Análisis de la Encuesta de Salud Nacional y Examen de Nutrición de Estados Unidos (NHANES) usando machine learning

María José Crespo Estévez
Máster Universitario en Ciencias de Datos (Data Science)
Área: Minería de datos y machine learning

Laia Subirats Maté
Jordi Casas Roma

Junio de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis de la Encuesta de Salud Nacional y Examen de Nutrición de Estados Unidos (NHANES) usando machine learning*
Nombre del autor:	*María José Crespo Estévez*
Nombre del consultor/a:	*Laia Subirats Maté*
Nombre del PRA:	*Jordi Casas Roma*
Fecha de entrega (mm/aaaa):	06/2019
Titulación::	*Máster Universitario en Ciencia de Datos (Data Science)*
Área del Trabajo Final:	*Minería de datos y machine learning*
Idioma del trabajo:	*Castellano*
Palabras clave	*medicina, NHANES, Machine learning*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

En este trabajo se usará el conjunto de datos de kaggle *National Health and Nutrition Examination Survey*. La finalidad será diseñar e implementar diferentes modelos no supervisados para identificar patrones, descubrir como tienden los datos a agruparse y si existen comorbilidades entre las enfermedades. También diseñaremos modelos predictivos para detectar si un paciente sufre hipertensión.

En los modelos de clustering, escogemos el parámetro n_neighbors con el método del codo y los parámetros de los modelos predictivos con el RandomizedSearchCV y después con GridSearchCV.

Se implementa un modelo de clustering con k-Means para el conjunto total de los datos y otro para las enfermedades del archivo *medications.* En el primero se concluye que la edad y las variables relacionadas con la salud dental son los más importantes para la determinación de los clústeres, en el segundo se obtienen unas posibles comorbilidades para las enfermedades.

Para los modelos predictivos se usan los algoritmos: *Support Vector Classification, Gradient Boosting Classifier, AdaBoost Classifier, Random Forest Classifier, Naive Bayes, Logistic Regression y k-NN* de la librería *sklearn.* El mejor modelo se obtiene con AdaBoost y una exactitud de 76.33, aunque el Naive Bayes ofrece un buen resultado del TPR de 62.69 al obtenerse la menor cantidad de falsos negativos entre todos los modelos.

Abstract (in English, 250 words or less):

We are going to work with the kaggle's dataset *named National Health and Nutrition Examination Survey* in this paper. The main purpose is to design and implement different unsupervised models to identify patterns, to discover how the data tends to group and if there are comorbidities among the diseases. We are also going to design predictive models to detect if a patient suffers from some of the diseases written in the dataset.

In the clustering models, we choose the parameter n_neighbors with the elbow method and the parameters of the predictive models with the RandomizedSearchCV and then with GridSearchCV.

A clustering model with k-Means is implemented for the total data set and another for diseases of the medications file. In the first, it is concluded that age and variables related to dental health are the most important for the determination of clusters, in the second, possible comorbidities for diseases are obtained.

For predictive models the algorithms are used: Support Vector Classification, Gradient Boosting Classifier, AdaBoost Classifier, Random Forest Classifier, Naive Bayes, Logistic Regression and k-NN from sklearn. The best model is obtained with AdaBoost and an accuracy of 76.33, although the Naive Bayes offers a good result of the TPR of 62.69 to obtain the lowest amount of false negatives among all models.

Índice

Lista de tablas

Lista de figuras

1. Introducción

1.1. Contexto y Justificación del Trabajo

Estamos viviendo un momento de cambio en la historia, en el que la tecnología está cambiando nuestras vidas, tanto a nivel profesional como personal. Estos cambios suponen ventajas y nuevos retos, ya que las organizaciones y empresas se ven obligadas a cambiar su forma de trabajar e implantar la tecnología adecuada en su entorno para generar ventaja competitiva en el mercado.
En la era de los datos masivos, éstos se capturan a través de: las redes sociales, sensores, dispositivos móviles, páginas web... También organizaciones institucionales y empresas públicas comparten datos abiertos que aportan confianza por transparencia y la reutilización de la información. Todos estos datos se transforman en productos finales que mejoran las experiencias de los usuarios personalizando los servicios o sirven como herramientas de apoyo para la toma de decisiones en cualquiera de los sectores que podamos plantear.

Uno de estas áreas es la medicina, que gracias al conocimiento que se pueden extraer de los datos, se realizan diagnósticos precoces de enfermedades que si se tratan en el momento adecuado sería más rápido de curar e incluso de salvar vidas de los enfermos con enfermedades graves. La aplicación de big data en la medicina da oportunidades desde enfermedades con menos visibilidad como pueden ser las enfermedades raras hasta las que son más conocidas como el cáncer.
Otra aplicación posible es la predicción de epidemias, creación de alertas y tomar medidas preventivas para reducir su alcance. Las aplicaciones son múltiples y variadas, y suponen para los especialistas una herramienta de apoyo para los diagnósticos y la búsqueda de tratamientos personalizados.

Todo esto es posible gracias a la digitalización de historiales clínicos de pacientes tratados, de la normalización de los datos médicos y la anonimización de ellos por tratarse de información sensible. Aún hay detalles que mejorar, ya que se está tratando de dotar de infraestructuras para centralizar los servicios y facilitar el acceso a información compartida [1]. Una vez que se dispongan de los datos se aplican técnicas de machine learning con sistemas de Inteligencia Artificial y redes neuronales y se obtienen resultados de los modelos aplicados dependiendo del problema que se esté resolviendo.

Vivimos en un país en el que la esperanza de vida es de las más elevadas en todo el mundo y que según el Bloomberg Healthiest Country Index [2] se ha alzado como el más saludable en 2019 por diversos factores: esperanza de vida, hábitos alimentarios, sistema de sanidad,...
Por todo esto, la cuestión de la salud la tenemos presente y es un tema de interés para cualquier persona. Además, al aumentar la esperanza de

vida hay garantizar la calidad de ella también, por lo que la generación de modelos predictivos que sean capaces de diagnosticar enfermedades antes de que se manifiesten, identificación de patrones de enfermedades y las comorbilidades entre ellas se convierte en un punto más importante si cabe.

Si hay algo que tenemos en común cualquier persona del mundo es la vida, por lo que cualquier avance en mayor o menor medida que se pueda hacer en la medicina que ayude a mejorar la calidad de vida y la salud de las personas merecerá la pena y puede ser inspiración para el trabajo e investigación de otras personas, que es mi motivación para realizar este trabajo. Si bien, el dataset que se va a usar pertenece a ciudadanos de otro país, sirve como muestra de lo que se puede realizar en el caso de que se disponga de un conjunto de datos similares características para cualquier país o región y obtener resultados propios.

Para este trabajo, usaremos el conjunto de datos *National Health and Nutrition Examination Survey* disponible en *Kaggle*, plataforma online con múltiples repositorios [3]
La Encuesta de Salud Nacional y Examen de Nutrición (National Health and Nutrition Examination Survey, NHANES), según se indica en la propia página [4], es un programa de estudios diseñados para evaluar la salud y estado nutricional de adultos y niños de Estados Unidos. NHANES es el mayor programa del Centro Nacional de Estadísticas de Salud (National Center for Health Statistics, NCHS). Este programa empezó en 1960 y examina una muestra representativa de 5000 personas cada año aproximadamente. En la encuesta se incluye: demografía, socio-economía, dieta y preguntas relacionadas con la salud. El componente de examen médico realizado consiste en medidas médicas, dentales y fisiológicas, y en tests de laboratorios por personal médico.

Los resultados que se quieren obtener se definirán en el apartado de objetivos siguiente.

1.2. Objetivos del Trabajo

Los objetivos principales de este trabajo serán:

- Diseñar un modelo de aprendizaje no supervisado y descubrir patrones que generen nuevo conocimiento.
- Diseñar un modelo predictivo que nos indique si un paciente es susceptible de tener una enfermedad de las incluidas en el dataset.
- Hallar comorbilidades entre enfermedades.

Como objetivos secundarios se podría definir:

- Explorar los datos en crudo y después de ser tratados.
- Reducir la dimensionalidad.
- Interpretar los resultados y comprobar que no se haya producido sobreentrenamiento.

- Evaluar y mejorar la precisión mediante un proceso iterativo del modelo.

1.3. Enfoque y método seguido

En este caso, se va a desarrollar un producto nuevo. Con el fin de alcanzar los objetivos, se seguirá un proceso iterativo para mejorar el modelo resultante: comprensión de los datos, procesamiento de los datos, modelado y evaluación de los modelos.
Diseñaremos y se analizarán diferentes modelos para conseguir el modelo más preciso y que mejor se adapte a los objetivos deseados.
Se usará como lenguaje de programación python con las librerías necesarias.

