TRABAJO FIN DE MASTER

COMPARATIVA DE ANÁLISIS DE IMPUTACIÓN DE DATOS FALTANTES CON ANÁLISIS DE CASOS COMPLETOS EN PRUEBAS DIAGNÓSTICAS

JULIO DE 2017

Álvaro Planchuelo Gómez

Tutora: Julia Amador Pacheco

Tutora: María Jesús López Herrero

Índice general

Índice de figuras

Índice de tablas

Resumen y objetivos

El objetivo principal de este trabajo es la comparación de diversos métodos de tratamiento de datos faltantes en bases de datos de pruebas diagnósticas generadas a partir de una base de datos, sin pérdida de datos, mediante los diferentes mecanismos de pérdida de datos (MCAR, MAR y MNAR). La base de datos original recoge información sobre tumores mamarios malignos y benignos en una muestra de mujeres. La comparación se basa en los efectos que tienen los métodos de tratamiento de datos faltantes, tanto en el modelo diagnóstico que se obtenga mediante regresión logística con sus errores estándar, como en las curvas ROC y los valores asociados a dichas curvas.

El objetivo secundario es obtener el mejor modelo de regresión logística para realizar el diagnóstico de los tumores. Se parte de la base de datos original sin datos faltantes, y el modelo final obtenido es el que se aplica para comparar los efectos de los métodos de tratamiento de datos faltantes.

Se explica la manera en la que se generan bases de datos con datos faltantes según cada mecanismo de pérdida de datos, tras detallar lo que representan las diferentes variables de la base de datos original. Se explican minuciosamente los métodos utilizados de tratamiento de datos faltantes con imputación múltiple, Missing Forest y MICE, y de manera más sencilla los demás métodos, el análisis de casos completos, el método de sustitución por la media y el uso de variables indicadoras de pérdida de datos.

Palabras clave

Regresión logística, curva ROC, datos faltantes, MICE, Missing Forest.

1. Introducción

En este capítulo, se va a proceder a describir el problema concerniente a este trabajo, a explicar brevemente la metodología que se emplea para el tratamiento de datos faltantes y a comentar la metodología matemática empleada en las pruebas diagnósticas.

1.1. Descripción del problema

Los datos faltantes o *missing data* son un problema importante a la hora de realizar estudios, ya que prácticamente todos los métodos estándar en Estadística asumen que la información está completa para todas las variables que están incluidas en el análisis [1]. Además, la falta de datos es un problema frecuente en todos los tipos de estudio, sin importar que el diseño sea muy estricto o que los investigadores traten de prevenirlo [2].

Según [3], los *missing data* se definen como los valores que no están disponibles pero que de otra manera serían significativos para el análisis si se pudieran observar. A pesar de la falta de datos, el objetivo es hacer inferencias sobre la población objetivo de la muestra completa, aunque no hay ningún método universal para tratar estos datos [4]. La causa de ello, es que aunque se suela conocer la selección de los sujetos de un estudio, se suele desconocer la causa de que sus observaciones se hayan perdido [4]. Esto obliga a que haya que hacer suposiciones adicionales sobre los datos para realizar el análisis, pero las mismas no se pueden validar únicamente a partir de los datos observados [4].

La importancia de estudiar y tratar los datos faltantes radica en que, en caso de no hacerlo, la ausencia de unas pocas observaciones en algunas variables puede reducir de manera drástica el tamaño de la muestra, afectando de manera muy significativa a los intervalos de confianza, a la potencia del estudio y provocando el posible sesgo de los parámetros estimados [1].

Hay que realizar la distinción entre mecanismos y patrones de pérdida de datos. Los patrones describen los valores que son observados y los que son perdidos, mientras que los mecanismos describen el proceso mediante el cual se produce la pérdida de los valores [4]. En prácticamente cualquier referencia sobre *missing data*,

se distinguen tres tipos de mecanismos de pérdida de datos:

- MCAR (*Missing completely at random*). Este mecanismo requiere que la probabilidad de pérdida de los datos de una variable Y no dependa de otras variables medidas, X, ni de la propia variable Y [5]. Entonces, los datos observados de la variable Y son una muestra aleatoria simple de los datos completos que se deberían haber observado en dicha variable [5]. Esto no implica que dicho mecanismo de pérdida no se pueda relacionar con el mecanismo de pérdida de los datos de otra variable [6]. Los datos de este mecanismo son los que se tratan con mayor facilidad [2]. Una situación que puede servir como ejemplo de este mecanismo es la anotación en papel de los datos de algunas variables de los pacientes al realizarles las pruebas con la posterior pérdida de los papeles en los que se hicieran las anotaciones.

- MAR (*Missing at random*). El mecanismo MAR tiene la misma base que el MCAR, pero es menos restrictivo. La diferencia entre ambos mecanismos consiste en que con MAR los datos que faltan no son una muestra aleatoria simple de todos los valores [4]. En este mecanismo la probabilidad de que se den datos faltantes en una variable Y depende de una o más de las demás variables medidas en el análisis, X, pero no depende de los valores de la propia variable Y [5]. Un ejemplo con esta situación podría darse cuando no hay posibilidad de hacer la medida de alguna variable a niños pequeños porque no quieren hacer una prueba y no se les pueda controlar; el mecanismo depende de datos observados (la edad) pero no de la propia variable en sí.

- MNAR (*Missing not at random*). En este mecanismo la probabilidad de pérdida de datos en una variable Y depende de los valores de dicha variable Y [5] y también puede depender de los valores observados de las demás variables [7]. Por ejemplo, en un ensayo clínico en el que se pruebe la eficacia de un medicamento contra la hipertensión y se realicen medidas a lo largo del tiempo, si al cabo de cierto tiempo el paciente se encuentra bien, puede decidir no acudir al lugar en el que se le toman las medidas de presión sanguínea, perdiéndose los datos de la parte final del estudio.

En la práctica, realizar estas distinciones entre los mecanismos de pérdida de datos no suele ser posible, sobre todo con MAR y MNAR, y salvo que haya información adicional, el tipo de mecanismo solamente se puede razonar o especular [2, 7].

Para comprender mejor lo que ocurre con la pérdida de datos en función del mecanismo, se va a poner un ejemplo extraído de [4]. En dicho ejemplo se considera un modelo de regresión lineal simple para modelar la relación entre las variables x e y, siguiendo la ecuación $y = x + \epsilon$ generando una muestra independiente de tamaño

300 y conociendo todos los valores de x e y. Los datos de x van a estar completos para cualquiera de las situaciones que se plantee. Se van a generar otras tres muestras para modelar los diferentes mecanismos de pérdida de datos en y. En la primera de ellas (MCAR), se pierden algunos datos de manera aleatoria. En la segunda (MAR), se pierden los datos en los que $x < 0$. En la tercera (MNAR), se pierden los datos en los que $y < 0$ y $x < 0$. Se muestran los resultados obtenidos de [4] en la Figura 1.

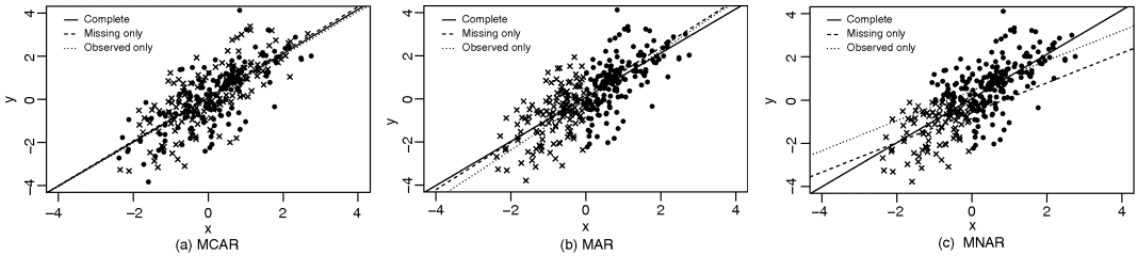


Figura 1: Gráficas de regresión lineal en mecanismos de pérdida de datos [4]

En la Figura 1, los puntos simbolizan la muestra observada y las cruces en forma de aspa la muestra con los datos de y que faltan. La línea continua representa la regresión lineal con todos los datos observados (muestra completa), la línea de puntos los resultados con solamente los datos observados y la línea discontinua los resultados con solamente los datos perdidos.

En la Tabla 1, obtenida de [4], se muestran los resultados de las regresiones lineales planteadas.

Valores	Datos completos	MCAR	MAR	MNAR
(Intercept)	0.08(0.06)	0.01(0.08)	−0.07(0.13)	0.44(0.09)
x	1.02(0.06)	1.01(0.08)	1.15(0.12)	0.70(0.09)
R^2	0.53	0.54	0.36	0.23
R^2 ajustado	0.53	0.53	0.35	0.23
Observaciones	300	148	163	201

Tabla 1: Resultados de los modelos de regresión lineal y pérdida de datos

En la Tabla 1, el paréntesis indica desviación estándar. Se puede comprobar que los resultados de MCAR respecto a los datos completos son prácticamente iguales. En cambio, en MAR y MNAR, los resultados difieren mucho respecto a los resultados con datos completos y hay sesgo en los mismos, especialmente con MNAR.

Respecto a los patrones de pérdida de datos, no existe el consenso que hay con los mecanismos de pérdida de datos, así que se van a exponer los patrones indicados y explicados en [4].

El primer patrón es el de una variable. Solamente faltan datos de una única variable mientras que el resto está completo.

Otro patrón es el monótono, que es muy común en estudios longitudinales. En este patrón, faltan los datos de varias variables de una parte muy localizada de la estructura de datos; en estudios longitudinales, por ejemplo, faltarían los datos de la parte final del estudio.

El patrón *file-matching* se da cuando dos o más variables no son medidas conjuntamente. Un ejemplo podría ser que no se pueden realizar dos pruebas molestas para los pacientes el mismo día y que de ambas pruebas se extraigan datos que se incluyen en un estudio. Es común en ensayos clínicos aleatorios cruzados en los que los resultados potenciales de un tratamiento no se pueden medir a la vez que los del placebo, siguiendo este caso una estructura del modelo causal de Rubin [8].

Y por último, en el patrón general no hay ninguna estructura de falta de datos que se pueda determinar con claridad. Es común en cuestionarios que no son completados por pacientes.

1.2. Metodología para el tratamiento de datos faltantes (*missing*)

A continuación, se van a explicar los diferentes métodos de tratamiento de datos faltantes a partir de las divisiones realizadas en [4].

1.2.1. Métodos que eliminan observaciones

Estos métodos consisten en eliminar la información de cualquier individuo en el que haya pérdida de datos, por lo menos, en una de las variables de estudio. Estos métodos son los más sencillos de implementar y los que menos recursos computacionales necesitan. Además, estos métodos reducen el tamaño de la muestra, lo que disminuye la eficiencia y conlleva que los errores estándar sean mayores en los parámetros de interés [4].

Los métodos más importantes son los de casos completos (*Complete-Case*), que pueden ser ponderados o no ponderados.

En el caso de no hacer ponderación, simplemente consiste en eliminar las observaciones en las que falte algún dato de alguna de las variables. La desventaja de este método es el sesgo que provoca en datos que no son MCAR, como ya se comentó a raíz de los resultados que se vieron en la Tabla 1, y la pérdida de precisión al tener un tamaño muestral menor. Se debería optar por este método en caso de que la pérdida de precisión y el sesgo sean despreciables o muy pequeños, pero son difíciles de cuantificar debido a que dependen de las diferencias en los parámetros de interés en los casos completos e incompletos [4]. Aunque no haya una regla universal establecida respecto a un porcentaje de fallo para aplicar este método de manera apropiada

[4], en [9] llegaron a la conclusión de que este método era más recomendable que la imputación múltiple (posteriormente se desarrollará) cuando el porcentaje de datos faltantes era aproximadamente del 10 % o menor.

El método ponderado se emplea especialmente con encuestas [4]. En caso de que falten datos solamente en una variable, se crea un modelo para predecir la ausencia de respuesta de la variable como una función del resto de variables disponibles y el inverso de las probabilidades predichas se puede utilizar como peso para hacer los casos completos más representativos [4]. En caso de que falten datos en dos o más variables, el método es más complicado y da problemas con los errores estándar cuando las probabilidades predichas son próximas a 0.

El último método consiste en eliminar las variables con una gran cantidad de datos faltantes, pero este método elimina información (variables) que puede ser muy importante para realizar modelos provocando sesgo y pérdida de precisión. Es un método que normalmente no se plantea aplicar.

1.2.2. Métodos que utilizan todos los datos disponibles

Estos métodos utilizan la información tanto de los casos completos como de los incompletos. En regresión multivariante, los casos con datos incompletos en algunas variables pueden dar información sobre la relación entre las variables respuesta y otras variables observadas [4]. Los modelos de efectos mixtos son una elección común a la hora de trabajar con datos longitudinales en los que faltan resultados [4]. Estos métodos dan mejores resultados en lo referente a la corrección del sesgo con datos faltantes MAR respecto a los métodos de casos completos [4] ya explicados. La principal desventaja de estos métodos es que si los datos faltantes difieren respecto a los casos observados, los resultados finales estarán sesgados [4].

Dentro de estos métodos caben destacar los que realizan inferencias basadas en métodos de máxima verosimilitud o *Maximum Likelihood* (ML), cuya idea principal es encontrar los valores más verosímiles a partir de los datos observados. Una de las técnicas más conocidas para maximizar la función de verosimilitud es el algoritmo EM (*Expectation-Maximization*), que se desarrolló en primer lugar en [10]. El primer paso es el E (esperanza), que en una iteración determinada calcula el valor esperado de log-verosimilitud de los datos completos, dados los valores de los datos observados y de las estimaciones de los parámetros en la iteración previa. El paso M (maximizar) encuentra la nueva estimación de los parámetros maximizando la esperanza obtenida del paso E.

1.2.3. Métodos que imputan los datos faltantes

La idea de estos métodos es que la información de los valores que faltan se puede extraer de las variables observadas [4]. Se van a distinguir dos clases de métodos en esta subsección: imputación simple e imputación múltiple.

1.2.3.1. Imputación simple

Los métodos de imputación simple reemplazan los datos faltantes por un único valor. En general, para cualquier enfoque de este tipo de métodos, se subestiman los errores estándar de las variables en las que faltan datos originalmente al tratar a los datos obtenidos como una muestra completa y no tener en cuenta las consecuencias del método [4].

El método más sencillo consiste simplemente en sustituir los datos faltantes de una variable por la media de los valores observados en dicha variable. En caso de que reemplazar por la media no tuviera sentido (por ejemplo, con el sexo), se podría reemplazar por la moda. Este método conlleva varios problemas aparte de la subestimación de varianzas y covarianzas. Uno de ellos es que se modifican las relaciones entre las variables debido a que las correlaciones tienden a ser nulas [11]. El sesgo no se elimina si los datos son MAR o MNAR [12]. Se incrementa la potencia estadística de manera espuria, así como la tasa del error de tipo I [12].

Otra aproximación similar pero más refinada es la imputación de la media condicionada, o también llamada imputación basada en regresión. En lugar de reemplazar cada valor faltante por un valor de media de la variable con observaciones perdidas, se sustituyen los diferentes valores por la media de la variable condicionada a las demás variables que se han observado completamente (sin datos perdidos).

Otros métodos de imputación simple son: *Last Observation Carried Forward* (LOCF), Sustitución por observaciones relacionadas, *Hot Deck* y Método de variables indicadoras, que pueden verse con detalle en [4, 11, 12].

Viendo los problemas de subestimación del error de los métodos anteriores, una estrategia para compensar dicho efecto puede ser añadir error para una mejor estimación de las varianzas y covarianzas y además poder obtener estimaciones precisas de los datos faltantes. El error añadido serviría para captar las diferentes fuentes de incertidumbre que tienen influencia en los datos reales [12]. Los problemas de este método son que las estimaciones son más imprecisas que en otros métodos ya comentados y que sigue habiendo subestimación de la variabilidad de los datos debido a que la varianza y las covarianzas de una muestra son parámetros que también deben ser estimados respecto a la población real [12].

1.2.3.2. Imputación múltiple

Los métodos de imputación múltiple consisten en reemplazar cada valor perdido por un conjunto de m valores obteniéndose así m conjuntos completos de datos, lo que da lugar a m estimaciones con sus respectivas varianzas o errores estándar [4]. Entonces, se combinan las estimaciones dando lugar a estimaciones e intervalos de confianza que incorporan la incertidumbre causada por la pérdida de datos [4]. Estos métodos permiten minimizar el sesgo o la pérdida de potencia estadística causada por la pérdida de datos con datos MCAR o MAR [12].

A priori, una desventaja que podría ser importante de estos métodos respecto a los de imputación simple es que son más difíciles de implementar, lo que además podría conllevar mayor gasto computacional. Para disminuir la dificultad que puede suponer la implementación de métodos complicados, ya existen paquetes en softwares como R para facilitar dicho trabajo, destacando, por ejemplo, el paquete mice [13].

En caso de que solamente haya una variable en la que haya pérdida de datos, el modelo de imputación puede ser una regresión logística para variables binarias o un modelo de regresión lineal para variables continuas [4].

Entonces, la estrategia más popular que se sigue es MICE (*Multiple Imputation with Chained Equations*) [13], que básicamente actualiza una a una las variables con datos faltantes según series completas de distribuciones condicionadas.

La Figura 2 muestra un esquema que sirve de resumen de los tres pasos principales de imputación múltiple: imputación, análisis y agrupación.

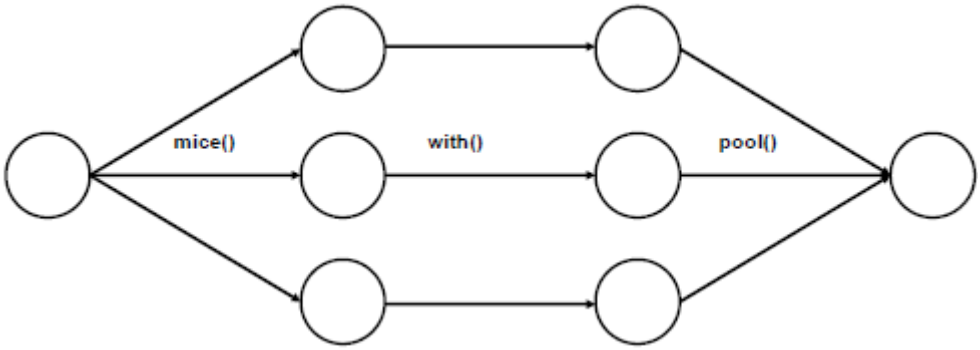


Figura 2: Pasos principales de la imputación múltiple [13]

En la Figura 2, la primera columna se corresponde con el conjunto de datos original, la segunda con los conjuntos de datos que se han imputado, la tercera con las estimaciones de las incógnitas de interés y la última con la agrupación final de los resultados de las estimaciones obtenidas.

Otro método de imputación múltiple, no paramétrico, es *Missing Forest*, basado en el método de clasificación o predicción *Random Forest*.

1.3. Pruebas diagnósticas

Las pruebas diagnósticas son exámenes que se utilizan para confirmar si los pacientes que se someten a dichos exámenes padecen ciertas enfermedades. En el caso de este trabajo, la evaluación de las células mamarias de las pacientes en un laboratorio es la prueba diagnóstica. Desde el punto de vista matemático, para la prueba diagnóstica de este trabajo, se van a emplear la regresión logística, para modelar el diagnóstico de tumores mamarios (determinar si son benignos o malignos) a partir de las características citológicas, y las curvas ROC (*Receiver Operating Characteristic*), que sirven para evaluar la calidad de los modelos de regresión logística.

1.3.1. Regresión logística

Los resultados binarios, como es el caso de las pruebas diagnósticas, por ejemplo, para confirmar la presencia de una determinada enfermedad en un individuo, se modelan típicamente con la regresión logística.

Partiendo de una respuesta y binaria que tiene valor 1 si, por ejemplo, un individuo presenta una enfermedad y 0 si no la presenta; y es una realización de una variable aleatoria Y que toma los valores 1 y 0 con probabilidades p y $1-p$ respectivamente, siguiendo una distribución de Bernoulli con parámetro p [14]. Si se denota Y, siguiendo con el mismo ejemplo, como el número de pacientes que padece una enfermedad concreta, toma valores desde 0 hasta n (número de observaciones) y, en caso de que las n observaciones sean independientes y tengan la misma probabilidad p de padecer la enfermedad, la distribución de Y es binomial con parámetros n y p [14].

Si p depende de un conjunto de covariables, lo más sencillo es tener un modelo de probabilidad lineal en el que p sea igual al producto de un vector de covariables por un vector de coeficientes de regresión, pero al ser p un valor que tiene que estar entre 0 y 1, esta ecuación no aseguraría un resultado en ese rango [14]. Para solucionar este problema, por una parte se usa el *odds* en lugar de p, que es igual al cociente de p entre $1-p$ (probabilidad de "éxito" entre la de "fracaso") y por otro, para arreglar que el límite inferior del *odds* es 0, se utiliza el logaritmo del *odds* [14], quedando así una ecuación de este tipo:

$$log\left(\frac{p}{1-p}\right) = logit(p) = X\beta, \tag{1.1}$$

siendo X el vector de covariables y β el vector de coeficientes de regresión. Aplicancdo la función exponencial a ambos lados de la ecuación, se puede despejar el

valor de p, siendo el resultado el siguiente:

$$p = \frac{exp(X\beta)}{1 + exp(X\beta)}, \tag{1.2}$$

obteniendo así diferentes valores de p, según el valor de las covariables en el individuo, comprendidos entre 0 y 1.

1.3.2. Curvas ROC (*Receiver Operating Characteristic*)

Una vez que se han obtenido los resultados de los individuos con el modelo de regresión logística, para ver la calidad del modelo se puede recurrir a las curvas ROC. Las curvas ROC permiten comparar los resultados del modelo respecto a los reales, que se obtienen mediante un *gold standard* o mediante una técnica cuya calidad muy alta haya sido demostrada.

Estas curvas permiten comparar resultados continuos u ordinales de pruebas diagnósticas, como es el caso que se obtiene con el modelo de regresión logística (valores de p entre 0 y 1). Para discernir los individuos que padecen una enfermedad de los que no (siguiendo con el ejemplo del anterior apartado), se toma un valor como límite (punto de corte de la curva ROC) y cuando la probabilidad (valor de p) esté por encima de ese valor, se diagnostica al individuo como enfermo y en caso contrario como sano.

La curva ROC tiene en el eje horizontal el valor de 1-especificidad y en el eje vertical el valor de la sensibilidad, estando ambos ejes comprendidos entre 0 y 1. La calidad del modelo se puede determinar mediante la medida resumen del área bajo la curva (AUC). El valor máximo de AUC es 1, que indicaría que el modelo o el test diagnóstico es perfecto. Si el valor del AUC es 0.5 (la curva sería una diagonal), el resultado del test sería equivalente al azar, por lo que se tendría un test o un modelo inútil.

Algunas propiedades interesantes de la curva ROC son que es creciente, que es invariante a transformaciones estrictamente crecientes de los valores de la prueba diagnóstica (esto puede ser útil, por ejemplo, para transformar distribuciones que no son normales a normales sin que cambie la curva ROC) y que es cóncava.

2. Metodología

En este capítulo se van a describir las distintas bases de datos con las que se va a trabajar, así como el estudio logístico que se va a llevar a cabo. En primer lugar, se describe la base de datos original en la que no hay datos faltantes. El resto de bases de datos, obtenidas a partir de la original, se han generado mediante los distintos mecanismos de datos faltantes.

2.1. Descripción de la base de datos original: B0

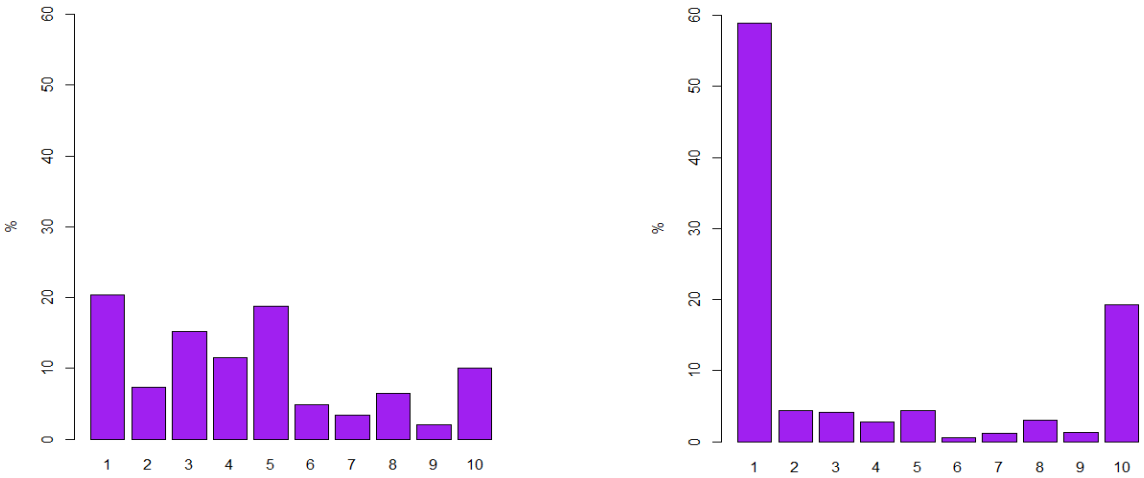
La base de datos que se ha escogido se puede encontrar en Orange [15] con el nombre breast-cancer-wisconsin, una de las bases de datos que incluye el software. Toda la información sobre los datos se puede encontrar en [16]. En este conjunto de datos hay 683 pacientes (444 pacientes con tumor benigno y 239 con tumor maligno) que tienen un tumor en el tejido mamario y se diagnostica si el tumor es benigno o maligno mediante punción-aspiración con aguja fina. Todas las variables del modelo son numéricas y tienen valores enteros del 1 al 10 salvo la variable Y, que es la variable respuesta (el diagnóstico) y tiene valor 2 si el tumor es benigno y 4 si es maligno, aunque para hacer el análisis se cambia 2 por 0 y 4 por 1. A continuación, se describen las 9 variables numéricas correspondientes a características citológicas y que han sido evaluadas en una escala de 1 a 10, donde 1 indica mayor proximidad a células benignas y 10 a células malignas.

- Clump thickness (C). La traducción sería similar a espesor de la masa. Hace referencia a si los agregados de células epiteliales eran monocapa o multicapa. En teoría, las células benignas tienden a agruparse en monocapa y las malignas en múltiples capas.
- Unif_Cell_Size (UCSi). Hace referencia a la uniformidad del tamaño de las células. Las células benignas tienden a ser más uniformes en tamaño.
- Unif_Cell_Shape (UCS). Hace referencia a la uniformidad de la forma de las células. Las células benignas tienden a ser más uniformes en forma.

- Marginal_Adhesion (MA). Hace referencia a la cohesión de las células periféricas de los agregados de células epiteliales. La pérdida de adhesión suele ser un rasgo distintivo de las células malignas.
- Single_Cell_Size (SCS). Hace referencia al tamaño de las células epiteliales. Las células malignas tienen en teoría un tamaño mayor que las benignas.
- Bare_Nuclei (BN). Hace referencia a la proporción de núcleos de células aisladas epiteliales desprovistos del citoplasma que los rodea. Es una característica propia de las células benignas que ocurra este suceso.
- Bland_Chromatin (BC). Hace referencia al grado de "suavidad" del núcleo. Las células malignas tiendes a ser más ásperas.
- Normal_Nucleoli (NN). Hace referencia al tamaño de los nucléolos. Suelen ser más grandes en células malignas.
- Mitoses (M). Hace referencia a la frecuencia de la mitosis. La tasa de replicación es mayor en las células malignas.

