Clasificador Multi-nivel para la identificación de individuos en riesgo de desarrollar sobrepeso
Multi-level classifier for identifying individuals at risk of developing overweight

Trabajo de Fin de Máster
Curso 2020–2021

Autor
Alberto Gutiérrez Gallego

Director
José Ignacio Hidalgo Pérez

Colaborador
Trabajo Financiado por la Comunidad de Madrid y Fondo Social Europeo a través del Proyecto GENOBIA-CM S2017/BMD-3773

Máster en Ingeniería Informática
Facultad de Informática
Universidad Complutense de Madrid

Clasificador Multi-nivel para la identificación de individuos en riesgo de desarrollar sobrepeso
Multi-level classifier for identifying individuals at risk of developing overweight

Trabajo de Fin de Máster en Ingeniería Informática
Departamento de Arquitectura de Computadores y Automática

Autor
Alberto Gutiérrez Gallego

Director
José Ignacio Hidalgo Pérez

Colaborador

Trabajo Financiado por la Comunidad de Madrid y
Fondo Social Europeo a través del Proyecto GENOBIA-CM
S2017/BMD-3773

Convocatoria: *Febrero* 2021
Calificación: 8

Máster en Ingeniería Informática
Facultad de Informática
Universidad Complutense de Madrid

10 de febrero de 2021

Colaboración y Financiación

Dedicatoria

A mis abuelos y mis padres por enseñarme a luchar en la vida. A mi novia Verónica por estar apoyándome y aguantándome día a día. Y por último y no menos importante a mi hermano que va siguiendo mis pasos.

Agradecimientos

A José Ignacio Hidalgo Pérez, por confiar en mí y darme la oportunidad de trabajar en el grupo ABSYS dentro del departamento de Arquitectura de Computadores y Automática de la Universidad Complutense de Madrid.

A Antonio López Farré por brindarnos la oportunidad de trabajar en el proyecto Genobia-CM, así como sus aportaciones y consejos durante el desarrollo de este trabajo, las cuales agilizaron el proceso de entendimiento del problema.

Resumen

Clasificador Multi-nivel para la identificación de individuos en riesgo de desarrollar sobrepeso

En este trabajo se ha desarrollado un sistema clasificador basado en técnicas de inteligencia artificial y aprendizaje automático. El sistema consiste en un novedoso mecanismo de clasificación Multi-nivel en el que se combinan hasta tres clasificadores diferentes. Éstos han sido seleccionados de un conjunto de 14 algoritmos de clasificación supervisada mediante un proceso previo de validación cruzada. El clasificador funciona colocando por niveles distintos algoritmos de clasificación y estableciendo unos umbrales para la aceptación de la respuesta de cada uno de los algoritmos. Los algoritmos de clasificación proporcionan tanto el grupo que asignan, como la probabilidad de que esa clasificación sea correcta. Si la probabilidad que proporciona el primer algoritmo es superior a este umbral, la decisión del clasificador se da por correcta, en caso contrario se solicita una clasificación al algoritmo del siguiente nivel y así sucesivamente.

Este clasificador Multi-nivel se ha diseñado específicamente para adaptarse a los datos del proyecto Genobia-CM, aunque su diseño permite aplicarlo a cualquier otro problema que utilice el formato de datos de entrada adecuado, que es el habitual en problemas de clasificación. Genobia es un proyecto participado por un consorcio de 20 instituciones, hospitales y empresas, financiado por el Fondo Social Europeo y la Comunidad de Madrid (genobia.es). El proyecto busca diseñar, utilizando inteligencia artificial, algoritmos predictivos para la identificación de personas en riesgo de desarrollar sobrepeso, obesidad y sus patologías asociadas. En este trabajo se ha utilizado una base de datos con 1179 individuos proporcionada por el Consorcio en el que se recoge información de los hábitos de vida y adherencia a la dieta mediterránea. Se ha implementado el clasificador Multi-nivel como algoritmo predictivo y de clasificación del riesgo de padecer sobrepeso, adaptándolo a los datos proporcionados donde el número de casos de obesidad es muy reducido. Mediante nuestra propuesta se consigue reducir el número de falsos negativos, lo que es fundamental dentro del problema en cuestión, ya que, al tratarse de salud pública, esto implica reducir el número de acciones clínicas erróneas u omitidas.

Los resultados obtenidos rondan el 80 % de tasa de exactitud y nuestro sistema está perfectamente preparado para aceptar los datos que proporcione el consorcio en el futuro. Estos datos incluirán información genética de cada individuo y esperamos que además incluya un mayor número de casos. Además, se han realizado otros tipo de clasificadores basados en árboles de decisión, así como un exhaustivo análisis de las variables, su influen-

cia en los modelos, redundancias y un estudio de sensibilidad de los modelos a las mismas.

Trabajo Financiado por la Comunidad de Madrid y Fondo Social Europeo a través del Proyecto GENOBIA-CM con referencia S2017/BMD-3773.

.

Palabras clave

- Clasificador Multi-nivel
- Aprendizaje Automático
- Sobrepeso
- Obesidad
- Dieta mediterránea
- Gradient Boosting
- Árboles de decisión

Abstract

Multi-level classifier for identifying individuals at risk of developing overweight

In this work, a classification system based on artificial intelligence and automatic learning techniques has been developed. The system consists of a novel multi-level classification mechanism in which up to three different classifiers are combined. These have been selected from a set of 14 supervised classification algorithms through a previous process of cross validation. The classifier works by placing by levels different classification algorithms and establishing thresholds for the acceptance of the response of each of the algorithms. The classification algorithms provide both the group they assign and the probability that the classification is correct. If the probability provided by the first algorithm is higher than this threshold, the decision of the classifier is considered to be correct, otherwise a classification is requested to the algorithm of the next level and so on.

This multi-level classifier has been specifically designed to adapt to the data of the Genobia-CM project, although its design allows it to be applied to any other problem using the appropriate input data format, which is the usual one in classification problems. Genobia is a project participated by a consortium of 20 institutions, hospitals and companies, financed by the European Social Fund and the Community of Madrid (genobia.es). The project seeks to design, using artificial intelligence, predictive algorithms for the identification of people at risk of developing overweight, obesity and their associated pathologies. In this work, a database with 1179 individuals provided by the Consortium has been used to collect information on living habits and adherence to the Mediterranean diet. The multi-level classifier has been implemented as a predictive and classification algorithm of the risk of suffering from overweight, adapting it to the data provided where the number of cases of obesity is very low. By means of our proposal, the number of false negatives is reduced, which is fundamental within the problem in question, since being a matter of public health, this implies reducing the number of erroneous or omitted clinical actions.

The results obtained are around 80% accurate and our system is perfectly prepared to accept the data provided by the consortium in the future. This data will include genetic information of each individual and we hope that it will also include a greater number of cases. In addition, other types of classifiers based on decision trees have been carried out, as well as an exhaustive analysis of the variables, their influence on the models, redundancies and a study of the sensitivity of the models to them.

Work financed by the Community of Madrid and the European Social Fund through the Project GENOBIA-CM with reference S2017/BMD-3773.

Keywords

- Multi-level classifier
- Machine Learning
- Overweight
- Obesity
- Mediterranean Diet
- Gradient Boosting
- Decision trees

Índice

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La OMS considera al sobrepeso como una epidemia global que constituye un problema de salud pública, fundamentalmente en los países desarrollados, aunque también está empezando a aparecer en países en vías de desarrollo. Es por ello necesario que las autoridades e instituciones públicas estén cada vez más concienciadas de que se trata de un problema serio y que es preciso reducirlo y evitar en la medida de lo posible que siga creciendo su incidencia en la población considerada sana. Hay una gran cantidad de factores y variables que participan en el desarrollo del sobrepeso y de la obesidad y no es una tarea fácil predecir de forma individualizada el riesgo de desarrollar estas patologías y sus comorbilidades asociadas (Organization, 2000).

Según un estudio realizado por el Instituto Nacional de Estadística (Ministerio de Sanidad, 2017) se afirma que, en los últimos 30 años, la prevalencia de obesidad en España se ha multiplicado por 2.4, pasando de 7.4 % en 1987 al 17.4 % en 2017. Analizando los datos se observa que los casos de obesidad y sobrepeso son superiores en los hombres frente a las mujeres. Respecto a la obesidad infantil, ya existe un 10 % de niños entre 2 y 17 años que la padecen, es por ello necesario prevenir casos futuros de sobrepeso u obesidad.

El gran desarrollo de la inteligencia artificial y aprendizaje automático abre una puerta a realizar de forma adecuada la tarea de indagar en las causas de la aparición y desarrollo del sobrepeso y la obesidad. Se entiende como inteligencia artificial el desarrollo de sistemas dotados de procesos intelectuales propios de los seres humanos. Entre estos procesos encontramos el razonamiento, la generalización, la mejora mediante experiencias pasadas y el descubrimiento de significados (Copeland, 2020) . Los avances en este sector son continuos y certeros, pero todavía queda mucho hasta alcanzar las marcas establecidas por el hombre. Este trabajo de Fin de Máster se encuadra en el marco del proyecto GENOBIA-CM cuyo objetivo principal es diseñar algoritmos predictivos y de clasificación para identificar personas en riesgo de desarrollar sobrepeso/obesidad y sus patologías asociadas.

A lo largo de este trabajo se han realizado pruebas con un gran número de algoritmos de clasificación supervisada. El objetivo de la clasificación es generar un modelo para poder predecir una clase dados unos valores de entrada. Estos valores de entrada son objetos caracterizados que pertenecen a diferentes clases. En este caso se está utilizando una clasificación binaria, ya que existen dos tipos de clase: 0 sin sobrepeso y 1 con sobrepeso. Algunos de los algoritmos de clasificación usados son Regresión Logística, Gradient Boos-

ting, árboles de decisión o Random Forest. La incorporación de estas y otras técnicas de aprendizaje en el ámbito médico cada vez es mayor.

Para la implementación del proyecto se ha empleado Python, un lenguaje de programación interpretado, no compilado, conocido por su flexibilidad, potencia y sencillez. Entre sus ventajas, también cuenta con la capacidad de soportar diferentes paradigmas como orientación a objetos, programación imperativa y estructurada, por poner un ejemplo. Otra de las ventajas es que Python nos permite recurrir a bibliotecas especializadas.

1.1. Motivación

La motivación fundamental de esta investigación es ayudar a los sistemas e instituciones de salud pública a reducir la incidencia del sobrepeso en la población de la Comunidad de Madrid. Debido a que se trata de un problema de salud pública, es de vital importancia clasificar correctamente aquellos usuarios que padecerán dichas afecciones, por ello este proceso requiere prestar especial atención a la seguridad con la que el modelo da sus predicciones. La inteligencia artificial en general y el aprendizaje automático en particular son las herramientas de las que disponemos y que nos permiten confiar en el éxito de nuestra tarea.

1.2. Objetivos

Los objetivos principales del trabajo son:

1. Diseñar un sistema de clasificación de sujetos en riesgo de padecer sobrepeso de la máxima precisión.
2. Proporcionar a los profesionales de la salud implicados en Genobia sistemas clasificadores basados en árboles de decisión que les permitan indagar en su funcionamiento sin la necesidad de tener un conocimiento profundo de las técnicas informáticas.

1.3. Plan de trabajo

Para conseguir los objetivos mencionados se va a abordar el siguiente plan de trabajo que constituye de manera ordenada los objetivos secundarios que se mencionan a continuación:

- Realizar un preprocesado y curado de los datos adecuado
- Analizar las dependencias entre las variables de entrada
- Estudiar el funcionamiento de los algoritmos clásicos de clasificación y su rendimiento.
- Realizar un análisis de sensibilidad e impacto de las variables sobre los modelos.
- Obtener información del resultado de la clasificación de los modelos aplicados que sea útil para los profesionales sanitarios.
- Diseñar un sistema de clasificación que recoja todas las conclusiones obtenidas de los pasos anteriores y combine de la manera más adecuada los algoritmos más utilizados en la literatura actual.

- Diseñar sistemas clasificadores basados en árboles de decisión.
- Particularizar los sistemas de clasificación para distintos segmentos de la población, separados por sexo y edad.

El desarrollo del proyecto ha seguido un sistema de entregas periódicas de resultados. Siguiendo esta metodología, se han realizado un total de seis lotes de experimentos. Inicialmente se ha realizado una etapa de preprocesamiento, análisis de variables y, finalmente, se ha realizado la generación de árboles de decisión. A continuación, se mencionan algunos detalles adicionales de este plan de trabajo.

- **Preprocesamiento**
  Antes de realizar las primeras pruebas era indispensable tratar los datos recibidos, modificando valores y eliminando variables innecesarias.
- **Análisis de sensibilidad con la herramienta SAlib**
  Medición del impacto de variables entre ellas y a posibles cambios.
- **Primer lote de experimentos**
  Primeras pruebas con modelos e interpretación de los mismos.
- **Segundo lote de experimentos**
  Estudio de las variables utilizadas mediante la herramienta SHAP y pruebas de los modelos con validación cruzada respecto a diferentes conjuntos de poblaciones en función del sexo y la edad.
- **Tercer lote de experimentos**
  Uso de cotas para medir la seguridad de los modelos en su toma de decisiones.
- **Cuarto lote de experimentos**
  Desarrollo del clasificador Multi-nivel y toma de decisiones respecto a las variables en uso.
- **Quinto lote de experimentos**
  Pruebas sobre el clasificador Multi-nivel.
- **Sexto lote de experimentos**
  Nueva selección de variables.
- **Generación de árboles**
  Se usa el modelo Random Forest Regressor para la generación de árboles y posteriormente su estudio por parte del equipo médico.

1.4. Organización de la memoria

El resto de la memoria está organizada de la siguiente manera:

- Capítulo 1: Introducción, objetivos y plan de trabajo.
- Capítulo 2: Breve análisis documental sobre artículos relacionados con el tema a tratar.
- Capítulo 3: Desarrollo de las tecnologías empleadas a lo largo del proyecto.

- Capítulo 4: Descripción y funcionamiento del clasificador multinivel.
- Capítulo 5: Documentación del trabajo realizado y los resultados obtenidos.
- Capítulo 6: Conclusiones generales y trabajo futuro.

