rics
Revista Ingeniería, Ciencias y Sociedad 05 (2023) 12–31

Análisis de datos de covid-19: imputación y rendimiento de modelos de aprendizaje supervisado

Covid-19 data analysis: imputing and performance of supervised learning models

● *Adrián Martínez Amarilla* [1]

Resumen

Este estudio investiga la eficacia de métodos de aprendizaje supervisado en la predicción del COVID-19 utilizando registros hospitalarios del departamento de Concepción durante el periodo 2020-2022. Se analiza el impacto de la imputación de datos faltantes en métricas de evaluación para varios modelos, incluyendo Máquinas de Vectores Soporte, Redes Neuronales Artificiales, Regresión Logística, Árbol de Decisión y Bosque Aleatorio. El preprocesamiento incluye la creación de dos conjuntos de datos: uno sin registros vacíos y otro con un 20% de datos faltantes por filas. La imputación se realiza mediante las técnicas de imputación por moda y bosque aleatorio en el conjunto con datos faltantes. La variable dependiente evaluada es la clasificación final de la enfermedad, confirmada o descartada por criterios laboratoriales. El modelo de Bosque Aleatorio destaca por su eficiencia superior en el conjunto de datos sin registros vacíos y muestra robustez ante la imputación de datos. Este estudio contribuye significativamente al proporcionar información sobre los efectos de la imputación de datos faltantes en el ámbito de la salud pública y su aplicación en la generación de modelos predictivos.

Palabras clave: *Análisis de datos, Imputación de valores perdidos, Modelos de aprendizaje supervisado.*

Abstract

This study conducts an investigation to determine the efficiency of supervised learning methods in predicting COVID-19 using hospital records from the Concepción department during the period 2020-2022. The impact of imputing missing data on evaluation metrics is also assessed for various models, including Support Vector Machines, Artificial Neural Networks, Logistic Regression, Decision Trees, and Random Forest. Data preprocessing involves creating two datasets: one without empty records and another with a 20% missing data rate per row. Imputation is carried out using mode imputation and random forest techniques in the dataset with missing data. The dependent variable evaluated is the final disease classification, confirmed or ruled out by laboratory criteria. The Random Forest model stands out for its superior efficiency in the dataset without empty records and demonstrates robustness in the face of data imputation. This research is practically relevant as it provides insights into the effects of frequent missing data imputation in the field of public health and its application in generating predictive models.

Keywords: *Data analysis, Missing value imputation, Supervised learning models.*

[1] Universidad Nacional de Concepción (UNC), Facultad de Ciencias Exactas y Tecnológicas (FACET), Paraguay, adrianmartinezamarilla@gmail.com.

DOI: https://doi.org/10.47133/facet-unc-2023-1-2b

1. INTRODUCCIÓN

La pandemia global del COVID-19 ha generado una profusión de datos sobre la propagación y los efectos de la enfermedad a nivel mundial. Estos datos representan una valiosa fuente de información para profesionales de la salud, epidemiólogos e investigadores que buscan comprender más profundamente la dinámica de la enfermedad. A pesar de su valor, trabajar con estos datos presenta desafíos, siendo la presencia de valores perdidos uno de los mayores obstáculos, ya que puede comprometer la precisión de los análisis y pronósticos, especialmente en modelos de aprendizaje supervisado.

El aprendizaje automático desempeña un papel esencial en la medicina, contribuyendo a la detección temprana de enfermedades como el COVID-19, la optimización de tratamientos y la predicción de resultados clínicos (Bhavsar et al., 2021). Además, en la industria, impulsa la automatización y la toma de decisiones inteligentes, mejorando la eficiencia y la productividad. Este enfoque versátil también encuentra aplicaciones en la investigación, identificando patrones en grandes conjuntos de datos y respaldando la toma de decisiones en tiempo real en sectores como finanzas, marketing y logística.

Es crucial destacar los dos enfoques fundamentales en el Aprendizaje Automático (AA): el aprendizaje supervisado (AS), que implica entrenar con datos previamente etiquetados y se utiliza comúnmente para clasificación o predicción; y el aprendizaje no supervisado (ANS), que trabaja con datos no etiquetados, buscando patrones que permitan la agrupación e identificación de similitudes. El aprendizaje supervisado ha demostrado ser efectivo en diversas aplicaciones, desde diagnósticos médicos (Kononenko, 2001; Mello-Román & Hernández, 2020) hasta la detección de fraudes (Alvarez, 2020).

Un estudio reciente de Andrade-Girón et al. (2023) compara la efectividad de diferentes modelos de Aprendizaje Supervisado (AS) en la detección de pacientes con COVID-19, evidenciando que las Máquinas de Vectores de Soporte y el Bosque Aleatorio destacan en la detección de la enfermedad. Otro estudio de Podder et al. (2021) utiliza aplicaciones de aprendizaje automático para el diagnóstico del COVID-19, con XGBoost y Regresión Logística demostrando ser los modelos más eficaces. Además de los estudios mencionados, el trabajo de Akhtar et al. (2021) resalta la relevancia del hemograma completo en la predicción de COVID-19 mediante algoritmos de aprendizaje automático, evaluando su desempeño mediante medidas como Exactitud, Recuperación, Precisión y Medida F, y subrayando la importancia de abordar la detección desde múltiples enfoques y métricas de evaluación.

El manejo de datos faltantes constituye además una problemática recurrente para los investigadores, siendo objeto de interés desde la década de 1970, gracias a los trabajos pioneros de investigadores como Dempster et al. (1977), Heckman (1979), Rubin (1976) y otros. Estos estudios marcaron el inicio de una relevante línea de investigación en este campo. Para abordar este desafío, los modelos de Aprendizaje Automático (AA) han sido empleados con éxito para imputar valores perdidos en bases de datos, evitando la eliminación de datos y las posibles consecuencias adversas que podrían afectar los resultados (Grillo et al., 2021).

En este contexto, este estudio se enfoca en analizar la base de datos de pacientes con COVID-19 en el departamento de Concepción en Paraguay durante el periodo 2020-2022. El objetivo es aplicar modelos de aprendizaje supervisado, como Máquinas de Vectores Soporte (MVS), Redes Neuronales Artificiales (RNA), Regresión Logística (RL), Bosque Aleatorio (BA) y Árbol de Decisión (AD), abordando simultáneamente los desafíos asociados con valores perdidos en el conjunto de datos. El preprocesamiento de datos se llevó a cabo por etapas, con el propósito de establecer dos conjuntos de datos, el primero sin valores perdidos y el segundo, con datos imputados por moda y bosque aleatorio. Solo se contemplaron casos cuya clasificación final se haya llevado a cabo con criterios laboratoriales.

La importancia social de esta investigación reside en la capacidad potencial para identificar de manera temprana la enfermedad del COVID-19 mediante el análisis de registros hospitalarios. La metodología desarrollada en este estudio se presenta como replicable en distintos contextos geográficos y temporales. Los descubrimientos obtenidos no solo podrían influir directamente en la toma de decisiones en salud pública, sino que también podrían impactar la efectividad de las intervenciones diseñadas para mitigar la propagación del virus. La consideración de los efectos de la imputación de valores perdidos en las métricas de evaluación añade relevancia práctica y metodológica, particularmente en el campo del aprendizaje automático, donde la variabilidad en la calidad de los registros plantea desafíos persistentes relacionados con datos incompletos.