A modo de ejemplo, se van a mostrar los histogramas para dos de estas variables, Clump y Bare_Nuclei, que serán las variables en las que posteriomente se tenga pérdida de datos.

En primer lugar, se muestran los histogramas para Clump y Bare_Nuclei con todos los datos en la Figura 3 y, posteriormente, los histogramas para dichas variables pero distinguiendo según si las pacientes tenían patologías benignas o malignas.



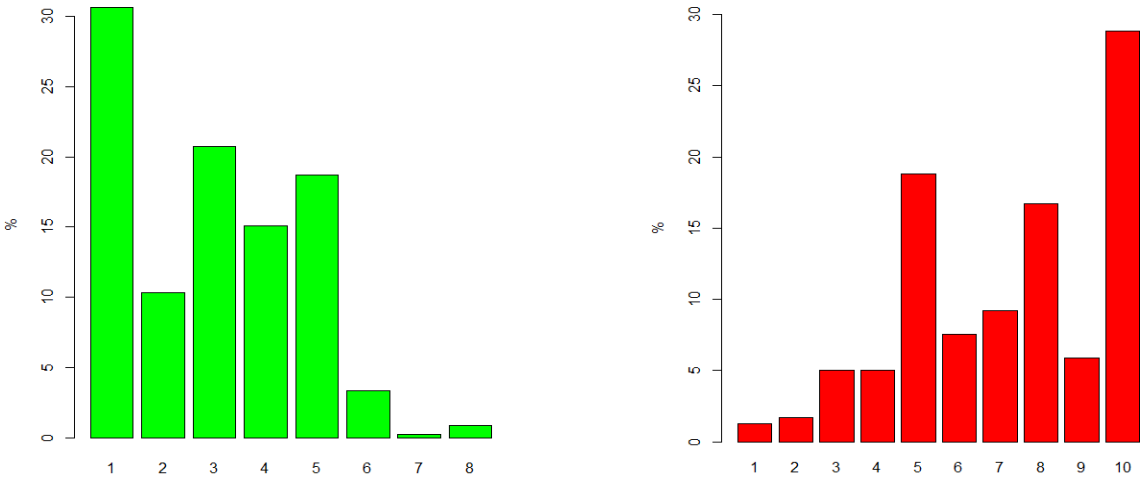
(a) Histograma de Clump

(b) Histograma de Bare_Nuclei

Figura 3: Histogramas con los valores de las variables Clump y Bare_Nuclei

En la Figura 3 se puede apreciar que en Clump los valores están bastante repartidos entre 1 y 10, habiendo más concentración de casos entre 1 y 5 y, en los

valores mayores, el mayor porcentaje se da con el valor 10 (10.1%). En cambio, con Bare_Nuclei, la gran parte de los casos se sitúa en 1 y, en segundo lugar, en 10 (porcentajes del 58.9% y del 19.3% respectivamente), habiendo un bajo porcentaje en el resto de valores (cerca del 5% o menor). Se muestran los histogramas distinguiendo según patologías benignas y malignas en Clump en la Figura 4.



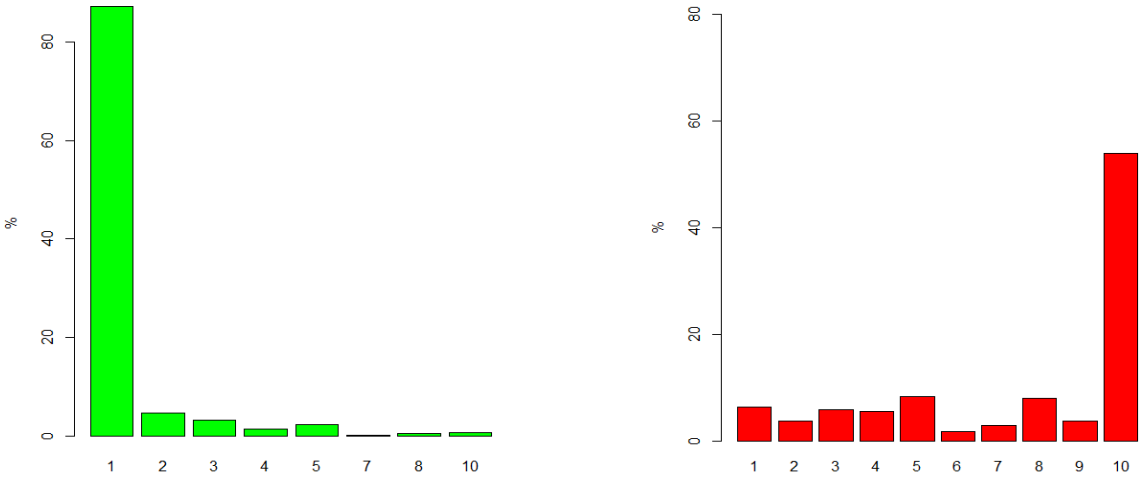
(a) Histograma de patologías benignas

(b) Histograma de patologías malignas

Figura 4: Histogramas con los valores de Clump distinguiendo según patologías benignas o malignas

En las patologías benignas con Clump se puede apreciar que el mayor porcentaje se encuentra entre los valores 1 y 5 y que los porcentajes en los valores superiores son menores al 5%, sin haber casos en los valores 9 y 10. Respecto a las patologías malignas, la mayor parte de casos se encuentra en los valores superiores a 4 (en los valores iguales o inferiores a 4, el porcentaje ronda el 5% o es incluso menor) y se da el menor número de casos para 1 y 2.

En la Figura 5 se muestran los histogramas con Bare_Nuclei.



(a) Histograma de patologías benignas

(b) Histograma de patologías malignas

Figura 5: Histogramas con los valores de Bare_Nuclei distinguiendo según patologías benignas o malignas

En las patologías benignas con Bare_Nuclei se puede apreciar que el mayor número de casos se encuentra en el valor 1 (porcentaje del 87.2%) y que hay pocos casos en el resto de valores, muy pocos en 7, 8 y 10, y ninguno en 6 ni 9. Respecto a las patologías malignas, el porcentaje es muy alto en el valor 10 (porcentaje del 54.0%) y el porcentaje es similar en el resto de valores (menor del 10%).

2.2. Descripción y planteamiento del estudio logístico

Se plantea realizar un modelo predictivo de regresión logística para predecir si los tumores de las pacientes son benignos o malignos en función de los resultados de las características citológicas comentadas en la sección anterior. Se procede a mostrar la ecuación de la que parte el modelo:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 * C + \beta_2 * UCSi + \beta_3 * UCS + \beta_4 * MA + \beta_5 * SCS + \beta_6 * BN + \beta_7 * BC + \beta_8 * NN + \beta_9 * M, \quad (2.1)$$

siendo p la probabilidad de que el tumor sea maligno, es decir, $p = P(Y = 1) = E[Y]$, y β_i, $i = 0, ..., 9$, los coeficientes de la ecuación de regresión logística asociados al término independiente (β_0) o a cada una de las variables predictoras del modelo.

La estrategia para obtener el modelo final va a ser "hacia atrás" o *backwards*, esto es, que se va a partir del modelo mostrado en la Ecuación 2.1 con todas las variables y se van a ir eliminando una a una aquéllas que no sean significativas, a un nivel de significación 0.05, comprobando el resultado del modelo después de eliminar cada variable que no sea significativa. Además, en cada paso en el que se vaya eliminando una variable, se compararán los valores AIC (Criterio de Información de Akaike) de los modelos resultantes. En caso de que al eliminar una variable, el AIC del modelo sea ligeramente mayor respecto al modelo que también contiene a dicha variable, al querer obtener un modelo predictivo se deja el modelo más sencillo, siguiendo así también el principio de parsimonia. Se utilizan modelos lineales generalizados siendo la función logit la función de enlace para obtener los resultados.

Una vez que se tenga el modelo ajustado con todas las variables significativas, se realiza la curva ROC para el mismo. La curva ROC sirve para poder apreciar gráficamente la capacidad de discriminación del modelo obtenido.

Se calculan los valores del AUC, que es la medida cuantitativa que sirve como medida de discriminación. Si el valor de AUC es muy alto, no se considerará introducir interacciones en el modelo para preservar un modelo más sencillo. En caso de obtener un valor de AUC que tenga margen de mejora, se probará a introducir

interacciones en el modelo y se preservarán en el mismo si son significativas y si se reduce el valor AIC de manera no despreciable respecto al modelo sin interacciones. También se obtiene el intervalo de confianza del valor AUC al 95 % con el método de DeLong [17].

El punto de corte óptimo, en este caso, es el valor de p condicionado a los valores de las covariables en la Ecuación 2.1; esto es, la probabilidad ajustada de clasificar un tumor nuevo como maligno o benigno. Cuando se quiere clasificar un nuevo caso, se compara el valor de p obtenido en función de los valores de las variables predictoras con el punto óptimo de corte, y si dicho valor es mayor que el punto óptimo, se clasifica el tumor como maligno. La obtención del punto óptimo de corte se realiza con el cálculo del índice de Youden, cuya expresión es la siguiente:

$$YI = max\left|Se(p) + Sp(p) - 1\right|, \tag{2.2}$$

siendo max el valor máximo, Se y Sp los valores de sensibilidad y especificidad respectivamente.

Los paquetes de R que se utilizan son pROC [18] y Epi [19]. Con el paquete pROC, se obtiene el intervalo de confianza al 95 % del valor AUC. Con el paquete Epi, se pueden ver en una misma figura la curva ROC, la estimación de los parámetros del modelo de regresión logística (sin el p-valor) con los errores estándar mediante modelos lineales generalizados (como se hace con la función glm en R), el valor AUC, parámetros epidemiológicos como la sensibilidad o la especificidad o el punto óptimo de corte obtenido a partir de la maximización de la sensibilidad y la especificidad.

2.3. Descripción de la obtención de nuevas bases de datos a partir de la original

Se van a generar nuevas bases de datos con datos *missing* según los mecanismos MCAR, MAR y MNAR, buscando que la pérdida de datos se sitúe entre el 20 y el 40 % de datos (el objetivo de manera puntual es la tercera parte) para las variables en la que haya pérdida.

Estas variables van a ser las dos más influyentes entre las que hayan quedado en el modelo final de regresión logística con la base de datos completa, que serán Clump y Bare_Nuclei. Las dos variables serán las mismas para cada mecanismo de pérdida de datos y el resto de variables no sufre pérdida de datos en ningún caso. El criterio para escogerlas ha sido seleccionar las dos más influyentes en el modelo logístico obtenido con la base de datos completa.

Una vez aplicado el procedimiento de pérdida de datos en cada caso, los datos de los individuos en los que se simula la pérdida se reemplazan por “NA”.

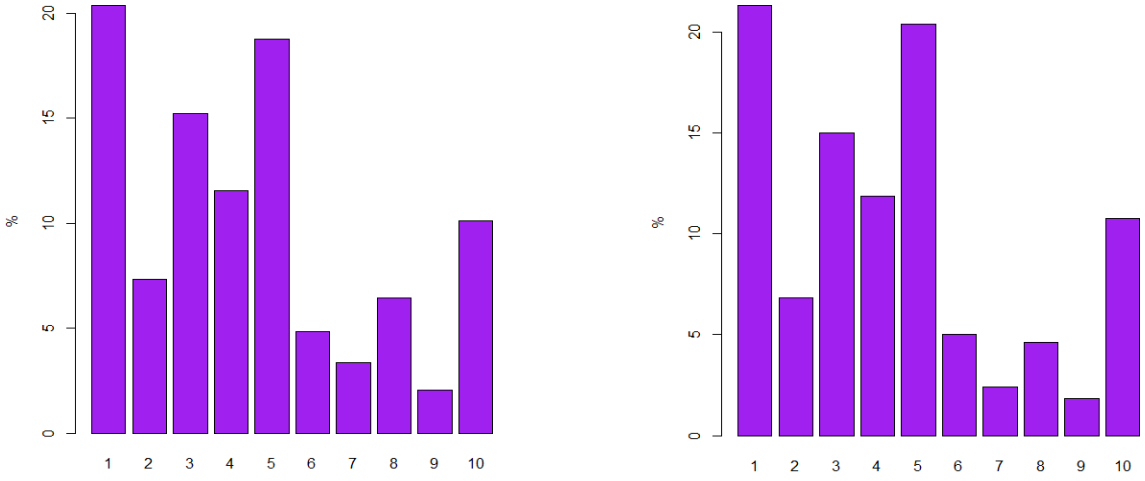
Los códigos de R para la obtención de las distintas bases de datos se pueden ver en los Anexos.

2.3.1. Base de datos con datos faltantes según mecanismo MCAR: B1

Para crear la base de datos MCAR se parte de la identificación de cada paciente, es decir, el orden de la fila que ocupa en la base de datos. Se asigna a cada paciente, independientemente de las demás, el valor de una variable aleatoria asociada a la pérdida de información, definida como $\{0, 1, 2, 3\}$, donde el valor "0" indica que no hay pérdida en la información de la paciente, el valor "1" indica que se pierde la información de la paciente en relación con la variable Clump, el valor "2" indica la pérdida del dato correspondiente a la variable Bare_Nuclei y, finalmente, el valor "3" indica que no se dispone de la información de la paciente en ninguna de las variables anteriores.

Para realizar la simulación se toma como criterio perder información en la tercera parte de las pacientes, que en el 30 % de los casos la información se pierda en ambas variables y el 70 % restante solo disponga de información en una de ellas, elegida al azar de manera equiprobable.

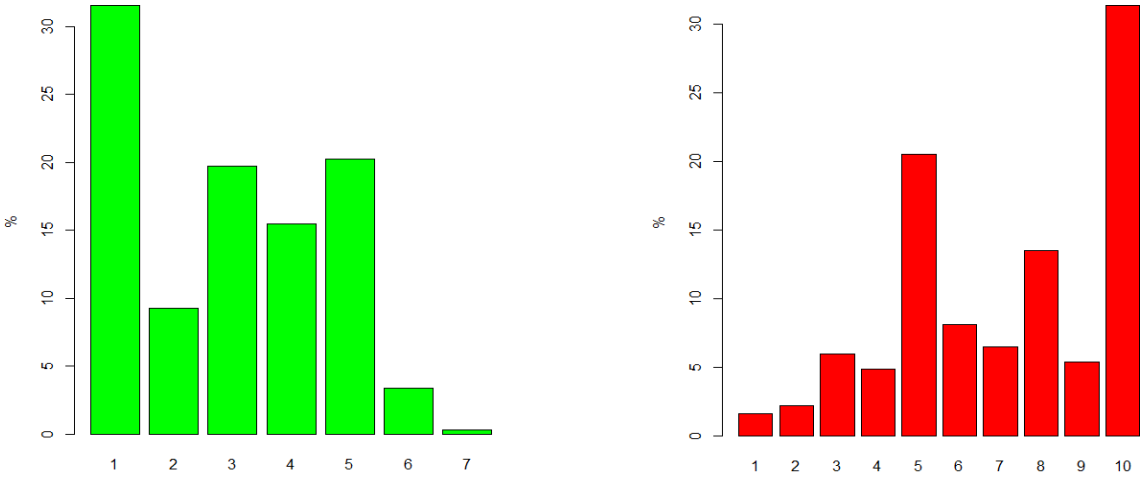
Se van a mostrar histogramas como en la Sección 2.1, para poder apreciar cómo afecta al comportamiento de las variables el mecanismo de pérdida de datos. Se muestran en primer lugar los histogramas de Clump para B0 y B1 en la Figura 6 y, posteriormente, distinguiendo según si las patologías son benignas o malignas con los datos de B1 en la Figura 7.



(a) Histograma de Clump para B0

(b) Histograma de Clump para B1

Figura 6: Histogramas con los valores de la variable Clump para B0 y B1



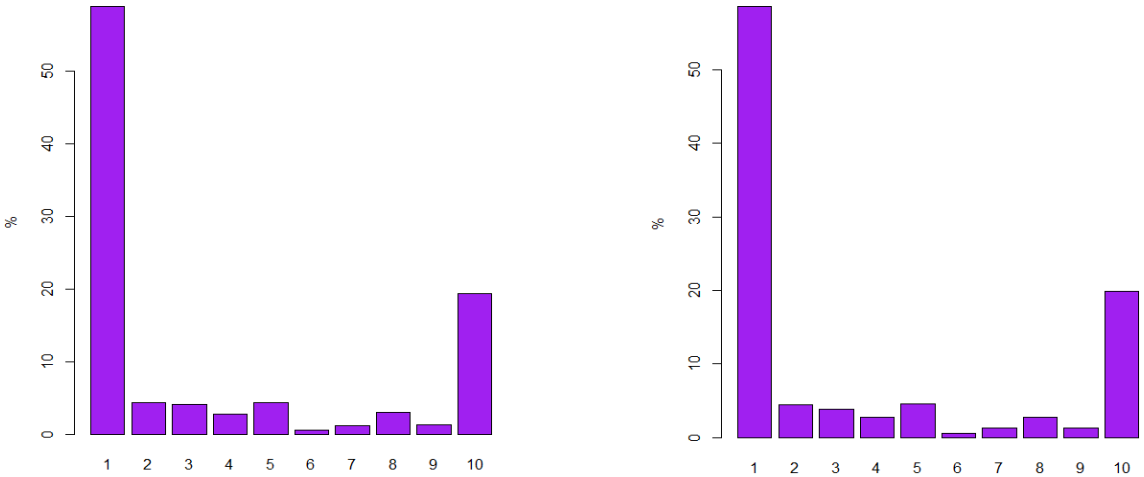
(a) Histograma de patologías benignas para B1

(b) Histograma de patologías malignas para B1

Figura 7: Histogramas con los valores de Clump en B1 distinguiendo según patologías benignas o malignas

Se puede apreciar en la Figura 6 que prácticamente el comportamiento de la variable Clump es el mismo entre B0 y B1, con alguna pequeña variación en el valor del porcentaje. Comparando las Figuras 4 y 7, se puede apreciar que el comportamiento de Clump también es muy similar, aunque en el caso de patologías benignas no hay casos para el valor 8 en la base de datos B1.

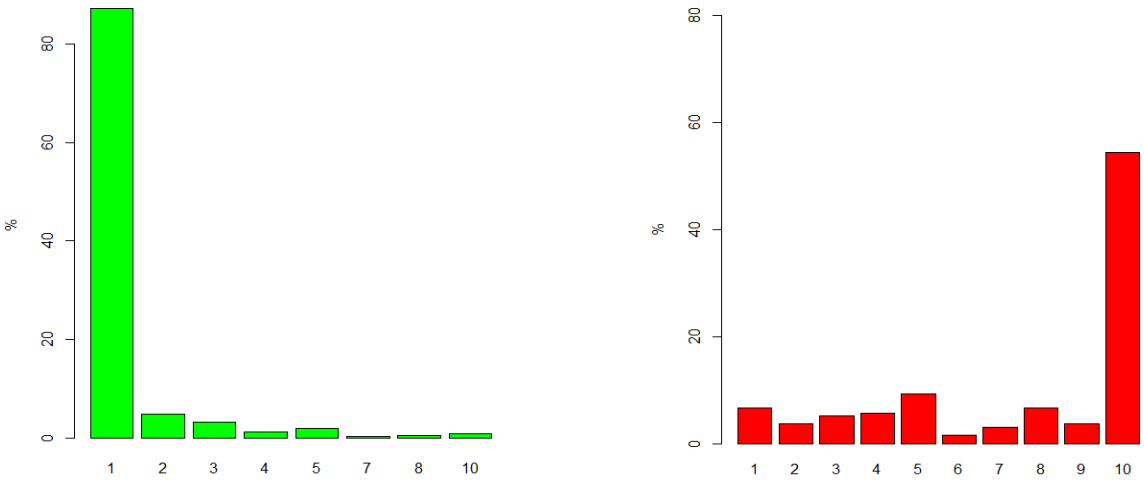
En las Figuras 8 y 9 se muestran las figuras equivalentes para Bare_Nuclei.



(a) Histograma de Bare_Nuclei para B0

(b) Histograma de Bare_Nuclei para B1

Figura 8: Histogramas con los valores de la variable Bare_Nuclei para B0 y B1



(a) Histograma de patologías benignas para B1

(b) Histograma de patologías malignas para B1

Figura 9: Histogramas con los valores de Bare_Nuclei en B1 distinguiendo según patologías benignas o malignas

Como ocurría con Clump, el comportamiento de Bare_Nuclei es prácticamente el mismo en B1 que en B0, habiendo diferencias muy pequeñas solamente en los valores de los porcentajes, tal como se puede observar en la Figura 8. Comparando las Figuras 5 y 9, no se aprecian apenas diferencias entre ellas a excepción de pequeñas variaciones de los valores de los porcentajes.

2.3.2. Base de datos con datos faltantes según mecanismo MAR: B2

Para ver el efecto de los métodos de tratamiento de datos faltantes con mecanismo MAR, se tiene que simular la pérdida de datos con dependencia de variables que entren en el modelo final. La pérdida de datos en cada variable seleccionada depende de los datos de alguna de las demás variables.

La variable de la que va a depender la pérdida de datos es Marginal_Adh. Los individuos en los que habrá pérdida de datos serán aquellos en los que en los que los valores de Marginal_Adh sean superiores a 2 (casi el 34 % de los individuos).

Por consiguiente, la simulación de pérdida de datos tendrá lugar sobre las variables Clump y Bare_Nuclei, en aquellas pacientes que presenten valor superior a 2 en la variable Marginal_Adh. En este sentido, la variable que asigna la pérdida tomará también los valores {0, 1, 2, 3} (con el mismo significado que el descrito en la construcción de la base B1). En este caso el valor "0" se asignará a las pacientes con Marginal_Adh inferior o igual a 2, y la asignación de valores {1, 2, 3} se asignará al resto de pacientes, independientemente, y con el mismo criterio descrito en la Sección 2.3.1.

Se van a mostrar histogramas siguiendo el esquema seguido en la sección anterior, para poder apreciar cómo afecta al comportamiento de las variables la pérdida de datos en función de Marginal_Adh. Se muestran los mismos histogramas que en las comparaciones realizadas en B1 y, además, los histogramas para Marginal_Adh, que serán los que se muestren en primer lugar, en las Figuras 10 y 11.

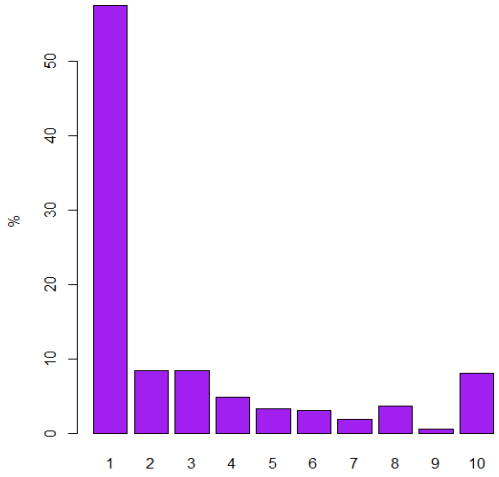
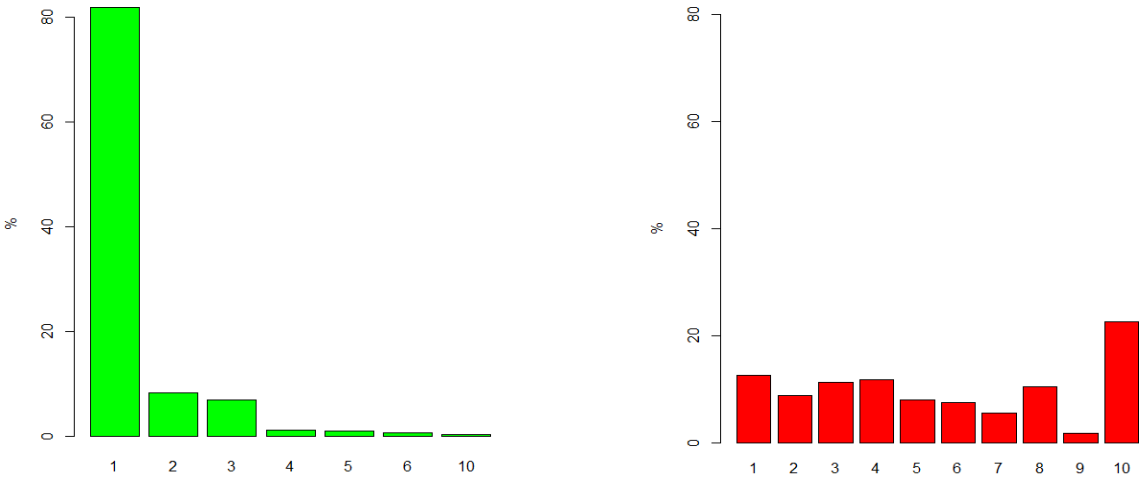


Figura 10: Histogramas con los valores de Marginal_Adh



(a) Histograma de patologías benignas

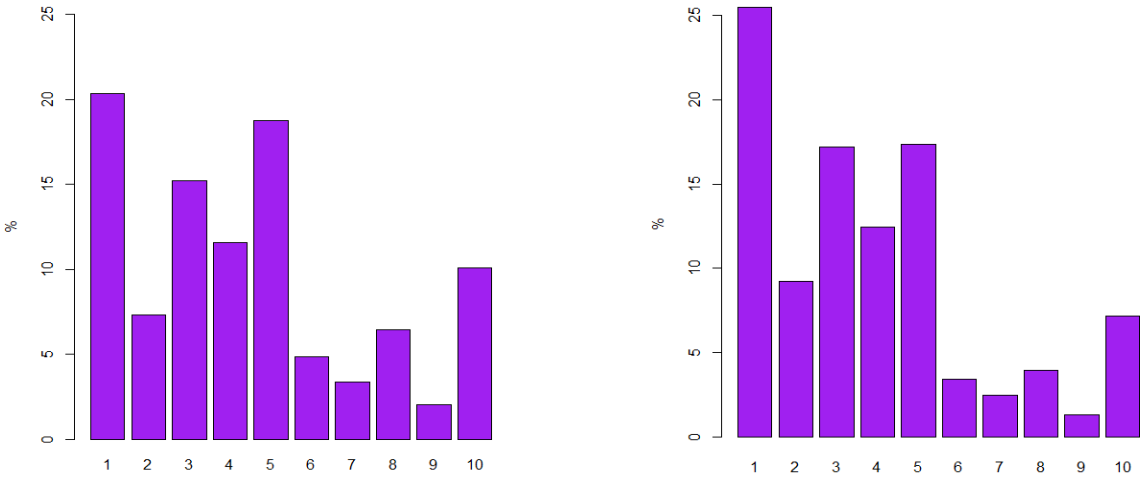
(b) Histograma de patologías malignas

Figura 11: Histogramas con los valores de Marginal_Adh según patologías benignas o malignas

En Marginal_Adh, la mayor parte de los casos se dan con el valor 1 (porcentaje del 57.4%) y, en segundo lugar, los valores más repetidos son 2, 3 y 10 (porcentajes próximos al 10% para los tres valores), tal como se puede ver en la Figura 10. En la Figura 11 se puede observar que, para las patologías benignas, se da la misma tendencia que a nivel general, aunque sin apenas casos para el valor 10, y el resto de casos solamente se dan entre los valores 1 y 6, habiendo un 81.8% en el valor 1. Con las patologías malignas, la mayoría de casos se sitúa en el valor 10 (porcentaje del 22.6%) y los demás casos están bastante repartidos entre los demás valores,

habiendo un porcentaje similar de casos en los valores 1, 3, 4 y 8, estando dicho porcentaje comprendido entre el 10 y el 15%.