Capítulo 2

Estado del arte

Hoy en día las diferentes técnicas de aprendizaje automático están presentes en muchos ámbitos, es por ello por lo que también están apareciendo en los dominios relacionados con la salud. Una de las técnicas en auge últimamente son las redes neuronales, las cuales permiten correlacionar los parámetros de entrada con los de datos de salida correspondientes. Existen bastantes casos prácticos de uso de redes neuronales en aplicaciones médicas, sobre todo relacionados con el análisis de imágenes para predecir riesgos de salud o enfermedades (Litjens et al., 2017). En este artículo en cuestión, se recogen y resumen alrededor de 300 contribuciones a dicho sector, desde la detección de objetos hasta clasificación de imágenes. Gracias al auge de este tipo de algoritmos, ha permitido hacer nuevos estudios en distintas áreas de aplicación: neuro, retiniana, pulmonar, patología digital, mamaria, cardíaca, abdominal y musculoesquelética por nombrar algunas de ellas.

Entre otros artículos destacables, encontramos el trabajo de Khalaf et al. (Khalaf et al., 2017), que trata el uso de varios algoritmos de aprendizaje automático para predecir la cantidad exacta de dosis de medicación necesaria para pacientes con anemia de células falciformes. Con este fin, realizaron pruebas para estudiar la precisión y el rendimiento de distintos algoritmos a la hora de dar respuestas a problemas relacionados con la medicina. Los resultados obtenidos en el documento señalan al Random Forest Classifier como la mejor de las opciones a la hora de realizar este trabajo.

Sobre estudios relacionados con la predicción de sobrepeso/obesidad no se encuentran un gran número de artículos con éxito y es por ello por lo que es un incentivo adentrarse en este campo. Algunos de los artículos relacionados encontrados tratan de identificar problemas de obesidad infantil, utilizando diferentes técnicas de aprendizaje automático. Cabe destacar que el IMC infantil[1] se calcula en este trabajo teniendo en cuenta también la edad y sexo, aparte del peso y la altura. En este caso (Singh y Tawfik, 2020) utilizan una fuente de datos con niños de 3, 5, 7 y 11 años para predecir el riego que corren de padecer sobrepeso u obesidad a los 14 años. Para ello usan métodos de aprendizaje automático como SVM o Random Forest obteniendo de media un 90 % de tasa de exactitud. Obtienen tan buenos resultados debido a que están introduciendo como variables de entrada los IMC de los niños de 3,5,7 y 11 años, lo cual conlleva a que están utilizando la edad, sexo, altura y peso de los mismos.

[1]IMC: Índice de Masa Corporal. El valor del IMC se utiliza para determinar si una persona tiene obesidad o no. El IMC se calcula como $Peso/Altura^2$. Si el IMC es mayor que 25 se considera sobrepeso.

Relacionados con árboles de decisión para predecir la obesidad se encuentra (De la Hoz Manotas et al., 2019), el cual utilizó datos de estudiantes entre 18 y 25 años. Recopilaron información relacionada con el sexo, edad, peso, frecuencia de actividad física e ingesta de alimentos, entre otros. Utilizando algoritmos como Random Forest y logistic regression obtienen una tasa de acierto entre el 91 % y 97 %. Obviamente obtienen estos resultados debido a que están introduciendo como variables de entrada para el modelo el peso e IMC de los individuos y por tanto el modelo conoce con exactitud dichos casos.

Por último, en (Muhamad Adnan et al., 2012) se propone el uso de Naïve Bayes para la predicción de la obesidad infantil. Este método ya se ha empleado en muchas ocasiones independientemente del dominio del problema, dando muy buenos resultados. Un problema al que se tuvieron que enfrentar fue la debilidad del predictor ante variables con valor 0. Como solución a este problema, se propuso el uso de un algoritmo genético para la optimización de parámetros. Para realizar la predicción emplean variables de 3 tipos, el ambiente familiar, el estilo de vida y datos del niño en cuestión. En sus resultados se muestra una mejora del 75 % en la precisión.

En resumen, no se encuentran muchos trabajos relacionados con la predicción de obesidad y sobrepeso, y los que existen contienen algunas decisiones de diseño bastante mejorables. Todos ellos aplican algoritmos de Machine Learning de manera individual y en ningún caso combinan varios de ellos. Además, los datos utilizados no suelen superar los 1000 individuos. En este trabajo se trata de abordar todas estas carencias mediante un clasificador Multi-nivel con una tasa de acierto fiable e identificación de los factores relacionados con el sobrepeso/obesidad. Además el proyecto Genobia proporciona una base de datos de 1179 individuos.

Capítulo 3

Tecnologías Empleadas

En este apartado se explicarán las diferentes tecnologías empleadas a lo largo del proyecto, así como las herramientas empleadas para ello. El Proyecto se ha realizado en su totalidad en el lenguaje Python. Este lenguaje se eligió debido a su flexibilidad y sencillez, permitiendo realizar pruebas en pocos minutos y proporcionando un gran número de bibliotecas con las que poder abordar los diferentes puntos del proyecto.

3.1. Librerías de python

3.1.1. Sklearn

Sklearn (Cournapeau, 2012) es una biblioteca para aprendizaje automático de software libre para Python. Esta incluye varios algoritmos de clasificación, regresión y análisis de grupos, además de tener diferentes técnicas de preprocesamiento de datos y estudio de los mismos. De dicha biblioteca se han utilizado los modelos citados a lo largo del proyecto, así como algunas técnicas de preprocesamiento y selección de variables.

3.1.2. Graphviz

Graphviz (Bilgin et al.) es un conjunto de herramientas software para el diseño de diagramas definido en el lenguaje descriptivo DOT. Estos diagramas se han usado sobre todo para generar los diferentes árboles obtenidos por los modelos Random Forest.

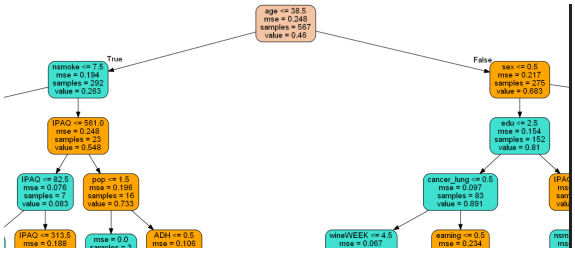

Figura 3.1: Ejemplo árbol con Graphviz

3.1.3. Numpy

Numpy (Oliphant, 1995) es una biblioteca para Python que permite crear vectores y matrices de gran tamaño y multidimensionales. A su vez cuenta con una gran cantidad de funciones matemáticas de alto nivel.

3.1.4. Pandas

Pandas (McKinney) es una biblioteca de software libre escrita como extensión de Numpy para manipulación y análisis de datos. Uno de los puntos fuertes de esta biblioteca es la creación de DataFrames (tablas de datos), los cuales nos permiten un manejo de datos más rápido y cómodo.

3.1.5. SHAP

SHAP (Lundberg y Lee, 2017a) es una herramienta basada en la teoría de juego cuya finalidad es dar una explicación al resultado de cualquier modelo de machine learning. El valor que proporciona SHAP mide cuanto contribuye cada característica a la puntuación. Cabe destacar que estos valores miden tanto la contribución positiva como la negativa (Lundberg y Lee, 2017b).

Un ventaja que ofrece SHAP es su capacidad de calcular sus valores para cualquier modelo basado en árboles, mientras que con otras tecnologías era necesario emplear modelos de regresión lineal o regresión logística como modelos sustitutos.
De todos los métodos que contiene SHAP se eligieron los dos siguientes

3.1.5.1. Beeswarm

Este gráfico permite ver las principales características y cómo afectan a la salida. Para interpretar correctamente el gráfico hay que tener presente el color y el impacto de la salida.

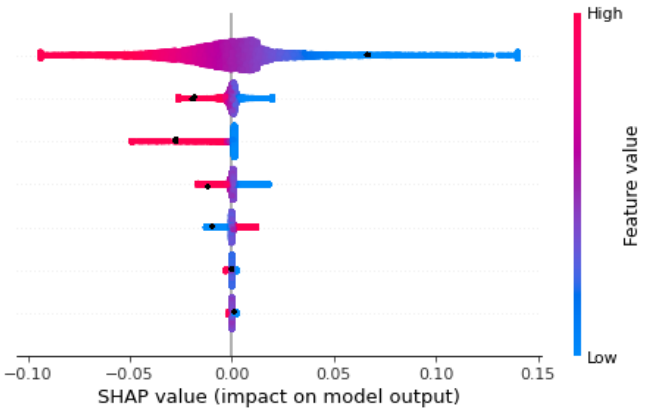

Figura 3.2: Ejemplo Beeswarm

El color identifica si el valor de la variable es alto o bajo, obteniendo una escala de colores desde el azul, siendo este el valor más bajo, hasta el rojo, que indica el valor más alto.
Por otro lado tendríamos el impacto, identificado por la posición de los puntos a lo largo de la gráfica, cuanto más a la izquierda más probable es que el resultado dé 0 y cuanto más

a la derecha más probable es que sea 1. Tomando como referencia la figura 3.2 podemos comprobar que la primera variable posee el mayor impacto, siendo aquella con mayor amplitud. Continuando con la primera variable se puede apreciar cómo a la izquierda se encuentran los tonos más intensos de rojos, esto se interpreta como que, a valores más altos de esta variable, tienden a darse casos cuyo resultado es 0. Por el contrario, los colores más azulados se encuentran al lado derecho, por lo que los valores más bajos de la variable implican un posible 1 en el resultado.

3.1.5.2. Waterfall

Con un objetivo similar al caso anterior, el diagrama en cascada busca mostrar la contribución de cada variable en la predicción, pero en este caso está diseñado para evaluar un individuo en concreto. Por cada característica se mostrará el peso que ha tenido en la clasificación y si ha sido de forma positiva o negativa. Este gráfico resulta muy interesante de cara a evaluar la clasificación de ciertos individuos.

En la figura 3.3 se aprecia cómo la edad y el síndrome metabólico representan un gran peso positivo para clasificar al usuario con sobrepeso, pero por el contrario variables como cal_ipaq o el ejercicio andando restan dicho peso. En la parte superior derecha de la gráfica, la función f(x) indica el resultado final obtenido. Si es superior a 0 se consideraría con sobrepeso y por el contrario si es inferior a 0 sin sobrepeso.

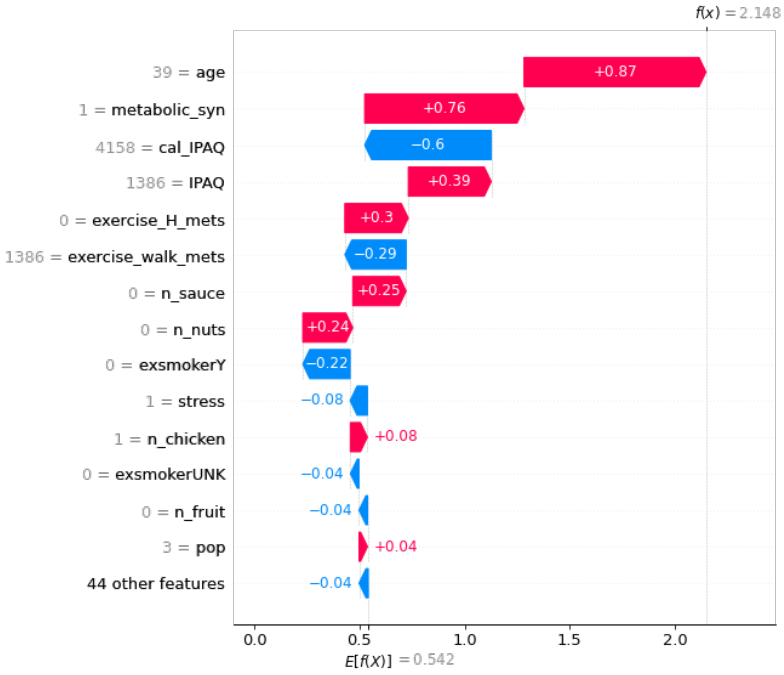

Figura 3.3: Ejemplo Waterfall

3.1.6. SALib

SALib (Herman et al.) es una librería de Python que permite realizar análisis de sensibilidad sobre un modelo. Se entiende por análisis de sensibilidad como el uso de técnicas para determinar cómo los valores de las variables independientes pueden afectar a una variable dependiente bajo un conjunto de supuestos. Para este proyecto se ha utilizado el método de Morris.

3.1.6.1. Método de Morris

El método de Morris (Iooss y Lemaître, 2015) permite clasificar las entradas en tres grupos: entradas con efectos insignificantes, entradas con efectos lineales sin interacciones y entradas con efectos no lineales y/o de interacción. Para realizar dicha tarea se realiza un número determinado de diseños " OAT ", *one-step-at-a-time*, en cada diseño OAT sólo un parámetro de entrada recibe un valor nuevo.
Este método devuelve cuatro variables por cada una de nuestras características:

- μ: Efecto elemental medio. Denotemos $E_j^{(i)}$ el efecto elemental de la jésima repetición, definida como:
$$E_j^{(i)} = \frac{f(X^{(i)} + \Delta e_j) - f(X^{(i)})}{\Delta}$$
Donde Δ es un múltiplo predeterminado de $\frac{1}{(n-1)}$ y e_j un vector de la base canónica.
- μ^*: Media del valor absoluto de los efectos elementales.
$$\mu_j^* = \frac{1}{r}\sum_{i=1}^{r} |E_j^{(i)}|$$
Siendo r el número de diseños OAT
- σ: Desviación estándar de los efectos elementales.
$$\sigma_j = \sqrt{\frac{1}{r}\sum_{i=1}^{r}\left(E_j^{(i)} - \frac{1}{r}\sum_{i=1}^{r} E_j^{(i)}\right)^2}$$
σ_j es una medida de los efectos no lineales y / o de interacción de la j-ésima entrada. Si σ_j tiene un valor bajo, los efectos elementales tienen variaciones bajas en el soporte de la entrada, por el contrario, si el valor es alto, menos probable es la hipótesis de la linealidad. Dicho esto, se considera que una variable con un σ_j grande, tiene efectos no lineales o está implicada en una interacción con al menos otra variable.

