UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA
DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

DESARROLLO DE UN ALGORITMO DE PREDICCIÓN DE RIESGO DE HOSPITALIZACIÓN, VENTILACIÓN MECÁNICA E INGRESO A UCI EN PACIENTES CON COVID-19

POR

Rocío Valentina Arriagada Arroyo

Memoria de Título presentada a la Facultad de Ingeniería de la Universidad de Concepción para optar al título profesional de Ingeniera Civil Biomédica

Profesor Guía
Rosa Figueroa
María Elena Lagos

Comisión Evaluadora
Jorge Pezoa

Enero 2022
Concepción (Chile)

Agradecimientos

A mis padres por el apoyo en esta etapa.

A Raúl por estar siempre.

A las amigas que me dejó el paso por esta facultad, Tiare e Isabella.

A mi profesora guía, Rosa Figueroa por su confianza.

Resumen

En diciembre de 2019 en la ciudad china de Wuhan comenzó un brote de neumonía asociada al coronavirus SARS-CoV-2. Rápidamente, el virus comenzó a esparcirse por el mundo, estando a la fecha presente en más de 189 países con un saldo aproximado de 397 millones de casos y 5,75 millones de fallecidos.

Este trabajo tiene como objetivo desarrollar un algoritmo de predicción de riesgo de hospitalización, ingreso a UCI y ventilación mecánica en pacientes chilenos con COVID-19. Se utilizó una base de datos que contiene 1.062 registros extraídos del Sistema de Vigilancia Epidemiológica Chileno EPIVIGILA del ministerio de salud. De las variables disponibles, se seleccionaron 52 variables predictoras incluyendo 5 variables correspondientes a signos vitales, 20 variables de biomarcadores, 11 variables del tipo comorbilidad y 14 variables del tipo síntomas, sexo y edad; dichas variables serán utilizadas para predecir las variables de salida de riesgo de hospitalización, ingreso a UCI y ventilación mecánica.

Sobre este conjunto de variables se realizó un proceso de selección de características usando regresión logística y transformaciones no lineales de splines sobre las variables continuas con el objetivo de identificar de mejor manera las posibles relaciones entre estas variables y las de salida. La selección de los subconjuntos de predictores para cada salida se realizó utilizando el estadístico Wald chi-cuadrado, método con el cual se puede valorar la significancia de un variable a la hora de realizar la predicción.

Posteriormente, se utilizaron 3 algoritmos de clasificación: regresión logística (RL), *Random Forest* (RF)y redes neuronales, se ajustaron sus parámetros con validaciones cruzadas y fueron entrenados. Para evaluar el desempeño de los modelos se utilizó la matriz de confusión y las métricas derivadas de esta, además se trazaron las curvas ROC y PRC para comparar entre modelos.

Con RF se identificó el riesgo de hospitalización con un 94% de sensibilidad y un 83,5% de precisión. Para ingreso a UCI se logró un 70,4% de sensibilidad y un 78,1% de precisión con perceptrón multicapa. Con este clasificador de logró identificar el riesgo de ventilación mecánica invasiva con un 69% de sensibilidad y un 80% de precisión.

A pesar de los desafíos propios de la clasificación con clases desbalanceadas y las limitaciones dadas por los datos, fue posible obtener modelos de predicción con buenos rendimientos, recordando siempre analizar este punto comparando distintas métricas y analizando la variación de estas para poder obtener una imagen más completa y real del rendimiento de los clasificadores.

Abstract

In December 2019, an outbreak of pneumonia associated with the SARS-CoV-2 coronavirus began in the Chinese city of Wuhan. Rapidly, the virus began to spread around the world, being to date present in more than 189 countries with an approximate balance of 397 million cases and 5.75 million deaths.

The aim of this work is to develop an algorithm for predicting the risk of hospitalization, admission to ICU and mechanical ventilation in Chilean patients with COVID-19. A database containing 1,062 records extracted from the Chilean Epidemiological Surveillance System EPIVIGILA of the Ministry of Health was used. Of the available variables, 52 predictor variables were selected, including 5 variables corresponding to vital signs, 20 biomarker variables, 11 comorbidity-type variables and 14 variables of the symptom, sex and age type; these variables will be used to predict the variables of outcome: hospitalization risk, admission to ICU and mechanical ventilation.

