

Evaluación de algoritmos de clasificación para la predicción de cáncer de mama basados en mamografías e implementación del modelo óptimo en una aplicación web



Universitat Oberta
de Catalunya

Alberto Rey Abelaira

Estadística y Aprendizaje Automático

Máster en Bioinformática y Bioestadística

Nombre del tutor/a de TF:
Romina Astrid Rebrij

Nombre del/de la PRA:
Carles Ventura Royo

16 de Enero de 2024



Esta obra esta sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada
<https://creativecommons.org/licenses/by-nc/3.0/es/>

*A Tania,
sin la cual este trabajo no habría visto la luz,
gracias. Por todo.*

A.R.

Ficha Del Trabajo Final

Título del trabajo:	Evaluación de algoritmos de clasificación para la predicción de cáncer de mama basados en mamografías e implementación del modelo óptimo en una aplicación web
Nombre del autor/a:	Alberto Rey Abelaira
Nombre del tutor/a de TF:	Romina Astrid Rebrij
Nombre del/de la PRA:	Carles Ventura Royo
Fecha de entrega:	16 de Enero de 2024
Titulación o programa:	Máster en Bioinformática y Bioestadística
Área del trabajo final:	Estadística y Aprendizaje Automático
Idioma del trabajo:	Castellano
Palabras clave:	Aprendizaje automático, cáncer de mama, clasificación

Resumen del trabajo

Este Trabajo Fin de Máster se centra en la aplicación de técnicas de aprendizaje automático para mejorar la detección de cáncer de mama a través del estudio de las características del tumor a través de mamografías. La motivación principal fue abordar la baja eficacia predictiva de las mamografías, que conlleva a un alto porcentaje de procedimientos diagnósticos innecesarios, incluyendo biopsias y otras pruebas.

En el estudio se evaluaron varios algoritmos de clasificación, incluyendo k Vecinos Más Cercanos (kNN), Máquinas de Vectores de Soporte (SVM), Naive Bayes y Regresión Logística, con énfasis en la comprensión y interpretabilidad del proceso de decisión. Se utiliza un conjunto de datos que contiene características de masas mamográficas extraídas de las imágenes, incluyendo edad del paciente y detalles morfológicos de las masas medidas en escala BIRADS.

El análisis reveló que el modelo de Regresión Logística, aplicado a datos procesados con validación cruzada fue el más eficaz, equilibrando precisión, sensibilidad y especificidad. Este modelo se integró en una aplicación web interactiva, proporcionando a los profesionales de la salud una herramienta práctica y accesible para la toma de decisiones informadas.

Las conclusiones del TFM destacan la importancia del preprocesamiento de datos y la validación cruzada en la mejora del rendimiento de los modelos de aprendizaje automático. El proyecto subraya el potencial del aprendizaje automático en la medicina predictiva y personalizada, mejorando la calidad del diagnóstico y reduciendo procedimientos médicos innecesarios.

Abstract

This Master's Thesis focuses on the application of machine learning techniques to improve breast cancer detection through the study of tumor characteristics in mammograms. The main motivation was to address the low predictive efficacy of mammograms, which leads to a high percentage of unnecessary diagnostic procedures, including biopsies and other tests.

In the study, several classification algorithms were evaluated, including k Nearest Neighbors (kNN), Support Vector Machines (SVM), Naive Bayes, and Logistic Regression, with an emphasis on understanding and interpreting the decision-making process. A dataset containing characteristics of mammographic masses extracted from images was used, including patient age and morphological details of the masses measured on the BIRADS scale.

The analysis revealed that the Logistic Regression model, applied to data processed with cross-validation, was the most effective, balancing accuracy, sensitivity, and specificity. This model was integrated into an interactive web application, providing healthcare professionals with a practical and accessible tool for informed decision-making.

The conclusions of the Master's Thesis highlight the importance of data preprocessing and cross-validation in improving the performance of machine learning models. The project underscores the potential of machine learning in predictive and personalized medicine, enhancing the quality of diagnosis and reducing unnecessary medical procedures.

Índice general

1. Introducción	8
1.1. Contexto y justificación del trabajo	8
1.2. Objetivos del trabajo	8
1.3. Impacto en sostenibilidad, ético-social y de diversidad	9
1.4. Enfoque y método seguido	10
1.5. Breve resumen de productos obtenidos	11
1.6. Breve descripción de la memoria	11
2. Estado del arte	12
3. Fundamentos teóricos	14
3.1. k Vecinos Más Cercanos	14
3.2. Máquinas de Vectores de Soporte	15
3.3. Naive Bayes	16
3.4. Regresión logística	18
3.5. Validación cruzada	19
4. Metodología	20
4.1. Materiales	20
4.2. Métodos	21
5. Resultados y discusión	26
5.1. Análisis exploratorio	26
5.2. Rendimiento de los modelos	30
5.2.1. kNN	30
5.2.2. SVM	31
5.2.3. Naive Bayes	33
5.2.4. Regresión logística	34
5.3. Discusión	35
5.4. Aplicación web	38
6. Conclusiones y trabajos futuros	41
7. Bibliografía	43

Índice de figuras

5.1. Histograma de la edad según el tipo de tumor	27
5.2. Distribución variables categóricas según el tipo de tumor	28
5.3. Mapa de calor entre variables	29
5.4. Gráficas rendimiento algoritmos kNN sin validación cruzada	31
5.5. Matrices de confusión de algoritmos SVM	32
5.6. Rendimiento de los algoritmos Naive Bayes	34
5.7. Rendimiento de los algoritmos Naive Bayes	35
5.8. Pestaña principal de la aplicación	39
5.9. Tabla de valores	39
5.10. Ventana emergente	39
5.11. Resto de funcionalidades pestaña principal	39
5.12. Paneles de entrada de datos con diferentes valores	40
5.13. Resultado para valores correctos	40
5.14. Entrada incorrecta	40
5.15. Posibles resultados al obtener una nueva clasificación	40

Índice de cuadros

4.1. Escala de valores variables categóricas	21
5.1. Cantidad de NA's por variable	26
5.2. Proporción de valores BIRADS - Tipo de tumor	26
5.3. Tabla con las métricas para todos los modelos ¹	37

Capítulo 1

Introducción

1.1. Contexto y justificación del trabajo

Según datos de la Organización Mundial de la Salud (OMS), en el año 2020 se diagnosticó cáncer de mama a un total de 2,3 millones de mujeres en todo el mundo, falleciendo en total unas 685 000 personas [1]. Dentro de los métodos de detección para este tipo de cáncer podemos encontrar las mamografías, ultrasonidos o resonancias magnéticas, entre otros, siendo entre todos ellos el método más eficaz la mamografía. Sin embargo, es una prueba con un bajo valor predictivo, lo que provoca que se realicen un 70 % innecesarias con un resultado benigno ([2], [3]).

En las últimas décadas se ha desarrollado una herramienta de garantía de calidad conocida como BIRADS. El objetivo principal de este proyecto es estandarizar los informes realizados sobre los profesionales médicos sobre el riesgo que tiene un paciente de contraer cáncer de mama. El sistema asigna un valor del 0 al 6 según la probabilidad, basada en unos parámetros a partir de las mamografías, de que el tumor estudiado sea benigno o maligno [4].

En Elter, M. R., & Schulz-Wendtland, T. W. [3] se evalúan tanto los algoritmos de árboles de decisión como redes neuronales artificiales, obteniendo resultados que pueden ayudar a reducir el número de biopsias innecesarias. Se desea conocer y evaluar el rendimiento de otros algoritmos de aprendizaje automático que no se abordan en este estudio, como los algoritmos de Vecinos Más Cercanos o Naive Bayes. Es posible que alguno de estos algoritmos ofrezca también resultados interesantes por su apariencia sencilla de interpretar e implementar, así como su capacidad de adaptarse mejor a conjuntos de datos de diferentes características [5].

Además, tampoco se plantea la posibilidad de acercar estos resultados a los profesionales médicos y por ello surge la idea de crear una aplicación web interactiva, donde introduciendo los datos del análisis de mamografía y los valores para el tumor utilizando el sistema BIRADS, se obtenga la predicción del tipo de tumor, facilitando y agilizando el proceso tanto para los profesionales como para los pacientes.

1.2. Objetivos del trabajo

A continuación enumeramos los objetivos que se pretenden conseguir durante la elaboración de este proyecto:

Objetivo general

- Desarrollar una aplicación web para implementar el modelo óptimo en la predicción de cáncer de mama entre los algoritmos seleccionados, que sirva como herramienta auxiliar para reducir el número de biopsias innecesarias.

Objetivos específicos

- Identificar las técnicas de clasificación más adecuadas para el problema de predicción de cáncer de mama basado en las características del tumor, obtenidas a partir del estudio de las mamografías.
- Determinar el modelo de clasificación más apropiado para la detección de cáncer de mama a partir de las características de las mamografías, con una precisión superior al 80 %.
- Desarrollar una aplicación web interactiva para dicho modelo, que pueda facilitar a los profesionales la clasificación de un tumor y acelerar las fases de diagnóstico y tratamiento.

1.3. Impacto en sostenibilidad, ético-social y de diversidad

Impacto en Sostenibilidad: Incorporar algoritmos de clasificación para la detección de cáncer de mama a través de un estudio de las mamografías conlleva un impacto positivo en la sostenibilidad, ya que contribuyendo a mejorar la eficiencia en el diagnóstico reducimos la necesidad en muchos casos de un análisis manual, minimizando el tiempo y la energía, con la consiguiente mejora en la eficiencia y la reducción de residuos, reduciendo la huella ecológica que suponen los procedimientos médicos. Estas tecnologías pueden integrarse con el Objetivo de Desarrollo Sostenible [6] ODS-9: Industria, Innovación e Infraestructura, ya que se fomenta la investigación para mejorar las infraestructuras del sector salud. Sin embargo, para que el consumo energético y el uso de recursos computacionales no suponga un impacto negativo al implementar estos sistemas, debemos tratar de minimizar estos dos aspectos.


Comportamiento Ético y Responsabilidad Social: Este proyecto aboga por la responsabilidad ética, debiéndose manifestar en la máxima confidencialidad y privacidad en el tratamiento de los datos sensibles de pacientes, con la necesidad de establecer protocolos para proteger la integridad de los datos, asegurando transparencia y ética en el uso de la inteligencia artificial. Todas estas consideraciones se pueden englobar con el ODS 16: Paz, justicia e instituciones sólidas, reforzando las prácticas éticas en el tratamiento de estos datos. No obstante, a pesar de que un manejo ético de la información es un aspecto positivo ineludible, hay un riesgo latente de que los algoritmos sean utilizados de manera inapropiada, suponiendo un impacto negativo en la confianza pública y la percepción en el uso de la tecnología. Por lo tanto, es crucial contar con vigilancia para prevenir un uso indebido de las tecnologías desarrolladas y mantener la transparencia acerca de sus usos.

Diversidad y Derechos Humanos: Un impacto positivo intrínseco de este proyecto es la posibilidad de ofrecer diagnósticos médicos de calidad a diversos grupos poblacionales, contribuyendo a un acceso equitativo y justo a la atención médica. Si estos métodos de detección temprana están disponibles independientemente del género, etnia u orientación sexual de los individuos, contribuimos con los objetivos ODS 5: Igualdad de género y ODS 10: Reducción de las desigualdades. A pesar de que la equidad en el acceso a los tratamientos médicos es clave para garantizar los derechos humanos, fomentando asimismo la inclusión, debemos ser precavidos para que los algoritmos se adecuen a esta premisa, evitando que éstos manifiesten posibles sesgos existentes en los datos de entrenamiento, pudiendo significar una distribución desigual de la tecnología a los diferentes grupos epidemiológicos.

Este proyecto, centrado en el uso de algoritmos de clasificación para la detección temprana de cáncer de mama a través de mamografías, se alinea estrechamente con el ODS 3: Salud y bienestar. Al mejorar la eficiencia y precisión en el diagnóstico, contribuye significativamente a la calidad de la atención médica, promoviendo además un acceso equitativo y justo a tratamientos de salud, lo que es fundamental para el bienestar global.

En términos generales, este proyecto tiene un enfoque para mejorar los procesos de diagnóstico del cáncer de mama, promoviendo innovaciones tecnológicas en el ámbito sanitario, fomentando prácticas éticas y asegurando la igualdad en el acceso a los recursos. Además, alineándose con varios Objetivos de Desarrollo Sostenible, refleja un compromiso integral con las prácticas éticas, sociales y sostenibles en la salud y la tecnología.

1.4. Enfoque y método seguido

Para realizar el trabajo se ha seleccionado el conjunto de datos *Mammographic Mass* del repositorio público UCI [7]. Este contiene los datos de 961 pacientes a los cuales se les ha realizado una mamografía. Las variables predictoras de las que se dispone son la edad del paciente y los datos en la escala del sistema BIRADS de la densidad, márgenes y forma del tumor. Adicionalmente se tiene una variable con la puntuación de la masa tumoral en la escala BIRADS, pero se descarta como variable predictora. Dejamos a continuación una breve descripción de las tareas realizadas para lograr el objetivo del proyecto, ya que los detalles de la metodología podrán verse en el capítulo 4 de la memoria. Todo el código necesario para lograr el objetivo final del proyecto se ha desarrollado en el lenguaje de programación  [8], a través de la versión 4.1.2 y la interfaz gráfica RStudio [9].

