ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

(ICAI)

Máster en Big Data: Tecnología y Analítica Avanzada

MACHINE LEARNING PARA LA SEGMENTACIÓN DE PACIENTES CRÓNICOS Y NIVELES DE CUIDADOS

Autor

Natalia Mirón Gracia

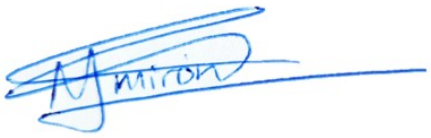
Dirigido por

Alberto Pardo Ortiz

Madrid

Junio 2022

Natalia Mirón Gracia, declara bajo su responsabilidad, que el Proyecto con título **MACHINE LEARNING PARA LA SEGMENTACIÓN DE PACIENTES CRÓNICOS Y NIVELES DE CUIDADOS** presentado en la ETS de Ingeniería (ICAI) de la Universidad Pontificia Comillas en el curso académico 2021/22 es de su autoría, original e inédito y no ha sido presentado con anterioridad a otros efectos. El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido tomada de otros documentos está debidamente referenciada.



Fdo.: Natalia Mirón Gracia Fecha: 05 / 06 / 2022

Autoriza la entrega:

El Director del Proyecto

Fdo.: Alberto Pardo Ortiz Fecha: ..... / 06 / 2022

V. B. del Coordinador de Proyectos

Fdo.: Carlos Morrás Ruiz-Falcó Fecha: ..... / ..... / .....

AUTORIZACIÓN PARA LA DIGITALIZACIÓN, DEPÓSITO Y DIVULGACIÓN EN RED DE PROYECTOS FIN DE GRADO, FIN DE MÁSTER, TESINAS O MEMORIAS DE BACHILLERATO

1º. Declaración de la autoría y acreditación de la misma.

El autor D. Natalia Mirón Gracia **DECLARA** ser el titular de los derechos de propiedad intelectual de la obra: **MACHINE LEARNING PARA LA SEGMENTACIÓN DE PACIENTES CRÓNICOS Y NIVELES DE CUIDADOS**, que ésta es una obra original, y que ostenta la condición de autor en el sentido que otorga la Ley de Propiedad Intelectual.

2º. Objeto y fines de la cesión.

Con el fin de dar la máxima difusión a la obra citada a través del Repositorio institucional de la Universidad, el autor **CEDE** a la Universidad Pontificia Comillas, los derechos de digitalización, de archivo, de reproducción, de distribución y de forma gratuita y no exclusiva, por el máximo plazo legal y con ámbito universal, de comunicación pública, incluido el derecho de puesta a disposición electrónica, tal y como se describen en la Ley de Propiedad Intelectual. El derecho de transformación se cede a los únicos efectos de lo dispuesto en la letra a) del apartado siguiente.

3º. Condiciones de la cesión y acceso

Sin perjuicio de la titularidad de la obra, que sigue correspondiendo a su autor, la cesión de derechos contemplada en esta licencia habilita para:

(a) Transformarla con el fin de adaptarla a cualquier tecnología que permita incorporarla a internet y hacerla accesible; incorporar metadatos para realizar el registro de la obra e incorporar "marcas de agua" o cualquier otro sistema de seguridad o de protección.

(b) Reproducirla en un soporte digital para su incorporación a una base de datos electrónica, incluyendo el derecho de reproducir y almacenar la obra en ser-

vidores, a los efectos de garantizar su seguridad, conservación y preservar el formato.

(c) Comunicarla, por defecto, a través de un archivo institucional abierto, accesible de modo libre y gratuito a través de internet.

(d) Cualquier otra forma de acceso (restringido, embargado, cerrado) deberá solicitarse expresamente y obedecer a causas justificadas.

(e) Asignar por defecto a estos trabajos una licencia Creative Commons.

(f) Asignar por defecto a estos trabajos un HANDLE (URL *persistente*).

4º. Derechos del autor.

El autor, en tanto que titular de una obra tiene derecho a:

(a) Que la Universidad identifique claramente su nombre como autor de la misma

(b) Comunicar y dar publicidad a la obra en la versión que ceda y en otras posteriores a través de cualquier medio.

(c) Solicitar la retirada de la obra del repositorio por causa justificada.

(d) Recibir notificación fehaciente de cualquier reclamación que puedan formular terceras personas en relación con la obra y, en particular, de reclamaciones relativas a los derechos de propiedad intelectual sobre ella.

5º. Deberes del autor.

El autor se compromete a:

(a) Garantizar que el compromiso que adquiere mediante el presente escrito no infringe ningún derecho de terceros, ya sean de propiedad industrial, intelectual o cualquier otro.

(b) Garantizar que el contenido de las obras no atenta contra los derechos al honor, a la intimidad y a la imagen de terceros.

(c) Asumir toda reclamación o responsabilidad, incluyendo las indemnizaciones por daños, que pudieran ejercitarse contra la Universidad por

terceros que vieran infringidos sus derechos e intereses a causa de la cesión.

(d) Asumir la responsabilidad en el caso de que las instituciones fueran condenadas por infracción de derechos derivada de las obras objeto de la cesión.

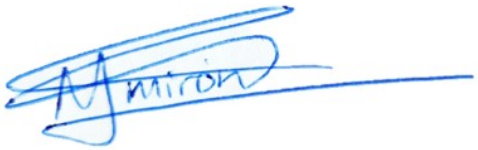
6º. Fines y funcionamiento del Repositorio Institucional.

La obra se pondrá a disposición de los usuarios para que hagan de ella un uso justo y respetuoso con los derechos del autor, según lo permitido por la legislación aplicable, y con fines de estudio, investigación, o cualquier otro fin lícito. Con dicha finalidad, la Universidad asume los siguientes deberes y se reserva las siguientes facultades:

- La Universidad informará a los usuarios del archivo sobre los usos permitidos, y no garantiza ni asume responsabilidad alguna por otras formas en que los usuarios hagan un uso posterior de las obras no conforme con la legislación vigente. El uso posterior, más allá de la copia privada, requerirá que se cite la fuente y se reconozca la autoría, que no se obtenga beneficio comercial, y que no se realicen obras derivadas.
- La Universidad no revisará el contenido de las obras, que en todo caso permanecerá bajo la responsabilidad exclusive del autor y no estará obligada a ejercitar acciones legales en nombre del autor en el supuesto de infracciones a derechos de propiedad intelectual derivados del depósito y archivo de las obras. El autor renuncia a cualquier reclamación frente a la Universidad por las formas no ajustadas a la legislación vigente en que los usuarios hagan uso de las obras.
- La Universidad adoptará las medidas necesarias para la preservación de la obra en un futuro.
- La Universidad se reserva la facultad de retirar la obra, previa notificación al autor, en supuestos suficientemente justificados, o en caso de reclamaciones de terceros.

Madrid, a 5 de junio de 2022

ACEPTA



Fdo.: **Natalia Mirón Gracia**

Motivos para solicitar el acceso restringido, cerrado o embargado del trabajo en el Repositorio Institucional:

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

(ICAI)

Máster en Big Data: Tecnología y Analítica Avanzada

MACHINE LEARNING PARA LA SEGMENTACIÓN DE PACIENTES CRÓNICOS Y NIVELES DE CUIDADOS

Autor

Natalia Mirón Gracia

Dirigido por

Alberto Pardo Ortiz

Madrid

Junio 2022

Agradecimientos

Este último año ha sido el más intenso de mi vida y en el que he aprendido más. Por ello, quiero agradecer a todo el mundo que me ha ayudado y apoyado en el camino.

Gracias a todos los profesores del Máster de Big Data en ICAI. En especial al director, Carlos Morrás, gracias por estar tan pendiente de nosotros todo el curso dirigiendo un máster tan especial. A ICAI, después de cinco años yendo diariamente a Alberto Aguilera, me toca despedirme estando muy orgullosa y agradecida de todo lo que esta facultad me ha aportado.

Gracias a mis compañeros de EY. En estos seis meses trabajando con el equipo del SAS me he sentido como una más. Muchas gracias a Alberto Pardo por ejercer de mi director de proyecto, y por apoyarme en la ejecución del mismo.

Gracias a mi familia. Papá, mamá, gracias por apoyarme en todas las decisiones que he tomado y por ser mi modelo de trabajo constante y esfuerzo diario. Todo es más fácil con vosotros al lado.

Gracias a todos mis compañeros de máster. Por las noches largas de entregas de proyectos a las 23.58, por la tercera fila del aula 408 y porque ha sido un año duro, pero con el apoyo de todos ha sido un gran año. A todos, ¡os deseo lo mejor!

¡Gracias a todo el mundo que ha hecho esto posible!

Resumen

El aumento de la esperanza de vida a nivel global ha supuesto una subida en España de 78,12 años en 1995 a 83,22 en 2022 y está causando un envejecimiento en la población. Este envejecimiento infiere un aumento de pacientes crónicos y crea una necesidad para un sistema de segmentación de estos pacientes, para su control y monitorización. En la actualidad, existen modelos de segmentación estáticos como el modelo de Káiser Permanente, donde la población crónica es repartida en tres niveles según su complejidad. Este proyecto explora un método para personalizar la segmentación de pacientes crónicos en base a variables socio-económicas, de hábitos de actividad y de índices médicos, específicas de la Comunidad de Andalucía. Un total de veinte variables críticas son seleccionadas para el modelo de Machine Learning, donde se realiza una reducción de variables con el método de EFA y una clusterización por el método de K-Means. Los resultados obtenidos son buenos, obteniendo un total de seis clusters distintivos que se ajustan a la población andaluza. A su vez, los resultados destacan la importancia de la georreferenciación, obteniéndose distinciones entre pacientes crónicos en diferentes zonas de Andalucía. Es por todo ello, que el modelo de segmentación de pacientes crónicos usando machine learning supone una mejora significativa a los modelos actuales de segmentación, con el objetivo de reducir el coste y mejorar la experiencia del cuidado del paciente.

Palabras Clave

Enfermedades no transmisibles (ENT), Análisis Factorial Exploratorio (EFA), K-Means, Bases de Datos (BBDD)

Abstract

The increase in life expectancy at a global level, assuming a rise in Spain from 78.12 years in 1995 to 83.22 in 2022, is causing an aging population. This aging infers an increase in chronic patients and creates a need for a segmentation system for these patients, for their control and monitoring. Currently, there are static segmentation models such as the Kaiser Permanente model, where the chronic population is divided into three levels according to their complexity. This project explores a method to personalise the segmentation of chronic patients based on socioeconomic variables, activity habits and medical indices, specific for the region of Andalusia. A total of twenty critical variables are selected for the Machine Learning model, where a reduction of variables is performed using the EFA method and a clustering using the K-Means method. The results obtained are good, obtaining a total of six distinctive clusters that fit the Andalusian population. In turn, the results highlight the importance of georeferencing, obtaining distinctions between chronic patients in different areas of Andalusia. For all these reasons, the chronic patient segmentation model using machine learning represents a significant improvement over current segmentation models.

Keywords

Non-communicable diseases (ENT), Exploratory Factorial Analysis (EFA), K-Means, Databases (BBDD)

Índice

Índice de Figuras

Índice de Tablas

1. Introducción y Enfoque del Proyecto

En el proyecto se explora el uso de técnicas de Machine Learning para personalizar la segmentación de los pacientes crónicos en base a las características específicas de la Comunidad Autónoma de Andalucía. Este Trabajo de Fin de Máster se basa en el caso de uso cuatro del proyecto desarrollado por EY-Solutia como aparece en el pliego del mismo [1].

Resaltar que las conclusiones y resultados obtenidos a lo largo del trabajo no son los oficiales concluidos por EY-Solutia, sino los resultados obtenidos de mis propios experimentos.

1.1. Descripción del Problema

Las enfermedades crónicas (o enfermedades no transmisibles, ENT), definidas como *enfermedades de más de seis meses y con una progresión lenta*, se han convertido en una amenaza global. Son la causa del 71 % de las muertes producidas en el mundo según la Organización Mundial de la Salud (OMS) [2].

Los pacientes con ENTs en España están aumentando debido al envejecimiento de la sociedad y al aumento de la esperanza de vida. De los últimos datos recogidos por la INE un 42 % de la población española (19 millones de personas) padece al menos una enfermedad crónica [3].

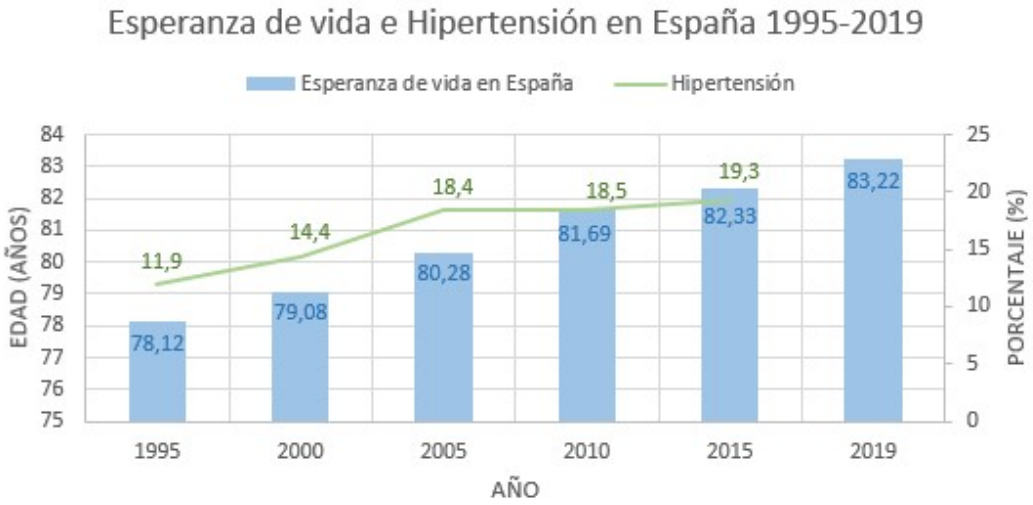


Figura 1.1: *Evolución de la esperanza de vida y la tasa de Hipertensión en España del 1995 a 2019 [4]*

Una de las enfermedades crónicas más comunes en España es la Hipertensión. Como se puede ver en la Figura 1.1, la tasa de la población española con hipertensión ha aumentado de 11,9 % en 1995 al 19,3 % de la población en 2015. A su vez, se aprecia como hay una tendencia positiva de aumento de esperanza de vida. Esta realidad requiere un nuevo enfoque sanitario a nuestro sistema actual, apoyándose de las nuevas tecnologías y el Big Data para centrarse en las nuevas necesidades [5].

Un paciente crónico es una persona con una o más enfermedades crónicas. Existen diversos perfiles de pacientes crónicos, como se puede leer en la sección 2.1, variando desde pacientes con diabetes a pacientes con padecimientos cardiovasculares. Cada tipo de paciente no requiere ni la misma atención, ni el mismo número de visitas a los servicios de Atención Especializada (AE) ni el mismo coste para la Salud Andaluza. Poder segmentar correctamente la población crónica en niveles de riesgo, optimizaría la gestión de los pacientes de mayor riesgo para, por ejemplo, anticiparse a un reingreso hospitalario y reducir los costes asociados [6].

Actualmente, la segmentación de pacientes crónicos se basa en la Pirámide de Káiser (Sección 2.1). Esta pirámide consiste en una estratificación de pacientes crónicos [1] en cuatro niveles estáticos [5]. Estos cuatro niveles son comunes para todos los hospitales, sin poder personalizar los grupos dependiendo de las diferencias sociales, geográficas o sanitarias. El uso de una segmentación con técnicas de Machine Learning consigue unos niveles dinámicos, cambiantes y específicos para la Comunidad de Andalucía, que ofrecen numerosas ventajas para reducir el coste y mejorar la experiencia del cuidado del paciente.

El proyecto tiene tres pilares teóricos fundamentales que son explicados con más detalle en el Capítulo 2 dedicado al estado del arte del proyecto. En primer lugar, las Enfermedades Crónicas (Sección 2.1), donde se explican los tipos de enfermedades crónicas existentes y los tipos de segmentación de pacientes crónicos actuales. Los últimos dos pilares técnicos consisten en las técnicas de Machine Learning utilizadas en el modelo: Análisis Factorial Exploratorio (EFA) (Sección 2.2) y el Algoritmo de K-Means (Sección 2.3).

[1] **Estratificación de pacientes crónicos** consiste en la clasificación de pacientes en grupos en base a su riesgo, morbilidad y complejidad

1.2. Motivación

La motivación principal detrás de la realización de este proyecto es ayudar a lograr el Objetivo 3.4 del tercer Objetivo de Desarrollo Sostenible (ODS).
La ONU tiene un total de 17 ODS para conseguir "un futuro mejor y más sostenible para todos" [7]. Cada uno de estos objetivos abarca un área específica, por ejemplo, el objetivo 1 es el fin de la pobreza y el objetivo número 5 es el de la igualdad de género, y dentro de cada objetivo hay metas específicas con fechas límites. Tal y como se puede ver en la Figura 1.2, el tercer ODS propone mejoras en el sector de la *Salud y el Bienestar*. Para 2030 se trabaja para reducir la mortalidad maternal a menos de 70 de cada 100.000 partos o reducir a la mitad el número de muertes en accidentes de tráfico en todo el mundo.

