Uso de datos craneométricos para la estimación del perfil biológico mediante *machine learning*

Autor: Manuel Jesús Oneto Fernández

Máster Universitario en Bioinformática y Bioestadística Antropología biológica

Tutor: Xavier Jordana Comin

PRA: David Merino Arranz

Febrero a junio de 2019

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Uso de datos craneométricos para la estimación del perfil biológico mediante "machine learning"*
Nombre del autor:	*Manuel Jesús Oneto Fernández*
Nombre del consultor:	*Xavier Jordana Comin*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega:	*06/2019*
Titulación::	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Antropología biológica*
Idioma del trabajo:	*Español*
Palabras clave	*Antropología forense; "machine learning"; perfil biológico*

Resumen del Trabajo*:*

La estimación del perfil biológico, y concretamente del sexo y el origen poblacional, es una de las tareas más relevantes a las que se enfrenta habitualmente la antropología forense. Estas tareas se realizan habitualmente con técnicas estadísticas convencionales, como el análisis linear discriminante o la regresión logística, pero los métodos de *machine learning* no han sido tan extensamente estudiados, habiendo demostrado sin embargo poder ofrecer modelos más precisos en algunos casos.

Los objetivos de nuestro trabajo han sido seleccionar un conjunto de técnicas de *machine learning* adecuadas para la posterior obtención de modelos predictivos para la estimación del sexo y el origen poblacional mediante datos craneométricos. Tras realizar una descripción de un conjunto de técnicas tanto convencionales como de *machine learning*, se seleccionaron finalmente como

métodos Bayes ingenuo, *Random forest* y redes neuronales artificiales, obteniéndose a partir de estos algoritmos modelos predictivos precisos para la tarea que nos propusimos, destacando en nuestro caso el modelo de redes neuronales artificiales.

Los métodos de *machine learning* son una alternativa factible a los métodos convencionales de construcción de modelos predictivos dentro del campo de la antropología forense. Aunque son más complejos, en algunos casos pueden llegar a tener una precisión mayor y por tanto deben ser tenidos en cuenta, sobre todo dada la trascendencia de la estimación correcta del perfil biológico.

Abstract:

Biological profile estimation, in particular sex and ancestry estimation, is one of the most relevant task that forensic anthropologists face. This task is usually performed with conventional statistical methods, like lineal discriminant analysis or logistic regression, but machine learning methods have not been thouroughly studied, although they have demostrated to be able to build in some cases models with higher accuracy.

Our aim was firstly to select a set of appropiate machine learning techniques to be able in second place of obtaining predictive models for the estimation of sex and ancestry by means of craniometric data. After describing both conventional and machine learning methods, we selected naïve Bayes, random forest and artificial neural networks as suitable to perform our task. Predictive models were obtained using this methods, with a special mention to the model obtained with artificial neural networks.

Machine learning is a feasible alternative to statistical conventional methods for the building of predictve models in the field of forensic anthropology, in particular in sex and ancestry estimation. They are more complex, although in some cases they may have higher accuracy. Therefore, they must be accounted, considering the importance of correct estimation of biological profile.

Índice

Índice de figuras

Índice de tablas

1. Introducción

1.1 Justificación

La antropología forense es una disciplina que aplica los conocimientos de la antropología biológica al proceso legal, principalmente en lo referido al estudio del esqueleto humano [1].

El conocimiento sobre el crecimiento, desarrollo, degeneración y variación esquelética permite obtener información sobre la edad, el sexo, el origen poblacional y la estatura de un individuo en el momento de su fallecimiento; lo que colectivamente se conoce como perfil biológico. Esto suele revestir especial relevancia para asistir en la identificación del fallecido y proveer si fuera posible con información sobre la circunstancias de su muerte, con el objeto de apoyar una investigación criminal o de otra índole [2].

Uno de los problemas más habituales es el grado de integridad y preservación en el que se encuentran los restos esqueléticos a partir de los cuales se estima el perfil biológico, lo que supone la principal limitación en el grado de precisión y exactitud de la estimación [3]. Los restos pueden encontrarse mezclados o estar quemados y fragmentados: fosas comunes, accidentes aéreos, explosiones… En muchas ocasiones no se dispone del esqueleto completo, por lo que es necesario poder realizar las estimaciones a partir de diferentes elementos esqueléticos.

La estimación del sexo y el origen poblacional tiene importancia por sí misma a la hora de construir el perfil biológico, pero además muchas de las estimaciones que se realizan posteriormente dependen de estos parámetros [3]. La valoración del sexo y el origen poblacional de un individuo se ha realizado clásicamente analizando características morfoscópicas del cráneo y la pelvis [4]. Sin embargo, la estimaciones basadas en métodos métricos son menos subjetivas y tienen

menor variabilidad intraobservador e interobservador, por lo que son de elección [5].

Dentro de la múltiples opciones existentes, los datos craneométricos se considerar habitualmente como uno de los mejores elementos del esqueleto humano para a estimación del sexo y el origen poblacional, aunque los elementos poscraneales han demostrado ser también de mucha utilidad para este fin [6].

Los métodos clásicos más usado para la estimación del sexo y el origen poblacional a partir de datos métricos son el análisis discriminante lineal y la regresión logística [7][8]. Sin embargo, en los últimos años se ha comenzado a usar un grupo de técnicas englobadas en el término *machine learning* al campos de la antropología forense para la construcción de modelos predictivos [3][9][10].

Se conoce como aprendizaje automático o *machine learning* (ML) a un campo de la computación que estudia los algoritmos y modelos estadísticos usados por los ordenadores para realizar una tarea específica, sin haber sido explícitamente programados para ello, permitiendo la mejoría a través de la experiencia. Es especialmente útil para el hallazgo de forma automática de patrones presentes en bases de datos con un gran número de muestras y variables, así como el hallazgo de algoritmos donde no existe un conocimiento desarrollado o que está en constante cambio. Implica el cálculo de una función objetivo, aprendida a partir de una experiencia de entrenamiento. Una vez obtenido el modelo, permite el cálculo de la función objetivo al entregarle al mismo un nuevo ejemplo [11].

A la hora de escoger un modelo predictivo, debemos tener en cuenta varios factores: supuestos, precisión, exactitud y simplicidad.

El análisis discriminante lineal asume normalidad multivariante y homocedasticidad de los datos, es decir, el valor constante de la varianza a lo largo de todo el rango de valores de las variables. Es además sensible a la

presencia de valores atípicos y a la multicolinealidad de los datos. Estas desventajas pueden ser paliada parcialmente mediante la regresión logística [9].

Los métodos de ML han demostrado ser capaces en algunos casos de construir modelos predictivos con mayor exactitud que los métodos estadísticos anteriormente mencionados. Cuentan con la ventaja además de no hacer ninguna asunción sobre los datos de trabajo. Como desventaja, algunos de sus métodos tienen mayor complejidad o costes computacionales elevados [3]. Algunos de los métodos usados son k-vecinos más cercanos, Bayes ingenuo, redes neuronales artificiales, árboles de clasificación y regresión o los modelos de conjunto.

La estimación del perfil biológico es una tarea de gran importancia en el campo de la antropología forense. El resultado de esta estimación puede orientar o incluso dar las claves de investigaciones en contexto de homicidios, desapariciones, crímenes de guerra, accidentes, atentados terroristas... Las consecuencias de un error en la clasificación de un individuo pueden ser muy graves. Es por ello que es de interés científico, social y judicial contar con las mejores herramientas posibles para la estimación del perfil biológico, lo que implica la creación de modelos predictivos precisos, exactos y simples en la medida de los posible, para lo que las técnicas de ML han demostrado una gran utilidad en otros campos y también dentro de la antropología forense.

1.2 Hipótesis y objetivos

Hipótesis de trabajo: los algoritmos de ML tienen la capacidad de crear modelos predictivos más precisos a la hora de estimar el sexo y el origen poblacional que los métodos estadísticos convencionalmente usados.

Objetivo general 1: Seleccionar los métodos de ML más adecuados para la construcción de modelos predictivos para la estimación del sexo y el origen poblacional a partir de datos craneométricos:

- **Objetivo específico 1.1:** Describir los métodos estadísticos convencionales.

- **Objetivo específico 1.2:** Describir los métodos de ML.

- **Objetivo específico 1.3:** Evaluar y comparar los métodos estudiados.

Objetivo general 2: obtener modelos predictivos para la estimación del sexo y el origen poblacional a partir de datos craneométricos:

- **Objetivo específico 2.1:** Analizar los datos craneométricos necesarios.

- **Objetivo específico 2.2:** Aplicación de los métodos seleccionados a los datos.

1.3 Método y planificación

Para desarrollar nuestro trabajo, es importante usar una metodología adecuada, y esto está en relación con las herramientas usadas. Es necesario que estas herramientas cuenten con la versatilidad y potencia suficientes para poder llevar a cabo todas las tareas que no proponemos para una correcta consecución de los objetivos:

- Bases de datos y buscadores para la revisión bibliográfica necesaria (mencionadas posteriormente en la tarea 1).

- El software estadístico necesario para el desarrollo de los modelos, en concreto *R* en su versión 3.6.0 "Planting of a Tree" y con ayuda de la interfaz gráfica *R* Studio en su versión 1.1.463. Ambos programas en su versión para Windows 10.

- El lenguaje de edición de documentos *LaTeX*, mediante *TeXworks* y el sistema de gestión de bibliografía integrado *JabRef*.

Para realizar de forma adecuada la planificación del trabajo, hubimos de considerar de manera realista las tareas a realizar, considerando su complejidad y dedicación necesaria y la dependencia que existe entre ellas. Todo ello, teniendo el cuenta el marco temporal en el que realizamos el proyecto y la duración que se nos exigía:

- El marco temporal del proyecto comprende el rango de fechas entre el 19/3/2019 y el 25/6/19, con una dedicación estimada de 300 horas.

- Consideramos las pruebas de evaluación continua (PEC) establecidas en el calendario de la asignatura como hitos, para establecer un marco en el que desarrollar las tareas necesarias: PEC2 (desarrollo del trabajo, fase 1), PEC3 (desarrollo del trabajo, fase 2), PEC4 (cierra de la memoria), PEC5a (elaboración de la presentación) y PEC6b (defensa pública).

- Consideramos nuestra capacidad de dedicación acorde a nuestras circunstancias personales y nuestro conocimiento y competencias previos sobre el tema del trabajo. Con esto, y teniendo en cuenta los hitos y la complejidad de las tareas y sus características, establecimos una temporalización de las mismas.

- Dentro de la PEC2, englobamos las siguientes tareas:

    - Tarea 1: consulta y lectura de bibliografía relacionada. Para la consecución del OG1 y en sus subapartados OE1.1, OE1.2 y OE 1.3, era fundamental una búsqueda extensa y profunda de bibliografía relacionada con antropología forense, métodos estadísticos clásicos y métodos de ML. Usamos herramientas para esta tarea como *Scopus,*

Pubmed, Google Academy y la biblioteca de la UOC. En nuestro caso, ya contábamos con competencias en la realización de búsqueda bibliográfica ya esta tarea se completó parcialmente durante las etapas principales del trabajo. Contábamos con formación estadística y básica en ML, y aunque los conceptos de antropología forense no eran totalmente nuevos por la formación previa recibida en anatomía humana y medicina legal, sí que han requerido una lectura más concienzuda. Es por ello que se le asignó a esta tarea una duración de una semana.

- Tarea 2: definición de las técnicas de estadística clásica y de ML. Esta tarea permitió la consecución del OE 1.1 y OE 1.2. Es fundamental tener una comprensión correcta de las técnicas estadísticas usadas en la construcción de modelos predictivos en antropología forense, ya que son las herramientas que utilizamos para la consecución del OG2. Fue necesario también una síntesis adecuada de estas técnicas para la comprensión del trabajo por personas más legas en la materia. Se le asignó a esta tarea una duración de dos semanas.

- Tarea 3: evaluación y comparación de los métodos. Para la consecución del OE1.3 fue necesario un adecuado entendimiento de los métodos y de sus respectivas ventajas y desventajas. De esta comprensión emana la justificación del uso de ML aplicado a la antropología forense y por tanto de nuestro trabajo. Además, seleccionamos los métodos específicos más adecuados para nuestro trabajo, por lo que fue una de las tareas más críticas para la correcta consecución del resto de objetivos. Se le asignó una duración de dos semanas.

- La realización de las siguientes tareas implicó la realización de la Tarea 3, engoblándose dentro de la PEC3:

- Tarea 4: descarga y estudio de datos. Los datos necesarios para nuestro trabajo los obtuvimos de bases de datos de libre acceso, en concreto "The William W. Howells Craniometric Data Set"[12][13][14][15]. El conjunto de datos craneométricos de Williams W. Howells consiste en una serie de hasta 82 medidas obtenidas de 2524 cráneos humanos de 30 poblaciones diferentes, abarcando todos los continentes. Los cráneos contienen además una estimación del sexo relacionada por el propio autor. Howells recopiló estos datos entre 1965 y 1980, y los publicó en tres monografías, compartiéndolos posteriormente con la comunidad científica.Esta es una tarea fundamental, porque además de la descarga es necesario además familiarizarse con las variables, su significado, tipo, valores, cómo están codificados los datos faltantes... Es por ello que le asignamos una duración de una semana.

- Tarea 5: desarrollo del código y obtención de los modelos. Es la tarea más importante del trabajo, pues de ella depende la obtención de resultados útiles que justificación de este trabajo. Necesaria para la consecución del OE 2.1 y OE 2.2.

A pesar de que existe un período de 10 semanas desde el comienzo hasta el final previsto del trabajo, hemos ajustado la duración de las tareas a 9 semanas. Esto nos permitirá dar cierta flexibilidad a la duración de las tareas en caso de que fuera necesario, lo que es muy probable teniendo en cuenta que por mi oficio a veces existe períodos de aumento del número de horas laborales. Todo con el objetivo de tener una planificación realista del proyecto.

- Tarea 6: redacción de la memoria. En relación con la PEC4. Aunque no referido a ningún objetivo concreto, esta tarea es necesaria para sintetizar y comunicar correctamente todo nuestro trabajo en un solo documento. Se le asignó una duración de dos semanas.

- Tarea 7: elaboración de la presentación. En relación con la PEC5a. Un paso necesario para comunicar de forma sucinta y gráfica los contenidos de la memoria. Duración de una semana.

- Tarea 8: defensa del trabajo: en relación con la PEC5b. Es el paso y final y supone exponer finalmente los resultados de nuestro trabajo ante el tribunal, con el objetivo de evaluar su calidad científica, así como de encontrar posibles áreas de mejora y direcciones futuras de la investigación. Duración de dos semanas.

