TÉCNICAS DE *MACHINE LEARNING* Y DESARROLLO DE MODELOS PREDICTIVOS APLICADOS A LA ANTROPOLOGÍA FORENSE

Noemi Álvarez Fernández
Master en Bioinformática y Bioestadística
Área 29: Antropología Biológica

Xavier Jordana Comin
David Merino Arranz

5 de junio del 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	Técnicas de *Machine Learning* y desarrollo de modelos predictivos aplicados a la Antropología Forense.
Nombre del autor:	Noemi Álvarez Fernández
Nombre del consultor/a:	Xavier Jordana Comin
Nombre del PRA:	David Merino Arranz
Fecha de entrega (mm/aaaa):	5 de junio del 2018
Titulación:	Máster de Bioinformática y Bioestadística
Área del Trabajo Final:	Área 29: Antropología Biológica
Idioma del trabajo:	Castellano
Palabras clave:	Redes neuronales artificiales, *Random Forest*, aprendizaje supervisado

Resumen del Trabajo

En este trabajo se estudia la aplicación de las técnicas de Machine Learning en el desarrollo de modelos de clasificación para la estimación del sexo de restos óseos humanos, dentro de un contexto forense y bioarqueológico. Estas técnicas permiten desarrollar algoritmos que puede aprender y esquematizar propiedades y patrones estructurales subyacentes de los datos, pudiendo utilizarse está información para entender y predecir fenómenos específicos. Por tanto, dada la importancia que tiene una estimación fiable del sexo y la utilidad de las técnicas de Machine Learning en el desarrollo de modelos de clasificación para la predicción de variables categóricas. Lo que se propone en este trabajo es la aplicación de algoritmos estadísticos clásicos y de Machine Learning para la construcción de modelos predictivos para la estimación del sexo a partir de estudios métricos de estos restos. Para ello, en primer lugar se realizó una revisión bibliográfica de las técnicas de Machine Learning y de estadística clásica utilizadas en el área de la antropología forense. Obteniéndose un texto donde se definen y comparan estos métodos. Una vez hecho hecho se evaluaron y se escogieron los que podían adaptarse mejor a nuestro problema. Por último, se obtuvieron tres modelos de clasificación con los métodos de regresión logística, redes neuronales artificiales y Random Forest, utilizando el software R. Los resultados obtenidos fueron una serie de potentes modelos predictivos. Por lo que se puede decir que las técnicas de Machine Learning son una prometedora alternativa a los métodos clásicos de clasificación.

Abstract

This paper studies the application of Machine Learning techniques in the development of models for estimating the sex of human beings, within a forensic and bioarchaeological context. These techniques allow the development of algorithms that can learn and schematize properties and underlying structural patterns of the data, and can be used for specific information and predict specific phenomena. Therefore, given the importance of a reliable estimation of sex and the usefulness of Machine Learning techniques in the development of classification models for the prediction of categorical variables. What is offered in this work is the application of classical algorithms and Machine Learning for the construction of predictive models for sex estimation from metric studies of these remains. For this, in the first place, a bibliographic review of the Machine Learning and Statistics Classics techniques in the area of forensic anthropology was carried out. Obtaining a text where these methods are defined and compared. Once done, we evaluated and chose those who stayed better to our problem. Finally, three classification models were obtained with the methods of logistic regression, artificial neural networks and Random Forest, using the R software. The results were a series of powerful predictive models. Therefore, it can be said that Machine Learning techniques are a primising alternative to classical classification methods.

Índice

1 Introducción **1**
1.1 Contexto y justificación del Trabajo 1
1.2 Objetivos del Trabajo 3
1.3 Enfoque y método seguido 4
1.4 Planificación del Trabajo 5
1.5 Breve sumario de productos obtenidos 9
1.6 Breve descripción de los otros capítulos de la memoria 10

2 Breve introducción al *Machine learning* **11**

3 Una perspectiva general de las técnicas **13**

4 Técnicas seleccionadas **19**

5 Selección de los datos **20**

6 Obtención de los modelos **21**
6.1 Análisis preliminar de los datos 21
6.2 El *software WEKA* 24
6.3 Regresión logística 25
6.4 Redes neuronales artificiales 29
6.5 *Random Forest* 33

7 Discusión **37**

8 Conclusiones **39**

9 Glosario **40**

10 Bibliografía **42**

11 Anexos **45**
11.1 ANEXO I: Código utilizado para los análisis en R 45
11.2 ANEXO II: Descripción de las variables 54

Índice de cuadros

1. Tabla de tareas 5
2. Técnicas - Referencias. 14
3. NAs por lateralidad de las medidas del fémur. 20
4. p-valores: modificación de Lillefors del test de Kolmogorov-Smirnov. . . 23
5. Matrices de confusión y estimadores de la bondad del ajuste del modelo LR con WEKA. 24
6. Matrices de confusión y estimadores de la bondad del ajuste del modelo ANN con WEKA. 24
7. Matrices de confusión y estimadores de la bondad del ajuste del modelo RF con WEKA. 24
8. Coeficientes, p-valores y VIF de cada variable de la LR. 25
9. Distancias de Cook. 25
10. Coeficientes, p-valores y VIF de cada variable de la LR modificada. . . . 27
11. Matrices de confusión y estimadores de la bondad del ajuste del modelo LR. 28
12. Matrices de confusión y estimadores de la bondad del ajuste del modelo ANN. 32
13. Matrices de confusión y estimadores de la bondad del ajuste del modelo RF. 36
14. Descripción de las variables. 55

Índice de figuras

1. Diagrama de Gantt 6
2. Proporciones de NAs para cada variable. 21
3. Histogramas y curvas de densidad de las medidas osteométricas del fémur. 22
4. Representación gráfica de los valores atípicos. 26
5. Curva ROC de LR. 27
6. Representación gráfica del modelo ANN. 29
7. Representación gráfica simplificada del modelo ANN. 30
8. Curva ROC de ANN. 31
9. Evolución de la tasa de error para el modelo RF. 33
10. Representación gráfica del modelo RF. 34
11. Importancia de las variables del modelo RF. 35
12. Curva ROC de RF. 36

1 Introducción

1.1 Contexto y justificación del Trabajo

En este trabajo se tratan las aplicaciones de las técnicas de *Machine Learning (ML)* en el desarrollo de modelos predictivos para la estimación del sexo de restos esqueléticos humanos dentro de un contexto forense y bioarquelógico.

Las técnicas de *Machine Learning* permiten desarrollar algoritmos que pueden aprender y esquematizar propiedades y patrones estructurales subyacentes de los datos. Esta información puede utilizarse posteriormente para entender y predecir un fenómeno específico [26].

La antropología forense representa la aplicación del conocimiento y las técnicas de la antropología física a problemas medicolegales. Sus objetivos son identificar restos humanos y determinar que les sucedió, especialmente ante casos en los que hay evidencias de una agresión. Habitualmente, el material examinado consiste en restos esqueléticos completos o prácticamente completos [33]. Sin emabargo, las técnicas que se tratarán en este trabajo cobran su mayor importancia cuando hablamos de restos no identificados. Este tipo de restos suelen proceder de desastres masivos y de explosiones de alta intensidad, donde los restos están severamente fragmentados, quemados y/o mezclados [1]. La antropología forense proporciona técnicas y experiencia en la interpretación de estos restos, así como una perspectiva comparada de la población mundial. Esta perspectiva es necesaria para evaluar adecuadamente las probabilidades involucradas y para evitar errores de interpretación [33].

La determinación del sexo en contextos forenses y bioarqueológicos es extremadamente importante ya que muchas de las estimaciones posteriores dependen de este parámetro [25], por lo que es necesario contar con buenos modelos. Si bien el objetivo de la determinación del sexo difiere entre los estudios paleontológicos y las investigaciones policiales, ambos se enfrentan con la misma biología, los mismos límites metodológicos y la necesidad de un alto nivel de fiabilidad. Esta fiabilidad y exactitud de la evaluación del sexo a partir de los restos esqueléticos depende de la región anatómica disponible [4], ya que la mayor limitación a la hora de evaluar cualquier parámetro del perfil biológico es el grado de integridad y preservación de los restos[[25].

Los análisis forenses modernos utilizan herramientas como la morfométrica geométrica y las imágenes médicas para adquirir, analizar y cuantificar la variación relacionada con el sexo en las estructuras esqueléticas humanas. Sin embargo, el efecto de la variación poblacional y los cambios seculares requieren de la investigación y actualización de los perfiles biológicos estándares de forma continua, para tener métodos que reflejen y expliquen la biología de la población a nivel regional y temporal [25].

Para analizar la información que aportan estas técnicas sobre los restos analizados se utilizan diferentes herramientas estadísticas. Aquí es donde cobran importancia las técnicas de *Machine Lerarning*, ya que en los últimos años se ha puesto de manifiesto que estas técnicas ayudan a mejorar los métodos clásicos discriminantes [6,8,9,14,15,19,23,25], tanto en el ámbito de la antropología forense como en otras disciplinas.

Podemos definir las técnicas estadísticas y de ML como un conjunto de herramientas de análisis de datos. Si tuviéramos que señalar una diferencia entre ellas, podríamos decir que la estadística se preocupa más por probar hipótesis, mientras que el ML se centra en la formulación de procesos de generalización como una búsqueda a través de posibles hipótesis [36].

Algunas herramientas aportadas por los métodos clásicos discriminantes son: las funciones discriminates, los modelos de regresión logística o la curva característica operativa del receptor (ROC). La diferencia esencial entre los dos primeros métodos es la forma en la que se ajusta la función lineal a los *training data* [9]. Pero a pesar de la utilidad de estas técnicas estadísticas, presentan algunas limitaciones. En el caso de los análisis de funciones discriminantes los datos deben cumplir algunas suposiciones llamadas normalidad multivariante y homocedasticidad de la matriz de varianzas y covarianzas. Cuando estas dos asunciones no se cumplen, algo habitual, es más adecuado el uso de los modelos de regresión logística. Además, existen diferentes rutinas matemáticas de optimización que se pueden utilizar para su ajuste [2,5,9,26].

Por otra parte, se ha demostrado que estos métodos pueden mejorarse con las técnicas de *Machine Learning* [25], algunas de ellas son [14,36]: las redes neuronales artificiales (*atificial neural networks(ANN)*), los árboles de clasificación y regresión (*classification and regression trees*) o los modelos de conjunto (*ensemble models*).

La mayor diferencia entre los métodos de *Machine Learning* y las técnicas de aprendizaje estadístico clásico es que los algoritmos de ML llevan a cabo un uso más exhaustivo de los *training data*. Generalmente, para aplicar los métodos de ML no es necesario que los datos cumplan determinadas asunciones estadísticas como la normalidad. Sin embargo, esta característica hace que sea más necesario un proceso de captación y validación de los datos más riguroso para evitar ajustes insuficientes o excesivos [36].

El desarrollo de nuevos modelos en este campo es relevante tanto para los expertos encargados de realizar los diagnósticos como para la ciudadanía general. Dado que al mejorarlos, estamos generando estudios más fiables y de mejor calidad que ayuden a avanzar en el ámbito científico. Pero también estamos prestando un servicio al conjunto de la sociedad, ya que su aplicación forense puede ser de ayuda tanto en casos de desapariciones o asesinatos, así como en casos de catástrofes naturales o fosas comunes dentro de contextos bélicos. En el caso de estos dos últimos supuestos es quizá más importante disponer de buenos modelos ya que la identificación es mucho más compleja al no disponer de todas las partes. De forma más indirecta su aplicación bioarqueológica puede ayudarnos a comprender mejor lo que pasó a lo largo de la historia, ayudando a mejorar ciertas situaciones del presente.

Dada la importancia que tiene una estimación fiable de esta variable y la utilidad de las técnicas de ML en el desarrollo de modelos predictivos, lo que se propone en este trabajo es la aplicación de algoritmos estadísticos y de aprendizaje automático para la construcción de modelos predictivos para la estimación del sexo a partir de estudios métricos de estos restos.

1.2 Objetivos del Trabajo

A continuación se detallan los objetivos generales (OG) y específicos (OE) planteados para este trabajo:

OG.1. Revisión bibliográfica sobre las técnicas de *Machine Learning* aplicadas en antropología forense y su contraste con las técnicas estadística clásicas utilizadas habitualmente en este campo.

Para ello fue necesaria:

OE.1.1. La consulta de diferentes fuentes de información.

OE.1.2. La síntesis de un texto en el que:

OE.1.2.1. Se definieron las técnicas estadísticas clásicas y de ML en el marco de la antropología forense.

OE.1.2.2. Se compararon paralelamente las dos clases de técnicas definidas.

OE.1.2.3. Se hizo una evaluación de las que se adaptan mejor a nuestro problema.

OG.2. Obtención de los modelos.

Para ello:

OE.2.1. Se obtuvieron los datos otseométricos necesarios de la base de datos *online*: *The Terry Collection Postcranial Osteometric Database*.

OE.2.2. Se obtuvieron los modelos empleando un *software* adecuado.

1.3 Enfoque y método seguido

Para la consecución del O.G.1. fue necesario:

OE.1.1. Consultar diferentes fuentes de información, tales como artículos y libros relacionados con la temática planteada. Utilizando diferentes buscadores como *Scopus* o el catálogo de la Biblioteca de la UOC. En este punto se plantearon dos opciones:

1. Hacer una revisión de todas las técnicas de ML que se podrían aplicar para mejorar las técnicas estadísticas clásicas que se aplican habitualmente en antropología forense.

2. Hacer una revisión de las técnicas de ML que se han aplicado en este campo con la finalidad de mejorar las técnicas estadísticas clásicas.

Este trabajo se abordó optando por la segunda opción.

O.E.1.2. Una vez que se dispuso de la información suficiente se abordó la elaboración del texto en el que se (O.E.1.2.1.) describieron y (O.E.1.2.2.) compararon brevemente estas técnicas. Sobre todo prestando atención en las mejoras que pueden aportar las técnicas de ML combinadas con la estadística clásica. A continuación (O.E.1.2.3.) se hizo una valoración de las técnicas que se adaptan mejor a nuestro problema y se justificó su elección. Este punto es especialmente importante debido a que condiciona el segundo objetivo general de este trabajo.

Para la obtención de los modelos (O.G.2.) fue necesario:

O.E.2.1. Obtener los datos osteométricos necesarios. Esto podría haberse hecho de dos formas:

1. Haciendo las mediciones nosotros mismo.

2. Utilizando la información disponible en una base de datos *online* como: *The Terry Collection Postcranial Osteometric Database*.

En este caso se escogió la segunda opción.

O.E.2.2. Obtención de los modelos: utilizando el *software R* [27] y *WEKA* [12].

1.4 Planificación del Trabajo

Los recursos informáticos utilizados para el desarrollo del trabajo fueron los siguientes:

- El *software* libre LaTeX - en su versión para el sistema operativo Mac OS X, MacTeX - para la redacción de los documentos. En concreto para el diseño del gráfico de *Gantt* se utilizó el paquete `pgfgantt` [32].
- El leguaje de programación *R version 3.4.1 (2017-06-30) – "Single Candle"* [27]. Utilizando el entorno de desarrollo integrado *RStudio*.
- El *software WEKA (Waaikato Enviroment for Knowledge Analysis)* [12].

En el cuadro.1 se detallan las tareas realizadas relacionándolas con los diferentes objetivos específicos. Además, se especifican las fechas de inicio y fin previstas para cada una de ellas, esta cronología se puede consultar de forma gráfica en la figura.1.

