El cáncer de mama. Análisis, representación y clasificación según varios modelos de Machine Learning sobre una base de datos.

Universidad Miguel Hernández de Elche

Facultad de Ciencias Sociales y Jurídicas de Elche

Grado en Estadística Empresarial

Trabajo de Fin de Grado

Alumno: Jose Javier Lucena Muñoz

Tutora: María Asunción Martínez Mayoral

Índice de contenidos

1. Resumen

Este informe estadístico se centra en el análisis de una base de datos que contiene información sobre las características de células tumorales mamarias. El objetivo principal es aplicar diversas técnicas de Machine Learning para clasificar con precisión el diagnóstico final de un tumor como benigno o maligno, en función del resto de características recopiladas, y comparar los resultados obtenidos, concluyendo sobre la mejor técnica para la predicción. Previamente se realiza un análisis descriptivo de la base de datos para conseguir un conocimiento pleno de cada una de las variables recopiladas en la muestra.

Después de obtener los resultados de ambos modelos de aprendizaje automático propuestos, se comparan respecto de diversas métricas de clasificación, para recomendar finalmente cuál proporciona mejores predicciones.

Este trabajo ha supuesto una primera toma de contacto con el análisis estadístico, y en particular los modelos de machine learning, desarrollados con el lenguaje de programación Python, que no se ha trabajado en el plan de estudios del grado.

2.Palabras clave

Las palabras clave que abarcan la generalidad del proyecto son: *cáncer de mama, tumor maligno, clasificación, estadística descriptiva, aprendizaje automático, regresión logística, Naïve Bayes, modelos de clasificación, Python, Machine-Learning*.

3. Objetivos

En este apartado presentamos los antecedentes de la base de datos utilizada en nuestro estudio y planteamos los objetivos sobre los que trabajamos.

3.1 Antecedentes

La base de datos trabajada en este informe ha sido objeto de estudio en numerosas ocasiones por parte de muchos equipos investigadores, dada la relevancia de la prevención del cáncer de mama y su detección temprana para evitar que se pueda convertir en una enfermedad mortal. En la base de datos “Wisconsin Diagnostic Breast Cancer (WDBC)”, creada conjuntamente por la Universidad de Wisconsin y el Centro de Ciencias Clínicas Madison (University of Wisconsin & Madison Clinical Sciences Center, 1996) se recopilaron un buen número de registros de tumores con información completa de diversas características de los tumores tomadas en diferentes secciones.

Algunos de los trabajos en los que se ha estudiado dicha base de datos, han permitido avanzar mucho en el campo de la prevención, permitiendo que actualmente la tasa de supervivencia a 10 años del cáncer de mama se encuentre, mundialmente, en el 84% de las afectadas.

Destacamos los siguientes estudios sobre esta base de datos:

- *Diagnosis of Breast Cancer Pathology on the Wisconsin Dataset with the Help of Data Mining Classification and Clustering Techniques*, de Singh et al. (2022), que realiza la clasificación de tumores a través de la minería de datos y de técnicas de clustering .
- *Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification*, de Haque et al. (2022), que propone un algoritmo de aprendizaje automático a través de los métodos de redes neuronales (perceptrón multicapa), k-vecinos cercanos, bosques aleatorios y programación genética .
- Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques, de Yazici et al.(2022), que aplica modelos de clasificación usando

las técnicas de bosques aleatorios, potenciación del gradiente, máquinas de vector soporte (SVM), redes neuronales artificiales y perceptrón multicapa.

3.2 Objetivos generales

El objetivo general de nuestro análisis consiste en la construcción de modelos de aprendizaje automático sobre la base de datos para predecir la clasificación correcta del tumor como benigno o maligno a partir del resto de variables observadas para cada tumor. Se propone además, la comparación entre los diversos modelos ajustados, para extraer conclusiones sobre cuál es más recomendable, proporcionando mayor tasa de aciertos, en esta base de datos.

Si bien los modelos que se van a aplicar ya se han aplicado por otros investigadores, el plantearlo en este trabajo viene justificado como una primera aproximación del estudiante al aprendizaje automático con Python, lenguaje en el que no se ha trabajado en estadística a lo largo del grado en Estadística Empresarial. Supone pues un reto en sí mismo por el dominio del lenguaje y de técnicas estadísticas que tampoco se han trabajado, al menos desde la perspectiva del aprendizaje automático, en asignatura alguna del grado.

3.3 Objetivos específicos

En cuanto a los objetivos específicos que planteamos en este estudio, destacamos:

1. Familiarizarse con Python y los cuadernos de Google Colab, con los que se ha programado este trabajo.
2. Resolver el preprocesado de los datos, adaptado al tipo de datos y tipo de modelos propuestos para realizar la clasificación de tumores.
3. Describir la base de datos, respondiendo a la tipología de cada una de las variables disponibles.

4. Aplicar el modelo de regresión logística, seleccionar las variables que explican la clasificación del tumor y extraer conclusiones sobre la bondad de dicha clasificación.
5. Aplicar el clasificador de Naïve-Bayes y extraer conclusiones sobre la bondad de la clasificación.
6. Comparar los dos modelos ajustados y extraer conclusiones.

4. Información disponible

La base de datos utilizada en este estudio fue localizada en la web de Kaggle, tiene por título "Breast Cancer Dataset", y se puede acceder desde la URL https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset (*Breast Cancer Dataset*, n.d.). Esta base de datos es de dominio público (CC0), la fecha de origen de la recogida de información data de enero de 1996; fue recopilada y actualizada por Yasser H. (2022). Ha sido analizada en múltiples ocasiones, como se aprecia en las contribuciones detalladas en los Antecedentes (en este trabajo) y también publicadas en Kaggle, con la finalidad, siempre, de encontrar un método óptimo para la clasificación de tumores a partir de la información disponible.

Contiene un total de 569 de registros de tumores de cáncer de mama en mujeres, clasificados como benignos y malignos, y un total de 10 variables numéricas relativas a características fisiológicas de los tumores (en la Tabla 1), medidas en diferentes partes de los mismos, y que se han resumido a modo de promedios (identificados como *nombrevariable_mean*), desviaciones típicas (*nombrevariable_se*) y peor valor o el más desfavorable hacia el diagnóstico (*nombrevariable_worst*). La base de datos contiene pues, un total de 32 columnas, de las cuales una se trata de un identificador anonimizado de los registros, otra al tipo de tumor y las restantes a los 3 descriptivos (media, desviación típica y peor valor) de cada una de las 10 características medidas.

Los nombres de las variables se muestran en la Tabla 1 con su nomenclatura original en la base de datos (en inglés), si bien a lo largo de este estudio han sido traducidos al español, especialmente cuando se proporcionan las discusiones de los resultados.

Tabla 1: Información disponible en la base de datos. Nombre de variables, tipología (categórica/numérica) y descripción.

Nombre	Tipo	Descripción
ID	Numérico	Documento identificativo de la paciente.
Diagnosis	Carácteres/ Variable dicotómica	Variable dicotómica que indica M si el diagnóstico tumoral es maligno y B si es benigno.
radius	Numérico	Radio de los lóbulos. (Distancia desde el centro hasta los puntos en el perímetro).
texture	Numérico	Textura de la superficie.
perimeter	Numérico	Perímetro de los lóbulos.
area	Numérico	Área de los lóbulos.
smoothness	Numérico	La lisura o suavidad de un contorno nuclear se cuantifica midiendo la diferencia entre la longitud de una línea radial y la longitud media de las líneas que lo rodean.
compactness	Numérico	Compacidad o ratio entre el volumen y la superficie. (perímetro²/area) -1
concavity	Numérico	La concavidad se obtiene al medir el número y la gravedad de las concavidades o hendiduras en un núcleo celular. Para ello, se dibujan puntos no adyacentes en la snake y se mide el grado en que el límite real del núcleo se encuentra en el interior de cada línea.
concave points	Numérico	Número de porciones cóncavas del contorno.
symmetry	Numérico	Similitud entre partes con respecto a ejes.

fractal_dimension	Numérico	Índice comparativo sobre el detalle de un patrón observado de células.

5. Metodología

En este apartado describimos los procedimientos utilizados en las diferentes etapas del análisis, desde el procesado hasta el ajuste con los modelos de aprendizaje automático utilizados, pasando por el análisis descriptivo realizado.

5.1 Preprocesado de los datos

El preprocesado de datos engloba todas las tareas relacionadas con el tratamiento de la base de datos en bruto, con la finalidad de convertirla en una base de datos eficiente y fácil de utilizar para su análisis estadístico con modelos de aprendizaje automático, y en particular en nuestro caso, con modelos de Machine Learning. En nuestro caso, dado que la base de datos ha sido trabajada de modo extensivo y depurada previamente, el preprocesado de datos consistirá en:

- identificar la existencia de valores faltantes, para en su caso, imputar valores o eliminar registros; en nuestro caso los valores faltantes ya habían sido tratados, y la base de datos no contenía valores faltantes;
- estandarización de las variables numéricas (básicamente todas las registradas salvo la respuesta), con el fin de trasladarlas todas a una escala común y así poder abordar de modo apropiado los modelos de aprendizaje;
- creación de variables dummy para las variables categóricas, esto es, para la variable respuesta que identifica el tipo de tumor;
- división de datos en muestras de entrenamiento (sobre los que ajustar el modelo) y test (para verificar la calidad del ajuste).

Respecto a la identificación de valores faltantes, utilizamos el comando *datos.isnull().sum()*. El resultado, como se comentó anteriormente, es nulo: no se encuentra ningún valor faltante. De hecho, la base de datos contiene todos los valores para los 569 registros de las 32 variables disponibles.

El proceso de estandarización consiste en transformar los datos de tipo numérico, para centrarlos en su media (eliminando así el valor medio de cada característica) y escalarlos dividiendo por su desviación estándar. Este proceso es necesario para aplicar de un modo eficiente las técnicas de clasificación automática, dado que las variables se han medido inicialmente en escalas dispares y no comparables. Esta estandarización se resuelve básicamente con la función *StandardScaler()* de la librería *Scikit-Learn*.

Para tratar con técnicas de clasificación automática la variable respuesta que identifica el tipo de tumor (maligno/benigno), hemos de crear una variable *dummy* numérica, a la que asignamos el valor 1 para los tumores malignos y 0 para los benignos. Este proceso se resuelve con la función *OneHotEncoder()* de la librería *Scikit-Learn*.

Una vez transformadas las variables con las que resolver el análisis, las almacenamos en una nueva base de datos que contiene la información en el formato necesario para abordar los análisis posteriores, y a la que accedemos ya directamente para llevarlos a cabo (disponible en *datos-estandarizados.csv*).