En la Figura 1 podemos apreciar una representación gráfica de la planificación mediante un diagrama de Gantt, desarrollado mediante el paquete *pgfgantt* del software de edición de documentos *LaTeX*:

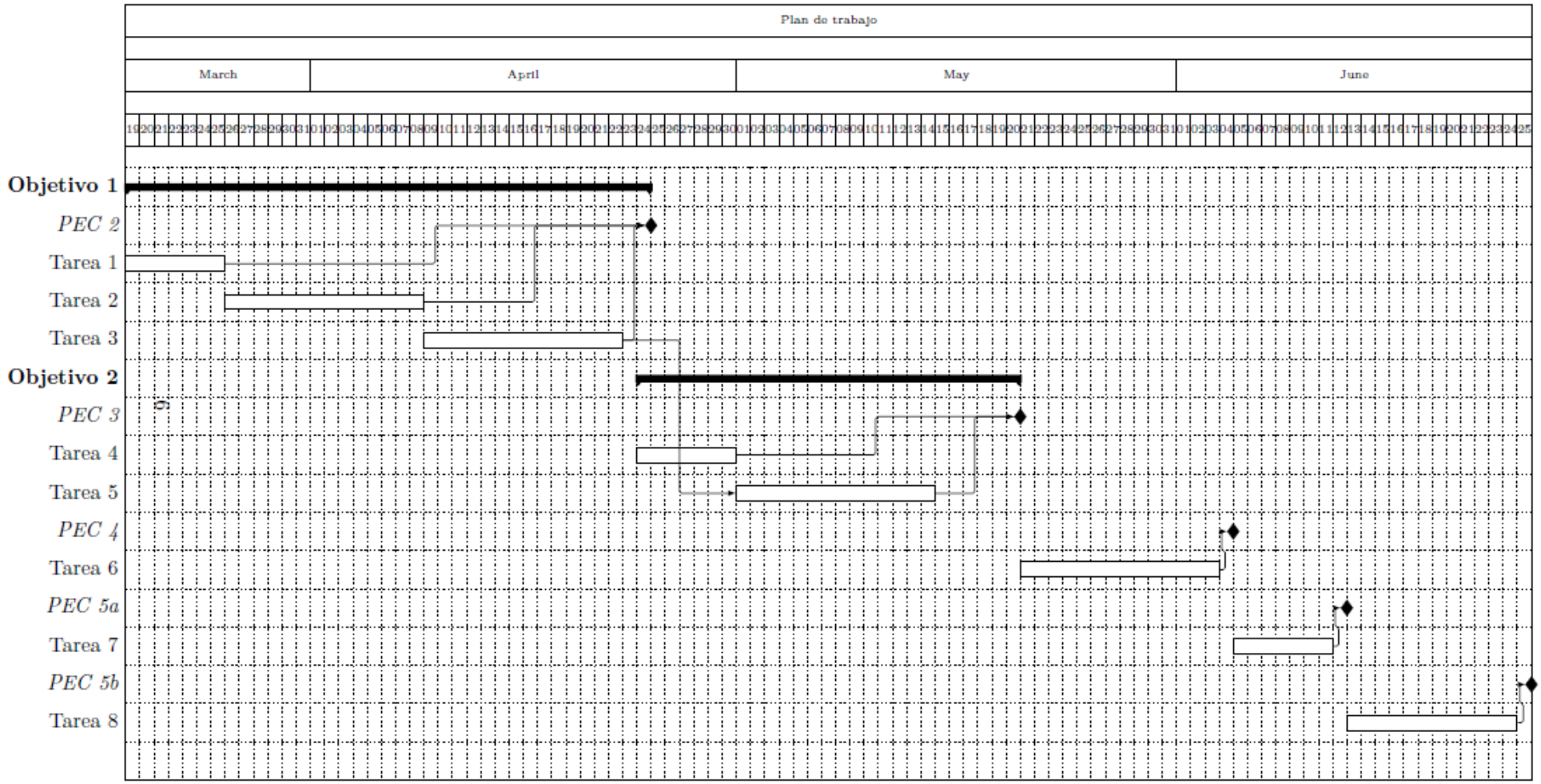

Figura 1: Diagrama de Gantt representativo de la planificación del trabajo

1.4 Resultados esperados

Como producto de este trabajo, hemos obtenido los siguientes resultados:

- Documento en el que se describe el trabajo y su justificación, los objetivos y la tareas a realizar y la temporalización.

- Monografía sobre los diferentes métodos existentes para la obtención de modelos predictivos aplicados al campo de la antropología forense, resaltando las diferencias entre los métodos habituales y los métodos de ML.

- Modelos predictivos obtenidos mediante *R* para la estimación del sexo y el origen poblacional a partir de datos craneométricos, junto con el código de programación usado para obtener dichos modelos.

- Presentación gráfica del trabajo que permita la fácil difusión de los contenidos del mismo.

- Documento de autoevaluación.

1.5 Descripción de los otros capítulos

En el capítulo 2 se realiza una descripción del concepto de ML, describiendo un conjunto de técnicas relacionadas con aplicabilidad en el trabajo que nos ocupa. Se realiza también una descripción de las técnicas estadísticas convencionales, realizando finalmente una selección de los métodos que vamos a aplicar.

También se estudian los datos de trabajo y se aplican los métodos seleccionados con anterioridad a dichos datos para la obtención de los modelos predictivos para la estimación del sexo y el origen poblacional a partir de datos craneométricos.

En el capítulo 3 se realiza la discusión de los resultados del trabajo, comparándolos con la literatura existente.

En el capítulo 4, se entregan las conclusiones de nuestro trabajo, incluyendo qué hemos aprendido, el grado de cumplimiento de los objetivos propuestos, el seguimiento de la metodología y la planificación y áreas futuras de desarrollo.

En el capítulo 5, se muestra la bibliografía en la que nos hemos basado para desarrollar nuestro trabajo.

En el capítulo 6, se anexa información complementaria del trabajo, como es el código de programación usado para su realización y una descripción de las variables con las que hemos trabajado.

2. Resultados

2.1 *Machine learning*

El aprendizaje automático o ML es un campo de la computación relacionado con la inteligencia artificial que estudia los métodos por los cuales un sistema informático realiza una tarea específica de forma efectiva, sin haber sido explícitamente programado para ello. Estos métodos deben permitir además su mejoría a través de la experiencia. Formalmente, se dice que un sistema informático es capaz de aprender a realizar una tarea si el rendimiento de dicho sistema en desarrollar la tarea mejora con la experiencia [11].

Las tareas que desarrollan estos sistemas son innumerables, pero las más comunes incluyen la clasificación, regresión y agrupamiento de datos de diversa índole.

El término fue acuñado en 1959 por Arthur L. Samuel mientras se encontraba en IBM, donde desarrolló un programa informático capaz de aprender a jugar a las damas [16]. Sin embargo, no sería hasta la década de los 90 cuando el ML se instauraría como un campo propio del conocimiento, gracias principalmente al desarrollo de la computación con el consiguiente aumento de la capacidad de cálculos de los computadores e Internet.

La experiencia de aprendizaje que se entrega al sistema informático consiste en conjunto de datos con unas características determinadas, conocido como datos de entrenamiento. En función de estas características, podemos distinguir tres tipos de aprendizaje [17]:

- Aprendizaje supervisado: la experiencia de aprendizaje del sistema contiene una serie de ejemplos, cada uno de ellos compuesto por unas

variables de entrada (que normalmente toman la forma de un vector de tantas dimensiones como número de variables) y una variable de salida asignada manualmente. Se dice por tanto que se tratan de ejemplos etiquetados. A partir de estos ejemplos, el algoritmo de aprendizaje infiere una función capaz de generalizar un valor de salida para unos valores de entrada nuevo no etiquetados entregados al algoritmo. Si la variable de salida es cualitativa, el sistema será capaz de clasificar el nuevo ejemplo, y si es cuantitativa, el sistema será capaz de realizar regresión sobre dicha variable de salida.

- Aprendizaje no supervisado: al contrario que en el caso anterior, los ejemplos solo contienen variables de entrada y no se encuentran etiquetados. Por tanto, la función del algoritmo será encontrar patrones subyacentes en la estructura de los datos que permita agrupar ejemplos similares y separar los diferentes, lo que se conoce como análisis de clusters. Otra posibilidad es este campo es la reducción de la dimensionalidad.

- Aprendizaje por refuerzo: en este tipo de algoritmos, el sistema se comporta dentro de un entorno en función de los resultados que provoca dicho comportamiento, adaptándose para conseguir el mejor resultado posible.

En el trabajo que nos ocupa, uno de los objetivos que nos hemos marcado es la construcción de modelos mediante ML que sean capaces de predecir el sexo y el origen poblacional de individuos a partir de datos craneométricos. Por tanto, los algoritmos de aprendizaje supervisado, capaces de realizar clasificación, son los adecuados para nuestra tarea.

Dentro del campo de ML, existen múltiples algoritmos clasificadores que serían capaces de enfrentarse al problema expuesto anteriormente. Sin embargo, el

rendimiento de cada algoritmo para un problema concreto es habitualmente diferente, siendo la única manera de que sea mejor si está especializada para dicho problema específico, lo que se desprende del teorema NFL [18].

Por tanto, sin ánimo de ser exhaustivo, en apartados posteriores daremos las bases y definiremos algunos de los algoritmos de clasificación más ampliamente usados y los que pensamos que por las características de nuestro problema son más adecuados.

2.1.1 Bayes ingenuo

El método de Bayes ingenuo o *Naïve Bayes* (NB) es una familia de algoritmos de clasificación de ML basados en modelos probabilísticos. Una de las aplicaciones más frecuentes de estos algoritmos se encuentra en la clasificación de texto, aunque sus usos no se quedan aquí y se han extendido a varios campos. Estos métodos se basan en el teorema de Bayes, es decir, en el hecho de que la probabilidad *a posteriori* de un evento se define de manera más precisa a medida que vamos conociendo más información [19].

El adjetivo de ingenuo se da porque una fuerte asunción de este teorema es que las variables predictoras de los sujetos son independientes entre sí. Sin embargo, a pesar de importantes violaciones de esta asunción, estos métodos han mostrado pese a su simplicidad una precisión similar o incluso superior en ocasiones a métodos muchos más complejos estadísticos o de ML [20]. Se ha observado que a pesar de que el cálculo de probabilidades de un sujeto para las diferentes clases dispuestas es a menudo erróneo, el algoritmo acierta siempre que la clase correcta sea al fin y al cabo la que cuente con mayor probabilidad *a posteriori*.

2.1.2 Árboles de decisión

Los árboles de decisión o *decision trees* (DT) son un grupo de algoritmos no paramétricos de ML, por lo que por definición no realizan asunciones sobre la estructura de los datos a estudio. Son útiles dentro del campo del aprendizaje supervisado tanto para regresión como para clasificación. Han tenido una alta difusión debido a la facilidad de entendimiento e interpretación de sus resultados, lo que que unido a lo anteriormente expuesto lo convierten en un interesante grupo de algoritmos de ML.

La estructura básica del algoritmo es un nodo de decisión, que aplica cierta prueba al sujeto en función de la información otorgada por las variables predictoras y divide el árbol en tantas ramas como resultados posibles de la prueba. Esta estructura básica se repite múltiples veces hasta alcanzar una hoja, que en el caso de los árboles de clasificación se corresponde con la clase predicha por el algoritmo. El algoritmo utiliza parámetros como la impureza de Gini, la ganancia de información o la reducción de la varianza para elegir la prueba que mejor divide los datos en cada caso [19].

Dado el volumen y diversidad habitual de los datos que se entregan a este tipo de algoritmos, la precisión de los algoritmos basados en árboles de decisión habitualmente no es muy elevada y tiene problemas de sobreajuste, categorizándose habitualmente dentro del concepto de aprendiz débil o *weak learner*, es decir, que es capaz de clasificar mejor que la asignación al azar de las clases pero cuyos resultados solo están ligeramente correlacionados con la verdadera clasificación.

Para superar las limitaciones anteriormente mencionadas, se han desarrollado estrategias conocidas como métodos de aprendizaje de conjunto o *ensemble learning* (EL) como el *boosting* y el *bagging* dentro de los que se engloba un algoritmo de árboles de decisión de interés particular en el presente trabajo

conocido como bosques aleatorios o *random forest* (RF). El RF consiste en la construcción de múltiples árboles de decisión individuales, construidos a partir de muestras aleatorias con reemplazo de los datos de entrenamiento (es decir, cada sujeto de entrenamiento puede ser seleccionado más de una vez para el mismo árbol). Las variables predictoras con las que se realiza la prueba en cada nodo de decisión también es seleccionada aleatoriamente, eligiéndose el punto de corte según los parámetros mencionados anteriormente. Al entregar un nuevo caso al RF, cada árbol emite una clasificación, siendo asignado finalmente el sujeto a la clase que haya votado por mayoría simple el algoritmo [21].

Este método cuenta con enormes ventajas con respecto a otros:

- Evita los problemas de sobreajuste típicos de estos algoritmos.

- Permite la estimación del error de clasificación con los datos de entrenamiento, por lo que elimina la necesidad de validación externa del algoritmo.

- La capacidad de cómputo necesaria es menor a otros métodos de EL con similar rendimiento.

- El rendimiento es independiente de la transformación realizada a los datos.

- No precisa de muchos parámetros ajustables por el usuario, no estando muy correlacionado el rendimiento final del algoritmo con dichos parámetros.

Este algoritmo ha sido usando con éxito en el campo de la antorpología forense para la estimación del sexo y el origen poblacional [22][23].

2.1.3 Máquinas de vectores de soporte

Las máquinas de vectores de soporte o *support vector machines* (SVM) son un conjunto de algoritmos de ML en el campo del aprendizaje supervisado que se basa en el cálculo de hiperplanos que separen de forma óptima sujetos en un espacio multidimensional pertenecientes a diferentes clases. Ante la entrega de un sujeto no etiquetado al algoritmo, se realiza una clasificación en función de la posición relativa del hiperplano a dicho sujeto [19].

Este método tiene habitualmente buen resultado cuando las variables predictoras tienen relaciones no lineales entre sí y ha sido poco estudiado dentro del campo de la antropología forense, probablemente por el alto grado de correlación que suelen presentar las variables en este campo.

2.1.4 Redes neuronales artificiales

Las redes neuronales artificiales o *artificial neural networks* (ANN) son estructuras compuestas por unidades funcionales conocidas como neuronas artificiales, en las que se pueden ejecutar múltiples algoritmos de ML []. Son uno de los métodos de ML más ampliamente estudiados y han sido aplicados en el caso concreto de la antropología forense, por lo que cuentan con especial interés.

El funcionamiento de las neuronas artificiales está basado en el funcionamiento de las neuronas biológicas:

- La neurona artificial recibe unos valores de entrada o *input*, bien directamente de los datos o bien de otras neuronas artificiales. Dichos input toman la forma de un vector de valores, cada uno de ellos con unos

pesos asignados. Las dendritas de las neuronas biológicas realizan una función similar.