PEC2	DESARROLLO DEL TRABAJO (Fase 1)		20/03/18	23/04/18
Tarea 1	Consulta y selección de fuentes de información	OE.1.1.	20/03/18	26/03/18
Tarea 2	Definición de las técnicas de estadística clásica y ML	OE.1.2.1.	27/03/18	10/04/18
Tarea 3	Comparación de las técnicas de estadística clásica y ML	OE.1.2.2.	11/04/18	23/04/18
PEC3	DESARROLLO DEL TRABAJO (Fase 2)		24/04/18	21/05/18
Tarea 4	Evaluación y selección de las técnicas ML	OE.1.2.3.	24/04/18	30/04/18
Tarea 5	Obtención de los datos	OE.2.1.	01/05/18	07/05/18
Tarea 6	Obtención de los modelos	OE.2.2.	08/05/18	21/05/18
PEC4	REDACCIÓN DE LA MEMORIA		22/05/18	05/06/18
PEC5a	ELABORACIÓN DE LA PRESENTACIÓN		06/06/18	13/06/18
PEC5b	DEFENSA PÚBLICA		14/06/18	25/06/18

Cuadro 1: Tabla de tareas

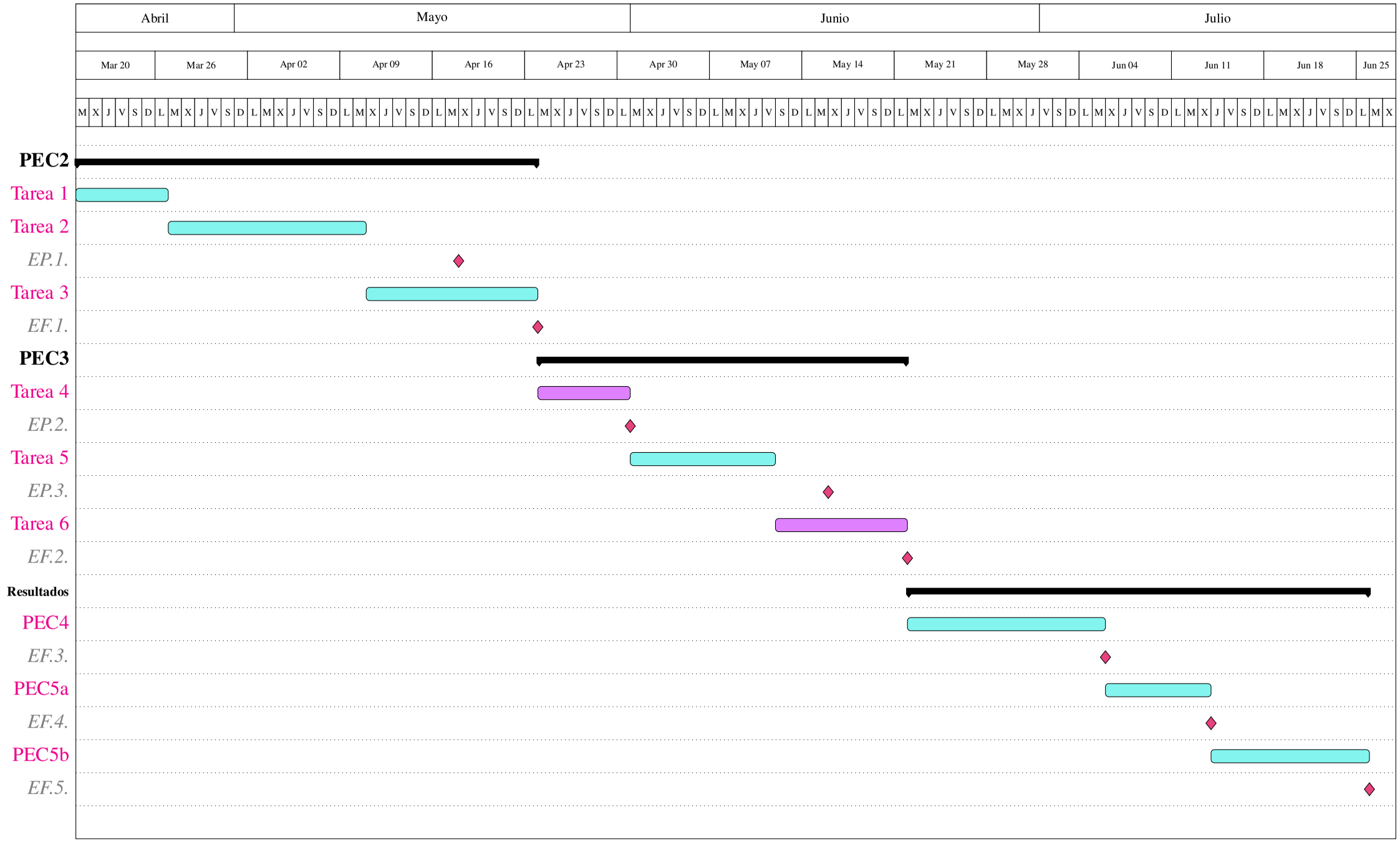

Figura 1: Diagrama de Gantt

E.F.: Entregas Finales. Corresponden con las entregas de las PECs.
E.P.: Entregas Preliminares.

Para la elaboración del diagrama de Gantt anterior (figura.1) se hizo la siguiente evaluación:

1. En primer lugar se tuvo en cuenta el tiempo establecido para el desarrollo de este proyecto, desde el 20/03/18 hasta el 25/06/18.

2. En segundo lugar se definieron los evento globales: PEC2, PEC3 y Resultados. Para establecer la cronología de estos eventos se utilizaron las fechas establecidas por el calendario de entregas de la UOC.

3. Los eventos anteriores sirven para englobar las tareas a realizar para poder finalizar el TFM. Estas tareas se describen en el cuadro.1. Para establecer su cronología se siguieron los siguientes puntos:

   a) La cronología de los tres eventos generales.

   b) La relevancia individual de los objetivos.

   c) El nivel de dificultad de cada tarea en relación con el nivel previo de conocimiento sobre los aspectos necesarios para su consecución.

   Teniendo en cuenta los puntos anteriores:

   3.1. Dentro del evento **PEC2** se englobaron las siguientes tareas:

      3.1.1. **Tarea 1**: consulta y selección de fuentes de información. Esta fue una tarea fundamental ya que sin ella no habría sido posible avanzar con el proyecto. Pero a pesar de su importancia no se consideró una tarea complicada, debido a que ya se hiciera una búsqueda preliminar de información para la elaboración de la PEC0 y PEC1 y a que el nivel en esta competencia personal no suponía un problema. Teniendo en cuenta todo esto y que en caso de ser necesario, mientras se desarrollaran las dos tareas siguientes, podría haberse ampliado la información recopilada: se le asignó una semana de duración.

      3.1.2. **Tarea 2**: definición de las técnicas de estadística clásica y ML. A pesar de no haberse considerado una de las tareas críticas del proyecto ni una de las más relevantes, ya que su importancia radica en que ayudará al lector a la compresión del trabajo y a que este sea más completo, pero no fue esencial para su desarrollo. En este caso se asignaron dos semanas, debido a la necesidad de comprensión de las técnicas para poder redactar una explicación sintética, completa y correcta.

      3.1.3. **Tarea 3**: comparación de las técnicas de estadística clásica y ML. Esta tarea fue una de las más importantes del trabajo ya que es necesaria para la consecución del OG.1. No se consideró una tarea crítica debido a que no presentaba una dificultad muy elevada. Sin embargo, debido a su importancia se planificó la entrega de un borrador (EP.1.) para poder realizar las modificaciones necesarias antes de la EF.1. Además, esta tarea condicionaba el correcto desarrollo de la siguiente.

   3.2. Dentro del evento **PEC3** se englobaron las siguientes tareas:

      3.2.1. **Tarea 4**: evaluación y selección de las técnicas de ML. Esta fue una de las tareas consideradas como críticas debido a que una mala elección podría haber supoesto un obstáculo a la hora de abordar la última tarea. A pesar

de esta importancia solo se le asignó una semana ya que se consideró que tras todo el trabajo previo debería ser suficiente. Por otra parte se planificó una entrega parcial (EP.2.) al final de esta tarea para asegurar que las técnicas escogidas fueran adecuadas.

3.2.2. **Tarea 5**: obtención de los datos. La relevancia de esta tarea está en que sin ella la implementación del ejemplo no habría sido posible, pero no se consideró crítica ya que su dificultad no es alta. Debido a esto se planificó una semana para su realización. Sin embargo, durante su desarrollo se produjo un incidente que obligó a ampliarla durante cuatro días, siendo necesario recortarlos de la siguiente tarea.

3.2.3. **Tarea 6**: obtención del los modelos. Esta fue una de las tareas más importantes del trabajo ya que supuso la consecución del OG.2. Además, fue una de las tareas críticas ya que los conocimientos previos eran escasos por lo que condicionaban su correcto desarrollo. Teniendo en cuenta estos factores se planificó una entrega parcial (EP.3) y se le asignaron dos semanas. Por otra parte en caso de que en alguna de las tareas anteriores se finalizasen antes de lo previsto ese tiempo *"sobrante"* se reasignaría a esta tarea. Desafortunadamente, no hubo exceso de tiempo en ninguna de las tareas previas, si no que como ya se comento en la descripción de la Tarea 5 se produjo un retraso de cuatro días obligando a una mayor inversión de horas durante la última semana de trabajo de esta tarea.

3.3. El evento **Resultados** engloba las tres PECs restantes, que serían los *"productos"* derivados del proyecto:

3.3.1. La redacción de la **memoria**.

3.3.2. La elaboración de la **presentación**.

3.3.3. La **defensa pública** del trabajo. Este punto no es exactamente una entrega, ya que no se corresponde con un documento de ningún tipo - texto o audiovisual - si no que consistirá en la defensa pública del proyecto ante un tribunal.

Al final de cada una de estas tareas se planificó una entrega final, correspondiente a las entregas de las PECs planificadas por la UOC.

1.5 Breve sumario de productos obtenidos

Como resultado del trabajo se obtuvieron:

1. Una revisión bibliográfica sobre las técnicas de *Machine learning*.
2. Tres modelos predictivos para la estimación del sexo a partir de las medidas del fémur izquierdo. Utilizando las técnicas de: LR, ANN y RF.
3. Un conjunto de *scripts* con el código necesario para obtener los tres modelos de predicción.

1.6 Breve descripción de los otros capítulos de la memoria

En el capítulo 2 se hace una breve introducción al *Machine Learning* en la que se habla de su historia y se describen los diferentes tipos de algoritmos: aprendizaje supervisado, aprendizaje semi-supervisado, aprendizaje no supervisado, aprendizaje reforzado. Por su parte, en el capítulo 3 se describen y comparan brevemente las diferentes técnicas de estadística clásica y *Machine Learning* utilizadas en el ámbito de la antropología forense. Para dar paso al capítulo 4 donde se seleccionan los métodos a utilizar para la implementación de los modelos, y se explica brevemente porque se tomó esta decisión. En el capítulo 5 se seleccionan los datos para la obtención de los modelos que se explican y desarrollan en el capítulo 6, para discutir sus resultados en el capítulo 7. Mientras que en el capítulo 8 se describen las conclusiones a las que se llegó tras la finalización del trabajo, así como el grado de consecución de los objetivos planteados y el ajuste del seguimiento de la planificación y la metodología propuestas al inicio de este. Para finalizar, se exponen brevemente posibles lineas de trabajo futuras que no se han podido incluir en este TFM. En el capítulo 9 se incluyen las definiciones de los término y acrónimos más relevantes utilizados en la memoria. Por último, en los anexos encontramos una descripción del código utilizado y la descripción de las variables utilizadas en los modelos.

2 Breve introducción al Machine learning

Antes de empezar a definir las diferentes técnicas vamos a hacer una breve introducción sobre el *Machine learning*. Como ya se ha comentado anteriormente, el ML es un método de análisis de datos que automatiza la construcción de modelos analíticos. Es una rama de la inteligencia artificial basada en la idea de que los sistemas pueden aprender de los datos, identificar patrones y tomar decisiones con la mínima intervención humana. Estos métodos nacieron de su aplicación al reconocimiento de patrones y de la idea de que las máquinas pueden aprender sin estar programadas para realizar tareas específicas. La aplicación de estas técnicas en diferentes disciplinas científicas surge del interés de algunos investigadores por ver si las máquinas podrían aprender de los datos que estos manejaban en sus investigaciones. Uno de los aspectos más importantes del ML es el hecho de que son métodos interactivos, es decir, a medida que los modelos se exponen a nuevos datos, pueden adaptarse de forma independiente. Aprenden de cómputos anteriores para producir decisiones y resultados confiables y repetibles [18]. Es una ciencia que no es nueva pero que ha ganado un nuevo impulso debido a los avances en el área de la informática, haciendo que cada vez más científicos muestren interés por utilizarlos y entender como funcionan y como podrían aplicarse a sus campos de investigación. Esta curiosidad llegó al ámbito de la antropología forense haciendo que, en los últimos años, se hayan aplicado estas herramientas con diferentes niveles de éxito a distintos problemas que requieren del uso de modelos de clasificación, como es el caso de la estimación del sexo a partir de restos óseos.

Dentro de los diferentes tipos de algoritmos de ML se puede hacer la siguiente clasificación [18]:

1. Aprendizaje supervisado: este tipo de algoritmos hacen predicciones en base a un conjunto de ejemplos. Con estas técnicas tenemos una variable de entrada que consta de un conjunto de *training data* etiquetados y una variable de salida deseada. Para analizar los *training data* utilizan un algoritmo para aprender la función que asigna la entrada a la salida. Esta función inferida mapea ejemplos nuevos y desconocidos generalizando a partir de los *training data* para anticipar resultados en situaciones no vistas. A su vez, dentro de esta clase de algoritmos se puede hacer la siguiente clasificación:

    1.1. Algoritmos de clasificación: cuando los datos se utilizan para predecir una variable categórica, el aprendizaje supervisado también se denomina de clasificación. Este es el caso cuando se asigna una etiqueta o indicador. Cuando solo hay dos etiquetas se llama clasificación binaria y cuando hay más de dos los problemas se llaman clasificación de clase múltiple.

    1.2. Algoritmos de regresión: al predecir valore continuos, los problemas se convierten en un problema de regresión.

    1.3. Algoritmos de pronóstico: este es el proceso de hacer predicciones sobre el futuro basadas en los datos pasados y presentes. Es el más comúnmente utilizado para analizar tendencias.

2. Aprendizaje semi-supervisado: el desafío que presentan las técnicas supervisadas es que los datos de etiquetado pueden ser costosos y llevar mucho tiempo obtenerlos.

Sin embargo, el aprendizaje semi-supervisado permite la utilización ejemplos sin etiqueta con una pequeña cantidad de datos etiquetados para mejorar la precisión de aprendizaje.

3. Aprendizaje no supervisado: en este caso la máquina se presenta con datos sin etiquetar. Se le pide que descubra los patrones intrínsecos que subyacen a los datos, como una estructura de agrupación, una variedad de baja dimensión o un árbol y un gráfico dispersos. Dentro de este tipo de algoritmos podemos hacer la siguiente clasificación:
   3.1. Agrupamiento: consisten en la agrupación de un conjunto de ejemplos de datos para que los ejemplos en un grupo sean más similares - según unos determinados criterios - que los de otros grupos. Esto a menudo se usa para segmentar el conjunto de datos completo en varios grupos. El análisis se puede realizar en cada grupo para ayudar a los usuarios a encontrar patrones intrínsecos.
   3.2. Reducción de dimensión: consiste en la reducción del número de variables consideradas. En muchas aplicaciones, los datos brutos tienen características dimensionales muy altas y algunas funciones son redundantes o irrelevantes para la tarea. Reducir la dimensionalidad ayuda a encontrar la verdadera relación latente.

4. Aprendizaje reforzado: estas técnicas analizan y optimizan el comportamiento de un agente en función de los comentarios del entorno. Las máquinas prueban diferentes escenarios para descubrir qué acciones producen la mayor recompensa, en lugar de que se les diga qué acciones tomar. La recompensa de prueba y error y retraso distingue el aprendizaje de refuerzo de otras técnicas.

A la hora de elegir uno de estos algoritmos se debe tener en cuenta la precisión, el tiempo de capacitación y la facilidad de uso [18].

3 *Una perspectiva general de las técnicas*

Desde finales de la década de 1960 la predicción del sexo dentro de un contexto forense ha generado una extensa literatura [8], donde predominan los análisis discriminantes para el desarrollo de modelos predictivos. Llegando al punto de que casi todos los huesos del esqueleto humano han sido clasificados utilizando esta técnica estadística.