En primer lugar se ha realizado un análisis descriptivo de los datos, haciendo uso de los paquetes `ggplot2`, `corrplot` y `MASS`. En él se ha tratado de buscar relaciones relevantes entre las variables, así como datos significativos que podamos extrapolar del conjunto de datos. A continuación se siguieron los pasos más comunes descritos y realizados en la mayoría de referencias, como por ejemplo en Lantz, 2015 [5], que incluyen el preprocesado de datos para adaptar los datos a las necesidades de cada algoritmo, el entrenamiento y la validación de los distintos modelos, evaluación del rendimiento y intentos de mejora en los modelos.

Una vez entrenados y evaluado el rendimiento de todos los modelos, se realizó una selección del modelo óptimo que mejor se ajusta a los datos en función de su rendimiento, interpretabilidad y sencillez, entre otros parámetros. Todo el proceso de entrenamiento, mejora y validación

se ha realizado con la librería `caret` [10]. Finalmente y conocido este modelo, se ha desarrollado una aplicación interactiva usando el paquete `shiny` [11] de R, que permite obtener la clasificación del tumor de un nuevo paciente a partir de las variables que contenga el modelo óptimo.

1.5. Breve resumen de productos obtenidos

El producto que se espera obtener de este proyecto es una aplicación web interactiva desarrollada con Shiny para la predicción de cáncer de mama, que permita a los profesionales de la salud obtener el tipo de tumor, diferenciado entre maligno y benigno, introduciendo los datos de la edad del paciente y las características del tumor siguiendo la escala de BIRADS.

1.6. Breve descripción de la memoria

A continuación mostraremos una breve explicación del resto de capítulos de esta memoria relacionados con el objetivo general del proyecto:

- **Fundamentos teóricos:** En este capítulo se plantean las ideas generales de los algoritmos de clasificación y sus modelos, así como las técnicas utilizadas para tratar de mejorar el rendimiento de los mismos, con el objetivo de que el lector consiga los conceptos básicos sobre los mismos.
- **Metodología:** Se describen y desarrollan los métodos seguidos para realizar la implementación de los modelos.
- **Resultados y discusión:** Se exponen los resultados del rendimiento de los diferentes modelos, y se hace una comparativa con el objetivo de escoger el modelo óptimo.
- **Desarrollo aplicación web:** Se detallan las diferentes funcionalidades de la aplicación web desarrollada.

Capítulo 2

Estado del arte

El cáncer de mama representa un desafío significativo en la salud pública mundial. La clave para un tratamiento efectivo es la detección temprana, en la cual las mamografías juegan un papel crucial. Más allá de la simple obtención de imágenes, la extracción de características clínicas y morfológicas de las mamografías mediante el uso de algoritmos de clasificación es un área de investigación en auge. En el análisis de mamografías, los médicos se enfocan en características como la densidad del tejido mamario, la presencia de masas, calcificaciones, y distorsiones arquitecturales. Debido a que la interpretación de estas características es manual y, por ende, subjetiva para cada profesional sanitario, cobran gran importancia los algoritmos de clasificación y la inteligencia artificial.

En cuanto al estudio de imágenes, en el estudio de Simonyan y Zisserman (2014) sobre redes neuronales convolucionales (CNN) para la clasificación de imágenes, se destaca la introducción de la arquitectura VGG, que marcó un hito por su profundidad y capacidad de reconocimiento [12]. El uso de múltiples capas con filtros de 3x3 permitió a las redes aprender características complejas y mejorar la precisión en tareas de reconocimiento visual. Este enfoque, validado en el conjunto de datos ImageNet, ha influenciado significativamente el desarrollo de modelos de aprendizaje profundo en visión por ordenador.

En el artículo de Budh, D. y Sapra, A. (2022) [2] se hace una descripción detallada de los métodos actualmente disponibles y en desarrollo para la detección de cáncer de mama, además de hacer una investigación sobre las causas, cuadro clínico e indicadores de la enfermedad. Pone de manifiesto el sistema BIRADS como alternativa plausible al estudio de las mamografías, y desprende la necesidad de mejorar las tecnologías de clasificación y predicción existentes.

Un estudio clave en este campo es el de McKinney et al. (2020) [13], que investigó la eficacia de los algoritmos de aprendizaje profundo para interpretar las mamografías. Los autores encontraron que la IA puede igualar o incluso superar a los expertos humanos en la identificación de cánceres a partir de estas características. Esto demuestra la capacidad de las técnicas avanzadas de IA para extraer y analizar información crítica a partir de las imágenes médicas.

Sin embargo, el desafío no se limita a la clasificación binaria de las imágenes como cancerosas o no cancerosas. La extracción detallada de características, como la forma y tamaño de las masas o la distribución de las calcificaciones, es fundamental para la determinación del tipo y la gravedad del cáncer. Elter y Schulz-Wendtland (2007) [3] destacaron la importancia de un proceso de decisión inteligible en la IA médica, lo cual es crucial para la aceptación clínica de estas herramientas.

A pesar de los avances en la IA y el aprendizaje automático, hay desafíos significativos. La calidad y cantidad de los datos, así como su anotación precisa, son cruciales para el entrenamiento de algoritmos efectivos. La variabilidad inter e intra-observador en la interpretación de las imágenes sigue siendo un desafío que necesita ser abordado. En conclusión, el análisis de las características extraídas de las mamografías mediante algoritmos de clasificación presenta un camino prometedor para mejorar la precisión diagnóstica en la detección del cáncer de mama. Este trabajo se basa en realizar un estudio comparativo entre diversos modelos obtenidos al ejecutar diferentes algoritmos de clasificación más clásicos, tratando así de facilitar el diagnóstico médico y en consecuencia mejorar la calidad de vida durante el tratamiento para los pacientes.

Capítulo 3

Fundamentos teóricos

En la siguiente sección haremos una introducción teórica a los algoritmos de clasificación que se desean implementar. Se desea que los conceptos, fórmulas y definiciones aquí descritos tengan un enfoque accesible a cualquier lector y están basados en las siguientes referencias:

- Bennet & Campbell (2000). [14]
- Bishop, C. M. (2006). [15]
- Harrell, F. E., Jr. (2001). [16]
- Hastie *et al.* (2001). [17]
- Hastie *et al.* (2013). [18]
- Lantz, B. (2015). [5]

3.1. k Vecinos Más Cercanos

El algoritmo k Vecinos Más Cercanos (kNN por sus siglas en inglés) destaca en el aprendizaje automático por su eficiencia y simplicidad ya que, al ser un método no paramétrico, no asume una estructura predefinida en los datos. Esto por un lado permite una amplia aplicabilidad en diferentes contextos, pero por otro lado no produce un modelo, por lo que puede ser difícil entender como se relacionan las variables de los datos con cada clase. Además, requiere una selección adecuada del parámetro k , pudiendo llevar a resultados inconsistentes y afectar a la precisión del algoritmo. El algoritmo de kNN se basa en la hipótesis de que instancias similares tienden a situarse de manera próxima en el espacio correspondiente. Para evaluar la proximidad, una de las distancias más comúnmente usada es la Euclidiana, $d_i = ||x_i - x_0||$, que se adapta muy bien a conjuntos de datos con características numéricas. Existen muchas otras que se recogen en Bishop (2006) [15], debiendo seleccionar aquella que mejor se adapte a nuestro conjunto de datos.

Ligadas a este concepto de distancia están las escalas de los datos, ya que diferentes rangos generarán una dependencia relevante del modelo a las variables con mayores rangos y valores. Es por esto que la normalización, el escalado de los datos o bien alguna codificación interpretable numéricamente son esenciales para garantizar la contribución equitativa de las variables en el

cálculo de la distancia. Una de las técnicas más comunes es normalizar los datos al intervalo $[0, 1]$ a partir de los valores máximo y mínimo de cada variable de la siguiente manera:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

Establecida la distancia y el conjunto de entrenamiento, se define la clasificación para un nuevo valor x como sigue:

$$\hat{Y}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$$

donde $N_k(x)$ se define como el conjunto de k puntos x_i más cercanos a x en la muestra de entrenamiento. En otras palabras, asignamos x a la clase con mayor proporción de observaciones dentro de las k muestras en el conjunto de entrenamiento más cercanas a x . La selección de k debe ser rigurosa ya que puede tener consecuencias en el rendimiento. Un valor bajo de k aumentará el efecto del ruido en el algoritmo, mientras que valores elevados acercarán la probabilidad de cada clase a la probabilidad real, pudiendo provocar una generalización.

Finalmente, en la búsqueda del valor óptimo para k , se combina el entrenamiento del algoritmo para diversos valores con técnicas de validación cruzada (capítulo 3.5), examinando la robustez del parámetro en distintos conjuntos. Sin embargo, este algoritmo no realiza un entrenamiento con los datos en sí mismo, sino que almacena los datos, por lo que requiere una preparación adecuada de los datos y una evaluación rigurosa para asegurar su eficacia.

3.2. Máquinas de Vectores de Soporte

Las Máquinas de Vectores de Soporte (SVM por sus siglas en inglés) son pilares en el aprendizaje automático supervisado, destacando en tareas de clasificación y regresión. Su eficacia radica en la capacidad para manejar datos complejos y ruidosos, aunque su interpretación puede ser desafiante debido a su naturaleza matemática compleja. En problemas de clasificación, SVM busca un hiperplano que maximice la distancia a los puntos más próximos de cada clase, lo cual es clave para su eficiencia en una variedad de contextos.

Si consideramos un conjunto de entrenamiento (x_i, y_i) , donde $x_i \in \mathbf{R}^n$ e $y_i \in \{1, 2, \dots, m\}$, siendo m es el número de clases y asumiendo, sin pérdida de generalidad, la clasificación binaria $y_i \in \{-1, 1\}$, el objetivo de SVM es encontrar un hiperplano, $\mathbf{w} \cdot \mathbf{x} - b = 0$, que optimice la separación de clases. Este hiperplano se diseña para maximizar el margen, es decir, la distancia mínima a los vectores de soporte. Si los datos son linealmente separables, podemos asumir que el margen entre las fronteras es $\frac{2}{\|\mathbf{w}\|}$, estableciendo dos hiperplanos paralelos y equidistantes. La formulación matemática del problema para minimizar esta distancia y conocer el hiperplano que separa de manera óptima los datos (es decir, maximiza el margen) puede escribirse como el problema de optimización:

$$\begin{aligned} &\text{minimizar}_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \\ &\text{sujeto a } y_i(\mathbf{w} \cdot x_i - b) \geq 1 \\ &\quad \forall x_i, \quad i = 1, \dots, p \end{aligned}$$

cuya solución da lugar al clasificador:

$$f(x) = \text{signo}(\mathbf{w} \cdot x_i - b).$$

Para casos donde los datos no son linealmente separables, se introduce una variable de holgura ϵ_i definida como la distancia de una observación situada en el lado contrario del hiperplano. Considerando una variable C de coste, podemos ajustar el modelo, de manera que busquemos un equilibrio entre maximizar el margen y minimizar el coste, es decir, los errores en la clasificación. Podemos formular este problema de la manera siguiente:

$$\begin{aligned} &\text{minimizar}_{\mathbf{w}, b, \epsilon} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \epsilon_i \\ &\text{sujeto a } y_i(\mathbf{w} \cdot \mathbf{x}_i - b) \geq 1 - \epsilon_i, \\ &\quad \forall x_i, \quad i = 1, \dots, p \end{aligned}$$

donde parámetro C es fundamental, ya que regula la compensación entre la complejidad del modelo y su capacidad para generalizar. Las ideas para encontrar la solución a los problemas de optimización se describen en Bennet (2000) [14].

Otra alternativa en el caso de que los datos no sean linealmente separables es introducir el uso de las funciones kernel, definidas como sigue:

$$K(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}),$$

siendo $\Phi(x)$ una función que transporta los datos a un espacio de dimensión superior. Los kernel permiten reescribir el hiperplano en este espacio:

$$\sum_{i=1}^l \alpha_i y_i K(x_i, \mathbf{x}) + b,$$

donde los valores α_i e y_i surgen de transformar el vector del hiperplano \mathbf{w} al nuevo espacio de dimensión superior. Uno de los kernels más extendidos es el kernel radial o gaussiano, haciendo que el algoritmo se adapte con facilidad a la mayoría de conjuntos de datos y que se define como sigue:

$$K(\mathbf{x}, \mathbf{z}) = e^{-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{2\sigma^2}}.$$

La selección adecuada de un kernel puede tener una relevancia crucial en el rendimiento del algoritmo, por lo que a menudo se combina con técnicas de validación cruzada y estudiando distintas configuraciones para los parámetros del modelo, asegurando la efectividad y capacidad de generalización del modelo.

En definitiva, las SVM son, a pesar de su complejidad y de la necesidad de una selección cuidadosa de parámetros, herramientas avanzadas y potentes en el ámbito del aprendizaje automático. Su habilidad para tratar eficazmente datos complejos las hace indispensables en una amplia variedad de aplicaciones prácticas.

3.3. Naive Bayes

El algoritmo de Naive Bayes (NB), fundamentado en el teorema de Bayes, es una técnica de clasificación supervisada conocida por su suposición de independencia entre predictores. Esta característica lo hace particularmente útil en la clasificación de texto y escenarios con un alto número de características, destacando por su simplicidad y eficiencia en el manejo de grandes volúmenes de datos.