1.4. Planificación del Trabajo

Para el desarrollo de este proyecto se estiman 36 horas semanales, repartidas en 6 días de la semana. Si bien, se contempla que si fuera oportuno se realizarán modificaciones de este número de horas dependiendo de la necesidad para poder cumplir los objetivos marcados dentro del plazo.

En esta planificación del trabajo, nos basaremos en las entregas de cada PEC marcadas, siendo éstas los hitos principales. Más concretamente, se muestra en esta tabla el número de horas que se van a dedicar a cada entrega parcial con sus fechas de entregas respectivas.
El tiempo dedicado a completar las 5 pruebas serían 612 horas en total.

	Descripción	Fecha	Horas
PEC 1	Definición y planificación del trabajo final	03/03/19	72
PEC 2	Estado del arte o análisis de mercado del proyecto	24/03/19	108
PEC 3	Diseño e implementación del trabajo	19/05/19	288
PEC 4	Redacción de la memoria	09/06/19	108
PEC 5	Presentación y defensa del proyecto	16/06/19	36
			612

Tabla 1. Planificación por horas.

La definición de las tareas a realizar en cada hito la mostraremos en las siguientes tres figuras que completan el diagrama de Gantt diseñado con el software ProjectLibre, [5]

	Nombre	Duracion	Inicio	Terminado
1	**Trabajo Final Máster**	**124 days**	**20/02/19 9:00**	**23/06/19 22:00**
2	**PEC1 - Definición y planificación del trabajo final**	**12 days**	**20/02/19 9:00**	**3/03/19 22:00**
3	Comprensión de los datos	2 days	20/02/19 9:00	21/02/19 22:00
4	Definición del contexto, del enfoque y de objetivos	6 days	22/02/19 9:00	27/02/19 22:00
5	Planificación del trabajo	4 days	28/02/19 9:00	3/03/19 22:00
6	Redacción de la entrega	3 days	1/03/19 9:00	3/03/19 22:00
7	**PEC2 - Estado del arte o análisis de mercado del proyecto**	**21 days**	**4/03/19 9:00**	**24/03/19 22:00**
8	Recolección de resultados de otros trabajos	17 days	4/03/19 9:00	20/03/19 22:00
9	Búsqueda de bibliografía	17 days	4/03/19 9:00	20/03/19 22:00
10	Selección inicial de herramientas y formación	5 days	20/03/19 9:00	24/03/19 22:00
11	Corrección detalles anteriores y revisión	4 days	21/03/19 9:00	24/03/19 22:00
12	Redacción de la entrega	3 days	22/03/19 9:00	24/03/19 22:00
13	**PEC3 - Diseño e implementación del trabajo**	**56 days**	**25/03/19 9:00**	**19/05/19 22:00**
14	Búsqueda de bibliografía	56 days	25/03/19 9:00	19/05/19 22:00
15	Preparación de los datos	6 days	25/03/19 9:00	30/03/19 22:00
16	Modelado de los datos	41 days	31/03/19 9:00	10/05/19 22:00
17	Evaluación de los modelos	38 days	3/04/19 9:00	10/05/19 22:00
18	Corrección detalles anteriores y revisión	9 days	11/05/19 9:00	19/05/19 22:00
19	Redacción de la entrega	6 days	14/05/19 9:00	19/05/19 22:00
20	**PEC4 - Redacción de la memoria**	**21 days**	**20/05/19 9:00**	**9/06/19 22:00**
21	Búsqueda de bibliografía	21 days	20/05/19 9:00	9/06/19 22:00
22	Desarrollo completo de la memoria y entrega final	21 days	20/05/19 9:00	9/06/19 22:00
23	Corrección detalles anteriores y revisión	7 days	3/06/19 9:00	9/06/19 22:00
24	**PEC5 - Presentación y defensa del proyecto**	**7 days**	**10/06/19 9:00**	**16/06/19 22:00**
25	Realización de la presentación	5 days	10/06/19 9:00	14/06/19 22:00
26	Revisión y entrega	3 days	14/06/19 9:00	16/06/19 22:00
27	**Defensa pública**	**7 days**	**17/06/19 9:00**	**23/06/19 22:00**
28	Responder a las preguntas del tribunal	7 days	17/06/19 9:00	23/06/19 22:00

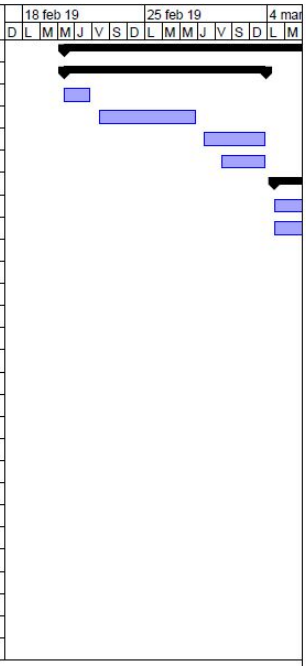

Figura 1. Diagrama de Gantt.

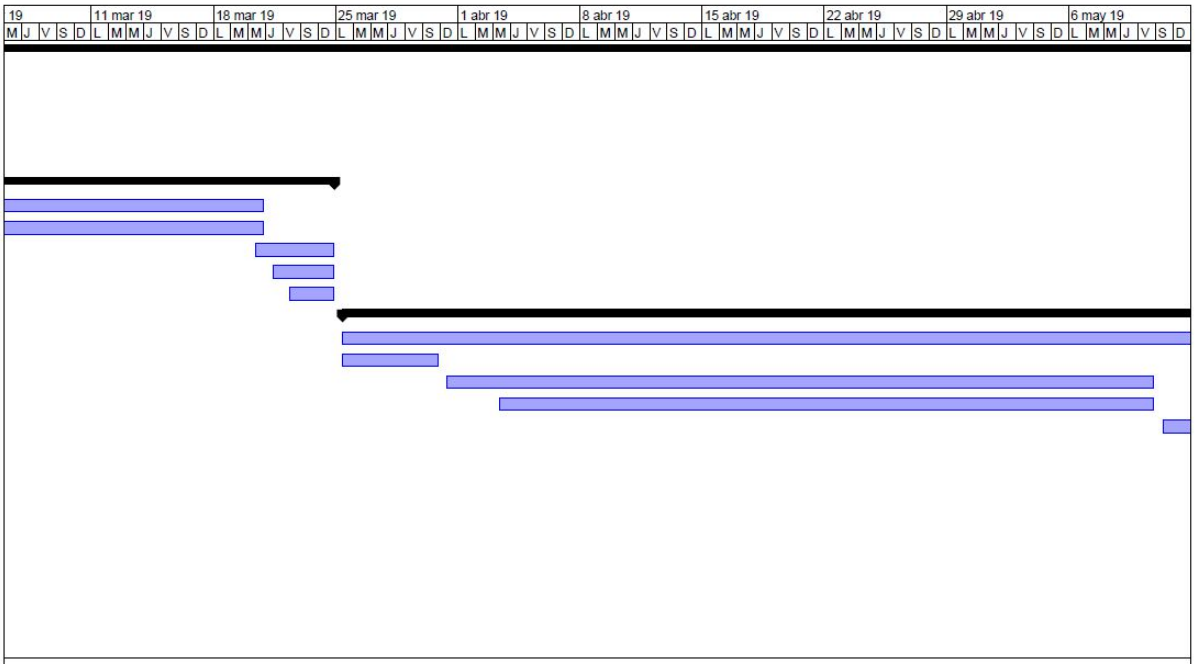

Figura 2. Diagrama de Gantt.