2.1. Análisis lineal discriminante

El análisis lineal discriminante (LDA) es un método estadístico descriptivo multivariante ampliamente utilizado en el análisis de resultados de variables categóricas - como el sexo - para el desarrollo de modelos de clasificación lineal. El LDA puede utilizarse para determinar qué variable discrimina entre dos o más clases, así como para derivar un modelo de clasificación para predecir la pertenencia a grupos de nuevas observaciones. El caso más simple de aplicación del LDA es el de dos grupos, aunque hay que tener en cuenta que puede aplicarse con un número mayor de estos. En el caso de dos grupos, para discriminar entre ellos se puede utilizar una función discriminante lineal (LDF) que pasa a través de los centroides de los grupos. Es importante resaltar que el modelo estándar de LDA asume que los datos presentan una distribución multivariada normal y una matriz de covarianzas común [26], ya que estas asunciones pueden suponer un problema, dado que en la mayoría de los casos no se cumplen.

Es interesante mencionar aquí a FORDISC, uno de los *softwares* más conocidos y utilizados en antropología forense. Esta herramienta permite realizar estimaciones del sexo, entre otras variables, utilizando el LDA [18]. Las estimaciones las hace empleando una base de datos de referencia, basada en el *American Forensic Data Bank (FDB)* y la *The Terry and Hamann Todd Collection*. Es necesario especificar que la fiabilidad de las estimaciones para muestras de individuos estadounidense es muy alta [24]. Mientras que si se utiliza con otras poblaciones, como pueden ser la europea o la asiática [7,15,24,30], esta fiabilidad queda entredicha. Esto se debe a la falta de datos para estas poblaciones en su base de datos. Este problema es extendible a cualquier análisis discriminante debido a que se necesita una población de referencia similar a la de nuestra muestra para realizar estimaiciones fiables [24].

Los antropólogos físicos utilizan comúnmente los análisis lineares discriminantes (LDA) para predecir y clasificar, pero raramente discuten sus limitaciones. En la mayoría de las aplicaciones del LDA, los investigadores suponen táctica o explícitamente que los datos son multivariantes normales y tienen matrices homogéneas de covarianza de grupo. Además, las implementaciones de LDA utilizadas con más frecuencia - es decir, SAS, SPSS y SYSTAT - están diseñadas solo para *suites* completas de mediciones en cada individuo, y eliminan cualquier caso en el que falte una o más variables de predicción. Para superar esto, algunos investigadores han utilizado la regresión múltiple para predecir los valores perdidos, o los han sustituido por la media [9]. Estos enfoques fueron desacreditados por Schafer (1997) [30] y Schimert et al. (2000) [31], ya que demostraron que el efecto de todos estos enfoques es la distorsión de la información de los datos restantes, dando lugar a respuestas potencialmente engañosas, o creando situaciones en las que es difícil discernir si los métodos funcionan o fallan [9]. Las opciones paramétricas para abordar este problema son limitadas, tal y como se detalla en el artículo de Feldesman (2002):

1. Analizar solo los casos con conjuntos de datos completos. Con el problema de que se reduce el tamaño comparativo de la muestra e introduce sesgos sutiles.

2. Formar subconjuntos de casos múltiples (y a veces superpuestos), cada uno con series de medidas ligeramente diferentes analizadas secuencialmente. En este caso se reduce el número de variables en un solo análisis y puede inflar los nivles de significación cuando hay comparaciones múltiples.

3. Encontrar, comprender y aplicar un algoritmo de imputación de datos diseñado tanto para "completar" los valores perdidos como para estimar los límites de confianza en las estimaciones estadísticas resultantes. Esta opción es problemática a menos que el número de valores perdidos sea relativamente pequeño y si no se tiene cuidado para evitar las singularidades de la matriz que resultan del sesgo introducido al "completar" los valores perdidos.

Además, la mayoría de los resultados obtenidos con estos modelos de predicción indican que la relación entre la probabilidad de que un individuo pertenezca a un determinado sexo y las variables explicativas no es lineal. Por esta razón algunos autores han desarrollado modelos de determinación del sexo utilizando métodos de clasificaición no lineales para eludir los límites impuestos por este método [8]. Una alternativa no paramétrica sería la regresión logística. Sin embargo, las técnicas de *Machine Learning* nos brindan otras opciones. Algunos de estos "novedosos" métodos de clasificación existen desde la década de 1960 [9] pero su uso se veía limitado por su dependencia de la velocidad y potencia de computación [14]. En este trabajo no se van a tratar todas las alternativas posibles, si no que nos centraremos en las que se han aplicado en el ámbito de la antropología forense. Entre estos métodos se incluyen técnicas como el análisis discriminante cuadrático (QDA), las redes neuronales artificiales (ANN), la partición recursiva binaria - más comúnmente conocida como "árboles de clasificación" (DT) -, los modelos de *random forest* (RFM) o las máquinas de vectores de soporte (SVM) [23]. En el cuadro.2 se muestra un pequeño resumen de algunos trabajos en los que se han aplicado estas técnicas. Estos métodos no trabajan directamente con parámetros de grupo como pueden ser la media o la desviación estándar. Si no que ajustan miles de puntos de corte aleatorios en la muestra para encontrar las formas más precisas de agrupar los individuos. Además, hay muchas transformaciones de datos posibles que se pueden aplicar a los datos originales y que pueden mejorar la precisión de la clasificación [13].

TÉCNICA	ARTÍCULO
TQDA	Moss (2011)
ANN	Du Jardin (2009)
DT	Feldesman (2011)
RFM	Hefner (2014)
SVM	Moss (2011)

Cuadro 2: Técnicas - Referencias.

2.2. Regresión logística

Una alternativa no paramétrica al LDA, utilizada frecuentemente, es la regresión logística (LR). Este método también se clasifica dentro de los de tipo estadístico multivariante y al igual que el LDA es adecuado para el desarrollo de modelos de clasificación lineal. También hay que comentar que se puede aplicar a casos con más de dos categorías. La diferencia entre el LDA y la LR radica en que la regresión lineal no asume la normalidad de la distribución de los datos. Sin embargo, esta técnica asume que la distribución de los errores que igualan la *Y* real menos la *Y* predicha es de tipo binomial [5]. El objetivo de la LR es el de encontrar el modelo más adecuado y parsimonioso para describir la relación entre los resultados - variables dependientes - y el conjunto de variables independientes. Este método es relativamente robusto, flexible y fácil de utilizar, además de que se presenta a una interpretación significativa. La diferencia entre el LDA y la LR está en la estimación de los coeficientes no en su forma funcional [26], estas diferencias se comentan en más detalle en el artículo de Maja Pohar et al. (2004).

2.3. Curvas ROC

Junto con estas técnicas se utilizan los análisis ROC (*Receiver operating characteristic*), una herramienta útil para evaluar el rendimiento de las pruebas de diagnóstico y, en general, para evaluar la precisión de un modelo estadístico que clasifica a los sujetos en una o dos categorías [35], como es el caso de los métodos descritos en las secciones anteriores. Las medidas fundamentales de precisión diagnóstica son la "proporción verdadera positiva" que se conoce generalmente como la "sensibilidad" de la prueba, y la "proporción positiva falsa" o "especificidad". El resultado de este tipo de análisis es una curva que representa un gráfico de la sensibilidad en el *eje-y* frente a (1 - especificidad) en el *eje-x*. EL área bajo la curva (AUC) es un resumen general de la precisión diagnóstica: cuando AUC es igual a 0.5 la curva ROC corresponde a la probabilidad aleatoria, mientras que es igual a 1.0 para una precisión perfecta [36]. Cuanto mayor es el valor de AUC, mejor es la capacidad de la ecuación de medición o de regresión logísitica para discriminar el sexo. Cuando la variable no puede distinguir entre los dos grupos, el AUC será igual a 0.5, por lo que la curva ROC coincidirá con la diagonal [35].

2.4. Análisis discriminante cuadrático

Como alternativas englobadas dentro del ML empezaremos comentando los análisis discriminantes cuadráticos (QDA), clasificado dentro de las técnicas supervisadas [22]. En primer lugar podemos destacar que los métodos discriminantes son herramientas intuitivas, ya que un individuo desconocido se clasificará en el grupo al que es más similar según las medias de grupo. Además, las relaciones se pueden graficar fácilmente. Tanto la opción del LDA como la del QDA presentan el requisito de que la distribución de las puntuaciones discriminantes se puedan aproximar a una normal. Esto siempre se cumple cuando los datos originales se distribuyen normalmente, como ocurre con algunas medidas esqueléticas [14]. Sin embargo hay muchos casos en los que esto no ocurre, por lo que se deben considerar otras alternativas. Como ya hemos comentado, el LDA es la técnica discriminante mejor conocida y su capacidad de clasificación es mejor cuando el nivel de variación es similar en todos los grupos. Cuando esto no es cierto se puede usar el QDA, pero la precisión de la clasificación es, a menudo, menor que cuando aplicamos el LDA,

y los requisitos de tamaño de muestra para QDA son más altos [14]. Debido a esto el QDA no es una alternativa muy atractiva al LDA, ya que no resuelve el problema de la asunción de normalidad y, además, es más exigente con el tamaño muestral.

2.5. Redes neuronales artificiales

Otra alternativa son las redes neuronales artificiales (ANN), consideradas una herramienta poderosa en áreas de reconocimiento de patrones y clasificación [18], se encuentran entre las técnicas supervisadas de ML [19]. Este método de aprendizaje se desarrolló paralelamente en el campo de la estadística y en el de la inteligencia artificial, basándose en modelos esencialmente idénticos. La idea central es la de extraer combinaciones lineales de las entradas como funciones derivadas y después modelar el objetivo como una función no lineal de estas características. El resultado es un robusto método de aprendizaje con aplicación en muchos campos [13], incluido el de la antropología forense. Este método utiliza un algoritmo de búsqueda para examinar subconjuntos múltiples de las variables independientes con varios pesos aleatorios asignados a cada uno, evitando encontrar resultados localmente óptimos que no se puedan generalizar a individuos fuera del conjunto de *training data*. Se generan una gran cantidad de modelos que luego se comparan para evaluar cuál se ajusta mejor a los datos. El uso de la predicción y el resultado real para asignar ponderaciones de variables relativas se conoce como propagación de retroalimentación directa, una de las ANN más comunes [14]. Recientemente algunos autores han comparado los resultados obtenidos con el LDA y el método de redes neuronales obteniendo resultados satisfactorios [6,8,13,14,19]. Este método a diferencia de los análisis discriminantes y la regresión logística no representa la relación entre las variables explicativas y la variable dependiente usando una ecuación. Si no que expresa esta relación como una matriz que contiene valores, también llamados nodos, que son similares a la gran red de neuronas en el cerebro. Otra diferencia con respecto a los análisis discriminantes es que no requiere asunciones sobre la distribución de las variables explicativas y es capaz de modelar todos los tipos de funciones no lineales entre la entrada y la salida de un modelo. Esta capacidad de aproximación y su habilidad para construir modelos parsimoniosos las convierten en una potente herramienta. De hecho, estas redes no solo son capaces de modelar las relaciones que un análisis discriminante no puede, sino que también pueden diseñar modelos que son más parsimoniosos que los construidos con otras técnicas no lineales. Por lo tanto, pueden construir modelos con la misma precisión que las técnicas tradicionales no lineales pero con menos parámetros ajustables o modelos con una precisión mucho mejor con el mismo número de parámetros [8]. Lo cual hace a esta técnica una alternativa muy atractiva al LDA.

2.6. Árboles de clasificación

Una tercera alternativa son los árboles de clasificación (DT), también clasificada dentro de las técnicas de ML de aprendizaje supervisado [18]. Este método emplea una serie secuencial de reglas para estimar la pertenencia a un grupo, comenzando con la regla o nodo más efectivo que separa individuos en dos o más submuestras que se clasifican con mayor precisión de acuerdo con la membresía grupal [14]. Esta técnica está entre las que existen desde la década de 1960 pero que han estado limitadas por los requisitos computacionales hasta hace poco. Actualmente, hay muchos algoritmos para formular estos árboles, sin embargo el algoritmo de partición recursivo binario desarrollado por

Breiman et al. (1984), conocido como BFOS, sigue siendo el más conocido, el más fiable y el más exhaustivamente probado. Las principales ventajas de este algoritmo sobre las técnicas como LDA, QDA o LR son que [9]:

1. No es paramétrico, por lo que las preguntas sobre la forma de distribución apropiada son discutibles.

2. No requiere selección anticipada de variable, porque las variables se seleccionan automáticamente por su eficacia para reducir los errores de clasificación, no se utilizan variables que hacen poca o ninguna contribución al éxito de la clasificación.

3. Es robusto para los valores atípicos, que rara vez definen puntos de división que clasifican correctamente un número significativo de casos.

4. Sus resultados son invariantes a transformaciones monótonas de variables independientes, por ejemplo: la transformación logarítmica no cambiará la estructura del árbol.

5. Puede usar cualquier combinación de variables pronóstico categóricas y continuas.

6. Maneja los valores perdidos en las variables predictoras mediante el desarrollo de reglas de división basadas en mediciones alternativas que exhiben una fuerte concordancia con la variable de división primaria en cualquier punto dado en el árbol.

7. Los casos con variables respuesta desconocidas e incognoscibles pueden colocarse en su propio grupo y participar en la construcción de árboles, lo que contrasta con el LDA, donde grupos con uno o solo unos pocos casos deben excluirse del cálculo porque no es posible calcular una matriz de covarianza significativa en clases que son pequeñas.

Por lo que se puede recomendar el uso de los árboles de clasificación como una alternativa o un complemento al LDA en los siguientes casos [9]:

1. Siempre que los datos multivariantes se aparten de la normalidad.

2. Cuando los árboles de clasificación proporcionen mejores predicciones cruzadas que el análisis discriminante.

3. Cuando las variables cruzadas tengan un mejor desempeño que las variables canónicas en un análisis estructurado de árbol combinado.

4. Siempre que los conjuntos de datos carezcan de información significativa.

2.7. Random forest

Otra de las alternativas al los LDAs es la clasificación de tipo *random forest*, englobada dentro de las técnicas de ML de aprendizaje supervisado [18]. Este método utiliza muchos subconjuntos aleatorios de las variables y el muestreo repetido de los datos originales para producir cientos de árboles de decisión, llamados conjuntos. Es el consenso del conjunto el que se usa para determinar las mejores reglas de clasificación. Los RFM generalmente pueden tolerar un gran número de variables simultáneamente, incluidas las

"ruidosas" [14], lo cual supone una ventaja. Esta es una de las técnicas menos utilizadas o evaluadas en el ámbito de la antropología biológica.

2.8. Maquina de vectores de soporte

La última técnica alternativa a los análisis lineares discriminantes que evaluaremos es la clasificación de la máquina de vectores de soporte (SMV), también clasificada dentro de las técnicas de ML de aprendizaje supervisado [18]. Este método identifica los límites entre los individuos que se encuentran cerca de los límites que separan los grupos y luego manipula las ponderaciones variables para producir los límites lineales que mejor separan a los individuos de los diferentes grupos. Las máquinas de vectores de soporte se comportan especialmente bien cuando hay muchas variables con relaciones no lineales entre sí [14]. Sin embargo, al igual que la clasificación *random forest* esté método está muy poco documentado en el área que nos atañe.

4 Técnicas seleccionadas

Tras la evaluación de las técnicas que se ha llevado a cabo en el capítulo anterior se decidió hacer una comparación entre:

A. Técnicas estadísticas clásicas:

A.1. Análisis lineal discriminante.

A.2. Regresión logística.

La selección de una u otra dependerá de las características de los datos seleccionados.

B. Técnicas de ML:

B.1. Redes neuronales artificiales.

B.2. *Random Forest*.

La selección de estas dos técnicas se debe a las ventajas, comentadas en el capítulo anterior, que aportan con respecto a las técnicas clásicas. Más concretamente la selección de las ANN se debe a que dentro de las técnicas de ML es una de las que existe más bibliografía en el ámbito de la antropología forense, lo cual facilitará su implementación e interpretación - ya que al disponer de material de referencia se podrán contrastar los resultados-. Por otra parte, los ejemplos de aplicación de la técnica de RFM son muy escasos por lo que es interesante su estudio debido a lo novedoso de su aplicación en esta área.