El teorema de Bayes se aplica para calcular la probabilidad posterior $P(c_j|\mathbf{x})$ de una clase c_j dado un vector de atributos $\mathbf{x} = (x_1, x_2, \dots, x_n)$:

$$P(c_j|\mathbf{x}) = \frac{P(c_j)P(\mathbf{x}|c_j)}{P(\mathbf{x})},$$

donde $P(c_j)$ es la probabilidad a priori de la clase c_j , y $P(\mathbf{x}|c_j)$ es la probabilidad condicional de observar \mathbf{x} en la clase c_j . La suposición de independencia entre los atributos simplifica $P(\mathbf{x}|c_j)$ al producto de probabilidades individuales de la manera siguiente:

$$P(\mathbf{x}|c_j) = \prod_{i=1}^n P(x_i|c_j).$$

En la clasificación binaria, se comparan $P(c_1|\mathbf{x})$ y $P(c_2|\mathbf{x})$, asignando la instancia a la clase con mayor probabilidad posterior. La regla de decisión se formula como:

$$\text{Clase asignada} = \begin{cases} c_1 & \text{si } P(c_1|\mathbf{x}) > P(c_2|\mathbf{x}), \\ c_2 & \text{en caso contrario.} \end{cases}$$

Un aspecto crítico en NB es el manejo de probabilidades cero, común cuando no existen ejemplos de una combinación específica de clase y atributo en el conjunto de entrenamiento. El suavizado de Laplace, que añade un valor constante α a cada frecuencia, mitiga este problema:

$$P(x_i|c_j) = \frac{N_{x_i,c_j} + \alpha}{N_{c_j} + \alpha n},$$

donde N_{x_i,c_j} y N_{c_j} son las frecuencias de x_i en la clase c_j y el total de instancias en c_j , respectivamente. Además, la eficacia de NB puede verse afectada cuando la independencia de atributos no se cumple. La elección de α en el suavizado de Laplace es crítica y puede influir significativamente en los resultados, motivos por el cual evaluar distintos valores de α junto con la validación cruzada es esencial para estimar su rendimiento y generalización en distintos contextos.

Por otro lado, este algoritmo es capaz de adaptarse muy bien a conjuntos de datos de diversas características si lo combinamos con el uso de kernels, de manera análoga a su uso en SVM. La mayoría de programas diseñados para entrenar este algoritmo llevan implementados diversos tipos de kernel para adaptarse a diferentes conjuntos de datos, por lo que no se profundiza más en esta sección, aunque pueden verse más detalles en Hastie (2023) [18].

En conclusión, Naive Bayes es una herramienta valiosa en el aprendizaje automático, destacando por su aplicabilidad en situaciones con altas dimensiones y su eficiencia computacional. Sin embargo, requiere una cuidadosa consideración en cuanto a la independencia de atributos y la selección de parámetros para maximizar su eficacia en la clasificación.

¹ Una función sigmoide es una función real de variable real definida para todos los parámetros de entrada, diferenciable, acotada, con primera derivada no negativa y que tiene un único punto de inflexión.

3.4. Regresión logística

La Regresión Logística (RL) es un método que, a pesar de su nombre, es ampliamente utilizado en el aprendizaje automático para problemas de clasificación. Se basa en la función logística, una función sigmoide¹ que convierte los valores de entrada en un rango de 0 a 1, lo que es ideal para modelar probabilidades en clasificación binaria.

La fórmula matemática de la RL se define como sigue:

$$P(Y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta \cdot \mathbf{x})}},$$

donde \mathbf{x} representa las características, β_0 es el término de intercepción, y β son los coeficientes que se estiman maximizando la verosimilitud del modelo. Una métrica habitual para evaluar el ajuste del modelo es el R cuadrado de McFadden:

$$R_{\text{McFadden}}^2 = 1 - \frac{\ln(\mathcal{L}_{\text{modelo completo}})}{\ln(\mathcal{L}_{\text{modelo nulo}})},$$

donde $\mathcal{L}_{\text{modelo completo}}$ y $\mathcal{L}_{\text{modelo nulo}}$ representan la verosimilitud del modelo con y sin predictores, respectivamente.

RL es apreciada por su capacidad interpretativa, donde los coeficientes β muestran la influencia de cada característica en la probabilidad de pertenencia a la clase 1. Sin embargo, su limitación principal es la dificultad para modelar relaciones no lineales sin una transformación previa de las características. Para abordar este posible sobreajuste y mejorar la generalización del modelo, se pueden incorporar penalizaciones como Ridge y Lasso. La penalización Ridge añade un término de regularización proporcional al cuadrado de la magnitud de los coeficientes:

$$\text{minimizar} \quad -\mathcal{L}(\beta_0, \beta) + \lambda \sum_{j=1}^p \beta_j^2,$$

mientras que Lasso penaliza el valor absoluto de los coeficientes:

$$\text{minimizar} \quad -\mathcal{L}(\beta_0, \beta) + \lambda \sum_{j=1}^p |\beta_j|.$$

En ambos casos, λ es el parámetro de regularización que controla el grado de penalización. Estos métodos son útiles para reducir la complejidad del modelo y evitar el sobreajuste, especialmente en situaciones con un alto número de predictores, siendo habitual que las matrices de diseño no sean de rango máximo.

En resumen, RL es un método efectivo y sencillo, ideal para contextos en los cuales las relaciones entre las características y la clase objetivo son aproximadamente lineales y se requiere de un modelo interpretativo. Además, introduciendo las penalizaciones Ridge y Lasso se añade un gran abanico de posibilidades para mejorar la robustez del modelo en situaciones con conjuntos de datos complejos.

3.5. Validación cruzada

La Validación Cruzada (CV por sus siglas en inglés) es una técnica indispensable en el aprendizaje automático para evaluar la capacidad de generalización de los modelos predictivos. Utilizada en una variedad de algoritmos como kNN, SVM, NB y RL, la CV implica dividir el conjunto de datos en varias partes para alternar su uso en entrenamiento y prueba, permitiendo así una evaluación integral del modelo.

Una metodología comúnmente empleada es la CV k-fold. Aquí, el conjunto de datos se divide en k partes iguales. En cada iteración, una parte se utiliza como conjunto de prueba y las restantes k como entrenamiento. La eficacia del modelo se calcula promediando los resultados de todas las iteraciones:

$$CV_{k\text{-fold}} = \frac{1}{k} \sum_{i=1}^k \text{Rendimiento}_{\text{modelo}_i},$$

donde $\text{Rendimiento}_{\text{modelo}_i}$ es la medida de desempeño en la i -ésima iteración.

La 10-fold CV es particularmente estándar debido a su equilibrio entre precisión de estimación y carga computacional, en la cual los datos se dividen en 10 partes, proporcionando una muestra suficientemente grande para la evaluación del modelo y, al mismo tiempo, manteniendo una carga computacional razonable. Se ha demostrado que 10 pliegues ofrecen una estimación robusta y fiable del rendimiento del modelo, lo que la convierte en una opción preferente en muchas aplicaciones.

Las ventajas de la CV, especialmente en su variante de 10 pliegues, incluyen una utilización eficiente de los datos, ya que cada observación se usa tanto para entrenamiento como para prueba, y una evaluación más precisa del rendimiento del modelo en comparación con una división simple de entrenamiento/prueba. Sin embargo, la CV puede ser computacionalmente costosa, especialmente para conjuntos de datos grandes y modelos complejos. Además, la elección de k puede afectar a la varianza y al sesgo en la estimación del rendimiento del modelo; un valor más alto de k generalmente reduce el sesgo, pero puede aumentar la varianza.

En resumen, la CV, y en particular la 10-fold CV, es una herramienta esencial en la validación de modelos en aprendizaje automático. Aunque resulta computacionalmente exigente, su capacidad para proporcionar una evaluación realista y confiable del rendimiento del modelo la hace indispensable, especialmente en contextos donde la precisión y la fiabilidad son cruciales.

Capítulo 4

Metodología

En este capítulo describiremos detalladamente los materiales y métodos introducidos en la sección 1.4. Veremos en profundidad los pasos seguidos en cada etapa y explicaremos los criterios tenidos en cuenta en cada caso.

4.1. Materiales

Para realizar este proyecto se ha seleccionado el conjunto de datos *Mammographic Mass* del repositorio UCI. Este contiene 961 registros con información obtenida a partir de mamografías digitales de masas tumorales, las cuales se realizaron en el Instituto de Radiología de la Universidad de Erlangen-Nuremberg entre los años 2003 y 2006. Este conjunto se compone de 6 variables, divididas en una variable no predictora (BIRADS), siguiendo el criterio de los creadores del conjunto de datos, cuatro variables predictoras y una variable respuesta (Type), las cuales describimos a continuación:


- **BIRADS:** Contiene la valoración en escala BIRADS de la masa tumoral y que ha sido obtenido después de una doble revisión del proceso. Este valor es un indicador de la probabilidad de malignidad, cuyas puntuaciones posibles son las siguientes:
 - **0:** Incompleto, siendo necesario repetir la mamografía o recoger datos adicionales.
 - **1:** Mamografía negativa, no se observan hallazgos sospechosos en ningún sentido.
 - **2:** Benigno.
 - **3:** Probablemente benigno, se considera una probabilidad de malignidad inferior al 2 %, aunque se recomienda seguimiento.
 - **4:** Sospechosamente maligno. Se incluye un rango dividido en 3 intervalos y una escala del 2 % al 95 % de sospecha.
 - **5:** Maligno con una probabilidad superior al 95 %.
 - **6:** Malignidad comprobada mediante biopsia.
- **Age:** Contiene la edad del paciente en escala numérica, con valores registrados entre 18 y 96 años.
- **Shape:** Variable categórica correspondiente al atributo en escala BIRADS a la forma de la masa tumoral.

- **Density:** Variable categórica correspondiente al atributo en escala BIRADS a la densidad de la masa tumoral.
- **Margin:** Variable categórica correspondiente al atributo en escala BIRADS de los márgenes o bordes de la masa tumoral.
- **Type:** Variable binaria que contiene la información real sobre la severidad del tumor.

La variable edad contiene registros para pacientes entre 18 y 96 años y la variable respuesta puede tomar los valores 0, que corresponde a masas benignas o 1, para masas malignas. Por otro lado, las variables categóricas toman valores entre y 5, cuyo significado resumimos en la siguiente tabla:

	1	2	3	4	5
Shape	Redondeado	Oval	Lobular	Irregular	N/A
Density	Alta	Iso	Baja	Contiene grasa	N/A
Margin	Circunscrito	Microlobulado	Indistinto	Mal definido	Espiculado

Cuadro 4.1: Escala de valores variables categóricas

El equipo utilizado para el desarrollo del proyecto ha sido un ordenador personal ASUS Vivobook con un procesador 12th Gen Intel® Core™ i5-1235U × 12, una memoria RAM de 16GB y a través del sistema operativo Ubuntu, en la versión 22.04.3 LTS (Jammy Jellyfish) de 64 bits . Tanto para entrenamiento y validación de los modelos, se ha usado el lenguaje de programación , a través de la versión 4.1.2 y la interfaz gráfica RStudio con la versión 2023.06.1+524.

Para finalizar esta sección, detallamos el conjunto de librerías específicas de R empleadas en cada uno de los apartados del proyecto, más allá de las funcionalidades básicas del sistema:

- Para realizar los gráficos se ha empleado el paquete `ggplot2` ??, junto con algunas funcionalidades adicionales para el cálculo de correlaciones con los paquetes `dplyr` y `reshape2`.
- Para el entrenamiento y validación de los modelos se ha utilizado el paquete `caret`, con sus diferentes métodos implementados. Adicionalmente, para las gráficas de rendimiento se ha utilizado la librería `pROC`.
- Por último, para el desarrollo de la aplicación web se han utilizado la librería `shiny` y el servidor de ShinyApps para alojar la aplicación.

4.2. Métodos

Un paso previo al entrenamiento y validación de los modelos es el análisis exploratorio de los datos. En este apartado comenzaremos evaluando la calidad de los datos, viendo si hay valores erróneos y comprobando que los datos estén bien balanceados. Después realizaremos un análisis descriptivo, para comprobar como se distribuyen los datos en las distintas variables, de manera que nos aseguremos que van a funcionar bien con los algoritmos que deseamos entrenar. Incluye las siguientes gráficas:

- Histograma superpuesto para la edad: En él, se realiza un histograma para la edad en cada una de las dos clases de la variable tipo de tumor y se superponen, de manera que comparemos a primera vista ambas distribuciones.
- Gráfico de barras para las variables categóricas: En este caso veremos simultáneamente la cantidad de registros para cada valor de las variables categóricas, divididos por clases. Es el gráfico análogo al anterior cuando las variables categóricas.
- Mapa de calor: Se visualizan las correlaciones entre variables, de manera que una relación más elevada se ilustra con un color más intenso.

Una vez comprobada la calidad del conjunto de datos y su adecuación para los modelos que debemos entrenar, preprocesamos los datos para que se ajusten lo mejor posible. Nuestro conjunto de datos contiene variables categóricas y numéricas, por lo que realizaremos los siguientes pasos:

1. Codificación one-hot para las variables categóricas. Esta codificación es una técnica de procesamiento de datos utilizada en aprendizaje automático que convierte variables categóricas en un formato numérico. Cada categoría se representa con una columna de valores binarios (0 o 1), donde cada columna corresponde a una categoría y solo una de ellas tiene el valor 1 para cada observación, de manera que transformamos en nuestro caso las 3 variables categóricas en un total de 13 variables binarias. Más detalles sobre esta técnica y sus aplicaciones pueden verse en Zheng & Casari (2018) [19].
2. Aplicaremos una normalización min-max (como se describe en el capítulo 3.1) a la variable Age. Debido a que ésta es numérica, la escalamos para que no se produzca un sobreajuste en los datos.