2. MATERIALES Y MÉTODOS

2.1. Enfoque de la investigación

La investigación en el campo del aprendizaje automático utiliza principalmente métodos de investigación cuantitativos, siendo el enfoque dominante el diseño de investigación experimental (Kamiri y Mariga, 2021). En la literatura científica relacionada con este campo, es habitual observar el uso de varios algoritmos para abordar un problema determinado, haciendo especial hincapié en la selección de características para optimizar su rendimiento y realizar una tarea, normalmente de naturaleza predictiva (Sripathi et al., 2024). La evaluación de la eficiencia del algoritmo se lleva a cabo comúnmente utilizando la matriz de confusión y sus derivados, al mismo tiempo que se considera el tiempo de procesamiento del algoritmo (Kröger, 2023; Mello-Román et al., 2021).

En ese sentido, la metodología adoptada en el presente estudio se configura primordialmente como experimental, focalizando su atención en el análisis cuantitativo y la aplicación de herramientas y técnicas específicas. Este enfoque reviste una importancia crucial para alcanzar el propósito central de la investigación, el cual se orienta a examinar la eficacia de diversos modelos de aprendizaje supervisado en la predicción de la incidencia del COVID-19 a través del análisis de registros hospitalarios. A través de esta metodología, se aspira a obtener propiedades y características del desarrollo de la pandemia del COVID-19 en el Departamento de Concepción durante el lapso comprendido entre los años 2020 y 2022. Adicionalmente, se propiciará una base matemática computacional que permitirá prever la presencia de la enfermedad a partir de variables registradas en el sistema sanitario, tales como los síntomas de los pacientes ingresados, variables de contexto, entre otros.

2.2. Población

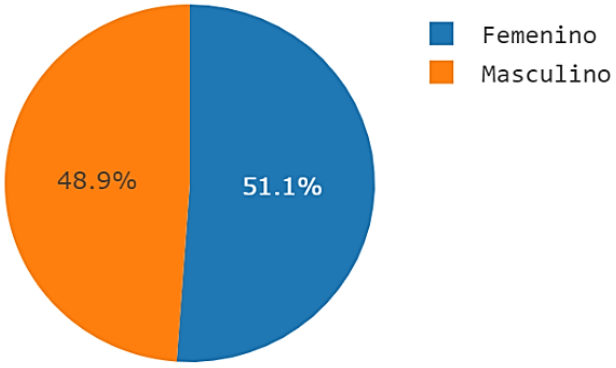
La población bajo estudio comprende un total de 33.028 individuos, correspondientes a registros hospitalarios de pacientes ingresados con sospecha de COVID-19 en centros de salud del Departamento de Concepción durante el periodo 2020-2022. Estos datos fueron proporcionados por la Dirección de Vigilancia Sanitaria del Ministerio de Salud Pública, y constituyen registros administrativos recopilados con el propósito de monitorear el desarrollo de la pandemia específicamente en el mencionado departamento.

Es importante señalar que el conjunto de datos puede no incluir a algunos pacientes de COVID-19 que no ingresaron oficialmente al sistema de salud pública. No obstante, se estima que este porcentaje sería insignificante y no tendría una incidencia significativa en la eficacia de los modelos implementados.

De la población total de 33.028 individuos, el 48,9 % pertenecía al sexo masculino, mientras que el 51,1 % era de sexo femenino, como se ilustra en la Figura 1. Se destaca la importancia de una distribución equitativa

de los datos en cuanto al género de los pacientes, con el objetivo de garantizar que las conclusiones obtenidas sean lo menos afectadas posible por las características individuales de hombres y mujeres.

Figura 1. Pacientes ingresados por COVID-19 según sexo

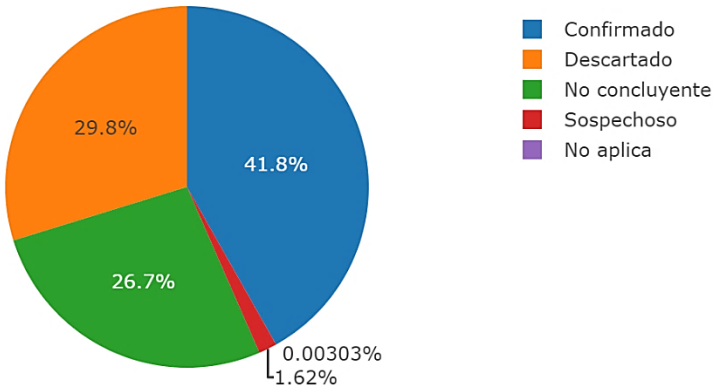


Fuente: elaboración propia.

A continuación, se presentan los valores descriptivos de la variable dependiente "Clasificación Final", la cual abarca cuatro categorías: Confirmado, Descartado, No concluyente o Sospechoso, con la clasificación basada en criterios médicos, pruebas de laboratorio (PCR y/o Antígeno), u otros. La Figura 2 ilustra que el 41,8 % de los casos (13.819 individuos) fueron clasificados como Confirmados, indicando infección por el virus SARS-CoV-2. En contraste, el 29,8 % del total (9.837 individuos) fueron clasificados como Descartados, representando pacientes ingresados al sistema sanitario, pero posteriormente diagnosticados como no infectados por el virus.

Adicionalmente, el 26,7 % (8.833 individuos) fueron considerados No concluyentes, reflejando que la evidencia disponible no permitió a los médicos realizar una determinación definitiva sobre el estado de infección por SARS-CoV-2. Solo el 1,62 % (535 casos) fueron clasificados como Sospechosos, indicando la necesidad de evaluaciones adicionales o pruebas para llegar a una conclusión definitiva sobre su estado, pero que aún no se han completado. Un caso, equivalente al 0,003%, fue categorizado como No Aplica, posiblemente debido a razones vinculadas al registro y/o procesamiento del caso.

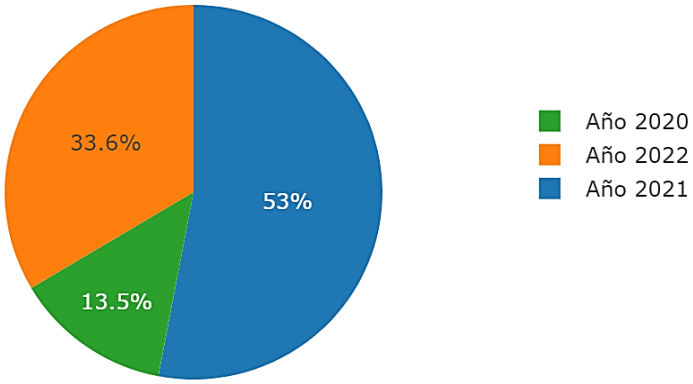
Figura 2. Clasificación final de pacientes ingresados por COVID 19



Fuente: elaboración propia.