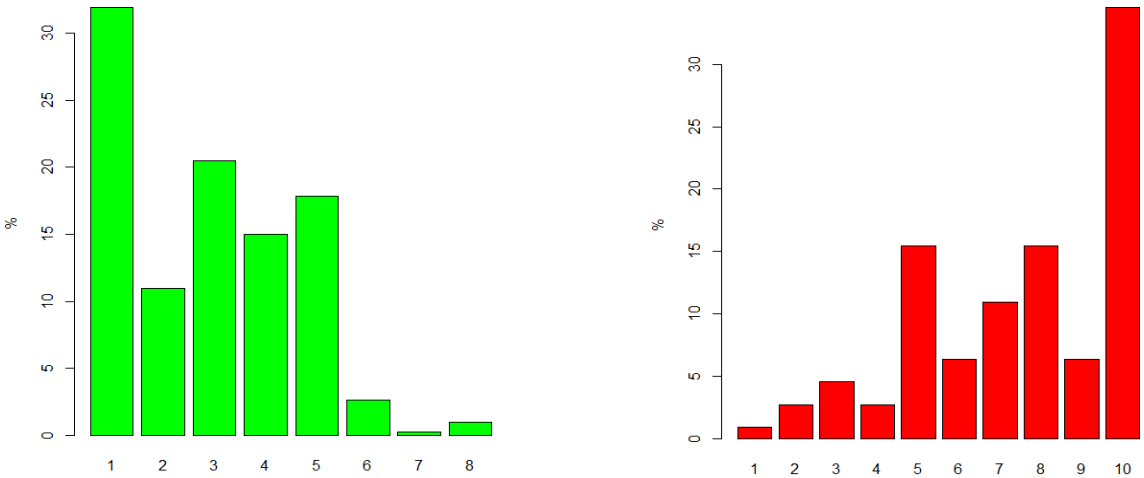
Se procede a mostrar en la Figura 12 los histogramas de Clump para B0 y B2 y, posteriormente, en la Figura 13 distinguiendo según si las patologías son benignas o malignas con los datos de B2.



(a) Histograma de Clump para B0

(b) Histograma de Clump para B2

Figura 12: Histogramas con los valores de la variable Clump para B0 y B2



(a) Histograma de patologías benignas para B2

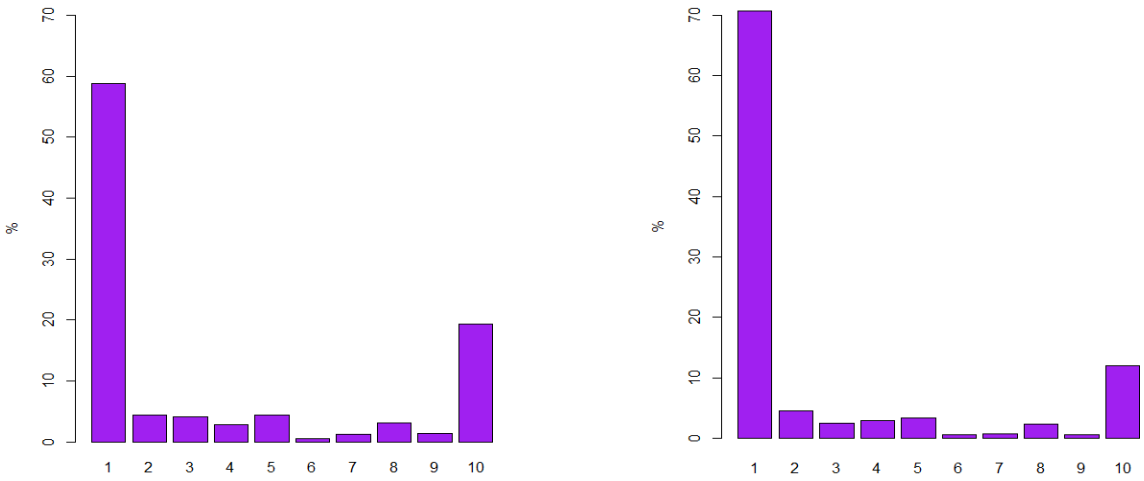
(b) Histograma de patologías malignas para B2

Figura 13: Histogramas con los valores de Clump en B2 distinguiendo según patologías benignas o malignas

Se puede apreciar en la Figura 12 que el comportamiento de la variable Clump cambia ligeramente entre B0 y B2, ya que la diferencia entre los valores más pequeños (del 1 al 5) y los valores mayores aumenta, lo que se puede apreciar comparando los valores 1, 3 o 4 (con 20.4%, 15.2% y 11.6% respectivamente, para B0, frente a 25.5%, 17.2% y 12.5%, respectivamente, para B2) contra 6, 7, 8 o 10 (con 4.8%, 3.4%, 6.4% y 10.1% respectivamente, para B0, frente a 3.4%, 2.5%, 4.0% y 7.2%,

respectivamente, para B2). Comparando las Figuras 4 y 13, se puede apreciar que el comportamiento de Clump es similar para los casos de las patologías benignas, aunque en el caso de las patologías malignas aumenta la diferencia entre el porcentaje en el valor 10 (el porcentaje en este valor aumenta del 28.9 al 34.5%) y los demás (por ejemplo, el porcentaje del valor 5 pasa de ser del 18.8% a ser 15.5%).

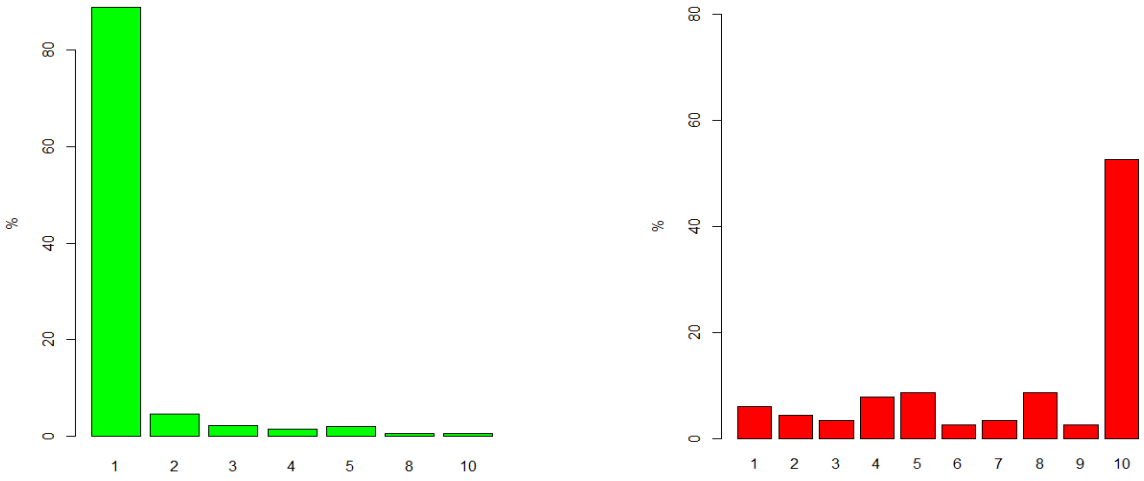
Se muestran las figuras equivalentes para Bare_Nuclei en las Figuras 14 y 15.



(a) Histograma de Bare_Nuclei para B0

(b) Histograma de Bare_Nuclei para B2

Figura 14: Histogramas con los valores de la variable Bare_Nuclei para B0 y B2



(a) Histograma de patologías benignas para B2

(b) Histograma de patologías malignas para B2

Figura 15: Histogramas con los valores de Bare_Nuclei en B2 distinguiendo según patologías benignas o malignas

En la Figura 14 se puede ver que el comportamiento de Bare_Nuclei es prácticamente el mismo en B2 que en B0, aunque aumenta claramente el porcentaje del valor 1 (pasa de ser 58.9% en B0 a ser 70.7% en B2) y disminuye el porcentaje del valor 10 (pasa de ser 19.3% en B0 a ser 11.9% en B2). Comparando las Figuras 5 y 15, el comportamiento es muy similar para las patologías benignas a excepción de

que con B2 no hay casos para el valor 7; lo mismo ocurre con las patologías malignas (sin perder la totalidad de los casos en alguno de los valores).

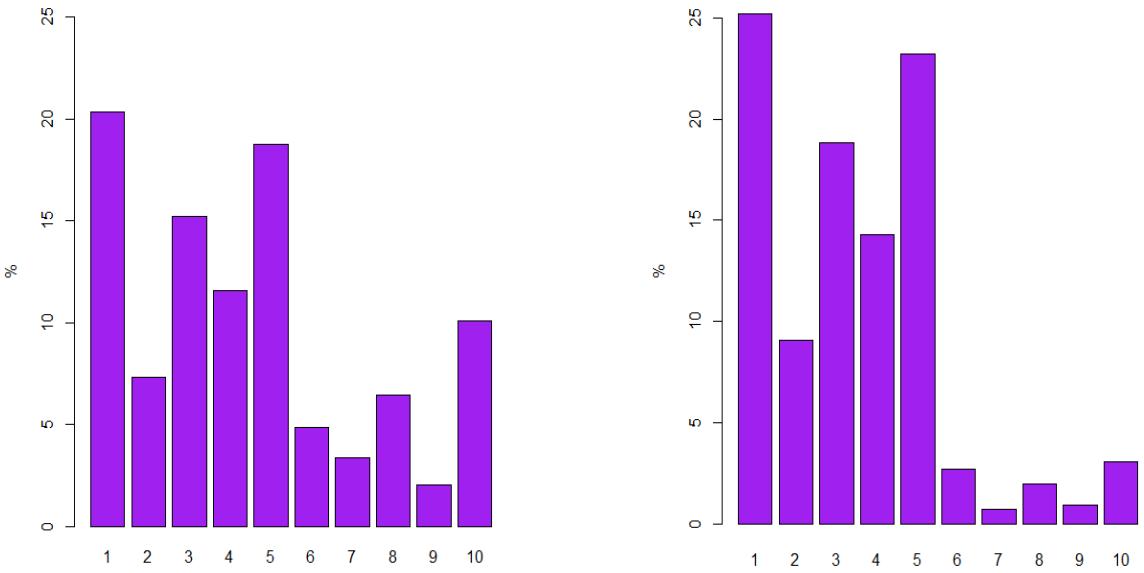
2.3.3. Base de datos con datos faltantes según mecanismo MNAR: B3

En esta situación, la pérdida de datos en cada variable del modelo logístico depende de cualquiera de las variables incluidas en el modelo. La pérdida de datos se va a realizar en función de una de las variables que no sufre pérdida y de las variables que sufren pérdida.

Para la creación de B3, se va a realizar la pérdida como en B2 utilizando la variable Marginal_Adh, siendo los individuos que van a perder datos aquéllos en los se cumple la condición de que los valores de Marginal_Adh son superiores a 2.

En los casos en los que se cumple la condición anterior, para Clump, la pérdida de datos tiene lugar cuando los valores de Clump en la base de datos original sean mayores a 5; y con Bare_Nuclei, la pérdida de datos se produce cuando los valores originales de dicha variable son superiores a 7.

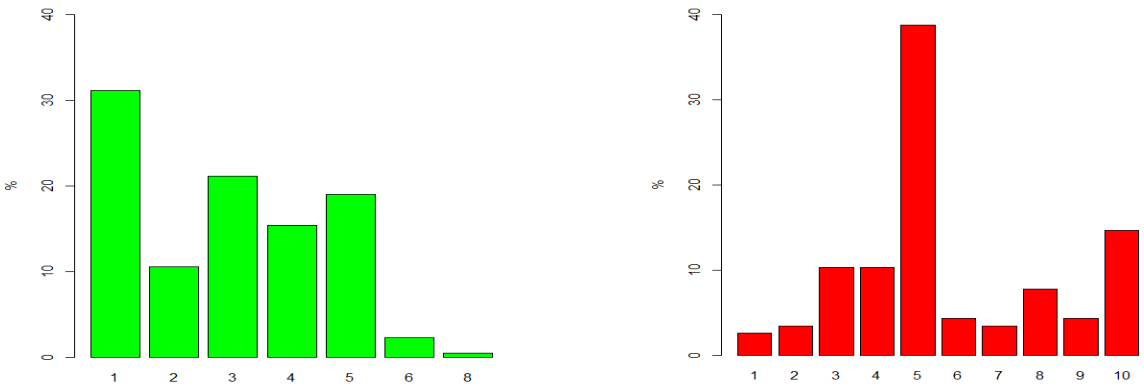
Se van a mostrar en las Figuras 16 y 17 los histogramas de Clump y Bare_Nuclei siguiendo el esquema seguido en la sección en la que se explicó la obtención de B1.



(a) Histograma de Clump para B0

(b) Histograma de Clump para B3

Figura 16: Histogramas con los valores de la variable Clump para B0 y B3



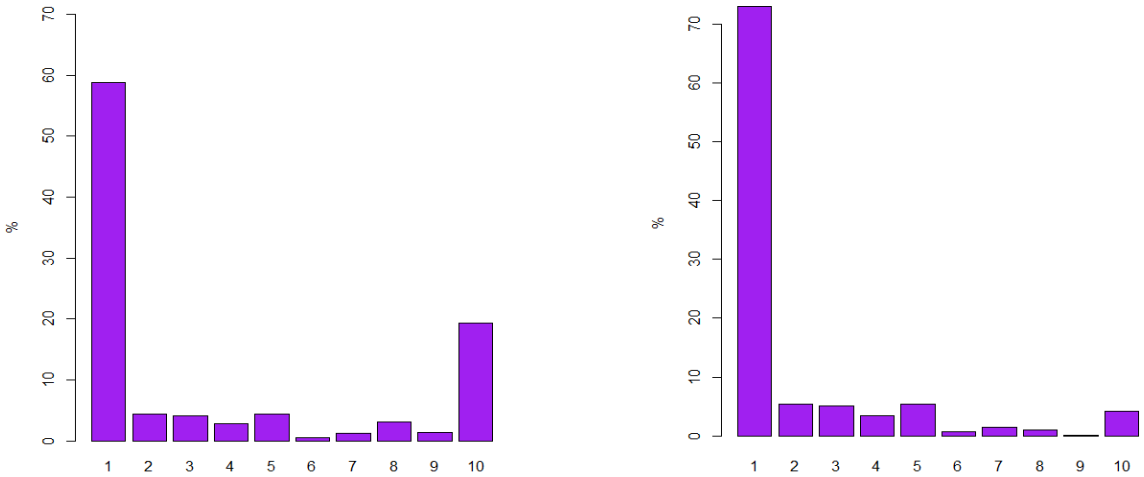
(a) Histograma de patologías benignas para B3

(b) Histograma de patologías malignas para B3

Figura 17: Histogramas con los valores de Clump en B3 distinguiendo según patologías benignas o malignas

Se puede apreciar en la Figura 16 que el comportamiento de la variable Clump cambia considerablemente entre B0 y B3, ya que el porcentaje en los valores superiores a 5 disminuye en gran medida, sin destacar ninguno de los valores sobre los demás, al pasar de estar el máximo del porcentaje de dichos valores en un 10.1% a estar en un 3.1%, preservándose aproximadamente el mínimo entre B0 y B3. Comparando las Figuras 4 y 17, se puede apreciar que el comportamiento de Clump es similar para los casos de las patologías benignas (aunque no hay casos en el valor 7 en B3) pero no para los casos de las malignas, ya que el porcentaje en los valores superiores a 5 se sitúa en un número similar al de los casos inferiores a 5, el porcentaje con el valor 5 es muy superior al del resto de valores (en B0 era 18.8% y pasa a ser 38.8% en B3) y el porcentaje con el valor 10 disminuye considerablemente (de 28.9% en B0 a 14.7% en B3).

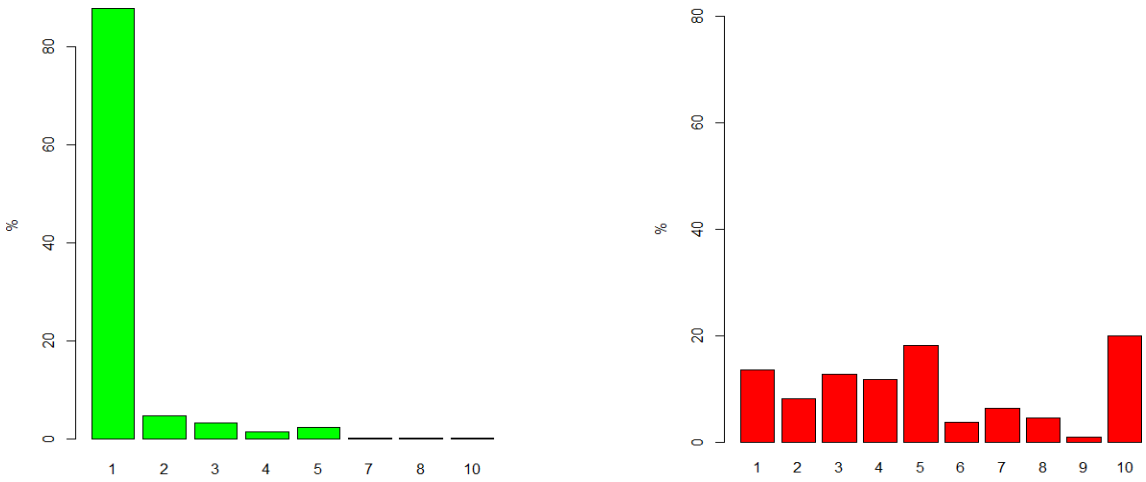
Se muestran las figuras equivalentes para Bare_Nuclei en las Figuras 18 y 19.



(a) Histograma de Bare_Nuclei para B0

(b) Histograma de Bare_Nuclei para B3

Figura 18: Histogramas con los valores de la variable Bare_Nuclei para B0 y B3



(a) Histograma de patologías benignas para B3

(b) Histograma de patologías malignas para B3

Figura 19: Histogramas con los valores de Bare_Nuclei en B3 distinguiendo según patologías benignas o malignas

El comportamiento de Bare_Nuclei cambia de B0 a B3, ya que el porcentaje para el valor 10 disminuye en B3 (de ser 19.3% en B0 a ser 4.2% en B3), y es parecido al porcentaje de los valores entre 2 y 5, y también disminuye en los valores 8 y 9 a la vez que aumenta de manera notable el porcentaje en el valor 1 (de ser 58.9% en B0 a ser 73.0% en B3). Comparando las Figuras 5 y 19, el comportamiento es muy similar para las patologías benignas, pero en las patologías malignas es diferente, ya que el porcentaje para el valor 10 es ligeramente superior al del valor 5 en B3 (descenso de 54.0% en B0 a 20.0% en B3 del valor 10 y aumento de entre 4 y 10 puntos porcentuales para cada valor entre 1 y 5) y el porcentaje entre los valores 6 y 9 es inferior al porcentaje en el resto de valores.

2.4. Descripción de los métodos empleados para trabajar con datos faltantes y estudio logístico

Los métodos que se van a utilizar para el tratamiento de datos faltantes son el análisis de casos completos, sustitución por la media, indicador de pérdida de datos con variables adicionales (*dummies*), MICE y *Missing Forest*. Para todos estos métodos el estudio de regresión logística que se llevará a cabo será el paralelo al realizado sin datos faltantes.

2.4.1. Análisis de casos completos

El análisis de casos completos consiste en eliminar de la base de datos a aquellos individuos que tengan pérdida de datos en cualquiera de las variables predictoras del modelo. Se ajusta el modelo de regresión logística con los individuos que tengan todos sus datos disponibles. A partir de ahora, se identificará este método con las siglas CC (Casos Completos).

2.4.2. Sustitución por la media

La sustitución por la media consiste en reemplazar los valores faltantes de cada una de las dos variables por sus correspondientes medias de los datos observados. A partir de ahora, se identificará este método con las siglas MI.

2.4.3. Uso de variables indicadoras de pérdida de datos

Este método de imputación simple consiste en introducir variables indicadoras de pérdida de datos. Se crea una variable indicadora para cada variable en la que haya pérdida de datos y los datos faltantes se reemplazan por 0. El valor de la variable indicadora es 1 en caso de que haya pérdida de datos en su variable predictora correspondiente para cada mujer de la base de datos y 0 en caso contrario. Un problema de esta estrategia es que se puede producir sesgo en las estimaciones de los coeficientes de otras variables predictoras del modelo debido a que se fuerza que la pendiente sea constante [11].

En este caso, el modelo de regresión logística contiene las mismas variables predictoras que en el modelo realizado para B0 más las variables indicadoras de pérdida de datos. A partir de ahora, se indentificará este método con las siglas VIP (Variable Indicadora de Pérdida).

2.4.4. Algoritmo MICE

El algoritmo MICE es un método de imputación múltiple basado en ecuaciones encadenadas. Puesto que este algoritmo se implementa como una concatenación de procedimientos de imputaciones simples o univariantes, se comienza describiendo el método del ajuste de la media predictiva, *predictive mean matching*, que será el que se utilice en este algoritmo MICE.

Para cada una de las dos variables predictoras con observaciones *missing*, se lleva a cabo un ajuste de regresión lineal por mínimos cuadrados, donde la variable dependiente es la que tiene datos faltantes a imputar y que denotamos por $Y = (Y_{obs}, Y_{mis})$, siendo Y_{obs} la parte correspondiente a las unidades observadas e Y_{mis} la correspondiente a las unidades con observaciones faltantes. Las predicciones de

las observaciones faltantes se van a realizar teniendo en cuenta las distribuciones a posteriori de los parámetros del modelo de regresión, vector de coeficientes de regresión β y varianza del error σ^2, tomando distribuciones a priori no informativas para estos parámetros.

El algoritmo para la búsqueda de la media predictiva tiene los siguientes pasos [20]:

1. Ajustar el modelo $Y_{obs} = X_{obs}\beta + e, e \sim N(0, \sigma^2 I)$, por mínimos cuadrados obteniendo $\hat{\beta}$, $\hat{\sigma}^2$, $\hat{e}$. X_{obs} es la matriz formada por los valores de las variables regresoras, sin observaciones faltantes, en las unidades en las que se han obtenido observaciones de la variable Y.

2. Obtener $\sigma^{2*} = \frac{\hat{e}'\hat{e}}{A}$, donde A es un valor simulado de una distribución $\chi^2_{n_{obs}-r}$, siendo n_{obs} el número de observaciones observadas de Y y r el número de coeficientes de regresión. Hay que tener en cuenta que la distribución a posteriori de σ^2 es $\chi^2_{n_{obs}-r}$.

3. Obtener β^* de la distribución normal multivariante $N(\hat{\beta}, \sigma^{2*}(X'_{obs}X_{obs})^{-1})$. Cabe considerar que la distribución a posteriori de β es $N(\hat{\beta}, \sigma^2(X'_{obs}X_{obs})^{-1})$.

4. Calcular $\hat{Y}_{obs} = X_{obs}\hat{\beta}$ y $\hat{Y}_{mis} = X_{mis}\beta^*$.

5. Para cada unidad i con Y_i no observado, calcular el vector $\Delta = |\hat{Y}_{obs} - \hat{Y}_{mis,i}|$.

6. Seleccionar aleatoriamente un valor de entre ($\Delta^{(1)}$,$\Delta^{(2)}$,$\Delta^{(3)}$), siendo estos tres valores los elementos más pequeños, respectivamente, del vector Δ, y tomar el correspondiente $Y_{obs,i}$ como imputación.

Una vez descrito el método del ajuste de la media predictiva pasamos a describir al algoritmo MICE, cuyos pasos son los siguientes [4]:

1. Examinar los patrones de datos faltantes e identificar las variables con datos perdidos que requieren imputación. Decidir la secuencia de la imputación.

2. Inicializar los valores *missing* con un método de imputación simple como sustituir por la media.

   Para cada iteración t, $t = 1, 2...$

3. Construir el modelo de imputación para la primera variable que requiere imputación y, en el caso de este trabajo, se utiliza un modelo de regresión lineal en el que la variable dependiente es la variable que requiere imputación, y, las variables independientes, las demás que hay en la base de datos original.

4. Reemplazar los valores faltantes para la variable Y_1 por las predicciones obtenidas por el método del ajuste de la media predictiva. Para los siguientes modelos de imputación que requieran los valores de la variable Y_1, se toman sus valores observados y los imputados, considerando dicha variable como completamente observada.

5. Repetir los pasos 3 y 4 para cada una de las restantes variables con datos faltantes, $Y_2, ..., Y_p$ (sólo Y_2 en este caso).

6. Repetir los pasos del 2-5 para obtener tantos conjuntos de datos como número de imputaciones que se haya elegido (m).

7. Obtener la estimación de parámetros con cada uno de los m conjuntos de datos (en este caso los parámetros correspondientes a la regresión logística) y agrupar los resultados para obtener el ajuste final.

Se usarán cinco iteraciones y cinco imputaciones para MICE, que son los valores por defecto que hay en R. Ya se comentó anteriormente que el paquete utilizado para este método en R es mice [13].

2.4.5. Missing Forest

Como último método de tratamiento de datos faltantes, se procede a explicar el método Missing Forest. Es un método no paramétrico que puede usarse cuando se trabaja con diferentes tipos de variables conjuntamente. Es un método iterativo y está basado en la imputación de los datos faltantes mediante la predicción de los mismos mediante el método Random Forest, consistente en la realización de múltiples árboles de decisión sobre muestras de un conjunto de datos. La idea es hacer muchas predicciones con menos variables y menos observaciones y al final quedarse con un "promedio" de las mismas.

El paquete de R que se utiliza para implementar este método es missForest [21]. A partir de ahora, este método se identificará con las siglas MF (Missing Forest).

Todos los métodos descritos en este epígrafe serán los se usarán para imputar los datos faltantes en cada una de las bases de datos B1, B2 y B3, descritas en la Sección 2.3.

3. Resultados

Se van a presentar los resultados obtenidos para los modelos logísticos y para las curvas ROC en la base de datos original y en las bases de datos obtenidas según los tres mecanismos de pérdida de datos y según cada método de tratamiento de datos faltantes.

3.1. Resultados del estudio diagnóstico basado en B0

Se van a proceder a explicar los pasos seguidos para obtener el modelo ajustado de regresión logística, mostrando su ecuación, el valor AUC de dicho modelo con su intervalo de confianza al 95 %, el punto óptimo de corte y la curva ROC con la estimación de parámetros del modelo. En la Tabla 2 se resumen los valores obtenidos de diversos parámetros en los pasos realizados para obtener el modelo final.