La razón para realizar este ajuste se basa en que los algoritmos de kNN y SVM, que se desean entrenar, tienen un componente de distancia importante, por lo que aplicar la codificación convierte las categóricas en un formato numérico compatible, y junto con el escalado de la edad podemos representar las diferencias entre categorías de manera más efectiva. Para el algoritmo de Naive Bayes debemos suponer independencia, lo cual se consigue mejor con variables numéricas discretas y por último, en la regresión logística no se hace tan necesaria, pero permite modelar el impacto de cada categoría de manera independiente.

En el análisis, se identificó que la matriz one-hot generada no poseía rango máximo, lo cual podría afectar la eficacia de los modelos predictivos debido a la multicolinealidad y la sobredimensión. Para abordar este problema, se implementaron técnicas de reducción de dimensionalidad utilizando la función `varImp` del paquete `caret` en R. Ésta evalúa la importancia de las variables mediante modelos de machine learning, permitiendo identificar y retener solo aquellas características más relevantes.

Para entrenar los modelos se ha trabajado con estos dos conjuntos de datos. Por un lado, el conjunto de datos crudos, con la variable edad sin normalizar y las variables categóricas con sus valores originales y, por otro lado, un conjunto de datos con la variable edad normalizada al intervalo [0,1] y las variables categóricas sometidas a la codificación one-hot. Podemos afirmar que ambos conjuntos de datos son válidos para los 4 algoritmos que se desean entrenar y además, permiten evaluar cuál se adapta mejor, ya que el conjunto con los datos crudos refleja la variabilidad natural de los datos, manteniendo las relaciones originales entre las variables, y

el conjunto procesado facilita el proceso de entrenamiento de estos algoritmos al estandarizar las escalas y transformar las categorías en un formato numérico binario, mejorando así la interpretación y comparación entre variables.

Se han entrenado los algoritmos mencionados, kNN, SVM, Naive Bayes y Regresión logística, explorando diferentes combinaciones de parámetros y validación cruzada. El algoritmo kNN se entrenó en primer lugar con un valor fijo de k para ambos conjuntos de datos, y posteriormente se comparó este rendimiento utilizando la validación cruzada para ambos métodos y modelos. El algoritmo SVM se entrenó utilizando dos tipos distintos de kernel, el lineal y el Gaussiano (o radial), todos ellos con la codificación one-hot. En ambos casos, se utilizó un modelo con los parámetros básicos C y σ , y un modelo con validación cruzada junto con distintos parámetros para valorar diferentes combinaciones. En este caso hemos utilizado tan solo la codificación one-hot ya que, al contrario que el algoritmo kNN, que funciona bien con datos crudos por su capacidad de adaptación, pero se decidió utilizar exclusivamente one-hot para SVM ya que transformando las variables categóricas en variables binarias, se evita que el modelo SVM asuma una relación de orden entre categorías que podría no existir. En cuanto al algoritmo de Naive Bayes, se entrenaron dos modelos básicos para ambos conjuntos, y se compararon con diversas configuraciones de suavizado de Laplace junto con la validación cruzada en ambos conjuntos, de manera que se evaluó el impacto de la regularización de Laplace en la capacidad predictiva del modelo bajo diferentes niveles de suavizado. Por último se ha entrenado el algoritmo de regresión logística con los datos crudos, ya que en este caso no es tan sensible a las diferencias de escala. Además, se ha entrenado este algoritmo con la matriz de diseño reducida, evaluando su rendimiento también con validación cruzada. En los casos donde existe multicolinealidad, las penalizaciones Ridge y Lasso son capaces de reducir esta variabilidad, por lo que se implementa el algoritmo de estos dos modelos junto con validación cruzada, entrenando un total de 6 modelos distintos de regresión logística. Todas estas combinaciones se han entrenado con la función `train` del paquete `caret`, cuyos parámetros básicos comentamos a continuación:

- **formula:** Parámetro para indicar la variable de clase y las variables predictoras que se usarán del conjunto de datos.
- **data:** Para indicar el conjunto de datos con el que se desea entrenar el modelo.
- **method:** Argumento que contiene el algoritmo que se desea entrenar. Las posibles opciones para los modelos que se han descrito son `knn`, `svmLinear`, `svmRadial`, `naive_bayes`, `glm` y `glmnet`.
- **trControl:** En nuestro caso, distinguimos si queremos aplicar o no validación cruzada (aunque admite numerosas configuraciones).
- **tuneGrid:** Argumento que indica las diferentes combinaciones de parámetros que se desean evaluar.
- **family:** Exclusivo en la regresión logística, se añade para indicar el tipo de regresión, con el valor `binomial`.

El siguiente paso que se realizó consiste en validar los modelos y evaluar su rendimiento, con el objetivo de discernir cual ha sido el que mejor se amolda al problema que se estudió con los datos disponibles. Se han obtenido las predicciones para cada modelo usando la función

`predict` para a continuación calcular las distintas métricas de rendimiento, que detallamos a continuación:

- Matriz de confusión: Muestra el resultado de la predicción, facilitando una visión de las discrepancias entre las clases predichas y las reales.
- Accuracy: Mide la proporción entre las predicciones correctas en ambas clases frente al total, siendo un indicador global de la eficacia del algoritmo. (En adelante se usará el termino en español precisión, que no debe confundirse con el término en inglés para el parámetro de rendimiento *precision*).
- Kappa: Constituye un ajuste de la precisión, donde se considera la precisión de un modelo que clasifica por azar. Difiere de ésta ya que se considera que algunas predicciones correctas sucedan por simple casualidad.
- AUC: Área bajo la curva (siglas del inglés *area under the curve*), es una medida para modelos de clasificación y binaria y representa la proporción entre la tasa de verdaderos positivos y la tasa de falsos positivos, aportando la capacidad que tiene el modelo de distinguir entre ambas clases.
- Sensibilidad: Indica la proporción de casos reales positivos que se han clasificado correctamente.
- Especificidad: Mide la proporción de casos negativos reales que se han clasificado correctamente.
- Valor predictivo positivo (PPV): Es la probabilidad de que una observación predicha como positiva sea realmente positiva.
- Ratio de descubrimiento falso (FDR): Refleja la proporción de falsos positivos dentro de las predicciones positivas del algoritmo. Es la métrica opuesta de la especificidad.

Todas estas métricas toman valores entre 0 y 1, buscando en cada una de ellas, excepto el FDR, obtener el resultado más cercano a 1 posible. Las únicas dos excepciones serían la métrica Kappa, que no debe ser superior a la precisión del modelo, ya que indicaría que un modelo que clasifica por azar sería mejor que el que estamos entrenando, lo que indicaría algún tipo de problema subyacente; y el FDR, que se buscamos que sea lo más cercano a cero posible. Dado que refleja la proporción de falsos positivos dentro del total de predicciones positivas, cuanto menor sea indicará que habrá un menor número de casos donde una clase negativa se clasifique erróneamente.

Se debe tener en cuenta cuando se hable de casos positivos o negativos, que esto es una cualidad arbitraria y nada tiene que ver con la propia definición y las métricas del modelo deberán ser analizadas bajo esta premisa. En este proyecto, se ha considerado la clase positiva la clase 1, es decir que un tumor sea maligno. Se escoge de esta manera ya que uno de los parámetros a estudiar y de más relevancia en los algoritmos de clasificación es el porcentaje de falsos positivos, es decir, la posibilidad de que una clase negativa, en este caso un tumor benigno, sea clasificado como positivo, es decir maligno. Este sería el peor escenario para el problema que estamos tratando, y ha sido una de las medidas que más se han tenido en cuenta la hora de seleccionar el mejor modelo.

Con la base de estas métricas y la naturaleza del problema, se ha realizado una discusión de los modelos par discernir cual ofrece un resultado óptimo. Por último de desarrolló la aplicación web con la librería `shiny`, que será alojada en el servidor de ShinyApps (un servicio en la nube que ofrece este paquete) para implementar el modelo seleccionado como óptimo y que mejor se ajusta a los datos y el problema estudiado.

Para finalizar este capítulo hablaremos sobre la reproducibilidad de los resultados obtenidos, ya que es un aspecto fundamental en la investigación científica ya que, además de permitir una validación independiente del estudio, sienta las bases para investigaciones futuras, pudiendo evitar por ejemplo la repetición de errores. En este TFM se ha fijado una semilla¹ en las siguientes situaciones:

- Antes de separar el conjunto de datos para entrenamiento y test, asegurando que entrenamos y validamos todos los modelos con los mismos datos. De esta manera cualquier diferencia en el rendimiento depende exclusivamente del tipo de modelo y sus parámetros y características.
- Antes de entrenar cada modelo, ya que es fundamental para que los resultados sean reproducibles, asegurando que las particiones de los datos sean consistentes en cada ejecución. Se ha seleccionado una semilla única para todos los modelos que no incluyen validación cruzada, y una semilla única y diferente a la anterior para aquellos modelos que sí incluyen validación cruzada, de manera que aseguramos solidez en la selección de características o al realizar subconjuntos.
- Antes de realizar la predicción de los modelos, para estabilizar los algoritmos que pueden incluir aleatoriedad en las predicciones.

La distinción en la semilla según se implemente la validación cruzada o no permite una comparación más precisa y controlada de los modelos, reforzando la integridad y robustez de los resultados.

¹ Una semilla es un número fijo usado para inicializar el generador de números aleatorios, asegurando la reproducibilidad de los resultados al generar datos aleatorios.

Capítulo 5

Resultados y discusión

En este capítulo detallaremos los resultados obtenidos siguiendo los pasos de la metodología anterior, así como la funcionalidades de la aplicación web descrita. Todo el código desarrollado y necesario para reproducir los resultados aquí descritos puede consultarse en el siguiente repositorio público: <https://github.com/alberabelaira/TFM>

5.1. Análisis exploratorio

En un primer análisis de la calidad de los datos se detectaron los siguientes datos erróneos u omitidos:

BIRADS	Age	Shape	Margin	Density	Type
2	5	31	48	76	0

Cuadro 5.1: Cantidad de NA's por variable

Estos registros se han descartado del conjunto de datos ya que se desconoce el motivo de esta omisión y se evitan posibles errores en la computación. Se observó también un registro para la variable BIRADS con el valor 55, descartado también por no ser un posible valor en esta variable. Para comprobar en última instancia la calidad y el balance de los datos, se estudió la relación entre las variables BIRADS, directamente relacionada con la malignidad de los tumores y la variable Type, que indica la clase real del tumor, resultados que mostramos en la siguiente tabla:

	0	1	2	3	4	5	6
Benigno	2	0	7	20	365	31	2
Maligno	3	0	0	4	103	285	7

Cuadro 5.2: Proporción de valores BIRADS - Tipo de tumor

El porcentaje de malignidad para los valores BIRADS 3, 4 y 5 es del 16 %, 22 % y 90 %, siendo algo elevado para la categoría 3, pues se considera aquí una probabilidad de malignidad inferior al 2 %, aunque no se ha considerado relevante debido al carácter subjetivo del estudio de

las radiografías. Sí se han descartado, al ser una incongruencia, los 2 registros de tipo benigno con la categoría 6, ya que esto indicaría que el tumor ha sido probado maligno mediante una biopsia. Se decidió mantener en el conjunto de datos, a pesar de que se desconoce cómo se ha obtenido este conjunto de datos al no especificarse en el estudio original (Elter & Schulz-Wendtland, 2007 [2]), los registros con categoría 6 y maligno, ya que son congruentes y pudieren ser de utilidad para evaluar el rendimiento de los modelos.

Finalmente, se obtuvo un conjunto de datos con un total de 827 observaciones de las cuales 425 son de clase benigna y 402 de clase maligna, lo que supone un 51.4% y 48.6% del total, respectivamente. Se ha considerado que estos datos tienen una calidad y balanceo válidos, así como un tamaño aceptable para entrenar los algoritmos seleccionados. Tan solo ha sido necesario transformar las variables categóricas y respuesta a tipo factor para proceder con el resto del análisis y el entrenamiento de los modelos.

A continuación se estudió como se distribuyen las distintas variables predictoras en función del tipo de tumor, cuyos resultados para la edad mostramos a continuación:

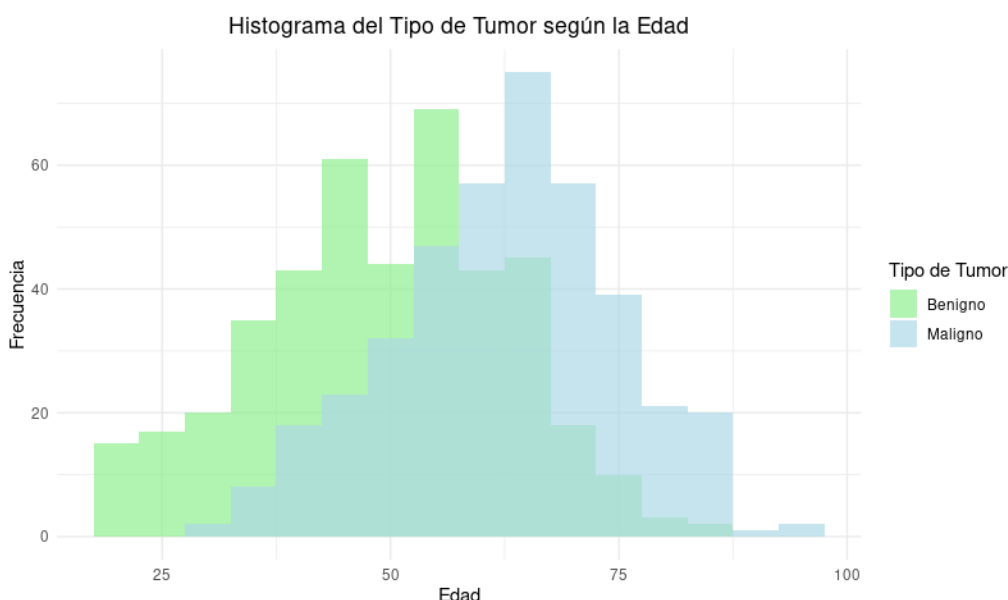


Figura 5.1: Histograma de la edad según el tipo de tumor

Se puede observar que la edad sigue una distribución aparentemente normal, independientemente del tipo de tumor, destacando el hecho de que los valores más bajos de la edad se asocian a tumores benignos, mientras que las edades más avanzadas se asocian a tumores malignos. Se percibe que ambas variables distribuyen sus datos de manera similar, con un ligero desplazamiento hacia edades más tempranas en tumores benignos y hacia edades más avanzadas en tumores malignos, lo que es acorde a la idea de que para pacientes más mayores, el riesgo de padecer cáncer aumenta [1].

