Mejora de la predicción en tareas de Aprendizaje Automático Supervisado: Imputación de valores perdidos.

by

Lucía Yanina Del Turco

A thesis submitted in conformity with the requirements
for the MSc in Economics, Finance and Computer Science

University of Huelva & International University of Andalucia

November 2021

Mejora de la predicción en tareas de Aprendizaje Automático Supervisado: Imputación de valores perdidos.

Lucía Yanina Del Turco

Máster en Economía, Finanzas y Computación

Supervisor: Antonio J. Tallón Ballesteros
Universidad de Huelva y Universidad Internacional de Andalucía

2021

Abstract

In this work, the recurring problems which databases usually present are analyzed and studied: missing data or missing values. Six imputation methods are assessed after its application to regression data sets are evaluated in the presence of missing values. In addition, the improvement percentages in the prediction according to Root Mean Squared Error (RMSE) and R^2 are determined for three different algorithms in eight databases, being the main goal to determine where there is an ideal imputation method with the best performance for all cases where there are missing values, or on the contrary, it is convenient to adopt different imputation procedures according to the applied regression method and the analyzed database.

Key words: Supervised Machine Learning, Missing Values, Value Imputation, Simple Imputation, Multiple Imputation.

Resumen

En el presente trabajo, se analiza y estudia la problemática recurrente que suelen presentar las bases de datos: los *missing data* o valores perdidos. Se evalúan seis métodos de imputación tras su aplicación a conjuntos datos de regresión en presencia de valores ausentes. Además, se determinan los porcentajes de mejora en la predicción según RMSE (raíz del error cuadrático medio) y R^2 para tres algoritmos diferentes en ocho bases de datos, siendo el objetivo central del mismo determinar si existe un método de imputación con mejor rendimiento para todos los casos donde existan valores perdidos, o por el contrario, resulta conveniente adoptar distintos procedimientos de imputación de acuerdo al método de regresión aplicado y a la base de datos analizada.

Palabras Clave: Aprendizaje Automático Supervisado, Valores Perdidos, Imputación de Valores, Imputación Simple, Imputación Múltiple.

Agradecimientos

Quiero agradecer en primer lugar a mi tutor, Antonio Tallón, por su gran ayuda y asesoramiento a la distancia y principalmente por haber depositado su confianza en mí.

A la Universidad Internacional de Andalucía por permitirme transitar esta gran aventura.

Y no podría olvidarme de mi familia, por su comprensión y cariño en todo momento.

Índice

Lista de Tablas

Lista de Figuras

1 Introducción

Es habitual enfrentarnos a la presencia de valores perdidos, es decir, registros en los que faltan valores para algunos de los atributos. Disponer de bases de datos completas es ideal, pero aplicar procedimientos inapropiados de sustitución de información para lograrlo, puede generar más problemas de los que resuelve, ya que introduce sesgos y reduce el poder explicativo de los métodos estadísticos, le resta eficiencia a la fase de inferencia y puede incluso, invalidar las conclusiones del estudio. Una forma de lidiar con los valores faltantes es imputarlos. La ventaja de la aplicación de técnicas de imputación radica en la recuperación de la información y la posibilidad de poder disponer de registros completos permitiendo el empleo de métodos que no son capaces de operar con valores perdidos o bien simplemente descartan objetos con valores perdidos.

Existe un enfoque estadístico del manejo de valores perdidos, donde podemos encontrar las opciones tradicionales para tratarlos, tales como: la eliminación de datos (*listwise*), el pareo de observaciones (*pairwise*), la imputación única, la imputación múltiple, el método de máxima verosimilitud. Sin embargo, también existe un enfoque de aprendizaje automático donde se proponen nuevos algoritmos como: arboles de decisión, k - vecinos más cercanos, redes neuronales, entre otros [1].

Los primeros aportes en imputación se realizaron en 1932 por Wilks [2], quien propuso el reemplazo de los datos faltantes por la media de los datos presentes de la variable.

Con el avance tecnológico de los sistemas computacionales se iniciaron investigaciones en las décadas de los setenta y ochenta. Entre los principales autores que han hecho grandes investigaciones referentes a imputación figuran Rubin [3][4], Little y Rubin [4] y Hemel [6]. Se resumen, brevemente, sus aportes fundamentales:

En 1983, Rubin expuso dos enfoques, uno basado en la aleatorización, enfoque aleatorio y otro basado en el modelo de superpoblaciones, enfoque Bayesiano [3].

En 1987, Little y Rubin [4], desarrollan una nueva técnica llamada Imputación Múltiple (IM), en la que los datos faltantes, son sustituidos por m>1 valores simulados. La IM permitió hacer un

uso eficiente de los datos, obtener estimadores no sesgados y reflejar adecuadamente la incertidumbre que la no respuesta parcial introduce en la estimación de parámetros [5].

En el mismo año, Hemel [6] aportó un método llamado Listwise, el cual, es usado cuando se tiene un gran conjunto de datos y se puede eliminar la fila o columna donde se encuentra el valor perdido, para obtener una base, aunque más pequeña, completa. Esta técnica es comúnmente usada hoy en día, pero no es recomendable, por la pérdida de información que ocasiona.

En la década de los 90 del siglo pasado, Todeschini [7] propuso *k*-vecinos más cercanos (*k*-NN, *k-Nearest Neighbours*) como método de estimación de valores perdidos; obteniéndose buenos resultados cuando se cuenta con información auxiliar.

A partir del año 2000, se ha implementado el uso de árboles de clasificación como mejora de los procedimientos de imputación [8].

Otras investigaciones más recientes han buscado maneras de mejorar las técnicas de imputación o crear nuevas, como las basadas en análisis factorial [9], en análisis de componentes principales [10] o basada en redes bayesianas [11], entre otras.

Aún son muchos los estudios que se deben llevar a cabo para resolver los problemas que se presentan en la mayoría de los métodos de imputación que existen hoy en día, haciendo distinción en los diferentes enfoques que puedan existir y dependiendo además del conjunto de datos a analizar.

En este trabajo, se analizan los fundamentos teóricos de un conjunto amplio de métodos de imputación. En la primera parte, se describe la teoría en la que se sustentan, haciendo énfasis en sus bondades y limitaciones, así como en los sesgos que se generan cuando se utilizan de manera acrítica. En la segunda parte, se aplican seis métodos de imputación con el objetivo de sustituir los valores omitidos de las diferentes variables de las bases de datos estudiadas. Posteriormente, se evalúa el porcentaje de mejora en la predicción, respecto a la base de datos original, del coeficiente de determinación (R^2) y del RMSE de los métodos utilizados en tres algoritmos de regresión diferentes: Regresión lineal múltiple, *k*-NN y máquinas de vector soporte (SVM, *Support Vector Machine*); y a modo de conclusión, se presentan algunas reflexiones que permiten afirmar que no existe un mejor método de imputación.

2 Marco Teórico

La minería de datos es un enfoque moderno para resolver muchos problemas complejos y del mundo real. Este término bastante auto explicativo es un proceso bien conocido y ampliamente utilizado que evoluciona con nuevas tecnologías. En la minería de datos, el preprocesamiento de datos es el más paso importante para garantizar la calidad de los datos y los resultados que conduce a decisiones fiables. Según Vivek [12], el preprocesamiento de datos es el proceso de transformación simple de datos en un formato comprensible. Especialmente el preprocesamiento de datos incluye las actividades de limpieza de datos, integración, transformación, reducción y discretización de datos como se muestra en la Figura 1.

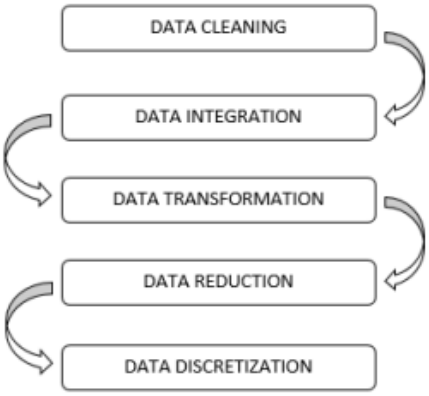


Figura 1: Tarea de minería de datos. *Fuente:* [12]

Una actividad crítica en el preprocesamiento de datos es lidiar con datos perdidos. Este proceso se incluye en la primera etapa de preprocesamiento de datos, que es la limpieza de datos. Esta primera etapa se preocupa por detectar datos incompletos, inexactos, inconsistentes y corruptos, y aplicar técnicas, modificar o eliminar estos datos espurios [12].

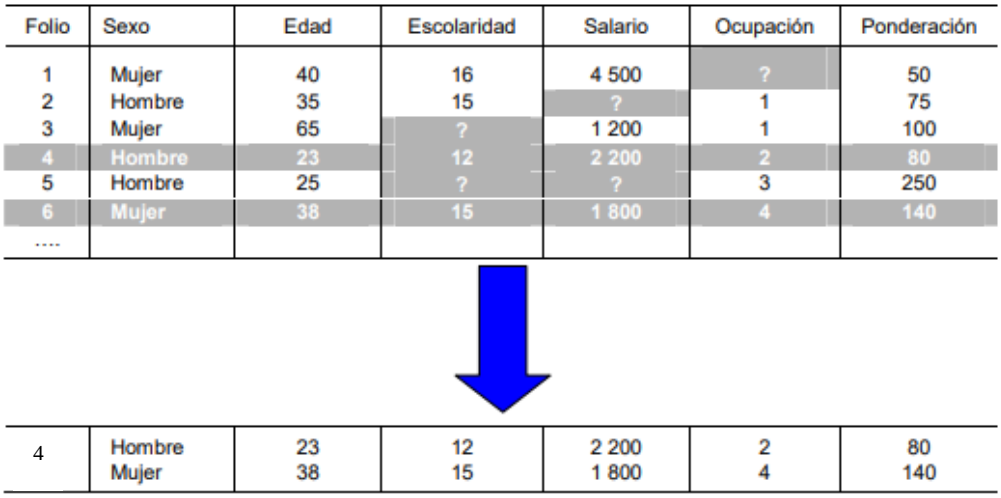
En cierta observación de interés, los valores faltantes se pueden definir como la ausencia de valor para una o varias instancias de un atributo.