- Una vez recibido el *input*, la neurona ejecuta una función conocida como función de activación con dicho *input*. Dependiendo de la función de activación elegida, la señal recibida por la neurona es procesada de determinada forma. Las ANN pueden estar compuestas por neuronas cuyas funciones de activación sean diferentes. El mecanismo biológico equivalente es la tasa de disparo del potencial de acción.

- Aplicada la función de activación determinada sobre el *input*, se obtiene un valor de salida o *output*. Este valor es propagado a otras neuronas artificiales de la red, componiendo el input de estas otras neuronas, o bien como resultado final del algoritmo. Es equivalente al axón de la neurona biológica.

Matemáticamente, la neurona biológica toma la forma de una función. Existen unas variables independientes, la función propiamente dicha, y una variable independiente, que se transmite como variable independiente hacia otras funciones [19].

Estas neuronas artificiales se estructuras en redes, que son las ANN propiamente dichas. Existen múltiples algoritmos que son aplicables a las ANN, pero el estudio de todos ellos se sale del alcance de esta revisión. Por este motivo nos centraremos en la redes alimentadas hacia delante con retropropagación o *feed-forward backpropagation*. Esta clase de algoritmos es ampliamente utilizada y es especialmente útil en los casos de aprendizaje supervisado.

Para su funcionamiento, el algoritmo asigna unos pesos en principio aleatorios a las entradas de las neuronas. Una vez obtenido el *output*, se calcula el error cometido y se transmite a todas las capas de la red. Mediante el algoritmo de

optimización del descenso del gradiente, se halla un mínimo de la función y se ajusta iterativamente el valor de los pesos de las neuronas [10].

2.2 Métodos convencionales

2.2.1 Análisis discriminante lineal

El análisis discriminante lineal o *linear discriminant analysis* (LDA) es un método estadístico desarrollado a principios del siglo XX por Fisher que tiene el propósito de discriminar en grupos una población mediante una función lineal de medidas realizadas sobre dicha población [25]. Antes de la publicación formal del método, fue aplicado a sugerencia del propio Fisher en dos trabajos de otros autores que versaban precisamente sobre antropología forense [26][27].

El método consiste en el cálculo de la probabilidad de pertenencia de los sujetos a cada grupo mediante la medida de la distancia de Mahalanobis existente entre dicho sujeto y los centroides de los distintos grupos. Los centroides de los grupos son calculados gracias a la existencia de un conjuntos de sujetos para los cuales el dato de pertenencia al grupo es conocida *a priori*, siendo conocidos estos sujetos como datos de entrenamiento por este motivo. En función de dichas probabilidades, es posible estimar la pertenencia más probable de un nuevo sujeto a un grupo concreto [7].

El LDA realiza ciertas asunciones sobre los datos que no siempre se cumplen. Sin embargo, se ha demostrado que la precisión del método es relativamente robusta a la violación de estas asunciones:

- Normalidad multivariante: las variables siguen una distribución normal multivariante.

- Homocedasticidad: la varianza de las variables es constante en todo el rango de valores de dichas variables.

- Independencia: los valores de las variables son independientes entre diferentes sujetos.

Las cualidades de este método lo hacen idóneo en el campo de la antropología forense para la estimación del sexo y el origen poblacional, que en este caso serían las variables de grupo. Las variables predictoras serían en este caso cualquier combinación de datos osteométricos que es posible obtener de los elementos esqueléticos. Sin embargo, dada la alta carga de cálculo necesaria, este método no sería usado de manera extensiva hasta principios de los 60, gracias al trabajo de Giles y Elliot en 1962 [28]. El trabajo de estos autores de uso de manera extensiva hasta principios de los 90, momento en el que se desarrolló el software *Fordisc*.

Fordisc es un *software* desarrollado por Jantz y Ousley, en la Universidad de Tennessee. Este *software*, que actualmente se encuentra en su versión 3.1, permite la estimación del sexo, origen poblacional y estatura de un individuo mediante el uso de LDA, regresión lineal y una extensa base de datos conocida como *Forensic Data Bank* (FDB). Dicha base de datos contiene información sobre sujetos americanos nacidos en el siglo XX o XXI, con medidas craneales y poscraneales y datos de edad, sexo, origen poblacional, altura, peso y causa de muerte obtenidos de forma tanto *premortem* como *posmortem*. Actualmente contiene información de más de 4000 individuos. *Fordisc* no utiliza métodos novedosos, sino que presenta los métodos comentados anteriormente en una interfaz gráfica para un manejo de estas técnicas más sencillo.

La principal limitación de *Fordisc* se encuentra en que la población presente en el FDB incluye únicamente sujetos del continente americano, y por tanto es menos preciso a la hora de estimar el perfil biológico de otras poblaciones con la europea

o la asiática. Este hecho se ha mejorado en las últimas versiones con la inclusión, por ejemplo de la base de datos de Howells, que contiene sujetos de varios continentes. Sin embargo, aún sigue existiendo margen de mejora en este aspecto [29].

2.2.2 Análisis discriminante cuadrático

El análisis discriminante cuadrático o *quadratic discriminant analysis* (QDA) es un método similar al LDA. La principal diferencia radica en que la superficie de separación entre los grupo son más complejas, formando superficie cuadráticas en vez de líneas o plano como en caso del LDA. Al contrario que en el LDA, asume que no existe una matriz de varianzas-covarianza homogénea, sino que existe una diferente para cada grupo en que es clasificable el sujeto, por lo que este método es más apropiado que el LDA en los casos en que es muy significativa la diferencia entre las matrices de varianzas-covarianzas de los grupos.

Sin embargo, cuando el número de variables predictoras es alto la cantidad de parámetros estimados por el QDA es muy grande, por que la varianza tiende a ser muy alta y perder precisión, por lo que para lograr una precisión similar que en el LDA, es necesario un tamaño muestral significativamente mayor. Por tanto es crucial conocer la estructura de los datos con los que realizamos las estimaciones para ser capaces de seleccionar la mejor alternativa en cada caso, ya que la elección del LDA en un caso poco favorable disminuiría la exactitud de la estimación y por otra parte realizar un QDA cuando la asunciones del LDA se cumplen nos haría perder precisión [30].

2.2.3 Regresión logística

La regresión logística o *logistic regression* (LR) es un modelo estadístico que se basa en la función logística. Se utiliza para predecir el resultado de una variable

categórica (y en el caso de la regresión logística multinomial, el de una variable nominal) en función de los valores de una o más variables continuas o categóricas. Hace uso del método de máxima probabilidad para encontrar el modelo que mejor clasifique los datos.

Cuenta con la ventaja con respecto al LDA en que no asume que las variables predictoras sigan una distribución normal multivariante, por lo que se puede usar en casos en los que no sea posible aplicar este modelo o cuando la desviación de esta asunción sea lo bastante grande que sesgue significativamente la exactitud de sus precisiones. Se ha visto sin embargo que en los casos en los que los supuestos del LDA se aplican, este último modelo ha mostrado se más exacto y preciso que la regresión logística [31].

2.3 Selección de clasificadores

Habiendo revisado algunas de las técnicas más ampliamente utilizadas en el campo de los algoritmos clasificadores, tanto basados en técnicas estadísticas como en ML, hemos de escoger cuáles son lo más adecuados para nuestro objetivo de construir modelos que permitan predecir el sexo y el origen poblacional a partir de datos craneométricos. Mostramos las técnicas elegidas y a continuación describimos los motivos:

- Ánalisis lineal discriminante: es probablemente el método más utilizados para la tarea que nos atañe dentro del campo de la antropología forense, habiendo sido estudiado a fondo y existiendo una extensa bibliografía al respecto. Por este motivo, creemos que cualquier clasificador dentro de la antropología forense debería compararse con el LDA, estableciendo una línea de base con la que los demás clasificadores puedan compararse.

- Bayes ingenuo: es una técnica de ML que cuenta con la ventaja de ser conceptual y computacionalmente simple, sin perder por esto rendimiento

en la clasificación, al haber demostrado que puede llegar a ser igual de precisa que otros métodos de ML más complejos como hemos comentado anteriormente. Por este motivo, creemos interesante construir un modelos mediante este método, para poder establecer si otros algoritmos de ML cuentan realmente con ventajas con respecto a un método más simple.

- Redes neuronales artificiales: son probablemente el método más utilizado de ML, por lo que existe una amplia bibliografía relacionada con este método y en concreto con su aplicación a la antropología forense.

- *Random forest*: este método no paramétrico ha demostrado tener un alto rendimiento como clasificador. Dentro de la antropología forense existe cierta bibliografía respecto al uso de este método, pero sin duda no tan extensa como con otros métodos, por lo que creemos que su estudio puede ofrecer aspectos novedosos en este campo.

2.4 Modelos clasificadores

2.4.1 Datos de trabajo

Pasamos a continuación a analizar el conjunto de datos, en inglés data set (DS). Las variables del DS consisten en:

- Un código de identificación del cráneo (ID).

- Sexo estimado del cráneo (Sex).

- Población a la que pertenece el cráneo y un código numérico asignado a dicha población (PopNum y Population).

- Medidas craneométricas, en total 82 (descritas posteriomente en el Anexo 2).

El código de identificación del cráneo y de las poblaciones es superfluo para nuestro análisis y ha sido eliminado del DS definitivo.

La estimación lo más precisa posible del sexo es uno de los objetivos de los clasificadores que vamos a construir, por lo que esta variable es de suma importancia.

En cuanto a la población, otro de nuestros objetivos de nuestros clasificadores es estimar lo más precisamente posible el origen poblacional de los cráneos. Dado que la eficacia de los clasificadores es menor cuando el número de niveles de la variable a estimar es alto, hemos decidido crear otra variable que describa el continente de origen de la población, reduciendo el número de niveles a predecir de 30 a 5. La nueva variable es Continent y sus niveles son: Africa (bosquimanos, dogones, egipcios y zulúes), America (arikara, esquimales, peruanos y santacruceros), Asia (ainu, andamaneses, de Anyang, atayal, buryats, hainan, del norte y del sur de Japón), Europe (de Berg, nórdicos y de Zalavaer) y Oceania (australianos, rapanui, guameños, de Mokapu, moriori, maoríes del norte y del sur, filipinos, palawa y tolai).

Las variables craneométricas son todas variables cuantitativas, obtenidas a partir de medidas directas sobre los cráneos, en total 82 diferentes como hemos comentado anteriormente. A partir de ellas vamos a desarrollar los clasificadores que nos permitan estimar el sexo y el origen poblacional de los individuos, por lo que son de suma importancia.

Adjuntamos en Tabla 1 una relación del número de cráneos que pertenece a cada clase que queremos estimar mediante nuestros modelos:

	África	América	Asia	Europa	Oceanía
Hombre	234	201	359	164	410
Mujer	250	188	256	153	309

Tabla 1: Distribución de la muestra en las variables de interés

Pasamos a continuación a analizar los datos faltantes dentro de dichas variables. Observamos que solo el 3.5% de los datos son faltantes. Sin embargo, esta falta de datos no se distribuye aleatoriamente, sino que se concentra en 12 de las 82 variables. Analizamos más detenidamente estas variablesen la Figura 2:

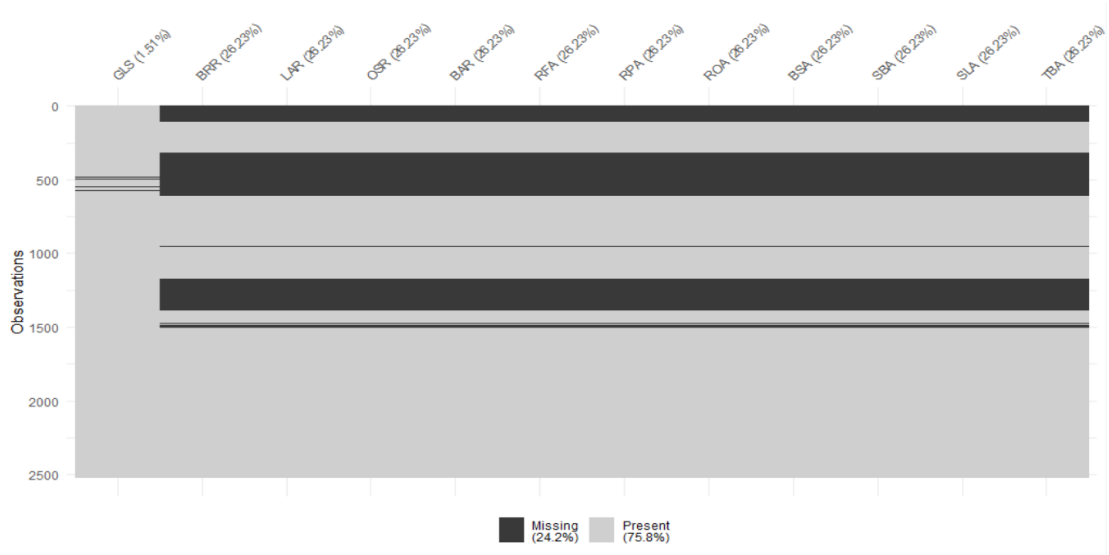

Figura 2: Datos faltantes en el DS original

En 11 de las 12 variables, los datos faltantes se producen en los mismos sujetos y tienen una proporción significativa. Dado que los cráneos están ordenados en el DS por poblaciones, inferimos que no se tomaron dichas medidas en esas poblaciones. Otra de las variables (GLS) cuenta con muchos menos datos faltantes y sí parece seguir una distribución más aleatoria a lo largo del DS.

El tratamiento de estos datos faltantes es un problema habitual dentro del análisis de datos. Una de las técnicas más comunes para lidiar con este problema es la imputación de un valor, normalmente la media o mediana o un valor obtenido mediante algoritmos como el de los k vecinos más cercanos o *k-nearest neighbors* (k-NN).

Precisamente es este último método mencionado el que vamos a utilizar para imputar los datos faltantes en nuestro DS. El algoritmo identifica un sujeto con datos faltantes y selecciona a continuación un número determinado de sujetos (k, en nuestro caso 7) considerados como los más cercanos o parecidos al sujeto problema. Esta cercanía se mide con parámetros como la distancia euclídea o de Manhattan (en nuesto caso, la primera). Observando cuál es el la variable con el dato faltante del sujeto problema, realiza la media del valor de dicha variable entre los sujetos más cercanos e imputa dicho valor a la variable del sujeto problema. De esta forma, podemos obtener un DS sin datos faltantes y con unas imputaciones probablemente más correctas que con otros métodos, como por ejemplo mediante estadístico descriptivos. De hecho, este algoritmo se considera una de las modalidades más básicas de ML [32].

