Métodos de aprendizaje automático para la predicción de la distribución anormal de grasa, masa magra y/o masa ósea en individuos infectados por VIH.

Manuel Ramírez
Máster en Bioinformática y Estadística

Nuria Pérez Álvarez
Carles Ventura Royo

02/01/2019

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Métodos de aprendizaje automático para la predicción de la distribución anormal de grasa, masa magra y/o masa ósea en individuos infectados por VIH.*
Nombre del autor:	*Manuel Ramírez Garrastacho*
Nombre del consultor/a:	*Nuria Pérez Álvarez*
Nombre del PRA:	*Carles Ventura Royo*
Fecha de entrega (mm/aaaa):	01/2019
Titulación::	*Máster en Bioinformática y Estadística*
Área del Trabajo Final:	Analisis de datos. M0.178 - TFM-Bioinformática y Bioestadística. Àrea 2 - aula 1.
Idioma del trabajo:	*Español*
Palabras clave	*machine learning, VIH, composición corporal*

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

El objetivo principal de este trabajo es el desarrollo de nuevos modelos de predicción de enfermedades relacionadas con la distribución de masa grasa, masa magra y tejido óseo en pacientes de VIH utilizando métodos de aprendizaje automático.

Debido a la mejora en las terapias contra el VIH, los pacientes que padecen esta enfermedad tienen en muchos casos una esperanza de vida similar a la de una persona sana. Esto ha provocado un creciente interés en el tratamiento de patologías secundarias derivadas de este síndrome o su tratamiento. Algunas de las patologías secundarias más frecuentes en pacientes de VIH son las relacionadas con la distribución de ciertos tejidos, como la lipodistrofia, la sarcopenia o la osteoporosis y osteopenia.

Para la realización del trabajo se ha comparado la eficiencia en la capacidad de predicción de estas enfermedades de un modelo tradicional de regresión logística frente a distintos modelos de aprendizaje automático. Los algoritmos utilizados son algunos de los más más usados en biomedicina: redes neuronales artificiales, *support vector machines* y *random forest*. Todos estos modelos se han implementado usando el lenguaje estadístico R.

Los resultados obtenidos muestran que los modelos generados a partir estos modernos algoritmos son en la mayoría de los casos tan eficaces como el modelo clásico. En ciertas ocasiones es incluso posible conseguir resultados mejores utilizando modelos de aprendizaje automático. Esto abre interesantes

posibilidades a la hora de mejorar la predicción de la aparición de estas patologías secundarias.

Abstract (in English, 250 words or less):

The objective of this work is the development of new prediction models for pathologies related to the distribution of fat mass, lean mass and bone tissue in HIV patients using machine learning methods.

Due to the improvement in anti-HIV therapies, patients suffering from this disease have often a life expectancy similar to a healthy person. This fact has led to a growing interest in the treatment of secondary pathologies derived from this syndrome or its treatment. Some of the most frequent secondary pathologies in HIV patients are those related to the distribution of certain tissues, such as lipodystrophy, sarcopenia or osteoporosis and osteopenia.

In this project the predictive efficiency of a traditional logistic regression model has been compared with different models created using machine learning. The algorithms used in this work are some of the most used in biomedicine: artificial neural networks, support vector machines and random forest. All of them have been implemented using the statistical language R.

The results obtained show that the models generated with these modern algorithms are in most cases as effective as the classical regression model. Sometimes it is even possible to achieve better results using machine learning models. This work shows very promising possibilities in the prediction of the appearance of secondary pathologies in HIV patients using these novel techniques.

Lista de figuras

- Figura 1: Distribución de variables en la base de datos.
- Figura 2: Distribución de las distintas enfermedades en los grupos *training* y *testing*.
- Figura 3: Resultados de los modelos de regresión logística para lipodistrofia y sarcopenia.
- Figura 4: Resultados de los modelos de regresión logística para osteopenia y osteoporosis.
- Figura 5: Modelo ANN de un solo nodo para lipodistrofia en hombres.
- Figura 6: Modelo ANN de un solo nodo con variables seleccionadas por Lasso para lipodistrofia en hombres
- Figura 7: Modificaciones del modelo ANN para lipodistrofia en hombres.
- Figura 8: Modelos ANN de uno y tres nodos para lipodistrofia en mujeres.
- Figura 9: Diversos modelos ANN para sarcopenia en hombres.
- Figura 10: Diversos modelos ANN para sarcopenia en mujeres.
- Figura 11: Modelos SVM sin transformación de datos para lipodistrofia en hombres.
- Figura 12: Modelos SVM con transformación de datos para lipodistrofia en hombres.
- Figura 13: Modelos SVM con modificaciones en la variable coste para lipodistrofia en hombres.
- Figura 14: Modelos SVM sin transformación de datos para lipodistrofia en mujeres.
- Figura 15: Modelos SVM con transformación de datos y modificaciones en la variable de coste para lipodistrofia en mujeres.
- Figura 16: Modelos SVM con transformación de datos y modificaciones en la variable de coste para sarcopenia en hombres.
- Figura 17: Modelos SVM con transformación de datos y modificaciones en la variable de coste para sarcopenia en mujeres.
- Figura 18: Matriz de confusión para el modelo SVM sin transformación de datos para enfermedades del tejido óseo en hombres.
- Figura 19: Matriz de confusión para el modelo SVM con transformación de datos para enfermedades del tejido óseo en hombres.
- Figura 20: Matriz de confusión para el modelo SVM con transformación radial para enfermedades del tejido óseo en mujeres.
- Figura 21: Modelos *random forest* para lipodistrofia en hombres.
- Figura 22: Modelos *random forest* con distinto número de árboles para lipodistrofia en hombres.
- Figura 23: Modelo *random forest* para lipodistrofia en mujeres.
- Figura 24: Modelos *random forest* con distinto número de árboles para sarcopenia en hombres.

- Figura 25: Modelo *random forest* para sarcopenia en mujeres.
- Figura 26: Matriz de confusión para el modelo *random forest* con 1000 árboles para enfermedades del tejido óseo en hombres.
- Figura 27: Matriz de confusión para el modelo *random forest* con 5000 árboles para enfermedades del tejido óseo en mujeres.
- Figura 28: Diagrama de barras con los resultados de AUC de los modelos de lipodistrofia en hombres.
- Figura 29: Diagrama de barras con los resultados de AUC de los modelos de lipodistrofia en mujeres.
- Figura 30: Diagrama de barras con los resultados de AUC de los modelos de sarcopenia en hombres.
- Figura 31: Diagrama de barras con los resultados de AUC de los modelos de sarcopenia en mujeres.
- Figura 32: Diagrama de barras con los resultados de AUC de los modelos de enfermedades del tejido óseo.
- Figura 33: Modelo *random forest* optimizado para lipodistrofia en hombres.
- Figura 34: Matrices de confusión para los modelos de regresión logística y *random forest* optimizado para lipodistrofia en hombres.
- Figura 35: Modelo *random forest* optimizado con modificación en el valor de *cutoff* para lipodistrofia en hombres.
- Figura 37: Modelo *random forest* optimizado para lipodistrofia en mujeres.
- Figura 37: Matrices de confusión de los modelos de regresión y *random forest* para lipodistrofia en mujeres
- Figura 38: Modelo *random forest* con valor *cutoff* modificado para lipodistrofia en mujeres.
- Figura 39: Modelo *random forest* optimizado para sarcopenia en hombres.
- Figura 40: Matrices de confusión de los modelos *random forest* con distintos valores de *cutoff* para sarcopenia en hombres.
- Figura 41: Modelos ANN optimizados para sarcopenia en mujeres.
- Figura 42: Análisis de las enfermedades del tejido óseo categorizando a los pacientes en dos posibles diagnósticos.

Índice

1. Introducción.....1
1.1 Contexto y justificación del Trabajo.....1
1.2 Objetivos del Trabajo.....2
1.3 Enfoque y método seguido.....2
1.4 Planificación del Trabajo.....3
1.5 Breve sumario de productos obtenidos.....5
1.6 Breve descripción de los otros capítulos de la memoria.....5
2. Resto de capítulos.....6
2.1 Análisis y preparación de los datos.....6
2.2 Regresión logística.....8
2.3 Redes neuronales artificiales (ANN).....10
2.4 *Support vector machines*.....16
2.5 *Random forest*.....23
2.6 Análisis de los resultados.....27
2.7 Optimización del algoritmo *random forest*.....30
3. Conclusiones.....37
4. Glosario.....40
5. Bibliografía.....41

1. Introducción

1.1 Contexto y justificación del Trabajo

Este trabajo busca desarrollar nuevos método para la predicción de patologías relacionadas con la distribución anormal de grasa o tejido óseo en pacientes infectados con virus VIH utilizando herramientas de aprendizaje automático de inteligencias artificiales (*machine learning*). En los últimos años la esperanza de vida de los pacientes con VIH ha aumentado significativamente, de forma que aquellos individuos que han sido diagnosticados y han comenzado el tratamiento retroviral durante la última década tienen una esperanza de vida que se aproxima a la de la población general (1). Esto ha hecho que se haya producido un creciente interés en el estudio de patologías derivadas de la enfermedad que pueden afectar a los pacientes a largo plazo. Uno de estos efectos es la anormal distribución de ciertos tejidos corporales. bien por causa de la medicación o bien por la carga viral en el organismo.

Una de estas patologías es la lipodistrofia, una enfermedad que provoca cambios físicos y metabólicos en las personas afectadas. Los cambios físicos se refieren a una anormal distribución de la grasa corporal. Ésta disminuye en los brazos, las piernas y la cara, y forma acúmulos en otras partes del cuerpo, especialmente el estómago, el cuello o nuca y los pechos. Los cambios metabólicos incluyen cambios en el nivel de grasas y ácido láctico en la sangre y algunos pacientes pueden desarrollar resistencia a la insulina. Algunos estudios afirman que más de la mitad de los infectados por VIH sufren lipodistrofia (2), aunque estos datos muestran una alta variabilidad entre distintos artículos y otros trabajos muestran cifras algo inferiores (3, 4).