Para las variables categóricas se han realizado gráficas análogas (Figura 5.2), con la diferencia que representamos las frecuencias en gráficos de barras. Los resultados son los siguientes:

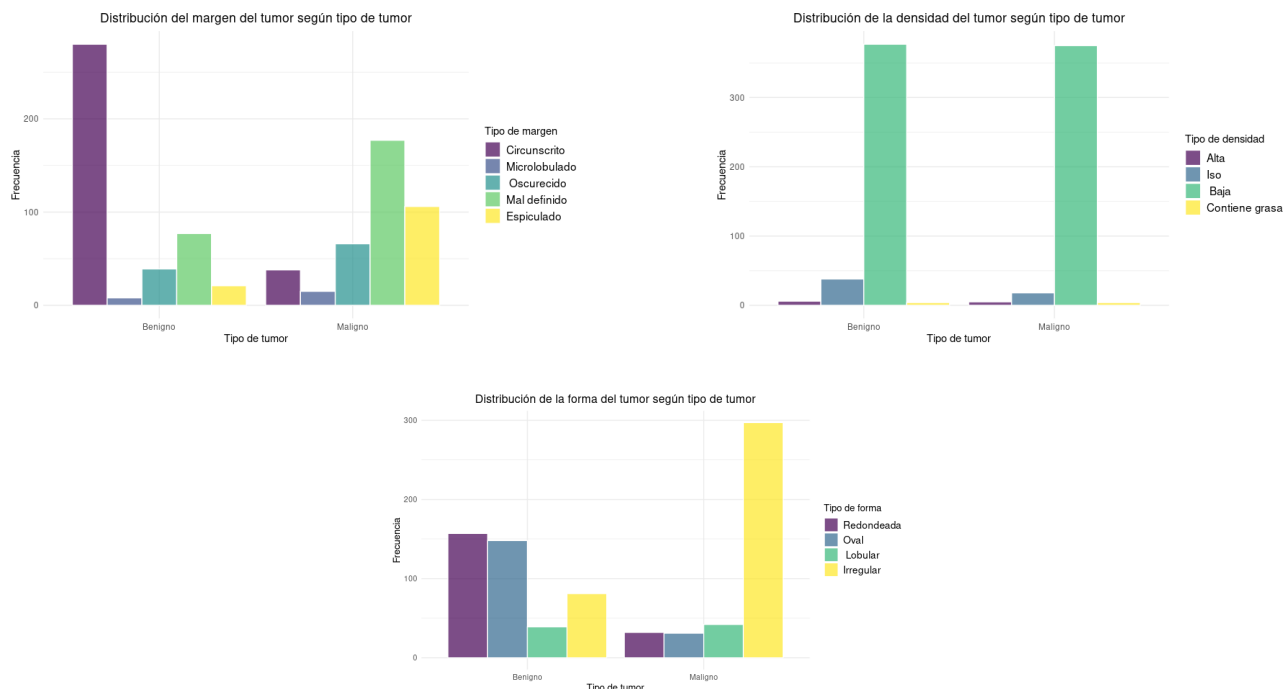


Figura 5.2: Distribución variables categóricas según el tipo de tumor

- Para el margen, un 65 % de los tumores benignos son de tipo circunscrito, mientras que más de la mitad de los tumores malignos son de tipo mal definido o espiculado, lo que implica una clara relación entre el margen del tumor con el tipo de tumor.
- La distribución de la densidad del tumor es prácticamente idéntica en ambas clases, siendo además densidad baja para ambos tumores en aproximadamente el 90 % de los casos.
- Destaca la forma irregular en el 73 % de los tumores malignos, siendo equitativo el reparto de esta variable para los tumores benignos.

De estas gráficas se puede extrapolar una cierta correlación entre el margen y la forma observadas en las mamografías en los distintos tipos de tumores, pero no se detectan diferencias significativas para la densidad en distintos tipos de tumores. Además, se prestó atención a la distribución de la densidad, poniendo el foco en que su distribución irregular no afecte al rendimiento de los modelos.

Para estudiar más a fondo la idea que se intuye acerca de las relaciones entre variables se ha realizado un mapa de calor, considerando las correlaciones entre las variables. Debido a que contamos con una variable numérica y varias categóricas, el coeficiente de correlación de Pearson, usado habitualmente en este contexto, no es válido. Utilizaremos en su lugar el coeficiente de Cramer [20], para la relación entre la variable numérica y las categóricas, y el coeficiente η^2 , obtenido a partir de un ANOVA sobre los datos. Los resultados se muestran en el siguiente mapa de calor, siendo aquellas variables con mayor correlación las que muestran un color más intenso:

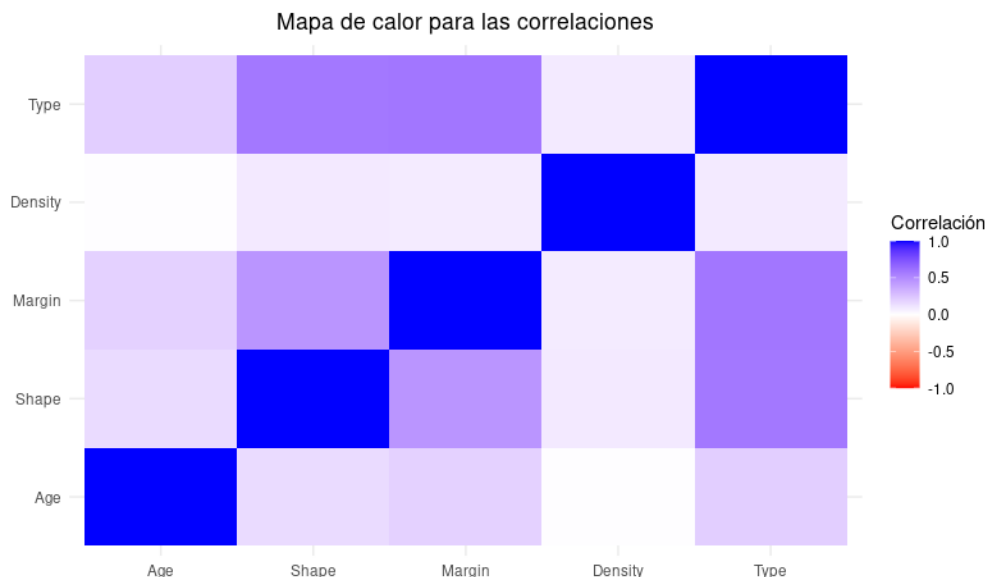


Figura 5.3: Mapa de calor entre variables

- La relación de la variable densidad con el resto es muy baja en todos los casos. Esto podría estar relacionado tanto con la distribución homogénea de sus valores con respecto al tipo de tumor, ya que el 90 % de los valores son de densidad baja.
- La relación del tipo de tumor con el resto de variables predictoras (excepto densidad) es moderada, pero no se ha considerado preocupante ya que conocemos de antemano que la edad tiene una influencia sobre el tipo de tumor, y es intrínseco a la naturaleza del cáncer que las características de una mamografía estén relacionadas con el tipo de tumor.
- Se obtiene una correlación relativamente elevada entre las variables margen y forma pero se ha descartado su relevancia de nuevo por la naturaleza del problema.

Globalmente, a pesar de obtener correlaciones relevantes para algún par de variables, ninguna supera el umbral del 75 % y ninguna variable predictora tiene una correlación superior al 60 % con la variable respuesta. Se ha decidido mantener todas las variables en el conjunto de datos, ya que no hay relaciones fuertes entre ellas, y a pesar de la relativa poca importancia de la variable densidad, no está apenas relacionada con las demás así que su inclusión podría ser interesante para los modelos.

En conclusión, nuestro análisis exploratorio de los datos subraya la calidad y adecuación del conjunto de datos para los diferentes modelos predictivos a evaluar para la detección de cáncer de mama. A pesar de la presencia inicial de valores erróneos y omitidos, el conjunto de datos final presenta un equilibrio adecuado entre las clases, con una distribución casi igual de casos benignos y malignos. La exploración detallada de las variables, incluyendo edad, forma, margen y densidad, ha permitido obtener conclusiones significativas sobre su distribución y relación con el tipo de tumor. Específicamente, se destaca la asociación entre el margen y la forma del tumor en las mamografías, aunque no se encontraron correlaciones excesivamente altas que pudieran indicar redundancias significativas entre las variables. Este balance y diversidad en los datos

refuerzan la confianza en su capacidad para entrenar modelos predictivos robustos y confiables, proporcionando así una base sólida para el entrenamiento de los algoritmos que veremos a continuación.

5.2. Rendimiento de los modelos

En esta sección mostraremos los resultados del rendimiento de los modelos, ofreciendo una comparación entre los modelos y sus posibles mejoras, tratando de discernir en primera instancia cual podría ser el que mejor se ajusta a los datos. Las métricas a las que se hace referencia se muestran en la sección siguiente en la Tabla 5.3. Se utilizar las métricas definidas en el capítulo 4.2, junto con los términos que se definen a continuación para facilitar la comprensión del lector:

- Verdadero positivo: Caso positivo real clasificado como positivo.
- Verdadero negativo: Caso negativo real clasificado como negativo.
- Falso negativo: Caso positivo real que se clasifica como negativo.
- Falso positivo: Caso negativo real que se clasifica como positivo.

Estas definiciones general se adecuarán a las necesidades específicas del problema en la sección siguiente (capítulo 5.3)

5.2.1. kNN

Se comenzó entrenando los algoritmos de vecinos más cercanos, para el cual se han utilizado tanto datos crudos como la codificación one-hot. Se entrenó el algoritmo para ambos conjuntos de datos tomando $k = 14$ como valor fijo, ya que es el valor más próximo a la raíz cuadrada del total de muestras en el conjunto de test, siendo una práctica habitual [5]. Posteriormente, se hizo la misma comparación del rendimiento implementando en ambos conjuntos de datos la validación cruzada y variando el valor de k de 3 a 25, obteniendo el modelo que mejor clasifica los datos par $k = 14$ con datos crudos y $k = 25$ con codificación one-hot, cuyos rendimientos comparamos en la Figura 5.4, donde se muestra la matriz de confusión y la comparación de las gráficas ROC. De los resultados destacamos lo siguiente:

- Los modelos entrenados con la codificación one-hot muestran métricas de rendimiento superiores en todos los casos, lo cual pone de manifiesto la necesidad de entrenar este tipo de modelos con los datos más adecuados posible.
- Ambos algoritmos implementados con one-hot ofrecen la misma clasificación a pesar de que el valor de k difiere. El rendimiento es prácticamente idéntico, difiriendo sólo en el valor de AUC, siendo ligeramente superior en el modelo con validación cruzada. Esto significa que el algoritmo Knn con codificación one-hot y validación cruzada tiene una mejor capacidad para adaptarse a distintos umbrales de decisión, según se desee minimizar más los falsos positivos o los falsos negativos.
- En ambos algoritmos entrenados con validación cruzada, los resultados de la clasificación son mucho mejores para el conjunto de datos codificado.