A continuación, se detallan distintos procedimientos para sustituir la falta de valores.

2.1 Métodos de Imputación

2.1.1 Análisis con datos completos (*Listwise o case deletion LD*)

Es habitual que los paquetes estadísticos trabajen -por defecto- sólo con información completa (*Listwise*), a pesar de que se reconoce que esta práctica no es la más apropiada, ya que genera sesgos en los coeficientes de asociación y de correlación [6][13]. Esta manera de proceder significa trabajar únicamente con las observaciones que disponen de información completa para todas las variables, tal como se observa en la Figura 2.



Folio	Sexo	Edad	Escolaridad	Salario	Ocupación	Ponderación
1	Mujer	40	16	4 500	?	50
2	Hombre	35	15	?	1	75
3	Mujer	65	?	1 200	1	100
4	Hombre	23	12	2 200	2	80
5	Hombre	25	?	?	3	250
6	Mujer	38	15	1 800	4	140
....						
4	Hombre Mujer	23 38	12 15	2 200 1 800	2 4	80 140

Figura 2. Análisis con datos completos (*Listwise*). *Fuente:* [14]

Al eliminar información se asume que la submuestra de datos excluidos tiene las mismas características que los datos completos, y que la falta de respuesta se generó de manera aleatoria lo cual en la mayoría de las situaciones prácticas no se cumple.

Cuando los datos analizados provienen de una muestra probabilística, eliminar observaciones no es correcto ya que se debe tener presente que las unidades fueron elegidas con un procedimiento aleatorio y con probabilidad de selección, conocida y distinta de cero, que no puede ser ignorada en el tratamiento de los datos ni en el cálculo de los estimadores y sus errores.

Si la eliminación de registros no se acompaña con el ajuste apropiado de los factores de expansión, los valores estimados por la muestra pueden ser incompatibles con los parámetros observados en la población de referencia. Es decir, se obtendrán estimadores sesgados de los parámetros poblaciones lo que podría invalidar las conclusiones.

2.1.2 Análisis con los datos disponibles (*Pairwise Deletion*)

El método de "*Pairwise Deletion*", también conocido como "*Available Case*", intenta mejorar, de algún modo, los problemas que se producen utilizando el método *listwise deletion*.

En la Figura 3 se observa información completa para distintos registros de las variables Salario y Escolaridad, por lo que es posible calcular la correlación entre ambas utilizando los folios 1, 4 y 6, en tanto que la relación entre el Salario y la Ocupación se podría determinar con los datos de los registros 1, 3, 4 y 6. Sin embargo, debido a la diferencia en el tamaño de muestra, no es posible comparar los valores de los coeficientes obtenidos por ambos procedimientos. Este método hace uso de toda la información disponible sin efectuar ningún tipo de corrección en los factores de expansión. Las observaciones que no tienen datos se eliminan, y los cálculos se realizan con diferentes tamaños de muestra lo que limita comparación de resultados.

Folio	Sexo	Edad	Escolaridad	Salario	Ocupación	Ponderación
1	Mujer	40	16	4 500	2	50
2	Hombre	35	15	?	1	75
3	Mujer	65	?	1 200	1	100
4	Hombre	23	12	2 200	2	80
5	Hombre	25	?	?	3	250
6	Mujer	38	15	1 800	4	140
....						

Figura 3. Análisis con datos disponibles (*Pairwise*). *Fuente:* [14]

2.1.3 Imputación de la media

Calcular la media, mediana o moda general es un método de imputación muy básico, es la única función probada que no aprovecha las características de la serie temporal o la relación entre las variables. Es muy rápida, pero tiene claras desventajas. Una desventaja es que la imputación de la media reduce la varianza en el conjunto de datos.

Este método implica reemplazar los valores perdidos de una variable individual con su media estimada general de los casos disponibles. Si bien este es un método simple y fácil de implementar para lidiar con los valores perdidos, tiene algunas consecuencias desafortunadas. El problema más importante con la imputación de la media, también llamada sustitución de la media, es que dará como resultado una reducción artificial de la variabilidad debido al hecho de que está imputando valores en el centro de la distribución de la variable. Esto también tiene la

consecuencia no deseada de cambiar la magnitud de las correlaciones entre la variable imputada y otras variables.

2.1.4 Imputación por regresión

Este método consiste en estimar los valores ausentes en base a su relación con otras variables mediante análisis de regresión, es decir, sirve para imputar información en la variable **Y**, a partir de un grupo de covariables ($\mathbf{X_1}$,$\mathbf{X_2}$,…, $\mathbf{X_p}$) correlacionadas. El procedimiento consiste en eliminar las observaciones con datos incompletos, y ajustar una ecuación de regresión para predecir los valores de $\hat{\mathbf{y}}$ que serán utilizados para sustituir los valores que faltan, de modo que el valor de $\hat{\mathbf{y}}$ se construye como una media condicionada de las covariables X's.

Las desventajas que presenta este método son las siguientes:

- Incrementa artificialmente las relaciones entre variables. Debido a que los valores reemplazados se predijeron a partir de otras variables, tienden a encajar "demasiado bien" y, por lo tanto, el error estándar se desinfla.
- Subestima la varianza de las distribuciones.
- Asume que las variables con datos ausentes tienen relación de alta magnitud con las otras variables, es decir, se debe suponer que existe una relación lineal entre las variables utilizadas en la ecuación de regresión cuando puede que no la haya.

2.1.5 Imputación *Hot-Deck*

Es un procedimiento imputación de datos no paramétrico a través de una distribución no condicionada, que consiste en ubicar registros completos (donantes) e incompletos (receptores) e identificar las características comunes entre donantes y receptores, consecutivamente se hace la imputación entre observaciones con características comunes, la selección de los donantes se realiza en forma aleatoria, así la suposición se basa en que dentro de cada grupo de clasificación los valores faltantes siguen la misma distribución como aquellos que no son faltantes. Si bien, la ventaja de este método es evitar que se introduzcan sesgos en el estimador de la varianza preservando las distribuciones conjuntas y marginales [15]. Podríamos mencionar como desventajas:

- Cuando hay muchos datos faltantes se reporta muchas veces el mismo valor duplicado.
- Distorsiona la relación con el resto de las variables.

- Carece de un mecanismo de probabilidad.
- Requiere tomar decisiones subjetivas que afectan a la calidad de los datos, lo que imposibilita calcular su confianza.
- Proporciona estimaciones sesgadas de la media.

2.1.6 Imputación Simple

Este método imputa valores a partir de un modelo de regresión. El algoritmo genera sólo una simulación y asigna un valor por cada valor faltante, tomando como referencia el valor de otras variables, generando así, una base de datos completa.

Se requiere especificar un modelo en donde las covariables estén altamente correlacionadas con la variable a imputar.

La ventaja de este método es que utiliza procedimientos estadísticamente robustos. Sin embrago, ocurren situaciones en que los supuestos del método no se cumplen. No es posible conocer el error estándar de los estimadores, ya que sólo efectúa una iteración.

2.1.7 Método de Imputación Múltiple

A diferencia de la imputación única, la IM crea múltiples copias del conjunto de datos, en las que un algoritmo imputa los datos faltantes en función de los datos disponibles, con diferentes estimaciones en cada copia del conjunto de datos.

La IM utiliza métodos de simulación de Monte Carlo y sustituye los datos faltantes a partir de un número ($m>1$) de simulaciones que, según Rubin, se ubica entre 3 y 10. La metodología consta de varias etapas, y en cada simulación se analizan la matriz de datos completos a partir de métodos estadísticos convencionales y posteriormente se combinan los resultados para generar estimadores robustos, su error estándar e intervalos de confianza. En la Figura 4 se esquematiza la metodología propuesta por Rubin [16].

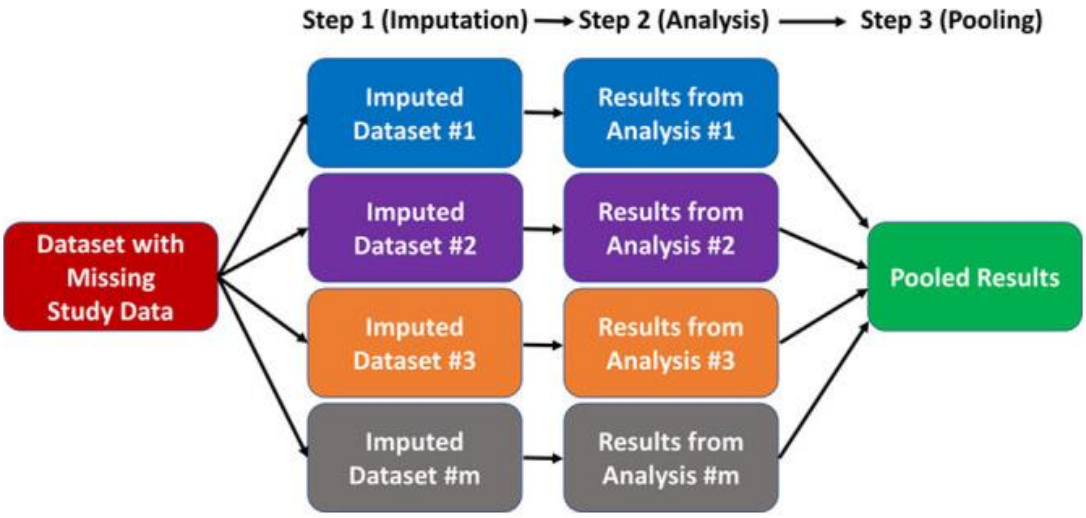


Figura 4: Pasos involucrados en la imputación múltiple de datos. *Fuente:* [17]

En la Figura anterior, se observa un resumen esquemático de los tres pasos involucrados en la IM de datos. En el paso 1, se crean múltiples conjuntos de datos (números 1, 2, 3… m), cada uno con diferentes estimaciones de los datos faltantes. En el paso 2, se analiza cada conjunto de datos imputados. En el paso 3, los resultados obtenidos en el paso 2 se combinan para obtener una estimación general.