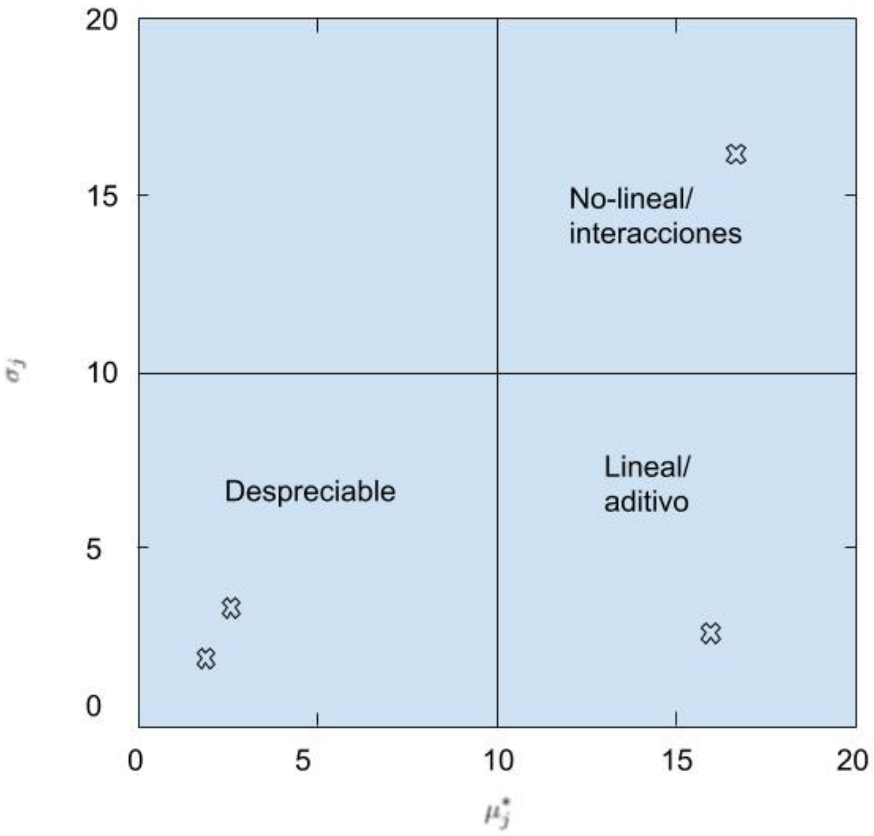

Figura 3.4: Diagrama interpretación mu_star y sigma

- μ^* conf: El intervalo de confianza de arranque

3.2. Aprendizaje automático

Stephen Marsland (Marsland, 2014) define el aprendizaje automático como *"el proceso en el cual la computadora trata de modificar o adaptar sus acciones, para que estas sean más precisas, donde la precisión se mide por que tan bien las acciones elegidas reflejan la correcta"*. Como bien se ha dicho anteriormente Sklearn nos ofrece una gran variedad de modelos tanto para predicción como para clasificación, en este caso tenemos un problema de clasificación ya que nuestro objetivo es clasificar si una persona va a tener o no sobrepeso. Para conseguir esto se deben seguir una serie de pasos de forma general:

- **1 - Definir el modelo**: Se define el modelo con sus hiperparámetros
- **2 - Dividir los datos en entrenamiento y test**: Normalmente el 75 % del conjunto de datos es destinado para entrenamiento del modelo y el 25 % restante para predicción

- **3 - Entrenar el modelo** : Se entrena el modelo con el conjunto de datos de entrenamiento, consiguiendo así que el modelo tenga una idea del problema.
- **4 - Test** : Una vez entrenado el modelo se procede a testear el aprendizaje del modelo con el conjunto de datos de test.
- **5 - Resultado** : Se comprueba la salida obtenida de la predicción con los valores reales del conjunto de test, obteniendo así un porcentaje de acierto, y comprobando la efectividad del modelo.

En este proyecto se han llegado a utilizar hasta 14 modelos, a continuación se habla de aquellos que han dado mejores resultados:

3.2.1. Gradient Boosting

El incremento de gradiente se usa para minimizar una función de pérdida. En cada ronda de entrenamiento se genera un modelo débil y se comparan sus predicciones con el resultado correcto esperado. La distancia entre la predicción y el valor correcto representa la tasa de error del modelo. A partir de estos errores se calcula el gradiente, el cual es la derivada parcial de la función de pérdida, por lo que describe la inclinación de la función de error (Glander, 2018).

Por lo tanto, se construye un modelo aditivo de manera progresiva hacia el escenario buscado, permitiendo la optimización de funciones de pérdida diferenciables arbitrarias. En cada etapa, un árbol de regresión se ajusta al gradiente negativo de la función de pérdida dada.

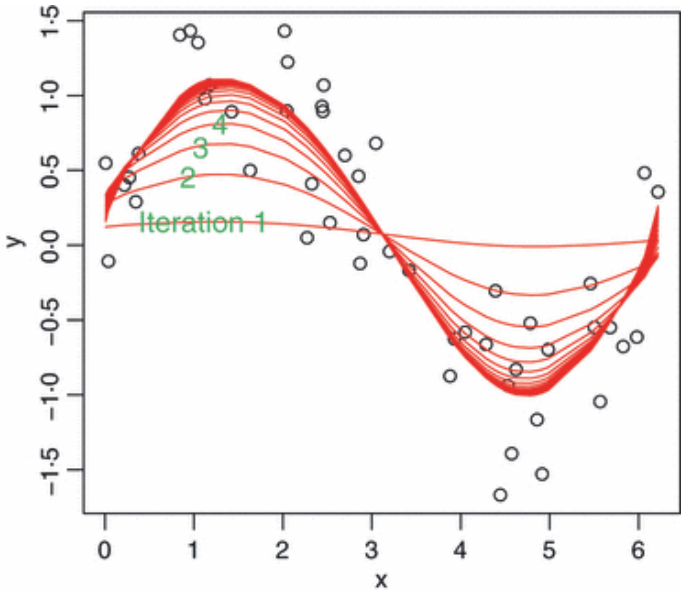

Figura 3.5: Evolución de Gradient Boosting entre iteraciones (Maloney et al., 2012)

3.2.2. Bagging Classifier

Bagging Classifier (Breiman, 1994) es un método que sirve para generar múltiples versiones de un predictor y usarlas para así obtener un predictor agregado. Para conseguirlo toma como base el entrenamiento con cada uno de los clasificadores con conjuntos de datos

aleatorios sobre el dataset original y posteriormente agrega dicha predicciones individuales (ya sea por votación o media) para generar una predicción final. Normalmente este tipo de modelo se suele utilizar para reducir la varianza de un modelo de caja negra (árbol de decisión), introduciendo la aleatorización en su procedimiento de construcción y luego haciendo un conjunto a partir de él.

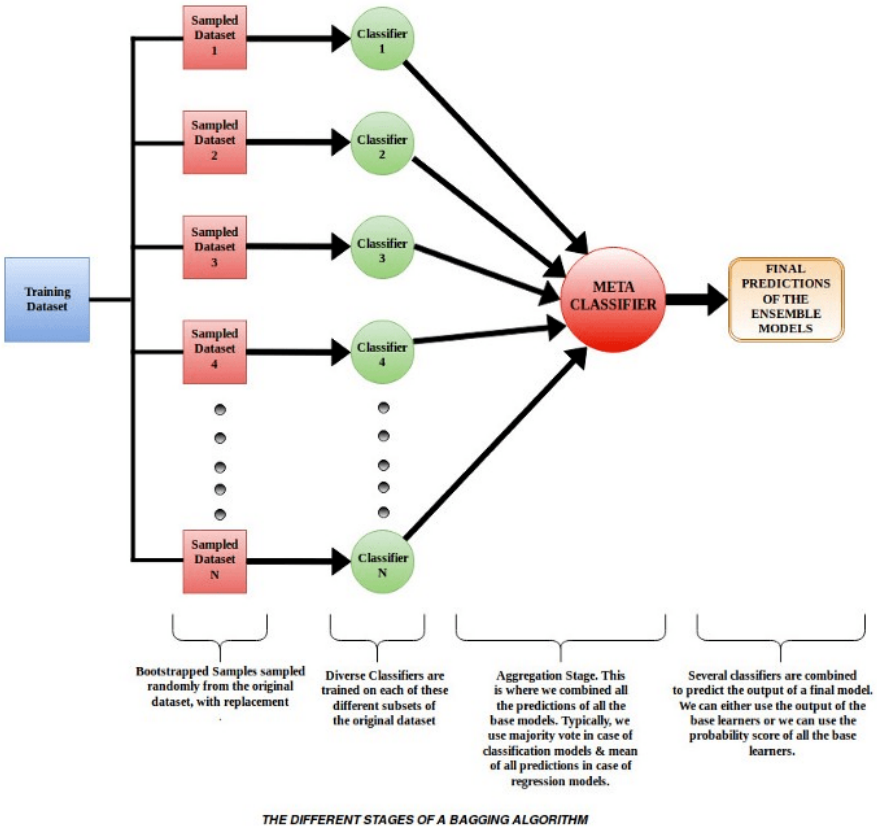

Figura 3.6: Estados Bagging Classifier (Paul, 2018)

3.2.3. Random Forest Regressor

Random Forest Regressor (Chakure, 2019) es un algoritmo de aprendizaje supervisado. Opera construyendo una multitud de árboles de decisión que se ejecutan en paralelo a la hora de realizar la etapa de entrenamiento y así generando las clases correspondientes, en el caso de clasificación, o la predicción media, en el caso de la regresión, para cada uno de los árboles individuales. Posteriormente combina el resultado de las múltiples predicciones.

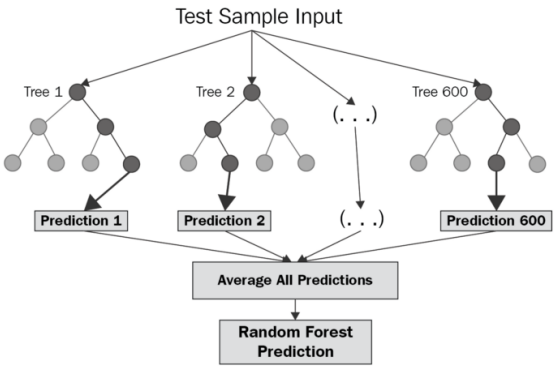

Figura 3.7: Proceso Random Forest Regressor (Chakure, 2019)

3.2.4. Logistic Regression

Logistic Regression (GeeksforGeeks, 2019) busca predecir la probabilidad de que una entrada de datos concreta pertenezca a un categoría. De forma similar a que la regresión lineal supone que los datos siguen una función lineal, la regresión logística modela los datos utilizando la función sigmoidea.

La regresión logística se puede convertir en una técnica de clasificación introduciendo un umbral de decisión. La configuración del valor umbral es uno de los aspectos mas importantes del modelo y depende del problema de clasificación en sí.

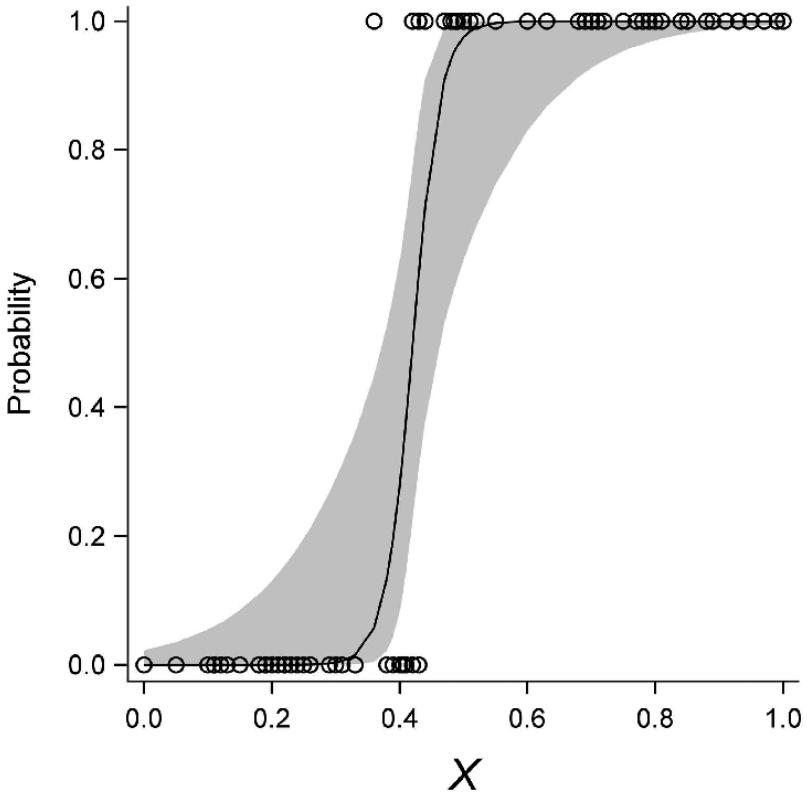

Figura 3.8: Logistic Regression (Wickham et al., 2018)

Capítulo 4

Clasificador Multinivel

Tras las diferentes etapas de filtrado de variables y preprocesamiento de datos (5.1), se empezaron a realizar pruebas con un amplio conjunto de modelos de aprendizaje automático. Para ello se utilizaron diferentes técnicas para evaluar los modelos como Cross-Validation y matrices de confusión para evaluar los resultados.

Rondando una tasa de acierto superior al 70 %, se observó que el número de falsos negativos, personas clasificadas como sanas pero que padecen sobrepeso, era bastante elevado y por ello se decidió establecer unas cotas a la hora de realizar la clasificación (5.5.1). Se fijó una cota del 60 %, esto quiere decir, que tanto para clasificar a un individuo como sano o con sobrepeso, al menos debe tener una probabilidad del 60 % para poder clasificarlo, si no, se descarta.

Aplicando estas cotas la tasa de acierto de los modelos ascendió y el número de falsos negativos disminuyó, pero ahora surgía otro problema, ¿qué hacer con los individuos descartados?. Así surgió la idea del clasificador Multi-nivel, para obtener el mayor número de individuos clasificados, con una tasa de acierto lo más alta posible y a su vez con un número de falsos negativos reducido. Para la evaluación de lo resultados obtenidos por los modelos se utilizó la matriz de confusión (5.3.2).