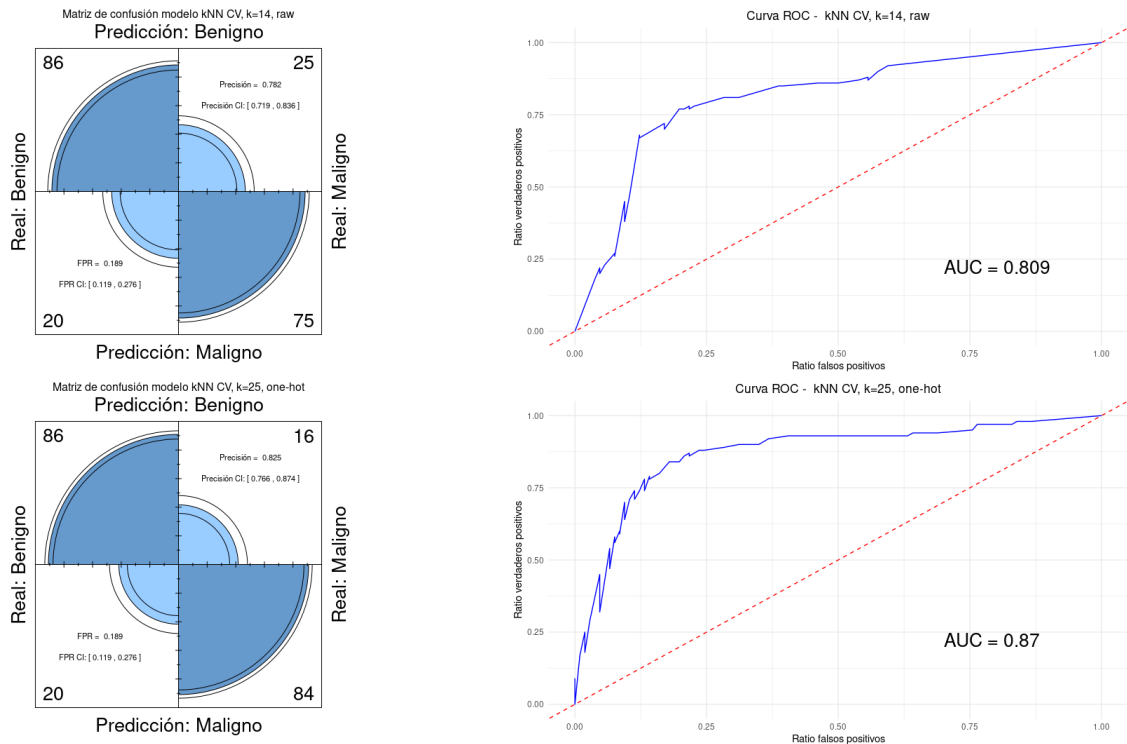


Figura 5.4: Gráficas rendimiento algoritmos kNN sin validación cruzada

Bajo estas premisas generales se ponen de manifiesto varios factores a tener en cuenta para este modelo. En primer lugar, la importancia de la selección de k en este algoritmo y la necesidad de probar diferentes configuraciones para encontrar el valor óptimo, ya que se inició probando con $k = 14$, pero un entrenamiento con validación cruzada indica que el valor que mejor adaptabilidad muestra es $k = 25$. Segundo, el mejor rendimiento del algoritmo con los datos procesados, lo que indica que un conjunto de datos interpretable numéricamente por el algoritmo permite que éste distinga mejor las categorías de las diferentes variables y además, reduce el efecto que puede tener la escala de la edad frente a las variables categóricas. Por último, la efectividad de la validación cruzada, ya que los modelos que utilizan esta técnica obtiene resultados ligeramente superiores. Esto puede indicar que la CV ayudar a mejorar la generalización del modelo, permitiéndole ajustarse mejor a variaciones en los datos y reduciendo el riesgo de sobreajuste. Concluimos así que la estrategia de implementar la codificación one-hot junto con diferentes configuraciones del modelo ha sido la mejor forma de proceder con este algoritmo, obteniendo el modelo que mejor rendimiento consigue y mejor se puede adaptar a distintas configuraciones del conjunto de datos el modelo con la codificación one-hot, validación cruzada y $k = 25$.

5.2.2. SVM

En el caso del algoritmo SVM, debido al mejor rendimiento conseguido con el conjunto de datos procesado y ya que las características de este algoritmo en cuanto a su dependencia a la distancia y la influencia de las categorías, se decidió evaluar el modelo exclusivamente con este conjunto de datos. Se entrenaron en primer lugar un modelo con kernel lineal y un parámetro

de costo $C = 1$ y un modelo con kernel lineal y explorando diversos parámetros de C junto con la validación cruzada, obteniendo el óptimo para $C = 0,1$. Se procedió de manera análoga para comparar estos resultados con un kernel radial, con parámetros óptimos $C = 0,25$, $\sigma = 0,22$ y $C = 0,1$, $\sigma = 0,2$ en los algoritmos sin CV y con CV, respectivamente. La cualidad común que más destaca en estos algoritmos es que, en contrario a lo que se espera teóricamente, los modelos más completos y complejos ofrecen una peor capacidad de predicción que los modelos más sencillos. Las métricas ROC son similares en todos los casos, por lo que vemos en la Figura 5.5 una comparativa de las matrices de confusión para los 4 modelos, destacando los siguientes aspectos:

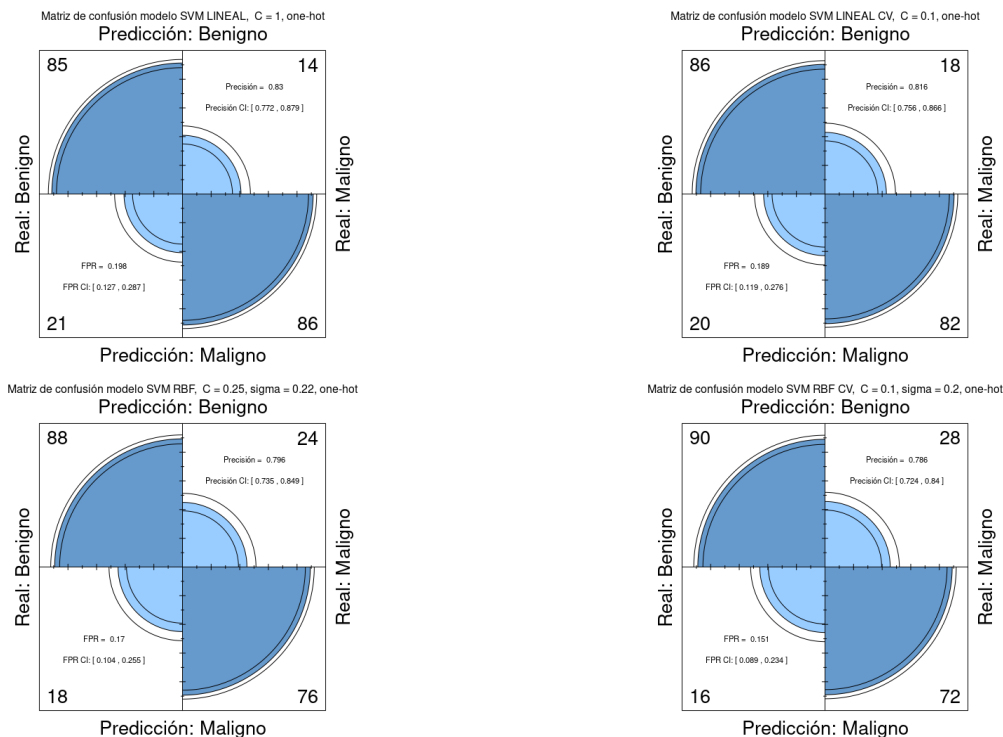


Figura 5.5: Matrices de confusión de algoritmos SVM

- A medida que aumenta la complejidad del modelo se reduce tanto el número de falsos positivos como el número de verdaderos positivos, lo que indica en el primer caso una mejor capacidad de predicción de la clase maligna y en el segundo una tendencia a identificar mejor los casos benignos.
- La cantidad de falsos negativos aumenta considerablemente en modelos complejos, lo cual es preocupante desde un punto de vista clínico, ya que estaríamos clasificando una muestra maligna como benigna.

Del análisis de las matrices de confusión se observa que los modelos con kernel radial son más efectivos identificando casos benignos, a costa de una menor capacidad de predicción para la clase maligna. Esta tendencia puede ser útil en casos donde es crucial minimizar los falsos positivos para diagnósticos benignos, pero teniendo en consideración el aumento de falsos negativos.

Globalmente, el rendimiento obtenido en estos algoritmos con la codificación one-hot es aceptable, resaltando de nuevo la importancia de procesar adecuadamente los datos al problema. Por otro lado, se observan diferencias significativas en el uso de los distintos kernels, ya que los modelos con kernel lineal tienen mejor precisión y mayor capacidad para predecir clases positivas, mientras que los modelos con kernel radial reducen significativamente el porcentaje de falsos positivos. Estos resultados sugieren que quizá una exploración de configuraciones diferentes de parámetros junto con la validación cruzada ofrezca un algoritmo con un mejor equilibrio entre sensibilidad y especificidad.

5.2.3. Naive Bayes

El algoritmo Naive Bayes es capaz de adaptarse con eficacia a conjuntos de datos con distintos tipos de variables, por lo que se ha evaluado el rendimiento con datos crudos y procesados. En primer lugar se entrenó el algoritmo con datos crudos, probando distintas configuraciones de suavizado de Laplace con la validación cruzada y se procedió de manera análoga con los datos procesados. Los rendimientos obtenidos son bastante diversos, por lo que se detallan los resultados de cada modelo:

- El modelo básico con datos crudos muestra un equilibrio razonable en la predicción de ambas clases (84 para la clase negativa y 77 para la positiva), aunque la precisión admite un margen de mejora considerable.
- El modelo óptimo para los datos crudos se obtiene sin suavizado de Laplace, introduciendo el uso de la función kernel (con el parámetro `adjust = TRUE` en el entrenamiento). Muestra una mejora notable en la precisión global y en la especificidad (capacidad de predicción correcta de negativos), reduciendo significativamente el número de falsos positivos.
- El modelo básico con datos procesados es el que peor capacidad de clasificación ofrece. Como se ve en la Figura 5.6 (gráfica superior), tiene un valor de AUC robusto (0.85), pero parece estar sesgado a identificar casos positivos, ya que el modelo tiene una sensibilidad del 80 % y un total de 30 falsos positivos, un ratio de 0.283, una puntuación demasiado elevada para el problema al que nos enfrentamos.
- Los parámetros óptimos con CV para datos procesados son sin suavizado de Laplace y el uso de kernel y ajuste con valor 1 (detalles en [10]). Aunque no es el que mejor precisión ofrece, cuenta con un AUC de 0.876, (Figura 5.6, gráfica inferior) que junto con valores similares para sensibilidad y especificidad es el que mejor equilibrio ofrece para la diferenciación entre ambas clases. Esto junto con su coeficiente Kappa de 0.612 lo hace un algoritmo consistente y fiable.

En términos generales, los algoritmos NB mejoran considerablemente al implementar la validación cruzada, posiblemente debido al uso de las funciones kernel, pues éstas tienen una mayor capacidad para procesar datos que no siguen una distribución normal en estos algoritmos. Destaca en el rendimiento, sin embargo, el modelo con procesamiento de datos y validación cruzada, ofreciendo un gran equilibrio para clasificar ambas clases.

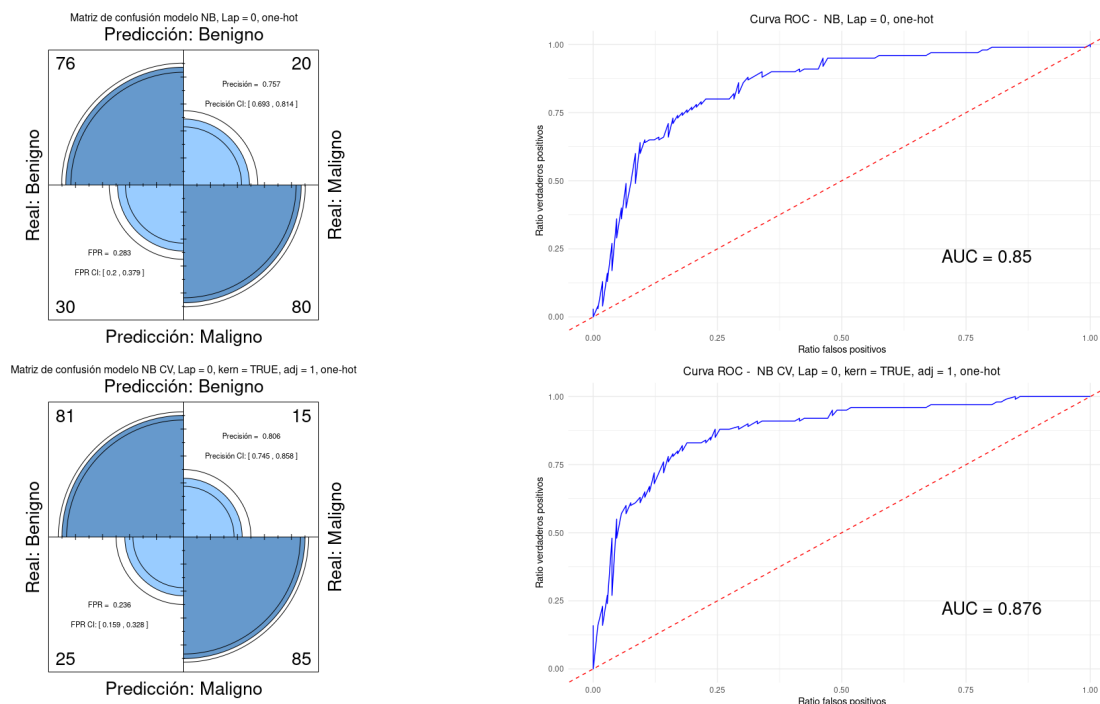


Figura 5.6: Rendimiento de los algoritmos Naive Bayes

5.2.4. Regresión logística

Para este algoritmo se han distinguido dos bloques diferentes. En primer lugar, se entrenan y comparan los resultados del modelo logístico simple, utilizando datos crudos y también datos procesados pero con la matriz de diseño reducida, evitando el problema de la multicolinealidad; y en segundo lugar se utiliza la matriz de diseño de la codificación one-hot completa, aplicando las penalizaciones Ridge y Lasso mencionadas en el capítulo 3.4. A continuación detallamos los resultados obtenidos para el primer bloque de modelos:

- El modelo básico con datos crudos ofrece una precisión aceptable de 0.811. Su valor AUC es fuerte, por lo que tiene una gran capacidad de distinción entre clases. Globalmente sus métricas son mejorables, aunque muestra una gran estabilidad en todas ellas.
- Utilizando la matriz de diseño reducida, se obtiene un rendimiento ligeramente superior, aunque no significativo, siendo bastante similar al modelo básico.
- En el modelo implementando la validación cruzada y datos procesados, aumenta significativamente la capacidad predictora del algoritmo para ambas clases. Un AUC de 0.876 muestra un gran equilibrio en este aspecto, y el valor de kappa de 0.670 aumenta la fiabilidad de los resultados.