Se afirma que el método IM es capaz de generar resultados robustos con un número pequeño de iteraciones. De acuerdo con Rubin [16], la eficiencia relativa de un estimador basado en *m* imputaciones comparada con la eficiencia obtenida con infinitas imputaciones viene dada por:

$$e = (1 + \lambda/m)^{-1}$$

donde λ es la proporción de información perdida de la variable de interés y *m* es el número de imputaciones. Rubin [18] argumenta que un valor pequeño de *m* es apropiado porque las simulaciones envueltas en la IM sólo tienen como objeto controlar las fracciones de información perdida, mientras que la información observada es manejada por los métodos de análisis utilizados con la base de datos completa. También señala, que para tasas de respuesta inusualmente altas sólo se requiere generar entre 5 y 10 imputaciones.

Un procedimiento que compite con los métodos de IM es el procedimiento de máxima verosimilitud (MV) que utiliza el algoritmo EM. Ambas propuestas aplican métodos numéricos y

simulaciones de Monte Carlo, y se demuestra que para tamaños de muestra grandes generan resultados similares. No obstante, también se ha comprobado que en el caso de muestras pequeñas la técnica IM produce resultados más robustos que el método de MV [19].

2.1.8 Método de Máxima Verosimilitud

Este método utiliza el algoritmo EM (*Expectation Maximization*) basado en la función de máxima verosimilitud, permite obtener estimaciones según la máxima verosimilitud de los parámetros cuando hay datos incompletos con unas estructuras determinadas. Resuelve de forma iterativa el cálculo del estimador máximo verosímil mediante dos pasos en cada iteración [4]. Este algoritmo tiene la ventaja de que puede resolver un amplio rango de problemas, incluyendo problemas no usuales que surgen de la pérdida o data incompleta, como lo es la estimación de los componentes de la varianza.

2.1.9 Árboles de decisión

Se define un árbol de decisión como una estructura en forma de árbol en la que las ramas representan conjuntos de decisiones. Estas decisiones generan sucesivas reglas para la clasificación de un conjunto de datos en subgrupos disjuntos y exhaustivos. Las ramificaciones se realizan de forma recursiva hasta que se cumplen ciertos criterios de parada.

El objetivo de estos métodos es obtener individuos más homogéneos con respecto a la variable que se desea discriminar dentro de cada subgrupo y heterogéneos entre los subgrupos. Para la construcción del árbol se requiere información de variables explicativas a partir de las cuales se va a realizar la discriminación de la población en subgrupos.

Dentro de los métodos basados en árboles se pueden distinguir dos tipos dependiendo de tipo de variable a discriminar:

- Árboles de clasificación: este tipo de árboles se emplea para variables categóricas, tanto nominales como ordinales.
- Árboles de regresión: este tipo de discriminación se aplica a variables continuas.

Teniendo en cuenta el tipo de variable con que estamos trabajando se calculan distintas medidas para el estudio de la homogeneidad. En todos los casos las variables explicativas son tratadas como variables categóricas. En particular en el caso de tener una variable explicativa continua,

salvo que haya sido categorizada previamente, será tratada como una variable categórica con el número de clases igual al número de valores distintos de la variable en el fichero de datos. Por esta razón el conjunto de datos requiere ser tratado previamente.

2.1.10 K-vecinos más cercanos (*k*-NN)

El método requiere la selección del número de vecinos más cercanos y una métrica de distancia. *k*-NN puede predecir atributos discretos (el valor más frecuente entre los k vecinos más cercanos) y atributos continuos (la media entre los k vecinos más cercanos). La métrica de distancia varía según el tipo de datos:

- Datos continuos: la métrica de distancia comúnmente utilizada para datos continuos es Euclidiana, y su promedio se usa como una estimación de imputación.

  Se define la distancia euclidiana D (a,b) entre los puntos a y b, ubicados sobre una recta, como la raíz cuadrada del cuadrado de las diferencias de sus coordenadas:

$$D(a,b) = \sqrt{\sum_{i=1}^{N}(b_i - a_i)^2}$$

- Datos categóricos: la distancia de *Hamming* se utiliza generalmente en este caso. Toma todos los atributos categóricos y para cada uno, cuenta uno si el valor no es el mismo entre dos puntos. La distancia de *Hamming* es entonces igual al número de atributos para los que el valor fue diferente.

Una de las características más atractivas del algoritmo *k*-NN es que es simple de entender y fácil de implementar. La naturaleza no paramétrica de *k*-NN le da una ventaja en ciertos entornos donde los datos pueden ser muy "inusuales". Uno de los inconvenientes obvios del algoritmo *k*-NN es que consume mucho tiempo al analizar grandes conjuntos de datos porque busca instancias similares en todo el conjunto de datos. Además, la precisión de *k*-NN puede verse seriamente degradada con datos de alta dimensión porque hay poca diferencia entre el vecino más cercano y el más lejano [20].

2.1.11 Red Bayesiana

Usar redes bayesianas para la imputación ofrece varias ventajas:

- La capacidad de manejar modelos de datos faltantes codifica dependencias entre todas las variables.

- Conserva la distribución de probabilidad conjunta de las variables que los métodos *k*-NN no prometen.

Desafortunadamente, este método no puede permitirse el lujo de admitir un gran tamaño de conjunto de datos ya que requiere aprender una red y discretización de todos los datos [11]. El enfoque bayesiano se basa en la recopilación de datos y luego calcula la probabilidad que los datos están significativamente relacionados con la información que fue extraída.

Los ingredientes clave de un análisis bayesiano son la función de verosimilitud, que refleja información sobre los parámetros contenidos en los datos, y la distribución previa, que cuantifica lo que se conoce sobre los parámetros antes de observar los datos. La distribución anterior y la probabilidad se pueden combinar fácilmente a partir de la distribución posterior, que representa el conocimiento total sobre los parámetros después de que se hayan observado los datos [21].

2.2 Algoritmos de Regresión

En este trabajo se analizan tres algoritmos de regresión diferentes, para evaluar en cada uno de ellos el porcentaje de mejora en la predicción tras aplicar distintos métodos de imputación de valores perdidos, cada uno de ellos de manera independiente. Estos son:

Regresión Lineal Múltiple:

La regresión lineal como forma de aprendizaje supervisado tiene como objetivo predecir valores continuos a partir de datos históricos etiquetados. Para aprender este tipo de modelos es necesario establecer la relación entre un cierto número de características X continuas y una variable objetivo Y continua, donde el comportamiento de una variable dependiente Y, se puede explicar a través de al menos una variable independiente X, lo que representamos mediante una recta $Y = f(X)$ [22][23], o sea, una o varias variable(s) dependientes pueden ser escritas en términos de una combinación lineal de las variables independientes [22]. Según sea el número de variables independientes xi estamos en presencia de una regresión lineal simple o múltiple.

La regresión lineal múltiple introduce más de una variable independiente X para predecir el comportamiento de la variable dependiente Y. Se fija la variable que se quiere predecir Y y se determina la relación con el resto de variables predictoras (independientes), por lo que tenemos un hiperplano de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e_i$$

donde x_n = es la enésima variable independiente, y_i = la variable dependiente, e_i = es el error observado de y_i de la línea $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$ para la i-ésima observación en la muestra, β_0 = el coeficiente de intersección (corte con el eje y), β_n = coeficiente de regresión de la enésima variable independiente *X*.

A modo de ejemplo, la Figura 5 muestra una regresión lineal múltiple con dos predictores **X_1 y X_2**, y una respuesta **Y**, donde la línea de regresión de mínimos cuadrados se convierte en un plano.

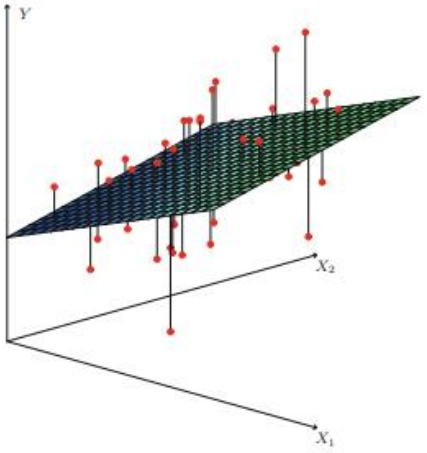


Figura 5: Regresión lineal múltiple. *Fuente:* [24]

Cuando desea ajustarse un modelo lineal a un conjunto de datos, el método de regresión usualmente más empleado es el de mínimos cuadrados. Este método es óptimo si la distribución de los residuos es gaussiana. Existen casos donde el supuesto de normalidad en los residuales no se cumple y se hace necesario el uso de métodos alternativos de regresión, como los de regresión no paramétrica, donde no se asume ninguna forma predefinida de la función de regresión y permiten obtener una mejor estimación de los parámetros del modelo, por ejemplo, SVM y *k*-NN.