A feature selection was performed on this set of variables using logistic regression and nonlinear spline transformations on the continuous variables. These transformations help to better identify the possible relationships between the continuous and output variables.

The selection of the subsets of predictors for each output was carried out using the Wald chi-square statistic, a method with which the significance of a variable can be assessed when making the prediction.

Subsequently, 3 classification algorithms were used: logistic regression (LR), Random Forest (RF) and neural networks, their parameters were adjusted with cross-validations and trained. To evaluate the performance of the models, the confusion matrix and the metrics derived from it were used, and ROC and PRC curves were plotted to compare between models.

With RF, the risk of hospitalization was identified with 94% sensitivity and 83.5% precision. For admission to the ICU, 70.4% sensitivity and 78.1% precision were achieved with multilayer perceptron. With this classifier we were able to identify the risk of invasive mechanical ventilation with 69% sensitivity and 80% precision.

Despite the challenges of classification with unbalanced classes and the limitations given by the data, it was possible to obtain predictive models with good performance, always remembering to analyze this point by comparing different metrics and analyzing the variation of these in order to obtain a more complete and truer picture of the performance of the classifiers.

Tabla de contenidos

Índice de figuras

Índice de tablas

Abreviaciones

SARS-CoV-2 : Coronavirus de tipo 2 causante del síndrome respiratorio agudo severo.

COVID-19 : Enfermedad por coronavirus 19.

MC : Matriz de confusión.

EPOC : Enfermedad pulmonar obstructiva crónica.

SVM : Máquinas de soporte vectorial, del inglés *support vector machine.*

LASSO : *Least absolute shrinkage and selection operator***,** por sus siglas en inglés.

XGBoost : Del inglés *Extreme gradient boost*.

RF : *Random Forest*, por sus siglas en inglés.

RL : Regresión logística.

IMC : Índice de masa corporal.

UCI : Unidad de cuidados intensivos.

UTI : Unidad de tratamientos intermedios.

MICE : *Multiple imputation by chained equations*, por sus siglas en inglés.

TGP : Alanina aminotransferasa.

pH : Potencial de hidrógeno.

PaO2 : Presión parcial de oxígeno.

FiO2 : Fracción inspirada de oxígeno.

VSG : Velocidad de segmentación globular.

TGO : Aspartato aminotransferasa.

CPK : Creatinina fosfoquinasa.

TTPa : Tiempo parcial de tromboplastina activada.

LDH : Lactato deshidrogenasa.

API : Interfaz de programación de aplicaciones, del inglés *Application Programming Interface.*

PMC : Perceptrón multi capa.

ROC : Característica Operativa del Receptor, por sus siglas en inglés.

VP : Verdaderos positivos.

FP : Falsos positivos.

FN : Falsos negativos.

VN : Verdaderos negativos.

PRC : Curva de precisión-sensibilidad, por sus siglas en inglés.

AUC : Área bajo la curva, por sus siglas en inglés.

Capítulo 1 Introducción

1.1 Introducción general

En diciembre de 2019 en la ciudad de Wuhan en China comenzó un brote de neumonía asociada al coronavirus de tipo 2 causante del síndrome respiratorio agudo severo SARS-CoV-2 (del inglés *severe acute respiratory síndrome coronavirus 2*). Rápidamente, el virus comenzó a esparcirse por todo el mundo, estando a la fecha presente en más de 188 países con un saldo de 397 millones de casos y 5,75 millones de fallecidos, afectando a personas de todas las edades, sexos y etnias. El desarrollo de la enfermedad por coronavirus 19 (COVID-19) varía desde casos asintomáticos, cursos leves, hasta cuadros graves con resultado de muerte [1].

Con el avance de la pandemia se ha logrado identificar grupos de riesgo, principalmente determinados por comorbilidades y rango etario. Los pacientes de mayor edad y quienes padecen enfermedades de base como la diabetes e hipertensión tienen mayor predisposición a cursar cuadros graves o tener un desenlace fatal [2][3]. Entre los estudios realizados alrededor del mundo, los autores han encontrado que las comorbilidades de mayor prevalencia son la hipertensión arterial, seguida de obesidad y diabetes tipo II [4][5][6], [7]. Además, se han estudiado resultados clínicos como mortalidad, enfermedad grave, hospitalización, duración de la hospitalización, necesidad de intubación y desarrollo de síndrome de distrés respiratorio agudo [8][9][10].