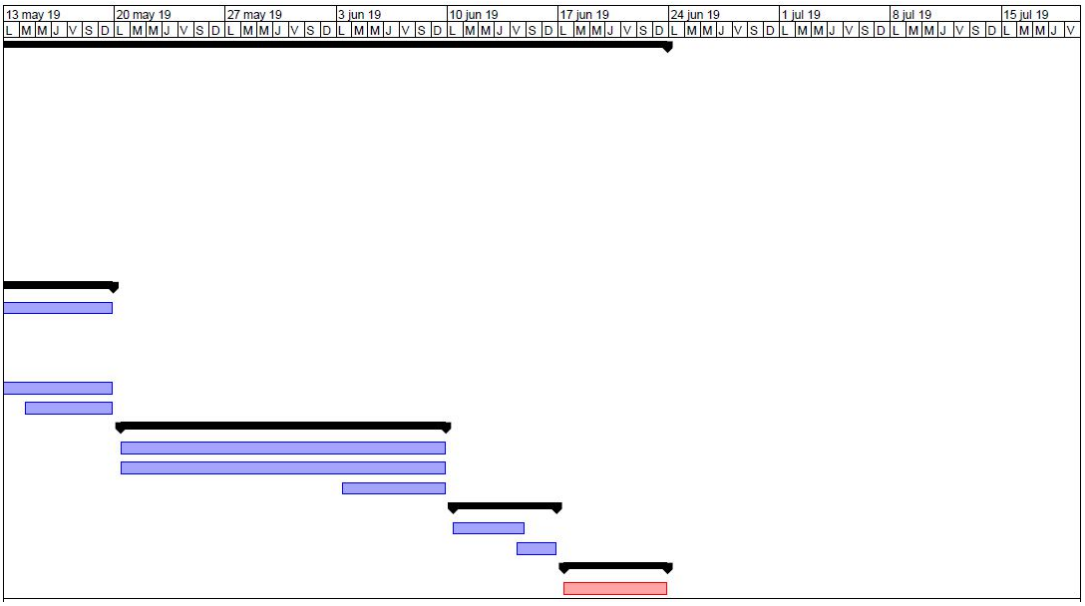

Figura 3. Diagrama de Gantt.

1.5. Breve Descripción de los otros Capítulos de la Memoria

Para la realización final del proyecto, se va a pasar por diferentes fases para la ejecución del mismo. En el capítulo 2, estado del arte, se hará un repaso de trabajos de investigación, donde se realizan técnicas similares a las que se van a usar y se aplica a problemáticas del área de la medicina con otros conjuntos de datos o con el de NHANES.

Tras este capítulo, que nos proporciona una visión del trabajo que tenemos que realizar, se pasará a la fase de implementación del trabajo. En esta fase se realiza la limpieza de los datos y se desarrolla el código de los modelos explicando los parámetros que se ajustan y como se han seleccionado.

Después de la implementación, se evalúan los modelos predictivos obtenidos con la matriz de confusión y las métricas que se obtienen de éstas.

Con los resultados de las matrices de confusión, se realiza la evaluación de los algoritmos predictivos. Con la evaluación, la interpretación de las métricas y de los clústeres obtenidos de los modelos de clustering se llegará a las conclusiones y las líneas de trabajo futura y mejoras.

2. Estado del Arte

En este apartado se repasarán artículos de investigación relacionados con el presente estudio. Estos artículos muestran como los diferentes autores plantean resolver diferentes problemáticas del área de la medicina con construcción de modelos predictivos o de agrupación y como se enfrentan a la preparación de los datos para poder conseguirlo.

Debido a todas las posibilidades de estudios que se podrían realizar, tanto por el área como por las variables que contienen el conjunto de datos de NHANES, es inviable plasmar todos los artículos relacionados con esta temática y todas las combinaciones de las variables que se podrían estudiar.

Por tanto, se mostrarán los estudios que, o bien usan el mismo conjunto de datos o que sin usarse este conjunto, se aplican algoritmos para conseguir objetivos similares a los que se definen en nuestro estudio.

Según la Organización Mundial de la Salud (OMS), *en 2008, en el mundo se había diagnosticado de hipertensión aproximadamente el 40% de los adultos mayores de 25 años; el número de personas afectadas aumentó de 600 millones en 1980 a 1000 millones en 2008* [6].

La hipertensión es también responsable de 9.4 millones de muertes de los 17 millones de muertes de enfermedades cardiovasculares que se producen por año y supone el 45% de las muertes por cardiopatías y el 51% de las muertes por enfermedades por accidente cerebrovascular. Tiene consecuencias de cardiopatías, accidentes cerebrovasculares, insuficiencia renal, mortalidad y discapacidad prematura. Debido a que esta enfermedad la sufre en mayor medida países con niveles de ingresos medios y bajos, también empobrece por los gastos médicos de hospitalizaciones y las complicaciones de enfermedades relacionadas con la hipertensión al no detectarse ni tratarse en una etapa temprana.
Merece la pena listar los factores de riesgo que pueden provocar la aparición de la enfermedad. Por una parte están los factores conductuales: dieta malsana, consumo de alcohol, consumo de tabaco, inactividad física, exposición al estrés. También influyen factores como el bajo nivel socioeconómico, vivienda y educación. Por último, factores como la edad, sobrepeso, padecer otras enfermedades (diabetes mellitus, hipercolesterolemia,...), factores genéticos y antecedentes familiares favorecen el desarrollo de la hipertensión [6].
Con el fin de obtener un modelo adecuado, tendremos en cuenta para la selección de estas características para el diseño del modelo predictivo y así ofrecer una herramienta e información para ayudar a detectar una enfermedad que normalmente no produce síntomas en sus primeras etapas [6]. Esto hace que sea difícil de diagnosticar y tratar y que pueda acarrear consecuencias peores.

2.1. Otros Estudios con el Conjunto de Datos NHANES

En el artículo [7] se usa el conjunto de datos NHANES entre 2007 y 2016 para crear un modelo de detección con elevadas probabilidades de enfermedades cardiovasculares. Para implementar el modelo se eliminan los valores nulos y se crea la variable objetivo definiendo que un paciente tenga hipertensión o no si la presión sistólica es superior a 140 mmHg, obteniéndose una variable indicadora binarizada.
Las variables independientes seleccionadas son: edad, sexo, raza, Índice de Masa Corporal (IMC), consumo de tabaco, presencia de enfermedades renales y diabetes. Estas variables son consecuentes con los factores de riesgo mencionados anteriormente según la OMS. El modelo que se desarrolla es una regresión logística de la librería scikit-learn de Python. De esta manera, se estiman los parámetros constantes y los coeficientes de la ecuación del modelo y se puede calcular la probabilidad de que un sujeto tenga hipertensión. Los autores proponen construir dos modelos después analizar si el p-valor asociado a las variables independientes es significativo o no para comparar los resultados, aunque en algunos casos de los que no son significativas las incluyen igualmente por considerarse importantes para la interpretación. Comparando estos dos modelos se puede concluir que no hay diferencias significativas entre las métricas obtenidas de sendos modelos.

Los resultados que se obtienen para este modelo son: sensibilidad del 77%, especificidad del 68%, precisión de la predicción positiva de 32% en la conjunto test y AUC del 73%. Se genera el conocimiento que los individuos con obesidad, edad entre 71 y 80 años, que sea de raza negra no hispana y hombres, tienen más probabilidad de tener hipertensión y que la diabetes. Enfermedades renales y fumadores no tienen relación con la hipertensión, según los parámetros y variables seleccionadas.

El siguiente estudio está enfocado a la diabetes y usa los datos de NHANES [8], que al igual que la hipertensión es una enfermedad que afecta a gran parte de la población y de diagnóstico complicado en etapas tempranas. También la metodología que se sigue en el artículo es de interés para el estudio que nos ocupa. Para completar el modelo se usan datos del NHANES de 2013-2014, compuesto por 10172 muestras y 54 características del archivo *questionnaire.* En el preprocesado se imputan los valores perdidos por la etiqueta más común de cada columna y los valores categóricos se transforman a variables numéricas. También se realiza selección a 24 variables por la alta dimensión para mejorar el modelo y reducir el coste de computación en base a la puntuación de importancia de la característica.
La construcción del modelo se lleva a cabo con la combinación de las probabilidades de predicción sin peso de diferentes modelos de machine learning con un *Ensemble Model*. Los modelos escogidos son: regresión logística, *kNN, Random Forests* y *Gradient Boosting*, todo implementado en python con la librería *Sckit-learn*. Cada modelo devuelve una probabilidad, el modelo de ensamblaje calcula una media de la

probabilidad p. Si es mayor o igual a 0.5 devuelve la etiqueta de clase diabética y si es menor que 0.5 a la no diabética. Para calibrar el modelo de ensamblaje, se emplea la técnica Majority Voting que es independiente de los parámetros tuneados.
La evaluación del modelo se realiza con el conjunto test, computando la media del error entre lo predicho y los valores reales.