SVM:

SVM se considera uno de los métodos de aprendizaje automático que introdujo originalmente Vapnik [25] basado en la teoría del aprendizaje estadístico y minimización de riesgos [26]. SVM es una de las técnicas supervisadas más populares de algoritmos de aprendizaje, que se pueden utilizar tanto para desafíos de regresión como de clasificación. El funcionamiento del algoritmo SVM se basa en encontrar el hiperplano óptimo para discriminar entre diferentes clases. Este hiperplano se conoce como kernel. En SVM, el parámetro de penalización C y el parámetro σ de

la función de base radial (RBF) tienen un impacto significativo en la complejidad y el rendimiento de SVM. Por lo general, estos parámetros se eligen al azar. Sin embargo, SVM es muy necesario para determinar los valores de parámetros óptimos para obtener el rendimiento de aprendizaje esperado.

El conjunto de valores de parámetros debe ser estipulado por el usuario antes de iniciar [27]. La elección de tales parámetros tiene un gran efecto en los resultados obtenidos y en el rendimiento del clasificador [28]. Necesitamos encontrar los parámetros que minimicen el error de generalización del algoritmo en cuestión.

La Figura 6 muestra un clasificador SVM ajustado a un pequeño set de observaciones. La línea continua representa el hiperplano y las líneas discontinuas el margen a cada lado. Las observaciones 2, 3, 4, 5, 6, 7, 9 y 10 se encuentran en el lado correcto del margen (también del hiperplano) por lo que están bien clasificadas. Las observaciones 1 y 8, a pesar de que se encuentran dentro del margen, están en el lado correcto del hiperplano, por lo que también están bien clasificadas. Las observaciones 11 y 12, se encuentran en el lado erróneo del hiperplano, su clasificación es incorrecta. Todas aquellas observaciones que, estando dentro o fuera del margen, se encuentren en el lado incorrecto del hiperplano, se corresponden con observaciones de entrenamiento mal clasificadas.

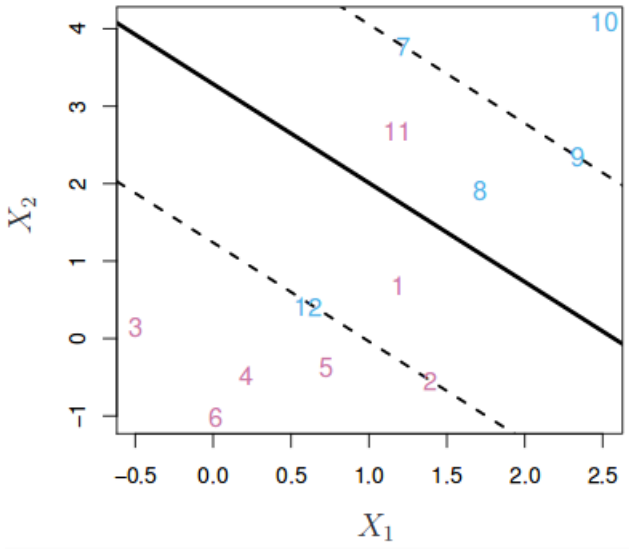


Figura 6: Algoritmo SVM. *Fuente:* [24]

Siendo C un parámetro de ajuste no negativo, es el que controla el número y severidad de las violaciones del margen (y del hiperplano) que se toleran en el proceso de ajuste. Si C = ∞, no se permite ninguna violación del margen y por lo tanto, el resultado es equivalente al Clasificador de Margen Máximo (solo es posible su aplicación si las clases son perfecta y linealmente separables). Cuando más se aproxima C a cero, menos se penalizan los errores y más observaciones pueden estar en el lado incorrecto del margen o incluso del hiperplano. C es a fin de cuentas el hiperparámetro encargado de controlar el balance entre bias y varianza del modelo.
El proceso de optimización tiene la peculiaridad de que solo las observaciones que se encuentran justo en el margen o que lo violan influyen sobre el hiperplano. A estas observaciones se les conoce como vectores soporte y son las que definen el clasificador obtenido. Esta es la razón por la que el parámetro C controla el balance entre bias y varianza. Cuando el valor de C es pequeño, el margen es más ancho, y más observaciones violan el margen, convirtiéndose en vectores soporte. El hiperplano está, por lo tanto, sustentado por más observaciones, lo que aumenta el bias pero reduce la varianza. Cuando mayor es el valor de C, menor el margen, menos observaciones serán vectores soporte y el clasificador resultante tendrá menor bias pero mayor varianza.

k-NN:

K-Vecinos más Cercanos (*k*-NN) es un algoritmo de aprendizaje supervisado (ya que a partir de un juego de datos inicial su objetivo será el de clasificar correctamente todas las instancias nuevas), de tipo no paramétrico y de aprendizaje perezoso (*Lazy Learning*). Lo que significa que no hace suposiciones sobre la distribución del conjunto de datos y no aprende explícitamente un modelo, simplemente memoriza las instancias de entrenamiento que se utilizan como "conocimiento" para la fase de predicción.
La simplicidad del algoritmo permite su utilización para realizar regresión (la tarea de pronóstico se puede plantear como un problema de regresión). El algoritmo asume que los datos se encuentran en un espacio de características y que los puntos de datos se pueden ubicar en un espacio métrico. Los datos pueden ser escalares o vectores multidimensionales, pero deben tener una noción de distancia; la métrica de la distancia Euclídea es la más comúnmente utilizada:

$$D(a,b) = \sqrt{\sum_{i=1}^{N}(b_i - a_i)^2}.$$

Sea la matriz $X = [x_1^T\ x_2^T\ ...\ x_N^T\ ]$ de valores de las variables independientes o regresores y el vector $Y = [y_1\ y_2\ ...\ y_N]$ de las variables dependientes. Dado un nuevo valor x_k , se seleccionan los K puntos de la matriz X que están más cerca de x_k y la predicción de $\hat{y}_k$ será una media ponderada de las *k*-salidas correspondientes a los *k*-vecinos.

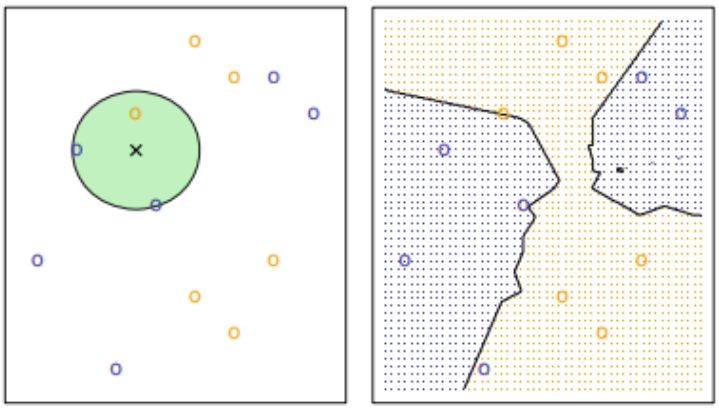


Figura 7: El enfoque *k*-NN en clasificación. *Fuente:* [24]

La Figura 7 muestra cómo funciona el algoritmo *k*-NN en clasificación. El gráfico de la izquierda, muestra un pequeño conjunto de datos de entrenamiento que consta de seis observaciones azules y seis naranjas. El objetivo es hacer una predicción para el punto marcado por la cruz negra. Si se toma K = 3. Entonces *k*-NN identificará primero las tres observaciones más cercanas a la cruz. Este vecindario se muestra como un círculo. Consta de dos puntos azules y un punto naranja, lo que da como resultado probabilidades estimadas de 2/3 para la clase azul y 1/3 para la clase naranja. Por lo tanto, *k*-NN predecirá que la cruz negra pertenece a la clase azul. En el gráfico de la derecha, se aplica el enfoque *k*-NN con *k* = 3 en todos los valores posibles para X1 y X2, y se ha dibujado el límite de decisión *k*-NN correspondiente. Hemos mostrado aquí una representación espacial del algoritmo *k*-NN en clasificación, a pesar de que el núcleo central de este TFM es regresión, debido a que visualmente puede ser más fácil entender dicho método en clasificación.

2.3 Métricas de bondad de ajuste

2.3.1 Error cuadrático medio (MSE):

El error cuadrático medio (MSE, *Mean Squared Error*) es una forma de evaluar la diferencia entre un estimador y el valor real de la cantidad que se quiere calcular. El MSE mide el promedio

al cuadrado del “error”, siendo el error el valor en la que el estimador difiere de la cantidad a ser estimada. Para cada punto, calcula la diferencia cuadrada entre las predicciones y el valor objetivo (real) y luego promedia esos valores.

Es quizás la métrica más simple y común para la evaluación de regresión, pero también es probablemente la menos útil. Se define por la ecuación:

$$MSE = \frac{\sum_{i=1}^{N}(yi - \hat{y}i)^2}{N}$$

donde y_i es el resultado real esperado y $\hat{y}_i$ es la predicción del modelo.

Cuanto mayor sea este valor, peor es el modelo. Nunca es negativo, ya que estamos cuadrando los errores de predicción individuales antes de sumarlos, pero sería cero para un modelo perfecto.

2.3.2 Raíz del error cuadrático medio (RMSE):

RMSE es la raíz cuadrada de MSE, se puede interpretar como la desviación estándar de la varianza inexplicada. La raíz cuadrada se introduce para hacer que la escala de los errores sea igual a la escala de los objetivos.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(yi - \hat{y}i)^2}{N}} = \sqrt{MSE}$$

Son similares en términos de sus minimizadores, cada minimizador de MSE es también un minimizador para RMSE y viceversa, ya que la raíz cuadrada es una función que no disminuye. Por ejemplo, si tenemos dos conjuntos de predicciones, A y B, y decimos que el MSE de A es mayor que el MSE de B, entonces podemos estar seguros de que RMSE de A es mayor que RMSE de B. Y también funciona en la dirección opuesta.

$$MSE(a) > MSE\ (b) \leftrightarrow RMSE(a) > RMSE(b)$$

Significa que, si la métrica objetivo es RMSE, aún podemos comparar nuestros modelos utilizando MSE, ya que MSE ordenará los modelos de la misma manera que RMSE.