Si bien la lipodistrofia no es una enfermedad mortal por sí misma, sus efectos disminuyen la calidad de vida del paciente y pueden llevarle a considerar el abandono del tratamiento antiviral. Además, en ciertos casos el exceso de grasa en la sangre puede aumentar el riesgo de sufrir una enfermedad cardiaca y la resistencia a la insulina favorece la aparición de diabetes.

Al igual que ocurre con la grasa, también el tejido muscular puede verse afectado en estos pacientes. La disminución del tejido muscular se denomina sarcopenia y se ha observado en diversos estudios. Por ejemplo, en (6) se observó que más de un 20% de los pacientes habían perdido una cantidad significativa de músculo después de un año de tratamiento.

Otro problema asociado con el VIH es la disminución de la densidad mineral ósea, es decir, la cantidad de calcio y otros minerales presentes en los huesos (7, 8). Se ha demostrado la pérdida de masa ósea en pacientes durante los primeros años de tratamiento con medicamentos antirretrovirales (9). Esta disminución puede provocar osteopenia, caracterizada por una baja densidad ósea, que pueda dar lugar a osteoporosis, una patología provocada por una disminución en la calidad del tejido ósea que lleva aparejado un aumento en las probabilidades de sufrir fracturas en los huesos. De hecho, un metaanálisis mostró que la prevalencia de osteoporosis en pacientes tratados con antirretrovirales es tres veces mayor que en un grupo control (10).

En los últimos años se ha producido un gran avance que el desarrollo del aprendizaje automático para inteligencias artificiales. Esto se ha debido principalmente a la mejora de los algoritmos, la mayor capacidad computacional y al registro y almacenamiento de una cantidad cada mayor de información. Si consideramos que una mayor cantidad de datos supone una mayor eficiencia en la eficacia de los modelos de *machine learning*, nos damos cuenta de que la biomedicina es uno de los campos que más puede favorecerse de este campo, ya que actualmente es posible recopilar una gran cantidad de datos sobre las características físicas de los pacientes.

Existen ya numerosas iniciativas encaminadas a desarrollar algoritmos de machine learning que faciliten el diagnóstico de diversas patologías. DeepMind Health es un proyecto de Google en colaboración con varios hospitales para desarrollar herramientas de este tipo. En agosto de 2018 publicaron su primer artículo en Nature Medicine (11). Otra iniciativa es la competición Camelyon (https://camelyon17.grand-challenge.org/), en la que distintos grupos compiten para desarrollar el mejor algoritmo de detección de cáncer de mama a partir del análisis automático de imágenes.

1.2 Objetivos del Trabajo

Para facilitar del desarrollo del trabajo decidí establecer una serie de objetivo con fechas concretas para la consecución de cada uno (especificadas en el punto 1.4 Planificación del trabajo). La idea detrás de mi planificación era utilizar en un principio las herramientas disponibles de forma genérica y, una vez observados los primeros resultados, usarlas de forma más específica para mi problema concreto.

Los objetivos que me tracé fueron:

-Objetivo 1:

Elegir, entre todas las herramientas disponibles, diversos algoritmos de aprendizaje automático y desarrollar diversos modelos para la predicción de la aparición de patologías relacionadas con la distribución anormal de grasa, músculo y tejido óseo en pacientes de VIH

- Objetivo 2:

Cuantificar la efectividad de dichos modelos para las distintas patologías y determinar cuál sería el más adecuado para su utilización como predictor de la aparición de enfermedades secundarias.

- Objetivo 3:

Mejora del modelo (o modelos) seleccionado para conseguir la mejor predicción posible.

1.3 Enfoque y método seguido

El propósito del análisis estadístico de datos clínicos es en muchos casos describir la relación entre dos o más variables. Por ejemplo, un estudio podría buscar la relación entre el peso de un paciente y su nivel de colesterol, o podría añadir más variables y analizar cómo afectan al colesterol la edad y el género de los pacientes. Con

frecuencia estos estudios buscan establecer la relación entre variables que pueden ser estudiadas en los pacientes y la posibilidad de padecer una determinada enfermedad. Siguiendo con el ejemplo anterior, se podría tratar de cuantificar la posibilidad de sufrir problemas cardiacos dependiendo del nivel de colesterol.

Los modelos clásicos de regresión se han usado durante décadas para realizar este tipo de análisis. De Fauw et al (12) presentan una revisión de estos métodos. Sin embargo, en los últimos años se ha experimentado un cambio hacia la utilización de modelos de aprendizaje automático (*machine learning*). En este tipo de algoritmos, el ordenador es capaz de "aprender" a realizar mejor una determinada función a través de sucesivas iteraciones. Muchos de estos procesos se denominan "de caja negra (*blackbox*)", ya que el modelo generado no es fácil de comprender por el usuario, en contraposición por ejemplo al resultado de un modelo de regresión cuya función se expresa fácilmente en términos matemáticos.

En los últimos años se han publicado multitud de artículos científicos en los que se comprueba la eficiencia de distintos métodos de aprendizaje automático aplicados a la predicción de la aparición o la tasa de supervivencia de distintas enfermedades (13 – 16 muestra una selección de publicaciones recientes). Entre todos los algoritmos utilizados he seleccionado tres que considero destacados tanto por la frecuencia de su uso como por los resultados que ofrecen:

- redes neuronales artificiales (ANN)
- *support vector machines* (SVM)
- *random forest*

Los distintos algoritmos de aprendizaje automático no son métodos fijos, sino que cada uno de ellos depende de ciertos parámetros o transformaciones de los datos que pueden ser modificados para optimizar los resultados. En este trabajo utilizaré R, una herramienta que posee diversos *packages* donde se incluyen las instrucciones para ejecutar determinadas funciones. Para algunos algoritmos de *machine learning* existen diferentes *packages* que funcionan de formas ligeramente diferentes y cuyas eficiencias podrían ser distintas.

El enfoque que me propuse desde un principio para este trabajo fue el de utilizar distintos algoritmos para diseñar modelos "sencillos" (es decir, utilizando los parámetros por defecto o realizando modificaciones sencillas en los algoritmos utilizados). Tras esta primera fase se compararía la eficiencia de los distintos y elegiría algoritmo o algoritmos más eficientes. En una última fase se profundizaría en ellos, modificando sus parámetros o buscando nuevos *packages* que utilicen el mismo tipo de modelo, con la intención de refinar el modelo y conseguir la mejor eficiencia posible.

1.4 Planificación del Trabajo

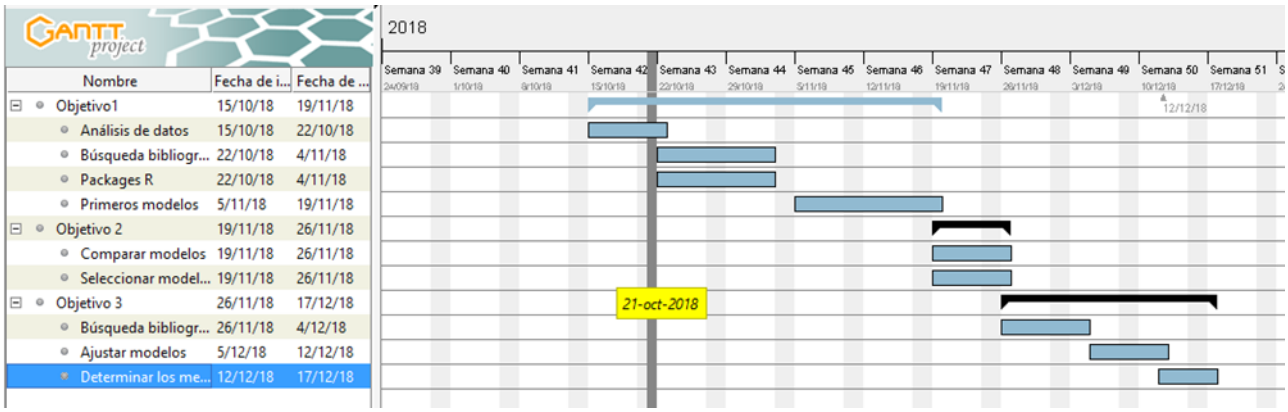
Antes de comenzar el trabajo con la base de datos realicé una asignación de tiempo en la que dividí cada uno de los objetivos en varias tareas y asigné a cada una de ellas un tiempo determinado para desarrollarla. Aquí muestro esa planificación

- Objetivo 1:
    - Análisis de los datos disponibles: 15/10/18 – 22/10/18
    - Búsqueda bibliográfica de posibles algoritmos de aprendizaje atomático: 22/10/18 – 4/11/18
    - Búsqueda de información sobre *packages* para R relacionados con los algoritmos seleccionados en el paso anterior: 22/10/18 – 4/11/18
    - Elaboración de los primeros modelos: 5/11/18 – 19/11/18

- Objetivo 2:
    - Comparación de los modelos generados: 19/11/18 – 26/11/18
    - Elección de los mejores modelos: 19/11/18 – 26/11/18

- Objetivo 3:
    - Búsqueda bibliográfica sobre técnicas para refinar los modelos que han mostrado un mejor resultado: 26/11/18 – 4/12/18
    - Refinar los modelos: 5/12/18 – 12/12/18
    - Determinar la eficiencia de los modelos refinados: 12/12/18 – 17/12/18

También creé un calendario para el proyecto usando el programa de distribución gratuita GranttProject. Aquí se puede ver una captura de pantalla del calendario completo:



La planificación se llevó a cabo en el orden establacido, aunque no dentro de los tiempos que se habían especificado. Concretamente, la elaboración de los primeros modelos llevó más tiempo del especificado debido a una mala selección de las variables en un principio, lo que producía errores en el código de R debido a variables con una correlación demasiado alta, y a la cantidad de opciones para la elaboración de distintos modelos, especialmente en el caso de SVM. Esto provocó que los objetivos 2 y 3 comenzaran más tarde de la fecha especificada y que los resultados finales de los modelos optimizados no estuvieran listos el 17 de diciembre, como se indica en la planificación, sino una semana después.

1.5 Breve sumario de productos obtenidos

Esta memoria se acompaña de un archivo de Markdown y el .pdf generado a partir de él. El archivo Markdown contiene todo el código en R utilizado para la realización de este trabajo, con comentarios sobre los distintos pasos realizados y los resultados obtenidos.