5 Selección de los datos

Los datos se obtuvieron de la base de datos *online* *The Terry Collection Postcranial Osteometric Database*, la cual recopila información de medidas de húmero, radio, ulna, fémur, tíbia, fíbula, clavícula y escápula para individuos de poblaciones blanca y negra. Como en este estudio solo vamos a utilizar individuos de una misma población y medidas de un solo hueso para una única lateralidad, se procedió al tratamiento de estos datos utilizando el *software* estadístico R, el código utilizado para ello puede consultarse en el Anexo I (11.1).

Tras un estudio preliminar de los datos se seleccionó la población negra debido al mayor número de registros, 329 frente a 164. El hueso por el que se optó fue el fémur, ya que los huesos largos son los que presentan un mayor grado de dimorfismo y entre estos este es del que se dispone de más datos. Por otra parte se escogieron los datos de lateralidad izquierda ya que en la mayoría de las medidas el número de valores ausentes (NAs) era mucho menor que en el lado derecho, cuadro. 3.

MEDIDA	LATERALIDAD	NAs
FemMxLng	L	3
	R	18
FemBiConLng	L	30
	R	90
FemTrocLngL	L	32
	R	119
FemSubTrAPDiaL	L	29
	R	91
FemSubTrMLDiaL	L	29
	R	91
FemAPDiaMidL	L	29
	R	89
FemMLDiaMidL	L	29
	R	89
FemHeadSIDiaL	L	34
	R	92
FemHeadHzDiaL	L	38
	R	121
FemAPLatCondL	L	34
	R	119
FemAPMedCondL	L	36
	R	121
FemEpicBrL	L	33
	R	93
FemBiConBrL	L	130
	R	128
FemNeckDiaL	L	39
	R	119
FemCircMidL	L	29
	R	89

Cuadro 3: NAs por lateralidad de las medidas del fémur.

Se pude consultar una descripción detallada de las variables en el Anexo II (11.2.)

6 Obtención de los modelos

6.1 Análisis preliminar de los datos

Una vez seleccionados los datos a utilizar, en primer lugar se trataron los valores ausentes (NAs). Para ello se utilizó una representación gráfica para una mejor visualización de la situación, figura.2, donde se muestran las proporciones de NAs para cada variable. En este gráfico se puede ver que la variable *FemBiConBrL* tiene 34 % de NAs y la *FemNeckDiaL* un 10 %. Debido a este alto porcentaje de valores ausentes se decidió prescindir de estas variables.

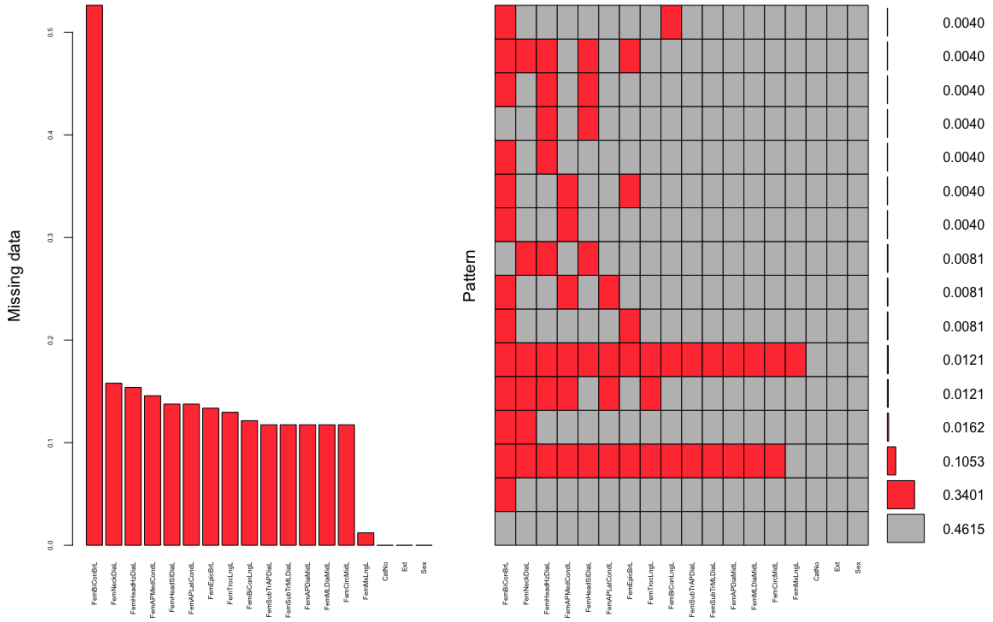

Figura 2: Proporciones de NAs para cada variable.

El siguiente paso fue la comprobación de su normalidad utilizando representaciones gráficas, figura. 3, además del la modificación de Lillefors del test de Kolmogorov-Smirnov, cuadro. 4. En los histogramas vemos como algunos de los datos parece que siguen una distribución normal mientras que en otros casos no parece tan claro. Para comprobar esta cualidad se recurrió a los p-valores del test de Lillefors, donde vemos como muchas de las medidas siguen esta distribución pero para otras no se puede aceptar la hipótesis nula - normalidad - ya que los p-valores son menores de 0.05 e incluso menores de 0.01, por lo que son claramente datos no paramétricos. Debido a esto, la técnica de estadística clásica escogida finalmente, para elaborar el modelo, fue la regresión logística (15).

Antes de proseguir con la implementación de los modelos se separó una submuestra aleatoria - *test data* - de los *training data* con la finalidad de testar los modelos, con un n=30.

Se puede consultar el código utilizado para la obtención de los resultados de este capítulo en el Anexo I (11.1.)

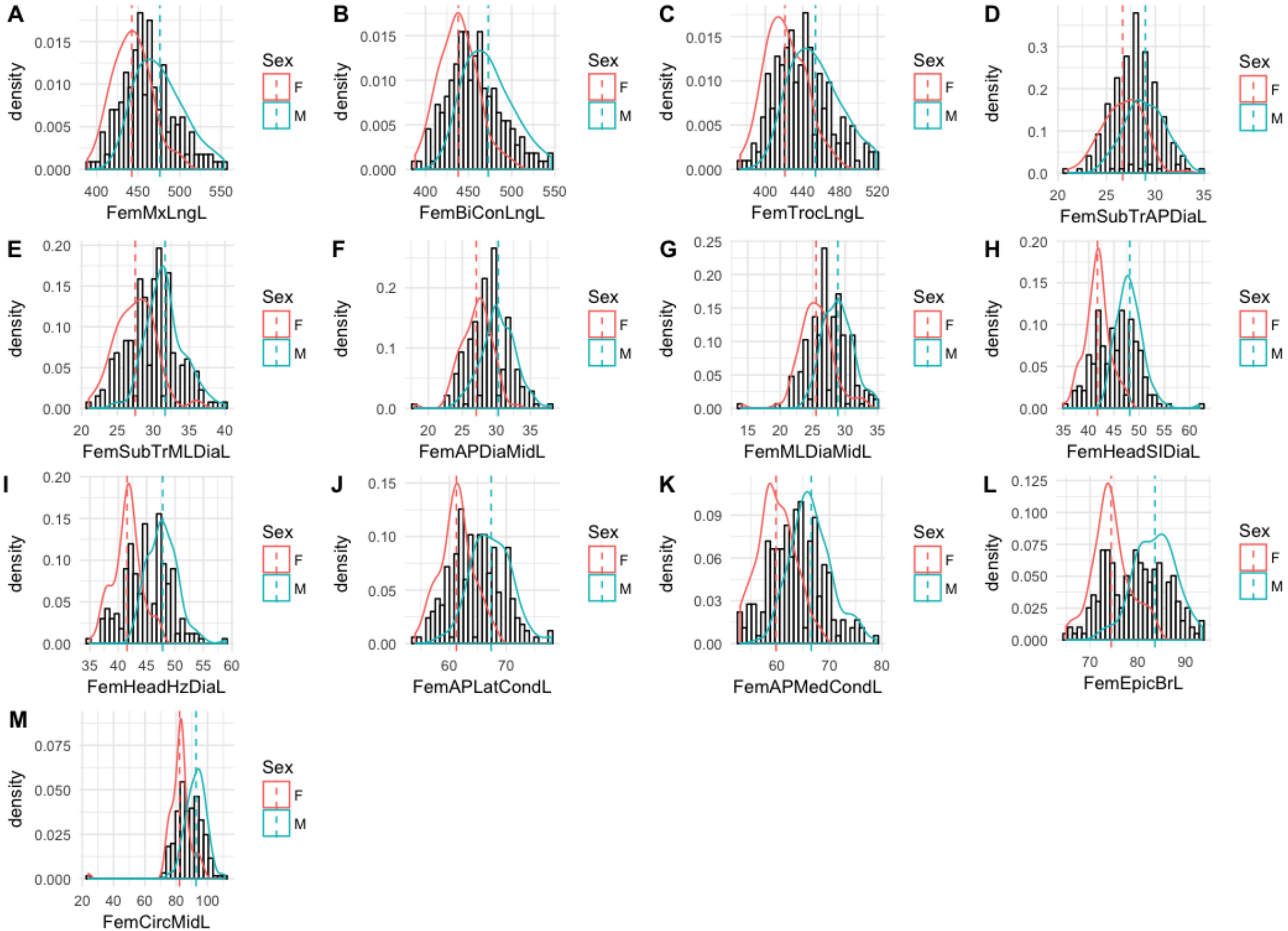

Figura 3: Histogramas y curvas de densidad de las medidas osteométricas del fémur.

MEDIDA	SEXO	p-valor
FemMxLngL	F	0.99
	M	**0.02**
FemBiConLngL	F	0.64
	M	0.11
FemTrocLngL	F	0.38
	M	0.11
FemSubTrAPDiaL	F	**0.01**
	M	**0.01**
FemSubTrMLDiaL	F	0.14
	M	$<$ **0.01**
FemMLDiaMidL	F	0.22
	M	$<$ **0.01**
FemSubTrAPDiaL	F	$<$ **0.01**
	M	$<$ **0.01**
FemHeadSIDiaL	F	**0.01**
	M	$<$ **0.01**
FemHeadHzDiaL	F	**0.01**
	M	$<$ **0.01**
FemAPLatCondL	F	**0.02**
	M	0.05
FemAPMedCondL	F	0.07
	M	$<$ **0.01**
FemEpicBrL	F	$<$ **0.01**
	M	0.29
FemCircMidL	F	$<$ **0.01**
	M	0.27

Cuadro 4: p-valores: modificación de Lillefors del test de Kolmogorov-Smirnov.

6.2 *El software WEKA*

El *software WEKA* permite probar de forma rápida diferentes métodos de ML y estadística clásica para compararlos entre sí [12]. En este trabajo se utilizó como punto de partida para la obtención de los tres modelos planteados. Los resultados muestran que se pueden obtener buenos modelos con los tres métodos escogidos, como se puede comprobar en los cuadros.5, 6, 7. Aunque es necesario matizar que WEKA utiliza el algoritmo de *backpropagation* para el modelo ANN, que también está disponible en R pero que no se utilizó porque proporcionaba peores resultados que con RPROP+. Los resultados son muy similares a los obtenidos posteriormente con R.

Prediccion / Referencia	F training	M training	F test	M test
F	80	8	76	12
M	4	110	8	106

DATOS	AUC	PRECISIÓN
training data	97.7 %	95.2 %
test data	95.5 %	90.5 %

Cuadro 5: Matrices de confusión y estimadores de la bondad del ajuste del modelo LR con WEKA.

Prediccion / Referencia	F training	M training	F test	M test
F	85	3	74	14
M	3	111	8	106

DATOS	AUC	PRECISIÓN
training data	97.1 %	96.6 %
test data	94.5 %	90.2 %

Cuadro 6: Matrices de confusión y estimadores de la bondad del ajuste del modelo ANN con WEKA.

Prediccion / Referencia	F training	M training	F test	M test
F	88	0	72	16
M	0	114	7	107

DATOS	AUC	PRECISIÓN
training data	100 %	100 %
test data	94.6 %	91.11 %

Cuadro 7: Matrices de confusión y estimadores de la bondad del ajuste del modelo RF con WEKA.

6.3 Regresión logística

En primer lugar se realizó un análisis univariante para cada una de las variables, utilizando la LR, obteniéndose que existe una relación significativa entre todas ellas y la variable respuesta, con un p-valor < 0.01.

Tras esta comprobación se generó el modelo de regresión logística obteniéndose como resultado que existe una relación significativa entre el sexo y la variable *FemHeadSIDiaL* (p- valor < 0.01), con un p-valor del conjunto del modelo de 0.01.

VARIABLE	Coeficiente	p-valor	VIF
FemMxLngL	-0.01	0.31	168.67
FemBiConLngL	< 0.01	0.80	204.62
FemTrocLngL	< 0.01	0.26	29.50
FemSubTrAPDiaL	< 0.01	0.95	2.83
FemSubTrMLDiaL	$<$ - 0.01	0.99	6.27
FemAPDiaMidL	0.02	0.15	3.99
FemMLDiaMidL	- 0.03	0.06	5.45
FemHeadSIDiaL	0.15	$<$ **0.01**	51.67
FemHeadHzDiaL	-0.05	0.20	48.55
FemAPLatCondL	$<$ - 0.01	0.46	6.32
FemAPMedCondL	$<$ - 0.01	0.75	6.42
FemEpicBrL	0.01	0.25	6.26
FemCircMidL	< 0.01	0.89	9.05

Cuadro 8: Coeficientes, p-valores y VIF de cada variable de la LR.

Antes de continuar con el análisis del modelo, se estudiaron los datos atípicos obteniéndose el gráfico de la figura.4, donde se ve como hay cuatro casos que sobresalen con respecto al resto (75, 88, 169, 184). Sin embargo, la distancia de Cook es menor que 1 para todos los casos, cuadro.9. Además, si se comparan los coeficientes del primer modelo con los de uno en el que se eliminaran los valores atípicos estos varían a partir del cuarto decimal, por tanto no se los considera influyentes para el modelo.

ID	CookD
75	0.06
88	< 0.01
169	0.01
184	0.04

Cuadro 9: Distancias de Cook.

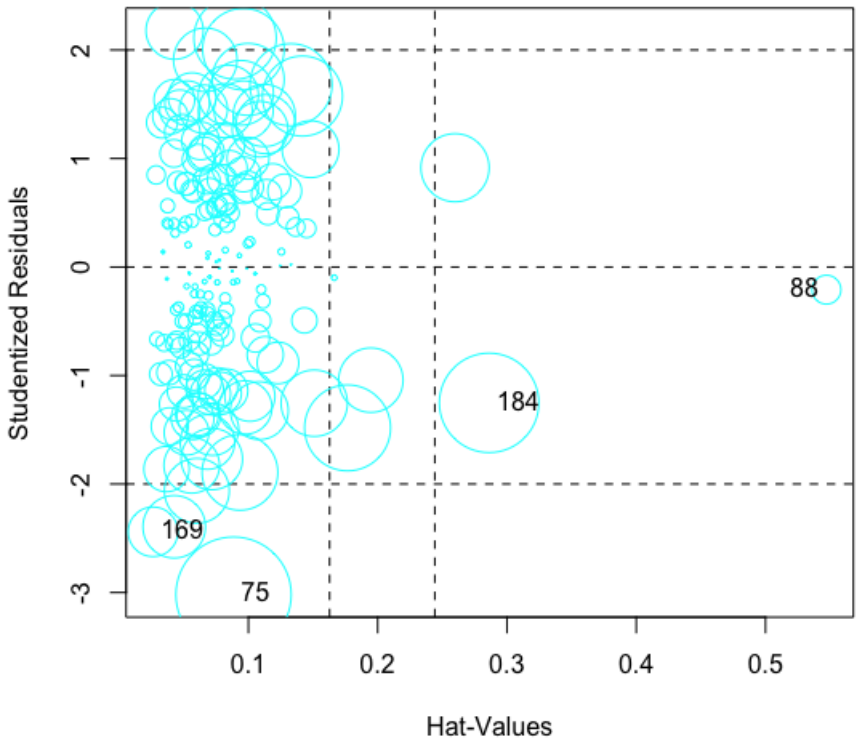

Figura 4: Representación gráfica de los valores atípicos.