La distribución de registros en el período de 2020 a 2022 se presenta en la Figura 3. Los registros del año 2020 constituyen el 13,5% del total, mientras que los del 2021 representan el 53%, y los del 2022, el 33,6%. La menor incidencia de casos de COVID-19 documentados en el sistema de salud durante 2020 se atribuye a las medidas de aislamiento implementadas por el gobierno paraguayo en respuesta a la pandemia. Estas medidas abarcaron protocolos de cuarentena en centros de salud, la transición temporal del sistema educativo al aprendizaje remoto, la adopción del trabajo remoto en oficinas públicas, entre otras medidas similares (Ramos et al., 2020; Rios-González, 2020). A lo largo de los años subsiguientes, estas restricciones se flexibilizaron de manera gradual.

Figura 3. Registros hospitalarios de COVID-19 por año



Fuente: Elaboración propia.

2.4. Etapas del preprocesamiento de datos

En el marco de la presente investigación, se establece como criterio de exclusión la restricción a la utilización exclusiva de registros hospitalarios cuya "Clasificación Final" sea Confirmado o Descartado mediante criterios laboratoriales, específicamente pruebas PCR y/o Antígenos. Se procede a excluir del análisis aquellos casos cuya confirmación o descarte se haya realizado a través de métodos diagnósticos alternativos, así como aquellos catalogados como Sospechosos o No Concluyentes. Este criterio reduce el conjunto de datos a la cantidad total de 23.595 registros. Seguidamente se describe el proceso de limpieza y preparación de los datos en cinco etapas:

Etapa 1: Como primera etapa se procedió a eliminar las columnas o variables con un porcentaje elevado de valores perdidos. Las variables con un porcentaje de valores perdidos superior al 20% fueron excluidos. Varias investigaciones sostienen que no imputar datos en variables con un elevado porcentaje de valores perdidos pone en riesgo la confiabilidad estadística pone en riesgo la confiabilidad estadística (Jannat-Khah et al., 2018; Mello-Román et al., 2019). En la Tabla 1 se citan las cincuenta variables que posterior a esta acción, fueron conservadas para entrenar los modelos: 49 (cuarenta y nueve) variables predictoras y 1 (una) variable dependiente.

Tabla 1. Variables conservadas para el entrenamiento de modelos de AS.

Variables de estudio		
Año	Congestión nasal	Diarrea
Edad	Tos	Factor de riesgo
Sexo	Dificultad para respirar	Cardiopatía crónica
Edad	Irritabilidad confusión	Asma
Semana de notificación	Dolor de cabeza	Enfermedad pulmonar crónica
Fiebre	Inyección conjuntival	Diabetes
Semana inicio fiebre	Disnea-Taquipnea	Enfermedad renal crónica
Zona	Nauseas – Vómitos	Enfermedad Hepática crónica
Asistencia respiratoria mecánica	Dolor abdominal	Inmunodeficiencia Enfermedad Tratamiento
Hospitalizado	Convulsiones	Enfermedad Neuronal
Fallecido	Auscultación pulmonar anormal	Síndrome de Down
Signos/Síntomas	Dolor de oído	Embarazada
Fiebre referida	Dolor de garganta	Viaje – Residencia
Temperatura>38	Mialgia	Contacto con personas
Coriza/Rinorrea	Postración	Contacto con infectados
Técnica	Semana de consulta	Grupo etario
Obesidad	Clasificación final	

Fuente: Elaboración propia.

Etapa 2: La siguiente etapa fue la codificación de registros en cada variable. Las variables conservadas para el entrenamiento de los modelos son en su mayoría de naturaleza dicotómica (Si = 1, No = 0), nominales u ordinales. De las variables que se citan en la Tabla 1 son variables numéricas: Edad, Semana de Notificación y Semana de Inicio de Fiebre.

Etapa 3: La siguiente etapa consistió en la exclusión de los registros cuya proporción de valores perdidos por filas excedía el 20%. En la revisión bibliográfica llevada a cabo, se han identificado diversos criterios respecto al porcentaje mínimo de valores faltantes que debe contener un caso o registro para ser considerado apto para la imputación de datos, sin menoscabo de la fiabilidad estadística del análisis. En este contexto, se ha adoptado como punto de referencia los estudios de Delisle Nyström et al. (2018) y Goicoechea (2002), quienes establecen dicho umbral como criterio.

De esta manera se genera un primer subconjunto de datos a ser imputados, con un total de 14.209 registros. Las técnicas de imputación seleccionadas para la presente investigación son la Imputación por Moda y la Imputación por Bosque Aleatorio. Esta elección se fundamenta en la naturaleza mayormente dicotómica de las variables conservadas y en el enfoque eminentemente experimental de la indagación. No obstante, es pertinente señalar que la aplicación de otras técnicas podría resultar en interpretaciones divergentes, lo que sugiere la apertura de futuras extensiones de este estudio o la realización de investigaciones análogas.

Etapa 4: En esta última etapa, se ha llevado a cabo la generación de un segundo conjunto de datos exento de valores perdidos. Este procedimiento implicó la eliminación de todos los casos que presentaban al menos un valor ausente. Este nuevo subconjunto contiene un total de 2,534 registros hospitalarios. La subdivisión de los datos en dos conjuntos distintos responde al objetivo de la investigación, que consiste en comparar las métricas de evaluación obtenidas para datos sin imputar frente a aquellos imputados mediante las técnicas de imputación por moda y bosque aleatorio. Se busca, asimismo, evaluar el impacto de la introducción de patrones artificiales en las tareas predictivas de los modelos de aprendizaje supervisado.

2.5. Implementación de técnicas

Se han seleccionado los siguientes modelos de aprendizaje supervisado para la presente investigación: Máquina de Vectores de Soporte (MVS), Bosque Aleatorio (BA), Árbol de Decisión (AD), Regresión Logística (RL) y Red Neuronal Artificial (RNA). Esta elección se sustenta en la capacidad de estos modelos para adaptarse a las características particulares del conjunto de datos, caracterizado mayoritariamente por variables independientes de naturaleza categórica. Cabe destacar que estos modelos han sido extensamente empleados en el ámbito del diagnóstico médico de enfermedades con naturaleza epidemiológica, específicamente en el contexto del COVID-19, como evidencian estudios previos (Kononenko, 2001; Andrade-Girón et al., 2023; Podder et al. 2021; Akhtar et al., 2021).

Para la aplicación de las técnicas, se procedió a la partición de cada conjunto de datos, asignando un 70% para las tareas de entrenamiento y un 30% para las tareas de prueba. El respaldo computacional empleado fue el software R, y las librerías utilizadas incluyeron e1071 (Meyer et al., 2023), randomForest (Breiman et al., 2022), rpart (Therneau et al., 2023) y nnet (Ripley & Venables, 2023). A continuación, se detalla la formulación matemática de los modelos de aprendizaje supervisado a implementar.