Tal y como se comentó anteriormente, uno de los factores de riesgo de la hipertensión según la OMS es el sobrepeso [6]. En el artículo que desglosaremos a continuación [9], se relacionan la obesidad y sus comorbilidades (hipertensión y diabetes mellitus tipo 2) con la exposición a metales pesados como el bario y el talio. El conjunto de datos escogido es NHANES desde 2003-2004 hasta 2013-2014. Con el fin de poder obtener resultados de las comorbilidades se eliminan participantes sin datos de hipertensión y diabetes mellitus tipo 2 (T2DM).
Primero se emplea una aproximación estadística con *elastic-net flexible* (AENET). Para analizar la alta dimensionalidad de los datos con la colinealidad del problema e identificar mezclas de metales pesados con su dependencia con el perímetro de la cintura. Se construye un entorno de riesgos para los valores principales y por pares de las interacciones de los metales seleccionados por AENET. La idea es construir un modelo de riesgo predictivo (ERS) como suma con pesos de las sustancias contaminantes. Finalmente se evalúa la asociación de ERS con otras medidas de obesidad y comorbilidades como hipertensión y T2DM. El ERS refleja la predicción de WC como mezcla de metales.
Se implementa regresión lineal para las asociaciones de ERS con las medidas de obesidad (IMC, grosor de la piel y grasa total corporal) y regresión logística para la hipertensión y T2DM en el software R.
En este estudio concluyen que la exposición a metales pesados está asociado con la obesidad y condiciones crónicas como hipertensión y la T2DM.

2.2. Técnicas de Interés con Otros Conjuntos de Datos

Uno de los objetivos que se han marcado es el de diseñar un modelo de aprendizaje no supervisado que muestre patrones que se escondan entre los datos. Con él, podremos ver como se agrupan y tener una exploración inicial de los datos que nos permita analizar los datos.

Según la medicina tradicional japonesa, Kampo, y la tradicional china, el color de la lengua puede indicar problemas mentales o físicos. En relación a poder detectar el color la lengua de los pacientes objetivamente se ha desarrollado un modelo no supervisado que con el conocimiento en medicina Kampo se podrá obtener de una forma más precisa la enfermedad asociada [10]. El estudio de resumidamente se explica a continuación.
Con un instrumento óptico médico, DS01-B, se toman imágenes de la lengua y se selecciona solo el área de la imagen donde se encuentra. DS01-B simula la luz natural, de manera que no altera el color de las

zonas a analizar de la lengua. Se recogen 1080 imágenes y se clasifican en 5 categorías según el color del cuerpo y en 6 según el color del recubrimiento por físicos con experiencia en la medicina Kampo.
Una vez que se eliminan las imágenes de la muestra que no son apropiadas, se diseña un modelo *con k-means* en MATLAB con k=3, k=4 y k=5, siendo el que mejor identifica las zonas k=4.
Con k=4 se diferencian con claridad las zonas del fondo de la lengua, recubrimiento, cuerpo y zonas de transición. *K-means* cuantifica la media del color de cada pixel, descompuesto en CIELAB en (L*, a*, b*) donde L* es la luminosidad de negro a blanco, a* va de rojo a verde y b* azul.
Los clústeres definen 5 colores en el cuerpo de la lengua (blanco claro, rojo claro, rojo, rojo oscuro y morado) y 6 para el recubrimiento (blanco, blanco amarillento, amarillo, marrón, gris y negro).
Se concluye que existen varios colores en el cuerpo de la lengua y la media de la información del color obtenida es insuficiente para tener una evaluación clínica más precisa.

En relación con los modelos predictivos para la hipertensión, tenemos el artículo [11]. El conjunto de datos de este estudio consiste en la recolección de la variabilidad de la frecuencia cardíaca a corto plazo de 30 voluntarios sanos y 41 pacientes que padecen de hipertensión arterial en dos posturas diferentes. Primero se toma la frecuencia con el individuo en posición horizontal durante 300 segundos y después en 70º por otros 300 segundos en el Sverdlovsk Clinical Hospital of Mental Diseases for Military Veterans (Yekaterinburg, Russian Federation) con un electroencefalograma. En estudios anteriores se muestran que la precisión es mejor en los modelos de clasificación realizados con las mediciones hechas en la posición de 70º, por lo que solamente se tienen en cuenta estos valores.
Se implementan diferentes modelos con librería *Sckit-learn* de python de estos datos para comparar los resultados obtenidos. Estos modelos son: análisis discriminante lineal y cuadrático (LDA y QDA), k-vecinos más cercanos (*kNN*), máquinas de soporte vectorial (SVM), árboles de decisión y clasificador *Naïve Bayes*. En la selección inicial de 53 características se analizan todas las posibles combinaciones de ellas y se escogen aquellas donde la correlación sea inferior a 0.25 para construir los modelos.
Una vez evaluados la eficacia del modelo con el conjunto test, comparando las etiquetas de los datos con las predichas por el modelo se concluye en este estudio que los mejores resultados se obtienen con LDA y QDA en general, siendo los mejores cuando se seleccionan 4 variables independientes. Esto es por la alta puntuación de clasificación y la baja desviación sobre diferentes realizaciones.

El siguiente artículo también está relacionado con la hipertensión [12]. En este caso los datos van a ser los recogidos del cuestionario de investigación epidemiológica de la población Han china de Beijing**.** Después de la limpieza de datos, se seleccionan 9 factores ambientales y 12 genéticos para construir el modelo de 1200 muestras, formado por 559

pacientes con hipertensión y 641 que no. Han sido excluidos previamente de esta muestra pacientes con enfermedades vasculares, enfermedades coronarias, historial con accidente cerebrovascular, diabetes, enfermedad vascular renal, fallos renales, enfermedad hepática grave, hipertensión secundaria e hipertensión de bata blanca. Se construyen tres modelos con SVM y función kernel radial en el software R, uno con solo factores ambientales (precisión de 72.8%), otro con solo factores genéticos (54.4%) y el último con ambos (76.3%) y otros tres de manera análoga con función kernel Laplaciana (76.9%, 57.7% y 80.1% respectivamente). La sensibilidad y especificidad también es superior con la Laplaciana (63.3% y 86.7%).

El último artículo que se va a comentar relaciona la hipertensión como efecto adverso de la hiponatremia [13]. Para construir el modelo del estudio se usan los datos de demografía, clínicos y datos de laboratorio de SPRINT (Systolic Blood Pressure Intervention Trial) de 9361 participantes con hipertensión y sin diabetes. Con el objeto de evaluar los resultados usan NHANES de 2005-2010.
Se implementa con gradient boosting machine en R con las siguientes descritas en el artículo y se imputan los valores que falten como la mediana de la columna. Se usa el quintil superior del colesterol HDL (HDL-C), (mayor o igual que el percentil 80).
HDL-C es un fuerte predictor, si es elevado es un factor de riesgo. Pacientes hipertensos con elevado HDL-C deberían monitorizarse para hiponatremia.

3. Proceso de Implementación

3.1. Exploración y Procesado de los Datos

El conjunto de datos se compone de seis archivos separados por comas: *demographic*, *diet*, *examination*, *labs*, *medications* y *questionnaire.*

El archivo *demographic* tiene 10175 registros y 47 variables. El cuestionario de *demographic* trata información individual y familia, composición familiar, género, edad, etnia, educación, estado civil,...

Diet tiene 9813 registros y 168 variables. En él se recoge información sobre el tipo de dieta de los participantes y nutrientes.

Examination tiene 9813 registros y 224 variables. Este archivo se compone de datos clínicos acerca de la salud del paciente (dentales, capacidad olfativa, alergias,...) y medidas corporales (pulsaciones, presión sanguínea, altura, peso, perímetro abdominal,...).

Labs tiene 9813 registros y 424 variables. Los datos que podemos encontrar en este archivo son los recogidos en diferentes pruebas analíticas.

Medications tiene 20194 registros y 13 variables. En *medications* encontramos las medicaciones que se toman los participantes: enfermedad, codificación, medicamento,...
En este archivo realizaremos eliminación de aquellos registros que tengan el campo *RXDRSD1* nulo, ya que *RXDRSC1* y *RXDRSD1* están relacionados al ser el código de una enfermedad y descripción respectivamente. Trabajaremos solo con los registros así, teniendo constancia de que enfermedad tiene el paciente sin tener interés el resto.

Questionnaire tiene 10175 registros y 953 variables. Se recoge diferente tipo de información: idioma, consumo de alcohol y tabaco, constancia de tener alguna enfermedad, hábitos alimenticios,...

Después de estas primeras exploraciones para conocer los archivos que vamos a tratar, visionamos los primeros registros, un resumen de cada variable y los valores vacíos o nulos de cada variable. A estos campos les imputaremos el valor el valor más frecuente en el caso de variable categórica y la mediana para variables numéricas.
En el caso de los datos que provienen del archivo medications, se observa que hay más de un registro de paciente, debido a que una persona puede ser diagnosticada con más de una enfermedad. Por ello, se agrupa por la variable SEQN tras descategorizar.

Para la variable RIDAGEMN (edad en meses) y RIDEXAGM (edad en meses en el momento del examen) se le asignará el valor de RIDAGEYR (edad en años) multiplicado por 12.