Para ajustar un modelo de Machine Learning basado en clasificación, hemos de dividir la base de datos en muestras de entrenamiento, con la que ajustamos el modelo, y de test, con la que testamos la calidad del mismo. Consideramos una partición aleatoria de los datos a razón de una proporción 75%-25%, respectivamente, para las muestras de entrenamiento y test. Esto se resuelve con la función train_test_split() de la librería Scikit-Learn.

5.2 Análisis descriptivo

El análisis descriptivo de la variable objetivo que identifica el tipo de tumor se resuelve mediante una tabla de frecuencias y un gráfico de barras que muestra el porcentaje de tumores de cada tipo.

El tratamiento descriptivo de las variables numéricas consiste en obtener como descriptivos numéricos el *mínimo, percentil 25, mediana, media, percentil 75 y máximo,* utilizando las mediciones realizadas en la escala original. Puesto que el objetivo principal del estudio es la diferenciación de los tumores benignos y malignos, estos descriptivos los calculamos de forma diferenciada para ambos tipos de tumor. El análisis descriptivo gráfico lo resolvemos a partir de diagramas de cajas, también diferenciando por tipo de tumor, para cada una de las características medidas. Estos gráficos se muestran por variable, agrupando para cada característica observada, los descriptivos para las medias, errores típicos y peores valores.

5.3 Modelos de clasificación

Con el objetivo de clasificar los tumores como benignos o malignos en función del resto de características observadas, existen diferentes métodos y muy variados: desde modelos de regresión hasta algoritmos sofisticados de aprendizaje automático en Machine Learning e incluso Deep Learning (redes neuronales). No obstante, nos centramos exclusivamente en dos técnicas englobadas dentro de lo que se denomina Machine Learning.

El primero a utilizar será el modelo de regresión logística, único modelo de regresión de entre los clasificadores que veremos. Seguidamente analizaremos los datos con el clasificador de Naïve Bayes.

5.3.1 Modelo logístico

En este tipo de modelos disponemos de una variable respuesta y, de tipo cualitativo con dos posibles respuestas que se codifican habitualmente con 0-1, donde el 0 indica "fracaso", en nuestro caso tumor Benigno, y el 1 indica la existencia de tumor Maligno. Contamos además, con una matriz X de variables predictoras, de tipo numérico y/o categórico (Borrás et al., 2023).

La información disponible pues, con p posibles predictores, viene representada por:

$$\{(y_i, x_{1i}, \ldots, x_{pi}),\ i = 1, \ldots, n\}$$

donde x_{ji} es el valor de la muestra i en la predictora j e y_i el valor de la respuesta (0/1) de la muestra i. En esta situación, intentamos relacionar la respuesta y con los predictores a través de un predictor lineal:

$$z = w_0 + w_1 x_1 + \ldots + w_p x_p,$$

donde cada w_j representa la pendiente o variación del predictor lineal con respecto a cada predictora, y w_0 representa el sesgo del modelo. Dado que el valor del predictor lineal z en general se mueve en el rango $(-\infty, +\infty)$, resulta necesario encontrar alguna transformación de la respuesta y que varíe en dicho rango.

Al tratarse la respuesta de una variable que toma valores 0/1, la modelización base es la que proporciona el modelo Bernoulli para explicar su variabilidad, con p la probabilidad de éxito, esto es, $p = Pr(y = 1)$:

$$y \sim Br(p)$$

El objetivo de la predicción lineal se convierte pues en conseguir estimar el valor esperado de la respuesta, esto es, p, en función de un predictor lineal z, construido con la matriz X. Ahora bien, puesto que $p \in [0, 1]$, y $z \in (-\infty, +\infty)$, se requiere una transformación que permita relacionarlos unívocamente, $z = f(p)$. Disponemos de

varias funciones matemáticas que nos permiten establecer una relación biunívoca, y una de ellas es la función logit, que viene dada por:

$$logit(p) = log(\frac{p}{1-p}) \in (-\infty, +\infty), \text{ para } p \in [0, 1]$$

Surge así el modelo logístico, que relaciona una función (logit) de la respuesta esperada, $E(y)$, con un predictor lineal z:

$$logit(p) = z = w_0 + w_1 X_1 + \ldots + w_p X_p$$

Inmediatamente, una vez ajustado este modelo, podemos recuperar la predicción de la probabilidad de tumor maligno $p = Pr(y = 1)$, mediante la función logística, dada por:

$$P(y = 1|X) = \frac{e^{w_0 + w_1 X_1 + \ldots + w_p X_p}}{1 + e^{w_0 + w_1 X_1 + \ldots + w_p X_p}}$$

Así pues, los odds a favor de un tumor maligno, esto es, cuántas veces es más probable un tumor maligno que uno benigno, o lo que es lo mismo, el riesgo de tumor maligno, se obtienen, en escala logarítmica (denominados entonces log-odds) como:

$$log(\frac{p}{1-p}) = w_0 + w_1 X_1 + \ldots + w_p X_p$$

Los principales elementos que hay que interpretar en un modelo de regresión logística son los siguientes coeficientes de los predictores:

- w_0 es la ordenada en el origen o interceptación. Se corresponde con el valor esperado del logaritmo de los odds cuando todos los predictores son cero.
- w_p son los coeficientes de regresión parcial de cada predictor e indican el cambio medio del logaritmo de los odds al incrementar en una unidad la variable predictora, manteniéndose constantes el resto de variables. Dado que la relación entre la probabilidad condicional y las predictoras no es lineal, los coeficientes de regresión no se corresponden con el cambio en la probabilidad

de la respuesta asociada con el incremento en una unidad de la predictora, sino con el cambio en el log-odds.

Como ocurría en los modelos lineales, la magnitud de cada coeficiente parcial de regresión depende de las unidades en las que se mida la variable predictora a la que corresponde, por lo que su magnitud no está asociada con la importancia de cada predictor. Sin embargo, para nuestro modelo con predictoras ya estandarizadas, cada coeficiente sí se interpreta como una magnitud asociada a la regresión. Esta "importancia" de cada coeficiente actúa directamente sobre la variable respuesta.

Una vez ajustado un modelo y la precisión (o error) obtenida sobre la muestra de entrenamiento, continuamos con el proceso denominado regularización, a través del cual buscamos penalizaciones a aplicar a los predictores para reajustar o corregir la correlación existente entre ellos, y en consecuencia identificar cuáles contribuyen de modo relevante a la clasificación, y reducir el peso en el modelo de aquellas cuya información ya está siendo compartida (aportada) por otros regresores.

Es interesante aludir al término de *parsimonia*, el cual hace referencia a que, el mejor modelo, es aquel capaz de explicar con mayor precisión la variabilidad observada en la variable respuesta empleando el menor número de predictores, por lo tanto, con menos asunciones.

A continuación, explicamos el método de regularización para seleccionar las variables del modelo final.

5.3.1.1 Regularización para la selección de variables

La regularización es un procedimiento común en aprendizaje automático y consiste en aplicar una serie de penalizaciones sobre los predictores, para reducir el efecto de la correlación entre ellos. Trabajamos con tres tipos de penalización que describimos a continuación:

- L1 o LASSO

- L2 o Ridge
- Elasticnet.

La regularización basada en la penalización **L1** o **LASSO** (Least Absolute Shrinkage and Selection Operator) introduce el "valor absoluto de magnitud" del coeficiente como penalización en la función de pérdida. Fuerza de algún modo a que los coeficientes de los predictores tiendan a cero. La regresión LASSO se reconoce muy útil para la selección de variables.

$$||w||_1 = \sum_{j=1}^{p} |w_j|$$

El segundo tipo de regularización, denominada **L2** o **Ridge**, es muy similar a la LASSO, pero utiliza el cuadrado del coeficiente en vez del valor absoluto. Es decir, en la función de pérdida se introduce como penalización el hiperparámetro multiplicado por el cuadrado del coeficiente, de modo que pretende reducir los coeficientes de modo proporcional, sin que lleguen a cero.

$$||w||_2^2 = \sum_{j=1}^{p} w_j^2$$

La regularización **Elasticnet** es una tipo de penalización que combina L1 y L2, es decir, incluye parte de la regularización LASSO y de la Ridge, proponiendo una penalización que pondera las correspondientes a LASSO y Ridge. Será labor del Data Scientist "jugar" con esta ponderación dando más peso al parámetro alpha en la fórmula:

$$w_{elasticnet} = \alpha||w||_1 + \frac{1}{2}(1 - \alpha)||w||_2^2$$

En definitiva, al proceder con la regularización en el modelo, conseguimos reducir la complejidad del mismo, provocando la estimación de coeficientes cero, y por lo tanto la

exclusión de ciertas variables que comparten información con las que sí quedan dentro del modelo. El algoritmo funciona de modo iterativo: cada vez que el modelo plantee incluir una nueva variable, decide si “compensa”, esto es, si esa nueva variable va a mejorar el rendimiento de una manera suficiente como para compensar la pérdida por la penalización; en caso afirmativo, dicha variable se incluye. Si la mejora es insignificante y la penalización es mayor a esa mejora, el modelo no la incluirá. Con esto se consigue que no haya variables irrelevantes en el modelo y que la complejidad del mismo sea la idónea.

5.3.1.2 Ventajas e inconvenientes del modelo de regresión logística

La regresión logística es una técnica muy empleada por los científicos de datos debido a su eficacia y simplicidad. No es necesario disponer de grandes recursos computacionales, tanto en entrenamiento como en ejecución. Además, los resultados son altamente interpretables, siendo esta una de sus principales ventajas respecto a otras técnicas de clasificación. El peso de cada una de las características determina la importancia que tiene en la predicción final.

El funcionamiento de la regresión logística, al igual que la regresión lineal, permite identificar qué atributos están relacionados con la respuesta, eliminando aquellos que no lo están. También, siguiendo el principio de parsimonia, se excluyen aquellas características que presentan multicolinealidad con otras y solapan por lo tanto la información que aportan sobre la respuesta. Sin embargo, cuando tenemos variables predictoras que no están relacionadas linealmente con la función de la media a predecir, en nuestro caso el logit de la probabilidad de éxito (tumor maligno), o bien las transformamos para corregir linealidad, o bien se excluyen al no poderse integrar en un predictor lineal.

5.3.2 Clasificador de Naïve Bayes

La estructura que se plantea para aplicar los clasificadores Naive Bayes es muy similar a la de los problemas de regresión logística, pero en este caso no se propone un modelo analítico sino que se trata de predecir directamente la probabilidad de cada etiqueta de la respuesta para una nueva muestra, en función de las variables predictoras registradas.