1.6 Breve descripción de los otros capítulos de la memoria

- Análisis y preparación de los datos

En este capítulo se describen las características del archivo de datos de pacientes con el que se ha trabajado en este proyecto y se comentan las variables que contiene. Además se describen las modificaciones realizadas sobre el archivo de datos antes de comenzar el análisis y se explican los distintos grupos (determinados por tipo de enfermedad y género) en los que se dividió el archivo.

- Regresión logística

Elaboración y evaluación de modelos de regresión logística para cada uno de los grupos en los que se ha dividido el trabajo.

- Redes neuronales artificiales (ANN)

Elaboración y evaluación de modelos utilizando el algoritmo de redes neuronales artificiales.

- *Support vector machines* (SVM)

Elaboración y evaluación de modelos utilizando el algoritmo de redes neuronales artificiales.

- *Random forest*

Elaboración y evaluación de modelos utilizando el algoritmo de redes neuronales artificiales.

- Análisis de los resultados

Comparación de los resultados obtenidos para cada una de las patologías estudiadas y para cada género.

- Optimización del algoritmo random forest

Búsqueda de los parámetros óptimos para modelos del algoritmo random forest, que había mostrado los mejores resultados. Comparación de los resultados obtenidos con random forest y regresión logística.

2. Resto de capítulos

Los siguientes capítulos contienen el análisis de los datos derivados de pacientes que se han utilizado en este trabajo: preparación de los datos, elaboración de los modelos y análisis de la eficiencia. En ellos no se especifica el código utilizado y no se muestran todas las gráficas y modelos generados durante el trabajo. El análisis completo se puede ver paso por paso en los archivos Markdown y .pdf adjuntos.

2.1 Análisis y preparación de los datos

El archivo de datos original posee observaciones de 82 variables para 1480 pacientes de VIH. Estas observaciones se corresponden a información general del paciente (sexo, altura, peso, etc), información específica sobre la masa grasa y masa ósea en distintos tejidos y partes del cuerpo e información sobre la presencia o ausencia de las patologías que se desean estudiar: lipodistrofia, sarcopenia, osteopenia y osteoporosis. El primer paso del trabajo consistió en el estudio y la preparación de los datos para la generación de los modelos.

En primer lugar eliminé las variables no útiles (identificación del paciente) o duplicadas (existían dos variables diferentes relacionadas con la edad y con el sexo).

Las columnas correspondientes a la presencia de enfermedades óseas ("minTscore" y "Tscore_3cat") estaban vacías. Sin embargo, los valores de estas columnas podían deducirse a partir de los valores de otras variables. Para cada paciente calculé su "T score" mínimo y le asigné un determinado grupo dentro de las enfermedades óseas: "normal", "osteopenia" u "osteoporosis".

Muchos modelos estadísticos tienen problemas para evaluar observaciones en las que faltan algunos de los datos. Comprobé si esto ocurría en mi base de datos y encontré 22 pacientes en los que alguna de las variables no mostraba ningún valor. Debido que este número representa una fracción muy pequeña del total de las observaciones decidí eliminar estos pacientes de la base de datos, con lo que trabajé sólo con las 1458 observaciones restantes.

La base de datos estaba compuesta por 1104 hombres y 354 mujeres. Hombres y mujeres tienen distinta fisiología y estas diferencias entre distintos sexos podrían ser más importantes que las diferencias existentes entre pacientes enfermos y sanos. Por lo tanto, decidí separar a los pacientes según su sexo y examinar cada grupo de forma independiente.

Con el fin de tener una idea de la composición de cada uno de estos grupos, representé en histogramas la distribución de las variables más sencillas: edad, peso, altura e índice de masa corporal **(Figura 1)**:

El resto de las variables indican la distribución de los distintos tejidos o el estado del paciente (enfermo o sano) para cada una de las enfermedades a analizar. Dentro de estas variables hay varias que están directamente relacionadas con cada una de las patologías. El objetivo del estudio es elaborar modelos que puedan predecir la presencia de una determinada enfermedad sin utilizar variables que se encuentran directamente relacionadas con dicha enfermedad. Por lo tanto, para analizar la lipodistrofia y la sarcopenia voy a utilizar únicamente las variables no relacionadas con la cantidad de masa grasa o masa magra, y para predecir la presencia de osteoporosis u osteopenia no voy a usar variables relacionadas con la distribución de tejido óseo.

Además, es necesario dividir los datos en dos grupos: *training* y *testing*. El primero servirá para generar el modelo y el segundo se usará para comprobar su eficacia. Esto es importante ya que si se utilizara el mismo grupo de datos para la generación del modelo y la comprobación de su efectividad se correría el riesgo de generar un modelo que se ajusta de forma excesiva a los datos que se le han suministrado y no puede ser extrapolado a datos independientes. En la literatura he visto que el grupo *training* está normalmente compuesto por al menos dos tercios de los observaciones de la base de datos, por lo que para generar mis grupos seleccioné aleatoriamente al 67% de las observaciones como *training* y al 33% restante como *testing*.

De esta forma generé doce grupos de datos, cuatro para cada una de las tres enfermedades a analizar (osteoporosis y osteopenia se analizaron juntas): dos grupos *training* (uno con los datos procedentes de hombres y otro con los de mujeres) y dos grupos *testing*. Antes de generar los modelos comprobé que el porcentaje de pacientes enfermos era similar en los grupos *training* y *testing* para cada una de las patologías y sexos (**Figura 2**):

Hombres	Mujeres
Lipodistrofia	
Training: 74% Testing: 77%	Training: 67% Testing: 67%
Sarcopenia	
Training: 21% Testing: 25%	Training: 32% Testing: 32%
Osteopenia y osteoporosis	
Training osteopenia: 59% Training osteoporosis: 13% Testing osteopenia: 50% Testing osteoporosis: 16%	Training osteopenia: 41% Training osteoporosis: 17% Testing osteopenia: 47% Testing osteoporosis: 13%

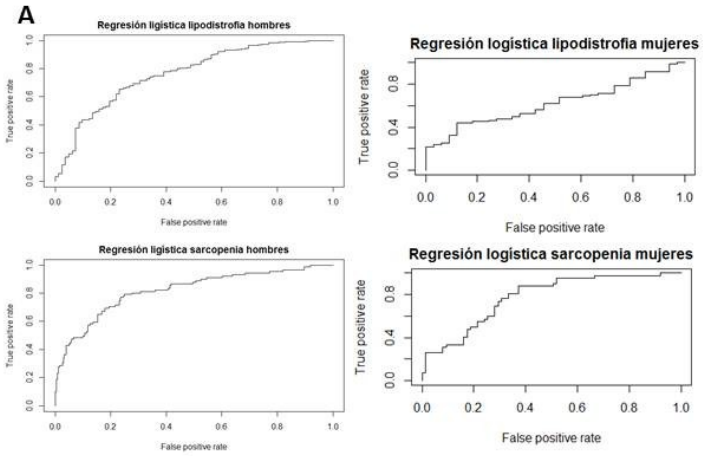
Figura 2: Porcentaje de pacientes enfermos en los grupos "training" y "testing" de cada patología y género.

2.2 Regresión logística.

Para poder determinar la eficiencia de los modelos construidos con herramientas de aprendizaje automático decidí generar primero un modelo de regresión logística. Este modelo serviría como referencia para comprobar si los algoritmos de *machine learning* eran capaces de generar modelos superiores al conseguido con una técnica clásica. Para determinar la eficiencia de cada modelo decidí utilizar la medida del área bajo la curva (AUC) de una curva ROC (acrónimo de *receiver operating characteristic*) (17). Esta medida es un estándar muy utilizado para la evaluación de la efectividad de métodos predictivos que se encuentran en la mayor parte de los artículos científicos (13-16). Los valores de AUC se encuentran en el rango de 0.5 (el resultado obtenido es igual al de asignar las categorías al azar) hasta 1 (todas las observaciones han sido predichas correctamente). Para la elaboración de estas curvas utilicé el *package* "ROCR" (18).

Incluso después de eliminar las variables directamente relacionadas con la enfermedad, el número de variables disponibles para realizar los análisis era bastante alto: 34 en el caso de lipodistrofia y sarcopenia y 40 en el caso de enfermedades óseas. Muchas de estas variables podrían tener una fuerte correlación entre ellas. Por ejemplo,

entre las variables que miden la masa grasa existía una variable para cada una de las extremidades del paciente, aunque es probable que estos valores estén relacionados entre ellos. Del mismo modo, dentro de las variables que medían la masa ósea había una variable diferente para cada hueso de las lumbares. El hecho de que dos o más variables aporten información similar puede afectar a la eficiencia de un modelo de regresión, así que antes de generarlo decidí seleccionar las variables más significativas usando el método Lasso (19), utilizando el *package* "glmnet". La **Figura 2** muestra los valores de AUC obtenidos para los modelos de regresión logística de lipodistrofia y sarcopenia.



B

	Hombres	Mujeres
Lipodistrofia	0.76	0.65
Sarcopenia	0.82	0.77

Figura 3: A: Curvas ROC de los modelos de regresión para lipodistrofia y sarcopenia. **B:** Valor de AUC para cada uno de los modelos.

En el caso de enfermedades relacionadas con el tejido óseo, el número de posibles diagnósticos es de tres en vez de dos, por lo que las técnicas anteriores (Lasso y curva ROC) no son aplicables. Para seleccionar las variables más significativas decidí calcular la correlación entre cada una de las variables y todas las demás y eliminar todas aquellas variables que tuvieran un valor de correlación de al menos 0.8 con al menos seis otras variables. Después generé un modelo utilizando la función multinom() del *package* "nnet" (21), que acepta realizar regresión logística con más de dos posibles resultados. Como medida de la eficacia del modelo generé una matriz de confusión usando el package "caret" (22). La **Figura 4** muestra los resultados obtenidos:

A

Predicción \ Pacientes	Normal	Osteopenia	Osteoporosis
Normal	28	8	3
Osteopenia	96	165	52
Osteoporosis	2	8	2

Precisión: 54%

B

Predicción \ Pacientes	Normal	Osteopenia	Osteoporosis
Normal	29	19	2
Osteopenia	17	33	12
Osteoporosis	1	3	1

Precisión: 54%

Figura 4: Matrices de confusión que muestran los resultados de la regresión logística para enfermedades del tejido óseo en hombres (A) y mujeres (B).