En el conjunto de datos que tenemos, las enfermedades con mayor frecuencia son:

Enfermedad	Código	Frecuencia Relativa
Essential (primary) hypertension	I10	14.18
Pure hypercholesterolemia	E78.0	9.79
Type 2 diabetes mellitus	E11	5.11
Gastro-esophageal reflux disease	K21	3.86
Major depressive disorder, single episode, unspecified	F32.9	3.55

Asthma	J45	3.09
Anxiety disorder, unspecified	F41.9	3.08

Tabla 2. Enfermedades con mayor frecuencia.

Veámoslo representado en un gráfico de tarta.

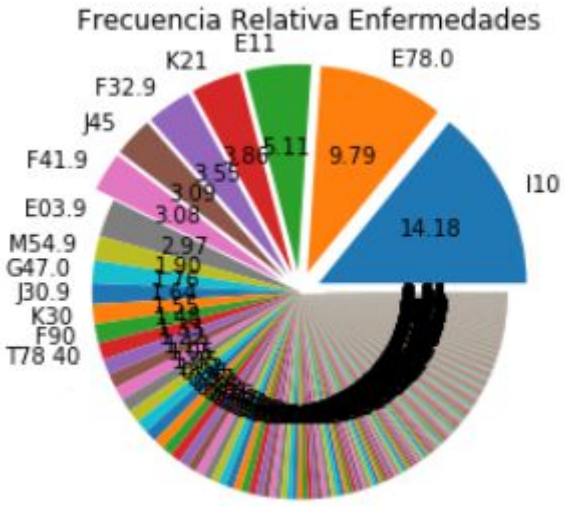

Figura 4. Gráfico de tarta con frecuencia relativa de las enfermedades.

Para conseguir el mejor modelo de agrupación se eliminarán las variables que están correladas (cuya correlación sea mayor o igual que 0.8 o menor o igual que - 0.8) para evitar variables de más. Se descategorizarán las variables que no son numéricas para poder usar el máximo de variables de las que disponemos con *get_dummies* de la librería Pandas.

Por último, se normalizarán los datos con el método *StandardScaler* de la librería *sklearn*.

3.2. Clustering

Con el objetivo de encontrar patrones escondidos entre los datos, vamos a aplicarle el modelo no supervisado de clustering. Se usará el algoritmo *k-Means* de la librería *sklearn* [14] de python.

***KMeans**(n_clusters=8, init='k-means++', n_init=10, max_iter=300, tol=0.0001, precompute_distances='auto', verbose=0, random_state=None, copy_x=True, n_jobs=None, algorithm='auto')*

Solamente se modifica el parámetro *n_clusters* a 6 en el clustering general y a 4 en el de las enfermedades, el resto son los valores por defecto especificados anteriormente.

3.2.1. Clustering General

Hemos implementado el algoritmo de la regla del codo para elegir el valor adecuado de clústeres, que será el que minimiza el error *SSE*. Tras analizar la gráfica de salida se escoge *k = 6*, ya que a partir de este valor el error es asumible y no especializaremos los clústeres que se obtengan.

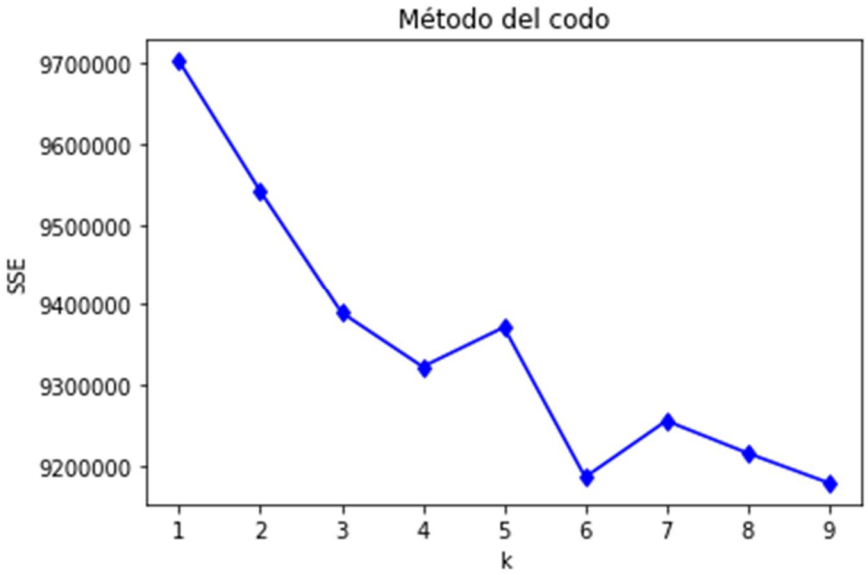

Figura 5. Método del Codo.

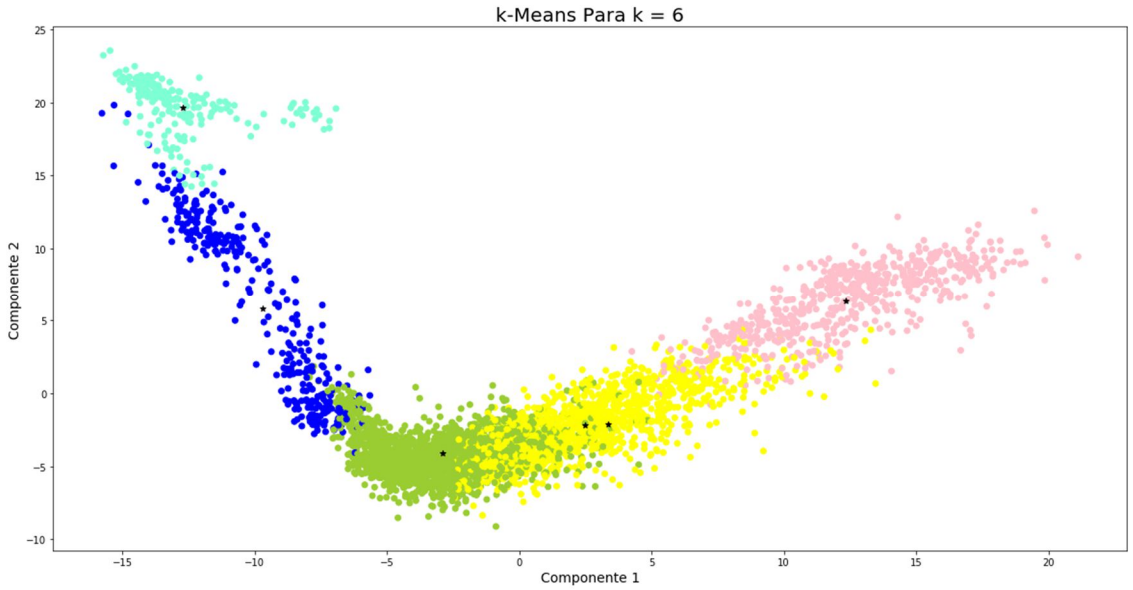

Figura 6. Modelo de clustering obtenido.

En la Figura 6 podemos ver solamente cinco grupos diferenciados por los colores del clúster, cuando el clustering es sobre *k = 6*. Esto es debido a que uno de los clústeres está compuesto solamente

por un elemento. Por lo tanto, coincide con el centroide y éstos están representados por la forma de estrella negra.

```
Tamaño del cluster1: 1
Tamaño del cluster2: 1683
Tamaño del cluster3: 611
Tamaño del cluster4: 1199
Tamaño del cluster5: 330
Tamaño del cluster6: 192
```

En la figura 7 se señala la posición de este clúster sobre la gráfica.