4.1. Estructura

El clasificador Multi-nivel tiene como base el fichero que contiene todos los datos filtrados hasta el momento, posteriormente se realiza una fase de selección de modelos, donde se les aplica CV a cada uno de ellos y se analizan los resultados. Una vez comparados, se escogen los 3 mejores modelos, ordenados de mejor a peor.
Como el objetivo es reducir el número de falsos negativos, se establecen una serie de cotas:

- **Cota clasificación sobrepeso**: 70 %
- **Cota clasificación no sobrepeso**: 80 %

Tras seleccionar los tres modelos y establecer las cotas, ya puede empezar a funcionar el clasificador Multi-nivel:

1. Se introducen los datos filtrados.

2. El primer clasificador procede a evaluar todos los elementos que ha recibido.

3. Se dividen los resultados en dos grupos, por un lado, aquellos casos en los que la probabilidad de clasificación no alcanza los límites establecidos y por el otro los que han podido ser clasificados.

4. Los casos descartados por el clasificador actual pasan al siguiente y se repite el proceso de evaluación y del apartado 3 hasta pasar por todos los clasificadores.

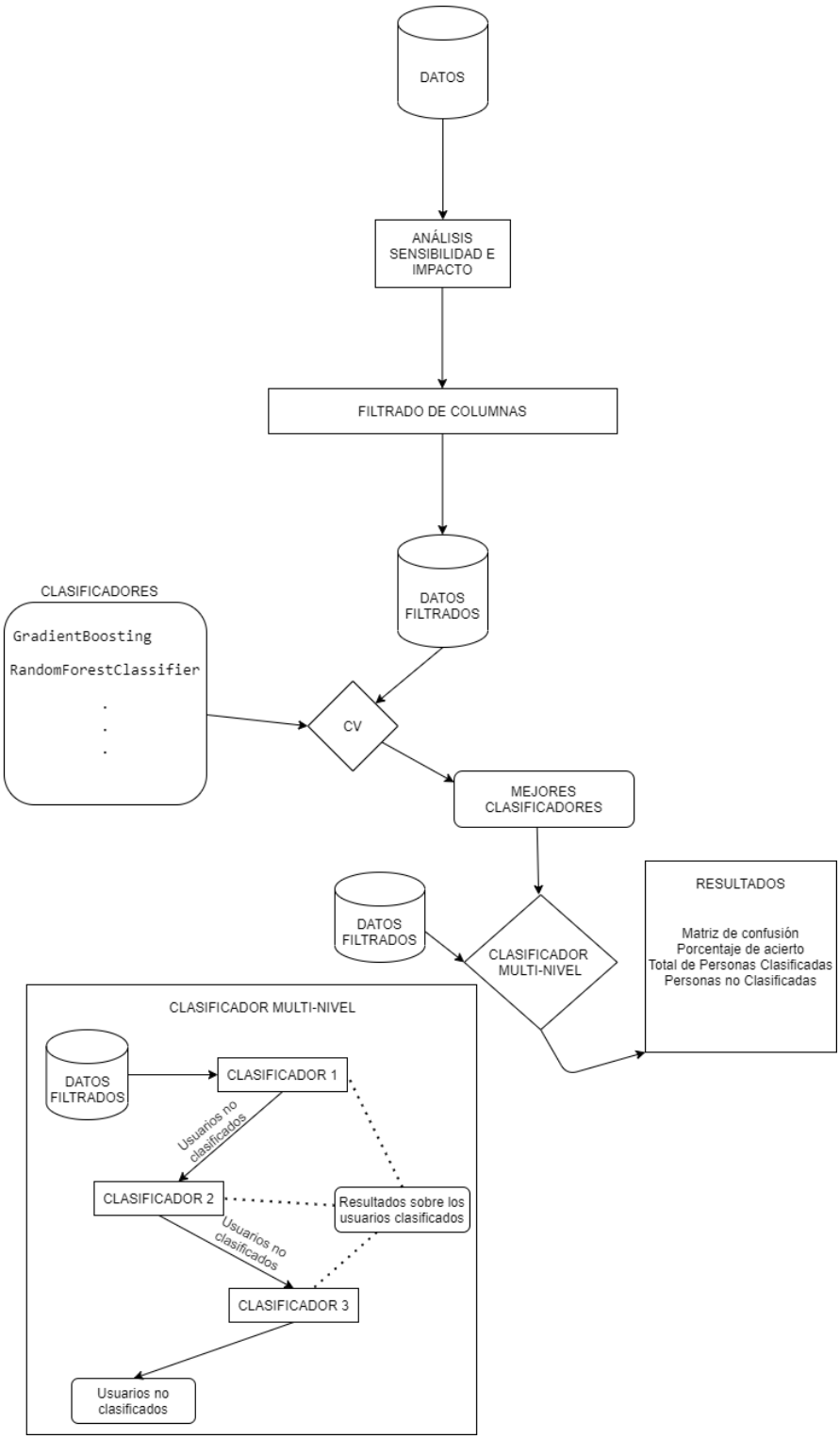

Figura 4.1: Diagrama del clasificador Multi-nivel

Este proceso nos dará finalmente dos resultados: un conjunto de individuos clasificados con su correspondiente tasa de acierto y otro conjunto de individuos los cuales no han conseguido ser clasificados.

4.2. Ejemplo de Funcionamiento

Tras tener la estructura montada se empezaron a realizar diferentes pruebas, inicialmente con las 37 variables obtenidas tras los diferentes filtrados explicados en el capítulo 5 y posteriormente, eliminando algunas variables de gran impacto para así evaluar el comportamiento del predictor.

Utilizando las 37 variables se obtuvo una tasa de acierto del 80 % clasificando un 65 % del total de individuos para test. Se aprecia cómo la precisión y el recall de ambas clases rondan el 80 %, esto indica que el número de falsos negativos y positivos es prácticamente bajo.

En uno de los resultados se aprecia a un hombre de 60 años con sobrepeso, se observa cómo por padecer apnea, ser exfumador, diabético, hombre y tener 60 años entre otros factores, se le clasifica como persona con sobrepeso/obesidad. En el caso de la apnea es común que una persona con sobrepeso padezca apnea, pero no que la apnea provoque sobrepeso. Los tres modelos utilizados para estas pruebas son: Gradient Boosting, RandomForest Classifier y Logistic Regression. A continuación, se muestran los resultados de los tres modelos, con su correspondiente tasa de acierto, matriz de confusión y número de personas clasificadas.

1. **Primer Modelo (Gradient Boosting)**

```
resultado eliminados --> 160
135
135
Correct classification rate: 0.8074074074074075
              precision    recall  f1-score   support

         0.0       0.85      0.72      0.78        65
         1.0       0.78      0.89      0.83        70

    accuracy                           0.81       135
   macro avg       0.81      0.80      0.80       135
weighted avg       0.81      0.81      0.81       135
```

	tn	fp	fn	tp
0	47	18	8	62

Figura 4.2: Resultado Primer Modelo del clasificador Multi-nivel

Este modelo es el encargado de realizar la clasificación más grande al ser el primer filtro de individuos. Posteriormente aquellos casos no clasificados, los cuales no están dentro de las cotas establecidas, se pasan al siguiente modelo. En este caso se obtiene una tasa de acierto del 80 % sobre un total de 135 individuos clasificados. Observando el número de falsos negativos, es realmente bajo, lo que conlleva que el recall de la clase sobrepeso/obesidad sea del 89 %, ya que apenas se escapan casos de individuos que padezcan sobrepeso/obesidad.

2. **Segundo Modelo (RandomForest Classifier)**

```
resultado eliminados --> 126
169
169
Correct classification rate: 0.7941176470588235
              precision    recall  f1-score   support

         0.0       0.82      0.86      0.84        21
         1.0       0.75      0.69      0.72        13

    accuracy                           0.79        34
   macro avg       0.78      0.77      0.78        34
weighted avg       0.79      0.79      0.79        34
```

	tn	fp	fn	tp
0	18	3	4	9

Figura 4.3: Resultado Segundo Modelo del clasificador Multi-nivel

En este caso, el número de individuos clasificados es menor, dado que la entrada de datos de dicho modelo son los individuos no clasificados del modelo anterior. Es por ello que el nivel de dificultad de clasificación según se avanza en el clasificador Multi-nivel se va incrementando.

Se han conseguido clasificar un total de 34 individuos de los 160 recibidos. Observando los resultados de la matriz de confusión son bastante buenos, ya que el número de falsos negativos sigue siendo bajo, con un recall de casi el 70 %.

Se observa en los resultados cómo, aparte de la edad, las variables de ejercicio físico o trabajo cogen una mayor importancia. Esto se debe al incremento de dificultad que presentan los individuos según se avanza en el clasificador Multi-nivel, provocando la aparición de ciertas variables dentro del modelo.

3. **Tercer Modelo (Logistic Regression)**

```
resultado eliminados --> 102
193
193
Correct classification rate: 0.8333333333333334
              precision    recall  f1-score   support

         0.0       0.86      0.67      0.75         9
         1.0       0.82      0.93      0.87        15

    accuracy                           0.83        24
   macro avg       0.84      0.80      0.81        24
weighted avg       0.84      0.83      0.83        24
```

	tn	fp	fn	tp
0	6	3	1	14

Figura 4.4: Resultado Tercer Modelo del clasificador Multi-nivel

Finalmente, este tercer modelo es el clasificador que menor número de individuos consigue clasificar, ya que los factores de los mismos no expresan con claridad al modelo ningún tipo de clasificación dentro de las cotas establecidas.

En este caso se han conseguido clasificar otros 24 individuos con una tasa de acierto del 83 %, y un recall excelente del 93 %, ya que tan solo se ha obtenido un falso negativo. Esto quiere decir que, de los 15 casos de sobrepeso, solo 1 se ha clasificado como sin sobrepeso. Recordemos, que se busca reducir el número de individuos con sobrepeso clasificados sin sobrepeso, ya que desde el punto de vista médico es más grave no diagnosticar el sobrepeso/obesidad a una persona que sí lo padece, que a una que no.

Llama la atención que existe un gran número de variables que apenas tienen peso en los modelos, esto se debe a que apenas existen casos con estas variables. No se han llegado a eliminar, ya que según se avanza en el clasificador Multi-nivel empiezan a tomar algún tipo de peso. Se espera que al incluir la genética de cada individuo y más casos con estas variables empiecen a tomar una mayor relevancia.

Más adelante, se realizan pruebas con un total de 22 variables escogidas tras una selección recursiva de las mismas en la sección 5.8.2. También se han realizado pruebas eliminando las variables edad y centro, dichas pruebas se encuentran en la sección 5.7.1

4.2.1. Clasificados

A partir de los resultados obtenidos se generó un CSV con los usuarios clasificados dentro de las cotas establecidas. El objetivo es apreciar los factores de cada individuo con su clasificación real y su predicción, así como el modelo que lo ha clasificado.

AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU
metabolic_syn	apnea	asthma	COPD	ADH	IPAQ	clasificador	real_values	prediction	Acierto
0	0	1	0	1	5172	GradientBoosting	1	0	FALSO
0	0	0	0	1	20650	GradientBoosting	0	0	VERDADERO
0	0	0	0	1	5502	GradientBoosting	0	0	VERDADERO
0	0	0	0	0	0	GradientBoosting	0	0	VERDADERO
0	0	0	0	0	438	GradientBoosting	1	1	VERDADERO
0	0	0	0	0	1386	GradientBoosting	1	1	VERDADERO
0	0	1	0	1	5652	GradientBoosting	1	1	VERDADERO
1	0	0	0	1	2799	GradientBoosting	1	1	VERDADERO
0	0	0	1	0	1386	GradientBoosting	1	1	VERDADERO
0	0	0	0	0	0	GradientBoosting	1	0	FALSO
0	0	0	0	0	924	GradientBoosting	0	0	VERDADERO
0	0	1	0	1	522	GradientBoosting	0	0	VERDADERO
0	0	0	0	1	3057	GradientBoosting	0	0	VERDADERO

Figura 4.5: Resultados obtenidos por el modelo 1

metabolic_sy	apnea	asthma	COPD	ADH	IPAQ	clasificador	real_valu	predictior	Acierto
0	0	0	0	0	165	RandomForestClassifier	1	0	FALSO
0	0	0	0	1	3093	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	1	1200	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	0	0	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	0	1398	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	1	318	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	0	1182	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	1	4830	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	1	11520	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	1	2706	RandomForestClassifier	1	0	FALSO
0	1	0	0	1	5946	RandomForestClassifier	0	1	FALSO
0	0	0	0	1	2556	RandomForestClassifier	0	0	VERDADERO
0	0	1	0	1	3245	RandomForestClassifier	0	0	VERDADERO
0	0	1	0	0	9492	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	0	1680	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	1	1920	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	0	2186	RandomForestClassifier	0	0	VERDADERO
0	0	0	0	0	360	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	0	975	RandomForestClassifier	1	1	VERDADERO
0	0	0	0	1	3804	RandomForestClassifier	0	0	VERDADERO

Figura 4.6: Resultados obtenidos por el modelo 2

metabolic_sy	apnea	asthma	COPD	ADH	IPAQ	clasificador	real_valu	predictior	Acierto
0	0	0	0	0	0	LogisticRegression	0	1	FALSO
0	0	0	0	1	720	LogisticRegression	1	1	VERDADERO
0	0	1	0	1	11982	LogisticRegression	0	0	VERDADERO
0	0	0	0	1	2068.5	LogisticRegression	0	0	VERDADERO
0	0	1	0	1	1080	LogisticRegression	1	0	FALSO
0	0	1	0	1	0	LogisticRegression	1	1	VERDADERO
0	0	0	0	1	1386	LogisticRegression	1	1	VERDADERO
0	0	0	0	1	1386	LogisticRegression	1	1	VERDADERO
0	0	0	0	0	2373	LogisticRegression	1	1	VERDADERO
0	0	0	0	0	346.5	LogisticRegression	0	1	FALSO
0	0	0	0	1	400	LogisticRegression	0	0	VERDADERO
0	0	0	0	0	5013	LogisticRegression	0	0	VERDADERO
0	0	0	1	1	5310	LogisticRegression	1	1	VERDADERO
0	0	0	0	1	693	LogisticRegression	1	1	VERDADERO
0	0	0	0	1	1059	LogisticRegression	0	0	VERDADERO
0	0	0	0	1	8718	LogisticRegression	0	0	VERDADERO
0	0	0	0	0	6186	LogisticRegression	1	1	VERDADERO
0	0	0	0	1	3066	LogisticRegression	1	1	VERDADERO

Figura 4.7: Resultados obtenidos por el modelo 3

Sobre todo estos archivos van destinados al equipo médico para que puedan identificar casos interesantes.