En un esfuerzo por apoyar el proceso diagnóstico y ayudar a disponer de forma óptima de los recursos de atención hospitalaria limitados, se ha desarrollado una cantidad importante de modelos de predicción, que van desde sistemas basados en reglas hasta avanzados modelos basados en *deep learning*. Estos modelos se pueden dividir en tres categorías principales: modelos para identificar personas en riesgo entre la población general, modelos para detectar COVID-19 en pacientes con sospecha de infección y modelos para predecir el pronóstico en pacientes con COVID-19 confirmado, utilizando como predictores datos demográficos, resultados de exámenes de laboratorio, información de signos vitales, exámenes de imagenología, entre otros [11].

Si bien los predictores considerados en los estudios revisados han sido considerados relevantes y sugeridos para modelos y análisis posteriores, los modelos desarrollados presentan alto riesgo de sesgo y sobreajuste, debido principalmente a selección de muestras no representativas y exclusión de

pacientes que no presentaran el evento de interés. Además, se considera que el reporte dichos estudios presenta falta de documentación o cantidad de datos limitada [11][12][13].

Con el objetivo de encontrar relaciones entre factores demográficos y clínicos que puedan influir en el riesgo de hospitalización, ingreso a unidad de cuidados intensivos (UCI) o ventilación mecánica al cursar COVID-19 en pacientes chilenos, en este trabajo se desarrollará un modelo de aprendizaje automático para lograr el objetivo anteriormente mencionado.

1.2 Objetivos

1.2.1 Objetivos generales

Desarrollar un algoritmo de predicción de riesgo de hospitalización, ingreso a UCI y necesidad de ventilación mecánica invasiva en pacientes con COVID-19 chilenos, utilizando datos extraídos desde el Sistema de Vigilancia Epidemiológica EPIVIGILA del Ministerio de Salud de Chile.

1.2.2 Objetivos específicos

- Preparar las bases de datos a través de un preprocesamiento y selección de variables predictoras.
- Aplicar los modelos de predicción de riesgo de hospitalización, ingreso a UCI y ventilación mecánica.
- Evaluar el desempeño de los modelos utilizando métricas de la matriz de confusión (MC).

1.3 Alcances

- Se analizarán los datos pertenecientes a un set de datos obtenidos de la plataforma EPIVIGILA con certificado de comité de ética.
- Los datos corresponden a los recogidos entre el 1 de enero de 2020 y el 12 de enero de 2021.
- Se excluirán del análisis los pacientes con alto índice de variables no reportadas.
- Los casos de COVID-19 con virus identificado o no identificado (de acuerdo al código CIE-10) comenzaron a ser registrados desde el 26 de febrero de 2020.

1.4 Limitaciones

- Los registros se encuentran incompletos para una gran cantidad de entradas.
- Algunas variables predictoras de interés encontradas en la literatura, presentan una densidad de datos muy baja.
- Se considerarán solo los registros que contengan información no nula para el biomarcador con mayor densidad de datos.

1.5 Materiales y métodos

- El análisis de los datos se realizará utilizando la plataforma Goole Colab Pro en lenguaje Python versión 3.7.10 y Rstudio en lenguaje R versión 4.1.1.
- Se revisará y estudiará literatura disponible relacionada con el tema y la problemática a solucionar, seleccionando solo documentos disponibles en fuentes confiables.

Capítulo 2 Revisión bibliográfica

2.1 Introducción

La revisión bibliográfica realizada se orientó a investigar sobre las comorbilidades asociadas al COVID-19 para contextualizar la enfermedad, encontrar las enfermedades de mayor prevalencia entre los individuos infectados, detectar diferencias entre las comorbilidades presentes en pacientes que desarrollan la enfermedad con distinta gravedad y aspectos relacionados con las comorbilidades presentes y los resultados clínicos observados. También se revisó literatura correspondiente a factores no clínicos que pudieran tener relación con el curso que toma la enfermedad. Finalmente, se exploró la literatura disponible sobre modelos de predicción desarrollados para ser utilizados con datos de pacientes COVID-19, principalmente los destinados a predicción de pronósticos clínicos y el uso de predictores tales como biomarcadores, factores clínicos y demográficos.