Las variables predictoras que han quedado en el modelo, que aportan información significativa, son Clump, Marginal_Adh, Unif_Cell_Shape, Bare_Nuclei, Bland_Chromatin y Normal_Nucleoli. En primer lugar, la variable que se elimina es Unif_Cell_Size debido a que tenía el p-valor más alto en el modelo con todas las variables (0.976). Posteriormente, se elimina la variable Single_Cell_Size al tener el p-valor más alto del modelo en el que no se incluye la anterior variable, que ya había sido eliminada, siendo dicho valor 0.537. La última variable en ser eliminada es Mitoses, cuyo p-valor era el único no significativo en el modelo con las variables restantes (0.102). El modelo que incluía Mitoses tiene un valor AIC ligeramente menor que el modelo con las variables predictoras mencionadas al principio del párrafo (119.27 contra 121.14), pero al buscar un modelo predictivo y al no haber gran diferencia entre los valores AIC, el modelo escogido ha sido el más simple, que tiene todas las variables significativas. Por lo tanto, el modelo ajustado de regresión logística resultante es:

$$ln\left(\frac{p}{1-p}\right) = -9{,}77 + 0{,}62 * C + 0{,}35 * UCSh + 0{,}34 * MA + 0{,}38 * BN + 0{,}47 * BC + 0{,}24 * NN. \quad (3.1)$$

Despejando p de esta ecuación y llamando A a la parte no logarítmica de la Ecuación 3.1, se puede obtener la expresión de la probabilidad de tener un tumor maligno:

$$p = \frac{exp(A)}{1 + exp(A)}. \quad (3.2)$$

Se recuerda que el valor de p viene ajustado por el valor de las covariables que se pueden ver en la Ecuación 3.1·

La Tabla 2 resume los pasos seguidos para obtener el modelo ajustado final.

Paso	Variable que se elimina	AIC	p-valor variable eliminada
1	Unif_Cell_Size	122.89	0.976039
2	Single_Cell_Size	120.89	0.536876
3	Mitoses	119.27	0.101598
4	Ninguna	121.14	No aplica

Tabla 2: Pasos realizados para obtener el modelo de regresión logística

En la Tabla 2, el paso 1 indica el resultado del modelo con todas las variable incluidas en la base de datos original. Cuando se indica la variable que se elimina en un paso, los resultados del modelo sin esa variable se muestran en el paso siguiente.

En la Figura 20 se muestra la curva ROC del modelo de regresión logística junto con la ecuación del modelo, la estimación de parámetros con el error estándar y el punto óptimo de corte.

El valor AUC del modelo es 0.996 y el intervalo de confianza al 95 % (0.9928,0.999). Al obtener un valor tan alto para el AUC, no se considera necesario añadir interacciones al modelo debido al escaso margen de mejora teniendo en cuenta que se obtendría un modelo más complejo que realmente no daría un beneficio mayor.

El punto óptimo de corte obtenido es 0.106. Esto significa que los tumores con probabilidades estimadas mediante la Ecuación 3.2 mayores que el valor de este punto de corte, se diagnosticarán como tumores malignos y en caso contrario como benignos.

Los p-valores obtenidos para cada uno de los parámetros fueron (siguiendo el orden de la Figura 20) 0 ($<$ 2e-16), $5.62{*}10^{-6}$, 0.034, 0.004, $5.45{*}10^{-5}$, 0.005 y 0.025.

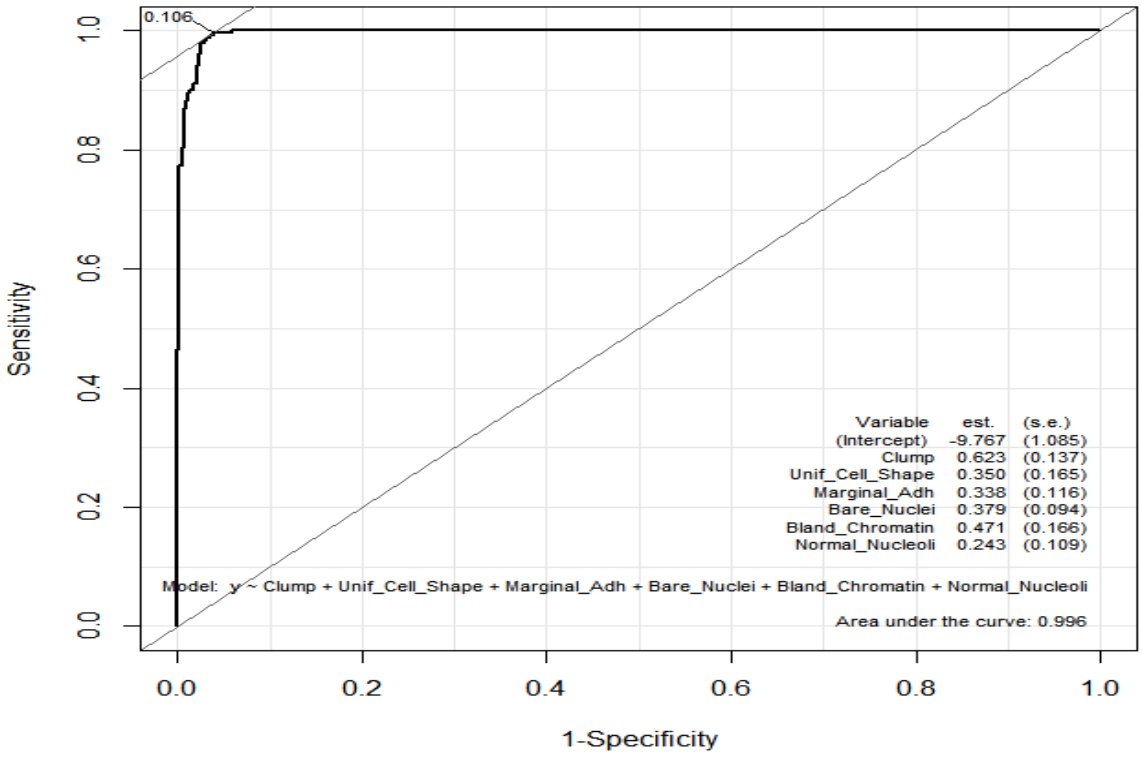


Figura 20: Curva ROC y estimación de parámetros del modelo logístico en B0

3.2. Resultados del estudio diagnóstico basado en B1

Se trabaja con la base de datos con datos faltantes según el mecanismo MCAR. Se realiza la regresión logística a cada una de las bases de datos con datos imputados (eliminados en el caso del análisis con casos completos) y, tras ajustar los modelos, se muestran los resultados correspondientes a los modelos ajustados. En concreto, se van a analizar por un lado, la estimación de los coeficientes con los errores estándar, el AIC del modelo obtenido en cada caso y se presta atención a las variables que no sean significativas (nivel de significancia 0.05) y a las dos variables con menor p-valor (las más significativas). Por otro lado, se analizan también las curvas ROC y los valores relacionados con ellas, que son los valores AUC (con su intervalo de confianza al 95 %) y los puntos óptimos de corte. En las Tablas 3 y 4, se resumen los resultados obtenidos de las bases de datos generadas a partir de B1. En las Figuras 21-25 se muestran las curvas ROC para los distintos métodos, con las estimaciones de los parámetros del modelo ajustado y sus errores estándares, así como los valores del AUC y del punto de corte óptimo.

Se resumen los resultados relacionados con las curvas ROC y con los modelos de regresión logística de las bases de datos generadas a partir de B1 en la Tabla 3. Antes de mostrar la tabla, cabe mencionar que los *dummies* de Clump y Bare_Nuclei

se denotarán como D_C y D_BN respectivamente. Para indicar las variables no significativas y las dos con menor p-valor (variables más significativas), se denotarán en las columnas como VNS y VMS. El punto de corte se denotará como PC. El intervalo de confianza al 95% se denotará como IC95.

Método	VNS	VMS	AIC	PC	AUC	IC95
CC	UCS, NN	C, MA	79.331	0.250	0.997	(0.9941, 0.9997)
MI	No	C, MA	126.12	0.442	0.995	(0.9924, 0.9985)
MICE	NN	C, BN	115.84	0.185	0.996	(0.9939, 0.9989)
MF	NN, UCS	C, MA	115.05	0.355	0.996	(0.9940, 0.9989)
VIP	D_BN	C, MA	128.81	0.136	0.995	(0.9924, 0.9986)

Tabla 3: Resultados de los modelos logísticos en B1

La Tabla 4 muestra los p-valores obtenidos para todas las covariables comunes en los cinco métodos en las bases de datos generadas a partir de B1:

Variables predictoras	CC	MI	MICE	MF	VIP
Término independiente	1.06e-11	<2e-16	<2e-16	<2e-16	<2e-16
Clump	9.15e-05	3.72e-06	1.41e-06	5.26e-07	8.93e-06
Unif_Cell_Shape	0.38516	0.000510	0.003587	0.050877	0.00066
Marginal_Adh	0.00127	2.99e-05	0.000457	0.000156	3.01e-05
Bare_Nuclei	0.00210	0.000636	0.000160	0.000264	0.00154
Bland_Chromatin	0.01322	0.001643	0.005288	0.011673	0.00117
Normal_Nucleoli	0.08350	0.022070	0.097873	0.063625	0.01513

Tabla 4: p-valores de los modelos logísticos en B1

Hay que añadir que, en el método con variables indicadoras de pérdida de datos, los p-valores para el *dummy* de Clump y para el de Bare_Nuclei son 0.001 y 0.525 respectivamente.

Tal como se puede observar en las Tablas 3 y 4, en todos los métodos la variable Clump es una de las dos más significativas y es la que tiene mayor aportación individual de información en el modelo de regresión logística, al igual que ocurría en el modelo ajustado con la base de datos original, es decir, en todos los casos el p-valor correspondiente al contraste individual de la variable Clump es el más pequeño. En cuanto a la segunda variable con mayor aportación individual (o menor p-valor para el correspondiente contraste individual), se encuentran diferencias entre los métodos. Para la base de datos original esta segunda variable era Bare_Nuclei, que coincide para el método MICE y sin embargo para los demás métodos, la variable es Marginal_Adh.

En las Tablas 3 y 4, el único método en el que todas las variables predictoras del

modelo ajustado son significativas es la sustitución por la media. En MICE hay una variable no significativa (Unif_Cell_Shape), así como en el método con variables indicadoras de pérdida (el *dummy* de Bare_Nuclei), mientras que en los demás métodos hay dos variables no significativas (Unif_Cell_Shape y Normal_Nucleoli).

Al comparar los valores AIC de los distintos modelos ajustados, se ve en la Tabla 3 que el valor AIC tiene su menor valor en el análisis de casos completos y su mayor valor en el método con variables indicadoras de pérdida de datos. El valor AIC del método que sustituye por la media es similar al mayor AIC, mientras que los métodos de imputación múltiple tienen valores intermedios, aunque no demasiado alejados de los valores AIC del análisis de casos completos. Se recuerda que el valor AIC para el modelo de regresión logística ajustado con la base de datos original fue de 121.14 (Tabla 2).

Como se puede ver en las Figuras 21-25, el análisis de casos completos es el método en el que los errores estándar son mayores para cada una de las covariables de los modelos ajustados de regresión logística. En el análisis de casos completos, destacan, por su bajo valor, el coeficiente de la covariable Unif_Cell_Shape (valor mucho más bajo al resto de coeficientes de dicha variable y del coeficiente de la variable en B0), y por su alto valor en relación a los coeficientes de los demás métodos, los coeficientes de Marginal_Adh y, sobre todo, de Bland_Chromatin (en esta covariable es el único método cuyo coeficiente es claramente superior al de B0). En todos los métodos, tanto el coeficiente como el error estándar de Clump tienen un valor mayor que el de B0, siendo el valor del coeficiente bastante similar para todos los casos de B1. El único método cuyo coeficiente de Unif_Cell_Shape tiene un valor muy próximo al de B0 es Missing Forest, siendo en los demás casos (salvo análisis de casos completos) el coeficiente superior; no hay grandes diferencias con el error estándar. Para Marginal_Adh, todos los métodos salvo el análisis de casos completos tienen coeficientes similares y mayores que el de B0, aunque el error estándar sí es muy similar. Con Bare_Nuclei no hay grandes diferencias entre los métodos tanto en el valor de los coeficientes como en el error estándar; y lo mismo ocurre con Bland_Chromatin con excepción del análisis de casos completos, y con Normal_Nucleoli. Se aprecia que, en casi todos los casos, el error estándar con los métodos de imputación múltiple es mayor respecto a los métodos de imputación simple.

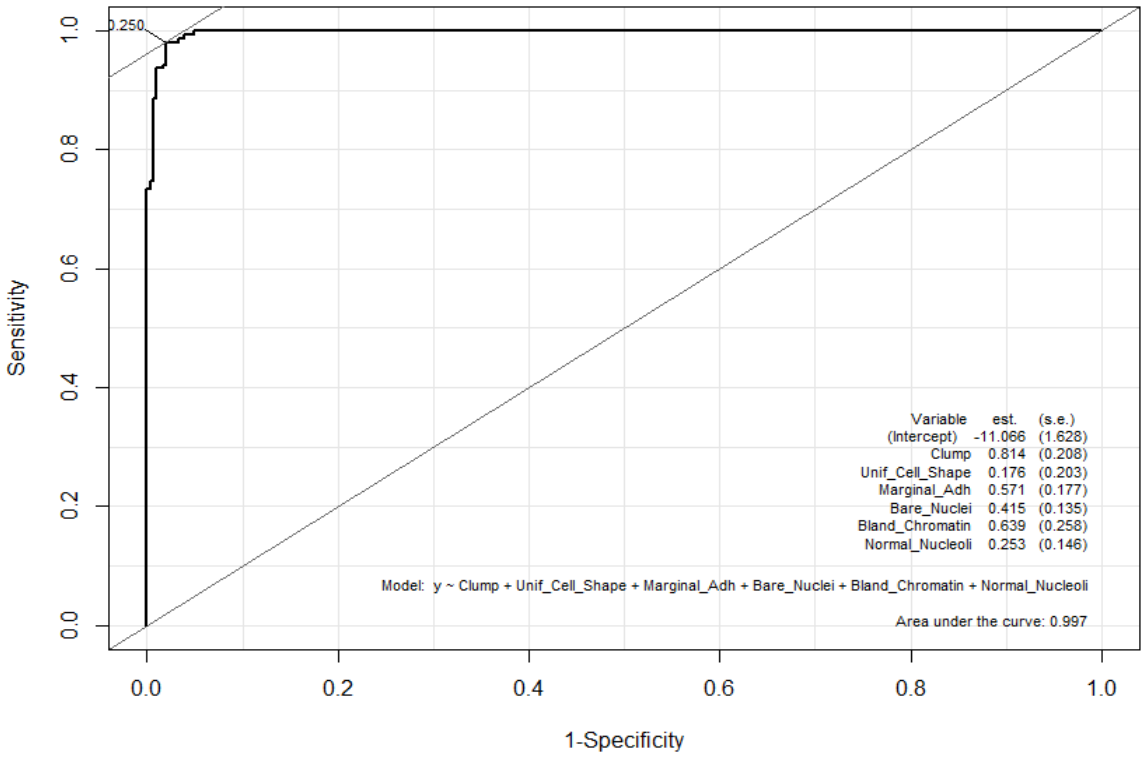


Figura 21: Curva ROC y estimación de parámetros del modelo logístico en B1 para CC

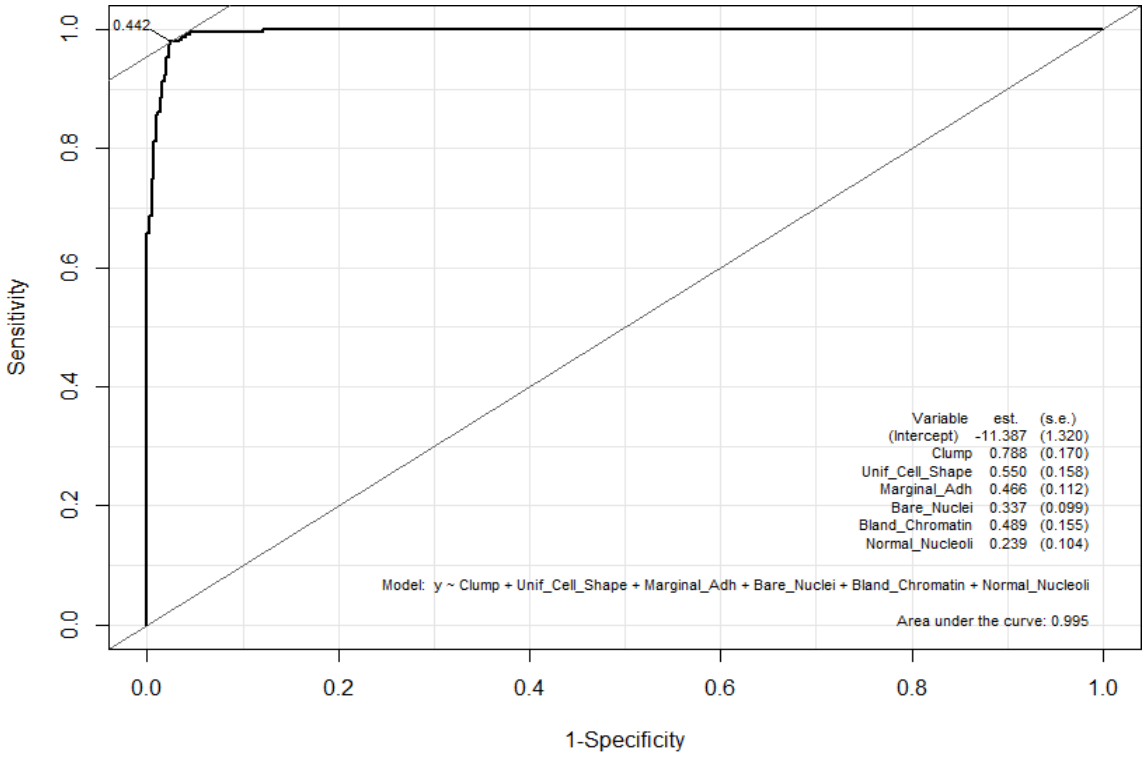


Figura 22: Curva ROC y estimación de parámetros del modelo logístico en B1 para MI

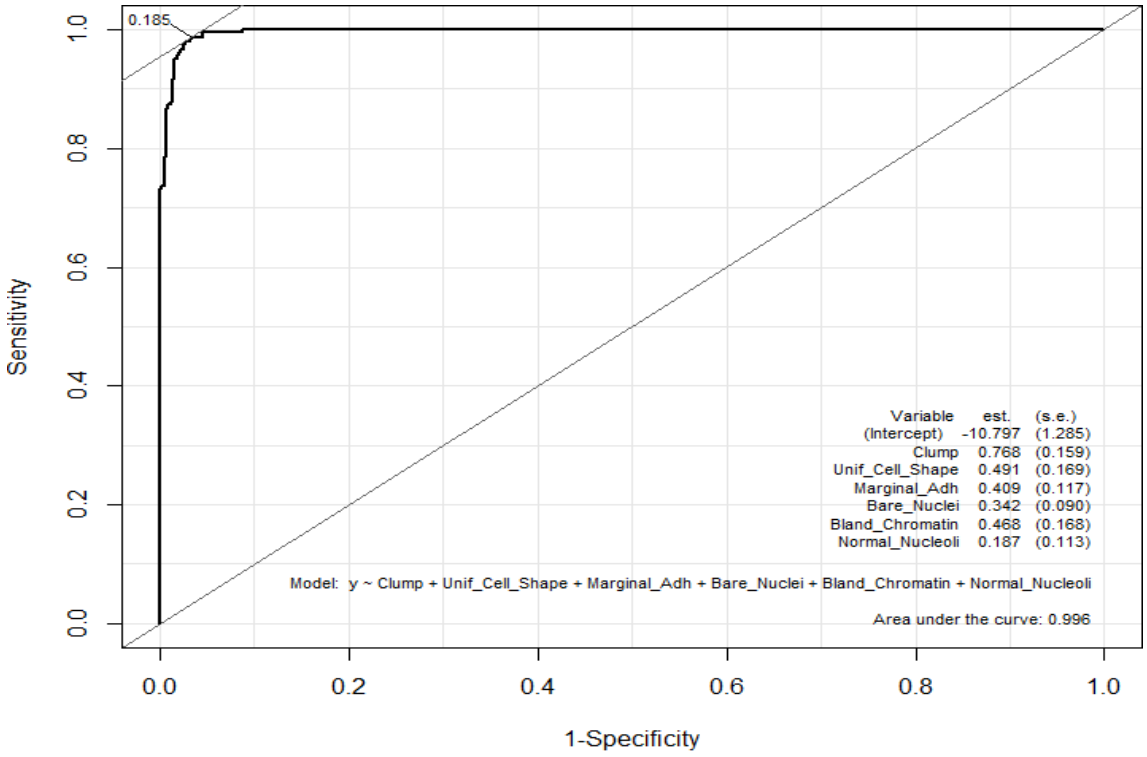


Figura 23: Curva ROC y estimación de parámetros del modelo logístico en B1 para MICE

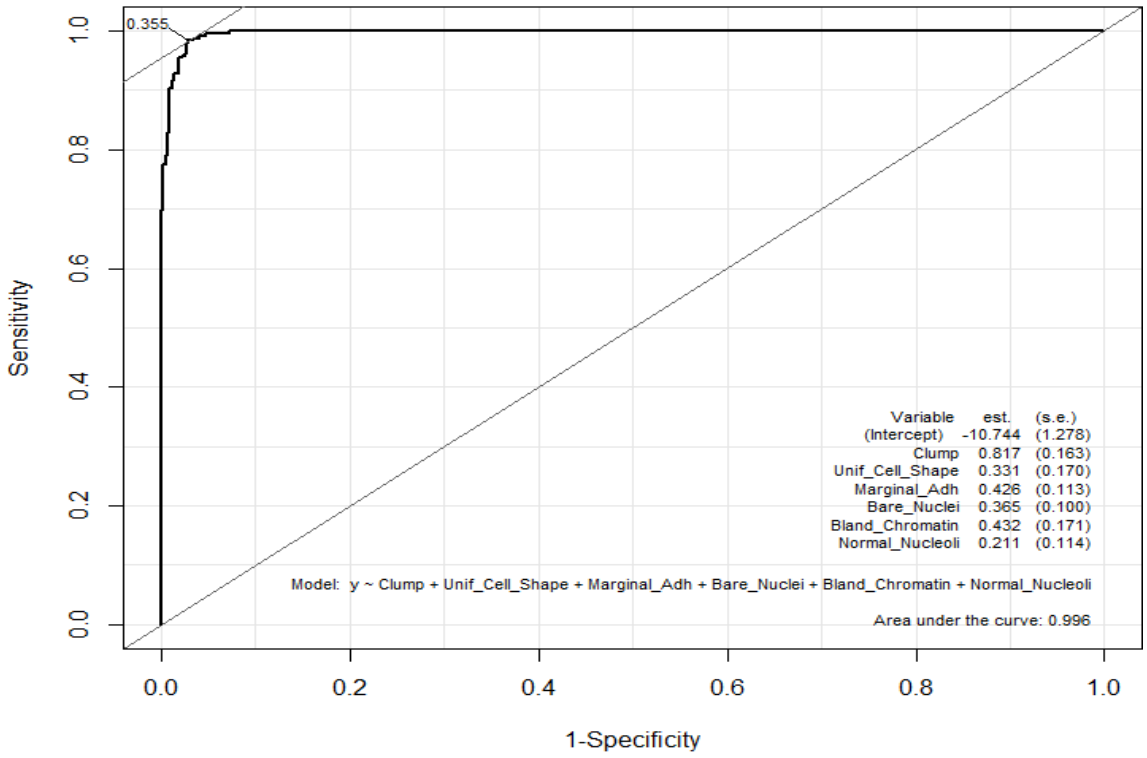


Figura 24: Curva ROC y estimación de parámetros del modelo logístico en B1 para MF

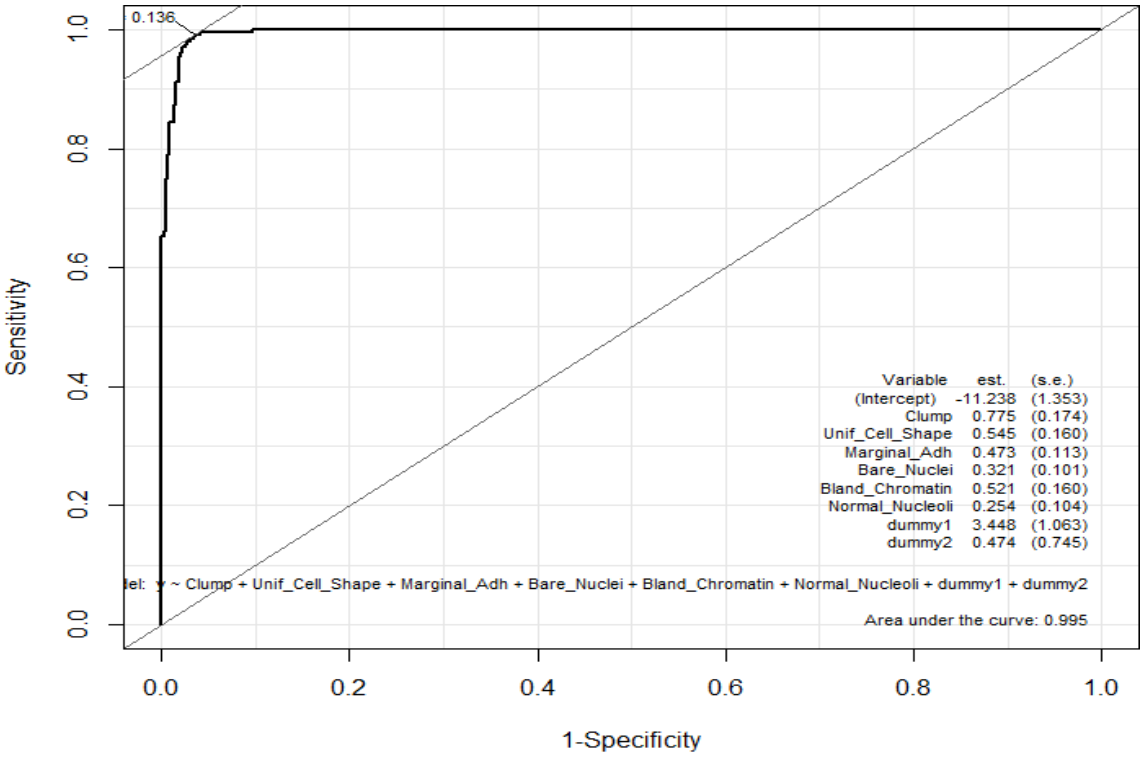


Figura 25: Curva ROC y estimación de parámetros del modelo logístico en B1 para VIP

Todas las curvas ROC son muy similares, como se puede ver en las figuras comprendidas entre la Figura 21 y la Figura 25. Todos los valores AUC y sus intervalos de confianza al 95 % también son extremadamente similares, como se puede observar en la Tabla 3, dándose los menores valores, tanto de manera puntual como en los extremos de los intervalos, en el método de sustitución por la media y en el método con varibles indicadoras de pérdida. Los valores de los demás métodos son muy similares, siendo los valores de los métodos de imputación múltiple casi idénticos.

Como se puede ver en la Tabla 3 o en las Figuras 21-25, los puntos de corte óptimos son bastante dispares y con valores superiores al de B0, que valía 0.106 (Figura 20). Los puntos de corte más similares al de B0, son los obtenidos en los métodos con variables indicadoras de pérdida y MICE. Los demás puntos de corte son mayores, especialmente en el método de sustitución por la media.

Además de las curvas ROC completas, también se muestra en la Figura 26 una ampliación del cuadrante superior izquierdo de la curva ROC con todas las curvas ROC superpuestas. En la Figura 26, el color negro representa el análisis de casos completos, el azul la sustitución por la media, el verde el algoritmo MICE, el rojo Missing Forest y el marrón el método con variables indicadoras de pérdida de datos. Analizando la Figura 26, la curva del análisis de casos completos está en valores más altos de sensibilidad, en la parte de la curva en la que no se superponen las cinco curvas, junto con la curva de Missing Forest. Para valores de sensibilidad próximos o menores a 0.9, las curvas que están en valores más bajos de sensibilidad son las

curvas de la sustitución por la media y la del método con variables indicadoras de pérdida de datos.