4.2.2. No Clasificados

Al igual que para el conjunto de clasificados, se creó un CSV para aquellos individuos que no cumplían las cotas de clasificación en ningún modelo del clasificador Multi-nivel. En el fichero se puede encontrar toda la información relacionada con estos individuos, para su posterior estudio por parte del equipo médico y obtener conclusiones.

sex	age	label	pop	edu	earning	job	stress	sleep.8	spirit	spiritWEEK	wine_beer	beerWEEK	wineWEEK	whiteWEEK	pinkWEEK	smoke	nsmoke	pipe	cigar	exsmokerY	exsmokerUN	canc
1	61	1	3	3	2	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0.03278689	1	
0	52	1	1	3	2	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	1	
1	60	0	3	3	2	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0.0125	1	
0	22	0	2	3	2	9	0	0	1	3	1	7	0	0	0	0	0	0	0	0	1	
0	52	0	3	2	2	10	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	21	0	3	2	1	9	0	0	1	3	1	10	0	0	0	1	10	0	0	0	0	
0	21	0	1	3	1	9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	61	1	3	2	2	11	0	0	0	0	1	0	3	0	0	0	0	0	0	0.44262295	1	
1	21	1	3	2	1	9	1	0	1	4	1	1	0	0	0	1	2	0	0	0	0	
1	64	0	1	3	2	1	0	0	0	0	1	1	1	0	0	0	0	0	0	0.390625	1	
1	64	0	3	3	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0.03776042	0	
1	32	0	2	2	0	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1	47	0	3	3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	33	0	3	3	2	9	0	0	0	0	1	3	0	0	0	1	7	0	0	0	0	
1	39	0	3	3	1	13	1	0	0	0	1	2	0	0	0	0	0	0	0	0	1	
1	51	1	3	2	1	11	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figura 4.8: CSV individuos no clasificados

Capítulo 5

Resultados Experimentales

El desarrollo del proyecto ha seguido un sistema de entregas periódicas de resultados, siguiendo esta metodología se han relazado un total de seis lotes de experimentos y una etapa de preprocesamiento, análisis de sensibilidad y generación de árboles:

5.1. Preprocesamiento

Para ser capaces de extraer información relevante de los datos, es necesario una etapa de preprocesado. En esta etapa el objetivo es filtrar, en la medida de lo posible, aquellos elementos que tengan potencial y darles la forma adecuada para poder trabajar con los modelos.

5.2. Análisis de sensibilidad con SAlib

Siguiendo los consejos de mi director, se decidió realizar un estudio de sensibilidad de las variables dentro de un modelo. En este caso se utilizó el método de Morris, el cual nos permite conocer los efectos insignificantes, lineales sin interacciones o no lineales y/o de interacción de las variables.
Lo primero que se debe definir es el problema en cuestión:

```
problem = {
'num_vars': len (columns),
'names': columns,
'bounds': bounds
}
```

Figura 5.1: Definición problema Morris

Para ello se deben especificar el número total de variables a usar, su nombre y los valores límites de cada una de ellas. Una vez definido el problema utilizaremos Morris para generar un conjunto de datos aleatorio a partir de las especificaciones del problema. Este DataSet se utilizará para realizar una predicción con un modelo de aprendizaje automático

y así obtener el resultado de las predicciones. Una vez obtenido las predicciones, se realiza el análisis con Morris, pasando los siguientes parámetros a la función:

- Especificación del problema
- Conjunto de Datos generados para la entrada del modelo
- Salida obtenida tras la predicción del modelo
- Nivel de confianza (0.95)

En este caso se ha utilizado el modelo de Gradient Boosting como referencia y observando los resultados de Mu y Sigma, se aprecia cómo destaca claramente la edad con un impacto alto en relación con otras variables. Esto ya indica que la edad será un factor clave como veremos posteriormente.

5.3. Primer lote de experimentos

5.3.1. Inicio clasificación

Tras la etapa de preprocesado el conjunto de datos está compuesto por un total de 58 Variables y 1179 casos. El objetivo inicial del proyecto era identificar individuos en riesgo de desarrollar obesidad, IMC igual o superior a 30, por lo que se planteó como un problema de clasificación, por un lado, tendríamos aquellos usuarios cuyo IMC estaría por encima de 30, clasificados como 1 y por otro los que estarían por debajo de dicho umbral, clasificados como 0.

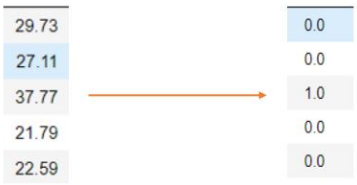

Figura 5.2: Clases Clasificación

Una vez estudiados los datos con detenimiento, el número de individuos con obesidad era muy pequeño como para trabajar con él, por este motivo se decidió añadir los usuarios con sobrepeso ,IMC entre 25 y 30, al grupo de 1s. Una vez establecida la columna a predecir (IMC), dividimos el conjunto en entrenamiento y test usando una división 65 % 35 % respectivamente.

Para probar el conjunto de datos se generó una plantilla con un total de 14 algoritmos de clasificación.
Tras esta etapa de pruebas, 6 de los modelos obtuvieron un resultado superior al 70 %, estos fueron:
Bagging classifier , ExtraTrees classifier , Random Forest classifier , Random Forest Regressor, Logistic Regression y Gradient Boosting.

5.3.2. Interpretación resultados - Matriz de confusión

Para entender mejor los resultados se generó una matriz de confusión para cada uno de ellos, ya que permite extraer una serie de métricas para medir el rendimiento del modelo.

	Positive Prediction	***Negative Prediction***
Positive Class	True Positive (TP)	False Negative (FN)
Negative Class	False Positive (FP)	True Negative (TN)

Tabla 5.1: Matriz de Confusión

- **Tasa de error (Misclassification Rate)**: porcentaje de datos clasificados incorrectamente.

$$TasaDeError = \frac{FP + FN}{Total} \tag{5.1}$$

- **Exactitud (Accuracy)**: porcentaje de datos clasificados correctamente.

$$Exactitud = \frac{TP + TN}{Total} \tag{5.2}$$

- **Sensibilidad (Recall)**: es la Tasa de verdaderos positivos, es decir, el porcentaje de datos que logra clasificar de la clase positiva y respectivamente con la clase negativa.

$$Sensibilidad = \frac{TP}{TP + FN} \tag{5.3}$$

- **Precisión (Precision)**: es el porcentaje de predicciones correctas que ha hecho.

$$Precision = \frac{TP}{TP + FP} \tag{5.4}$$

Analizando la matriz de confusión a partir de los resultados de estos seis modelos, todos los modelos presentaban el mismo problema, la cantidad de falsos negativos siempre fue mayor que la de falsos positivos, es decir, existían un gran número de casos de usuarios con sobrepeso que los modelos no detectaban.

	tn	fp	fn	tp
0	168	47	64	134

Figura 5.3: Matriz de confusión para Gradient Boosting

	tn	fp	fn	tp
0	166	42	70	135

Figura 5.4: Matriz de confusión para Random Forest Classifier

Una vez conocidas las variables que tenían más peso (5.2) y tras comprobar los casos en los que los modelos fallaban más, se observó que había dos grupos bien diferenciados donde se producían los fallos, por un lado los usuarios jóvenes que padecían sobrepeso y por otro los usuarios de mayor edad que no padecían de esta afección. Esto apuntaba a cierto sesgo en los datos.

5.4. Segundo lote de experimentos

Analizando una vez más los datos se observó que el número de usuarios jóvenes con sobrepeso/obesidad era muy bajo, siendo en los usuarios de mayor edad donde se encontraban los casos de obesidad.

- Menores de 35 años: 140 casos de sobrepeso
- Mayores de 34 años: 427 casos de sobrepeso

5.4.1. Análisis de impacto de variables con SHAP

En esta etapa el punto central de la investigación fueron las variables en las que se basó el predictor para su toma de decisiones. Para esta tarea se utilizó la herramienta SHAP, que ofrece una serie de valores para cuantificar la relevancia de cada variable.

Las primeras pruebas se realizaron empleando el método de Beeswarm, cuyo funcionamiento ya se ha explicado en el capítulo 3. Como podemos observar en la siguientes imágenes, el sexo ocupa una de las posiciones más altas, al igual que la edad, por lo que se decidió utilizar esta variable para realizar la división con el fin de profundizar en los factores que presentan mayor relevancia en cada uno de los casos.

Con los datos recopilados a lo largo de esta etapa se puede afirmar que los modelos pueden llegar a considerar que un usuario no va a padecer obesidad observando en gran medida el sexo y la edad del individuo.

Hasta ahora todas las pruebas realizadas se aplicaban sobre la población general, sin tener en cuenta diferenciaciones por edad y sexo. Para comprobar qué modelos obtenían mejores resultados se aplicaba CV, *cross-validation*, obteniendo así una media de accuracy. Pero cuidado, no siempre obtener un buen accuracy implica que se estén realizando correctamente las clasificaciones, tal y como veremos en los siguientes apartados. Se observa que algunos modelos como el Gradient Boosting destacan sobre otros, es por ello, que se realizó una matriz de confusión para observar mejor los resultados de CV.

```
Accuracy: 0.70 (+/- 0.13) [Logistic Regression]
Accuracy: 0.72 (+/- 0.10) [BagginC ]
Accuracy: 0.71 (+/- 0.10) [GradientBoosting]
Accuracy: 0.69 (+/- 0.08) [RandomForestClassifier]
Accuracy: 0.67 (+/- 0.12) [GaussianNB]
Accuracy: 0.64 (+/- 0.05) [DecisionTreeClassifier]
Accuracy: 0.67 (+/- 0.10) [BernoulliNB]
Accuracy: 0.63 (+/- 0.05) [AdaBoostClassifier]
Accuracy: 0.70 (+/- 0.10) [ExtraTreesClassifier]
Accuracy: 0.72 (+/- 0.11) [Ensemble]
```

Figura 5.5: CV con Población General

```
Accuracy: 0.71 (+/- 0.10) [GradientBoosting]
              precision    recall  f1-score   support

         0.0       0.71      0.75      0.73       612
         1.0       0.71      0.68      0.69       567

   micro avg       0.71      0.71      0.71      1179
   macro avg       0.71      0.71      0.71      1179
weighted avg       0.71      0.71      0.71      1179
```

	tn	fp	fn	tp
0	457	155	183	384

Figura 5.6: CV Gradient Boosting - Población General

Observando la imagen (Figura: 5.6), se aprecia que tanto la precisión y el recall están muy equilibrados entre sí y no existen grandes diferencias, confirmando así el buen resultado obtenido en el Accuracy tras CV.

5.4.2. Pruebas con Población Hombres

Dado que la edad y el sexo eran dos variables con un gran impacto dentro de los modelos, se decidió separar la población en hombres y mujeres y posteriormente por edad.

5.4.2.1. Sin tener en cuenta la edad

En este caso se trata la población de los hombres, donde se encuentran un total de 569 individuos, de los cuales 235 no tienen sobrepeso y 334 sí.

```
Accuracy: 0.71 (+/- 0.17) [Logistic Regression]
Accuracy: 0.70 (+/- 0.15) [BagginC ]
Accuracy: 0.68 (+/- 0.13) [GradientBoosting]
Accuracy: 0.66 (+/- 0.13) [RandomForestClassifier]
Accuracy: 0.69 (+/- 0.15) [GaussianNB]
Accuracy: 0.63 (+/- 0.12) [DecisionTreeClassifier]
Accuracy: 0.68 (+/- 0.15) [BernoulliNB]
Accuracy: 0.62 (+/- 0.10) [AdaBoostClassifier]
Accuracy: 0.68 (+/- 0.14) [ExtraTreesClassifier]
Accuracy: 0.70 (+/- 0.15) [Ensemble]
```

Figura 5.7: CV Población Hombres

```
Accuracy: 0.71 (+/- 0.17) [Logistic Regression]
              precision    recall  f1-score   support

         0.0       0.63      0.73      0.68       235
         1.0       0.79      0.70      0.74       334

   micro avg       0.71      0.71      0.71       569
   macro avg       0.71      0.72      0.71       569
weighted avg       0.72      0.71      0.72       569
```

	tn	fp	fn	tp
0	172	63	100	234

Figura 5.8: CV Logistic Regression - Población Hombres

En este modelo (Figura: 5.8) se puede observar cómo la precisión para detectar casos de individuos que no padecen sobrepeso es un 16 % inferior. Esto se debe principalmente a que casi el 60 % de los hombres tratados padecen sobrepeso/obesidad y por tanto el modelo tiene menos casos para identificar los factores de los hombres sin sobrepeso.