Figura 3: Datos faltantes tras imputación mediante k-NN

Como podemos observar en la Figura 3, tras la imputación de los datos, las variables que presentaban datos faltantes ya cuentan con un 100% de casos completos.

Para lidiar con la alta dimensionalidad de los datos y los costes computacionales asociados, hemos decidido aplicar en algunos casos un tratamiento previo al uso a estos datos, en concreto mediante un análisis de componentes principales o *principal component analysis* (PCA). El PCA es un método por el cual las variables

de un DS se transforman en variables ortogonales, llamadas componentes principales o *principal component* (PC), no correlacionadas entre sí. Estas variables aportan además un porcentaje conocido de la varianza del DS original, por lo que es posible seleccionar siguiendo diferentes métodos el número adecuado de PC. Este método permite por tanto además reducir la dimensionalidad del DS de forma importante sin perder apenas información [33]. No solo eso, si no que existen datos que apoyan que este tratamiento previo puede aumentar la precisión de los modelos predictivos creados a partir de estos datos [34]. Teniendo en cuenta este objetivo, nos hemos quedado con las PC que aportan el 95% de la varianza, habiendo conseguido reducir las dimensiones de nuestro DS a menos de la mitad, consiguiendo además PC no correlacionadas entre sí. En la figura 4 observamos este fenómeno:

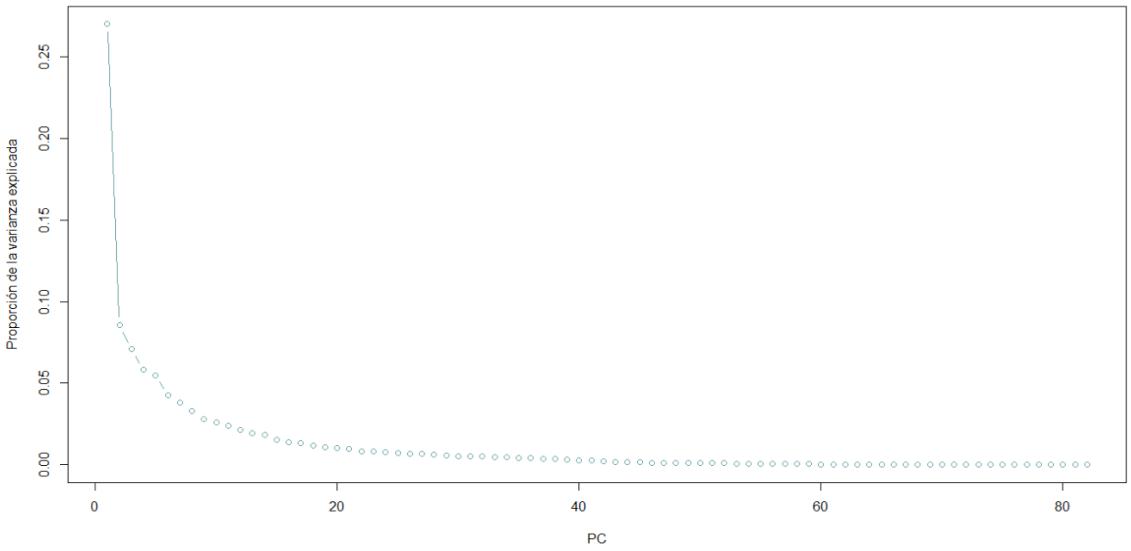

Figura 4: Proporción de la varianza explicada según el número de PC

Por último, antes de pasar a la construcción propiamente dicha de los clasificadores, pasamos a describir lo que se conoce como validación cruzada, en inglés *cross-validation* (CV). Se conoce como CV a un grupo de técnicas cuyo objetivo es evaluar cómo de preciso y válido es el clasificador a la hora de enfrentarse a nuevos casos. Estas técnicas surgen por el hecho de que no es recomendable evaluar la validez del modelo con los datos de entrenamiento que

se han usado para construir el modelo, pues esto lleva a un problema conocido como sobreajuste que conlleva una sobreestimación de las capacidades del clasificador. Es por ello que es preciso el uso de datos independientes, conocidos como datos de validación. Existen varios métodos para obtener estos datos de validación. El más simple consiste en reservar parte de los datos, dejarlos fuera de la construcción de los modelos y usarlos como datos de validación. La desventaja evidente de este método, pese a su simplicidad, es el hecho de que perdemos una gran cantidad de información, y que el azar puede jugar un papel importante en el sentido de cómo se separan los datos. Nosotros hemos decidido usar el método conocido como validación cruzada de K iteraciones, en inglés *k-fold cross-validation*. Este método es no exhaustivo, al contrario que otros métodos como la validación cruzada dejando uno fuera, en inglés *leave-one-out cross-validation* (LOOCV) en el sentido que aprende y valida con K iteraciones de los datos, normalmente 10 que es el valor que vamos a usar nosotros. Es por tanto una forma precisa y computacionalmente más eficiente de evaluar la validez de nuestros modelos [35]. En la Figura 5 podemos apreciar un ejemplo de cómo se realiza la validación cruzada con este método.

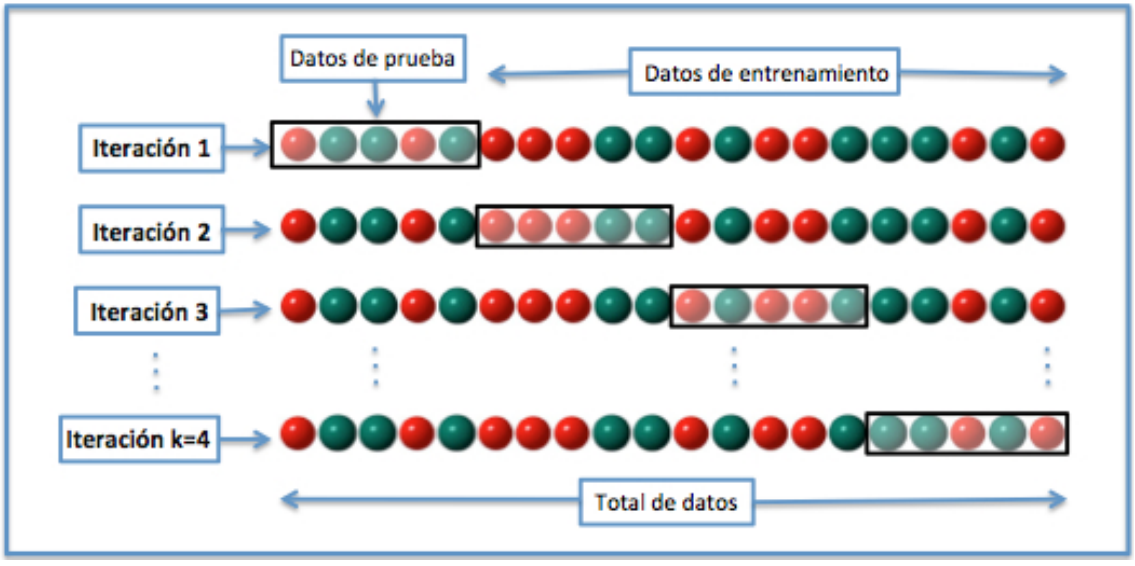

Figura 5: Ejemplo de validación cruzada de 4 iteraciones. Fuente: Wikimedia Commons

Como medida de la calidad del modelo usaremos la precisión, es decir, el porcentaje de veces que el modelo es capaz de predecir correctamente el verdadero valor de la variable predictora, pero también el coeficiente Kappa de Cohen, que es una medida similar de precisión pero que está corregida con las probabilidades que tiene el clasificador de acertar por azar, por lo que se suele considerar como una medida más robusta de la calidad del clasificador.

2.4.2 Construcción de los modelos

Para la construcción de los modelos, hemos recurrido al *software* estadístico *R*. El paquete clave en nuestro estudio es *caret (classification and regression training)*, una serie de funciones creadas para automatizar el proceso de creación de modelos predictivos [36].

2.4.3 Análisis lineal discriminante

Comenzamos creando el modelo mediante análisis lineal discriminante, que como hemos comentado en la secciones anteriores de nuestro trabajo, es el método usado habitualmente en el campo de la antropología forense como clasificador, existiendo una amplia literatura al respecto.

Hemos creado los modelos usando tanto las variables originales como los componentes principales. Como podemos comprobar en la Figura 6 y Figura 7, existen diferencias significativas en la precisión de los modelos de una manera y de otra. El modelo construido con las variables originales es más preciso en los dos casos que el construido con las componentes principales. Teniendo en cuenta además que el coste computacional del LDA no es muy alto incluso para DS con

un alto número de variables, hemos decidido quedarnos con las variables originales.

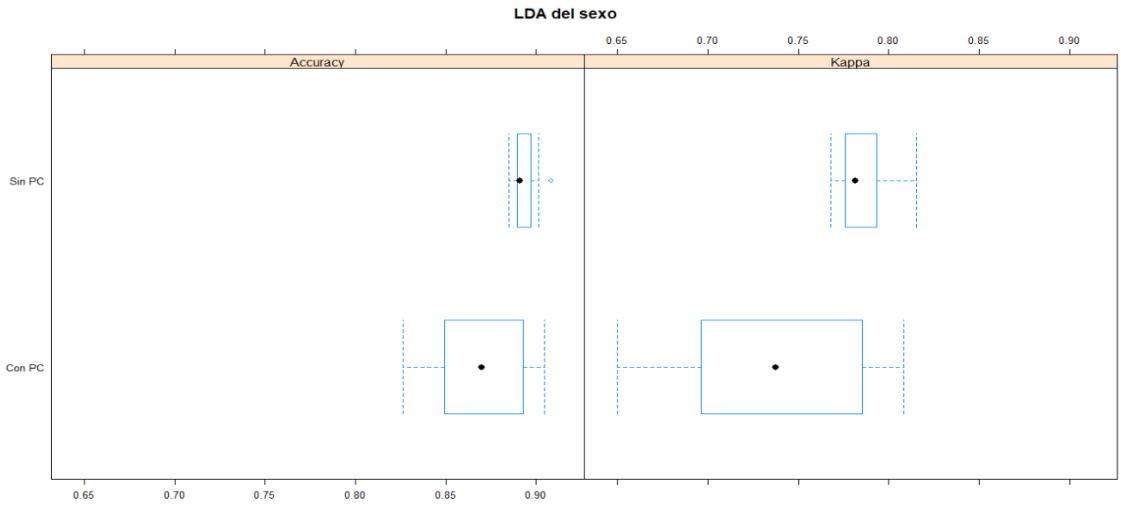

Figura 6: LDA del sexo con variables originales y con PC

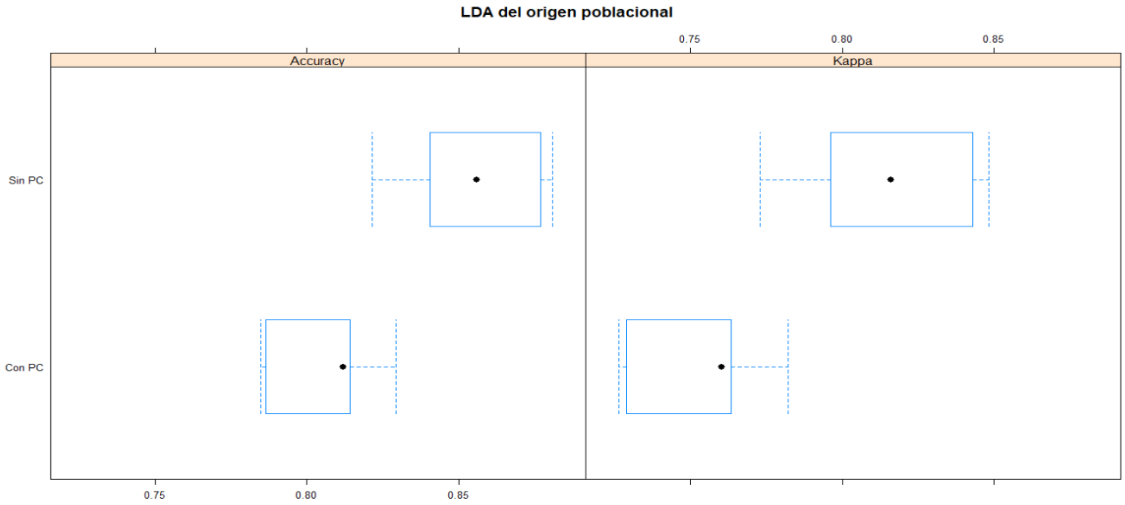

Figura 7: LDA del origen poblacional con variables originales y con PC

En el caso de la estimación del sexo, el mejor modelo tiene una precisión del 89,20% y un coeficiente kappa de 0.79. En el caso de la estimación del origen poblacional, los resultados son algo menores en precisión (85,70%), pero cuentan con un coeficiente kappa mayor que en el caso anterior (0.82). Esto es debido a que, a pesar de que la precisión es algo menor, la probabilidad de que el clasificador acierte por simple azar es menor, dado que la variable a predecir

cuenta con un mayor número de niveles (en este caso 5 en vez de 2). Es por ello que podemos decir que, como modelo predictivo, presenta mayor calidad el segundo que el primero.

