos tienden a bajar hacia la derecha, al contrario que en el caso anterior, hay una relación negativa, es decir, cuando una variable aumenta la otra disminuye. Si los puntos están dispersos sin una dirección clara, puede no haber una relación lineal aparente entre las variables.

```
kmeans = KMeans(n_clusters=3, init='k-means++', max_iter=300, n_init=10, random_state=0)
pred_y = kmeans.fit_predict(normalizado)

plt.scatter(normalizado[:,0], normalizado[:,1])
plt.scatter(kmeans.cluster_centers_[:, 0], kmeans.cluster_centers_[:, 1], s=300, c='red')
plt.xlabel('EDAD')
plt.ylabel('FÍSICAS')
plt.title('Clusters')

cluster_data = pd.DataFrame({'EDAD': data['EDAD'], 'FÍSICAS': normalizado[:,1], 'Cluster': pred_y})
```

Figura 7.27: Ejemplo de código de asignación de datos a cada centroide en K-means mediante gráfico de dispersión.

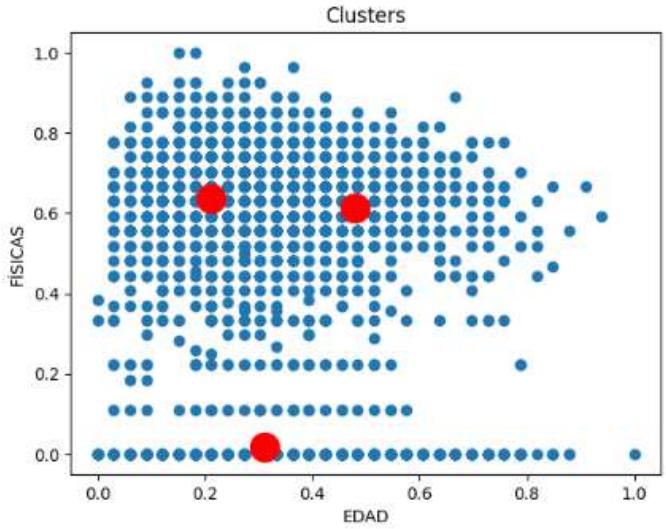


Figura 7.28: Ejemplo de visualización de asignación de datos a cada centroide en K-means mediante gráfico de dispersión.

- **Código del gráfico de dispersión por colores de las muestras y sus centroides**: Este código, además de permitir lo mismo que el del apartado anterior al representar un gráfico de dispersión, permite visualizar que cada cluster se representa con un color diferente, facilitando la distinción visual de los diferentes grupos formados por el algoritmo K-means y permitiendo observar cómo el algoritmo ha segmentado los datos.

```
colors = ['blue', 'green', 'red', 'purple', 'orange', 'gray']

k = len(np.unique(pred_y))

for i in range(k):
    cluster_i = cluster_data[cluster_data['Cluster'] == i]
    plt.scatter(cluster_i['EDAD'], cluster_i['FÍSICAS'], color=colors[i])
plt.xlabel('EDAD')
plt.ylabel('FÍSICAS')
plt.title('Clusters')
plt.show()
```

Figura 7.29: Ejemplo de código de asignación de datos a cada centroide en K-means por colores mediante gráfico de dispersión.

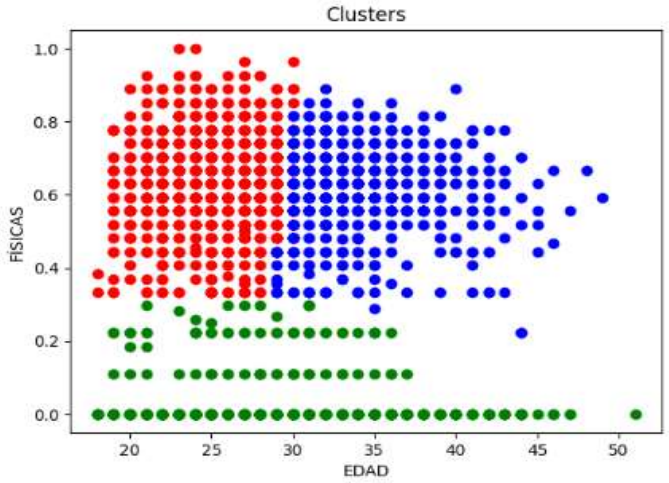


Figura 7.30: Ejemplo de código de visualización de asignación de datos a cada centroide en K-means por colores.

- **Código del gráfico de del método pairplot**: Sns.pairplot es una función de la librería seaborn que se utiliza para crear una matriz de gráficos de pares, de ahí su nombre "pair plots". Los gráficos de pares son útiles para visualizar las relaciones entre múltiples variables en un conjunto de datos y cómo estas relaciones varían entre diferentes grupos o clusters. Dicho gráfico de pares muestra relaciones emparejadas entre las variables de un DataFrame, facilitando la visualización de distribuciones y relaciones bivariadas. En resumen, estos gráficos permiten explorar visualmente cómo se estructuran los datos en relación con las variables, y cómo el algoritmo de clustering ha segmentado estos datos, pudiendo ser útil para validar la calidad del clustering y para entender mejor las características de cada grupo identificado.

```
sns.pairplot(cluster_data, hue='Cluster', vars=['EDAD', 'FÍSICAS'])
```

```
sns.pairplot(cluster_data, hue='Cluster', vars=['EDAD', 'Cluster'])
```

Figura 7.31: Código de asignación de datos a cada centroide en K-means método sairplot.

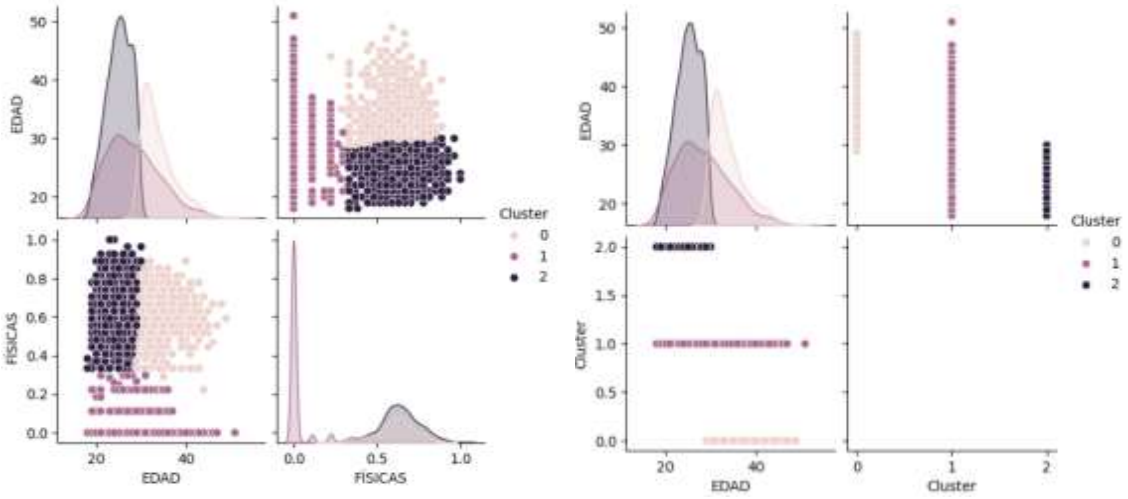


Figura 7.32: Ejemplo de visualización de asignación de datos a cada centroide en K-means método sairplot.

Este proceso se aplicará a todos los ficheros csv que forman nuestro dataset, obteniendo el resultado que se puede apreciar en las siguientes imágenes en función de cada archivo csv que forman el dataset del presente proyecto:

- **Fichero archivoFisicas.csv:**

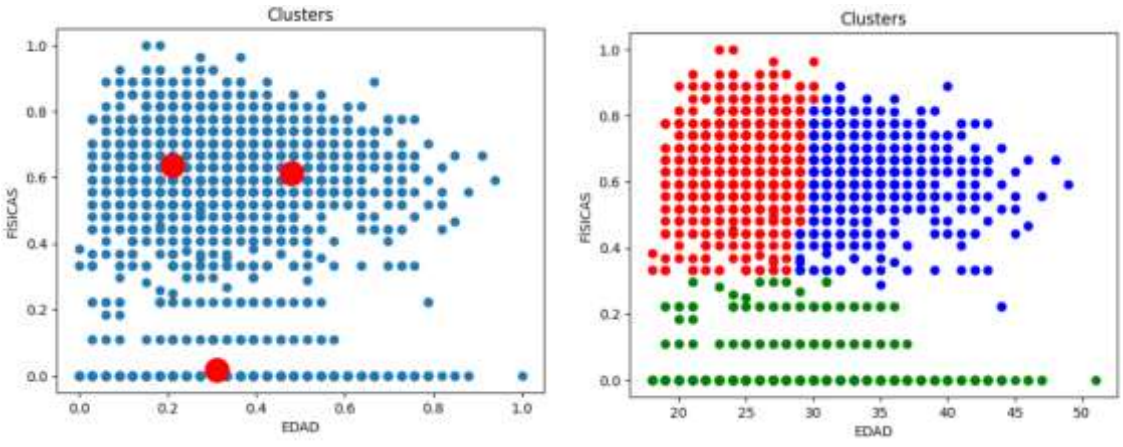


Figura 7.33: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Pruebas Físicas.

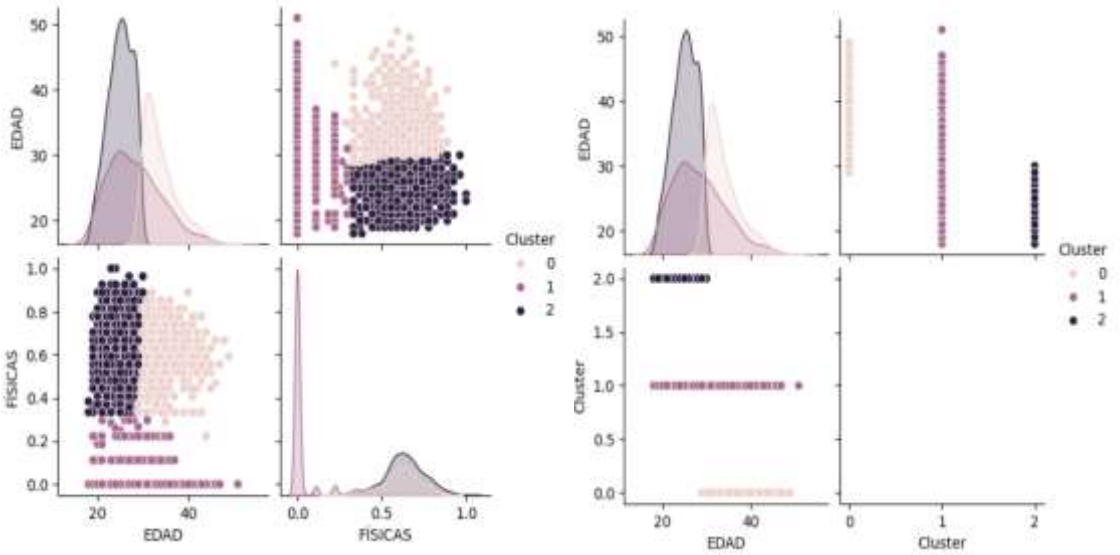


Figura 7.34: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Pruebas Físicas.

