



UNIVERSIDAD
DE MURCIA



TRABAJO FIN DE GRADO

Grado en Biología

Facultad de Biología – Universidad de Murcia

**Aplicación de razonamiento bayesiano y
machine learning a la taxonomía: Clasificando
algunas especies de cítricos (rutáceas)**



Autor: Pablo Manuel López Ruiz

Curso 2023-2024

Ilustraciones de la portada procedentes de Risso, Antoine y Pierre Antoine Poiteau. 1818-1822. Histoire naturelle des orangers. Paris. Impr. de Mme Hérisson Le Doux. <https://gallica.bnf.fr/ark:/12148/bpt6k1512210b>

Agradecimientos

Quiero agradecer a mi tutor, Diego Rivera Núñez y a José Antonio Palazón Ferrando por su ayuda, orientación y dedicación en todo el trascurso del trabajo, he aprendido mucho de vosotros. A mi familia por aguantarme estos años y a mis amigos por, a menudo, acompañarme en este viaje.

Contenidos

Agradecimientos	i
Resumen	1
Abstract	3
1 Introducción	5
1.1 Importancia de los cítricos	5
1.2 Antecedentes	6
1.3 Variación morfológica	7
2 Objetivos	9
3 Metodología	11
3.1 Recogida de muestras	11
3.2 Prensado de las hojas	11
3.3 Escaneado de las muestras	11
3.4 Medición con el software ImageJ	12
3.5 Creación de variables secundarias con R	12
3.6 Modelado de los datos	13
4 Resultados	15
4.1 Dificultades en las mediciones	15
4.2 Variedad del espacio muestral	15
4.3 Diferencias entre los distintos grupos de cítricos	15
4.4 Regresión logística	17
4.4.1 Conjunto de entrenamiento: Naranjas y mandarinas	17
4.4.2 Conjunto de entrenamiento: Limones	18
4.4.3 Comparación	18
4.4.4 Establecimiento del umbral de exigencia óptimo	19
4.5 <i>Random forest</i>	19
4.5.1 Conjunto de entrenamiento: Naranjas y mandarinas	19
4.5.2 Conjunto de entrenamiento: Limones	19
4.5.3 Comparación	20
4.5.4 Establecimiento del umbral de exigencia óptimo	20
4.6 <i>XGBoost</i>	20
4.6.1 Conjuntos de entrenamiento: Naranjas y mandarinas y limones	20
4.6.2 Comparación	21

4.6.3	Establecimiento del umbral de exigencia óptimo	21
4.7	Unión de los modelos	21
4.7.1	Observaciones problemáticas	23
4.8	Creación de modelos con las hojas ancestrales para observar las posibilidades de las hibridadas	23
4.9	Estudio de hojas fósiles	25
5	Discusión	27
5.1	Comprendiendo los resultados	27
5.1.1	Diferencias entre los modelos	27
5.1.2	Diferencias entre los conjuntos de entrenamiento	28
5.1.3	Observaciones problemáticas	28
5.1.4	Inferencia desde los ancestrales	29
5.1.5	Catalogación de hojas fósiles	29
5.2	Contexto del trabajo	30
5.3	Limitaciones y futuras investigaciones	30
6	Conclusiones	33
Relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS)		35
7	Referencias	37
Anexo A		41
1	Código ejecutado y tablas de datos	41
2	Esquemas en relación con los modelos	41
2.1	Regresión logística	41
2.2	Random forest	41
2.3	<i>XGBoost</i>	41

Relación de figuras

1.1 Kumquat y especies ancestrales principales de <i>Citrus</i> con los frutos de sus sucesivos cruzamientos. Modificado de Luro et al. (2017).	7
4.1 Diagrama de cajas de las medidas para los distintos grupos de cítricos, se indica el número de hojas perteneciente a cada grupo. Abreviaturas: max, <i>C. maxima</i> (8); med, <i>C. medica</i> (21); medXmax, <i>C. lumia</i> (26); medXretXmax, <i>C. limon</i> (162); mic, <i>C. micrantha</i> (3); retXmax, <i>C. aurantium</i> (252); otras (18).	16
4.2 Diagrama de cajas de las medidas de las hojas de los distintos grupos de cítricos.	17
4.3 Diagrama de cajas de las variables alométricas para las hojas de los distintos grupos de cítricos seleccionados.	18
4.4 Diagrama de cajas del resto de variables alométricas para las hojas de los distintos grupos de cítricos seleccionados.	19
4.5 Distribuciones de certidumbres de la regresión logística para el conjunto de entrenamiento de las naranjas y mandarinas. 44 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.	20
4.6 Distribuciones de certidumbres de la regresión logística para el conjunto de entrenamiento de los limones. 26 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.	20
4.7 Comparación de la cantidad de estimaciones y la media de acierto con respecto al umbral de clasificación para los dos grupos del modelo de regresión lineal.	21
4.8 Distribuciones de certidumbres del modelo de <i>random forest</i> para el conjunto de entrenamiento de las naranjas y mandarinas. 43 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.	22
4.9 Distribuciones de certidumbres del modelo de <i>random forest</i> para el conjunto de entrenamiento de los limones. 23 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.	22
4.10 Comparación de la cantidad de estimaciones y la media de acierto con respecto al umbral de clasificación para los dos grupos del modelo de <i>random forest</i>	23
4.11 Distribuciones de certidumbres del modelo de <i>XGBoost</i> para el conjunto de entrenamiento de las naranjas y mandarinas. 52 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.	24

4.12 Distribuciones de certidumbres del modelo de <i>XGBoost</i> para el conjunto de entrenamiento de los limones. 42 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.	24
4.13 Comparación de la cantidad de estimaciones y la media de acierto con respecto al umbral de clasificación para los dos grupos del modelo de <i>XGBoost</i>	25
4.14 Hojas identificadas por la integración de los modelos como parte de alguno de los grupos. La certidumbre para Figura 4.14a y Figura 4.14b es mayor al 95 % mientras que para Figura 4.14c es menor al 5 %	26
A.1 Resumen de los modelos de regresión logística para el conjunto de entrenamiento de naranjas y mandarinas (Izquierda) y limones (Derecha). Se observa que el modelo de limones tiene un AIC menor y una variable más importante que el modelo de naranjas y mandarinas.	42
A.2 Gráfico de importancia relativa de las variables que se han tenido en cuenta a la hora de la clasificación con <i>random forest</i> con los limones (Izquierda) y Árbol de decisión de ejemplo que se ha seguido en esta clasificación (Derecha). Se observa que la variable más importante y que separa las observaciones en dos grandes grupos es el area de la lámina entre la longitud del peciolo.	42
A.3 Gráfico de importancia relativa de las variables que se han tenido en cuenta a la hora de la clasificación con <i>random forest</i> con las naranjas y mandarinas (Izquierda) y Árbol de decisión de ejemplo que se ha seguido en esta clasificación (Derecha). Una vez más la variable más importante a la hora de la clasificación vuelve a ser el area de la lámina entre la longitud del peciolo.	43
A.4 Importancia de las variables a la hora de la clasificación con el modelo <i>XGBoost</i>	43
A.5 Árbol de nodos creado en la clasificación con <i>XGBoost</i>	44

Relación de tablas

4.1	Estadísticos de la variabilidad de todas las mediciones realizadas (valores en cm).	15
4.2	Frecuencias iniciales y finales de cada uno de los grupos trabajados.	22
4.3	Intersecciones entre los priores y los modelados. Las asignaciones <i>a priori</i> se muestran en las columnas y las asignaciones inferidas por los modelos en las filas.	22
4.4	Estadísticos de los modelos basados en las especies ancestrales. Umbral de certidumbre: 0.7.	24
4.5	Certidumbres de pertenencia a los grupos ancestrales para las hojas fósiles.	25

Resumen

La taxonomía de *Citrus* ha sido estudiada por muchos autores a lo largo de los años, empleando distintos enfoques y técnicas. Sin embargo, la creciente disponibilidad de herramientas de inteligencia artificial aplicada al análisis de datos abre un abanico de posibilidades para las clasificaciones taxonómicas, ya que gracias a estos algoritmos podemos capturar relaciones antes no detectadas.

Además, debido a la base bayesiana a la que orbitan estos modelos, la integración de las inferencias realizadas por los mismos con el resto de las clasificaciones taxonómicas previamente establecidas parece lo más adecuado para obtener una clasificación robusta y precisa, que no solo se base en una única fuente de información y que sea lo más holista posible.

En este estudio se evalúa el potencial de este enfoque para la generación de una clasificación taxonómica del género mencionado, tomando como base una clasificación genética previa. Para la evaluación, se han clasificado hojas de los árboles disponibles del jardín de las Hespérides, en el campus de Espinardo de la Universidad de Murcia, en función de si pertenecían al grupo de los limones, de las naranjas y mandarinas en conjunto o pertenecían a otra clasificación.

Sobre estas muestras se midieron distintas variables morfológicas utilizando técnicas de análisis de imagen y se aplicaron distintos modelos predictivos para realizar una clasificación de las muestras. Los modelos utilizados son: Regresión logística, *random forest* y *XGBoost*. Estas inferencias se integraron gracias a el paquete **opera** de R para obtener una clasificación final.

Los resultados revelan diferencias significativas entre los grupos de cítricos establecidos, reproduciendo en gran medida lo establecido por la clasificación previa y destacando la capacidad de estos modelos para capturar las relaciones complejas. Sin embargo, esta clasificación difiere en algunos aspectos de la previa, sugiriendo que esta debería de ser revisada y actualizada.

palabras clave: *Citrus*, taxonomía, inteligencia artificial, razonamiento bayesiano, clasificación, morfología foliar.

Abstract

The taxonomy of *Citrus* has been studied by many authors over the years, using different approaches and techniques. However, the increasing availability of artificial intelligence tools applied to data analysis opens up a range of possibilities for taxonomic classifications, as these algorithms allow us to capture previously unrecognized relationships.

Moreover, due to the Bayesian basis on which these models operate, the integration of the inferences made by these models with the rest of the previously established taxonomic classifications seems to be the most appropriate way to obtain a robust and accurate classification that is not based on a single source of information and is as holistic as possible.

This study evaluates the potential of this approach for generating a taxonomic classification of the genus, based on a previous genetic classification. For the evaluation, different *Citrus* leaves from the Hesperides garden, located at the Espinardo campus of the University of Murcia, were classified according to whether they belonged to the lemon group, to oranges and mandarins group, or to another classification.

Several morphological variables were measured on the samples using image analysis techniques and various predictive models were applied to classify them. The models used are: Logistic regression, *random forest* and *XGBoost*. The inferences of these models were then integrated using the R package *opera* to obtain a final classification.

The results show significant differences between the defined *Citrus* groups, largely reproducing the previous classification and highlighting the ability of these models to capture complex relationships. However, this classification differs in some aspects from the previous one, suggesting that it should be reviewed and updated.

keywords: *Citrus*, taxonomy, artificial intelligence, Bayesian reasoning, classification, leaf morphology.

1. Introducción

1.1 Importancia de los cítricos

El género *Citrus*, perteneciente a la familia de las rutáceas. Tuvo su origen en la región comprendida entre el noreste indio y el sureste de China (Swingle y Reece 1967; Talon et al. 2020; Tolkowsky 1938) aunque existen evidencias fósiles que extienden su área al occidente de Eurasia durante el Plioceno (hace unos 4 millones de años) (Fischer y Butzmann 1998) en donde se extinguiría durante las glaciaciones del cuaternario. Las plantas del género ocupan un hueco muy especial en la cultura murciana, siendo objeto de postres, platos y por supuesto condimento gastronómico. Pero esta cultura de los cítricos se remonta en tiempo y espacio hasta los tiempos del mítico primer emperador chino Yu. En estas leyendas, lejos de ser un recurso al alcance de todos, los cítricos servían como tributo del más alto nivel para la corte imperial (Deng et al. 2020; Liu, Heying, y Tanumihardjo 2012).

Desde aquel entonces, el uso de los cítricos se ha extendido por todos los continentes, siendo el primero en ser conocido por nosotros el cidro (*Citrus medica* L., 1753), introducido en la región mediterránea por los griegos en tiempos de Alejandro Magno en el siglo IV aC. La primera descripción se encuentra en la “Investigación sobre las Plantas” de Teofrasto de Eressos (en torno al 310 aC) el cual lo bautizó como la “manzana persa” o la “manzana medica” (por Media, la región entre el mar Caspio y el río Tigris)(Deng et al. 2020; Isaac 1959; Langgut 2017).

Otra fuente de entrada sin duda fueron las rutas comerciales que Al-Ándalus y el sur de Italia mantenían con China, siendo responsables de la introducción de los naranjos, de la cimboba y posiblemente del limón y la lima entre otros, expandiendo estas frutas por todo el mediterráneo (Barbera 2023; Cascales 1873; Ramón-Laca 2003; Tolkowsky 1938). Además de su relevancia histórica, las frutas producidas por el género tienen una gran importancia económica, habiendo sido la segunda fruta más producida del mundo en el año 2021, solo por detrás de las bananas (Pereira-Gonzatto y Scherer-Santos 2023). Países como China, Brasil, Estados Unidos, México y España lideran la producción mundial de naranjas, mandarinas, limas y limones entre otros, habiendo llegado a producir en nuestro país seis millones de toneladas en el año 2019 (FAO 2021; Spreen et al. 2020).

La versatilidad de estos frutos es destacable, ya que pueden ser consumidos como fruta, parte de platos elaborados, zumo, mermelada, aceite esencial, comida para ganado e incluso como ingredientes en medicina tradicional. Esta última aplicación se remonta a obras antiguas como “Bencao Gangmu” o “Compendio de la Materia Médica”, escrito por Shizheng Li en 1596 durante la dinastía Ming, donde se describen algunos de los usos medicinales que se les daban a los cítricos (Deng et al. 2020). En la actualidad, el interés por las propiedades saludables y medicinales de los cítricos ha experimentado un aumento. Compuestos como el ácido ascórbico (vitamina C), además de prevenir el escorbuto (Halliwell y Gutteridge 2015), cuenta con propiedades antioxidantes que se parecen vincular con una

menor incidencia de enfermedades cardiovasculares (Ma et al. 2020; Morelli et al. 2020) y cáncer (Ma et al. 2020; Mastrangelo et al. 2018).