Por otra parte se calculó la colinearidad entre las variables, observándose VIFs $>$ 10 (*variance influence factor*, cuadro.8), para conseguir VIFs $<$ 10 se eliminaron secuecialmente las variables no significativas con los VIFs de mayor tamaño, recalculando el modelo y los VIFs antes de seleccionar la siguiente variable a eliminar. Como resultado de este proceso se eliminaron las siguientes variables, en el orden en el que se citan: *FemBiConLngL, FemHeadHzDiaL, FemTrocLngL.*

Al hacer esto se obtiene un modelo significativo con un p-valor $<$ 0.01 en el que existe una relación significativa entre las variables *FemHeadSIDiaL* (p-valor $<$ 0.01), *FemMxLngL* (p-valor de 0.02) y la variable respuesta. Los coeficientes del modelo para cada variable junto con sus p-valores y VIFs se pueden consultar en el cuadro.10.

VARIABLE	Coeficiente	p-valor	VIF
FemMxLngL	< - 0.01	**0.02**	2.45
FemSubTrAPDiaL	< - 0.01	0.87	2.74
FemSubTrMLDiaL	< - 0.01	0.92	6.15
FemAPDiaMidL	0.02	0.20	3.87
FemMLDiaMidL	-0.03	0.12	5.24
FemHeadSIDiaL	0.10	< **0.01**	7.30
FemAPLatCondL	< 0.01	0.58	6.01
FemAPMedCondL	< - 0.01	0.75	6.33
FemEpicBrL	0.01	0.20	5.99
FemCircMidL	< 0.01	0.59	8.34

Cuadro 10: Coeficientes, p-valores y VIF de cada variable de la LR modificada.

Para comprobar el poder predictivo del modelo se utilizó la curva ROC (15), figura. 5, y se obtuvo un valor AUC de 0.97, para los *training data*. Por tanto, podemos decir que su valor predictivo es elevado y en consecuencia que contamos con un buen modelo.

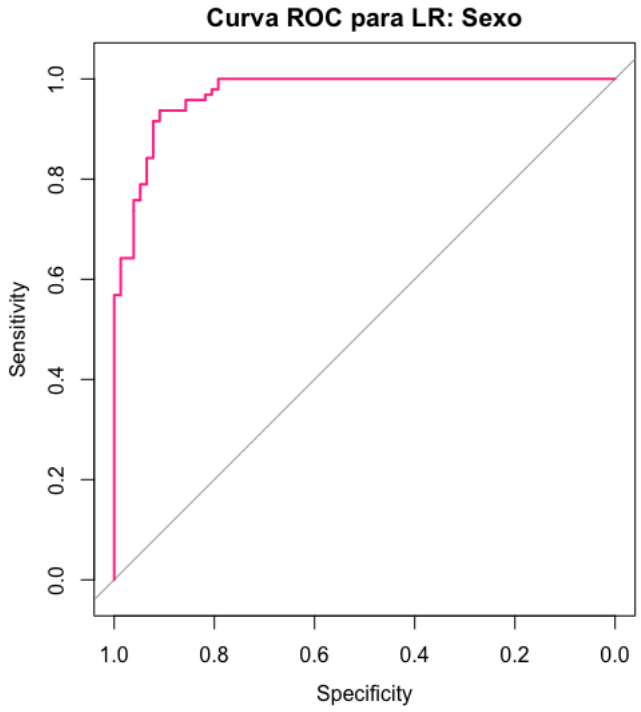

Figura 5: Curva ROC de LR.

Para estimar la bondad del ajuste del modelo se obtuvo la matriz de confusión para los *training data* obteniendo los valores de precisión, sensibilidad y especificidad de este. Sin embargo, utilizar los mismo datos con los que se estimó el modelo para obtener estos valores no es una practica recomendable ya que suelen sobreestimarse. Por ello, también se utilizó la validación cruzada, empleando los *test data*. El resultado de ambas estimaciones se muestran en el cuadro.11.

Prediccion / Referencia	**F** $_{\text{training}}$	**M** $_{\text{training}}$	**F** $_{\text{test}}$	**M** $_{\text{test}}$
F	70	6	10	1
M	7	89	1	18

DATOS	PRECISIÓN	SENSIBILIDAD	ESPECIFICIDAD
training data	99.44 %	90.91 %	93.68 %
test data	93.33 %	90.91 %	94.74 %

Cuadro 11: Matrices de confusión y estimadores de la bondad del ajuste del modelo LR.

Observando estos datos y el valor AUC podemos concluir diciendo que se ha conseguido un buen modelo para la estimación del sexo utilizando la regresión logística, con un alto poder de clasificación.

6.4 *Redes neuronales artificiales*

Antes de entrenar una ANN es una buena práctica normalizar los datos, ya que evitarlo puede llevar a resultados inútiles o a un proceso de entrenamiento muy difícil - en la mayoría de los casos el algoritmo no convergerá antes de la cantidad máxima de iteraciones permitidas -. Esto se debe a que si no se ajustan los valores a una escala común no se podrán comparar con precisión los valores predichos por la ANN con los valores reales. En este caso, se normalizaron por medio del escalado de variables - *Feature Scaling o MinMax Scaler* -, y se binarizó la variable *Sex*. Una vez hecho esto se procedió a la aplicación de este método utilizando el algoritmo RPROP+ - *resilient backpropagation with weight backtracking* - con el cual se calcularon 20 modelos de los cuales se escogió el que mejor se ajustaba nuestros *training data*, su estructura está representada en la figura.6. Es importante señalar que para la obtención de este modelo se utilizaron todas las variables seleccionadas en el análisis preliminar.

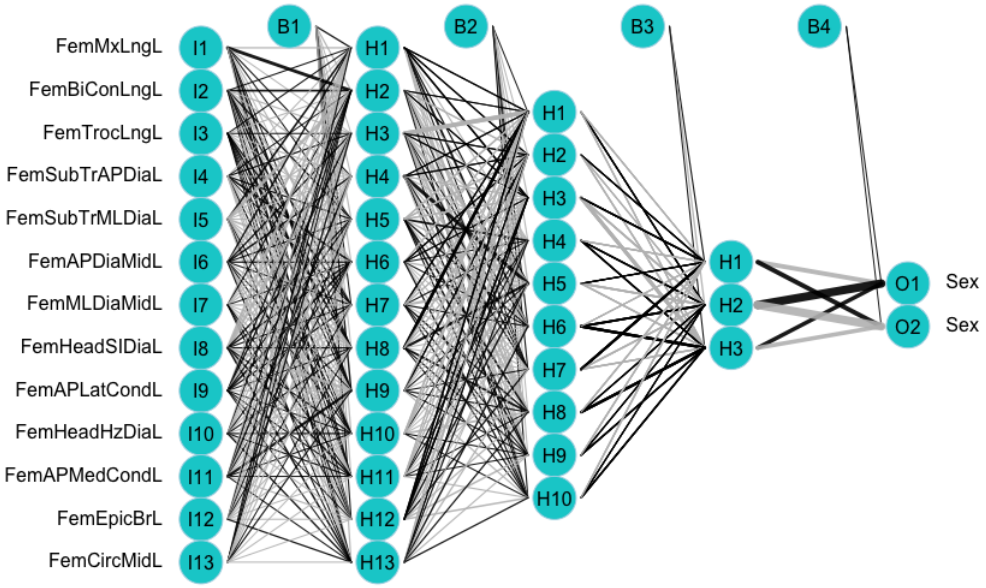

Figura 6: Representación gráfica del modelo ANN.

Para explicar la estructura representada en la figura.6 es mejor utilizar una representación de una ANN más simple como la de la figura figura.7, dónde se ven claramente las diferentes capas del modelo y sus relaciones:

1. La primera capa - *imput layer* - está formada por cuatro nodos que representan las variables predictivas - I_i -. El número de nodos de esta capa varía en función del número total de variables que se utilicen en el entrenamiento del modelo.

2. La segunda capa - *hidden layer* - es donde se utiliza el algoritmo RPROP+ para optimizar los pesos de las variables de entrada con el fin de mejorar el poder predictivo del modelo, en este caso está formada por dos nodos - H_j -. El número de nodos y de *hidden layers* varían en función de la complejidad del modelo. Cuando se utiliza R es necesario especificar cuantas *hidden layers* se quieren utilizar y su número de nodos.

3. La última capa representa la salida del modelo, es decir, las predicciones - *output* -. En este caso existen dos posibles salidas - O_k -, M y F, ya que el número de nodos de la capa *output* depende del número de etiquetas diferentes que contenga la variable respuesta.

4. Además de estas capas se representan las relaciones entre los nodos de las diferentes capas mediante flechas con los valores de los pesos de cada interacción, w_{ij}^1 y w_{jk}^2.

5. Por otra parte, también se incluye una capa - S_{ij} y S_{jk}, en este caso - donde cada valor s_{ij}^1 y s_{jk}^2 representa el sesgo incluido en el modelo para cada nodo de las diferentes capas.

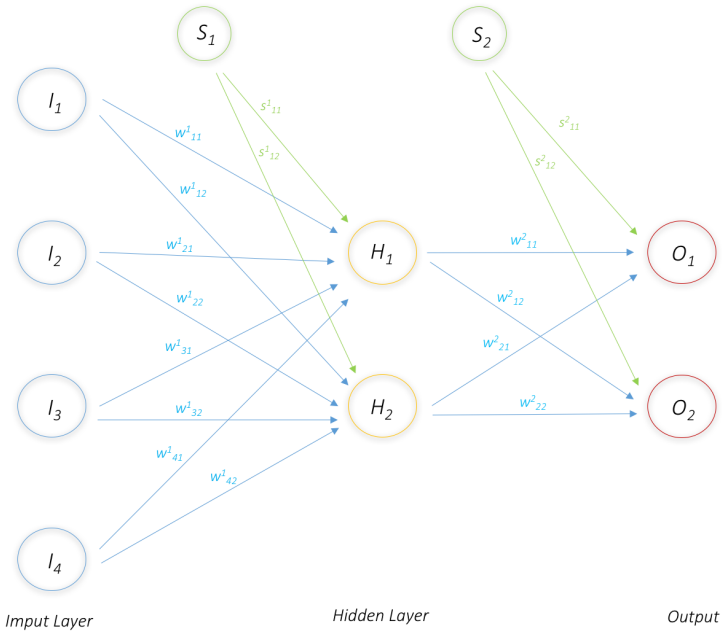

Figura 7: Representación gráfica simplificada del modelo ANN.

La representación del modelo ANN que se puede obtener con la función `plot(ann, rep="best")` - no se adjunta en el documento ya que el número de interacciones es muy alto y se genera un gráfico que hay que ampliar para poder ver los valores - sería él gráfico equivalente a la figura.6, ya que la figura.7 representa las mismas relaciones pero sin incluir los valores correspondientes a las diferentes interacciones.

La red es esencialmente una caja negra por lo que no se puede decir mucho más sobre el ajuste, los pesos y el modelo. En este punto lo que sabemos es que el algoritmo de entrenamiento ha convergido y, por tanto, el modelo está listo para ser utilizado.

El siguiente paso fue la comprobación del poder predictivo del modelo, para lo que se utilizó la curva ROC obtenieńdose un valor AUC de 0.97, para los *training data*. Por tanto, podemos decir que su valor predictivo es elevado y en consecuencia que contamos con un buen modelo.

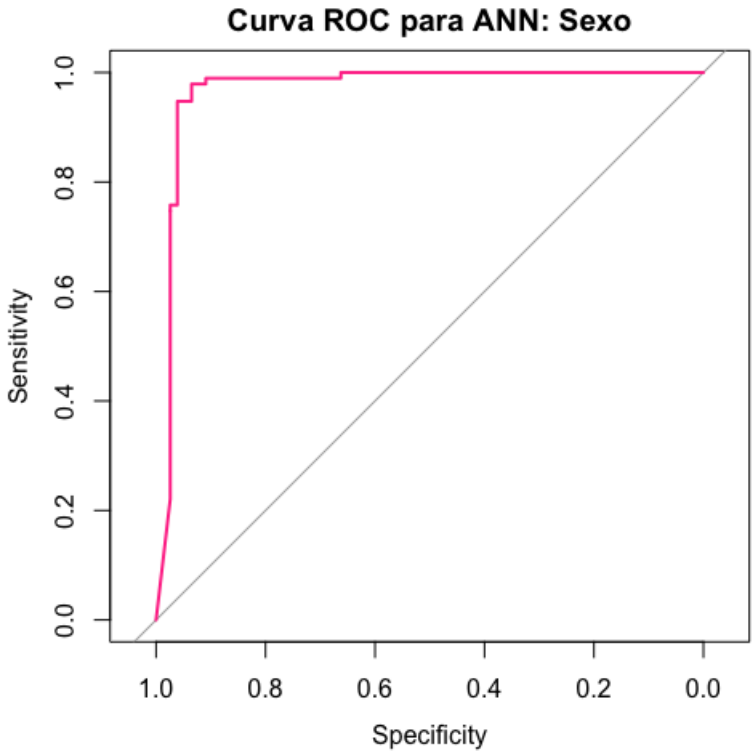

Figura 8: Curva ROC de ANN.

Para estimar la bondad del ajuste del modelo se obtuvo la matriz de confusión para los *training data* y los *test data* obteniendo los valores de precisión, sensibilidad y especificidad de este. El resultado de ambas estimaciones se muestran en el cuadro.12.

Prediccion / Referencia	**F** training	**M** training	**F** test	**M** test
F	70	10	10	2
M	7	85	1	17

DATOS	PRECISIÓN	SENSIBILIDAD	ESPECIFICIDAD
training data	94.76 %	93.51 %	95.79 %
test data	90.00 %	81.81 %	94.74 %

Cuadro 12: Matrices de confusión y estimadores de la bondad del ajuste del modelo ANN.

Observando estos datos y el valor AUC podemos concluir diciendo que se ha conseguido un buen modelo para la estimación del sexo utilizando la técnica de redes neuronales artificiales, con un alto poder de clasificación. Sin embargo, los valores obtenidos en la matriz de confusión para los *test data* son menores que los obtenidos con el modelo de LR.

6.5 *Random Forest*

En este modelo se utilizaron todas las variables seleccionadas durante el análisis preliminar de los datos. Se generaron 600 árboles utilizando dos características para la construcción de cada nodo. La elección de las dos variables y 600 árboles se debe a que son los valores mínimos para lograr el mínimo error de clasificación, según las pruebas realizadas. En la figura.9 se ve la evolución del error para la clasificación de la variable *Sex* en función del número de árboles generados (la línea roja representa el error para la clasificación de *M* la negra para *F* y la verde el error general). Por otra parte, en la figura.10 se puede consultar la representación gráfica de la estructura del árbol seleccionado.

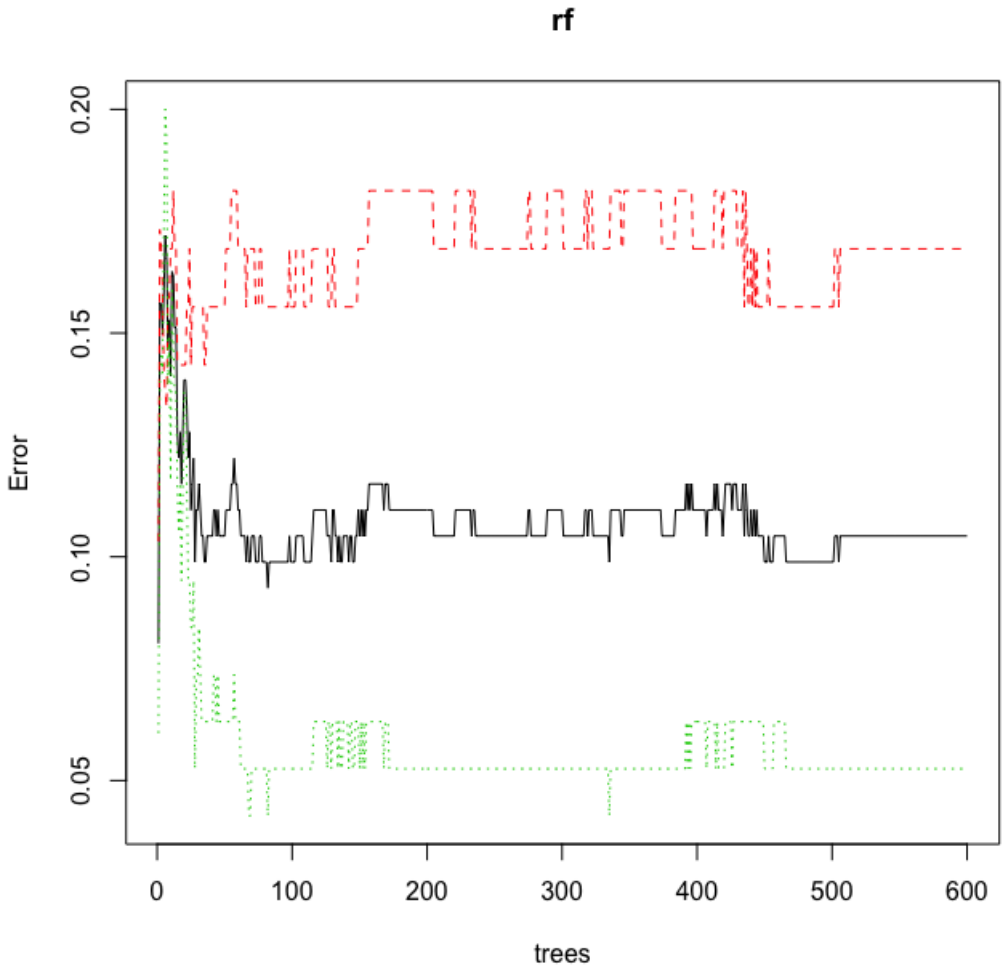

Figura 9: Evolución de la tasa de error para el modelo RF.