2.3.3 Error absoluto medio (MAE):

En MAE, el error se calcula como un promedio de diferencias absolutas entre los valores objetivo y las predicciones. El MAE es una puntuación lineal, lo que significa que **todas las diferencias individuales se ponderan por igual** en el promedio. Por ejemplo, la diferencia entre 10 y 0 será el doble de la diferencia entre 5 y 0. Sin embargo, lo mismo no es cierto para RMSE. Matemáticamente, se calcula utilizando esta fórmula:

$$MAE = \frac{\sum_{i=1}^{N} \left| yi - \hat{y}i \right|}{N}$$

Lo importante de esta métrica es que **penaliza errores enormes no tan mal como lo hace MSE.** Por lo tanto, no es tan sensible a los valores atípicos como el MSE.

2.3.4 Coeficiente de determinación o R al cuadrado (R^2):

Se entiende como una versión estandarizada del MSE (al ser independiente de la escala de y), que proporciona una mejor interpretación del rendimiento del modelo. Si definimos RSS como RSS = $\sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$ (la suma total de las diferencias entre un valor predicho y un valor observado o conocido, al cuadrado) y TSS como $\sum_{i=1}^{N} (y_i - \tilde{y}_i)^2$ (la suma total de los cuadrados), donde $\tilde{y}$ es la media de todos los valores de y observados; técnicamente, el R^2 representa la varianza de la respuesta capturada por el modelo:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

TSS mide la varianza total en la respuesta Y, y puede ser considerado como la cantidad de variabilidad inherente a la respuesta antes de realizar la regresión. Por el contrario, RSS mide la cantidad de variabilidad que queda sin explicación después de realizar la regresión. Por lo tanto, TSS – RSS mide la cantidad de variabilidad en la respuesta que se ha explicado (o eliminado) realizando la regresión, y R^2 mide la proporción de variabilidad en Y que puede ser explicada mediante X. Esto lleva muchas veces a definir la ecuación anterior como sigue:

$$R^2 = 1 - \frac{Varianza\ final}{Varianza\ inicial}$$

La palabra inicial y final implica la aplicación o no de la regresión.

El mejor resultado posible es 1, y ocurre cuando la predicción coincide con los valores de la variable objetivo, indica que una gran proporción de la variabilidad en la respuesta ha sido explicada por la regresión. R^2 puede tomar valores negativos pues la predicción puede ser arbitrariamente mala, pero cuando la predicción coincide con la esperanza de los valores de la variable objetivo, el resultado de R^2 es 0.

De acuerdo a Young [22], R^2 es una medida de evaluación muy popular pero no debe utilizarse únicamente para evaluar la idoneidad de un modelo de regresión sin una justificación adicional, por ejemplo:

1. El valor de R^2 es muy sensible al tamaño de la muestra y aumenta al agregar más predictores al modelo. También puede provocar un aumento del MSE para tamaños de muestra pequeños.
2. R^2 está influenciado por el rango de los predictores, en el sentido de que, si el rango de X aumenta o disminuye, R^2 aumenta o disminuye, respectivamente.
3. La magnitud de las pendientes no se mide por R^2.
4. Solo mide la fuerza del componente lineal de un modelo.
5. Un nivel alto o bajo de R^2 no necesariamente indica la previsibilidad del modelo.

En el presente trabajo fin de máster se parte de la premisa de que se puede presentar la problemática de la presencia de valores perdidos en los conjuntos de datos asociados a técnicas de regresión de aprendizaje automático. Por tanto, el objetivo es llevar a cabo la imputación de valores perdidos y con ello poder disponer de registros completos e información lo más fiel posible a los datos originales, que por algún motivo han podido no almacenarse, extraerse o captarse correctamente y adicionalmente, intentar mejorar la predicción.

3 Análisis de los Resultados

3.1 Experimentación

En esta sección se describen, en primer lugar, los métodos de imputación utilizados en cada base de datos, señalando los comandos empleados de la herramienta STATA (versión MP 16.1) así como los valores de sus parámetros y, en un segundo lugar, los valores de los parámetros de los tres algoritmos empleados para evaluar las diferentes medidas de rendimiento obtenidas.

3.1.1 Métodos de imputación utilizados

La Tabla 1 muestra los comandos, así como los valores de sus parámetros de los seis métodos de imputación utilizados.

Método de imputación	**Comando Stata**	**Parámetro**	**Descripción**	**Valor**
Medias	mean	mean	Media	de la variable a imputar
Hot-Deck con regresión	hotdeck	imput	Número de imputaciones que se agregarán	5
		command	Análisis realizado en cada conjunto de datos imputado	reg
		parms	Parámetros de interés para el análisis	subconjunto de variables
k-NN	mi impute pmm	level	Nivel de confianza, predeterminado	95
		knn	Número de observaciones más cercanas para extraer	20
		add	Número de imputaciones que se agregarán	5
Por Regresión	impute	-	-	-
Simple	uvis	-	-	-
Múltiple - MICE	mi impute	add	Número de imputaciones que se agregarán	5

Tabla 1: Comandos y parámetros de cada método de imputación utilizado. *Fuente: Elaboración propia.*

3.1.2 Algoritmos de regresión utilizados

Los valores perdidos de las ocho bases de datos descriptas en el punto 2, fueron imputados con los diferentes métodos mencionados arriba, para luego ser evaluados a través del cálculo de las métricas de RMSE y R^2 en tres algoritmos de regresión, cuyos parámetros se describen a continuación:

Algoritmo	Parámetro	Descripción	Valor
Regresión Lineal Múltiple	level	Nivel de confianza, predeterminado	95
	vce	Tipo de error estándar informado, predeterminado	ols
SVM	C	Peso en el margen de error	100
	gamma	Factor de escala para la parte lineal del núcleo	1
	eps	Margen de error permitido	1
	type	Tipo de modelo para estimar	svr (Support Vector Regression)
k-NN	mlmodel	Algoritmo de aprendizaje automático a estimar	nearestneighbor
	seed	Semilla entera	10

Tabla 2: Valor de los parámetros utilizados en los algoritmos de regresión. *Fuente: Elaboración propia.*

3.1.3 Bases de Datos

Las bases de datos de regresión, con valores faltantes, que se analizan en este trabajo son: Auto93, Autompg, Cholesterol, Pharynx, Player_Performance, Schlvote, Sleep y Stroke, las cuales se han extraído de los repositorios UCI de aprendizaje automático, mantenido por la Universidad de California en Irvine [29][30].

La Tabla 3 muestra una descripción de la estructura de datos (básica) contenida en los conjuntos de datos mencionados previamente.

Base de datos	Instancias	Atributos	Número de atributos con valores perdidos	Variable a predecir
auto93	93	23	3	class
autompg	398	8	1	class
cholesterol	303	14	2	chol
pharynx	195	12	2	class
player_performance	1340	20	1	gamesplayed
schlvote	38	6	1	wealth_per_student
sleep	62	8	3	total_sleep
stroke	5110	18	2	bmi

Tabla 3: Estructura de datos básica de las bases de datos utilizadas. *Fuente: Elaboración propia.*

En la tabla anterior, la columna “Instancias” se refiere a la cantidad de observaciones de la base de datos. Luego la columna “Atributos” se refiere a la cantidad de variables; “Número atributos con valores perdidos” se refiere a la cantidad de atributos con valores faltantes en la base de datos; y por último la columna “Variable a predecir” indica cual es la variable que se predecirá con los diferentes algoritmos de regresión.

A todas y cada una de estas bases de datos se le aplicó la técnica de validación cruzada *hold-out*. Consiste en separar el conjunto de datos disponibles en dos subconjuntos, uno utilizado para entrenar el modelo, y otro para realizar el test de validación [31]. De esta manera, se crea un modelo únicamente con los datos de entrenamiento. Con el modelo creado se generan datos de salida que se comparan con el conjunto de datos reservados para realizar la validación (que no han sido utilizados en el entrenamiento, por lo que no han sido utilizados para generar el modelo). Los estadísticos obtenidos con los datos del subconjunto de validación son los que nos dan la validez del método empleado en términos de error [32].

Para el caso particular de este trabajo, el 75% de las observaciones de cada base de datos conforman el conjunto de entrenamiento, mientras que el 25% restante, el de test.

3.2 Resultados

Para cada una de las bases de datos se obtuvo el valor de las métricas RMSE y R^2 antes de imputarlas y después de aplicar cada método de imputación de manera independiente. Esto nos permitió calcular posteriormente, los porcentajes de mejora en las predicciones de las bases de datos imputadas, con respecto a las bases de datos originales, tanto para la regresión lineal múltiple como para SVM y *k*-NN. Estos resultados se ven reflejados en las tablas 4, 6 y 8, respectivamente.

En las tablas 5, 7 y 9 se pueden ver la cantidad de victorias, empates y derrotas de cada método de imputación, según sea el algoritmo de regresión aplicado. Se definen como victorias, las veces que el porcentaje de mejora (ya sea de RMSE y/o R^2) es mayor que cero, lo que implica que se han recuperado valores perdidos e incluso se obtuvo una mejor predicción de la base de datos; derrotas, son las veces que el porcentaje de mejora es menor que cero, mientras que empate existe cuando el mismo es igual a cero.