1.2 Antecedentes

Varios enfoques han intentado abarcar la diversidad del género, con el fin de llegar a una clasificación taxonómica certera, no obstante esto es una tarea complicada debido, precisamente a la falta de documentación y la larga historia de hibridaciones de *Citrus* (Luro et al. 2017). Hoy en día existen diversidad de opiniones en torno a esta cuestión debido a la diferencia en los caracteres utilizados históricamente para la distinción de las especies, generando las distintas clasificaciones que observamos en la actualidad.

Históricamente, destacan dos clasificaciones principales: la propuesta por Swingle y Reece (1967), que identifica 16 especies, y la propuesta por Tanaka y Furuta (1977), que postula la existencia de 162. Esta gran disparidad surge por las diferentes estrategias y enfoques empleados en la generación de las clasificaciones. Pese a que ambas difieren en cuanto al número de especies, realmente se trata más de una cuestión de fronteras y por tanto metodológica que de una cuestión taxonómica, ya que ambos utilizan características morfológicas (principalmente del fruto) como marcadores para la taxonomía, además, los resultados de estos autores se encontraban sesgados por la muestra que utilizaron en su clasificación.

Recientemente Ollitrault, Curk, y Krueger (2020) propusieron una taxonomía con base filogenética, plasmada en Rivera et al. (2022) con la publicación de *Citrus × limon* var. *limetta* (Risso) Ollitrault, Curk & R.Krueger, pese a que todavía está pendiente de consolidación, esta clasificación simplifica la taxonomía, unificando en función de una historia genética compartida a especies que anteriormente se consideraban diferentes como es el caso de *Citrus × aurantium* L., 1753, y *Citrus deliciosa* Ten., 1840, que con esta hipótesis pasarían a ser *C. × aurantium* var. *aurantium* y *C. × aurantium* var. *deliciosa* respectivamente. Además de esta unificación también tiene en cuenta las complejas hibridaciones que han ocurrido en la historia del género y que han dado lugar al grueso de especies actuales, la mayoría a partir de únicamente cuatro especies parentales: *C. medica*, *Citrus reticulata* Blanco., 1837, *Citrus maxima* (Burm.) Merr., 1917, y *Citrus micrantha* Wester., 1915; también llamado en la literatura como *Citrus hystrix* DC., 1813 (Figura 1.1). Esta taxonomía será la base de la que partiremos en este trabajo.

El avance en este campo ha sido cambiante, ya que las observaciones realizadas se utilizaron para crear nuevas hipótesis sobre las clasificaciones previas, cambiando los taxones a la vista de nuevas evidencias. La morfología inicialmente proporcionó una comprensión superficial, que con la llegada de los marcadores moleculares, adquirió mayor detalle, permitiendo explorar las relaciones genéticas que se dan entre especies.

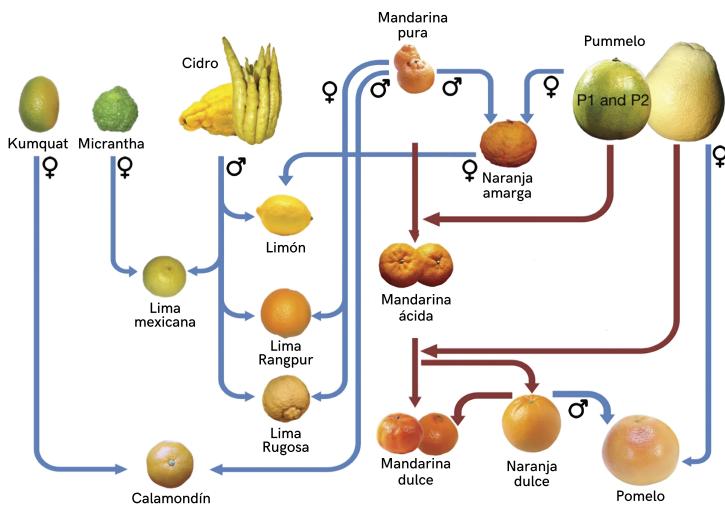


Figura 1.1: Kumquat y especies ancestrales principales de *Citrus* con los frutos de sus sucesivos cruzamientos. Modificado de Luro et al. (2017).

En la actualidad, la búsqueda de nueva información continúa. Adoptando nuevos puntos de vista enfocados en la continua observación y en la actualización de nuestras hipótesis. Estas premisas fundamentan el razonamiento bayesiano, que se basa en una mentalidad de actualización continua. Este enfoque aborda de manera activa la incertidumbre inherente a la taxonomía, integrando sistemáticamente la evidencia disponible con nuestras creencias previas. formando hipótesis más certeras y minimizando sesgos. Además, al considerar múltiples fuentes de evidencia, no solo enriquecemos nuestras conclusiones sino que también promovemos un enfoque más integral y colaborativo en el estudio taxonómico.

En este contexto se han buscado nuevas variables que puedan sumar a la clasificación y que contribuyan a una mejor representación de la realidad. Debido a la accesibilidad y la gran cantidad de información que se puede obtener, se ha decidido centrar la atención en la morfología foliar. Pese a que difiere del carácter genético que ha sido últimamente destacado (Munankarmi et al. 2023; Shi et al. 2023) e históricamente se ha visto relegada en favor del fruto, cuenta con una gran cantidad de valores medibles y cuantificables que la convierten en un posible elemento de gran valor en el estudio taxonómico.

1.3 Variación morfológica

La larga y enrevesada historia de este género con nuestra cultura, ha dado lugar a una amplia variedad de híbridos con características morfológicas distintivas y diferenciadas. Estas suelen presentar hábito arbóreo, con copas que van desde formas achataadas hasta ovaladas, las ramas que conforman el árbol pueden variar en la densidad de espinas, desde la ausencia hasta la presencia de espinas largas y numerosas (IPGRI 1999).

Sus flores son típicamente pequeñas, blancas o moradas y con un marcado olor, con cinco pétalos y pequeños sépalos rodeando a estos. Los estambres de color blanco y amarillo, rodeando el pistilo en

posición central. El fruto varía en muchos aspectos: el zumo va desde dulce a extremadamente ácido, y la corteza y membranas comprenden sabores dulces y amargos, el tamaño es muy variado, la forma es oblonga, piriforme, ovoide, redondeada o aplanada, entre otros (Liu, Heying, y Tanumihardjo 2012). Con colores que van desde el verde oscuro hasta el rojo anaranjado, algo que depende de factores genéticos y ambientales.

Esta gran variedad de características también se encuentra en las hojas de las distintas especies. Encontrándonos hojas que abarcan desde los 4 cm^2 a los 40 cm^2 . El pecíolo puede estar ausente, presente sin alas o presente alado, la lámina cuenta con forma y borde variable (IPGRI 1999).

Las hojas de *C. reticulata*, según Swingle y Reece (1967) presentan una forma lanceolada-elíptica, con un tamaño aproximado de $6.8 \times 2.5 \text{ cm}$ de largo y ancho respectivamente. Su base se estrecha de manera aguda hacia un ápice obtuso, mientras que destacan alrededor de 10 pares de venas laterales. El pecíolo es mínimamente alado.

Por otro lado, descrito por Wester (1915), las hojas de *C. micrantha* varían en tamaño, siendo de 9 a 12 cm de largo y de 2.7 a 4 cm de ancho. Con carácter elíptico-ovalado, estas se caracterizan por un pecíolo alado de hasta 6 cm de longitud y 4 cm de ancho.

Con relación a *C. medica*, según Swingle y Reece (1967), las hojas muestran una apariencia lisa y un contorno elíptico-ovalado u ovado-lanceolado. Con un ápice redondeado o ligeramente puntiagudo, estas hojas se unen al pecíolo, que carece de alas o presenta márgenes estrechos.

Según los mismos autores en el mismo año (Swingle y Reece 1967), las hojas de *C. maxima* son grandes, desde ovaladas a elíptico-ovaladas, con un ápice romo y base redondeada. Pueden superponerse ligeramente al pecíolo alado, que presenta márgenes más o menos cordados.

Para los híbridos encontramos dos categorías de principal importancia: *C. medica* × *C. maxima* × *C. reticulata*, que engloba a especies como el limón común (*Citrus × limon* (L.) Osbeck var. *limon*) o la bergamota (*C. × limon* var. *bergamia* (Loisel.)). Estas presentan hojas de forma elíptica, con un ápice redondeado y una base cuneada, con un pecíolo alado de 1 a 2 cm de longitud (Ollitrault, Curk, y Krueger 2020).

Y *C. maxima* × *C. reticulata*, que no establece categorías completamente discretas si no que forma un gradiente de cruzamientos donde se distinguen grupos muy distintos de especies en función de la proporción de cada uno de los parentales, englobando cosas tan distintas como los pomelos (*Citrus × aurantium* L. var. *paradisi*), las naranjas (*Citrus × aurantium* L. var. *aurantium* o *Citrus × aurantium* L. var. *sinensis*) o las mandarinas (*Citrus × aurantium* var. *clementina*) entre otras) (Ollitrault, Curk, y Krueger 2020). Sus hojas cuentan con tamaños muy variados, pecíolos con ala o sin ella y forma menos lanceolada que la del grupo de los limones, debido a la gran diversidad intragrupal se hace difícil establecer una descripción general.

2. Objetivos

El objetivo principal de este trabajo es el de estudiar la morfología de las hojas de *Citrus* y comprender su papel en la determinación taxonómica. Creando un modelo que clasifique a las hojas en función de su morfología foliar, obteniendo así una clasificación con base genética pero con nuevos datos morfológicos. Este se basa en la clasificación realizada por Ollitrault, Curk, y Krueger (2020).

Además se busca evaluar la aplicabilidad de la metodología escogida, comprobando su potencia y facilidad a la hora de la interpretación de los resultados, teniendo en cuenta la consistencia entre los distintos modelos metodológicos seleccionados.

3. Metodología

3.1 Recogida de muestras

En la recogida de muestras, se ha usado como espacio de muestreo, el jardín de las hespérides de la Universidad de Murcia, con coordenadas 38°01'10"N 1°10'02"W. En este jardín encontramos una gran variedad de árboles de cítricos.

Para la recogida se ha seguido un patrón de muestreo aleatorio, recogiendo de entre 3 y 8 hojas maduras de la región media-superior del árbol. Para la minimización del sesgo, las hojas fueron recogidas intentando abarcar la totalidad de la copa del árbol, para que cuestiones como la irradiancia solar o la humedad no influyeran en el modelo.

Estas hojas son posteriormente almacenadas en sobres previamente rotulados con un código que identifica la localización del árbol del que se han extraído y la fecha del muestreo. Todas las hojas se recogieron preservando tanto la lámina en su totalidad como el peciolo, para conseguir un conjunto robusto de datos que permita el análisis propuesto. No se distinguió entre las hojas sanas y las hojas que presentaban algún tipo de deficiencia o enfermedad siempre y cuando estas no tuvieran regiones necróticas.

3.2 Prensado de las hojas

En el prensado, se han usado prensas de herbario que permiten el secado de las hojas en un periodo de tiempo de 3-5 días, para absorber la humedad se dispusieron las hojas sobre papel de periódico y almohadillas higroscópicas.

Una de las dificultades que supone el prensado de las hojas es la pérdida de color o la aparición de manchas debido a la humedad retenida. No obstante esto no ha sido un problema de grave para el proyecto ya que, en su mayoría, los secados se han realizado con rapidez y en los casos en los que la hoja se ha manchado, la morfología no ha sido alterada.

Tras el prensado de las hojas, se almacenaron en un lugar seco hasta la digitalización en conjunto de todas ellas. Es aquí donde se ha presentado otra dificultad en esta fase, muchas hojas preservaban cierta humedad y esto ha provocado que desde el momento que se prensaron hasta la digitalización, se hayan abarquillado, inutilizándolas para la misma.

3.3 Escaneado de las muestras

El escaneado se llevó a cabo con un escáner Arcus 1200, con una resolución de 300 dpi, en RGB y las imágenes se almacenaron en formato "tif". Se escaneó el envés de entre 2 y 6 hojas de cada muestra prensada, seleccionando primero de forma aleatoria de cada muestra y corroborando después que las

hojas no presentaban modificaciones morfológicas aberrantes que pudieran dar a cabo un análisis erróneo.

3.4 Medición con el software ImageJ

La medición de los parámetros en sí fue realizada con el software de código abierto Fiji (ImageJ). Se trabajó sobre las imágenes escaneadas conseguidas y se midieron los siguientes parámetros:

1. Longitud del peciolo (lp)
2. Longitud de la lámina, con el raquis como eje de la misma (l1)
3. Ancho máximo del peciolo (wp)
4. Ancho mínimo del peciolo (sp)
5. Ancho máximo de la lámina (wl)
6. Distancia entre el punto máximo de la lámina y el ápice de la misma. (hl)
7. Tipología del ápice de la lámina (Presencia o ausencia) (p1)
8. Tipología del margen de la lámina (Crenado, dentado, entero, sinuado u otro (IPGRI 1999)) (ml)

3.5 Creación de variables secundarias con R

Se analizaron 490 hojas provenientes de más de 150 árboles distintos de *Citrus*. Los datos se organizaron utilizando tanto R como Excel, creando una tabla de datos o *dataframe* con la que posteriormente se generaría el modelo. Gracias al análisis relativo de los antepasados de un gran número de especies de este género realizado por Rivera et al. (2022), se creó una variable en función de la filogenia (fil), teniendo en cuenta la participación de las especies antecesoras en el cruce actual, con esta variable se generaron 7 grupos: *C. maxima*, *C. medica*, *C. aurantium*, *C. limon*, *C. micrantha*, *Citrus × lumia* Risso & Poit. y otros parentales.

También se generaron las siguientes variables secundarias:

1. Relación entre el ancho máximo de la lámina y la longitud de la lámina (wlDl1)
2. Relación entre el ancho máximo del peciolo y el ancho máximo de la lámina (wpDwl)
3. Relación entre el ancho máximo del peciolo y la longitud del peciolo (wpDlp)
4. Relación entre la longitud del peciolo y la longitud de la lámina (lpDl1)
5. Relación entre el ancho mínimo del peciolo y el ancho máximo del peciolo (spDwp)
6. Relación entre el ancho máximo de la lámina y la longitud desde el punto de máxima anchura de la lámina hasta el ápice (wlDhl)
7. Relación entre el ancho máximo de la lámina y la longitud del punto de máxima anchura (l1MhlDl1)
8. Relación entre el área de la lámina y la longitud del peciolo (alDlp)
9. El área aproximada de la lámina (al)

3.6 Modelado de los datos

El análisis se realizó con tres tipos de enfoque distintos. Por un lado se utilizó una regresión logística con el paquete **stats** que forma parte de R (R Core Team 2013) en el marco de RStudio, para comprobar si la morfología de las hojas se ajustaba a algún tipo de clusterización que correspondiese con la clasificación taxonómica. También se empleó un enfoque de aprendizaje automático o *machine learning*, concretamente el algoritmo del paquete **Random Forest** (Breiman et al. 2024) para asignar predicciones a las especies a partir de la morfología de sus hojas. Y por último, se utilizó el paquete **XGBoost** (T. Chen 2024) con el cual se intentó complementar los resultados obtenidos con el resto de modelos.

