Trabajo Fin de Máster - Curso 2022/ 2023

Desarrollo de modelos predictivos para la detección de cáncer de mama

Autor: **Julián Guilló Vigo**

Tutor: ANTONIO LÓPEZ QUÍLEZ

Facultat de Ciències Matemàtiques

Máster en Bioestadística

ÍNDICE DE CONTENIDOS

RESUMEN

El cáncer de mama es una enfermedad grave, cuya supervivencia depende en gran medida de una rápida detección. En este trabajo tratamos de mejorar la precisión en el diagnóstico del cáncer, mediante el ajuste de clasificadores binarios que puedan traducirse en herramientas para la asistencia en el diagnóstico. El objetivo es comparar diferentes tipos de modelos predictivos y tratar de mejorar la capacidad predictiva citada en trabajos previos que utilizan el mismo banco de datos.

Se ajustan múltiples modelos, tanto del ámbito de la estadística tradicional como del ámbito del aprendizaje automático, utilizando métodos de validación cruzada para estimar la exactitud, sensibilidad y especificidad de cada modelo.

La mayoría de los modelos ajustados superaron el umbral del 90% de exactitud, llegando en algunos casos hasta el 98%. Aunque esto tan sólo supone una ligera mejora respecto a los resultados obtenidos en estudios previos, este trabajo ha sido muy enriquecedor desde el punto de vista del aprendizaje y la ampliación de conocimientos en el campo de la estadística aplicada a problemas de clasificación.

Palabras clave: diagnóstico de cáncer de mama, núcleos celulares, modelos estadísticos, análisis de componentes principales, clasificación binaria.

1. INTRODUCCIÓN Y MARCO TEÓRICO

1.1. Cáncer de mama: situación y contexto actuales

A escala mundial, el cáncer se ha convertido en una importante preocupación y un peligro para las personas de todo el mundo, y actualmente es, junto a las enfermedades cardiovasculares, la principal causa de muerte en la mayoría de los países (World Health Organization, 2018). Es una afección que se produce en el organismo como resultado de una proliferación celular aberrante. Los tejidos que componen un órgano están formados por células que funcionan juntas como una unidad. En condiciones normales, las células nuevas sustituyen a las enfermas mediante la replicación de células sanas, conocido como el proceso de proliferación celular, que es universal para todas las células. Los tejidos o bultos de células se forman debido a que estas células se replican de forma incontrolada, lo que lleva a la creación de tumores o cánceres, que son masas o bultos de células que han crecido de forma incontrolada y han superado su entorno original. El cáncer de pulmón, el de hígado, el colorrectal, el de estómago y el de mama son los tipos de cáncer más frecuentes (Winters et al., 2017).

El cáncer de mama es el tipo de tumor más frecuente en mujeres, y una enfermedad muy compleja a nivel molecular. Durante los últimos 10-15 años los tratamientos han evolucionado mucho, intentando abordar la complejidad del

problema con tratamientos dirigidos biológicamente y tratando de rebajar los tratamientos para reducir los efectos adversos de los mismos.

El cáncer de mama temprano, donde el tumor queda contenido dentro de la mama o tan sólo se ha propagado a los nódulos linfáticos axilares, se considera curable. Las mejoras en los últimos tratamientos desarrollados han aumentado las probabilidades de cura en un 80-90% de las pacientes aproximadamente. Por contra, el cáncer metastático avanzado todavía se considera incurable con los métodos de tratamiento de los que disponemos actualmente. Aun así, este tipo de cáncer sí es tratable, siendo el objetivo de la terapia el prolongar la vida de la paciente y controlar los síntomas con terapias de bajo impacto en cuanto a toxicidad (Harbeck et al., 2019).

1.1.1 Epidemiología y demografía

En 2018, alrededor de 2.1 millones de mujeres fueron diagnosticadas con cáncer de mama, aproximadamente un nuevo caso cada 18 segundos. Además de eso, 629.679 mujeres murieron por cáncer de mama (Bray et al., 2018). La incidencia global del cáncer se ha ido incrementando anualmente en un 3.1% entre 1980 y 2010, con 641.000 casos en 1980 y más de 1.6 millones en 2010 (Bray et al., 2015).

La incidencia varía a lo largo el planeta, siendo esta mayor en regiones con altos ingresos en comparación con regiones de ingresos más bajos (92 por cada 100.000 habitantes en Norteamérica, frente a 27 por cada 100.000 habitantes en África Central y Asia del Este) (Harbeck et al., 2019). Estos patrones reflejan tanto los factores de riesgo de la enfermedad como la disponibilidad de la mamografía en cada región (y por tanto el número de cánceres detectados); las mayores incidencias se dan en Norteamérica, Australia, Nueva Zelanda y Europa del Norte y del Oeste. A pesar de la menor incidencia, la mortalidad por cáncer de mama es mayor en la mayoría de los países de renta media y baja, debido al retraso en la presentación de síntomas, el diagnóstico incorrecto o tardío y el acceso limitado al tratamiento. En los países de renta alta el cáncer de mama se diagnostica a menudo en una fase temprana, y el pronóstico suele ser bueno. Sin embargo, en los países de ingresos bajos y medios, el cáncer de mama se suele diagnosticar en una fase más tardía y, por tanto, se asocia a una menor supervivencia, hecho que se refleja en las estadísticas de mortalidad. Además de esto, las pacientes de países en desarrollo diagnosticados con cáncer de mama son aproximadamente 10 años más jóvenes que los diagnosticados en países desarrollados (Winters et al., 2017).

Alrededor del 10% de los casos de cáncer de mama tienen una componente hereditaria y se asocian a un historial familiar, aunque factores como los medioambientales y los culturales, el estilo de vida, y las campañas nacionales de prevención y concienciación tienen un impacto muy superior a la hora de determinar si una persona va a padecer de cáncer de mama o no. Desde 1980 hasta los 2000 se produjo un gran aumento de la incidencia de la enfermedad a nivel global, y la incidencia ha continuado en crecimiento hasta la fecha. Esto se debe probablemente al incremento en la edad materna del primer embarazo, ya que la sensibilidad de las glándulas mamarias a grandes secreciones hormonales se ve alterada con la edad; y a la mayor concienciación mundial y su consecuente aumento de los cribados por mamografía (Colditz et al., 2006).

1.1.2 Diagnóstico y prevención

El diagnóstico del cáncer de mama está basado principalmente en una prueba triple que abarca la examinación clínica mediante palpación, las técnicas de imagen como la mamografía y/o ultrasonografía, y la biopsia por aspiración con aguja. En casos excepcionales donde el diagnóstico es incierto se realiza una biopsia quirúrgica invasiva, aunque este método se ha venido utilizando cada vez menos durante los últimos años, en favor de métodos de aspirado por punción (Street et al., 2000). Una evaluación adecuada ayuda a discriminar con precisión entre quienes padecen de cáncer de mama y quienes tienen condiciones benignas (como un fibroadenoma) o cambios mamarios normales, que pueden ser manejados con seguridad mediante un seguimiento, evitando así la necesidad de una intervención quirúrgica.

En la mayoría de los países desarrollados se ha implementado un cribado por mamografía. Este cribado de la población tiene como objetivo detectar la enfermedad en una fase temprana para la que existe un tratamiento eficaz, mediante una prueba no invasiva y lo suficientemente precisa para realizar un cribado rápido. En conjunto los ensayos controlados aleatorios por mamografía (Figura 1) han demostrado reducir significativamente la mortalidad por cáncer de mama en un riesgo relativo del 20% para las personas invitadas al cribado. La eficacia del cribado mamográfico depende de la edad; es más evidente en las mujeres de 50-69 años, pero el beneficio es menos claro en las mujeres fuera de este rango de edad (Nelson et al., 2016).

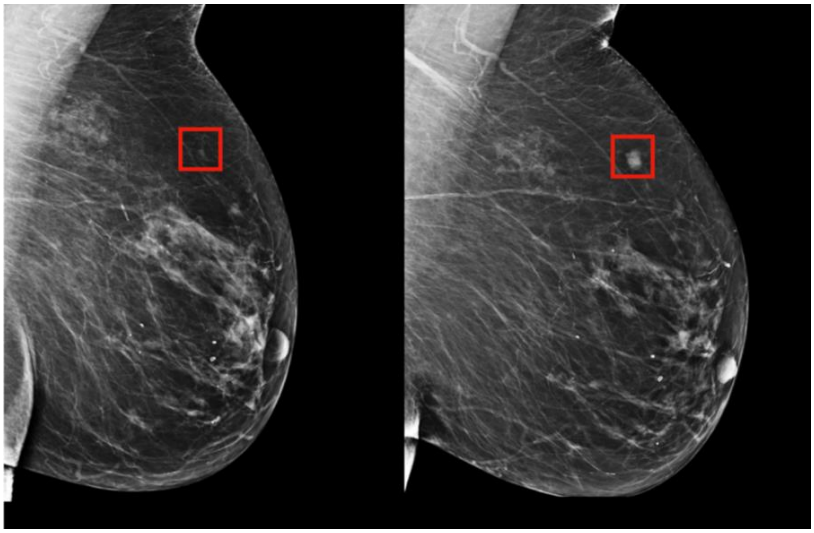


Figura 1. Imágenes laterales de un pecho obtenidas mediante mamografía. Rodeada por un cuadrado rojo se encuentra una masa sospechosa de ser cancerosa (Conner-Simons et al., 2019).

Teniendo en cuenta lo que acabamos de comentar, debemos abordar el diagnóstico desde dos ejes diferentes. El primero es la precisión; cualquier sistema predictivo debe de ser lo suficientemente preciso para poder ser utilizado con confianza en un entorno clínico. El segundo eje es el perjuicio a la paciente. El procedimiento para determinar si la masa es benigna o maligna debería ser lo menos invasivo posible. En este sentido, podemos ver el espectro de técnicas diagnósticas que van desde la mamografía, que es no invasiva, pero proporciona información diagnóstica imperfecta, hasta el examen patológico de masas extirpadas, que es máximamente invasivo, pero resuelve completamente la cuestión del diagnóstico.

También existen métodos a medio camino, como la aspiración con aguja fina (FNA, del inglés *Fine Needle Aspiration*) (Figura 2). Este procedimiento mínimamente invasivo implica la inserción de una aguja de calibre pequeño en una masa mamaria localizada y la extracción de una pequeña cantidad de material celular.

Antes de la aparición de herramientas computacionales para el apoyo al diagnóstico, la sensibilidad reportada (es decir, la capacidad de diagnosticar correctamente el cáncer cuando la enfermedad está presente) de la mamografía variaba del 68% al 79% (Fletcher et al., 1992), y la de la FNA con interpretación visual del 65% al 98% (Giard et al., 1992), siendo la biopsia quirúrgica la única opción cercana al 100%. Por lo tanto, la mamografía carecía de sensibilidad, la sensibilidad de FNA variaba ampliamente, y la biopsia quirúrgica, aunque precisa, es invasiva, consume tiempo y es costosa. Con la mejora de las capacidades computacionales y la aparición de modelos predictivos potentes, estos números han cambiado drásticamente las últimas dos décadas, siendo actualmente la sensibilidad de la mamografía cercana al 90% (Maitra et al., 2012) y la de la FNA y otras biopsias no invasivas por punción cercana al 100% (Hickman et al., 2016). En los últimos años se ha venido sustituyendo la FNA por otras técnicas de punción muy similares pero que utilizan agujas de un calibre ligeramente superior, para garantizar que la muestra obtenida contiene células de la masa que se desea analizar. No obstante, el procedimiento es prácticamente idéntico, tanto en la punción como en el posterior análisis del tejido recogido (American Cancer Society, 2022).

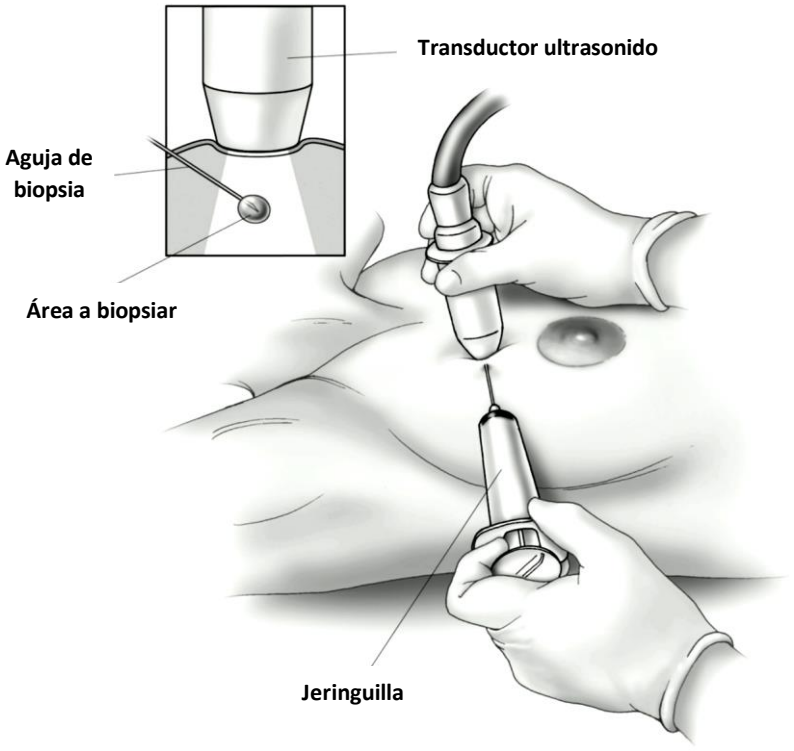


Figura 2. Proceso de aspiración de una masa sospechosa mediante punción con aguja fina, guiada por imagen de ultrasonido (American Cancer Society, 2022). Imagen modificada al castellano.

1.2. Herramientas para el apoyo al diagnóstico del cáncer

Las diferencias principales entre un tejido normal y uno canceroso son la forma y tamaño de sus células y núcleos celulares. Debido a su crecimiento descontrolado, las células cancerosas varían más en tamaño y forma que las normales. Respecto a los núcleos celulares, estos suelen ser más oscuros, grandes e irregulares (Figura 3). Estas diferencias físicas son utilizadas por los expertos a la hora de clasificar un tejido.

El procedimiento habitual cuando se realiza la biopsia de una masa extraña en un órgano consiste en la evaluación detallada mediante microscopio por parte de un patólogo experto, que utilizando sus conocimientos en histopatología determina si hay presencia de células cancerosas en la muestra. Antes de la inspección microscópica de la muestra de tejido por parte del patólogo, se obtiene una pequeña muestra de la masa y se coloca en un portaobjetos, tras lo cual se somete a una serie de procesos de tinción para permitir una óptima visualización del tejido y sus células (Ghayumizadeh et al., 2012). Las muestras histopatológicas típicas constan de un elevado número de células y estructuras que están rodeadas y dispersas de forma aleatoria por una gran variedad de tipos de tejidos (Figura 4).

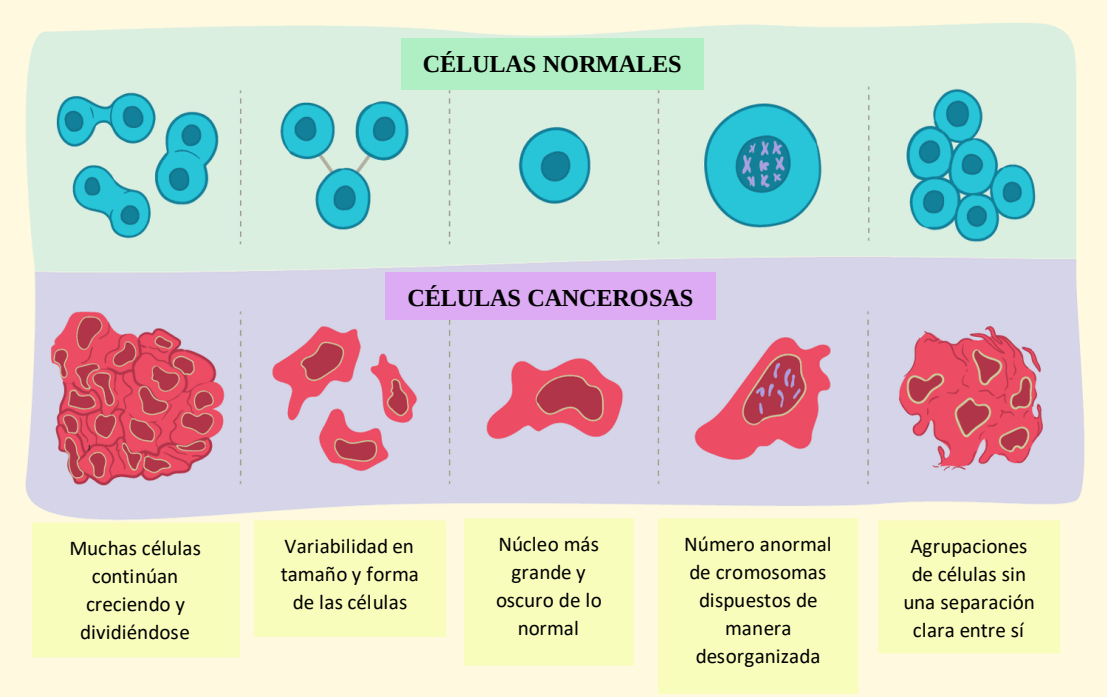


Figura 3. Principales diferencias en las morfología celular y nuclear en tejidos normales y tejidos cancerosos. Modificada de (https://www.verywellhealth.com/cancer-cells-vs-normal-cells-2248794)

La interpretación física de las imágenes histopatológicas, así como el recorrido visual a través de estas (al utilizar grandes aumentos en el microscopio, el terreno a abarcar es muy grande a pesar del reducido tamaño muestral), lleva tiempo. Requiere años de experiencia y pericia. Además, debido al factor humano, existe cierta variabilidad en la interpretación de las imágenes, lo cual se traduce en diferencias en la precisión del diagnóstico.

Para aumentar la capacidad analítica y predictiva de estas imágenes, en los 90 comenzaron a surgir herramientas para el análisis de imágenes asistido por ordenador, utilizando como criterio de clasificación características físicas de células y núcleos como las que acabamos de comentar (Street et al., 2000). Estas herramientas, aparte de mejorar la precisión del diagnóstico y reducir la variabilidad entre expertos, también contribuyen a la eficiencia de los patólogos, ofreciendo una segunda opinión fiable, lo que aumenta su productividad, permitiendo agilizar los diagnósticos y, como resultado, reduciendo la tasa de mortalidad y la carga de los patólogos.

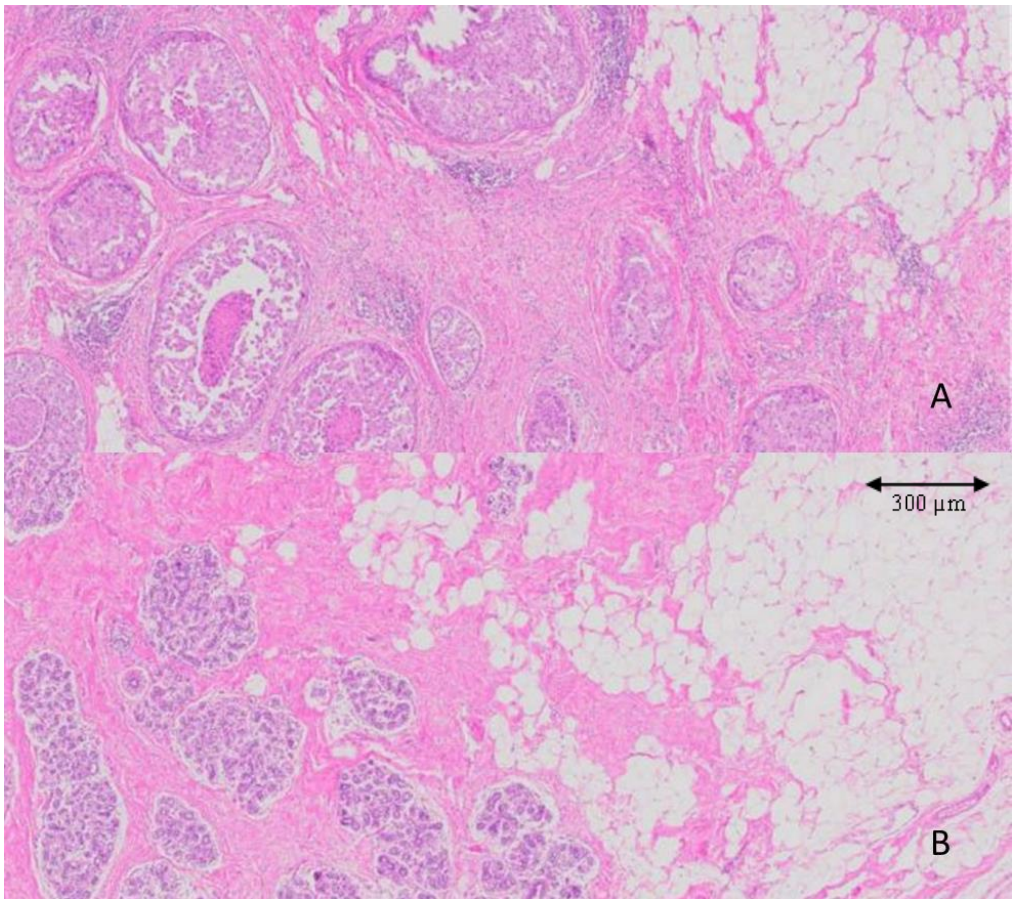


Figura 4. Imágenes histológicas de tejido mamario, obtenidas mediante tinción con hematoxilina-eosina y observación en microscopio óptico. A) tejido canceroso. B) tejido normal. Se puede apreciar la mayor dispersión de las células en el tejido canceroso, siendo menos claro el límite entre ellas. (Imágenes modificadas de la web de la empresa VitroVivo Biotech LLC, https://vitrovivo.com/)

1.3. Descripción de los datos y trabajo previo

En esta memoria se ha trabajado con el banco de datos *Wisconsin Diagnostic Breast Cancer*, que contiene información de los núcleos celulares de 569 pacientes. El banco de datos se obtuvo a través de la plataforma *Kaggle*, una comunidad online con cursos y competiciones en las que se resuelven problemáticas reales utilizando modelos estadísticos avanzados. Fue distribuido originalmente por el Dr. William H. Wolberg, en el Departamento General de Cirugía de la Universidad de Wisconsin-Madison, Estados Unidos; se puede acceder a él a través del siguiente link (https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

(último acceso el 28 de Marzo de 2023)). Los datos se generaron utilizando muestras de masas sólidas en senos de pacientes, obtenidas mediante aspirado por aguja fina. Tras el procesamiento de las muestras y la generación de las imágenes histológicas, se empleó un software llamado *Xcyt* para escanearlas y generar las variables utilizadas en este estudio (Street et al., 2000). Este software utiliza un algoritmo que devuelve la media, el peor valor y el error estándar de 10 características físicas de los núcleos de cada muestra. Las características son las siguientes (a lo largo del estudio se han mantenido los nombres de las variables en inglés, tal y como se obtuvieron del banco de datos):

- **Radio (radius)** - distancia media del centro a puntos en el perímetro
- **Textura (texture)** - desviación estándar de valores de color en escala de grises
- **Perímetro (perimeter)**
- **Área (area)**
- **Suavidad (smoothness)** - variación local en la longitud de los radios
- **Compacidad (compactness)** - calculada como $\frac{perímetro^2}{área - 1}$
- **Concavidad (concavity)** - severidad de los puntos cóncavos en el perímetro
- **Puntos cóncavos (concave points)** - número de puntos cóncavos en el contorno nuclear
- **Simetría (symmetry)**
- **Dimensión fractal (fractal dimension)** - índice para caracterizar la complejidad del núcleo.

Tenemos por tanto un total de 30 covariables con las que trabajar. También están presentes en el banco de datos un número identificador de la paciente (id) y el diagnóstico (diagnosis). El id no ha sido utilizado a la hora de ajustar los modelos. Nuestra variable respuesta es el diagnóstico, una variable binaria que nos dice si la masa es benigna (B) o maligna (M).

En el artículo original, Street et al. (2000) utilizaron el banco de datos generado por *Xcyt* para ajustar varios modelos que clasificaran correctamente las muestras como benignas o malignas. Para este sistema de diagnóstico utilizaron una variante del método multisuperficie (MSM, del inglés Multi-Surface Method) (Mangasarian et al., 1968), llamada MSM-Tree (MSM-T) (Bennett et al., 1992). El método consiste en lo siguiente: se guarda un vector maligno *m* en una matriz n-dimensional *m x n* A, y un vector benigno *b* en una matriz *b x n* B. El modelo MSM-T va a buscar el hiperplano que mejor separe los puntos de las matrices A y B, generando un árbol de decisión cuyos nodos son planos que mejoran la separación de A y B, hasta llegar o bien a la separación total de A y B o a un criterio de parada, como un número mínimo de observaciones por nodo.

Tras realizar una búsqueda global manual del mejor subconjunto de variables, Street et al. afirman haber obtenido una exactitud del 97.5% mediante validación cruzada (Stone et al., 1974), una mejora considerable respecto a la mayoría de las interpretaciones visuales por expertos. Una vez obtenido dicho modelo, decidieron

integrarlo en el software *Xcyt* para que devolviera el diagnóstico directamente a partir de una imagen histológica. También decidieron hacer la herramienta pública para su uso en la *World Wide Web*. Por desgracia, el enlace mencionado en el artículo ha quedado obsoleto y ya no se puede acceder a él.

Hoy en día ya existen técnicas que permiten analizar la imagen directamente sin tener que procesar previamente los datos, utilizando distintas arquitecturas de redes neuronales y otras técnicas de visión por computador (Reshma et al., 2022). Además de esto, se han desarrollado en el laboratorio múltiples técnicas histoquímicas con moléculas afines a células cancerosas, que permiten clasificar los tejidos en distintos tipos de cáncer e incluso dar una prognosis detallada a la paciente (Hammond et al., 2010). Aun así, en los países menos desarrollados no se tiene acceso a este tipo de reactivos debido a su elevado precio, por lo que se sigue optando por tinciones histológicas comunes cuya interpretación es más complicada.