Figura 1.2: *Objetivo de Desarrollo Sostenible 3 de las Naciones Unidas [7]*

Otra meta dentro del tercer ODS es el objetivo 3.4, siendo la motivación de este proyecto. Consiste en reducir en un tercio la mortalidad prematura por ENTs a través de la prevención y el tratamiento, y promover la salud mental y el bienestar para el 2030. [8] Utilizar las nuevas tecnologías y el Big Data afecta de manera directa a la detección de enfermedades crónicas, así como realizar un seguimiento a su evolución.
La segunda motivación del proyecto es puramente técnica. Se busca estudiar el comportamiento del método K-Means con datos estandarizados y analizar la importancia de las variables elegidas. Se quiere confirmar el poder de los datos para casos médicos y responder a la pregunta: ¿pueden los datos usarse para mejorar el sistema sanitario?

1.3. Objetivos del Proyecto

El proyecto tiene objetivos funcionales y analíticos. Los objetivos funcionales corresponden con el objetivo final del proyecto. Este objetivo funcional es identificar similitudes y agrupar a pacientes crónicos para personalizar la atención sanitaria y optimizar la utilización de recursos. Los objetivos analíticos son los objetivos específicos de las tareas individuales necesarias para obtener el objetivo funcional. Los objetivos analíticos del proyecto son los siguientes:

- Obtener un mínimo de **15 variables** para la segmentación de pacientes crónicos
- **Estandarizar las variables** escogidas normalizándolas para la clasificación del algoritmo de K-Means
- Desarrollar modelo que utilice las coordenadas de posicionamiento de pacientes para obtener **variables sociales y económicas por zonas**
- Desarrollar un modelo capaz de segmentar a los pacientes en un **mínimo de 5 clústers** obteniendo un valor de **silhouette medio mínimo de 0.6** en función de las variables seleccionadas
- Encontrar patrones y una **interpretación lógica de del segmentos identificados** y justificar cada clúster a un perfil de paciente

Al utilizar para la segmentación el algoritmo de K-Means que pertenece a la clasificación no supervisada, no se obtienen valores numéricos de la precisión de la segmentación de forma directa. Para determinar la precisión de la segmentación de pacientes crónicos, se podría probar la primera iteración en hospitales y registrar la diferencia en costes e ingresos a urgencias con respecto a la segmentación anterior.

1.4. Metodología

La metodología seguida para la realización de este proyecto sigue el concepto de metodología cuantitativa. La metodología cuantitativa consiste en llegar a conclusiones con datos numéricos en lugar de la metodología cualitativa, que se centra en la realización de encuestas.

La planificación del proyecto se puede ver en la sección 5, en el apartado 5.2. Se basa en las diferentes fases a seguir para cumplir los objetivos citados en el apartado anterior, el apartado 1.3.

1.5. Estructura del Documento

Este trabajo de fin de máster está compuesto por un total de seis capítulos. El capítulo actual es la introducción del proyecto, donde se han explicado las motivaciones y los objetivos del proyecto. El capítulo 2 incluye una descripción detallada de áreas, técnicas y tecnologías que han sido utilizadas durante el desarrollo de este proyecto. En este caso se explican las enfermedades crónicas y su historia, así como aspectos técnicos del proyecto como son las técnicas de Machine Learning. El capítulo 3 abarca una descripción detallada del proceso de análisis y diseño del trabajo. Se analiza el modelo construido así como las variables elegidas y sus transformaciones. En el capítulo 4 se exponen los resultados y el análisis de la reducción de variables y de la clusterización final. En el capítulo 5 se describe el ciclo de vida que ha sido elegido para el desarrollo de este trabajo. Finalmente, el capítulo 6 se explican las conclusiones obtenidas y se presentan posibles trabajos futuros.

2. Estado del Arte

Este proyecto requiere el conocimiento de dos áreas, por un lado el área sanitaria y médica y por otro lado, el área técnica de análisis de datos. Por ello, en este Capítulo 2 se incluye una descripción detallada de ambos.

2.1. Enfermedades Crónicas y su segmentación

Tal y como se introdujo en el apartado 1.1, en los últimos años ha habido cambios demográficos en todo el mundo que están haciendo replantearse la organización de los servicios sanitarios. El primer cambio es el aumento de la esperanza de vida, envejeciendo a la población. El porcentaje de personas mayores de 65 años en la unión Europea (UE) pasará del 16,1 %, en el año 2000, al 27,5 % en el año 2050 [9]. A su vez, está aumentando el porcentaje de fallecimientos a causa de las enfermedades crónicas. Las enfermedades crónicas eran la causa de 40 % de las muertes en el mundo en 2000, y actualmente suponen el 71 % [2]. Estas cifras explican el impacto de estas enfermedades, el deterioro de calidad de vida de los pacientes crónicos y la importancia del correcto tratamiento de pacientes con estas. Sin embargo, ¿a qué se considera una enfermedad crónica? Una enfermedad crónica es una enfermedad con una duración mínima de seis meses y de progresión lenta, esto significa que no aparece de forma abrupta, sino progresivamente. Estas enfermedades pueden ser incurables, pero para la mayoría hay tratamientos para reducir sus efectos. El origen de las enfermedades crónicas proviene de la genética, de malos hábitos o por infecciones [10].

Los tipos de enfermedades crónicas más comunes en España se pueden ver se ven en la Figura 2.1, así como ejemplos de enfermedades de cada tipo. El tipo más común son las enfermedades cardiovasculares, estas son todas las patologías que afectan al corazón y los vasos sanguíneos. Estas enfermedades son en su mayor parte de los casos prevenibles, siendo factores de su aparición la alimentación, el peso, el ejercicio o fumar. El segundo tipo más común, como se ve en la Figura 2.1, son las enfermedades respiratorias, mayormente causadas por el tabaco aunque también por anomalías genéticas, como puede ser el caso del asma. El tercer tipo de enfermedad crónica es el cáncer. Los demás tipos se pueden ver en la Figura 2.1, que engloban enfermedades como el SIDA (infecciosa), diabetes (endocrina) o

la celiaquía (autoinmune) [10].

Figura 2.1: *10 tipos de Enfermedades Crónicas más comunes en España [10]*

Observando los tipos de enfermedades crónicas, se detectan diferencias entre ellos. Un paciente con cáncer de pulmón requiere un nivel de cuidado significativamente mayor que un paciente con diabetes. Por ello, existen diversos modelos que los centros sanitarios usan para segmentar a los pacientes crónicos, con el objetivo de mejorar la salud de los pacientes crónicos, aumentar el nivel de conocimiento del paciente y disminuir los ingresos hospitalarios innecesarios. El modelo de segmentación más utilizado en Europa, Estados Unidos, Australia y España es el modelo de Kaiser Permanente [9].

2.1.1. Modelo de Kaiser Permanente

El modelo poblacional de Kaiser Permanente fue construido por el médico Sidney Garfield y fundado por el empresario industrial Henry J Kaiser. Este modelo se inauguró en agosto de 1942 en el hospital de Campo Kaiser Richmond en Estados Unidos y sigue utilizándose en la actualidad.[11]

El modelo optimiza la salud poblacional, centrándose en la prestación de servicios en el nivel de atención que resulte más efectivo económicamente. Se potencia la

capacidad resolutiva en el nivel de Atención Primaria (AP) y se minimizan los ingresos hospitalarios que se identifican como "fallos en el sistema"[9]. El modelo consiste en una pirámide con tres niveles en base a la complejidad del paciente, como se puede ver en la Figura 2.2.

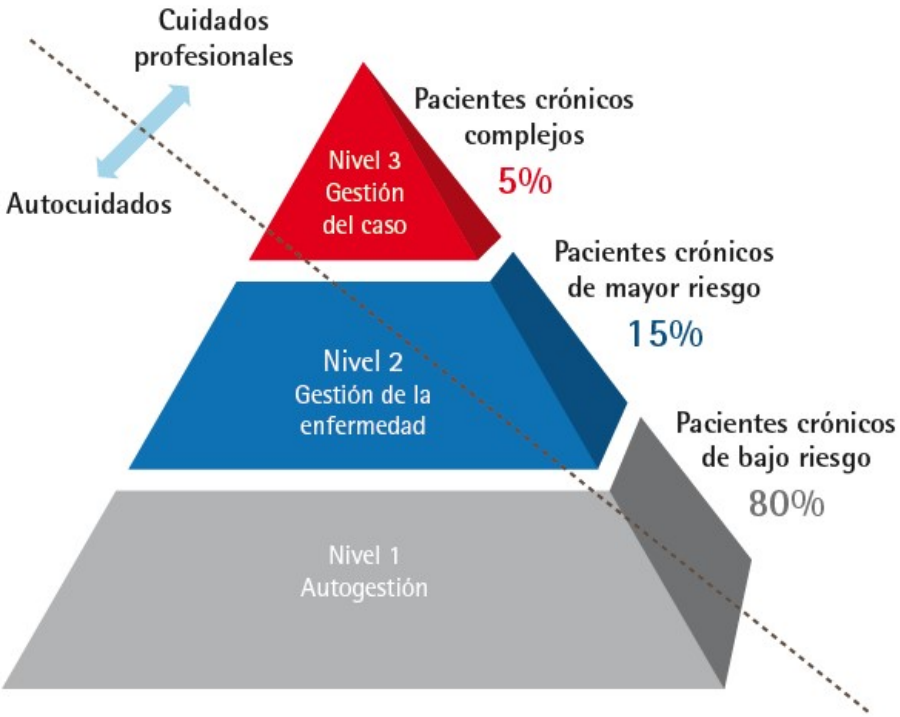


Figura 2.2: *Pirámide de Kaiser Permanente [11]*

La pirámide se separa en tres niveles de intervención y donde se asume que el paciente es crónico [11]. La gestión de la enfermedad por autogestión o por cuidados profesionales es un diferenciador importante en los tres niveles que se explican a continuación:

- **Nivel 1**: 85 % de la población crónica. Son pacientes con riesgo bajo y complejidad baja. La gestión de la enfermedad requerida es fundamentalmente de autogestión.
- **Nivel 2**: 15 % de la población crónica. Son pacientes con riesgo alto pero una complejidad baja, y con consumo de recursos intermedio. Requieren gestión de la enfermedad que combine autogestión y cuidado profesional.
- **Nivel 3**: 5 % de la población crónica. Son los pacientes más complejos y con

consumo de recursos más alto. Son pacientes de alto riesgo que requieren de intervenciones complejas, con frecuente comorbilidad [2] y requieren gestión de la enfermedad con cuidado fundamentalmente profesional.

Existe un cuarto nivel correspondiente a la población general sin patologías crónicas llamado *Prevención y Promoción*. Los objetivos con la segmentación de Kaiser Permanente son evitar que la población sana enferme, fortaleciendo la prevención y promoción de la salud, controlar a los pacientes enfermos y tratar a los pacientes complejos y/o con varias patologías adecuadamente y que se asegure la coordinación de todos los profesionales involucrados [9].

2.2. Análisis Factorial Exploratorio

La segunda parte del estado del arte se centra en Análisis Factorial Exploratorio (EFA o FA para abreviar), una técnica que se utiliza para reducir variables dentro de un problema de Machine Learning. La técnica más común para usar para la reducción de variables es el Análisis de Componentes Principales (PCA). Ambas técnicas son parecidas, pero difieren en la forma de conseguir la reducción de variables. Mientras EFA toma ciertas suposiciones para estimar los factore, PCA solo observa las relaciones lineales existentes y cuánto cada variable aporta a la relación [12].
En análisis de factores se buscan tendencias y correlaciones entre las variables, estas tendencias luego se utilizan para inferir de los datos variables *latentes* o factores [12]. La variable latente no se puede medir directamente con una variable, sino que se calcula con las relaciones entre las variables. La Ecuación 1 representa el modelo EFA:

$$\mathbf{x} = \Lambda\boldsymbol{\xi} + \delta, \tag{1}$$

donde x es un vector de p x 1 con datos de un individuo en p variables, Λ es una matriz p x m con los *loadings* relacionando las p variables a los m factores, $\boldsymbol{\xi}$ es un vector de m x 1 con las variables latentes, δ es un vector de p x 1 con información específica de los *scores* de cada factor. La Ecuación 2 muestra la estructura de la

[2] **Comorbilidad** se entiende como la presencia de más de una enfermedad en una persona al mismo tiempo

covarianza, partiendo de la Ecuación 2:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Psi}, \tag{2}$$

donde $\mathbf{\Sigma}$ es una matriz de covarianza p x p, Λ es igual que para la Ecuación 2, $\mathbf{\Phi}$ es una matriz simétrica con varianzas y covarianzas de las variables latentes, $\mathbf{\Psi}$ es una matriz diagonal con varianzas de factores específicos. Los parámetros en Λ, $\mathbf{\Phi}$ y $\mathbf{\Psi}$ se estiman con los datos observados [13].

La motivación principal para la optimización de la reducción de variables por EFA es la selección del número de variables latentes m. Existen tres métodos diferentes para la elección de m:

- Método basado en **autovalores**
- Método basado en la verosimilitud
- Método basado en la generalización

El método usado en este proyecto es el **método basado en autovalores**, debido a que es el método más interpretable de los tres. Los autovalores explican la cantidad de varianza total explicada por una variable latente. Por lo tanto, cuanto mayor sea el autovalor de una variable latente, más varianza es explicada por este factor. El criterio a seguir para determinar el número óptimo de variables latentes es elegir las variables latentes cuyo autovalor sea mayor a 1. Una vez se ha elegido el valor de m, con la matriz de covarianza de la Ecuación 2 se calculan los valores de correlaciones para cada variable en las nuevas variables latentes [14]. Para entender el funcionamiento de EFA se va a explicar con un ejemplo.

Ejemplo EFA

La Tabla 2.1 es una tabla de correlaciones de siete programas de la televisión española. Al ser un ejemplo didáctico, los datos no son reales. Esta tabla es la matriz $\mathbf{\Sigma}$ en la Ecuación 2. Cada valor en la tabla representa una correlación, que equivale a la relación entre dos variables. Por ejemplo, hay una correlación del 0.7 entre Final Roland Garros y Premio Formula 1. Esto significa que la probabilidad que una persona vea la final de Roland Garros y también el Premio de Formula 1 es alta, del 70 %. A su vez se aprecia que hay correlaciones bajas entre otros programas,

como entre el telediario y la Final de Roland Garros. Se pueden observar que hay correlaciones altas entre personas que ven contenido deportivo y entre personas que ven programas de concursos o diarios. Como hipótesis se predice que se puedan crear dos factores que agrupen las variables de esta manera.

Tabla 2.1: Ejemplo FA - Correlaciones en visualizaciones de programas

	Final Roland Garros	El Hormiguero	Deportes Cuatro	Premio Formula 1	Pasapalabra	Me resbala	Telediario
Final Roland Garros		0,4	0,6	0,7	0,1	0,2	0,1
El Hormiguero	0,4		0,2	0,1	0,4	0,5	0,2
Deportes Cuatro	0,6	0,2		0,5	0,1	0,1	0,4
Premio Formula 1	0,7	0,1	0,5		0,1	0,2	0,2
Pasapalabra	0,1	0,4	0,1	0,1		0,6	0,2
Me resbala	0,2	0,5	0,1	0,2	0,6		0,2
Telediario	0,1	0,2	0,4	0,2	0,2	0,2	

El ejemplo contiene siete variables, y se quiere reducir este número para explicar la mayor cantidad de varianza en el mínimo número de factores. Siguiendo los parámetros de la Ecuación 1, $p = 7$. El siguiente paso es determinar el número de factores óptimo a reducir, m.

Para elegir el valor de m, se deben obtener los autovalores. En la Tabla 2.2, se observan estos valores para el ejemplo. El Factor 1 tiene un autovalor de 3,05, y al tener un valor de variables (p) de 7:

$$\%varianza = \frac{autovalor * 100}{p} = \frac{3,05 * 100}{7} = 43,57\,\%$$

Haciendo el mismo cálculo para todos los autovalores se obtiene la información de la Tabla 2.2. Usando las siete variables iniciales no se pierde información (% Acumulada = 100). Usando dos factores se pierde información, pero se simplifica el problema, y se explica un 63 % de información.

Tabla 2.2: Ejemplo FA - Autovalores y Varaianza Explicada por Factor

Factor	Autovalores	Varianza %	% Acumulada
1	3,05	43,57	43,57
2	1,35	19,29	62,86
3	0,87	12,43	75,29
4	0,59	8,43	83,71
5	0,44	6,29	90,00
6	0,38	5,43	95,43
7	0,32	4,57	100,00

Siguiendo el criterio de los autovalores explicado anteriormente, se determina $m = 2$, ya que solo las dos primeras variables latentes tienen su valor de autovalor mayor a 1. Se va a explicar un 63 % de la varianza del problema original. Ya se puede calcular la matriz $\mathbf{\Lambda\Phi\Lambda'}$, que contiene las correlaciones existentes entre las variables de entrada (las 7 emisiones de televisión) y las variables latentes (los dos factores resultantes). Se observa el resultado en la Tabla 2.3.

Tabla 2.3: Ejemplo FA - Matriz Correlaciones entre Variables y Factores $\mathbf{\Lambda\Phi\Lambda'}$

	Factor 1	Factor 2
Final Roland Garros	0,7	0
El Hormiguero	0,1	0,5
Deportes Cuatro	0,4	0,2
Premio Formula 1	0,8	0,1
Pasapalabra	0,1	0,6
Me resbala	0,2	0,7
Telediario	0,1	0,3

Los resultados de la tabla explican lo observado al principio del ejemplo. El primer factor explica las variables de emisiones deportivas como son la final de Roland Garros, Deportes Cuatro y el premio de Formula 1. El segundo factor explica los

programas de concursos.

El beneficio principal de haber reducido las variables es la simplificación del problema. En muchas ocasiones, antes de clusterizar datos en grupos se reducen las dimensiones del problema.