Previo a la utilización de los modelos, estos deben de “entrenar”, para ello, se introduce un conjunto de datos con todas las variables (*training set*) con el cual el modelo va a ajustar sus parámetros para luego extrapolar al resto de datos. Se utilizaron distintos tipos de subconjuntos para evitar el sesgo de una selección dada. Finalmente se usó el subconjunto formado por las naranjas y mandarinas (*C. reticulata x C. maxima*) y el formado por los limones (*C. medica x C. reticulata x C. maxima*)

El modelo de regresión logística busca la relación entre una variable dependiente categórica (pertenencia o ausencia a un grupo) y varias variables independientes (variables medidas y calculadas) a través de un proceso iterativo ajustando los coeficientes de las variables independientes para maximizar la verosimilitud de los datos observados (LaValley 2008). Atribuyendo valores de certidumbre de pertenencia o ausencia a las hojas en función de las variables medidas.

El modelo de *random forest* se basa en un algoritmo de aprendizaje supervisado que utiliza un conjunto de variables predictoras para estimar la clasificación de un evento a través de la construcción de un gran número de árboles de decisión (Biau y Scornet 2016). Cada uno de estos árboles se entrena con un subconjunto aleatorio de las observaciones y variables del conjunto de entrenamiento o *training set* establecido, esto se realiza para minimizar el sobreajuste, término que se utiliza para referirse a los modelos que predicen muy bien un conjunto de muestras pero son muy poco extrapolables. En nuestro caso, el evento sería la clasificación de una hoja en una de las categorías establecidas, y las variables predictoras serían variables primarias y secundarias medidas.

Por último, el modelo de *XGBoost* es también un algoritmo de aprendizaje supervisado que se basa en la creación de un conjunto de árboles de decisión, pero a diferencia de *random forest*, en este caso los árboles se construyen de manera secuencial, es decir, cada árbol se construye en función de los errores del anterior, mejorándolo en cada iteración (T. Chen y Guestrin 2016). También cuenta con técnicas para la reducción del sobreajuste.

Para comprobar la certeza de las predicciones se utilizaron dos estadísticos:

- **Media de acierto:** Conjunto de predicciones correctas sobre el número total de predicciones

realizadas.

$$\frac{P_u \cap C}{P_u}$$

- **Representatividad:** conjunto de predicciones correctas sobre el número total de muestras pertenecientes al grupo.

$$\frac{P_u \cap C}{C}$$

Donde:

- P_u es el conjunto de muestras con una certidumbre de pertenencia al grupo mayor o igual al umbral u
- C es el conjunto de muestras pertenecientes al grupo

Con los tres modelos contamos con aproximaciones que proporcionan una variable con la certidumbre de pertenencia a un determinado grupo de entrenamiento, teniendo en cuenta desde relaciones simples con la regresión logística hasta relaciones más complejas con los modelos de *random forest* y *XGBoost*. Estos resultados se combinaron para optimizar las inferencias, para la agregación se utilizaron los paquetes `metrics` (Hamner, Frasco, y LeDell 2024) y `opera` (Gaillard et al. 2024) de R. La agregación, plenamente bayesiana integra la evidencia de múltiples modelos, mejorando la precisión y confiabilidad de las predicciones.

El *script* con el código de R que fue utilizado para todas las tareas descritas se incluye en el anexo.

4. Resultados

4.1 Dificultades en las mediciones

En cuanto al estudio del margen de la hoja (*m1*), se presentaron dificultades, particularmente en la atribución a uno u otro de los tipos reconocidos en el manual para la descripción de los cítricos. lo que llevó finalmente a descartar esta medida del análisis. La visibilidad limitada del carácter y la posibilidad de errores en la toma de medidas podrían haber comprometido la calidad de los datos y, por tanto, la validez de los resultados obtenidos.

Por otro lado, la naturaleza del carácter ápice (*p1*), en el que se intentaba medir la presencia de un hueco en el extremo apical de la lámina, también fue descartado, ya que este carácter se encontraba presente de una forma u otra en la mayoría de hojas y por tanto no aportaba información relevante al estudio.

4.2 Variedad del espacio muestral

Los valores máximos y mínimos muestreados, así como otros estadísticos, se presentan en la Tabla 4.1 para cada una de las variables primarias.

Tabla 4.1: Estadísticos de la variabilidad de todas las mediciones realizadas (valores en cm).

Variable	Mínimo	Máximo	Media	σ
Longitud del peciolo	0.14	4.92	1.00	0.53
Anchura máxima del peciolo	0.04	3.56	0.25	0.31
Anchura mínima del peciolo	0.04	0.26	0.13	0.04
Longitud de la lámina	1.35	15.57	8.01	2.22
Anchura máxima de la lámina	0.69	9.59	3.97	1.39
Distancia entre el punto de anchura máxima y el ápice de la lámina	0.54	9.03	4.22	1.33

4.3 Diferencias entre los distintos grupos de cítricos

Se compararon las medidas de los distintos grupos generados por la variable filogenética mencionada (*fil*) a través de diagramas de cajas para observar la variabilidad de las mismas. Los resultados se presentan en la Figura 4.1.

Los grupos “max”, “med” y “mic” que pertenecen respectivamente al pummelo, al cidro y a la lima kaffir, contaban con muy poco número de muestras, de forma que pese a que se incluyen en el modelo, es difícil realizar inferencias sobre ellas. En el grupo de “otras” encontramos especies diversas que provienen de cruzamientos de cítricos con especies que se encuentran fuera del rango de las 4 especies ancestrales principales consideradas y por esto no se incluye en la caracterización.

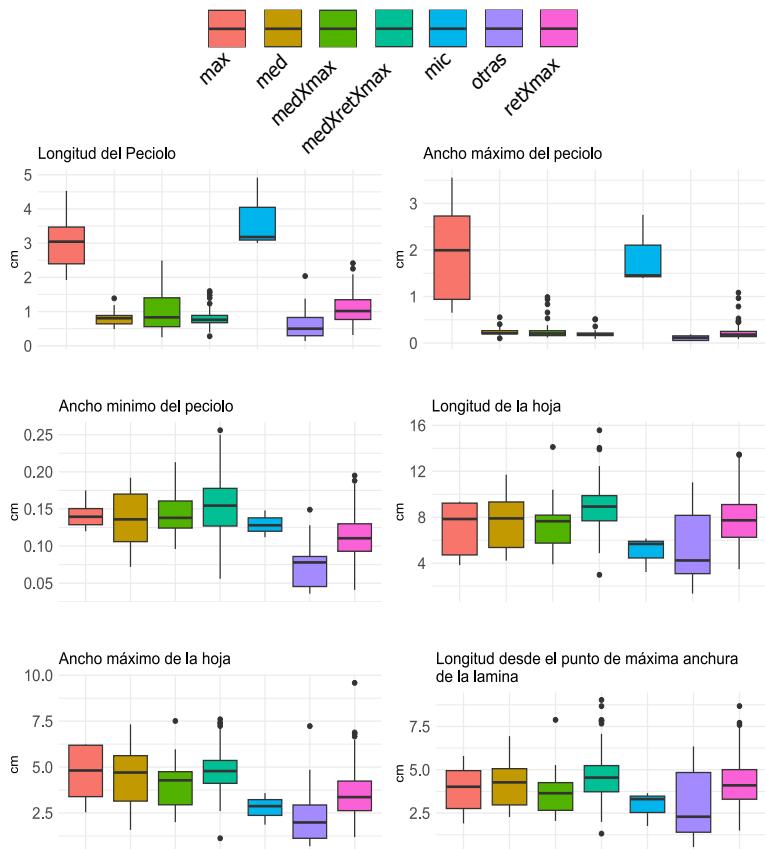


Figura 4.1: Diagrama de cajas de las medidas para los distintos grupos de cítricos, se indica el número de hojas perteneciente a cada grupo. Abreviaturas: max, *C. maxima* (8); med, *C. medica* (21); medXmax, *C. lumia* (26); medXretXmax, *C. limon* (162); mic, *C. micrantha* (3); retXmax, *C. aurantium* (252); otras (18).

Para “medXmax” contamos con las lumias (*C. lumia*), estas son poco numerosas y cuentan con orígenes dispares (Curk et al. 2016), por tanto no se incluirá en los conjuntos de entrenamiento. En el subconjunto “medXretXmax” encontramos a los limones y en “retXmax” se ven representados las naranjas, las mandarinas hibridas y los pomelos. Estos dos últimos grupos van a ser los que finalmente se utilicen para la modelación debido a que presentan homogeneidad filogenética y un alto número de individuos muestreados. Tras estas aclaraciones podemos, de nuevo, formar el diagrama de cajas separando los grupos que nos interesan, es decir, los limones por un lado y las naranjas junto con las mandarinas por otro.

Observamos (Figura 4.2) que las mayores diferencias se dan en la longitud y anchura mínima del pectíolo, así como en el ancho máximo de la lámina, no obstante, incluso en las variables más diferenciadas, el grado de solapamiento entre los dos grupos es muy grande. Para mejorar los descriptores y teniendo en cuenta la gran variabilidad de tamaños, se crearon los índices o variables secundarias mencionados en la metodología, ahora compararemos estos entre los dos grupos formados.

En las Figuras 4.3 y 4.4 se observa que, aunque persiste el solapamiento en muchas medidas, las variables secundarias en general se separan más entre los dos grupos, aumentando su potencial descriptivo. Los índices más diferenciados son el área de la lámina dividido por la longitud del pectíolo,

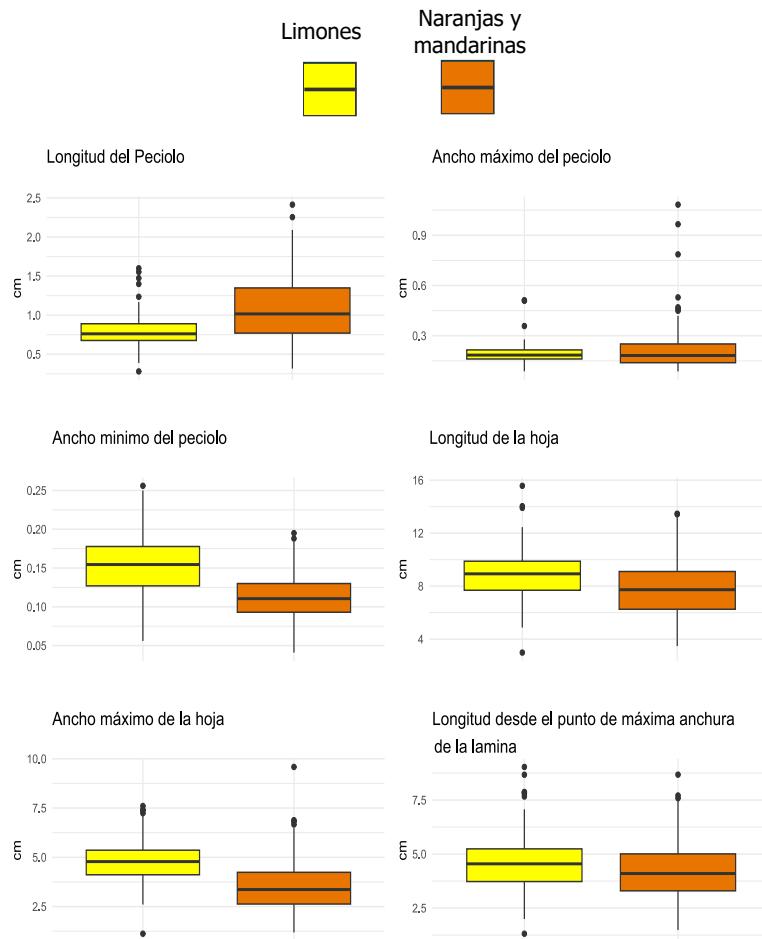


Figura 4.2: Diagrama de cajas de las medidas de las hojas de los distintos grupos de cítricos.

el ancho mínimo del pecíolo dividido por su ancho máximo y la longitud del pecíolo dividido por la de la lámina.

Ahora sí, con las 414 hojas de naranjas, mandarinas y limones para formar los grupos de entrenamiento, es el momento de realizar la clasificación usando los modelos mencionados en el apartado de la metodología.

4.4 Regresión logística

4.4.1 Conjunto de entrenamiento: Naranjas y mandarinas

Para el modelo de regresión logística se han incluido todas las variables primarias y secundarias. La distribución resultante con el conjunto de naranjas y mandarinas (*C. reticulata* × *C. maxima*) se recoge en la Figura 4.5. Esta representa la certidumbre de pertenencia a este grupo para todas las observaciones, ordenadas de menor a mayor.