Base de datos / Método de imputación	Regresión lineal múltiple			
	RMSE	R^2	Mejora (%) RMSE	Mejora (%) R^2
auto93				
Sin imputar los valores	40,8700	0,8785	-	-
Imputación de la media	35,6800	0,8717	12,6988	-0,7740
Hot-deck con regresión	33,6900	0,8811	17,5679	0,2960
k-NN	34,3267	0,8682	16,0100	-1,1725
Por regresión	36,8847	0,8729	9,7512	-0,6375
Imputación Simple	36,7321	0,8719	10,1245	-0,7513
Imputación Múltiple - MICE	34,3267	0,8682	16,0100	-1,1725
autompg				
Sin imputar los valores	12,1634	0,8274	-	-
Imputación de la media	12,1703	0,8269	-0,0567	-0,0604
Hot-deck con regresión	11,9500	0,8269	1,7544	-0,0604
k-NN	12,3036	0,8250	-1,1526	-0,2901
Por regresión	12,1935	0,8268	-0,2475	-0,0725
Imputación Simple	12,1860	0,8268	-0,1858	-0,0725
Imputación Múltiple - MICE	12,3036	0,8250	-1,1526	-0,2901
cholesterol				
Sin imputar los valores	2285,4600	0,1203	-	-
Imputación de la media	2282,5741	0,1209	0,1263	0,4988
Hot-deck con regresión	2127,9781	0,1144	6,8906	-4,9044
k-NN	2215,6033	0,1166	3,0566	-3,0756
Por regresión	2281,9202	0,1210	0,1549	0,5819
Imputación Simple	2285,4684	0,1203	-0,0004	0,0000
Imputación Múltiple - MICE	2215,6033	0,1166	3,0566	-3,0756
pharynx				
Sin imputar los valores	77025,0278	0,5495	-	-
Imputación de la media	76521,7138	0,5503	0,6534	0,1456
Hot-deck con regresión	90692,8370	0,5187	-17,7446	-5,6051
k-NN	96448,6100	0,5162	-25,2172	-6,0601
Por regresión	76594,6251	0,5506	0,5588	0,2002
Imputación Simple	77025,0270	0,5495	0,0000	0,0000
Imputación Múltiple - MICE	96448,6100	0,5162	-25,2172	-6,0601

player_performance				
Sin imputar los valores	176,4999	0,4626	-	-
Imputación de la media	176,2980	0,4620	0,1144	-0,1297
Hot-deck con regresión	176,2928	0,4638	0,1173	0,2594
k-NN	176,0976	0,4616	0,2279	-0,2162
Por regresión	176,3083	0,4619	0,1086	-0,1513
Imputación Simple	176,3187	0,4619	0,1027	-0,1513
Imputación Múltiple - MICE	176,0976	0,4446	0,2279	-3,8911
schlvote				
Sin imputar los valores	760741422995,5500	0,2476	-	-
Imputación de la media	785860877511,1100	0,2472	-3,3020	-0,1616
Hot-deck con regresión	743024525226,6660	0,3080	2,3289	24,3942
k-NN	815767146382,2220	0,2759	-7,2332	11,4297
Por regresión	760741422995,5500	0,2476	0,0000	0,0000
Imputación Simple	760741422995,5500	0,2476	0,0000	0,0000
Imputación Múltiple - MICE	696119677795,5500	0,2761	8,4946	11,5105
sleep				
Sin imputar los valores	3,9271	0,6221	-	-
Imputación de la media	4,2642	0,5274	-8,5839	-15,2226
Hot-deck con regresión	16,3367	0,6903	-315,9991	10,9629
k-NN	8,6153	0,5134	-119,3807	-17,4731
Por regresión	4,5393	0,5908	-15,5891	-5,0313
Imputación Simple	4,7860	0,5341	-21,8711	-14,1456
Imputación Múltiple - MICE	8,5572	0,5241	-117,9013	-15,7531
stroke				
Sin imputar los valores	41,1103	0,2314	-	-
Imputación de la media	41,1642	0,2241	-0,1311	-3,1547
Hot-deck con regresión	40,0237	0,2321	2,6431	0,3025
k-NN	41,1296	0,2181	-0,0469	-5,7476
Por regresión	41,0878	0,2336	0,0547	0,9507
Imputación Simple	41,2044	0,2197	-0,2289	-5,0562
Imputación Múltiple - MICE	41,2112	0,2144	-0,2454	-7,3466

Tabla 4: Resultados obtenidos para regresión lineal múltiple. *Fuente: Elaboración propia.*

Método de Imputación		s/ RMSE	s/ R^2
Imputación de la media	W	4	2
	T	0	0
	L	4	6
Hot-deck con regresión	W	6	5
	T	0	0
	L	2	3
k-NN	W	3	1
	T	0	0
	L	5	7
Por regresión	W	5	3
	T	1	1
	L	2	4
Imputación Simple	W	2	0
	T	3	3
	L	3	5
Imputación Múltiple - MICE	W	4	1
	T	0	0
	L	4	7

Notas: s/: Según, W: *win* (victoria), T: *tie* (empate), L: *loss* (derrota).

Tabla 5: Resumen de resultados por cada método de imputación para regresión lineal múltiple. *Fuente: Elaboración propia.*

Como se desprende de la Tabla 5, para el algoritmo de Regresión Lineal Múltiple:

Tanto para el método de Imputación de la media, como el de Imputación múltiple – MICE, se consigue recuperar valores perdidos sin empeorar el rendimiento de RMSE. Con dichos métodos el coeficiente de determinación rara vez mejora.

En el método de *Hot–Deck* con regresión se observa tanto que RMSE como R^2 se consiguen resultados bastantes aceptables. Para el algoritmo de regresión múltiple, éste es el mejor método de imputación, siempre que gana con RMSE también gana con R^2, se observa una asociación en términos de rendimiento.

En la Imputación *k*-NN se puede ver que, si se recuperan los valores perdidos, el rendimiento no es nada satisfactorio, desde el punto de vista de las dos métricas disponibles.

En cuanto a la Imputación por Regresión se observa que respecto a RMSE los resultados son aceptables, no se puede decir lo mismo con respecto a R^2.

En el método de Imputación Simple, no logramos conseguir mejorías con respecto a RSME y con respecto a R^2 estamos bastante lejos de conseguir buenos resultados.

Esto puede verse representado gráficamente en la Figura 8.

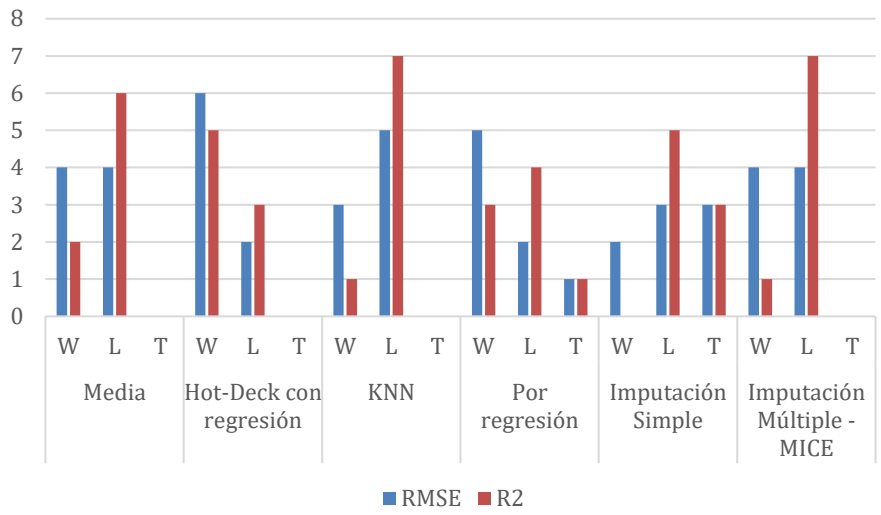


Figura 8: Mejora de la predicción en regresión lineal múltiple. *Fuente: Elaboración propia.*

Base de datos / Método de imputación	**SVM**			
	RMSE	R^2	Mejora (%) RMSE	Mejora (%) R^2
auto93				
Sin imputar los valores	67,6500	0,9915	-	-
Imputación de la media	67,2633	0,9906	0,5716	-0,0908
Hot-deck con regresión	67,6500	0,9915	0,0000	0,0000
k-NN	67,2633	0,9906	0,5716	-0,0908
Por regresión	67,2633	0,9906	0,5716	-0,0908
Imputación Simple	67,6500	0,9915	0,0000	0,0000
Imputación Múltiple - MICE	67,2633	0,9906	0,5716	-0,0908

autompg				
Sin imputar los valores	59,6167	0,9844	-	-
Imputación de la media	50,5658	0,9844	15,1818	-0,0041
Hot-deck con regresión	60,5795	0,9843	-1,6150	-0,0091
k-NN	59,5658	0,9843	0,0854	-0,0091
Por regresión	59,6167	0,9844	0,0000	0,0000
Imputación Simple	59,6167	0,9843	0,0000	-0,0091
Imputación Múltiple - MICE	59,5658	0,9844	0,0854	-0,0041
cholesterol				
Sin imputar los valores	2359,8864	0,9041	-	-
Imputación de la media	2353,2425	0,9996	0,2815	10,5630
Hot-deck con regresión	2258,2518	0,9996	4,3068	10,5630
k-NN	2352,2715	0,9996	0,3227	10,5630
Por regresión	2353,4244	0,9996	0,2738	10,5630
Imputación Simple	2351,9448	0,9996	0,3365	10,5630
Imputación Múltiple - MICE	2352,2715	0,9996	0,3227	10,5630
pharynx				
Sin imputar los valores	142275,6586	0,2370	-	-
Imputación de la media	142727,3903	0,2339	-0,3175	-1,3163
Hot-deck con regresión	142727,3903	0,2339	-0,3175	-1,3163
k-NN	142275,6749	0,2370	0,0000	0,0000
Por regresión	142275,6742	0,2370	0,0000	0,0000
Imputación Simple	142275,6586	0,2370	0,0000	0,0000
Imputación Múltiple - MICE	142275,6749	0,2370	0,0000	0,0000
player_performance				
Sin imputar los valores	274,6336	0,9968	-	-
Imputación de la media	274,6336	0,9968	0,0000	0,0000
Hot-deck con regresión	286,8841	0,9968	-4,4607	0,0000
k-NN	274,6413	0,9968	-0,0028	0,0000
Por regresión	274,6262	0,9968	0,0027	0,0000
Imputación Simple	274,6264	0,9968	0,0026	0,0000
Imputación Múltiple - MICE	274,6336	0,9968	0,0000	0,0000
schlvote				
Sin imputar los valores	114350740167,1110	0,1452	-	-
Imputación de la media	111388022670,2220	0,1345	2,5909	-7,3682
Hot-deck con regresión	118860064881,7770	0,1342	-3,9434	-7,5170
k-NN	114350740167,1110	0,1452	0,0000	0,0000
Por regresión	111388022670,2220	0,1315	2,5909	-9,4369
Imputación Simple	118860064881,7770	0,1315	-3,9434	-9,4369
Imputación Múltiple - MICE	114350740167,1110	0,1452	0,0000	0,0000