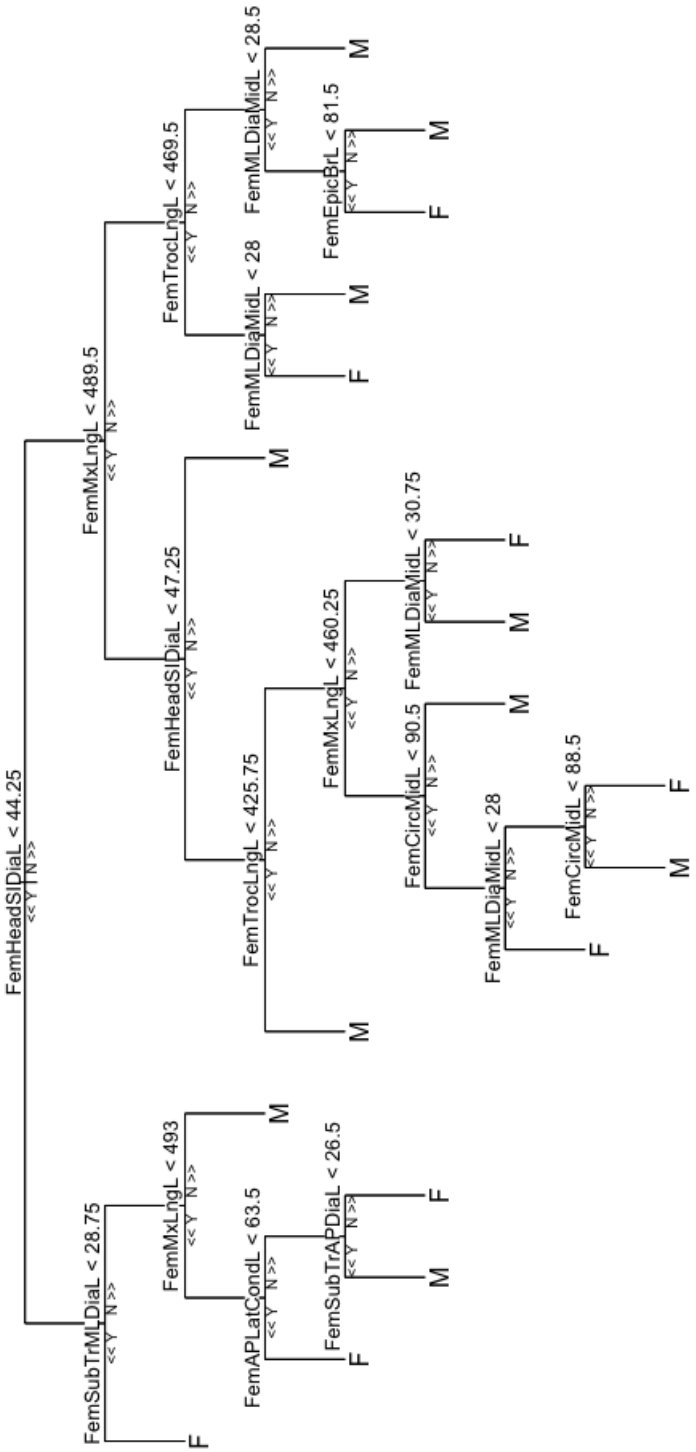

Figura 10: Representación gráfica del modelo RF.

En la figura.11 se puede ver el peso de cada variable en el modelo de clasificación, este peso se evalúa según dos criterios:

1. La precisión media de disminución (*mean decrease accuracy*): que proporciona una estimación aproximada de la pérdida en el rendimiento de predicción cuando la variable se omite del conjunto de entrenamiento.

2. La reducción media de Gini (*mean drecrease Gini*): que es una medida de impureza de nodo. Cuando la pureza es mayor significa que cada nodo contiene solo elementos de una única clase.

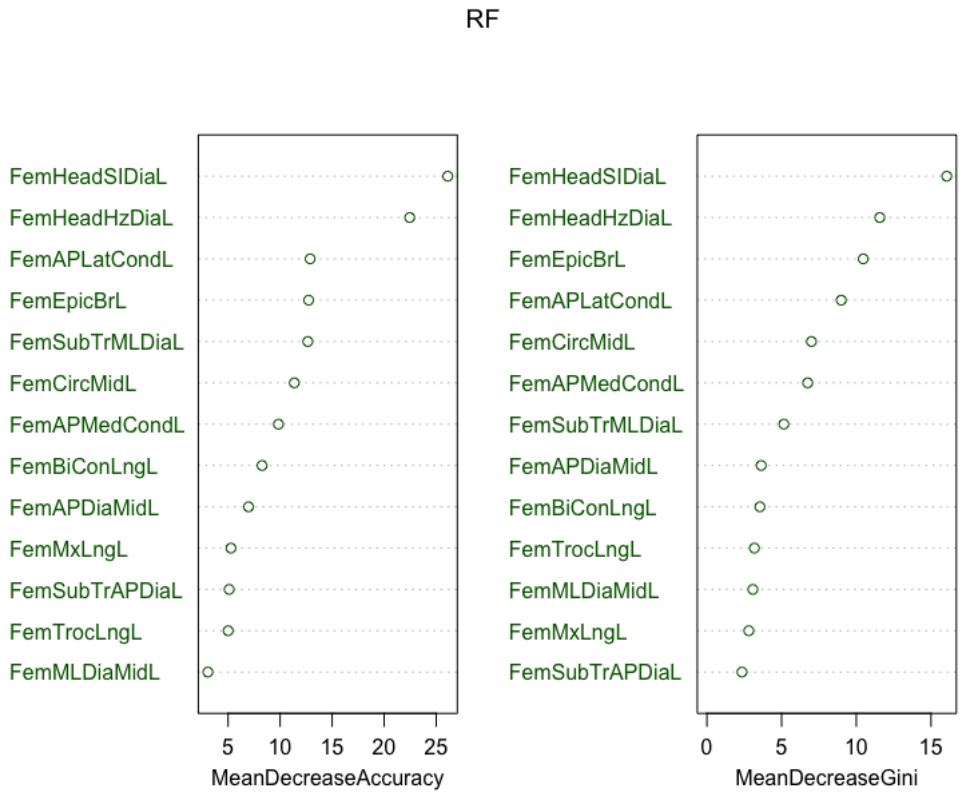

Figura 11: Importancia de las variables del modelo RF.

Como se puede ver, las dos variables con un mayor peso en este modelo son *FemHeadSIDiaL y FemHeadHzDiaL*. La primera de ellas se desmarca notablemente de las demás para los dos valores de clasificación. Mientras que para la reducción media de Gini la *FemHeadHzDiaL* se mantiene en el segundo puesto pero siguiendo la tendencia ascendiente de las demás variables.

Para comprobar el poder predictivo del modelo se utilizó la curva ROC, y se obtuvo un valor AUC de 0.94, para los *training data*. Por tanto, podemos decir que su valor predictivo es elevado y en consecuencia que contamos con un buen modelo.

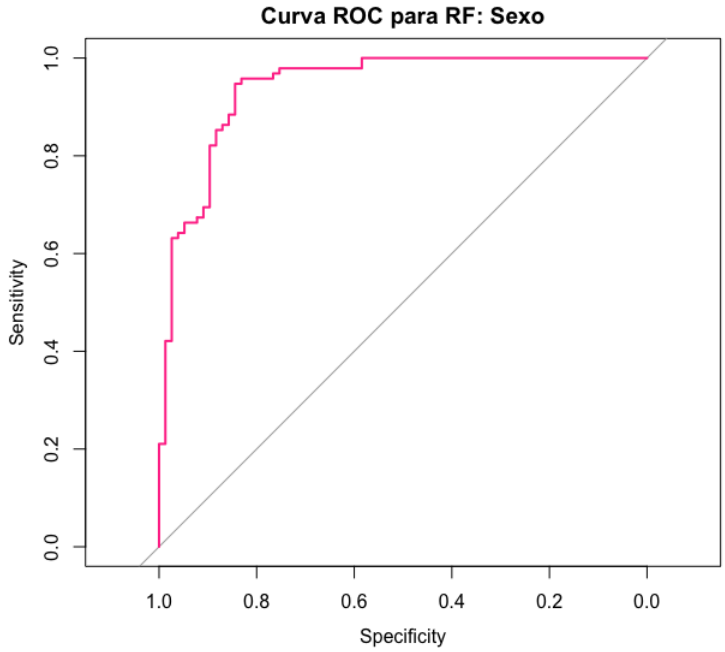

Figura 12: Curva ROC de RF.

En último lugar, para estimar la bondad del ajuste del modelo se obtuvo la matriz de confusión para los *training data* y los *test data* obteniendo los valores de precisión, sensibilidad y especificidad de este. El resultado de ambas estimaciones se muestran en el cuadro.13.

Prediccion / Referencia	**F** training	**M** training	**F** test	**M** test
F	77	0	10	0
M	0	95	1	19

DATOS	PRECISIÓN	SENSIBILIDAD	ESPECIFICIDAD
training data	100 %	100 %	100 %
test data	96.67 %	90.90 %	100 %

Cuadro 13: Matrices de confusión y estimadores de la bondad del ajuste del modelo RF.

Observando estos datos y el valor AUC podemos concluir diciendo que se ha conseguido un buen modelo para la estimación del sexo utilizando el método *Random Forest*, con un alto poder de clasificación.

7 Discusión

El objetivo de este trabajo no es comparar las diferentes herramientas disponibles para la implementación de modelos utilizando métodos de ML. Si embargo, ya que hemos empleado dos *softwares* diferentes y se han comentado los resultados obtenidos con ambos, haremos una breve comparación entre ellos. En primer lugar podemos decir que el *software* WEKA es una herramienta que proporciona una forma rápida y sencilla de comprobar los resultados que se pueden obtener empleando diferentes métodos de ML y también de estadística clásica. Este aspecto de WEKA es muy atractivo, y se debe a que en R es necesario escribir el código para efectuar estos análisis mientras que en WEKA se puede escoger entre hacerlo o no. Sin embargo, WEKA limita las modificaciones que se pueden implementar a la hora de obtener los diferente modelos, como puede ser el algoritmo escogido para el modelo de ANN. Mientras que R, en este aspecto, ofrece una gama de posibles modificaciones mucho más amplia, para adaptar los parámetros del método a las características de los datos que se quieren analizar. Así como todas las comprobaciones posteriores que se pueden efectuar con R y no con WEKA. Además, es importante destacar que los gráficos que pueden obtener con R son de mayor calidad que los que nos ofrece WEKA. Por lo que podemos decir que WEKA es una buena opción para realizar una evaluación preliminar del método que mejor modela nuestros datos. Sin embargo, R sería la mejor opción para mejorar su ajuste, así como para realizar un análisis posterior del modelo obtenido más completo.

Una vez hecha esta reflexión sobre estos dos *softwares*, vamos a evaluar los diferentes modelos obtenidos. En primer lugar, como ya se comentó en los diferentes apartados del capítulo.6, se obtuvieron buenos modelos de clasificación utilizando los tres métodos planteados. Sin embargo, podemos destacar que la LR en comparación con las ANN y RF requiere de una mayor inversión de tiempo en el análisis preliminar de los datos y del comportamiento de las variables. Esto se debe a que este método es sensible a las relaciones de multicolinearidad mientras que los métodos de ML buscan más allá de estas relaciones por lo que no se ven afectados. En nuestro caso, debido a las fuertes relaciones de colinearidad entre variables fue necesario ajustar el modelo eliminado las variables con VIFs más altos, recalculando el modelo antes de cada eliminación, hasta conseguir que todos los VIFs fueran menores de diez. Por tanto, vemos como los métodos de ML en este aspecto son más eficientes que los métodos de estadística clásica.

Por otra parte, si comparamos las ANN con RF, sin ninguna duda, el método más sencillo de implementar e interpretar es RF, además de que el modelo obtenido fue el que mejores predicciones proporcionó. Si profundizamos en esta afirmación, buscando las razones que nos llevan a ella, nos encontramos con que en primer lugar las ANN necesitan un ajuste de las variables antes de realizar el análisis, ya que si los valores de estas no está escalados los modelos obtenidos no son buenos, debido a que la ANN no es capaz de converger al no poder comparar los pesos que le asigna a cada variable. Además, el ajuste del modelo a nivel de número de capas y nodos por capa requiere la realización de numerosas pruebas hasta encontrar la combinación con la que se obtiene un mejor modelo. En segundo lugar, una vez obtenido el modelo, la interpretación de la salida de la ANN

es mucho más compleja que la de RF. Debido a que con la ANN se obtienen una serie de matrices con los valores de los pesos de las interacciones entre los nodos de cada capa. En el caso de una ANN sencilla, es decir, con pocas variables y pocas interacciones la interpretación no es complicada. Sin embargo, cuando nos enfrentamos a una ANN más compleja, como la obtenida para nuestros datos, su interpretación se hace impracticable. Por lo que solo se puede comentar el modelo a nivel de las predicciones que se obtienen, en nuestro caso a nivel de la matriz de confusión.

En el caso del modelo de RF su salida es mucho más limpia e informativa que la de la LR o las ANN. Además de que a la hora de ajustar el modelo es mucho más sencillo, ya que no es necesario ir compilando y comprobando el ajuste del modelo una y otra vez por cada cambio que se hace - en al caso de las ANN son modificaciones que, además, se hacen un poco a ciegas -. Si no que una vez escogido un número de repeticiones y de variables a utilizar en cada nodo, se puede obtener el gráfico de la figura.9 que sirve de guía para las modificaciones a hacer en el código para el nuevo modelo. Otra ventaja de RF frente a las ANN es la posibilidad de evaluar cuales son las variables que tienen un mayor peso en el modelo de una forma sencilla. En el caso de las ANN, si solo existiera una salida posible podríamos utilizar una función de R que permite obtener los gráficos con los pesos asignados a cada variable. Desafortunadamente, en nuestro caso esto no es posible, por lo tanto, en el caso de que se quisieran evaluar estos pesos habría que perderse entre la densa red de conexiones existente entre la capa de entrada y de salida del modelo.

Teniendo todo esto en cuenta podemos decir que, a pesar de que todos los métodos proporcionaron buenos modelo, el mejor es el obtenido con RF. Además de ser el método más rápido y claro.

8 Conclusiones

Tras la obtención y evaluación de los tres modelos obtenidos podemos concluir diciendo que las técnicas de ML proporcionan valiosas herramientas para la implentación de potentes modelos de clasificación, en este caso para la predicción del sexo a partir de medidas osteométricas del fémur. Sin embargo, a pesar de que las dos técnicas de ML analizadas proporcionaron muy buenos resultados, hay que destacar que el método de RF resultó mucho más potente y sobre todo más claro, tanto a la hora de su entrenamiento como en la interpretación de su resultados. Por lo que, aunque las dos son técnicas muy potentes, desde un punto de vista práctico los análisis RF parecen una herramientas más clara para su interpretación.