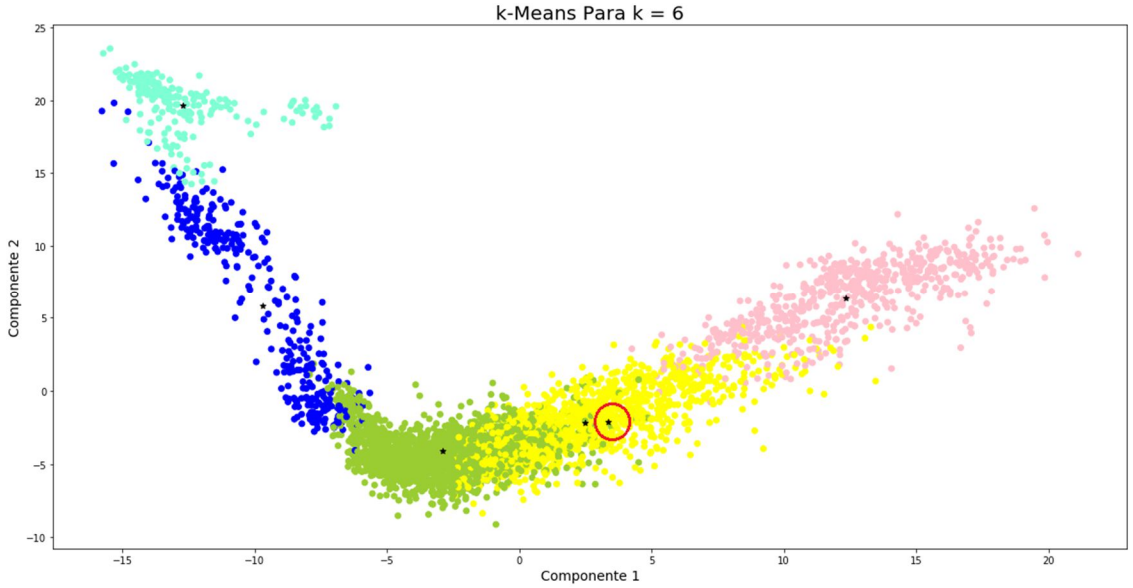

Figura 7. Detalle de cluster1.

Para poder entender de qué forma se organizan los datos, vamos a hallar un modelo predictivo con *Random Forest* cuya variable objetivo será la etiqueta del clúster al que pertenece. Con este modelo podemos obtener la importancia de las variables que deciden a que clúster pertenece un registro.

Las diez variables más importantes son:

Descripción	Código Variable	Importancia
Edad en años	RIDAGEYR	0.016688
Buen estado del 2º premolar arriba derecha permanente	OHX04CTC_S	0.013943
Buen estado del 2º premolar abajo derecha permanente	OHX29CTC_S	0.01197

Buen estado del 2º premolar arriba izquierda permanente	OHX13CTC_S	0.011789
Buen estado del 2º premolar abajo izquierda permanente	OHX20CTC_S	0.011515
Estado 2º molar abajo izquierda	OHX18TC	0.010943
Canino arriba izquierda perdido por enfermedad dental y reemplazado	OHX11CTC_P	0.010511
Estado incisivo abajo izquierda	OHX23TC	0.010231
Buen estado del incisivo central arriba izquierda permanente	OHX09CTC_S	0.01016
2º molar arriba derecha sin erupciones	OHX02CTC_U	0.010122

Tabla 3. Importancia de las variables para el Clustering.

Estas variables son las más importantes para decidir al clúster que va cada punto. Por el resultado de las diez mejores variables, se concentra la decisión en la edad del paciente y en el estado de salud dental.

3.2.2. Clustering de las Enfermedades

Se realiza un segundo clustering con únicamente las enfermedades diagnosticadas por paciente con *kMeans* de *sklearn* [14].

En el método del codo obtenido para este clustering se visualiza que $k = 4$ es un buen parámetro para obtener el modelo *k-Means*.

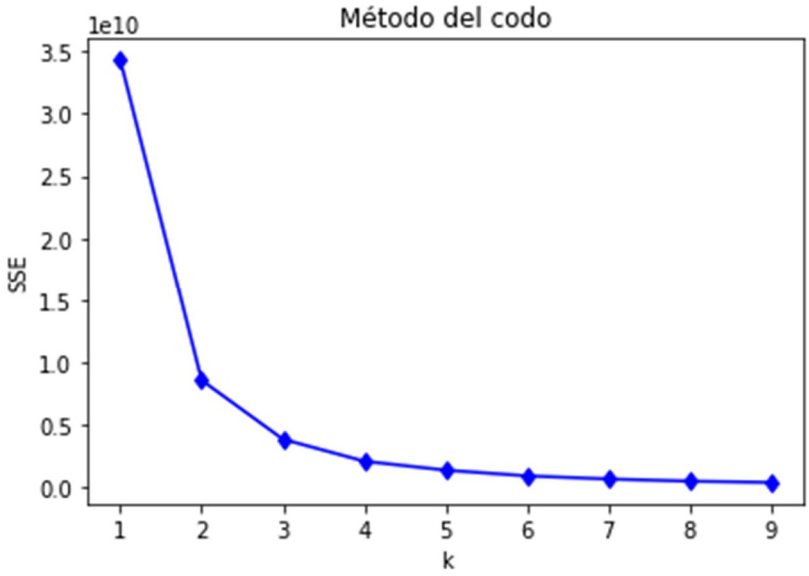

Figura 8. Método del codo para Clustering de enfermedades.

El modelo de clustering obtenido, lo podemos ver en la figura 9.

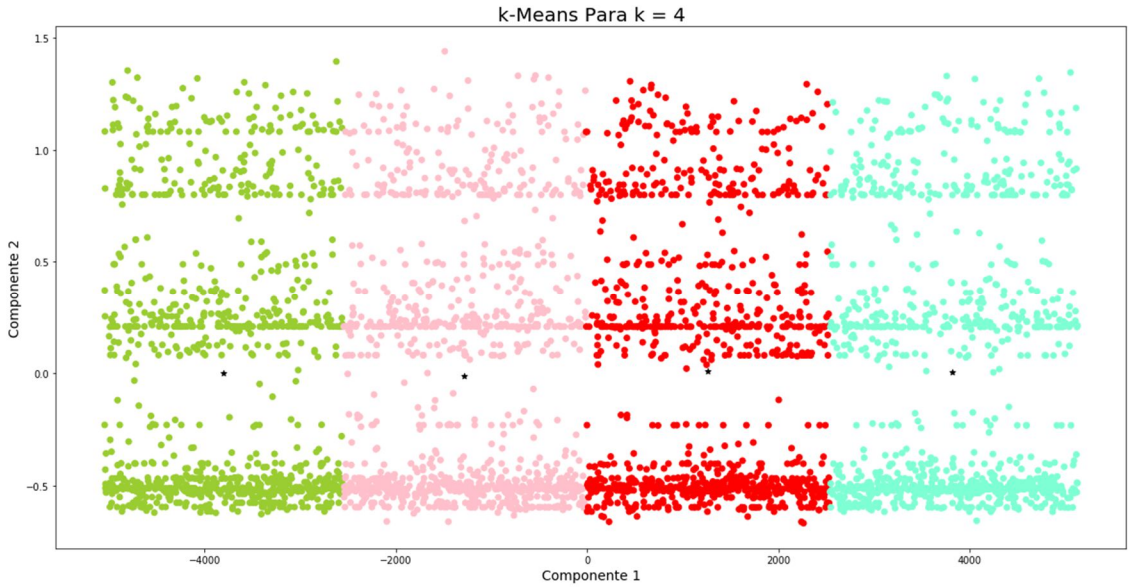

Figura 9. Clustering para enfermedades obtenido.

Cuando se calcula la matriz de correlación, se obtienen las siguientes relaciones:

```
[['RXDRSC1_A44.9', 'RXDRSC1_A69.2'], 0.81639486],
[['RXDRSC1_C74', 'RXDRSC1_R20.0'], 1.0],
[['RXDRSC1_C74', 'RXDRSC1_Z51.11'], 1.0],
[['RXDRSC1_D25.P', 'RXDRSC1_Z94.0'], 1.0],
[['RXDRSC1_D35.2', 'RXDRSC1_E27.40'], 1.0],
[['RXDRSC1_I49.1', 'RXDRSC1_K91.5'], 1.0],
[['RXDRSC1_K12.2', 'RXDRSC1_M25.52'], 1.0],
[['RXDRSC1_R20.0', 'RXDRSC1_Z51.11'], 1.0]]
```

Tras identificar y contar las enfermedades que hay por clúster, comprobamos si las enfermedades que resultan como correladas, están en el mismo clúster.

Se obtiene efectivamente, que "Bartonellosis, unspecified" (RXDRSC1_A44.9) y "Lyme disease" (RXDRSC1_A69.2) están relacionadas con el mismo paciente en el clúster 3 y clúster 4 de las dos y tres veces respectivas que están estas enfermedades en los datos.

También "Malignant neoplasm of adrenal gland" (RXDRSC1_C74) y "Anesthesia of skin" (RXDRSC1_R20.0) se encuentran relacionadas en el mismo sujeto en el clúster 2. De igual manera "Malignant neoplasm of adrenal gland" (RXDRSC1_C74) y "Encounter for antineoplastic chemotherapy" (RXDRSC1_Z51.11) están relacionadas con el mismo sujeto y se encuentran en el clúster 2. Por otra parte "Anesthesia of skin" (RXDRSC1_R20.0) y "Encounter for antineoplastic chemotherapy" (RXDRSC1_Z51.11) se relacionan en el clúster 2.