2.2 Revisión de literatura

Iniciamos esta revisión con el estudio de W. H. Ng et al, quien describe un meta análisis de estudios sobre comorbilidades en pacientes con SARS-CoV-2. Este estudio incluyó 375.859 participantes de 14 países, entre los cuales las comorbilidades de mayor prevalencia fueron hipertensión, obesidad y diabetes, presentes en el 21.3%, 18.3% y 18.1% de los pacientes[14]. Por otra parte, en el trabajo de J. Liu et.al, se encontró que las comorbilidades más frecuentes entre pacientes con SARS-CoV2 fueron hipertensión (26.1 %), diabetes (12.2%) y enfermedad cardíaca crónica (7.3%). También, se identificaron otras comorbilidades presentes en menor porcentaje como enfermedad pulmonar obstructiva crónica (EPOC) en el 1.9%, enfermedad renal crónica en el 2.6%, enfermedad hepática crónica en el 3.4%, accidente cerebrovascular en el 3.3%, Cáncer en 2.9%, inmunosupresión en 2.0% y tuberculosis en el 1.3% de los pacientes. En A. Sanyaolu et.al, identificaron como comorbilidades más frecuentes entre pacientes con COVID-19 hipertensión (15.8%), condiciones cardiovasculares y cerebrovasculares (11.7%) y diabetes (9.4%).

El asma no ha sido considerado como factor de riesgo en el desarrollo de enfermedad grave, sin embargo, haber sufrido exacerbaciones agudas durante el año previo al cuadro de COVID-19 está asociado a una mayor mortalidad, especialmente en personas edad avanzada y varones[15].

Comparando diversos estudios, se encontró que los pacientes con algún tipo de cáncer como comorbilidad tienen un 63% más de probabilidad de morir por COVID-19. La enfermedad renal crónica incrementó 3.6 veces el riesgo de muerte en pacientes que la padecen. También se determinó que padecer diabetes incrementa el riesgo de muerte asociado a COVID-19 en un 94%, en relación con pacientes sin la enfermedad, en tanto la hipertensión aumenta en 2.1 veces el riesgo de mortalidad por COVID-19, finalmente, la obesidad incrementa en un 58% el riesgo de muerte en pacientes COVID-19 [14].

Por otro lado, en un estudio en el cual se realizaron autopsias a 26 fallecidos por COVID-19, se indicó a la hipertensión arterial como la enfermedad de mayor prevalencia entre el grupo de estudio, presente en el 65.4% de los casos, seguida por obesidad (38.5%), cardiopatía isquémica crónica (34.6%), fibrilación auricular(26.9%) y EPOC (23.1%) [7].

Entre los modelos de predicción revisados se encontraron modelos basados en regresión logística, máquinas de soporte vectorial (SVM), ensambles de árboles de decisión y deep learning. Los métodos de selección de características utilizados en el desarrollo de los modelos incluyeron técnicas como LASSO, del inglés *Least absolute shrinkage and selection operator***,** análisis de

coeficientes de regresiones y criterios de importancia de árboles de decisión. En[16] los autores realizaron una selección de características usando LASSO para luego ingresar las características obtenidas en un modelo basado en *multi-tree extreme gradient boosting* (XGBoost) con el cual ranquearon la importancia de las características obtenidas con LASSO, seleccionaron las 10 mejores y entrenaron un modelo de predicción de riesgo de muerte en pacientes hospitalizados utilizando un *simple-tree* XGBoost. En [17], [18]utilizaron criterios de importancia de árbol de decisión para seleccionar variables predictoras. En ambos trabajos se entrenó un modelo basado en *Random Forest* (RF) y en el último se comparó el rendimiento del modelo con otros basados en SVM, regresión logística (RL), árbol de decisión y redes neuronales. Boot et al.2021 utilizaron regresión logística para selección de características, considerando los coeficientes asignados a cada una de ellas como medida de importancia e identificaron las 5 con mayor peso para lograr un modelo simple y parsimonioso. Teorizando que las variables seleccionadas podrían tener relaciones no lineales, eligieron un modelo basado en SVM con kernel radial para llevar a cabo sus análisis posteriores. En [12] se entrenó con regresión logística un modelo completo utilizando transformaciones no lineales con *splines* en las variables continuas, luego seleccionaron las más importantes de acuerdo al estadístico Wald y con estas se entrenó un modelo parsimonioso de regresión logística utilizando solo 5 variables.