El resultado coincide con el esperado, ya que se aprecian dos grupos muy numerosos extremos que representan los que no pertenecen (cercaos a 0) y pertenecen con mayor confianza (cercaos

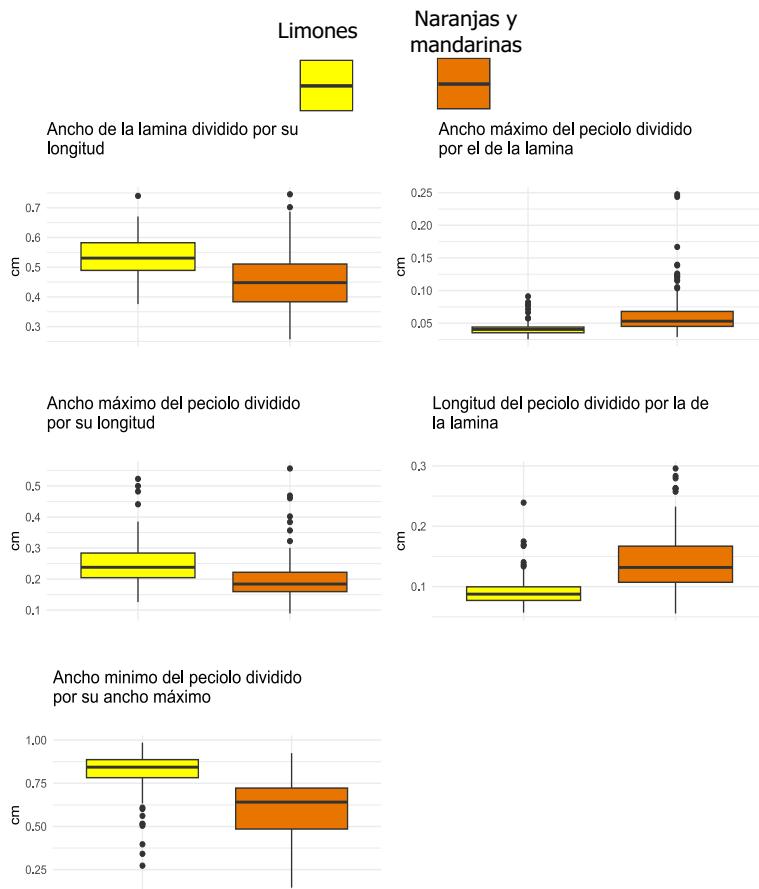


Figura 4.3: Diagrama de cajas de las variables alométricas para las hojas de los distintos grupos de cítricos seleccionados.

a 1). Además de un grupo menor central, que representa las hojas que no se pueden clasificar con seguridad.

4.4.2 Conjunto de entrenamiento: Limones

Para el conjunto de entrenamiento de los limones (*C. medica* × *C. reticulata* × *C. maxima*), se prosiguió de forma idéntica a la anterior, en este caso los resultados que se obtuvieron fueron distintos. En la Figura 4.6 se observa que pese a que la distribución es parecida, la cantidad de muestras certeras es menor debido principalmente al menor tamaño muestral. La cantidad de hojas que no se pueden clasificar con seguridad ha resultado similar.

4.4.3 Comparación

También se ha hecho una comparación del rendimiento de los conjuntos de entrenamiento con respecto a la certidumbre de la clasificación establecida. Se ha calculado la media de acierto del subconjunto y su representatividad (Figura 4.7). Los gráficos superiores corresponden con una curva logarítmica que indica la media de aciertos. En los inferiores observamos curvas inversas, relativas a la caída de la representatividad por la subida de exigencia de la clasificación. Se dan diferencias entre ambos

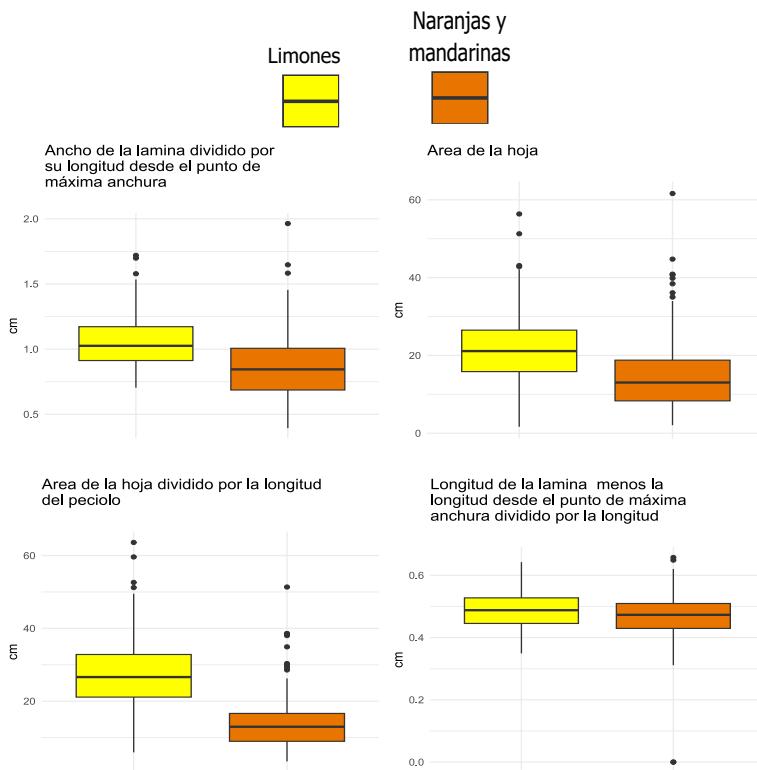


Figura 4.4: Diagrama de cajas del resto de variables alométricas para las hojas de los distintos grupos de cítricos seleccionados.

subconjuntos, la curva de las naranjas y mandarinas se aleja más de la diagonal debido a que la mayor cantidad de hojas que pertenecen al grupo retrasa la caída de la representatividad.

4.4.4 Establecimiento del umbral de exigencia óptimo

Estudiando las curvas, el umbral óptimo para ambos conjuntos se establece en 0.70, en este punto se tiene representado el 78 % de la población de naranjas y mandarinas y el 67 % de los limones con una tasa de acierto del 91 % y del 85 % respectivamente.

4.5 Random forest

4.5.1 Conjunto de entrenamiento: Naranjas y mandarinas

Para *random forest*, utilizando el subconjunto de las naranjas y mandarinas se obtuvo el reparto observando en la Figura 4.8. Este es muy parecido al observado para la regresión lineal (Figura 4.5), con leves diferencias en las frecuencias pero similar a grandes rasgos.

4.5.2 Conjunto de entrenamiento: Limones

Con los limones, la distribución observada se aprecia en la Figura 4.9. En este caso la variación con Figura 4.6 la encontramos en la forma de la distribución, no obstante, a grandes rasgos los resultados son muy parecidos.

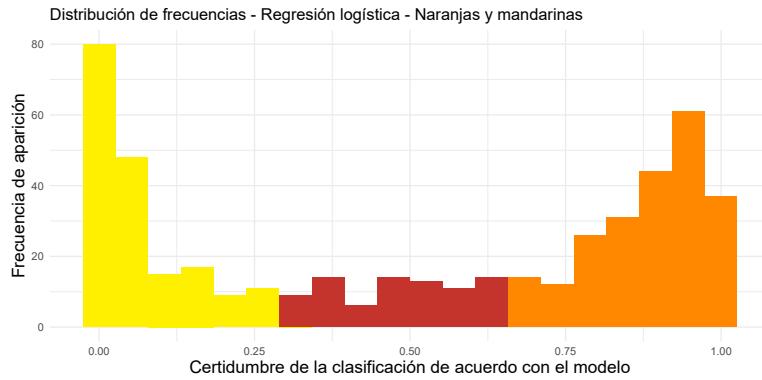


Figura 4.5: Distribuciones de certidumbres de la regresión logística para el conjunto de entrenamiento de las naranjas y mandarinas. 44 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.

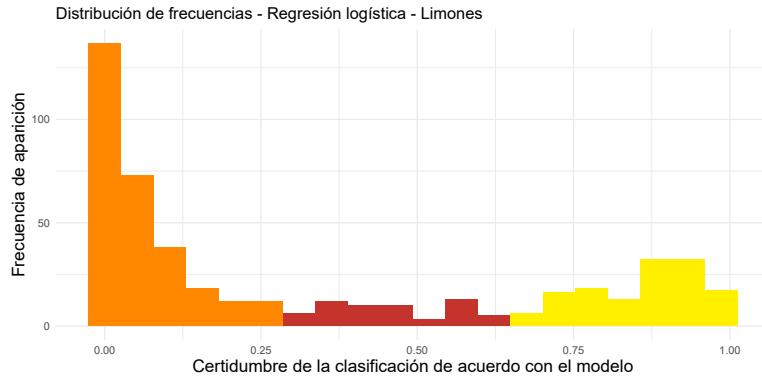


Figura 4.6: Distribuciones de certidumbres de la regresión logística para el conjunto de entrenamiento de los limones. 26 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.

4.5.3 Comparación

También se ha realizado una comparativa del rendimiento usando los mismos parámetros empleados anteriormente (Figura 4.10). Estas curvas son prácticamente idénticas a las observadas en el modelo anterior, la única diferencia apreciable se da en que las curvas son algo más abruptas, especialmente para el grupo de los limones, posiblemente por el menor número de muestras de este subconjunto, además de la propia metodología del modelo.

4.5.4 Establecimiento del umbral de exigencia óptimo

Observando estas curvas, y con el fin de homogeneizar los resultados, se establece nuevamente el umbral óptimo en 0.70, para este valor, se cuenta con una representatividad del 75 % de la población de naranjas y mandarinas y un 64 % de los limones con una tasa de acierto del 91 % en ambos casos.

4.6 XGBoost

4.6.1 Conjuntos de entrenamiento: Naranjas y mandarinas y limones

Para este modelo, debido a que no está sujeto a la dicotomía de los anteriores (pertenencia o no pertenencia), se creó un único modelo multiclasa con los dos conjuntos de entrenamiento. Las

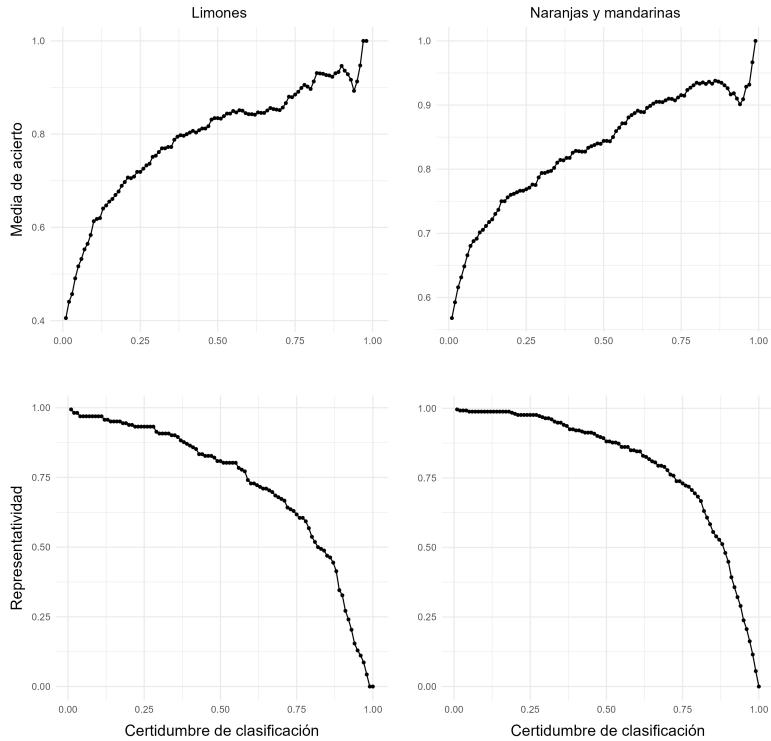


Figura 4.7: Comparación de la cantidad de estimaciones y la media de acierto con respecto al umbral de clasificación para los dos grupos del modelo de regresión lineal.

distribuciones de las certidumbres de estos dos grupos se pueden observar en las Figura 4.11 y Figura 4.12.

4.6.2 Comparación

La comparativa de rendimiento observada en la Figura 4.13, muestra medias de acierto más irregulares, con especial atención a las inferidas para las naranjas y mandarinas, pero representatividades muy altas hasta umbrales avanzados, representando que la mayoría de las hojas catalogadas se encuentran en los extremos de la clasificación.

4.6.3 Establecimiento del umbral de exigencia óptimo

Con el umbral en 0.70, el modelo consiguió una representatividad del 94 % de la población de naranjas y mandarinas con una tasa de acierto del 79 % y un 86 % de los limones acertando el 79 % de las ocasiones.

4.7 Unión de los modelos ¹

Para dar mayor peso a la hipótesis propuesta de clasificación, se integraron los tres modelos siguiendo un esquema de actualización bayesiana, el cual nos deja con los niveles poblacionales de las Tablas 4.2 y 4.3. Con el modelo integrado y usando el mismo umbral, la media de acierto es del 89 % para las naranjas y mandarinas y del 86 % para los limones, las representatividades son del 89 % y

¹en el anexo se detallan los resultados obtenidos por los modelos empleados, así como las variables más importantes para la clasificación de las hojas entre otros aspectos de interés.

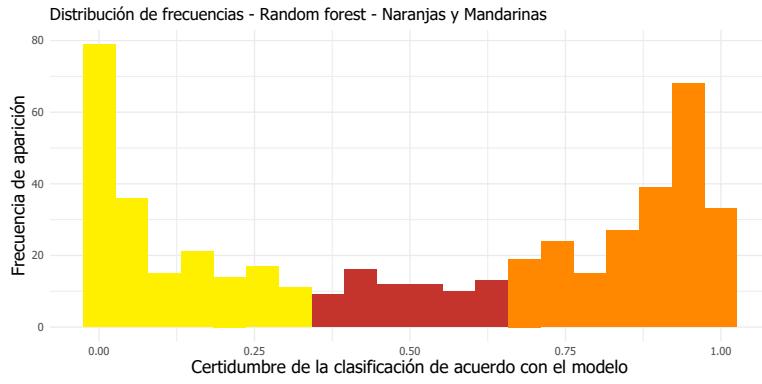


Figura 4.8: Distribuciones de certidumbres del modelo de *random forest* para el conjunto de entrenamiento de las naranjas y mandarinas. 43 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.

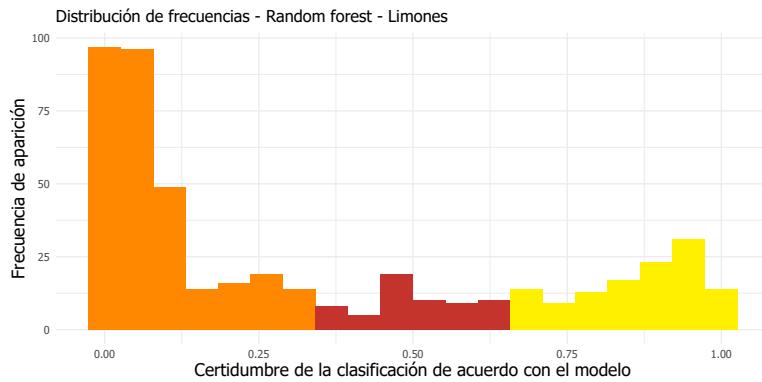


Figura 4.9: Distribuciones de certidumbres del modelo de *random forest* para el conjunto de entrenamiento de los limones. 23 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.