Los valores de precisión obtenidos con estos modelos son muy bajos. Hay que tener en cuenta que el hecho de que haya tres posibles diagnósticos es una dificultad añadida al modelo, pero una precisión tan pobre haría que este modelo fuera poco efectivo en la predicción de patologías.

2.3 Redes neuronales artificiales (ANN)

El algoritmo Redes Neuronales Artificiales (ANN) modela la relación entre unos determinados valores de entrada y uno o más valores de salida utilizando una aproximación que trata de imitar un cerebro biológico. Para ello estos modelos poseen varias capas de nodos que reciben información, la modifican y transmiten al siguiente nodo, de la misma forma que el cerebro biológico posee capas de neuronas interconectadas entre ellas. En un principio el algoritmo asigna aleatoriamente un determinado peso a cada unidad de información (variable). A lo largo de sucesivos

ciclos de entrenamiento el peso de cada una de las variables se va modificando para perfeccionar el modelo y acercarse a los resultados reales. Estos cambios en la importancia relativa de las variables, que simulan el fortalecimiento o deterioro de las conexiones neuronales durante el desarrollo del cerebro, son la clave de la eficiencia del algoritmo. Una revisión del tema se puede encontrar en Abiodun OI *et al* (23).

Para la generación de los modelos ANN utilicé el *package* "neuralnet" (24). Al igual que he hecho anteriormente en el caso de la regresión, voy a aplicar este método a cada una de las patologías a estudiar.

- **Lipodistrofia**

La función neuralnet() requiere que los datos estén normalizados. Por lo tanto antes de generar el modelo apliqué sobre los datos la normalización "min-max": a cada dato se le resta el valor mínimo de esa variable y el resultado se divide entre la diferencia entre el valor máximo y el mínimo de la variable. De esta forma todos los resultados están comprendidos entre 0 y 1.

Para comenzar generé un modelo sencillo con una sola capa y un solo nodo. La **Figura 5** muestra el resultado de aplicar este modelo a los datos procedentes de hombres. La representación del modelo ANN contiene demasiada información sobre las variables en un espacio reducido, por lo que no es posible apreciar el peso asignado a cada una de ellas:

El resultado es muy inferior a lo obtenido con regresión logística. Es posible que esto se deba a que utilicé demasiadas variables que están correlacionadas entre sí. Para comprobarlo generé un segundo modelo utilizando sólo las variables que habían sido seleccionadas por el método Lasso en el apartado anterior (**Figura 6**):

El valor de AUC mejoró sensiblemente al usar menos variables, aunque siguió siendo inferior al obtenido con regresión logística. Para comprobar las posibilidades del algoritmo generé distintos modelos aumentando el número de nodos o capas usando sólo las variables seleccionadas por Lasso (**Figura 7**):

El modelo con 2 capas de un solo nodo consiguió un mejor resultado que otros modelos más complejos, aunque su efectividad fue todavía levemente inferior a la del modelo de regresión logística. Posteriores intentos de aumentar la complejidad del modelo fracasaron al no tener suficiente capacidad computacional.

Repetí el proceso para los datos procedentes de mujeres utilizando únicamente las variables seleccionadas por Lasso. Muestro el resultado para el modelo simple y el modelo de 3 nodos, que fue de nuevo el que dio el mejor resultado (**Figura 8**):

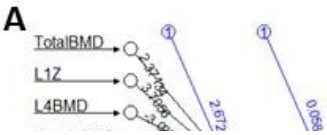
En este caso los modelos ANN han dado un resultado muy pobre.

- **Sarcopenia**

A la vista de los resultados anteriores decidí utilizar sólo las variables seleccionadas con Lasso para generar los modelos de sarcopenia. Al igual que hice en el apartado anterior, comienzo mostrando los datos más representativos que obtuve para los diferentes modelos generados a con los datos obtenidos a partir de pacientes masculinos (**Figura 9**):

Los tres modelos han dado resultados muy similares, pero al igual que ocurría con la lipodistrofia en varones la eficiencia de los modelos ANN es inferior a la obtenida con el modelo de regresión logística.

La **Figura 10** muestra los resultados que obtuve al aplicar los mismos algoritmos a los datos de sarcopenia obtenidos a partir de pacientes femeninos. En este caso puede observarse como al incrementar la complejidad del modelo disminuye su eficiencia. El modelo más sencillo, de un solo nodo, consigue un resultado ligeramente superior al obtenido con regresión logística.



- Osteoporosis y osteopenia

La función neuralnet() requiere que todos los datos que se le suministren sean numéricos. En el caso de los datos relacionados con patologías del tejido óseo la variable a predecir era categórica, así que para solucionarlo codifiqué la misma información en tres columnas denominadas “normal”, “osteopenia” y “osteoporosis”. Asigné a cada una de estas columnas el valor 0 o 1 dependiendo del estado del paciente.

Al generar las matrices de confusión para comprobar la precisión de cada uno de los modelos en hombres me encontré con un error en RStudio. Leyendo la documentación sobre esta función vi que ese error aparece cuando uno de los posibles resultados no ha sido predicho por el modelo, lo que provoca que la predicción y el resultado real tengan un número distinto de niveles. Comprobé los resultados de la predicción y vi que tanto el modelo sencillo (1 nodo) como el modelo con dos capas y un nodo en cada capa habían predicho que todos los pacientes sufrirían osteopenia. El modelo con tres nodos repartía sus predicciones entre osteoporosis y osteopenia, pero no encontraba ningún paciente que no presentara la enfermedad. Al analizar los datos referidos a mujeres encontré un problema similar, aunque en este caso el error se producía porque ninguno de los modelos había predicho que alguno de los pacientes sufriría osteoporosis. Estos resultados me llevaron a descartar el algoritmo ANN como un buen predictor para las patologías relacionadas con el tejido óseo.

2.4 *Support vector machines* (SVM)

Support vector machines (SVM) son algoritmos que clasifican a cada una de las observaciones asignándoles una posición en un espacio multidimensional y después trazando hiperplanos que separan las distintas categorías. El número de dimensiones del espacio es igual al número de variables analizadas. Este método es muy fácil de visualizar si pensamos en un plano bidimensional, en el que el funcionamiento sería equivalente a trazar una línea que separara los individuos pertenecientes a cada categoría. Nuestro cerebro no es capaz de imaginar espacios mayores a tres dimensiones, pero el principio del funcionamiento del algoritmo es el mismo. Un buen ejemplo de aplicaciones para SVM se encuentra en (25).

Las SVM presentan un gran número de posibilidades a la hora de generar los modelos. Al utilizar este sistema podemos encontrarnos con que las muestras son linealmente separables o no. Linealmente separables significa que es posible trazar un plano recto que delimite perfectamente los distintos grupos de muestras. En el caso de grupos no linealmente separables el sistema requiere pasos adicionales. Una posibilidad es definir una variable de coste que permite que un cierto número de individuos estén posicionados en el grupo incorrecto cuando se selecciona el MMH. Otra solución es realizar una transformación de los datos añadiendo nuevas características (y por lo tanto nuevas dimensiones) para lograr que las distintas categorías se vuelvan linealmente separables. Para implementar estos modelos utilicé el *package* "kernlab" de R (26).

- **Lipodistrofia**

Para comenzar voy a usar los parámetros por defecto para generar dos modelos utilizando los datos derivados de hombres. Para uno de los modelos usaré todas las variables y para el otro sólo aquéllas seleccionadas por el método Lasso (**Figura 11**). En este caso no es posible obtener una representación gráfica del modelo, así que muestro sólo el resultado de las curvas ROC:

En este caso la diferencia en el valor de estos dos modelos fue menor a 0.01, por lo que no estaba claro si reducir el número de variables era beneficioso. Para analizar esto en más profundidad generé dos nuevos modelos utilizando dos transformaciones de los datos: radial (rbf) y polinómica. En cada caso utilicé tanto todas las variables como sólo las seleccionadas por Lasso. Muestro los resultados en el la **Figura 12**:

La transformación de los datos no ha supuesto un gran cambio en la eficiencia de los modelos excepto en el caso de la transformación radial con las variables seleccionadas por Lasso, que es inferior a los demás modelos.

Como he mencionado anteriormente, otra de las posibilidades que ofrecen los SVM es modificar la variable de coste. Este valor está relacionado con el número de muestras que pueden clasificarse incorrectamente cuando se desarrolla el modelo. Aumentar la variable de coste puede dar lugar a modelos menos específicos para el grupo de *training* y que funcionan mejor con otros datos independientes. El valor de esta variable es por defecto 1. En la **Figura 13** muestro el resultado para modelos similares a los anteriores pero aumentando la variable de coste a 3.