2.3. Clustering de K-Means

El análisis de clusters consiste en una serie de algoritmos no-supervisados que agrupan objetos similares en grupos llamados clusters. El objetivo final de hacer un clustering es el de obtener clusters que sean diferentes unos de otros, y donde los objetos dentro de cada cluster sean similares. En la Figura 2.3, se observa como los datos de dos dimensiones (x e y) se han separado en cuatro clusters, claramente distintivos unos de otros [15].

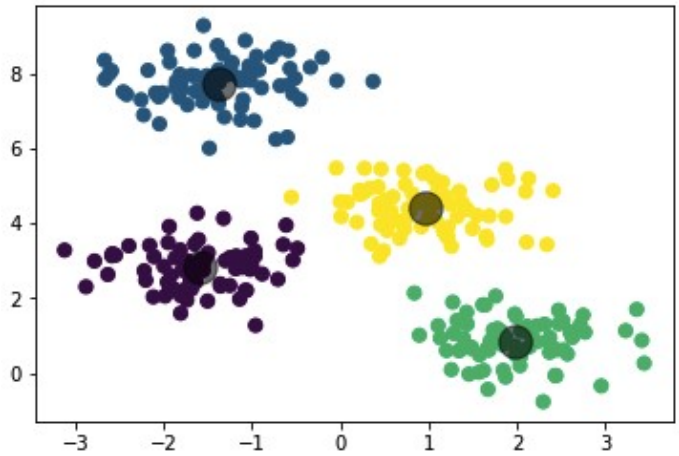


Figura 2.3: *Ejemplo cuatro clusters en datos [16]*

Hay dos algoritmos de clustering principales: clustering Jerárquico y clustering de K-Means. Las principales diferencias entre ambos algoritmos se explican en la Tabla 2.4. La principal diferencia es que para clusterizar usando K-Means es necesario conocer previamente el número de clústers por los que quieres dividir los datos, K. En el caso del clustering Jerárquico esto no es necesario, siendo una ventaja para utilizar este algoritmo de clustering [17].

Tabla 2.4: Tabla de Características de K-Means y Clustering Jerárquico [17]

K-Means	Jerárquico
Necesita saber número de clústers (K)	No necesita número de clusters previamente
Se garantiza la convergencia	Clusters anidados organizados como un árbol
Clusters de tamaños y formas diferentes	Requiere mucho almacenamiento computacional

Debido a que el proyecto es de Big Data, el algoritmo de clustering elegido es K-Means, principalmente por el problema de almacenamiento y latencia que originaría utilizar clustering jerárquico con millones de datos. K-Means se puede utilizar para agrupar datos que no han sido etiquetados explícitamente.

¿Cómo funciona el algoritmo de K-Means?

Se asume que se tienen datos de entrada con p variables de entrada y N observaciones: $\{(x_{1e}, \cdots, x_{pe})\}_{e=1,N}$, como los datos de la Figura 2.4.

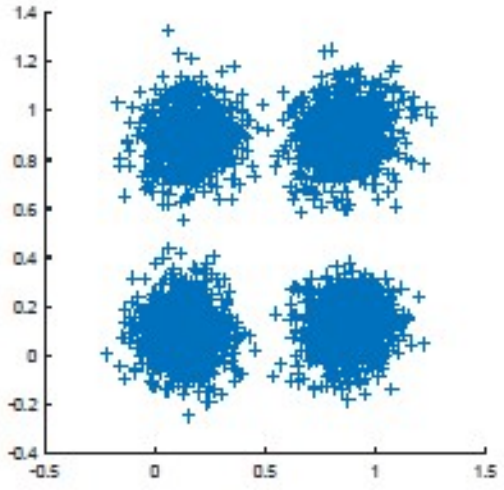


Figura 2.4: *Datos para hacer Clustering por K-Means [18]*

El primer paso es escalar todas las variables. Esto es crítico en K-Means, todas las variables deben estar normalizadas, para que los clusters creados no estén descompensados por que una variable pese más que otra.

El siguiente paso, es seleccionar el número de clusters. Para ello, se itera el algoritmo de K-Means con diversas opciones de K. En el ejemplo se ha iterado con K de 2 a

5, como se puede observar en la Figura 2.5. Se observan unos puntos naranjas que son los *centroides* de los datos, hay tantos como valor K.

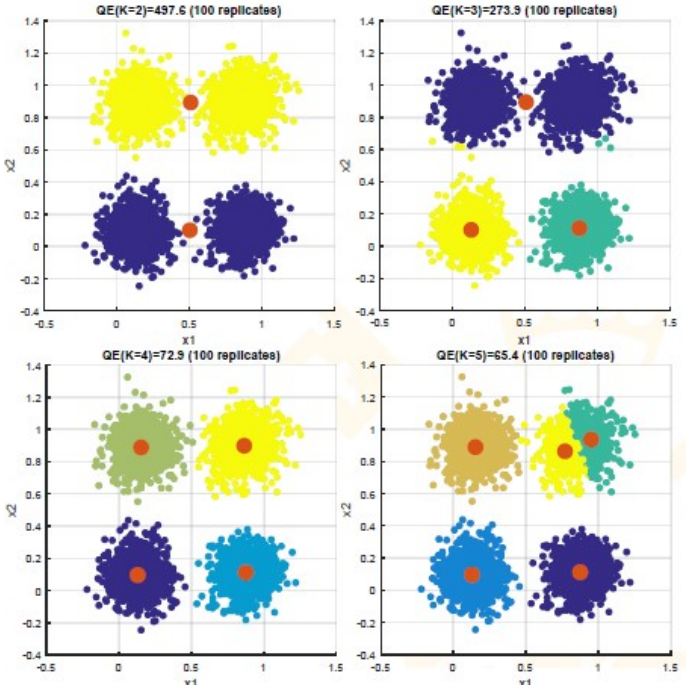


Figura 2.5: *K-Means con distinto número de K [18]*

En este ejemplo, se ve que el número de clústers óptimo es cuatro. Para confirmar la hipótesis se debe calcular el error de cuantización (QE) para cada valor de K. El valor óptimo de K será el valor que minimice el error, con el método del codo. El error de cuantización es la suma de la distancia entre cada punto y su centroide, como se ve en la Ecuación 3.

$$QE = \sum_{i=1,K} QE_i = \sum_{i=1,K} \sum_{\substack{e \in L,S \\ e \in \text{ Clusteri}}} \|\mathbf{x}_e - \mathbf{c}_i\|^2 \tag{3}$$

$\mathbf{C}_i$: Prototipo de cluster (centroide)

El valor de K óptimo es 4, como se ve en la Figura 2.6. Según el **método del codo**, se debe elegir el valor de K donde se estabilice el QE. En este caso, el valor de K es muy claro usando el método del codo y QE, sin embargo hay casos donde este método es subjetivo y no presenta una única solución posible. Por ello, este método sirve para hacerse una idea de cuantos clusters podrían tener los datos, y aparece el análisis de *Silhouette*.

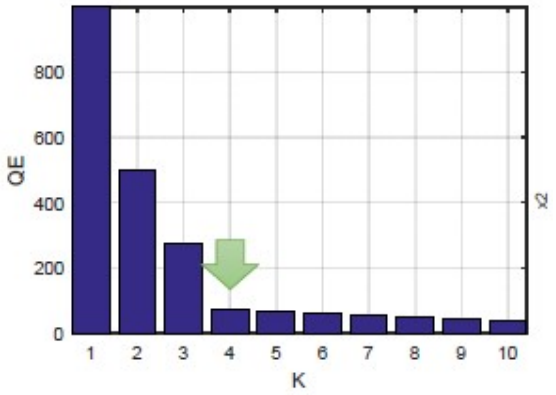


Figura 2.6: *Método del codo y QE para la elección de K [18]*

El **análisis de Silhouette** determina el grado de separación entre clusters. Se obtienen valores en el rango [-1,1], 0 si los datos están muy cerca del cluster vecino, 1 si los datos están muy lejos del cluster vecino y -1 si lo datos se han asignado al cluster incorrecto. Por lo tanto, el valor óptimo de K será aquel que tenga un valor medio de Silueta máximo. En el caso del ejemplo de la Figura 2.4, se obtiene un valor medio de Silhouette de 0.79 con K = 4. Este valor es el máximo, y por lo tanto el número de clusters elegido es cuatro.

Una vez se han escalado los datos y se ha elegido un número óptimo para K, se realiza el K-Means a los datos. El **algoritmo de K-Means** sigue los siguientes pasos [15]:

1. Asignar a cada observación un cluster aleatoriamente
2. Calcular los centroides de los clusters, que se solaparán prácticamente debido a que los clusters han sido asignados aleatoriamente a cada observación
3. Cada observación se le asigna el centroide más cercano, usando como medida la distancia euclídea, que se calcula como muestra la Ecuación 4

$$D\left(\mathbf{x}_u, \mathbf{x}_v\right) = \sqrt{\sum_{j=1,p} \mid d_{x_j}\left(x_{ju}, x_{jv}\right) \mid^2} \tag{4}$$

donde $\mathbf{x}_u$ es una observación y $\mathbf{x}_v$ es un centroide. El subíndice j se refiere a una variable, por lo tanto se suma la distancia media entre dos puntos en cada una de sus variables. En el ejemplo, habrá dos iteraciones, ya que $p = 2$ (x1 y x2)

4. Repetir el punto 2 y 3, hasta que los centroides no se muevan. En este momento se habrá convergido el problema, cada punto están en el cluster que le corresponde

5. Se ha acabado la clusterización de K-Means

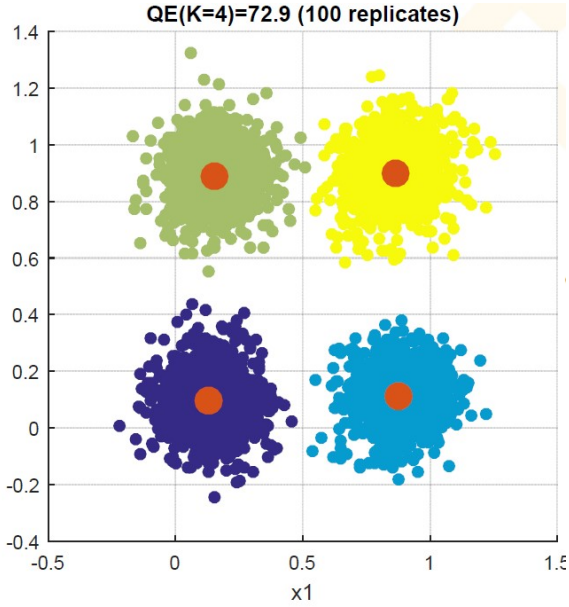


Figura 2.7: *K-Means terminado con K=4 para los datos de entrada [18]*

Con esto, se han explicado las dos técnicas de Machine Learning y el aspecto médico del proyecto. En el siguiente apartado se desarrollará el proyecto.

3. Descripción del Trabajo

En este capítulo se va a describir de forma detallada el proceso de análisis y diseño del proyecto. Se empezará haciendo una introducción al Proyecto, seguido de una descripción del Pipeline y Arquitectura del proyecto y finalizando con un análisis del sistema. El tercer y último apartado es el más extenso, donde se describen todos los procedimientos realizados para la obtención del modelo final y resultados, que se explican en el Cápítulo 4.

3.1. Introducción y Bases del Trabajo

El trabajo busca segmentar la población andaluza en un número finito de clusters para diferenciar distintos grupos de cronicidad y niveles de cuidado. El Servicio Andaluz de Salud (SAS) tiene diversas bases de datos (BBDD) con datos desde el 2001. Las bases de datos están organizadas por área sanitaria o utilidad y contienen una gran cantidad de datos. Una tabla dentro de una base de datos puede llegar a contener 15 millones de registros.

Para acceder a los datos de las BBDD del SAS se ha usado OracleSQL Developer. El proyecto, tal y como se ve explicado en el Pliego ([1] - Caso 4) ingesta los datos en la plataforma de Stratio, para luego realizar las técnicas de Machine Learning y las visualizaciones finales ahí. En este Trabajo de Fin de Máster, se va a segmentar una muestra de 6917 pacientes del Servicio de Salud Andaluz elegida aleatoriamente y anonimizada. Esta muestra es una representación de la población total de Andalucía, y los procesos aplicados en la muestra se podrán ejecutar en una cantidad mayor de datos.

En el siguiente apartado se analiza el diseño del sistema del proyecto y el Pipeline que siguen los datos.

3.2. Diseño del sistema

Esta sección se separa en dos apartados. En el primer apartado de la descripción general del sistema se explican las cinco fases necesarias para la ejecución del proyecto. El segundo apartado explica la arquitectura del sistema, centrándose en el software usado para cada fase.

3.2.1. Arquitectura del sistema

La arquitectura del sistema definida a alto nivel es la mostrada en la Figura 3.1. Hay una distinción de cajas de distintos colores. Las cajas rosas (punto 1) explican los procesos previos a la ingesta de datos. Las cajas en verde (los puntos 4 y 5) corresponden con resultados y serán desarrolladas en la Sección 4 para mostrar los resultados. En este apartado se explicarán cada uno de los procesos.

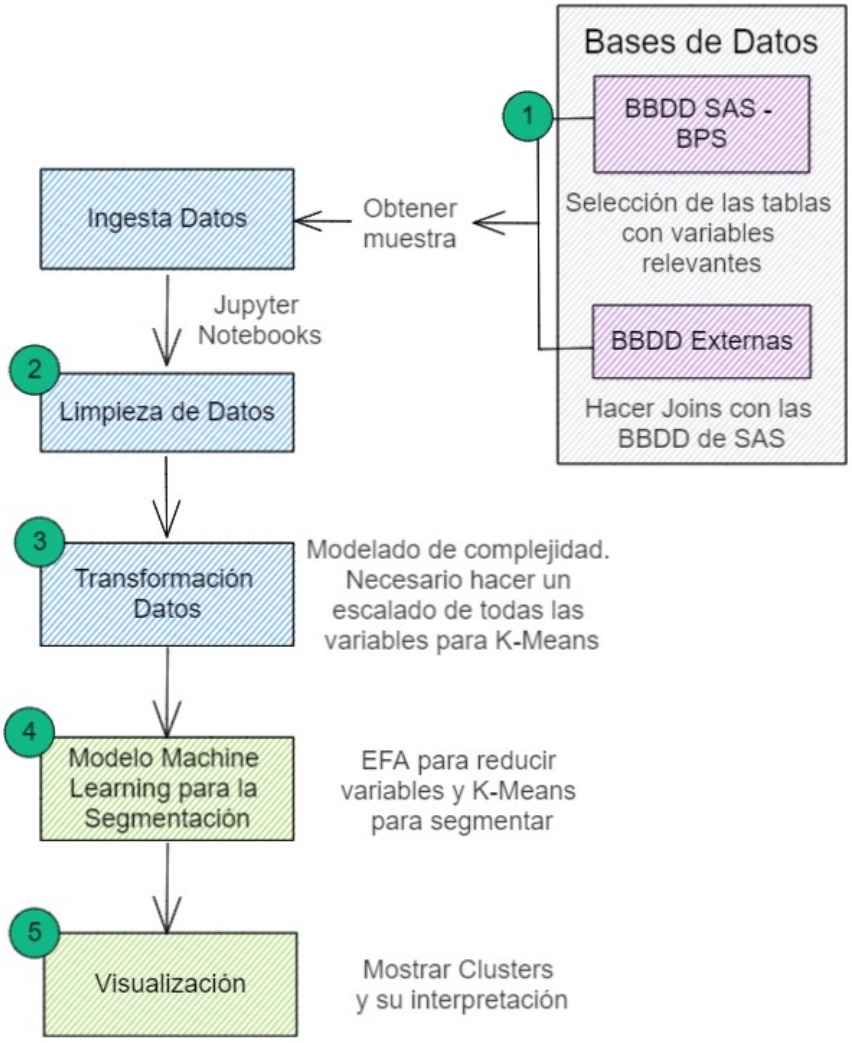


Figura 3.1: *Esquema Resumen Arquitectura del Trabajo*

A continuación, se explican brevemente cada uno de los procesos:

1. **Proceso 1**: Seleccionar las BBDD necesarias. Para ello se analizan las variables interesantes y se buscan tablas relevantes dentro de las BBDD del SAS. En el caso de no obtener toda la información interesante dentro de las BBDD del

SAS, se procede a buscar BBDD externas. Esta parte del proyecto es una de las más largas, se debe estudiar qué variables van a ser críticas para segmentar pacientes crónicos y encontrar estas variables dentro de distintas BBDD. Se utilizará OracleSQL Developer y Confluence para ver los modelos y tablas de cada BBDD.

2. **Proceso 2**: Análisis Exploratorio de los datos. Una vez los datos han sido ingestados de diversas BBDD, se hace una limpieza de datos para asegurar que los datos no contienen un elevado número de nulos o anomalías.

3. **Proceso 3**: Modelado de Complejidad. En este proceso se transforman los datos para prepararlos para la fase de Machine Learning. Es necesario escalar todas las variables para que estén todas dimensionadas en la misma escala y el clustering por K-Means sea correcto.

4. **Proceso 4**: Data Analytics. Se hacen modelos de reducción de variables y de clustering usando los datos colectados y transformados. Se usará el lenguaje de Python, y al usarse una muestra pequeña (en términos de Big Data) se ejecuta en local, usando Anaconda Jupyter Lab.

5. **Proceso 5**: Visualización. Se usa Tableau y Python para la representación de los resultados y el análisis de ellos. El análisis de resultados se explicará en la Sección 4.