1.4. Objetivos del trabajo

A pesar de que hoy en día existan modelos y técnicas avanzados que permiten no solo obtener un diagnóstico preciso, sino realizar una clasificación y una caracterización de la tipología de cáncer mucho más descriptivas, vamos a centrarnos en la primera versión de *Xcyt*, mediante la cual se afirmaba conseguir una exactitud del 97.5% en el diagnóstico. El objetivo de este trabajo es doble: por un lado, utilizando el mismo banco de datos, un ordenador portátil actual y modelos similares a los estudiados a lo largo del máster, investigar si podemos igualar o mejorar esa exactitud. Por otro lado, aprender a ajustar nuevos tipos de modelos estadísticos y de aprendizaje automático, así como a automatizar tanto su validación como la selección de hiperparámetros óptimos.

2. METODOLOGÍA

2.1. Análisis exploratorio

Antes de proceder al ajuste de modelos predictivos, realizamos una exploración de nuestros datos, con el objetivo de comprender mejor la distribución de las variables del banco de datos, las diferencias en dicha distribución entre los dos tipos de tumores, y la correlación entre todas las covariables. Utilizando esta información podremos decidir qué variables incluir en nuestros modelos predictivos y cómo preprocesar los datos antes de ajustarlos. Para todas las representaciones gráficas, se ha utilizado una paleta de colores accesibles para personas con daltonismo u otras alteraciones en la percepción del color, combinando tonos morados con tonos amarillos y naranjas.

Comenzamos observando la distribución de nuestra variable respuesta, el diagnóstico. Esta variable representa un conjunto de diagnósticos tipo Bernoulli en los que se determina si la paciente padece de un tumor maligno (M) o benigno (B). Representamos un diagrama de barras y una tabla de frecuencias para determinar si hay una representación suficiente de ambos grupos en nuestro banco de datos.

Para explorar el resto de covariables y sus correlaciones utilizamos gráficos de dispersión bivariables y matrices de correlación. Esta información nos será de utilidad más adelante a la hora de mejorar el ajuste de nuestros modelos mediante la eliminación de variables muy correlacionadas entre sí.

A la hora de determinar las diferencias en la distribución de las covariables entre tumores malignos y benignos, utilizamos dos tipos de gráficos a partir de los cuales se obtiene una información similar. Por un lado, representamos las densidades de cada covariable separadas por diagnóstico; por otro, diagramas de caja, también para cada covariable y separados por diagnóstico. Estas representaciones nos van a permitir discernir si existe una diferencia significativa en la distribución de las covariables entre los dos grupos de tumores.

Para mayor rigurosidad de este análisis por grupos, se realizan 30 test-t para comparar las medias de los dos grupos de tumores en todas las covariables. A pesar de que muchas de estas covariables no parecen seguir una distribución normal, sabemos que, por el teorema del límite central, tomando las medias de muchas submuestras de una variable, obtendremos que esta sí que sigue una distribución normal, por lo que en este caso es lícito utilizar este tipo de prueba. Al ser las muestras tumorales de pacientes aleatorias sin relación entre sí, utilizaremos un test-t no pareado. Las covariables que no muestren una diferencia significativa entre las medias de ambos grupos se descartarán y no se tendrán en cuenta a la hora de modelizar.

2.1.1. Análisis de Componentes Principales

Al ser nuestro banco de datos de tamaño mediano, con 30 covariables y 569 individuos, empieza a resultar complicado encontrar patrones generales en los datos mediante la visualización de las variables por separado, por lo que una reducción de

la dimensionalidad podría resultar beneficiosa. Procedemos por tanto a realizar un análisis de componentes principales (PCA, del inglés *Principal Component Analysis*).

Esta técnica de reducción de la dimensionalidad es utilizada ampliamente para simplificar conjuntos de datos complejos. Su objetivo principal es transformar un banco de datos de múltiples dimensiones en un espacio de menor dimensionalidad, manteniendo la mayor cantidad posible de la información original (Jolliffe et al., 2016).

El PCA logra esto creando un conjunto de nuevas variables, llamadas componentes principales, que son combinaciones lineales de las variables originales. Estas componentes se construyen de tal manera que la primera componente captura la mayor varianza posible en los datos, la segunda componente captura la segunda mayor varianza, y así sucesivamente. En nuestro caso vamos a trabajar con la matriz de correlaciones en lugar de la de varianzas, ya que las escalas de las varianzas en nuestras covariables con muy diferentes entre sí, y esto altera su importancia relativa en las componentes principales.

Al proyectar los datos originales en estas componentes principales, podemos reducir de manera efectiva la dimensionalidad del conjunto de datos. Esta reducción se logra al retener solo las primeras k componentes que explican la mayor parte de la varianza en los datos, descartando las componentes con varianzas más pequeñas. La elección del número de componentes a retener depende del nivel deseado de reducción de dimensionalidad y la cantidad de información que estemos dispuestos a sacrificar.

En nuestro caso realizamos dos representaciones, una con una única componente principal, y otra con las dos primeras componentes principales. Esto nos permite visualizar e interpretar los datos de una manera más manejable y comprensible, así como proporcionarnos información sobre la estructura subyacente de los datos, lo cual puede revelar patrones, grupos o valores atípicos que no son tan fáciles de observar analizando todas las variables por separado.

2.2. Modelización

Una vez finalizado el análisis descriptivo de los datos, procedemos al ajuste de varios clasificadores binarios con el objetivo de discriminar lo más precisamente posible entre tumores malignos y benignos. A continuación, se describen los modelos utilizados, así como las métricas utilizadas para validarlos y evaluarlos.

2.2.1. Validación y métricas de evaluación de los modelos

En este trabajo se ajustarán los modelos buscando maximizar la exactitud del diagnóstico, definida como la proporción de individuos clasificados correctamente (verdaderos positivos y verdaderos negativos) sobre el total de individuos. En el lenguaje castellano coloquial se utilizan de forma indiscriminada los términos exactitud y precisión, pero a la hora de evaluar modelos estadísticos es importante diferenciar entre ambos, siendo la precisión la proporción de predicciones positivas

correctamente clasificadas (Stallings et al., 1971). Se ha decidido utilizar esta métrica de evaluación por las razones detalladas a continuación.

En primer lugar, la exactitud proporciona una medida sencilla y fácil de entender del rendimiento del modelo. Representa la proporción de predicciones correctas entre todas las predicciones realizadas por el modelo. Esta simplicidad lo hace conveniente para comunicar e interpretar los resultados a las partes interesadas no técnicas, como podrían ser los médicos u otros profesionales de la salud interesados en el uso de una herramienta automática de diagnóstico.

En segundo lugar, el banco de datos con el que se trabaja está suficientemente balanceado (véase figura 13). Cuando la distribución de clases es lo bastante equilibrada, la exactitud tiende a proporcionar una medida confiable del rendimiento general del modelo. Otorga el mismo peso a las predicciones correctas para ambas clases y evalúa la capacidad del modelo para clasificar los individuos correctamente en todo el conjunto de datos. En este caso es tan importante diagnosticar correctamente a una paciente de cáncer, para poder tratarla en un estadio temprano, como detectar la ausencia de cáncer, ya que el tratamiento del cáncer de mama es un procedimiento invasivo y doloroso para la paciente, y un diagnóstico positivo erróneo podría traer un sufrimiento innecesario a esta.

En tercer lugar, Street et al. (2000) utilizan la exactitud como métrica de evaluación a la hora de ajustar el modelo descrito en su artículo, utilizando el mismo banco de datos para el ajuste; esto permite realizar una comparación directa entre los modelos ajustados en este trabajo y el ajustado por ellos. No obstante, una vez ajustados los modelos tratando de optimizar la exactitud, realizaremos una matriz de confusión para conocer la sensibilidad y especificidad de estos, ya que estas métricas nos brindan información muy importante que complementa adecuadamente la brindada por la exactitud (figura 5).

<table>
<tr><td colspan="2" rowspan="2">Matriz de confusión</td><td colspan="2">Estimado por el modelo</td><td colspan="2" rowspan="2"></td></tr>
<tr><td>Negativo (N)</td><td>Positivo (P)</td></tr>
<tr><td rowspan="2">Real</td><td>Negativo</td><td>a: (TN)</td><td>b: (FP)</td><td></td><td></td></tr>
<tr><td>Positivo</td><td>c: (FN)</td><td>d: (TP)</td><td>Precisión
("precision")
Porcentaje predicciones positivas correctas:</td><td>d/(b+d)</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>Sensibilidad, exhaustividad
("Recall")
Porcentaje casos positivos detectados</td><td>Especifidad
(Specifity)
Porcentaje casos negativos detectados</td><td colspan="2">Exactitud
("accuracy")
Porcentaje de predicciones correctas</td></tr>
<tr><td>d/(d+c)</td><td>a/(a+b)</td><td colspan="2">(a+d)/(a+b+c+d)</td></tr>
</table>

Figura 5. Esquema de una matriz de confusión y las distintas métricas obtenidas a partir de esta. Tabla obtenida de (https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/)

La sensibilidad, también conocida como tasa de verdaderos positivos, es la proporción de casos positivos que el modelo clasifica correctamente. En otras palabras, es la capacidad del modelo para identificar de manera adecuada los casos verdaderamente positivos. Se calcula dividiendo el número de verdaderos positivos entre la suma de verdaderos positivos y falsos negativos:

$$Sensibilidad = Verdaderos\ positivos\ /\ (Verdaderos\ positivos + Falsos\ negativos)$$

Por otro lado, la especificidad es la proporción de casos negativos que el modelo clasifica correctamente. Es la capacidad del modelo para identificar de manera adecuada los casos verdaderamente negativos. Se calcula dividiendo el número de verdaderos negativos entre la suma de verdaderos negativos y falsos positivos:

$$Especificidad = Verdaderos\ negativos\ /\ (Verdaderos\ negativos + Falsos\ positivos)$$

Combinando exactitud, sensibilidad y especificidad obtenemos una evaluación más completa de la utilidad de nuestros modelos para el diagnóstico, ya que cada una de estas medidas proporciona información valiosa sobre aspectos específicos de la capacidad del modelo, complementándose entre sí.

Con el objetivo de estimar el valor real de estas métricas de evaluación para cada uno de los modelos ajustados, han sido utilizados métodos de validación cruzada múltiple de k-iteraciones. La validación cruzada consiste en dividir el banco de datos en un número determinado de subconjuntos del mismo tamaño, y proceder a ajustar el modelo utilizando k-1 subconjuntos cada vez, y el subconjunto restante para calcular la métrica deseada. Se realiza el proceso k veces de forma que todos los subconjuntos han sido utilizados una vez para validación y k-1 veces para el ajuste, y posteriormente se calcula el valor estimado de la métrica como la media de todas las iteraciones (figura 6). De este modo todos los individuos del banco de datos son utilizados en la estimación, dando lugar a una estimación de las métricas de evaluación más precisa que dividiendo el banco de datos en un único subconjunto de ajuste y un único subconjunto de validación (Kuhn et al., 2013).

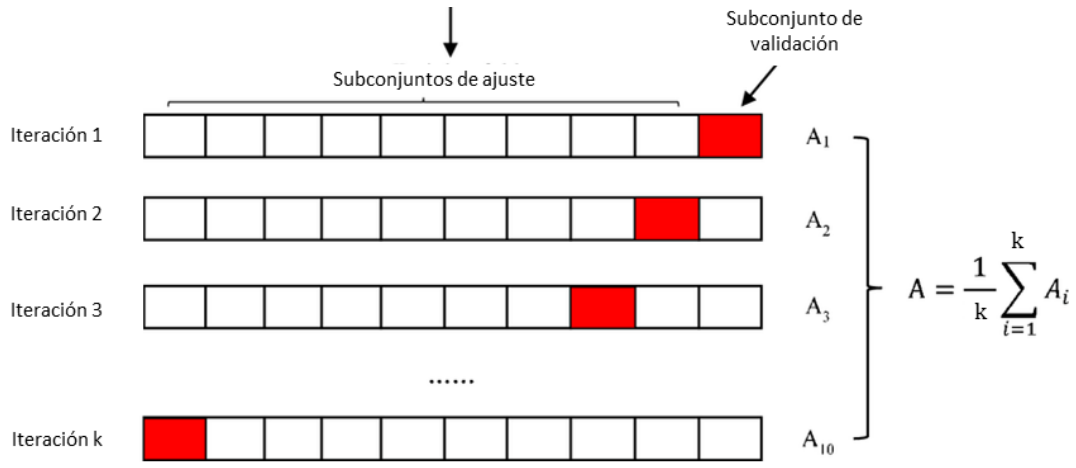


Figura 6. Esquema del funcionamiento de la estimación de la exactitud de un modelo utilizando validación cruzada de k iteraciones. Imagen modificada de Pu et al., 2019.

La validación cruzada múltiple consiste en repetir el proceso mencionado n veces, realizando un muestreo aleatorio previo, es decir, generando k subconjuntos de individuos aleatorios diferentes cada vez (Tougui et al., 2021). En este trabajo hemos realizado una validación cruzada de diez iteraciones, repetida tres veces (k = 10; n = 3).

Al comienzo de la modelización, se validaron los primeros modelos utilizando validación cruzada *leave-one-out*, que consiste en utilizar cada individuo como un subconjunto, dando lugar a una validación cruzada con tantas iteraciones como número de individuos tenga el banco de datos. Esto suponía realizar 569 iteraciones para cada modelo, multiplicado por el número de modelos ajustados. El resultado fue un coste computacional muy elevado, con unos tiempos de ejecución muy altos debido al elevado número de modelos ajustados. Tras leer literatura al respecto, se optó por estimar las métricas de evaluación de todos los modelos utilizando el método de validación cruzada múltiple ya descrito, ya que los resultados se aproximan mucho a los obtenidos por validación cruzada *leave-one-out*, siendo sustancialmente más precisos que los obtenidos mediante validación cruzada sin repeticiones (Tougui et al., 2021).

Para una mayor consistencia y reproducibilidad de los análisis realizados, se utiliza una semilla a la hora de realizar cualquier cálculo o proceso que conlleve generar números pseudoaleatorios, como a la hora de realizar las particiones en la validación cruzada. En el lenguaje de programación R, esto se consigue mediante la función $set.seed(x)$, donde x sería la semilla. En este trabajo se utiliza el valor 100 como semilla.

Para tener una idea de la complejidad de cada modelo, se han medido también sus tiempos de ejecución. Todos ellos han sido ajustados utilizando el mismo equipo, en una sola sesión, siendo RStudio el único programa ejecutándose en la máquina. Se utilizan la misma semilla y el mismo método de validación para que tanto las métricas estimadas como los tiempos de ejecución sean comparables entre modelos.

2.2.2. Regresión logística

Uno de los modelos más utilizados cuando se trata con variables dicotómicas es la regresión logística. Es un tipo de modelo lineal generalizado que se centra en modelar la relación entre las covariables (predictores) y la probabilidad de que ocurra un evento o la probabilidad de que una observación pertenezca a una categoría específica.

El concepto fundamental detrás de la regresión logística se basa en la idea de la transformación de las odds o logit. No existe una traducción directa al castellano de la palabra odds; dada una probabilidad p, las odds correspondientes a dicha probabilidad se calculan como $p/(1-p)$. La función logit es el logaritmo de las odds:

$$logit(x) = \log\left(\frac{x}{1-x}\right)$$

Esta transformación convierte la probabilidad de un evento en una relación lineal con los predictores, lo que nos permite aplicar técnicas de regresión a un problema de clasificación.

Matemáticamente, el modelo de regresión logística se puede expresar de la siguiente manera:

$$logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_\mathrm{p} X_\mathrm{p},$$

donde $logit(p)$ es el log-odds del evento, p es la probabilidad de que ocurra el evento, β_0 es el intercepto, y $\beta_1, \beta_2, \ldots, \beta_\mathrm{p}$ son los coeficientes de los predictores $X_1, X_2, \ldots, X_\mathrm{p}$, respectivamente.

Los coeficientes ($\beta_1, \beta_2, \ldots, \beta_\mathrm{p}$) en la ecuación de regresión logística representan la relación entre cada predictor y el log-odds del resultado. Indican el cambio en el log-odds por un cambio unitario en el predictor correspondiente, asumiendo que todos los demás predictores se mantienen constantes. El signo y la magnitud de los coeficientes proporcionan información sobre la dirección y la fuerza de la asociación entre los predictores y el resultado.

Para estimar los coeficientes, se utiliza un proceso llamado estimación de máxima verosimilitud. El objetivo es encontrar los valores de los coeficientes (β) del modelo que hacen que los datos observados sean más probables, es decir, encontrar el conjunto de coeficientes que maximiza la función de verosimilitud. Para lograr esto, se utiliza un enfoque iterativo. Inicialmente, se selecciona un conjunto de valores iniciales para los coeficientes. Luego, se calcula la función de verosimilitud del modelo con esos coeficientes y se busca una forma de ajustar los coeficientes para aumentar la verosimilitud.

Una vez que se ajusta el modelo de regresión logística, se puede utilizar para hacer predicciones. Las probabilidades predichas se obtienen al insertar los valores de los predictores en la ecuación de regresión logística y aplicar la función sigmoide,

$$S(x) = \frac{1}{1 + e^{-x}}$$

para deshacer la transformación logit previamente mencionada. Luego se elige un umbral para clasificar las probabilidades predichas en categorías específicas, según la naturaleza del problema (Gareth et al., 2013). En nuestro caso se elige un umbral de 0,5, lo que significa que cualquier probabilidad superior a este número se tomará como éxito (o como diagnóstico positivo para cáncer en este caso), y cualquier probabilidad menor como fracaso o diagnóstico negativo para cáncer.

La regresión logística ofrece varias ventajas. Proporciona una interpretación directa de la relación entre los predictores y la probabilidad de un evento, lo que la hace útil para obtener conclusiones significativas y realizar predicciones. Puede

manejar tanto predictores categóricos como continuos, y puede tener en cuenta las interacciones entre los predictores.

Sin embargo, la regresión logística también tiene limitaciones. Asume una relación lineal entre los predictores y los log-odds, lo que puede no ser válido en todos los casos. Es sensible a los valores atípicos y puede estar influenciada por la multicolinealidad entre los predictores, aunque este último problema puede mitigarse haciendo una correcta selección de covariables (Sperandei, 2014).

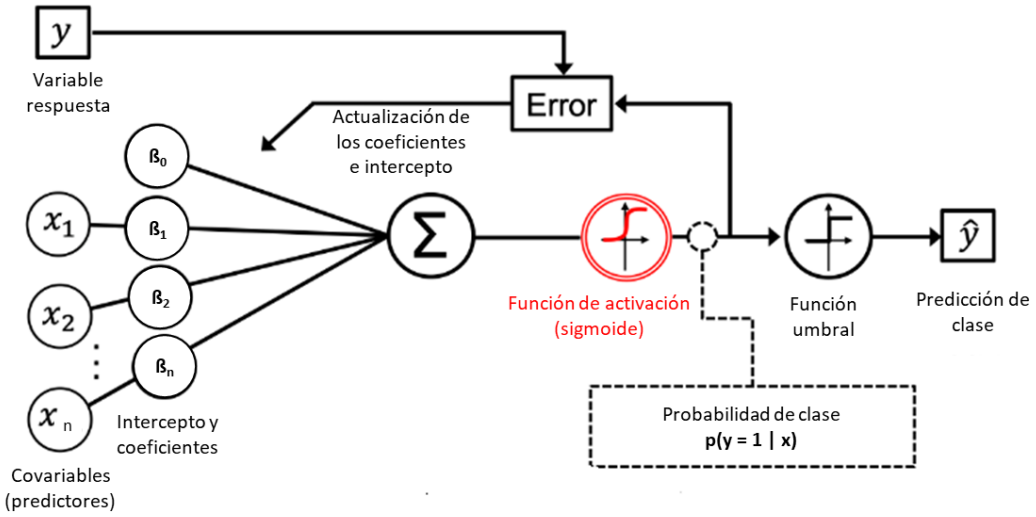


Figura 7. Esquema del funcionamiento de la regresión logística. Imagen modificada de (https://vitalflux.com/python-train-model-logistic-regression/).

En este caso se ajustan múltiples modelos de regresión logística, comenzando por un modelo con todas las covariables, para posteriormente utilizar varios métodos de eliminación de covariables con el objetivo de eliminar ruido del modelo y mejorar su exactitud.

2.2.2.1. Sobre todas las variables

El primer modelo ajustado es una regresión lineal con todas las covariables, que tomaremos como base ya que es el más sencillo de implementar, aunque también será el que más problemas de ruido y colinealidad posee. Si el banco de datos y el número de covariables fuera lo suficientemente pequeño, se podría encontrar el mejor subconjunto de predictores ajustando todas las posibles combinaciones de covariables, llegando al final al modelo de regresión logística con la mayor exactitud posible (sin realizar ninguna transformación a ninguna covariable). En nuestro banco de datos esto supondría un coste computacional inasumible, ya que teniendo 30 covariables tendríamos que ajustar $2^{30} \approx 1000$ millones de modelos. Por lo tanto, vamos a necesitar emplear otros métodos para simplificar y mejorar la exactitud de nuestro modelo.

2.2.2.2. Utilizando las componentes principales como covariables

Dado que las primeras componentes principales parecen recoger una gran cantidad de información del banco de datos en unas pocas covariables, y que a simple vista podemos observar una separación muy buena con tan sólo las dos primeras componentes, vamos a comenzar a ajustar una regresión logística sobre las componentes principales.

La ventaja principal de este planteamiento es que se reduce la dimensionalidad y se facilita la selección de variables, ya que, al examinar la contribución de cada componente principal a la varianza general, se vuelve más fácil priorizar las componentes que explican la mayor variación en los datos y seleccionarlas como predictores en el modelo. Por tanto, hacer una selección de covariables se vuelve tan sencillo como empezar con un modelo con la primera componente principal, es decir, la componente que mayor varianza de los datos explique, e ir añadiendo de forma secuencial las siguientes componentes principales, en orden de varianza explicada. Una vez ajustados todos los modelos, se selecciona aquel que dé lugar a una mayor exactitud, quedándonos con un subconjunto de componentes principales menor al conjunto inicial de covariables.

El uso de componentes principales en lugar de covariables también contribuye a la reducción de la multicolinealidad en los bancos de datos cuyos predictores estén altamente correlacionados, ya que, debido a la naturaleza de la PCA, se genera un nuevo conjunto de variables ortogonales entre sí, reduciendo así la interdependencia entre los predictores.

2.2.2.3. Utilizando métodos *stepwise*

Al no poder realizar una búsqueda exhaustiva del mejor subconjunto de covariables, debemos optar por buscar algún criterio de eliminación de covariables, o algún método automático de selección de estas que no sea tan costoso computacionalmente. Uno de los métodos más conocidos de selección automática de covariables son los métodos *stepwise*. En este tipo de métodos se agregan o eliminan sistemáticamente predictores del modelo en función de su significancia estadística. Hay dos tipos principales de procedimientos paso a paso: selección hacia adelante y eliminación hacia atrás (*stepwise forward* y *stepwise backward* en inglés, respectivamente).

La métrica utilizada habitualmente para decidir la adición o eliminación de covariables en este tipo de modelos es el Criterio de Información de Akaike (AIC, del inglés *Akaike Information Criterion*). Esta es una medida que combina el ajuste del modelo a los datos y la complejidad del modelo, buscando encontrar el equilibrio óptimo entre estos dos aspectos. Su objetivo es seleccionar el modelo que mejor se ajuste a los datos con la menor cantidad de covariables posible, evitando el sobreajuste. Se basa en maximizar la función de verosimilitud del modelo, penalizando por el número de covariables incluidas. Cuanto menor sea el valor del

AIC, mejor se considera el modelo. Al comparar varios modelos utilizando el AIC, se selecciona aquel con el valor más bajo como el modelo preferido (Bozdogan et al., 1987).

La selección hacia adelante consiste en comenzar con un modelo sin covariables e ir realizando iteraciones sobre todas las covariables. Se prueba a ajustar un modelo con la primera covariable, después una con la segunda, así hasta haber probado todos los posibles modelos de una covariable. Se selecciona la covariable que haya dado lugar a un mejor AIC, y se repite el mismo proceso para ver cual debe ser la siguiente covariable a añadir. Tan sólo se añade una covariable si esa adición da lugar a una mejora en el modelo. Si termina una iteración a través de todas las posibles covariables a añadir y ninguna ha conllevado una reducción del AIC, el algoritmo finaliza, y se toma como mejor modelo el último modelo seleccionado con menor AIC.

La eliminación hacia atrás sigue el mismo proceso, pero a la inversa, partiendo del modelo con todas las covariables y eliminando sistemáticamente covariables de una en una. El algoritmo termina cuando, tras terminar una iteración a través de todas las posibles covariables a eliminar, ninguna ha conllevado a una reducción del AIC (Gareth et al., 2013).

Estos dos algoritmos suelen dar lugar a modelos diferentes, por lo que ejecutamos ambos. Es importante destacar que el AIC es una medida del ajuste del modelo, no de su poder predictivo, por lo que el hecho de que un modelo posea un AIC reducido no tiene por qué suponer una mejora en la exactitud de este.

El objetivo de este trabajo es la obtención de la mayor exactitud posible en el diagnóstico, pero los paquetes de R utilizadas no permiten el uso de la exactitud como criterio de selección. Debido a esto, se programó el algoritmo *stepwise* manualmente en R, pero utilizando la exactitud como criterio de selección en lugar del AIC. Se programaron tanto el método de selección hacia adelante como el de eliminación hacia atrás.

2.2.2.4. Eliminando variables muy correlacionadas

Se debe tener en cuenta que utilizando métodos *stepwise*, a pesar de ajustarse un número considerable de modelos, no es habitual que este método de lugar al mejor subconjunto de variables, ya que dejamos sin ajustar un porcentaje elevado de combinaciones. Por eso se deben probar otros enfoques, como eliminar o seleccionar covariables siguiendo distintos criterios. En los siguientes modelos ajustados se eliminaros las covariables que estuvieran altamente correlacionadas.