- **Fichero archivoTeoria.csv:**

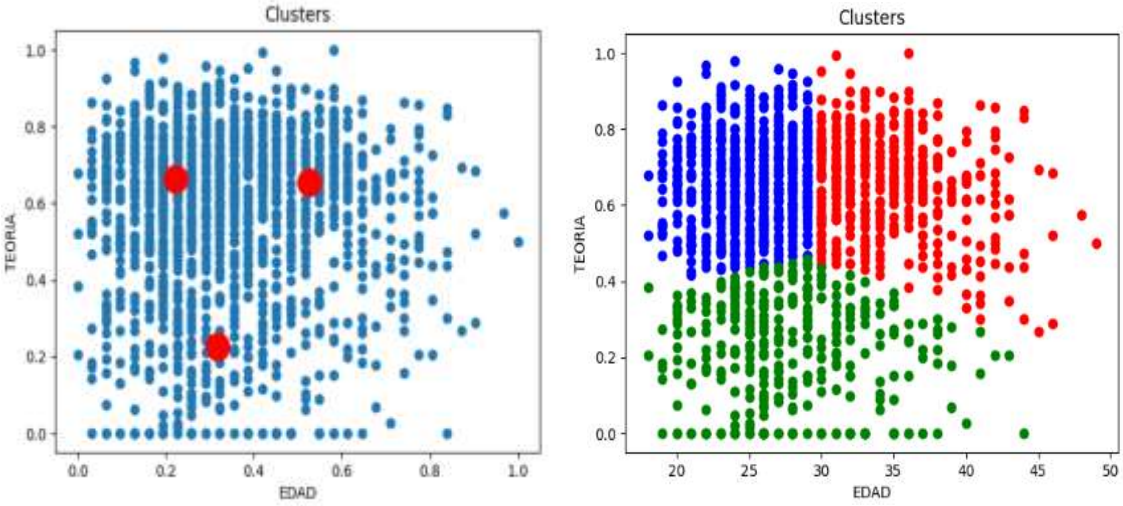


Figura 7.35: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Prueba Teórica.

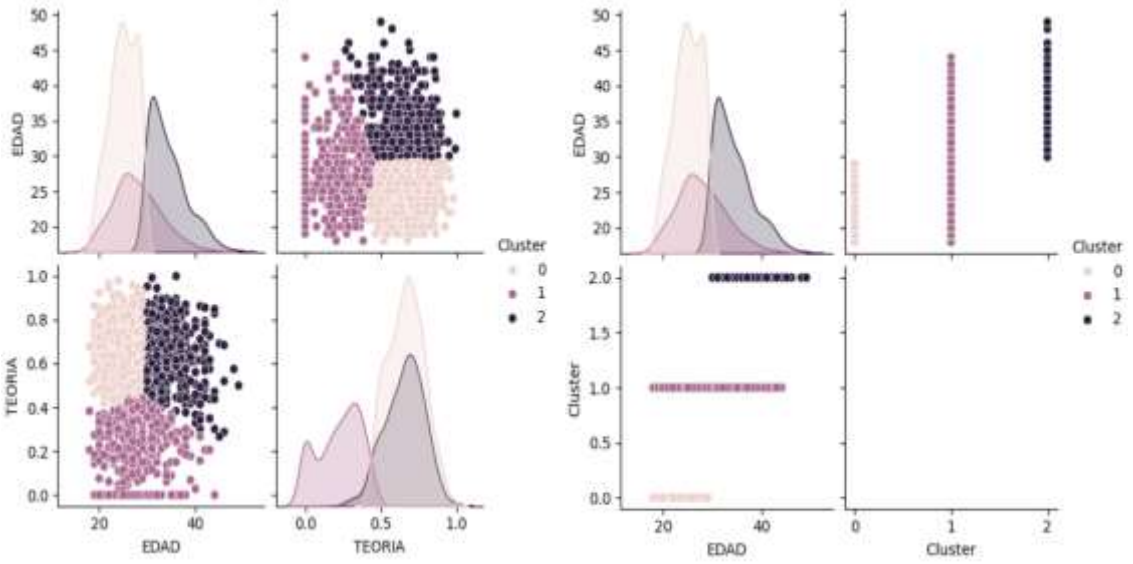


Figura 7.36: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Prueba Teórica.

- **Fichero archivoOrto.csv:**

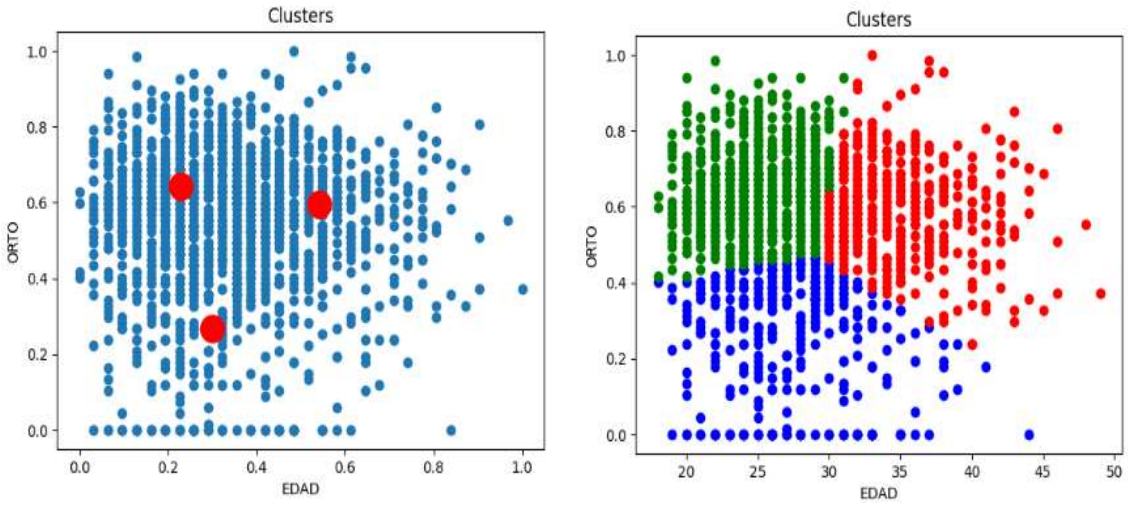


Figura 7.37: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Prueba de Ortografía.

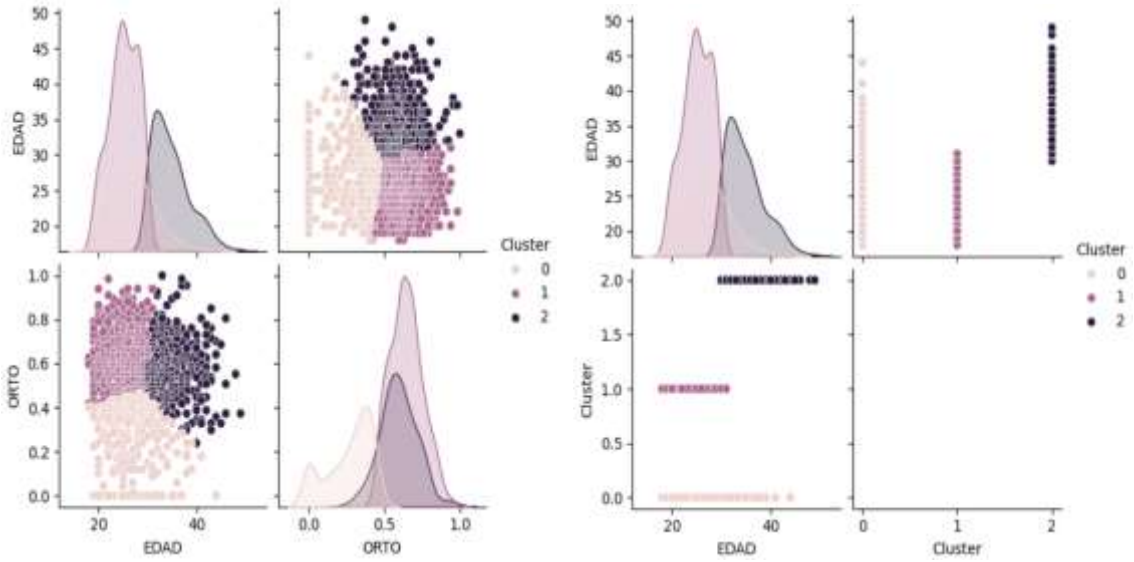


Figura 7.38: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Prueba de Ortografía.

- **Fichero archivoTeoriaOrto.csv:**

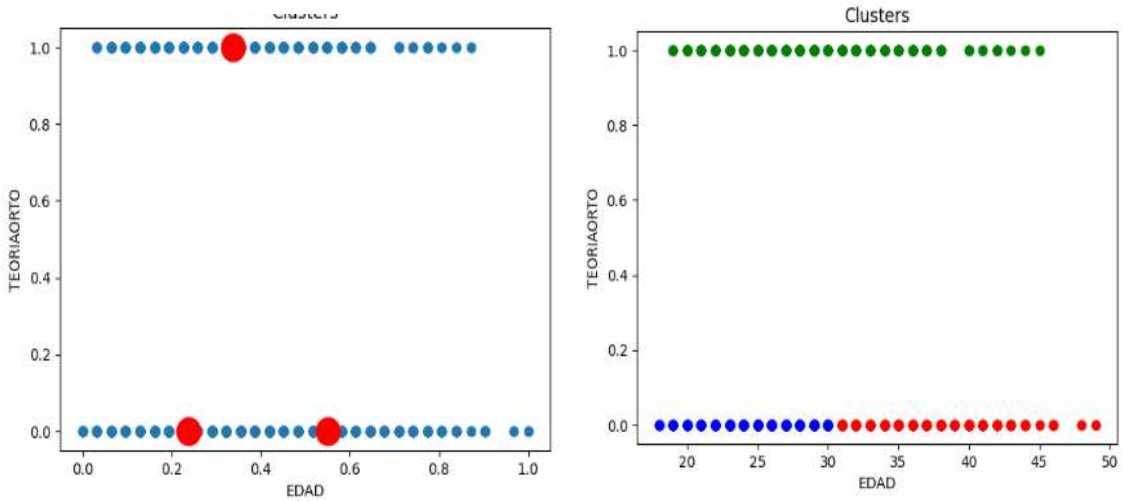


Figura 7.39: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Pruebas de Teoría de Ortografía conjunta.

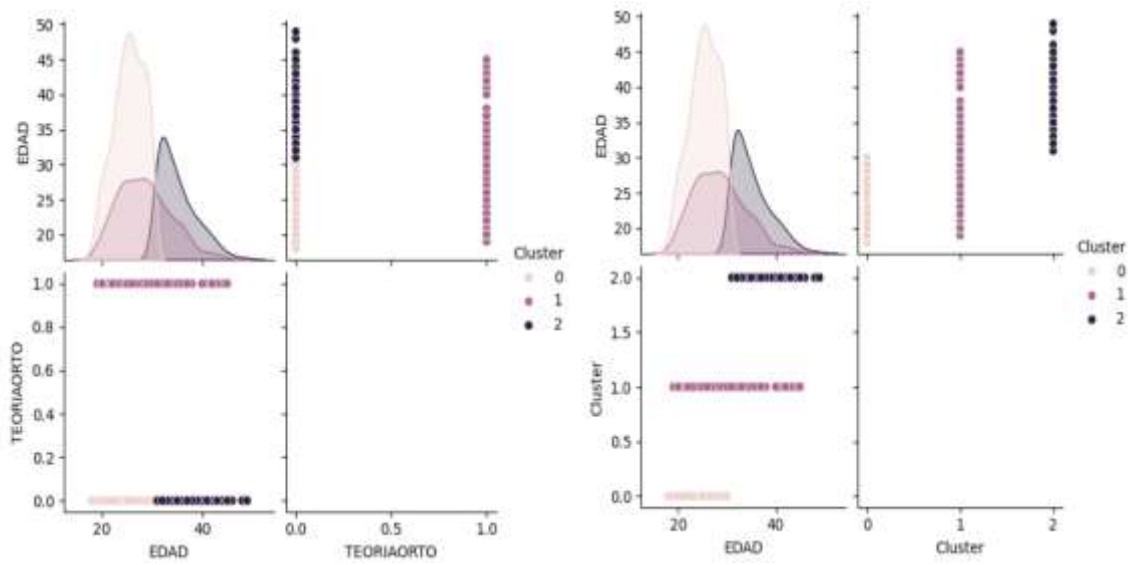


Figura 7.40: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Pruebas de Teoría de Ortografía conjunta.