A continuación mostramos las matrices de confusión de los modelos más precisos construidos mediante LDA:

		Referencia	
		Hombre	Mujer
Predicción	Hombre	40.8%	5.8%
	Mujer	5.0%	48.4%

Tabla 2: Matriz de confusión para el modelo LDA para el sexo

Precisión para clase hombre: 89,08%

Precisión para clase mujer: 89,30%

		Referencia				
		África	América	Asia	Europa	Oceanía
Predicción	África	16.4%	0.1%	1.1%	0.7%	0.5%
	América	0.1%	13.1%	0.6%	0.4%	0.6%
	Asia	0.4%	0.9%	20.9%	0.5%	2.6%
	Europa	1.7%	0.3%	0.6%	10.7%	0.2%
	Oceanía	0.6%	1.0%	1.2%	0.3%	24.6%

Tabla 3: Matriz de confusión para el modelo LDA para el origen poblacional

Precisión para clase África: 85,41%

Precisión para clase América: 85,06%

Precisión para clase Asia: 85,66%

Precisión para clase Europa: 84,92%

Precisión para clase Oceanía: 86,32%

2.4.4 Bayes ingenuo

En este caso, la calidad del modelo es significativamente mayor cuando usamos las variables rotadas en vez de las variables originales (al contrario que en el caso anterior), como podemos observar en Figura 8 y Figura 9:

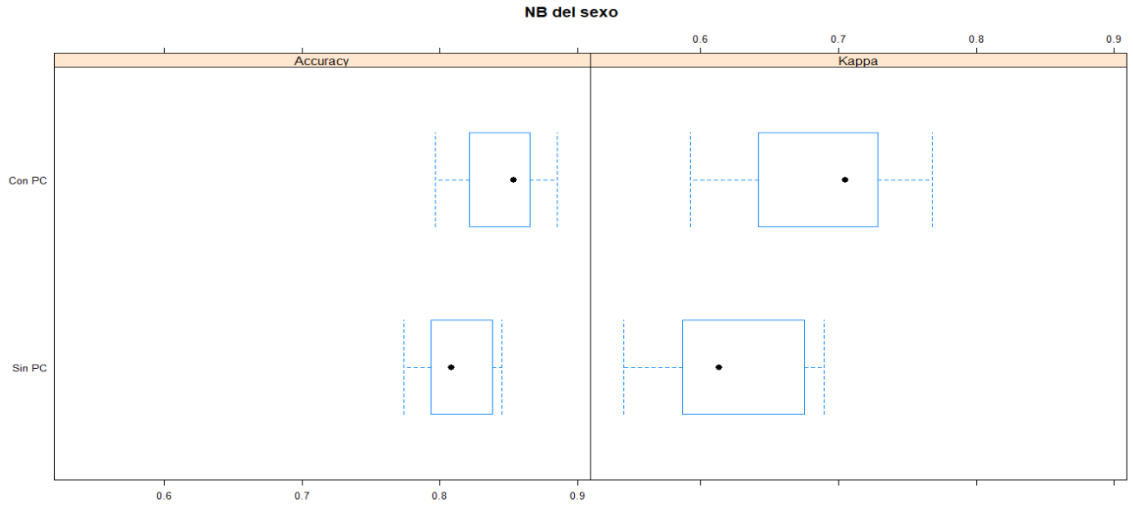

Figura 8: NB del sexo con variables originales y con PC

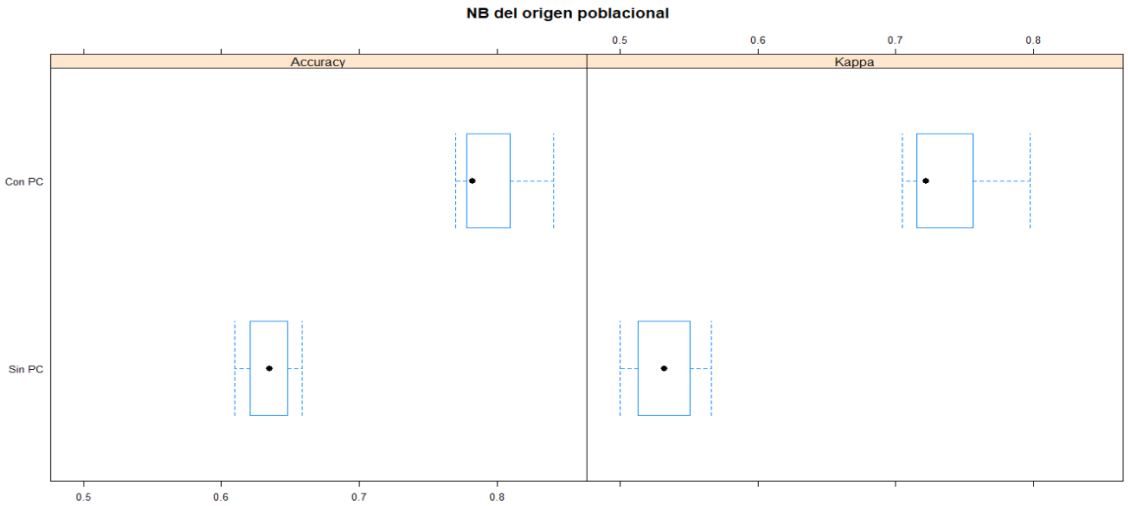

Figura 9: NB del origen poblacional con variables originales y con PC

Por el funcionamiento de este algoritmo clasificador, descrito con anterioridad, el modelo calcula las probabilidades de que un probando pertenezca a una determinada clase con la ayuda de una distribución estadística determinada. Estas opciones son habitualmente la distribución normal y la distribución *kernel*. Esta última distribución se usa cuando no se quieren realizar asunciones sobre la distribución de una determinada variable aleatoria. Dado que de antemano desconocemos de qué forma nuestro modelo tendrá más calidad, vamos a realizar una comparación de los modelos de las dos formas y seleccionaremos la que mejor resultados nos muestre.

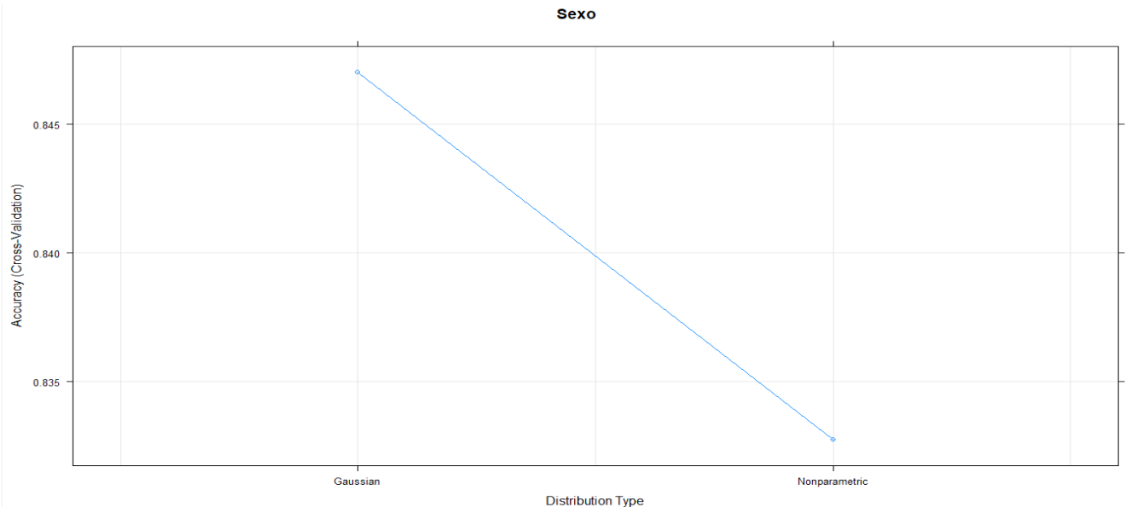

Figura 10: Comparación de las dos distribuciones para el modelo NB para el sexo

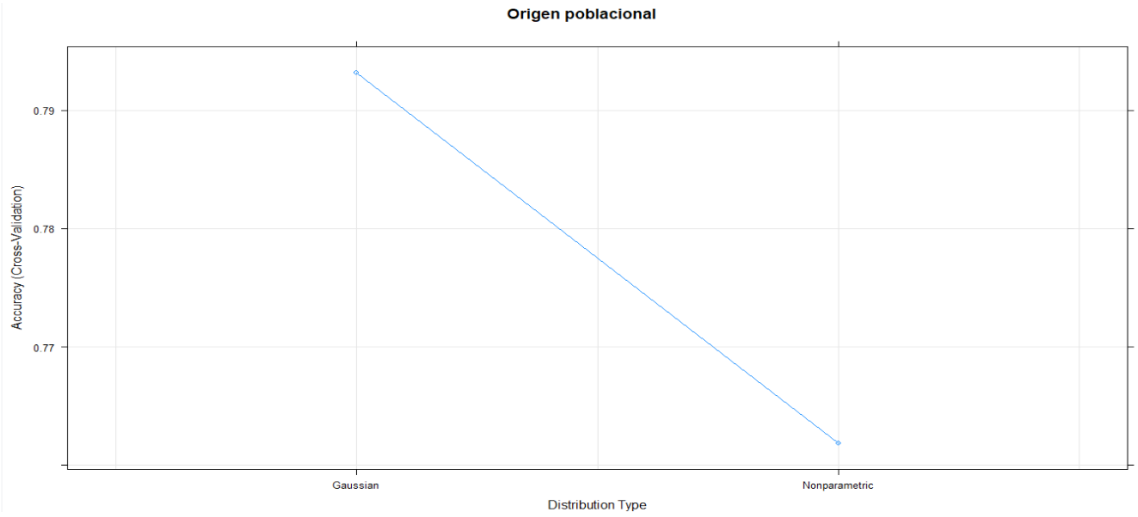

Figura 11: Comparación de las dos distribuciones para el modelo NB para el origen poblacional

Observamos en Figura 10 y Figura 11 que los modelos son más precisos cuando usan la distribución normal como comparador, aunque el efecto no sea muy grande.

La precisión del mejor modelo para predecir el sexo mediante NB tiene una precisión del 84.70%, con un coeficiente Kappa de 0.69. El mejor modelo para predecir el origen poblacional tiene una precisión del 79.32%, con un coeficiente kappa de 0.74. En Tabla 3 y Tabla 4 proporcionamos las matrices de confusión de los modelos.

		Referencia	
		Hombre	Mujer
Predicción	Hombre	37.9%	7.4%
	Mujer	7.9%	46.8%

Tabla 4: Matriz de confusión para el modelo NB para el sexo

Precisión para clase hombre: 82,75%

Precisión para clase mujer: 86,35%

		Referencia				
		África	América	Asia	Europa	Oceanía
Predicción	África	15.1%	0.2%	1.3%	0.8%	1.0%
	América	0.2%	11.5%	1.2%	0.7%	0.7%
	Asia	1.3%	1.7%	19.4%	0.6%	3.1%
	Europa	1.7%	0.6%	0.4%	9.8%	0.2%
	Oceanía	0.9%	1.3%	2.0%	0.8%	23.5%

Tabla 5: Matriz de confusión para el modelo NB para el origen poblacional

Precisión para clase África: 78,65%
Precisión para clase América: 75,16%
Precisión para clase Asia: 79,84%
Precisión para clase Europa: 77,17%
Precisión para clase Oceanía: 82,46%

2.4.5 *Random forest*

El algoritmo RF cuenta con hiperparámetros, es decir, valores que se especifican a un algoritmo determinado antes de comenzar el proceso de aprendizaje, y cuya variación cambia la forma en que el clasificador aprende y por tanto también su poder de predicción. Dado que no se conoce de antemano qué combinación de hiperparámetros va a otorgar los mejores resultados, es necesario desarrollar un proceso empírico conocido como optimización de hiperparámetros para encontrar la mejor combinación de valores para nuestro clasificador.

En concreto, el algoritmo RF cuenta principalmente con tres hiparámetros significativos: profundidad del árbol, número de variables seleccionadas en cada nodo y el número de árboles. De estas tres, vamos a ajustar el número de árboles y el número de variables seleccionadas en cada nodo. De este último hiperparámetro, es necesario mantener un equilibrio: valores muy bajos limitan la probabilidad de encontrar una buen valor para separar, pero valores muy altos

hace que los árboles estén muy correlacionados. Una regla general es usar la raíz cuadrada del número de variables totales del DS, en nuestro caso se correspondería con unas 9. También hemos comprobado el efecto que tiene el número de árboles construidos en cada modelo en la calidad de las predicciones. En principio, el aumento del número de árboles debería conllevar consigo un aumento de la calidad del modelo.

Dado el alto coste computacional de calcular todas las posibles combinaciones de modelos, vamos a seleccionar los mejores hiperparámetros mediante la estimación del *out-of-bag error* (OOB), un método usado en los métodos de ML que usan árboles de decisión para validar los modelos. Valores más bajos de error OOB se corresponderán con modelos más precisos y por tanto elegiremos la combinación de hiperparámetros que minimice este valor para cada modelo. En la Figura 12 lo representamos:

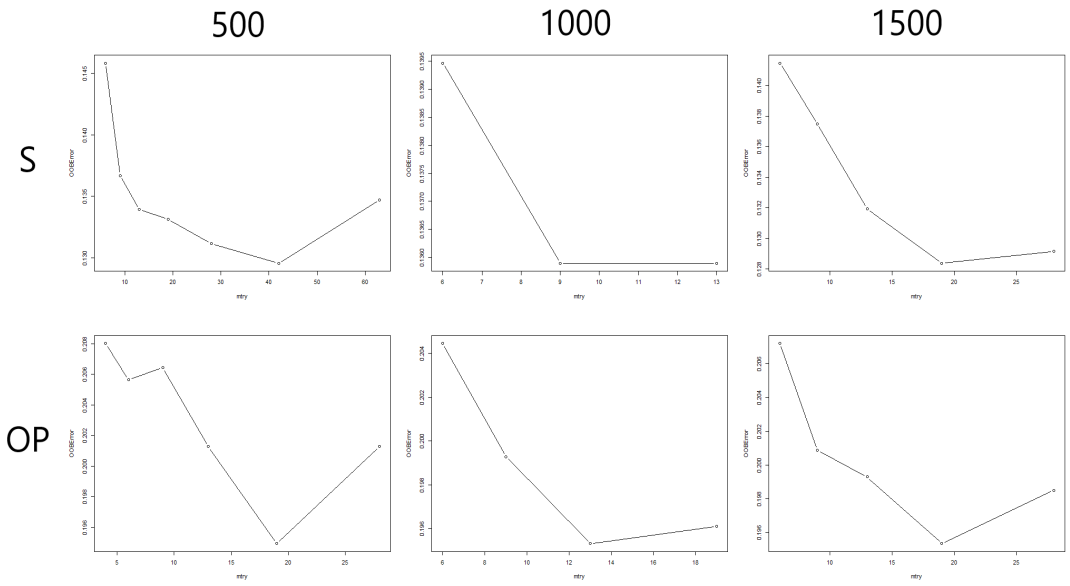

Figura 12: Estimación del error OOB para diferentes números de árboles (500, 1000 y 1500) y diferentes número de variables seleccionadas por nodo (en el eje de abscisas). Tanto para el sexo (S) como origen poblacional (OP)

Para el caso del modelo del sexo, la combinación de hiperparámetros que otorga el resultado más bajo de error OOB es 1500 árboles y 18 variables seleccionadas por nodo, y en el caso del origen poblacional 500 árboles y 19 variables por nodo.

Ahora que hemos seleccionado la mejor combinación de hiperparámetros para nuestros modelos, creamos los modelos con dichos hiperparámetros mediante el paquete *caret* como hemos estado haciendo hasta ahora, con el objetivo de medir mediante validación cruzada los estimadores que hemos estado calculando en los modelos anteriores.