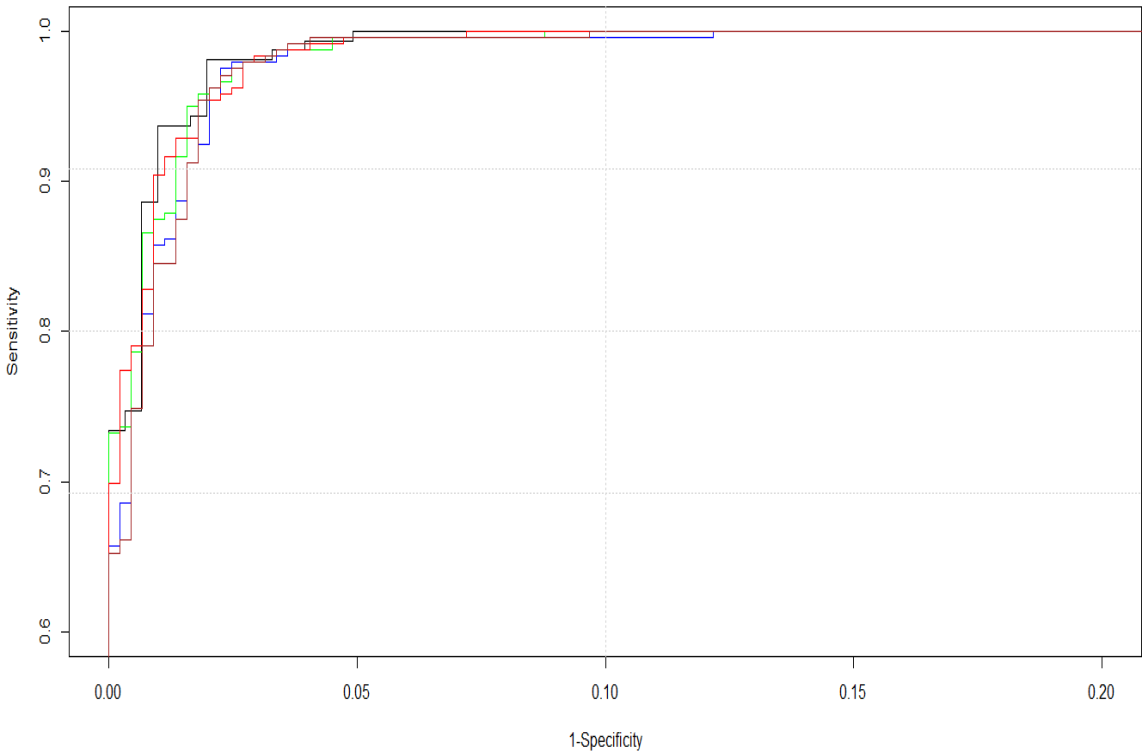


Figura 26: Detalles de las curvas ROC de B1

3.3. Resultados del estudio diagnóstico basado en B2

Se trabaja con la base de datos con datos faltantes según el mecanismo MAR. Se va a proceder a realizar un análisis como el que se realizó en la Sección 3.2, siguiendo la misma estructura en el análisis. En las Tablas 5 y 6, se resumen los resultados obtenidos de las bases de datos generadas a partir de B2. En las Figuras 27-31 se muestran las curvas ROC para los distintos métodos, con las estimaciones de los parámetros del modelo ajustado y sus errores estándares, así como los valores del AUC y del punto de corte óptimo.

Se resumen los resultados relacionados con las curvas ROC y con los modelos de regresión logística de las bases de datos generadas a partir de B2 en la Tabla 5:

Método	VNS	VMS	AIC	PC	AUC	IC95
CC	UCS, MA	C, BN	36.536	0.036	0.999	(0.9979, 1)
MI	NN	C, MA	132.96	0.125	0.995	(0.9913, 0.9983)
MICE	UCS	C, BN	96.371	0.381	0.998	(0.9958, 0.9995)
MF	UCS, MA	C, BN	110.07	0.162	0.997	(0.9939, 0.9993)
VIP	NN, D_BN	C, BN	134.26	0.335	0.995	(0.9914, 0.9983)

Tabla 5: Resultados de los modelos logísticos en B2

La Tabla 6 muestra los p-valores obtenidos para todas las covariables comunes en los cinco métodos en las bases de datos generadas a partir de B2:

Variables predictoras	CC	MI	MICE	MF	VIP
Término independiente	0.00169	<2e-16	7.43e-16	<2e-16	<2e-16
Clump	0.00611	9.89e-05	3.45e-07	1.57e-06	4.79e-05
Unif_Cell_Shape	0.37176	0.000347	0.27616	0.2672	0.001033
Marginal_Adh	0.15652	5.84e-05	0.03810	0.3594	0.020118
Bare_Nuclei	0.01301	0.000648	1.05e-06	3.25e-06	0.000538
Bland_Chromatin	0.02801	0.000724	0.00191	0.0255	0.001292
Normal_Nucleoli	0.02978	0.057500	0.00701	0.0414	0.080928

Tabla 6: p-valores de los modelos logísticos en B2

Hay que añadir que, en el método con variables indicadoras de pérdida de datos, los p-valores para el *dummy* de Clump y para el de Bare_Nuclei son 0.001 y 0.060 respectivamente.

Tal como se puede observar en las Tablas 5 y 6, en todos los métodos la variable Clump es una de las dos más significativas y es la que tiene mayor aportación individual de información en el modelo de regresión logística. En cuanto a la segunda variable con mayor aportación individual, se encuentran diferencias entre los métodos. Esta segunda variable es Bare_Nuclei para todos los métodos salvo la sustitución por la media, donde dicha variable es Marginal_Adh.

En las Tablas 5 y 6, no hay ningún método en el que todas las variables predictoras del modelo ajustado sean significativas. En MICE hay una variable no significativa (Unif_Cell_Shape), así como en el método de sustitución por la media (Normal_Nucleoli), mientras que en los demás métodos hay dos variables no significativas, que son Unif_Cell_Shape y Marginal_Adh para Missing Forest y el análisis de casos completos, y el *dummy* de Bare_Nuclei y Normal_Nucleoli en el método con variables indicadoras de pérdida.

Al comparar los valores AIC de los distintos modelos ajustados, se ve en la Tabla 5 que el valor AIC tiene su menor valor en el análisis de casos completos y su mayor

valor en el método con variables indicadoras de pérdida de datos. El valor AIC del método que sustituye por la media es similar al mayor AIC, mientras que los métodos de imputación múltiple tienen valores intermedios.

Como se puede ver en las Figuras 27-31, el análisis de casos completos es el método en el que los errores estándar son mayores para cada una de las covariables de los modelos ajustados de regresión logística. En el análisis de casos completos, los valores de los coeficientes son muy diferentes respecto a los valores de los demás métodos. Para dicho método, el término independiente es mucho mayor (en valor absoluto), así como los coeficientes de todas las covariables salvo Unif_Cell_Shape, cuyo coeficiente tiene un valor negativo, lo que no tiene sentido. El valor del coeficiente de Clump en los métodos de imputación múltiple es mayor que el de los demás métodos (sin contar el análisis de casos completos), e incluso el coeficiente del método de sustitución por la media es menor respecto al coeficiente en B0, siendo el coeficiente del método con variable indicadoras de pérdida muy similar al de B0; el error estándar es menor respecto al error correspondiente en B0. Respecto al valor del coeficiente de Unif_Cell_Shape, los coeficientes de los métodos de imputación múltiple son muy pequeños y los de los métodos de imputación simple son bastante grandes en comparación con el coeficiente correspondiente en B0; en cambio, los errores estándar son menores en los métodos de imputación simple. Para Marginal_Adh, ocurre algo similar a lo que sucede con Unif_Cell_Shape, aunque los métodos con las variables indicadoras de pérdida de datos y MICE tienen valores del coeficiente bastante similares entre sí y al coeficiente correspondiente en el modelo de B0 (valor exacto con el método con las variables indicadoras de pérdida de datos); el error estándar es mayor en todos los casos respecto al de B0. Con Bare_Nuclei sucede lo contrario, esto es, que los coeficientes en los métodos de imputación múltiple son mayores que los de los métodos de imputación simple; los errores estándar son muy similares y mayores que el del modelo de B0. Con Bland_Chromatin, los coeficientes tienen valores similares excepto en Missing Forest, que tiene un valor menor; el error estándar es mayor en los métodos de imputación múltiple. Con Normal_Nucleoli, los métodos de imputación múltiple tienen coeficientes mayores que los métodos de imputación simple; los errores estándar son bastante similares (mayores en los métodos de imputación múltiple).

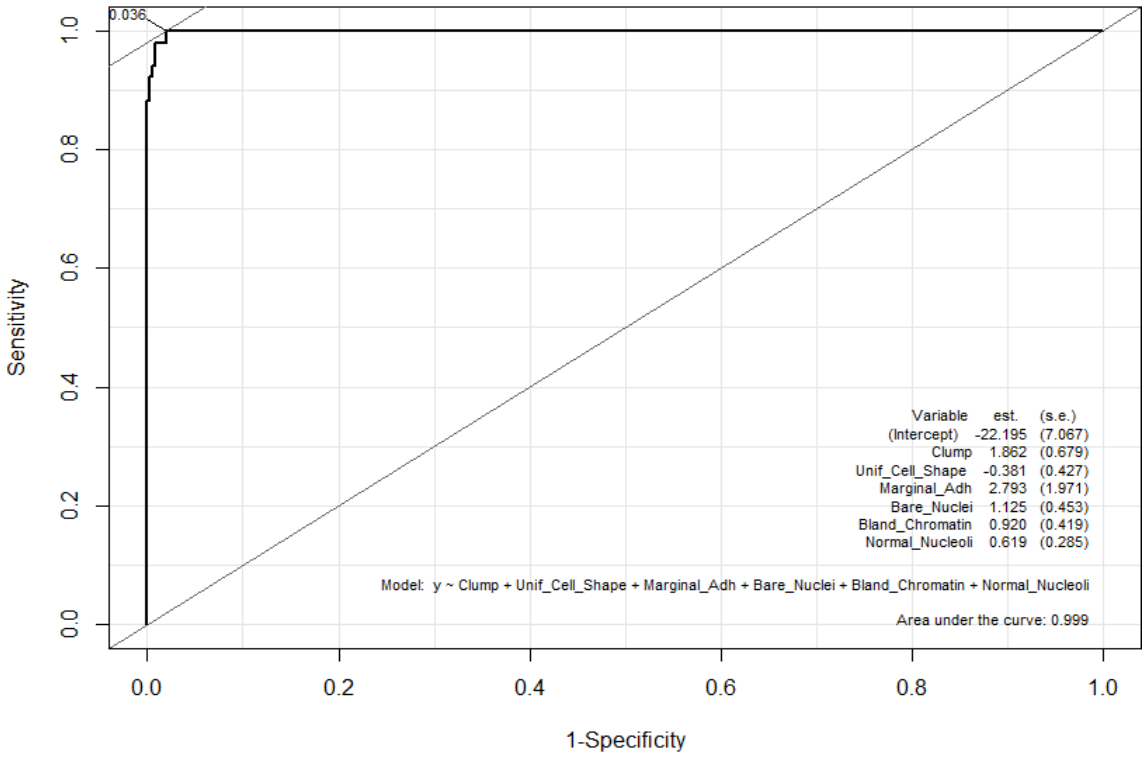


Figura 27: Curva ROC y estimación de parámetros del modelo logístico en B2 para CC

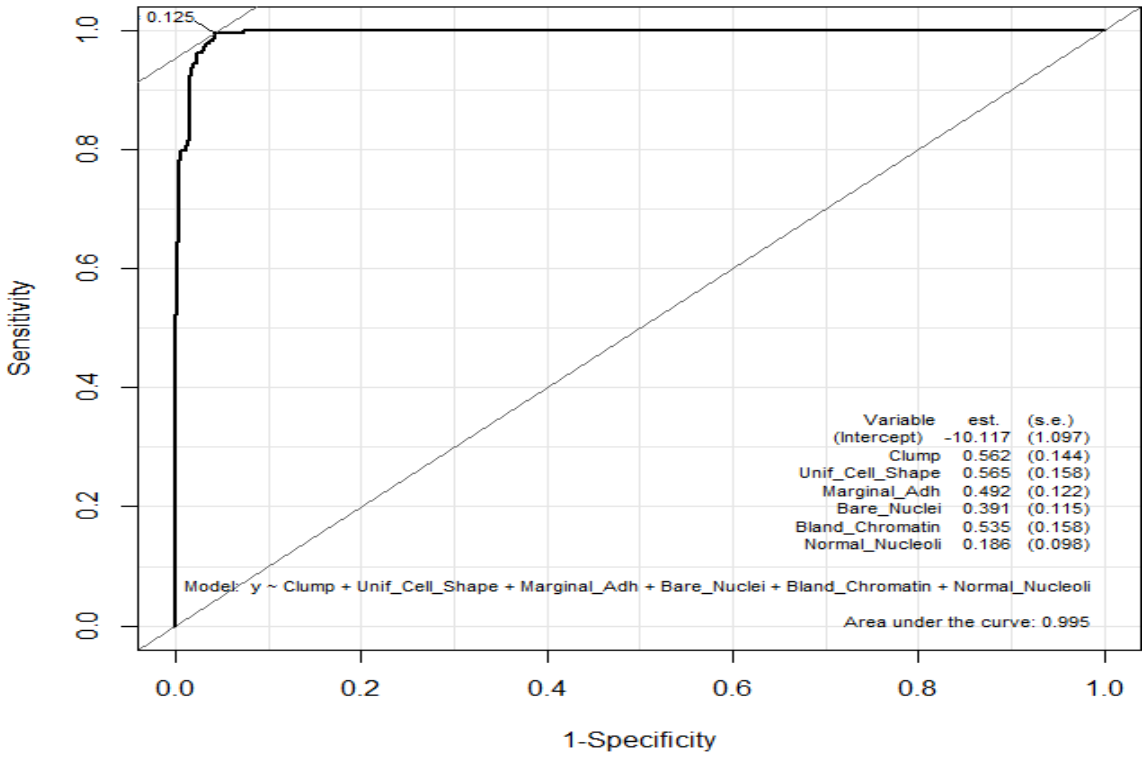


Figura 28: Curva ROC y estimación de parámetros del modelo logístico en B2 para MI

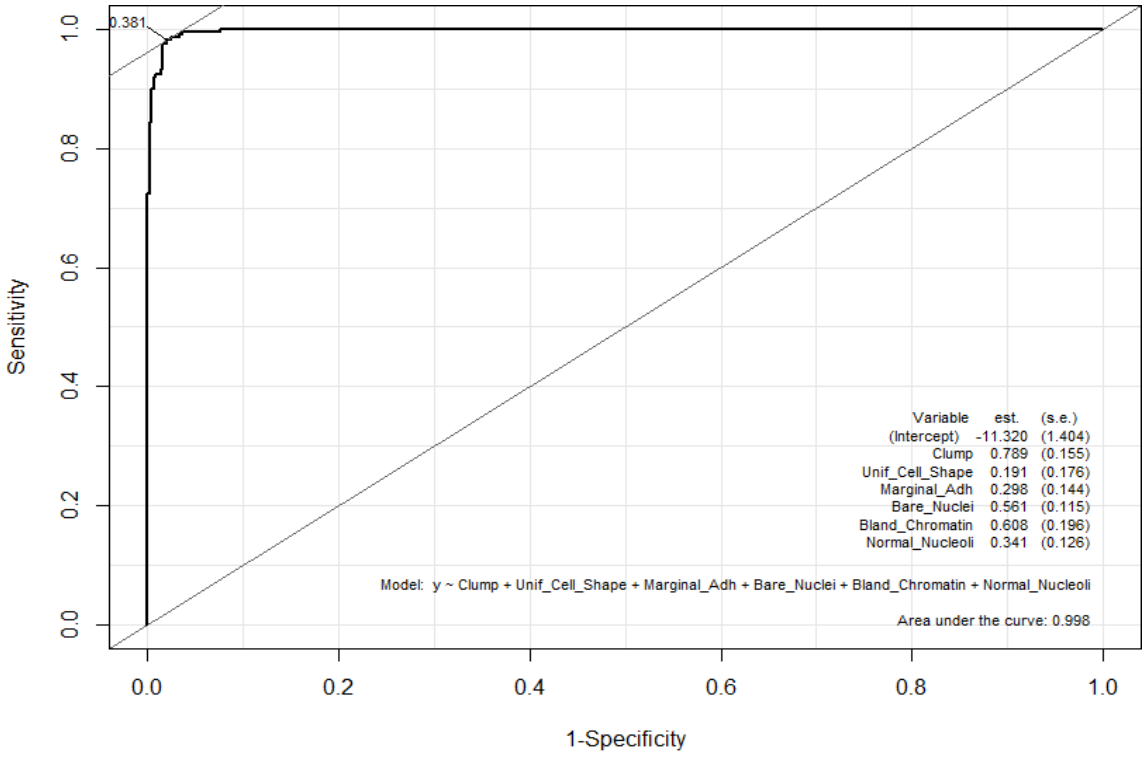


Figura 29: Curva ROC y estimación de parámetros del modelo logístico en B2 para MICE

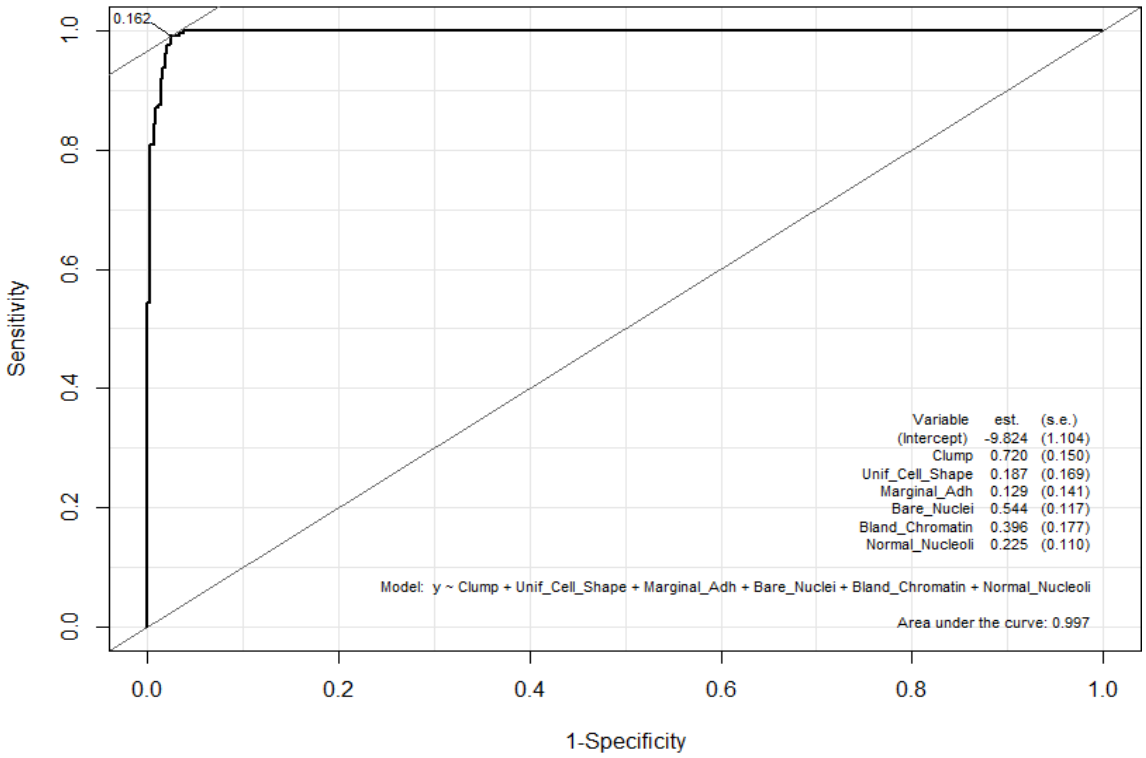


Figura 30: Curva ROC y estimación de parámetros del modelo logístico en B2 para MF

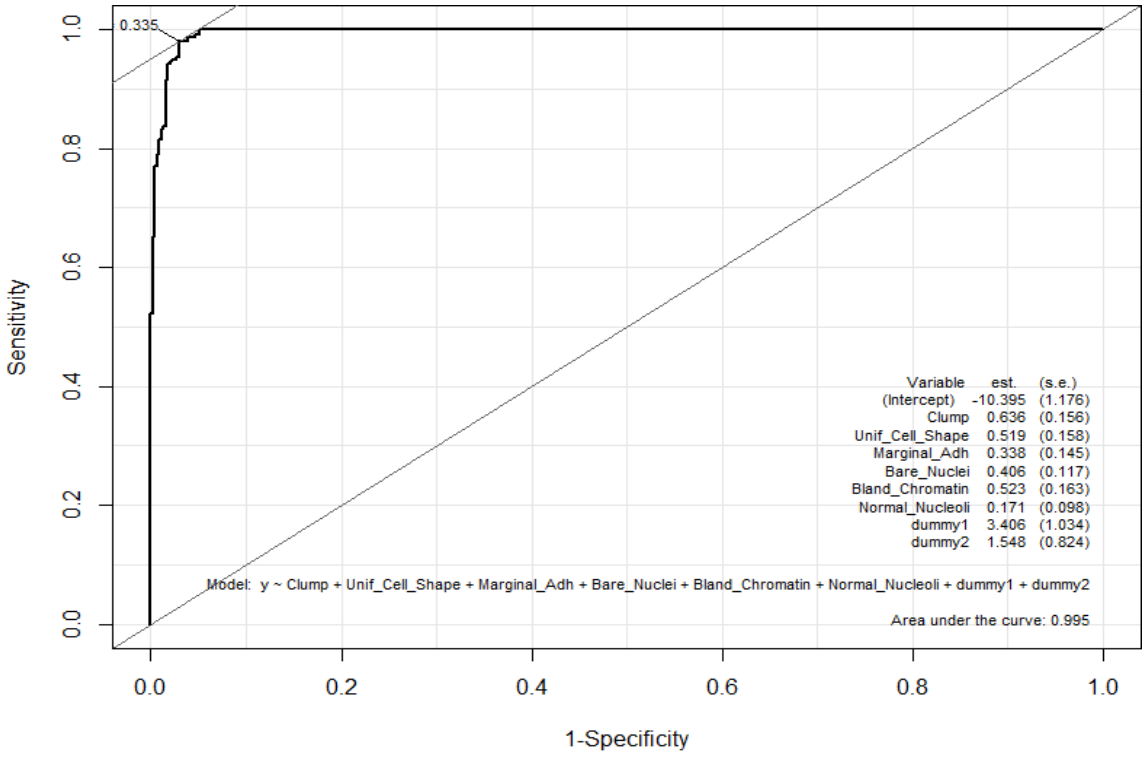


Figura 31: Curva ROC y estimación de parámetros del modelo logístico en B2 para VIP

Todas las curvas ROC son muy similares, como se puede ver en las figuras comprendidas entre la Figura 27 y la Figura 31. Todos los valores AUC y sus intervalos de confianza al 95 % también son extremadamente similares, como se puede observar en la Tabla 5, dándose los menores valores, tanto de manera puntual como en los extremos de los intervalos, en el método de sustitución por la media y en el método con varibles indicadoras de pérdida, con valores casi idénticos. Los valores mayores se dan en el análisis de casos completos, llegando el extremo superior a 1. Los métodos de imputación múltiple tienen valores mayores que los de imputación simple.

Como se puede ver en la Tabla 5 o en las Figuras 27-31, los puntos de corte óptimos son bastante dispares. El punto de corte del análisis de casos completos es extremadamente pequeño, mientras que los puntos de corte de Missing Forest y el método de sustitución por la media son bastante similares al obtenido con B0, y los puntos de corte de los otros dos métodos son bastante mayores.

En la Figura 32 se muestra una ampliación de las curvas ROC para los distintos métodos, tal y como se hizo en la Figura 26, y con las mismas indicaciones de colores que se hicieron en dicha figura. Analizando la Figura 32, la curva del análisis de casos completos está en valores más altos de sensibilidad, en la parte de la curva en la que no se superponen las cinco curvas. En valores de 1-Especificidad menores a los valores en los que se superponen las cinco curvas, las curvas que están en valores más bajos de sensibilidad son las curvas de la sustitución por la media y la del método con variables indicadoras de pérdida de datos.

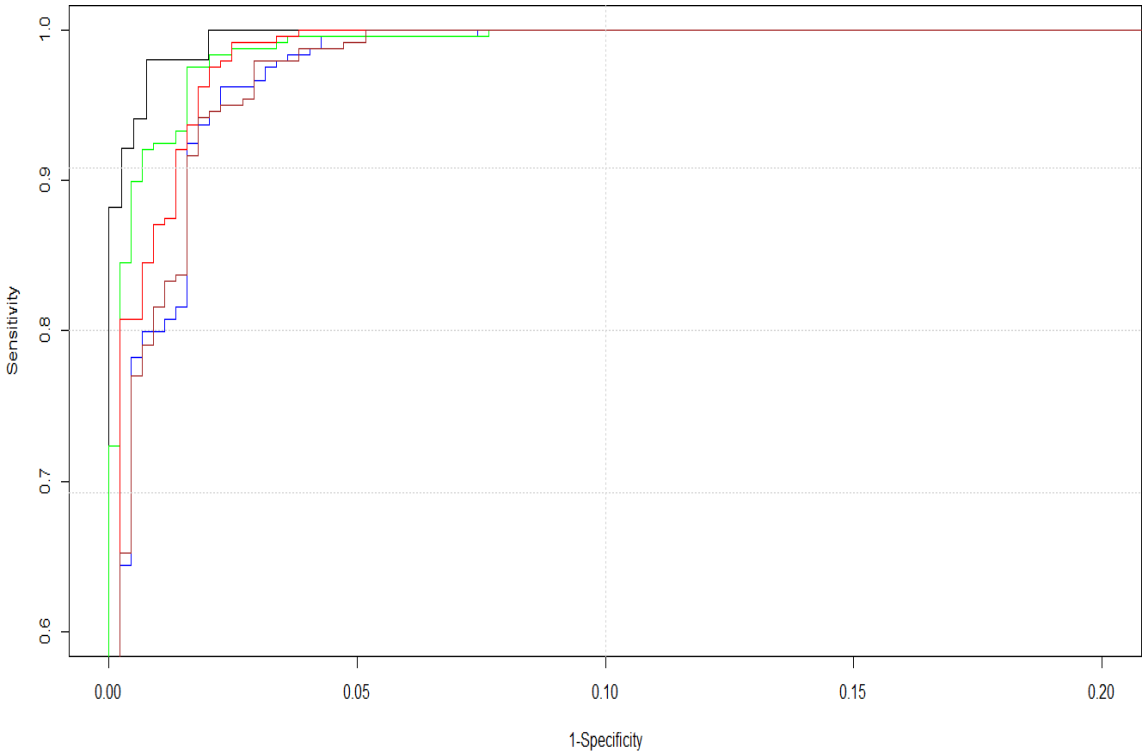


Figura 32: Detalles de las curvas ROC de B2

3.4. Resultados del estudio diagnóstico basado en B3

Se trabaja con la base de datos con datos faltantes según el mecanismo MNAR. Se va a proceder a realizar un análisis como el que se realizó en las Secciones 3.2 y 3.3, siguiendo la misma estructura en el análisis. En las Tablas 7 y 8, se resumen los resultados obtenidos de las bases de datos generadas a partir de B3. En las Figuras 33-37 se muestran las curvas ROC para los distintos métodos, con las estimaciones de los parámetros del modelo ajustado y sus errores estándares, así como los valores del AUC y del punto de corte óptimo.