3.2.2. Descripción general del sistema

En la Sección 3.3 se profundizará en la selección de variables y tablas. En este apartado se explican las BBDD utilizadas y su proveniencia. Existen dos tipos de bases de datos:

- **BBDD SAS**: Se explicará a fondo en la siguiente sección, pero para el trabajo bastará con la utilización de una de las BBDD del SAS, llamada BPS (Base Poblacional de Salud). La Base de Datos Contiene información de cada paciente, donde la *primary key* es el NUHSA (Número único de Historia de Salud de Andalucía). Cada NUHSA pertenece a un paciente.

- **BBDD Externas**: Para el trabajo se usará una BBDD externa. La BBDD es un archivo de la Distribución Espacial de la Población en Andalucía, utilizado

para obtener información socioeconómica de los pacientes. Se utilizará la información geográfica obtenida en la BBDD BPS de cada paciente para machearla con información en cada zona de Andalucía de la BBDD externa. La información básica de la BBDD externa es:

- **Nombre BBDD**: mallapob
- **Tipo archivo**: .xls
- **Descripción**: Contiene información estadística sobre número de parados, afiliados, pensiones, ingresos o nacionalidades en cada *celda* [3] de Andalucía
- **Número observaciones**: 53243
- **url**: `https://www.juntadeandalucia.es/institutodeestadisticaycartografia/distribucionpob/descargahoja.htm`

3.3. Análisis del sistema

En la sección anterior se han explicado los cinco procesos a desarrollar en el trabajo. A continuación se van a desarrollar uno a uno cada proceso detalladamente, empezando por la selección de variables.

3.3.1. Selección de variables

El primer proceso es seleccionar las variables que se van a utilizar en el modelo. Este proceso tiene dos partes. La primera es determinar qué variables o factores son críticos en la determinación de pacientes crónicos. La segunda parte es buscar las variables dentro de las BBDD disponibles.

Analizando los causantes de enfermedades crónicas, los hábitos de salud del paciente son una importante causa. Estos son los hábitos de actividad física, de dieta, de alcohol o de tabaco. Un paciente con hábitos saludables tiene una menor probabilidad en tener una enfermedad crónica. Otro factor relevante según estudios es la edad. Cuánto mayor es una persona, más probabilidad tiene de tener una

[3] **celda**: Porción de espacio de 250m x 250m, el suelo Andaluz está separado en una malla, y a cada celda se la identifica

enfermedad crónica, y más probabilidad de tener efectos más fuertes. Vinculado con la edad, otras variables a analizar son los diagnósticos y valoraciones de enfermería y resultados de pruebas. Otro factor relevante es la geografía, la zona donde viva el paciente. La información obtenida serán datos socio-económicos del paciente. La primera iteración de variables a explorar dentro de las BBDD se resumen en la Figura 3.2.

GRUPO	VARIABLE A INCLUIR
Constantes	Peso
	Talla
	IMC
	TA
Hábitos de vida (actividad física, dieta, tabaco, alcohol, tóxicos)	Actividad física
	Dieta
	Tabaco
	Alcohol
	Tóxicos
Resultados de pruebas analíticas	Extracciones
	Valor del resultado
Si está siendo dializado y el número de sesiones	Diálisis
Condiciones **socio-económicas**	Ingresos familiares
	IF: Vivienda
	PS: Riesgo social familiar
	Problemas sociales
Datos **medio-ambientales**	Geográfico
	Geográfico: Código ambiental vivienda

GRUPO	VARIABLE A INCLUIR
Edad y sexo	Edad
	Sexo
Diagnósticos médicos	Asistencia Primaria
	CMBDS Hospitalización
	CMA
	HDM
	Urgencias AH
	Consultas AH
Diagnósticos y valoraciones de enfermería	Diagnóstico
	NANDA
	NIC
	NOC
	Índices de Barthel
	Pfeiffer
	Norton
	Braden
	Índice de esfuerzo del cuidador
	Enfermedades raras

Figura 3.2: *Variables seleccionadas para explorar en BBDD*

Para la búsqueda de variables dentro de las BBDD del SAS se utilizan los modelos de cada base en Confluence. Estos modelos contienen la estructura de la base de datos, explicando las relaciones entre cada tabla y las variables dentro de cada tabla. El trabajo es mirar una a una en todas las Bases de Datos (hay más de 20), y estudiar cada tabla (hay cientas en cada BBDD) hasta encontrar una variable dentro de una tabla para cada *variable a elegir* de la Figura 3.2. Después de analizar todas las BBDD del SAS, se obtienen variables (una columna de una tabla) para todas las *variables a elegir* menos para las condiciones socio-económicas y datos medio-ambientales. Para estas variables se necesita una BBDD externa. El proceso de obtención de estas variables se explica a continuación.

Variables Socio-Económicas

La BBDD del SAS contiene información de datos de urgencia, de vacunaciones, de número de patologías entre otros. Sin embargo, no contiene información socio-económica directa por paciente. Una de las BBDD del SAS es la base de datos BDU

que contiene la información de contacto de cada paciente con su identificador, el NUHSA. Una variable de esta BBDD es geográfica, con la coordenadas de cada paciente. Como se muestra en la Figura 3.3, para obtener la información socio-económica se va a hacer un *join* por el campo de las coordenadas geográficas con una BBDD externa.

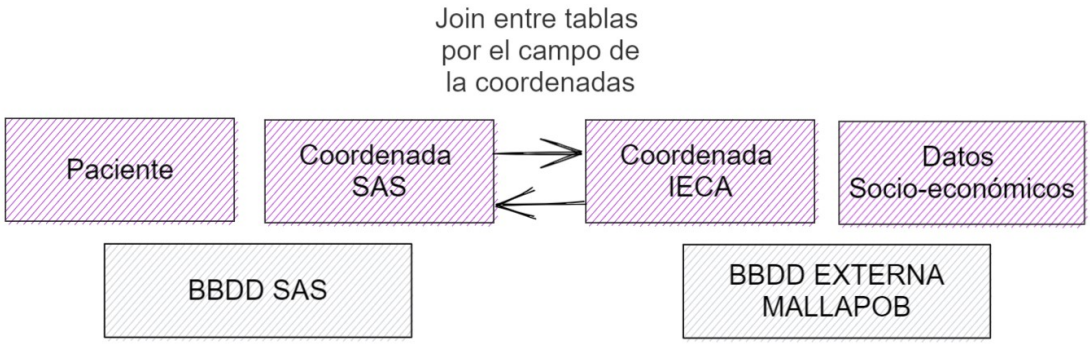
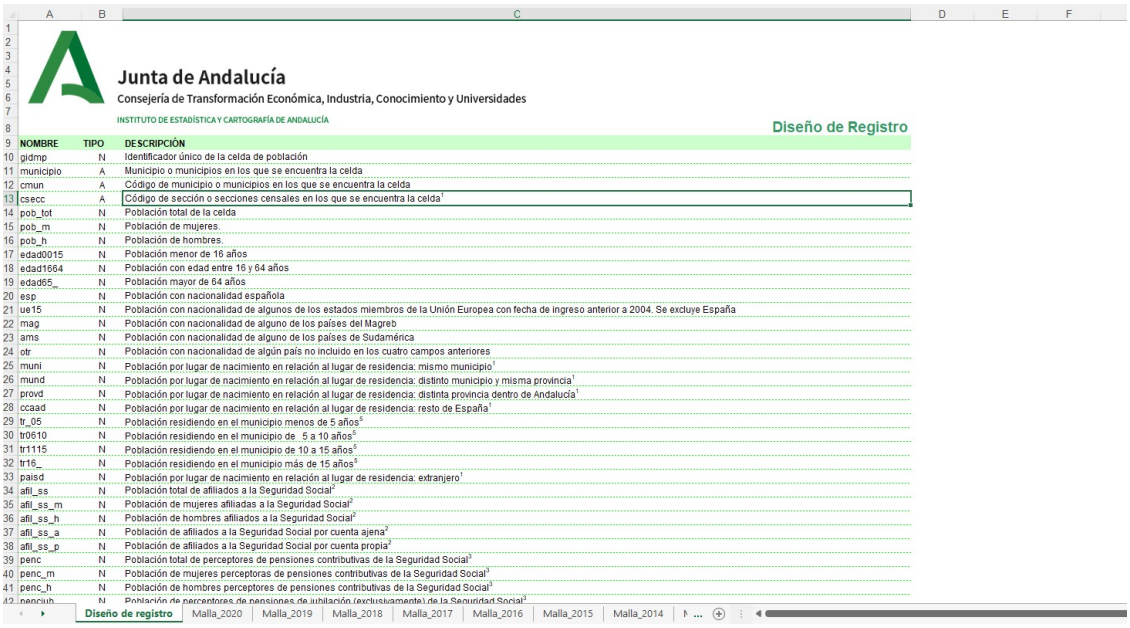


Figura 3.3: *Arquitectura para la obtención de variables socio-económicas*

La BBDD elegida, como se explicó en 3.3.3, es **Mallapob**. Esta BBDD se descarga en formato Excel y contiene datos públicos distribuidos por la Instituto de Estadística y Cartografía de Andalucía (IECA - por ello *Coordenada IECA en 3.3*) que se actualizan cada año, Figura 3.4.



Junta de Andalucía
Consejería de Transformación Económica, Industria, Conocimiento y Universidades
INSTITUTO DE ESTADÍSTICA Y CARTOGRAFÍA DE ANDALUCÍA

Diseño de Registro

NOMBRE	TIPO	DESCRIPCIÓN
gidmp	N	Identificador único de la celda de población
municipio	A	Municipio o municipios en los que se encuentra la celda
cmun	A	Código de municipio o municipios en los que se encuentra la celda
csecc	A	Código de sección o secciones censales en los que se encuentra la celda[1]
pob_tot	N	Población total de la celda
pob_m	N	Población de mujeres.
pob_h	N	Población de hombres.
edad0015	N	Población menor de 16 años
edad1664	N	Población con edad entre 16 y 64 años
edad65_	N	Población mayor de 64 años
esp	N	Población con nacionalidad española
ue15	N	Población con nacionalidad de algunos de los estados miembros de la Unión Europea con fecha de ingreso anterior a 2004. Se excluye España
mag	N	Población con nacionalidad de alguno de los países del Magreb
ams	N	Población con nacionalidad de alguno de los países de Sudamérica
otr	N	Población con nacionalidad de algún país no incluido en los cuatro campos anteriores
muni	N	Población por lugar de nacimiento en relación al lugar de residencia: mismo municipio[1]
mund	N	Población por lugar de nacimiento en relación al lugar de residencia: distinto municipio y misma provincia[1]
provd	N	Población por lugar de nacimiento en relación al lugar de residencia: distinta provincia dentro de Andalucía[1]
ccaad	N	Población por lugar de nacimiento en relación al lugar de residencia: resto de España[1]
tr_05	N	Población residiendo en el municipio menos de 5 años[5]
tr0610	N	Población residiendo en el municipio de 5 a 10 años[5]
tr1115	N	Población residiendo en el municipio de 10 a 15 años[5]
tr16_	N	Población residiendo en el municipio más de 15 años[5]
paisd	N	Población por lugar de nacimiento en relación al lugar de residencia: extranjero[1]
afil_ss	N	Población total de afiliados a la Seguridad Social[2]
afil_ss_m	N	Población de mujeres afiliadas a la Seguridad Social[2]
afil_ss_h	N	Población de hombres afiliados a la Seguridad Social[2]
afil_ss_a	N	Población de afiliados a la Seguridad Social por cuenta ajena[2]
afil_ss_p	N	Población de afiliados a la Seguridad Social por cuenta propia[2]
penc	N	Población total de perceptores de pensiones contributivas de la Seguridad Social[3]
penc_m	N	Población de mujeres perceptoras de pensiones contributivas de la Seguridad Social[3]
penc_h	N	Población de hombres perceptores de pensiones contributivas de la Seguridad Social[3]
pencjub	N	Población de perceptores de pensiones de jubilación (exclusivamente) de la Seguridad Social[3]

Diseño de registro | Malla_2020 | Malla_2019 | Malla_2018 | Malla_2017 | Malla_2016 | Malla_2015 | Malla_2014

Figura 3.4: *Captura BBDD Mallapob - Pestaña inicial*

En las pestañas del Excel están las pestañas de los datos. La pestaña utilizada para el proyecto es la pestaña *Malla_2020*, con la información actualizada al 2020. La Figura 3.5 muestra el formato de los datos. Mallapob contiene información del número de afiliados en cada zona, del número de parados o de la media de ingresos. La primera columna contiene el código *gidmp* y la última (AY) el código *grd_inspir_1k*, las referencias geográficas de la tabla.

	A	B	C	D	E	F	G	AY
1	gidmp	municipio	cmun	csecc	pob_tot	pob_m	pob_h	grd_inspir_1k
2	1	Guadalcanal	41048	4104801(	296	152	144	1kmN1812E2934
3	2	Guadalcanal	41048	4104801(	59	29	30	1kmN1812E2934
4	3	Guadalcanal	41048	4104801(	61	32	29	1kmN1812E2934
5	4	Guadalcanal	41048	4104801(	38	17	21	1kmN1812E2934
6	5	Guadalcanal	41048	4104801(	95	52	43	1kmN1812E2934
7	6	Guadalcanal	41048	4104801(	303	147	156	1kmN1812E2934
8	7	Guadalcanal	41048	4104801(	281	132	149	1kmN1812E2934
9	8	Guadalcanal	41048	4104801(	-1	-1	-1	1kmN1812E2934

Figura 3.5: *Captura BBDD Mallapob - Pestaña Datos*

La tabla Mallapob contiene información socio-económica interesante pero no sigue el sistema de georreferenciación común de GPS, lo cual se debe solucionar. Para ello, se deben entender los códigos gidmp y grd_inspir_1k y ver qué relaciones tienen con el código GPS.

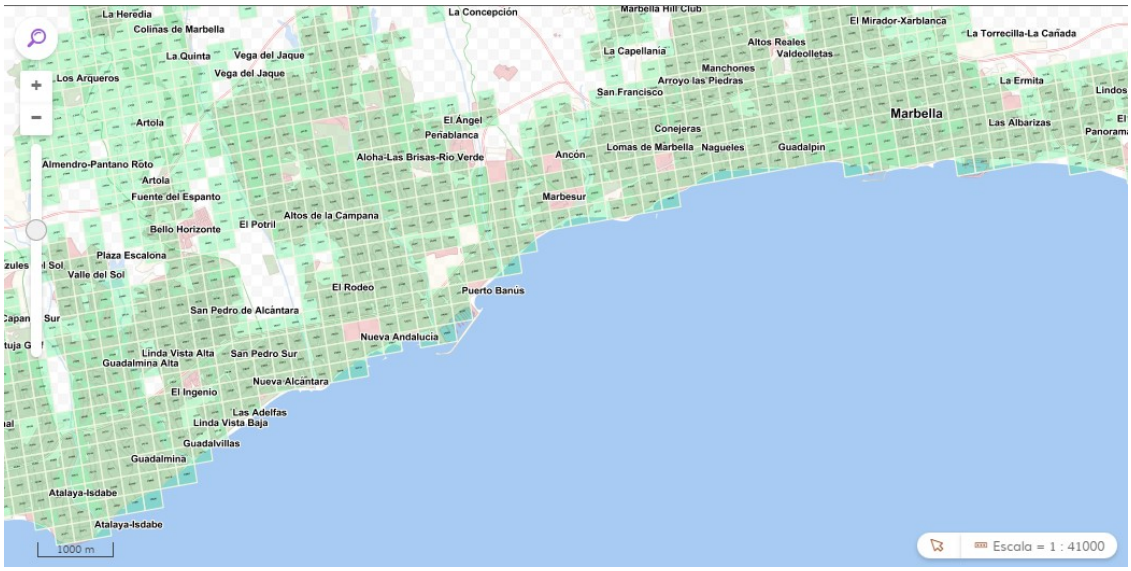


Figura 3.6: *Malla de celdas gidmp sobre Marbella [19]*

Según explicado en [19], la Comunidad de Andalucía está cubierta con una malla de cuadrados, donde cada uno cubre una superficie de 250m x 250m. Cada uno de

estos cuadrados de la malla se le asigna un valor de 1 a 1406473 y el código de cada celda se le denomina código **gidmp**. En la Figura 3.6 se muestra la malla cubriendo la zona de Marbella. Como se aprecia en la imagen, el código gidmp no se refiere a un punto como hace el código GPS, sino a un área. Por lo tanto, una celda gidmp contiene muchos puntos GPS.

A su vez, se debe entender el código grd_inspir_1k o **INSPIRE1K**. De manera parecida al código gidmp, el código INSPIRE1K representa un área en un mapa, pero de $1km^2$ y es un sistema usado para representar a Europa. El código INSPIRE1K proviene de el sistema de georreferenciación EPSG:3035 o ETRS89, mientras que el código GPS proviene del sistema EPSG:4326. Las coordenadas GPS tienen una latitud y una longitud y su eje de referencia es el punto (0,0), el punto donde el Meridiano de Greenwich se cruza con el Ecuador. Las coordenadas del sistema de INSPIRE1K tienen un cambio de eje de referencia al punto 52^{o}N, 10^{o}E con *false easting*: x_0=4321000m y *false northing*: y_0=3210000m. Los cambios de eje de referencia son para centrar al sistema de INSPIRE1K en el centro de Europa, ya que el sistema EPSG:3035 representa a Europa.