78 % respectivamente. En la Figura 4.14 se pueden observar las morfologías de algunas de las hojas clasificadas.

Tabla 4.2: Frecuencias iniciales y finales de cada uno de los grupos trabajados.

	Limones	Naranjas y mandarinas	Otros
Asignación <i>a priori</i>	162	252	76
Asignación según modelo	137	221	132

Tabla 4.3: Intersecciones entre los priores y los modelados. Las asignaciones *a priori* se muestran en las columnas y las asignaciones inferidas por los modelos en las filas.

Intersecciones entre previos y modelos	Limones <i>a priori</i>	Naranjas y mandarinas <i>a priori</i>	Otros <i>a priori</i>
Limones inferidos	115	3	19
Naranjas y mandarinas inferidas	4	199	18
Otros inferidos	43	50	39

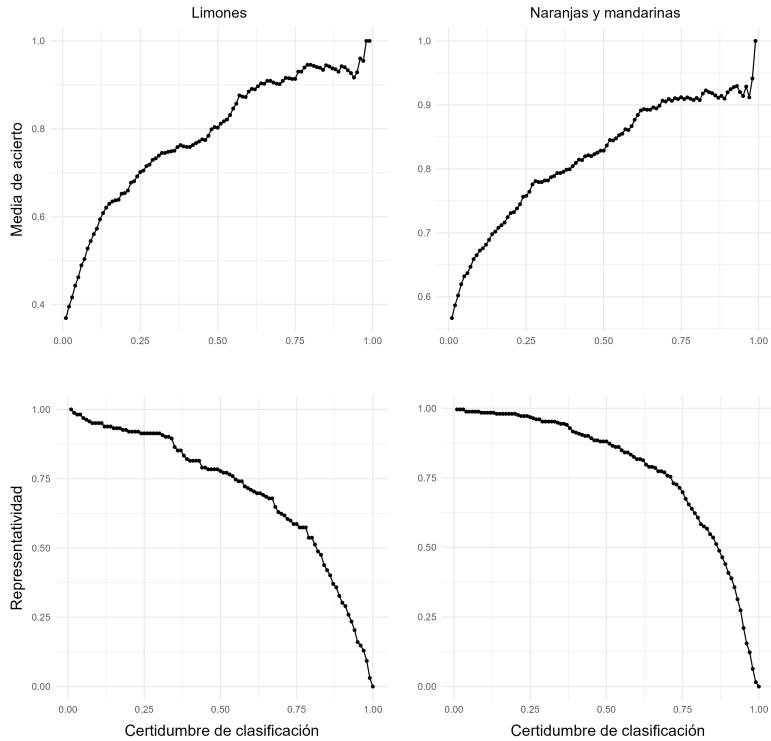


Figura 4.10: Comparación de la cantidad de estimaciones y la media de acierto con respecto al umbral de clasificación para los dos grupos del modelo de *random forest*.

4.7.1 Observaciones problemáticas

Para la catalogación de muestras problemáticas solamente se tendrán en cuenta aquellas muestras que cuentan con dos o más hojas con catalogación *a priori* equivocada.

1. Los modelados como limones sin ser limones *a priori*:

- 2 *Citrus × limonimedica* Lush
- 1 *Citrus × latifolia* (Yu. Tanaka) Tanaka
- 1 *Citrus medica*

2. Los modelados como naranjas y mandarinas sin ser naranjas y mandarinas *a priori*:

- 3 *Citrus australasica* F. Muell.
- 1 *Citrus medica*
- 2 *Citrus × lumia*
- 1 *Citrus × bergamia* Risso.

4.8 Creación de modelos con las hojas ancestrales para observar las posibilidades de las hibridadas

En un intento de afrontar el problema con un enfoque distinto se generaron modelos que no se basaban en la filogenia de las especies, sino en los ancestrales que intervienen en la formación de los grupos

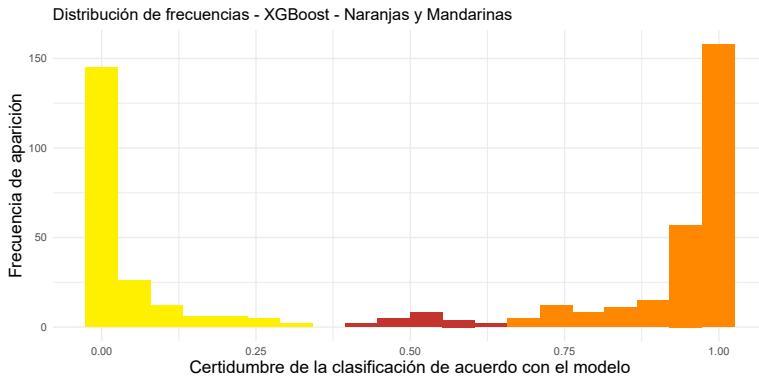


Figura 4.11: Distribuciones de certidumbres del modelo de *XGBoost* para el conjunto de entrenamiento de las naranjas y mandarinas. 52 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.

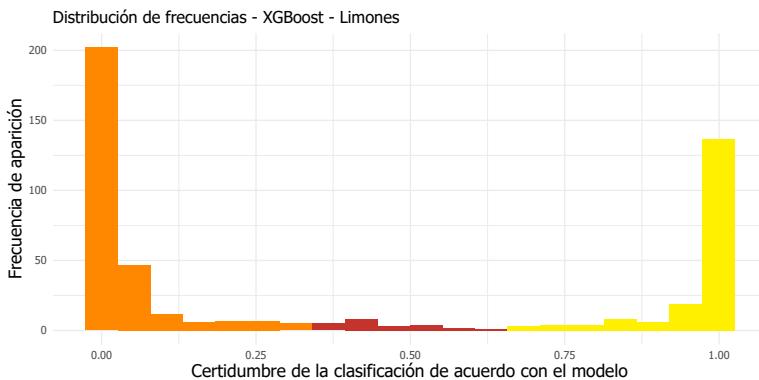


Figura 4.12: Distribuciones de certidumbres del modelo de *XGBoost* para el conjunto de entrenamiento de los limones. 42 % de las hojas superan el umbral del 70 %. Distinguimos tres grupos de baja, media y alta certidumbre de pertenencia al grupo del entrenamiento con distintos colores.

híbridos. Los modelos con los que se trabajó fueron *random forest* y regresión logística y la integración de inferencias se realizó de la misma forma que en el modelo anterior (Tabla 4.4).

Además de los ya mencionados, también se intentó extrapolar estas posibilidades para la creación de nuevas predicciones para los híbridos *C. × aurantium var. aurantium* y *C. × limon var. limon* a través del producto de las probabilidades de pertenencia a los grupos de sus ancestros.

Tabla 4.4: Estadísticos de los modelos basados en las especies ancestrales. Umbral de certidumbre: 0.7.

Grupo	Media de acierto	Representatividad
<i>C. medica</i>	0.96	0.74
<i>C. maxima</i>	0.96	0.99
<i>C. micrantha</i>	0.00	0.00
<i>C. reticulata</i>	0.93	0.96
<i>C. aurantium</i>	0.92	0.70
<i>C. limon</i>	0.87	0.67

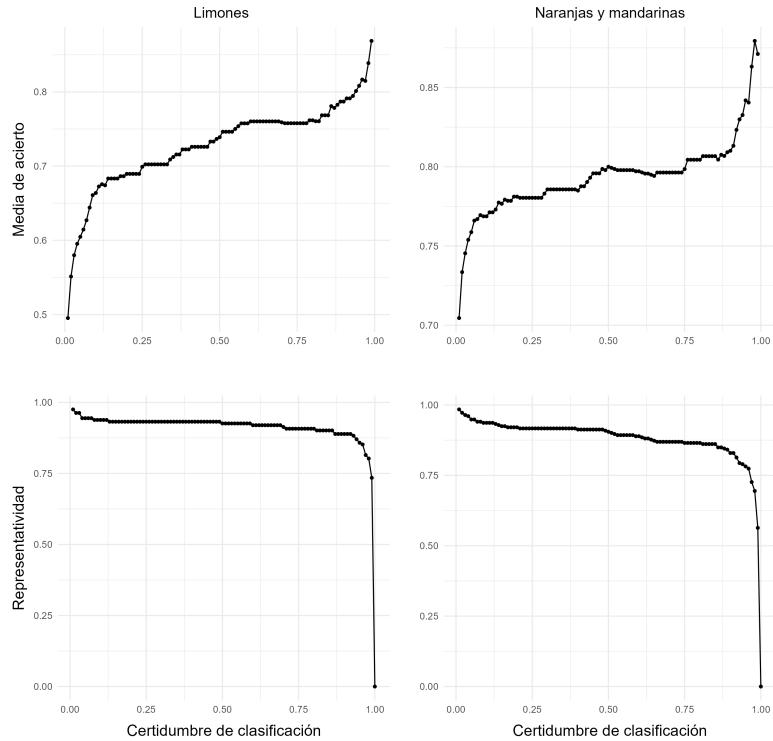


Figura 4.13: Comparación de la cantidad de estimaciones y la media de acierto con respecto al umbral de clasificación para los dos grupos del modelo de *XGBoost*.

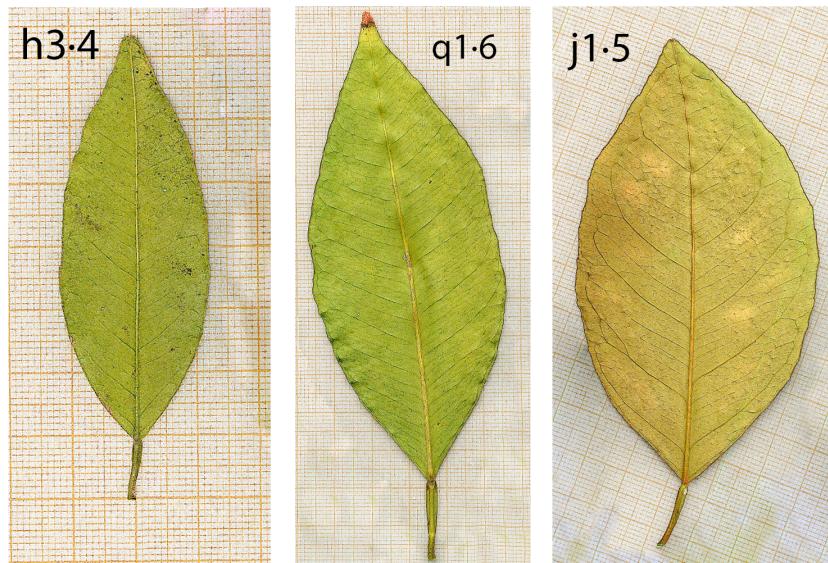
4.9 Estudio de hojas fósiles

Por último también se clasificaron las hojas fósiles encontradas en el artículo de Fischer y Butzmann (1998) y Xie et al. (2013), denominadas *Citrus meletensis*, *Hesperidophyllum senogallienense* A.Massal y *Citrus linczangensis*. respectivamente. Algunas de las variables fueron estimadas para poder realizar la clasificación correctamente.

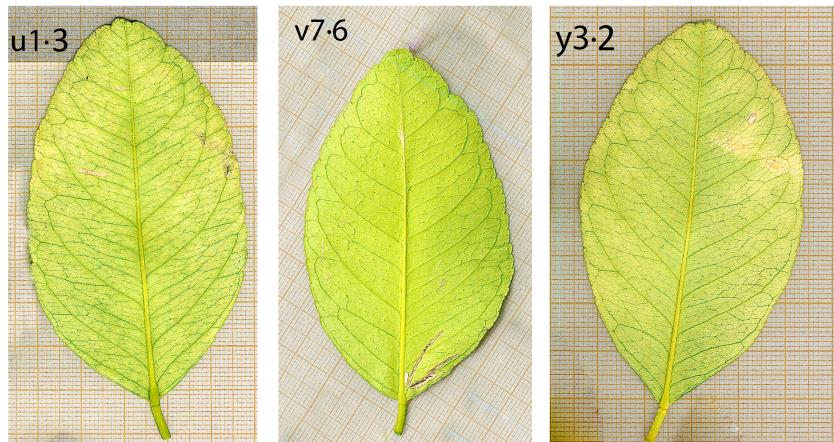
Debido a las fechas a las que pertenecían estas hojas (entre 2.6 y 23 millones de años) se considera que son anteriores a las hibridaciones actuales y por tanto se usó el modelo basado en las ancestrales para la clasificación (Tabla 4.5).

Tabla 4.5: Certidumbres de pertenencia a los grupos ancestrales para las hojas fósiles.

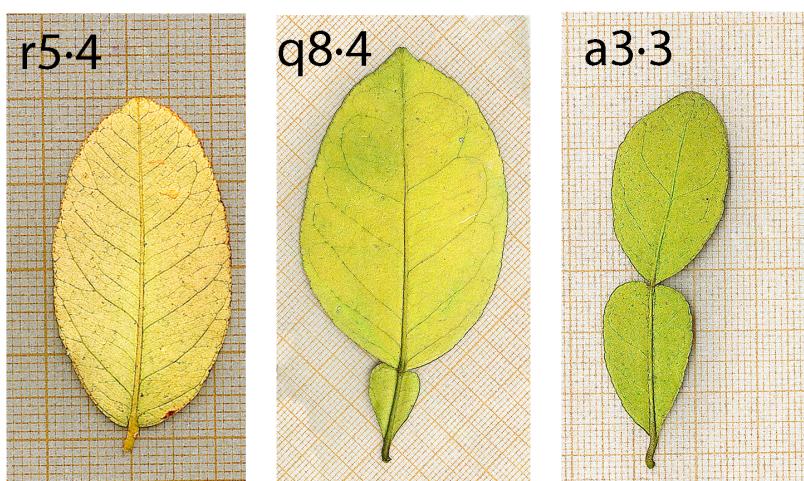
Fósiles	<i>C. medica</i>	<i>C. maxima</i>	<i>C. micrantha</i>	<i>C. reticulata</i>
Fósil 1 (Fischer y Butzmann 1998)	0.56	0.24	0.63	0.27
Fósil 2 (Fischer y Butzmann 1998)	0.06	0.61	0.13	0.70
Fósil 3 (Xie et al. 2013)	0.10	0.40	0.11	0.03



(a) Identificado como naranjas y mandarinas. La hoja h3·4 corresponde a *C. × aurantium var. sinensis* "Washington Navel" y las hojas q1·6 y j1·5 corresponden a *C. × aurantium var. sinensis* "Tarocco".