- **Fichero archivoMedico.csv:**

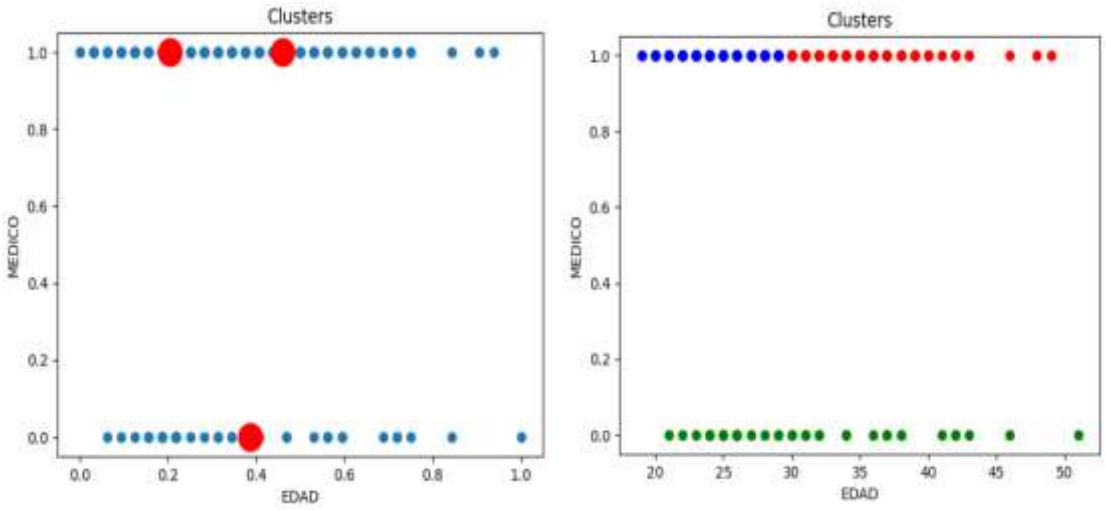


Figura 7.41: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Prueba de Reconocimiento Médico.

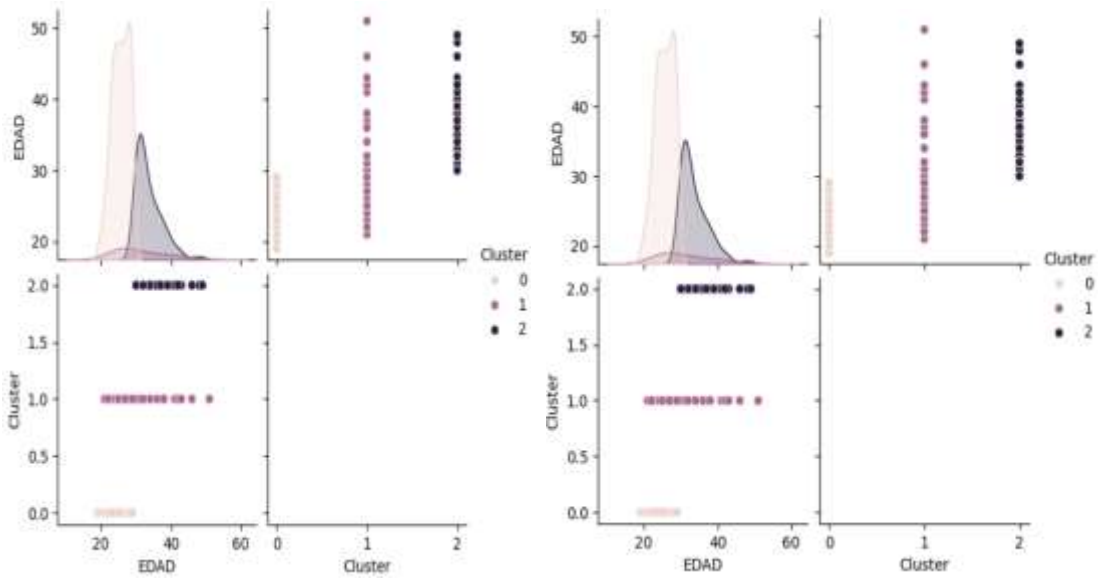


Figura 7.42: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Pruebas Reconocimiento Médico.

- **Fichero archivoEntrevista.csv:**

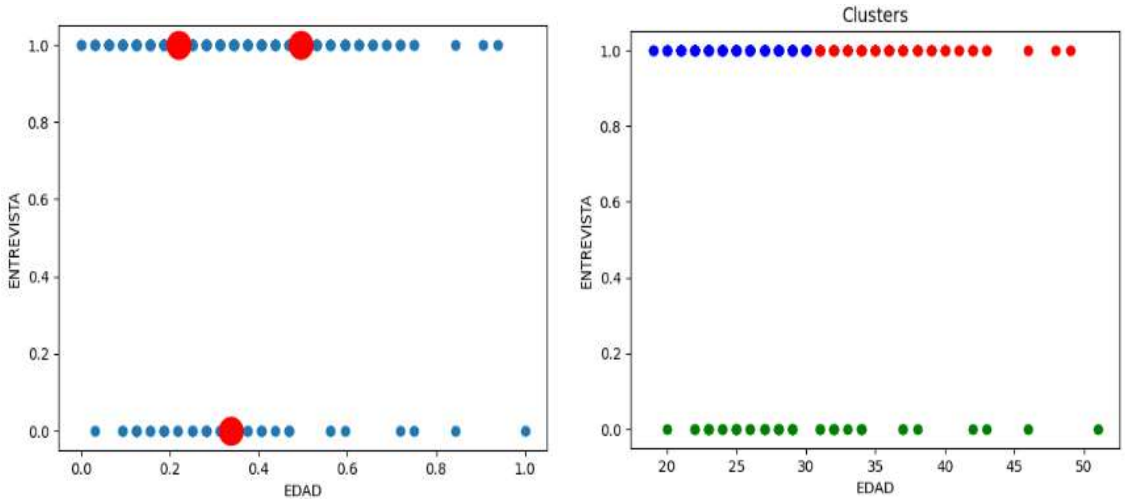


Figura 7.43: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Prueba de Entrevista.

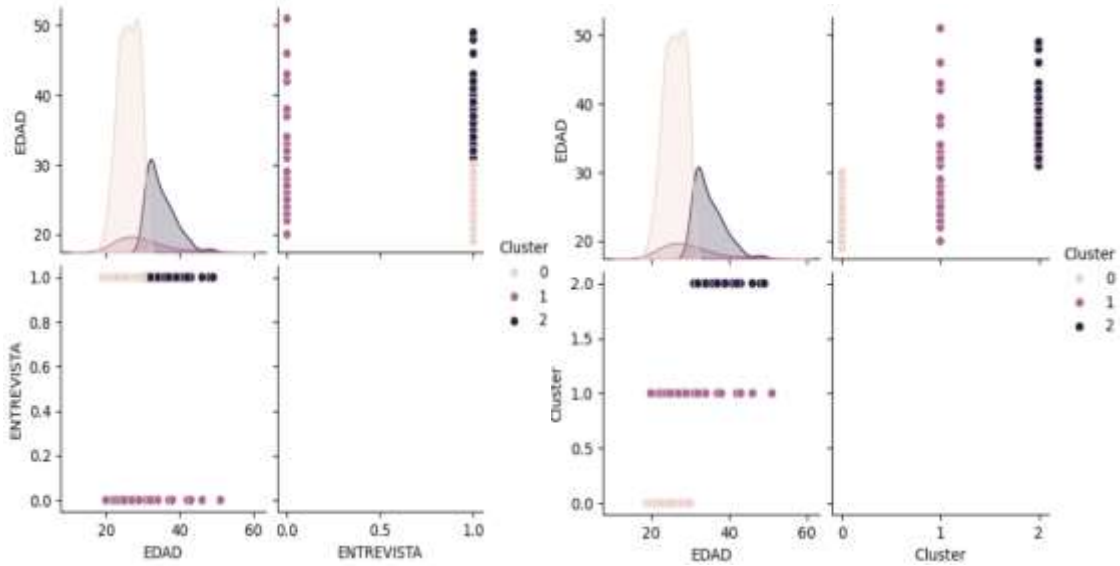


Figura 7.44: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Pruebas Entrevista.

- **Fichero archivoPsico.csv:**

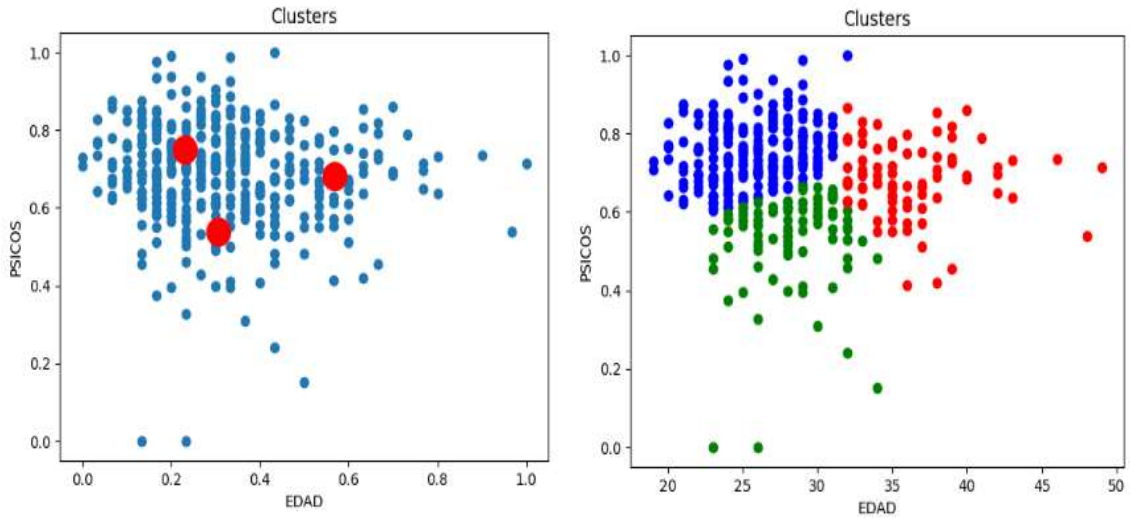


Figura 7.45: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Prueba de Psicotécnicos.

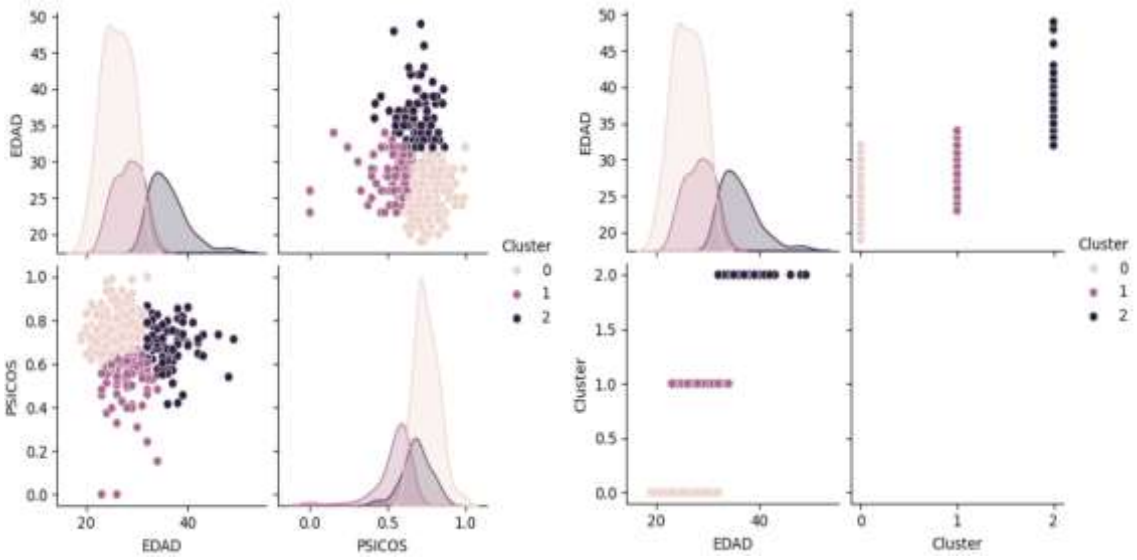


Figura 7.46: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Pruebas Psicotécnicos.