Observando la matriz de correlaciones de todas las covariables, se filtran las parejas de covariables con una alta correlación entre sí. Para ello, se seleccionaron dos umbrales, 0.9 y 0.95, se obtuvieron todas las parejas de covariables cuya correlación superase dicho umbral, y se ajustaron modelos de regresión logística eliminando una covariable de cada una de las parejas. La hipótesis es que esta eliminación de

covariables reducirá considerablemente la multicolinealidad presente en el modelo con todas las covariables, dando lugar a una mejora en la capacidad predictiva.

2.2.2.5. Utilizando las covariables de mayor importancia en el PCA

Como ya se ha comentado en apartados anteriores, utilizar las componentes principales como predictores en lugar de las covariables originales tiene una serie de ventajas, y es una buena opción para considerar. Sin embargo, uno de los mayores problemas que presenta es que se pierde capacidad de interpretación del modelo, y resulta más complicado de explicar la influencia individual de cada covariable en el modelo final.

Teniendo esto en cuenta, se idea una estrategia en la que utilizar el análisis de componentes principales para ayudar a seleccionar covariables, escogiendo aquellas covariables que hayan tenido una importancia relativa mayor en las primeras componentes principales. Se prueba a seleccionar las covariables cuyos coeficientes superen un valor absoluto de 0.25 en las primeras 9 componentes principales. Se decide optar por este criterio de selección de forma arbitraria, tras probar varias combinaciones posibles manualmente.

2.2.2.6. Regresión con penalización

Una alternativa a la regresión paso por paso, y que habitualmente da lugar a mejores resultados en bancos de datos de alta dimensionalidad, es la regresión penalizada o regularizada. Esta es una extensión de la regresión logística que incorpora un término de penalización al modelo. La regularización ayuda a controlar la complejidad del modelo al modificar la función de pérdida agregando un término de penalización, desalentando así al modelo de depender demasiado de alguna variable predictora en particular. Dos técnicas de regularización comúnmente utilizadas son la regresión Ridge y la regresión Lasso, y existe una tercera, Elastic Net, que combina las dos anteriores.

En la regresión ridge, el objetivo es minimizar la función de pérdida más un término de penalización proporcional a la suma de los coeficientes al cuadrado multiplicados por un hiperparámetro de regularización (λ). Es importante diferenciar entre parámetro e hiperparámetro. Los parámetros son la parte del modelo que se ajusta en función de los datos. Los hiperparámetros son parámetros externos al modelo, que determinan las distintas posibles configuraciones de este. Su valor no puede conocerse a priori, y que deben ser seleccionados utilizando valores genéricos, valores que hayan funcionado bien en problemas similares anteriores, o buscando la mejor opción mediante prueba y error, como es nuestro caso (Claesen et al., 2015).

El hiperparámetro de la regresión Ridge encoge los coeficientes de regresión hacia 0 (pero no exactamente a 0), y ha demostrado tener un buen desempeño en escenarios con predictores correlacionados. La notación matemática del término de penalización es la siguiente:

$$\lambda \sum_{j=1}^{p} \beta^2{}_j$$

La regresión Lasso (del inglés *least absolute shrinkage and selection operator*) impone una restricción a la suma del valor absoluto de los coeficientes de regresión, y el término de penalización es el siguiente:

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

Lasso fue originalmente desarrollado para la selección de variables en análisis de datos de alta dimensionalidad cuando el número de covariables (p) es mucho mayor que el tamaño de la muestra ($p >> n$). En la regresión Lasso, el modelo puede llevar hasta 0 el valor de los coeficientes si λ es lo suficientemente alto, y, por lo tanto, puede realizar selección de covariables.

Elastic net es una combinación de Ridge y Lasso. Tiene una penalización con componentes de ambos:

$$\lambda \sum_{j=1}^{p} \left(\alpha \beta^2{}_j + (1 - \alpha) |\beta_j| \right)$$

El α se puede considerar como un hiperparámetro de mezcla que describe la contribución relativa de Ridge y Lasso al término de penalización. Si $\alpha = 0$, estamos ante una penalización tipo Ridge, y si $\alpha = 1$, ante una de tipo Lasso. Elastic net combina las fortalezas de Ridge y Lasso: puede producir modelos más parsimoniosos que Ridge al realizar selección de variables, al tiempo que tiende a seleccionar u omitir predictores altamente correlacionados como grupo (Yan et al., 2022).

Utilizamos el paquete *glmnet* para ajustar modelos de regresión logística con los 3 tipos de penalización. Este paquete nos permite variar los hiperparámetros α y λ del término de penalización de elastic net. Se describe el procedimiento seguido para ajustar los modelos Ridge y Lasso a continuación.

Mediante la función *cv.glmnet* realizamos una primera búsqueda del valor de λ que minimiza la deviance del modelo, así como del valor más alto de λ que esté a un error estándar o menos del primer λ, es decir, el valor que mayor penalización nos aporte sin desviarse demasiado del valor óptimo estimado. Una vez encontrados estos valores de λ, se realiza una búsqueda más exhaustiva a través de valores de λ comprendidos entre estos dos, pero buscando optimizar la exactitud del modelo en lugar de la deviance.

Una vez ajustados los modelos Ridge y Lasso, se procede a ajustar modelos elastic net, utilizando como referencia los valores de λ obtenidos en Ridge y Lasso, y buscando a través de una batería de valores de α comprendidos entre 0 y 1. Utilizando

la función *expand.grid* de R, a partir de una matriz conteniendo los valores de alfa y lambda se genera un dataframe con todas las posibles combinaciones de ambos hiperparámetros, y se ajusta un modelo elastic net para cada una de estas.

Se prueba también a ajustar estos modelos eliminando previamente variables muy correlacionadas, como se describe en el apartado 2.2.2.4., por si esto mejorase la exactitud del modelo final.

2.2.3. Análisis discriminante lineal

El siguiente clasificador binario utilizado es el análisis discriminante lineal (LDA, del inglés *Linear Discriminant Analysis*). Este es un tipo de modelo del ámbito del reconocimiento de patrones, utilizado tanto para clasificación como para reducción de dimensionalidad de datos. El objetivo principal del LDA es encontrar una transformación lineal que maximice la separación entre clases y minimice la variabilidad dentro de cada clase.

En el LDA, se considera un conjunto de datos etiquetados, donde cada instancia pertenece a una clase específica. El LDA busca encontrar una combinación lineal de las covariables que permita distinguir de manera óptima entre las diferentes clases.

El LDA se basa en dos suposiciones fundamentales: las clases tienen distribuciones normales multivariadas y las matrices de covarianza de ambas clases son similares. Con estas suposiciones, se puede calcular una proyección lineal de los datos en un espacio de menor dimensión, donde la separación entre clases se maximiza.

El proceso del LDA implica varios pasos. Primero, se calculan los vectores medios de cada clase y las matrices de dispersión inter- e intraclase. Luego, se obtienen los vectores y los valores propios de la combinación de estas matrices. Los vectores propios representan las direcciones en las que los datos se proyectarán, y los valores propios indican la importancia de cada dirección.

Los vectores propios se ordenan según los valores propios asociados, y se seleccionan los k vectores principales que mejor expliquen la varianza de los datos. Estos vectores principales forman una matriz de transformación, que se utiliza para proyectar los datos originales en un espacio de menor dimensión.

Para clasificar nuevos individuos utilizando el LDA, estos se proyectan en este espacio de menor dimensión y se utiliza una función discriminante para asignar a estos individuos sus respectivas clases (Xanthopoulos et al., 2013).

A la hora de ajustar este tipo de modelo, se ha seguido un procedimiento muy similar al ya descrito en la regresión logística. En primer lugar, se ajustó un modelo con todas las covariables para tomarlo como punto de partida; luego se utilizaron las componentes principales como covariables; después se utilizaron algoritmos *stepwise forward* y *backward* para generar modelos con subconjuntos de covariables que dieran

lugar a una mayor exactitud; posteriormente se ajustaron modelos eliminando de forma manual las covariables altamente correlacionadas; y por último se ajustó un modelo seleccionando únicamente las covariables con una importancia relativa mayor en las primeras componentes principales.

2.2.4. K-vecinos más cercanos

Tras ajustar todos los modelos LDA, pasamos a ajustar modelos del tipo k-vecinos más cercanos (KNN, del inglés *k-nearest neighbours*). Estos son un tipo de modelos no paramétricos, muy utilizados en problemas de clasificación por su simplicidad y flexibilidad. Muchos algoritmos de recomendación de películas funcionan basándose en modelos de este tipo (Jayalakshmi et al., 2022).

Para ajustar un modelo KNN, partimos de un banco de datos previamente clasificado. Cada individuo consiste en un conjunto de variables y una etiqueta de clase correspondiente. Primero se representan todos los individuos del banco de datos en un espacio p-dimensional, siendo p el número de covariables. Cuando se presenta un nuevo individuo sin etiquetar, el algoritmo busca los k vecinos más cercanos en ese espacio p-dimensional, en función de una medida de distancia. La distancia más comúnmente utilizada es la distancia euclídea, aunque existen muchos otros tipos de distancias que es posible utilizar, como por ejemplo la distancia Manhattan (figura 8).

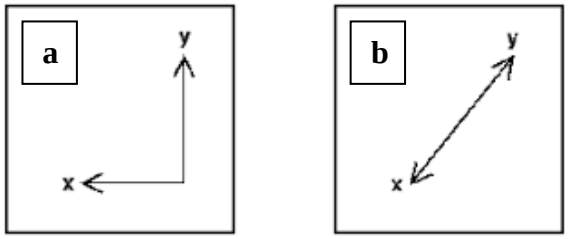


Figura 8. Representación gráfica bidimensional de las distancias Manhattan (a) y euclídea (b). Imagen obtenida de (https://medium.com/analytics-vidhya/euclidean-and-manhattan-distance-metrics-in-machine-learning-a5942a8c9f2f)

Una vez que se han identificado los vecinos más cercanos, el KNN clasifica el nuevo individuo asignándole la etiqueta de la clase más común entre los k vecinos. Por ejemplo, si los vecinos más cercanos son tres elementos de la clase A y dos de la clase B, el nuevo individuo se clasificaría como clase A. Este procedimiento encaja muy bien con el conocido refrán “dime con quién andas y te diré quién eres”.

El valor de k en KNN es un hiperparámetro que se selecciona antes de aplicar el algoritmo. Un valor pequeño de k puede llevar a una clasificación más sensible al ruido en los datos, mientras que un valor grande de k puede suavizar las fronteras de decisión y llevar a una clasificación más generalizada.

El KNN tiene varias ventajas. En primer lugar, es fácil de entender e implementar. Además, no asume ninguna distribución específica de los datos, lo que lo hace útil en situaciones donde los datos siguen patrones complejos. Sin embargo, también tiene algunas limitaciones. Por ejemplo, puede ser computacionalmente costoso en

conjuntos de datos grandes, ya que se requiere calcular la distancia entre el nuevo individuo y todos los individuos utilizados en el ajuste del modelo. En este caso esto no debería preocuparnos, ya que nuestro banco de datos no es lo suficientemente grande como para que esto sea un problema (Gareth et al., 2013).

Se ajustan modelos KNN probando con valores de k entre 1 y 30, y probando a utilizar como distancias las ya mencionadas euclídea y Manhattan. Se comenzó utilizando el paquete *knn*, que utiliza la distancia euclídea para ajustar los modelos, y más adelante se implementaron los modelos con la distancia Manhattan utilizando el paquete *kknn*, la cual utiliza la distancia Minkowski, una generalización que abarca la distancia Manhattan y la euclídea, entre otras. Este paquete nos permite modificar el parámetro p de esta distancia, resultando en la distancia Manhattan cuando $p = 1$, y en la euclídea cuando $p = 2$. La expresión matemática de la distancia de Minkowski es

$$\left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Tras ajustar los modelos sobre todas las covariables, se repitió el proceso eliminando variables muy correlacionadas, del mismo modo al descrito en el apartado 2.2.2.4.

2.2.5. Árboles de decisión

Una vez finalizados los modelos KNN, se procedió a trabajar con árboles de decisión. Los modelos de árboles de decisión son un tipo de algoritmo de aprendizaje automático que se utiliza para realizar tareas de clasificación y regresión. Se basan en la idea de dividir recursivamente el conjunto de datos en subconjuntos más pequeños y homogéneos, utilizando reglas de decisión basadas en características o variables predictoras. Estas divisiones se representan mediante una estructura en forma de árbol, donde cada nodo representa una pregunta o condición sobre una característica y cada rama representa una posible respuesta o resultado.

La construcción de un árbol de decisión se realiza en base a un proceso de aprendizaje supervisado, utilizando un conjunto de entrenamiento que contiene individuos etiquetados con sus respectivas clases o valores objetivo. El objetivo es encontrar las reglas de decisión que maximicen la precisión de la clasificación o minimicen el error en el caso de la regresión.

El proceso de construcción del árbol se realiza de manera iterativa y se basa en algoritmos que evalúan todas las covariables y seleccionan la que mejor separa los individuos en subconjuntos más homogéneos. Esta evaluación se realiza utilizando medidas de impureza, como la entropía o el índice Gini, que cuantifican la mezcla de clases en un subconjunto. En nuestro caso los modelos se ajustaron utilizando el índice de Gini. Este índice indica la probabilidad de que un individuo seleccionado aleatoriamente en un conjunto de datos sea clasificado incorrectamente cuando se le

asigna una clase al azar, según la distribución de las clases en el conjunto. Un índice Gini de 0 indica una pureza total, lo que significa que todas las muestras en el conjunto de datos pertenecen a la misma clase. En cada partición, el árbol busca entre todas las posibles particiones la que más disminuya el índice Gini.

Una vez seleccionada la covariable óptima, se crea un nodo de decisión en el árbol y se generan ramas correspondientes a las posibles respuestas. Luego, el proceso se repite de forma recursiva en cada uno de los subconjuntos creados, bien hasta que se produce una separación total de los grupos (figura 9), o hasta que se alcanza un criterio de parada, como una profundidad máxima predefinida, o que ninguna de las posibles siguientes divisiones supere un umbral de mejora mínimo.

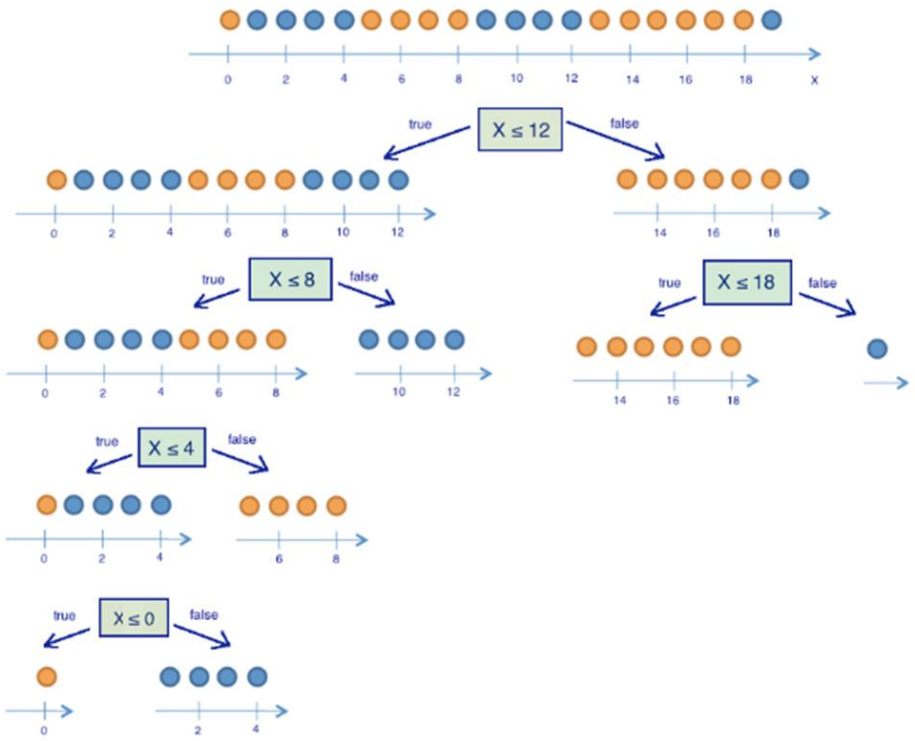


Figura 9. Ejemplo simple del ajuste de un árbol de decisión sobre un banco de datos, con una sola covariable y dos posibles clases. En este caso el algoritmo continúa hasta la separación total de las clases, lo cual no es habitual en problemas reales por problemas de sobreajuste. Imagen obtenida de (https://mlcourse.ai/book/topic03/topic03_decision_trees_kNN.html#useful-resources)

Tras finalizar la construcción del árbol, se utiliza para realizar predicciones sobre nuevos datos. Cada nueva instancia recorre todo el árbol desde la raíz hasta llegar a una hoja, siguiendo las reglas de decisión establecidas en cada nodo. La clase asignada la hoja correspondiente se toma como predicción para esa instancia (Gareth et al., 2013).

Los árboles de decisión tienen varias ventajas, como la capacidad de manejar datos mixtos (categóricos y numéricos), ser fácilmente interpretables y capaces de capturar relaciones no lineales. Sin embargo, también pueden ser propensos al

sobreajuste si no se controla adecuadamente la complejidad del árbol. Por esta razón, es común aplicar técnicas de poda o utilizar estrategias de ensamblaje de árboles (más información sobre esto en el apartado 2.2.6.), para mejorar la generalización y el rendimiento del modelo.

En este trabajo ajustamos los árboles utilizando el paquete *rpart*, que nos permite seleccionar entre dos criterios de parada o poda: un parámetro de complejidad (cp) y la profundidad máxima del árbol (maxdepth). El parámetro de complejidad determina la mejora mínima que se debe dar en la uniformidad de los grupos tras una partición. El árbol termina cuando la siguiente partición supera el umbral de mejora seleccionado. Se ajustan por un lado modelos probando valores de cp comprendidos entre 0 y1, y por otro lado probando profundidades máximas de 2 a 8 niveles. Como ambos casos dieron lugar a árboles finales prácticamente idénticos, los siguientes árboles se ajustaron siguiendo como criterio la profundidad máxima, ya que este hiperparámetro es más sencillo de interpretar. También se probó a ajustar estos modelos realizando una eliminación previa de variables muy correlacionadas.

Para observar la importancia relativa de cada covariable en el modelo, se representa en un gráfico de barras la importancia de cada covariable, estimada por el árbol de decisión más exacto. Esta se mide como el decrecimiento del índice de Gini que se obtiene al realizar una partición del banco de datos.

El problema de los árboles de decisión es que buscan siempre la partición óptima que disminuya al máximo el índice de Gini o maximice la ganancia de información, pero a veces la combinación de particiones "subóptimas" da como resultado un árbol más preciso que el árbol que comienza con la partición óptima. Para intentar mitigar esta limitación, se escogen grupos de variables aleatorios y se ajustan nuevos árboles con ellos. Al no ser viable computacionalmente probar todas las combinaciones de covariables, se realizó una búsqueda reducida, siguiendo un algoritmo aleatorizado. Este algoritmo selecciona desde 3 hasta 12 covariables, seleccionando 10 posibles combinaciones para cada número de covariables propuesto (10 combinaciones de 3 variables aleatorias, después 10 combinaciones de 4, etc.). Para cada combinación se ajustan 4 árboles, variando el parámetro maxdepth de 2 a 5.

2.2.6. *Bagging* y *boosting*

Como se acaba de mencionar, los árboles de decisión pueden presentar problemas de sobreajuste o de no seleccionar la combinación óptima de particiones. La solución de generar árboles aleatorios puede llevarse un paso más allá, y realizar ensamblajes de varios árboles para llegar a modelos más precisos. Los dos métodos más empleados son el ensamblado por votación (*bagging* en inglés, o ***b**ootstrap **agg**regation*), y los árboles secuenciales reforzados (*boosting trees* en inglés).

2.2.6.1. *Bagging*. Bosques aleatorios

Cuando se ensamblan varios árboles mediante *bagging*, al algoritmo se le conoce como bosque aleatorio (*random forest* en inglés). En lugar de depender de un solo

árbol de decisión, los bosques aleatorios construyen una colección de árboles y promedian las predicciones individuales para obtener una predicción final.

Cada árbol de decisión en el bosque se construye utilizando un subconjunto aleatorio de covariables. Esto se conoce como el muestreo aleatorio con reemplazo. Al seleccionar covariables de manera aleatoria, los árboles se vuelven menos propensos a sobreajustarse a patrones específicos en los datos de ajuste y capturan diferentes aspectos del problema.

Como se ha comentado en el apartado anterior, durante la construcción de los árboles, en cada nodo se elige la mejor división entre un subconjunto aleatorio de características. Esto permite una mayor diversidad en los árboles y evita que el bosque se sesgue hacia características dominantes, como ocurre con árboles individuales. Una vez que se han construido todos los árboles, las predicciones se obtienen promediando las predicciones de cada árbol para la clasificación o tomando la media ponderada para la regresión (Gareth et al., 2013) (figura 10).

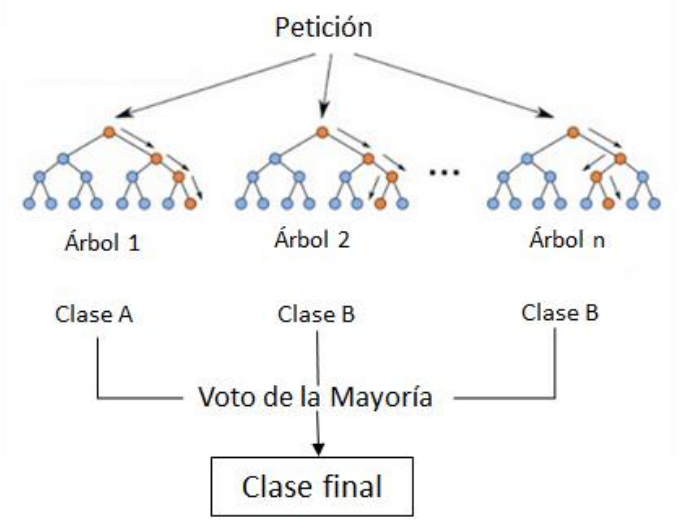


Figura 10. Esquema simplificado del funcionamiento del algoritmo de bosques aleatorios. Imagen obtenida de (https://medium.com/@hpumah/bosques-aleatorios-482163ace92e)

Por defecto, el modelo *random forest* integrado en caret solo permite variar el parámetro mtry, que es el número de variables seleccionadas al azar en cada corte. Para poder probar a variar los valores de más hiperparámetros, se genera función customizada para ajustar modelos *random forest* desde caret, que nos permita probar todas las combinaciones posibles entre:

- mtry: número de variables seleccionado al azar en cada corte.
- ntree: número de árboles de decisión que se van a ajustar.
- nodesize: tamaño de nodo mínimo en todos los árboles.

Como en apartados anteriores, se comienza con un rango de valores para cada hiperparámetro, y se genera una tabla con todas las posibles combinaciones de los

dichos valores. Después caret procede a ajustar todos los modelos y a seleccionar el que mejor exactitud dé como resultado. En función de los resultados obtenidos, se van variando los valores de los hiperparámetros para probar nuevas combinaciones. De nuevo ajustamos un modelo sobre todas las covariables, y otros eliminando variables muy correlacionadas o seleccionando las variables de mayor importancia en las primeras componentes principales.

En un bosque aleatorio los árboles de decisión se calculan seleccionando un número de covariables al azar para cada partición, por lo que se obtiene una estimación de la importancia de las covariables diferente a la obtenida en el modelo de un único árbol de decisión, ya que en este último se escoge siempre entre todas las covariables del banco de datos. Representamos un nuevo gráfico de barras para observar los cambios con respecto al modelo de un único árbol.

2.2.6.2. *Boosting trees*

Los árboles secuenciales de refuerzo, más conocidos en inglés como *boosting trees*, son otro tipo de algoritmo de aprendizaje automático, que combina múltiples árboles de decisión en una secuencia para mejorar gradualmente la precisión del modelo. A diferencia de otros métodos de ensamblaje, como el *bagging*, el *boosting* se enfoca en corregir los errores cometidos por los árboles anteriores en lugar de simplemente combinar sus predicciones.

El proceso de *boosting* comienza con la construcción de un primer árbol de decisión utilizando los datos de ajuste o entrenamiento. A medida que se construyen más árboles, se da mayor importancia a los individuos clasificados incorrectamente en los árboles anteriores. Esto significa que los árboles posteriores se centran en aprender de los errores y mejorar la predicción de los individuos difíciles de clasificar.

Cada árbol se construye de forma secuencial, y su contribución al modelo final se pondera en función de su rendimiento en los datos de entrenamiento. Los árboles que hacen una predicción más precisa tienen un mayor peso en la combinación final de predicciones. Este proceso de construcción secuencial y ponderación adaptativa permite que el modelo se ajuste mejor a los datos y mejore su capacidad de predicción a medida que se construyen más árboles (figura 11).

El *boosting* es especialmente efectivo para problemas en los que los datos son desequilibrados o ruidosos. Al enfocarse en los ejemplos difíciles, los árboles posteriores pueden capturar patrones más sutiles y mejorar la generalización del modelo. Sin embargo, existe el riesgo de sobreajuste si se construyen demasiados árboles, por lo que es importante encontrar un equilibrio adecuado entre la complejidad del modelo y su capacidad para generalizar (Gareth et al., 2013).