Máquinas de Vectores Soporte (MVS):

Este modelo de aprendizaje supervisado se genera en base a una transformación lineal $\Phi: R^m \rightarrow H$, donde R^m representa los vectores de entrada y $\Phi(x) \in H$ un nuevo espacio de características. La función Kernel es una medida de similitud y se obtiene por medio del producto escalar entre dos vectores dados en el

espacio transformado $\Phi(u) \cdot \Phi(v) = K(u, v)$ (Shawe-Taylor, Cristianini et al., 2004). A continuación, se citan las tres funciones Kernel a implementar:

- Kernel Gaussiano

$$K(u, v) = \exp(-\sigma \|u - v\|^2), \quad (1)$$

- Kernel Lineal

$$K(u, v) = u \cdot v, \quad (2)$$

- Kernel Polinomial

$$K(u, v) = (u \cdot v + b)^d, \quad (3)$$

donde u y v son vectores en un espacio transformado, d representa el grado del polinomio y σ es el hiperparámetro del Kernel Radial o Gaussiano

Bosque Aleatorio (BA):

El modelo del Bosque Aleatorio se basa en una técnica conocida como Bootstrap, en la que, dado un conjunto de entrenamiento que consta de n observaciones, una variable dependiente Y , y variables predictoras $X_1, X_2, \cdots, X_p$, se realizan muestreos repetidos B veces (Hernán, 2023). El resultado final $\hat{f}$, representa la predicción del modelo del Bosque Aleatorio y se obtiene promediando las predicciones individuales de los B árboles. Es decir:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^{B} f_b(x'), \quad (4)$$

donde $\hat{f}$ es la predicción final del modelo, B representa el número de árboles en el bosque aleatorio, y es $f_b(x')$ la predicción del árbol individual b para una entrada x'. Este enfoque combina múltiples árboles de decisión para mejorar la precisión y la capacidad de generalización del modelo.

El Bosque Aleatorio, conocido en inglés como Random Forest (RF), se compone de predictores de árboles, donde cada árbol depende del valor de un vector aleatorio seleccionado de manera independiente. La distribución de este vector es la misma para todos los árboles en el bosque (Breiman, 2001; A. Cutler et al., 2012). En el caso de problemas de clasificación, una función de pérdida comúnmente utilizada es la entropía cruzada, cuya ecuación se define como sigue:

$$H(y, \hat{y}) = -\sum_{i=1}^{C} y_i \log(\hat{y}), \quad (5)$$

donde y es el vector de etiquetas reales, $\hat{y}$ es el vector de probabilidades predichas y C es el número de clases. La entropía cruzada mide la divergencia entre la distribución real y la predicha, y se busca minimizarla para lograr una mayor precisión (D. R. Cutler et al., 2007; Liaw & Wiener, 2002).

Árbol de Decisión (AD):

En este modelo de aprendizaje supervisado, el proceso de clasificación de un objeto comienza en un nodo raíz, donde el algoritmo de decisión examina cada nodo (i) hasta alcanzar un nodo hoja (Agrusti et al., 2020). El algoritmo de decisión utilizado se conoce como Detector Automático de Interacción Chi-Cuadrado (CHAID) (Kass, 1980). Esta técnica emplea la prueba de independencia Chi-cuadrado para determinar la regla de división en cada nodo, cuya ecuación se representa mediante el estadístico Chi-cuadrado de Pearson:

$$\chi^2 = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - \hat{m}_{ij})}{\hat{m}_{ij}}, \tag{6}$$

donde n_{ij} es la frecuencia de celda observa y $\hat{m}_{ij}$ es la frecuencia de celda esperada, además se considera también el p-valor calculado de la siguiente forma $p = P(\chi^2 > X^2)$, donde χ^2es la distribución con grado de libertad $g = (J-1)(I-1)$.

Regresión Logística (RL):

La regresión logística es una técnica de aprendizaje supervisado utilizada para abordar problemas de clasificación, donde el objetivo es predecir la pertenencia a una categoría binaria. Se emplea para modelar la probabilidad de que un evento pertenezca a una de las dos clases (Segura et al.,2022).

Utiliza una función logística F para transformar la combinación lineal de las características del conjunto de datos en una probabilidad entre 0 y 1, y que puede expresarse como la siguiente ecuación:

$$F(\mathrm{t}) = \frac{\exp(t)}{1 + \exp(t)}, \tag{7}$$

donde $t = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ es denominado modelo logit.

Red Neuronal Artificial (RNA):

La Red Neuronal Artificial como modelo de aprendizaje supervisado, se aplica a problemas de clasificación mediante el uso del algoritmo de propagación inversa de errores. El objetivo principal de este algoritmo es minimizar el error de estimación mediante el cálculo y la actualización sistemática de todos los pesos de la red. Esto se realiza con el fin de alcanzar la configuración óptima de la red neuronal (Kayri, 2015). La ecuación matemática propuesta viene dada por:

$$\begin{aligned} U_k &= \sum\nolimits_{j=1}^{n} W_{kj} X_j, \\ Y_k &= f(U_k + b_k), \end{aligned} \tag{8}$$

donde U_k representa el combinador lineal, X_j son señales de entrada, W_{kj} son los pesos para la neurona k y b_k el valor del sesgo. Luego f es la función de activación, Y_k es la señal de salida de la neurona (Haykin, 1998).

La configuración y elección de los hiperparámetros de los modelos de aprendizaje supervisado presentados, así como su influencia en su rendimiento, se integran y describen en la sección de resultados, como paso previo a la implementación definitiva de dichos modelos.

2.6. Métricas de evaluación

Las métricas de evaluación seleccionadas para este estudio son: Exactitud, Sensibilidad y Especificidad, comúnmente utilizadas para evaluar el rendimiento de los modelos de aprendizaje supervisado en términos de su capacidad predictiva. Desde una perspectiva epidemiológica, la Exactitud representa la proporción de pacientes que fueron clasificados correctamente como enfermos o sanos (Nahm, 2022).

$$\text{Exactitud} = \frac{VP + VN}{VP + FN + FP + VN}, \tag{9}$$

La Sensibilidad se relaciona con la probabilidad de clasificar de manera precisa a un individuo que está enfermo. En otras palabras, es una métrica que calcula la probabilidad de obtener un resultado positivo para un individuo enfermo (Castro Capelo, 2022).

$$\text{Sensibilidad} = \frac{VP}{VP + FN}, \tag{10}$$

La Especificidad se refiere a la probabilidad de clasificar correctamente a un individuo que está sano, según las pruebas diagnósticas realizadas.

$$\text{Especificidad} = \frac{VN}{VN + FP}, \tag{11}$$

3. RESULTADOS

En esta sección, se exponen los resultados derivados de la implementación de modelos de aprendizaje supervisado, así como las métricas de evaluación y las curvas ROC, con el propósito de evaluar el desempeño de cada uno en la tarea de prever la presencia de la enfermedad COVID-19 a partir de registros hospitalarios. Se detallan primeramente los resultados de cada modelo entrenado en los datos sin imputar. Estos resultados se examinan minuciosamente, ya que se consideran exentos de patrones artificiales que podrían ser introducidos mediante la imputación de datos. El conjunto de datos sin imputar consta de un total de 2,534 registros, 49 variables predictoras y 1 variable dependiente llamada "Clasificación Final".

En el apartado final, se presenta una comparación de las métricas de evaluación al implementar estos modelos tanto en subconjuntos de datos imputados por moda como imputados mediante el uso de bosques aleatorios.