A su vez, el campo de coordenadas proveniente de la BBDD del SAS tampoco es el código GPS. En este caso es sistema de georreferenciación elegido es EPSG:25380. Para no perderse entre tantas referencias se resumen las diferentes geo-referencias vistas en la Tabla 3.1.

Tabla 3.1: Resumen sistemas de georreferencia para cada BBDD

	BBDD EXTERNA			BBDD SAS
	GPS	gidmp	INSPIRE1K	SAS
Valor EPSG	4326	-	3035	25380

En conclusión, se debe generar un modelo que haga de diccionario de coordenadas, para poder hacer el join entre la BBDD del SAS y la BBDD Mallapob. Para cada coordenada de la BBDD del SAS, el modelo la traduce a un punto referenciado al sistema de INSPIRE1K, lo escribe en el formato que aparece en el Excel de Mallapob y obtiene y devuelve los datos socio-económicos para ese punto geográfico. El modelo creado se explica en el apartado 3.3.4.

Variables de Enfermería

Las variables socio-económicas ya se pueden obtener con la base externa de Mallapob. Ahora, se van a explicar Las variables de enfermería encontradas. De las variables de *Diagnósticos y valoraciones de enfermería* de la Figura 3.2 solo se van a usar cinco: índice de Barthel, índice de Norton, índice de Pfeiffer, índice de Braden y el índice del esfuerzo del cuidador. Las variables NIC, NOC y NANDA se han descartado al ser índices de enfermería que se han valorado como no relacionadas directamente con los pacientes crónicos.

Los cinco índices que sí se van a usar son índices médicos sobre el riesgo, la dependencia física y el deterioro cognitivo, donde el médico examina al paciente y le da una puntuación dentro de un rango de posibilidades. Para el modelo de segmentación se van a utilizar dos variables de cada índice:

1. **Último índice**: para cada paciente (NUHSA) se busca el índice con fecha más cercana a la actual
2. **Índice máximo**: para cada paciente (NUHSA) se busca el índice máximo del historial de índices

Al obtener estas dos variables se entiende el progreso del paciente. Un paciente con un valor alto de un índice máximo pero un valor bajo para el último índice, puede ser positivo y cambiar de un clúster a otro en menos riesgo.

En el siguiente apartado se analizan las 20 variables a usar y se hace el modelado de complejidad, donde se transforman las variables.

3.3.2. Modelado de Complejidad

En este apartado se utilizan las variables elegidas y encontradas en el apartado 3.3.1, transformándolas para que todas las variables estén escaladas entre 0 y 1. El proceso de escalado es necesario para el clustering de K-Means. Se van a explicar las transformaciones hechas para cada una de las variables individualmente. En el Anexo A, se encuentra una tabla resumen de todas las transformaciones. Además de escalar todas las variables entre 0 y 1, las transformaciones tienen como objetivo orientar las variables. Para ello, en las transformaciones hay veces que se resta la transformación y otras no. De esta manera, los valores transformados cercanos a 0 se refieren a valores de menos riesgo y aquellas cercanas a 1 de mayor riesgo.

Índice de Barthel

Índice de 0 a 100, donde una puntuación de 100 es la de una persona sana, y una puntuación de 0 de una persona con dependencia física total.

Para el modelado se usan dos variables:

1. Último índice de Barthel
2. Índice de Barthel máximo

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 100, & x_i \notin Z^+ \end{cases}, \quad t_i = 1 - \frac{v_i}{100}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, que se asume v_i=100 (paciente sano). La transformación necesita ser restada ya que una persona en riesgo tiene una puntuación de 0, cuando en la transformación debe voltearse, una persona sana tener un valor transformado de 0.

Índice de Braden

Evalúa el riesgo de que un paciente desarrolle una úlcera de presión. Índice de 0 a 24, donde una puntuación de 24 es la de una persona sana sin riesgo, y una puntuación de 0 de una persona en riesgo.

Para el modelado se usan dos variables:

3. Último índice de Braden
4. Índice de Braden máximo

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 24, & x_i \notin Z^+ \end{cases}, \quad t_i = 1 - \frac{v_i}{24}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, que se asume v_i=24 (paciente sin riesgo).

Índice de Norton

Al igual que el índice de Braden, el índice de Norton evalúa el riesgo de que un paciente desarrolle una úlcera por presión. Índice de 0 a 20, donde una puntuación

de 20 es la de una persona sin riesgo, y una puntuación de 0 de una persona en riesgo.

Para el modelado se usan dos variables:

5. Último índice de Norton
6. Índice de Norton máximo

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 20, & x_i \notin Z^+ \end{cases}, \quad t_i = 1 - \frac{v_i}{20}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, que se asume v_i=20 (paciente sin riesgo).

Índice de Pfeiffer

El índice de Pfeiffer consiste en 10 preguntas para analizar el deterioro cognitivo de un paciente. Índice de 0 a 10, donde una puntuación de 10 es la de una persona con signos de deterioro cognitivo, y una puntuación de 0 de una persona sin riesgo. Se suma un punto por cada pregunta fallada.

Para el modelado se usan dos variables:

7. Último índice de Pfeiffer
8. Índice de Pfeiffer máximo

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}, \quad t_i = \frac{v_i}{10}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, que se asume v_i=0 (paciente sin riesgo). En este caso, la variable transformada no es restada, ya que el valor de un paciente sano ya es 0.

Valoración del Esfuerzo del Cuidador

Índice de 0 a 14, donde una puntuación de 14 es la de una persona que requiere cuidador y una puntuación de 0 de una persona sana.

Para el modelado se usan dos variables:

9. Ultima valoración de esfuerzo del cuidador

10. Valoración de esfuerzo del cuidador máxima

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}, \quad t_i = \frac{v_i}{14}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, que se asume v_i=0 (paciente sin riesgo).

Indicador de Actividad Física y de Hábitos Tóxicos

En el caso del indicador de actividad física, hay 16 tipos de actividad física, cada tipo corresponde a un código específico. Por ejemplo, ACT_FISICA_1.0 corresponde a una persona sedentaria. PAra el indicador de hábitos tóxicos se tienen 46 códigos diferentes. Para ambos casos, se va a hacer One-Hot encoding, que consiste en crear tantas variables como códigos hay y asignar un 1 si aparece el código en el paciente y un 0 si no aparece.

Para el modelado se usan dos variables:

11. Último Indicador de Actividad Física

12. Último Indicador de Hábitos Tóxicos

$$v_i = \begin{cases} x_i, & x_i \notin nan \\ NA, & x_i \in nan \end{cases}, \quad t_i = OneHot(v_i)$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, en este caso se asume que t_i será una categoría vacía.

Indicador de Consumo de Alcohol y Tabaco

La variable de consumo de alcohol representa los gramos de alcohol en sangre. La variable de consumo de tabaco el número de cigarrillos consumidos en una semana. Un valor de 0 representa una persona mas sana, en cuanto a hábitos tóxicos.

Para el modelado se usan dos variables:

13. Último Indicador de Consumo de Alcohol

14. Último Indicador de Consumo de Tabaco

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}, \quad t_i = \frac{v_i}{max(x)}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, en este caso se asume que v_i=0. En la transformación se divide por el máximo valor del dataset.

Edad

Para la edad, se obtiene el último registro de la edad del paciente:

15. Edad registrada por última vez

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ mean(x), & x_i \notin Z^+ \end{cases}, \quad t_i = \frac{v_i}{120}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable menos en el caso que no se encuentren mediciones para un paciente, en este caso se asume que v_i sea la media de las edades del dataset. La transformación se divide por 120, al asumir que no habrá ningún paciente con más de 120 años.

Peso

Para la peso, se obtiene el último registro del peso del paciente:

16. Última medición de constante de peso

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ \{ \begin{matrix} PC(x_i), & PC(x_i) < 65 \\ 65, & PC(x_i) \geq 65 \end{matrix} & x_i \notin Z^+ \end{cases}, \quad t_i = \frac{v_i}{max(x)}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable, menos en el caso que no se encuentren mediciones para un paciente. En este caso se asume que v_i sea un peso calculado a partir de la edad: $PC(x_i) = 9 + (2,5 * edad_i)$, con una cuota superior máxima de 65. La transformación tiene en cuenta el máximo de todos los pesos del dataset.

Comorbilidades

Para las comorbilidades, se calcula el número total de comorbilidades (patologías)

de un paciente:

17. Máximo de comorbilidades

$$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}, \quad t_i = \frac{v_i}{100}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable, menos en el caso que no se encuentren mediciones para un paciente, donde se asume que el paciente no tiene comorbilidades $v_i = 0$.

Variables Socio-económicas

Como se ha explicado en el apartado 3.2.1, hay tres variables socio-económicas:

18. Mediana de ingresos de perceptores de pensiones en su comunidad
19. Población total parada en su comunidad
20. Población total afiliada a la seguridad social en su comunidad

$$v_i = \begin{cases} x_i, & x_i \notin nan \\ mean(x), & x_i \in nan \end{cases}, \quad t_i = \frac{v_i}{100}$$

El dato transformado es t_i, el dato de entrada es v_i. v_i es el valor de la variable, menos en el caso que no se encuentren mediciones para un paciente, donde se asigna el valor medio del dataset. La transformación tiene en cuenta el máximo del dataset. El modelo usado para la transformación y obtención de estas variables se desarrolla en el apartado 3.3.4.

Al realizar la transformación de las 20 variables de entrada, se obtienen 80 variables transformadas. Se van a estudiar ahora las variables en un análisis exploratorio.

3.3.3. Análisis Exploratorio de Datos

En el apartado anterior se han transformado todas las variables para la primera iteración de la segmentación de pacientes crónicos. La exploración de datos se va a hacer en una nuestra de 6917 observaciones, y 81 columnas (una columna del NUHSA y las 80 columnas de las variables transformadas). De estas 81 variables, 19 son variables continuas y 62 son variables discretas. Las variables discretas toman los valores 0 o 1, tras el encoding One-Hot. Las variables continuas toman valores

entre 0 y 1 tras las las transformaciones explicadas en el aparatado 3.3.2. El dataset no contiene valores nulos ni valores extremos.

En primer lugar, dataset se encuentra en el mapa de Andalucía, obteniéndose la distribución de puntos de la Figura 3.7. Se aprecian puntos concentrados en las zonas urbanas de las capitales de provincia.

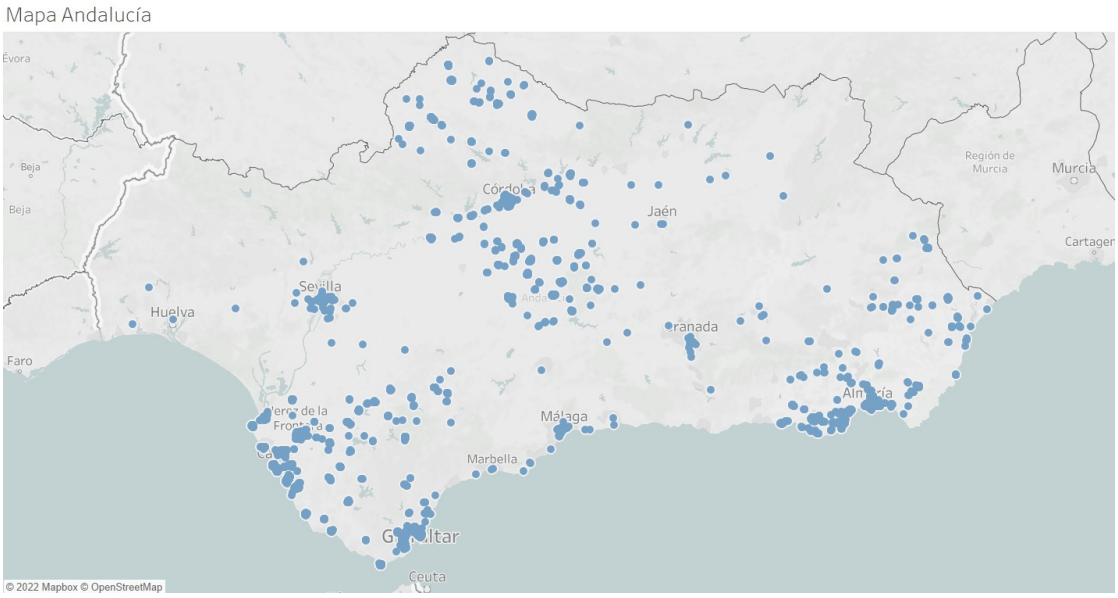


Figura 3.7: *Mapa Andalucía del Dataset estudiado*

Para este análisis inicial se muestran diversos gráficos. En primer lugar, se observa en la Figura 3.8, como hay una relación positiva entre la edad y el número de comorbilidades. Esto es esperado, pues una persona mayor, tiene mayor probabilidad de tener problemas de salud y por lo tanto un mayor número de patologías.

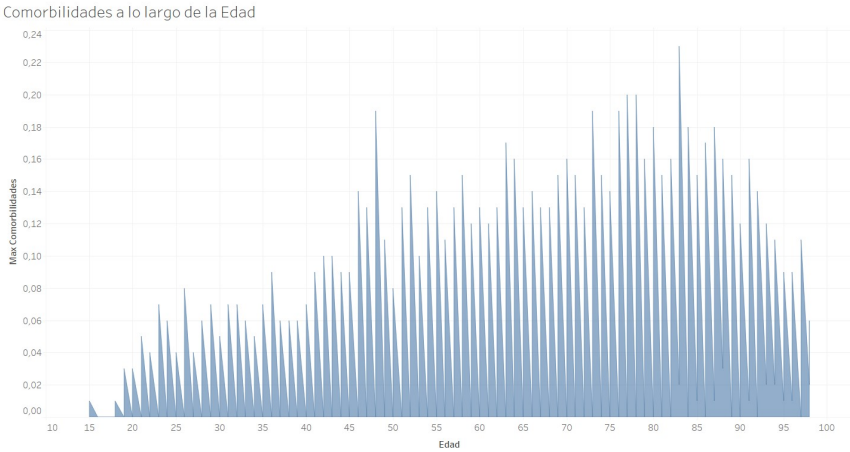


Figura 3.8: *Número de comorbilidades con respecto a la edad de los pacientes*

La edad media del dataset es de 50 años. Para determinar la relevancia de la localización de los pacientes se estudian ciertas variables por provincia. En primer lugar, como muestra la Figura 3.9, está la edad media de los pacientes por provincia. En este caso, la edad media máxima del dataset corresponde a Córdoba con una media de 57,9 años y la edad media mínima a Huelva con una media de 48,7 años. La diferencia de más 9 años de edad media entre provincias intuye que la localización sí es relevante.

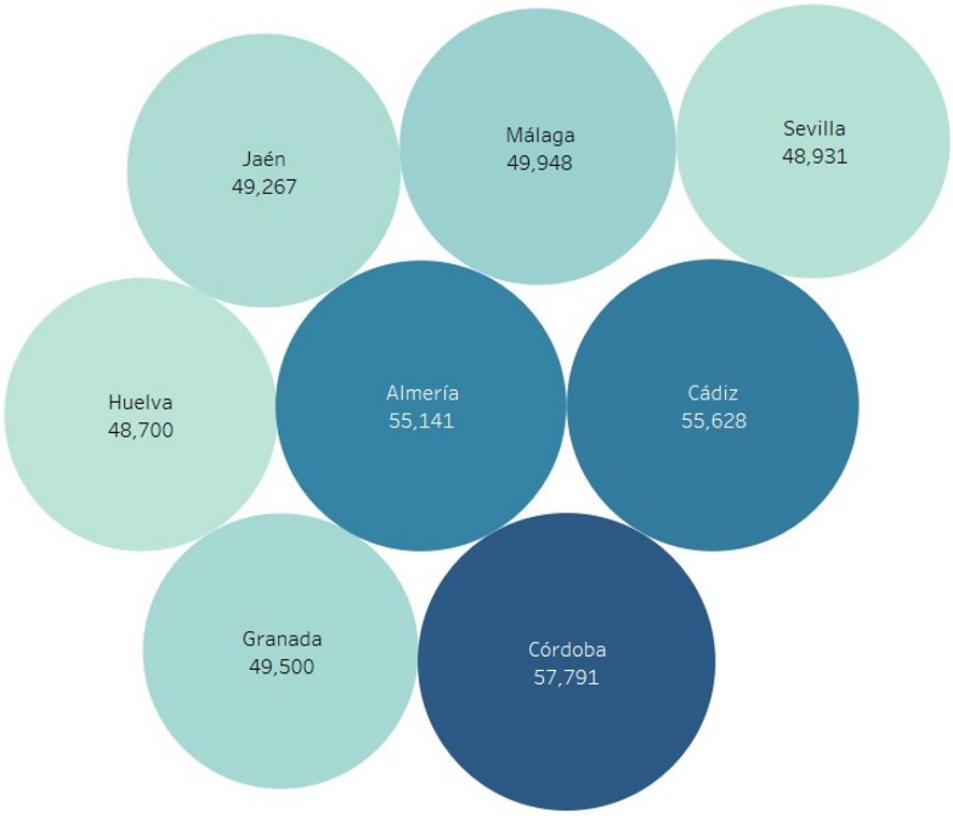


Figura 3.9: *Análisis Exploratorio - Edad Media por Provincia*

Siguiendo con la diferencia por provincias, se analizan las variables socio-económicas. Estas variables se espera sean claramente distintivas entre provincias, ya que dependen de las coordenadas geográficas. En la Figura 3.10 se muestra la población parada media por provincia. Al haberse transformado los datos, los valores medios son índices de 0 a 1, siendo un valor de 0 una zona sin población parada. De esta forma, Cádiz es la provincia con mayor número de parados y Huelva la provincia con el menor número de parados dentro de dataset. En esta variable socio-económica se observan diferencias altas entre las provincias.