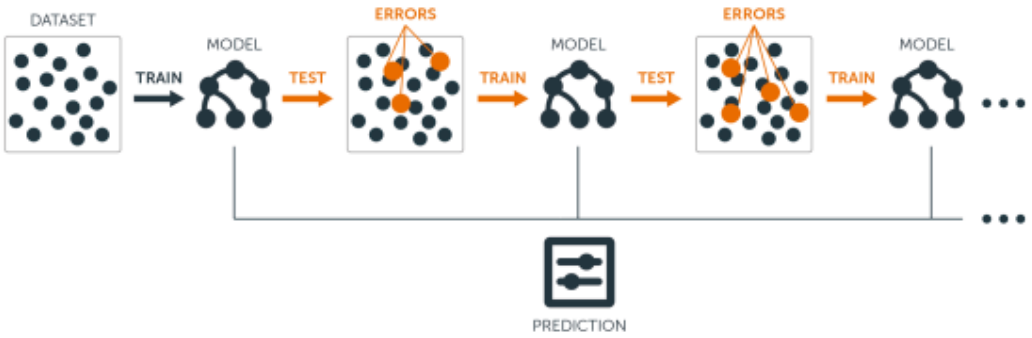


Figura 11. Esquema simplificado del funcionamiento del algoritmo de bosques aleatorios. Imagen obtenida de (https://blog.bigml.com/2017/03/14/introduction-to-boosted-trees/)

El paquete caret permite variar los valores de 4 hiperparámetros:

- n.trees: número de árboles ajustados
- interaction.depth: profundidad máxima de cada árbol
- shrinkage: tasa de aprendizaje. Reduce el impacto de cada árbol adicional en el modelo final. Canto mayor su valor menor es el cambio provocado por cada nuevo árbol en el modelo final.
- n.minobsinnode: número mínimo de observaciones en los nodos terminales u hojas de cada árbol.

Seguimos el mismo proceso de selección de hiperparámetros que con los bosques aleatorios. También ajustamos los modelos siguiendo las mismas estrategias de reducción de covariables.

Respecto a la importancia relativa de las variables, esta se va actualizando secuencialmente a medida que se ajustan nuevos árboles, por lo que el resultado final debería ser diferente al obtenido por un árbol único o un modelo de bosque aleatorio. Representamos las importancias un gráfico de barras para poder apreciar cuan diferentes son los tres modelos a pesar de sus similitudes.

2.2.7. Máquinas de vector soporte (SVM)

El último tipo de modelo ajustado sobre los datos son las Máquinas de Vector Soporte (SVM, del inglés *Support Vector Machine*). Estas se basan en el concepto del clasificador de margen máximo, que a su vez se basa en el concepto de hiperplano.

En un espacio p-dimensional, un hiperplano se define como un subespacio plano de dimensiones p-1. Cuando tenemos un conjunto de n observaciones con p predictores y una variable respuesta binaria, es posible utilizar hiperplanos para construir un clasificador que permita predecir a qué grupo pertenece una observación según sus predictores. Si las observaciones pueden ser perfectamente separadas en las dos clases mediante un hiperplano, el clasificador más sencillo es asignar cada observación a una clase según el lado del hiperplano en el que se encuentre. Esto se

conoce como hiperplano de margen máximo, llamado así porque maximiza el espacio entre las observaciones más cercanas al hiperplano y el propio hiperplano.

Sin embargo, en la mayoría de los casos reales, los datos no se pueden separar perfectamente mediante un hiperplano, lo que significa que no existe un hiperplano de separación de clases. En estas situaciones, se puede extender el concepto de hiperplano de margen máximo para obtener un hiperplano que casi separe las clases, permitiendo algunos errores. A este tipo de hiperplano se le conoce como Clasificador de Vectores de Soporte o Margen Suave.

El Clasificador de Vectores de Soporte o Margen Suave busca encontrar un hiperplano que maximice el margen entre las observaciones de las clases y, al mismo tiempo, permita un margen de error para clasificar correctamente las observaciones que no se pueden separar de manera perfecta. Esto se consigue mediante el uso de un hiperparámetro de complejidad C. Cuando C es infinito, significa que no se permite ningún error de clasificación, y por tanto es equivalente al hiperplano de margen máximo. Cuanto más se aproxima a 0, menos penaliza los errores y más observaciones pueden estar en el lado incorrecto del margen o incluso del hiperplano (Amat, 2017).

El Casificador de Vectores de Soporte consigue buenos resultados únicamente cuando el límite de separación entre clases es aproximadamente lineal. Para hacer frente a problemas donde el límite de separación no es lineal, se utilizan estrategias que consisten expandir las dimensiones del espacio. Las Máquinas de Vector Soporte (SVM) son una extensión del Clasificador de Vectores Soporte obtenida al aumentar la dimensión de los datos. Para ello se utilizan kernels, que son funciones que devuelven el resultado del producto escalar entre dos vectores, realizado en un nuevo espacio dimensional distinto al espacio original en el que se encuentran los vectores. La ecuación matemática mediante la cual se resuelve el problema de optimización de los Clasificadores de Vectores Soporte contiene un producto escalar. Si se reemplaza este producto escalar por un kernel, se obtienen los vectores soporte y el hiperplano en la dimensión del kernel. Esto se conoce como truco del kernel, y permite obtener resultados en cualquier dimensión con solo una pequeña modificación del problema original (Gareth et al., 2013).

Algunos de los kernels más utilizados son el lineal, el polinómico y el radial (figura 12). Al utilizarse un kernel lineal, la máquina de vector soporte obtenida es equivalente a clasificador de vectores soporte.

$$K(x, x') = x \cdot x'$$

En el caso de los vectores polinómicos, si el grado es 1 y la escala es 0, el resultado es un kernel lineal. Si aumentamos el grado, el límite de decisión deja de ser lineal.

$$K(x, x') = (x \cdot x' + c)^d$$

En los kernels radiales, se utiliza el parámetro γ para controlar la flexibilidad del modelo. Cuando γ se aproxima a 0 es equivalente a un kernel lineal.

$$K(x, x') = \exp\left(-\gamma||x - x'||^2\right)$$

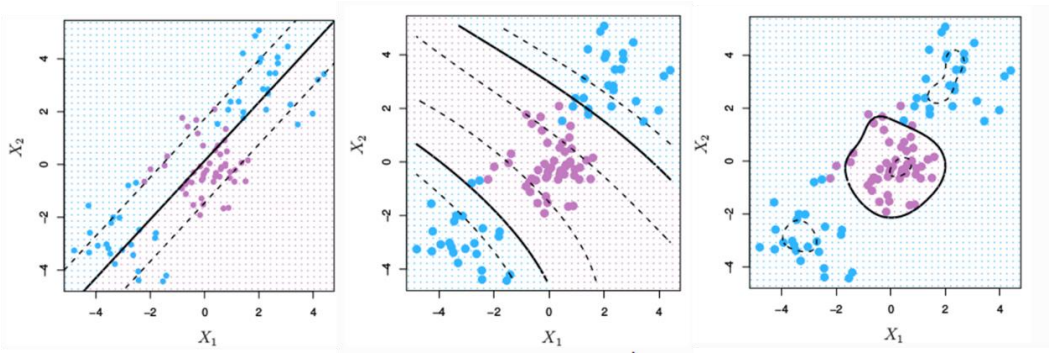


Figura 12. De izquierda a derecha, límite de decisión de una SVM con kernel lineal, polinómico y radial, respectivamente. Imagen obtenida de Gareth et al., 2013.

El paquete kernlab nos permite ajustar modelos de máquinas de vector soporte con kernels de tipo lineal, polinomial y radial, variando el hiperparámetro C en todos ellos, la escala y el grado en el kernel polinomial, y el valor de γ en el kernel radial. Ajustamos SVMs de los 3 tipos siguiendo la estrategia de selección de hiperparámetros ya mencionada anteriormente. Se ajustan modelos con todas las covariables y eliminando variables muy correlacionadas.

Street et al. (2000) comentan en su artículo que, tras una búsqueda global a través de todas las covariables, su mejor modelo ajustado es un único hiperplano que separa basándose en 3 covariables: el área más extrema (área_worst), la suavidad más extrema (smoothness_worst) y la textura media (textura_mean). Ajustamos una máquina de vector soporte con estas covariables para tratar de obtener un hiperplano lo más cercano posible al mencionado en el artículo, y poder compararlo así con el resto de los modelos ajustados, ya que en el artículo original, aunque se menciona que se utiliza validación cruzada para estimar la exactitud del modelo, no se menciona el número de particiones ni los individuos presentes en cada partición, por lo que la estimación del 97.5% de exactitud en su modelo no se puede comparar de forma con nuestros modelos. Probamos con un kernel lineal, uno polinómico y otro radial para ver qué hiperplano da mejor resultado.

2.2.8. Límite de decisión de los modelos ajustados

Con el objetivo de observar cómo los modelos ajustados separan nuevos individuos, vamos a dibujar el límite de decisión de cada uno de ellos sobre las dos primeras componentes principales. Para ello, se representa sobre las dos primeras componentes principales una rejilla de 150 puntos de ancho por 100 puntos de alto, abarcando un rango desde el mínimo hasta el máximo valor de estas componentes para nuestro banco de datos. Después se realiza el proceso inverso a la PCA para reconstruir las variables originales a partir de los valores de estas dos componentes principales. Una vez se tiene la reconstrucción de las variables originales para todos los puntos de la rejilla, se utiliza uno de los modelos ajustados para predecir la clase

de cada punto. Finalmente, se representan todos los puntos en un gráfico asignando un color diferente dependiendo de la predicción del modelo. Se representan también los puntos originales del banco de datos para observar qué individuos han quedado bien o mal clasificados.

Se debe tener en cuenta que al reconstruir las covariables originales a partir de únicamente dos componentes principales se pierde algo de información, por lo que la reconstrucción, y por ende el límite de decisión dibujado, serán una aproximación del límite real. Sin embargo, tras probar a reconstruir el banco de datos original a partir de las dos primeras componentes se obtuvo un banco de datos muy similar al original, por lo que consideramos que el límite obtenido es lo suficientemente informativo como para hacernos una idea del ajuste del modelo sobre los datos.

2.3. Paquetes utilizados

Una herramienta indispensable para este trabajo ha sido el paquete *caret*, Este paquete ha sido utilizado para realizar la partición de los datos y controlar de manera más cómoda los hiperparámetros de los modelos. También permite elegir el método de validación del modelo, así como la métrica a utilizar para evaluarlo. Caret integra dentro de sí una gran cantidad de paquetes de regresión y clasificación, proporcionando una capa de abstracción que simplifica el proceso de ajuste, la selección de hiperparámetros y la estimación de múltiples métricas de evaluación. Gracias a su uso se han ahorrado horas de trabajo y se ha mejorado la reproducibilidad de los análisis.

El resto de los paquetes empleados en el trabajo, así como sus respectivas versiones, están representados en la tabla 1, junto con un pequeño resumen de su funcionalidad.

Tabla 1. Versión y funcionalidad de los paquetes empleados en el trabajo, ordenados alfabéticamente.

Paquete	Versión	Funcionalidad
base	4.1.0	funciones de R base
class	7.3-19	modelos de clasificación (knn en este caso)
caret	6.0-90	validación cruzada y ajuste de hiperparámetros
dplyr	1.0.7	manipulación de tablas
gbm	2.1.8	modelos boosting trees
GGally	2.1.2	generar rejillas de gráficas
ggcorrplot	0.1.3	matriz de correlación en ggplot2
ggplot2	3.3.5	gráficas
glmnet	4.1-3	modelos penalizados/regularización
kernlab	0.9-29	máquinas de vector soporte
kknn	1.3.1	k-vecinos más cercanos (distancia Manhattan)

MASS	7.3-54	selección de variables mediante métodos stepwise
randomForest	4.6-14	modelos de bosques aleatorios
renv	0.17.3	manejo de dependencias
rpart	4.1.16	árboles de decisión
rpart.plot	3.1.0	gráficas de árboles de decisión
tibble	3.1.2	manipulación de tablas
tidyr	1.1.3	manipulación de tablas

3. RESULTADOS Y DISCUSIÓN

3.1. Análisis exploratorio

3.1.1. Variable respuesta

Como se muestra en la figura 13, el banco de datos contiene información de 212 tumores malignos y 357 benignos. El desbalance entre clases no es muy pronunciado, por lo que no debería de dar problemas a la hora de ajustar modelos predictivos. En caso de que la diferencia hubiese sido más pronunciada, se podría haber recurrido a técnicas de sobremuestreo para balancear las proporciones antes de ajustar modelos.

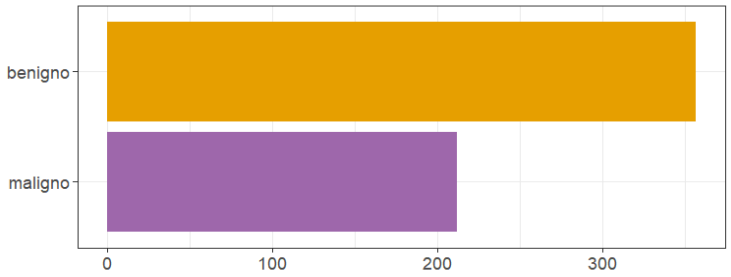


Figura 13. Gráfico de barras representando el número de individuos perteneciente a cada clase.

3.1.2. Relación entre covariables, entre sí y con la variable respuesta

Tanto las matrices de correlación como los diagramas de dispersión muestran la existencia de mucha información redundante en el banco de datos, habiendo un elevado número de covariables con una correlación del 90% o superior.

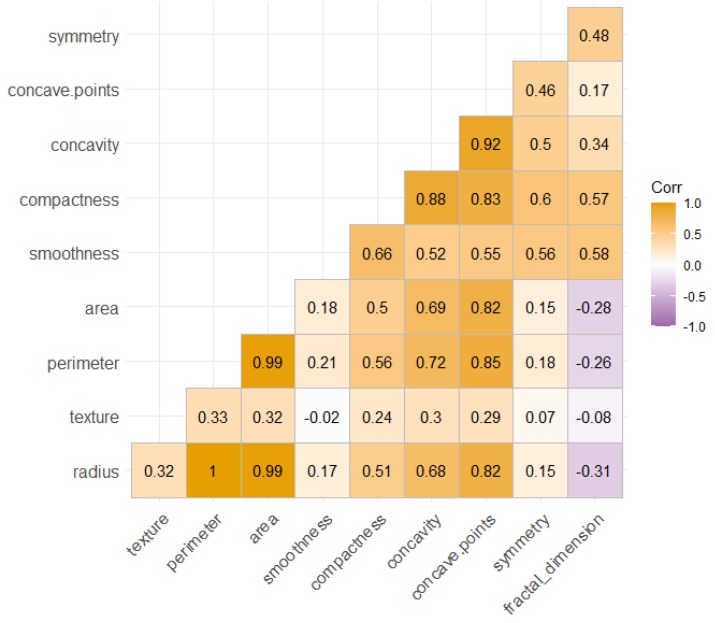


Figura 14. Matriz de correlación de las 10 primeras covariables, correspondiente a la media de las 10 medidas tomadas de los núcleos de cada muestra. Se puede acceder al resto de gráficas en el apartado “Gráficas y tablas extra” del Anexo.

Al representar en colores diferentes los tumores benignos y los malignos, los diagramas de dispersión muestran una clara diferencia entre grupos en los valores de la mayoría de covariables (Figura A2 del anexo). Al comparar la distribución de todas las covariables separando por clase esto se hace aún más evidente, poseyendo por lo general los tumores malignos medias más altas y campanas de distribución más aplanadas (figura 15).

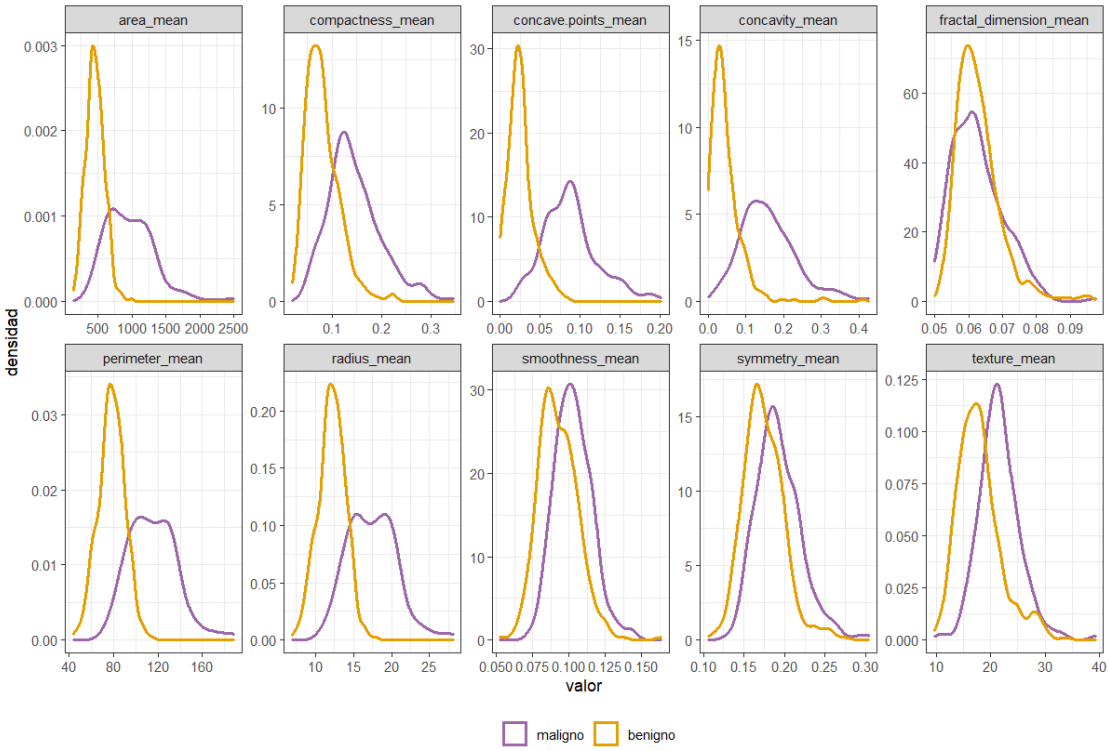


Figura 15. Gráfico de densidad de las primeras 10 covariables del banco de datos, separadas por diagnóstico.

Utilizando test-t, se compara la media de las covariables entre grupos benigno y maligno, dando lugar a 4 covariables cuya media no difiere de forma significativa entre grupos (nivel de confianza 95%). Estas variables son la dimensión fractal promedio (fractal_dimension_mean), el error estándar de la textura (texture_se), la suavidad (smoothness_se) y la simetría de los núcleos (symmetry_se), y no serán utilizadas para el ajuste de los posteriores modelos.

La primera conclusión a la que llegamos en base a lo observado hasta ahora es que por lo general en los aspirados de masas malignas los núcleos son más grandes e irregulares, y que las diferencias en características físicas entre los núcleos son mayores. Como es complicado analizar e interpretar 30 covariables por separado, vamos a tratar de analizar las primeras componentes principales.

3.1.3. Análisis de Componentes Principales

En la tabla 2 se puede apreciar que las dos primeras componentes principales acumulan el 63% de la varianza, por lo que estas son una buena fuente de información

que nos permitirá observar buena parte de las características del banco de datos en tan sólo dos dimensiones.

Tabla 2. Proporción de varianza explicada por las primeras 5 componentes principales

	Proporción de varianza explicada	Proporción de varianza acumulada
Componente 1	0.44	0.44
Componente 2	0.19	0.63
Componente 3	0.09	0.72
Componente 4	0.07	0.79
Componente 5	0.05	0.84

La primera componente principal podría explicarse como lo grande e irregular que es el núcleo de las células, siendo esta mayor cuanto más grandes sea la célula y mayor sea el número y la severidad de los puntos cóncavos en su superficie (tabla A1 en el anexo). En general, nuestra hipótesis será que las muestras con células cancerosas poseen valores más altos de esta componente que las que posean células sanas.

La segunda componente es más complicada de analizar. Las variables que influyen más positivamente son el radio, área y perímetro, por lo que aspirados con núcleos celulares grandes serán positivos para esta componente. Pero en este caso la dimensión fractal afecta de manera muy negativa, lo que nos dice que cuanto más compleja sea la forma del núcleo menor va a ser su valor de la segunda componente principal. Por otro lado, el error estándar de la suavidad, compacidad y dimensión fractal también afectan de manera negativa, por lo que en aspirados donde haya mucha variabilidad en la complejidad y en lo lisos y compactos que sean los núcleos, la segunda componente será más negativa.

Representando las proyecciones de todos los individuos sobre las primeras componentes principales, se aprecia a simple vista una clara separación entre masas malignas y benignas (figura 16). Se confirma además nuestra hipótesis de que las muestras con células malignas poseen valores más altos de la primera componente principal. Respecto a la segunda componente, ambas clases parecen poseer valores bastante centrales, sin desviarse en exceso del 0. Esto podría tener una explicación diferente para cada clase. En el caso de las muestras benignas, estas poseen valores centrales porque los núcleos celulares sanos poseen un tamaño promedio y no son excesivamente irregulares. En el caso de las muestras malignas, las covariables que afectan positivamente y las que afectan negativamente a esta componente podrían estar anulándose entre sí, ya que, como hemos comentado, un gran tamaño del núcleo, característica típica de células cancerosas, aumenta el valor de la segunda componente. En cambio, la alta variabilidad en la complejidad y rugosidad del núcleo, también propias de las células cancerosas, influyen de manera negativa en esta componente.

Se puede apreciar también que los puntos correspondientes a muestras benignas se encuentran más aglutinados entre sí en el plano generado por las dos primeras componentes, mientras que los de muestras malignas se encuentran más dispersos. Esto también puede ser un indicativo de que los núcleos de las células cancerosas son más irregulares y diferentes entre sí, y que los núcleos de células sanas son más uniformes, lo cual se corresponde con lo ya mencionado en la introducción (figura 3).

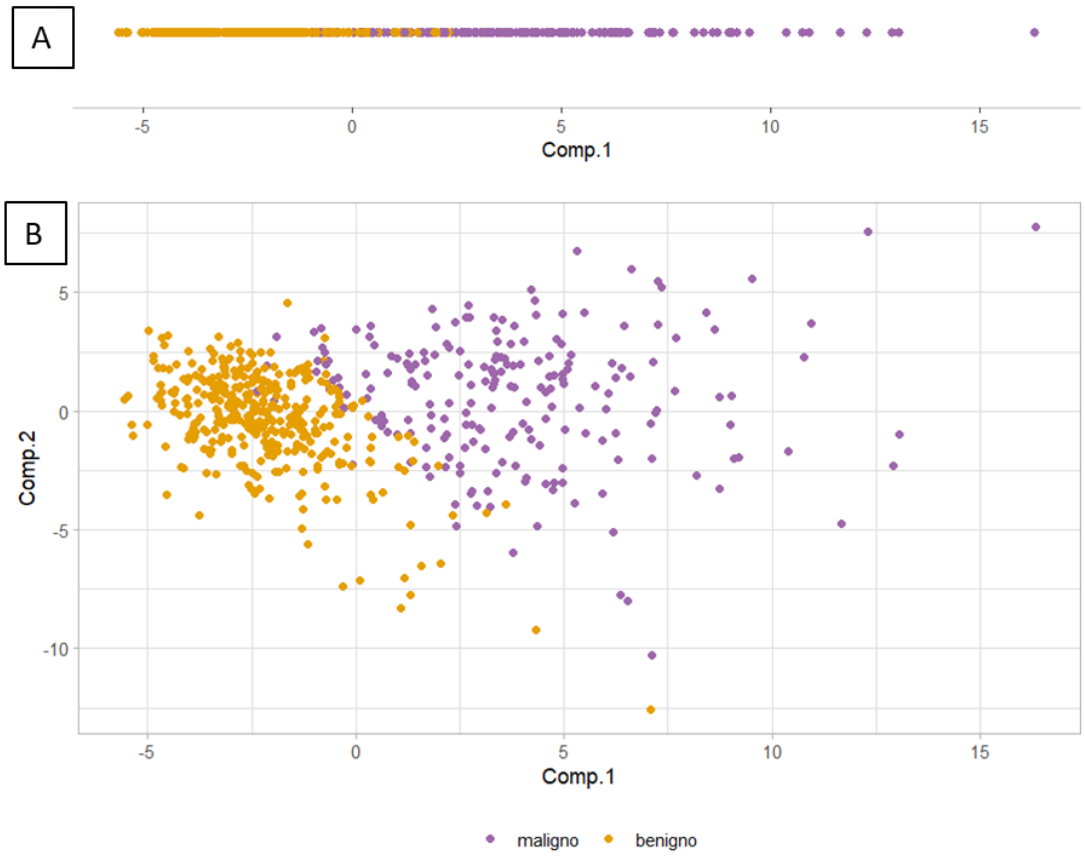


Figura 16. Proyección de los datos sobre: A) la primera componente principal. B) Las dos primeras componentes principales.

3.2. Regresión logística

En general, todos los modelos de regresión logística superaron el 90% de exactitud. El modelo con todas las covariables, que se tomó como base, logró una exactitud de 0,958.

En cuanto al uso de componentes principales como predictores en lugar de las covariables originales, estas demostraron ser una buena forma de reducir la dimensionalidad, consiguiéndose la mejor exactitud con tan sólo 9 componentes (figura 17). Al principio la exactitud aumenta a medida que se utilizan más componentes principales, pero a partir de las 10 componentes principales esta comienza a estancarse y a empeorar, probablemente debido a la redundancia entra predictores.

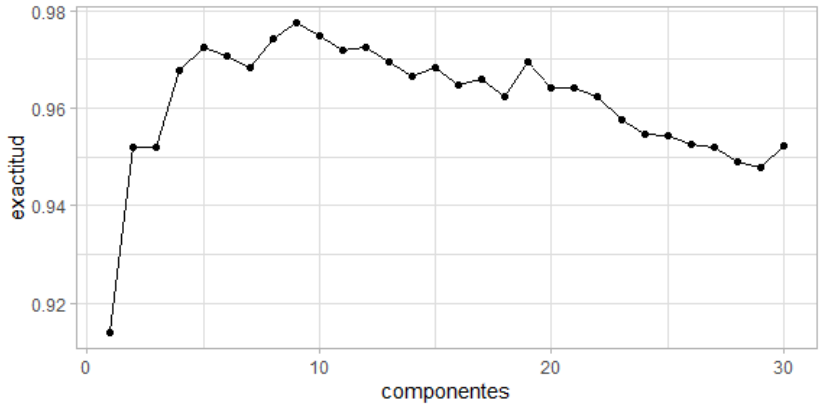


Figura 17. Exactitud de los modelos de regresión logística dependiendo del número de componentes principales utilizadas como predictores.