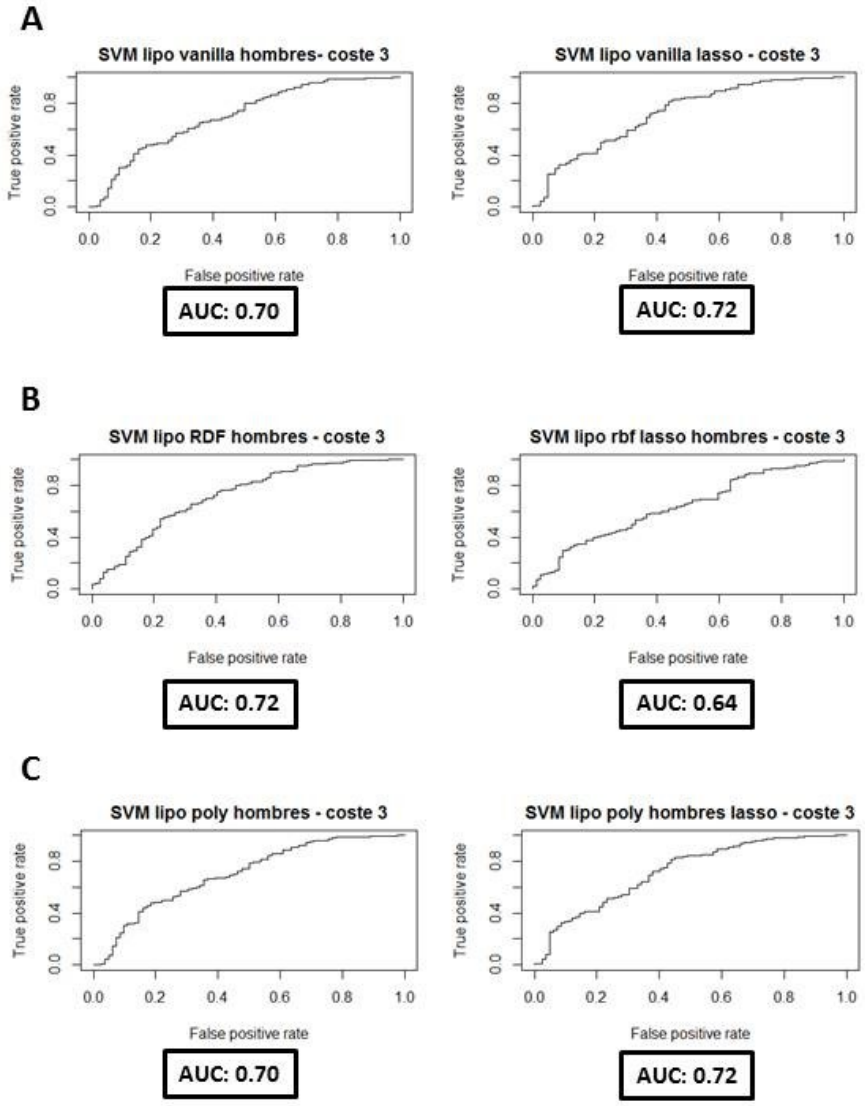
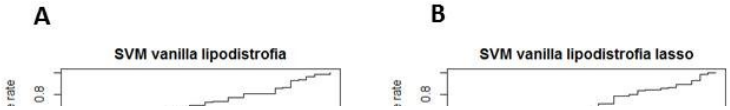


Figura 13: Curvas ROC de las distintas transformaciones de datos para modelos SVM de lipodistrofia para hombres con coste 3. Los paneles de la izquierda muestran los resultados obtenidos con todas las variables y los de la derecho los de las variables seleccionadas por Lasso. **A:** Sin transformar **B:** Transformación radial **C:** Transformación polinómica.

Al igual que ocurría con el algoritmo ANN, los resultados de los modelos de aprendizaje automático para lipodistrofia en hombres no se mostraron superiores a la regresión logística. Repetí el análisis usando los datos referidos a mujeres. Para ello primero generé dos modelos sin transformación usando todas las variables o sólo las seleccionadas por Lasso (**Figura 14**).



El modelo con menos variables es inferior al que tiene todas. Ya que en el caso de los datos procedentes de hombres utilizar sólo las variables seleccionadas por Lasso tampoco demostró tener ninguna ventaja cuando se utilizan SVM, a partir de momento decidí continuar generando únicamente modelos utilizando todas las variables.

Al igual que en el caso anterior, probé a modificar el modelo mediante la transformación de los datos o modificando la variable coste (**Figura 15**).

Muchos de los modelos generados para estos datos consiguen una mejor eficiencia que el modelo clásico de regresión.

- **Sarcopenia**

A la vista de los resultados obtenidos en el análisis de la lipodistrofia, realicé la predicción de la presencia de sarcopenia en los pacientes generando únicamente modelos en los que utilizo todas las variables. La **Figura 16** muestra los resultados obtenidos para los datos procedentes de hombres.

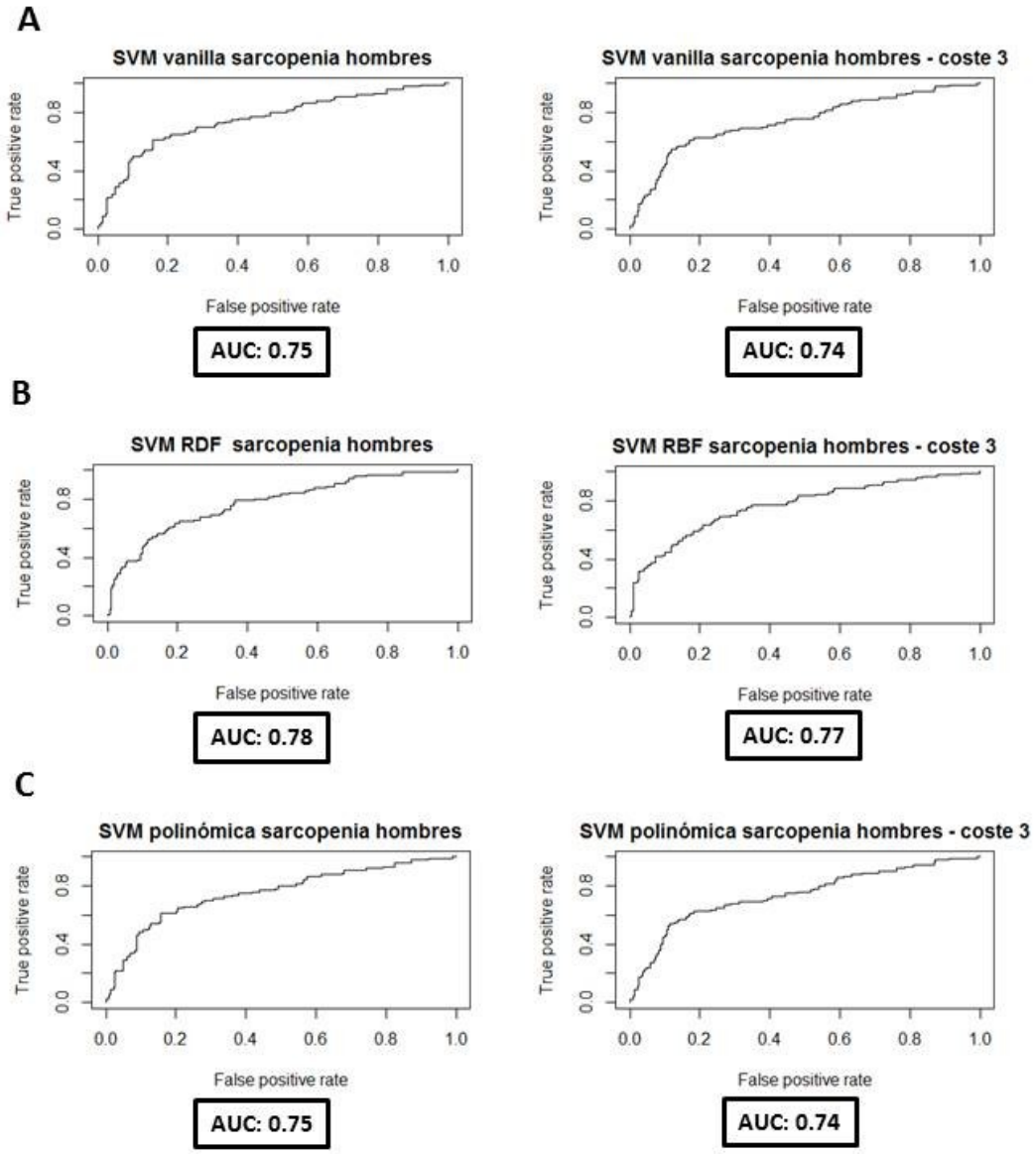


Figura 16: Curvas ROC de las distintas transformaciones de datos para modelos SVM de sarcopenia para hombres. Los paneles de la izquierda muestran los resultados obtenidos con coste = 1 (valor por defecto) y los de la derecha con coste = 3. **A:** Sin transformar **B:** Transformación radial **C:** Transformación polinómica.

Al igual que ocurría en el caso de la lipodistrofia, estos modelos SVM muestran valores de AUC muy similares y en ninguno de los casos han llegado al valor obtenido para la regresión logística (en este caso AUC = 0.82).

Muestro en la **Figura 17** los resultados del mismo análisis realizado con los datos procedentes de pacientes femeninas.

De nuevo los modelos generados muestran valores de AUC muy similares. Ninguno de ellos se comporta mejor que el modelo clásico de regresión logística.

- **Osteoporosis y osteopenia**

Según la documentación del package "kernlab", los modelos SVM pueden utilizar variables categóricas con más de dos posibles resultados sin realizar ningún tipo de transformación. Generé un modelo estándar con los datos derivados de hombres y comprobé su efectividad con una matriz de confusión (**Figura 18**).

Aunque esta vez no he obtenido un error como ocurrió con los modelos ANN, el resultado no es tampoco positivo. El modelo apenas predice un solo paciente dentro del grupo "osteoporosis" (y además es un resultado confundido) y predice de forma exagerada que la mayoría de los pacientes sufren osteopenia. Comprobé si el modelo mejoraba con alguna de las transformaciones (**Figura 19**):

A pesar de la transformación de los datos sigo encontrando el mismo problema. Para comprobar si era un efecto derivado del *package* que estaba usando, repetí el análisis con la función svm() del *package* "e1071" (27). Los resultados fueron muy similares en el sentido de que persistía el problema de que ninguno de los modelos predecía los pacientes aquejados de osteoporosis y siempre se sobrepredecía la osteopenia. La única excepción fue la transformación sigmoidea, que sí que predecía la osteoporosis en ciertos pacientes, pero su número de fallos era mayor que en los demás modelos y por lo tanto su precisión era la más baja.

El análisis de los datos procedentes de mujeres mostró resultados muy similares. Como ejemplo en la **Figura 20** se puede ver el resultado de modelos generado sin realizar ningún tipo de transformación en los datos. Los resultados de los demás modelos fueron similares.

2.5 *Random forest*

Existen un tipo de algoritmos denominados de árboles de clasificación porque generan modelos que se asemejan al patrón de crecimiento de las ramas de un árbol. El modelo comienza con un nodo central donde se encuentras todas las muestras. Tras pasar por un nodo de decisión las muestras se dividen en distintas ramas. Cada una de estas ramas posee otros nodos de decisión que vuelven a subdividir los datos en nuevas ramas, etc. Al final el modelo acaba con nodos "hoja" o terminales, donde se toma una decisión de clasificación. Una de las principales ventajas de estos algoritmos es que si es necesario seguir la lógica que el modelo ha utilizado para clasificar una muestra en un determinado grupo, esto es mucho más sencillo que en modelos generados con ANN o SVM, incluso sin tener grandes nociones de matemáticas o estadística.