5.4.2.2. Mayores de 50 años

En este caso se trata la población de los hombres mayores de 50, que abarca el gran número de casos de sobrepeso/obesidad. Existen 196 casos de sobrepeso/obesidad frente a 47 sin sobrepeso, esto indica que el 80 % de los hombres mayores de 50 años del DataSet padecen sobrepeso/obesidad.

```
Accuracy: 0.75 (+/- 0.09) [Logistic Regression]
Accuracy: 0.78 (+/- 0.04) [BagginC ]
Accuracy: 0.75 (+/- 0.07) [GradientBoosting]
Accuracy: 0.75 (+/- 0.05) [RandomForestClassifier]
Accuracy: 0.74 (+/- 0.05) [GaussianNB]
Accuracy: 0.69 (+/- 0.08) [DecisionTreeClassifier]
Accuracy: 0.73 (+/- 0.06) [BernoulliNB]
Accuracy: 0.70 (+/- 0.07) [AdaBoostClassifier]
Accuracy: 0.78 (+/- 0.03) [ExtraTreesClassifier]
Accuracy: 0.78 (+/- 0.04) [Ensemble]
```

Figura 5.9: CV - Población Hombres mayores de 50 años

```
Accuracy: 0.78 (+/- 0.04) [BagginC ]
              precision    recall  f1-score   support

         0.0       0.12      0.02      0.04        47
         1.0       0.80      0.96      0.88       196

   micro avg       0.78      0.78      0.78       243
   macro avg       0.46      0.49      0.46       243
weighted avg       0.67      0.78      0.71       243
```

	tn	fp	fn	tp
0	1	46	7	189

Figura 5.10: CV Bagging Classifier - Población Hombres mayores de 50 años

En este caso se observa claramente cómo se tiene una tasa de exactitud de casi el 80 % con el modelo Bagging Classifier, pero si observamos la matriz de confusión se aprecia como la precisión y recall para clasificar hombres mayores de 50 años sin sobrepeso/obesidad es pésima. Al estar tan descompensados los casos de hombres con sobrepeso/obesidad frente a los que no, esto provoca que el modelo clasifique a casi todos los individuos como sobrepeso y así obtiene una buena tasa de acierto global, pero pésima dentro de la clase de individuos sin sobrepeso, ya que ante la duda los clasifica con sobrepeso.

5.4.2.3. Menores de 30 años

En este intervalo se encuentran un total de 231 hombres, de los cuales 150 no padecen sobrepeso y 75 sí. Una vez más, los resultados vuelven a mostrar lo sesgados que están los datos.

```
Accuracy: 0.64 (+/- 0.11) [Logistic Regression]
Accuracy: 0.69 (+/- 0.10) [BagginC ]
Accuracy: 0.66 (+/- 0.10) [GradientBoosting]
Accuracy: 0.61 (+/- 0.08) [RandomForestClassifier]
Accuracy: 0.56 (+/- 0.12) [GaussianNB]
Accuracy: 0.61 (+/- 0.10) [DecisionTreeClassifier]
Accuracy: 0.64 (+/- 0.09) [BernoulliNB]
Accuracy: 0.59 (+/- 0.11) [AdaBoostClassifier]
Accuracy: 0.62 (+/- 0.09) [ExtraTreesClassifier]
Accuracy: 0.69 (+/- 0.09) [Ensemble]
```

Figura 5.11: CV - Población Hombres menores de 30 años

```
Accuracy: 0.66 (+/- 0.10) [GradientBoosting]
              precision    recall  f1-score   support

         0.0       0.73      0.79      0.76       156
         1.0       0.47      0.39      0.42        75

   micro avg       0.66      0.66      0.66       231
   macro avg       0.60      0.59      0.59       231
weighted avg       0.64      0.66      0.65       231
```

	tn	fp	fn	tp
0	123	33	46	29

Figura 5.12: CV Gradient Boosting - Población Hombres menores de 30 años

Al igual que pasó con los hombres mayores de 50 años, en este caso (Figura: 5.12) ocurre al revés, ya que el número de casos sin sobrepeso es el doble que con sobrepeso. Es por ello que se aprecia cómo la precisión y el recall de la categoría de hombres con sobrepeso tiene valores muy poco óptimos.

5.4.2.4. Entre 30-50 años

Llegados a este punto analizamos el último rango de edad, en el que se encuentra el menor número de hombres, tan solo 95, de los cuales 63 tienen sobrepeso/obesidad y 32 no.

```
Accuracy: 0.62 (+/- 0.10) [Logistic Regression]
Accuracy: 0.63 (+/- 0.08) [BagginC ]
Accuracy: 0.72 (+/- 0.12) [GradientBoosting]
Accuracy: 0.71 (+/- 0.07) [RandomForestClassifier]
Accuracy: 0.59 (+/- 0.14) [GaussianNB]
Accuracy: 0.68 (+/- 0.14) [DecisionTreeClassifier]
Accuracy: 0.60 (+/- 0.16) [BernoulliNB]
Accuracy: 0.70 (+/- 0.13) [AdaBoostClassifier]
Accuracy: 0.67 (+/- 0.15) [ExtraTreesClassifier]
Accuracy: 0.66 (+/- 0.09) [Ensemble]
```

Figura 5.13: CV - Población Hombres entre 30 y 50 años

```
Accuracy: 0.72 (+/- 0.12) [GradientBoosting]
              precision    recall  f1-score   support

         0.0       0.60      0.47      0.53        32
         1.0       0.76      0.84      0.80        63

   micro avg       0.72      0.72      0.72        95
   macro avg       0.68      0.66      0.66        95
weighted avg       0.70      0.72      0.71        95
```

	tn	fp	fn	tp
0	15	17	10	53

Figura 5.14: CV Gradient Boosting - Población Hombres entre 30 y 50 años

Se puede observar (Figura: 5.14) cómo con la escasez de datos, la precisión y recall de ambas clases fluctúan, siendo mejor la clasificación de hombres con sobrepeso en este caso, pero sin llegar a ser fiable debido al número de casos.

5.4.3. Pruebas con Población Mujeres

Se realizan las mismas pruebas que con la población de hombres, mostrando una imagen de los resultados obtenidos por los modelos con CV y posteriormente analizando uno de ellos.

5.4.3.1. Sin tener en cuenta la edad

El número de casos de mujeres es similar al de hombres, teniendo un total de 610 mujeres frente a 569 hombres, de las cuales 377 mujeres no padecen sobrepeso/obesidad y 233 sí. En cuanto a los hombres sucede al revés, se dan más casos de hombres con sobrepeso/obesidad.

```
Accuracy: 0.70 (+/- 0.12) [Logistic Regression]
Accuracy: 0.68 (+/- 0.11) [BagginC ]
Accuracy: 0.68 (+/- 0.08) [GradientBoosting]
Accuracy: 0.67 (+/- 0.09) [RandomForestClassifier]
Accuracy: 0.69 (+/- 0.12) [GaussianNB]
Accuracy: 0.59 (+/- 0.06) [DecisionTreeClassifier]
Accuracy: 0.68 (+/- 0.10) [BernoulliNB]
Accuracy: 0.61 (+/- 0.05) [AdaBoostClassifier]
Accuracy: 0.69 (+/- 0.09) [ExtraTreesClassifier]
Accuracy: 0.69 (+/- 0.09) [Ensemble]
```

Figura 5.15: CV - Población Mujeres

```
Accuracy: 0.70 (+/- 0.12) [Logistic Regression]
              precision    recall  f1-score   support

         0.0       0.79      0.69      0.74       377
         1.0       0.58      0.70      0.64       233

   micro avg       0.70      0.70      0.70       610
   macro avg       0.69      0.70      0.69       610
weighted avg       0.71      0.70      0.70       610
```

	tn	fp	fn	tp
0	260	117	69	164

Figura 5.16: CV Logistic Regression - Población Mujeres

En este caso (Figura: 5.16) se obtiene una exactitud del 70 % y observando ambas clases en la matriz de confusión, se puede apreciar que la precisión relacionada con las mujeres que padecen sobrepeso es inferior. Esto se debe a que el número de mujeres sin sobrepeso es mayor y por lo tanto el modelo puede identificarlas de forma más fácil.

5.4.3.2. Mayores de 50 años

Si subimos el rango de edad, mujeres mayores de 50 años, existen un total de 227 mujeres, de las cuales 136 padecen sobrepeso. Esto indica que del conjunto de datos existente, el 60 % de las mujeres mayores de 50 años padece sobrepeso/obesidad. Al igual que en los hombres, en las mujeres con más edad se dan más casos de sobrepeso/obesidad,

```
Accuracy: 0.68 (+/- 0.07) [Logistic Regression]
Accuracy: 0.58 (+/- 0.13) [BagginC ]
Accuracy: 0.56 (+/- 0.12) [GradientBoosting]
Accuracy: 0.55 (+/- 0.07) [RandomForestClassifier]
Accuracy: 0.52 (+/- 0.09) [GaussianNB]
Accuracy: 0.56 (+/- 0.18) [DecisionTreeClassifier]
Accuracy: 0.57 (+/- 0.12) [BernoulliNB]
Accuracy: 0.55 (+/- 0.17) [AdaBoostClassifier]
Accuracy: 0.60 (+/- 0.15) [ExtraTreesClassifier]
Accuracy: 0.58 (+/- 0.12) [Ensemble]
```

Figura 5.17: CV - Población Mujeres mayores de 50 años

```
Accuracy: 0.68 (+/- 0.07) [Logistic Regression]
              precision    recall  f1-score   support

         0.0       0.58      0.68      0.63        91
         1.0       0.76      0.68      0.72       136

   micro avg       0.68      0.68      0.68       227
   macro avg       0.67      0.68      0.67       227
weighted avg       0.69      0.68      0.68       227
```

	tn	fp	fn	tp
0	62	29	44	92

Figura 5.18: CV Logistic Regression - Población Mujeres mayores de 50 años

.

En este caso (Figura: 5.18) se puede observar cómo el modelo alcanza una tasa de acierto del 68 %. Comparando ambas clases se observa que se obtienen mejores resultados con la clase de mujeres de sobrepeso. Ocurre al igual que en los hombres, ya que el número de mujeres mayores de 50 años con sobrepeso es superior que el de mujeres sin sobrepeso. Esto provoca que el modelo conozca mejor dichos factores para su correcta clasificación.

5.4.3.3. Menores de 30 años

Al seleccionar al grupo de mujeres con edad inferior a 30 años, observamos una gran diferencia entre el número de casos de mujeres con sobrepeso/obesidad, poco más de 53, frente a las 219 que no lo padecen.

```
Accuracy: 0.71 (+/- 0.11) [Logistic Regression]
Accuracy: 0.76 (+/- 0.04) [BagginC ]
Accuracy: 0.77 (+/- 0.04) [GradientBoosting]
Accuracy: 0.79 (+/- 0.03) [RandomForestClassifier]
Accuracy: 0.69 (+/- 0.13) [GaussianNB]
Accuracy: 0.65 (+/- 0.09) [DecisionTreeClassifier]
Accuracy: 0.77 (+/- 0.04) [BernoulliNB]
Accuracy: 0.64 (+/- 0.08) [AdaBoostClassifier]
Accuracy: 0.79 (+/- 0.03) [ExtraTreesClassifier]
Accuracy: 0.78 (+/- 0.04) [Ensemble]
```

Figura 5.19: CV - Población Mujeres menores de 30 años

```
Accuracy: 0.79 (+/- 0.03) [ExtraTreesClassifier]
              precision    recall  f1-score   support

         0.0       0.80      0.98      0.88       219
         1.0       0.00      0.00      0.00        53

   micro avg       0.79      0.79      0.79       272
   macro avg       0.40      0.49      0.44       272
weighted avg       0.65      0.79      0.71       272
```

	tn	fp	fn	tp
0	214	5	53	0

Figura 5.20: CV ExtraTreesClassifier - Población Mujeres menores de 30 años

Observando la figura 5.20, podríamos pensar que los resultados han sido buenos, si solo nos fijamos en el valor de *accuracy*, sin embargo, al prestar atención a la matriz de confusión, podemos comprobar que el modelo ha clasificado a prácticamente todo el conjunto de muestra como ceros, a excepción de 5 casos, en los que ha fallado también.

5.4.3.4. Entre 30-50 años

Para finalizar, tenemos el conjunto formado por mujeres de entre 30 a 50 años. Como pasaba al analizar los hombres en este mismo tramo de edad, el número de casos es muy pequeño, contando con un total de 111 casos y entre ellos 44 de sobrepeso/obesidad.

```
Accuracy: 0.76 (+/- 0.17) [Logistic Regression]
Accuracy: 0.65 (+/- 0.13) [BagginC ]
Accuracy: 0.69 (+/- 0.14) [GradientBoosting]
Accuracy: 0.71 (+/- 0.12) [RandomForestClassifier]
Accuracy: 0.66 (+/- 0.18) [GaussianNB]
Accuracy: 0.52 (+/- 0.11) [DecisionTreeClassifier]
Accuracy: 0.68 (+/- 0.13) [BernoulliNB]
Accuracy: 0.53 (+/- 0.12) [AdaBoostClassifier]
Accuracy: 0.68 (+/- 0.18) [ExtraTreesClassifier]
Accuracy: 0.70 (+/- 0.13) [Ensemble]
```

Figura 5.21: CV - Población Mujeres entre 30 y 50 años

Accuracy: 0.76 (+/- 0.17) [Logistic Regression]

	precision	recall	f1-score	support
0.0	0.84	0.76	0.80	67
1.0	0.68	0.77	0.72	44
micro avg	0.77	0.77	0.77	111
macro avg	0.76	0.77	0.76	111
weighted avg	0.77	0.77	0.77	111

	tn	fp	fn	tp
0	51	16	10	34

Figura 5.22: CV Logistic Regression - Población Mujeres entre 30 y 50 años

En la figura 5.22, la exactitud del modelo es del 76 %. Debido a la falta de casos de sobrepeso/obesidad, el modelo no ha sido capaz de aprender a clasificar la clase de los 1s (individuos con sobrepeso) por completo, pero los resultados son aceptables.

5.5. Tercer lote de experimentos

5.5.1. Creación de cotas

Llegado este punto, se decidió estudiar la seguridad con la que el modelo tomaba sus decisiones tratando de descartar aquellos casos en los que la respuesta era prácticamente aleatoria. Se estableció que los resultados que rondasen el intervalo de predicción ente el 40 % y 60 % se debían a datos insuficientes y se realizaron nuevas pruebas.

Al eliminar aquellos casos en los que el modelo no tenía una predicción fiable, la tasa de éxito alcanzó valores cercanos al 80 %. En la siguiente imagen se puede ver un ejemplo de Gradient Boosting en el que se evalúan un total de 233 individuos y se descartan 62 de un total de 295.

```
resultado eliminados --> 62
Correct classification rate: 0.8540772532188842
              precision    recall  f1-score   support

         0.0       0.84      0.89      0.87       122
         1.0       0.87      0.81      0.84       111

   micro avg       0.85      0.85      0.85       233
   macro avg       0.86      0.85      0.85       233
weighted avg       0.86      0.85      0.85       233
```

	tn	fp	fn	tp
0	109	13	21	90

Figura 5.23: Resultados Gradient Boosting con cotas 40-60

Como hemos comprobado, los resultados son bastante buenos, pero el número de falsos negativos es algo elevado en comparación con el de falsos positivos y es por ello que posteriormente se incrementarán estas cotas para reducir aquellas clasificaciones más aleatorias o indecisas por parte del modelo. Recordemos que un falso negativo penaliza mucho más que un falso positivo, ya que implica que un paciente con riesgo de padecer sobrepeso / obesidad no va a ser atendido correctamente. Este problema será tratado más adelante.