Las regresiones logísticas eliminando variables muy correlacionadas mejoraron ligeramente la exactitud de la regresión logística con todas las covariables. Utilizando un umbral de 0,95, se eliminan 6 covariables; utilizando un umbral de 0,9, se eliminan 10. La regresión con las covariables de mayor influencia en las primeras 9 componentes principales también dio lugar a una ligera mejora respecto al modelo completo, seleccionando 13 covariables. Estos subconjuntos de covariables se utilizaron para ajustar otro tipo de modelos más adelante.

En cuanto a las regresiones utilizando métodos *stepwise*, estas mantuvieron la misma exactitud que la regresión logística sobre todas las covariables, pero utilizando tan solo 11 covariables en la de tipo *forward* y 17 en la de tipo *backward*. Es lógico que la exactitud no mejorase, ya que este método utiliza el AIC como métrica para seleccionar o eliminar covariables, que es una medida de ajuste y no de capacidad predictiva. En cambio, los algoritmos *stepwise forward* y *backward* programados a mano para utilizar la exactitud como métrica de decisión sí que dieron lugar a una mejora notable en la exactitud, utilizando 18 y 4 covariables respectivamente. El modelo forward destaca especialmente, ya que se llega a un modelo en el que con tan sólo 4 covariables se obtiene una exactitud del 97,6%. Las covariables empleadas por el modelo fueron el perímetro y la suavidad más extremos, la textura promedio y el error estándar en la dimensión fractal. Dos de estas covariables coinciden con las seleccionadas como mejores en el trabajo de Street et al. (2000), y el área extrema, la única de las 3 covariables del trabajo de Street et al. que no está presente en el subconjunto de nuestro modelo, posee una correlación del 98% con el perímetro extremo, que sí está presente. Concluimos por tanto que nuestro algoritmo ha dado lugar a un subconjunto de covariables casi idéntico al del estudio original.

Se podría haber ahondado más en este subconjunto de covariables y utilizarlo para ajustar otros modelos, pero el algoritmo *stepwise* utilizando la exactitud como selector fue uno de los últimos modelos ajustados en el trabajo, y por tanto no ha habido tiempo de profundizar más en este subconjunto. En cuanto al uso del algoritmo *stepwise* para el ajuste de otro tipo de modelos, en la mayoría de los casos el coste computacional era muy elevado, principalmente debido a que la mayoría de los

modelos ajustaos realiza un ajuste de múltiples submodelos probando distintas combinaciones de hiperparámetros, por lo que sólo se aplica a la regresión logística y al análisis discriminante lineal.

De todos los modelos de regresión logística ajustados, los que mayor poder predictivo mostraron fueron los modelos con penalización. Tanto la penalización Ridge como la Lasso dieron lugar a modelos que superaban el 97% de exactitud, pero el mejor modelo se obtuvo utilizando la penalización Elastic Net, que permite una mezcla entre ambas. Aplicar este tipo de modelo sobre todas las covariables resultó en un 98.24% de exactitud, siendo el tercer mejor modelo de todos los ajustados en el trabajo, prácticamente a la par de los dos primeros. Este modelo seleccionó 24 covariables. En la tabla 3 se muestran todos los modelos de regresión logística ajustados junto a su exactitud, sensibilidad, especificidad y tiempo de ejecución. Para conocer el valor de los hiperparámetros de todos los modelos ajustados, véase la tabla A2 del Anexo.

Tabla 3. Resultados de todos los modelos de regresión logística ajustados.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)
logística completa	0,95773	0,95418	0,95976	1,48
logística 9 Componentes Principales	0,97765	0,96364	0,98595	44,84
logística correlación <90%	0,96593	0,95426	0,97283	1,09
logística correlación <95%	0,96241	0,95267	0,96815	1,11
logística covariables PCA	0,96472	0,94163	0,97844	1,50
logística step forward AIC	0,95884	0,93853	0,97095	31,49
logística step backward AIC	0,95892	0,95418	0,96167	79,64
logística step forward exactitud	0,97651	0,95736	0,98786	136,77
logística step backward exactitud	0,97537	0,96385	0,98222	449,71
logística Ridge	0,973	0,93391	0,99624	213,5
logística Ridge <90%	0,95488	0,88838	0,99442	220,59
logística Ridge <95%	0,96605	0,91364	0,99722	230,93
logística Lasso	0,97418	0,9544	0,98598	163,54
logística Lasso <90%	0,9777	0,95115	0,99347	153,18
logística Lasso <95%	0,97712	0,9544	0,99066	154,09
logística Elastic Net	0,9824	0,96999	0,98974	89,22
logística Elastic Net <90%	0,98064	0,95902	0,99347	83,76
logística Elastic Net <95%	0,97597	0,94495	0,99439	86,12

En la figura 18 se representa el límite de decisión del mejor modelo de regresión logística ajustado. Se puede apreciar que todo el plano queda separado por una única línea recta.

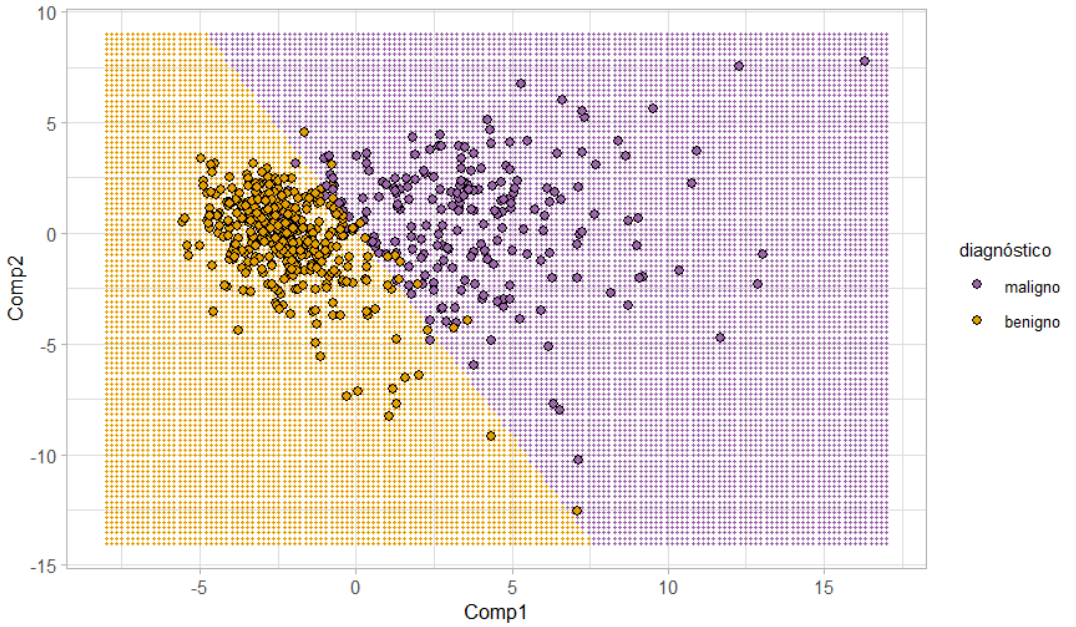


Figura 18. Proyección sobre las primeras dos componentes del límite de decisión del modelo de regresión logística con penalización Elastic Net utilizando todas las covariables. Se representan sobre la rejilla de puntos los individuos del banco de datos, coloreados según su diagnóstico observado.

3.3. Análisis discriminante lineal

En general, los modelos de análisis discriminante lineal son muy similares a los mencionados en regresión logística, aunque comparando el mismo enfoque en ambos modelos, el modelo de análisis discriminante es ligeramente inferior. La única excepción son los modelos ajustados mediante el algoritmo *stepwise* que utiliza la exactitud como criterio, que obtuvieron un poder predictivo prácticamente idéntico a los modelos de regresión logística que seguían el mismo enfoque.

En cuanto al análisis discriminante utilizando componentes principales como covariables, se encontró que el número óptimo de componentes principales eran 19, dando lugar a un modelo con una exactitud de 0,961 (figura 19).

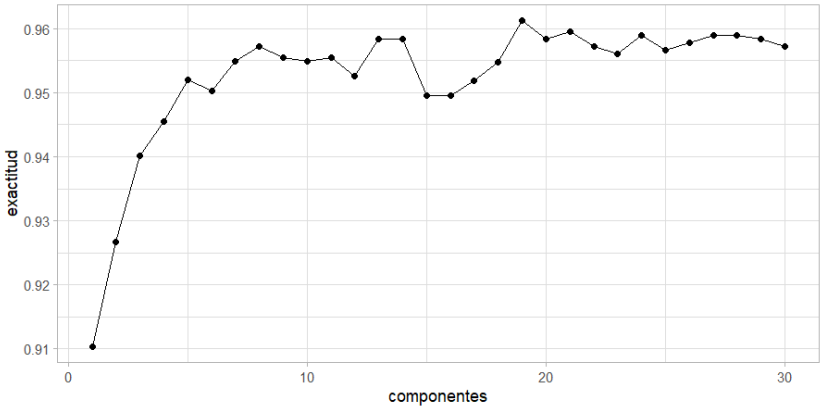


Figura 19. Exactitud de los modelos LDA dependiendo del número de componentes principales utilizadas como predictores.

En la tabla 4 se muestran todos los modelos de análisis discriminante ajustados, junto a su exactitud, sensibilidad, especificidad y tiempo de ejecución.

Tabla 4. Resultados de todos los modelos de análisis discriminante lineal ajustados.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)
LDA completo	0,95786	0,89646	0,99437	1,03
LDA 19 Componentes Principales	0,96124	0,90382	0,99534	32,43
LDA <90%	0,95657	0,88824	0,9972	0,78
LDA <95%	0,94728	0,85859	1	0,84
LDA covariables PCA	0,949	0,87071	0,99529	1,08
LDA step forward exactitud	0,97772	0,94502	0,9972	231,63
LDA step backward exactitud	0,97007	0,9246	0,9972	372,07

El mejor modelo ajustado es el stepwise forward, con una exactitud de 0,978. Este modelo posee tan sólo 9 covariables. Representamos su límite de decisión en la figura 20. Se puede observar en la gráfica que el modelo sacrifica sensibilidad para predecir correctamente casi todos los tumores benignos. Esto sucede con todos los modelos LDA, que poseen una sensibilidad notablemente inferior a la reportada en los modelos de regresión logística.

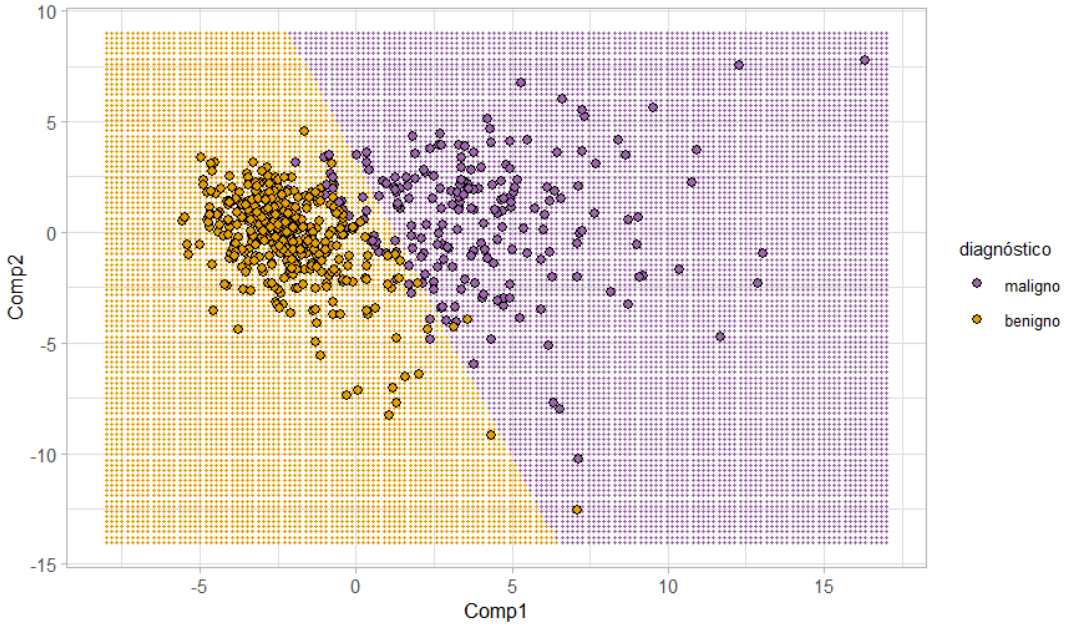


Figura 20. Proyección sobre las primeras dos componentes del límite de decisión del modelo LDA stepwise forward. Se representan sobre la rejilla de puntos los individuos del banco de datos, coloreados según su diagnóstico observado.

3.4. K-vecinos más cercanos

El uso de la distancia Manhattan o euclídea no tuvo una gran repercusión en la exactitud de los modelos KNN, obteniéndose modelos muy similares para ambos tipos de distancia (tabla 5). La eliminación de covariables muy correlacionadas afectó

negativamente a la capacidad predictiva de los modelos, siendo el mejor modelo un KNN sobre todas las covariables, con k = 3 y una exactitud de 0.97 (figura 21).

Tabla 5. Resultados de todos los modelos KNN ajustados.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)
KNN	0,97066	0,93867	0,98974	16,84
KNN <90%	0,95082	0,90267	0,97944	15,31
KNN <95%	0,96017	0,9215	0,9832	15,79
KNN Manhattan	0,96888	0,94488	0,98315	17,89
KNN Manhattan <90%	0,95953	0,91955	0,9832	15,14
KNN Manhattan <95%	0,96545	0,93095	0,98601	16,68

El paquete kknn, que utilizamos para ajustar modelo KNN con distancia Manhattan, permite especificar el número máximo de k que se quiere probar, y ajusta k modelos, pero tan solo guarda en memoria el que da lugar a la mejor exactitud, por lo que no nos permite representar una gráfica similar a la figura 21 utilizando la distancia Manhattan en lugar de la euclídea.

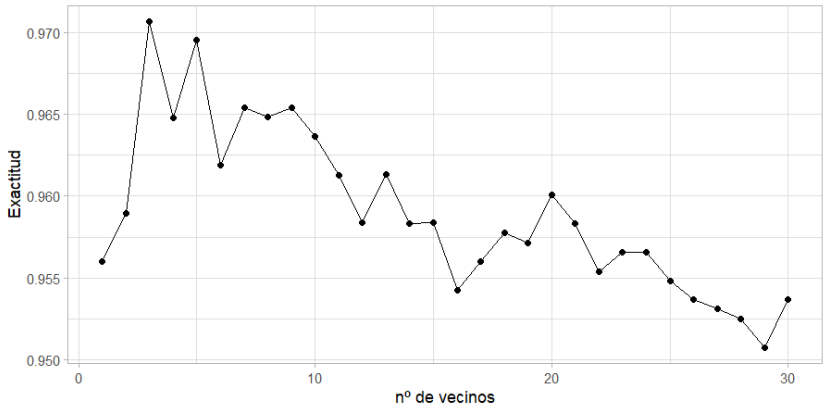


Figura 21. Exactitud de los modelos KNN con distancia euclídea dependiendo del número de vecinos, utilizando todas las covariables como dimensiones para el cálculo de la distancia.

Representamos en la figura 22 el límite de decisión del mejor modelo KNN ajustado. Este parece haber ajustado bien todos los casos claros de tumores benignos y malignos, clasificando incorrectamente tan sólo algunos de los casos más extremos.

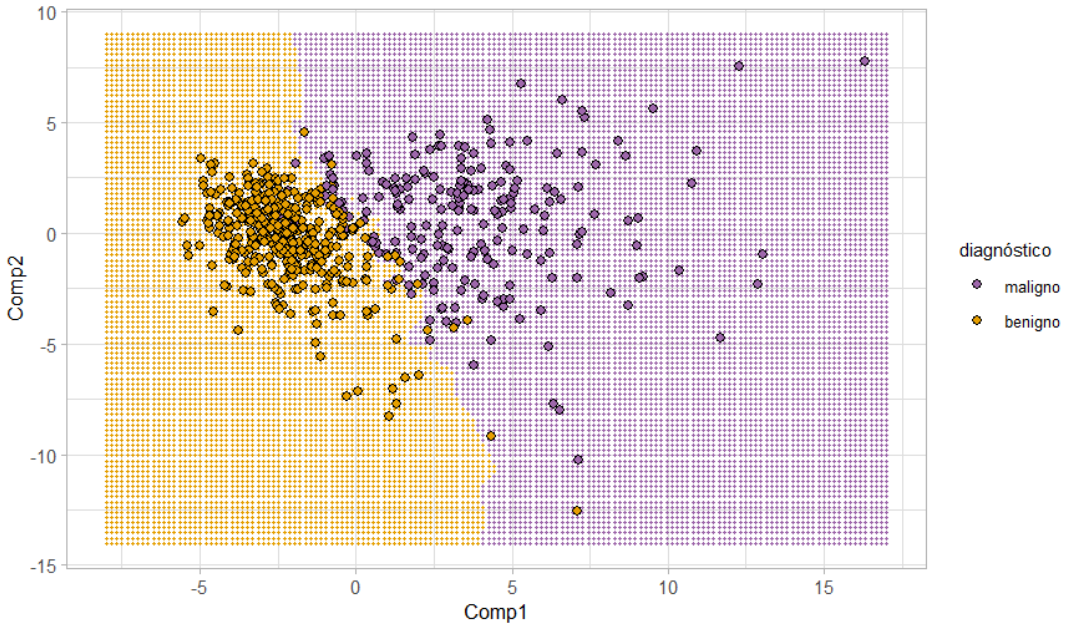


Figura 22. Proyección sobre las primeras dos componentes del límite de decisión del modelo KNN utilizando todas las covariables, empleando la distancia euclídea. Se representan sobre la rejilla de puntos los individuos del banco de datos, coloreados según su diagnóstico observado.

3.5. Árboles de decisión

El árbol de decisión ajustado por defecto y el ajustado probando diferentes valores del parámetro de complejidad (cp) dieron lugar al mismo árbol de decisión, con una exactitud de 0,92 (figura 23A). En general, los árboles de decisión dieron lugar a un poder predictivo menor que el resto de los modelos (tabla 6), probablemente debido a la tendencia al sobreajuste de este tipo de modelos, y a la multicolinealidad del banco de datos. Los dos tipos de árboles ajustados, seleccionando el mejor hiperparámetro de complejidad y de profundidad máxima respetivamente, dieron lugar a árboles de decisión prácticamente idénticos, siendo ligeramente más simple y con mayor exactitud el modelo que utiliza la profundidad máxima como hiperparámetro (figura 23B).

Tabla 6. Resultados de todos los modelos de árboles de decisión ajustados.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)
árbol decisión	0,92024	0,8728	0,94844	4,81
árbol decisión cp	0,92024	0,8728	0,94844	9,05
árbol decisión profundidad	0,92134	0,87727	0,94757	3,72
árbol decisión profundidad <90%	0,94304	0,91955	0,9569	3,48
árbol decisión profundidad <95%	0,93313	0,90404	0,95034	3,71
árboles aleatorios mejor árbol	0,94186	0,93059	0,94844	315,36

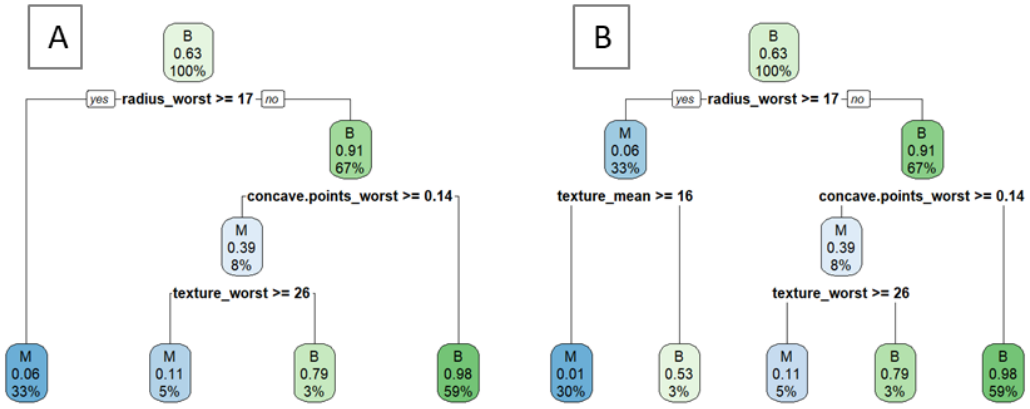


Figura 23. Árboles de decisión tomando como hiperparámetro A) la complejidad (cp = 0). B) la profundidad máxima (maxdepth = 3). Ajustados sobre todas las covariables.

La eliminación de covariables muy correlacionadas mejora la exactitud del modelo, siendo el mejor modelo ajustado el árbol de decisión ajustado sobre las covariables con una correlación menor al 90% (figura 24A). Este árbol separa utilizando tan sólo 3 covariables: el área extrema, el mayor número de puntos cóncavos, y la textura extrema. Observando los nodos del árbol, vemos una de las particiones, que separa según la textura más extrema, da lugar a una hoja con tan sólo el 1% de los datos, lo cual probablemente es un sobreajuste a los datos. Esta tendencia al sobreajuste parece estar presente en todos los árboles ajustados.

En cuanto a los árboles de decisión aleatorizados, el mejor árbol obtenido mediante este método es muy similar al obtenido eliminando covariables correlacionadas en más de un 90%, con la excepción de que separa en función del promedio de puntos cóncavos en lugar de en función del número máximo de estos (figura 24B).

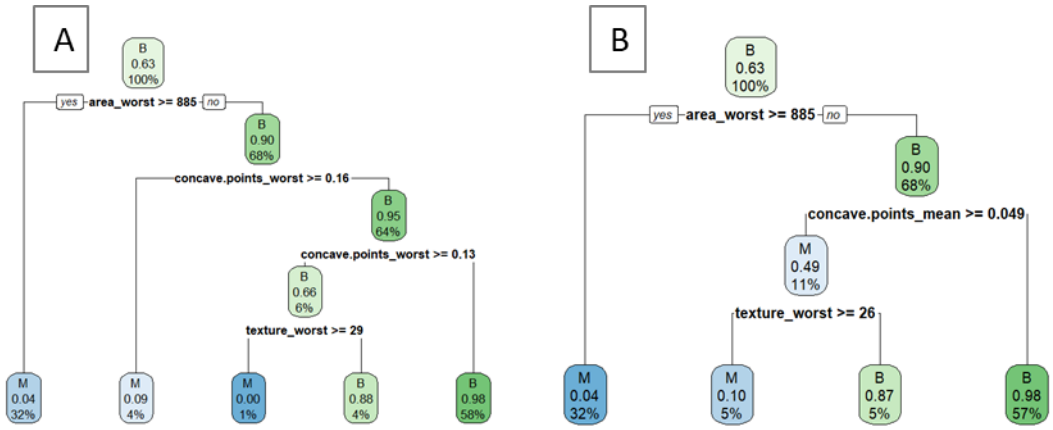


Figura 24. Mejores árboles de decisión ajustados A) con una profundidad máxima de 4 y ajustado eliminando covariables con una correlación superior a 0,9. B) con una profundidad máxima de 4, ajustado sobre 11 covariables seleccionadas al azar.

Los árboles de decisión determinan la mejor partición en función del decrecimiento del índice de Gini. De este modo se puede cuantificar la importancia de cada covariable estimada por el árbol de decisión. Representamos la importancia de cada covariable, estimada por el árbol de decisión con todas las covariables con una profundidad máxima de 3 (figura 25).

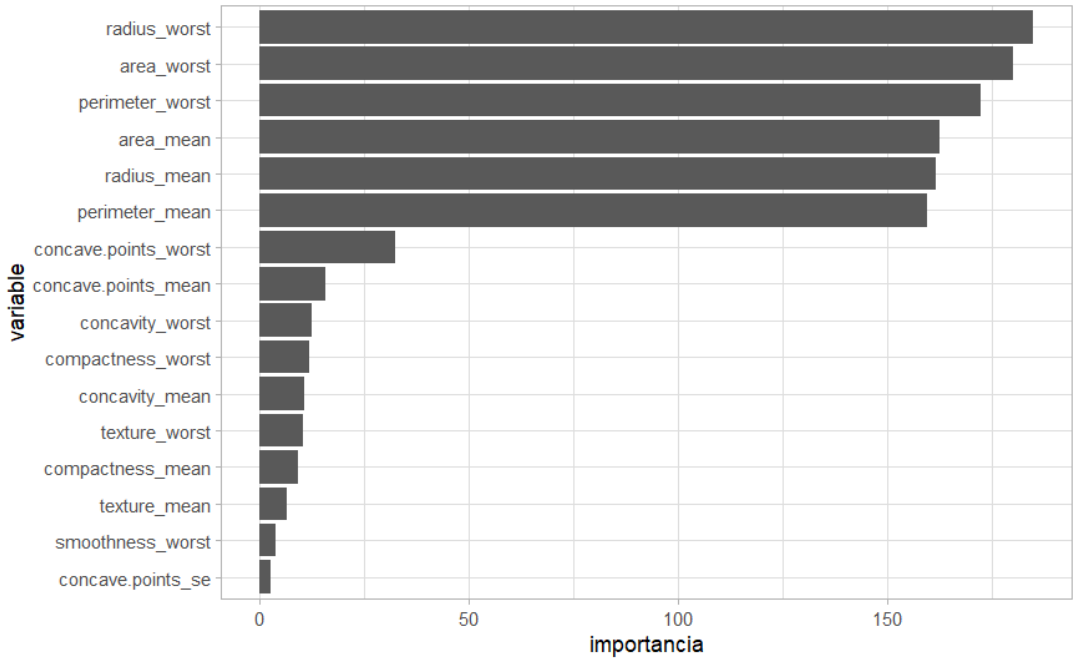


Figura 25. Importancia relativa de las covariables, según el árbol de decisión con todas las covariables con una profundidad máxima de 3. Se representan las 15 covariables más importantes.

Del gráfico se extrae que el tamaño del núcleo celular posee una gran importancia a la hora de realizar particiones, utilizándose siempre alguna de sus covariables representativas como separador. Esta desproporción en la importancia de las covariables, así como la redundancia de información en las covariables, puede estar afectando negativamente al modelo, dándonos a entender que quizás un único árbol de decisión no es capaz de abarcar toda la complejidad y ruido del banco de datos.