- **Fichero archivoIdioma.csv:**

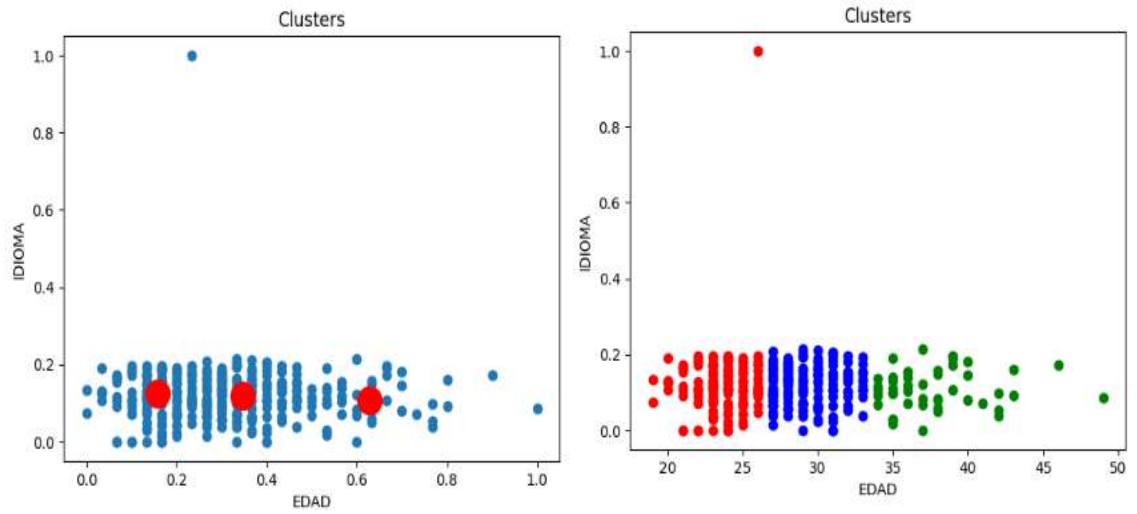


Figura 7.47: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Prueba de Idioma.

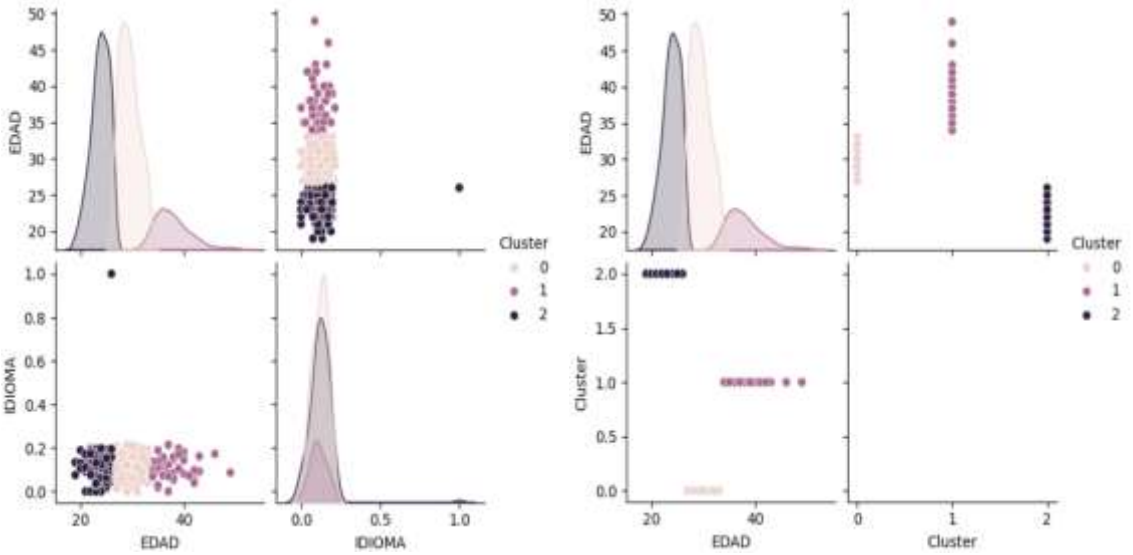


Figura 7.48: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Pruebas Psicotécnicos.

- **Fichero archivoAptosNota.csv:**

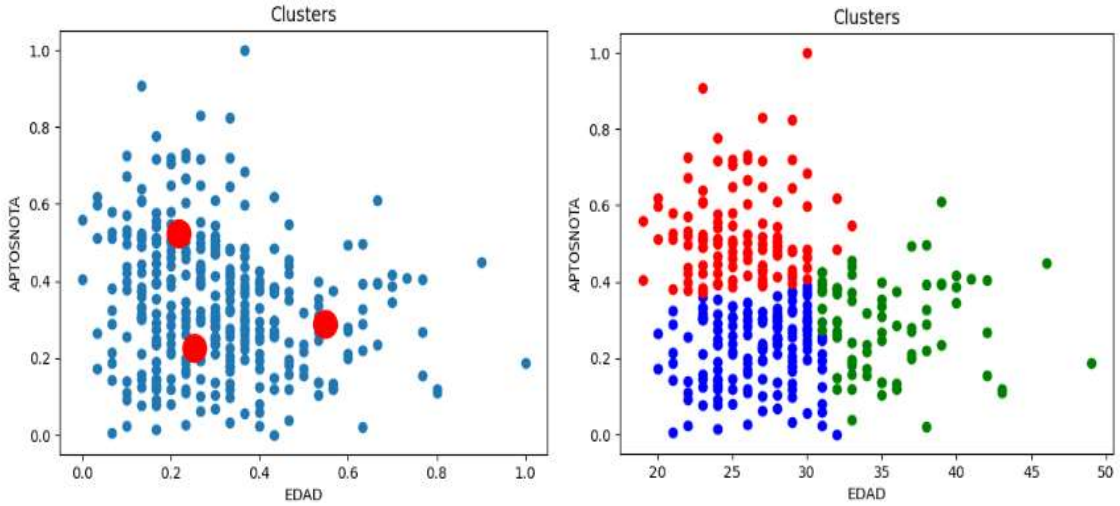


Figura 7.49: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Aptos Nota.

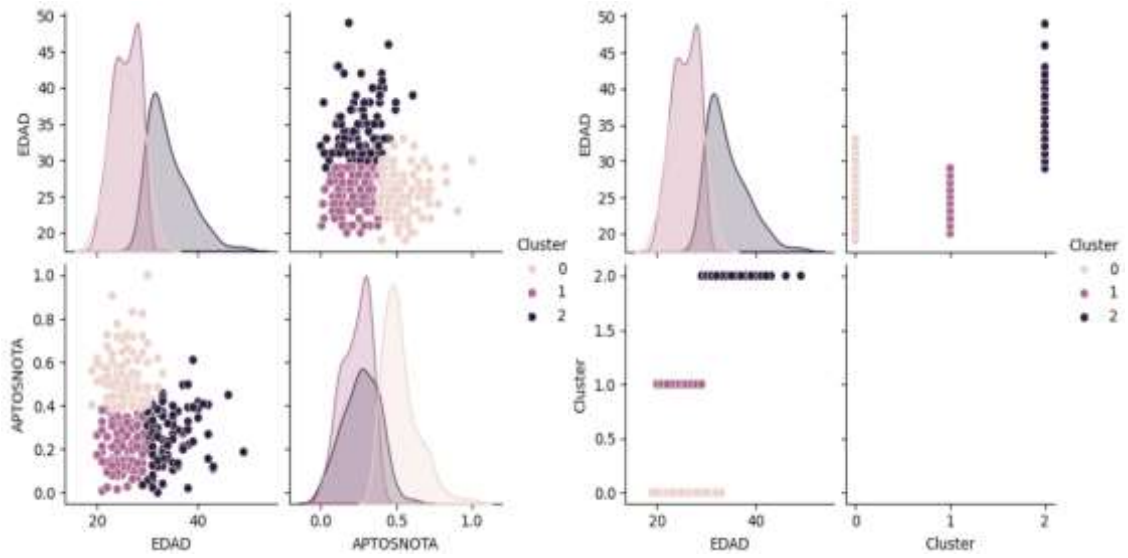


Figura 7.50: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo de Aptos Nota.

- **Fichero archivoAptosEsca.csv:**

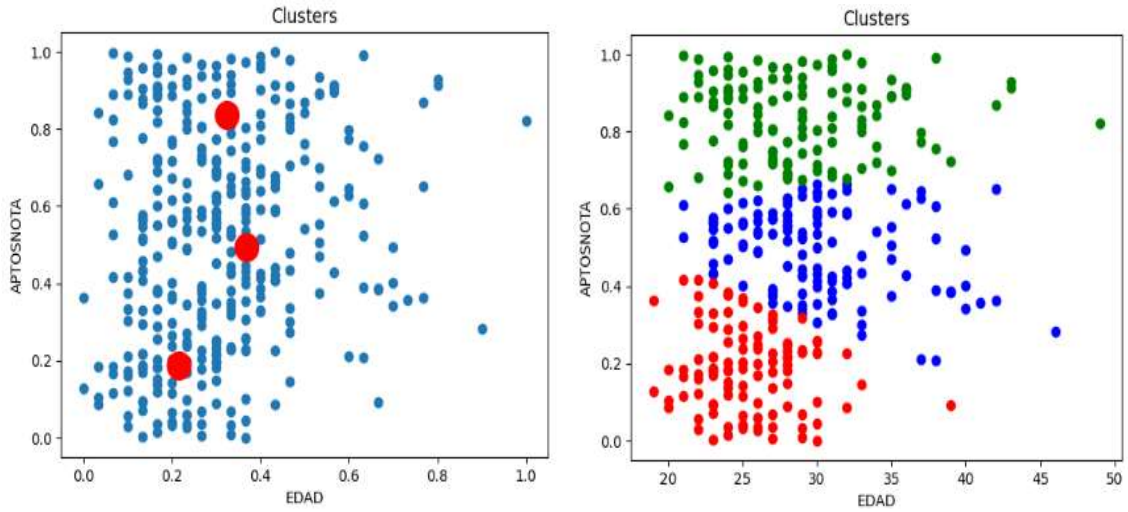


Figura 7.51: Gráfico de dispersión con y sin colores de asignación de datos a cada centroide para el archivo de Escalafón.

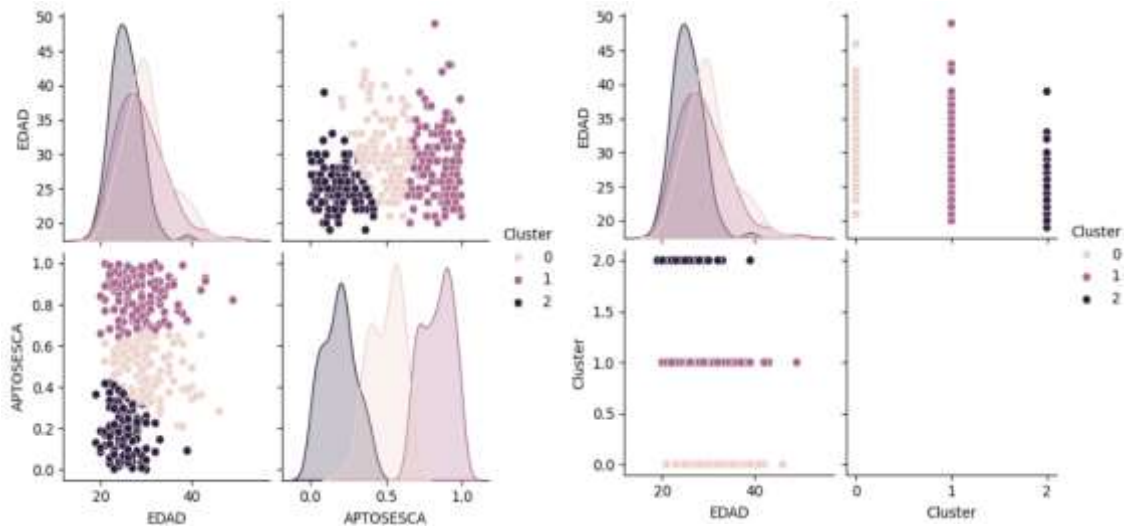


Figura 7.52: Gráfico del método pairplot de asignación de datos a cada centroide para el archivo Escalafón.