sleep				
Sin imputar los valores	25,4988	0,9578	-	-
Imputación de la media	24,4775	0,9592	4,0053	0,1462
Hot-deck con regresión	24,6206	0,9590	3,4441	0,1253
k-NN	24,4146	0,9587	4,2520	0,0940
Por regresión	24,7636	0,9591	2,8833	0,1357
Imputación Simple	24,9491	0,9590	2,1558	0,1253
Imputación Múltiple - MICE	25,4988	0,9578	0,0000	0,0000
stroke				
Sin imputar los valores	50,9952	0,9836	-	-
Imputación de la media	50,9635	0,9831	0,0622	-0,0508
Hot-deck con regresión	55,6380	0,9854	-9,1044	0,1830
k-NN	55,0925	0,9854	-8,0347	0,1830
Por regresión	50,9285	0,9834	0,1308	-0,0203
Imputación Simple	50,9323	0,9835	0,1233	-0,0102
Imputación Múltiple - MICE	55,0925	0,9854	-8,0347	0,1830

Tabla 6: Resultados obtenidos para SVM. *Fuente: Elaboración propia*.

Método de Imputación		s/ RMSE	s/ R^2
Imputación de la media	W	6	2
	T	1	1
	L	1	5
Hot-deck con regresión	W	2	3
	T	1	2
	L	5	3
k-NN	W	4	3
	T	3	3
	L	1	2
Por regresión	W	5	2
	T	3	3
	L	0	3
Imputación Simple	W	3	2
	T	4	3
	L	1	3
Imputación Múltiple - MICE	W	3	2
	T	4	4
	L	1	2

Notas: s/: Según, W: *win* (victoria), T: *tie* (empate), L: *loss* (derrota).

Tabla 7: Resumen de resultados por cada método de imputación para SVM. *Fuente: Elaboración propia.*

De la Tabla 7, se concluye que:

Desde el punto de vista de la métrica del RMSE, el método de Imputación de la media e Imputación por regresión funcionan bastante bien con el algoritmo SVM.

El método de Imputación simple, Imputación múltiple – MICE e Imputación por *k*-NN tienen un rendimiento aceptable, mientras que *Hot-Deck* con regresión es el que peor funciona para SVM.

Y respecto a la métrica R^2, se observa que en el método de la media no hay mejorías, siendo muy leve para el método de imputación *k*-NN. El método de *Hot-Deck* con regresión y MICE consiguen recuperar valores perdidos sin empeorar su rendimiento. Mientras que en el resto de los métodos (Imputación Simple y por Regresión) prácticamente se mantiene el valor de R^2.

Lo comentado arriba, se observa en la Figura 9:

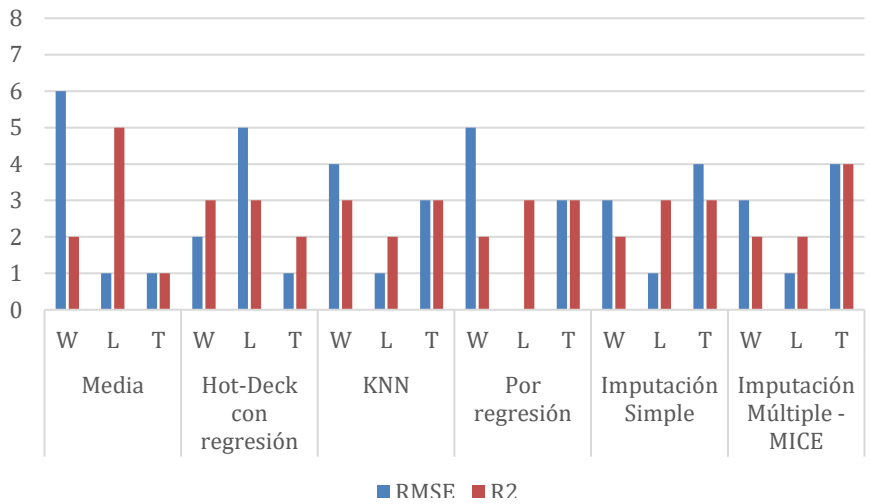


Figura 9: Mejora de la predicción en SVM. *Fuente: Elaboración propia.*

Base de datos / Método de imputación	***k*-NN**			
	RMSE	R^2	Mejora RMSE	Mejora R^2
auto93				
Sin imputar los valores	4,6228	0,8785	-	-
Imputación de la media	4,2819	0,8717	7,3743	-0,7740
Hot-deck con regresión	4,9167	0,8770	-6,3576	-0,1707
k-NN	4,4481	0,8733	3,7791	-0,5919
Por regresión	4,2662	0,8728	7,7139	-0,6488
Imputación Simple	4,2648	0,8727	7,7442	-0,6602
Imputación Múltiple - MICE	4,4090	0,8755	4,6249	-0,3415
autompg				
Sin imputar los valores	3,2850	0,8273	-	-
Imputación de la media	3,2740	0,8268	0,3349	-0,0604
Hot-deck con regresión	3,3069	0,8257	-0,6667	-0,1934
k-NN	3,2814	0,8284	0,1096	0,1330
Por regresión	3,2850	0,8273	0,0000	0,0000
Imputación Simple	3,2850	0,8273	0,0000	0,0000
Imputación Múltiple - MICE	3,2850	0,8273	0,0000	0,0000
cholesterol				
Sin imputar los valores	51,3105	0,1202	-	-
Imputación de la media	51,0644	0,1209	0,4796	0,5824
Hot-deck con regresión	51,1569	0,1260	0,2994	4,8253
k-NN	51,1900	0,1210	0,2348	0,6656
Por regresión	51,0609	0,1210	0,4865	0,6656
Imputación Simple	51,3105	0,1202	0,0000	0,0000
Imputación Múltiple - MICE	51,3105	0,1210	0,0000	0,6656
pharynx				
Sin imputar los valores	302,6517	0,5495	-	-
Imputación de la media	301,8091	0,5502	0,2784	0,1274
Hot-deck con regresión	313,3817	0,5187	-3,5453	-5,6051
k-NN	302,6517	0,5495	0,0000	0,0000
Por regresión	302,6517	0,5495	0,0000	0,0000
Imputación Simple	302,6517	0,5495	0,0000	0,0000
Imputación Múltiple - MICE	302,6517	0,5495	0,0000	0,0000

player_performance				
Sin imputar los valores	13,0710	0,4625	-	-
Imputación de la media	13,0426	0,4619	0,2173	-0,1297
Hot-deck con regresión	13,0563	0,4630	0,1125	0,1081
k-NN	13,0621	0,4625	0,0681	0,0000
Por regresión	13,0432	0,4619	0,2127	-0,1297
Imputación Simple	13,0442	0,4618	0,2050	-0,1514
Imputación Múltiple - MICE	13,0453	0,4639	0,1966	0,3027
schlvote				
Sin imputar los valores	1356974,7717	0,2475	-	-
Imputación de la media	1327610,1597	0,2472	2,1640	-0,1616
Hot-deck con regresión	1400699,1551	0,2711	-3,2222	9,4911
k-NN	1356974,7717	0,2475	0,0000	-0,0404
Por regresión	1327147,5080	0,2498	2,1981	0,8885
Imputación Simple	1327928,9811	0,2550	2,1405	2,9887
Imputación Múltiple - MICE	1356974,7717	0,2475	0,0000	-0,0404
sleep				
Sin imputar los valores	3,1080	0,6220	-	-
Imputación de la media	3,3428	0,5273	-7,5547	-15,2251
Hot-deck con regresión	3,4801	0,5750	-11,9723	-7,5563
k-NN	3,1080	0,6647	0,0000	6,8650
Por regresión	3,1493	0,5907	-1,3288	-5,0322
Imputación Simple	3,0890	0,6177	0,6113	-0,6913
Imputación Múltiple - MICE	3,1080	0,6220	0,0000	0,0000
stroke				
Sin imputar los valores	7,0048	0,2314	-	-
Imputación de la media	6,9032	0,2240	1,4504	-3,1979
Hot-deck con regresión	7,0390	0,2326	-0,4882	0,5186
k-NN	7,0048	0,2314	0,0000	0,0000
Por regresión	6,8831	0,2336	1,7374	0,9507
Imputación Simple	7,0390	0,2181	-0,4882	-5,7476
Imputación Múltiple - MICE	7,0048	0,2314	0,0000	0,0000

Tabla 8: Resultados obtenidos para *k*-NN. *Fuente: Elaboración propia.*

Método de Imputación		s/ RMSE	s/ R^2
Imputación de la media	W	7	2
	T	0	0
	L	1	6
Hot-deck con regresión	W	2	4
	T	0	0
	L	6	4
k-NN	W	4	3
	T	4	3
	L	0	2
Por regresión	W	5	3
	T	2	2
	L	1	3
Imputación Simple	W	4	1
	T	3	3
	L	1	4
Imputación Múltiple - MICE	W	2	2
	T	6	4
	L	0	2

Notas: s/: Según, W: *win* (victoria), T: *tie* (empate), L: *loss* (derrota).