Se resumen los resultados relacionados con las curvas ROC y con los modelos de regresión logística de las bases de datos generadas a partir de B3 en la Tabla 7:

Método	VNS	VMS	AIC	PC	AUC	IC95
CC	UCS, NN	C, BN	75.874	0.069	0.996	(0.9934, 0.9994)
MI	NN	UCS, MA	128.90	0.211	0.995	(0.9911, 0.9983)
MICE	No	C, MA	110.79	0.190	0.997	(0.9944, 0.9990)
MF	NN	C, MA	113.70	0.320	0.996	(0.9924, 0.9990)
VIP	No	C, BN	120.06	0.369	0.996	(0.9931, 0.9990)

Tabla 7: Resultados de los modelos logísticos en B3

La Tabla 8 muestra los p-valores obtenidos para todas las covariables comunes en los cinco métodos en las bases de datos generadas a partir de B3:

Variables predictoras	CC	MI	MICE	MF	VIP
Término independiente	2.68e-12	<2e-16	5.68e-16	8.20e-16	7.18e-16
Clump	3.87e-05	4.43e-05	4.39e-06	1.72e-05	1.61e-05
Unif_Cell_Shape	0.23223	1.47e-05	0.001051	0.00465	0.014055
Marginal_Adh	0.00144	5.48e-06	2.37e-06	1.73e-05	0.000311
Bare_Nuclei	0.00028	0.003351	0.000197	9.47e-05	0.000196
Bland_Chromatin	0.01078	0.000333	0.005753	0.00407	0.006336
Normal_Nucleoli	0.05055	0.063392	0.038999	0.07453	0.047418

Tabla 8: p-valores de los modelos logísticos en B3

Hay que añadir que, en el método con variables indicadoras de pérdida de datos, los p-valores para el *dummy* de Clump y para el de Bare_Nuclei son 0.0006 y 0.0004 respectivamente.

Tal como se puede observar en las Tablas 7 y 8, en todos los métodos la variable Clump es una de las dos más significativas en todos los métodos excepto en el de la sustitución por la media, método en el que las dos variables más significativas son Unif_Cell_Shape y Marginal_Adh. En los métodos de imputación múltiple, la otra variable más significativa es Marginal_Adh y en el método con variables indicadoras de pérdida y el análisis de casos completos sí es Bare_Nuclei.

En las Tablas 7 y 8, los métodos en los que todas las variables predictoras del modelo ajustado son significativas son el método con variables indicadoras de pérdida de datos y MICE. En el análisis de casos completos hay dos variables no significativas y en los métodos de sustitución por la media y Missing Forest, una variable. En los tres métodos recientemente comentados, la covariable Normal_Nucleoli no es significativa, y en el análisis de casos completos, además, la variable que no es significativa es Unif_Cell_Shape.

Al comparar los valores AIC de los distintos modelos ajustados, se ve en la Tabla 7 que el valor AIC tiene su menor valor en el análisis de casos completos y su mayor valor en el método de sustitución por la media. El valor AIC de los métodos de imputación simple es mayor que el AIC de los métodos de imputación múltiple, pero los valores son muy similares.

Como se puede ver en las Figuras 33-37, el análisis de casos completos es el método en el que los errores estándar son mayores para cada una de las covariables de los modelos ajustados de regresión logística. El valor del coeficiente de Clump es mayor que el coeficiente del modelo obtenido con B0 en todos los casos y en los métodos de imputación simple es menor que el de los demás métodos, aunque el coeficiente

del método con variables indicadoras de pérdida es muy similar al de MICE y el coeficiente con mayor valor es el del análisis de casos completos; el error estándar es mayor en todos los casos respecto al error correspondiente en B0. Respecto al valor del coeficiente de Unif_Cell_Shape, en el análisis de casos completos es claramente menor mientras que en el método de sustitución por la media es claramente mayor, estando también los coeficientes del otro método de imputación simple y de los métodos de imputación múltiple por encima del coeficiente obtenido en el modelo con B0; todos los errores estándar son mayores que el error correspondiente obtenido con B0. Para Marginal_Adh, no hay grandes diferencias entre los valores de los coeficientes, aunque todos son mayores respecto al coeficiente obtenido con B0, siendo el coeficiente de mayor valor el del método MICE y el de menor valor el del método con las variables indicadoras de pérdida; el error estándar es mayor en todos los casos respecto al de B0. Con Bare_Nuclei el coeficiente del método de sustitución por la media tiene un valor similar al coeficiente equivalente con B0, pero los demás métodos tienen coeficientes con valores claramente superiores, especialmente con el análisis de casos completos y Missing Forest; los errores estándar son muy similares y mayores que el del modelo de B0. Con Bland_Chromatin, los coeficientes tienen valores similares y no lejanos al valor del coeficiente que se obtuvo con B0, siendo el coeficiente de mayor valor el obtenido con el análisis de casos completos y el de valor menor y casi idéntico al de B0, el obtenido con el método con variables indicadoras de pérdida; el error estándar es mayor en los métodos de imputación múltiple (después del error del análisis de casos completos). Con Normal_Nucleoli, sucede algo similar a lo que ocurre con Bland_Chromatin, aunque el valor de los coeficientes es aún más cercano respecto al obtenido con B0, y siendo el coeficiente de mayor valor el obtenido con el análisis de casos completos y el de menor valor el obtenido el método de sustitución por la media; los errores estándar son bastante similares (mayor en el análisis de casos completos).

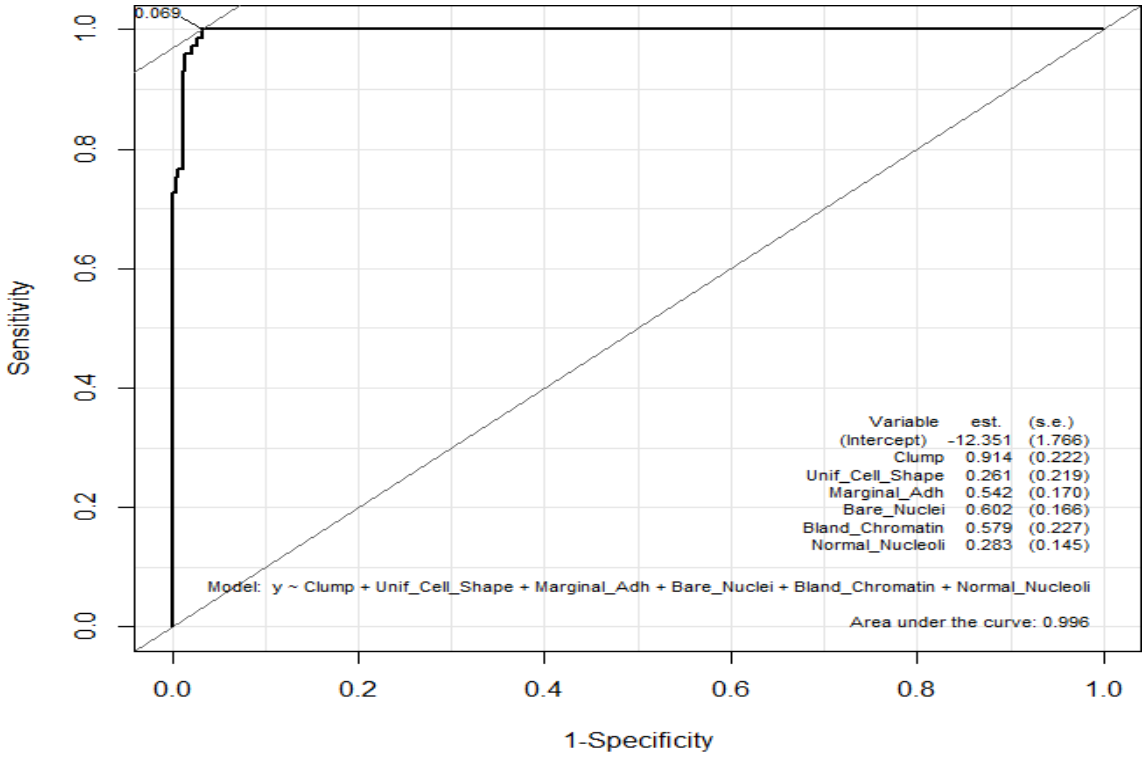


Figura 33: Curva ROC y estimación de parámetros del modelo logístico en B3 para CC

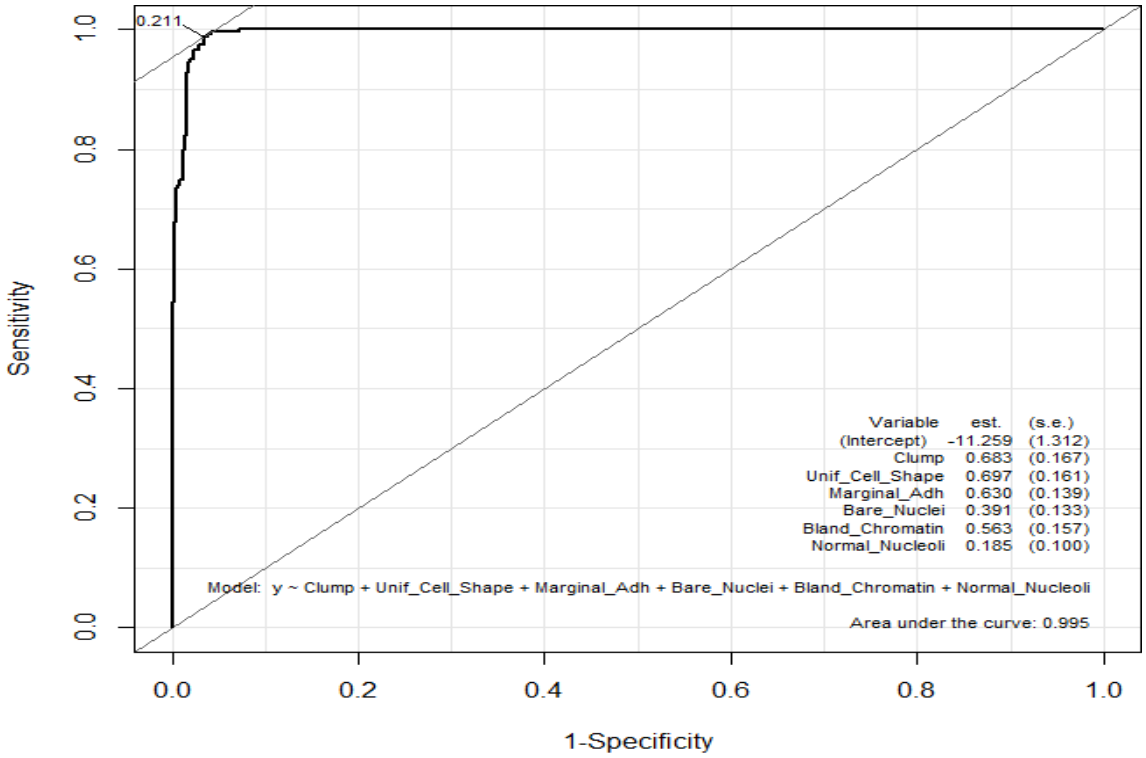


Figura 34: Curva ROC y estimación de parámetros del modelo logístico en B3 para MI

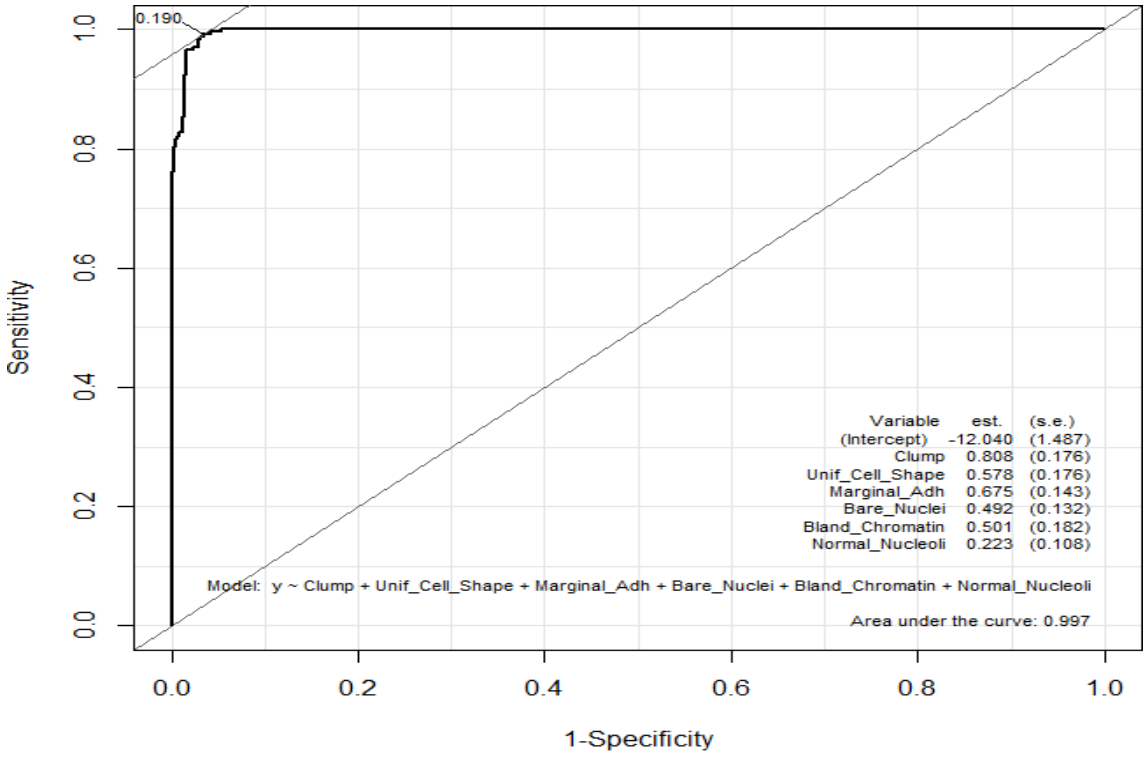


Figura 35: Curva ROC y estimación de parámetros del modelo logístico en B3 para MICE

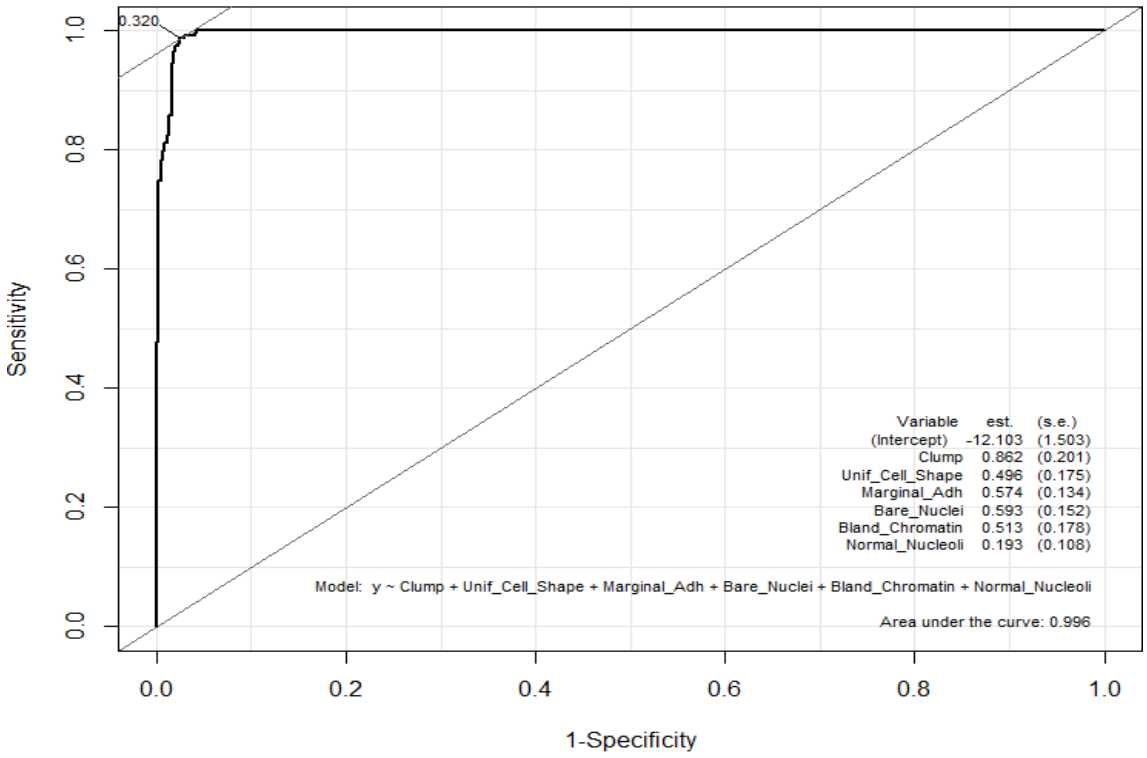


Figura 36: Curva ROC y estimación de parámetros del modelo logístico en B3 para MF

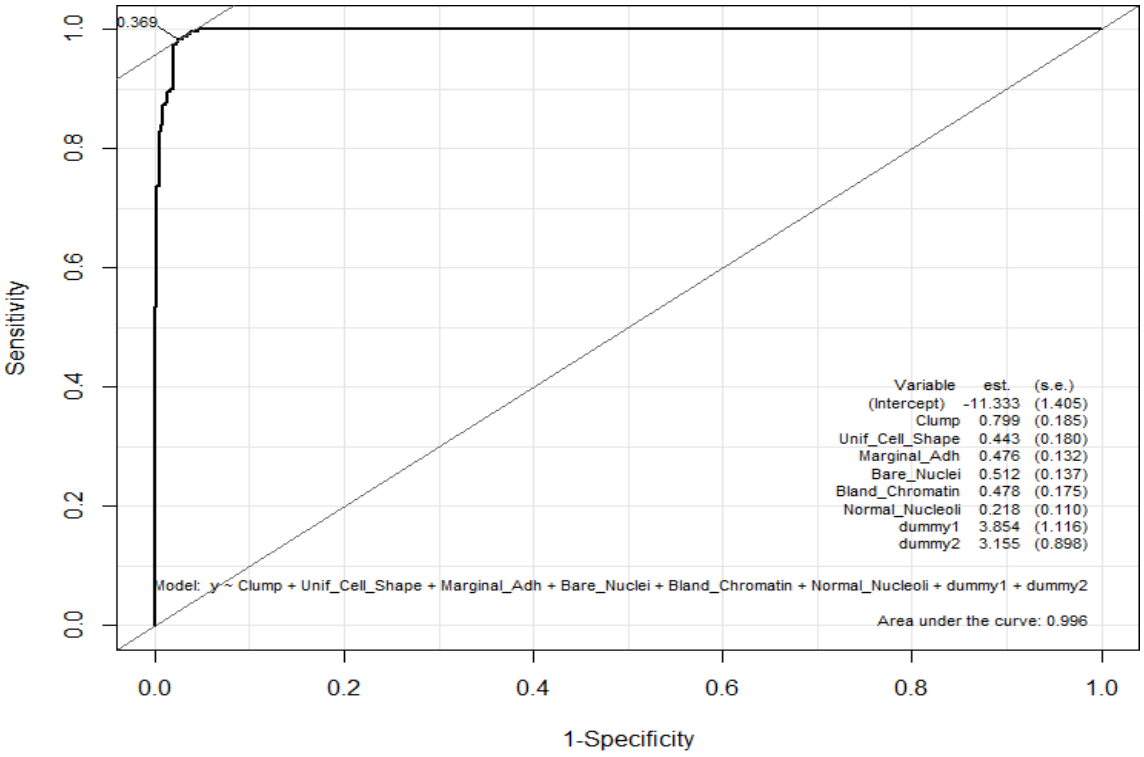


Figura 37: Curva ROC y estimación de parámetros del modelo logístico en B3 para VIP

Todas las curvas ROC son muy similares, como se puede ver en las figuras comprendidas entre la Figura 33 y la Figura 37. Todos los valores AUC y sus intervalos de confianza al 95 % también son similares, dándose los menores valores, tanto de manera puntual como en los extremos de los intervalos, en el método de sustitución por la media. Los valores mayores se dan en el análisis de casos completos en el límite superior del intervalo de confianza y en MICE para el límite inferior del intervalo de confianza y el valor puntual.

Como se puede ver en la Tabla 7 o en las Figuras 33-37, los puntos de corte óptimos son bastante dispares. El punto de corte del análisis de casos completos es extremadamente pequeño. Los puntos de corte de Missing Forest y del método con variables indicadoras de pérdida son los más altos y los puntos de corte de los métodos de sustición por la media y MICE tienen valor intermedios, aunque más cercanos a los valores más altos.

En la Figura 38 se muestra una ampliación de las curvas ROC para los distintos métodos, tal y como se hizo en la Figura 26 y en la Figura 32, y con las mismas indicaciones de colores que en dichas figuras. Analizando la Figura 38, en valores de 1-Especificidad menores a los valores en los que se superponen las cinco curvas, la curva que está en valores más bajos de sensibilidad es la curva de la sustitución por la media, mientras que no hay una curva predominante que se encuentre por encima.

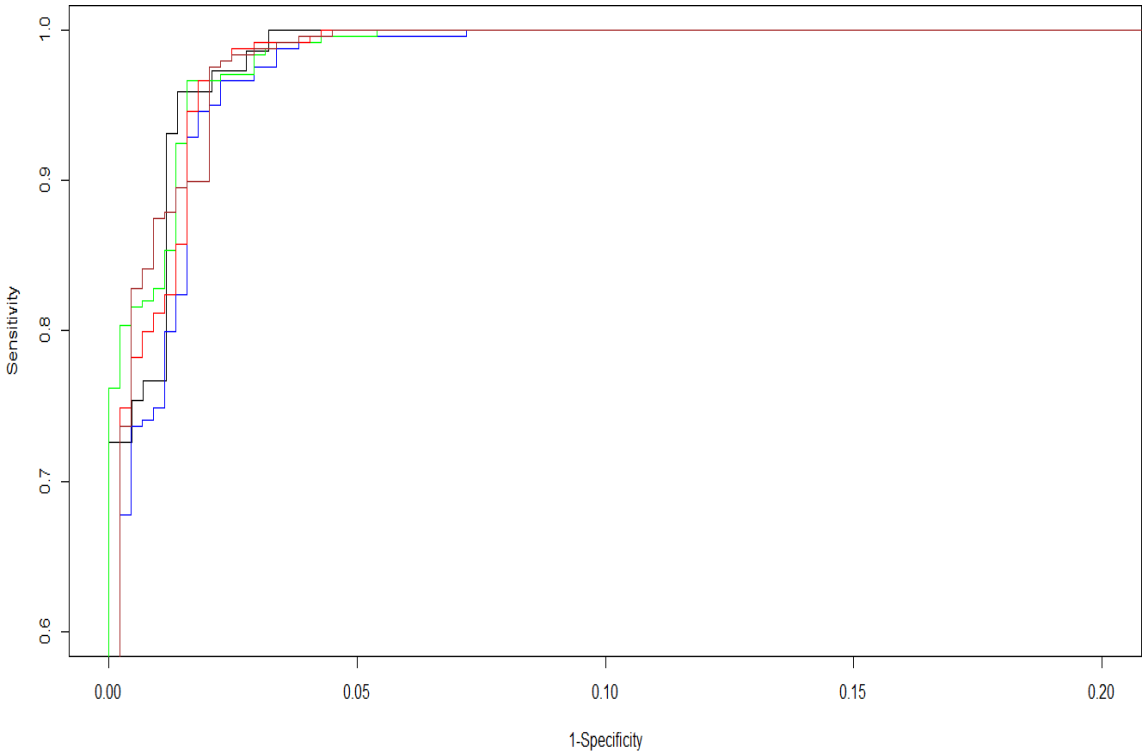


Figura 38: Detalles de las curvas ROC de B3

3.5. Análisis comparativo de los resultados obtenidos

Se va a realizar un análisis basado en la comparación de los resultados descritos para las diferentes bases de datos.

Todas las curvas ROC son muy similares, lo que se refleja en que el valor AUC es muy similar en todos los casos y las diferencias en los intervalos de confianza al 95 % son de milésimas entre métodos y entre las diferentes bases de datos, superponiéndose todos los intervalos de confianza. Esto muestra que el modelo es muy sólido y que la base de datos da unos resultados que se acercan a una situación ideal.

Respecto al punto óptimo de corte, los valores varían de manera considerable entre los métodos de tratamiento de datos faltantes para cada base de datos y para cada método de una base de datos a otra. Con los métodos de casos completos y MICE, hay dos casos en los que el punto de corte se sitúa a menos de una décima respecto al punto de corte en B0. Aunque no se aprecia una tendencia clara en los valores de los puntos óptimos de corte, se puede intuir que los métodos Missing Forest y de las variables indicadoras de pérdida tienden a dar puntos de corte más altos, mientras que el método de análisis de casos completos sería el que da puntos de corte más bajos en estos casos.

Respecto a los valores AIC, se puede observar que los métodos de imputación

simple son los que tienen valores AIC más altos para las diferentes bases de datos. Respecto a los valores AIC en el método con variables indicadoras de pérdida, hay que considerar que hay dos variables más que en el resto de modelos, lo que influye en el aumento del AIC. En todas las bases de datos, el método con menor AIC es el de casos completos, teniendo un valor especialmente bajo en B2, lo que puede deberse a la eliminación de pacientes que son outliers. En los dos métodos de imputación múltiple, los valores AIC son similares en todos los casos salvo en MICE para B2, donde el valor es ligeramente menor. En el método de la sustitución por la media, todos los valores AIC son muy similares, dándose el mayor en B2 y el menor en B1. Esto muestra que, a partir de un criterio basado en el valor AIC, los métodos de tratamiento de datos faltantes de imputación múltiple dan mejores resultados que los métodos de imputación múltiple.

Respecto a las variables más significativas, en todos los casos salvo para el método de la sustitución por la media en B3, Clump es una de esas variables. No hay ningún método en el que en las tres bases de datos tengan como las dos variables más significativas las mismas que en B0 (Clump y Bare_Nuclei), aunque con MICE y el método con variables indicadoras de pérdida se da en dos casos, y con el análisis de casos completos y Missing Forest en uno. Con el método de la sustitución por la media, no hay coincidencia de las dos variables más significativas en ningún caso. Cuando una de las dos variables más significativas no es Bare_Nuclei, dicha variable es Marginal_Adh, aunque cabe recordar que en el método de sustitución por la media en B3, ninguna de las variables más significativas es Clump o Bare_Nuclei. Entonces, B3 sería la base de datos con menor coincidencia con B0 en cuanto a las variables significativas y B2 la que tiene mayor coincidencia.