Máquinas de Vectores Soporte (MVS)

Las MV fueron implementadas para tres funciones Kernel: Lineal, Gaussiano y Polinomial. Previamente se realizó el ajuste de los hiperparámetros por medio de funciones de calibración en las librerías utilizadas. Tanto los hiperparámetros seleccionados para cada función kernel, como las métricas de evaluación obtenidas se recogen en la Tabla 2.

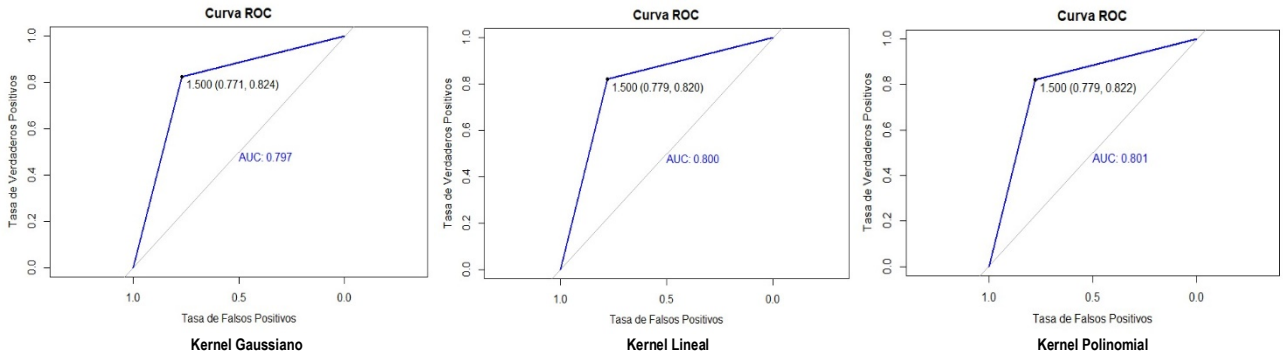
Tabla 2. Rendimiento de MVS con diferentes funciones de Kernel

Kernel	Penalización	Parámetros Kernel	Sensibilidad	Especificidad	Exactitud
Lineal	$C = 10$	-	0,820	0,779	0,807
Gaussiano	$C = 5$	$\sigma = 0{,}001$	0,824	0,771	0,807
Polinomial	$C = 10$	$d = 1\ y\ b = 0{,}5$	0,822	0,779	0,808

Fuente: Elaboración propia

Los resultados de la Tabla 2 indican un desempeño similar de las MV para las tres funciones kernel, con una Exactitud alrededor del 81%, Sensibilidad del 82% y Especificidad entre el 77% y 78%. A fin de complementar el análisis se presentan las Curvas ROC obtenidas para las tres funciones kernel en la Figura 4. Se observa que el AUC (Área Bajo la Curva) para el Kernel Gaussiano fue de 0,797, para el Kernel Lineal de 0,800 y para el Kernel Polinomial 0,801.

Figura 4. Curvas ROC para MVS Kernel Gaussiano, Lineal y Polinomial



Fuente: Elaboración propia

Árbol de Decisión (AD)

La Tabla 3 presenta un resumen de los diversos parámetros empleados en el proceso de entrenamiento del modelo de Árbol de Decisión. Este modelo fue entrenado utilizando 10 validaciones cruzadas con un criterio de entropía. Se seleccionó el mejor parámetro para el número de poda (Cp), siendo este 0.01. El árbol experimentó 6 divisiones (Nsplit), lo que resultó en un error relativo de precisión del modelo de 0,561. Estos valores también señalan que un menor error de validación cruzada (Xerror), en este caso, 0.618, representa un rendimiento mejorado del modelo.

Tabla 3. Evaluación del modelo AD mediante Poda: Métricas de Error y Complejidad

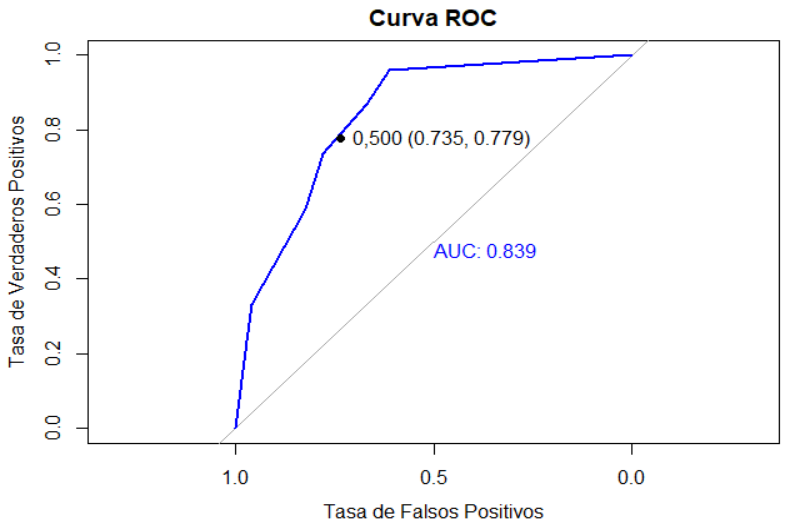
Cp	Nsplit	Error relativo	Xerror	Xstd	Sensibilidad	Especificidad	Exactitud
0,308	0	1	1	0,033	1	0	0,642
0,024	1	0,692	0,692	0,030	0,779	0,735	0,763

0,010	6	0,561	0,618	0,029	0,779	0,735	0,763

Fuente: Elaboración propia

Las métricas de evaluación obtenidas fueron: Sensibilidad 77,9%, Especificidad 73,5% y Exactitud del 76,3%. La curva ROC se presenta en la Figura 5. El AUC o área bajo la curva obtenido fue de 0,839, lo que indica una buena capacidad de discriminación del modelo.

Figura 5. Curva ROC para Árbol de Decisión.

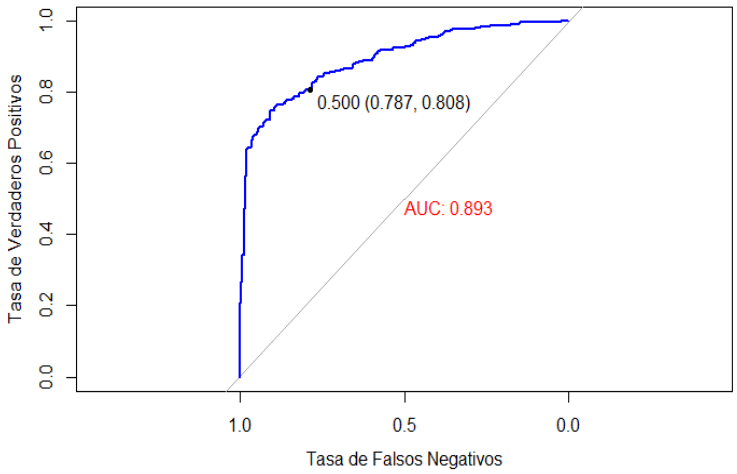


Fuente: Elaboración propia

Regresión Logística (RL):

Antes de llevar a cabo el entrenamiento con Regresión Logística, se llevaron a cabo 10 validaciones cruzadas con el objetivo de asegurar la confiabilidad de la clasificación. Durante este proceso, la función utilizada aplicó una penalización automática con un valor de 1553,5, y el modelo logró converger en tan solo 2 iteraciones. Las métricas de evaluación resultantes fueron las siguientes: una Sensibilidad del 80,8%, una Especificidad del 78,7%, y una Exactitud del 80%. La Figura 6 presenta la curva ROC, y el área bajo la curva (AUC) obtenida fue de 0.893.