Figura 3.10: *Análisis Exploratorio - Población Parada media por Provincia*

En la siguiente figura, la Figura 3.11 se han estudiado las 3 variables socio-económicas del modelo por provincia: mediana de pensiones, población total parada y población total afiliada. Hay mucha variación entre provincias para las 3 variables, lo cual es muy positivo para utilizar estas variables dentro del modelo de segmentación. Sevilla es la provincia con una media mayor de pensiones y población afiliada. Estos números elevados de Sevilla pueden ser justificados por ser la capital de la comunidad y la provincia andaluza con mayor número de habitantes. A su vez, Huelva tiene las menores medias para las tres variables, siendo la provincia con el menor número de habitantes de Andalucía.

Con este análisis exploratorio de las variables transformadas se concluyen varios puntos importantes. Por un lado, las transformaciones de todas las variables se han escalado correctamente y el dataset está limpio. Por otro lado, se han apreciado relaciones interesantes y variación en las variables. Esto es muy positivo para la

reducción de factores en el modelo de Machine Learning.

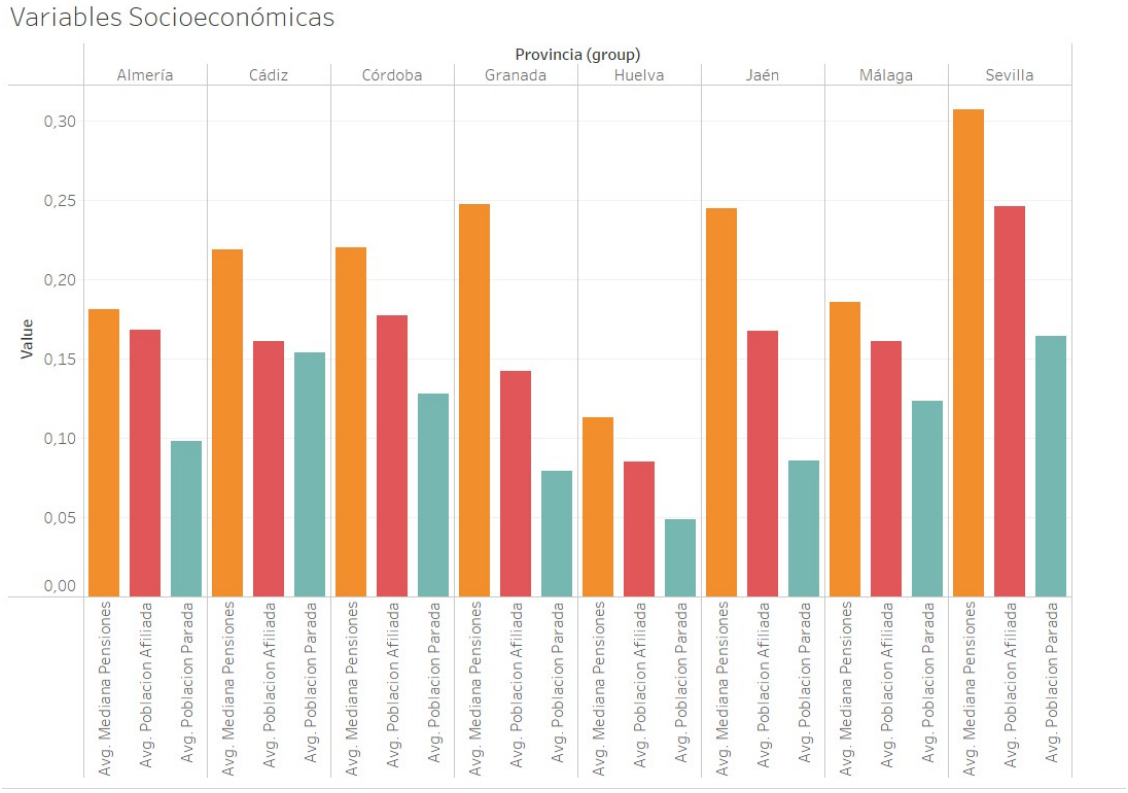


Figura 3.11: *Análisis Exploratorio - Variables Socioeconómicas por Provincia*

En el siguiente apartado, se analizan los modelos desarrollados en Python.

3.3.4. Modelos Desarrollados

Se han desarrollado dos modelos para la ejecución del proyecto. El primer modelo es para la transformación de los datos y el segundo para las técnicas de Machine Learning ejecutadas en el dataset de datos transformados. Del primer modelo destaca la transformación de datos socio-económicos, la cual se va a explicar a continuación.

Modelo de georreferenciación

El modelo de georreferenciación es necesario para la obtención de las variables socio-económicas. Es un script de Python que sirve de punto de conexión entre las dos bases de datos, la del SAS y la de Mallapob. De la base de Datos de SAS se obtienen los datos geográficos como se muestra en la Figura 3.12, coordenada de latitud,

coordenada de longitud y código postal. Si el paciente no ha dado sus datos de dirección, aparece como NaN o -1.

COD_POSTAL	COD_LONGITUD	COD_LATITUD
4008	548368.83	4078207.71
4007	NaN	NaN
4720	NaN	NaN
4860	-1	-1
4610	-1	-1
...	...	...
14550	NaN	NaN
14001	344358.93	4195296.64
14011	342120,12	4195286,13
14100	NaN	NaN
14900	367688.19	4141501.05

Figura 3.12: *Datos geográficos de entrada SAS*

Como se ha explicado en el apartado 3.3.1, las coordenadas provenientes de la BBDD del SAS siguen el sistema EPSG 25830 y las de la BBDD de Mallapob EPSG 3035. Para poder convertir las coordenadas del SAS a las de la BBDD de Mallapob se crea una función diccionario con la librería pyproj de Python. La función es muy sencilla, de entrada se tienen las coordenadas de la Figura 3.12, y de salida las coordenadas en la nueva referencia. La función se encuentra en el Anexo B para referencia adicional. A su vez, la función devuelve las coordenadas como un punto, mientras el código INSPIRE1K se refiere a un área de 1km^2. Las transformaciones necesarias para pasar al formato de la BBDD de Mallapob también se encuentran en el Anexo B.

Una vez se obtienen todas las coordenadas traducidas y formateadas se buscan en la tabla de Mallapob los valores objetivo. Las variables de Mallapob seleccionadas son:

18. **Mediana de ingresos de perceptores de pensiones en su comunidad**
    Variable de Mallapob: penc
    Descripción: Población total de perceptores de pensiones contributivas de la Seguridad Social

19. **Población total parada en su comunidad**

Variable de Mallapob: demp_pr

Descripción: Población total parada del Servicio Andaluz de Empleo

20. **Población total afiliada a la seguridad social en su comunidad**
    Variable de Mallapob: afil_ss
    Descripción: Población total de afiliados a la Seguridad Social

Al hacer el modelo se dieron varios problemas y se solucionaron de la siguiente manera:

1. **INSPIRE1K vacíos**: Algunos pacientes no tienen coordenadas en la BBDD del SAS y por lo tanto no se tienen coordenadas INSPIRE1K. Para resolver esto se hicieron dos iteraciones:

    a) En la primera iteración, a los pacientes sin INSPIRE1K se les asigna la media de los valores del dataset

    b) En la segunda iteración, se utiliza el código postal para obtener un código INSPIRE1K

2. **Varios códigos gidmp por el mismo código INSPIRE1K**: En la BBDD Mallapob se organizan los datos por celda gidmp, que es más pequeña que la celda INSPIRE1K. Por lo tanto, aparecen repetidos códigos INSPIRE1K en la BBDD Mallapob. La pregunta es ¿qué valores de pensiones, número de afiliados y parados se deben elegir? Para el modelo, se escogen los valores máximos.

El modelo para obtener las variables socio-económicas con el campo de georreferencia se ha terminado de explicar. El segundo modelo, el modelo de Machine Learning se va a explicar a continuación.

Modelo de Machine Learning

Este segundo modelo es la clave del proyecto, es el modelo que forma los clusters de pacientes crónicos. Este modelo se expande en la siguiente Sección 4, que continúa con el desarrollo del proyecto pero centrándose en los experimentos y resultados. Se va a explicar el orden del modelo:

1. **Leer el dataset** de las 80 variables transformadas

2. **Realizar la reducción de variables**

   El método para la reducción de variables es el análisis Factorial Exploratorio. Se debe elegir el número de factores óptimo.

3. **Realizar el Clustering por K-Means**

   Una vez se haya reducido las dimensión del dataset, se elige el valor K del número de clusters óptimo y se realiza la clusterización.

4. **Generar Visualizaciones**

   El modelo genera diversas visualizaciones, tanto de la sección de EFA como de K-Means para ayudar en las tareas analíticas.

De esta manera se termina de explicar la descripción del proyecto, pero no se termina el desarrollo del mismo. En la siguiente sección se desarrollan las técnicas de Machine Learning realizadas al dataset generado.

4. Experimentación

En la sección 3 se ha explicado el proyecto: las bases de datos utilizadas, las variables seleccionadas, las transformaciones a cada variable y los pasos para segmentar la población en grupos de pacientes crónicos. En esta sección, se van a mostrar los resultados obtenidos, así como analizarlos.

4.1. Análisis de la Reducción de Variables

Una vez se han elegido las variables y estas han sido transformadas, se debe proceder a una reducción de variables para su posterior segmentación. Se ha elegido hacer una reducción por análisis de factores. Para la elección del número de factores, se ha seguido el criterio de Kasiser, donde si el autovalor es mayor a 1 se considera al factor un factor significativo. Por lo tanto, se van a formar tres factores, reduciendo las variables de 80 variables a tres.

Para el análisis de factores se usa la librería de *factor_analyzer* de Python. Se debe asegurar que cada una de las ochenta variables tiene una desviación típica mayor a 0, esto es que cada variable no sea una repetición de valores. Como se puede ver en la línea 8 del código, una vez se eliminan las variables sin varianza, quedan 30 variables. La mayor parte de las variables eliminadas son variables de códigos de hábitos tóxicos, al contener todo 0s. La reducción de variables se va a desarrollar con 30 variables.

```
1  # Cargar csv con todas las variables transformadas
2  df = pd.read_csv('Transformacion_resultados.csv')
3  # Eliminar las variables con desviaci n t pica = 0
4  nunique = csv.nunique()
5  cols_to_drop = nunique[nunique == 1].index
6  df = df.drop(cols_to_drop, axis=1) # 50 variables, quedan 30
7  # Hacer el analisis de Factores con 3 factores
8  nfact = 3
9  fa = FactorAnalyzer(rotation=None, n_factors = nfact)
10 fa.fit(csv)
11 rotator = factor_analyzer.rotator.Rotator()
12 res = rotator.fit_transform(fa.loadings_)
```

Listing 1: Factor Analyzer en Python

Como se ve en el código superior, se han elegido un total de tres factores. La elección del número de factores (m) es crítica. Para la elección se ha seguido el criterio de los autovalores explicado en el apartado 2.1 del estado de arte. Los tres primeros autovalores tenían un valor superior a 1, y por ello se han elegido tres factores para la realización del EFA.

Una vez ya se han obtenido los loadings de los tres factores, se va a analizar qué variables son las más relevantes para el modelo y qué correlaciones existen para cada factor Para ello, en la figura 4.1 se muestran las correlaciones entre las diferentes variables y los factores. En la tabla aparecen solo aquellas variables que tienen una correlación con algún factor superior a 0.1, apareciendo un total de 16 variables. Las variables que no aparecen no son relevantes para el modelo.

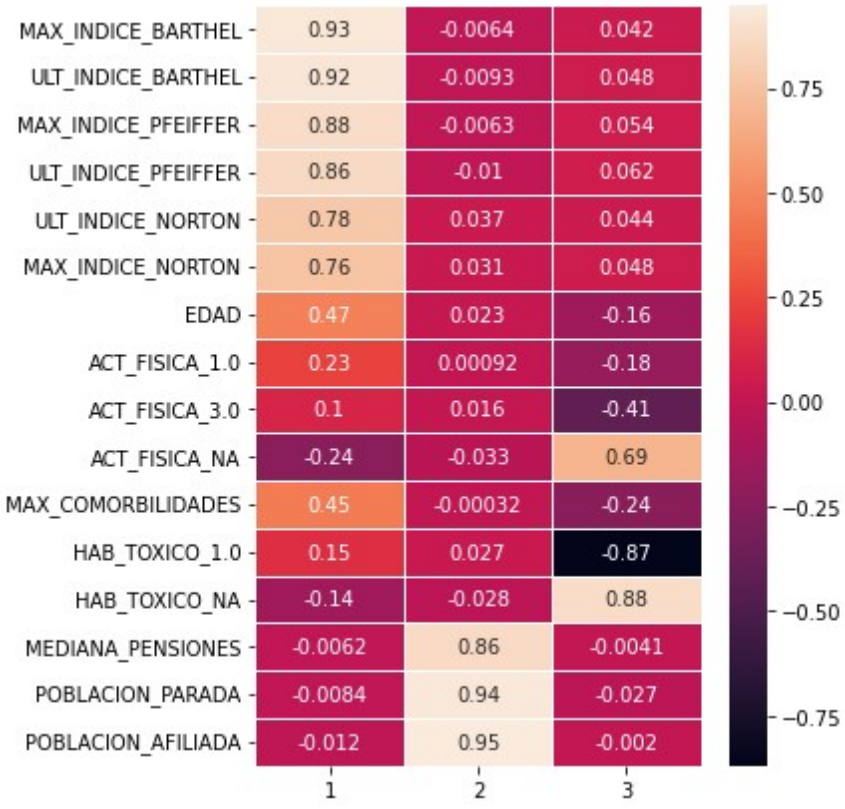


Figura 4.1: *Mapa de calor de correlaciones entre variables y los factores*

Los resultados obtenidos del análisis de factores son muy interesantes. Se aprecian diferencias entre los 3 factores, obteniéndose correlaciones altas diferentes en cada factor. En el primer factor (grupo 1) las variables con alta correlación (y por lo tanto gran importancia) son *MAX_INDICE_BARTHEL*, *ULT_INDICE_BARTHEL*,

MAX_INDICE_PFEIFFER, *ULT_INDICE_PFEIFFER*, *MAX_INDICE_NORTON*, *ULT_INDICE_NORTON*, *EDAD* y *MAX_COMORBILIDADES*. Estas 7 variables definen el estado de salud del paciente y están fuertemente relacionadas con la edad. A este primer factor se le denomina el factor de **Índices**, referiéndose a los diferentes índices de salud presentes.

El segundo factor está formado por tres variables con correlaciones muy altas (0.86, 0.94 y 0.95) y sin ninguna otra variable por encima de 0.03 como valor de correlación. Este grupo engloba a las tres variables socioeconómicas y geográficas de *MEDIANA_PENSIONES*, *POBLACION_PARADA* y *POBLACION_AFILIADA*. A este grupo se le denomina como factor de **Entorno**, ya que engloba ciertos parámetros referenciados al entorno o zona geográfica del paciente.

El tercer factor contiene cuatro variables, tres de ellas con correlaciones altas. Estas variables son: *HAB_TOXICO_NA* (0.88), *HAB_TOXICO_01* (-0.87), *ACT_FISICA_NA* (0.69) y *ACT_FISICA_3.0* (-0.41). Este factor contiene variables de hábitos tóxicos y de actividad física, refiriéndose a los hábitos cotidianos del paciente. A este último factor se le denomina como factor de **Hábitos**.

Según la figura 4.1, se concluye que hay tres factores principales que explican la mayor varianza posible y son factores de:

1. **Índices**: Factor ligado a la edad y los resultados de pruebas.
2. **Entorno**: Factor ligado a variables socio-económicas y geográficas.
3. **Hábitos**: Factor ligado a los hábitos de actividad y consumo.

Para entender los factores, éstos se van a representar en dos dimensiones. En la Figura 4.2 se muestra un gráfico del factor Entorno contra el factor Índices. Las líneas azules muestran posibles clusters. Los puntos están repartidos por todo el área, aunque predominan los valores cerca de los ejes.

De la misma manera que para la combinación de factores Entorno-Índices, en la Figura 4.3 se muestran las demás: Hábitos-Índices y Hábitos-Entorno. El factor de Hábitos parece formar cuatro grupos horizontales, que podrían formar clusters.

Una vez se ha analizado la reducción de dimensionalidad en el número de variables usando el método de EFA, se procede con la segmentación en clusters por el método de clustering de K-Means.