Centrándonos en los objetivos académicos planteados para este trabajo, los O.G. que nos propusimos alcanzar fueron:

1. La redacción de un texto, tras una revisión bibliográfica, en el que se describieran y contrastaran las diferentes técnicas de estadística clásica y ML aplicadas en el ámbito de la antropología forense.

2. La obtención de unos modelos de clasificación utilizando una serie de técnicas escogidas tras la revisión bibliográfica.

Como se puede comprobar tras la lectura de este documento, el primer objetivo general se refleja en los capítulos 2, 3 y 4. Donde en primer lugar se hace una breve introducción al ML, centrando el tema del trabajo. Para después proceder a la descripción, comparación y evaluación de las diferentes técnicas, tanto de estadística clásica como de ML, aplicadas en esta área, y por último se hace una selección de las escogidas para la obtención de los modelos. Mientras que el segundo se corresponde con los capítulos 5, 6 y 7. Dónde, primero se seleccionan los datos que se analizarían posteriormente. Después se obtienen los modelos utilizando los diferentes métodos seleccionados y por último se hace una evaluación de los resultados obtenidos con cada uno de ellos.

Para la consecución de estos objetivos, se propuso un plan de trabajo y una metodología que se cumplieron en todos los caso a excepción de la tarea 5 que, como ya se comentó en la sección.1.4 de la introducción, sufrió un retraso de cuatro días. Por lo demás, se cumplió tanto el calendario como los métodos propuestos.

Como reflexión final, se puede decir que que se obtuvieron unos buenos resultados y que sería interesante hacer una evaluación de otros métodos de ML aplicándolos en este campo de la biología, ya que muchos de ellos están prácticamente ausentes en la bibliografía relacionada con la antropología forense y su evaluación podría arrojar resultados muy útiles.

9 Glosario

- **Algoritmo:** conjunto ordenado y finito de operaciones que permite hallar la solución de un problema.
- **ANN:** Redes neuronales artificiales.
- **Antropología física:** rama de la antropología que estudia al ser humano considerando su naturaleza.
- **Antropología forense:** subdisciplina de la antropología física que se encarga de la identificación de restos óseos humanos.
- **AUC:** área bajo la curva, resumen general de la precisión diagnóstica.
- **Bioarqueológía:** subdisciplina de la antropología física que estudia las poblaciones ancestrales por medio del análisis de restos óseos.
- **Centroide:** punto donde se produce la intersección de la medianas que forman parte de un triángulo.
- **DT:** árboles de clasificación.
- **Especificidad:** proporción positiva falsa.
- **Homocedasticidad:** homogeneidad de varianzas.
- **LDA:** análisis discriminate lineal.
- **LDF:** función discriminante lineal.
- **LR:** regresión logística.
- **Matriz de confusión:** método de evaluación de los modelos de clasificación.
- **ML:** *Machine Learning*.
- **NA:** valor ausente.
- **Perfil biológico:** conjunto de parámetros que definen las características morfológicas de un individuo.
- **Precisión:** dispersión del conjunto de valores obtenidos de mediciones repetidas de una magnitud.
- **QDA:** Análisis discriminante cuadrático.
- **ROC:** curva característica operativa del receptor. Herramienta útil para evaluar el rendimiento de las pruebas de diagnóstico.
- **RF:** *Random Forest*.

- **Sensibilidad:** proporción verdadera positiva.
- **SMV:** maquína de vectores de soporte.
- ***Software*:** conjunto de programas, instrucciones y reglas informáticas para ejecutar determinadas tareas en una computadora.
- ***Test data*:** submuestra, extraída del conjunto de datos inicial para testar la bondad del ajuste de los modelos.
- **Tiempo de capacitación:** tiempo que tarda en converger un algoritmo.
- ***Training data*:** datos de entrenamiento del modelo.
- **Variable categórica:** son aquellas que contienen un número finito de categorías o grupos distintos.
- **Variable respuesta:** aquella que se quiere estimar.
- **Variable predictiva:** aquella que se utiliza para hacer la predicción de la variable respuesta.
- **VIF:** factor de influencia de la varianza.

10 Bibliografía

1. Bidmos MA, Gibbon VE, Strkalj G (2010) Recent advances in sex identification of human skeletal remains in South Africa. South Africa Journal Sciences 238:1-6.

2. Bishop CM (2006) Pattern recognition and machine learning.Springer, New York.

3. Bruzek J (2002) A method for visual determination of sex, using the human hip bone. American Journal of Physical Anthropology 117:157-168.

4. Bruzek J, Murail P (2006) Forensic anthropology and medicine. Chapter 9. Methodology and Reliability of Sex Determination From the Skeleton. Humana Press, New Jersey.

5. Chao-Ying Joanne Peng T-SHS (2002) Logistic regression analysis and reporting: a primer. Underst Stat Education 31-70.

6. Corsini MM, Schmitt A, Bruzek J (2005) Aging process variability on the human skeleton: artificial network as an appropriate tool for age at death assessment. Forensic Science International 148:163-167.

7. Dudzik B and Jantz RL (2016) Misclassifications of Hispanics Using Fordisc 3.1.: Comparing Crenial Morphology in Asian and Hispanic Populations. Journal of Forensic Sciences 61(5):1311-1318.

8. Du Jardin P, Ponsaillé J, Alunni-Perret V, Quatrehomme G (2009) A comparison between neural network and other metric methods to determine sex from the upper femur in a modern French population. Forensic Science International 192:e1?e6.

9. Felderman MR (2002) Classification Trees as an Alternativa to Linear Discriminant Analysis. American Journal of Physical Anthropology 119:257-275.

10. Fielding A (2007) Cluster and classification techniques for the bio-sciences. Cambridge University Press, Cambridge.

11. Günther F and Fritsch S (2010) neuralnet: Training of Neural Networks. The R Journal (2) 1:30-38.

12. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. ACM SIGKDD Explorations Newsletter 11:10-18

13. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, second edition. Springer, New York.

14. Hefner JT, Ousley SD (2014) Statistical classification methods for estimating ancestry using morphoscopic traits. Journal of Forensic Sciences 59:883-890.

15. Hefner JT, Spradley KM, Anderson B (2014) Ancestry Assessment Using Random Forest Modeling. Journal of Forensic Sciences 59:583-589.

16. Jellinghaus K, Hoeland K, Hachmann C, Prescher A, Bohnert M, Jantz R (2017) Crenial secular change form the nineteenth to the twentieth century in modern German individuals compared to modern Euro-American individuals. International Journal of Legal Medicine 0937-9827.

17. Kazzazi SM, Kranioti EF (2018) Sex estimation using cervical dental measurements in an archeological population from Iran. Archeological and Anthropological Sciences 10:439-448.

18. Li H (2017) Which machine Learning algorithm should I use? [Consultado: 26-03-2018] Disponible en: https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

19. Mahfouz M, Badawi A, Merkl B, Fatah EEA, Pritchard E, Kesler K, Moore M, Jantz R, Jantz L (2007) Patella sex determination by 3D statistical shape models and nonlinear classifiers. Forensic Science International 173:161-170.

20. Manthey L, Jantz RL, Bohnert M, Jellinghaus K (2017) Secular change of sexually dimorphic crenial variables in Euro-American and Germans. International Journal of Legal Medicine 131:1113-1118.

21. Mitchell TM (1997) Machine learning. McGraw Hill, Burr Ridge.

22. Morais CLM and Lima KMG (2018) Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry. Jpurnal of the Brazilian Chemical Society 29(3) 472-481.

23. Moss GP, Shah AJ, Adams RG, Davey N, Wilkinson SC, Pugh WJ, Sun Y (2012) The application of discriminant analysis and machine learning methods as tools to identify and classify compounds with potential as transdermal enhancers. European Joutnal of Pharmaceutical Sciences 45:116-127.

24. Murail P, Bruzek J, Houët F, Cunha E (2005) DSP: A tool for probabilistic sex diagnosis using worldwide variability in hip-bone measurements. Bulletins et Mémories de la Société D?Anthropologie de Paris 17:167-176.

25. Navega D, Vicente R, Vieira DN, Ross AH, Cunha E (2015) Sex estimation from the tarsal bones in a Portuguese sample: a machine learning approach. International Journal of Legal Medicine 129:651-659.

26. Pohar M, Blas M and Turk S (2004) Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki zvezki 1:143-161.

27. Ramsthaler F, Kreutz K, Verhoff MA (2007) Accuracy of metric sex analysis of skeletal remains using Fordisc®based on a recent skull collection. International Journal of Medicine 121:477-482.

28. R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/ [Consultada: 9-03-2018].

29. Saldías E, Malgosa A, Jordana X, Isidro A (2016) Sex estimation from the navicular bone in Spanish contemporary skeletal collections. Forensic Science International 267:229.e1-229.e6.

30. Schafer J (1997) Analysis of incomplete multivariate data. New York: Chapman and Hall.

31. Schimert J, Schafer JL, Hesterberg TM, Fraley C, Clarkson DB. (2000) Analyzing data with missing values in S-Plus. Seattle: Insightful Corp.

32. Skala W (2018) Drawing Gantt Charts in LaTeX with TikZ. The pgfgantt Package. URL: http://bay.uchicago.edu/CTAN/graphics/pgf/contrib/pgfgantt/pgfgantt.pdf [Consultada: 12-03-2018].

33. Uberlaker DH (2006) Forensic anthropology and medicine. Chapter 1. Introduction to Forensic Anthropology. Humana Press, New Jersey.

34. Urbanová P, Ross AH, Jurda M, Nogueira MI (2014) Testing the reliability of software tools in sex and ancestry stimation in a multi-ancestral Brazilian sample. Legal Medicine 16:264:273.

35. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. Elsevier, New York.

36. Zou KH, O'Malley AJ, Mauri L (2007) Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Test and Predictive Models. American Heart Association 115:654-657.