Los modelos de *random forest* utilizan este principio para generar un conjunto de árboles de decisión que después se combinan en un sólo árbol. Esto da lugar a modelos con una mayor capacidad predictiva. (28) es una revisión del uso de este algoritmo en un problema concreto (clasificación de imágenes en pacientes de Alzheimer). Para generar los modelos utilicé el *package* "randomForest" (29).

- **Lipodistrofia**

Al igual que he hecho con los algoritmos anteriores, comencé comparando un modelo que utiliza todas las variables con otro que usa sólo aquellas seleccionadas por el método Lasso (**Figura 21**).

El valor de AUC para ambos modelos es muy similar, por lo que decidí utilizar todas las variables para los siguientes modelos. Los resultados mostrados en **Figura 21** se lograron con modelos de *random forest* construidos a partir de 1000 árboles. Para ver si aumentar el número de árboles mejoraba la predicción construí modelos con 2000 y 5000 árboles (**Figura 22**).

Aumentar el número de árboles no supuso una mejora en la eficacia del modelo. Esto está de acuerdo con la literatura existente sobre este modelo, que dice una vez alcanzado el número necesario de árboles los resultados son consistentes en un amplio rango de esta variable. Los resultados son inferiores a lo obtenido con regresión logística.

Analicé entonces los datos relativos a mujeres Utilizando tanto un conjunto de 1000 árboles como uno de 5000 para generar los modelos (**Figura 23**):

A B

De nuevo ambos modelos muestran un resultado muy similar. La eficacia de los modelos es superior a lo que conseguí anteriormente con el modelo de regresión logística. .

- **Sarcopenia**

Realicé un análisis similar para la base de datos de los pacientes de sarcopenia. Muestro en la **Figura 24** los resultados obtenidos al generar modelos a partir de 1000 y 5000 árboles.

Los resultados son muy similares a lo observado en el caso de la lipodistrofia: la eficiencia del modelo no varía al aumentar el número de árboles y el resultado es un poco inferior al obtenido en la regresión logística.

Analicé a continuación los datos procedentes de mujeres creando un modelo a partir de 1000 árboles (**Figura 25**).

Los resultados conseguidos con *random forest* para sarcopenia en mujeres son inferiores a los que he obtenido anteriormente con otros modelos.

- **Osteoporosis y osteopenia**

Al igual que he hecho con las otras dos patologías. Comienzo generando un modelo con 1000 árboles para los datos procedentes de hombres. La **Figura 26** muestra la matriz de confusión con el resultado obtenido.

Aunque en este caso no es tan pronunciado como en modelos anteriores, sigue existiendo un problema de sobrepredicción de la osteopenia. La capacidad de predicción del modelo, medida por el valor de precisión, sigue siendo muy baja. Aumentar el número de árboles hasta 5000 produjo exactamente los mismos resultados.

Viendo que el número de árboles no parece tener un efecto mesurable en la eficiencia del modelo, utilicé el algoritmo en los datos procedentes de mujeres únicamente con 5000 árboles (**Figura 27**).

El problema persiste también en este caso. Ninguno de los modelos que he probado ha conseguido una buena predicción de las enfermedades del tejido óseo.

2.6 Análisis de los resultados

A lo largo de este trabajo he generado un gran número de modelos para la predicción de la lipodistrofia, sarcopenia, osteopenia y osteoporosis utilizando el método clásico de regresión logística y tres métodos de aprendizaje automático: redes neuronales artificiales, *support vector machines* y *random forest*. Para decidir cuál es el modelo más eficiente comparé los valores de AUC de las curvas ROC. En la **Figura 28** se recogen los valores de AUC obtenidos para los modelos de lipodistrofia generados con los datos procedentes de hombres.

En el diagrama de barras puede verse que el modelo de regresión logística consiguió el mejor resultado. Entre los algoritmos de aprendizaje automático, ANN tiene un comportamiento errático con grandes variaciones en su efectividad dependiendo de los parámetros de entrada. SVM y *random forest* son más consistentes, siendo este último el que consigue los mejores resultados. Los modelos generados con random forest tienen un valor de AUC que es apenas dos centésimas inferior al conseguido por la regresión logística.

En la **Figura 29** se pueden ver los resultados obtenido con los datos derivados de mujeres en lipodistrofia. En este caso hay menos resultado que en el diagrama anterior ya que descarté algunos modelos después de observar el comportamiento de ciertos algoritmos en los modelos generados con los datos provenientes de hombres. En este caso casi todos los modelos generados con ANN muestran unos resultados muy pobres, mientras que tanto SVM como *random forest* se comportan mejor que la regresión logística. SVM con transformación radial consigue el mejor valor de AUC, pero la diferencia con los modelos de *random forest* es muy pequeña.

Los resultados de sarcopenia en hombres (**Figura 30**) son muy similares a lo que ocurría en con la lipodistrofia, aunque en este caso los modelos ANN mostraron un mejor comportamiento. El método de regresión logística consiguió el mejor resultado de AUC, aunque algunos de los modelos de aprendizaje automática consiguen un rendimiento que se acerca al método clásico. El método que más se aproximó al resultado del modelo de regresión es el *random forest.*

Al analizar los resultados obtenidos para sarcopenia en los datos procedentes de mujeres (**Figura 31**) encontamos resultados muy diferentes. En este caso el modelo ANN más simple (con un solo nodo) fue el que consiguió el mejor resultado con un valor AUC ligeremente superior al obtenido por la regresión logística. Los demás modelos tuvieron un comportamiento mucho menos eficiente.

La predicción de las patologías relacionadas con el tejido ósea (osteoporosis y osteopenia) no ha funcionado tan bien como las dos anteriores. La **Figura 32** muestra los valores de precisión obtenidos por los distintos modelos aplicados tanto a los datos procedentes de hombres como a los de mujeres. Existe muy poca diferencia entre los distintos modelos y los mejores valores se encuentran en torno al 54%. Para poner este valor en perspectiva con respecto a lo conseguido por otros modelos, el modelo de *random forest* para lipodistrofia en hombres de 1000 árboles, (AUC = 0.74), tiene una precisión del 80%. Mientras que este modelo de lipodistrofia no es completamente fiable, pero sí podría utilizarse como indicador de una posible enfermedad o tal vez como predictor combinado con otro tipo de modelos, el algoritmo que predice

osteoporosis y osteopenia no proporciona resultados fiables que puedan ser utilizados en clínica.

2.7 Optimización de los algoritmos

Viendo el resumen de los resultados del apartado anterior, considero que el algoritmo que ha mostrado un mejor comportamiento es el *random forest*. Este algoritmo ha conseguido los mejores valores de AUC de entre los métodos de aprendizaje automático en la predicción de la lipodistrofia y la sarcopenia en hombres. En esta última parte del trabajo busqué la manera de optimizar este algoritmo para conseguir los mejores resultados posibles.

El *package* "mrl" (30) posee diversas funciones para determinar los parámetros más adecuados para distintos algoritmos de clasificación. Lo usé para determinar los mejores parámetros para determinar si los pacientes sufrían lipodistrofia o sarcopenia. Dado el mal resultado obtenido con las enfermedades de tejido óseo consideré que debía descartar ese análisis.

Los parámetros más importantes de la función randomForest() son:

- ntree: número de árboles. Es el único que modifiqué en las pruebas anteriores y no pareció tener un gran efecto sobre el modelo.
- mtry: número de variables seleccionadas aleatoriamente en cada nodo para realizar la clasificación
- nodesize: tamaño mínimo del último nodo. Cuando más alto sea este número antes se detendrá la clasificación y los árboles tendrás "ramas" más cortas
- cutoff: porcentaje de votos que debe tener cada resultado para que una determinada observación se clasifique dentro de esa categoría

- **Lipodistrofia**

El *package* “mrl” tiene funciones que permiten determinar el valor más adecuado para ntree, mtry y nodesize. El ajuste de los parámetros en los datos procedente de pacientes varones determinó que los valores óptimos eran:

- ntree: 1215
- mtry: 8
- nodesize: 42

Con estos parámetros contruí un nuevo modelo (**Figura 33**).

Fi
“
li

El valor de AUC de este modelo es mejor que los obtenidos anteriormente con otros modelos de aprendizaje automático, pero es aún inferior al que conseguí por regresión logística. Un aspecto importante en la predicción clínica es que no todos los errores son iguales: diagnosticar como enfermo a un paciente sano no tiene el mismo coste (económico y humano) que diagnosticar como sano a un paciente enfermo. Voy a analizar el tipo de errores que se producen en el modelo de regresión logística y en este último modelo con una matriz de confusión (**Figura 34**).

Ambos modelos muestran la misma precisión, aunque los valores de sensibilidad y especificidad son distintos.. En este caso podemos ver que el modelo de regresión logística se está comportando mejor, ya que asigna a muy pocos pacientes enfermos en la categoría "sano". Modifiqué el valor de *cutoff* del modelo *random forest* para que una determinada observación necesitara el 60% de los votos para ser considerada como sana y sólo el 40% para ser considerada enferma (**Figura 35**).

Este último modelo tiene valores de AUC y precisión muy similares a los otros modelos, pero con valores más extremos tanto para sensibilidad (muy baja) como para especificidad (muy alta).

Repetí el proceso usando los datos procedentes de mujeres. El modelo óptimo tendría los siguientes parámetros:

- ntree: 1309
- mtry: 8
- nodesize: 17

Con los datos obtenidos generé un nuevo modelo (Figura 36):

Como en el caso anterior, comparé las matrices de confusión del nuevo modelo y del modelo de regresión logística (**Figura 37**).

El modelo generado por *random forest* es más preciso que el obtenido por regresión logística clásica. Modifiqué el valor de *cutoff* del modelo de aprendizaje automático (**Figura 38**).