5.6. Cuarto lote de experimentos

5.6.1. Clasificador Multi-nivel

Posteriormente, con base en los resultados anteriores, se optó por construir un clasificador Multi-nivel tomando los 3 mejores modelos obtenidos mediante Cross-Validation. En estos modelos se establecieron límites para asegurarse de que las predicciones no eran aleatorias, siendo 70 % para afirmar si una persona iba a tener sobrepeso/obesidad y 80 % para lo contrario. Con esta diferencia de porcentajes se busca reducir el número de falsos negativos
A continuación se enumeran los distintos pasos del clasificador Multi-nivel:

1. Se introducen los datos filtrados.

2. El primer clasificador procede a evaluar todos los elementos que ha recibido.

3. Se dividen los resultados en dos grupos, por un lado, aquellos casos en los que la probabilidad de clasificación no alcanza los límites establecidos y por el otro los que han podido ser clasificados.

4. Los casos no clasificados del clasificador actual pasan al siguiente y se repite el proceso de evaluación y del apartado 3 hasta pasar por todos los clasificadores.

Tras realizar diferentes pruebas con el clasificador Multi-nivel, en todos los casos fue capaz de predecir aproximadamente dos tercios de los casos con una tasa de clasificación en torno al 80 % y manteniendo el número de falsos negativos entre 8 y 18.

5.7. Quinto lote de experimentos

5.7.1. Pruebas eliminando edad y centro para el clasificador Multi-nivel

En esta ocasión se decidió probar las capacidades del clasificador Multi-nivel eliminando alguna de sus Variables clave, para ello se tomó la decisión de eliminar el campo con mayor importancia en ese momento, la edad.

Los resultados fueron mejores de los esperado, alcanzando una tasa de clasificación del 79.47 %, con un total de 8 falsos negativos, y logrando clasificar más del 50 % de los casos (151 de 295).

Gracias a la eliminación de la edad los demás campos cobraron mayor relevancia, sorprendentemente el campo que ostentaba la posición más alta era el centro.
Algunos de los centros donde se realizaron las encuestas eran universidades, aunque también se recogieron datos de gente adulta en estos centros, la mayoría procedía de gente joven, pudiendo considerarse que introducían cierto nivel de sesgo, por este motivo se decidió eliminar este campo.

5.8. Sexto lote de experimentos

Tras estudiar los resultados del apartado anterior, se procede a eliminar el campo "centro" de la lista de variables utilizadas, quedando un total de 37 Variables y manteniendo el número de casos.

5.8.1. Selección de variables

Con el objetivo de mejorar los resultados del clasificador Multi-nivel, se tomó la decisión de profundizar en las variables que se estaban utilizando en ese momento. En primer lugar, se realizó un estudio empleando la función SelectKbest sobre el modelo Gradient boosting, estableciendo un número de Variables para el filtrado. Este método presentaba un inconveniente, no se podía afirmar cuál era la cantidad óptima de variables para los modelos con los que se trabajaba.

Para solventar este problema, se optó por usar una función para la selección de características de forma recursiva, dicha función tiene como base un modelo de predicción y, mediante validación cruzada, estima cuáles son las variables más importantes junto con el número óptimo a emplear.

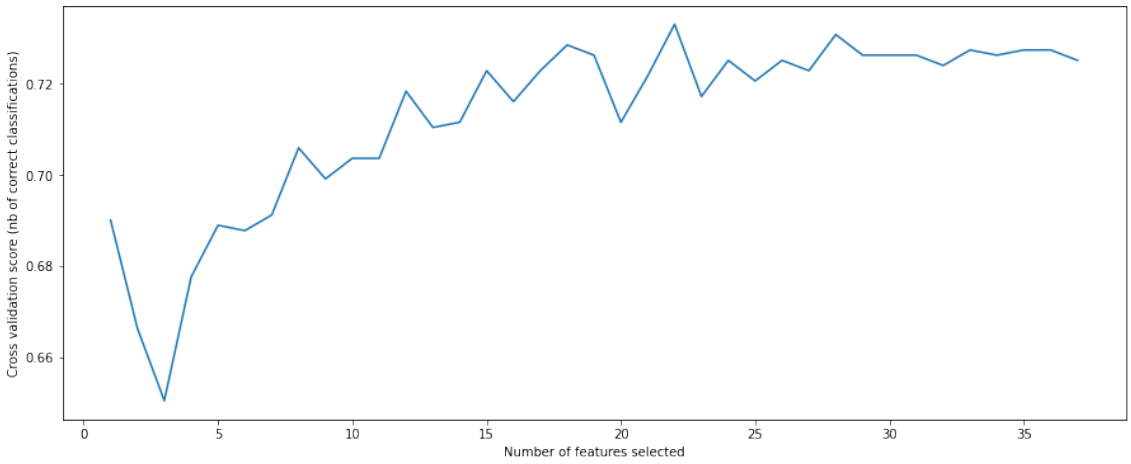

Figura 5.24: Gráfica de la selección de características para Gradient Boosting

Como se aprecia en la imagen, el número óptimo de variables es 22, alcanzando el 73 %. La diferencia entre el conjunto total y el uso de estas 22 es despreciable, tratándose de una diferencia de puntuaciones por debajo del 1 %. Es destacable el hecho de que con 1 o 2 variables se alcanza casi el 70 %, debido muy probablemente a las variables sexo y edad. A continuación, se ofrece el listado con la clasificación de variables.

Tras una conversación con el personal médico se optó por mantener el total de variables puesto que la diferencia era mínima y podrían ser útiles en algunos casos.

5.8.2. Pruebas clasificador Multi-nivel con 22 variables

A continuación, se muestran los resultados obtenidos por el clasificador Multi-nivel usando las 22 variables obtenidas en el paso anterior.

```
Personas no clasificadas : 106
Personas clasificadas : 189
Correct classification rate: 0.8148148148148148
              precision    recall  f1-score   support

         0.0       0.83      0.79      0.81        95
         1.0       0.80      0.84      0.82        94

    accuracy                           0.81       189
   macro avg       0.82      0.81      0.81       189
weighted avg       0.82      0.81      0.81       189
```

	tn	fp	fn	tp
0	75	20	15	79

Figura 5.25: Resultado clasificador Multi-nivel 22 variables

Como se observa en la figura 5.25 la tasa de acierto obtenida por el clasificador Multi-nivel apenas es un 1 % mejor que los resultados obtenidos en el capítulo 4. Comparando ambos resultados, y como se lleva indicando a lo largo del proyecto, se decide mantener las 37 variables debido a la escasa diferencia entre ellos.

5.9. Generación de árboles

Aun rondando ya una tasa de acierto del 80 % y con un número de falsos negativos bastante reducido, los médicos solicitaron la generación de árboles de decisión para observar su comportamiento con las 37 variables finales. La generación de estos árboles era muy interesante para ellos ya que podían observar gráficamente el comportamiento del modelo. A continuación, se describe cómo interpretar un árbol.

Los árboles están formados por nodos, cada nodo presenta 4 apartados:

- Variable y condición: Al inicio del nodo aparece una variable con una condición. Si se cumple la condición se debe seguir por la rama de la izquierda, si no se cumple por la de la derecha.
- MSE: Error cuadrático medio
- Samples: El número de individuos que cumplen las condiciones anteriores hasta el nodo actual.
- Value : Cuanto más cercano al 1, más probabilidades de padecer sobrepeso, cuanto más cercano al 0, menos probabilidades de padecer sobrepeso.

Desde la raíz toda rama conduce hasta un nodo hoja, donde el modelo estima la clase correspondiente a esa rama, ya sea 1 o 0, es decir con o sin sobrepeso. Para ello se utilizó el modelo Random Forest Regressor, obteniendo los siguientes resultados:

Correct classification rate: 0.7186440677966102

	precision	recall	f1-score	support
0.0	0.72	0.73	0.72	148
1.0	0.72	0.71	0.71	147
micro avg	0.72	0.72	0.72	295
macro avg	0.72	0.72	0.72	295
weighted avg	0.72	0.72	0.72	295

	tn	fp	fn	tp
0	108	40	43	104

Figura 5.26: Resultado Random Forest Regressor para la Población General

Como se observa en la matriz de confusión, los resultados obtenidos rondan en torno al 71 % de acierto, pero a su vez el número de falsos negativos se ha incrementado considerablemente respecto al clasificador Multi-nivel.

En algunos de los árboles generados se observa cómo la edad forma parte de la raíz, y por lo tanto es el primer filtro. Posteriormente en los siguientes niveles se tienen en cuenta factores como la edad, trabajo, apnea y así sucesivamente hasta llegar a un nodo hoja donde se clasifica al individuo.

Utilizando la población de mujeres, y como ocurría en las pruebas de CV, se obtiene una tasa de acierto baja del 65 %. Esto viene provocado por la pésima precisión y recall obtenidos a la hora de clasificar mujeres con sobrepeso/obesidad. Se observa que la muestra está descompensada ya que el número de mujeres sin sobrepeso es el doble.

Capítulo 6

Conclusiones y Trabajo Futuro

6.1. Conclusiones

El sobrepeso y la obesidad son problemas que año tras año afectan a más personas. En una de las últimas encuestas realizadas por el Instituto Nacional de Estadística se afirma que, en los últimos 30 años, la prevalencia de obesidad en España se ha multiplicado por 2.4, pasando de 7.4 % en 1987 al 17.4 % en 2017. En este estudio de encuestas (Ministerio de Sanidad, 2017) se afirma que la obesidad afecta más frecuentemente a hombres (18.2 %) que a mujeres (16.7 %), esta diferencia se hace mucho más notable al comparar los números relacionados con el sobrepeso, llegando al 44.3 % para los hombres frente al 30 % de las mujeres. Otra estadística que sigue avanzando, aunque en menor medida, es la de la obesidad infantil, alcanzando a más del 10 % entre los 2 y 17 años. Prevenir futuros casos de sobrepeso/obesidad parece ser la opción más correcta tras observar estos datos.

Con este objetivo en mente, se han implementado un conjunto de algoritmos de aprendizaje automático para clasificación de personas en riesgo de padecer sobrepeso. Las conclusiones obtenidas de este trabajo podemos resumirlas en los siguientes puntos:

- En líneas generales podemos afirmar que los resultados obtenidos en este trabajo han sido francamente buenos, pese a la dificultad del problema.
- Las características con más peso en los modelos son la edad y el sexo del individuo, lo que concuerda con los documentos citados.
- Si seguimos avanzando en el documento (Ministerio de Sanidad, 2017) descubrimos que muchos de los elementos que se mencionan como relevantes para desarrollar sobrepeso/obesidad, están reflejados en las variables con las que se alimentan los modelos.
- Partiendo de un conjunto inicial de modelos con unos resultados rondando el 70 % de tasa de acierto, se ha llegado a obtener un clasificador Multi-nivel que ronda el 80 %. Estos resultados son muy buenos y es por ello que se espera que añadiendo la genética de cada individuo y más casos se obtengan mejores resultados.
- A lo largo del proyecto han surgido ciertos casos prácticos bastante curiosos, como por ejemplo, que a partir de la edad y el sexo los modelos obtengan casi una tasa de acierto del 70 %, esto se debe a que los datos están sesgados.

- Existen un gran número de casos de sobrepeso/obesidad entre gente de mediana y avanzada edad, por el contrario, la gran mayoría de casos sin sobrepeso/obesidad se encuentran en la gente joven. Igual ocurre con el sexo, ya que hay más casos de sobrepeso/obesidad en los hombres. Esto provoca que ciertos casos de sobrepeso de gente joven sean considerados como gente sin sobrepeso. Es por ello que se han solicitado más casos de sobrepeso en gente joven y viceversa.

- Se tienen en cuenta un total de 37 variables, es cierto que algunas tienen un peso insignificante dentro de los modelos, pero se mantienen debido a que pueden ser de importancia tras añadir la genética de los individuos. Algunas de estas variables son la insuficiencia cardiaca o infartos de miocardio, las cuales apenas tienen peso en los modelos por la falta de casos dentro de los datos.

- Si observamos los resultados obtenidos en el apartado 5.7.1, en uno de los casos se decide eliminar la edad para evaluar la respuesta del clasificador Multi-nivel. Los resultados obtenidos dan un gran peso a variables como la educación o el trabajo, aparte del sexo o ejercicio físico. Esto se debe a que el nivel de estudios y el salario de un individuo afectan de forma directa a sus hábitos alimenticios y diarios.

Todos estos estudios sobre el impacto de las variables en los modelos fueron facilitados al equipo médico, el cual a día de hoy está muy contento con el trabajo realizado. Estos a su vez, solicitaron la generación de árboles de decisión. Dichos árboles tienen una gran importancia para los médicos, ya que los han trabajado con anterioridad y son muy visibles e interpretables para ellos,

Por último, dada la complejidad del problema los resultados obtenidos son francamente buenos y junto al TFM realizado por mi compañero de Departamento, Daniel Parra Rodríguez, sobre selección de variables con computación evolutiva, se espera obtener una selección óptima de las mismas.

6.2. Trabajo Futuro

Como se ha visto a lo largo del proyecto, las distintas variables sobre las que se sustenta el clasificador Multi-nivel han sido fundamentales en el desarrollo del mismo. Por ello es necesario conocer mejor dichas variables a través de un estudio más exhaustivo y continuar las reuniones con el personal médico, siendo estas últimas uno de los elementos que más ayudó en la evolución del proyecto.

Otro elemento presente a lo largo de este trabajo es la predisposición de los modelos por dar más peso a la edad y al sexo, si bien son dos características muy importantes a la hora de evaluar el problema, es posible que la falta en la variedad de casos aumente esta brecha respecto al resto de variables, por lo que sería muy interesante contar con un volumen de datos mayor.