(b) Identificado como limones. La hoja u1·3 corresponde a *C. × limon var. limon* "lunario", la hoja v7·6 a *C. × limon var. limon* "Portogheso" y la hoja y3·2 a *C. × limon var. limon* "Quatre Saisons".



(c) No identificado como ninguno de los grupos. La hoja r5·4 corresponde a *C. × lumia* "Limone pera", la hoja q8·4 a *C. obovoidea* Takahashi. "Pompelmo Kinkoji" y la hoja a3·3 a *C. micrantha* "lima kaffir".

Figura 4.14: Hojas identificadas por la integración de los modelos como parte de alguno de los grupos. La certidumbre para Figura 4.14a y Figura 4.14b es mayor al 95 % mientras que para Figura 4.14c es menor al 5 %

5. Discusión

5.1 Comprendiendo los resultados

Como se ha observado, la variabilidad que hay en las muestras es muy grande, abarcando tamaños desde los milímetros hasta las decenas de centímetros, pese a esto y gracias a la capacidad de los modelos utilizados, se ha conseguido clasificar con una confianza mayor al 70 % la mayoría de las muestras observadas.

En las Figuras 4.7, 4.10 y 4.13 de forma general, para la media de aciertos se observa que esta aumenta rápidamente al inicio, debido a que las hojas con muy bajas certidumbres no pertenecen al subconjunto y eliminarlas supone un aumento de los aciertos. Aproximadamente para valores mayores de 25 % de certidumbre, el aumento decelera hasta el 75 % donde en los dos primeros modelos, se observan comportamientos extraños y en el tercero sube hasta alcanzar el máximo.

Este comportamiento corresponde con que, a medida que se incrementa la exigencia, la proporción de hojas que pertenecen a la clasificación aumenta, mientras que la parte no perteneciente disminuye, repercutiendo cada vez menos la eliminación de estas en la mejora de la tasa de acierto. No obstante, en el caso de *XGBoost* la curva es algo distinta, al comienzo, las medidas con muy baja certidumbre se pierden rápidamente al aumentar la exigencia, seguidamente se llega a un periodo de estabilidad y por último la tasa de acierto se dispara, debido a que hay un grupo de observaciones con certidumbres muy altas de pertenencia y que *a priori* pertenecen al grupo.

Para la representatividad en los tres modelos ocurre lo mismo, al inicio disminuye muy poco, debido a que, con certidumbre baja, la mayoría de las hojas que se eliminan no pertenecen a la clasificación y conforme se acerca a la certidumbre absoluta disminuye en mayor medida, ya que estas hojas sí que pertenecen al grupo.

5.1.1 Diferencias entre los modelos

En el primer modelo ambos conjuntos de entrenamiento han conseguido resultados similares, una capacidad media alta de predicción con estadísticos moderadamente buenos. Estos reflejan la existencia de patrones claros en las muestras que diferencian los grupos. No obstante, como se comentó, este modelo puede contar con el problema del sobreajuste. Esto ocurre cuando un modelo se ajusta muy definidamente a los datos de entrenamiento, perdiendo capacidad de generalización. Pero lo que se observa aquí dista de ser una caracterización perfecta (Figura 4.7) e incluso algunas de las muestras han sido catalogadas como el grupo opuesto, lo cual no quita para que, dada la limitación de las muestras, el ajuste no sea completamente extrapolable a otros conjuntos completamente distintos a los estudiados.

Random forest, consigue superar por poco la tasa de acierto del modelo de regresión logística, siendo la diferencia media entre las predicciones de los dos modelos de menos de 0.1 %. Además, gracias a el

proceso previo de aleatorización, se podría decir que la representatividad de los resultados puede llegar a ser mayor. En nuestro caso ocurre así, ya que destaca como descriptor más valioso la misma variable en ambos subconjuntos de entrenamiento, lo cual es un indicativo de que el modelo ha encontrado un patrón determinante en los datos. A cambio, por este mismo proceso, es posible que un pequeño porcentaje de los resultados se aleje, incluso de manera notable de los valores reales.

El último de los tres modelos utilizados resulta muy interesante. En la Figura 4.13 se observa que a diferencia con los otros modelos, las certidumbres asignadas se encuentran concentradas principalmente en los extremos aunque sin cambiar proporción de aciertos. Por esto parece que el modelo tiene una naturaleza polarizante, que acierta en la mayoría de los casos pero que se equivoca clasificando como seguros observaciones que no pertenecían inicialmente al grupo. Con todo, consigue representatividades no observadas en el resto. La unión de los modelos, no obstante, es la representación más certera, habiendo mejorado los estadísticos de forma general, además de contar con las distintas perspectivas ofrecidas por los tres modelos metodológicos utilizados.

5.1.2 Diferencias entre los conjuntos de entrenamiento

Los conjuntos de entrenamiento también reflejan diferencias en los resultados arrojados, siendo más bajos los valores de representatividad de los limones con respecto a los de las naranjas y mandarinas para el umbral de exigencia establecido. Esto se relaciona con la mayor incertidumbre asociada a la clasificación, el motivo de este comportamiento puede estar relacionado con la menor cantidad de observaciones de los limoneros con respecto al otro grupo.

Los dos sets de entrenamiento, con carácter general se han diferenciado principalmente en función de algunas de las variables secundarias, siendo la principal el área de la lámina entre la longitud del pecíolo (*a1D1p*) que tiene en cuenta todas las medidas tomadas sobre la lámina y la compara con la longitud del pecíolo. Por su parte, las variables primarias pertenecientes al pecíolo (longitud y anchura mínima) también han tenido una importancia significativa en el caso del limón debido a su peculiar naturaleza corta y ala completamente ausente. El resto de las variables primarias, como se observa en el conjunto de datos adjuntado en el anexo, no han contado con un papel representativo en la clasificación, siendo las relaciones entre las variables (variables secundarias) por tanto, los parámetros más descriptivos encontrados. Queda, sin embargo para futuros análisis, la consideración para el estudio de subconjuntos más detallados.

5.1.3 Observaciones problemáticas

Un número elevado de observaciones no han sido correctamente catalogadas *a priori*, no obstante esta clasificación mejora de forma general la taxonomía original de la que partimos inicialmente. En el caso de los limones, de los 4 individuos mal clasificados *a priori*, 3 pertenecían a hibridaciones con limones (*C. latifolia* y *C. limonimedica*) y el restante correspondía con uno de sus parentales (*C. medica*). En el caso de las naranjas y mandarinas, de los 7 individuos erradamente clasificados según la hipótesis inicial,

cuatro correspondían con especies sin relación de parentesco con estos (*C. australasica* y *C. medica*) y el resto contaban con uno de sus parentales en la ecuación (*C. lumbia* y *C. bergamia*)

En definitiva, para los limones encontramos resultados coherentes de clasificación, con una tendencia a clasificar como tal sus hibridaciones y algunos de sus parentales, esto puede guardar relación con que los primeros cruzamientos son los que resultan más determinantes y por tanto, pese a que los híbridos continúen cruzándose, seguirán manteniendo algo relacionado con la morfología que los primeros parentales determinaron. Para el grupo de las naranjas y mandarinas por otro lado, los resultados son más heterogéneos y sin relación aparente con el grupo.

De las muestras que han entrado en esta categoría, las hojas que no han seguido esta clasificación mayoritariamente se encontraban indeterminadas, con certidumbres bajas de pertenencia a todos los grupos, aunque un grupo más pequeño sí que se clasificaba como limones. Estos resultados deben de revisarse en futuras investigaciones ya que si antes encontrábamos que el número de ejemplares muestreados para el limón eran insuficientes para conformar un grupo de tamaño suficiente, para el segundo grupo definido, posiblemente la heterogeneidad intragrupal también haya sido un factor limitante, como se puede inferir por lo encontrado en este apartado.

5.1.4 Inferencia desde los ancestrales

Para abordar el problema desde otro punto de vista se intentó realizar una aproximación distinta, observada en la Tabla 4.4. Los resultados fueron sorprendentes: Tanto para las variables de pertenencia a *C. maxima* como a *C. reticulata*, la media de acierto y la representatividad superan a las vistas en el modelo general utilizado. Incluso para *C. medica* los valores de representatividad pese a ser menores que los del modelo general, resultan muy cercanos. No obstante, debido a la baja cantidad de muestras pertenecientes a *C. micrantha* no se encontraron valores relevantes para esta especie.

Se intentó extraer las certidumbres generadas por este modelo para los grupos naranja y mandarina (*C. reticulata* × *C. maxima*) y limón (*C. medica* × *C. reticulata* × *C. maxima*). Se encontraron representatividades mayores al 60 % y pese a ser significativas estas resultan menores que las observadas con el principal modelo desarrollado. Es por esto que se considera que futuras investigaciones con este enfoque pueden conseguir interesantes resultados.

Precisamente un enfoque parecido fue el que empleó Traband et al. (2023), este encontró que efectivamente, las formas híbridas contaban con una morfología aproximadamente intermedia entre las especies progenitoras que habían participado en el cruce, llegando a la conclusión de que no solamente podríamos diferenciar las distintas especies actuales sino que también las especies antecesoras subyacentes al cruce.

5.1.5 Catalogación de hojas fósiles

Según los resultados obtenidos Tabla 4.5 y teniendo de nuevo en cuenta que las inferencias realizadas para *C. micrantha* se basan en muy pocos datos, para *C. linczangensis*, la única especie medianamente

cercana es *C. maxima*, llegando una deducción similar a la de Xie et al. (2013). Esto se deba probablemente por la distinguible forma del peciolo de este ancestral, no obstante algunos caracteres son similares a otras especies distintas y por tanto se asigna como una especie distinta propia.

El fósil *C. meletensis* por otro lado se encuentra medianamente cercano a *C. medica* mientras que *H. senogallienne* se encuentra más cerca de *C. maxima* y *C. reticulata*, aunque está catalogado como un género distinto. La conclusión obtenida es que pese a el parecido entre *C. meletensis* y *H. senogallienne*, y el parecido de ambos con el aspecto de *C. aurantium var. aurantium* mencionado en Fischer y Butzmann (1998), los resultados obtenidos no son concluyentes con esta idea. Situando a la hoja fósil 2 medianamente cercana a los parentales de la naranja pero dejando la hoja fósil 1 alejada de los mismos. Es por esto que no se puede asegurar la catalogación de *C. meletensis* como perteneciente a *Citrus*.

5.2 Contexto del trabajo

Otros trabajos enfocados en la morfología de las hojas de los cítricos como clasificador a través de algún tipo de modelamiento se encuentran mayormente asociados al ámbito de la agricultura, ya sea para medir enfermedades o mejorar de alguna manera la producción y almacenaje de los frutos (Y. Chen et al. 2021; Dou et al. 2023; Luaibi, Salman, y Miry 2021).

Por otro lado Ballve, Medina-Filho, y Bordignon (1997), consideran un enfoque parecido al presente, en este se utilizaron mediciones del ala del peciolo para la clasificación de individuos pertenecientes a *Citrus × limonia* Osbeck × *C. aurantium*, *Citrus sunki* hort. ex. Tanaka × *C. aurantium*, y *C. sunki* × *C. sinensis*. Encontraron que efectivamente, estas medidas podían separar estos grupos el 90 % de las ocasiones, siendo el 10 % restante áreas de solapamiento. Esta idea es similar a la que se encontró en esta investigación, donde se pudo separar los grupos propuestos para cítricos pero siempre con cierto grado de indeterminación debido principalmente a la variabilidad de la muestra y de las especies.

Otros trabajos más modernos ya mencionados como los de Oza, Desai, y Raole (2021) o Traband et al. (2023), han estudiado la morfología de las hojas a través de variables continuas, que tienen en cuenta la totalidad de la hoja. Con estas medidas se formaron distintos grupos según patrones y se encontraron resultados similares a los aquí descritos. En última instancia, llegando a la conclusión de que la morfología de las hojas es un indicador suficientemente informativo como para crear una clasificación en este género.

5.3 Limitaciones y futuras investigaciones

A la hora de realizar el trabajo hemos encontrado varias limitaciones, la primera de ellas es el bajo número de observaciones con el que se ha trabajado, ya que los modelos predictivos utilizados se caracterizan por necesitar grandes cantidades de datos (Luan et al. 2020). No obstante aumentar el

conjunto de datos se hizo imposible debido al marco muy limitado de tiempo y recursos que supone un TFG, pero es algo que podría llevarse a cabo en ulteriores investigaciones.

Por otro lado cabe preguntarse la relevancia del jardín utilizado como espacio muestral, ya que aunque cuenta con un gran número de variedades, para ciertos taxones como puede ser el grupo de los limones o los grupos parentales, se encuentran algo limitados, mientras que para otros grupos como las naranjas y mandarinas, contábamos con un grupo mucho más amplio y por tanto partimos con diferencias en la representación de los grupos, pese a esto se han encontrado claras diferencias entre ambos grupos con el procedimiento desarrollado.

También, el enfoque metodológico de presencia-ausencia, puede haber resultado limitante debido a las diferencias intragrupales que encontramos en conjunto mencionado de las naranjas y mandarinas, donde la variabilidad comprende desde porcentajes muy altos de *C. maxima* hasta porcentajes mayoritarios de *C. reticulata*. Para futuras investigaciones este problema debe de ser abordado, con clasificaciones filogenéticas más detalladas. Además de la integración usando la lógica bayesiana de los parámetros genéticos con otro tipo de variables morfológicas observables como las medidas en el trabajo presente.

Por último se debería de incluir el posible efecto de otro tipo de variables en el análisis como por ejemplo, de tipo ambiental, además de realizar estudios con nuevas herramientas de análisis para luego integrarlos en el modelo aquí propuesto y seguir actualizando nuestra hipótesis con respecto a la taxonomía de este grupo.

6. Conclusiones

Los resultados obtenidos implican la existencia de suficiente variabilidad morfológica foliar asociada a los distintos taxones como para diferenciar distintos grupos de cítricos.

La importancia del peciolo con respecto a la lámina en cuanto a la clasificación, siendo esta parte al parecer más informativa sobre la filogenia de la planta, al menos a los niveles estudiados.