Respecto a las variables que no resultan significativas en los diferentes modelos, en casi todos los casos se da alguna variable no significativa. Las excepciones son MICE en B3, el método con variables indicadoras de pérdida en B3 y el método de la sustitución por la media en B1. Con el método de casos completos, hay dos variables no significativas en cada base de datos, que son Unif_Cell_Shape (en las tres bases de datos), Normal_Nucleoli (B1 y B3) y Marginal_Adh (B2). En casi todos los casos en los que una variable no es significativa, dicha variable es Unif_Cell_Shape o Normal_Nucleoli (las variables con mayor p-valor en el modelo obtenido con B0), salvo en B2 para Missing Forest y el análisis de casos completos, y el método con variables indicadoras de pérdida, donde el *dummy* de Bare_Nuclei no es significativo ni en B1 ni en B2. Entonces, la tendencia es que hay más variables no significativas con B2 y menos con B3.

Se procede a realizar el análisis comparativo con la estimación de parámetros. Cuando se compare el valor de los parámetros, se realiza en valor absoluto salvo que se especifique lo contrario. Con el método de casos completos, los errores estándar son

los más altos en comparación con los demás métodos para todas las bases de datos, lo que era previsible al eliminar individuos. Con el método de casos completos, además, se puede apreciar que los parámetros del modelo en B2 tienen un valor mucho mayor al de todos los demás casos, excepto un parámetro (el de Unif_Cell_Shape), que tiene valor negativo, lo que con estos datos no tiene sentido. También se vio que la tendencia es que el análisis de casos completos es el método en el que la diferencia del valor de los coeficientes respecto al valor de los coeficientes en el modelo obtenido con B0 sea mayor.

Con el método de la sustitución por la media, en B1 la diferencia de valores entre el parámetro de Clump (valor que debería ser el más alto a partir de los resultados de B0) y los demás valores sí se mantiene en B1, pero en B2 y B3 deja de ser el parámetro con mayor valor para que pase a ser Unif_Cell_Shape la variable con parámetros estimados más altos, cuando en B0 su valor era claramente menor al parámetro de Clump. En B2 y B3, algunos parámetros (variables Unif_Cell_Shape, Marginal_Adh y Bland_Chromatin) tienden a estar claramente sobreestimados en comparación con los resultados obtenidos en B0, mientras que Normal_Nucleoli está subestimado en dichas bases de datos y los parámetros estimados para Bare_Nuclei tienen valores similares respecto a B0, mientras que con Clump no hay una tendencia clara. Respecto a los errores estándar, no hay una clara sobreestimación o subestimación de los mismos (se dan ambos casos), como ocurría con el método de casos completos.

Con MICE, el parámetro de Clump está sobreestimado en todos los casos. Respecto a los demás parámetros, en algunas bases de datos están sobreestimados y en otras subestimados o con valores similares a los de B0, sin haber un patrón claro. Respecto a Bare_Nuclei, el parámetro estimado tiene un valor similar (ligeramente inferior) al de B0 en el resultado obtenido en B1, mientras que en B2 y B3 está sobreestimado. Los errores estándar son similares a los originales en los resultados obtenidos con B1 (diferencia de milésimas salvo para Clump) y algo superiores con B2 y B3 (salvo en Normal_Nucleoli con B3, donde el error estándar es muy similar al de B0, siendo una milésima menor).

Con Missing Forest, los resultados obtenidos son similares respecto a los obtenidos con MICE, especialmente los errores estándar para todas las bases de datos. La estimación de parámetros es muy parecida a la de MICE en B1 (aunque la estimación de Unif_Cell_Shape es claramente mayor en MICE), habiendo alguna diferencia más en las otras bases de datos. En B2, todos los parámetros estimados tienen menor valor respecto a MICE. En B3, los parámetros con pérdida de datos tienen un valor mayor que los de MICE (mayor sobreestimación) mientras que los demás, salvo Bland_Chromatin (con valor muy similar al de MICE), tienen un valor menor que los de MICE.

Con el método con variables indicadoras de pérdida, lo que se aprecia con claridad es que el valor del *dummy* de Clump se mantiene más o menos estable para todos los casos, subiendo un poco el valor en B3, mientras que el de Bare_Nuclei sube claramente su valor de B1 a B2 y de B2 a B3. Los errores estándar son similares a los obtenidos con el método de sustitución por la media en B1 y B2, mientras que en B3 los errores de los métodos son más similares a los errores obtenidos con los métodos de imputación múltiple. No se aprecia un patrón de estimación de parámetros claramente diferente al de los métodos de imputación múltiple, estando los valores estimados del método con variables indicadoras de pérdida entre medias de los valores obtenidos con los métodos de imputación múltiple, notoriamente por encima o notoriamente por debajo, dándose estos tres casos en cada base de datos simultáneamente excepto en B3, donde no hay ningún parámetro estimado que esté por encima de los parámetros estimados con los dos métodos de imputación múltiple al mismo tiempo.

También cabe comentar que los errores estándar tienden a ser más altos (sobre-estimados) en la base de datos con datos faltantes según patron MNAR (B3).

4. Conclusiones

En este trabajo se ha obtenido un modelo de regresión logística a partir de variables obtenidas mediante la extracción de características de las células tumorales para predecir si los tumores mamarios de mujeres son malignos o benignos. Se han generado bases de datos según los mecanismos de pérdida de datos (MCAR, MAR y MNAR), se han aplicado diversos métodos de tratamiento de datos faltantes y se han comparado los resultados de los modelos logísticos obtenidos a partir de las bases de datos resultantes de aplicar los mecanismos de pérdida y los métodos de tratamiento de datos faltantes. Para la comparación de los resultados, se han utilizado curvas ROC y valores cuantitativos asociados a las curvas (AUC y su intervalo de confianza al 95 %) y la estimación de los parámetros con los errores estándar de los modelos logísticos y su valor AIC, observando las variables más significativas y las no significativas en los modelos obtenidos.

Observando las curvas ROC y los valores AUC con sus intervalos de confianza al 95 %, no se observan grandes diferencias entre los diversos resultados obtenidos. Todas las curvas ROC y valores AUC obtenidos se acercan a resultados ideales de clasificación de tumores. Estos excelentes resultados suponen una complicación a la hora de sacar conclusiones respecto a qué metodos de tratamiento de datos faltantes son mejores debido a que las diferencias en las comparaciones basadas en los resultados diagnósticos con curvas ROC son muy pequeñas.

Con los resultados relativos a la estimación de parámetros y los errores estándar sí se vieron diferencias. Se pudo apreciar que el método de casos completos sobreestimaba los errores estándar y la estimación de parámetros podía cambiar mucho según el mecanismo de pérdida de datos, dando resultados claramente sesgados. El método de la sustitución por la media daba en algún caso estimaciones de parámetros bastante alejadas de la estimación con los datos originales. El resto de métodos daba resultados similares, aunque el método con variables indicadoras de pérdida tiene la desventaja de que los *dummies* añaden variables al modelo logístico y no tienen interpretación clínica. Entonces, los mejores resultados, se obtendrían con los métodos de imputación múltiple. Entre MICE y Missing Forest, los resultados de MICE, utilizando los criterios de AIC, variables no significativas y variables más significativas, son ligeramente mejores. La diferencia entre los resultados obtenidos

se nota más en los mecanismos MAR y MNAR que en el mecanismo MCAR.

Como línea futura de investigación, podrían aplicarse métodos de tratamiento de datos faltantes Bayesianos, que podrían ser especialmente ventajosos si se han realizado estudios sobre las variables de interés con pérdida de datos. También sería interesante analizar el resultado de métodos en bases de datos con variables cualitativas.

Aunque el algoritmo MICE sí utiliza Estadística Bayesiana, los valores que imputa no pueden ser diferentes a uno de los valores de la base de datos sobre la que se aplica el algoritmo. Por ejemplo, con datos perdidos según el mecanismo MNAR, si hubiera un rango de valores de alguna variable en particular que se perdiera, con MICE no se podría imputar valores de ese rango. En cambio, con metología Bayesiana, se podría saber a partir de distribuciones a priori, que podrían darse los valores que se han perdido y, por lo tanto, se podrían imputar.

5. Bibliografía

[1] Soley-Bori M. Dealing with missing data: Key assumptions and methods for applied analysis. Technical Report No. 4: Boston University School of Public Health, Department of Health Policy & Management; 2013.

[2] van der Heijden GJMG, Donders ART, Stijnen T, Moons KGM. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. J Clin Epidemiol. 2006;59(10):1102-9.

[3] Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012; 367(14):1355–60.

[4] Zhou XH, Zhou C, Lui D, Ding X. Applied Missing Data Analysis in the Health Sciences. Hoboken, New Jersey: John Wiley & Sons; 2014.

[5] Enders CK. Applied Missing Data Analysis. New York: The Guildford Press; 2010.

[6] Briggs A, Clark T, Wolstenholme J, Clarke P. Missing.... presumed at random: cost-analysis of incomplete data. Health Econ. 2002;12(5):377-92.

[7] Schafer JL, Graham JW. Missing Data: Our View of the State of the Art. Psychol Methods. 2002;7(2):147-77.

[8] Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol. 1974;66(5):688-701.

[9] Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. BMC Med Res Methodol [Internet]. 2010 [cited 2016 Nov 28];10(7):[about 16 p.]. Available from: http://www.biomedcentral.com/1471-2288/10/7

[10] Dempster AP, Laird NM, Rubin DB. Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B Stat Methodol. 1977;39(1):1-38.

[11] Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Model. New York: Cambridge University Press; 2006.

[12] Baguley T, Andrews M. Handling Missing Data. In: Robertson J, Kaptein M, editors. Modern Statistical Methods for HCI. Cham, Switzerland: Springer; 2016. p. 57-82.

[13] van Buuren S, Groothuis-Oudshoorn K. MICE: multivariate imputation by chained equations in R. J Stat Softw. 2011;45(3):1-67.

[14] Rodríguez G. Chapter 3: Logit Models for Binary Data. In: Rodríguez G. Lecture Notes on Generalized Linear Models. Princeton, New Jersey: Princeton University Press. 2007. p. 1-50.

[15] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, et al. Orange: Data Mining Toolbox in Python. J Mach Learn Res. 2013;14(1):2349–53.

[16] Wolberg WH, Mangasarian OL. Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc Natl Acad Sci U S A. 1990;87(23):9193-6.

[17] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837-45.

[18] Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011;12:77.

[19] Carstensen B, Plummer M, Laara E, Hills M. Epi: A Package for Statistical Analysis in Epidemiology. R package version 2.12 [Internet]. CRAN. 2017 [cited 2017 May 15]. Available from: https://CRAN.R-project.org/package=Epi

[20] Vink G, Frank LE, Pannekoek J, van Buuren S. Predictive mean matching imputation of semicontinuous variables. Stat Neerl. 2014;68(1):61-90.

[21] Stekhoven DJ, Bühlmann P. MissForest: non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.

6. Anexos

6.1. Bases de datos B0-B3

Se van a mostrar parcialmente las bases de datos utilizadas en el trabajo, tanto la original como las simuladas para cada mecanismo de pérdida de datos. En concreto, se van a mostrar los valores entre las filas 100 y 111 para las cuatro bases de datos.

	Clump	Unif_Cell_Size	Unif_Cell_Shape	Marginal_Adh	Single_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	y
100	2	3	4	4	2	5	2	5	1	1
101	4	1	2	1	2	1	3	1	1	0
102	8	2	3	1	6	3	7	1	1	1
103	10	10	10	10	10	1	8	8	8	1
104	7	3	4	4	3	3	3	2	7	1
105	10	10	10	8	2	10	4	1	1	1
106	1	6	8	10	8	10	5	7	1	1
107	1	1	1	1	2	1	2	3	1	0
108	6	5	4	4	3	9	7	8	3	1
109	1	3	1	2	2	2	5	3	2	0
110	8	6	4	3	5	9	3	1	1	1
111	10	3	3	10	2	10	7	3	3	1

Figura 37: Base de datos original (B0)

	Clump	Unif_Cell_Size	Unif_Cell_Shape	Marginal_Adh	Single_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	y
100	2	3	4	4	2	5	2	5	1	1
101	NA	1	2	1	2	1	3	1	1	0
102	8	2	3	1	6	3	7	1	1	1
103	10	10	10	10	10	1	8	8	8	1
104	7	3	4	4	3	3	3	2	7	1
105	10	10	10	8	2	NA	4	1	1	1
106	1	6	8	10	8	10	5	7	1	1
107	1	1	1	1	2	1	2	3	1	0
108	NA	5	4	4	3	NA	7	8	3	1
109	NA	3	1	2	2	2	5	3	2	0
110	8	6	4	3	5	9	3	1	1	1
111	10	3	3	10	2	10	7	3	3	1

Figura 38: Base de datos con datos faltantes que siguen el patrón MCAR (B1)

	Clump	Unif_Cell_Size	Unif_Cell_Shape	Marginal_Adh	Single_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	y
100	2	3	4	4	2	NA	2	5	1	1
101	4	1	2	1	2	1	3	1	1	0
102	8	2	3	1	6	3	7	1	1	1
103	NA	10	10	10	10	1	8	8	8	1
104	NA	3	4	4	3	NA	3	2	7	1
105	NA	10	10	8	2	NA	4	1	1	1
106	NA	6	8	10	8	NA	5	7	1	1
107	1	1	1	1	2	1	2	3	1	0
108	NA	5	4	4	3	NA	7	8	3	1
109	1	3	1	2	2	2	5	3	2	0
110	NA	6	4	3	5	NA	3	1	1	1
111	NA	3	3	10	2	NA	7	3	3	1

Figura 39: Base de datos con datos faltantes que siguen el patrón MAR (B2)

	Clump	Unif_Cell_Size	Unif_Cell_Shape	Marginal_Adh	Single_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses	y
100	2	3	4	4	2	5	2	5	1	1
101	4	1	2	1	2	1	3	1	1	0
102	8	2	3	1	6	3	7	1	1	1
103	NA	10	10	10	10	1	8	8	8	1
104	NA	3	4	4	3	3	3	2	7	1
105	NA	10	10	8	2	NA	4	1	1	1
106	1	6	8	10	8	NA	5	7	1	1
107	1	1	1	1	2	1	2	3	1	0
108	NA	5	4	4	3	NA	7	8	3	1
109	1	3	1	2	2	2	5	3	2	0
110	NA	6	4	3	5	NA	3	1	1	1
111	NA	3	3	10	2	NA	7	3	3	1

Figura 40: Base de datos con datos faltantes que siguen el patrón MNAR (B3)

6.2. Códigos de programación

En primer lugar, se muestra el código empleado para obtener el modelo de regresión logística con la base de datos original, junto con la obtención de la curva ROC del modelo obtenido finalmente (incluyendo la parte ampliada de la curva en la Figura 20), el punto óptimo de corte y el intervalo de confianza al 95 % para AUC.

```
wisconsin <- read.csv("wisconsin.csv",sep = ";")

wisconsin$y[wisconsin$y==2] <- 0
wisconsin$y[wisconsin$y==4] <- 1

#Modelos glm
M1 <- glm(y~Clump+Unif_Cell_Size+Unif_Cell_Shape
          +Marginal_Adh+Single_Cell_Size+Bare_Nuclei
          +Bland_Chromatin+Normal_Nucleoli+Mitoses,
```

```
          family=binomial(link="logit"),data=wisconsin)
summary(M1)

M2 <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
          +Single_Cell_Size+Bare_Nuclei+Bland_Chromatin
          +Normal_Nucleoli+Mitoses,
          family=binomial(link="logit"),data=wisconsin)
summary(M2)

M3 <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh+Bare_Nuclei
          +Bland_Chromatin+Normal_Nucleoli+Mitoses,
          family=binomial(link="logit"),data=wisconsin)
summary(M3)

M4 <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh+Bare_Nuclei
          +Bland_Chromatin+Normal_Nucleoli,
          family=binomial(link="logit"),data=wisconsin)
summary(M4)

library(Epi)
#calcula los coeficientes del modelo con glm
rc <- ROC(form = y~Clump+Unif_Cell_Shape+Marginal_Adh
          +Bare_Nuclei+Bland_Chromatin+Normal_Nucleoli,
          plot="ROC",data = wisconsin, PV = FALSE)
x <- 1-rc$res$spec
y <- sort(rc$res$sens,decreasing = T)
plot(x,y,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
#Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt <- which.max(rowSums(rc$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc$res$lr.eta[opt]

require(pROC)
roc1 <- roc(wisconsin$y,predic,plot = TRUE)
```

```
print(roc1)
ICaucd<- ci.auc(roc1, method="d")
print(ICaucd)
```

A continuación, se muestra el código empleado para obtener la base de datos B1, el conjunto de métodos de tratamiento de datos faltantes y los resultados correspondientes con cada base de datos obtenida a partir de cada método. Además, entre la generación de datos de tipo MCAR y el código del primer método de tratamiento de datos faltantes (análisis de casos completos), se muestra el código que se ha utilizado para obtener los histogramas en los que se mostró anteriormente el comportamiento de las variables Clump y Bare_Nuclei tanto en B0 como en B1.

```
wisconsin <- read.csv("wisconsin.csv",sep = ";")

wisconsin$y[wisconsin$y==2] <- 0
wisconsin$y[wisconsin$y==4] <- 1
num_indi <- nrow(wisconsin)

require(pROC)
require(mice)
require(missForest)
require(Epi)

#Procedemos a generar datos de tipo MCAR
set.seed(3332119)
clasif <- sample(c(0,1,2,3),num_indi,replace=TRUE,
          prob=c(2/3,0.7/6,0.7/6,0.1))
wisc_mcar <- wisconsin
#columnas de la base de datos correspondientes a Clump y
#Bare_Nuclei (siguiente línea)
var_miss <- c(1,6)
wisc_mcar$Clump[clasif==1 | clasif==3] <- NA
wisc_mcar$Bare_Nuclei[clasif==2 | clasif==3] <- NA

#Descriptivo
ben <- wisconsin[wisconsin$y==0,]
mal <- wisconsin[wisconsin$y==1,]
mcar_ben <- wisc_mcar[wisc_mcar$y==0,]
mcar_mal <- wisc_mcar[wisc_mcar$y==1,]
```

```
#Clump (datos originales)
barplot(table(wisconsin$Clump)/num_indi*100,
        main = "Clump",xlab = 'Clump',ylab = '%',
        col = 'purple') #también se hace con ylim=c(0,60)
barplot(table(ben$Clump)/nrow(ben)*100,
        main = "Clump en patologías benignas",
        xlab = 'Clump',ylab = '%',col = 'green')
barplot(table(mal$Clump)/nrow(mal)*100,
        main = "Clump en patologías malignas",
        xlab = 'Clump',ylab = '%',col = 'red')
        #también se hace con ylim=c(0,30)

#Clump (MCAR)
barplot(table(wisc_mcar$Clump)/sum(table(
        wisc_mcar$Clump))*100,main = "Clump (MCAR)",
        xlab = 'Clump',ylab = '%',col = 'purple')
        #también se hace con ylim=c(0,60)
barplot(table(mcar_ben$Clump)/sum(table(
        mcar_ben$Clump))*100,
        main = "Clump en patologías benignas (MCAR)",
        xlab = 'Clump',ylab = '%',col = 'green')
barplot(table(mcar_mal$Clump)/sum(table(
        mcar_mal$Clump))*100,
        main = "Clump en patologías malignas (MCAR)",
        xlab = 'Clump',ylab = '%',col = 'red')

summary(wisconsin$Clump)
summary(wisc_mcar$Clump)

#Bare_Nuclei (datos originales)
barplot(table(wisconsin$Bare_Nuclei)/num_indi*100,
        main = "Bare_Nuclei",xlab = 'Bare_Nuclei',
        ylab = '%',col = 'purple')
        #también se hace con ylim=c(0,60)
barplot(table(ben$Bare_Nuclei)/nrow(ben)*100,
        main = "Bare_Nuclei en patologías benignas",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'green')
barplot(table(mal$Bare_Nuclei)/nrow(mal)*100,
        main = "Bare_Nuclei en patologías malignas",
```

```
        xlab = 'Bare_Nuclei',ylab = '%',col = 'red')

#Bare_Nuclei (MCAR)
barplot(table(wisc_mcar$Bare_Nuclei)/sum(table(
        wisc_mcar$Bare_Nuclei))*100,
        main = "Bare_Nuclei (MCAR)",xlab = 'Bare_Nuclei',
        ylab = '%',col = 'purple')
barplot(table(mcar_ben$Bare_Nuclei)/sum(table(
        mcar_ben$Bare_Nuclei))*100,
        main = "Bare_Nuclei en patologías benignas (MCAR)"
        ,xlab = 'Bare_Nuclei',ylab = '%',col = 'green')
barplot(table(mcar_mal$Bare_Nuclei)/sum(table(
        mcar_mal$Bare_Nuclei))*100,
        main = "Bare_Nuclei en patologías malignas (MCAR)"
        ,xlab = 'Bare_Nuclei',ylab = '%',col = 'red')
        #también se hace con ylim=c(0,80)

summary(wisconsin$Bare_Nuclei)
summary(wisc_mcar$Bare_Nuclei)

#Casos completos
wisc_mcar_cc <- wisc_mcar[clasif==0,]
M_MCAR_CC <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                 +Bare_Nuclei+Bland_Chromatin
                 +Normal_Nucleoli,
                 family=binomial(link="logit"),
                 data=wisc_mcar_cc)
summary(M_MCAR_CC)

rc_MCAR_CC <- ROC(form = y~Clump+Unif_Cell_Shape
                  +Marginal_Adh+Bare_Nuclei
                  +Bland_Chromatin+Normal_Nucleoli,
                  plot="ROC",data = wisc_mcar_cc,
                  PV = FALSE)
#Ampliación de zona
x1 <- 1-rc_MCAR_CC$res$spec
y1 <- sort(rc_MCAR_CC$res$sens,decreasing = T)
plot(x1,y1,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
```

```
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MCAR_CC <- which.max(
rowSums(rc_MCAR_CC$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MCAR_CC$res$lr.eta[opt_MCAR_CC]

predic_MCAR_CC <- predict(M_MCAR_CC)
roc_MCAR_CC <- roc(wisc_mcar_cc$y,predic_MCAR_CC,
               plot = FALSE)
ICauc_MCAR_CC <- ci.auc(roc_MCAR_CC, method="d")
print(ICauc_MCAR_CC)

#Sustituir por media de manera
wisc_mcar2 <- wisc_mcar
media_mcar2 <- c(mean(wisc_mcar2[,var_miss[1]],na.rm = T),
                 mean(wisc_mcar2[,var_miss[2]],na.rm = T))
#reemplazamos por la media
wisc_mcar2[(clasif==1 | clasif==3),
           var_miss[1]] <- media_mcar2[1]
wisc_mcar2[(clasif==2 | clasif==3),
           var_miss[2]] <- media_mcar2[2]

M_MCAR_media2 <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                     +Bare_Nuclei+Bland_Chromatin
                     +Normal_Nucleoli,
                     family=binomial(link="logit"),
                     data=wisc_mcar2)
summary(M_MCAR_media2)

rc_MCAR_media2 <- ROC(form = y~Clump+Unif_Cell_Shape
                      +Marginal_Adh+Bare_Nuclei
                      +Bland_Chromatin+Normal_Nucleoli,
                      plot="ROC", data = wisc_mcar2,
                      PV = FALSE)
#Ampliación de zona
```

```
x2 <- 1-rc_MCAR_media2$res$spec
y2 <- sort(rc_MCAR_media2$res$sens,decreasing = T)
plot(x2,y2,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MCAR_media2 <- which.max(
rowSums(rc_MCAR_media2$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MCAR_media2$res$lr.eta[opt_MCAR_media2]

predic_MCAR_media2 <- predict(M_MCAR_media2)
roc_MCAR_media2 <- roc(wisc_mcar2$y,predic_MCAR_media2,
                       plot = FALSE)
ICauc_MCAR_media2 <- ci.auc(roc_MCAR_media2, method="d")
print(ICauc_MCAR_media2)

#mice
imp <- mice(wisc_mcar, seed = 3332119)
wisc_mcar_mice <- complete(imp)

M_MCAR_mice <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                   +Bare_Nuclei+Bland_Chromatin
                   +Normal_Nucleoli,
                   family=binomial(link="logit"),
                   data=wisc_mcar_mice)
summary(M_MCAR_mice)

rc_MCAR_mice <- ROC(form = y~Clump+Unif_Cell_Shape
                    +Marginal_Adh+Bare_Nuclei
                    +Bland_Chromatin+Normal_Nucleoli,
                    plot="ROC", data = wisc_mcar_mice,
                    PV = FALSE)
#Ampliación de zona
x3 <- 1-rc_MCAR_mice$res$spec
y3 <- sort(rc_MCAR_mice$res$sens,decreasing = T)
```

```
plot(x3,y3,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MCAR_mice <- which.max(
rowSums(rc_MCAR_mice$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MCAR_mice$res$lr.eta[opt_MCAR_mice]

predic_MCAR_mice <- predict(M_MCAR_mice)
roc_MCAR_mice <- roc(wisc_mcar_mice$y,predic_MCAR_mice,
                     plot = FALSE)
ICauc_MCAR_mice <- ci.auc(roc_MCAR_mice, method="d")
print(ICauc_MCAR_mice)

#DESCRIPTIVO MICE (primero con NA)
summary(wisc_mcar$Clump)
summary(wisc_mcar_mice$Clump)
summary(wisc_mcar$Bare_Nuclei)
summary(wisc_mcar_mice$Bare_Nuclei)

#missing Forest
wisc_mcar_mf <- missForest(wisc_mcar)
wisc_mcar_mf <- as.data.frame(wisc_mcar_mf$ximp)

M_MCAR_mf <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                 +Bare_Nuclei+Bland_Chromatin
                 +Normal_Nucleoli,
                 family=binomial(link="logit"),
                 data=wisc_mcar_mf)
summary(M_MCAR_mf)

rc_MCAR_mf <- ROC(form = y~Clump+Unif_Cell_Shape
                  +Marginal_Adh+Bare_Nuclei
                  +Bland_Chromatin+Normal_Nucleoli,
                  plot="ROC", data = wisc_mcar_mf,
```

```
                    PV = FALSE)
#Ampliación de zona
x4 <- 1-rc_MCAR_mf$res$spec
y4 <- sort(rc_MCAR_mf$res$sens,decreasing = T)
plot(x4,y4,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MCAR_mf <- which.max(
rowSums(rc_MCAR_mf$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MCAR_mf$res$lr.eta[opt_MCAR_mf]

predic_MCAR_mf <- predict(M_MCAR_mf)
roc_MCAR_mf <- roc(wisc_mcar_mf$y,predic_MCAR_mf,
                   plot = FALSE)
ICauc_MCAR_mf <- ci.auc(roc_MCAR_mf, method="d")
print(ICauc_MCAR_mf)

#DESCRIPTIVO MF (primero con NA)
summary(wisc_mcar$Clump)
summary(wisc_mcar_mf$Clump)
summary(wisc_mcar$Bare_Nuclei)
summary(wisc_mcar_mf$Bare_Nuclei)

#con dummy
wisc_mcar_dummy <- wisc_mcar
wisc_mcar_dummy$dummy1 <- 0
wisc_mcar_dummy$dummy2 <- 0
na1 <- which(is.na(wisc_mcar[,var_miss[1]]))
na2 <- which(is.na(wisc_mcar[,var_miss[2]]))
wisc_mcar_dummy[na1,var_miss[1]] <- 0
wisc_mcar_dummy[na2,var_miss[2]] <- 0
wisc_mcar_dummy$dummy1[na1] <- 1
wisc_mcar_dummy$dummy2[na2] <- 1
```

```
M_MCAR_dummy <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                    +Bare_Nuclei+Bland_Chromatin
                    +Normal_Nucleoli+dummy1+dummy2,
                    family=binomial(link="logit"),
                    data=wisc_mcar_dummy)
summary(M_MCAR_dummy)

rc_MCAR_dummy <- ROC(form = y~Clump+Unif_Cell_Shape
                     +Marginal_Adh+Bare_Nuclei
                     +Bland_Chromatin+Normal_Nucleoli
                     +dummy1+dummy2, plot="ROC",
                     data = wisc_mcar_dummy, PV = FALSE)
#Ampliación de zona
x5 <- 1-rc_MCAR_dummy$res$spec
y5 <- sort(rc_MCAR_dummy$res$sens,decreasing = T)
plot(x5,y5,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MCAR_dummy <- which.max(
rowSums(rc_MCAR_dummy$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MCAR_dummy$res$lr.eta[opt_MCAR_dummy]

predic_MCAR_dummy <- predict(M_MCAR_dummy)
roc_MCAR_dummy <- roc(wisc_mcar_dummy$y,predic_MCAR_dummy,
                      plot = FALSE)
ICauc_MCAR_dummy <- ci.auc(roc_MCAR_dummy, method="d")
print(ICauc_MCAR_dummy)

#Representación conjunta de las zonas ampliadas
plot(x1,y1,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
lines(x2,y2, col="blue")
lines(x3,y3, col="green")
lines(x4,y4, col="red")
```

```
lines(x5,y5, col="brown")
grid(nx=2,ny=4)
```