Tabla 9: Resumen de resultados por cada método de imputación para *k*-NN. *Fuente: Elaboración propia.*

Como se desprende de la Tabla 9, en el algoritmo de regresión *k*-NN, se amplifica el rendimiento comparado con SVM, tanto en los mejores, como en los peores escenarios.

La imputación por *k*-NN, Imputación simple y múltiple tienen un comportamiento bastante similar al comentado en SVM.

Lo comentado, se observa gráficamente en la Figura 10.

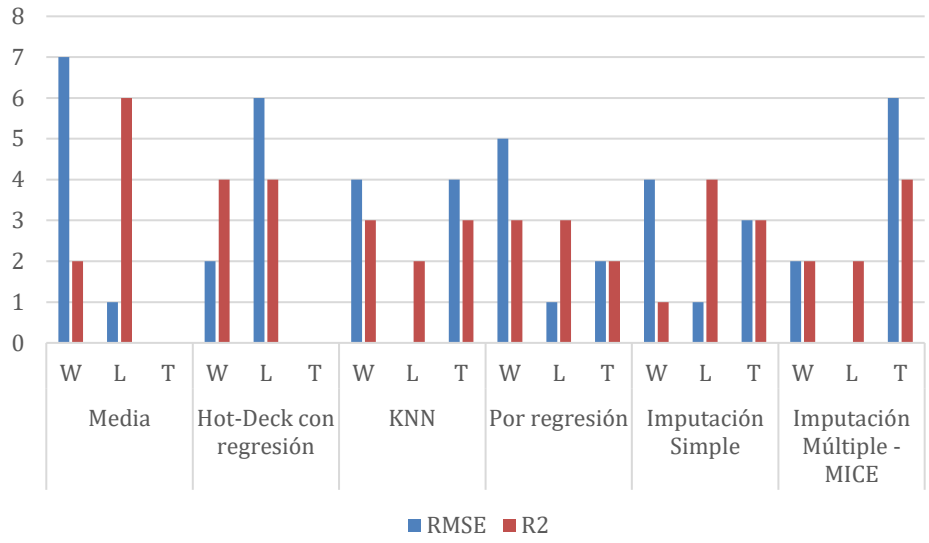


Figura 10: Mejora de la predicción en *k*-NN. *Fuente: Elaboración propia.*

4 Conclusiones

Se usaron seis métodos de imputación para completar ocho bases de datos con valores perdidos; dichos conjuntos de datos presentan diferentes niveles de valores perdidos. Se analizó el rendimiento de dichos métodos comparando tres algoritmos de aprendizaje automático, utilizando el RMSE y R^2 como métricas de análisis, donde se llegó a concluir que no existe un mejor método de imputación, se muestra que hay casos donde el que resulta ser mejor según RMSE, es el que menor porcentaje de mejora consigue según R^2.

Es posible afirmar que el mejor método de imputación es el que no se aplica, lo que sugiere agotar todos los recursos para minimizar la pérdida de valores. La técnica de imputación que genera los mejores resultados y preserva la verosimilitud de los datos, es la que se sustenta en información recabada en el terreno.

No es posible afirmar de forma categórica que el algoritmo de IM genera siempre mejores resultados que los métodos simples. Esto se puede observar si se comparan los porcentajes de mejora conseguidos en IM, frente a los conseguidos en la imputación de la media, siendo este último método de imputación el que logra mayor rendimiento en 2 de los 3 algoritmos analizados.

No existe un mejor método de imputación. Cada situación es diferente y la elección del procedimiento de sustitución de datos depende de la variable de estudio, del porcentaje de datos faltantes, del tipo de algoritmo a aplicar y del uso que se hará de la información imputada. Por ello, no se aconseja elegir un procedimiento de imputación y aplicarlo de forma generalizada para todas las variables de todas las bases de datos.

Se sugiere encarar el análisis datos sin la elección -a priori- de un método de imputación. El análisis exploratorio y la consistencia de la información, darán la pauta para elegir el método que genera los estimadores más eficientes. Lo que funcionó para una base de datos, no necesariamente generará buenos resultados en otras.

Se consiguió recuperar la información de todas las bases de datos, proporcionando registros completos que pueden ser usados en trabajos posteriores. Así mismo, esto permitirá poder usarlas con otros algoritmos o métodos que no sean capaces de trabajar con valores perdidos.

Referencias

[1] Abidin, Nueva Zelanda, Ismail, AR y Emran, NA (2018). Análisis de rendimiento de algoritmos de aprendizaje automático para la imputación de valores perdidos. *Revista Internacional de Aplicaciones y Ciencias Informáticas Avanzadas (IJACSA),* 9 (6), 442-447.

[2] Wilks, S. (1932). *Moments and distributions of estimates of population parameters from fragmentary simple,* Annals of Mathematical Statistics, B, 163-195.

[3] Rubin, D. (1983). *Panel of incomplete data in sample surveys.* En Madow, W.G., Olkin, I. y Rubin, D.B. Incomplete Data in Simple Surveys. Vol 2,12, 123-145. Report and Case Studies. New York. Academic Press.

[4] Little, R. y Rubin, D. (1987). *Statistical Analysis with Missing Data. Series in Probability and Mathematical Statistics*. John Wiley & Sons, Inc. New York.

[5] Goicoechea, P. (2002). *Imputación basada en árboles de clasificación.* EUSTAT.

[6] Hemel, J. y otros (1987). *Stepwise deletion: a technique for missing data handling in multivariate analysis.* Analytical Chemical Acta 193 255-268.

[7] Todeschini, R. (1990). *Weighted k-nearest neighbour method for the calculation of missing values.* Chenometrics and Intelligent Laboratory Systems 9.201-205.

[8] Mesa, D. Tsai, P. y Chambers, R. (2000). *Using Tree-Based Models For Missing Data Imputation: An Evaluation Using Uk Census Data,* Reporte Técnico. Proyecto AUTIMP. Recuperado 20, septiembre del 2004 en: http://www.cbs.nl/en/service/autimp/CART-Dutch%20Data-(AUTIMP).pdf.

[9] Geng, Z. y Li, K. (2003). *Factorization of posteriors and partial imputation algorithm for graphical models with missing data.* Statistics and probability letters. 64, 369-379

[10] Gleason, T. y Staelin, R. (1975). *A proposal for handling missing data.* Psychometrika. Vol 40, 2. 229-252.

[11] Liu, Yuzhe y Gopalakrishnan, Vanathi (2017). *An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data*, Data, 2(1):8.

[12] Agarwal Vivek, *Research on Data Preprocessing and Categorization Technique for Smartphone Review Analysis*, International Journal of Computer Applications, 131(4):30–36, 2015.

[13] Kalton, G. y Kasprzyk, D. (agosto de 1982). Imputación de respuestas faltantes a la encuesta. *En Actas de la sección sobre métodos de investigación de encuestas, Asociación Estadounidense de Estadístic*a Vol. 22, 31. Asociación Estadounidense de Estadística de Cincinnati.

[14] Medina, F., & Galván, M. (2007). *Imputación de datos: teoría y práctica*. Cepal.

[15] Ávila G, Carlos. (2002) *Una aplicación del procedimiento Hot Deck como método de imputación.* Capítulo III. Método de imputación hot deck. Trabajo Monográfico Para optar el Título Profesional de Licenciado en Estadística. Universidad Mayor de San Marcos, Perú.

[16] Rubin, D.B. (1987), *Multiple imputation for non-response in surveys.* New York, Wiley.

[17] Schober, P. y Vetter, TR (2020). Datos faltantes y métodos de imputación. Anestesia y analgesia , 131 (5), 1419-1420. https://doi.org/10.1213/ANE.0000000000005068

[18] Rubin, DB (1996). *Multiple imputation after 18+ years.* Journal of the American statistical Association, 91(434), 473-489.

[19] Schafer, JL (1999). *Multiple imputation: a primer.* Métodos estadísticos en la investigación médica, 8 (1), 3-15.

[20] Alvira Swalin (2018). *How to Handle Missing Data*. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

[21] Glickman, ME y Van Dyk, DA (2007). *Métodos bayesianos básicos.* Temas de bioestadística , 319-338.

[22] Young, D. S. (2018) *Manual de métodos de regresión.* Prensa CRC.

[23] Olive, D. J. (2017). Multiple linear regression. In *Linear regression* (pp. 17-83). Springer, Cham.

[24] James, G., Witten, D., Hastie, T. y Tibshirani, R. (2013). Una introducción al aprendizaje estadístico (Vol. 112, p. 18). Nueva York: springer.

[25] Vapnik, V.(1995). *The nature of statistical learning theory.* Informat. Sci. Stat. Springer, New York

[26] Lin, S., Ying, K., Chen, S., Lee, Z. (2008). *Evolutionary tuning of svm parameter values in multiclass problems*. Neurocomputing 71(4), 3326–3334

[27] Sayed, G., Ali, M., Gaber, T., Hassanien, A., Sansel, V. (2015). *Interphase cells removal from metaphase chromosome images based on meta-heuristic grey wolf optimizer*. In: 11th International Computer Engineering Conference (ICENCO). IEEE, 261–266. Egypt, Cairo.

[28] Muller, K.R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B. (2001). *An introduction to kernel-based learning algorithms.* IEEE Trans. Neural Netw. 12(2), 181–201

[29] Dua, D., & Graff, C. (2017). UCI machine learning repository.

[30] Dua, D., & Graff, C. (2019). UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, CA, 2019.

[31] Arlot, S. y Celisse, A. (2010). Un estudio de los procedimientos de validación cruzada para la selección de modelos. *Encuestas estadísticas*, 4 , 40-79.

[32] Hawkins, DM, Basak, SC y Mills, D. (2003). Evaluación del ajuste del modelo mediante validación cruzada. *Revista de información química y ciencias de la computación*, *43* (2), 579-586.