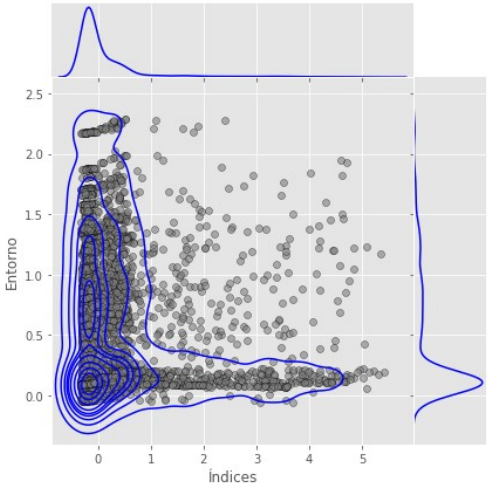
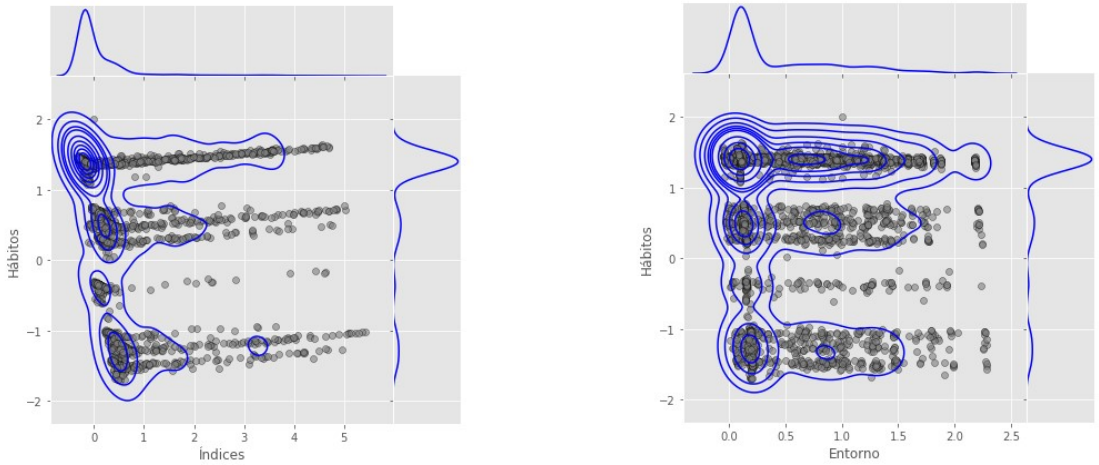


Figura 4.2: *Gráfico del Factor Entorno contra el Factor Índices*



(a) Hábitos-Índices (b) Hábitos-Entorno

Figura 4.3: *Gráficos de Factores tras el Factor Analysis (EFA)*

4.2. Análisis y Visualización de los Clusters

Para la clusterización de pacientes crónicos se utiliza la técnica de K-Means, como ha sido explicado anteriormente. Para obtener el número óptimo de clusters se utiliza el análisis de la silueta (silhouette analysis).

En la Figura 4.4 se muestra el resultado del análisis de silueta. El valor máximo de la silueta ocurre con 2 clusters. Sin embargo, se elige el siguiente valor máximo ya que se considera que dos clusters con es suficientemente explicativo. Con un valor de 0.6148 se elige un número de clusters de 6.

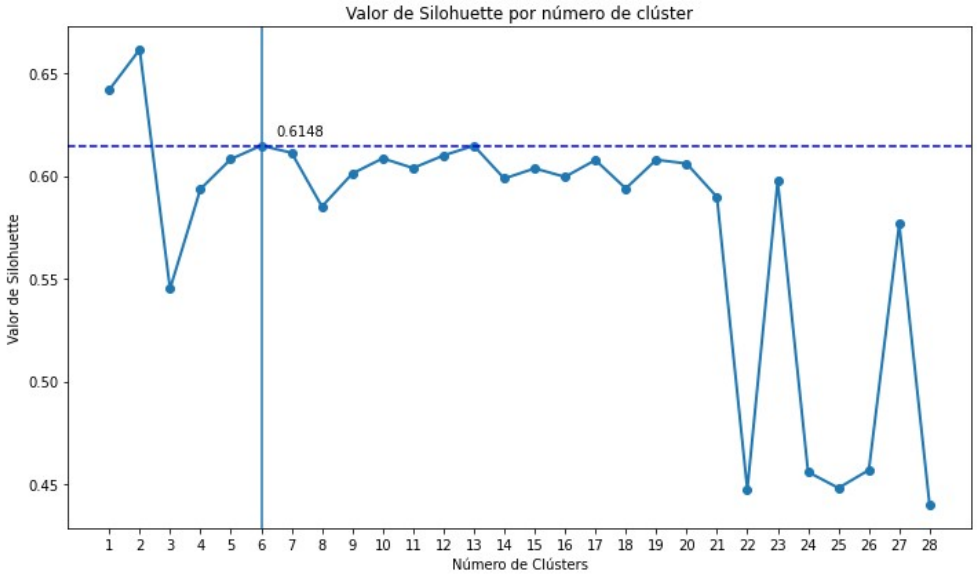


Figura 4.4: *Silhouette Analysis: Elección de número de clusters*

Para una primera visualización de los clusters, se muestra en la Figura 4.5 la segmentación de los seis grupos en dos dimensiones.

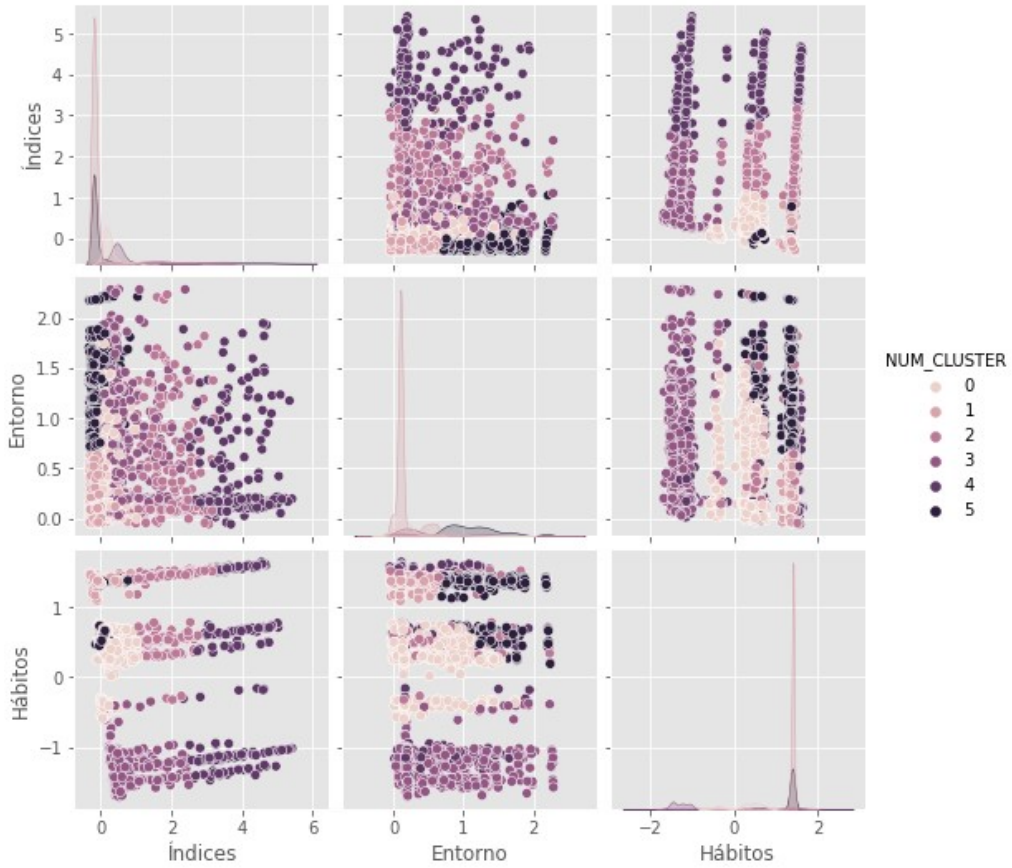


Figura 4.5: *Segmentación de los clusters en 2-D*

4.3. Discusión de los Resultados

En dos dimensiones, el resultado no resulta tan lógico ni visual. Por ello, se representa en la Figura 4.6 la clusterización en tres dimensiones. Se aprecian 3 niveles distintos de hábitos debido a los clusters.

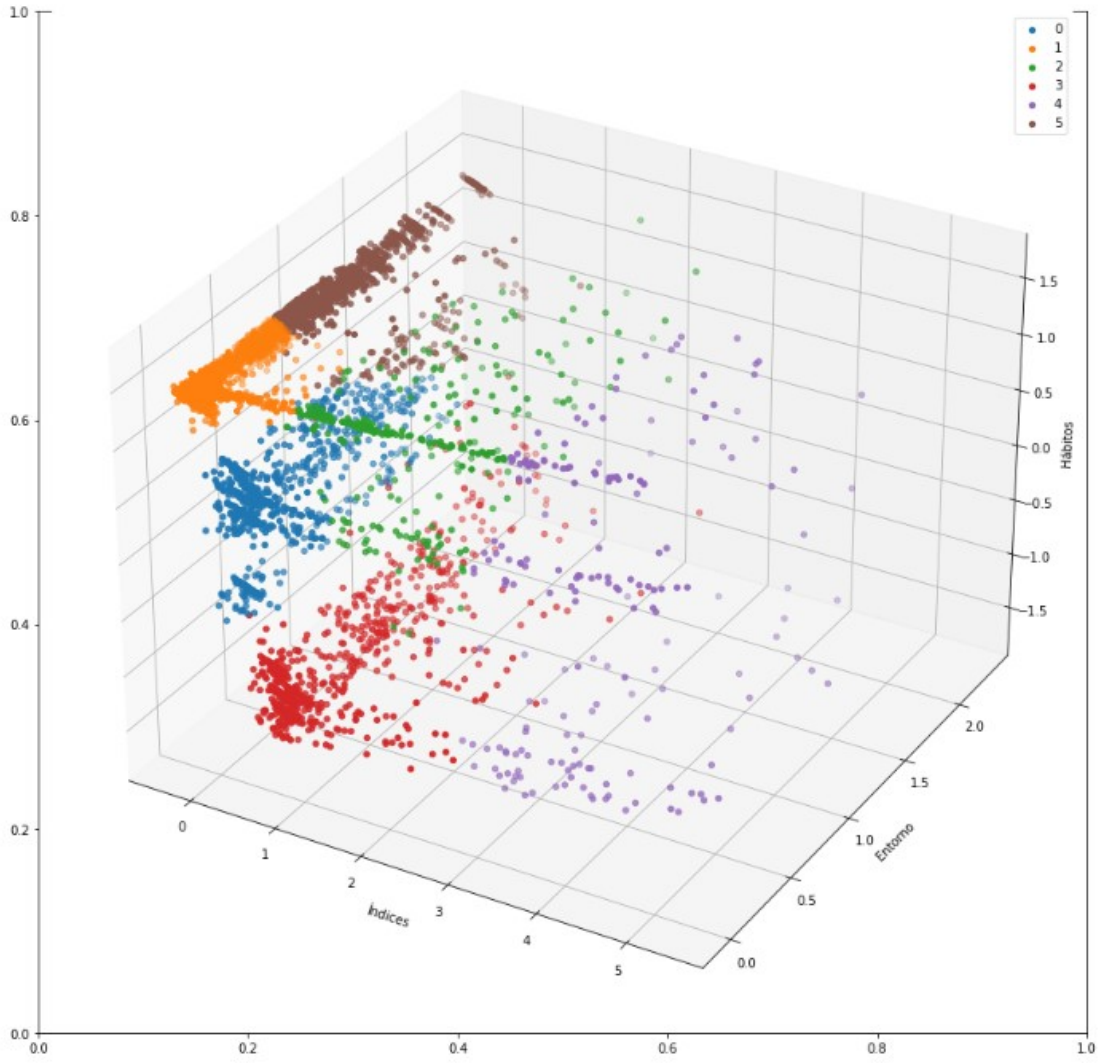


Figura 4.6: *Segmentación de los clusters en 3-D*

Se han obtenido seis grupos para segmentar a la población y entender los niveles de cronicdad de un paciente nuevo. En esta sección se analiza el perfil de cada uno de los 6 clusters. Se empieza por analizar la segmentación en base a la Figura 4.6.

1. **Num cluster 0**: Índices bajos, hábitos medios y entorno alto y bajo
2. **Num cluster 1**: Índices bajos, hábitos altos y entorno bajo

3. **Num cluster 2**: Índices medios, hábitos altos y entorno bajo
4. **Num cluster 3**: Índices bajos, hábitos bajo y entorno alto y bajo
5. **Num cluster 4**: Índices altos, hábitos altos y bajos, entorno bajo
6. **Num cluster 5**: Índices bajos, hábitos altos y entorno alto

Los seis clusters son grupos muy diferentes de la población. Aquellos clusters con el factor de índices, entorno y hábitos bajo son pacientes en menos riesgo. Un paciente del cluster rojo (cluster 3) está en menos riesgo que un paciente en el cluster lila (cluster 5). Por esta razón, se estima que aquellos clusters con un factor de Índices alto son personas mayores, que necesitan un cuidado profesional mayor. En la Figura 4.7 se muestran las edades medias de los pacientes por cluster.

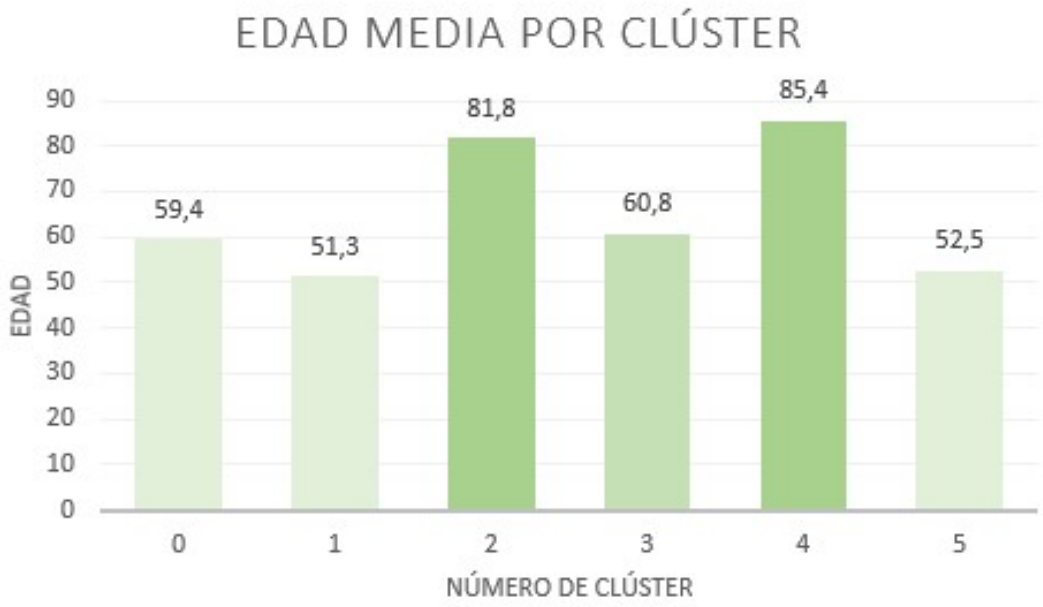


Figura 4.7: *Edad media de los pacientes por cluster*

El factor de Índices está fuertemente relacionado con la edad, como se ha podido comprobar. El clúster 4, con el factor de Índices más alto es el cluster con edad media más alta. De hecho, los clusters 2 y 4 tienen una edad media muy superior a la media del dataset.

Aparte del factor de índices, el factor de hábitos es muy interesante. Se aprecian tres niveles de hábitos. Esto podría indicar que un paciente con un factor de Hábitos alto tiene hábitos tóxicos y de actividad malos.

Para finalizar, se analiza el último factor, el factor de Entorno. La Tabla 4.1 resume los valores medios de cada una de las variables por cluster.

Tabla 4.1: Valores medios de variables socio-económicas por cluster

	MEDIA VARIABLES		
NUM_CLUSTER	MEDIANA_PENSIONES	POBLACIÓN_PARADA	POBLACION_AFILIADA
0	0,164	0,097	0,126
1	0,086	0,046	0,078
2	0,199	0,12	0,164
3	0,237	0,165	0,192
4	0,203	0,121	0,147
5	0,553	0,369	0,416

El cluster con los valores medios máximos en las tres variables es el cluster 5. Un paciente en este cluster vive en una zona con una alta tasa de paro, afiliación a la seguridad social y ingresos por pensiones. Observando la Figura 4.6, el cluster 5 tiene un factor de Índices alto.

Con la información obtenida de la segmentación, se hace un primer ranking de complejidad de paciente en categorías de nivel de intervención visto en la Figura 4.8. Para confirmar el orden el la complejidad de los clusters se debe usar la segmentación.

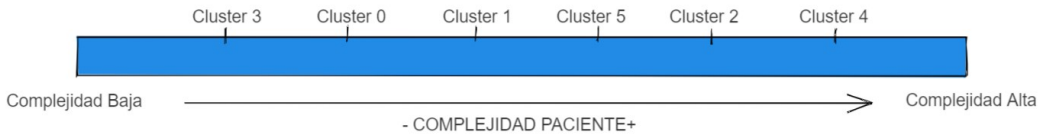


Figura 4.8: *Ranking de los clusters por nivel de complejidad del paciente*

Los clusters con edad media máxima y mayor valor en el factor de Índices se estiman como los clusters que requerirán mayor cuidado profesional. Siguiendo los hábitos, se han organizado los siguientes clusters. Esta nueva agrupación es la *Pirámide de Káiser* actualizada, dinámica y específica para la población andaluza.

5. Gestión del Proyecto

En este apartado de se describe el ciclo de vida que ha sido elegido para el desarrollo de este trabajo, realizando una descripción de cada una de las fases. A su vez, se añade un apartado con la planificación del trabajo de fin de máster.

5.1. Descripción de las fases del Proyecto

El proyecto se ha compuesto por cinco fases, explicadas a continuación:

1. **Fase de Estudio**: En esta fase se estudia el problema y las necesidades del cliente. Se estudian los factores principales del paciente crónico y las segmentaciones de pacientes crónicos existentes en la actualidad.