Dentro de las predictores más utilizados se encuentran las comorbilidades, edad, sexo, recuento de linfocitos, proteína C-reactiva, temperatura, creatinina y registros de imagenología [11]. En Wynants et al. se propone considerar los predictores incluidos en los estudios analizados como candidatos a predictores en nuevas investigaciones. Esta idea es recogida por van Klavern et al.2021 quienes seleccionan variables candidatas a predictores basándose en la literatura disponible, las cuales se pueden dividir en características del paciente (sexo, edad, IMC), signos vitales (saturación de oxígeno, presión sistólica, frecuencia cardiaca, frecuencia respiratoria, temperatura) y resultados de análisis de sangre(proteína C- reactiva, deshidrogenasa láctica, dímero D, leucocitos, linfocitos, monocitos, neutrófilos, eosinófilos, volumen corpuscular medio, albúmina, bicarbonato, sodio, creatinina, urea) para luego realizar una regresión logística sobre estas y elegir las de mayor relevancia para el modelo.

Capítulo 3 Marco teórico

3.1 Introducción

A los datos extraídos desde EPIVIGILA se les aplica distintos preprocesamientos hasta obtener un set de datos con el que sea posible trabajar. Entre estos preprocesamientos se utilizarán técnicas de procesamiento natural del lenguaje, corroboración de formatos, unidades de medida, imputación de valores faltantes, dependiendo del tipo de dato que se esté trabajando. Una vez obtenido el set de datos a utilizar, se realizará una selección de características usando regresión logística con transformaciones no lineales de splines sobre las variables continuas, técnica que ayuda a captar de mejor manera las posibles relaciones entre estas variables y las de salida. Luego, utilizando el estadístico Wald chi-cuadrado se analiza la significancia de las variables predictoras y se obtiene un sub set de características, las cuales integrarán el modelo.

3.2 Descripción de la Base de Datos

Los datos adquiridos corresponden a una base de datos extraída de EPIVIGILA con previa autorización de comité ético científico. Este set de datos contiene datos demográficos, clínicos, epidemiológicos, biomarcadores, entre otros. Esta base de datos se encuentra anonimizada para proteger la identidad del paciente.

El set de datos disponible contiene 6.764.619 filas y 263 columnas. Cada fila representa a un paciente que fue ingresado al sistema en el contexto epidemiológico de COVID-19 en nuestro país, entre el 1 de enero de 2020 y el 15 de enero de 2021. Cada columna corresponde a una variable que puede ser demográfica, clínica, resultado de laboratorio o de la vida diaria, relevante para el manejo epidemiológico de la enfermedad. El detalle de las columnas incluidas se encuentra en el Anexo A.

3.2.1 Variables demográficas

Del total de pacientes 3.533.557 son hombres, 3.229.919 son mujeres. 1073 fueron reportados como desconocido y 70 como indeterminado. Esto se visualiza en la Fig. 3-1. La distribución de edades presenta una mayor densidad entre los 25 y 35 años aproximadamente como se puede apreciar en la Fig. 3-2 .