La relación entre "Prevent uterine fibroids" (RXDRSC1_D25.P) y "Kidney transplant status" (RXDRSC1_Z94.0) es visible, en que se encuentran en el clúster 1 y que no vuelve a aparecer más en los datos. De igual manera ocurre con "Benign neoplasm of pituitary gland" (RXDRSC1_D35.2) y "Unspecified adrenocortical insufficiency" (RXDRSC1_E27.40) en el clúster 2, con "Atrial premature depolarization" (RXDRSC1_I49.1) y "Postcholecystectomy syndrome" (RXDRSC1_K91.5) en el clúster 3 y con "Cellulitis and abscess of mouth" (RXDRSC1_K12.2) y "Pain in elbow " (RXDRSC1_R20.0) en el clúster 4.

3.3. Modelo Predictivo

Para implementar el modelo predictivo, previamente se han seleccionado variables del dataframe según el criterio de riesgos mencionados en capítulos anteriores. Estos riesgos son: enfermedades cardiovasculares, enfermedades renales, ingresos, hábitos saludables como alimentación y ejercicio físico, consumo de alcohol y tabaco, edad, género, sobrepeso, antecedentes familiares, etc.

La enfermedad diagnosticada con mayor frecuencia en los pacientes del conjunto de datos es la hipertensión, cuyo código en la variable *RXDRSC1* corresponde a *I10.*

Se realizarán varios modelos predictivos con distintos métodos para poder comparar las métricas de la predicción obtenida respectivamente.

Los modelos desarrollados son: *Support Vector Classification, Gradient Boosting Classifier, AdaBoost Classifier, Random Forest Classifier, Naive Bayes, Logistic Regression y k-NN*. Para todos ellos se hace una primera aproximación de los parámetros con *RandomizedSearchCV* y con los mejores resultados se busca el mejor modelo con *GridSearchCV*, en ambos casos se realiza con una validación cruzada *cv = 10.*

3.3.1. Support Vector Classification

Según la documentación de *sklearn* [15] estos son los siguientes parámetros que tiene el algoritmo y los valores que usa por defecto:

***SVC**(C=1.0, kernel='rbf', degree=3, gamma='auto_deprecated', coef0=0.0, shrinking=True, probability=False, tol=0.001, cache_size=200, class_weight=None, verbose=False, max_iter=-1, decision_function_shape='ovr', random_state=None)*

Los parámetros escogidos para este modelo han sido*:*
C = 10 y gamma = 20.
Se obtiene una precisión de *58.25%*

3.3.2. Gradient Boosting Classifier

Para este modelo, la documentación según *sklearn* [16] indica que los parámetros son:

***GradientBoostingClassifier**(loss='deviance', learning_rate=0.1, n_estimators=100, subsample=1.0, criterion='friedman_mse', min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_depth=3, min_impurity_decrease=0.0, min_impurity_split=None, init=None, random_state=None, max_features=None, verbose=0, max_leaf_nodes=None, warm_start=False, presort='auto', validation_fraction=0.1, n_iter_no_change=None, tol=0.0001)*

Los parámetros modificados para este modelo han sido*:*
learning_rate = 0.1 y n_estimators = 102.
Se obtiene una precisión de *75.95%*

3.3.3. AdaBoost Classifier

Para *AdaBoost* el modelo y los parámetros expuestos en la documentación [17] son:

AdaBoostClassifier*(base_estimator=None, n_estimators=50, learning_rate=1.0, algorithm='SAMME.R', random_state=None)*

Los parámetros que se han cambiado para este modelo han sido*:*
learning_rate = 0.3 y n_estimators = 107.
Se obtiene una precisión de *76.33%*

3.3.4. Random Forest Classifier

En este caso, la documentación de *sklearn* [18] indica que los parámetros del Random Forest son:

RandomForestClassifier*(n_estimators='warn', criterion='gini', max_depth=None, min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False, class_weight=None)*

Los parámetros cambiados para este modelo han sido*:*
max_depth = 50 y n_estimators = 200.
Se obtiene una precisión de *75.65%*

3.3.5. Naïve Bayes

Los parámetros del modelo bayesiano según sklearn [19] son:

GaussianNB*(priors=None, var_smoothing=1e-09)*

El parámetro escogido para este modelo ha sido*:*
var_smoothing = 9.81e-09.
Se obtiene una precisión de *60.34%*

3.3.6. Logistic Regression

En *Logistic Regression*, los parámetros que se pueden modificar según la documentación [20] son:

LogisticRegression*(penalty='l2', dual=False, tol=0.0001, C=1.0, fit_intercept=True, intercept_scaling=1, class_weight=None, random_state=None, solver='warn', max_iter=100, multi_class='warn', verbose=0, warm_start=False, n_jobs=None, l1_ratio=None)*

Los parámetros escogidos para este modelo han sido*:*
penalty = l1 y C = 0.16.
Se obtiene una precisión de *73.41%*

3.3.7. k-NN

Para *KNeighborsClassifier* los parámetros personalizables [21] son:

KNeighborsClassifier*(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None, **kwargs)*

Los parámetros escogidos para este modelo han sido*:*
n_neighbors = 8 y n_weights = distance.
Se obtiene una precisión de *67.06%*

4. Evaluación

De cada uno de los modelos predictivos se obtiene la matriz de confusión asociada:

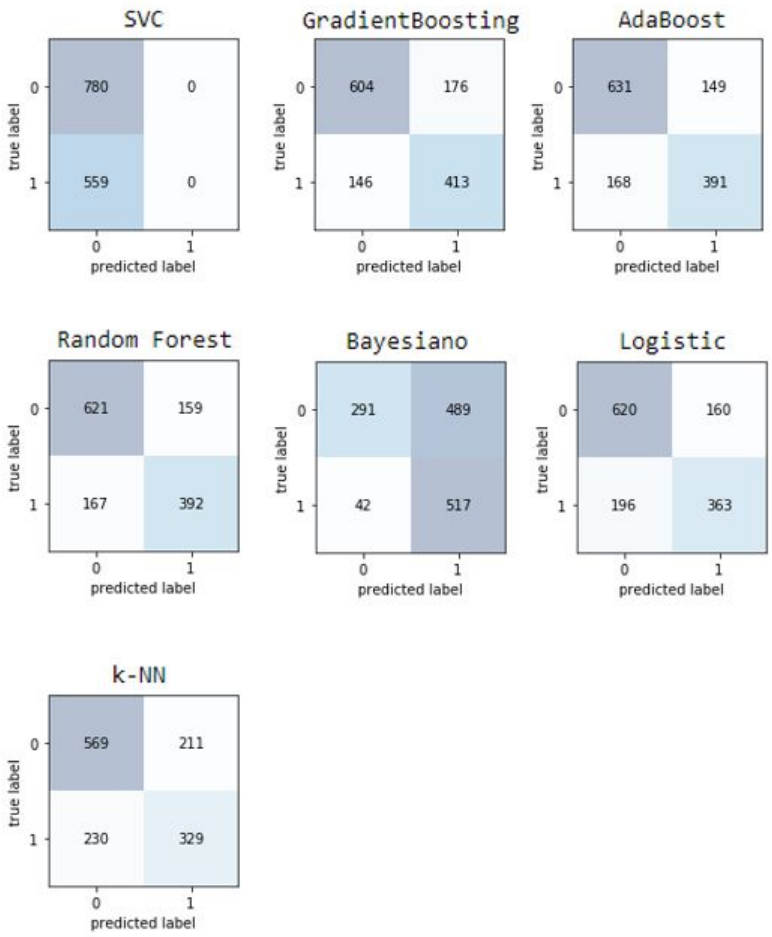

Figura 10. Matriz de confusión de los modelos.

También vamos a comparar las métricas derivadas de la matriz de confusión en la tabla 4 en base a poder compararlas y discutir los resultados obtenidos.

Modelo	ACC	ERR	TPR	FPR	PRE	SPE	F1
Support Vector Classification	58.25	41.75	0	0	-	100	-
Gradient Boosting Classifier	75.95	24.05	73.88	22.56	70.12	77.44	71.95
AdaBoost Classifier	76.33	23.67	69.95	19.10	72.41	80.90	71.16
Random Forest Classifier	75.65	24.35	70.13	20.38	71.14	79.61	70.63
Naive Bayes	60.34	39.66	92.49	62.69	51.39	37.31	66.07
Logistic Regression	73.41	26.58	64.94	20.51	69.41	79.49	67.10
k-NN	67.06	32.94	58.86	27.05	60.93	72.95	59.87

Tabla 4. Métricas de los modelos.