En este grupo de modelos observamos la importancia de eliminar la multicolinealidad para conseguir un buen rendimiento en el modelo. En la Figura 5.7 observamos el rendimiento de estos dos modelos, con una muy buena precisión, y un número de falsos positivos y falsos negativos equilibrados, lo que se refleja en la estabilidad de sus curvas ROC. Esto refleja de nuevo la importancia de utilizar un conjunto de datos adecuado, ya que los modelos ofrecen un rendimiento superior.

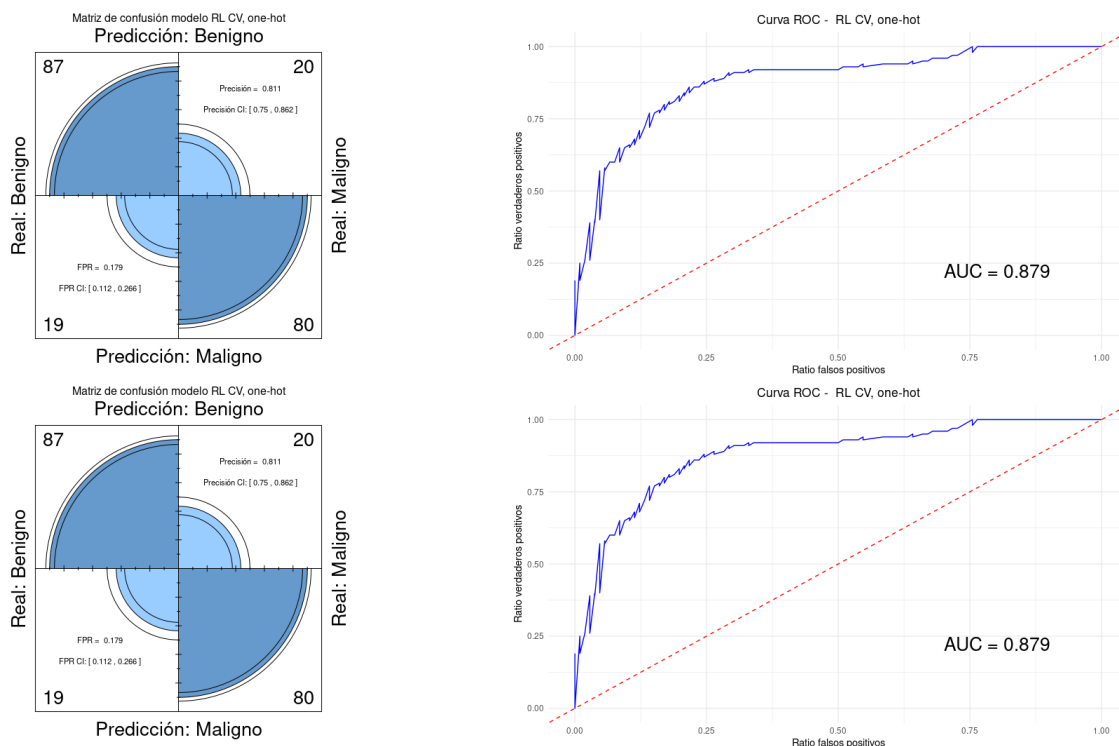


Figura 5.7: Rendimiento de los algoritmos Naive Bayes

Para el segundo bloque de modelos se utilizó la matriz de diseño original, cuyo rango es inferior al número de variables. En estos casos la práctica más habitual es entrenar los algoritmos descritos previamente, aunque se decidió evaluar el rendimiento de estas técnicas ya que intrínsecamente producen una reducción de dimensionalidad. La capacidad predictiva de los modelos se ve reflejada en la Figura 5.7 a través de las matrices de confusión. Aunque los tres modelos ofrecen rendimientos similares, destacando la penalización Lasso por su precisión y elevada especificidad, que lo hará clasificar correctamente los casos malignos. La penalización Ridge no ofrece mejoras destacables, y la penalización *elastic*, cuyo valor óptimo es $\alpha = 0,25$ con la CV, ofrece mejor equilibrio y precisión, pero no mejora en este caso la capacidad predictiva de la penalización Lasso.

En conclusión, los modelos cuyo conjunto de datos teóricamente se adapta mejor al algoritmo, es decir aquellos con rango máximo ofrecen mejores rendimientos en general. Esta elección junto con la validación cruzada, asegurando la adaptabilidad del modelo a diferentes muestras y ofreciendo un gran equilibrio entre la correcta clasificación de clase positiva y negativa. Además, las penalizaciones que ofrece la regresión logística no reducen la colinealidad de manera más efectiva.

5.3. Discusión

En esta sección ofreceremos una comparativa global del rendimiento en todos los modelos, tratando de vislumbrar cual es el modelo óptimo que mejor se ajusta a los datos, sin dejar de lado el contexto del problema que queremos estudiar. El objetivo en este problema es la reducción en

el número de biopsias innecesarias, es decir, más allá de obtener la mejor capacidad predictora, buscamos reducir al máximo el número de falsos positivos y falsos negativos, de la manera más equitativa posible. Debido a que la selección de clase positiva y negativa es arbitraria dependiendo del problema a estudiar, se ha seleccionado en este caso la clase positiva como maligno, de manera que un falso positivo sería un tumor benigno que el modelo clasifica como maligno, y un falso negativo sería un paciente con un tumor canceroso que no se identifica como tal. Para reducir al máximo el número de biopsias innecesarias debemos tratar de minimizar estas cantidades de manera equilibrada.

Nos basaremos para ello tanto en la tabla de rendimiento de los modelos (Tabla 5.3) como en las siguientes descripciones sobre los términos, para ajustar las definiciones de las métricas más importantes en el contexto del problema:

- Sensibilidad: Mide la proporción real de verdaderos positivos, es decir, los casos de cáncer reales que se clasifican correctamente.
- Especificidad: Mide la proporción real de verdaderos negativos, es decir, los tumores benignos que el algoritmo identifica correctamente. Maximizar esta métrica llevaría a un modelo que reduce en gran medida las biopsias innecesarias.
- Valor Predictivo Positivo (PPV): Indica la proporción de casos positivos que son verdaderos positivos. Más allá de la sensibilidad es esencial para garantizar que cuando el modelo predice cáncer, es probable que realmente lo sea.

Realizando una comparación entre algoritmos kNN y SVM, ya que ambos tienen una dependencia del conjunto de variables y su interpretación de manera numérica además de una buena capacidad de adaptación, observamos que el mejor modelo para obtener una buena proporción de falsos negativos y falsos positivos es el algoritmo kNN con $k = 25$, obtenido con CV. Sin embargo en este caso, la sencillez y adaptabilidad del algoritmo no son suficientes para clasificar con una precisión adecuada los datos.

Los modelos resultantes del algoritmo SVM presentan una gran descompensación en estas medidas ya que con el kernel lineal obtenemos un ratio de falsos negativos muy bajos, y en el kernel radial se reducen al máximo los falsos negativos pero provocando que aumenten los falsos positivos. Siendo esto una característica a evitar. Este desajuste en ambos casos hace descartar estos algoritmos como aceptables para la clasificación que se desea obtener.

En los modelos NB obtenemos una situación similar a la que ocurre en SVM. El rendimiento de las distintas configuraciones es bastante diverso, y considerablemente inferior en general comparado con el resto de algoritmos. Se consigue una precisión aceptable del 81.1 % del algoritmo con los datos originales y validación cruzada, pero se percibe una tendencia a identificar la clase negativa (especificidad de 0.849) en detrimento de clasificar correctamente los tumores malignos, con tan sólo 77 casos malignos clasificados correctamente, uno de los que peor funciona en este aspecto. Es probable que la codificación implementada y el conjunto de datos originales no sean los que más fácilmente interpreta el algoritmo, ya que ninguna configuración óptima implementa el suavizado de Laplace, algo habitual para reducir la dimensionalidad. De igual manera que sucede con SVM, no parece que las configuraciones entrenadas rindan de manera óptima con los datos que se disponen.

Los modelos entrenados con algoritmos de regresión logística destacan al equilibrar sensibilidad y especificidad, idóneo para este contexto, además de ofrecer una precisión suficientemente

Alg.	Dat. / CV	Param.	Acc.	Kap.	AUC	Sen.	Esp.	PPV	FDR
kNN	raw	$k = 14$	0.743	0.483	0.784	0.67	0.811	0.770	0.189
kNN	one-hot	$k = 14$	0.825	0.651	0.859	0.84	0.811	0.808	0.189
kNN	raw, CV	$k = 14$	0.782	0.562	0.809	0.75	0.811	0.789	0.189
kNN	one-hot, CV	$k = 25$	0.825	0.651	0.870	0.84	0.811	0.808	0.189
SVM LIN	one-hot	$C = 1$	0.830	0.661	0.872	0.86	0.802	0.804	0.198
SVM LIN	one-hot, CV	$C = 0,1$	0.816	0.631	0.873	0.82	0.811	0.804	0.189
SVM RAD	one-hot	$C = 0,25,$ $\sigma = 0,22$	0.796	0.591	0.870	0.76	0.830	0.809	0.170
SVM RAD	one-hot, CV	$C = 0,1,$ $\sigma = 0,2$	0.786	0.571	0.851	0.72	0.849	0.818	0.151
RL	raw, CV	NA	0.811	0.621	0.879	0.80	0.821	0.808	0.179
RL	one-hot	NA	0.816	0.631	0.878	0.81	0.821	0.810	0.179
RL	one-hot, CV	NA	0.835	0.670	0.876	0.84	0.830	0.824	0.170
RL	one-hot, CV	$\alpha = 1,$ $\lambda = 0,023$	0.835	0.670	0.877	0.83	0.840	0.830	0.160
RL	one-hot, CV	$\alpha = 0,$ $\lambda = 0,023$	0.816	0.631	0.878	0.81	0.821	0.810	0.179
RL	one-hot, CV	$\alpha = 0,25,$ $\lambda = 0,1$	0.830	0.660	0.878	0.83	0.830	0.822	0.170
NB	raw	lap=0	0.782	0.563	0.803	0.77	0.792	0.778	0.208
NB	raw, CV	lap=0, adj- kern=2	0.811	0.620	0.875	0.77	0.849	0.828	0.151
NB	one-hot	lap=0	0.757	0.516	0.850	0.80	0.717	0.727	0.283
NB	one-hot	lap=0, adj.kern=1	0.806	0.612	0.876	0.85	0.764	0.773	0.236

Cuadro 5.3: Tabla con las métricas para todos los modelos¹

buena. Ofrecen todos ellos un rendimiento superior global que en el resto de algoritmos, lo que establece una gran consistencia de este algoritmo sobre diferentes conjuntos de datos y configuraciones. El modelo con datos crudos tiene métricas aceptable en comparación al resto de algoritmos con datos sin procesar, lo que indica que la RL se ve menos afectadas por problemas de escala o diferencias significativas en la distribución de las categorías. En definitiva, los algoritmos de RL se presentan como la alternativa más viable para este estudio ya que, más allá de su justificación estadística, favorecen la detección de tumores malignos correctamente (alta sensibilidad) y son efectivos descartando tumores benignos (especificidad), reduciendo así el número de intervenciones o pruebas adicionales necesarias. Como valoración global y teniendo en cuenta todas las métricas y análisis, el modelo de Regresión Logística con datos procesados

¹ Alg. (algoritmo) - Dat. (conjunto de datos utilizado) - Param. (parámetros del modelo) - Acc. (precisión) - Kap. (kappa) - Sen. (sensibilidad) - Esp. (especificidad) - FDR (Ratio de Descubrimiento Falso) - raw (datos crudos, sin procesar)

y validación cruzada emerge como el más adecuado para este problema. Este no solo ofrece una precisión elevada sino que también logra un excelente equilibrio entre sensibilidad y especificidad. La alta sensibilidad garantiza que los casos reales de cáncer se identifiquen correctamente, mientras que la alta especificidad reduce el número de biopsias innecesarias al evitar falsos positivos. Además, un PPV alto en este modelo aumenta la confianza en que las predicciones positivas son verdaderamente casos de cáncer, contribuyendo de igual manera a reducir el número de biopsias.

Este modelo es particularmente adecuado en el contexto clínico para la detección del cáncer de mama, donde la precisión en la identificación de casos malignos y la minimización de intervenciones innecesarias son igualmente críticas. La implementación de la validación cruzada asegura la robustez y la generalización del modelo a nuevas muestras de datos, mientras que la codificación one-hot facilita una interpretación más efectiva de las variables por parte del modelo. En conclusión, la selección de la Regresión Logística con estas características se justifica plenamente tanto por su rendimiento estadístico como por su relevancia en el contexto clínico, subrayando la importancia de un enfoque meticuloso en la selección de técnicas de pre-procesamiento, validación y elección de modelo para optimizar el rendimiento en aplicaciones médicas.

5.4. Aplicación web

En esta sección mostraremos la aplicación web desarrollada, que supone el producto final del TFM. Veremos las funcionalidades que presenta, cómo usarla y qué alternativas ofrece al usuario. La aplicación se encuentra alojada en el servidor de ShinyApps, pudiendo acceder a ella en el siguiente enlace: <https://albertoreyabelaira.shinyapps.io/breastcancerprediction/>. Está diseñada con el propósito principal de facilitar la predicción de cáncer de mama a los profesionales médicos, utilizando un modelo avanzado de regresión logística para analizar y clasificar características de mamografías. Es el producto de este TFM, que se centra en aplicar técnicas avanzadas de análisis de datos en el campo de la biomedicina. Su desarrollo demuestra prácticamente cómo los modelos de aprendizaje automático, en este caso, regresión logística, pueden ser utilizados para facilitar decisiones clínicas importantes, como la predicción de la malignidad de masas mamográficas.