La precisión del modelo final para el sexo es 86,90%, con un coeficiente kappa de 0.74. El modelo final para el origen poblacional tiene una precisión del 80.35% y un coeficiente kappa de 0.75. Lo representamos en Figura 13 y 14:

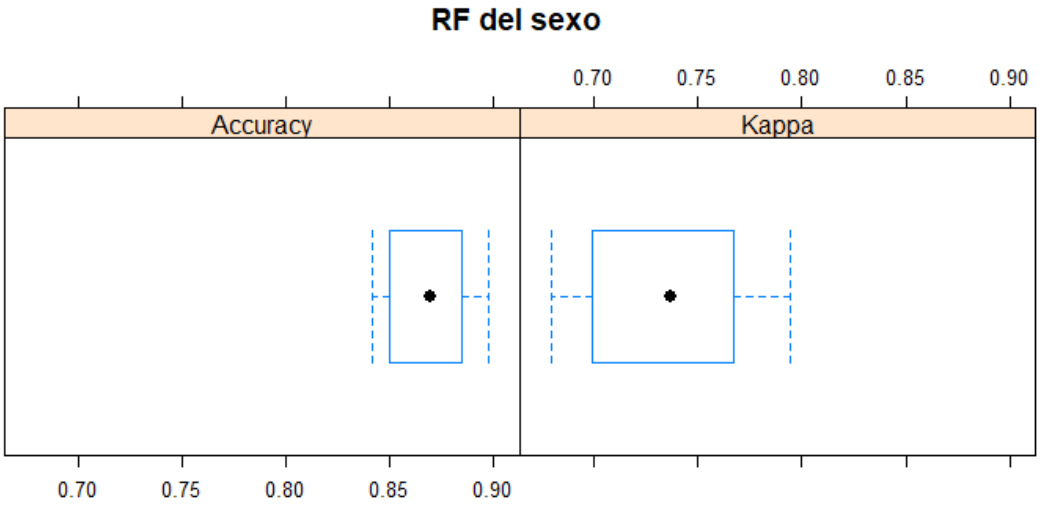

Figura 13: RF del sexo

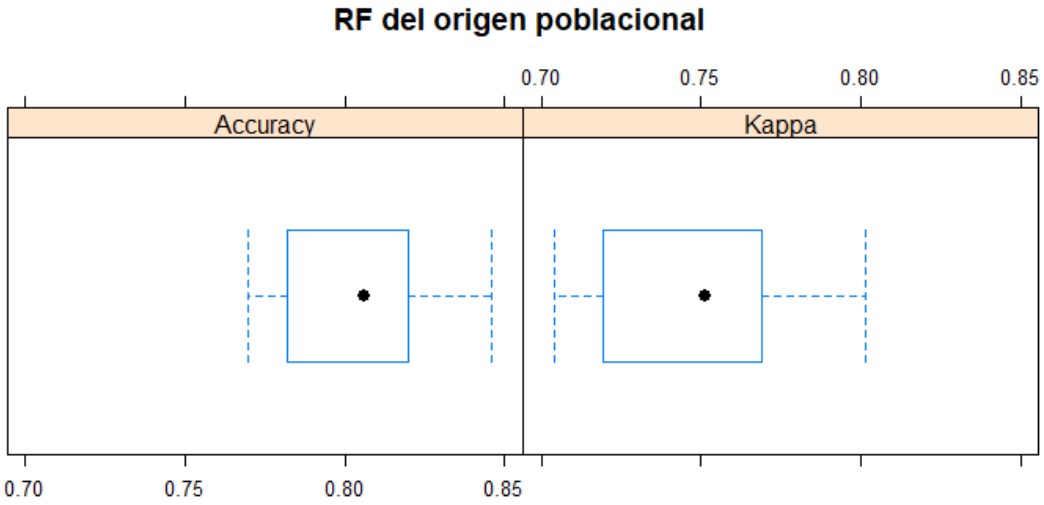

Figura 14: RF del origen poblacional

Otorgamos también las matrices de confusión correspondientes en Tabla 5 y Tabla 6:

		Referencia	
		Hombre	Mujer
Predicción	Hombre	39.8%	7.1%
	Mujer	6.0%	47.1%

Tabla 6: Matriz de confusión para el modelo RF para el sexo

Precisión para clase hombre: 86,89%

Precisión para clase mujer: 86,90%

		Referencia				
		África	América	Asia	Europa	Oceanía
Predicción	África	15.7%	0.2%	1.2%	1.1%	0.6%
	América	0.1%	11.1%	0.5%	0.4%	0.4%
	Asia	1.1%	1.6%	20.3%	1.0%	3.2%
	Europa	1.3%	0.7%	0.4%	9.2%	0.1%
	Oceanía	1.0%	1.9%	2.0%	0.9%	24.1%

Tabla 7: Matriz de confusión para el modelo RF para el origen poblacional

Precisión para clase África: 81,77%
Precisión para clase América: 71,61%
Precisión para clase Asia: 83,20%
Precisión para clase Europa: 73,02%
Precisión para clase Oceanía: 84,86%

2.4.6 Redes neuronales artificiales

Antes de comenzar a trabajar con nuestros datos para la creación de modelos mediante ANN, es muy recomendable aplica una normalización y centralización de los datos. Valores muy variables y diferentes en las variables de entrenamiento pueden provocar inestabilidad durante el proceso de aprendizaje y una pobre calidad de las predicciones. De forma similar al caso anterior, vamos a realizar un ajuste de los hiperparámetros con los que cuenta un modelo de ANN. En concreto, vamos a considerar el número de capas de neuronas ocultas, el número de neuronas en cada capa y la tasa de caída de los pesos en el proceso de aprendizaje.

Sin duda, los valores más críticos se corresponden con la estructura de la capa o capas ocultas. Dado el alto número de estructuras posibles, es necesario alguna regla que nos permita estimar al menos el número aproximadamente adecuado de neuronas ocultas. Una de estas posibles reglas es la raíz cuadrada del producto del número de neuronas de la capa de entrada (que se corresponde con el número de variables predictoras y el número de neuronas de la capa de salida (que se corresponde con el número de clases de la variables a predecir) [37]. Siguiendo esta regla, hemos agrupado el número de neuronas correspondientes de diferentes maneras, en diferente número de capas ocultas, seleccionando finalmente el modelo que mejor precisión ha otorgado.

En cuanto al modelo para el sexo, finalmente hemos seleccionado el compuesto por tres capas ocultas, de 5, 4 y 4 neuronas (13 en total), respectivamente. La

caída de los pesos la hemos fijado en 0.0001. La precisión de dicho modelo es del 87.87%, con un coeficiente kappa de 0.76. Representamos la red y su precisión en Figura 15 y 16:

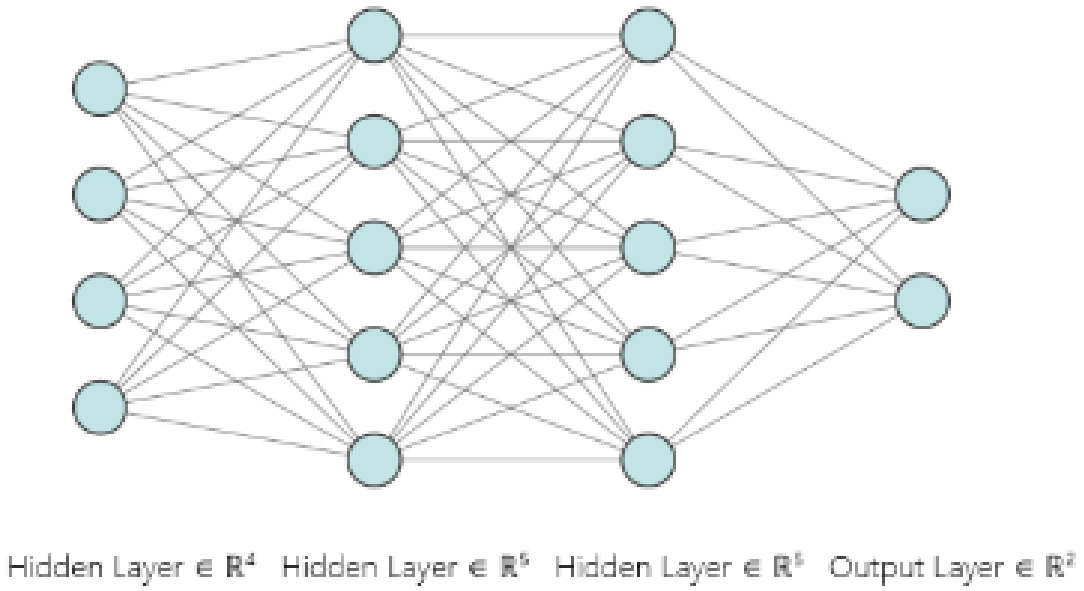

Figura 15: Estructura de la ANN para el modelo del sexo (capa de entrada no representada)

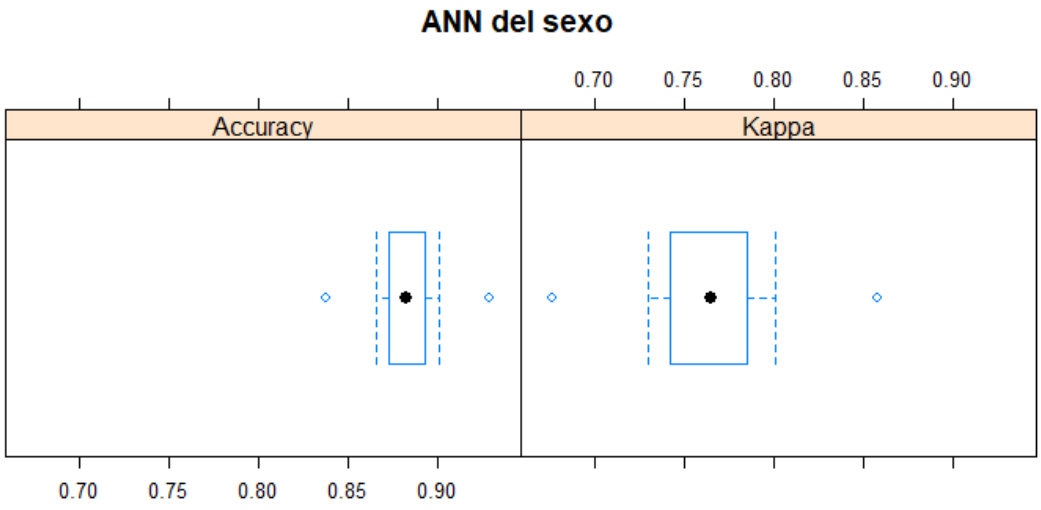

Figura 16: ANN del sexo

El modelo predictivo para el origen poblacional cuenta con un mayor número de neuronas ocultas recomendadas que en el anterior caso. En este caso, las 20 neuronas que hemos usado la hemos distribuido uniformemente en una sola capa, con un valor de caída de los pesos de 0. La precisión final del modelo es del 86,92%, con un coeficiente kappa de 0.83. Representamos la red y su precisión en Figura 17 y 18:

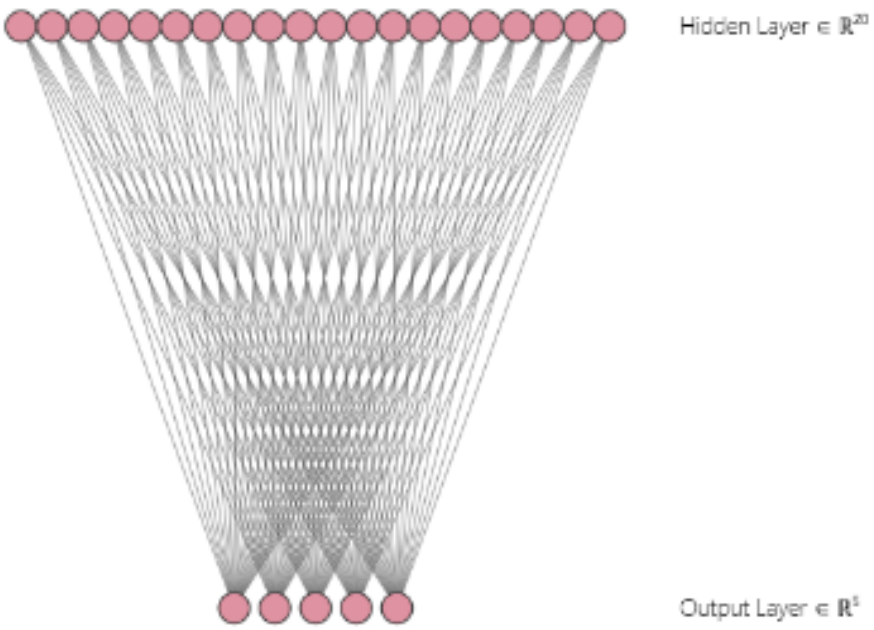

Figura 17: Estructura de la ANN para el modelo del origen poblacional (capa de entrada no representada)

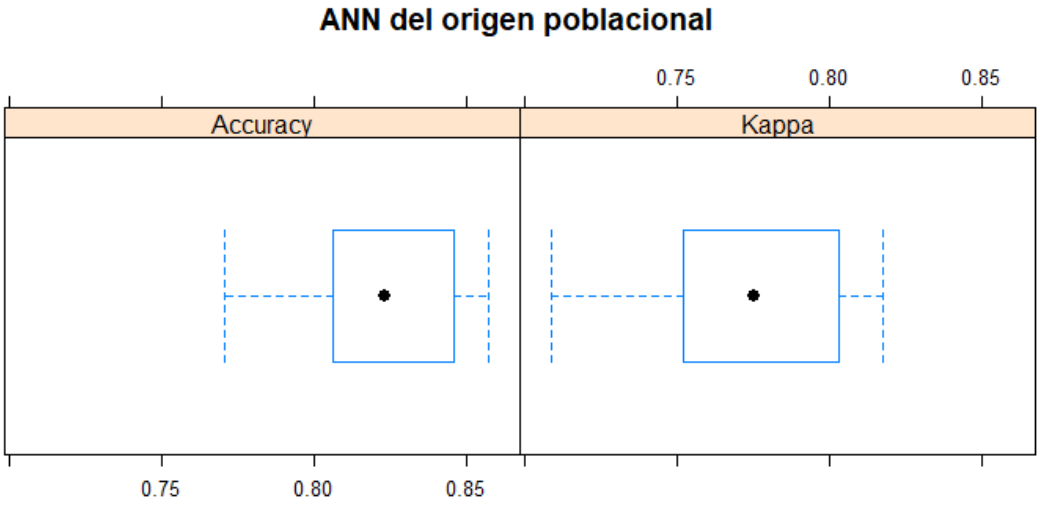

Figura 18: ANN del origen poblacional

Adjuntamos las correspondiente matrices de confusión en Tabla 7 y Tabla 8:

		Referencia	
		Hombre	Mujer
Predicción	Hombre	40.3%	6.7%
	Mujer	5.5%	47.5%

Tabla 8: Matriz de confusión para el modelo ANN para el sexo

Precisión para clase hombre: 87,99%

Precisión para clase mujer: 87,64%

		Referencia				
		África	América	Asia	Europa	Oceanía
Predicción	África	17.2%	0.0%	0.7%	0.6%	0.6%
	América	0.0%	13.6%	0.4%	0.5%	0.5%
	Asia	0.4%	0.6%	20.4%	0.4%	2.5%
	Europa	0.8%	0.4%	0.5%	10.9%	0.1%
	Oceanía	0.7%	0.8%	2.3%	0.2%	24.8%

Tabla 9: Matriz de confusión para el modelo ANN para el origen poblacional

Precisión para clase África: 90,05%

Precisión para clase América: 88,31%

Precisión para clase Asia: 83,95%

Precisión para clase Europa: 86,51%

Precisión para clase Oceanía: 87,02%

2.4.7 Ejemplo de uso de los modelos

El objetivo principal de nuestro trabajo es la construcción de modelos predictivos mediante ML que permitan estimar correctamente el sexo y el origen poblacional de un individuo a partir de datos craneométricos.