La relevancia de la herencia de los caracteres parentales en los híbridos. Estos, como se ha estudiado, cuentan también con más que suficiente capacidad descriptiva para establecer clasificaciones del género. Además, se puede inferir que son precisamente los primeros cruzamientos entre parentales los que actúan como un nexo a partir del cual, las sucesivas hibridaciones crean nuevas variedades en torno al tipo central establecido, generando diferencias pero sin alejarse sensiblemente de este.

Por último, implica, la idea de que la integración bayesiana de modelos para la clasificación aumenta la capacidad descriptiva de los mismos. Además de permitir la integración de nuevas evidencias de la misma manera que se ha realizado en este trabajo, tomando los resultados aquí obtenidos como hipótesis y combinándolos con nuevos datos para así actualizar nuestra clasificación, representando de una forma más completa la filogenia de los cítricos. Aunque siempre lidiando con algún nivel de incertidumbre que se encuentra dentro de la naturaleza propia de las observaciones.

Relación del trabajo con los Objetivos de Desarrollo Sostenible (ODS)

El presente trabajo se enmarca en varios de los ODS por los que se rige la ONU. En primer lugar, el ODS 15, que busca entre otros objetivos detener la pérdida de biodiversidad, por aumentar el cuerpo de conocimiento y darles valor adicional a las especies de *Citrus*. No solamente a las que se encuentran muy presentes en nuestra vida actual, si no variedades que se encuentran relegadas geográficamente y que no son tan conocidas como el resto, aunque sí tienen valor intrínseco y de uso futuro. Ya que siendo el género *Citrus* tan diverso y contando con tantos hechos de hibridación, puede que en un futuro estas especies relegadas den lugar a híbridos dignos de estudio, explotación o admiración.

En el marco del ODS número 4, este trabajo, que busca descubrir nuevas formas metodológicas de enfrentar problemas clásicos como los de clasificación de especies, puede ser de gran utilidad para la educación y la formación de las nuevas generaciones de científicos. Abriendo las puertas a nuevas y no tan populares herramientas en biología que pueden facilitar y posibilitar la investigación en este campo.

7. Referencias¹

- Ballve, Rosa M. L., Herculano Penna Medina-Filho, y Rita Bordignon. 1997. «Identification of reciprocal hybrids in citrus by the broadness of the leaf petiole wing». *Brazilian Journal of Genetics* 20 (4): 697-702. <https://doi.org/10.1590/s0100-84551997000400023>.
- Barbera, Giuseppe. 2023. *Una storia del mondo*. Editado por Il Saggiatore.
- Biau, Gérard, y Erwan Scornet. 2016. «A random forest guided tour». *TEST* 25 (2): 197-227. <https://doi.org/10.1007/s11749-016-0481-7>.
- Breiman, Leo, Adele Cutler, Andy Liaw, y Matthew Wiener. 2024. «randomForest: Breiman and Cutler's Random Forests for Classification and Regression». <https://CRAN.R-project.org/package=randomForest>.
- Cascales, Francisco. 1873. «Discursos históricos de la muy noble y muy leal ciudad de Murcia y su reino por el licenciado Francisco Cascales.»
- Chen, Tianqi. 2024. «xgboost: Extreme Gradient Boosting». <https://CRAN.R-project.org/package=xgboost>.
- Chen, Tianqi, y Carlos Guestrin. 2016. «XGBoost: A Scalable Tree Boosting System». En *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. ACM. <https://doi.org/10.1145/2939672.2939785>.
- Chen, Yaohui, Xiaosong An, Shumin Gao, Shanjun Li, y Hanwen Kang. 2021. «A Deep Learning-Based Vision System Combining Detection and Tracking for Fast On-Line Citrus Sorting». *Frontiers in Plant Science* 12 (febrero). <https://doi.org/10.3389/fpls.2021.622062>.
- Curk, Franck, Frédérique Ollitrault, Andres Garcia-Lor, François Luro, Luis Navarro, y Patrick Ollitrault. 2016. «Phylogenetic origin of limes and lemons revealed by cytoplasmic and nuclear markers». *Annals of Botany* 117 (4): 565-83. <https://doi.org/10.1093/aob/mcw005>.
- Deng, Xiuxin, Xiaoming Yang, Masashi Yamamoto, y Manosh K. Biswas. 2020. «Domestication and history». En *The Genus Citrus*, 33-55. Elsevier. <https://doi.org/10.1016/b978-0-12-812163-4.00003-6>.
- Dou, Shiqing, Lin Wang, Donglin Fan, Linlin Miao, Jichi Yan, y Hongchang He. 2023. «Classification of Citrus Huanglongbing Degree Based on CBAM-MobileNetV2 and Transfer Learning». *Sensors* 23 (12): 5587. <https://doi.org/10.3390/s23125587>.
- FAO. 2021. «Citrus Fruit Statistical Compendium 2020. Rome».
- Fischer, Thilo C., y Rainer Butzmann. 1998. «Citrus meletensis (Rutaceae), a new species from the Pliocene of Valdarno (Italy)». *Plant Systematics and Evolution* 210 (1-2): 51-55. <https://doi.org/10.1007/bf00984727>.
- Gaillard, Pierre, Yannig Goude, Laurent Plagne, Thibaut Dubois, y Benoit Thieurnel. 2024. «opera: Online Prediction by Expert Aggregation». <https://CRAN.R-project.org/package=opera>.
- Halliwell, Barry, y John M. C. Gutteridge. 2015. *Free Radicals in Biology and Medicine*. Oxford

¹Se ha utilizado el estilo de Chicago 17th Edition para las referencias bibliográficas.

- University Press. <https://doi.org/10.1093/acprof:oso/9780198717478.001.0001>.
- Hamner, Ben, Michael Frasco, y Erin LeDell. 2024. «Metrics: Evaluation Metrics for Machine Learning». <https://CRAN.R-project.org/package=Metrics>.
- IPGRI. 1999. *Descriptors for Citrus*. International Plant Genetic Resources Institute, Rome, Italy.
- Isaac, Erich. 1959. «Influence of Religion on the Spread of Citrus: The religious practices of the Jews helped effect the introduction of citrus to Mediterranean lands». *Science* 129 (3343): 179-86. <https://doi.org/10.1126/science.129.3343.179>.
- Langgut, Dafna. 2017. «The history of Citrus medica (citron) in the Near East: Botanical remains and ancient art and texts». *Agrumed. Archaeology and history of citrus fruit in the Mediterranean: Acclimatization, diversifications, uses.*, 84-94.
- LaValley, Michael P. 2008. «Logistic Regression». *Circulation* 117 (18): 2395-99. <https://doi.org/10.1161/circulationaha.106.682658>.
- Liu, YuQiu, Emily Heying, y Sherry A. Tanumihardjo. 2012. «History, Global Distribution, and Nutritional Importance of Citrus Fruits». *Comprehensive Reviews in Food Science and Food Safety* 11 (6): 530-45. <https://doi.org/10.1111/j.1541-4337.2012.00201.x>.
- Luaibi, Ahmed R., Tariq M. Salman, y Abbas H. Miry. 2021. «Detection of citrus leaf diseases using a deep learning technique». *International Journal of Electrical and Computer Engineering (IJECE)* 11 (2): 1719. <https://doi.org/10.11591/ijece.v11i2.pp1719-1727>.
- Luan, Jing, Chongliang Zhang, Binduo Xu, Ying Xue, y Yiping Ren. 2020. «The predictive performances of random forest models with limited sample size and different species traits». *Fisheries Research* 227 (julio): 105534. <https://doi.org/10.1016/j.fishres.2020.105534>.
- Luro, Francois, Franck Curk, Yann Froelicher, y Patrick Ollitrault. 2017. «Recent insights on Citrus diversity and phylogeny». *AGRUMED: Archaeology and history of citrus fruit in the mediterranean. Naples: Publications du Centre Jean Bérard*. <https://doi.org/10.4000/books.pcjb.2169>.
- Ma, Gang, Lancui Zhang, Minoru Sugiura, y Masaya Kato. 2020. «Citrus and health». En *The Genus Citrus*, 495-511. Elsevier. <https://doi.org/10.1016/b978-0-12-812163-4.00024-3>.
- Mastrangelo, Domenico, Elvira Pelosi, Germana Castelli, Francesco Lo-Coco, y Ugo Testa. 2018. «Mechanisms of anti-cancer effects of ascorbate: Cytotoxic activity and epigenetic modulation». *Blood Cells, Molecules, and Diseases* 69 (marzo): 57-64. <https://doi.org/10.1016/j.bcmd.2017.09.005>.
- Morelli, Marco B., Jessica Gambardella, Vanessa Castellanos, Valentina Trimarco, y Gaetano Santulli. 2020. «Vitamin C and Cardiovascular Disease: An Update». *Antioxidants* 9 (12): 1227. <https://doi.org/10.3390/antiox9121227>.
- Munankarmi, Nabin N., Neesha Rana, Bal K. Joshi, Tribikram Bhattacharai, Sujan Chaudhary, Bikash Baral, y Sangita Shrestha. 2023. «Characterization of the genetic diversity of Citrus species of Nepal using simple sequence repeat (SSR) markers». *South African Journal of Botany* 156: 192-201. <https://doi.org/10.1016/j.sajb.2023.03.014>.
- Ollitrault, Patrick, Franck Curk, y Robert Krueger. 2020. «Citrus taxonomy». En *The Genus Citrus*,

- 57-81. Elsevier. <https://doi.org/10.1016/b978-0-12-812163-4.00004-8>.
- Oza, Kavi K., Rinku J. Desai, y Vinay M. Raole. 2021. «Digital Morphometrics: A Tool for Leaf Morpho-Taxonomical Studies». *Indian Journal of Advanced Botany* 1 (2): 1-7. <https://doi.org/10.54105/ijab.b2001.101221>.
- Pereira-Gonzzatto, Mateus, y Júlia Scherer-Santos. 2023. «Introductory Chapter: World Citrus Production and Research». En *Citrus Research - Horticultural and Human Health Aspects*. IntechOpen. <https://doi.org/10.5772/intechopen.110519>.
- R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.
- Ramón-Laca, Luis. 2003. «The Introduction of Cultivated Citrus to Europe via Northern Africa and the Iberian Peninsula». *Economic Botany* 57 (4): 502-14. [https://doi.org/10.1663/0013-0001\(2003\)057%5B0502:tiocct%5D2.0.co;2](https://doi.org/10.1663/0013-0001(2003)057%5B0502:tiocct%5D2.0.co;2).
- Rivera, Diego, Antonio Bermúdez, Concepción Obón, Francisco Alcaraz, Segundo Ríos, Jorge Sánchez-Balibrea, Pedro P. Ferrer-Gallego, y Robert Krueger. 2022. «Analysis of “Marrakesh limetta” (*Citrus × limon* var. *limetta* (Risso) Ollitrault, Curk & R.Krueger) horticultural history and relationships with limes and lemons». *Scientia Horticulturae* 293 (febrero): 110688. <https://doi.org/10.1016/j.scienta.2021.110688>.
- Shi, Wenbo, Weicai Song, Jin Liu, Chao Shi, y Shuo Wang. 2023. «Comparative chloroplast genome analysis of Citrus (Rutaceae) species: Insights into genomic characterization, phylogenetic relationships, and discrimination of subgenera». *Scientia Horticulturae* 313 (abril): 111909. <https://doi.org/10.1016/j.scienta.2023.111909>.
- Spreen, Thomas H., Zhifeng Gao, Waldir Fernandes, y Marisa L. Zansler. 2020. «Global economics and marketing of citrus products». En *The Genus Citrus*, 471-93. Elsevier. <https://doi.org/10.1016/b978-0-12-812163-4.00023-1>.
- Swingle, Walter T., y Philip C. Reece. 1967. «History, world distribution, botany, and varieties». En *Reuther, W., Webber, H.J., Batchelor, L.D. (Eds.), The Citrus Industry. Revised second ed*, 1:190-430.
- Talon, Manuel, Guohong A. Wu, Frederick G. Gmitter, y Daniel S. Rokhsar. 2020. «The origin of citrus». En *The Genus Citrus*, 9-31. Elsevier. <https://doi.org/10.1016/b978-0-12-812163-4.00002-4>.
- Tanaka, Tomio, y Masami Furuta. 1977. *Journal of the Japan Society of Engineering Geology* 18 (1/2): 1-12. <https://doi.org/10.5110/jjseg.18.1>.
- Tolkowsky, Shmuel. 1938. *Hesperides: A History of the Culture and Use of Citrus Fruits*. J. Bale, sons & Curnow, Limited.
- Traband, Ryan C., Xuesong Wang, Jill Lui, Lei Yu, Yoko Hiraoka, Ira A. Herniter, Christian Bowman, et al. 2023. «Exploring the Phylogenetic Relationship among Citrus through Leaf Shape Traits: A Morphological Study on Citrus Leaves». *Horticulturae* 9 (7): 793. <https://doi.org/10.3390/horticulturae9070793>.

- Wester, Peter J. 1915. *Citrus fruits in the Philippines*. Editado por The Philippine agricultural review. Vol. 8. 1. Bureau of Agriculture.
- Xie, Sanping, Steven R. Manchester, Kenan Liu, Yunfeng Wang, y Bainian Sun. 2013. «Citrus linczangensisssp. n., a Leaf Fossil of Rutaceae from the Late Miocene of Yunnan, China». *International Journal of Plant Sciences* 174 (8): 1201-7. <https://doi.org/10.1086/671796>.

Anexo A

1 Código ejecutado y tablas de datos

Las tablas de datos, al igual que el código se incluyen en el repositorio de github del proyecto: <https://github.com/pabloLopezRuiz/Trabajo-fin-de-grado>.

2 Esquemas en relación con los modelos

Algunos resultados obtenidos de los diferentes análisis realizados que pueden ser de interés para profundizar en distintos aspectos.

2.1 Regresión logística

En este caso, los resúmenes nos informan tanto del ajuste de la clasificación (“AIC”) como de la importancia de las variables (“Pr”). En primer lugar se observa el conjunto de entrenamiento de Naranjas y mandarinas y en segundo lugar el de limones (Figura A.1).

2.2 Random forest

Para este modelo se presenta un esquema de la importancia de las variables y un árbol de decisión formado (Figuras A.2 y A.3).