La aplicación está estructurada en tres pestañas, entre las cuales se puede navegar usando el panel superior y seleccionando aquella que se desea visualizar (Figura 5.8), facilitando una experiencia de usuario fluida y eficiente. La primera pestaña, que se visualiza al acceder a la aplicación, contiene información importante, incluyendo la funcionalidad básica de la aplicación y sus usos; el modelo que lleva implementado, el rendimiento que se obtuvo entrenándolo y qué conjunto de datos se utilizó, la referencia al repositorio donde se aloja el trabajo, las gráficas de rendimiento de modelo y los datos del creador y contacto. Esta pestaña se muestra al usuario como en la Figura 5.8, que vemos a continuación:

Para facilitar su visualización, el resto de la pestaña se muestra en la Figura 5.9, que incluye una tabla de valores con la descripción de los parámetros de entrada. La última funcionalidad de esta pestaña es la ventana emergente de la Figura 5.10 que obtenemos pulsando el botón *aquí* de la Figura 5.9 y muestra al usuario los pasos a seguir para utilizar la aplicación correctamente y obtener una nueva predicción.

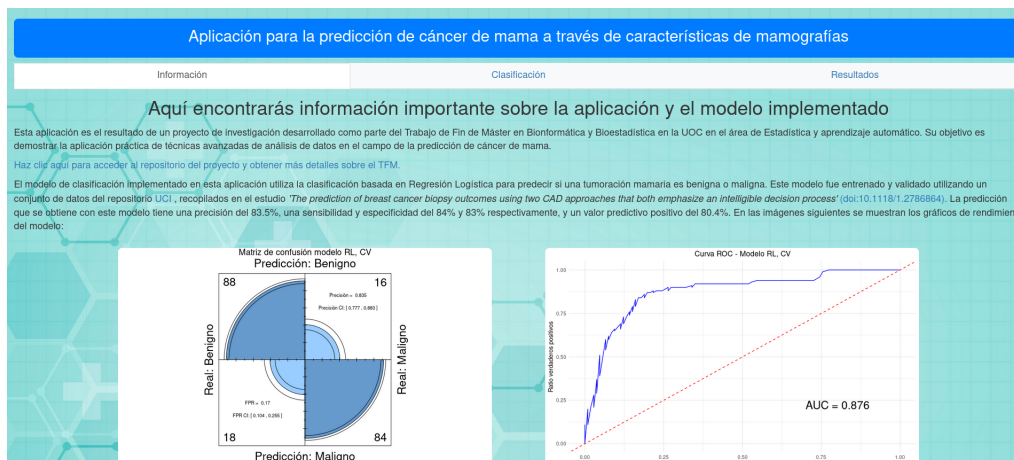


Figura 5.8: Pestaña principal de la aplicación

A continuación tenemos la pestaña principal y central de la aplicación, que corresponde a la clasificación del modelo. La funcionalidad de esta ventana es permitir al usuario introducir los datos para obtener la predicción que ofrece el modelo. Para ello cuenta con un panel para introducir los parámetros de entrada, donde seleccionamos el valor para las variables categóricas de un menú desplegable con los correspondientes a cada variable. Para la variable edad permite introducir un valor numérico manualmente, o seleccionar con las flechas de navegación los valores directamente inferior y superior. Este panel con diferentes casos lo vemos en la Figura 5.12 a continuación.

El último paso que queda es obtener la predicción, para lo cual debemos pulsar en el botón *Obtener clasificación* de la Figura 5.12. A continuación se presentan dos escenarios:

- Si se introducen correctamente los datos para todas las variables, se abrirá automáticamente la pestaña de Resultados (Figura 5.13) donde podremos ver el resultado de la predicción. Además de ésta, se muestran al usuario los parámetros de entrada con los que se calcula la predicción, para poder asegurar que son correctos.
- En caso de que alguno de los parámetros de entrada no sean correctos, bien porque no se ha seleccionado un valor para las variables categóricas o bien porque el valor de la edad está fuera del rango, se abrirá una ventana emergente como en la Figura 5.14 para informar al usuario de este problema.

Haz clic [aquí](#) para detalles sobre cómo obtener una predicción. En la siguiente tabla puedes consultar los parámetros de entrada.

Variable	Descripción	Tipo	Rango
Edad	Edad del sujeto	Númerica	18 - 96
Forma	Forma de la tumoración	Categórica	1 - 4
Margen	Margen de la tumoración	Categórica	1 - 5
Densidad	Densidad de la tumoración	Categórica	1 - 4

© 2024, Alberto Rey Abela - albertorey@uoc.edu

Figura 5.9: Tabla de valores

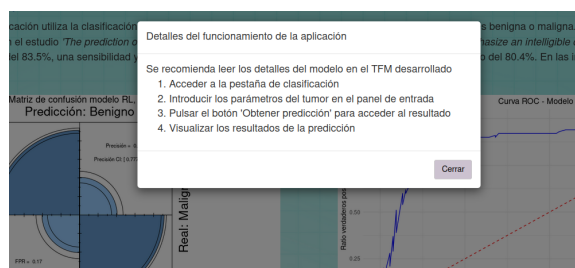


Figura 5.10: Ventana emergente

Figura 5.11: Resto de funcionalidades pestaña principal

Figura 5.12: Paneles de entrada de datos con diferentes valores

Figura 5.13: Resultado para valores correctos

Figura 5.14: Entrada incorrecta

Figura 5.15: Posibles resultados al obtener una nueva clasificación

Finalmente, para que la navegación y la obtención de una nueva predicción sea lo más intuitiva posible, una vez se obtiene la predicción y se accede de nuevo a la pestaña Clasificación, el panel de entrada de datos se encuentra limpio y preparado para introducir de nuevo los datos. Además, la pestaña de Resultados no se actualiza mientras no se obtenga una nueva predicción, facilitando al usuario el historial previo de uso.

La aplicación web se ha desarrollado de manera que resulte lo más natural, intuitiva y funcional posible en el contexto de obtener una clasificación. La predicción se obtiene de manera reactiva al pulsar el botón de la pestaña de Clasificación, mostrando automáticamente el resultado de la predicción. El equilibrio que supone la implementación de una técnica de análisis de datos avanzada junto con una interfaz sencilla e intuitiva permite que usuarios de diferentes niveles de habilidad técnica utilicen la herramienta eficazmente, acercando la ayuda que supone en la predicción de cáncer de mama a un mayor rango de profesionales de la salud.

Por último, a pesar de que la aplicación implementa un modelo con una capacidad de predicción considerable, no debe ser utilizada en ningún caso como diagnóstico médico profesional y, en aquellos casos que se desee utilizar por especialistas en la materia, se debe tomar como un consejo añadido al resto de pruebas y parámetros disponibles para ofrecer un diagnóstico (ofreciendo esta información en la pestaña de Resultados, Figura 5.13). Por último, los datos ingresados en la aplicación no se almacenan ni se utilizan para otros fines, respetando así la privacidad y confidencialidad de los pacientes.

Capítulo 6

Conclusiones y trabajos futuros

En este proyecto se abordó la tarea de predecir la malignidad de los tumores mamarios aplicando algoritmos de aprendizaje automático a las características que se pueden extraer del estudio de las mamografías, suponiendo un desafío constante en la medicina y la salud pública. Las conclusiones más significativas que podemos extraer con las siguientes:

- **Papel clave del tratamiento de los datos:** El preprocesado de datos, que incluye la codificación one-hot para variables categóricas junto con la normalización de la edad ha demostrado ser crucial en la mejora del rendimiento de los modelos, especialmente en el modelo óptimo seleccionado.
- **Selección óptima del modelo:** Enfocándonos en el objetivo del proyecto y la naturaleza del problema, la regresión logística ha emergido como el método de clasificación más adecuado. Este modelo no solo proporciona una alta precisión en la predicción del cáncer de mama, sino que también equilibra de manera óptima la detección de tumores malignos y benignos, un aspecto crítico en el ámbito clínico.
- **Aplicación práctica en el contexto clínico:** La creación de una aplicación web interactiva proporciona a los profesionales de la salud una herramienta práctica para la toma de decisiones informadas en la detección del cáncer de mama, resaltando el valor de la tecnología en entornos clínicos.
- **Eficacia en la predicción de cáncer de mama:** Los resultados generales de este trabajo han demostrado la aplicabilidad y eficacia de los métodos de aprendizaje automático en la detección y clasificación de masas mamográficas, resaltando la importancia de estas técnicas en el ámbito de la medicina predictiva.
- **Responsabilidad ética y consciencia de las limitaciones:** A pesar de la eficacia del modelo y la aplicación, es crucial reconocer sus limitaciones. La herramienta no reemplaza en ningún contexto el juicio clínico profesional y debe utilizarse exclusivamente como un complemento a las pruebas y evaluaciones médicas estándar.

Relizando una valoración global del grado de cumplimiento de los objetivos del proyecto (capítulo 1.2), se ha alcanzado completamente el objetivo general con el desarrollo de la aplicación web y la implementación en esta del modelo considerado como óptimo para el problema. En cuanto a los objetivos específicos, se han cumplido los dos segundos en su totalidad, ya que se engloban en el objetivo general. El primer objetivo específico no se puede determinar si se

ha cumplido al 100 %, ya que quizá haya técnicas que no se han explorado en este trabajo que resultan más adecuadas para la predicción de cáncer de mama.

Siguiendo esta línea de exploración de modelos, junto con otros aspectos que surgieron a lo largo del proyecto, se establecen las siguientes líneas de futuro:

- **Alternativas en el conjunto de datos:** Investigar con conjuntos de datos más amplios y diversificados podría mejorar aún más la robustez del modelo. Además, la correlación fuerte entre las variables de forma y margen del tumor, junto con la baja relación de la densidad con todas las variables, sugiere la posibilidad de que un modelo que prediga el cáncer basándose en la edad, la forma y el margen del tumor podría ser igualmente eficaz.
- **Evaluación clínica y validación:** Realizar estudios clínicos para validar la eficacia de la aplicación en entornos reales sería un paso esencial para su adaptación al entorno profesional.
- **Incorporación de nuevas técnicas de análisis y variables:** Explorar la inclusión de otras variables clínicas relevantes y la aplicación de técnicas analíticas más avanzadas, como el aprendizaje profundo, podría proporcionar una comprensión más profunda y mejorar la precisión del modelo.

Como reflexión final y valoración global del proyecto, se obtiene un balance positivo en los resultados. A pesar de que el rendimiento en el modelo de Regresión Logística es más que aceptable, aflora la posibilidad de que el conjunto de datos sea algo escaso para el problema a tratar, ya que los resultados de la clasificación son similares en muchos modelos. La principal consecuencia de esto es la duda razonable de haber estudiado los algoritmos con las configuraciones más adecuadas para los datos que se disponían. Aún así, los resultados obtenidos son reseñables y ponen de manifiesto las sutilezas, entresijos y potencial del aprendizaje automático.

Capítulo 7

Bibliografía

- [1] Organización Mundial de la Salud. *Cáncer de mama*. World Health Organization. <https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer>
- [2] Budh, D. P., & Sapra, A. (2022). *Breast Cancer Screening*. [Updated 2022 Oct 22]. In: *StatPearls [Internet]*. Treasure Island (FL): StatPearls.
- [3] Elter, M. R., & Schulz-Wendtland, T. W. (2007). *The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process*. *Med. Phys.*, **34**, 4164–4172.
- [4] American College of Radiology (ACR) (2003). *Breast Imaging Reporting and Data System Atlas (BIRADS Atlas)*. Reston, Va: American College of Radiology.
- [5] Lantz, B. (2015). *Machine Learning with R*. Birmingham: PACK Publishing.
- [6] Naciones Unidas. (s/f). *Objetivos de Desarrollo Sostenible (ODS) - Naciones Unidas*. Recuperado de <https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/>
- [7] Elter, M. (2007). *Mammographic Mass*. UCI Machine Learning Repository. <https://doi.org/10.24432/C53K6Z>
- [8] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2021. <https://www.R-project.org/>
- [9] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC., Boston, MA, 2020. <http://www.rstudio.com/>.
- [10] Max Kuhn. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software*, 28(5):1–26, 2008. DOI: 10.18637/jss.v028.i05. URL <https://www.jstatsoft.org/index.php/jss/article/view/v028i05>.
- [11] Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., ... Borges, B. (2023). *Shiny: Web Application Framework for R*. R package version 1.7.4.9002.

- [12] Simonyan, K. y Zisserman, A. (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556*.
- [13] McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafi, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, 577(7788), 89–94.
- [14] Bennett, K., & Campbell, C. (2000, diciembre). *Support Vector Machines: Hype or Hallelujah?* *ACM SIGKDD Explorations Newsletter, 2*, 1-13. <https://doi.org/10.1145/380995.380999>
- [15] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. USA: Springer.
- [16] Harrell, F. E., Jr. (2001). *Regression Modeling Strategies*. USA: Springer.
- [17] Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning*. USA: Springer.
- [18] Hastie, T., et al. (2013). *An Introduction to Statistical Learning: with applications in R*. USA: Springer.
- [19] Zheng, Alice y Casari, Amanda. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media. ISBN: 978-1491953242.
- [20] Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.