7.4.5. Histograma y estimación de la densidad kernel (KDE) de la variable

En este apartado se realiza un estudio de las variables del dataset, en concreto utilizando un histograma y un estimación de la densidad Kernel (KDE).

El histograma nos ofrece una representación gráfica dividiendo los datos en intervalos y cuenta el número de observaciones en cada intervalo, mostrando los datos en una forma discreta, ya que se basa en intervalos específicos.

La estimación de la densidad Kernel (KDE) es una técnica no paramétrica para estimar la función de densidad de probabilidad de una variable aleatoria, y en luugar de agrupar los datos en intervalos utiliza una función de kernel para suavizar la distribución de los mismos.

Uniendo los dos métodos se muestra un histograma con una superposición de la estimación de densidad de kernel (KDE) para distribución de los datos, representando en el eje horizontal (eje x) los valores de la variable en cuestión, y en el eje vertical (eje y) la frecuencia con la que aparecen esos valores. Cada barra del histograma representa la frecuencia de los valores dentro de un intervalo específico.

Esta combinación se aplica a las dos variables que forman cada uno de los datasets, para poder visualizar la distribución de los datos de cada una de ellas.

A continuación, se muestra el código utilizado para realizar la combinación de los métodos anteriormente descritos, uno por cada variable que forma nuestrso dataset:

```
sns.histplot(data=cluster_data, x='FÍSICAS', kde=True)
plt.show()
```

Figura 7.53: Ejemplo del código que grafica el histograma y la curva de densidad kernel.

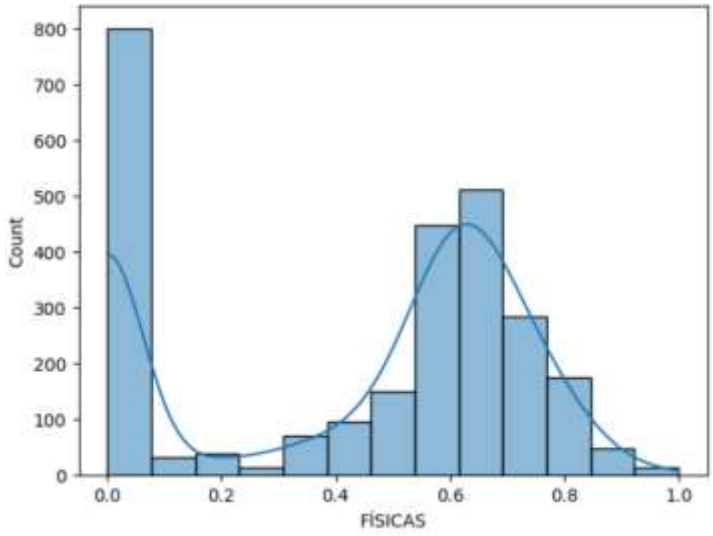


Figura 7.54: Ejemplo de las gráficas del histograma y de la curva de densidad kernel.

Seguidamente se aplica este código a cada uno de los archivos csv que forman nuestro dataset, con la intención de poder observar la distribución de los datos mediante la agrupación de estos en intervalos discretos que nos ofrece el histograma, junto con la estimación más suave que nos ofrece la función de densidad de probabilidad kernel.

- **Fichero archivoFisicas.csv**: A continuación, se muestran las gráficas de las variables 'EDAD' y `FÍSICAS' en las que se puede apreciar que la edad mayoritaria de los opositores en esta prueba es de 26 años, disminuyendo el número de éstos cuando son menores de 20 años y mayores de 35 años. Y con respecto a la nota obtenida en la prueba, se aprecia gran número de opositores que no se han presentado a la prueba, probablemente por ser la primera prueba del proceso, estando en la nota media la mayoría de los que se han presentado.

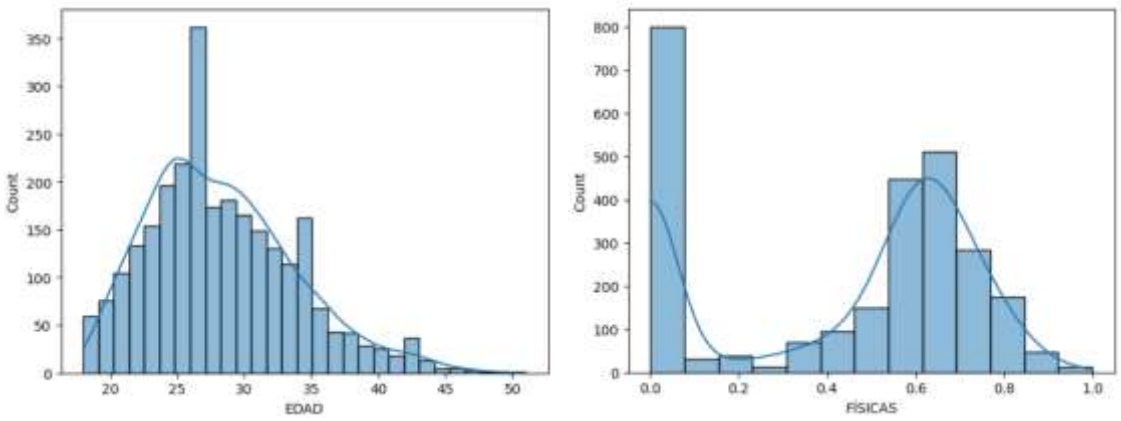


Figura 7.55: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Físicas del archivo Físicas.

- **Fichero archivoTeoria.csv**: En las gráficas que se muestran a continuación, se puede apreciar que los opositores de 26 y de 29 años son los mayoritarios en esta prueba, siendo alta la gran mayoría de notas obtenidas en la misma.

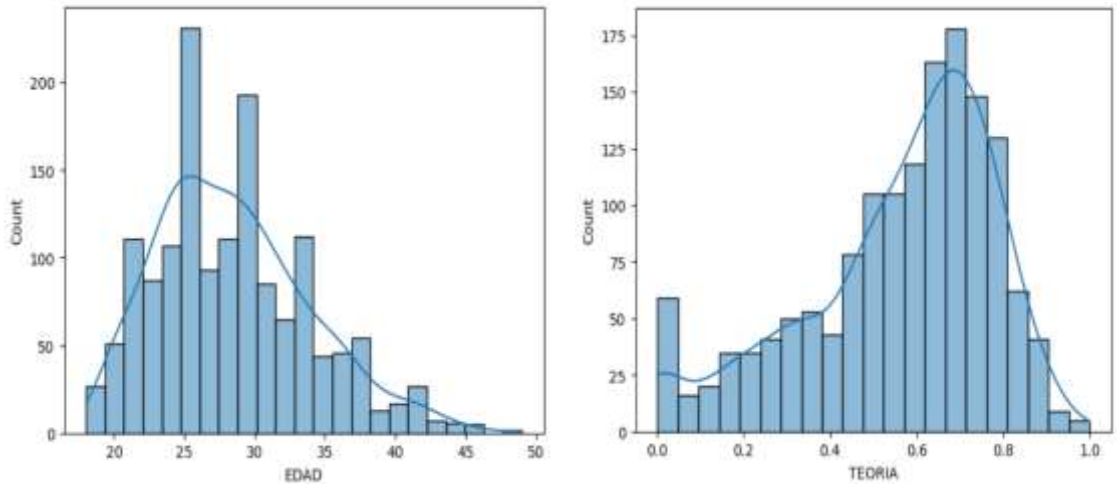


Figura 7.56: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Teoría del archivo Teoría.

- **Fichero archivoOrto.csv**: Se pude observar en las gráficas que se muestran a continuación como la distribución de los datos de la edad es la misma que con el fichero de Teoría porque ambas pruebas se realizan al mismo tiempo, así que son los mismos opositores los que las realizan. Y son respecto a la nota de ortografía, se

observa como la gran mayoría ha obtenido notas altas.

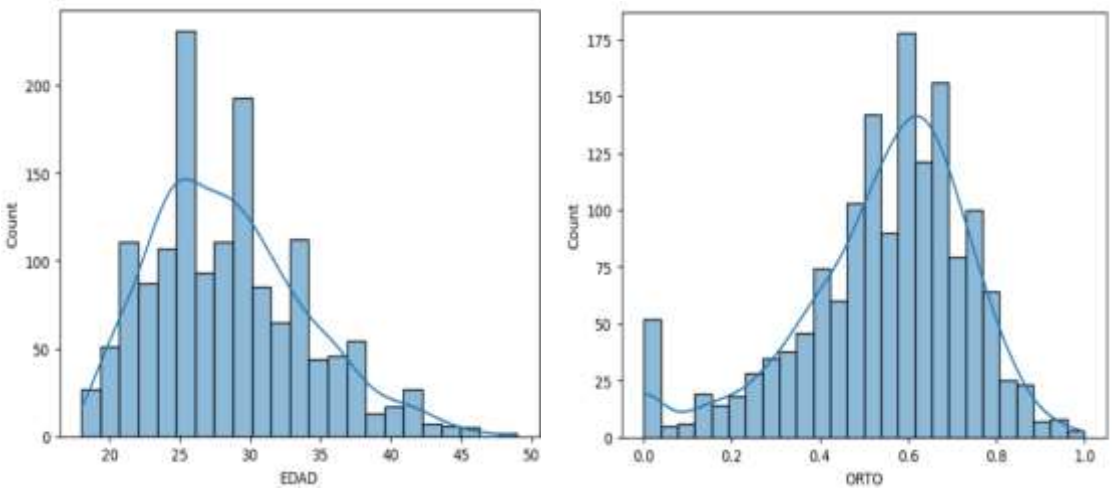


Figura 7.57: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Orto del archivo Ortografía.

- **Fichero archivoTeoriaOrto.csv**: En las siguientes gráficas se observa cómo se mantiene la edad de los dos apartados anteriores, dado que se analiza el mismo número de opositores que se han sometido a ambas pruebas de Teoría y de Ortografía. Sin embargo, en la gráfica llamada "TEORIAORTO" se puede observar que hay muchos opositores que no han supera ambas pruebas conjuntamente, pudiéndose apreciar que éstos son más del doble de los que las han superado.

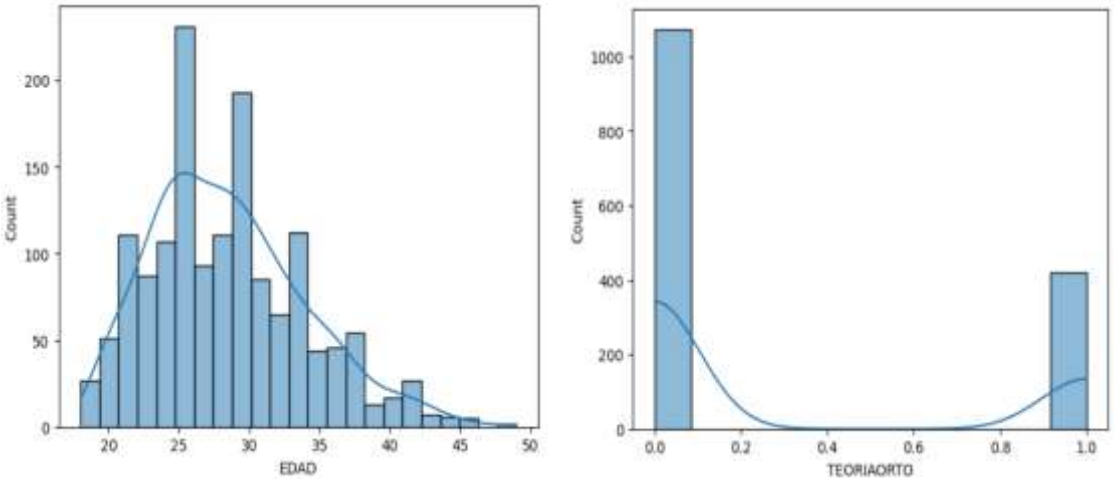


Figura 7.58: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Teoriaorto del archivo Teoría y ortografía.