Como se explica en (Borrás et al., 2023), los clasificadores Naïve Bayes son posiblemente los algoritmos de clasificación más básicos frecuentemente utilizados como modelo de base o partida en los problemas de clasificación. Son algoritmos extremadamente rápidos y sencillos, y suelen ser adecuados para conjuntos de datos de muy alta dimensión (con muchas variables).

Los clasificadores Bayes se basan en el teorema de Bayes, que describe la relación de las probabilidades condicionales entre dos conjuntos de sucesos. En nuestro caso estamos interesados en determinar la probabilidad del conjunto posible de respuestas, en función del conjunto de predictoras observadas:

$$P(y_l|x_1, \dots, x_p) = \frac{P(x_1, \dots, x_p|y_l)P(y_l)}{P(x_1, \dots, x_p)}$$

donde $P(x_1, \dots, x_p|y_l)$ es la verosimilitud para una etiqueta dada, $P(y_l)$ es la probabilidad previa de cada clase antes de la toma de datos, $P(x_1, \dots, x_p)$ es la información marginal aportada por los datos, y $P(y_l|x_1, \dots, x_p)$ representa la distribución posterior de la clase l dada la información recogida. La distribución posterior cuantifica la probabilidad de cada clase dado el conjunto de datos observado. Para evaluar dicho cociente los algoritmos Naïve Bayes asumen independencia entre las observaciones.

5.3.2.1 Tipos de clasificadores Naïve Bayes

Dentro de los clasificadores Naïve Bayes encontramos tres tipos principales, en función de las características de la variable respuesta y las predictoras. Dichos tipos son: **Naïve Bayes Bernouilli**, **Naïve Bayes Multinomial**, y **Naïve Bayes Gaussiano**.

El algoritmo *Naïve Bayes Bernouilli* se utiliza cuando tanto la respuesta como las predictoras tienen únicamente dos etiquetas o categorías.

El algoritmo *Naïve Bayes Multinomial* se utiliza cuando la variable respuesta tiene dos o más etiquetas posibles, y las variables predictoras son de tipo categórico multinomial.

Es por ello que nosotros utilizaremos el clasificador **gaussiano**, pues es el único algoritmo que utiliza variables predictoras de tipo numérico y continuo.

Este algoritmo hace uso de las medias y desviaciones estándar de las predictoras en cada una de los niveles de respuesta de la variable objetivo, para obtener la probabilidad de clasificación de cada etiqueta de la respuesta.

En primer lugar, segmentamos los datos por el nivel de respuesta, y a continuación, calculamos la media y la varianza de X_i en cada nivel. Donde μ_y es la media de X_i asociado a la clase y, σ_y^2 es la varianza de X_i asociado a la clase y. Entonces, la densidad de probabilidad de un cierto valor dada una clase, $P(X_i|y)$, se puede calcular agregando X_i en la ecuación de una distribución Normal con parámetros μ_y y σ_y^2. Es decir:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right), \qquad i = 1,\dots,30;\ y = \{M, B\}$$

donde los parámetros μ_y y σ_y^2 se estiman por máxima verosimilitud a partir de la información contenida en la muestra de entrenamiento, es decir, las medias y varianzas muestrales de cada predictora continua cuando estamos en la clase y.

5.3.2.2 Calibración de probabilidades

Un modelo está perfectamente calibrado cuando, para cualquier valor p, la clasificación predicha con una confianza (probabilidad) de p es correcta el 100∗p por ciento de las veces. Por ejemplo, si se seleccionan las observaciones cuya probabilidad predicha es $\widehat{p} = 0.8$, es de esperar que el porcentaje de esas observaciones bien clasificadas sea del 80%.

El proceso más empleado para saber si un modelo está bien calibrado es generar la curva de calibrado o *reliability plot.* La curva de calibrado queda por encima de la diagonal si el modelo tiende a infravalorar las probabilidades y por debajo si las sobrevalora.

Aunque las curvas de calibración aportan información detallada, es interesante disponer de una métrica que permita cuantificar con un único valor la calidad de la calibración del modelo. El coeficiente de Brier es la diferencia cuadrática media (mean squared difference) entre la probabilidad estimada por el modelo y la probabilidad real (1 para los tumores malignos y 0 para los benignos). Cuanto menor es su valor, mejor calibrado está el modelo. Esta métrica es adecuada solo para clasificaciones binarias.

Calibrar un clasificador consiste en ajustar un regresor que asigna la salida del clasificador a una probabilidad calibrada en [0,1]. Si denotamos la salida del clasificador para una muestra dada por f_i, el calibrador intenta predecir $p(y_i = 1|f_i)$.

Los calibradores habituales utilizados son:

- **Calibrador sigmoide:** basado en el modelo de regresión de Platt dado por la ecuación:

$$P(y_i = 1|f_i) = \frac{1}{1+exp(Af_i+B)},$$

donde y_i es el verdadero valor para la muestra i, f_i es el valor obtenido por el calibrador, y A y B son valores que se obtienen cuando ajustamos el regresor por máxima verosimilitud.

En general, este método es más eficaz cuando el modelo no calibrado está infravalorado y tiene errores de calibración similares para las salidas altas y bajas.

- **Calibrador isotónico:** ajusta un regresor no paramétrico que produce una función escalonada no decreciente, buscando el mínimo valor de:

$$\sum_{i=1}^{n} (y_i - \hat{f}_i)^2,$$

con la restricción de que $\hat{f}_i \geq \hat{f}_j$ siempre que $f_i \geq f_j$.

Este método es más general en comparación con el "sigmoide", ya que la única restricción que impone es que la función de calibración sea monótonamente creciente. Por tanto, es más potente, ya que puede corregir cualquier distorsión monótona del modelo no calibrado. Sin embargo, es más propenso a sobreajustarse, especialmente en conjuntos de datos pequeños. En general, esta calibración funcionará tan bien o mejor que la "sigmoide" cuando haya suficientes datos (más de ~ 1000 muestras) para evitar el sobreajuste.

5.3.2.3 Ventajas e incovenientes del clasificador Naïve Bayes

Los clasificadores Naïve Bayes tienden a funcionar especialmente bien en cualquiera de las siguientes situaciones:

- Cuando las clases de la respuesta están bien separadas, es decir, la distribución de probabilidad posterior de las clases en función de las predictoras son diferentes. Sobre una respuesta dicotómica, como es el caso

que nos ocupa, estamos hablando de que la probabilidad de tumor maligno está distanciada del valor 0.5.

- Cuando disponemos de una gran cantidad de predictoras y la complejidad del modelo predictivo no es relevante.

El clasificador Naïve Bayes tiene además las siguientes ventajas computacionales:

- Es extremadamente rápido tanto para el entrenamiento como para la predicción, y por tanto tiene un coste de cálculo muy bajo.
- Proporciona una predicción probabilística directa.
- Puede trabajar eficazmente en un gran conjunto de datos.

Entre las desventajas de este algoritmo podemos mencionar:

- La hipótesis de la independencia condicional no siempre se cumple. En la mayoría de las situaciones, las variables predictoras muestran alguna forma de asociación.
- El problema de la probabilidad cero hace referencia a las situaciones en las que en la muestra test tenemos valores de la respuesta que no estaban en la muestra de entrenamiento. Esto provoca automáticamente que la probabilidad de esa clase sea siempre cero. Por ese motivo hay seleccionar con cuidado la muestra de entrenamiento para asegurar que se dispone de valores en todas las clases de la respuesta.

Es un algoritmo habitual dentro del ámbito de la salud donde para cada sujeto hay mucha información disponible, ya que el clasificador Naïve Bayes tiene en cuenta la evidencia de todos los atributos considerados para determinar la probabilidad de que el sujeto padezca o no cierta enfermedad, proporcionando una herramienta muy sencilla para la toma de decisiones

5.4 Método de aprendizaje en clasificación

En esta sección resumimos la metodología general de los modelos de aprendizaje automático para clasificar a los sujetos en las distintas clases de la variable respuesta.

5.4.1 Pasos a seguir para ajustar el modelo y validar la clasificación

Para comenzar a analizar modelos de aprendizaje automático, la primera premisa es el requisito de que las variables numéricas deben estar estandarizadas, ya que si cada una de ellas permaneciera con su propia escala se generarían distorsiones y provocarían clasificaciones no óptimas. El siguiente requisito concierne a la separación de los datos en muestras aleatorias de entrenamiento, en nuestro caso del 75% de la base de datos original, y de test, del 25%.

Una vez asumido esto, los modelos propuestos comparten una metodología general, diferenciándose solo en la algoritmia para el ajuste del modelo o la selección de variables. Dicha metodología consta de los siguientes pasos:

1. Crear el modelo con todas las variables originales y calcular el error (o exactitud) del modelo con la muestra de entrenamiento:
    - Para el ajuste de la regresión logística utilizamos la función: `linear_model.LogisticRegression(solver='saga',max_iter=500)`
    - Para el ajuste del método de Naïve Bayes Gaussiano usamos el código:

      ```
      GaussianNB().fit(X1_train, y1_train)
      ```

2. Una vez ajustado el modelo se busca optimizar los hiperparámetros estimados.
    - Para regresión logística aplicamos la regularización, esto es, distintas penalizaciones, y decidimos cuál simplifica el modelo con la menor pérdida de exactitud desechando así las variables irrelevantes.
    - Para el método de Naïve Bayes recalibramos y comprobamos si se mejoran también los resultados.

3. Una vez decidido cuál es el mejor modelo, se ajusta con nuevas muestras de entrenamiento y test.
4. A continuación se calcula la matriz de confusión con la muestra test (consistente en 143 registros en nuestro caso), con la función `ConfusionMatrixDisplay`, y se evalúan las 4 métricas de clasificación con la muestra test, utilizando el comando `classification_report`:
    - ***Accuracy***: precisión general de la predicción.
    - ***Precision***: ratio de tumores malignos bien clasificados.
    - ***Recall***: recuerdo o ratio de tumores malignos bien clasificados sobre todos los tumores bien clasificados (benignos y malignos).
    - ***F1-score***: la puntuación F1 corresponde a 2x((recall * precision)/(recall + precision)).
5. Se diseña la curva ROC (o Característica Operativa del Receptor), consistente en una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. La interpretación de la curva ROC se realiza comparando la proporción de los verdaderos positivos (tumores malignos correctamente clasificados), o sensibilidad (en el eje Y) con respecto al ratio de falsos positivos (tumores benignos clasificados como malignos sobre el total de benignos), o especificidad (en el eje X). El indicador más utilizado en muchos contextos es el área bajo la curva ROC o AUC. Este índice se puede interpretar como la probabilidad de que un clasificador ordene o puntúe una instancia positiva elegida aleatoriamente más alta que una negativa. Valores cercanos a 1 indican gran precisión del modelo seleccionado.
6. Se analiza a continuación la estabilidad del modelo con el método de validación cruzada, consistente en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones, con el fin de asegurar que los resultados obtenidos no dependan de muestras de entrenamiento y test específicas. Se calcula observando las métricas de la función `cross_val_score`.