Como en los apartados anteriores, representamos el límite de decisión del mejor árbol ajustado (figura 26). Se puede observar cómo el modelo no termina de predecir bien del todo el límite real entre malignos y benignos.

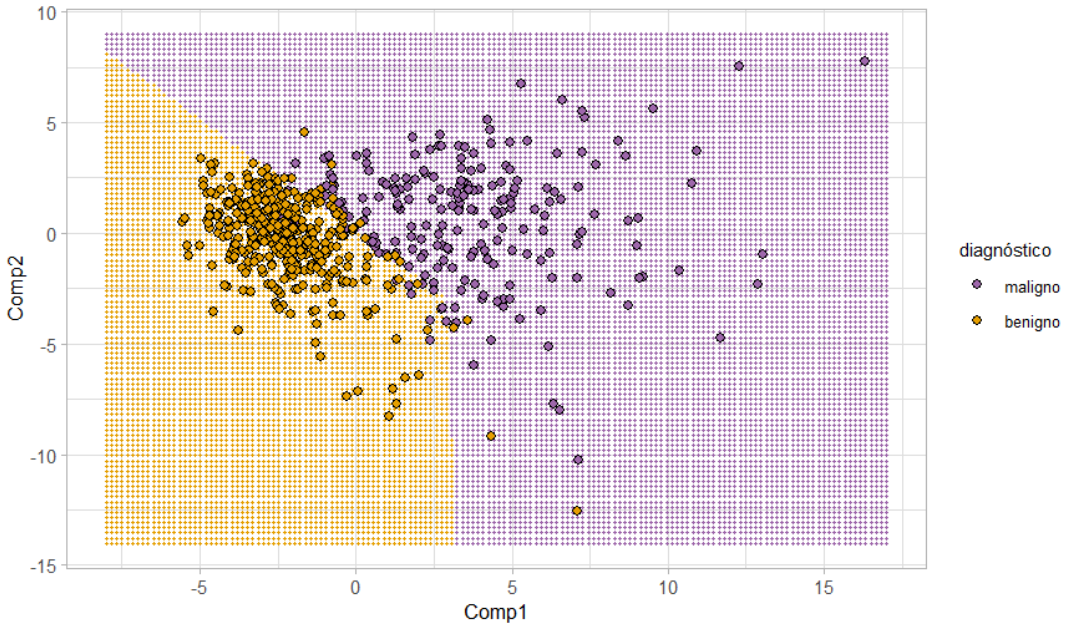


Figura 26. Proyección sobre las primeras dos componentes del límite de decisión del árbol de decisión utilizando una profundidad máxima de 4, seleccionando sólo aquellas covariables con una correlación menor al 90%. Se representan sobre la rejilla de puntos los individuos del banco de datos, coloreados según su diagnóstico observado.

3.6. Bosques aleatorios y *boosting trees*

Los ensamblajes de múltiples árboles de decisión dieron lugar a modelos con un poder predictivo notablemente superior al de los árboles de decisión individuales. Todos los modelos ajustados dieron lugar a una exactitud superior al 96%. En general, los modelos de bosques aleatorios fueron ligeramente inferiores a los modelos *boosting trees* (tabla 7). Esto puede deberse al hecho de que los modelos de bosques aleatorios ensamblan árboles aleatorios mediante el voto por mayoría, mientras que los *boosting trees* lo hacen de manera secuencial, tratando de corregir los errores de los árboles anteriores. Se observa también que los tiempos de ejecución de este tipo de modelos son considerablemente más altos que los del resto de modelos ajustados

Tabla 7. Resultados de todos los ensamblajes de árboles de decisión ajustados.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)
random forest	0,96477	0,94632	0,97563	180,97
random forest <90%	0,96595	0,93088	0,98688	208,33
random forest <95%	0,96946	0,94473	0,98413	196,81
random forest covariables PCA	0,95959	0,92482	0,98042	124,53
boosting trees	0,97824	0,96068	0,98873	451,67
boosting trees <90%	0,97647	0,95267	0,99061	353,05
boosting trees <95%	0,98295	0,96378	0,99434	421,51
boosting trees covariables PCA	0,97127	0,94661	0,98603	364.79

Como se aprecia en las tablas de la 3 a la 7, la mayoría de los modelos ajustados poseen una especificidad notablemente superior a la sensibilidad. Esto probablemente se debe a la decisión de utilizar la exactitud para seleccionar el mejor submodelo en un banco de datos que no está completamente balanceado, ya que de este modo se favorecen modelos que predicen mejor la clase mayoritaria que la minoritaria. En este caso esta característica podría ser algo deseable, ya que si la especificidad es muy cercana al 100% nos aseguramos de no realizar diagnósticos erróneos de cáncer de mama, manteniendo aun así una sensibilidad superior al 95% en los mejores modelos ajustados. Aun así, visto en retrospectiva tal vez habría sido más acertado utilizar otras métricas, como la exactitud balanceada, que es la media entre la sensibilidad y la especificidad. De este modo hubiéramos obtenido modelos con un equilibrio entre falsos positivos y falsos negativos. Lo bueno de haber ajustado los modelos con el paquete *caret* es que quedan guardados en el objeto del modelo todas las combinaciones de hiperparámetros probadas, con las métricas correspondientes al submodelo de cada combinación, por lo que si se quisiera se podrían reajustar todos los modelos para maximizar otra métrica de manera más o menos trivial.

Volviendo a los ensamblajes de árboles de decisión, tanto para los bosques aleatorios como para los *boosting trees* eliminar las variables correlacionadas en más de un 95% dio lugar a la mejor exactitud. El mejor modelo de bosques aleatorios dio lugar a una exactitud del 97%. La combinación óptima de hiperparámetros de este modelo fueron 500 árboles, 10 covariables seleccionadas al azar en cada partición, y un tamaño mínimo de nodo de 2 individuos (figura 27).

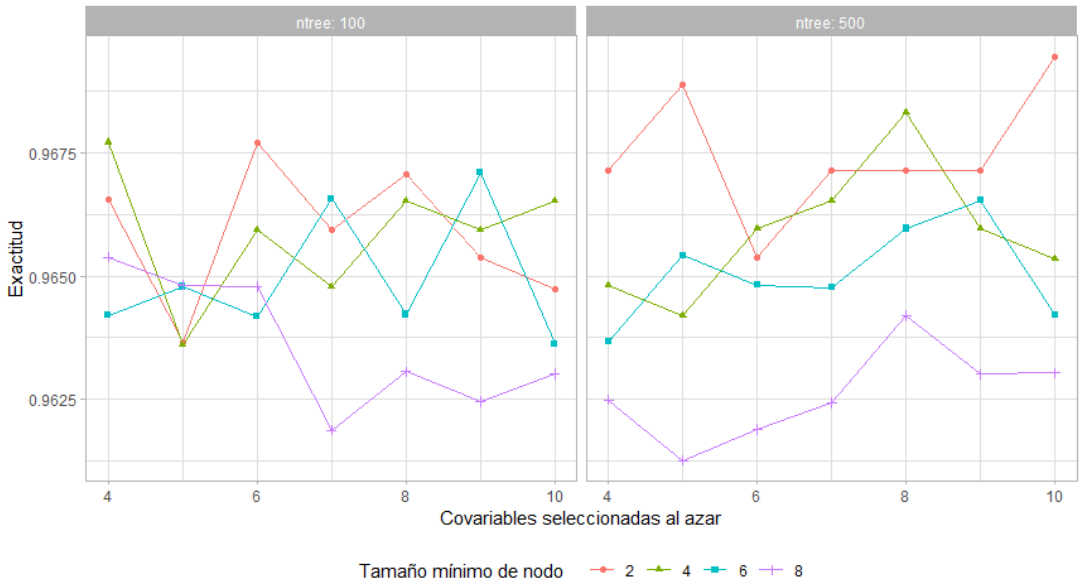


Figura 27. Exactitud de todos los submodelos ajustados buscando la mejor combinación de hiperparámetros para el modelo de bosques aleatorios eliminando covariables correlacionadas en más de un 95%.

El modelo *boosting trees* fue el modelo con la segunda mejor exactitud en todo el trabajo, con un 98,3%. La combinación óptima de hiperparámetros de este modelo fueron una tasa de aprendizaje (shrinkage) de 0.2, una profundidad máxima (interaction.depth) de 2, 1000 árboles ajustados, y un tamaño mínimo de nodo de 10 individuos (figura 28).

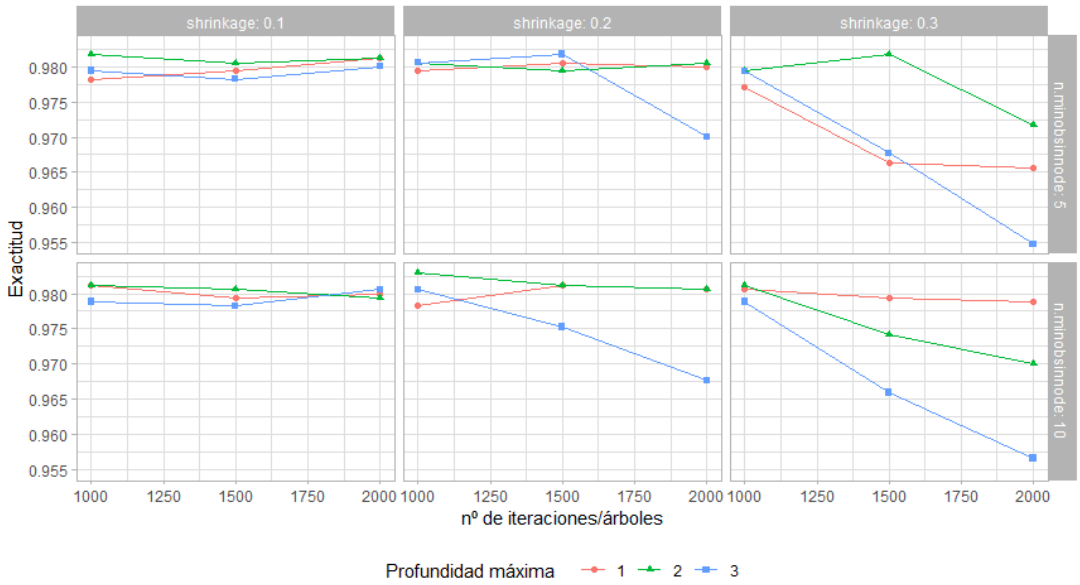


Figura 28. Exactitud de todos los submodelos ajustados buscando la mejor combinación de hiperparámetros para el modelo *boosting trees* eliminando covariables correlacionadas en más de un 95%.

Proyectando el límite de decisión del mejor modelo sobre las dos primeras componentes principales (figura 29), este parece similar al del modelo KNN de la figura 22, aunque clasifica de forma diferente puntos cuyas covariables dan lugar a una segunda componente principal elevada. Haría falta obtener nuevos datos que cubrieran dicha área para saber cuál de los dos modelos es más acertado en esa zona.

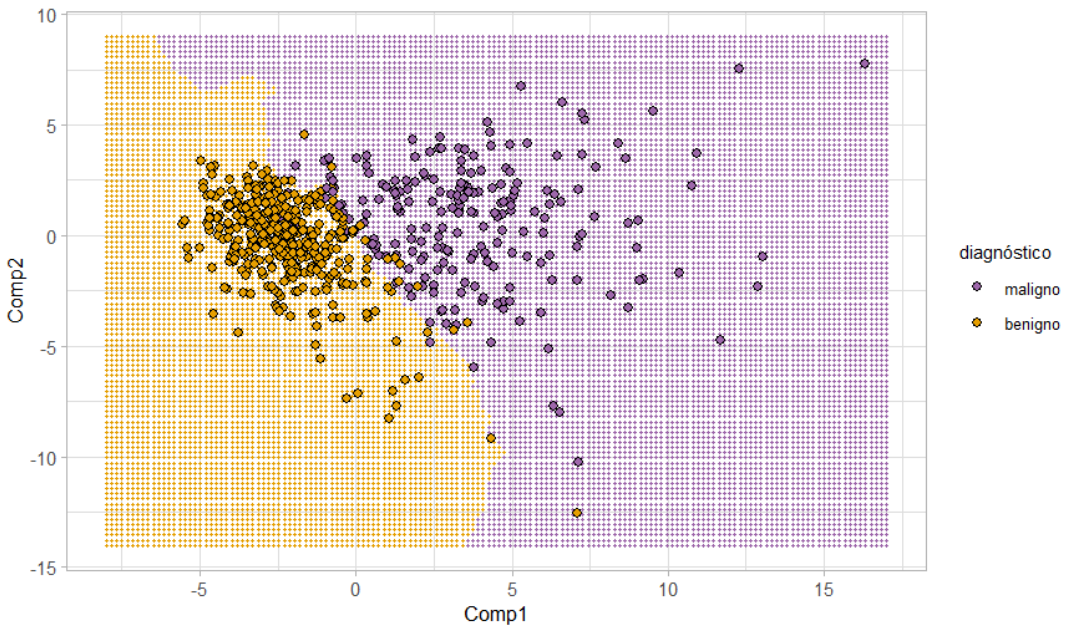


Figura 29. Proyección sobre las primeras dos componentes del límite de decisión del modelo *boosting trees* utilizando una tasa de aprendizaje de 0.2, una profundidad máxima de 2, 1000 árboles, y un tamaño mínimo de nodo de 10 individuos; seleccionando sólo aquellas covariables con una correlación menor al 95%. Se representan sobre la rejilla de puntos los individuos del banco de datos, coloreados según su diagnóstico observado.

Como se comenta en el apartado 2.2.6., las estimaciones de la importancia de las covariables deberían ser distintas entre los modelos de un único árbol de decisión, los bosques aleatorios y los *boosting trees*, ya que aunque la importancia se estime como la reducción del índice de Gini, en los bosques aleatorios esta se calculará como un promedio de la estimación dada por todos los árboles ajustados con covariables seleccionadas al azar; y en los *boosting trees* la importancia de las covariables se actualiza de forma secuencial a medida que se van ajustando nuevos árboles. Las figuras 30 y 31 confirman esta hipótesis, eliminando la redundancia de covariables relacionadas con el tamaño nuclear que veíamos en los árboles de decisión individuales. Esta eliminación de redundancia podría ser parte de la explicación de por qué mejoran la exactitud con respecto a los árboles individuales. Tanto los bosques aleatorios como los *boosting trees* dan una gran importancia a los puntos cóncavos y al área del núcleo, y en general asignan a una importancia bastante similar a las covariables.

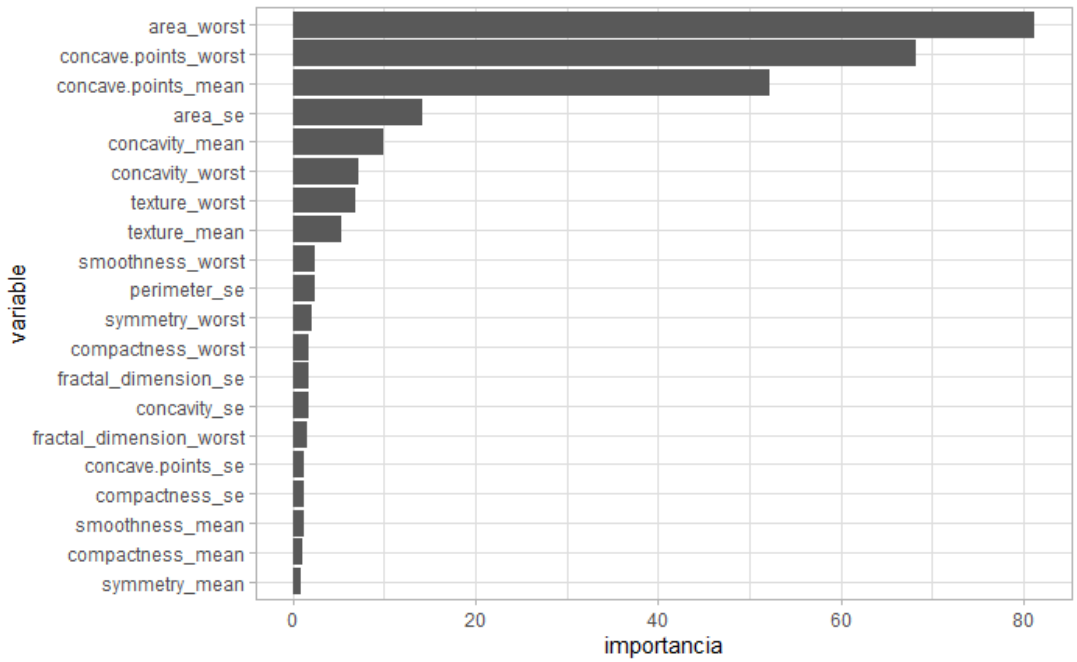


Figura 30. Importancia relativa de las covariables, según el mejor modelo de bosque aleatorio. Se representan las 15 covariables más importantes.

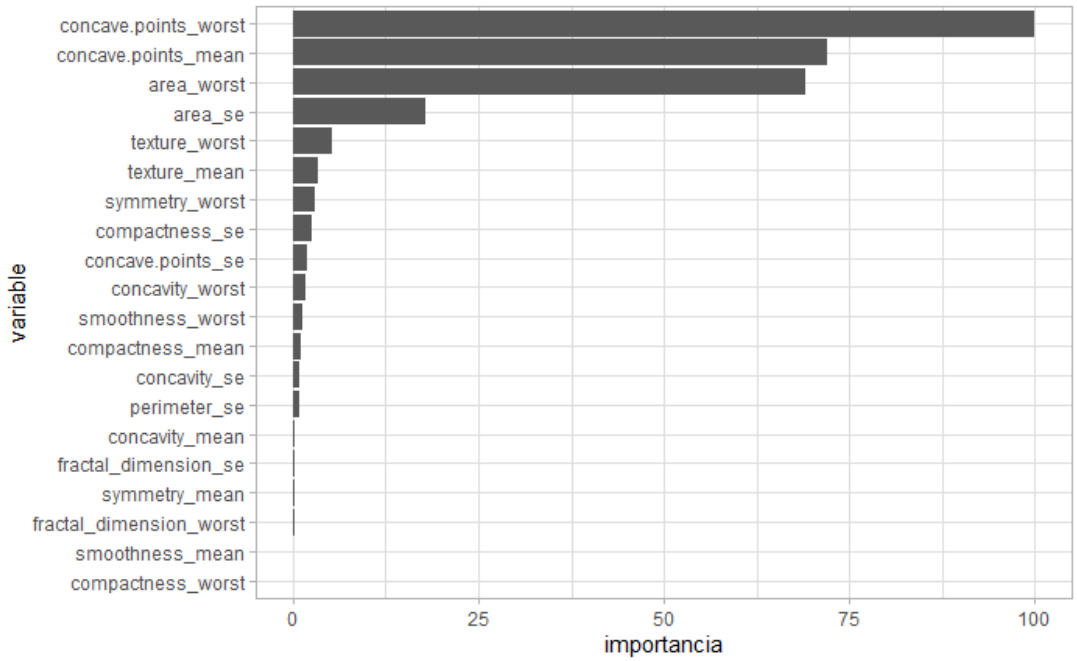


Figura 31. Importancia relativa de las covariables, según el mejor modelo *boosting trees*. Se representan las 15 covariables más importantes.

3.7. Máquinas de vector soporte

Las máquinas de vector soporte también dieron lugar a algunos de los modelos más precisos de todos los ajustados, probablemente debido a que los puntos del banco de datos ofrecen facilidad de separación mediante hiperplanos (figura A2 del anexo).

Prácticamente todas las máquinas de vector soporte, independientemente del kernel utilizado, superaron el 97% de exactitud (tabla 8). La máquina de vector soporte con un kernel radial ajustada sobre todas las covariables dio lugar a la mejor exactitud de todos los modelos ajustados en el trabajo, con un 98,3%. El valor del parámetro γ de este modelo fue 0,013, un número muy cercano a 0, indicándonos que el kernel no dista mucho del lineal. En el caso de las SVM polinómicas el grado óptimo seleccionado fue bajo, entre 2 y 3, por lo suponemos que el hiperplano de separación final fue también cercano al obtenido por las SVM lineales (para observar los hiperparámetros seleccionados en cada modelo, véase tabla A2 del anexo).

En cuanto a las SVM ajustadas con las 3 covariables seleccionadas como mejores por Street et al. (2000), estas dan lugar a exactitudes muy cercanas a las obtenidas con todas las covariables. Las SVM polinómica y radial reportaron una exactitud del 97,4%, prácticamente idéntica a la del hiperplano propuesto por Street et al. (2000) en su artículo, lo cual da fe de la reproducibilidad de su trabajo.

Estos modelos de 3 covariables, junto a la regresión logística obtenida mediante el algoritmo *stepwise* forward, con 4 covariables, nos indican que el banco de datos contiene mucha información redundante que apenas mejora la capacidad predictiva de los modelos.

Tabla 8. Resultados de todos los modelos SVM ajustados.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)
SVM lineal	0,97769	0,95584	0,99066	54,19
SVM lineal <95%	0,97596	0,94488	0,99442	61,39
SVM lineal <90%	0,97539	0,94336	0,99442	59,79
SVM polinómica	0,97999	0,95743	0,99341	436,73
SVM polinómica <95%	0,97709	0,95108	0,99254	427,51
SVM polinómica <90%	0,97948	0,9544	0,99442	336,84
SVM radial	0,98352	0,9684	0,99251	102,05
SVM radial <90%	0,97885	0,95281	0,99437	77,14
SVM radial <95%	0,98063	0,95584	0,99534	74,14
SVM lineal covariables artículo	0,9671	0,93059	0,98881	107,97
SVM polinómica covariables artículo	0,97416	0,94949	0,98881	645,87
SVM radial covariables artículo	0,97417	0,95108	0,98788	193,62

Al representar la proyección del límite de decisión de la SVM radial sobre las dos primeras componentes principales, observamos un límite de decisión muy similar al obtenido por el modelo *boosting trees*, con la misma curvatura en la zona inferior de la segunda componente principal.

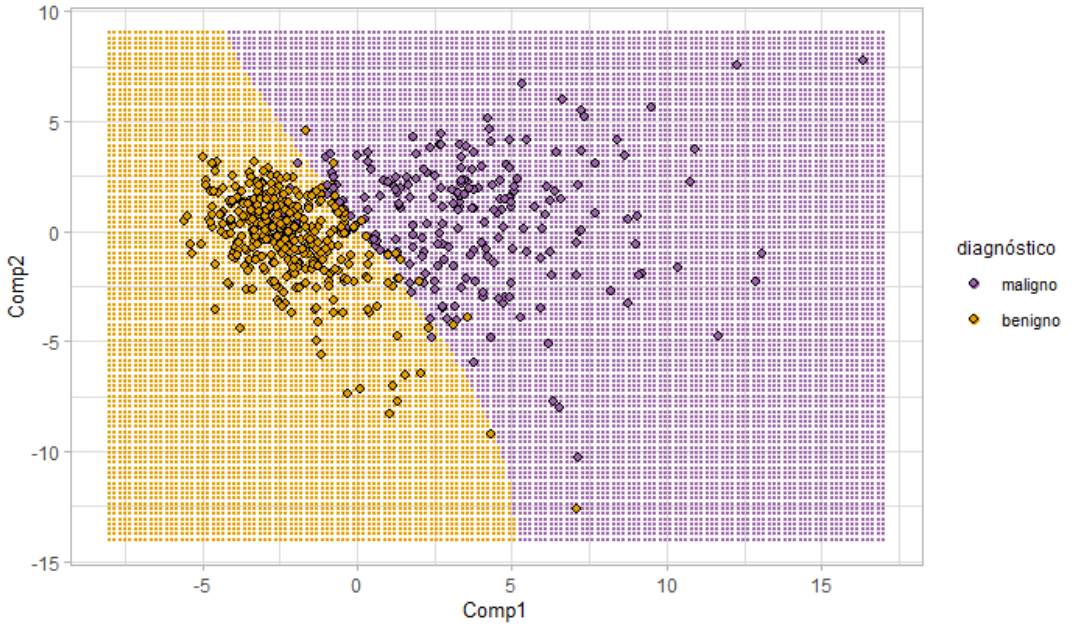


Figura 24. Límite de decisión de la SVM radial utilizando parámetro de complejidad de 0,452, una escala de 0,112, y un grado de polinomio de 2; ajustado sobre todas las covariables.

4. CONCLUSIONES Y PERSPECTIVAS DE FUTURO

De todos los modelos ajustados, nos quedamos como mejores modelos la regresión logística con penalización *Elastic Net* y la máquina de vector soporte radial ajustadas sobre todas las covariables, ya que son los modelos que mejor exactitud ofrecen junto al modelo *boosting trees*, pero con un mayor equilibrio entre sensibilidad y especificidad. Se eligen estos modelos también por su simplicidad, ya que son mucho más fáciles de interpretar que los 1000 árboles de decisión ajustados secuencialmente en el modelo *boosting trees*. Todos ellos superan el umbral del 97.5% de exactitud reportado por Street et al., aunque sólo sea por un pequeño margen.

Si el objetivo fuese obtener un modelo lo más sencillo posible que tuviera una buena capacidad predictiva, sin duda los candidatos principales serían la regresión logística de 4 covariables obtenida mediante el algoritmo *stepwise* y las máquinas de vector soporte que generaron hiperplanos basados en únicamente las 3 covariables seleccionadas como mejores por Street et al. (2000). Ambos modelos superaron el 97% de exactitud utilizando un subconjunto de covariables muy reducido. Sin embargo, el número de covariables que tenga el modelo final no es tan importante, ya que en un escenario real las mediciones se harían de forma automática utilizando herramientas de segmentación que calculan de forma automática las características físicas de los núcleos celulares.

Se podrían haber ajustado muchos más modelos, desde nuevos tipos de modelos a combinaciones de los ya ajustados. Por ejemplo, podría haberse probado a seleccionar el subconjunto obtenido por *Elastic Net*, o el subconjunto obtenido por el método stepwise que maximiza la exactitud, y emplear estos subconjuntos para ajustar otro tipo de modelos, como una SVM. También se podría haber probado otros enfoques y clasificadores, como diferentes arquitecturas de redes neuronales, pero el tiempo es limitado y ya se han conseguido diagnósticos muy precisos con los modelos ajustados, por lo que se da por bueno lo obtenido hasta ahora.