Se procede a mostrar el código análogo al mostrado recientemente, pero con la base de datos B2. Además de mostrar el código utilizado para obtener los histogramas de Clump y Bare_Nuclei en B0 y en B2, se muestra el código empleado para obtener los histogramas de Marginal_Adh.

```
wisconsin <- read.csv("wisconsin.csv",sep = ";")

wisconsin$y[wisconsin$y==2] <- 0
wisconsin$y[wisconsin$y==4] <- 1
num_indi <- nrow(wisconsin)

require(pROC)
require(mice)
require(missForest)
require(Epi)

#Procedemos a generar datos de tipo MAR
clasif <- numeric(num_indi) #se inicializa la variable que
#indica dónde se generan los NA (vector con 0 por defecto)
wisc_mar <- wisconsin
set.seed(3333129) #semilla diferente respecto a B1
for(i in 1:num_indi){
  if(wisconsin$Marginal_Adh[i] > 2){
    clasif[i] <- sample(c(1,2,3), 1,
                        prob=c(0.7/2,0.7/2,0.3))
    if(clasif[i]==1) wisc_mar$Clump[i] <- NA
    else if(clasif[i]==2) wisc_mar$Bare_Nuclei[i] <- NA
    else{
      wisc_mar$Clump[i] <- NA
      wisc_mar$Bare_Nuclei[i] <- NA
    }
  }
}
var_miss <- c(1,6) #columnas de la base de datos con NA

#Descriptivo
ben <- wisconsin[wisconsin$y==0,]
```

```
mal <- wisconsin[wisconsin$y==1,]
mar_ben <- wisc_mar[wisc_mar$y==0,]
mar_mal <- wisc_mar[wisc_mar$y==1,]

#Marginal_Adh (datos originales)
barplot(table(wisconsin$Marginal_Adh)/num_indi*100,
        main = "Marginal_Adh",xlab = 'Marginal_Adh',
        ylab = '%',col = 'purple')
barplot(table(ben$Marginal_Adh)/nrow(ben)*100,
        main = "Marginal_Adh en Patologías benignas",
        xlab = 'Marginal_Adh',ylab = '%',col = 'green')
barplot(table(mal$Marginal_Adh)/nrow(mal)*100,
        main = "Marginal_Adh en patologías malignas",
        xlab = 'Marginal_Adh',
        ylab = '%',col = 'red')
        #para comparar se hace con ylim=c(0,80)

#Clump (datos originales)
barplot(table(wisconsin$Clump)/num_indi*100,main = "Clump"
        ,xlab = 'Clump',ylab = '%',col = 'purple')
barplot(table(ben$Clump)/nrow(ben)*100,
        main = "Clump en patologías benignas",
        xlab = 'Clump',ylab = '%',col = 'green')
barplot(table(mal$Clump)/nrow(mal)*100,
        main = "Clump en patologías malignas",
        xlab = 'Clump',ylab = '%',col = 'red')
        #también se hace con ylim=c(0,25)

#Clump (MAR)
barplot(table(wisc_mar$Clump)/sum(table(
        wisc_mar$Clump))*100,main = "Clump (MAR)",
        xlab = 'Clump',ylab = '%',col = 'purple')
barplot(table(mar_ben$Clump)/sum(table(
        mar_ben$Clump))*100,
        main = "Clump en patologías benignas (MAR)",
        xlab = 'Clump',ylab = '%',col = 'green')
barplot(table(mar_mal$Clump)/sum(table(
        mar_mal$Clump))*100,
        main = "Clump en patologías malignas (MAR)",
```

```
        xlab = 'Clump',ylab = '%',col = 'red')

summary(wisconsin$Clump)
summary(wisc_mar$Clump)

#Bare_Nuclei (datos originales)
barplot(table(wisconsin$Bare_Nuclei)/num_indi*100,
        main = "Bare_Nuclei",xlab = 'Bare_Nuclei',
        ylab = '%',col = 'purple')
barplot(table(ben$Bare_Nuclei)/nrow(ben)*100,
        main = "Bare_Nuclei en patologías benignas",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'green')
barplot(table(mal$Bare_Nuclei)/nrow(mal)*100,
        main = "Bare_Nuclei en patologías malignas",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'red')

#Bare_Nuclei (MAR)
barplot(table(wisc_mar$Bare_Nuclei)/sum(table(
        wisc_mar$Bare_Nuclei))*100,
        main = "Bare_Nuclei (MAR)",xlab = 'Bare_Nuclei',
        ylab = '%',col = 'purple')
barplot(table(mar_ben$Bare_Nuclei)/sum(table(
        mar_ben$Bare_Nuclei))*100,
        main = "Bare_Nuclei en patologías benignas (MAR)",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'green')
barplot(table(mar_mal$Bare_Nuclei)/sum(table(
        mar_mal$Bare_Nuclei))*100,
        main = "Bare_Nuclei en patologías malignas (MAR)",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'red')
        #también se hace con ylim=c(0,80)

summary(wisconsin$Bare_Nuclei)
summary(wisc_mar$Bare_Nuclei)

#Casos completos
wisc_mar_cc <- wisc_mar[clasif==0,]
M_MAR_CC <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                +Bare_Nuclei+Bland_Chromatin
                +Normal_Nucleoli,
```

```
                family=binomial(link="logit"),
                data=wisc_mar_cc)
summary(M_MAR_CC)

rc_MAR_CC <- ROC(form = y~Clump+Unif_Cell_Shape
                 +Marginal_Adh+Bare_Nuclei+Bland_Chromatin
                 +Normal_Nucleoli, plot="ROC",
                  data = wisc_mar_cc, PV = FALSE)
#Ampliación de zona
x1 <- 1-rc_MAR_CC$res$spec
y1 <- sort(rc_MAR_CC$res$sens,decreasing = T)
plot(x1,y1,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MAR_CC <- which.max(
rowSums(rc_MAR_CC$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MAR_CC$res$lr.eta[opt_MAR_CC]

predic_MAR_CC <- predict(M_MAR_CC)
roc_MAR_CC <- roc(wisc_mar_cc$y,predic_MAR_CC,
                  plot = FALSE)
ICauc_MAR_CC <- ci.auc(roc_MAR_CC, method="d")
print(ICauc_MAR_CC)

#Sustituir por media de manera
wisc_mar2 <- wisc_mar
media_mar2 <- c(mean(wisc_mar2[,var_miss[1]],na.rm = T),
                mean(wisc_mar2[,var_miss[2]],na.rm = T))
#reemplazamos por la media
wisc_mar2[(clasif==1 | clasif==3),
          var_miss[1]] <- media_mar2[1]
wisc_mar2[(clasif==2 | clasif==3),
          var_miss[2]] <- media_mar2[2]
```

```
M_MAR_media2 <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                    +Bare_Nuclei+Bland_Chromatin
                    +Normal_Nucleoli,
                    family=binomial(link="logit"),
                    data=wisc_mar2)
summary(M_MAR_media2)

rc_MAR_media2 <- ROC(form = y~Clump+Unif_Cell_Shape
                     +Marginal_Adh+Bare_Nuclei
                     +Bland_Chromatin+Normal_Nucleoli,
                     plot="ROC", data = wisc_mar2,
                     PV = FALSE)
#Ampliación de zona
x2 <- 1-rc_MAR_media2$res$spec
y2 <- sort(rc_MAR_media2$res$sens,decreasing = T)
plot(x2,y2,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MAR_media2 <- which.max(
rowSums(rc_MAR_media2$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MAR_media2$res$lr.eta[opt_MAR_media2]

predic_MAR_media2 <- predict(M_MAR_media2)
roc_MAR_media2 <- roc(wisc_mar2$y,predic_MAR_media2,
                      plot = FALSE)
ICauc_MAR_media2 <- ci.auc(roc_MAR_media2, method="d")
print(ICauc_MAR_media2)

#mice
imp <- mice(wisc_mar, seed = 3333129)
wisc_mar_mice <- complete(imp)

M_MAR_mice <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                  +Bare_Nuclei+Bland_Chromatin
```

```
                    +Normal_Nucleoli,
                    family=binomial(link="logit"),
                    data=wisc_mar_mice)
summary(M_MAR_mice)

rc_MAR_mice <- ROC(form = y~Clump+Unif_Cell_Shape
                    +Marginal_Adh+Bare_Nuclei
                    +Bland_Chromatin+Normal_Nucleoli,
                    plot="ROC", data = wisc_mar_mice,
                    PV = FALSE)
#Ampliación de zona
x3 <- 1-rc_MAR_mice$res$spec
y3 <- sort(rc_MAR_mice$res$sens,decreasing = T)
plot(x3,y3,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MAR_mice <- which.max(
rowSums(rc_MAR_mice$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MAR_mice$res$lr.eta[opt_MAR_mice]

predic_MAR_mice <- predict(M_MAR_mice)
roc_MAR_mice <- roc(wisc_mar_mice$y,predic_MAR_mice,
                    plot = FALSE)
ICauc_MAR_mice <- ci.auc(roc_MAR_mice, method="d")
print(ICauc_MAR_mice)

#DESCRIPTIVO MICE (primero con NA)
summary(wisc_mar$Clump)
summary(wisc_mar_mice$Clump)
summary(wisc_mar$Bare_Nuclei)
summary(wisc_mar_mice$Bare_Nuclei)

#missing Forest
wisc_mar_mf <- missForest(wisc_mar)
```

```
wisc_mar_mf <- as.data.frame(wisc_mar_mf$ximp)

M_MAR_mf <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                +Bare_Nuclei+Bland_Chromatin
                +Normal_Nucleoli,
                family=binomial(link="logit"),
                data=wisc_mar_mf)
summary(M_MAR_mf)

rc_MAR_mf <- ROC(form = y~Clump+Unif_Cell_Shape
                 +Marginal_Adh+Bare_Nuclei+Bland_Chromatin
                 +Normal_Nucleoli,
                 plot="ROC", data = wisc_mar_mf,
                 PV = FALSE)
#Ampliación de zona
x4 <- 1-rc_MAR_mf$res$spec
y4 <- sort(rc_MAR_mf$res$sens,decreasing = T)
plot(x4,y4,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MAR_mf <- which.max(
rowSums(rc_MAR_mf$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MAR_mf$res$lr.eta[opt_MAR_mf]

predic_MAR_mf <- predict(M_MAR_mf)
roc_MAR_mf <- roc(wisc_mar_mf$y,predic_MAR_mf,
                  plot = FALSE)
ICauc_MAR_mf <- ci.auc(roc_MAR_mf, method="d")
print(ICauc_MAR_mf)

#DESCRIPTIVO MF (primero con NA)
summary(wisc_mar$Clump)
summary(wisc_mar_mf$Clump)
summary(wisc_mar$Bare_Nuclei)
```

```
summary(wisc_mar_mf$Bare_Nuclei)

#con dummy
wisc_mar_dummy <- wisc_mar
wisc_mar_dummy$dummy1 <- 0
wisc_mar_dummy$dummy2 <- 0
na1 <- which(is.na(wisc_mar[,var_miss[1]]))
na2 <- which(is.na(wisc_mar[,var_miss[2]]))
wisc_mar_dummy[na1,var_miss[1]] <- 0
wisc_mar_dummy[na2,var_miss[2]] <- 0
wisc_mar_dummy$dummy1[na1] <- 1
wisc_mar_dummy$dummy2[na2] <- 1

M_MAR_dummy <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                   +Bare_Nuclei+Bland_Chromatin
                   +Normal_Nucleoli+dummy1+dummy2,
                   family=binomial(link="logit"),
                   data=wisc_mar_dummy)
summary(M_MAR_dummy)

rc_MAR_dummy <- ROC(form = y~Clump+Unif_Cell_Shape
                    +Marginal_Adh+Bare_Nuclei
                    +Bland_Chromatin+Normal_Nucleoli
                    +dummy1+dummy2,
                    plot="ROC", data = wisc_mar_dummy,
                    PV = FALSE)
#Ampliación de zona
x5 <- 1-rc_MAR_dummy$res$spec
y5 <- sort(rc_MAR_dummy$res$sens,decreasing = T)
plot(x5,y5,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MAR_dummy <- which.max(
rowSums(rc_MAR_dummy$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
```

```
rc_MAR_dummy$res$lr.eta[opt_MAR_dummy]

predic_MAR_dummy <- predict(M_MAR_dummy)
roc_MAR_dummy <- roc(wisc_mar_dummy$y,predic_MAR_dummy,
                     plot = FALSE)
ICauc_MAR_dummy <- ci.auc(roc_MAR_dummy, method="d")
print(ICauc_MAR_dummy)

#Representación conjunta de las zonas ampliadas
plot(x1,y1,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
lines(x2,y2, col="blue")
lines(x3,y3, col="green")
lines(x4,y4, col="red")
lines(x5,y5, col="brown")
grid(nx=2,ny=4)
```

Se procede a mostrar el código análogo al mostrado recientemente, pero con la base de datos B3. El código utilizado para obtener los histogramas en este caso corresponde a B0 y a B3.

```
wisconsin <- read.csv("wisconsin.csv",sep = ";")

wisconsin$y[wisconsin$y==2] <- 0
wisconsin$y[wisconsin$y==4] <- 1
num_indi <- nrow(wisconsin)

require(pROC)
require(mice)
require(missForest)
require(Epi)

#Procedemos a generar datos de tipo MNAR
clasif <- numeric(num_indi) #se inicializa la variable que
#indica dónde se generan los NA (vector con 0 por defecto)
wisc_mnar <- wisconsin

for(i in 1:num_indi){
  if(wisconsin$Marginal_Adh[i] > 2){
    if(wisc_mnar$Clump[i]>5 & wisc_mnar$Bare_Nuclei[i]<=7)
```

```
      {wisc_mnar$Clump[i] <- NA
      clasif[i] <- 1
      }
    else if(wisc_mnar$Clump[i]<=5 &
            wisc_mnar$Bare_Nuclei[i]>7){
      wisc_mnar$Bare_Nuclei[i] <- NA
      clasif[i] <- 2
    }
    else if(wisc_mnar$Clump[i]>5 &
            wisc_mnar$Bare_Nuclei[i]>7){
      wisc_mnar$Clump[i] <- NA
      wisc_mnar$Bare_Nuclei[i] <- NA
      clasif[i] <- 3
    }
  }
}
var_miss <- c(1,6) #columnas de las variables con missing

#Descriptivo
ben <- wisconsin[wisconsin$y==0,]
mal <- wisconsin[wisconsin$y==1,]
mnar_ben <- wisc_mnar[wisc_mnar$y==0,]
mnar_mal <- wisc_mnar[wisc_mnar$y==1,]

#Clump (datos originales)
barplot(table(wisconsin$Clump)/num_indi*100,
        main = "Clump",xlab = 'Clump',ylab = '%',
        col = 'purple') #ylim=c(0,25) para comparar
barplot(table(ben$Clump)/nrow(ben)*100,
        main = "Clump␣en␣patologías␣benignas",
        xlab = 'Clump',ylab = '%',col = 'green')
barplot(table(mal$Clump)/nrow(mal)*100,
        main = "Clump␣en␣patologías␣malignas",
        xlab = 'Clump',ylab = '%',col = 'red')

#Clump (MNAR)
barplot(table(wisc_mnar$Clump)/sum(table(
        wisc_mnar$Clump))*100,main = "Clump␣(MNAR)",
        xlab = 'Clump',ylab = '%',col = 'purple')
```

```
barplot(table(mnar_ben$Clump)/sum(table(
        mnar_ben$Clump))*100,
        main = "Clump␣en␣patologías␣benignas␣(MNAR)",
        xlab = 'Clump',ylab = '%',col = 'green')
        #ylim=c(0,40) para comparar
barplot(table(mnar_mal$Clump)/sum(table(
        mnar_mal$Clump))*100,
        main = "Clump␣en␣patologías␣malignas␣(MNAR)",
        xlab = 'Clump',ylab = '%',col = 'red')
        #ylim=c(0,40) para comparar

summary(wisconsin$Clump)
summary(wisc_mnar$Clump)

#Bare_Nuclei (datos originales)
barplot(table(wisconsin$Bare_Nuclei)/num_indi*100,
        main = "Bare_Nuclei",xlab = 'Bare_Nuclei',
        ylab = '%',col = 'purple')
        #ylim=c(0,70) para comparar
barplot(table(ben$Bare_Nuclei)/nrow(ben)*100,
        main = "Bare_Nuclei␣en␣patologías␣benignas",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'green')
barplot(table(mal$Bare_Nuclei)/nrow(mal)*100,
        main = "Bare_Nuclei␣en␣patologías␣malignas",
        xlab = 'Bare_Nuclei',ylab = '%',col = 'red')

#Bare_Nuclei (MNAR)
barplot(table(wisc_mnar$Bare_Nuclei)/sum(table(
        wisc_mnar$Bare_Nuclei))*100,
        main = "Bare_Nuclei␣(MNAR)",xlab = 'Bare_Nuclei',
        ylab = '%',col = 'purple')
barplot(table(mnar_ben$Bare_Nuclei)/sum(table(
        mnar_ben$Bare_Nuclei))*100,
        main = "Bare_Nuclei␣en␣patologías␣benignas␣(MNAR)"
        ,xlab = 'Bare_Nuclei',ylab = '%',col = 'green')
barplot(table(mnar_mal$Bare_Nuclei)/sum(table(
        mnar_mal$Bare_Nuclei))*100,
        main = "Bare_Nuclei␣en␣patologías␣malignas␣(MNAR)"
        ,xlab = 'Bare_Nuclei',ylab = '%',col = 'red')
```

```
        #ylim=c(0,80) para comparar

summary(wisconsin$Bare_Nuclei)
summary(wisc_mnar$Bare_Nuclei)

#Casos completos
wisc_mnar_cc <- wisc_mnar[clasif==0,]
M_MNAR_CC <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                 +Bare_Nuclei+Bland_Chromatin
                 +Normal_Nucleoli,
                 family=binomial(link="logit"),
                 data=wisc_mnar_cc)
summary(M_MNAR_CC)

rc_MNAR_CC <- ROC(form = y~Clump+Unif_Cell_Shape
                  +Marginal_Adh+Bare_Nuclei
                  +Bland_Chromatin+Normal_Nucleoli,
                  plot="ROC",data = wisc_mnar_cc,
                  PV = FALSE)
#Ampliación de zona
x1 <- 1-rc_MNAR_CC$res$spec
y1 <- sort(rc_MNAR_CC$res$sens,decreasing = T)
plot(x1,y1,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MNAR_CC <- which.max(
rowSums(rc_MNAR_CC$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MNAR_CC$res$lr.eta[opt_MNAR_CC]

predic_MNAR_CC <- predict(M_MNAR_CC)
roc_MNAR_CC <- roc(wisc_mnar_cc$y,predic_MNAR_CC,
                   plot = FALSE)
ICauc_MNAR_CC <- ci.auc(roc_MNAR_CC, method="d")
print(ICauc_MNAR_CC)
```

```
#Sustituir por media de manera
wisc_mnar2 <- wisc_mnar
media_mnar2 <- c(mean(wisc_mnar2[,var_miss[1]],na.rm = T),
                 mean(wisc_mnar2[,var_miss[2]],na.rm = T))
#reemplazamos por la media
wisc_mnar2[(clasif==1 | clasif==3),
           var_miss[1]] <- media_mnar2[1]
wisc_mnar2[(clasif==2 | clasif==3),
           var_miss[2]] <- media_mnar2[2]

M_MNAR_media2 <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                     +Bare_Nuclei+Bland_Chromatin
                     +Normal_Nucleoli,
                     family=binomial(link="logit"),
                     data=wisc_mnar2)
summary(M_MNAR_media2)

rc_MNAR_media2 <- ROC(form = y~Clump+Unif_Cell_Shape
                      +Marginal_Adh+Bare_Nuclei
                      +Bland_Chromatin+Normal_Nucleoli,
                      plot="ROC", data = wisc_mnar2,
                      PV = FALSE)
#Ampliación de zona
x2 <- 1-rc_MNAR_media2$res$spec
y2 <- sort(rc_MNAR_media2$res$sens,decreasing = T)
plot(x2,y2,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MNAR_media2 <- which.max(
rowSums(rc_MNAR_media2$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MNAR_media2$res$lr.eta[opt_MNAR_media2]

predic_MNAR_media2 <- predict(M_MNAR_media2)
```

```
roc_MNAR_media2 <- roc(wisc_mnar2$y,predic_MNAR_media2,
                       plot = FALSE)
ICauc_MNAR_media2 <- ci.auc(roc_MNAR_media2, method="d")
print(ICauc_MNAR_media2)

#mice
imp <- mice(wisc_mnar, seed = 3333129)
wisc_mnar_mice <- complete(imp)

M_MNAR_mice <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                   +Bare_Nuclei+Bland_Chromatin
                   +Normal_Nucleoli,
                   family=binomial(link="logit"),
                   data=wisc_mnar_mice)
summary(M_MNAR_mice)

rc_MNAR_mice <- ROC(form = y~Clump+Unif_Cell_Shape
                    +Marginal_Adh+Bare_Nuclei
                    +Bland_Chromatin+Normal_Nucleoli,
                    plot="ROC", data = wisc_mnar_mice,
                    PV = FALSE)
#Ampliación de zona
x3 <- 1-rc_MNAR_mice$res$spec
y3 <- sort(rc_MNAR_mice$res$sens,decreasing = T)
plot(x3,y3,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MNAR_mice <- which.max(
rowSums(rc_MNAR_mice$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MNAR_mice$res$lr.eta[opt_MNAR_mice]

predic_MNAR_mice <- predict(M_MNAR_mice)
roc_MNAR_mice <- roc(wisc_mnar_mice$y,predic_MNAR_mice,
                     plot = FALSE)
```

```
ICauc_MNAR_mice <- ci.auc(roc_MNAR_mice, method="d")
print(ICauc_MNAR_mice)

#DESCRIPTIVO MICE (primero con NA)
summary(wisc_mnar$Clump)
summary(wisc_mnar_mice$Clump)
summary(wisc_mnar$Bare_Nuclei)
summary(wisc_mnar_mice$Bare_Nuclei)

#missing Forest
wisc_mnar_mf <- missForest(wisc_mnar)
wisc_mnar_mf <- as.data.frame(wisc_mnar_mf$ximp)

M_MNAR_mf <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                 +Bare_Nuclei+Bland_Chromatin
                 +Normal_Nucleoli,
                 family=binomial(link="logit"),
                 data=wisc_mnar_mf)
summary(M_MNAR_mf)

rc_MNAR_mf <- ROC(form = y~Clump+Unif_Cell_Shape
                  +Marginal_Adh+Bare_Nuclei
                  +Bland_Chromatin+Normal_Nucleoli,
                  plot="ROC", data = wisc_mnar_mf,
                  PV = FALSE)
#Ampliación de zona
x4 <- 1-rc_MNAR_mf$res$spec
y4 <- sort(rc_MNAR_mf$res$sens,decreasing = T)
plot(x4,y4,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MNAR_mf <- which.max(
rowSums(rc_MNAR_mf$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MNAR_mf$res$lr.eta[opt_MNAR_mf]
```

```
predic_MNAR_mf <- predict(M_MNAR_mf)
roc_MNAR_mf <- roc(wisc_mnar_mf$y,predic_MNAR_mf,
                   plot = FALSE)
ICauc_MNAR_mf <- ci.auc(roc_MNAR_mf, method="d")
print(ICauc_MNAR_mf)

#DESCRIPTIVO MF (primero con NA)
summary(wisc_mnar$Clump)
summary(wisc_mnar_mf$Clump)
summary(wisc_mnar$Bare_Nuclei)
summary(wisc_mnar_mf$Bare_Nuclei)

#con dummy
wisc_mnar_dummy <- wisc_mnar
wisc_mnar_dummy$dummy1 <- 0
wisc_mnar_dummy$dummy2 <- 0
na1 <- which(is.na(wisc_mnar[,var_miss[1]]))
na2 <- which(is.na(wisc_mnar[,var_miss[2]]))
wisc_mnar_dummy[na1,var_miss[1]] <- 0
wisc_mnar_dummy[na2,var_miss[2]] <- 0
wisc_mnar_dummy$dummy1[na1] <- 1
wisc_mnar_dummy$dummy2[na2] <- 1

M_MNAR_dummy <- glm(y~Clump+Unif_Cell_Shape+Marginal_Adh
                    +Bare_Nuclei+Bland_Chromatin
                    +Normal_Nucleoli+dummy1+dummy2,
                    family=binomial(link="logit"),
                    data=wisc_mnar_dummy)
summary(M_MNAR_dummy)

rc_MNAR_dummy <- ROC(form = y~Clump+Unif_Cell_Shape
                     +Marginal_Adh+Bare_Nuclei
                     +Bland_Chromatin+Normal_Nucleoli
                     +dummy1+dummy2,
                     plot="ROC", data = wisc_mnar_dummy,
                     PV = FALSE)
#Ampliación de zona
x5 <- 1-rc_MNAR_dummy$res$spec
```

```
y5 <- sort(rc_MNAR_dummy$res$sens,decreasing = T)
plot(x5,y5,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
grid(nx=2,ny=4)
## Combinación óptima para maximizar sensibilidad y
#especificidad (equivalente a índice de Youden)
#Se devuelve el índice en opt
opt_MNAR_dummy <- which.max(
rowSums(rc_MNAR_dummy$res[, c("sens", "spec")]))
#Valor óptimo para el punto de corte en la regresión
#logística
rc_MNAR_dummy$res$lr.eta[opt_MNAR_dummy]

predic_MNAR_dummy <- predict(M_MNAR_dummy)
roc_MNAR_dummy <- roc(wisc_mnar_dummy$y,predic_MNAR_dummy,
                      plot = FALSE)
ICauc_MNAR_dummy <- ci.auc(roc_MNAR_dummy, method="d")
print(ICauc_MNAR_dummy)

#Representación conjunta de las zonas ampliadas
plot(x1,y1,type = "l",xlim=c(0,0.2), ylim=c(0.6,1),
     xlab="1-Specificity",ylab = "Sensitivity")
lines(x2,y2, col="blue")
lines(x3,y3, col="green")
lines(x4,y4, col="red")
lines(x5,y5, col="brown")
grid(nx=2,ny=4)
```