Figura 6. Curva ROC para Regresión Logística



Fuente: Elaboración propia

Bosque Aleatorio (BA):

En el modelo de Bosque Aleatorio, según Bartz et al. (2023), no hay un acuerdo pleno sobre los valores que puede tomar "mtry" (número de aleatoriedad). Por lo tanto, se han seleccionado previamente diferentes valores, teniendo en cuenta el criterio de entropía. Cabe señalar que los valores de "n" representan el número de variables predictoras, en este caso, 49 variables.

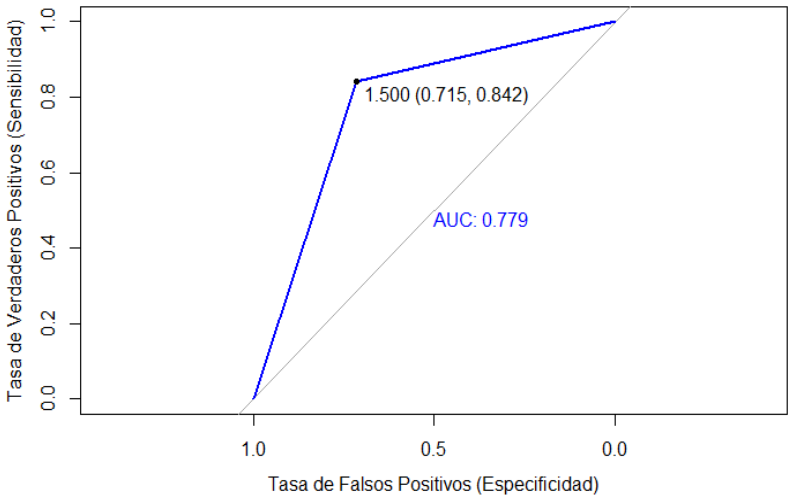
Tabla 4. Evaluación del modelo BA para diferentes valores de mtry

Parámetro	Valor mtry	Exactitud	Sensibilidad	Especificidad
$\boldsymbol{mtry = \sqrt{n}}$	7	0,805	0,868	0,685
$\boldsymbol{mtry = n/2}$	25	0,793	0,834	0,715
$\boldsymbol{mtry = n/3}$	16	0,799	0,842	0,715
$\boldsymbol{mtry = log_2(n) + 1}$	6	0,795	0,878	0,635

Fuente: Elaboración propia

Con mtry =16, el modelo de Bosque Aleatorio logró una exactitud del 79,9%, una sensibilidad del 84,2%, y una especificidad del 71,5%. Para este valor de parámetro, el modelo registró un área bajo la curva (AUC) más alto, siendo este igual a 0,779. La curva ROC correspondiente se muestra en la Figura 7.

Figura 7. Curva ROC para Bosque Aleatorio



Fuente: Elaboración propia

Red Neuronal Artificial (RNA):

La Tabla 5 resume las variaciones de parámetros, capas ocultas y pesos en la red que se realizaron durante el entrenamiento del modelo de Red Neuronal Artificial. Se llevaron a cabo 10 iteraciones de validación cruzada en formato K-Fold. La máxima exactitud 77,6% se logró al emplear 3 unidades ocultas y un peso de 0e+00.

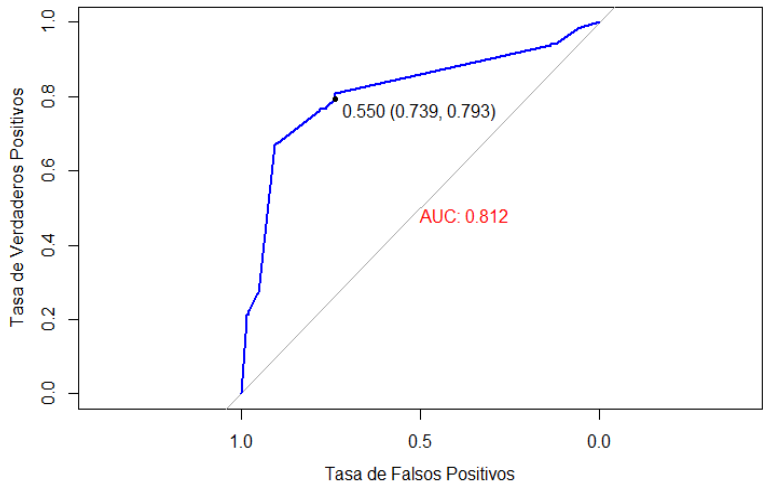
Tabla 5: Evaluación del modelo RNA: Variaciones de parámetros.

Unidades ocultas	Decaimiento de los pesos	Exactitud
1	0e+00	0,768
1	1e-04	0,767
1	1e-01	0,769
3	0e+00	0,776
3	1e-04	0,767
3	1e-01	0,761
5	0e+00	0,750
5	1e-04	0,738
5	1e-01	0,760

Fuente: Elaboración propia

En esta configuración de parámetros, la sensibilidad del modelo fue del 79,3% y la especificidad del 73,9%. Se obtuvo un área bajo la curva (AUC) de 0,812, como se puede apreciar en la Figura 8.

Figura 8. Curva ROC para Red Neuronal Artificial



Fuente: Elaboración propia

En el contexto de datos sin imputar, el Bosque Aleatorio (RF) sobresalió como el modelo óptimo, exhibiendo una sensibilidad destacada del 84.2% y una especificidad aceptable del 71.5%, lo que resultó en una notable exactitud del 79.9%. La elección de este modelo se fundamenta en la importancia crítica de la sensibilidad, particularmente crucial en el ámbito de datos epidemiológicos donde el enfoque se centra en el diagnóstico de enfermedades como el COVID-19. La sólida combinación de sensibilidad y especificidad del Bosque Aleatorio resalta su eficacia en la clasificación precisa de casos, consolidándolo como la opción más robusta en el contexto de datos sin imputar.

Imputación de valores perdidos: Impacto en las métricas de evaluación.

En secciones anteriores, se presentaron los resultados de la implementación de cinco modelos de aprendizaje supervisado: Máquina de Vectores de Soporte (MVS), Bosque Aleatorio (BA), Árbol de Decisión (AD), Regresión Logística (RL) y Red Neuronal Artificial (RNA), exclusivamente para los datos sin imputar. En esta sección, se lleva a cabo una comparación de los resultados de estos mismos modelos en tres escenarios de datos distintos: Datos sin imputar, Datos imputados por Moda y Datos imputados por Bosque Aleatorio.

El conjunto de datos imputados por Moda y Bosque Aleatorio consta de un total de 14,209 registros, con casos que presentaban valores perdidos en menos del 20% del total de variables predictoras, y que fueron remplazados mediante las técnicas mencionadas anteriormente. La Tabla 6 presenta las métricas de evaluación para los diferentes modelos, permitiendo su comparación en los tres escenarios.