7. Estudiamos la curva de aprendizaje del modelo, que analiza la variación de la exactitud en la validación a medida que aumentamos el porcentaje de muestra entrenada.

5.5 Software y hardware

El lenguaje de programación utilizado en este proyecto ha sido Python, programado con los cuadernos Jupyter de Google Colab, que permiten el trabajo colaborativo en línea. Para almacenar datos, se ha utilizado el repositorio público de Github. Para desarrollar el informe se han utilizado Documentos de Google en línea.

Las librerías de Python utilizadas han sido:

- *Pandas*, utilizada para el trabajo con matrices.
- *Numpy,* que permite realizar operaciones numéricas con vectores y matrices.
- *Matplotlib,* para representar descriptivos y funciones gráficamente.
- *Seaborn,* que sirve para crear gráficos estadísticos.
- *Data Frame Image,* para guardar matrices creadas como un archivo de imagen (.png)
- *Scikit-Learn,* base para los modelos de Machine Learning trabajados.

Puesto que se ha trabajado con software en línea, los requisitos de hardware son mínimos. En el caso de nuestro estudio se ha utilizado un equipo Lenovo™ Ideapad™ 110 con sistema operativo Windows 10 y procesador Intel Core i7.

6. Resultados

6.1 Resultados del análisis descriptivo

A continuación presentamos los resultados obtenidos en el análisis descriptivo realizado sobre la base de datos original. La variable respuesta se describe de modo univariado, y las explicativas se describen diferenciando por los dos niveles de respuesta: tumores malignos y benignos.

6.1.1 Tipo de tumor

Al obtener la tabla de frecuencias (ver Tabla 2) y mostrar los conteos con diagramas de barras (ver Figura 1), destacamos la superioridad de tumores benignos sobre malignos en la base de datos, que representan un 62.7% del total de tumores observados.

Tabla 2: Tabla de frecuencias para la variable objetivo Tipo de tumor.

	BENIGNO	MALIGNO	TOTAL
Nº de registros	357	212	569
Porcentaje	62.74%	37.26%	100%

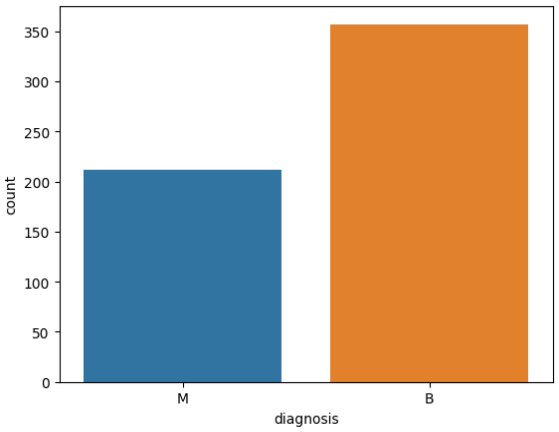


Fig 1. Gráfico de barras con los conteos de tumores benignos (B) y malignos (M).

6.1.2 Análisis descriptivo de las características observadas por tipo de tumor

Para mostrar los descriptivos numéricos de las características fisiológicas medidas, agrupamos la información disponible por medias (Tabla 3), desviaciones típicas (Tabla 4) y peor valor (Tabla 5), diferenciando según el tipo de tumor (M=maligno y B=benigno). Los gráficos correspondientes se muestran en la Figura 2.

Tabla 3: Descriptivos numéricos con los valores (por columnas, de izquierda a derecha) de mínimo, percentil 25, mediana, media, percentil 75 y máximo dadas las variables indicadoras de medias, según la variable respuesta B (benigno) o M (Maligno)

		min	p25	median	mean	p75	max
B	area_mean	143.5000	378.2000	458.4000	462.7902	551.1000	992.1000
B	compactness_mean	0.0194	0.0556	0.0753	0.0801	0.0976	0.2239
B	concave points_mean	0.0000	0.0150	0.0234	0.0257	0.0325	0.0853
B	concavity_mean	0.0000	0.0203	0.0371	0.0461	0.0600	0.4108
B	fractal_dimension_mean	0.0518	0.0585	0.0615	0.0629	0.0658	0.0958
B	perimeter_mean	43.7900	70.8700	78.1800	78.0754	86.1000	114.6000
B	radius_mean	6.9810	11.0800	12.2000	12.1465	13.3700	17.8500
B	smoothness_mean	0.0526	0.0831	0.0908	0.0925	0.1007	0.1634
B	symmetry_mean	0.1060	0.1580	0.1714	0.1742	0.1890	0.2743
B	texture_mean	9.7100	15.1500	17.3900	17.9148	19.7600	33.8100
M	area_mean	361.6000	705.3000	932.0000	978.3764	1203.7500	2501.0000
M	compactness_mean	0.0460	0.1096	0.1324	0.1452	0.1724	0.3454
M	concave points_mean	0.0203	0.0646	0.0863	0.0880	0.1032	0.2012
M	concavity_mean	0.0240	0.1095	0.1513	0.1608	0.2030	0.4268
M	fractal_dimension_mean	0.0500	0.0566	0.0616	0.0627	0.0671	0.0974
M	perimeter_mean	71.9000	98.7450	114.2000	115.3654	129.9250	188.5000
M	radius_mean	10.9500	15.0750	17.3250	17.4628	19.5900	28.1100
M	smoothness_mean	0.0737	0.0940	0.1022	0.1029	0.1109	0.1447
M	symmetry_mean	0.1308	0.1740	0.1899	0.1929	0.2098	0.3040
M	texture_mean	10.3800	19.3275	21.4600	21.6049	23.7650	39.2800

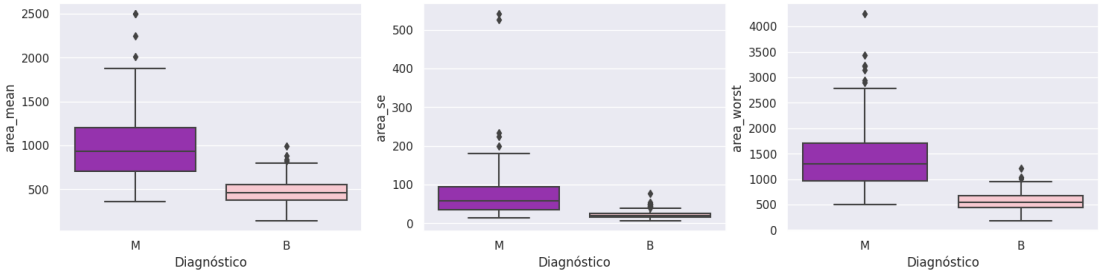
Tabla 4: Descriptivos numéricos con los valores (por columnas, de izquierda a derecha) de mínimo, percentil 25, mediana, media, percentil 75 y máximo dadas las variables indicadoras de errores estándar, según la variable respuesta B (benigno) o M (Maligno)

		min	p25	median	mean	p75	max
B	area_se	6.8020	15.2600	19.6300	21.1351	25.0300	77.1100
B	compactness_se	0.0023	0.0113	0.0163	0.0214	0.0259	0.1064
B	concave points_se	0.0000	0.0064	0.0091	0.0099	0.0119	0.0528
B	concavity_se	0.0000	0.0110	0.0184	0.0260	0.0306	0.3960
B	fractal_dimension_se	0.0009	0.0021	0.0028	0.0036	0.0042	0.0298
B	perimeter_se	0.7570	1.4450	1.8510	2.0003	2.3880	5.1180
B	radius_se	0.1115	0.2073	0.2575	0.2841	0.3416	0.8811
B	smoothness_se	0.0017	0.0052	0.0065	0.0072	0.0085	0.0218
B	symmetry_se	0.0095	0.0156	0.0191	0.0206	0.0241	0.0615
B	texture_se	0.3602	0.7959	1.1080	1.2204	1.4920	4.8850
M	area_se	13.9900	35.7625	58.4550	72.6724	94.0000	542.2000
M	compactness_se	0.0084	0.0197	0.0286	0.0323	0.0389	0.1354
M	concave points_se	0.0052	0.0114	0.0142	0.0151	0.0175	0.0409
M	concavity_se	0.0110	0.0270	0.0371	0.0418	0.0504	0.1438
M	fractal_dimension_se	0.0011	0.0027	0.0037	0.0041	0.0049	0.0128
M	perimeter_se	1.3340	2.7155	3.6795	4.3239	5.2063	21.9800
M	radius_se	0.1938	0.3904	0.5472	0.6091	0.7573	2.8730
M	smoothness_se	0.0027	0.0051	0.0062	0.0068	0.0080	0.0311
M	symmetry_se	0.0079	0.0146	0.0177	0.0205	0.0221	0.0790
M	texture_se	0.3621	0.8928	1.1025	1.2109	1.4292	3.5680