11 Anexos

11.1 ANEXO I: Código utilizado para los análisis en R

En este anexo se describe el código utilizado en R para la obtención de los modelos detallados en este trabajo. Además, también están disponibles los *scripts* en el repositorio: MEMORIA. Por lo que, para ejecutar este código puede hacerse bien copiándolo de este documento o descargando los *scripts* desde el repositorio, siempre modificando las rutas hasta los ficheros.

```
#LECTURA DE LOS DATOS:
#Previa descarga desde la BBDD
library(gdata)
demography <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 9)
humerus <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 1)
radius <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 2)
ulna <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 3)
femur <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 4)
tibia <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 5)
fibula <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 6)
clavicle <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 7)
scapula <- read.xls("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/terry.xlsx", sheet = 8)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#MODIFICACIÓN DE LOS DATOS:
#Se renombraron las variables 1 y 3 de demography,
#para que tengan los mismos nombres que en los demás conjuntos de datos
#(es decir, los data.frame para cada hueso)
#por comodidad a la hora de manejar los datos:
names(demography)[1] <- paste('CatNo')
names(demography)[2] <- paste('Ext')

#Selección de la raza de los individuos a utilizar,
#para lo que se obtuvo el número de individuos blancos y negros:
sum(demography$Race == 'BL')
sum(demography$Race == 'WH')
#Debido al mayor número de registros de individuos negros
#se optó por utilizar los datos de estos, excluyendo los de los individuos blancos:
demography.BL <- demography[demography$Race == 'BL',]

#Se analizaron las dimensiones de los datos para los huesos largos,
#ya que son los que presentan un mayor dimorfismo:
dim(humerus)
dim(radius)
dim(ulna)
dim(femur)
dim(tibia)
dim(fibula)
#El hueso con un mayor número de medidas y de registros es el Fémur,
#por lo que fue el hueso escogido.

#Se añadieron a los datos de las medidas del fémur las variables de demography,
#para poder identificar el sexo de cada individuo:
femur <- merge(demography.BL, femur, by = c("CatNo", "Ext"))

#Analizando los datos podemos ver como el número de NAs
#para las medidas del lado derecho es mucho mayor que para el lado izquierdo:
sapply(femur, function(x) sum(is.na(x)))

#Por lo que analizaremos los datos con lateralidad izquierda,
#para ello se eliminaron las columnas con las medidas de lateralidad derecha del data.frame:
names(femur)
```

```
femur <- femur[, -c(4, 5, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30, 32, 34, 36, 37)]

#Se generó un .csv con los datos modificados para el fémur:
write.csv(femurL, file='femurL.csv')

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#LECTURA de los datos modificados:
femurL <- read.table("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/femurL.csv",
                     header = TRUE, sep = ",")
#En primer lugar se eliminaron las variables
#que no se utilizarían en el análisis:
names(femurL)
femurL <- femurL[, -c(1:3)]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#ANÁLISIS DE LOS DATOS:
#Análisi de los NAs por variable:
library(VIM)
na_plot <- aggr(femurL,
                col=c('gray', 'brown1'),
                numbers=TRUE,
                sortVars=TRUE,
                labels=names(femur),
                cex.axis=.3,
                gap=3,
                ylab=c("Missing data","Pattern"))

#Se eliminaron las variables 5 y 17, debido al gran porcentage de NAs,
#además de los NAs de todas las variables:
femurLNNA <- na.omit(femurL[,-c(14, 15)])

#Se creo un nuevo archivo con los datos que se emplearán para los modelos:
write.csv(femurLNNA, file='femurLNNA.csv')

#El siguiente paso fue comprobar la normalidad de nuestros datos:
#Histograma y curva normal:
#Femur:
library(plyr)
mu.fe1 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemMxLngL))
mu.fe2 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemBiConLngL))
mu.fe3 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemTrocLngL))
mu.fe4 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemSubTrAPDiaL))
mu.fe5 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemSubTrMLDiaL))
mu.fe6 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemAPDiaMidL))
mu.fe7 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemMLDiaMidL))
mu.fe8 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemHeadSIDiaL))
mu.fe9 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemHeadHzDiaL))
mu.fe10 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemAPLatCondL))
mu.fe11 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemAPMedCondL))
mu.fe12 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemEpicBrL))
mu.fe13 <- ddply(femurLNNA, "Sex", summarise, grp.mean=mean(FemCircMidL))

library(ggplot2)
fe1 <- ggplot(femurLNNA, aes(x=FemMxLngL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe1, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe2 <- ggplot(femurLNNA, aes(x=FemBiConLngL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe2, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe3 <- ggplot(femurLNNA, aes(x=FemTrocLngL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
```

```
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe3, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe4 <- ggplot(femurLNNA, aes(x=FemSubTrAPDiaL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe4, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe5 <- ggplot(femurLNNA, aes(x=FemSubTrMLDiaL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe5, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe6 <- ggplot(femurLNNA, aes(x=FemAPDiaMidL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe6, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe7 <- ggplot(femurLNNA, aes(x=FemMLDiaMidL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe7, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe8 <- ggplot(femurLNNA, aes(x=FemHeadSIDiaL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe8, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe9 <- ggplot(femurLNNA, aes(x=FemHeadHzDiaL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe9, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe10 <- ggplot(femurLNNA, aes(x=FemAPLatCondL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe10, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe11 <- ggplot(femurLNNA, aes(x=FemAPMedCondL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe11, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe12 <- ggplot(femurLNNA, aes(x=FemEpicBrL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe12, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()
fe13 <- ggplot(femurLNNA, aes(x=FemCircMidL, color=Sex)) +
  geom_histogram(aes(y=..density..), colour="black", fill="white", position="dodge")+
  geom_density(alpha=.2) +
  geom_vline(data = mu.fe13, aes(xintercept=grp.mean, color = Sex), linetype="dashed") +
  theme(legend.position="top") +
  theme_minimal()

library('cowplot')
plot_grid(fe1, fe2, fe3, fe4, fe5, fe6, fe7, fe8, fe9, fe10, fe11, fe12, fe13,
          labels = c(LETTERS[1:13]), align = "v")
```

```
#Modificación de Lillefors del test de Kolmogorov-Smirnov:
library(nortest)
lillie.test(femurLNNA$FemMxLngL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemMxLngL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemBiConLngL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemBiConLngL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemTrocLngL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemTrocLngL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemSubTrAPDiaL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemSubTrAPDiaL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemSubTrMLDiaL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemSubTrMLDiaL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemMLDiaMidL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemMLDiaMidL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemSubTrAPDiaL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemSubTrAPDiaL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemHeadSIDiaL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemHeadSIDiaL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemHeadHzDiaL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemHeadHzDiaL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemAPLatCondL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemAPLatCondL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemAPMedCondL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemAPMedCondL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemEpicBrL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemEpicBrL[femurLNNA$Sex=="M"])

lillie.test(femurLNNA$FemCircMidL[femurLNNA$Sex=="F"])
lillie.test(femurLNNA$FemCircMidL[femurLNNA$Sex=="M"])

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#LECTURA de los datos para el análisis:
femurLNNA <- read.table("/Users/noemi/Documents/BIO_INF_EST/TFM/DATOS/femurLNNA.csv",
                        header = TRUE, sep = ",")

#Al leer los datos se genera una columna "X" indicando el número de filas,
#como no es necesaria se elimina para tener un data.frame más limpio:
femurLNNA <- femurLNNA[, -1]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#TRAINING y TEST DATA:
#Se utilizó la función set.seed()
#para fijar que cada vez que se genere la submuestra esta sea siempre la misma,
#asegurando la reproducibilidad de los modelos:
set.seed(88888)
submuestra <- sample(1:nrow(femurLNNA), size = 30, replace = FALSE)
submuestra
test <- femurLNNA[submuestra,]
training <- femurLNNA[-submuestra,]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#REGRESIÓN LOGÍSTICA:
#ANALISIS UNIVARIANTE (LR):
lr.1 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemMxLngL , data=training)
summary(lr.1)
```

```
lr.2 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemBiConLngL , data=training)
summary(lr.2)
lr.3 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemTrocLngL , data=training)
summary(lr.3)
lr.4 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemSubTrAPDiaL , data=training)
summary(lr.4)
lr.5 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemSubTrMLDiaL , data=training)
summary(lr.5)
lr.6 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemAPDiaMidL , data=training)
summary(lr.6)
lr.7 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemMLDiaMidL , data=training)
summary(lr.7)
lr.8 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemHeadSIDiaL , data=training)
summary(lr.8)
lr.9 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemHeadHzDiaL , data=training)
summary(lr.9)
lr.10 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemAPLatCondL , data=training)
summary(lr.10)
lr.11 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemAPMedCondL , data=training)
summary(lr.11)
lr.12 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemEpicBrL , data=training)
summary(lr.12)
lr.13 <- glm(formula = ifelse(Sex == "M", 1, 0) ~ FemCircMidL , data=training)
summary(lr.13)

#ANALISIS MULTIVARIANTE (LR):
#Primer modelo:
lr <- glm(formula = ifelse(Sex == "M", 1, 0) ~ ., data=training)
summary(lr)

#Evaluación del modelo:
#Diferencia de residuos
dif_residuos <- lr$null.deviance - lr$deviance

#Grados libertad
df <- lr$df.null - lr$df.residual

#p-value
p_value <- pchisq(q = dif_residuos,df = df, lower.tail = FALSE)
round(p_value, 4)
#El conjunto del modelo es significativo.

#Sin embargo antes de continuar se hicieron algunas comprobaciones:
#Out-liers:
library(mvinfluence)
#Generación de gráficos:
#1. Este no se muestra en la memoria, pero es una alternativa al escogido:
influenceIndexPlot(lr)

#2. Este el el gráfico que se escogió para incluir en este trabajo,
#a parte del gráfico, genera un output con las distancias de Cook:
influencePlot(lr, col="cyan")

#Con las siguientes líneas se comparan los coeficientes del modelo LR y
#de un modelo en el que se hubieran eliminado los out-liers:
lr.out <- update(lr, subset=-c(184, 88, 169, 75))
compareCoefs(lr, lr.out)

#VIF:
library(car)
vif(lr)
#Donde vemos como la colinearidad es muy alta entre variables.
#Por lo que se fueron eliminando variables desde las que presentaban un mayor VIF
#y no eran significativas.
#Hasta reducir la colinearidad a VIFs <10.

#Para ello se ejecutaron las siguientes líneas de código
#modificando los valores del argumento:
#data=training[, -c('columna de la variable a eliminar').
```

```
#Obteniéndose el segundo modelo:
lr.mod <- glm(formula = ifelse(Sex == "M", 1, 0) ~ ., data=training[, -c(3,10,4)])
summary(lr.mod)

#Evaluación del modelo:
#Diferencia de residuos
dif.residuos.mod <- lr.mod$null.deviance - lr.mod$deviance

#Grados libertad
df.mod <- lr.mod$df.null - lr.mod$df.residual

# p-value
p.value.mod <- pchisq(q = dif.residuos.mod,df = df.mod, lower.tail = FALSE)
round(p.value.mod, 4)

#VIFs:
vif(lr.mod)

#Una vez obtenido el modelo se obtuvo su poder predictivo utilizando la curva ROC:
require(pROC)
#Se otiene una predicción utilizado el modelo lr.mod y los training data:
p <- predict(lr.mod, newdata = subset(training,
                                      select = c(2, 5:9, 11:14),
                                      type = "response" ))

#Se obtiene la curva ROC:
lr.mod.roc<-roc(training$Sex, p)

#Se grafíca:
plot(lr.mod.roc, col="violetred1", lwd=2, main="Curva␣ROC␣para␣LR:␣Sexo")

#Se obtiene el valor AUC:
auc(lr.mod.roc)

#Ratio de clasificación utilizando una matriz de confusión:
library(e1071)
library(caret)

#training data:
#Se obtiene una predicción utilizando el modelo lr.mod y los training data:
predicted.lr.1 <- predict(lr.mod, training)

#Se genera la matriz de confusión:
pf.lr.1 <- as.factor(ifelse(round(predicted.lr.1) == 1, "M", "F"))
confusionMatrix(data=pf.lr.1, reference=training$Sex)

#test data:
#Se obtiene una predicción utilizando el modelo lr.mod y los test data:
predicted.lr.2 <- predict(lr.mod, test)

#Se genera la matriz de confusión:
pf.lr.2 <- as.factor(ifelse(round(predicted.lr.2) == 1, "M", "F"))
confusionMatrix(data=pf.lr.2, reference=test$Sex)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#ANN:
#PROCESADO DATOS:
#Normalización: Max-Min
#Se define la siguiente función para normalizar los datos siguiendo este método:
normalize <- function(x) {
  return ((x - min(x)) / (max(x) - min(x)))
}

#Se normalizan los datos y se guardan en el data.frame nnet_femur:
nnet_femur <- as.data.frame(lapply(femurLNNA[, -1], normalize))

#Binarización de la variable categórica Sex:
#Ya que la función neuralnet() no permite argumentos factoriales
```

```
#para las variables utilizadas en el modelo
nnet_femur <- cbind(nnet_femur, femurLNNA$Sex == "M")
names(nnet_femur)[14] <- "M"

nnet_femur <- cbind(nnet_femur, femurLNNA$Sex == "F")
names(nnet_femur)[15] <- "Fe"

nnet_femur <- cbind(nnet_femur, femurLNNA$Sex)
names(nnet_femur)[16] <- "Sex"

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#TRAINING y TEST DATA:
#Hay que volver a obtener un training y un test data set con las variables binarizadas
#Pero como la semilla, fijada con set.seed, es la misma
#Los datos seleccionados serán los mismo que para el modelo LR
set.seed(88888)
submuestra <- sample(1:nrow(nnet_femur), size = 30, replace = FALSE)
submuestra
nnet_test <- nnet_femur[submuestra,]
nnet_training <- nnet_femur[-submuestra,]

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#ANN:
#Obtención del modelo mediante la función neuralnet
#Para consultar dudas sobre los argumentos de esta función
#puede utilizarse el comando: ?neuralnet
#En este caso:
#1. La función se corresponde con: las variables M y Fe en función del resto
#2. El conjunto de datos: todos los datos del data.frame nnet_training
#menos la última columa, ya que se corresponde con la variable Sex (sin binarizar).
#3. El argumento hidden, hacer referencia al numero de capas ocultas y sus nodos.
#En este caso tenemos tres capas ocultas con 13, 10 y 3 nodos cada una, respectivamente.
#4. El argumento rep, se corresponde con el número de ANNs que se quieren ejecutar,
#20 en este caso.
#5. El argumento act.fct en este caso es 'logistic'.
#Para la correcta comprensión de este argumento, se recomienda consultar su descripción
#con el comando ?neuralnet y la referencia bilbiográfica [11].
#6. El argumento linear.output es igual a FALSE,
#ya que queremos que se aplique el argumento anterior.
library(neuralnet)
ann <- neuralnet(M+Fe~FemMxLngL+FemBiConLngL+FemTrocLngL+FemSubTrAPDiaL+FemSubTrMLDiaL+
                   FemAPDiaMidL+FemMLDiaMidL+FemHeadSIDiaL+FemAPLatCondL+FemHeadHzDiaL+
                   FemAPMedCondL+FemEpicBrL+FemCircMidL,
                 nnet_training[, -c(16)],
                 hidden = c(13, 10, 3),
                 rep = 20,
                 act.fct = "logistic",
                 linear.output = FALSE)

#Con esta función se puede ver la estructura de la ANN y las puntuaciones
#de cada interacción:
#plot(ann, rep="best")

#La siguiente función nos permite obtener un gráfico con la estructura de la ANN
#pero sin las puntuaciones de las interacciones:
library(NeuralNetTools)
plotnet(ann, rep="best", y_names="Sex", alpha=0.9, circle_col = "darkturquoise")

#Poder predictivo, curva ROC:
#Predicciones para training data:
output.ann.1 <- compute(ann, nnet_training[, c(1:13)])
p.ann.1 <- output.ann.1$net.result

#Predicciones para test data:
output.ann.2 <- compute(ann, nnet_test[, c(1:13)])
p.ann.2 <- output.ann.2$net.result
```

```
#Se calcula la curva ROC:
require(pROC)
#Es necesario desactivar las librerías neuralnet y NeuralNetTools
#para que no interfieran con la librería pROC.
detach("package:neuralnet")
detach("package:NeuralNetTools")
rf.roc<-roc(nnet_training$Sex, p.ann.1[,1])

#Se grafica la curva ROC:
plot(rf.roc, col="violetred1", lwd=2, main="Curva␣ROC␣para␣ANN:␣Sexo")

#Se obtiene el valor AUC:
auc(rf.roc)

#Ratio de clasificación utilizando una matriz de confusión:
#training data:
library(e1071)
library(caret)
pf.ann.1 <- as.factor(ifelse(round(p.ann.1[,1]) == 1, "M", "F"))
confusionMatrix(data=pf.ann.1, reference=nnet_training$Sex)

#test data:
pf.ann.2 <- as.factor(ifelse(round(p.ann.2[,1]) == 1, "M", "F"))
confusionMatrix(data=pf.ann.2, reference=nnet_test$Sex)

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

#RANDOM FOREST:
#Se obtiene el modelo utilizando la función randomForest,
#los argumentos son:
#1. La fórmula: Variable Sex en fucnión de las demás.
#En este caso nuestra variable ya está en formato factor,
#pero en caso de que fuera una variable numérica habría que transformarla,
#ya que la función randomForest no admite variables numéricas
#para este argumento de la fórmula.
#2. El argumento ntree es de 600 árboles ha ejecutar.
#3. El argumento importance es igual a TRUE,
#ya que queremos que se evalúe la importancia de los predictores.
#Igual que en el caso de la ANN para comprender mejor los argumentos de esta función se
#recomienda ejecutar el comando ?randomForest.
library(randomForest)
rf <- randomForest(Sex ~ ., training, ntree = 600,
                   mtry=2, importance=TRUE)

#Se printa la salida del modelo:
rf

#Se obtiene el gráfico del error:
plot(rf)

#Se obtiene el gráfico de la importancia de las variables en el modelo:
varImpPlot(rf, col="darkgreen", main = "RF")

#Se obtiene la representación del árbol con la mejor predicción:
options(repos='http://cran.rstudio.org')
have.packages <- installed.packages()
cran.packages <- c('devtools','plotrix','randomForest','tree')
to.install <- setdiff(cran.packages, have.packages[,1])
if(length(to.install)>0) install.packages(to.install)

library(devtools)
if(!('reprtree' \%in\% installed.packages())){
  install_github('araastat/reprtree')
}

for(p in c(cran.packages, 'reprtree')) eval(substitute(library(pkg), list(pkg=p)))

library(reprtree)
reprtree:::plot.getTree(rf, k=500)
```

```
#Poder predictivo, curva ROC:
#Se calcula la curva ROC:
require(pROC)
rf.roc<-roc(training$Sex,rf$votes[,1])

#Se obtiene la gráfica para la curva:
plot(rf.roc, col="violetred1", lwd=2, main="Curva␣ROC␣para␣LR:␣Sexo")

#Se obtiene el valor AUC:
auc(rf.roc)

#Ratio de clasificación utilizando una matriz de confusión:
#training data:
library(e1071)
library(caret)
pred.rf.1 <- predict(rf, training[, c(2:14)])
confusionMatrix(data=pred.rf.1, reference=training$Sex)

#test data:
pred.rf.2 <- predict(rf, test[, c(2:14)])
confusionMatrix(data=pred.rf.2, reference=test$Sex)
```

11.2 ANEXO II: Descripción de las variables

Para consultar más información sobre las variables se pude consultar la página web de la base de datos *online* *The Terry Collection Postcranial Osteometric Database*. Donde se puede descargar un archivo *PDF* con toda la información de las variables.

VARIABLE	DESCRIPCIÓN
FemMxLng	**Longitud máxima**. Distancia desde el punto más superior en la cabeza del fémur hasta el punto más inferior en los cóndilos distales.
FemBiConLng	**Longitud Bicondilar**. Distancia desde el punto más superior en la cabeza a un plano dibujado a lo largo de las superficies inferiores de los cóndilos distales.
FemTrocLng	**Longitud trocantérica**. La mayor distancia entre el borde superior del trocánter mayor y el cóndilo lateral.
FemSubTrAPDia	**A-P (sagital) diámetro subtrocantérico.** Distancia entre las superficies anterior y posterior en el extremo proximal de la diáfisis, medida perpendicular al diámetro medial-lateral.
FemSubTrMLDia	**M-L (transversa) diámetro subtrocantérico**. Distancia entre las superficies medial y lateral del extremo proximal de la diáfisis en el punto de su mayor expansión lateral por debajo de la base del trocánter menor.
FemAPDiaMid	**A-P (sagital) diámetro del eje medio**. Distancia entre las superficies anterior y posterior medida en el punto medio de la diáfisis, en la elevación más alta de la línea áspera.
FemMLDiaMid	**M-L (transversa) diámetro del eje medio**. Distancia entre las superficies medial y lateral en el eje mediano, medida perpendicular al diámetro anteroposterior.
FemHeadSIDia	**Diámetro máximo vertical de la cabeza**.El mayor diámetro superior inferior perpendicular al plano que pasa por el eje del cuello.
FemHeadHzDia	**Diámetro máximo horizontal (trnasverso) de la cabeza**. El diámetro máximo, medido perpendicular al diámetro vertical de la cabeza.
FemAPLatCond	**A-P diámetro lateral del cóndilo**. La distancia proyectada entre el punto más posterior en el cóndilo lateral y el labio de la superficie rotuliana tomada perpendicularmente al eje del eje.
FemAPMedCond	**A-P diámetro medial del cóndilo**. La distancia proyectada entre el punto más posterior en el cóndilo medial y el labio de la superficie rotuliana tomada perpendicular al eje del eje.
FemEpicBr	**Amplitud epicondilar**. La mayor amplitud de los puntos más sobresalientes de los epicóndilos, paralelos al plano infracondíleo.
FemBiConBr	**Amplitud bicondilar**. La amplitud más grande a través de los cóndilos (anchura condilar transversal) tomada en un punto en el medio de cada cóndilo (posteriormente).
FemNeckDia	**Diámetro vertical del cuello**. Diámetro vertical mínimo del cuello femoral.
FemCircMid	**Circunferencia de la parte media**. Circunferencia medida al nivel de los diámetros del eje central. Si la línea aspera exhibe una fuerte proyección que no se expresa uniformemente en una gran parte de la diáfisis, entonces esta medida se registra aproximadamente a 10 mm por encima del eje central.

Cuadro 14: Descripción de las variables.