Tabla 6: Métricas de evaluación de modelos para datos sin imputar y datos imputados

	Datos sin imputar			Datos imputados por Moda			Datos imputados por Bosque Aleatorio		
Modelos	Sen.	Esp.	Exa.	Sen.	Esp.	Exa.	Sen.	Esp.	Exa.
MVS-L	0,820	0,779	0,807	0,570	0,949	0,747	0,570	0,947	0,746
MVS-R	0,824	0,771	0,807	0,583	0,959	0,757	0,583	0,960	0,758
MVS-P	0,822	0,779	0,808	0,583	0,954	0,755	0,583	0,952	0,754
RNA	0,790	0,740	0,780	0,726	0,800	0,760	0,697	0,843	0,764
AD	0,779	0,735	0,763	0,565	0,975	0,756	0,565	0,975	0,756
RL	0,808	0,787	0,801	0,647	0,907	0,767	0,6464	0,911	0,769
BA	0,842	0,715	0,799	0,753	0,820	0,784	0,756	0,826	0,788

* Sen.= Sensibilidad; Esp.=Especificidad; Exa.= Exactitud

Fuente: Elaboración propia.

Los resultados de la Tabla 6 podrían resumirse de la siguiente manera:

Máquina de Vectores de Soporte (MVS)

En el escenario de datos sin imputar, el modelo MVS-L (Kernel Lineal) demostró un rendimiento robusto con una sensibilidad del 82,0%, especificidad del 77,9% y exactitud del 80,7%. Sin embargo, al imputar datos por Moda o Bosque Aleatorio, la sensibilidad disminuyó significativamente a 57,0%, mientras que la especificidad se mantuvo alta. Los resultados indican que el modelo es sensible a la imputación de datos, especialmente en términos de sensibilidad.

En cuanto al modelo MVS-R (Kernel Gaussiano), en el escenario de datos sin imputar, exhibió un rendimiento sólido con una sensibilidad del 82,4%, especificidad del 77,1% y exactitud del 80,7%. Al imputar datos por Moda o Bosque Aleatorio, la sensibilidad disminuyó, pero la especificidad aumentó, manteniendo la exactitud alrededor del 75-76%. La robustez del modelo se mantiene, pero con variaciones en sensibilidad y especificidad.

En el escenario de datos sin imputar, el modelo MVS-P (Kernel Polinomial) destacó con una sensibilidad del 82,2%, especificidad del 77,9% y exactitud del 80,8%. La imputación por Moda o Bosque Aleatorio llevó a una disminución de la sensibilidad, pero la especificidad se mantuvo alta. En general, el modelo muestra consistencia en el rendimiento, aunque afectado por la imputación.

Red Neuronal Artificial (RNA):

Para datos sin imputar, la RNA mostró una sensibilidad del 79,0%, especificidad del 74,0% y exactitud del 78,0%. Con datos imputados, se observa una disminución en la sensibilidad, pero un aumento en la especificidad, afectando la exactitud. La red neuronal parece ser sensible a la imputación de datos, mostrando variaciones en el rendimiento.

Árbol de Decisión (AD):

En datos sin imputar, el árbol de decisión presentó una sensibilidad del 77,9%, especificidad del 73,5% y exactitud del 76,3%. La imputación de datos afectó significativamente la sensibilidad, aunque la especificidad se mantuvo alta. El modelo muestra cierta sensibilidad a la imputación de datos, especialmente en términos de sensibilidad.

Regresión Logística (RL):

Para datos sin imputar, la regresión logística exhibió sensibilidad del 80,8%, especificidad del 78,7% y exactitud del 80,1%. La imputación de datos resultó en una disminución de la sensibilidad, pero un aumento en la especificidad, afectando ligeramente la exactitud. El modelo muestra consistencia, aunque con cierta sensibilidad a la imputación de datos.

Bosque Aleatorio (RF):

En el escenario de datos sin imputar, el Bosque Aleatorio mostró una sensibilidad del 84,2%, especificidad del 71,5% y exactitud del 79,9%. La imputación de datos afectó la sensibilidad, pero mantuvo una especificidad razonable, y la exactitud se mantuvo en niveles aceptables. El modelo demuestra robustez, aunque muestra variaciones en la sensibilidad con la imputación de datos.

Observando los resultados proporcionados, el modelo que muestra una menor variación en la exactitud entre datos sin imputar y datos imputados es también el Bosque Aleatorio (BA), con una Exactitud del 79,9 % para datos sin imputar, 78,4%, para datos imputados por Moda y 78,8%, para datos imputados por Bosque Aleatorio. La diferencia en la exactitud entre datos sin imputar y datos imputados es relativamente pequeña para el Bosque Aleatorio, indicando una menor variación en su rendimiento en comparación con otros modelos en diferentes escenarios de imputación. Esto sugiere que el Bosque Aleatorio es más robusto ante la imputación de datos en este conjunto de datos específico.

4. CONCLUSIONES

En conclusión, los hallazgos de esta investigación poseen implicaciones significativas tanto para la evaluación de la eficacia de modelos de aprendizaje supervisado en la predicción del COVID-19 como para la aplicación de técnicas de imputación de datos faltantes en el análisis de datos de salud pública. La evaluación de varios modelos, incluyendo Máquinas de Vectores de Soporte (MVS), Redes Neuronales Artificiales (RNA), Regresión Logística (RL), Bosque Aleatorio (BA), y Árbol de Decisión (AD), reveló que el Bosque Aleatorio destacó como el modelo más eficiente y robusto para la predicción de la incidencia del COVID-19 en este estudio.

La problemática de datos faltantes fue abordada, subrayando la importancia de la imputación de valores perdidos mediante modelos de aprendizaje automático. Se estableció un criterio riguroso de exclusión basado en la fiabilidad de los registros hospitalarios, específicamente aquellos con "Clasificación Final" Confirmado o Descartado mediante pruebas PCR y/o Antígenos. Se evidenció la importancia práctica y metodológica de considerar los efectos de la imputación de valores perdidos en las métricas de evaluación, especialmente en el contexto del aprendizaje automático, donde la variabilidad en la calidad de los registros presenta desafíos persistentes vinculados a datos incompletos.

No obstante, es esencial reconocer las posibles limitaciones de este estudio, como la regionalización de los resultados y la evaluación de un conjunto específico de modelos de aprendizaje supervisado. Se sugiere futura investigación que replique el estudio en diferentes regiones o países, explore otras técnicas de imputación y considere variables adicionales para mejorar la precisión de los modelos de predicción. Además, la incorporación del análisis de series temporales podría proporcionar una comprensión más profunda de la evolución de la incidencia del COVID-19 y su relación con diversas variables a lo largo del tiempo. Estas ampliaciones propuestas contribuirían a una comprensión más completa y robusta de la eficacia de los modelos de aprendizaje supervisado en la predicción del COVID-19, así como en la gestión de datos faltantes en el contexto de la salud pública.