Este último modelo mantiene un valor similar de AUC aumentando su precisión. Sin embargo, al analizar los datos se puede ver que tiende a clasificar a los pacientes como enfermos**.**

- **Sarcopenia**

Repetí el análisis anterior calculando los parámetros óptimos para el modelo de random forest para esta patología:

- ntree: 1441
- mtry: 7
- nodesize: 24

El resultado se muestra en la **Figura 39**.

A

Random forest sarcopenia hombres

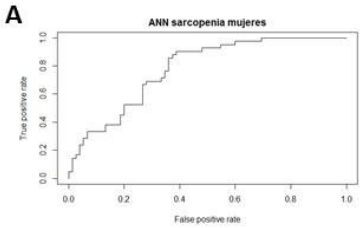
B

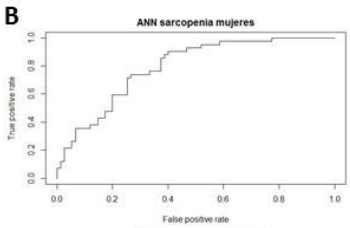
	Pacientes	
Predicción	Sano	Enfermo

El valor de AUC es superior a modelos anteriores de *random forest.* Como ha ocurrido con otros modelos, tiene el problema que diagnostica como sanos a un número relativamente algo de pacientes enfermos. La **Figura 40** muestra cómo cambios en los valores de *cutoff* cambian esta clasificación

- **Algoritmo ANN**

Cuando analicé la predicción de la sarcopenia en mujeres vi que el algoritmo que mejor se comportaba era el ANN. Por lo tanto voy a intentar optimizar este modelo para ese grupo concreto de datos. Lamentablemente no existe una función que optimice los parámetros de la función neuralnet(), pero he encontrado en la literatura que el *package* "caret" puede buscar los mejores parámetros para una red neuronal que puede implementarse con el *package* "nnet" (31). La **Figura 41** muestra los modelos generados con este procedimiento usando todas las variables o únicamente las variables seleccionadas por Lasso.





C

Predicción	Pacientes: Sano	Pacientes: Enfermo
Sano	66	28
Enfermo	9	14

Precisión: 68%
Sensibilidad: 33%
Especificidad: 88%

D

Predicción	Pacientes: Sano	Pacientes: Enfermo
Sano	70	29
Enfermo	5	13

Precisión: 71%
Sensibilidad: 31%
Especificidad: 93%

Figura 41: Modelos ANN optimizados para sarcopenia en mujeres. **A:** Curva ROC del modelo con variables seleccionadas por Lasso. **B:** Curva ROC del modelo con todas las variables. **C:** Matri de confusión del modelo con las variables seleccionadas por Lasso **D:** Matriz de confusión del modelo con todas las variables.

- **Osteoporosis y osteopenia**

Ninguno de los modelos generados para tratar de predecir la presencia de osteoporosis u osteopenia en los pacientes ha mostrado una gran eficiencia. Es posible que uno de los principales motivos sea que en este caso existían tres posibles diagnósticos en vez de dos como en las demás enfermedades. Para comprobar si esto era cierto decidí agrupar ambas patologías en el mismo grupo y calificar a las observaciones como "sano" o "enfermo". Tras esta modificación el modelo de regresión logística muestra un valor de AUC similar a las otras enfermedades tanto en hombres como en mujeres (Figura 42.A). El algoritmo ANN siguió mostrando dificultades con este grupo de datos y no fue posible generar ni siquiera el modelo más simple, pero tanto SVM como *random forest* obtuvieron resultados razonables (Figura 42.B).

El mejor resultado de entre los modelos de aprendizaje automático se logró con SVM tanto para hombres como para mujeres. Con el fin de tratar de mejorar la predicción generé diferentes modelos incrementando el valor de la variable "coste" (Figura 42.C). En ambos casos logré una mayor eficiencia, incluso superando el valor obtenido por la regresión en el caso de los datos de hombres.

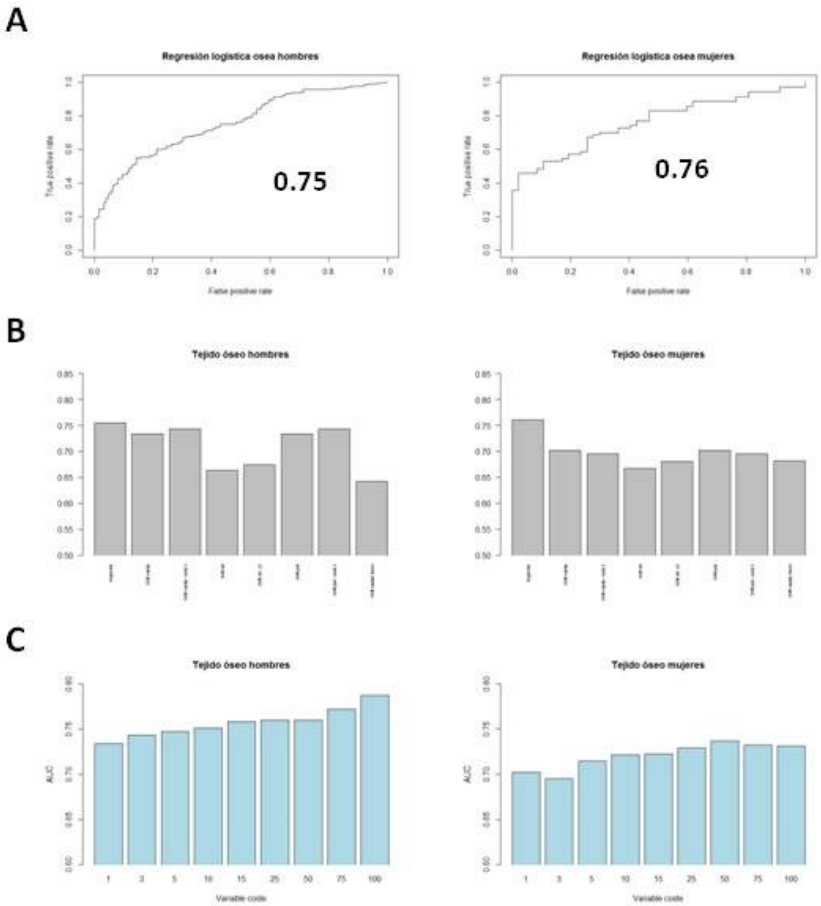


Figura 42: Análisis de las enfermedades óseas agrupadas en sólo dos diagnósticos. Las figuras de la izquierda se refieren a hombres y las de la derecha a mujeres. **A:** Curvas ROC para los modelos de regresión logística. **B:** Diagrama de barras que representa los valores de AUC obtenidos por los diferentes modelos de aprendizaje automático. **C:** Valores AUC del modelo SVM sin transformar los datos para los diferentes valores de la variable coste.

3. Conclusiones

El objetivo de este trabajo era abordar la utilización de herramientas de aprendizaje automático en el diagnóstico de enfermedades relacionadas con una distribución anormal de tejido graso, muscular y óseo en pacientes de VIH. Para su realización se han explorado cuatro algoritmos diferentes: el método de regresión logística y tres métodos de *machine learning*: *redes neuronales artificiales, support vector machines* y *random forest.*

Los resultados obtenidos en este trabajo se pueden dividir el trabajo en tres partes:

- resultados para lipodistrofia en hombres y mujerse y sarcopenia en hombres
- resultados para sarcopenia en mujeres
- resultados para enfermedades del tejido óseo (osteoporosis y osteopenia)

Los modelos generados para tratar de predecir lipodristrofia en hombres y mujeres y la sarcopenia en hombres tienen en común que el algoritmo *random forest* ha sido el modelo de aprendizaje automático que ha dado los mejores resultados. Existe una diferencia entre ellos, ya que mientras en el caso de los varones estos modelos no superaban la eficiencia del modelo clásico de regresión, en los datos procedentes de mujeres el modelo *random forest* mostraba la eficiencia más alta de todos los analizados. Estos modelos se generaron sin intentar ningún tipo de optimización más allá de la modificación en el número de árboles utilizados para generar el modelo final, algo que se ha visto que tiene una importancia marginal en el resultado. En el último capítulo se mostró que era posible mejorar ligeramente todos los modelos utilizando funciones de optimización, pero que los resultados obtenidos no estaban tan lejos de lo conseguido con los modelos que utilizaban los parámetros por defecto.

En el caso de la predicción de la sarcopenia en mujeres, el mejor resultado correspondió al modelo ANN con un solo nodo. Este algoritmo se ha mostrado, con los datos utilizados en este trabajo, como el más dependiente de los parámetros del modelo. Por ejemplo, aumentar en este caso la complejidad utilizando un modelo de 3 nodos disminuía enormemente el valor de AUC. Esto hace que haya que ser más cuidadoso al usar ANN que al usar los otros métodos de aprendizaje automático, en los que cambios en los parámetros de entrada producían efectos mínimos en la eficiencia del modelo. Al igual que en los casos anteriores, una estrategia de optimización dio lugar a modelos ligeramente más eficientes.

Finalmente, el análisis de las enfermedades del tejido óseo ha constituido uno de los mayores problemas de este trabajo. Cuando traté de clasificar a los pacientes utilizando tres posibles diagnósticos los resultados han estado muy lejos de lo que debería ser aceptable para un método de diagnóstico. Esto me llevó a cambiar la estrategia y clasificar a los pacientes únicamente como “sanos” o “enfermos”, lo que hace que se pierda parte de la utilidad diagnóstica del modelo ya que existe una diferencia importante entre padecer una u otra de las dos enfermedades del tejido óseo analizadas. Esta modificación permitió generar modelos eficientes tanto con regresión logística como con algoritmos de aprendizaje automático, destacando los resultados obtenidos con SVM. El análisis de modificaciones más extremas de la variable “coste” mostró que aumentar notablemente su valor producía mejores modelos. En el caso de los datos derivados de hombres incluso superiores al modelo de regresión. Los valores

utilizados para esta variable eran mucho mayores que los que había usado en otras patologías, por lo que cabía la posibilidad de que este tipo de modelos también funcionaran muy bien en lipodistrofia o sarcopenia. Sin embargo, un rápido análisis realizado con los datos derivados de hombres mostró que este efecto era específico de la osteoporosis y osteopenia ya que el aumento del valor de la variable "coste" no mejoraba el modelo en las otras dos enfermedades.