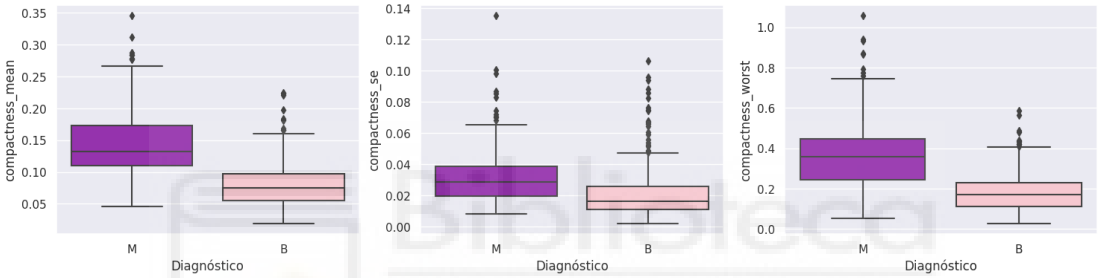
Tabla 5: Descriptivos numéricos con los valores (por columnas, de izquierda a derecha) de mínimo, percentil 25, mediana, media, percentil 75 y máximo dadas las variables indicadoras del peor valor, según la variable respuesta B (benigno) o M (Maligno)

		min	p25	median	mean	p75	max
B	**area_worst**	185.2000	447.1000	547.4000	558.8994	670.0000	1210.0000
B	**compactness_worst**	0.0273	0.1120	0.1698	0.1827	0.2302	0.5849
B	**concave points_worst**	0.0000	0.0510	0.0743	0.0744	0.0975	0.1750
B	**concavity_worst**	0.0000	0.0771	0.1412	0.1662	0.2216	1.2520
B	**fractal_dimension_worst**	0.0552	0.0701	0.0771	0.0794	0.0854	0.1486
B	**perimeter_worst**	50.4100	78.2700	86.9200	87.0059	96.5900	127.1000
B	**radius_worst**	7.9300	12.0800	13.3500	13.3798	14.8000	19.8200
B	**smoothness_worst**	0.0712	0.1104	0.1254	0.1250	0.1376	0.2006
B	**symmetry_worst**	0.1566	0.2406	0.2687	0.2702	0.2983	0.4228
B	**texture_worst**	12.0200	19.5800	22.8200	23.5151	26.5100	41.7800
M	**area_worst**	508.1000	970.3000	1303.0000	1422.2863	1712.7500	4254.0000
M	**compactness_worst**	0.0513	0.2445	0.3564	0.3748	0.4478	1.0580
M	**concave points_worst**	0.0290	0.1528	0.1820	0.1822	0.2107	0.2910
M	**concavity_worst**	0.0240	0.3264	0.4049	0.4506	0.5562	1.1700
M	**fractal_dimension_worst**	0.0550	0.0763	0.0876	0.0915	0.1026	0.2075
M	**perimeter_worst**	85.1000	119.3250	138.0000	141.3703	159.8000	251.2000
M	**radius_worst**	12.8400	17.7300	20.5900	21.1348	23.8075	36.0400
M	**smoothness_worst**	0.0882	0.1305	0.1434	0.1448	0.1560	0.2226
M	**symmetry_worst**	0.1565	0.2765	0.3103	0.3235	0.3592	0.6638
M	**texture_worst**	16.6700	25.7825	28.9450	29.3182	32.6900	49.5400

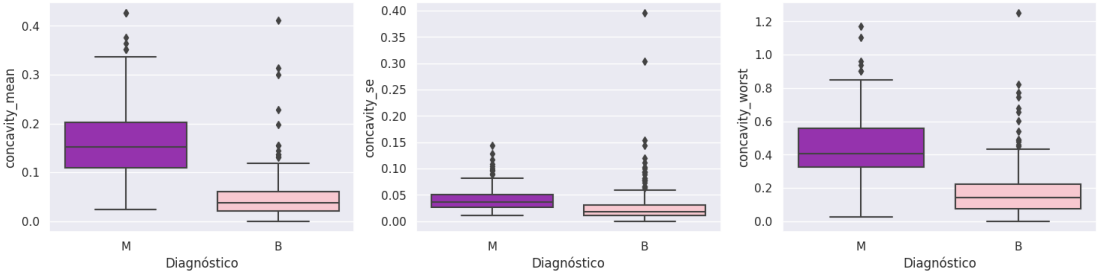
(a) Área del tumor



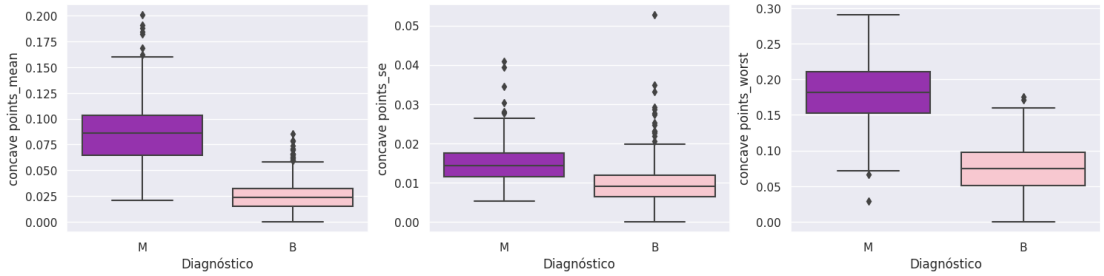
(b) Compacidad



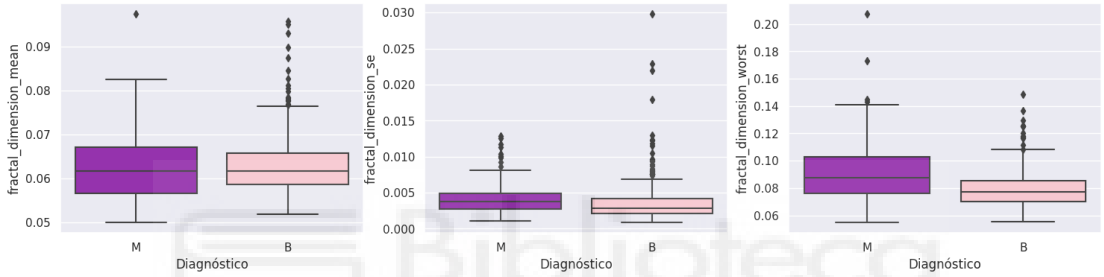
(c) Concavidad



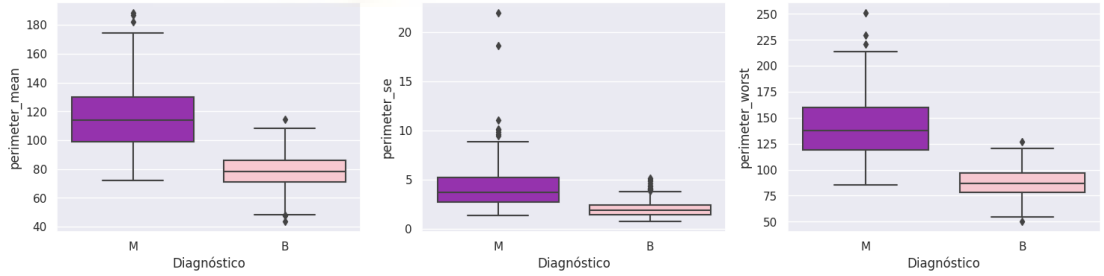
(d) Puntos cóncavos



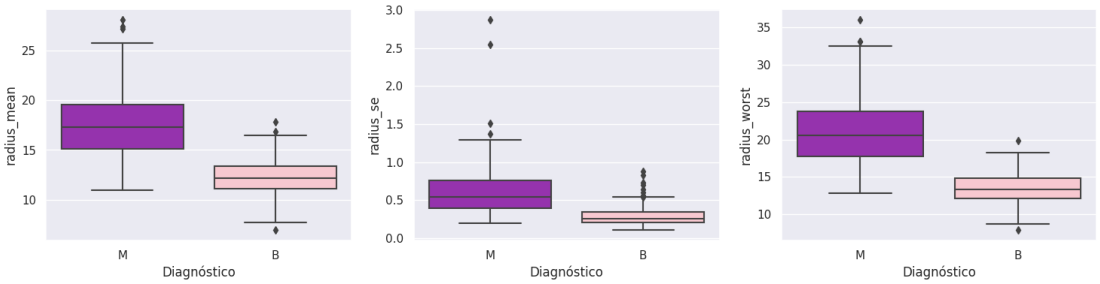
(e) Dimensión fractal



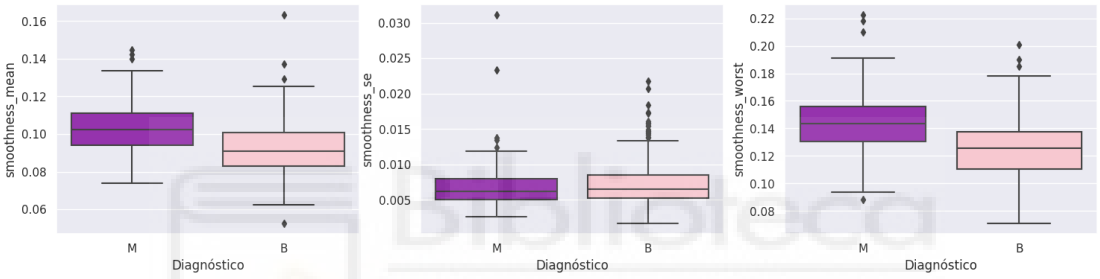
(f) Perímetro



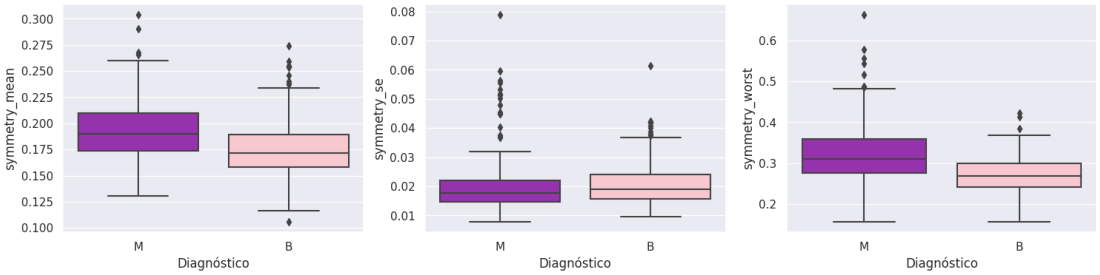
(g) Radio



(h) Suavidad



(i) Simetría



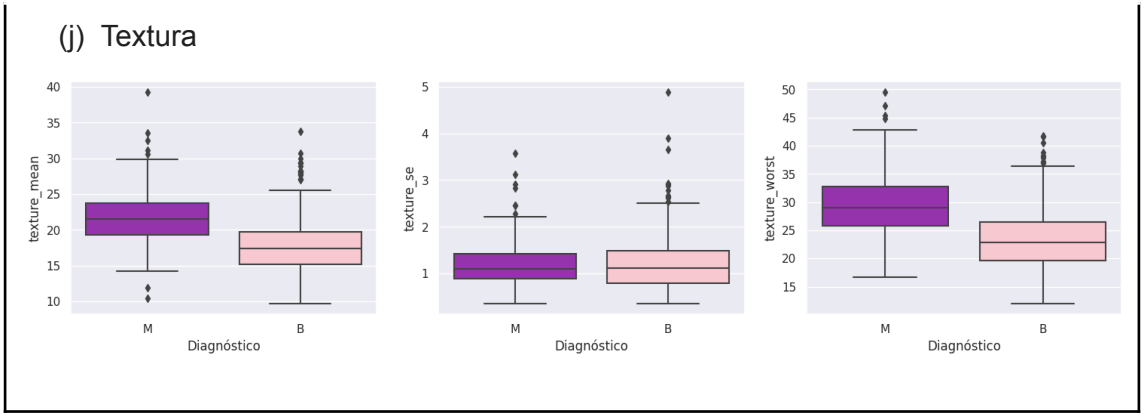


Figura 2: Gráficos de cajas para las características fisiológicas medidas, diferenciadas por tipo de tumor (B=benigno, M=maligno). De izquierda a derecha: medias, desviaciones típicas y peor valor.