Es importante mencionar que, gracias a la drástica mejora en los procesadores para uso personal, se está dando una democratización de métodos estadísticos avanzados y costosos computacionalmente, quedando al alcance de contribuidores individuales la creación de herramientas de código libre para la solución de todo tipo de problemas. En este caso concreto, al trabajar con un banco de datos lo suficientemente grande pero no excesivamente grande, se ha podido experimentar con muchos tipos de modelos diferentes utilizando únicamente un ordenador portátil, siendo esta una experiencia de aprendizaje muy enriquecedora de cara a un futuro trabajo como consultor estadístico, o como ingeniero de *machine learning* optimizando modelos para después embeberlos en una herramienta de software.

En cuanto a los siguientes pasos que realizar una vez obtenido el modelo final, lo ideal sería combinar el modelo con algoritmos de segmentación por computador que automaticen la medición y el cálculo de las covariables utilizadas por el modelo predictivo final. De esta forma se obtendrían herramientas muy interesantes en las que el médico o el analista de laboratorio tan sólo tendría que introducir las imágenes del microscopio y el programa daría un diagnóstico directo. Una herramienta que tener

en cuenta en este ámbito es EBImage, un paquete de Bioconductor que permite segmentar núcleos celulares y obtener características de los mismos (Pau et al., 2010).

Un posible producto final sería una aplicación Shiny que integre los mejores modelos. La aplicación aceptaría como input imágenes al microscopio de tejido mamario, con unos aumentos y tinción previamente determinados; utilizaría el paquete EBImage para procesar las imágenes y calcular las covariables necesarias para el ajuste del modelo, y devolvería la probabilidad de la existencia de células cancerosas en la muestra. Para evaluar la calidad de esta herramienta sería necesaria una base de datos de imágenes histológicas con buena resolución de muestras de masas mamarias.

5. BIBLIOGRAFÍA

Allemani, C., Weir, H. K., Carreira, H., Harewood, R., Spika, D., Wang, X. S., ... & CONCORD Working Group. (2015). Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2). The Lancet, 385(9972), 977-1010.

Amat, J., Máquinas de Vector Soporte (Support Vector Machines, SVMs). https://www.cienciadedatos.net/documentos/34_maquinas_de_vector_soporte_support_vector_machines

American Cancer Society (2022). Breast Cancer Early Detection and Diagnosis. American Cancer Society.

Arbyn, M., Weiderpass, E., Bruni, L., de Sanjosé, S., Saraiya, M., Ferlay, J., & Bray, F. (2020). Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. The Lancet Global Health, 8(2), e191-e203.

Barbosa, A. M., & Martel, F. (2020). Targeting glucose transporters for breast cancer therapy: The effect of natural and synthetic compounds. Cancers, 12(1), 154.

Bennett, K.P. (1992), "Decision Tree Construction via Linear Programming," Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference, pp. 97-101.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. Psychometrika, 52(3), 345-370.

Bray, F., Ferlay, J., Laversanne, M., Brewster, D. H., Gombe Mbalawa, C., Kohler, B., ... & Forman, D. (2015). Cancer Incidence in Five Continents: inclusion criteria, highlights from Volume X and the global status of cancer registration. International journal of cancer, 137(9), 2060-2071.

Claesen, M., & De Moor, B. (2015). Hyperparameter search in machine learning. arXiv preprint arXiv:1502.02127.

Colditz, G. A., Sellers, T. A., & Trapido, E. (2006). Epidemiology—identifying the causes and preventability of cancer? Nature Reviews Cancer, 6(1), 75-83.

Conner-Simons, A., & Gordon, R. (2019). Using AI to predict breast cancer and personalize care. Retrieved from MIT News website.

Fletcher, S.W., Black, W., Harris, R., Rimer, B.K. and Shapiro, S. (1992). Report of the International Workshop on Screening for Breast Cancer. Journal of the National Cancer Institute, vol. 85, pp. 1644-1656.

Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger.

Ghayumizadeh, H., Pakdelazar, O., Haddadnia, J., Rezai, R. G., & Mohammad, Z. M. (2012). Diagnosing breast cancer with the aid of fuzzy logic based on data mining of a genetic algorithm in infrared images.

Giard, R.W. and Hermans, J. (1992). The Value of Aspiration Cytologic Examination of the Breast. A Statistical Review of the Medical Literature. Cancer, vol. 69, pp. 2104-2110.

Hammond, M.E.H. et al. (2010) American Society of Clinical Oncology/College of American Pathologists guideline recommendations for immunohistochemical testing of estrogen and progesterone receptors in breast cancer. Journal of Clinical Oncology 28, 2784–2795

Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., ... & Cardoso, F. (2019). Breast cancer. Nature reviews: Disease primers, 5(1), 1-31.

Hickman, D., Davis, E., Meyer, S., & Schechtel, M. (2016) Core-Needle Biopsy for Breast Abnormalities. Agency for Healthcare Research and Quality.

Jayalakshmi, S., Ganesh, N., Čep, R., & Senthil Murugan, J. (2022). Movie recommender systems: Concepts, methods, challenges, and future directions. Sensors, 22(13), 4904.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Kuhn, M., & Johnson, K. (2013). Applied predictive modeling (Vol. 26, p. 13). New York: Springer.

Maitra, I. K., Nag, S., & Bandyopadhyay, S. K. (2012). Technique for preprocessing of digital mammogram. Computer methods and programs in biomedicine, 107(2), 175-188.

Mangasarian, O.L. (1968). Multisurface Method of Pattern Separation, IEEE Transactions on Information Theory, vol. IT-14, pp. 801-807.

Nelson, H. D., Fu, R., Cantor, A., Pappas, M., Daeges, M., & Humphrey, L. (2016). Effectiveness of breast cancer screening: systematic review and meta-analysis to update the 2009 US Preventive Services Task Force recommendation. Annals of internal medicine, 164(4), 244-255.

Pau, G., Fuchs, F., Sklyar, O., Boutros, M., & Huber, W. (2010). EBImage - an R package for image processing with applications to cellular phenotypes. Bioinformatics, 26(7), 979-981.

Pu, Y., Apel, D. B., & Wei, C. (2019). Applying machine learning approaches to evaluating rockburst liability: a comparation of generative and discriminative models. Pure and Applied Geophysics, 176(10), 4503-4517.

Reshma, V. K., Arya, N., Ahmad, S. S., Wattar, I., Mekala, S., Joshi, S., & Krah, D. (2022). Detection of breast cancer using histopathological image classification dataset with deep learning techniques. BioMed Research International, 2022.

Sperandei, S. (2014). Understanding logistic regression analysis. Biochemia medica, 24(1), 12-18.

Stallings, W. M., & Gillmore, G. M. (1971). A note on "accuracy" and "precision". Journal of Educational Measurement, 8(2), 127-129.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions," Journal of the Royal Statistical Society (Series B), vol. 36, pp. 111-147.

Street, W. N. (2000). Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer. Series in Machine Perception and Artificial Intelligence, 39, 297-326.

Tougui, I., Jilbab, A., & El Mhamdi, J. (2021). Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthcare informatics research, 27(3), 189-199.

Winters, S., Martin, C., Murphy, D., & Shokar, N. K. (2017). Breast cancer epidemiology, prevention, and screening. Progress in molecular biology and translational science, 151, 1-32.

World Health Organization. (2018). World health statistics 2018: monitoring health for the SDGs, sustainable development goals. World Health Organization.

Xanthopoulos, P., Pardalos, P. M., Trafalis, T. B., Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. Robust data mining, 27-33.

Yan, Y., Yang, Z., Semenkovich, T. R., Kozower, B. D., Meyers, B. F., Nava, R. G., ... & Puri, V. (2022). Comparison of standard and penalized logistic regression in risk model development. JTCVS open, 9, 303-316.

ANEXO

Hardware y software utilizados

Se ha utilizado un ordenador portátil de la marca Dell, modelo Inspiron 5593, con las siguientes características:

Procesador: Intel(R) Core(TM) i7-1065G7 CPU. 1.30GHz - 1.50 GHz.

RAM instalada: 16 GB.

Sistema operativo: Windows 10 Pro, versión 21H2.

Versión de RStudio: 2022.12.0+353, "Elsbeth Geranium" para Windows.

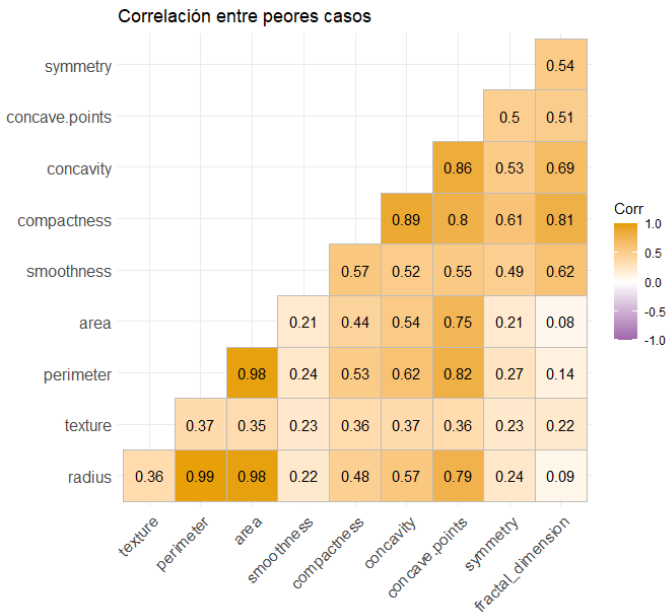
Código utilizado en la realización de la memoria

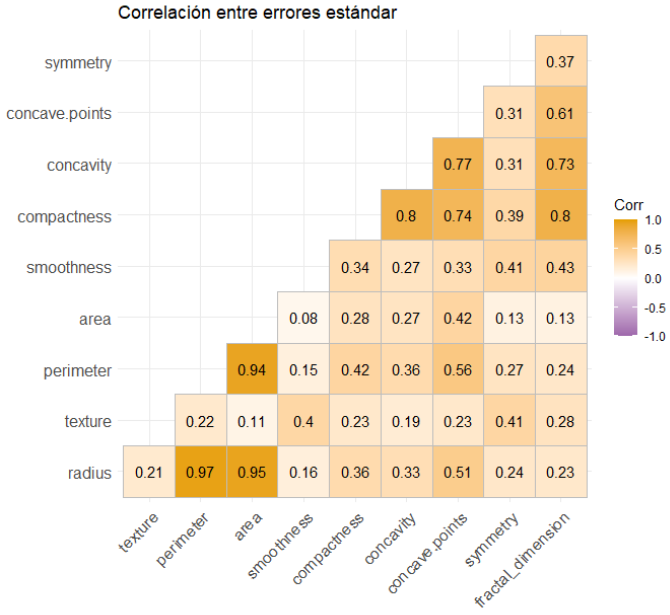
Disponible en el archivo 'codigo_TFM_Julian_Guillo.Rmd'. Se puede acceder al mismo, junto con un archivo .RData que contiene todas las gráficas y los modelos ajustados, en el siguiente enlace:

https://github.com/Julian-Guillo/breast_cancer_diagnosis

Gráficas y tablas extra

Figura A1. Matrices de correlación para todas las covariables




Correlación entre errores estándar
symmetry
concave.points
concavity
compactness
smoothness
area
perimeter
texture
radius
0.37
0.31 0.61
0.77 0.31 0.73
0.8 0.74 0.39 0.8
0.34 0.27 0.33 0.41 0.43
0.08 0.28 0.27 0.42 0.13 0.13
0.94 0.15 0.42 0.36 0.56 0.27 0.24
0.22 0.11 0.4 0.23 0.19 0.23 0.41 0.28
0.21 0.97 0.95 0.16 0.36 0.33 0.51 0.24 0.23
texture
perimeter
area
smoothness
compactness
concavity
concave.points
symmetry
fractal_dimension
Corr
1.0
0.5
0.0
-0.5
-1.0

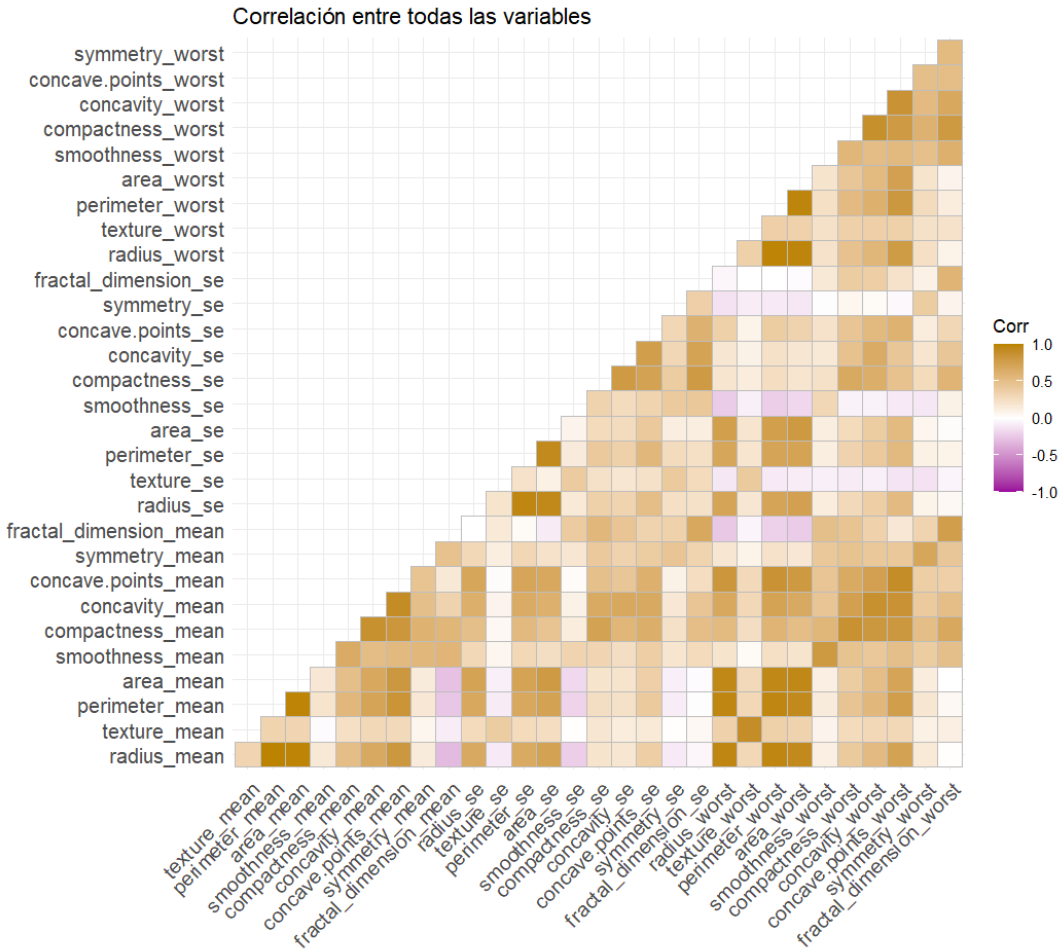
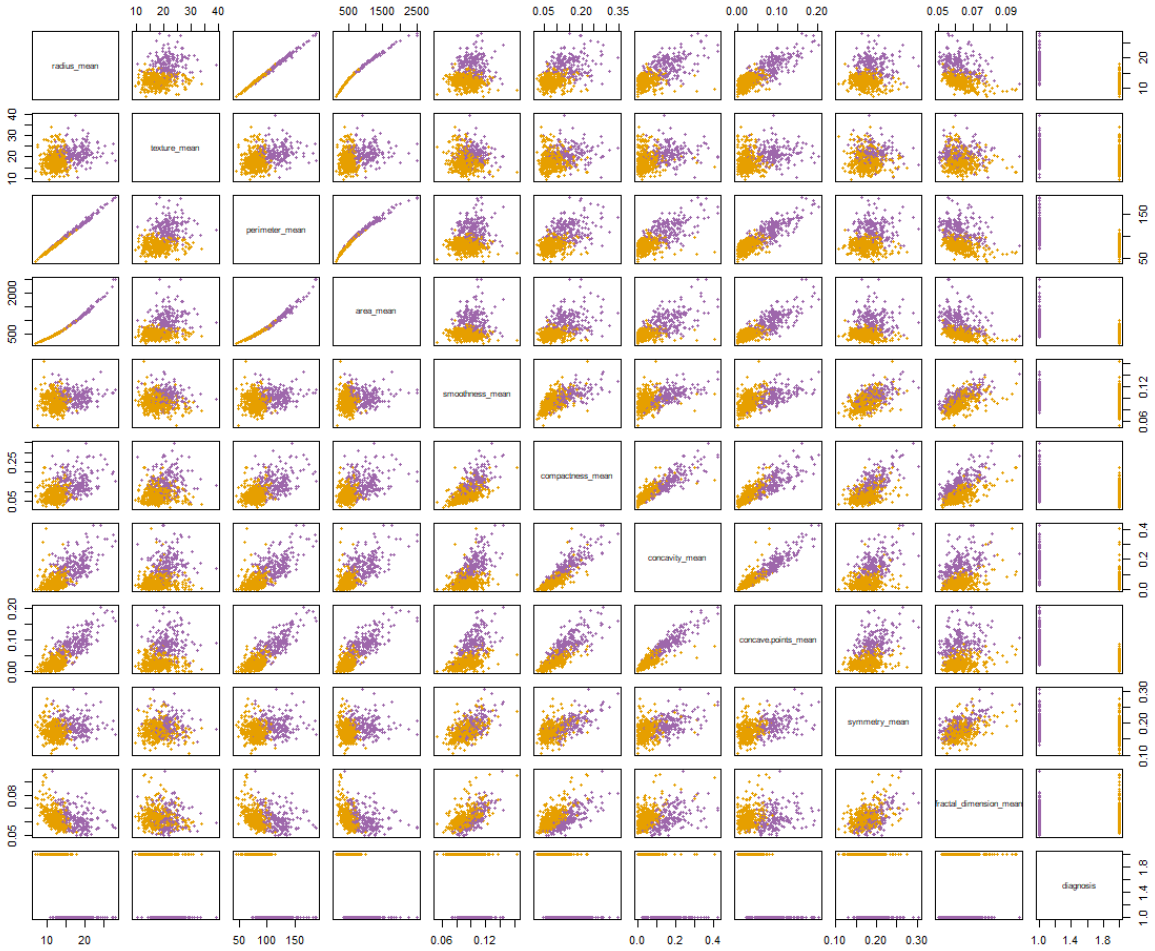
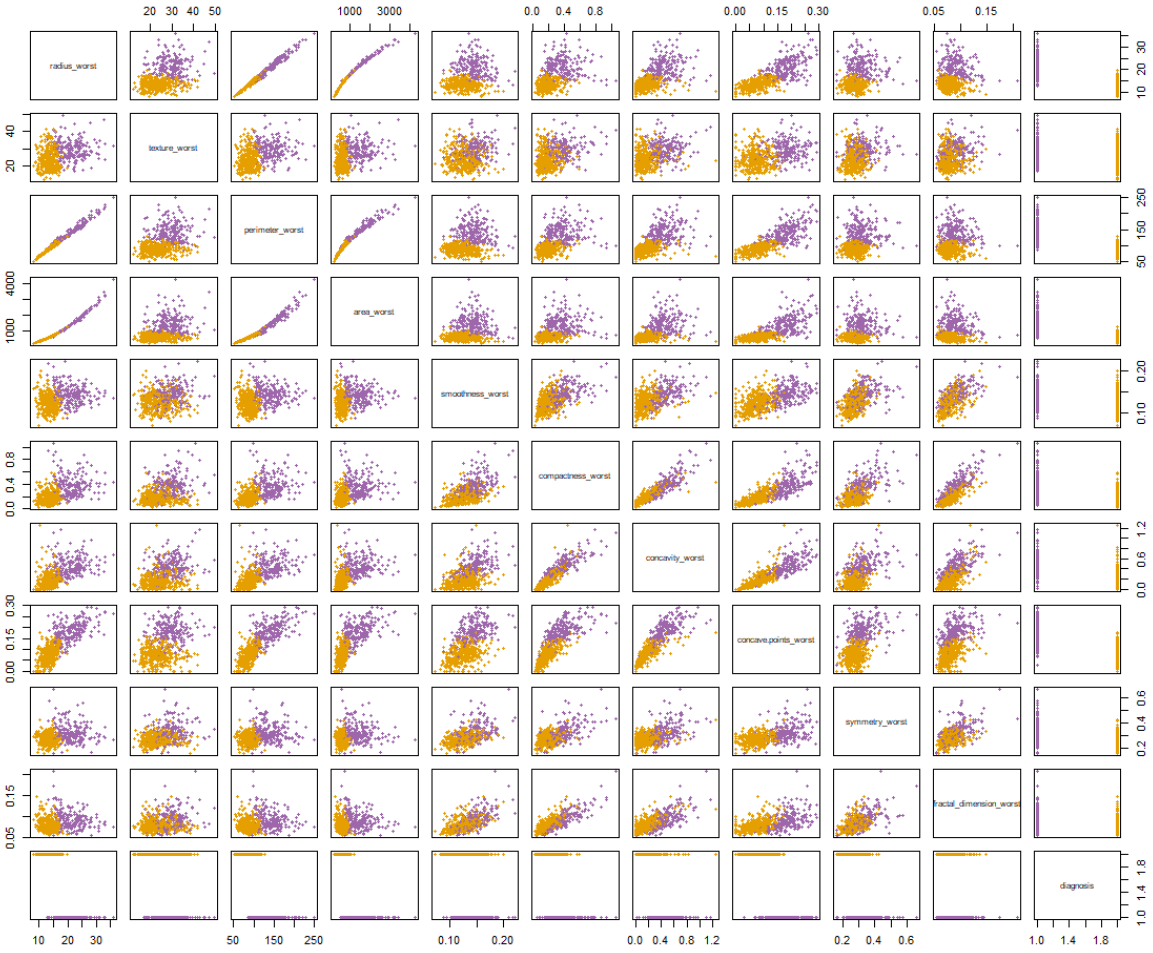

Correlación entre todas las variables
symmetry_worst
concave.points_worst
concavity_worst
compactness_worst
smoothness_worst
area_worst
perimeter_worst
texture_worst
radius_worst
fractal_dimension_se
symmetry_se
concave.points_se
concavity_se
compactness_se
smoothness_se
area_se
perimeter_se
texture_se
radius_se
fractal_dimension_mean
symmetry_mean
concave.points_mean
concavity_mean
compactness_mean
smoothness_mean
area_mean
perimeter_mean
texture_mean
radius_mean
texture_mean
perimeter_mean
area_mean
smoothness_mean
compactness_mean
concavity_mean
concave.points_mean
symmetry_mean
fractal_dimension_mean
radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave.points_se
symmetry_se
fractal_dimension_se
radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave.points_worst
symmetry_worst
fractal_dimension_worst
Corr
1.0
0.5
0.0
-0.5
-1.0

Figura A2. Gráficas de dispersión para todas las variables




radius_worst
texture_worst
perimeter_worst
area_worst
smoothness_worst
compactness_worst
concavity_worst
concave.points_worst
symmetry_worst
fractal_dimension_worst
diagnosis