- **Fichero archivoMedico.csv**: En esta prueba el rango mayoritario de edad que se han presentado a la misma oscila entre los 25 y los 30 años, destacando los opositores con edades de 26 y 28 años. El Reconocimiento Médico ha sido superado por la gran mayoría de los opositores que se presentaron al mismo, tal y como se puede apreciar en la gráfica "MEDICO".

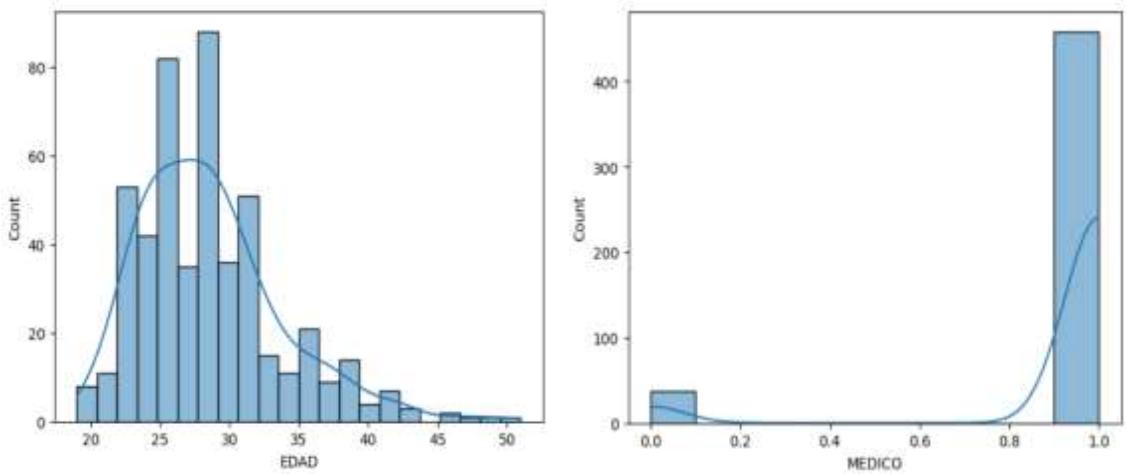


Figura 7.59: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Médico del archivo Médico.

- **Fichero archivoEntrevista.csv**: Como se puede observar en las siguientes gráficas, la prueba de la Entrevista ha sido superada por la mayoría de los opositores que se presentaron a la misma, manteniéndose el mismo gráfico de edad como en el apartado anterior dado que los opositores se presentan conjuntamente a ambas pruebas.

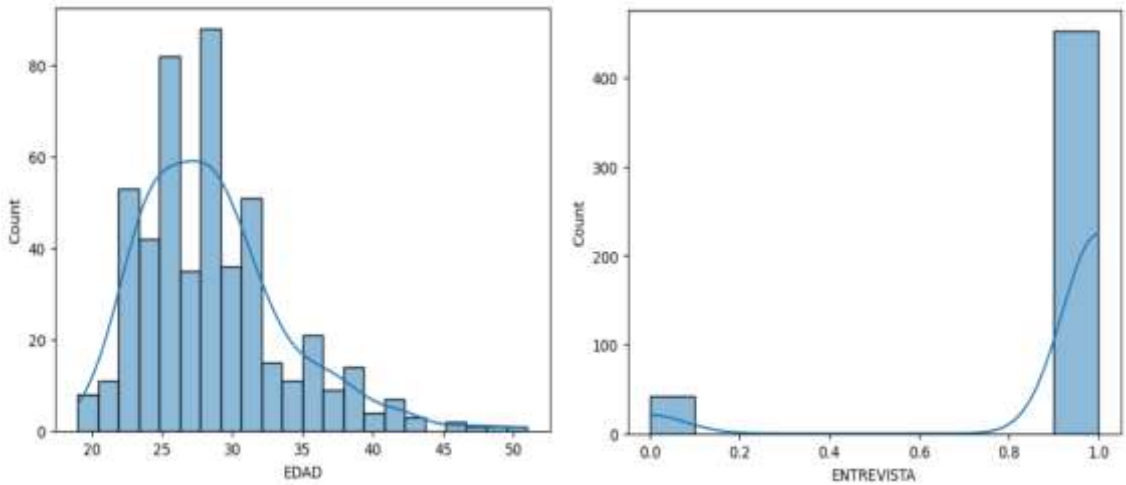


Figura 7.60: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Entrevista del archivo Entrevista.

- **Fichero archivoPsico.csv**: La variación del rango de edad de los opositores que se han presentado a la prueba de Psicotécnicos con respecto a las dos pruebas anteriores (pruebas de Reconocimiento Médico y de Entrevista) es prácticamente inapreciable, tal y como se aprecia en la gráfica siguiente, dado que la mayoría de opositores han superado dichas dos pruebas anteriores, Sin embargo, las notas obtenidas por lo aspirantes en la prueba de Psicotécnicos son bastantes varias, encontrándose la mayoría en un rango entre 0,6 y 0,8 puntos.

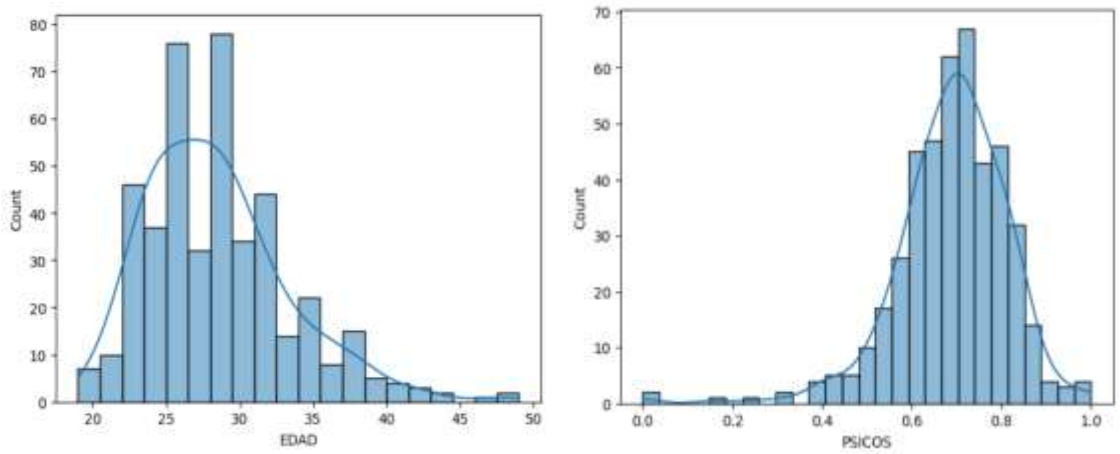


Figura 7.61: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Psicos el archivo Médico.

- **Fichero archivoIdioma.csv**: En las siguientes gráficas se puede observar como la edad no ha variado con respecto a la prueba anterior dado que son los mismos aspirantes que se presentaron a las tres pruebas anteriores (Reconocimiento Médico, Entrevista y Psicotécnicos). Respecto a la puntuación obtenida en le prueba de Idioma, en la siguiente gráfica se observa que la gran mayoría de los opositores han obtenido una puntuación bastante baja.

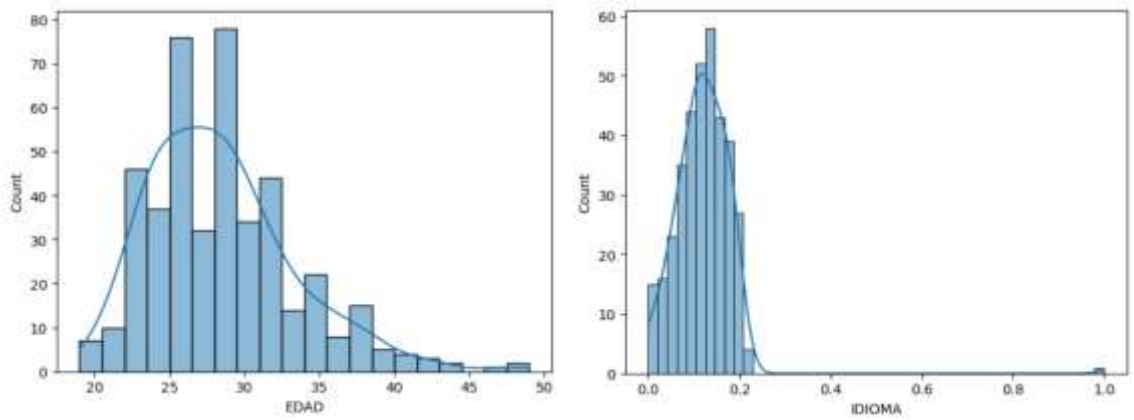


Figura 7.62: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Idioma del archivo Idioma.

- **Fichero archivoAptosNota.csv**: En las gráficas siguientes se observan los aspirantes que han supera la primera fase del proceso selectivo, siendo los opositores que vayan a la Escuela Nacional de Policía de Ávila para continuar con el proceso selectivo y de formación al mismo tiempo. La gráfica de Edad es prácticamente igual a las de las últimas pruebas, pudiéndose apreciar el rango de edad de los opositores que han supera esta primera fase del proceso discurre entre los 22 y los 33 años. Observando la gráfica de Aptos por Nota obtenida en esta primera fase del proceso selectivo, se aprecia que la mayoría de los opositores ha obtenido una puntuación inferior a la mitad de la máxima puntuación que se puede obtener.

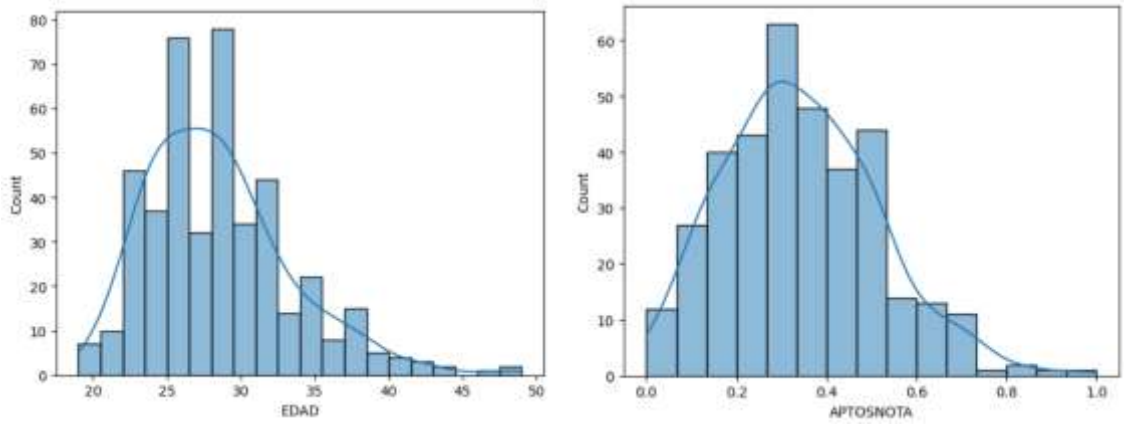


Figura 7.63: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Aptosnota del archivo Aptos Nota.

- **Fichero archivoAptosEsca.csv**: Al igual que el apartado anterior, la gráfica de Edad contiene los mimos datos porque son los mismos opositores. Sin embargo, de la gráfica del Escalafón siguiente se obtiene la conclusión que los aspirantes se encuentran repartidos de manera muy similar en el rango establecido para el escalafón.