A la vista de los descriptivos realizados, remarcamos en qué variables se observan las diferencias más acusadas entre tumores malignos y tumores benignos:

- Área

  Respecto de las mediciones del área de los tumores (medidas en mm^2) en distintas zonas de este, observamos una clara diferencia entre tumores malignos (M) y benignos (B), para las tres medidas sumario (*Media, SE y Worst)*. En todos ellos se aprecian diferencias claras en los rangos intercuartílicos (las cajas en los boxplot están desencajadas en las tres medidas, hecho que es patente obviamente al comparar las medianas (en el área media: 458.4 mm^2 para tumores benignos B, frente a 932 mm^2 para tumores malignos M) y cuartiles en la primera imagen de la Figura 2). La dispersión (relacionada con la longitud de los bigotes de la caja) es mayor en las tres medidas para los tumores malignos (M) que para los benignos (B).

- Compacidad

  En cuanto a compacidad volvemos a observar la diferencia entre células malignas y benignas. Las cajas, que representan la distribución de los datos para tumores malignos y benignos, se ubican en regiones no solapadas, lo que nos dice que hay diferencias claras entre tumores malignos y benignos para esta variable (especialmente para el valor medio, y el peor valor). Si bien en la variable que mide los errores estándar no se observa tanta diferencia entre los dos niveles, por lo que posiblemente sería una variable descartada como predictora en un modelo a ajustar; las que miden medias y peores mediciones sí parecen aportar diferencias entre los niveles de respuesta.

- Concavidad

  De la misma manera que en la variable *compacidad*, observamos una diferencia notoria en tamaño y dispersión, con valores mayores para las células malignas que para las benignas, tanto en las mediciones para la *concavidad media* como para las de *peor valor*. Ambas variables podrían ser relevantes en un modelo de clasificación, pero no tanto la variable *concavity_se*.

- Puntos cóncavos

  Si hablamos de puntos cóncavos en una zona del núcleo celular, estamos apreciando lo mismo que la concavidad, con la diferencia de que en este caso solo contamos el número de puntos y no la magnitud de las zonas cóncavas del contorno. Pero se observa la misma diferencia, bastante notoria, en el número de puntos cóncavos entre células malignas y benignas, siendo mayor en las dañinas, para los descriptivos *media* y *worst*.

- Dimensión fractal

  La variable dimensión fractal, al igual que con todas las características de forma, un valor más alto se corresponde con un contorno menos regular. Sin embargo en nuestro conjunto de datos apreciamos que no es excesiva la diferencia respecto de esta medición entre los tumores malignos y benignos (de hecho, son casi iguales las cajas y los bigotes de los gráficos).

- Perímetro

  Observamos de nuevo una clarísima diferencia entre las regiones intercuartílicas entre los tumores malignos y benignos. Las malignas a priori poseen un perímetro algo mayor que las benignas.

  También numéricamente son relevantes las diferencias: las medianas para las *medias* son de 78.18mm y 114.20mm para células benignas y malignas respectivamente; para desviaciones típicas las medianas son 1.85 y 3.68, y para el peor valor son de 86.92 y 138.00. Los rangos intercuartílicos también difieren, por lo que en principio podríamos anticipar que estas variables sí serán válidas en el modelo de clasificación.

- Radio

  Respecto de las mediciones del radio de los tumores (medidos en mm) en distintas zonas de la célula, observamos de nuevo, como en el caso del área, una diferencia entre tumores malignos (M) y benignos (B), para las tres medidas consideradas. Se aprecian diferencias claras en los rangos intercuartílicos; de hecho, las cajas en los boxplot están desencajadas en las tres medidas, hecho que se corrobora al comparar las medianas en medias

12.20 mm para benignos frente a 17.32 mm para malignos. La dispersión (relacionada con la longitud de los bigotes de la caja) es mayor en las tres medidas para los tumores malignos (M) que para los benignos (B).

- Suavidad

  La suavidad resulta ligeramente mayor en *media* y en *worst* en las células malignas que en las benignas, pero no de modo relevante; de hecho, las cajas se solapan para las tres medidas.

- Simetría

  No se aprecian diferencias en simetría entre los tumores malignos y benignos en ninguna de las tres mediciones, por lo que estas variables, en principio, se prevén poco relevantes para los modelos de clasificación.

- Textura

  Atendiendo a la literatura existente, se reconoce que los tumores malignos provocan piel escamosa alrededor del pezón y la areola, lo que repercute en valores superiores de textura para las células malignas. De hecho, todos los cuartiles son superiores en el grupo de tumores malignos que en el de benignos, tanto para medias como para peor valor. No es el caso para las mediciones de *errores estándar*, en las que no se aprecian diferencias entre ambos tipos de tumores.

6.1.2.1 Correlación entre variables predictoras

A continuación, en la Figura 3, se muestra el gráfico de correlaciones para todas las variables en la base de datos. La correlación existente entre las variables no es especialmente alta, oscilando en un rango de valores entre -0.3 y 0.3. Es destacable que en su mayoría las correlaciones son positivas.

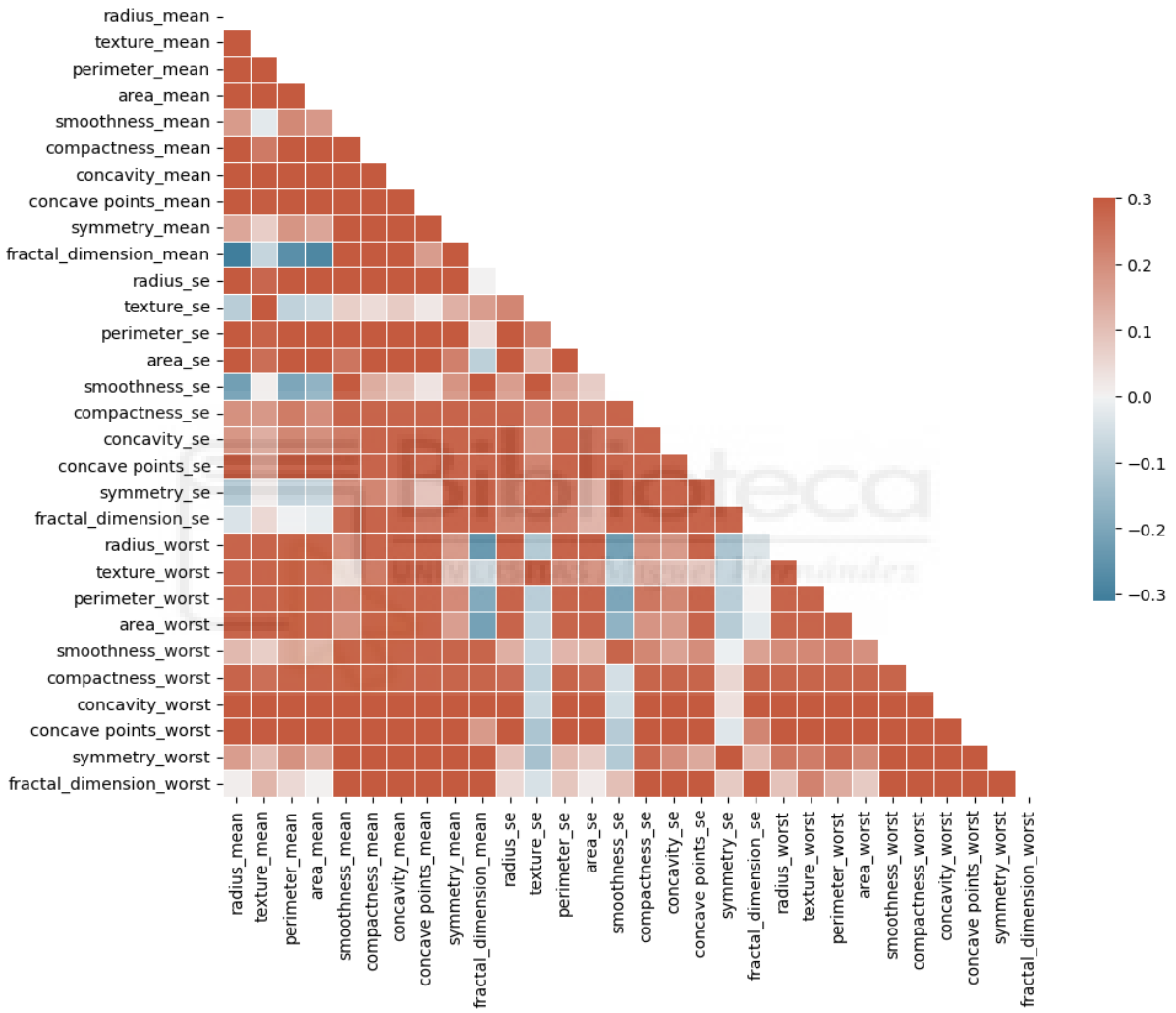


Figura 3: Correlación entre las variables del conjunto de datos

6.1.2.2 Conclusiones sobre las variables predictoras

Una vez realizado el análisis descriptivo diferenciando por tipo de tumor, hemos podido apreciar diferencias entre los tumores malignos y benignos para las variables:

radius_mean, radius_se, radius_worst, texture_mean, texture_worst, perimeter_mean, perimeter_se, perimeter_worst, area_mean, area_se, area_worst, compactness_mean, compactness_worst, concavity_mean, concavity_worst, concave points_mean y *concave points_worst*.

Es esperable que estas variables intervengan en los modelos predictivos de clasificación, hecho que corroboraremos más tarde después de ajustar el modelo logístico y proceder con la selección de variables vía regularización.

6.2 Resultados

En esta sección presentamos los resultados obtenidos al ajustar los modelos de aprendizaje automático estudiados en la sección 5.3 Modelos de clasificación. Compararemos los ajustes obtenidos en términos de métricas de evaluación.

6.2.1 Modelo logístico

Siguiendo con los pasos presentados en el apartado 5.4.1 Pasos a seguir para ajustar el modelo y validar la clasificación, los resultados que se presentan son los siguientes:

Partimos del modelo logístico ajustado con todas las variables, que proporciona una medida de exactitud sobre la muestra de entrenamiento de 0.99.