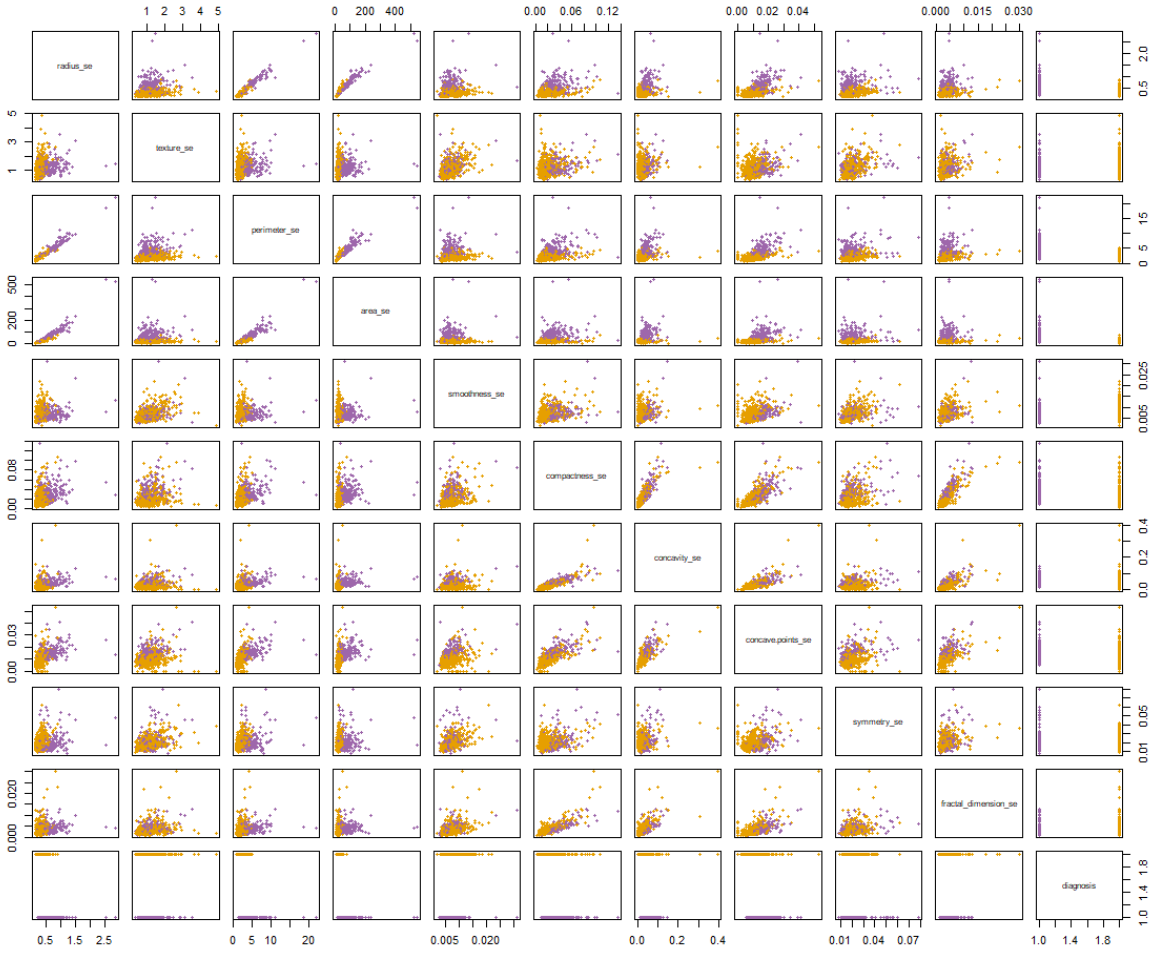
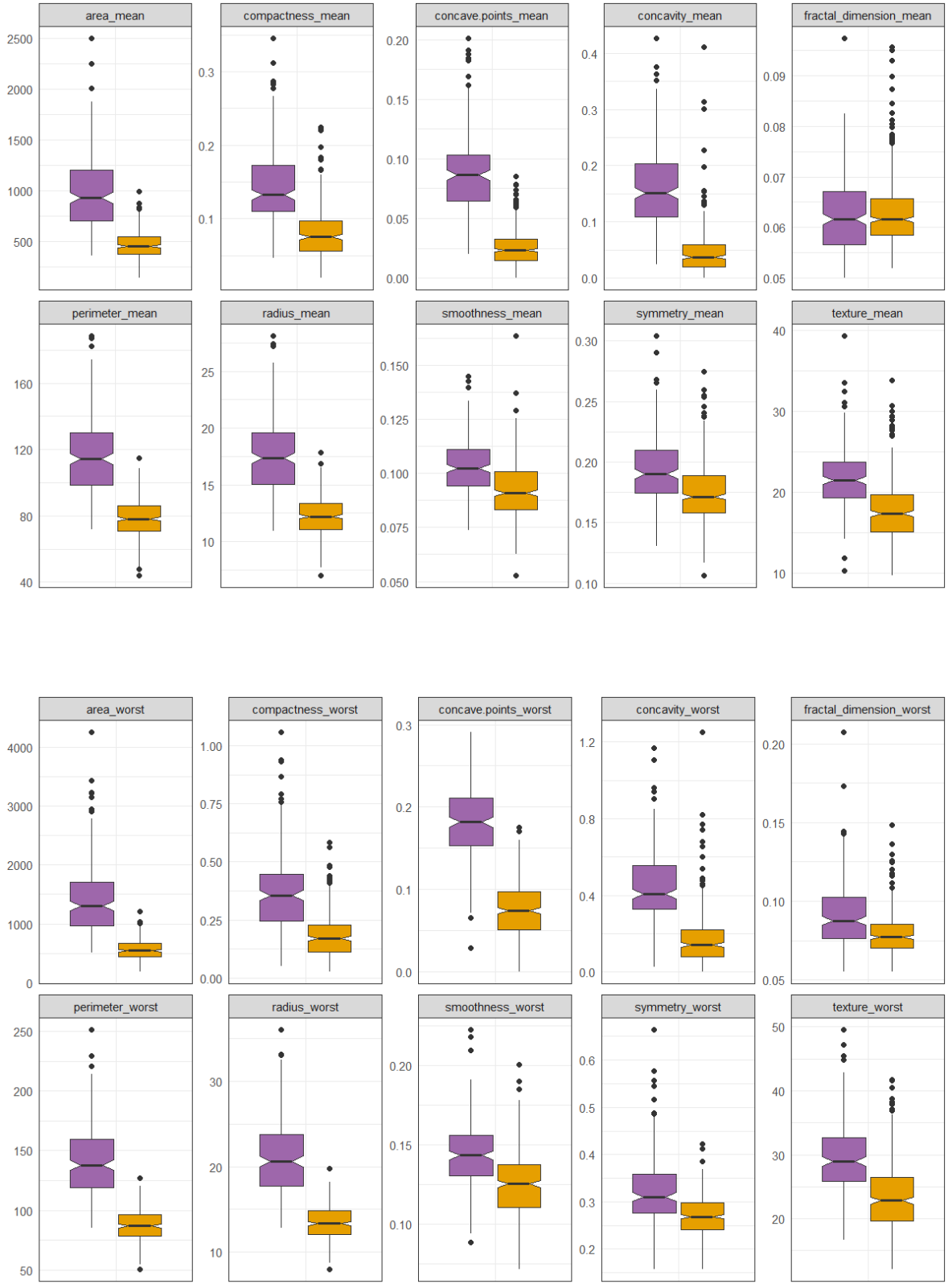

radius_se
texture_se
perimeter_se
area_se
smoothness_se
compactness_se
concavity_se
concave.points_se
symmetry_se
fractal_dimension_se
diagnosis

Figura A3. Diagramas de caja para todas las variables, separando por diagnóstico



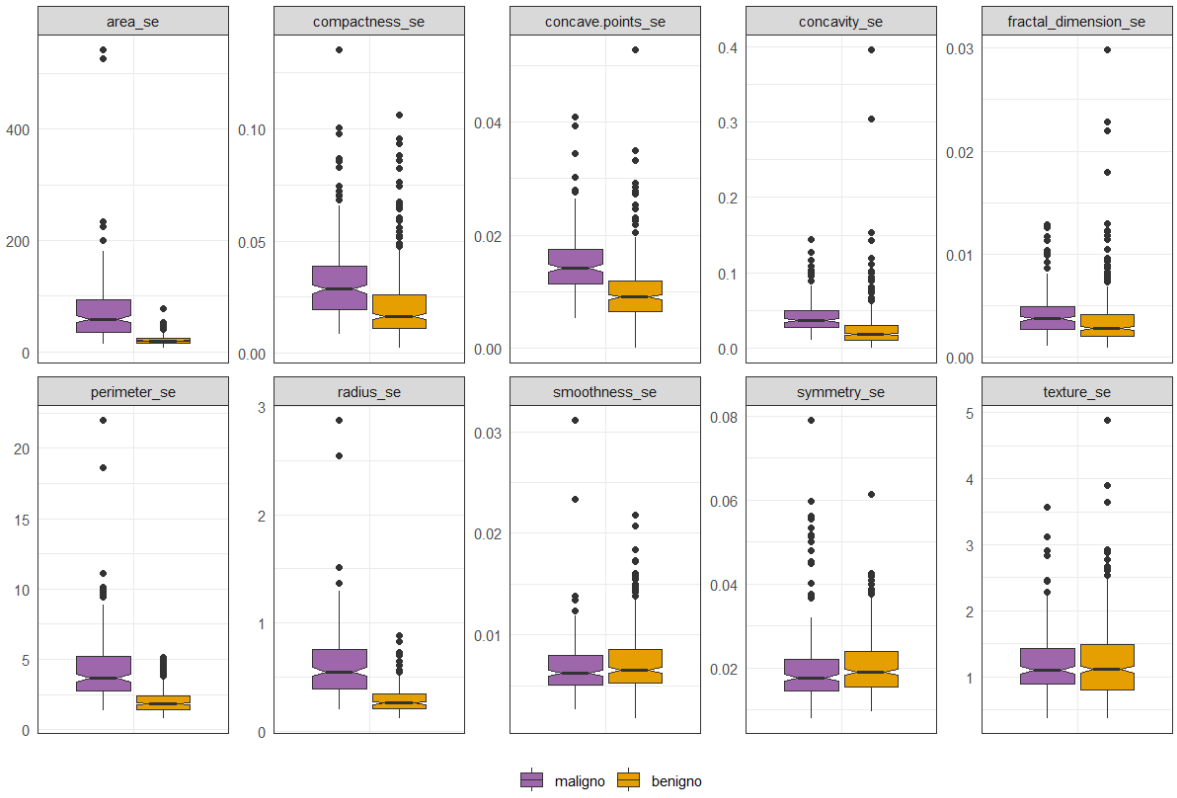
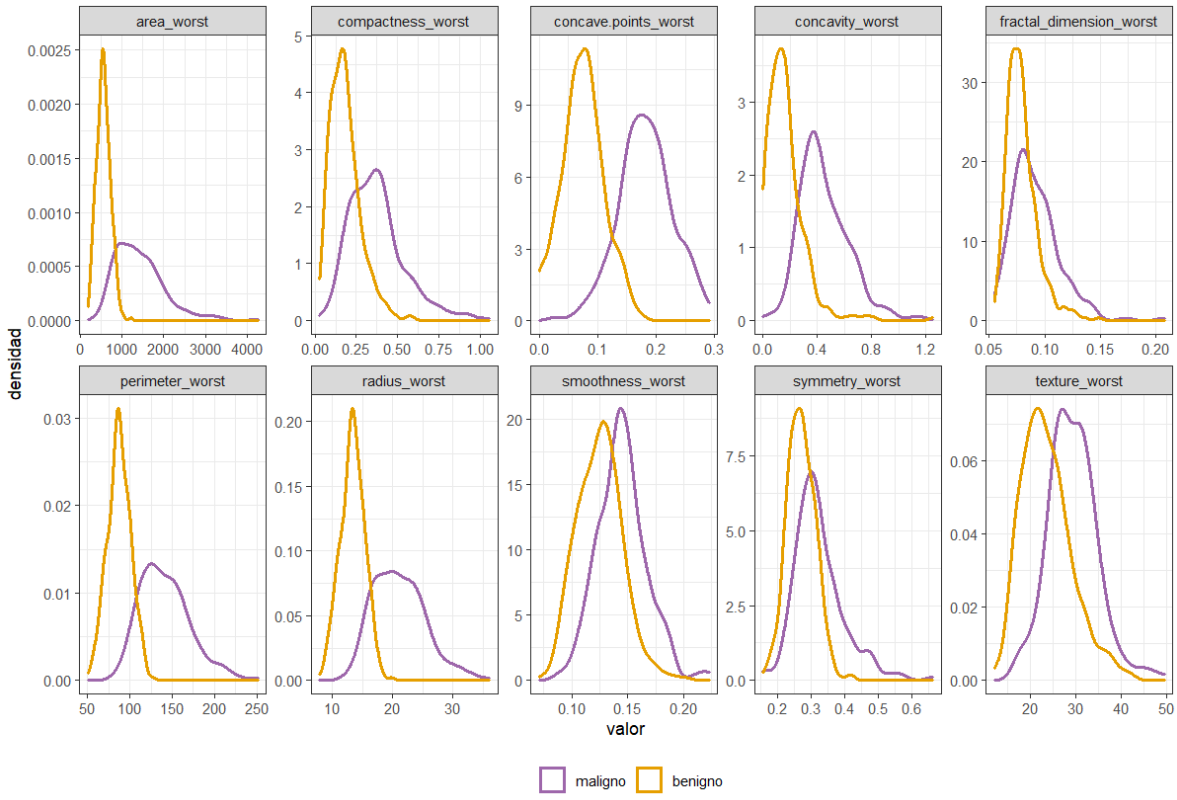


Figura A4. Gráficos de densidad para todas las variables, separando por diagnóstico



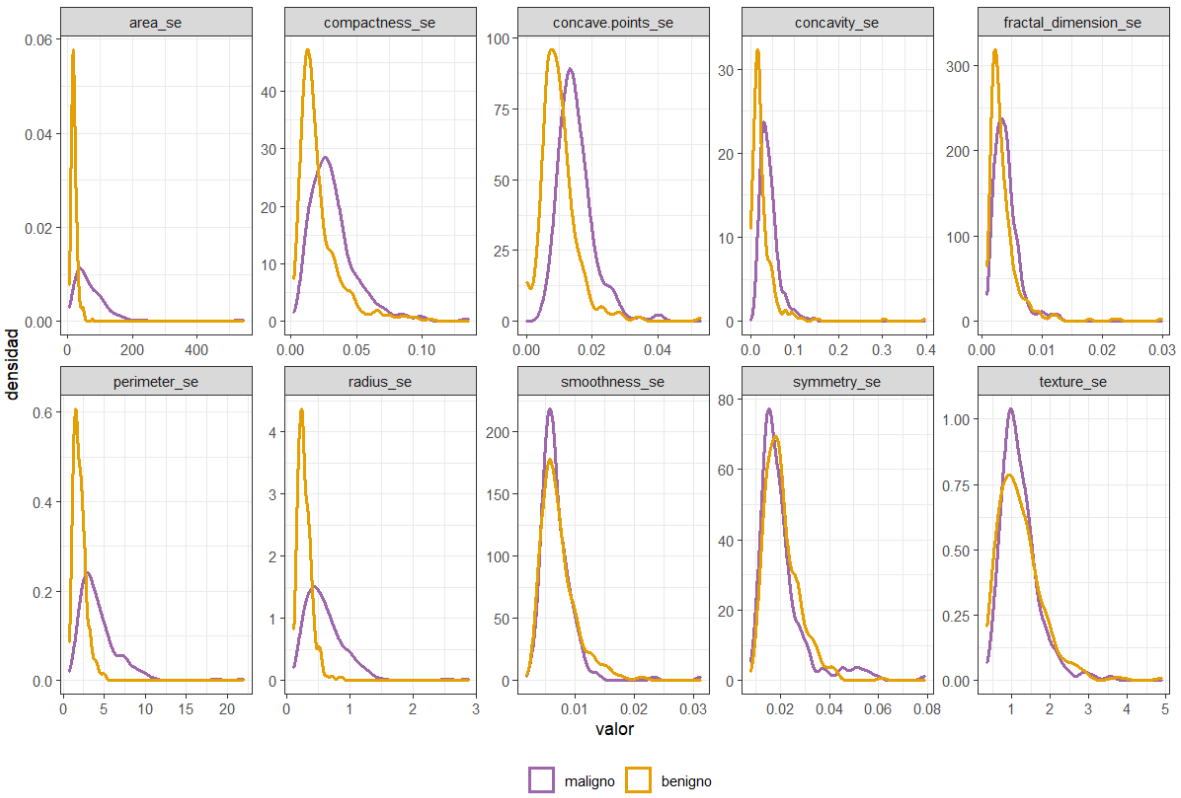


Tabla A1. Importancia de las covariables en las primeras dos componentes principales. Seleccionadas solo aquellas cuyo valor absoluto supera 0.2 en alguna de las dos componentes.

Covariable	Componente 1	Componente 2
concave.points_mean	0,261	0,035
concavity_mean	0,258	-0,06
concave.points_worst	0,251	0,008
compactness_mean	0,239	-0,152
perimeter_worst	0,237	0,2
concavity_worst	0,229	-0,098
perimeter_mean	0,228	0,215
radius_worst	0,228	0,22
area_worst	0,225	0,219
area_mean	0,221	0,231
radius_mean	0,219	0,234
compactness_se	0,17	-0,233
fractal_dimension_worst	0,132	-0,275
fractal_dimension_se	0,103	-0,28
fractal_dimension_mean	0,064	-0,367
smoothness_se	0,015	-0,204

Todos los modelos ajustados

Tabla A2. Información de todos los modelos ajustados, agrupados por tipo de modelo.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)	Hiperparámetro/s seleccionado/s
logística completa	0,95773	0,95418	0,95976	1,48	-
logística 9 Componentes Principales	0,97765	0,96364	0,98595	44,84	-
logística correlación <90%	0,96593	0,95426	0,97283	1,09	-
logística correlación <95%	0,96241	0,95267	0,96815	1,11	-
logística covariables PCA	0,96472	0,94163	0,97844	1,50	-
logística step forward AIC	0,95884	0,93853	0,97095	31,49	-
logística step backward AIC	0,95892	0,95418	0,96167	79,64	-
logística step forward exactitud	0,97651	0,95736	0,98786	136,77	-
logística step backward exactitud	0,97537	0,96385	0,98222	449,71	-
logística Ridge	0,973	0,93391	0,99624	213,5	λ= 0,0394
logística Ridge <90%	0,95488	0,88838	0,99442	220,59	λ= 0,0392
logística Ridge <95%	0,96605	0,91364	0,99722	230,93	λ= 0,039
logística Lasso	0,97418	0,9544	0,98598	163,54	λ= 0,0023
logística Lasso <90%	0,9777	0,95115	0,99347	153,18	λ= 0,0036
logística Lasso <95%	0,97712	0,9544	0,99066	154,09	λ= 0,0027
logística Elastic Net	0,9824	0,96999	0,98974	89,22	α = 0,3; λ= 0,002
logística Elastic Net <90%	0,98064	0,95902	0,99347	83,76	α = 0,05; λ= 0,0005
logística Elastic Net <95%	0,97597	0,94495	0,99439	86,12	α = 0,05; λ= 0,003
LDA completo	0,95786	0,89646	0,99437	1,03	-
LDA 19 Componentes Principales	0,96124	0,90382	0,99534	32,43	-
LDA <90%	0,95657	0,88824	0,9972	0,78	-
LDA <95%	0,94728	0,85859	1	0,84	-
LDA covariables PCA	0,949	0,87071	0,99529	1,08	-

LDA step forward exactitud	0,97772	0,94502	0,9972	231,63	-
LDA step backward exactitud	0,97007	0,9246	0,9972	372,07	-
KNN	0,97066	0,93867	0,98974	16,84	k = 3
KNN <90%	0,95082	0,90267	0,97944	15,31	k = 7
KNN <95%	0,96017	0,9215	0,9832	15,79	k = 6
KNN Manhattan	0,96888	0,94488	0,98315	17,89	k = 8
KNN Manhattan <90%	0,95953	0,91955	0,9832	15,14	k = 5
KNN Manhattan <95%	0,96545	0,93095	0,98601	16,68	k = 5
árbol decisión	0,92024	0,8728	0,94844	4,81	cp = 0
árbol decisión cp	0,92024	0,8728	0,94844	9,05	cp = 0
árbol decisión profundidad	0,92134	0,87727	0,94757	3,72	maxdepth = 3
árbol decisión profundidad <90%	0,94304	0,91955	0,9569	3,48	maxdepth = 4
árbol decisión profundidad <95%	0,93313	0,90404	0,95034	3,71	maxdepth = 4
árboles aleatorios mejor árbol	0,94186	0,93059	0,94844	315,36	maxdepth = 4
random forest completo	0,96477	0,94632	0,97563	180,97	mtry = 3; ntree = 100; nodesize = 4
random forest <90%	0,96595	0,93088	0,98688	208,33	mtry = 2; ntree = 100; nodesize = 2
random forest <95%	0,96946	0,94473	0,98413	196,81	mtry = 10; ntree = 500; nodesize = 2
random forest covariables PCA	0,95959	0,92482	0,98042	124,53	mtry = 1; ntree = 500; nodesize = 2
boosting trees completo	0,97824	0,96068	0,98873	451,67	shrinkage = 0,2; interaction.depth = 2; n.minobsinnode = 5; n.trees = 1000
boosting trees <90%	0,97647	0,95267	0,99061	353,05	shrinkage =0,2; interaction.depth = 2; n.minobsinnode = 10; n.trees = 2000
boosting trees <95%	0,98295	0,96378	0,99434	421,51	shrinkage = 0,2; interaction.depth = 2; n.minobsinnode = 10; n.trees = 1000
boosting trees covariables PCA	0,97127	0,94661	0,98603	364.79	shrinkage = 0,1; interaction.depth = 1; n.minobsinnode = 10; n.trees = 1000
SVM lineal completa	0,97769	0,95584	0,99066	54,19	C = 0,277
SVM lineal <95%	0,97596	0,94488	0,99442	61,39	C = 1,788
SVM lineal <90%	0,97539	0,94336	0,99442	59,79	C = 0,216

SVM polinómica completa	0,97999	0,95743	0,99341	436,73	C = 0,452; scale = 0,112; degree = 2
SVM polinómica <95%	0,97709	0,95108	0,99254	427,51	C = 0,452; scale = 0,112; degree = 2
SVM polinómica <90%	0,97948	0,9544	0,99442	336,84	C = 1,558; scale = 0,112; degree = 3
SVM radial completa	0,98352	0,9684	0,99251	102,05	C = 5,56; γ = 0,013
SVM radial <90%	0,97885	0,95281	0,99437	77,14	C = 4,45; γ = 0,021
SVM radial <95%	0,98063	0,95584	0,99534	74,14	C = 3,34; γ = 0,023
SVM lineal covariables artículo	0,9671	0,93059	0,98881	107,97	C = 0,395
SVM polinómica covariables artículo	0,97416	0,94949	0,98881	645,87	C = 0,894; scale = 0,667; degree = 2
SVM radial covariables artículo	0,97417	0,95108	0,98788	193,62	C = 10; γ = 0,078

Tabla A3. Información de todos los modelos ajustados, ordenados de mayor a menor exactitud.

Modelo	Exactitud	Sensibilidad	Especificidad	Tiempo (s)	Hiperparámetro/s seleccionado/s
SVM radial completa	0,98352	0,9684	0,99251	102,05	C = 5,56; γ = 0,013
boosting trees <95%	0,98295	0,96378	0,99434	421,51	shrinkage = 0,2; interaction.depth = 2; n.minobsinnode = 10; n.trees = 1000
logística Elastic Net	0,9824	0,96999	0,98974	89,22	α = 0,3; λ= 0,002
logística Elastic Net <90%	0,98064	0,95902	0,99347	83,76	α = 0,05; λ= 0,0005
SVM radial <95%	0,98063	0,95584	0,99534	74,14	C = 3,34; γ = 0,023
SVM polinómica completa	0,97999	0,95743	0,99341	436,73	C = 0,452; scale = 0,112; degree = 2
SVM polinómica <90%	0,97948	0,9544	0,99442	336,84	C = 1,558; scale = 0,112; degree = 3
SVM radial <90%	0,97885	0,95281	0,99437	77,14	C = 4,45; γ = 0,021
boosting trees completo	0,97824	0,96068	0,98873	451,67	shrinkage = 0,2; interaction.depth = 2; n.minobsinnode = 5; n.trees = 1000
LDA step forward exactitud	0,97772	0,94502	0,9972	231,63	-
logística Lasso <90%	0,9777	0,95115	0,99347	153,18	λ= 0,0036
SVM lineal completa	0,97769	0,95584	0,99066	54,19	C = 0,277
logística 9 Componentes Principales	0,97765	0,96364	0,98595	44,84	-
logística Lasso <95%	0,97712	0,9544	0,99066	154,09	λ= 0,0027
SVM polinómica <95%	0,97709	0,95108	0,99254	427,51	C = 0,452; scale = 0,112; degree = 2
logística step forward exactitud	0,97651	0,95736	0,98786	136,77	-
boosting trees <90%	0,97647	0,95267	0,99061	353,05	shrinkage =0,2; interaction.depth = 2; n.minobsinnode = 10; n.trees = 2000
logística Elastic Net <95%	0,97597	0,94495	0,99439	86,12	α = 0,05; λ= 0,003
SVM lineal <95%	0,97596	0,94488	0,99442	61,39	C = 1,788
SVM lineal <90%	0,97539	0,94336	0,99442	59,79	C = 0,216
logística step backward exactitud	0,97537	0,96385	0,98222	449,71	-
logística Lasso	0,97418	0,9544	0,98598	163,54	λ= 0,0023
SVM radial covariables artículo	0,97417	0,95108	0,98788	193,62	C = 10; γ = 0,078

SVM polinómica covariables artículo	0,97416	0,94949	0,98881	645,87	C = 0,894; scale = 0,667; degree = 2
logística Ridge	0,973	0,93391	0,99624	213,5	λ= 0,0394
boosting trees covariables PCA	0,97127	0,94661	0,98603	364.79	shrinkage = 0,1; interaction.depth = 1; n.minobsinnode = 10; n.trees = 1000
KNN	0,97066	0,93867	0,98974	16,84	k = 3
LDA step backward exactitud	0,97007	0,9246	0,9972	372,07	-
random forest <95%	0,96946	0,94473	0,98413	196,81	mtry = 10; ntree = 500; nodesize = 2
KNN Manhattan	0,96888	0,94488	0,98315	17,89	k = 8
SVM lineal covariables artículo	0,9671	0,93059	0,98881	107,97	C = 0,395
logística Ridge <95%	0,96605	0,91364	0,99722	230,93	λ= 0,039
random forest <90%	0,96595	0,93088	0,98688	208,33	mtry = 2; ntree = 100; nodesize = 2
logística correlación <90%	0,96593	0,95426	0,97283	1,09	-
KNN Manhattan <95%	0,96545	0,93095	0,98601	16,68	k = 5
random forest completo	0,96477	0,94632	0,97563	180,97	mtry = 3; ntree = 100; nodesize = 4
logística covariables PCA	0,96472	0,94163	0,97844	1,50	-
logística correlación <95%	0,96241	0,95267	0,96815	1,11	-
LDA 19 Componentes Principales	0,96124	0,90382	0,99534	32,43	-
KNN <95%	0,96017	0,9215	0,9832	15,79	k = 6
random forest covariables PCA	0,95959	0,92482	0,98042	124,53	mtry = 1; ntree = 500; nodesize = 2
KNN Manhattan <90%	0,95953	0,91955	0,9832	15,14	k = 5
logística step backward AIC	0,95892	0,95418	0,96167	79,64	-
logística step forward AIC	0,95884	0,93853	0,97095	31,49	-
LDA completo	0,95786	0,89646	0,99437	1,03	-
logística completa	0,95773	0,95418	0,95976	1,48	-
LDA <90%	0,95657	0,88824	0,9972	0,78	-
logística Ridge <90%	0,95488	0,88838	0,99442	220,59	λ= 0,0392
KNN <90%	0,95082	0,90267	0,97944	15,31	k = 7
LDA covariables PCA	0,949	0,87071	0,99529	1,08	-

LDA <95%	0,94728	0,85859	1	0,84	-
árbol decisión profundidad <90%	0,94304	0,91955	0,9569	3,48	maxdepth = 4
árboles aleatorios mejor árbol	0,94186	0,93059	0,94844	315,36	maxdepth = 4
árbol decisión profundidad <95%	0,93313	0,90404	0,95034	3,71	maxdepth = 4
árbol decisión profundidad	0,92134	0,87727	0,94757	3,72	maxdepth = 3
árbol decisión	0,92024	0,8728	0,94844	4,81	cp = 0
árbol decisión cp	0,92024	0,8728	0,94844	9,05	cp = 0