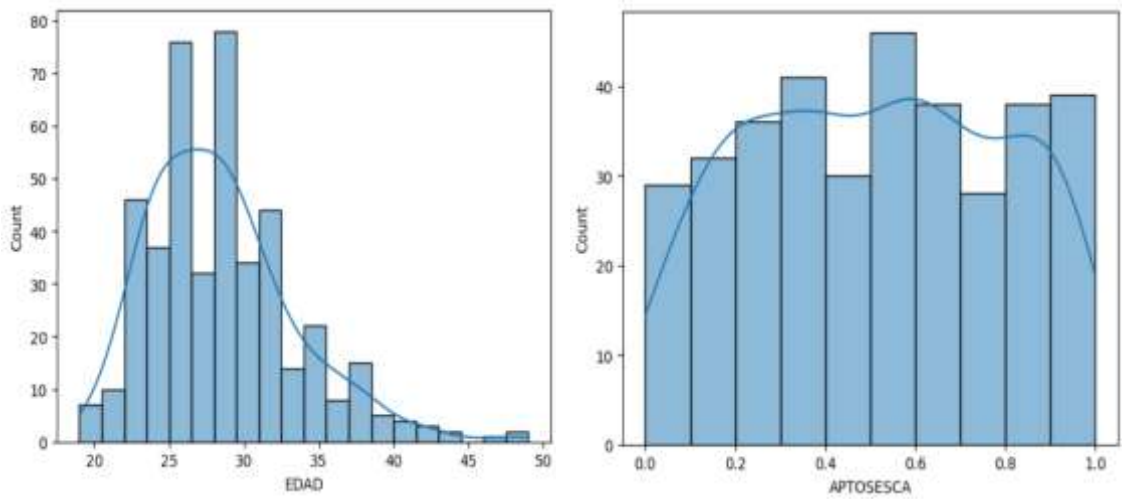


Figura 7.64: Gráficas del histograma y de la curva de densidad kernel de la variable Edad y Aptosesca del archivo Aptos Escalafón.

7.4.6. Gráfico de dispersión con una línea de regresión lineal

Para finalizar el análisis de los datos del dataset, en el presente apartado se realiza un estudio de las variables del dataset representadas mediante un gráfico de dispersión a la vez que se representa una línea de regresión lineal ajustada, permitiendo examinar la tendencia general entre las dos variables que se van a estudiar en cada caso en concreto.

A continuación, se muestra el código utilizado para realizar la combinación de los métodos anteriormente descritos, uno por cada variable que forma nuestrso dataset:

```
sns.regplot(data=cluster_data, x='EDAD', y='FÍSICAS')
plt.show()
```

Figura 7.65: Ejemplo del código que grafica gráfico de dispersión con una línea de regresión lineal.

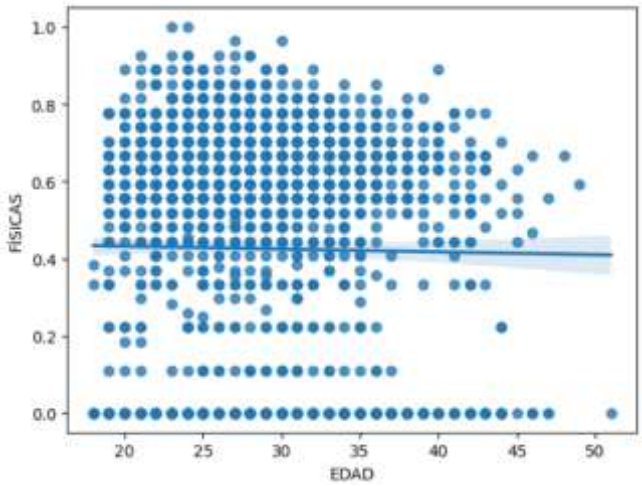


Figura 7.66: Ejemplo de las gráficas del gráfico de dispersión con una línea de regresión lineal.

Seguidamente se aplica este código a cada uno de los archivos csv que forman nuestro dataset, con la intención de poder observar la relación entre las dos variables en cada caso concreto, siendo esta combinación de métodos la que más información gráfica nos ofrece entre la relación de ambas variables.

- **Fichero archivoFisicas.csv**: En la siguiente gráfica se representa la relación entre las variables de 'EDAD' y 'FÍSICAS', observando la agrupación de los datos en la mitad superior de la nota de físicas, significando esto la mayoría de los presentados aprobaron la prueba de aptitud física. Además, se puede apreciar la presencia la presencia de algunos outliers. Y con respecto a la línea de regresión lineal, esta muestra inclinación descendente muy ligera, casi inapreciable, lo que nos indica que no existe ninguna relación entre las notas obtenidas en la prueba de Aptitud Física y la edad de los aspirantes.

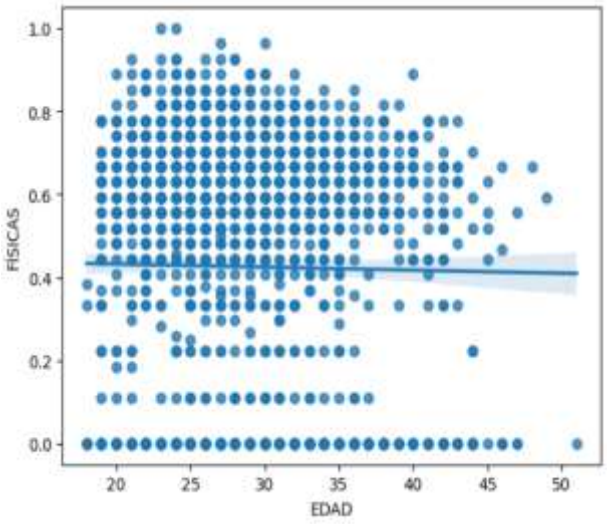


Figura 7.67: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Físicas.

- **Fichero archivoTeoria.csv**: En la gráfica siguiente se puede apreciar una mayor agrupación de los datos en la mitad superior del gráfico, por lo que hay mayor número de aspirantes que han obtenido 5 puntos en esta prueba. Así mismo, se pueden observar algunos outliers, no siendo éstos gran cantidad. Se puede apreciar una leve inclinación ascendente de la línea de regresión lineal, indicando ésta que a medida que la edad de los aspirantes aumenta, también lo hacen las notas obtenidas en esta prueba de Teoría.

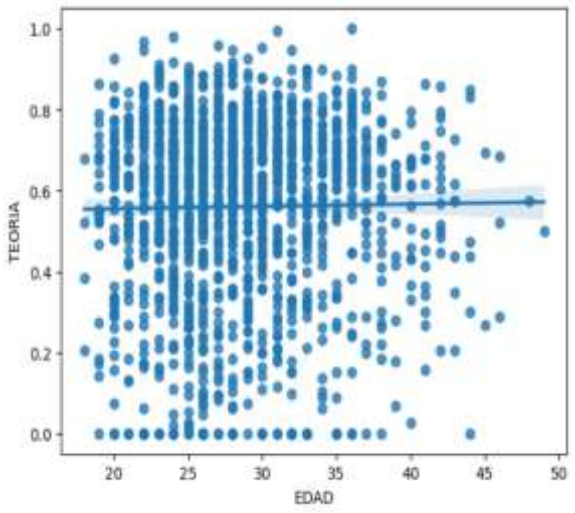


Figura 7.68: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Teoría.

- **Fichero archivoOrto.csv**: En la gráfica de más abajo, se puede apreciar una agrupación de los datos en la mitad superior del gráfico, indicando esta que la mayoría de los opositores han obtenido una puntuación alta. En la misma, se pueden apreciar outliers, pero sobre tos en la mitad inferior del gráfico. Y respecto a la línea de regresión lineal, no se aprecia inclinación alguna de ésta, lo que nos indica que no hay ninguna relación entre la edad y la nota de ortografía.

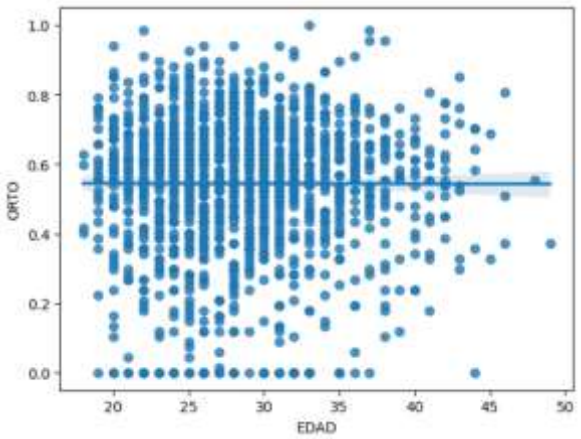


Figura 7.69: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Ortografía.

- **Fichero archivoTeoriaOrto.csv**: En esta gráfica solamente aparecen dos líneas de puntos, la superior representa los aspirantes que han superado las pruebas de Teoría y Ortografía conjuntamente, y la inferior los que no las han superado. Al igual que en los análisis por separado de la línea de regresión lineal que se han hecho más arriba de ambas pruebas, no existe relación alguna con la edad dado que dicha línea no muestra ninguna inclinación.

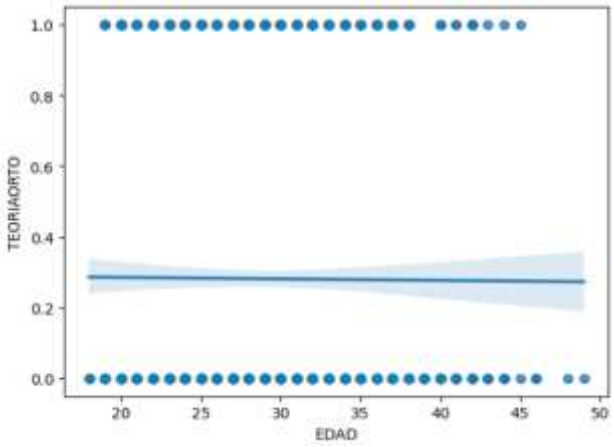


Figura 7.70: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo de Teoría y Ortografía.

- **Fichero archivoMedico.csv**: Al igual que en el caso anterior, la siguiente gráfica muestra dos únicas líneas de datos, representando la de la parte superior del grafico los opositores que has sido aptos en esta prueba, al contrario de lo que representa la de la parte inferior. Como se puede observar, la línea de regresión lineal en este caso muestra una inclinación inferior bastante pronunciada, lo que nos indica que a medida que la edad aumenta, las personas que han superado el reconocimiento médico en promedio disminuyen.

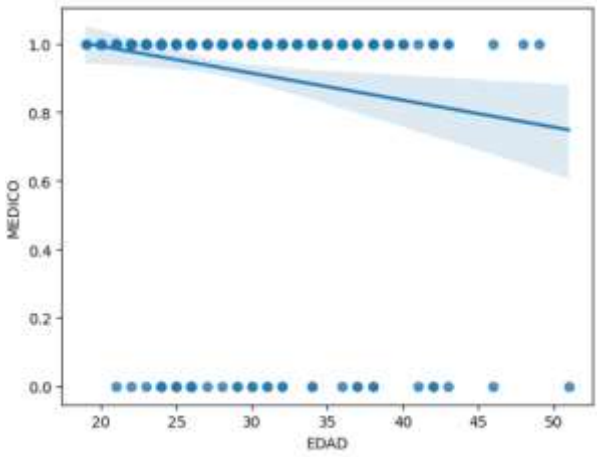


Figura 7.71: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Médico.

- **Fichero archivoEntrevista.csv**: En la siguiente gráfica, también observamos únicamente dos líneas, la de la parte superior nos vuelve a indicar los que han supera la prueba de entrevista y la de la parte inferior de la gráfica los que no. Además, podemos observar claramente que es mayor el número de aspirantes que han superado la prueba de Entrevista. La línea de regresión lineal aparece claramente con una inclinación descendente, lo que indica que a medida que la edad aumenta, las notas de la Entrevista en promedio disminuyen.