Tras la regularización, conseguimos un modelo con una precisión de 0.99, también sobre la muestra de entrenamiento. En dicha regularización hemos optado por la penalización L1 porque simplifica el modelo, eliminando hasta 16 variables, manteniendo una precisión de predicción sobre la muestra test bastante alta. Las variables independientes que resultan finalmente relevantes en este modelo son: *radius_se, radius_worst,*

texture_mean, texture_worst,

perimeter_worst,

area_worst,

compactness_se,

smoothness_worst,

concavity_worst,

concave points_mean, concave points_worst,

symmetry_mean, symmetry_worst,

fractal_dimension_se.

Estos resultados corroboran total/parcialmente las conclusiones que extrajimos del análisis descriptivo. En concreto, coinciden en todas las variables incluidas en el modelo logístico, salvo por las variables que incluyen **la simetría y la dimensión fractal**, donde en un principio no parecían ser muy influyentes según la Figura 2. Ocurre lo mismo con variables como ***compactness_se*** y ***smoothness_worst***.

Los coeficientes estimados en el modelo se muestran en la Tabla 6, de la que se concluye que las variables que más peso tienen sobre la predicción de la probabilidad de que un tumor sea maligno son *radius_se, concave points_worst* y *radius_worst* (con coeficientes positivos, luego influencia directa sobre el tumor maligno), y *compactness_se* y *fractal_dimension_se* (con coeficientes negativos indicando la influencia inversa, es decir, a mayor valor de las variables, menor es la probabilidad de malignidad en el tumor).

Tabla 6: Coeficientes del modelo logístico resultante de aplicar la penalización L1

Variable predictora	Coeficiente
radius_se	2.081
radius_worst	1.255
texture_mean	0.609
texture_worst	0.839
perimeter_worst	1.215
area_worst	1.211
compactness_se	-0.891
smoothness_worst	0.389
concavity_worst	1.041
concave points_mean	1.071
concave points_worst	1.187
symmetry_mean	0.464
symmetry_worst	0.437

fractal_dimension_se	-0.613

Las métricas de evaluación del modelo sobre la muestra test son (ver Tabla 7); al comparar con el modelo base, apreciamos las probabilidades de detección de un tumor maligno según cada métrica son:

La precisión (*accuracy*) general de predicción es de 0.97,

La precisión (*precision*) de predicción de tumores malignos es de 0.96,

El ratio de verdaderos positivos o recuerdo (*recall*) de predicción de tumores malignos es de 0.94,

La puntuación F1 (*f1-score*) de predicción de tumores malignos es de 0.95,

indicando los altísimos porcentajes de acierto del modelo para clasificar un tumor de manera efectiva.

Tabla 6: Métricas de clasificación para el modelo logístico

	Precision	Recall	F1-score	Accuracy
0 (Tumor benigno)	0.97	0.98	0.97	0.97
1 (Tumor maligno)	0.96	0.94	0.95	0.97

El área bajo la curva ROC es de 0.994, lo que nos indica que el modelo es muy preciso, casi exacto.

Los resultados de la validación cruzada se muestran en la Tabla 7. Podemos afirmar que la solución es muy estable ya que la desviación estándar es bastante pequeña (0.025).

Tabla 7: Validación cruzada sobre el modelo logístico

	Count	Media	Desviación estándar	Min	25%	50%	75%	Max
Score	10	0.986	0.025	0.9302	0.9826	1	1	1

Respecto a la curva de aprendizaje, observamos en la Figura 8 que con un 20% de muestra de entrenamiento, la precisión del modelo es de más de un 95.5%. Varía en menos de un 2.5% la exactitud de la validación al coger la muestra de entrenamiento completa.

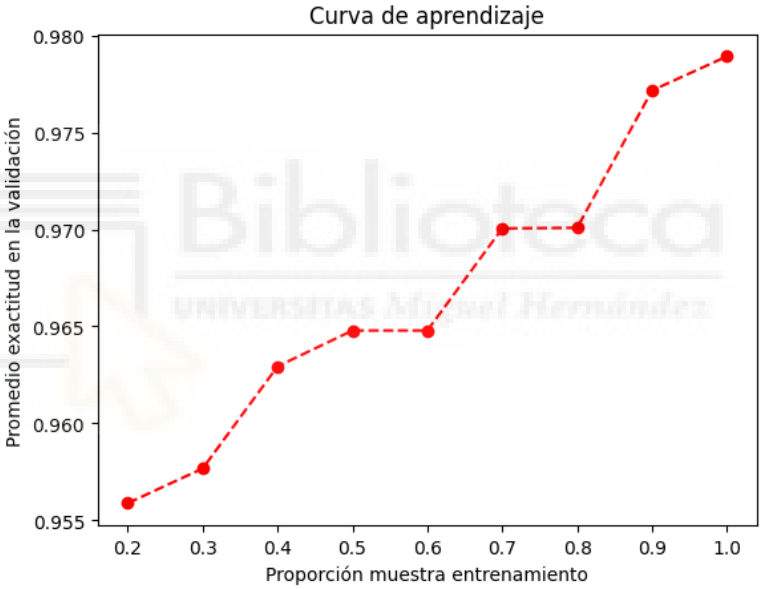


Figura 8: Gráfico con curva de aprendizaje según la proporción de muestra de entrenamiento

Se puede concluir, a raíz de todos estos resultados, que el modelo logit escogido es apropiado para predecir con precisión la probabilidad de tener un tumor maligno o benigno a través de las variables significativas, que corresponden a los niveles medios de textura, de puntos cóncavos y de simetría de la mama. También se tienen en

cuenta los errores típicos para valores de radio, compacidad y dimensión fractal; así como los peores valores de radio, textura, concavidad, puntos cóncavos y simetría de las células mamarias.

6.2.2 Clasificador de Naïve Bayes

De nuevo evaluamos los 5.4.1 Pasos a seguir para ajustar el modelo y validar la clasificación, en el que obtenemos los siguientes resultados para la clasificación por el método de Naïve Bayes:

Partiendo del clasificador de Naïve Bayes Gaussiano ajustado con todas las variables, obtenemos una medida de exactitud sobre la muestra de entrenamiento del 95%, así como una exactitud del 92% para el conjunto de ejecución.

Tras observar la curva de calibración del modelo seleccionado, se aprecia mucha infravaloración de las probabilidades. Optamos por realizar la calibración del modelo con los dos métodos disponibles (calibración isotónica y sigmoide) y obtenemos unos Brier Score de 0.052 y 0.068, respectivamente. Son valores semejantes, por eso optamos por evaluar las métricas de clasificación para cada uno de los tres modelos, donde los resultados se aprecian en las tablas 8, 9 y 10 . Se observa una mayor precisión general, así como mejores métricas para el modelo con calibración sigmoide que para el modelo con calibración isotónica. Por ello, nos quedamos con el modelo que corrige la calibración, además de replicar las exactitudes de las predicciones obtenidas en el modelo original.

Tabla 8: Métricas de clasificación para el modelo Naïve Bayes sin calibrar

	Precisión	Recuerdo	F1-score	Exactitud
0 (Tumor benigno)	0.93	0.93	0.93	0.92
1 (Tumor maligno)	0.89	0.89	0.89	0.92

Tabla 9: Métricas de clasificación para el modelo Naïve Bayes con calibración isotónica

	Precisión	Recuerdo	F1-score	Exactitud
0 (Tumor benigno)	0.95	0.89	0.92	0.90
1 (Tumor maligno)	0.83	0.92	0.88	0.90

Tabla 10: Métricas de clasificación para el modelo Naïve Bayes con calibración sigmoide

	Precisión	Recuerdo	F1-score	Exactitud
0 (Tumor benigno)	0.93	0.93	0.93	0.92
1 (Tumor maligno)	0.89	0.89	0.89	0.92

El área bajo la curva ROC es de 0.985, lo que nos indica que el modelo es muy preciso.

Los resultados de la validación cruzada se muestran en la Tabla 11. La exactitud media se sitúa en el 94.1% pero hay cierta dispersión ya que observamos un diferencia de casi un 10% entre el valor mínimo y el máximo (0.884 frente a 0.977). Aunque el clasificador proporciona buenos resultados en general vemos cierta inestabilidad debida a la muestra de entrenamiento utilizada.

Tabla 11: Validación cruzada sobre el modelo Naïve Bayes seleccionado

	Count	Media	Desviación estándar	Min	25%	50%	75%	Max
Score	10	0.941417	0.029428	0.883721	0.928987	0.952381	0.953488	0.976744

Respecto a la curva de aprendizaje, observamos en la Figura 9 que para tamaños pequeños de la muestra de entrenamiento, la exactitud de la validación es muy elevada, incluso aumenta esa exactitud entre el 20% y el 30% de la muestra de entrenamiento, pero a partir de ahí, la evolución del aprendizaje es decreciente, aunque mínimamente, a medida que aumentemos el tamaño de la muestra.

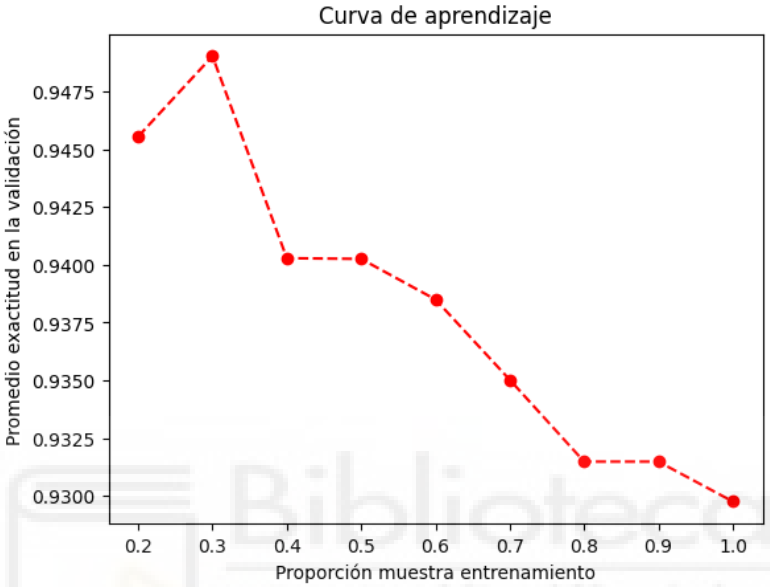


Figura 9: Gráfico con curva de aprendizaje según la proporción de muestra de entrenamiento

Por todo lo obtenido anteriormente podemos afirmar que el clasificador Naïve Bayes Gaussiano que hemos ajustado clasifica de forma precisa el conjunto de datos.

6.3 Conclusiones

Para finalizar con el informe, resaltamos las conclusiones básicas del estudio, relacionándolas con los objetivos planteados.