Si bien en los apartados anteriores hemos desarrollado dichos modelos, en concreto mediante NB, RF y ANN, nos gustaría mostrar a continuación un ejemplo el uso práctico que pueden tener.

Para ello, hemos seleccionado un conjunto de cráneos recogidos por Howells en otro DS, con una identificación del sexo, el origen poblacional y las mismas medidas craneométricas que hemos usado anteriormente.

Para poder realizar una predicción con los modelos construidos, hemos de entregar a nuestros modelos un conjunto de variables de entrada con los mismos nombres que las variables con las que han sido construidos. Es importante que no existen datos faltantes en estos datos de entrada, por lo que en estos casos sería recomendable realizar una imputación como hicimos anteriormente:

```
#Lectura de los nuevos datos
HoT←read.csv("HowellTest.csv")
#Selección de un caso con todos los valores de las variables presentes
HoT[18,]
#Uso de la función predict para clasificar el cráneo en un sexo y origen poblacional
determinado
predict(list(NB=nb.s.model2,RF=rf.s.model,ANN=ann.s.model),newdata=HoT
[18,8:89])
```

```
$NB
[1] F
Levels: F M
$RF
[1] M
Levels: F M
$ANN
[1] M
Levels: F M
predict(list(NB=nb.op.model2,RF=rf.op.model,ANN=ann.op.model),newdata=
HoT[18,8:89])
$NB
[1] Europe
5 Levels: Africa America … Oceania
$RF
[1] America
5 Levels: Africa America … Oceania
$ANN
[1] America
5 Levels: Africa America … Oceania
#Verdaderas clases del cráneo
HoT[18,2]
[1] M
Levels: F M
HoT[18,3]
[1] Peru
Levels: Abri Pataud … Zulu
#Ejemplo de otro caso
predict(list(NB=nb.s.model2,RF=rf.s.model,ANN=ann.s.model),newdata=
HoT[363,8:89])
$NB
[1] M
Levels: F M
$RF
[1] M
Levels: F M
$ANN
[1] M
Levels: F M
predict(list(NB=nb.op.model2,RF=rf.op.model,ANN=ann.op.model),newda
ta=HoT[363,8:89]
$NB
[1] Asia
5 Levels: Africa America … Oceania
$RF
[1] Asia
5 Levels: Africa America … Oceania
$ANN
[1] America
5 Levels: Africa America … Oceania
HoT[363,2]
[1] M
Levels: F M
HoT[363,3]
[1] Korea
Levels: Abri Pataud … Zulu
```

En el primer caso, el cráneo es de un individuo varón de procedencia peruana, y por tanto americano. Observamos que los modelos construidos con NB fallan en clasificarlo (mujer, europea), pero los modelos de RF y de ANN aciertan.

En el segundo caso, el cráneo es de un individuo varón de procedencia coreana, y por tanto asiático. Todos los modelos predicen correctamente el sexo, mientras que el origen poblacional es predicho correctamente por los modelos de NB y de RF, fallando el modelo de ANN al clasificarlo como americano.

3. Discusión

La estimación de las distintas partes del perfil biológico a partir de elementos esqueléticos mediante algoritmos de ML es un área de investigación emergente dentro del campo de la antropología forense. El motivo del interés en esta área es la hipótesis de que los algoritmos de ML pueden proporcionar modelos predictivos de mayor calidad que los métodos estadísticos convencionalmente usados para este fin, teniendo en cuenta la importancia de la correcta estimación del perfil biológico desde el punto de vista forense.

En este trabajo nos hemos centrado principalmente en la estimación a partir del cráneo de dos de los aspectos probablemente más importantes del perfil biológico: el sexo y el origen poblacional.

Dadas las condiciones en las que muchas veces se encuentran los restos humanos, es importante contar con modelos predictivos que usen diferentes elementos esqueléticos en caso de que nos encontremos con esqueletos incompletos o en un estado de conservación que impidan la correcta medición de los elementos. En la literatura encontramos trabajos como el de Navega et al. [3], du Jardin et al. [9] y Mahfouz et al. [10] que han usado elementos poscraneales como los huesos del tarso, el fémur y la rótula, respectivamente para la estimación del sexo principalmente mediante diferentes algoritmos de ML como son ANN y RF.

Existen algunos trabajos como el de Navega et al. (2) [22] y Santos et al. [38]que han usado el cráneo para esta tarea. Sin embargo, en el trabajo de Santos et al. se crearon modelos para la estimación únicamente del sexo mediante SVM, siendo además los cráneos de poblaciones seleccionadas y por tanto con poca variabilidad en su origen. En el trabajo de Navega et al. (2) ocurrió de forma similar y se usó únicamente el método de RF.

Por tanto, nuestro trabajo es el primero que existe en la bibliografía hasta donde sabemos que ha usado cráneos de múltiples poblaciones diferentes como elemento esquelético para la creación de modelos predictivos mediante métodos de ML habitualmente usados como NB, RF y ANN para la estimación del sexo y el origen poblacional.

En general, la precisión de los modelos fue alta, similar a la encontrada por otros autores. Navega et al. fue capaz de clasificar correctamente el sexo del 88,3% de los individuos mediante los huesos tarsales y un modelo de RF, du Jardin et al. tuvo una precisión del 93,4% usando el fémur y Mahfouz et al. Del 93,51% usando la rótula y ANN, en comparación con el 84,70%, 86,90% y 87,87% clasificado por nosotros con el cráneo y NB, RF y ANN, respectivamente. Santos et al. consiguió una precisión similar a nosotros usando el cráneo y SVM.

En cuanto a la estimación del origen poblacional mediante nuestros modelos, observamos que existe una tendencia a la no confusión entre sujetos africanos y americanos y una tendencia a la confusión entre sujetos asiáticos y oceánicos, debido a la disimilitud y similitud de los cráneos de dichas poblaciones.

Existen datos escasos como hemos mencionado anteriormente sobre los métodos de ML para estimar el origen poblacional mediante el cráneo. El trabajo de Navega et. al en concreto solo analiza poblaciones africanas y europeas mediante RF. Nuestro modelo, concretamente el construido mediante ANN, alcanzó un grado alto de precisión, 86,92%, superior al obtenido mediante LDA para clasificar en hasta cinco orígenes poblacionales diferentes, lo que muestra el potencial de estas técnicas para la construcción de modelos predictivos precisos.

Como limitaciones encontradas en nuestro trabajo, destacar el alto coste computacional de las técnicas de ML, en concreto de RF y ANN. Centrándonos en las ANN, observamos que cuentan con un alto número de hiperparámetros ajustables (número de neuronas, número de capas ocultas, tasa de caída de los peso), lo que hace que existan infinidad de formas de construir un modelo

mediante este método. Nosotros seguimos una reglas generales para seleccionar un número de neuronas y las estructuramos de diferente forma, pero el alto coste computacional de calcular los modelos junto con la necesidad de validación cruzada superaba con creces las capacidades de nuestro ordenador personal, por lo que no podemos asegurar que no existen mejores modelos al que hemos mostrado aquí.

Como línea de investigación futura, sería interesante seguir profundizando en el método que nos ha dado mejores resultado, las ANN. En concreto, sería interesante probar otros algoritmos basados en ANN y avanzar en el uso del aprendizaje profundo o *deep learning*, es decir, ANN con un gran número de capas ocultas con probada eficacia en otros campos del conocimiento. También sería muy interesante demostrar el potencial en estas tareas del *stacking*, una técnica que consiste en la creación de un metamodelo que combina el resultado de diferentes clasificadores, tanto convencionales como de ML, para la otorgar una clasificación final, a menudo más precisa que los modelos por separado.

4. Conclusiones

La estimación del perfil biológico es una de las tareas más importantes a las que enfrenta la antropología forense. Por ello, contar con las herramientas lo más precisas posibles es fundamental para este fin. Si bien los métodos habituales usados, como el LDA, otorgan buenos resultados, en los últimos años el profundo desarrollo que han experimentado las técnicas de ML, con su implementación en todo tipo de escenarios, hace que sea necesario estudiar su empleo en el campo que nos ocupa.

Mediante este trabajo, hemos estudiado la aplicabilidad de este tipo de técnicas, en concreto NB, RF y ANN para la construcción de modelos predictivos para la estimación del sexo y el origen poblacional a partir de datos craneométricos obtenidos de un DS ampliamente conocido y estudiado, que cuenta con la ventaja además de contener individuos de todo el mundo. Dichas técnicas han permitido obtener modelos altamente precisos, destacando sobre todo los modelos obtenidos mediante ANN. Sin embargo, también es evidente que son técnicas computacionalmente más exigente y que requieren muchas veces una transformación previa de los datos y un ajuste de hiperparámetros, por lo que son más complejas que los métodos convencionales. En definitiva, estas técnicas son una alternativa factible a los métodos convencionales de estimación del perfil biológico y en algunos casos pueden superarlos.

Como objetivos del trabajo, nos planteamos en primer lugar la selección de una serie de técnicas de ML adecuadas, para lo que se redactó un texto en el que se describían algunas de dichas técnicas junto con las técnicas convencionales, lo que permitió su evaluación y comparación. Una vez seleccionadas dichas técnicas, se procedió al siguiente objetivo, que no era otro que la aplicación de dichas técnicas en la construcción de modelos predictivos para la estimación del sexo y el origen poblacional mediante datos craneométricos.

El primer objetivo se cumplió, estimándose como técnicas adecuadas para la construcción de los modelos, además del LDA, NB, RF y ANN.

El segundo objetivo se cumplió, al obtener los modelos predictivos mencionados anteriormente mediante los métodos seleccionados con el primer objetivo.

La metodología y planificación seguida durante el desarrollo del trabajo ha sido la adecuada para la correcta consecución de los objetivos. No hubo que realizar cambios significativos en este aspecto.

Como línea de investigación futura en este campo, creo que sería interesante el estudio de diferentes tipos de ANN para esta tarea, con la aplicación del concepto de aprendizaje profundo o *deep learning*. También sería interesante ver cómo desempeñan estos algoritmos su papel en la estimación de otros factores del perfil biológico como son la edad y la estatura, pues los algoritmos deberían realizar en estos casos tareas de regresión, no de clasificación como han desarrollado en este trabajo.

5. Bibliografía

[1] American Board of Forensic Anthropology. General information about the field of forensic anthropology [Internet]. 2018 [citado 3 de junio de 2019]. Disponible en: http://theabfa.org/faq/

[2] Langley NR, Tersigni-Tarrant MTA. Forensic Anthropology in the United States. Past and Present. En: Forensic Anthropology A comprehensive guide. 2.ª ed. Boca Raton: Taylor & Francis Group; 2017. p. 7-22.

[3] Navega D, Vicente R, Vieira DN, Ross AH, Cunha E. Sex estimation from the tarsal bones in a Portuguese sample: a machine learning approach. Int J Legal Med. mayo de 2015;129(3):651-9.

[4] Phenice TW. A newly developed visual method of sexing the os pubis. American Journal of Physical Anthropology. 1969;30(2):297-301.

[5] Adams BJ, Byrd J. Interobserver Variation of Selected Postcranial Skeletal Measurements. Journal of forensic sciences. 1 de diciembre de 2002;47:1193-202.

[6] Spradley MK, Jantz RL. Sex Estimation in Forensic Anthropology: Skull Versus Postcranial Elements. Journal of Forensic Sciences. 2011;56(2):289-96.

[7] Moore MK. Sex estimation and assesment. En: Research methods in human skeletal biology. 1.ª ed. Cambridge: Academic Press; 2012. p. 91-116.

[8] Hefner JT, Spradley MK, Anderson B. Ancestry assessment using random forest modeling. J Forensic Sci. mayo de 2014;59(3):583-9.

[9] du Jardin P, Ponsaillé J, Alunni-Perret V, Quatrehomme G. A comparison between neural network and other metric methods to determine sex from the upper femur in a modern French population. Forensic Sci Int. 20 de noviembre de 2009;192(1-3):127.e1-6.

[10] Mahfouz M, Badawi A, Merkl B, Fatah EEA, Pritchard E, Kesler K, et al. Patella sex determination by 3D statistical shape models and nonlinear classifiers. Forensic Sci Int. 20 de diciembre de 2007;173(2-3):161-70.

[11] Mitchell TM. Machine learning. 1.ª ed. New York City: McGraw-Hill; 1997.

[12] Auerbach BM. The William W. Howells craniometric data set [Internet]. 2014 [citado 3 de junio de 2019]. Disponible en: https://web.utk.edu/~auerbach/HOWL.htm

[13] Howells WW. 1973. Cranial Variation in Man. A Study by Multivariate Analysis of Patterns of Differences Among Recent Human Populations. Papers of the Peabody Museum of Archeology and Ethnology, vol. 67, pp. 259. Cambridge: Peabody Museum.

[14] Howells WW. 1989. Skull Shapes and the Map. Craniometric Analyses in the Dispersion of Modern Homo. Papers of the Peabody Museum of Archaeology and

Ethnology, vol. 79, pp. 189. Cambridge: Peabody Museum.

[15] Howells WW. 1995. Who's Who in Skulls. Ethnic Identification of Crania from Measurements. Papers of the Peabody Museum of Archaeology and Ethnology, vol. 82, pp. 108. Cambridge: Peabody Museum.

[16] Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM Journal of Research and Development. julio de 1959;3(3):210-29.

[17] James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning with applications in R. 1.ª ed. New York City: Springer; 2013.

[18] Ho YC, Pepyne DL. Simple Explanation of the No-Free-Lunch Theorem and Its Implications. Journal of Optimization Theory and Applications. 1 de diciembre de 2002;115(3):549-70.

[19] Ramasubramanian K, Singh A. Machine learning theory and practice. 2.ª ed. New York City: Apress; 2019.

[20] Bickel PJ, Levina E. Some Theory for Fisher's Linear Discriminant Function, «Naive Bayes», and Some Alternatives When There Are Many More Variables than Observations. Bernoulli. 2004;10(6):989-1010.