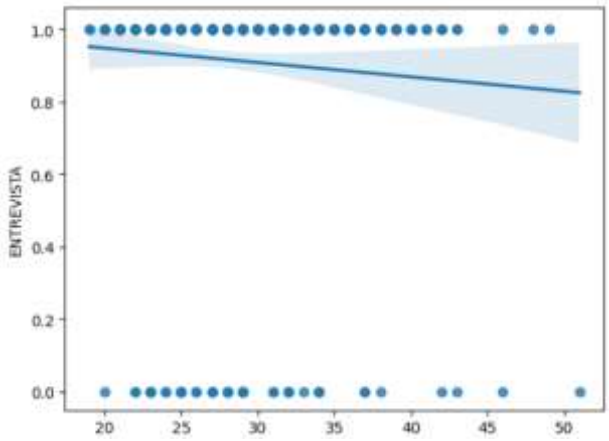


Figura 7.72: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Entrevista.

- **Fichero archivoPsico.csv**: La siguiente gráfica nos muestra que la mayoría de los opositores han obtenido una nota alta en la prueba de Psicotécnicos, así como la presencia de algunos outliers. Se puede apreciar que a medida que la edad aumenta las notas de los psicotécnicos en promedio disminuyen, aunque tampoco de una manera muy clara, dada la inclinación descendente que muestra la línea de regresión lineal no es muy pronunciada.

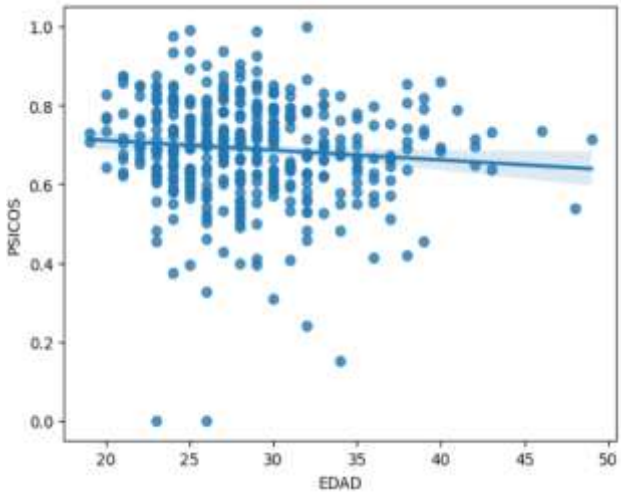


Figura 7.73: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Psicotécnicos.

- **Fichero archivoIdioma.csv**: A continuación, se puede apreciar como la inmensa mayoría de notas en la prueba de idioma es bastante baja, apreciándose algunos outliers, no demasiados. Respecto a la línea de regresión lineal, no indica ningún tipo de relación entre la nota de idioma y la edad, dado que no muestra inclinación alguna.

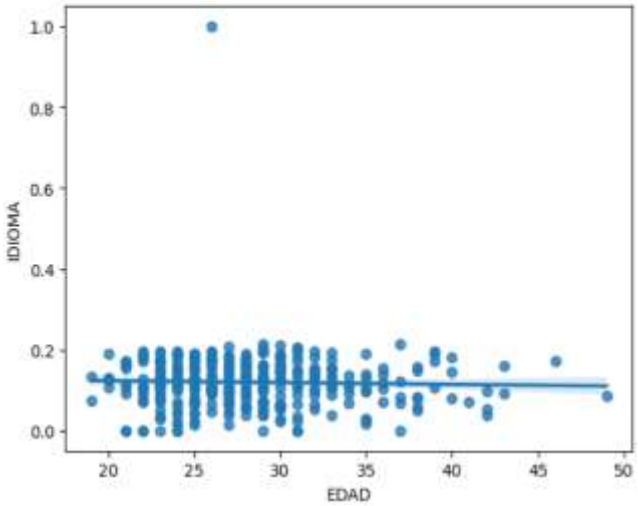


Figura 7.74: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Idioma.

- **Fichero archivoAptosNota.csv**: En la siguiente gráfica se observa que la mayoría de los datos se encuentran en la mitad inferior de la gráfica, mostrando la misma ciertos outliers. La línea de regresión lineal, dada su inclinación descendente, muestra que a medida que la edad aumenta, la nota obtenida en el proceso selectivo en promedio disminuye.

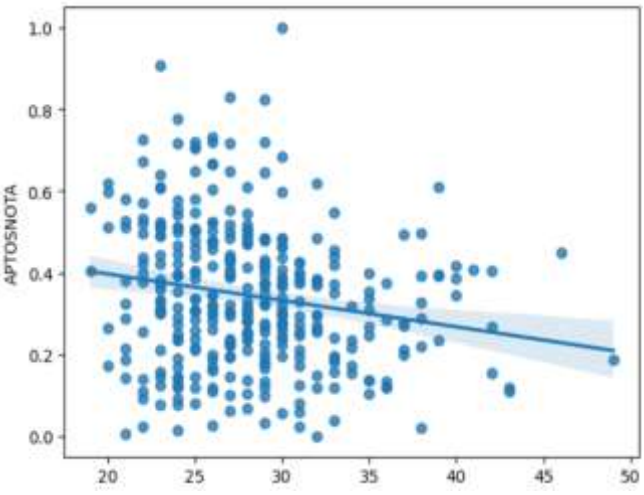


Figura 7.75: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Aptos Nota.

- **Fichero archivoAptosEsca.csv**: Como se puede apreciar en la siguiente gráfica, los datos se encuentran muy repartidos por la misma, lo que nos indica que los aspirantes que han superado esta primera parte del proceso selectivo se encuentran en posiciones muy variadas de escalafón. La línea de regresión lineal, muestra una pronunciada inclinación ascendente, lo que indica que a medida que la edad aumenta, el escalafón en promedio disminuye.

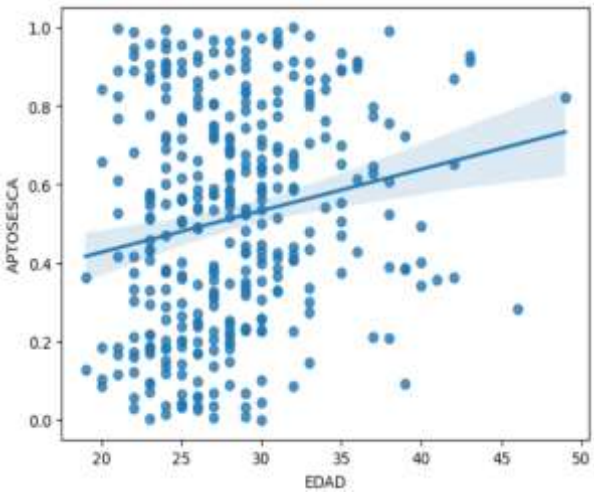


Figura 7.76: Gráfica del del gráfico de dispersión con una línea de regresión lineal del archivo Aptos Escalafón.

Capítulo 8

Conclusiones y trabajo futuro

Una vez terminado el proyecto habiendo alcanzado los objetivos propuestos al inicio del mismo, en este capítulo se describen las conclusiones a las que se han llegado.

8.1. Conclusión

Antes de realizar el proyecto, se tenía la duda de si la edad de los aspirantes podía afectar de algún modo, es decir, positiva o negativamente, al desarrollo de la preparación del proceso selectivo para optar a una plaza como Funcionario Público del Estado del Cuerpo Nacional de Policía, motivo por el cual se pensó en utilizar las herramientas que hoy en día nos ofrece la Inteligencia Artificial, concretamente utilizando métodos y modelos de Machine Learning que, a partir de unos datos ya procesados, pudieran ofrecer una respuesta a lo inicialmente planteado.

Una vez realizado el análisis de las puntuaciones obtenidas en todas y cada una de las pruebas de las que consta esta primera fase de oposición al Cuerpo Nacional de Policía, y utilizando tanto las opciones que nos ofrece la aplicación Excel, como los métodos y modelos de Machine Learning, se ha llegado a la conclusión de que la edad de los aspirantes no influye, ni positiva ni negativamente, en dicha puntuación, es decir, no se ha encontrado en este estudio una relación lo suficientemente clara para determinar que la edad del aspirante podría influir en la puntuación obtenida en las pruebas.

8.2. Experiencias y aprendizajes personales

Este proyecto me ha supuesto un gran reto debido tanto a factores personales, como profesionales, como de conocimiento. Con respecto al ámbito personal, me ha supuesto un esfuerzo importante dado que he tenido que armonizar la ejecución del proyecto con el tiempo y las obligaciones que conlleva ser padre de familia y, en muchas ocasiones, dedicar tiempo de estar con la familia y de descanso personal a la realización del mismo. A este aspecto hay que sumarle que, además, he tenido que compatibilizarlo con mi carrera profesional dentro del Cuerpo Nacional de Policía, lo que en ocasiones se ha tornado más complicado si cabe, teniendo en cuenta los turnos de trabajo que a veces hacen más costoso mantener una rutina de trabajo que me permitiera cumplir con los objetivos propuestos. Y por último, lo que ha contribuido a aumentar el reto ha sido el escaso conocimiento previo en la materia de Inteligencia Artificial, lo que me ha conllevado una gran cantidad de tiempo y estudio para alcanzar un nivel apropiado que me permitiera la elaboración del presente proyecto.

En cuanto al aprendizaje obtenido en la elaboración de este proyecto, cabe destacar los conocimientos adquiridos en el campo de Machine Learning, aportándome una experiencia en la combinación de la Inteligencia Artificial y el procesamiento de datos, dándome la oportunidad de hacer uso de los mismos para el estudio de futuros análisis en mi sector profesional de la Fuerzas y Cuerpos de Seguridad.

Bibliografía - Webgrafía

Yogesh Agrawal. “K-Means Clustering on Cars Dataset using Seaborn Visualization”. url: https://www.kaggle.com/code/yugagrawal95/k-means-clustering-using-seaborn-visualization

Fernando Berzal. “Clustering basado en particiones”. En: DECSAI, Universidad de Granada. Clustering.ppt. url: https://es.slideshare.net/slideshow/clusteringppt-259140531/259140531

Ricardo Moya. “Selección del número óptimo de Clusters”. En: Jarroba. url: https://jarroba.com/seleccion-del-numero-optimo-clusters/.

Kira Paulin. “Clustering”. En: SlidePlayer (2015). url: https://slideplayer.com/slide/2735334/

Manuel Trigas Gallego. “Metodología Scrum”. url: https://openaccess.uoc.edu/bitstream/10609/17885/1/mtrigasTFC0612memoria.pdf

J. S. Hicks M. Foster. “Adapting Scrum to Managing a Research Group”. url: https://drum.lib.umd.edu/handle/1903/10743.

Synapptica. Dar los primeros pasos en SCRUM. url: https://synapptica.net/metodologia-scrum.html

Tutorial Trello. by Yago González. url: https://yagogonzalez.com/tutorial-trello/

Glassdoor. Sueldos para Analista Junior. url: https:/ / www.glassdoor.es /Sueldos/analista-junior-sueldo-SRCH_KO0,15.htm

Parte II
Apéndices

Apéndice A

Contenido adjunto

A continuación, se detallan los directorios que se adjuntan junto a la memoria del presente proyecto desarrollado:

- **ANÁLISIS EXCEL**: Carpeta la cual contiene los ficheros del análisis realizado en Excel.

- **DATASETS**: Carpeta que contiene todos los datasets que han sido analizados con Clustering en el proyecto, tanto en formato Excel (.xlsx) como .csv.

- **IMPLEMENTACIÓN:** Carpeta con todos los ficheros o notebooks con el formato .ipynb, en los cuales se encontrará todo el código correspondiente a los métodos y modelos de aprendizaje usados en el proyecto.

Estos directorios se compartirán y almacenarán a través del repositorio en Onedrive proporcionado por la Escuela de Ingeniería Informática de Segovia.