El objetivo principal del estudio es la construcción de modelos de aprendizaje automático sobre la base de datos para clasificar correctamente un tumor a la hora de dar un diagnóstico, a partir de una serie de variables asociadas a este. Además, la comparación entre los diversos modelos ajustados, para extraer conclusiones sobre el modelo más recomendable.

Más específicamente, los objetivos propuestos se han completado de manera satisfactoria:

El manejo con cuadernos de Google Colab ha sido una forma de aprendizaje muy atractiva para el alumno, por su sencillez y rapidez de ejecución, facilitando resultados en lenguaje Python visiblemente llamativos. En particular a mí me ha supuesto un gran aprendizaje y dominio de diferentes librerías de Python para aprender a analizar datos.

Para el problema de preprocesar los datos, utilizando los mecanismos adecuados, no ha supuesto una gran dificultad, pues la base de datos importada presenta registros y tipos de datos de fácil manejo, homogéneos y sin valores faltantes. Jugar con las variables para estandarizar los datos numéricos, además de codificar una variable categórica en una variable de tipo dummy, no han supuesto un aprieto para el alumno.

La descripción de la base de datos estudiada es, cuanto menos, interesante, ya que permite al lector conocer sobre las células mamarias y, en detalle, acerca de características medibles como la suavidad o la concavidad. En términos estadísticos, dichas variables de tipo numérico son simples de describir, así como todas las demás. Nuestra variable respuesta, de tipo categórica, con dos tipos de respuesta, tiene fácil síntesis, pues arroja un diagnóstico positivo o negativo (tumor maligno o benigno).

Seguidamente, y ya entrando en modelos de aprendizaje automático, aplicamos el modelo de regresión logística, obteniendo conclusiones sobre las variables que mejor explican el tipo de respuesta, o el tipo de tumor en este caso. Estos resultados corroboran parcialmente las conclusiones que se extraen del análisis descriptivo. En concreto, todas las variables que en un principio pensábamos que serían incluidas en el modelo por su relación con la respuesta, se encuentran en este, a excepción de variables como la media y el peor valor de la simetría, el error estándar de la dimensión fractal, el error estándar de la compacidad, o el peor valor de la suavidad (los descriptivos gráficos no nos aportaron tales conclusiones). Luego, variables como radius_mean, texture_mean, texture_worst, perimeter_mean, perimeter_se, area_mean, area_se, compactness_mean, compactness_worst o concavity_mean que presumiblemente tenían gran relación con el tipo de respuesta, no han sido incluidas en el modelo logístico final (debido a correlaciones con las demás variables). Por último, se comenta el acierto en la aplicación de dicho modelo, ya que la bondad de las clasificaciones obtenidas ha sido bastante precisa.

Acerca del modelo de clasificación por el método de Naïve Bayes, las conclusiones que podemos extraer son claras: un modelo sin calibración no nos aporta información auténtica, pues se infravalora mucho la probabilidad de acertar a la hora de clasificar el tumor. Es por ello que se aborda el estudio de los dos modelos de calibración sobre la base de datos, en el que uno de ellos tiene mejores métricas de predicción, luego clasificamos con él (calibración sigmoide). Resulta atractivo el estudio de este método, ya que no consiste en un modelo de regresión al que aplicar coeficientes a las variables predictoras; si no que consiste en calcular directamente probabilidades de clasificar un tumor como maligno o benigno dadas unas características. Es común encontrar este modelo en estudios de filtrado de spam o similares, para que cuando la probabilidad de que un mensaje entrante sea considerado como spam es muy alta, marcarlo automáticamente como mensaje sospechoso. En esto consisten las técnicas de aprendizaje automático, en aprender del trabajo precedente, para ser cada vez más efectivos a medida que avanza el tiempo. En mi opinión, me ha satisfecho aprender

este método de clasificación para variables de tipo categórico, ya que no había estudiado este tipo de modelos de aprendizaje automático.

Una vez analizados los modelos propuestos, el último de nuestros objetivos consiste en comparar las métricas de clasificación de tumores malignos de ambos modelos para, finalmente, estar en condiciones de recomendar uno u otro (véase Tabla 12).

Tabla 12: Comparativa de métricas de clasificación para los modelos estudiados

	Exactitud	Precisión	Recuerdo	F1-score
Regresión logística	**0.965**	**0.962**	**0.943**	**0.952**
Naïve Bayes	**0.916**	**0.887**	**0.887**	**0.887**

Como observamos, sobre nuestra base de datos es más recomendable ejecutar el modelo de **regresión logística** propuesto, pues todas sus métricas de clasificación son superiores a las del modelo de clasificación por el método Naïve Bayes (a pesar de las buenas métricas observadas para dicho modelo).

Con esta parte concluimos nuestro cuaderno, y por ende, el estudio que nos concierne acerca de la clasificación de tumores en las dos categorías de la variable 'diagnóstico' de nuestra base de datos. Comentar que todas estas prácticas ya se han puesto sobre la mesa en materia de investigación, y es por ello que con la ayuda de métricas estadísticas, se pueden lograr avances para encontrar la cura contra este cáncer tan perjudicial en la población mundial. En el año 2023, la tasa de supervivencia por este cáncer está por encima del 80% cuando se diagnostica a tiempo. Esto hace años no hubiera sido posible, pero con las técnicas de estudio estadístico y el avance tecnológico de los últimos años en materia de oncología ha sido posible lograrlo.

Finalmente, cabe agradecer la ayuda y confianza depositada por la doctora en Ciencias Matemáticas, Estadística e Investigación Operativa, Mª Asunción Martínez Mayoral; parte colaboradora del proyecto junto a un servidor, con el fin de estudiar nuevos modelos de aprendizaje automático sobre esta base de datos.

7. Bibliografía

Borrás, F., Botella, F., Hernández, I., Martínez, M. A., Moltó, J., & Morales, J. (2023, May 9). *Clasificador Naïve Bayes: Introducción y algoritmo para dos grupos*. Retrieved June 12, 2023, from https://colab.research.google.com/drive/1CPffo-FW9r6clHW09DxI5u2kYqrg10e0#scrollTo=-7whhErFzY9J

Borrás, F., Botella, F., Hernández, I., Martínez, M. A., Moltó, J., & Morales, J. (2023, May 9). *40-Primeros pasos con Scikit-Learn II*. Retrieved June 12, 2023, from https://colab.research.google.com/drive/14izQbel0eo9Ae04Fjb9lh5XYWS0LlVtf

Borrás, F., Botella, F., Hernández, I., Martínez, M. A., Moltó, J., & Morales, J. (2023, May 9). *Modelos de regresión para respuesta cualitativa dicotómica*. Retrieved June 12, 2023, from https://colab.research.google.com/drive/1WsJGCgvYDNVIFDZZjX0DBxXc1dBSS2pm#scrollTo=-7whhErFzY9J

Borrás, F., Botella, F., Hernández, I., Martínez, M. A., Moltó, J., & Morales, J. (2023, May 9). *90-El libro de recetas matplotlib: visualización de datos*. Retrieved June 12, 2023, from https://colab.research.google.com/drive/1jk31qhQ3yyrDqFZglJuhRgv0EXMlrpvI?usp=sharing

Borrás, F., Botella, F., Hernández, I., Martínez, M. A., Moltó, J., & Morales, J. (2023, May 9). *30-Primeros-pasos-con-Scikit-Learn I*. Retrieved June 12, 2023, from

https://colab.research.google.com/drive/1aiVZtm6vSPqxNCOmBzfQnfrHGdol-Ykj#scrollTo=YUiwA25wC4eV

Breast Cancer Dataset. (n.d.). Kaggle. Retrieved May 2, 2023, from https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset

Breast Cancer - Various Models. (n.d.). Kaggle. Retrieved May 9, 2023, from https://www.kaggle.com/code/fareselmenshawii/breast-cancer-various-models#Model-Evaluation

Guidance on Statistical Reporting | EFSA. (2014, December 2). EFSA. Retrieved May 2, 2023, from http://www.efsa.europa.eu/en/efsajournal/pub/3908

Haque, M. N., Tazin, T., Almalki, F. A., & Ashokkumar, N. (2022, July 7). *Tree-Based and Machine Learning Algorithm Analysis for Breast Cancer Classification*. Hindawi. Retrieved June 12, 2023, from https://www.hindawi.com/journals/cin/2022/6715406/

Martínez Mayoral, M. A. (2001). *Modelos lineales generalizados*. Universidad Miguel Hernández.

Rodríguez, D. (2018, July 23). *La regresión logística*. Analytics Lane. Retrieved June 12, 2023, from https://www.analyticslane.com/2018/07/23/la-regresion-logistica/

Scorsetti, M. (2020, January 7). *Ingeniería de Variables con Pivot Table (Pandas) | by Matias Scorsetti*. Medium. Retrieved May 9, 2023, from https://medium.com/@matiasscorsetti/ingenier%C3%ADa-de-variables-con-pivot-table-pandas-5f4a0e0a8454

Singh, S., Jangir, S. K., Kamal, M., & Ashokkumar, N. (2022, April 1). *Diagnosis of Breast Cancer Pathology on the Wisconsin Dataset with the Help of Data*

Mining Classification and Clustering Techniques. Hindawi. Retrieved June 12, 2023, from https://www.hindawi.com/journals/abb/2022/6187275/

sklearn.datasets.load_breast_cancer — scikit-learn 1.2.2 documentation. (n.d.). Scikit-learn. Retrieved May 2, 2023, from https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html

Tapia, C., Ulloa, C., & Cieza, J. P. (2021, June 26). *Cancer de Mama*. RPubs. Retrieved May 9, 2023, from https://rpubs.com/Cieza/785685

University of Wisconsin & Madison Clinical Sciences Center. (1996, Enero). *Index of /math-prog/cpo-dataset/machine-learn/cancer/WDBC*. cs.wisc.edu. Retrieved June 12, 2023, from https://ftp.cs.wisc.edu/math-prog/cpo-dataset/machine-learn/cancer/WDBC/

Visualize and export full Pandas DataFrames as images. (2021, December 23). Random Data Science. Retrieved May 9, 2023, from https://randomds.com/2021/12/23/visualize-and-save-full-pandas-dataframes-as-images/

Yazici, H., Aydin, M. A., & Ashokkumar, N. (2022, August 16). *Predicting Breast Cancer Leveraging Supervised Machine Learning Techniques*. Hindawi. Retrieved June 12, 2023, from https://www.hindawi.com/journals/cmmm/2022/5869529/

8. Anexos

El código utilizado para el desarrollo de este estudio está íntegramente contenido en el cuaderno Jupyter de Google Colab accesible en mi repositorio GitHub. También se encuentran disponibles los datos originales y los procesados por el alumno en este estudio.