[21] Deschamps B, McNairn H, Shang J, Jiao X. Towards operational radar-only crop type classification: comparison of a traditional decision tree with a random forest classifier. Canadian Journal of Remote Sensing. 13 de enero de 2012;38(1):60-8.

[22] Navega D, Coelho C, Vicente R, Ferreira MT, Wasterlain S, Cunha E. AncesTrees: ancestry estimation with randomized decision trees. Int J Legal Med. septiembre de 2015;129(5):1145-53.

[23] Hefner JT, Spradley MK, Anderson B. Ancestry assessment using random forest modeling. J Forensic Sci. mayo de 2014;59(3):583-9.

[24] Li H. Which machine learning algorithm should I use? [Internet]. 2017 [citado 3 de junio de 2019]. Disponible en: https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

[25] Fisher RA. The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics. 1936;7(2):179-88.

[26] Barnard MM. The Secular Variations of Skull Characters in Four Series of Egyptian Skulls. Annals of Eugenics. 1935;6(4):352-71.

[27] Martin ES. A Study of an Egyptian Series of Mandibles, with Special Reference to Mathematical Methods of Sexing. Biometrika. 1 de junio de 1936;28(1-2):149-78.

[28] Giles E, Elliot O. Race identification from cranial measurements. J Forensic Sci. 1962.

[29] Jantz RL, Ousley SD. Introduction fo Fordisc 3 and human variation statistics. En: Forensic Anthropology A comprehensive guide. Boca Raton: Taylor & Francis Group; 2017. p. 255-70.

[30] Dixon SJ, Brereton RG. Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure. Chemometrics and Intelligent Laboratory Systems. 15 de enero de 2009;95(1):1-17.

[31] Perme M, Blas M, Turk S. Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study. Metodološki Zvezki. 1 de enero de 2004;1:143-61.

[32] Marwala T. Computational intelligence for missing data imputation, estimation, and management. 1.ª ed. Hershey: Information Science Reference; 2011.

[33] Xiangyu K, Hu C, Duan Z. Principal component analysis networks and algorithms. 1.ª ed. Beijing: Science Press; 2017.

[34] Howley T, Madden MG, O'Connell M-L, Ryder AG. The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data. Knowledge-Based Systems. 1 de septiembre de 2006;19(5):363-70.

[35] Brownlee J. A gentle introduction to K-fold cross-validation [Internet]. 2018 [citado 3 de junio de 2019]. Disponible en: https://machinelearningmastery.com/k-fold-cross-validation/

[36] Kuhn M. The caret package [Internet]. 2019 [citado 3 de junio de 2019]. Disponible en: http://topepo.github.io/caret/index.html

[37] Masters T. Practical neural network recipies in C++. 1.ª ed. San Diego: Academic Press; 1993.

[38] Santos F, Guyomarc'h P, Bruzek J. Statistical sex determination from craniometrics: Comparison of linear discriminant analysis, logistic regression, and support vector machines. Forensic Sci Int. diciembre de 2014;245:204.e1-8.

6. Anexos

6.1 Anexo 1: código de programación en *R*

```
#Carga de los paquetes usados en nuestro análisis
library(visdat)
library(DMwR)
library(randomForest)
library(RSNNS)
library(caret)
library(e1071)
#Carga de los datos de trabajo
Ho←read.csv("Howell.csv")
#Eliminación de algunas de las variables
Ho$Continent←Ho$PopNum
Ho←Ho[,-c(1,3,4)]
#Recodificación de la variable Continent
for (i in 1:nrow(Ho)){if (Ho$Continent[i]==1|Ho$Continent[i]==2|
Ho$Continent[i]==3){Ho$Continent[i]←"Europe"}}
for (i in 1:nrow(Ho)){if (Ho$Continent[i]==16|Ho$Continent[i]==17|
Ho$Continent[i]==18|Ho$Continent[i]==19|Ho$Continent[i]==24|
Ho$Continent[i]==25|Ho$Continent[i]==26|Ho$Continent[i]==28)
{Ho$Continent[i]←"Asia"}}
for (i in 1:nrow(Ho)){if (Ho$Continent[i]==4|Ho$Continent[i]==5|
Ho$Continent[i]==6|Ho$Continent[i]==22|Ho$Continent[i]==23)
{Ho$Continent[i]←"Africa"}}
for (i in 1:nrow(Ho)){if (Ho$Continent[i]==13|Ho$Continent[i]==14|
Ho$Continent[i]==15|Ho$Continent[i]==27){Ho$Continent[i]←"America"}}
for (i in 1:nrow(Ho)){if (Ho$Continent[i]==7|Ho$Continent[i]==8|
Ho$Continent[i]==9|Ho$Continent[i]==10|Ho$Continent[i]==11|
Ho$Continent[i]==12|Ho$Continent[i]==20|Ho$Continent[i]==21|
Ho$Continent[i]==29|Ho$Continent[i]==30){Ho$Continent[i]<-"Oceania"}}
Ho$Continent←factor(Ho$Continent)
#Recodificación de NA y muestra de los mismos
Ho[Ho==0]←NA
vis_miss(Ho[,which(colSums(is.na(Ho))>0)])
#Ausencia de NA tras imputación mediante k-NN
Ho←knnImputation(Ho,k=7)
vis_miss(Ho[,c(37,58,60:62,77:83)])
#Muestra de las variables rotadas mediante PC
Ho.prcomp<-prcomp(Ho[,c(-1,-84)],scale=T)
plot((Ho.prcomp$sdev)^2/(sum((Ho.prcomp$sdev)^2)),xlab="PC",ylab="Propo
rción de la varianza explicada",type="b",col="cadetblue")
#Elección del método de validación cruzada
train←trainControl(method="cv",number=10)
#Creación de los modelos para el sexo y el origen poblacional mediante LDA, con
PC y sin ellas
lda.s.model<-train(x=Ho[,c(-1,-
84)],y=Ho[,1],method="lda",preProcess="pca",trControl=train)
lda.s.model2←train(x=Ho[,c(-1,-
84)],y=Ho[,1],method="lda",trControl=train)lda.op.model←
lda.op.model<-train(x=Ho[,c(-1,-
84)],y=Ho[,84],method="lda",preProcess="pca",trControl=train)
lda.op.model2←train(x=Ho[,c(-1,-
84)],y=Ho[,84],method="lda",trControl=train)
```

```
#Comparación de ambos modelos de LDA
com.lda<-resamples(list("Con PC"=lda.s.model,"Sin PC"=lda.s.model2))
com.lda2<-resamples(list("Con PC"=lda.op.model,"Sin PC"=lda.op.model2))
par(mfrow=c(1,2))
bwplot(com.lda,main="LDA del sexo")
bwplot(com.lda2,main="LDA del origen poblacional")
#Creación de los modelos para el sexo y el origen poblacional mediante NB, con PC y sin ellas
nb.s.model<-train(x=Ho[,c(-1,-
84)],y=Ho[,1],method="nb",preProcess="pca",trControl=train)
nb.s.model2←train(x=Ho[,c(-1,-
84)],y=Ho[,1],method="nb",trControl=train)nb.op.model←nb.op.model<-
train(x=Ho[,c(-1,-
84)],y=Ho[,84],method="nb",preProcess="pca",trControl=train)
nb.op.model2←train(x=Ho[,c(-1,-
84)],y=Ho[,84],method="nb",trControl=train)
#Comparación de ambos modelos de NB
com.nb<-resamples(list("Con PC"=nb.s.model,"Sin PC"=nb.s.model2))
com.nb2<-resamples(list("Con PC"=nb.op.model,"Sin PC"=nb.op.model2))
par(mfrow=c(2,1))
bwplot(com.nb,main="NB del sexo")
bwplot(com.nb2,main="NB del origen poblacional")
#Comparación de los modelos de NB con distintas distribuciones
par(mfrow=c(2,1))
plot(nb.s.model,main="Sexo")
plot(nb.op.model,main="Origen poblacional")
#Estimación de la combinación de hiperparámetros que otorga el menor error OOB para los modelos de RF
tune.1←tuneRF(Ho[,c(-1,-84)],Ho[,1],stepFactor=1.5,improve=1e-
5,ntree=500)
tune.2←tuneRF(Ho[,c(-1,-84)],Ho[,1],stepFactor=1.5,improve=1e-
5,ntree=1000)
tune.3←tuneRF(Ho[,c(-1,-84)],Ho[,1],stepFactor=1.5,improve=1e-
5,ntree=1500)
tune.4←tuneRF(Ho[,c(-1,-84)],Ho[,84],stepFactor=1.5,improve=1e-
5,ntree=500)
tune.5←tuneRF(Ho[,c(-1,-84)],Ho[,84],stepFactor=1.5,improve=1e-
5,ntree=1000)
tune.6←tuneRF(Ho[,c(-1,-84)],Ho[,84],stepFactor=1.5,improve=1e-
5,ntree=1500)
#Creación de un método de RF que permite el ajuste de los hiperparámetros de interés
customRF <- list(type = "Classification", library = "randomForest", loop =
NULL)
customRF$parameters <- data.frame(parameter = c("mtry", "ntree"), class
= rep("numeric", 2), label = c("mtry", "ntree"))
customRF$grid <- function(x, y, len = NULL, search = "grid") {}
customRF$fit <- function(x, y, wts, param, lev, last, weights,
classProbs, ...) { randomForest(x, y, mtry = param$mtry,
ntree=param$ntree, ...) }
customRF$predict <- function(modelFit, newdata, preProc = NULL,
submodels = NULL) predict(modelFit, newdata)
customRF$prob <- function(modelFit, newdata, preProc = NULL, submodels
= NULL) predict(modelFit, newdata, type = "prob") customRF$sort <-
function(x) x[order(x[,1]),]
customRF$levels <- function(x) x$classes
```

```
#Creación de los modelos de RF con dicho método
rf.s.model←train(x=Ho[,c(-1,-
84)],y=Ho[,1],method=customRF,trControl=train,tuneGrid=expand.grid(mtr
y=19,ntree=1500))rf.op.model←
rf.op.model←train(x=Ho[,c(-1,-
84)],y=Ho[,84],method=customRF,trControl=train,tuneGrid=expand.grid(mt
ry=19,ntree=500))
#Muestra de los modelos de RF
com.rf<-resamples(list(" "=rf.s.model," "=rf.s.model))
com.rf2<-resamples(list(" "=rf.op.model," "=rf.op.model))
bwplot(com.rf,main="RF del sexo")
bwplot(com.rf2,main="RF del origen poblacional")
#Creación de los modelos de ANN
ann.s.model←train(x=Ho[,c(-1,-
84)],y=Ho[,1],method="mlpWeightDecayML",preProcess=c("center","scal
e"),trControl=train,tuneGrid=expand.grid(layer1=5,layer2=4,layer3=4
,decay=0.0001))ann.op.model←
ann.op.model←train(x=Ho[,c(-1,-
84)],y=Ho[,84],method="mlpWeightDecayML",preProcess=c("center","sca
le"),trControl=train,tuneGrid=expand.grid(layer1=20,layer2=0,layer3
=0,decay=0))
#Muestra de los modelos de ANN
com.ann<-resamples(list(" "=ann.s.model," "=ann.s.model))
com.ann2<-resamples(list(" "=ann.op.model," "=ann.op.model))
bwplot(com.ann,main="ANN del sexo")
bwplot(com.ann2,main="ANN del origen poblacional")
#Muestra de las matrices de confusión de los modelos
confusionMatrix(lda.s.model)
confusionMatrix(lda.op.model)
confusionMatrix(nb.s.model2)
confusionMatrix(nb.op.model2)
confusionMatrix(rf.s.model)
confusionMatrix(rf.op.model)
confusionMatrix(ann.s.model)
confusionMatrix(ann.op.model)
```

6.2 Anexo 2: variables del DS

Definimos a continuación las variables craneométricas del DS de trabajo:

GOL: Longitud glabelo-occipital
NOL: Longitud nasion-occipital
BNL: Longitud basion-nasion
BBH: Altura basion-bregma
XCB: Anchura craneal máxima
XFB: Anchura frontal máxima
ZYB: Anchura bicigomática
AUB: Anchura biauricular
WCB: Anchura craneal mínima
ASB: Anchura biastérica
BPL: Longitud Basion-prostion
NPH: Altura nasion-prostion
NLH: Altura nasal
JUB: Anchura biyugal
NLB: Anchura nasal
MAB: Anchura máxilo-alveolar
MDH: Altura mastoidea izquierda
MDB: Anchura mastoidea izquierda
OBH: Altura orbitaria izquierda
OBB: Anchura orbitaria izquierda
DKB: Anchura interorbitaria
NDS: Subtensa naso-dacrial
WNB: Cuerda simótica
SIS: Subtensa simótica
ZMB: Anchura bimaxilar
SSS: Subtensa bimaxilar
FMB: Anchura bifrontal
NAS: Subtensa nasio-frontal
EKB: Anchura biorbitaria
DKS: Subtensa dacrial
IML: Longitud malar inferior
XML: Longitud malar máxima
MLS: Subtensa malar
WMH: Altura mínima del pómulo
SOS: Proyección supraorbital
GLS: Proyección glabelar
STB: Anchura biestefánica
FRC: Cuerda frontal
FRS: Subtensa frontal
FRF: Fracción frontal
PAC: Cuerda parietal
PAS: Subtensa parietal
PAF: Fracción parietal
OCC: Cuerda occipital
OCS: Subtensa occipital
OCF: Fracción occipital
FOL: Longitud del agujero occipital
NAR: Radio del nasion
SSR: Radio del subespinal

PRR: Radio del prostion
DKR: Radio del dacrion
ZOR: Radio del cigoorbital
FMR: Radio del frontomalar
EKR: Radio del ectocoquion
ZMR: Radio del cigomaxilar
AVR: Radio molar alveolar
BRR: Radio del bregma
VRR: Radio del vértex
LAR: Radio del lambda
OSR: Radio del opistion
BAR: Radio del basion
NAA: Ángulo del nasion (basion-prostion)
PRA: Ángulo del prostion
BAA: Ángulo del basion (nasion-prostion)
NBA: Ángulo del nasion (basion-bregma)
BBA: Ángulo del basion (nasion-bregma)
BRA: Ángulo del bregma
SSA: Ángulo cigomaxilar
NFA: Ángulo nasio-frontal
DKA: Ángulo dacrial
NDA: Ángulo naso-dacrial
SIA: Ángulo simótico
FRA: Ángulo frontal
PAA: Ángulo parietal
OCA: Ángulo occipital
RFA: Ángulo radio-frontal
RPA: Ángulo radio-parietal
ROA: Ángulo radio-occipital
BSA: Ángulo basal
SBA: Ángulo subbregmático
SLA: Ángulo sublambdoideo
TBA: Ángulo basal transverso