AGRADECIMIENTOS

Agradezco a la Dirección de Vigilancia Sanitaria del Ministerio de Salud Pública y Bienestar Social y la Primera Región Sanitaria del Departamento de Concepción por facilitar el conjunto de datos debidamente anonimizado, para uso exclusivo con fines académicos.

REFERENCIAS

Agrusti, F., Mezzini, M., & Bonavolontà, G. (2020). Deep learning approach for predicting university dropout: A case study at Roma Tre University. Journal of e-Learning and Knowledge Society, 16(1), 44-54.

Akhtar, A., Akhtar, S., Bakhtawar, B., Kashif, A. A., Aziz, N., & Javeid, M. S. (2021). COVID-19 detection from CBC using machine learning techniques. *International Journal of Technology, Innovation and Management (IJTIM), 1*(2), 65-78.

Alvarez, F. (2020). Machine Learning en la detección de fraudes de comercio electrónico aplicado a los servicios bancarios. *Ciencia y tecnología*, 81-95.

Andrade-Girón, D., Carreño-Cisneros, E., Mejía-Dominguez, C., Marín-Rodriguez, W., & Villarreal-Torres, H. (2023). Comparación de Algoritmos Machine Learning para la Predicción de Pacientes con Sospecha de COVID-19. *Salud, Ciencia y Tecnología*, 3, 336-336.

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., & Mersmann, O. (2023). Hyperparameter Tuning for Machine and Deep Learning with R: A Practical Guide. Springer Nature.

Bhavsar, KA, Abugabah, A., Singla, J., AlZubi, AA y Bashir, AK (2021). Una revisión exhaustiva sobre el diagnóstico médico mediante aprendizaje automático. Computadoras, Materiales y Continua , 67 (2), 1997.

Breiman, L. (2001). Random forests. Machine learning, 45, 5-32

Breiman, L., Cutler, A., Liaw, A., & Wiener, M. (2022). randomForest: Breiman and Cutler's Random Forests for Classification and Regression [R package version 4.7-1.1]. Recuperado de https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

Castro Capelo, R. M. (2022). Curvas ROC

Chiapella, L. (2020). Impacto de estrategias para el tratamiento de información faltante sobre la estimación de modelos de regresión de Cox.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. Ensemble machine learning: Methods and applications, 157-175.

Delisle Nyström, C., Barnes, J. D., & Tremblay, M. S. (2018). An exploratory analysis of missing data from the Royal Bank of Canada (RBC) Learn to Play–Canadian Assessment of Physical Literacy (CAPL) project. *BMC Public Health*, *18*(2), 1-9.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society: series B (methodological), 39(1), 1-22.

Goicoechea, A. P. (2002). Imputación basada en árboles de clasificación. *Eustat. Available in: http://www. eustat. es/documentos/datos/ct*, *4*.

Grillo, S. A., Román, J. C. M., Mello-Román, J. D., Noguera, J. L. V., García-Torres, M., Divina, F., & Sotomayor, P. E. G. (2021). Adjacent Inputs With Different Labels and Hardness in Supervised Learning. *IEEE Access*, *9*, 162487-162498.

Haykin, S. (1998). Neural networks: a comprehensive foundation. Prentice Hall PTR.

Heckman, J. J. (1979). Sample selection bias as a specification error. Econometrica: Journal of the econometric society, 153-161.

Hernán, F. (2023). Random Forests. Consultado el 20 de octubre de 2023, Recuperado de https://fhernanb.github.io/libro_mod_pred/rand-forests.html

Jannat-Khah, D. P., Unterbrink, M., McNairy, M., Pierre, S., Fitzgerald, D. W., Pape, J., & Evans, A. (2018). Treating loss-to-follow-up as a missing data problem: a case study using a longitudinal cohort of HIV-infected patients in Haiti. BMC public health, 18, 1-11.

Kamiri, J., & Mariga, G. (2021). Research methods in machine learning: A content analysis. *International Journal of Computer and Information Technology (2279-0764)*, *10*(2), 78-91.

Kass, G. (1980). Una técnica exploratoria para investigar grandes cantidades de datos categóricos. Revista de la Royal Statistical Society: Serie C (Estadísticas aplicadas), 29(2), 119-127

Kayri, M. (2015). An intelligent approach to educational data: Performance comparison of the multilayer perceptron and the radial basis function artificial neural networks. Educational Sciences: Theory & Practice, 15(5).

Kononenko, I. (2001). Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1), 89-109.

Kröger, H. (2023). 8. Predictive machine learning approaches–possibilities and limitations for the future of life course research. *Handbook of Health Inequalities Across the Life Course*, 112.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

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

Mello-Roman, J. D., & Hernandez, A. (2020). KPLS optimization with nature-inspired metaheuristic algorithms. *IEEE Access*, *8*, 157482-157492.

Mello-Román, J. D., Hernández, A., & Mello-Román, J. C. (2021). Improved Predictive Ability of KPLS Regression with Memetic Algorithms. *Mathematics*, *9*(5), 506.

Mello-Román, J. D., Mello-Román, J. C., Gomez-Guerrero, S., & García-Torres, M. (2019). Predictive models for the medical diagnosis of dengue: a case study in Paraguay. *Computational and mathematical methods in medicine*, *2019*.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2023). e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien [R package version 1.7-13]. Recuperado de https://cran.r-project.org/web/packages/e1071/index.html

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Korean journal of anesthesiology*, 75(1), 25-36.

Podder, P., Bharati, S., Mondal, M. R. H., & Kose, U. (2021). Application of machine learning for the diagnosis of COVID-19. En Data science for COVID-19 (pp. 175-194). Elsevier

Ramos, P., Silva, E., Canese, J., & Velázquez, G. (2021). Epidemiologia de los casos de COVID-19 diagnosticados en albergues sanitarios del gran Asunción, Paraguay (2020). *Memorias del Instituto de Investigaciones en Ciencias de la Salud*, *19*(2), 69-77.

Rios-González, C. M. (2020). Conocimientos, actitudes y prácticas hacia COVID-19 en paraguayos el periodo de brote: una encuesta rápida en línea. *Revista de salud publica del Paraguay*, *10*(2), 17-22.

Ripley, B., & Venables, W. (2023). nnet: Feed-Forward Neural Networks and Multinomial LogLinear Models [R package version 7.3-19]. Recuperado de https://cran.r-project.org/web/packages/nnet/nnet.pdf

Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581-592

Segura, M., Mello, J., & Hernández, A. (2022). Machine Learning Prediction of University Student Dropout: Does Preference Play a Key Role? *Mathematics, 10*(18), 3359

Shawe-Taylor, J., Cristianini, N., et al. (2004). Kernel methods for pattern analysis. Cambridge university press.

Sripathi, K. N., Moscarella, R. A., Steele, M., Yoho, R., You, H., Prevost, L. B., ... & Haudek, K. C. (2024). Machine learning mixed methods text analysis: An illustration from automated scoring models of student writing in biology education. *Journal of mixed methods research*, *18*(1), 48-70.

Therneau, T., Atkinson, B., & Ripley, B. (2023). rpart: Recursive Partitioning and Regression Trees [R package version 4.1.21]. Recuperado de https://cran.r-project.org/web/packages/rpart/rpart.pdf