2.3 XGBoost

Igual que el modelo anterior, este cuenta con un tipo de esquemas parecidos, no obstante, debido a que se ha entrenado a la vez con ambos subconjuntos de entrenamiento, se presenta un único gráfico (Figura A.4) y un único árbol de decisión (Figura A.5).

```

Call:
glm(formula = retXmax ~ ., family = binomial, data = df_sub_1)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-2.5744 -0.5777 -0.1085  0.6415  2.9125 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) 6.36649   6.18008  1.030  0.302934  
lp          6.09517   2.72787  2.234  0.024546 *  
wp         -6.73992   3.70555 -1.819  0.068931 .  
sp        -16.97226  10.83908 -1.566  0.117386  
ll          1.86623   0.60855  3.067  0.002164 **  
wl         -3.63730   1.09107 -3.334  0.000857 ***  
hl         -0.99652   0.89515 -1.113  0.265608  
al          0.08429   0.07883  1.069  0.284926  
wldll     13.21061   9.39458  1.406  0.159666  
wpdw1    -22.60763  17.06276 -1.325  0.185181  
wpd1p     3.25372   6.27173  0.519  0.603906  
lpd1l     -18.10607  15.95373 -1.135  0.256412  
spdpw     -4.77553   1.87609 -2.545  0.010913 *  
wldhl     1.20243   3.44487  0.349  0.727052  
ald1p     -0.06576   0.09601 -0.685  0.493371  
llmhd1ll -14.79358  9.72645 -1.521  0.128269  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Dispersion parameter for binomial family taken to be 1)

Null deviance: 788.33 on 570 degrees of freedom
Residual deviance: 465.01 on 555 degrees of freedom
AIC: 497.01

Number of Fisher Scoring iterations: 7

Call:
glm(formula = medXretXmax ~ ., family = binomial, data = df_sub_2)

Deviance Residuals:
    Min      1Q   Median      3Q     Max 
-2.4801 -0.4454 -0.2276  0.3640  3.6022 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -8.328842  7.890069 -1.056  0.291146  
lp          -7.835456  3.195425 -2.452  0.014203 *  
wp         8.122220  3.392715  2.394  0.016665 *  
sp        39.512933 11.339640  3.484  0.000493 ***  
ll          0.769755  0.748840  1.028  0.303983  
wl         1.906955  1.103162  1.729  0.083876 .  
hl         -1.090852  1.160824 -0.940  0.347360  
al          -0.211064  0.092157 -2.290  0.022006 *  
wldll     13.726958 10.903032  1.259  0.208029  
wpdw1    9.550013 20.799081  0.459  0.646122  
wpd1p    -17.760492  7.045865 -2.521  0.011712 *  
lpd1l     4.310985  22.769875  0.189  0.849835  
spdpw     1.634133  2.139239  0.764  0.444936  
wldhl     -6.249630  3.857588 -1.620  0.105213  
ald1p     0.004271  0.082050  0.052  0.958486  
llmhd1ll  6.613293 13.792969  0.479  0.631605  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Dispersion parameter for binomial family taken to be 1)

Null deviance: 697.10 on 570 degrees of freedom
Residual deviance: 360.57 on 555 degrees of freedom
AIC: 392.57

Number of Fisher Scoring iterations: 6

```

Figura A.1: Resumen de los modelos de regresión logística para el conjunto de entrenamiento de naranjas y mandarinas (Izquierda) y limones (Derecha). Se observa que el modelo de limones tiene un AIC menor y una variable más importante que el modelo de naranjas y mandarinas.

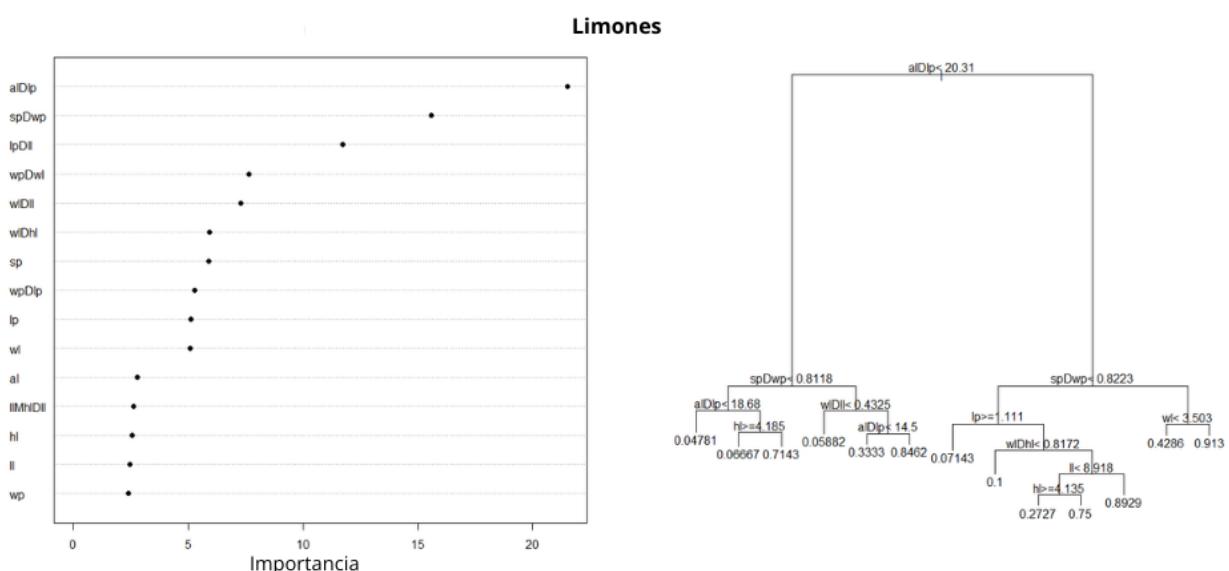


Figura A.2: Gráfico de importancia relativa de las variables que se han tenido en cuenta a la hora de la clasificación con *random forest* con los limones (Izquierda) y Árbol de decisión de ejemplo que se ha seguido en esta clasificación (Derecha). Se observa que la variable más importante y que separa las observaciones en dos grandes grupos es el área de la lámina entre la longitud del pecíolo.

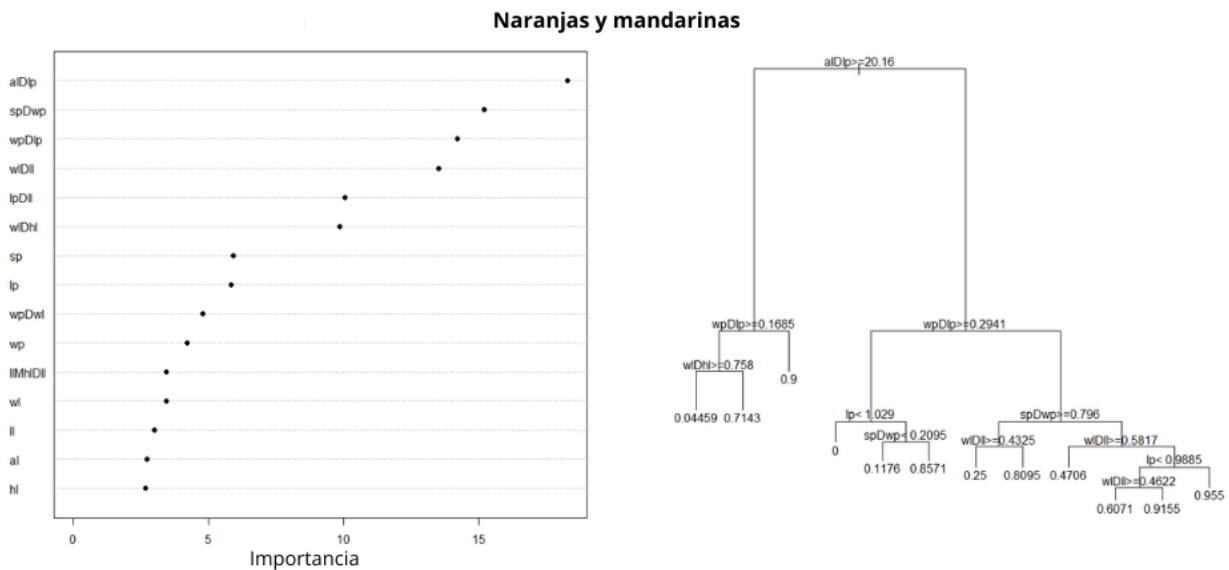


Figura A.3: Gráfico de importancia relativa de las variables que se han tenido en cuenta a la hora de la clasificación con *random forest* con las naranjas y mandarinas (Izquierda) y Árbol de decisión de ejemplo que se ha seguido en esta clasificación (Derecha). Una vez más la variable más importante a la hora de la clasificación vuelve a ser el área de la lámina entre la longitud del peciolo.

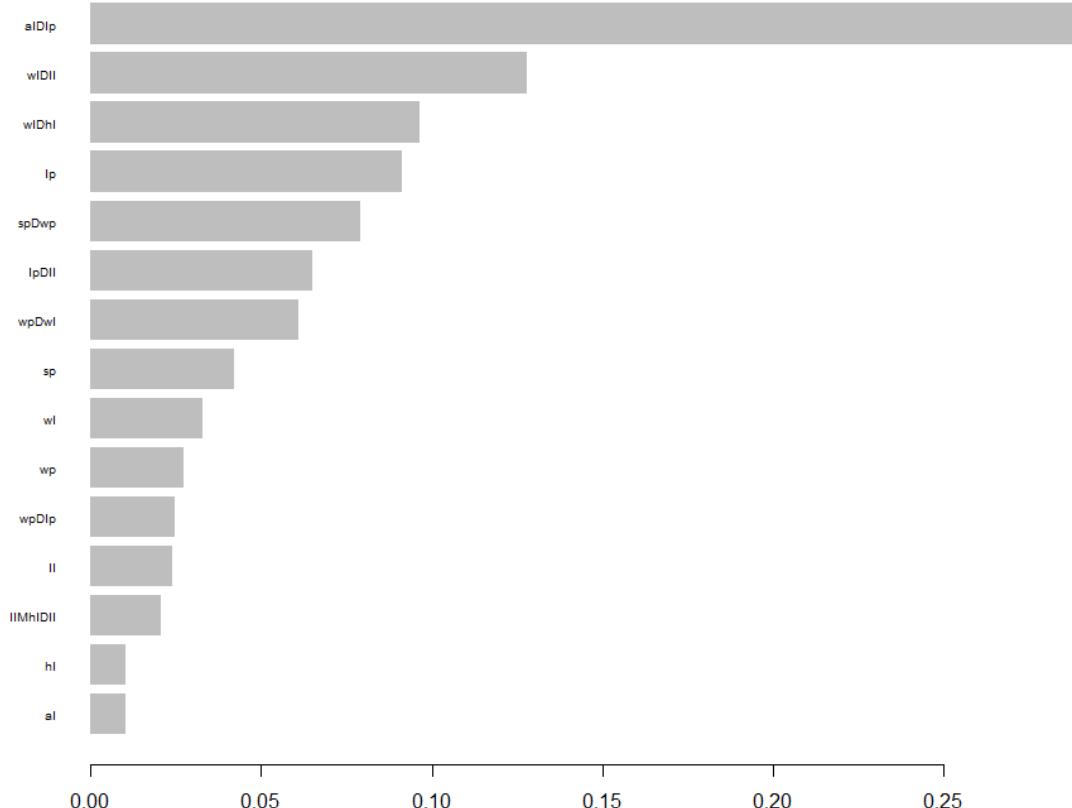


Figura A.4: Importancia de las variables a la hora de la clasificación con el modelo *XGBoost*.

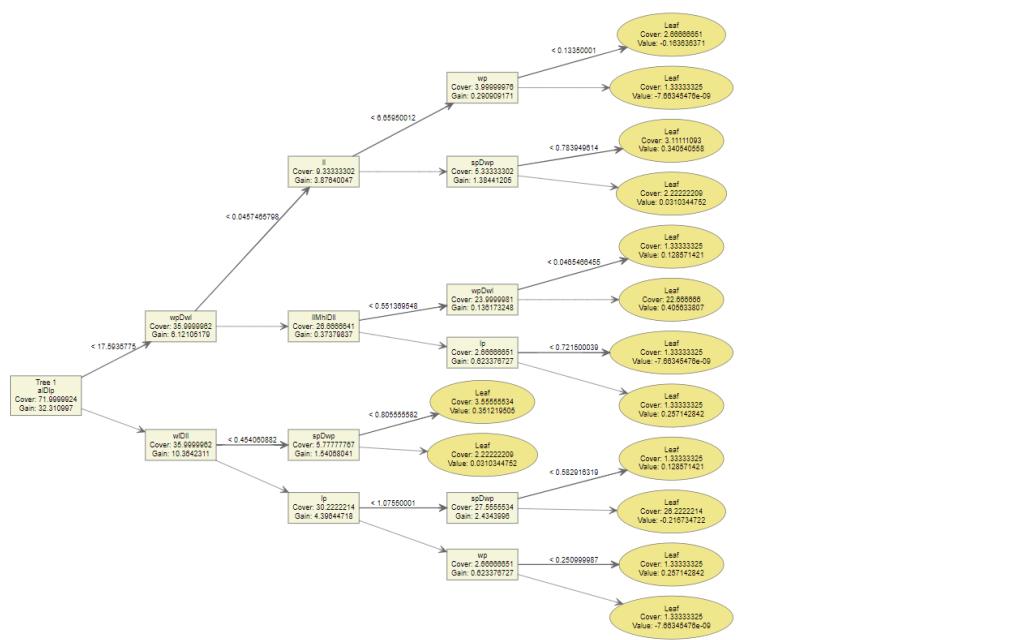


Figura A.5: Árbol de nodos creado en la clasificación con *XGBoost*.

TRABAJOS FIN DE GRADO (TFG) DE LA FACULTAD DE BIOLOGÍA

D. Pablo Manuel López Ruiz, con DNI nº: 24458369P, estudiante del Grado en Biología de la Facultad de Biología de la Universidad de Murcia,

DECLARO:

Que el Trabajo de Fin de Grado que presento para su exposición y defensa titulado: Aplicación de razonamiento bayesiano y machine learning a la taxonomía: Clasificando algunas especies de cítricos (rutáceas)

y cuyo tutor es

D. Diego Rivera Núñez

es original y que todas las fuentes utilizadas para su realización han sido debidamente citadas en el mismo.

Murcia, a 29 de Mayo de 2024.

Firma



Declaración de originalidad del Trabajo Fin de Grado