2. **Fase de Búsqueda de Variables**: Esta segunda fase se corresponde a la elección de las variables interesantes para el modelo, y a su búsqueda en todas las BBDD disponibles del SAS y externas. Esta fase es larga, y se acaban seleccionando veinte variables, 17 variables de la base de datos BPS de SAS y 3 variables de la base externa de Mallapob.

3. **Fase de Georreferencia**: La fase de georreferencia se solapa con la anterior. En esta fase se estudia como vincular la base externa con la base del SAS, analizando los sistemas de coordenadas geográficas y usando este campo para unirlas. Se crea un script en python para la obtención de las 3 variables socio-económicas: Mediana de ingresos de perceptores de pensiones en su comunidad, Poblacion total parada en su comunidad y Población total afiliada a la seguridad social en su comunidad.

4. **Fase Modelado de Complejidad**: En esta fase se obtienen las variables seleccionadas y se transforman, escalándolas a valores entre 0 y 1. Las veinte variables pasan a ser ochenta al encoding *one-hot* utilizado en las variables de Actividad Física y de Hábitos Tóxicos. A su vez, el dataset es limpiado y preparado para la realización de las técnicas de Machine Learning en la siguiente fase.

5. **Machine Learning y Análisis de datos**: La quinta fase utiliza un dataset con ochenta variables (representando a las veinte iniciales) y se crea un modelo

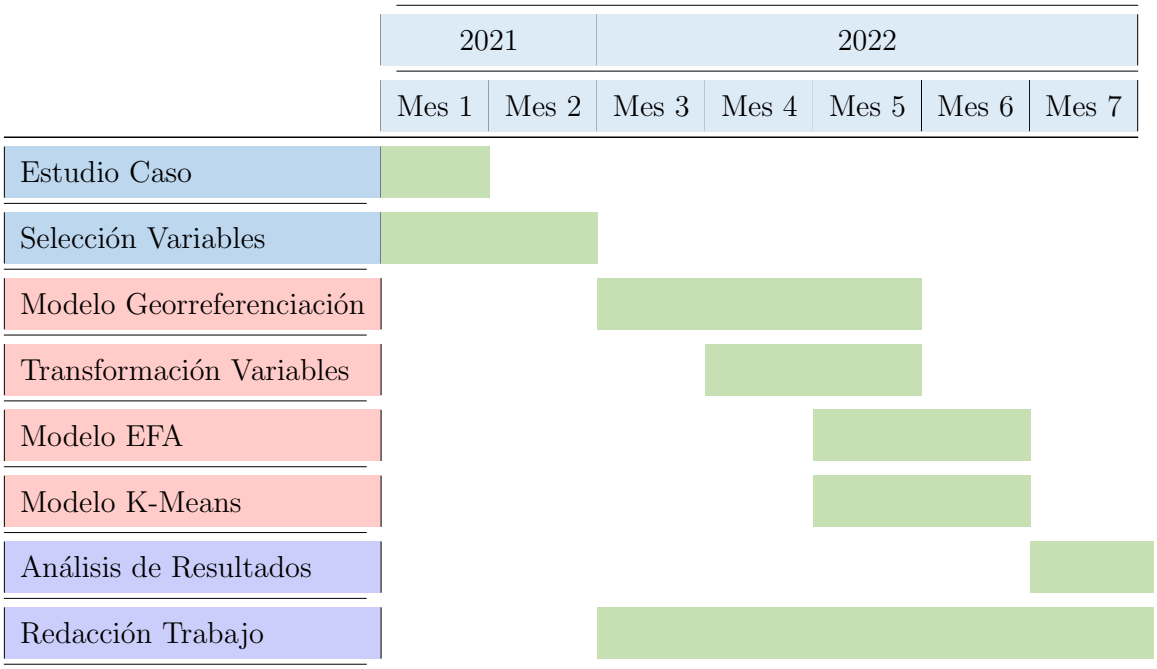
que realiza una reducción de variables por EFA, seguida de un clustering por K-Means. Se escogen adecuadamente el número de factores para EFA y el valor K del número de clústers para el modelo. Finalmente, se analizan los resultados obtenidos y se obtienen conclusiones.

Este orden es el seguido en el proyecto, aunque varias fases se han solapado y a su vez se ha retrocedido en alguna fase para realizar cambios, como la introducción de nuevas variables o mejoras en el modelo de georreferenciación.

5.2. Planificación

La ejecución de este proyecto abarca desde noviembre de 2021 a junio de 2022, siendo la duración del proyecto un total de siete meses. Las fases explicadas anteriormente resumen las distintas tareas realizadas a lo largo del proyecto. La Tabla 5.1 muestra la planificación del proyecto.

Tabla 5.1: Planificación del Trabajo de Fin de Máster



En la tabla, los bloques en azul (Estudio Caso y Selección de Variables) corresponden a tareas más funcionales y de data engineering, los bloques en rosa corresponden a las tareas de Machine Learning y codificación y finalmente los bloques en morado representan las tareas analíticas.

6. Conclusiones y Trabajos Futuros

6.1. Conclusiones

El proyecto propone un método de segmentación de pacientes crónicos y niveles de cuidados mediante el uso de variables de hábitos de actividad, hábitos tóxicos, variables socio-económicas, constantes e índices médicos. Es han usado los métodos de Machine Learning de EFA para la reducción de variables y K-Means para la clusterización de los datos. Se cumple el objetivo de obtener al menos 15 variables para la segmentación, se obtienen 20.

Los datos obtenidos son buenos. Se obtienen tres factores tras el Análisis Factorial Exploratorio. Estos factores corresponden con los grupos que se han nombrado Hábitos, Entorno e Índices. De las variables seleccionadas algunas no han sido críticas como puede ser el peso o el índice de Pfeiffer. El modelo de georreferenciación para obtener variables sociales y económicas por zonas ha dado un resultado muy bueno, al formar su propio factor en la reducción de variables.

Para el clustering, con el método de K-Means se han obtenido 6 clusters, obteniendo un valor del Análisis de Silueta de 0,6348. Con este valor, se cumple el objetivo relativo a K-Means. Se esperaba un mayor número de clusters, para la obtención de un número mayor se deben usar más variables que ayuden a explicar el problema mejor. Sin embargo, 6 clusters es un valor adecuado y manejable para su uso en un entorno laboral.

En conclusión, el método es correcto y ha segmentado la población bien, creando grupos claros y distintivos.

6.2. Trabajos Futuros

Las posibilidades para trabajos futuros partiendo de este proyecto son numerosas, dado el impacto que el mundo de Big Data tiene sobre la sanidad a día de hoy. Por ello, se plantean dos trabajos a futuro a continuación.

El primer trabajo a futuro sería mejorar el modelo actual, analizando nuevas variables que puedan ser relevantes para el caso. En la segunda iteración del proyecto se analizarían variables del número de ingresos hospitalarios recientes de los pacientes y de análisis de patologías. Añadir nuevas variables al modelo añadiría más información, y crearía nuevos clusters que explicarían los grupos de pacientes

crónicos con una nueva perspectiva.

El segundo trabajo futuro, sería llevar todo a Big Data y hacer unos *dashboards* finales para el cliente. De esta manera, médicos y pacientes podrán interactuar con los clústers y mostrar el progreso si un paciente se mueve entre clústers. Se usaría Stratio para este proceso, pudiendo hacer las visulaizaciones y entregables finales desde la plataforma.

A su vez, los entregables finales usarían un control de saltos entre categorías con *cadenas de Markov*, para activar medidas preventivas a estos pacientes. Se podría usar una APIRest que pregunte a la plataforma en que clúster se encuentra un paciente en cierto momento.

Referencias

[1] Parga, David Cierco Jiménez de.
Pliego de prescripciones técnicas que regirán la realización del contrato de "Servicio para la Implantación de una solución corporativa de analítica avanzada, basada en tecnologías Big Data, para el sistema sanitario público de Andalucía". red.es, 2021.

[2] OMS. *Enfermedades no transmisibles*. 2021. URL: https://www.who.int/es/news-room/fact-sheets/detail/noncommunicable-diseases#:~:text=Los%5C%20principales%5C%20tipos%5C%20d%5Ce%5C%20ENT,el%5C%20asma)%5C%20y%5C%20la%5C%20diabetes.. (accessed: 29/05/22).

[3] Minaya, Denisse Cepeda. *El reto de atender a los pacientes crónicos sin mirar su código postal*. 2022. URL: https://cincodias.elpais.com/cincodias/2017/10/18/companias/1508339430_989761.html. (accessed: 28/05/22).

[4] Sanidad, Ministerio de. *Informe anual del Sistema Nacional de Salud 2015*. Ministerio de Sanidad, Servicios Sociales e Igualdad, 2015.

[5] IIC. *Segmentación de Pacientes crónicos*. 2022. URL: https://www.iic.uam.es/soluciones/salud/servicio-segmentacion-de-cronicos/#:~:text=La%5C%20segmentaci%5C%C3%5C%B3n%5C%20de%5C%20pacientes%5C%20cr%5C%C3%5C%B3nicos%5C%20permite%5C%20predecir%5C%20la%5C%20posible%5C%20evoluci%5C%C3%5C%B3n,utilizaci%5C%C3%5C%B3n%5C%20de%5C%20servicios%5C%20y%5C%20costes.. (accessed: 27/04/22).

[6] Pozo, Javier Segura del. *La "estratificación" de la atención de pacientes crónicos y sus determinantes sociales*. 2013. URL: https://saludpublicayotrasdudas.wordpress.com/2013/04/27/la-estratificacion-de-la-atencion-a-pacientes-cronicos-y-sus-determinantes-sociales/. (accessed: 25/05/22).

[7] Unidas, Naciones. *Objetivos de Desarrollo Sostenible*. 2022. URL: https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/. (accessed: 30/05/22).

[8] Bennett, James E. *NCD Countdown 2030: worldwide trends in non-communicable disease mortality and progress towards Sustainable Development Goal target 3.4*. 2018. URL: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(18)31992-5/fulltext. (accessed: 29/05/22).

[9] Sanidad, Consejería de. *Estrategia de Atención a Pacientes con Enfermedades Crónicas en la Comunidad de Madrid*. Ministerio de Sanidad, Servicios Sociales e Igualdad, 2013.

[10] Prieto, Pol Bertran. *Los 10 tipos de enfermedades crónicas (y características)*. 2022. URL: https://medicoplus.com/medicina-general/tipos-de-enfermedades-cronicas. (accessed: 01/06/22).

[11] unir. *La pirámide de Kaiser en enfermería: aplicación, características y beneficios para el paciente*. 2021. URL: https://www.unir.net/salud/revista/piramide-kaiser/. (accessed: 31/05/22).

[12] Bock, Tim. *Factor Analysis and Principal Component Analysis: A Simple Explanation*. 2022. URL: https://www.displayr.com/factor-analysis-and-principal-component-analysis-a-simple-explanation/. (accessed: 02/06/22).

[13] Preacher, Kristopher J. y col. *Choosing the Optimal Number of Factors in Exploratory Factor Analysis: A Model Selection Perspective*. Vanderbilt University, 2013. DOI: 10.1080/00273171.2012.710386.

[14] UCLA. *PRINCIPAL COMPONENTS (PCA) AND EXPLORATORY FACTOR ANALYSIS (EFA) WITH SPSS*. 2021. URL: https://stats.oarc.ucla.edu/spss/seminars/efa-spss/. (accessed: 02/06/22).

[15] Bock, Tim. *What is k-Means Cluster Analysis?* 2022. URL: https://www.displayr.com/what-is-k-means-cluster-analysis/. (accessed: 31/05/22).

[16] T, Muthu Krishnan. *Mathematics behind K-Mean Clustering algorithm*. 2021. URL: https://muthu.co/mathematics-behind-k-mean-clustering-algorithm/. (accessed: 30/05/22).

[17] Gupta, Abhishek. *Difference between K means and Hierarchical Clustering*. 2021. URL: https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/#:~:text=A%5C%20hierarchical%5C%20clustering%5C%20is%5C%20a,the%5C%20clusters%5C%20is%5C%20hyper%5C%20spherical.. (accessed: 31/05/22).

[18] Úbeda, Eugenio Sánchez. *Clustering - Estadística II*. Universidad Pontificia de Comillas, 2020.

[19] Andalucía, Junta de. *Clasificación Grado de Urbanización*. 2022. URL: https://www.juntadeandalucia.es/institutodeestadisticaycartografia/gradourbanizacion/descargahoja.htm. (accessed: 28/03/22).

A. Manual de Transformaciones de Variables

En la siguiente tabla se resumen las variables y las transformaciones realizadas, como explicado en el apartado 3.3.2

N	Variable	Imputación	Transformación
1	Último índice de Barthel	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 100, & x_i \notin Z^+ \end{cases}$	$t_i = 1 - \frac{v_i}{100}$
2	Índice de Barthel máximo	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 100, & x_i \notin Z^+ \end{cases}$	$t_i = 1 - \frac{v_i}{100}$
3	Último índice de Braden	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 24, & x_i \notin Z^+ \end{cases}$	$t_i = 1 - \frac{v_i}{24}$
4	Índice de Braden máximo	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 24, & x_i \notin Z^+ \end{cases}$	$t_i = 1 - \frac{v_i}{24}$
5	Último índice de Norton	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 20, & x_i \notin Z^+ \end{cases}$	$t_i = 1 - \frac{v_i}{20}$
6	Índice de Norton máximo	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 20, & x_i \notin Z^+ \end{cases}$	$t_i = 1 - \frac{v_i}{20}$
7	Último índice de Pfeiffer	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{10}$
8	Índice de Pfeiffer máximo	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{10}$
9	Última valoración de esfuerzo del cuidador	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{14}$
10	Máxima valoración de esfuerzo del cuidador	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{14}$
11	Último indicador de actividad física	$v_i = \begin{cases} x_i, & x_i \neq nan \\ "NA", & x_i = nan \end{cases}$	$t_i = OneHot(v_i)$
12	Último indicador de consumo de alcohol	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{\max(x)}$
13	Último indicador de consumo de tabaco	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{\max(x)}$
14	Último indicador de hábitos tóxicos	$v_i = \begin{cases} x_i, & x_i \neq nan \\ "NA", & x_i = nan \end{cases}$	$t_i = OneHot(v_i)$
15	Edad registrada por última vez	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ mean(x), & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{120}$
16	Última medición de constante de peso	$v_i = \begin{cases} x_i, & x_i \neq nan \\ \begin{cases} PC(x_i), & PC(x_i) < 65 \\ 65, & PC(x_i) \geq 65 \end{cases}, & x_i = nan \end{cases}$ $PC(x_i) = 9 + (2{,}5 \times edad_i)$	$t_i = \frac{v_i}{\max(x)}$
17	Máximo de comorbilidades	$v_i = \begin{cases} x_i, & x_i \in Z^+ \\ 0, & x_i \notin Z^+ \end{cases}$	$t_i = \frac{v_i}{100}$

N	Variable	Imputación	Transformación
18	Mediana de ingresos de perceptores de pensiones en su comunidad	$v_i = \begin{cases} x_i, & x_i \neq nan \\ mean(x), & x_i = nan \end{cases}$	$t_i = \frac{v_i}{\max(x)}$
19	Población total parada en su comunidad	$v_i = \begin{cases} x_i, & x_i \neq nan \\ mean(x), & x_i = nan \end{cases}$	$t_i = \frac{v_i}{\max(x)}$
20	Población total afiliada a la seguridad social en su comunidad	$v_i = \begin{cases} x_i, & x_i \neq nan \\ mean(x), & x_i = nan \end{cases}$	$t_i = \frac{v_i}{\max(x)}$

Donde

- Variables 1 – 17 tienen como fuente la BBDD del SAS
- Variables 18 – 19 tienen como fuente la BBDD del IECA - Mallapob

B. Función Diccionario Coordenadas

Debajo se muestra el código para la función diccionario de sistemas de coordenadas geográficas. También se incluye el código para la creación de la columna de códigos INSPIRE1K. Para la explicación del código, se desarrolla en el apartado 3.3.4.

```
import pandas as pd
import numpy as np
from pyproj import Proj, transform

def geobdu_a_inspire(x1, y1):
    """
    Funcion Diccionario que transforma de codigos con referencia ETRS89
    (EPSG: 25830) a coordenadas INSPIRE (EPSG = 3035).

    Params
    ------
    x2 : array
        Coordenada X de la direcci n. ETRS89 - EPSG: 25830.
    y2 : array
        Coordenada Y de la direcci n. ETRS89 - EPSG: 25830.


    Return
```

```
18      --------
19      x2 : array
20          Vector de la coordenada de LONGITUD. INSPIRE1K, EPSG = 3035
21      y2 : array
22          Vector de la coordenada de LATITUD. INSPIRE1K, EPSG = 3035
23      """
24      x1 = np.array(x1)
25      y1 = np.array(y1)
26      # Transforma las coordenadas
27      inProj = Proj('epsg:25830')  # ETRS89
28      outProj = Proj('epsg:3035')  # INSPIRE
29      x2, y2 = transform(inProj, outProj, x1, y1)
30
31      return x2, y2
32
33  ### CREAR COLUMNA CON CODIGO INSPIRE
34  x2 = np.array(x1)
35  y2 = np.array(y1)
36  inspire = np.array(y1)
37  for i in range(len(x1)):
38      if float(x1[i]) > 0:
39          x2[i], y2[i] = geobdu_a_inspire(x1[i], y1[i])
40          inspire[i] = "1kmN{}E{}".format(int(round(x2[i], -3)/1000), int
        (round(y2[i], -3)/1000))
41      else:
42          x2[i], y2[i] = (0, 0)
43          inspire[i] = 0
```

Listing 2: Función Diccionario Python entre referencias geográficas