Una de las decisiones más difíciles a la hora de seleccionar el mejor método para diagnosticar una determinada enfermedad es encontrar un equilibrio entre la sensibilidad y la especificidad. La sensibilidad es la probabilidad de que un individuo enfermo sea efectivamente detectado como enfermo por el modelo. La especificidad es la probabilidad de que un individuo sano se clasifique como sano usando el modelo. Estos dos valores están íntimamente relacionados y modificaciones de un modelo que supongan un aumento de la sensibilidad provocan una bajada de la especificidad. Los dos tipos de errores derivados (diagnosticar como sano a un individuo enfermo o diagnosticar como enfermo a un individuo sano) no tienen las mismas consecuencias. Un error del primer tipo provocaría que una persona no reciba el tratamiento necesario para una enfermedad que padece o va a padecer, lo que dependiendo de la patología que se está tratando puede suponer un problema más o menos grave, incluso la muerte en casos extremos. Por lo tanto, en un principio podría parecer que una alta sensibilidad es la característica deseable para un buen modelo. Pero hay que tener en cuenta que un número muy alto de pacientes sanos diagnosticados como enfermos suponen no sólo un gasto excesivo para el sistema sanitario, sino un estrés y preocupaciones innecesarios para el paciente. En el último apartado hemos visto como modificaciones en el valor de *cutoff* del modelo de *random forest* provocan variaciones importantes en la sensibilidad y especificidad del modelo. Determinar en esos casos cuál es el mejor modelo es un problema complejo y que necesita de la opinión de profesionales de la salud, y por lo tanto fuera del objetivo de este estudio.

Cuando realicé la planificación de este trabajo decidí dedicar la primera mitad a probar los algoritmos seleccionados en la predicción de todas las enfermedades y durante la segunda mitad comparar los resultados y tratar de optimizar aquellos algoritmos que hubieran mostrado un mejor comportamiento. A la vista de los resultados obtenidos en el caso de las enfermedades del tejido óseo, considero que una mejor aproximación hubiera sido tratar lipodistrofia y sarcopenia por un lado, ya que tienen diagnósticos similares y usan el mismo grupo de variables, y las enfermedades óseas por otro, buscando algoritmos o *packages* específicos que estén descritos en la literatura como específicos para modelos con más de dos posibles resultados. Parece evidente que las herramientas utilizadas, a pesar de que en la literatura están descritas como válidas para modelos con más de dos diagnósticos, en la práctica no han funcionado con la eficiencia esperada.

Los resultados de este trabajo muestran que es posible generar modelos de predicción eficientes para las enfermedades analizadas utilizando herramientas de aprendizaje automático. Hay que tener en cuenta que el rango de este estudio era muy amplio: tres enfermedades diferentes analizadas en los dos sexos y utilizando tres algoritmos distintos. Considero que hay un amplio margen de mejora en estas técnicas y que un estudio más enfocado, que se centre en sólo una de las patologías o uno de los algoritmos, podría posiblemente desarrollar modelos aún más eficientes. También hay que tener en cuenta que en trabajos de este tipo analizar un solo grupo de datos no es

suficiente, sino que son necesarios varios estudios usando bases de datos independientes para contrastar los resultados. Por ejemplo, en este caso el porcentaje de pacientes con lipodistrofia dentro de la muestra era sensiblemente más alto a lo esperado en la población objetivo del estudio (los enfermos de VIH), mientras que el porcentaje de pacientes con sarcopenia era mucho más bajo, posiblemente más acorde con el dato real en dicha población. Repetir este análisis con un grupo de sujetos con una distinta proporción entre pacientes enfermos y sanos para las diferentes patologías sería útil para obtener más información acerca de cuáles son las mejores herramientas de construir modelos predictivos.

Como conclusión final, creo que el trabajo confirma una vez más que los algoritmos de aprendizaje automático son herramientas de diagnóstico muy poderosas. Algunos centros de salud ya han incluido este tipo de modelos en sus rutinas de diagnóstico y es casi seguro que muchos otros seguirán este camino a corto plazo. Los modelos generados en este trabajo utilizando datos no relacionados con la enfermedad (principalmente datos obtenido en el análisis del tejido óseo) para diagnosticar lipodistrofia y sarcopenia con una elevada precisión. Estos modelos u otros similares resultarían muy útiles para los profesionales de la salud a la hora de diagnosticar dichas enfermedades.

4. Glosario

Algoritmos de aprendizaje automático (*machine learning*): Algoritmos generados por un ordenador a través de un "aprendizaje" basado en sucesivas iteraciones de un determinado problema.

ANN: Redes neuronales artificiales. Algoritmo de aprendizaje automático que simula el funcionamiento de un cerebro biológico.

AUC: A*rea under curve*. Valor utilizado para comparar diferentes curvas ROC.

Lipodistrofia: Enfermedad provocada por la distribución anormal del tejido adiposo.

Osteoporosis y osteopenia: Enfermedades provocados por la pérdida de masa ósea.

Random forest: Algoritmo de aprendizaje automático que genera un modelo combinando un cierto número de árboles de decisión.

Regresión logística: Método clásico de clasificación en el que se genera una curva de regresión para distinguir entre distintos resultados de tipo categórico.

ROC: *Receiver operator characteristic*. Curvas utilizadas para medir la eficiencia de una predicción en biomedicina.

Sarcopenia: Enfermedad provocada por la pérdida de masa muscular.

SVM: *Support vector machines*. Algoritmo de aprendizaje automático que clasifica las observaciones según su posición en un hiperplano.

5. Bibliografía

(1): Trickey A *et al.* Survival of HIV-positive patients starting antiretroviral therapy between 1996 and 2013: a collaborative analysis of cohort studies. Lancet HIV. Agosto 2017; 4(8): e349-e356.

(2): Grunfeld C *et al*. Regional adipose tissue measured by MRI over 5 years in HIV-infected and control participants indicates persistence of HIV-associated lipoatrophy. AIDS. Julio 2010, vol. 24: 1717 – 26.

(3) van Griensven J *et al*. High prevalence of lipoatrophy among patients on stavudine-containing first-line antiretroviral therapy regimens in Rwanda. Trans R Soc Trop Med Hyg. Agosto 2007; 101(8): 793-8.

(4) Diehl LA *et al*. Prevalence of HIV-associated lipodystrophy in Brazilian outpatients: relation with metabolic syndrome and cardiovascular risk factors. Arq Bras Endocrinol Metabol. Junio 2008;52(4): 658-67.

(5) www.aidsinfonet.org (19/11/2018)

6. Wanke CA *et al*. Weight loss and wasting remain common complications in individuals infected with human immunodeficiency virus in the era of highly active antiretroviral therapy. Clin Infect Dis. Septiembre 2000; 31(3): 803-5.

(7) Knobel H *et al*. Osteopenia in HIV-infected patients: is it the disease or is it the treatment? AIDS. Abril 2001; 15(6): 807-8.

(8) Arnsten JH *et al*. Decreased bone mineral density and increased fracture risk in aging men with or at risk for HIV infection. AIDS. Marzo 2007. 21(5): 617-23.

(9) Assoumou L *et al*. Changes in bone mineral density over a 2-year period in HIV-1-infected men under combined antiretroviral therapy with osteopenia, AIDS. Septiembre 2013; 27(15): 2425-30.

(10) Brown TT, Qaqish RB. Antiretroviral therapy and the prevalence of osteopenia and osteoporosis: a meta-analytic review. AIDS. Noviembre 2006; 20(17): 2165-74.

(11) A. Schneider *et al*. Linear Regression Analysis. Dtsch Arztebl Int. Noviembre 2010; 107(44): 776–782.

(12) Jeffrey De Fauw *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature Medicine. Septiembre 2018; 24(9): 1342–50.

(13) Lynch CM *et al*. Prediction of lung cancer patient survival via supervised machine learning classification techniques. International Journal of Medical Informatics. Diciembre 2017; 108: 1-8

(14) Tsao *et al*. Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. BMC Bioinformatics. Agosto 2018; 9(Suppl 9): 283

(15) Artetxe A *et al*. Predictive models for hospital readmission risk: A systematic review of methods. Computer methods and programs in biomedicine. Octubre 2018; 164: 49-64

(16) Bisaso *et al*. A survey of machine learning applications in HIV clinical research and care. Computers in Biology and Medicine. Diciembre 2017; 91: 366-371

(17) Hoo ZH *et al*. What is an ROC curve? Emerg Med J. Junio 2017;34(6): 357-9.

(18) https://cran.r-project.org/web/packages/ROCR/ROCR.pdf (21/10/18)

(19) Tibshirani, R. Regression Shrinkage and Selection via the lasso. Journal of the Royal Statistical Society. Series B (methodological). 1996; volumen 58(1).

(20) https://cran.r-project.org/web/packages/glmnet/glmnet.pdf (21/10/18)

(21) https://cran.r-project.org/web/packages/nnet/nnet.pdf (21/10/18)

(22) https://cran.r-project.org/web/packages/caret/caret.pdf (25/10/18)

(23) Abiodun OI *et al*. State-of-the-art in artificial neural network applications: A survey. Heliyon. Noviembre 2018; 4(11):e00938.

(24) https://cran.r-project.org/web/packages/neuralnet/neuralnet.pdf (2/11/18)

(25) Wang MFZ, Fernandez-Gonzalez R (Machine-)Learning to analyze in vivo microscopy: Support vector machines. Biochim Biophys Acta Proteins Proteom. Noviembre 17;1865(11 Pt B): 1719-1727.

(26) https://cran.r-project.org/web/packages/kernlab/kernlab.pdf (16/11/18)

(27) https://cran.r-project.org/web/packages/e1071/e1071.pdf (21/12/18)

(28) Sarica A *et al*. Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review. Front Aging Neurosci. Octubre 2017; 9: 329.

(29)https://cran.r-project.org/web/packages/randomForest/randomForest.pdf (20/11/18)

(30) https://cran.r-project.org/web/packages/mlr/index.html (21/12/18)

(31) https://cran.r-project.org/web/packages/nnet/nnet.pdf (23/12/18)

6. Anexos