Se pretende unir este TFM junto el TFM de mi compañero de departamento Daniel Parra Rodríguez, con el fin de mejorar los resultados obtenidos.

En el momento de la redacción de este documento queda pendiente de recibir la información genética de los usuarios, y a su vez comprobar cuánto mejorará el modelo contando

con esta nueva información.

Este Trabajo ha sido financiado por la Comunidad de Madrid y Fondo Social Europeo a través del Proyecto GENOBIA-CM con referencia S2017/BMD-3773.

Chapter 7

Introduction

The WHO considers overweight to be a global epidemic that constitutes a public health problem, mainly in developed countries, although it is also beginning to appear in developing countries. That is why public authorities and institutions are increasingly aware that this is a serious problem and that it must be reduced and prevented as far as possible to continue growing its incidence in the population considered healthy. There are a large number of factors and variables involved in the development of overweight and obesity and it is not an easy task to predict individually the risk of developing these pathologies and their associated comorbidities (Organization, 2000).

According to a study conducted by the National Institute of Statistics (Ministerio de Sanidad, 2017) it is stated that, in the last 30 years, the prevalence of obesity in Spain has multiplied by 2.4, going from 7.4% in 1987 to 17.4% in 2017. Analyzing the data, it can be seen that the cases of obesity and overweight are higher in the case of men than women. With regard to child obesity, there are already 10% of children between 2 and 17 years old who suffer from it, so it is necessary to prevent future cases of overweight or obesity.

The great development of the artificial intelligence and automatic learning opens a door to realize in a suitable way the task of investigating in the causes of the appearance and development of the overweight and the obesity. Artificial intelligence is understood as the development of systems equipped with intellectual processes that are characteristic of human beings. Among these processes we find reasoning, generalization, improvement through past experiences and the discovery of meanings (Copeland, 2020) . The advances in this sector are continuous and accurate but there is still a long way to go before reaching the marks established by man. This end-of-master's work is part of the GENOBIA-CM project, whose main objective is to design predictive and classification algorithms to identify people at risk of developing overweight/obesity and its associated pathologies.

Throughout this work, tests are carried out with a large number of supervised classification algorithms. The objective of the classification is to generate a model to predict a class given some input values. These input values are characterized objects that belong to different classes. In this case a binary classification is being used, since there are two class types: 0 without overweight and 1 with overweight. Some of the classification algorithms used are Logistic Regression, Gradient Boosting, decision trees or Random Forest. The incorporation of these and other learning techniques in the medical field is increasing.

For the implementation of the project, Python has been used, a programming language interpreted, not compiled, known for its flexibility, power and simplicity. Among its advantages, it also has the ability to support different paradigms such as object-oriented, imperative and structured programming, to give an example. Another advantage is that Python allows us to use specialized libraries.

7.1. Motivation

The main motivation for this research is to help public health systems and institutions to reduce the incidence of overweight in the population of the Community of Madrid. Because it is a public health problem, it is of vital importance to correctly classify those users who will suffer from these conditions, so this process requires special attention to the safety with which the model gives its predictions. Artificial intelligence in general and machine learning in particular are the tools we have at our disposal and which allow us to be confident in the success of our task.

7.2. Objetives

The main objectives of the work are:

1. To design a classification system of subjects at risk of suffering from overweight of the highest precision.
2. To provide the health professionals involved in Genobia with classification systems based on decision trees that allow them to investigate its operation without the need to have a deep knowledge of computer techniques.

.

7.3. Workplan

In order to achieve the aforementioned objectives, the following work plan will be addressed, which constitutes the following secondary objectives in an orderly manner:

- Perform proper data pre-processing and curing
- Analyze dependencies between input variables
- Study the operation of classical classification algorithms and their performance.
- To carry out a sensitivity analysis and impact of the variables on the models
- To obtain information about the result of the classification of the applied models that is useful for health professionals.
- To design a classification system that collects all the conclusions obtained from the previous steps and combines in the most appropriate way the most used algorithms in the current literature.
- To design classification systems based on decision trees.

- To particularize the classification systems for different segments of the population, separated by sex and age.

The development of the project has followed a system of regular deliveries of results. Following this methodology, a total of six batches of experiments have been carried out. Initially a pre-processing stage has been carried out, as well as a variable analysis and finally the generation of decision trees. Some additional details of this work plan are mentioned below.

- **Data preprocessing**
  Before performing the first tests it was essential to treat the data received, modifying values and eliminating unnecessary columns.
- **Sensitivity analysis with the SAlib tool**
  Measuring the impact of variables on each other and on possible changes
- **First batch of experiments**
  First tests with models and their interpretation.
- **Second batch of experiments**
  Study of the variables used by the SHAP tool and tests of the models with cross-validation with respect to different sets of populations based on sex and age..
- **Third batch of experiments**
  Use of heights to measure the security of the models in your decision making.
- **Fourth batch of experiments**
  Development of the multi-level classifier and decision making regarding the variables in use.
- **Fifth batch of experiments**
  Tests on the Multi-Level Classifier.
- **Sixth batch of experiments**
  New selection of variables.
- **Generation trees**
  The Random Forest Regressor model is used for the generation of trees and later their study by the medical team.

7.4. Memory organization

The rest of the memory is organized as follows:

- Chapter 1: Introduction, objectives and work plan.
- Chapter 2: Brief documentary analysis of articles related to the topic to be discussed.
- Chapter 3: Development of the technologies used throughout the project.
- Chapter 4: Description and operation of the multilevel classifier.
- Chapter 5: Documentation of the work done and the results obtained.
- Chapter 6: General conclusions and future work.

Chapter 8

Conclusions and Future Work

8.1. Conclusions

Overweight and obesity are problems that year after year affect more people. One of the latest surveys conducted by the National Institute of Statistics states that in the last 30 years, the prevalence of obesity in Spain has multiplied by 2.4, going from 7.4% in 1987 to 17.4% in 2017. In this study of surveys (Ministerio de Sanidad, 2017) it is stated that obesity affects men more frequently (18.2%) than women (16.7%), this difference becomes much more noticeable when comparing the numbers related to overweight, reaching 44.3% for men against 30% for women. Another statistic that continues to advance, although to a lesser extent is that of child obesity, reaching more than 10% between the ages of 2 and 17. To prevent future cases of overweight/obesity seems to be the most correct option after observing these data..

With this objective in mind, a set of automatic learning algorithms have been implemented to classify people at risk of being overweight. The conclusions obtained from this work can be summarized in the following points:

- In general we can say that the results obtained in this work have been frankly good, despite the difficulty of the problem.
- The characteristics with the most weight in the models are the age and sex of the individual, which is consistent with the documents cited.
- If we continue to advance in the document we discover that many of the elements mentioned as relevant to developing overweight/obesity are reflected in the variables that feed the models. Some of them are food (ADH and those related to beverages), working life (the work field) and how they spend their free time (IPAQ from the point of view of the exercise they do in those time periods).
- It should be noted that the variables mentioned are some of the most important in the models, without taking into account age and sex.
- Starting from an initial set of models with results around 70% of success rate, we have obtained a multi-level classifier that is around 80%. These results are very good and that is why it is expected that by adding the genetics of each individual and more cases, better results are obtained.

- Throughout the project some rather curious case studies have emerged, for example, that from age and sex the models obtain almost a 70% success rate, this is due to the fact that the data are biased.

- There are a large number of cases of overweight/obesity among middle-aged and older people; in contrast, the vast majority of cases without overweight/obesity are found in young people. The same occurs with sex, since there are more cases of overweight/obesity in men. This causes certain cases of overweight young people to be considered as not overweight. This is why more cases of overweight young people have been requested and vice versa.

- A total of 37 variables are taken into account, it is true, that some have an insignificant weight within the models, but they are maintained because they can be of importance after adding the genetics of the individuals. Some of these variables are heart failure or myocardial infarction, which barely have weight in the models due to the lack of cases within the data.

- If we look at the results obtained in the section 5.7.1, in one of the cases it is decided to eliminate the age to evaluate the predictor response. The results obtained give great weight to variables such as education or work, apart from sex or physical exercise. This is due to the fact that the level of education and salary of an individual directly affects his or her eating and daily habits.

All these studies on the impact of the variables in the models were provided to the medical team, which today is very happy with the work done. They, in turn, requested the generation of decision trees. These trees are of great importance to physicians, since they have worked on them before and are very visible and interpretable to them.

Finally, given the complexity of the problem, the results obtained are frankly good, and together with the TFM carried out by my colleague in the Department, Daniel Parra Rodríguez, on variable selection with evolutionary computing, it is hoped that an optimal selection of the same will be obtained.

8.2. Future Work

As has been seen throughout the project, the different variables on which the Multilevel classifier is based have been fundamental in its development. Therefore, it is necessary to better understand these variables through a more exhaustive study and continue the meetings with the medical staff, the latter being one of the elements that most helped in the evolution of the project.

Another element present throughout this work is the predisposition of the models to give more weight to age and sex, although these are two very important characteristics when evaluating the problem, it is possible that the lack in the variety of cases increases this gap with respect to the rest of the variables, so it would be very interesting to have a greater volume of data.

It is intended to unite this TFM with the TFM of my department colleague Daniel Parra Rodríguez, in order to improve the results obtained.

At the time of writing this document is pending to receive genetic information from users, so it remains to check how much will improve the model with this new information.

This work has been financed by the Community of Madrid and the European Social Fund through the Project GENOBIA-CM with reference S2017/BMD-3773.

Bibliografía

Bilgin, A., Ellson, J., Gansner, E., Hu, Y., North, S. y contribuciones. Graphviz. https://graphviz.org/, ???? Accedido 04-01-2021.

Breiman, L. Bagging predictors. vol. 421, páginas –20, 1994.

Chakure, A. Random forest regression. https://medium.com/swlh/random-forest-and-its-implementation-71824ced454f, 2019. Accedido 04-01-2021.

Copeland, B. Artificial intelligence. *Encyclopædia Britannica*, (2), páginas –24, 2020. Accedido 15-01-2021.

Cournapeau, D. Scikit-learn. https://scikit-learn.org/stable, 2012. Accedido 04-01-2021.

GeeksforGeeks. Understanding logistic regression. https://www.geeksforgeeks.org/understanding-logistic-regression/, 2019. Accedido 04-01-2021.

Glander, S. Machine learning basics - gradient boosting xgboost. https://www.shirin-glander.de/2018/11/ml_basics_gbm/, 2018. Accedido 04-01-2021.

Herman, J., Usher, W., Mutel, C., Trindade, B., Hadka, D., Woodruff, M., Rios, F., Hyams, D. y xantares. Salib - sensitivity analysis library in python. https://salib.readthedocs.io/en/latest/, ???? Accedido 04-01-2021.

De la Hoz Manotas, A., De la Hoz Correa, E., Mendoza, F., Morales, R. y Sanchez, B. Obesity level estimation software based on decision trees. *Journal of Computer Science*, vol. 15, páginas –10, 2019.

Iooss, B. y Lemaître, P. *A Review on Global Sensitivity Analysis Methods*, páginas 101–122. Springer US, Boston, MA, 2015. ISBN 978-1-4899-7547-8.

Khalaf, M., Hussain, A. J., Keight, R., Al-Jumeily, D., Fergus, P., Keenan, R. y Tso, P. Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models. *Neurocomputing*, vol. 228, páginas 154 – 164, 2017. ISSN 0925-2312. Advanced Intelligent Computing: Theory and Applications.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B. y Sánchez, C. I. A survey on deep learning in medical image analysis. *Medical Image Analysis*, vol. 42, páginas 60 – 88, 2017. ISSN 1361-8415.

Lundberg, S. M. y Lee, S.-I. Shapley additive explanations. https://shap.readthedocs.io/en/latest/, 2017a. Accedido 04-01-2021.

Lundberg, S. M. y Lee, S.-I. A unified approach to interpreting model predictions. En *Advances in Neural Information Processing Systems* (editado por I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan y R. Garnett), vol. 30, páginas 4765–4774. Curran Associates, Inc., 2017b.

Maloney, K. O., Schmid, M. y Weller, D. E. Applying additive modelling and gradient boosting to assess the effects of watershed and reach characteristics on riverine assemblages. *Methods in Ecology and Evolution*, vol. 3(1), páginas 116–128, 2012.

Marsland, S. *Machine Learning: An Algorithmic Perspective, Second Edition*. Chapman and amp; Hall/CRC, 2nd edición, 2014. ISBN 1466583282.

McKinney, W. Pandas. https://pandas.pydata.org/, ???? Accedido 04-01-2021.

Muhamad Adnan, M. H. B., Husain, W. y Abdul Rashid, N. A hybrid approach using naïve bayes and genetic algorithm for childhood obesity prediction. En *2012 International Conference on Computer Information Science (ICCIS)*, vol. 1, páginas 281–285. 2012.

Oliphant, T. Numpy. https://numpy.org/, 1995. Accedido 04-01-2021.

Organization, W. H. *Obesity: preventing and managing the global epidemic*. 2000.

Paul, S. Ensemble learning — bagging, boosting, stacking and cascading classifiers in machine learning using sklearn and mlextend libraries. https://medium.com/@saugata.paul1010/ensemble-learning-bagging-boosting-stacking-and-cascading-classifiers-in-machine-learning-9c66cb271674, 2018. Accedido 15-01-2021.

Ministerio de Sanidad, C. y. B. S. Encuesta nacional de salud. españa 2017. https://www.mscbs.gob.es/estadEstudios/estadisticas/encuestaNacional/encuestaNac2017/ENSE2017_notatecnica.pdf, 2017. Accedido 15-01-2021.

Singh, B. y Tawfik, H. Machine learning approach for the early prediction of the risk of overweight and obesity in young people. En *Computational Science – ICCS 2020* (editado por V. V. Krzhizhanovskaya, G. Závodszky, M. H. Lees, J. J. Dongarra, P. M. A. Sloot, S. Brissos y J. Teixeira), páginas 523–535. Springer International Publishing, Cham, 2020. ISBN 978-3-030-50423-6.

Wickham, J., Stehman, S. y Homer, C. Spatial patterns of the united states national land cover dataset (nlcd) land-cover change thematic accuracy (2001–2011). *International Journal of Remote Sensing*, vol. 39, páginas 1729–1743, 2018.