5. Conclusiones

Según los objetivos principales hemos obtenido dos modelos: un modelo de clustering y uno clasificatorio para predecir si un paciente sufre de hipertensión.
Para hallar el modelo predictivo se han probado diferentes algoritmos y ajustado los parámetros de éstos.

Aparte de estos modelos, hemos implementado un segundo modelo de clustering con el archivo *medications* solamente. En *medications* es donde están recogidas las enfermedades diagnosticadas.

En el modelo de clustering es, principalmente las variables importantes son la edad del paciente y las variables relacionadas con el estado de salud dental. Aunque se aprecian grupos diferenciados en relación a la edad, hay ciertos rangos en los que coinciden los clústeres.
Por una parte, es evidente que el resto de variables hacen un papel importante al agrupar y asignan otros clústeres. Por otra parte, la imputación de los valores nulos altera la agrupación y se imputan valores en algunos casos concretos que no son reales. Por ejemplo, en niños de 5 a 8 años se asigna el valor 2 a la variable *OHX18TC,* cuando debería ser 1 o 4.
El valor 2 significa que tiene el 2º molar de abajo e izquierda permanente, y este valor tiene que ser 1 (diente de leche) o 4 (no tiene diente). Al estar el conjunto de datos formado por pacientes adultos mayoritariamente, la mediana es 2 y se imputa este valor incorrectamente. De esta manera el algoritmo decide que está más cercano a niños más mayores (hasta 16 años) en vez de incluirlo en el grupo de niños que tienen esa edad pero tienen el valor correcto al tener que imputarle el valor.

Para el otro modelo de clustering, el de las enfermedades, en vista de los resultados discutidos en el apartado 3.3.2 con los datos que hemos trabajado, se podría pensar que existen las siguientes comorbilidades:

- "Malignant neoplasm of adrenal gland", "Anesthesia of skin" y "Encounter for antineoplastic chemotherapy".
- "Bartonellosis, unspecified" y "Lyme disease".
- "Prevent uterine fibroids" y "Kidney transplant status".
- "Benign neoplasm of pituitary gland" y "Unspecified adrenocortical insufficiency".
- "Atrial premature depolarization" y "Postcholecystectomy syndrome".
- "Cellulitis and abscess of mouth" y "Pain in elbow ".

En el modelo predictivo, según se puede ver en la tabla 4, el que más exactitud (accuracy) ha mostrado ha sido el Adaboost (76.33%).
A pesar que el modelo Naive Bayes sea de los menos exactos, aporta una métrica buena en el contexto que estamos trabajando. En el área de la medicina es especialmente importante minimizar los falsos negativos, es decir, que un paciente que tenga una enfermedad sea diagnosticado como que no la tiene, sobre todo si es una enfermedad grave. Este modelo es el que mejor TPR presenta (62.69%), que está inversamente relacionado con los falsos negativos.

$$TPR = \frac{TP}{FN + TP}$$

Maximizar los falsos positivos también puede suponer un coste económico innecesario, por lo que también es importante minimizarlo. Este modelo es el que peor precisión tiene, y está relacionado con los falsos positivos inversamente. En este sentido es un mal resultado por el coste médico que puede ocasionar.

$$PRE = \frac{TP}{TP + FP}$$

Por lo tanto, se han conseguido los objetivos principales marcados en el proyecto.

6. Línea de Trabajo Futura

Debido a todas las variables que disponemos en el conjunto de datos inicial, se podrían realizar más modelos predictivos basados en otras enfermedades del mismo: diabetes, hipercolesterolemia, cáncer.
En esta línea, una mejora del modelo predictivo sería obtener uno más complejo que generalizara correctamente a cualquier enfermedad.

Otra mejora posible, sería la obtención de un modelo de clustering con la eliminación de los registros nulos o la imputación de valores con otro tipo de métrica segmentada. Así, se conseguiría un modelo de agrupación que tuviera una interpretación más sencilla.

Para el clustering de las enfermedades, aplicaría el modelo a otro conjunto de datos con el que poder fundamentar las comorbilidades obtenidas y que puedan ser base de una futura investigación en el área de la medicina.

7. Código

El código implementado se puede ver en github [22].

8. Glosario

- **AENET.** Adaptive Elastic-Net (Elastic-Net flexible).
- **BMI.** Body Mass Index.
- **ERS.** Environmental Risk Score (Calificación de entorno de riesgo).
- **HDL-C.** Colesterol HDL
- **IMC.** Índice de Masa Corporal.
- **kNN.** K-nearest neighbor (K-vecinos más cercanos).
- **LDA.** Linear Discriminant Analysis (Análisis Discriminante Lineal).
- **NCHS.** National Center for Health Statistics.
- **NHANES.** National Health and Nutrition Examination Survey. Programa de estudios de la National Center for Health Statistics diseñados para evaluar la salud y estado nutricional de adultos y niños de Estados Unidos.

- **OMS.** Organización Mundial de la Salud.
- **PCA**. Principal Component Analysis (Análisis de Componentes Principales).
- **QDA.** Quadratic Discriminant Analysis (Análisis Discriminante Cuadrático).
- **SSE.** Error Sum of Square (Suma de Cuadrados de los Errores).
- **SVM**. Support Vector Machines (Máquinas de Soporte Vectorial).
- **T2DM.** Type-2 Diabetes Mellitus
- **WC.** Waist Circumference (Perímetro de la cintura).

9. Bibliografía y Referencias

[1] Instituto de ingeniería del conocimiento [Consulta: 20 de Marzo de 2019]
< http://www.iic.uam.es/lasalud/big-data-en-medicina-aplicaciones-utiles/>

[2] Bloomberg [Consulta: 29 de Febrero de 2019]
<https://www.bloomberg.com/news/articles/2019-02-24/spain-tops-italy-as-world-s-healthiest-nation-while-u-s-slips?srnd=premium-europe>

[3] Kaggle [Consulta: 20 de Febrero de 2019]
<https://www.kaggle.com/cdc/national-health-and-nutrition-examination-survey>

[4] Center for Disease Control and Prevention [Consulta: 27 de Febrero de 2019]
<https://www.cdc.gov/Nchs/Nhanes/about_nhanes.htm>

[5] ProjectLibre [Consulta: 25 de Febrero]
< https://www.projectlibre.com/>

[6] Organización Mundial de la Salud (2013). “Información general sobre la hipertensión en el mundo”. Ginebra.

[7] López-Martínez F., Schwarcz.MD A., Núñez-Valdez E.R. & García-Díaz V. (2018). “Machine Learning Classification Analysis for a Hypertensive Population as a Function of Several Risk Factors”. *Expert Systems With Applications* (vol. 110, pág. 206-215).

[8] Husain, A. & Khan M.H. (2018). “Early Diabetes Prediction Using Voting Based Ensemble Learning”. *Communications in Computer and Information Science* (vol. 905, pág. 95-103). Singapur: Springer.

[9] Wang, X., Mukherjee, B. & Park, S.K. (2018). “Associations of cumulative exposure to heavy metal mixtures with obesity and its comorbidities among U.S. adults in NHANES 2003–2014”. *Environment International* (vol. 121, pág 683–694).

[10] Kawanabe T., Kamarudin N.D., Ooi C.Y., Kobayashi F., Mi X., Sekine S., Wakasugi A., Odaguchi H. & Hanawa T. (2016). “Quantification of tongue colour using machine learning in Kampo medicine”. *European Journal of Integrative Medicine* (núm. 8, pág. 932-941).

[11] Kublanov V.S., Dolganov A.Y, Belo D. & Gamboa H. (2017). “Comparison of Machine Learning Methods for the Arterial Hypertension Diagnostics”. *Applied Bionics and Biomechanics* (vol. 2017, Article ID 5985479, 13 pages).

[12] Pei, Z., Liu, J., Liu, M., Zhou, W., Yan, P., Wen, S. & Chen, C. (2018). “Risk-Predicting Model for Incident of Essential Hypertension Based on Environmental and Genetic Factors with Support Vector Machine”. *Interdisciplinary Sciences: Computational Life Sciences.* (vol. 10, pág. 126-130). Alemania.

[13] Israel, A. & Grossman, E. (2017). “Elevated High Density Lipoprotein Cholesterol is associated with hyponatremia in hypertensive patients”. *The American Journal of Medicine* (vol. 130, pág. 1324.e7-1324.e13).

[14] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

[15] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html >

[16] Scikit-learn [Consulta: 7 de Junio de 2019]<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>

[17] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>

[18] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

[19] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html>

[20] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html>

[21] Scikit-learn [Consulta: 7 de Junio de 2019] <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html>

[22] Github [Consulta: 09 de Junio de 2019] <https://github.com/marcreest1/TFM>