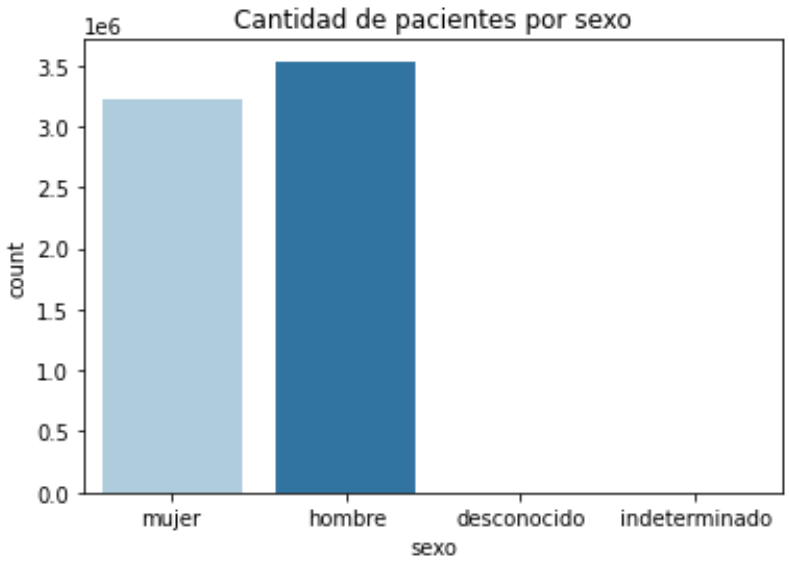


Fig. 3-1 Cantidad de pacientes por sexo.

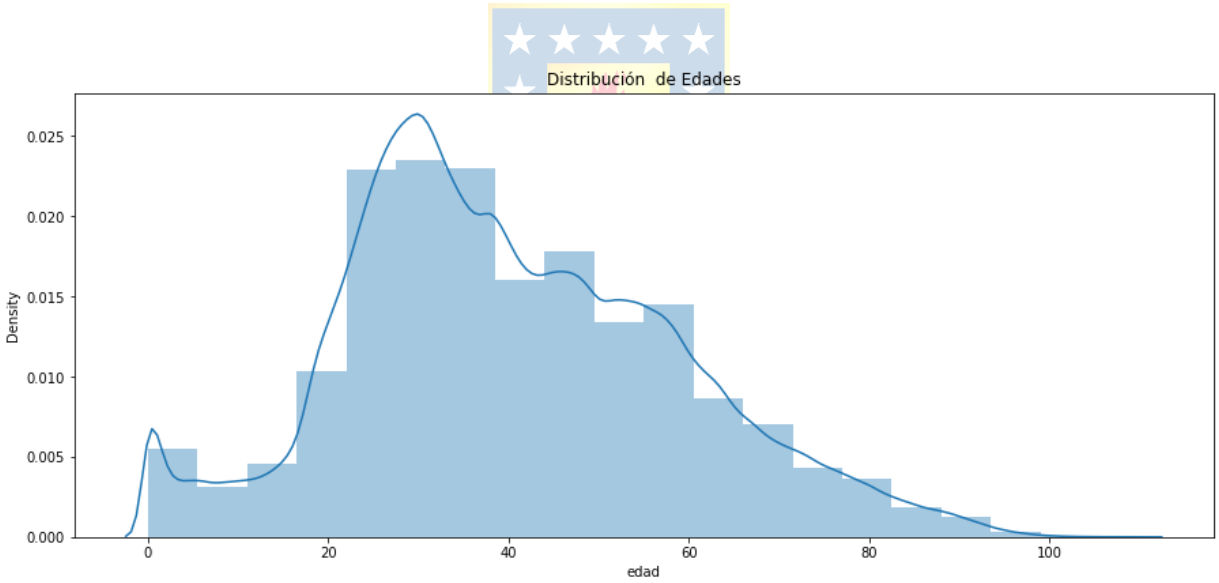


Fig. 3-2 Distribución de densidad por edades.

3.2.2 Características clínicas

En la Tabla 3-1se presentan algunas características clínicas presentes en la base de datos. Es importante considerar que en la plataforma EPIVIGILA se registran todos los casos de interés epidemiológico, en este caso, respecto de la pandemia por COVID-19. Debido a lo anterior, el 75,97% de los casos figura como descartados. Sin embargo, al filtrar los datos de acuerdo al estatus de hospitalización, aún se puede encontrar en la variable etapa clínica de estos pacientes el valor

"descartada". Dado lo anterior, no se puede filtrar asumiendo a priori que solo se encontraran pacientes que cursaron la enfermedad entre quienes aparecen como caso confirmado.

Al revisar la variable de presentación clínica se observa un gran porcentaje de pacientes fueron categorizados como asintomáticos. Esto no significa que el 63,67% de los registros correspondan a personas que cursaron una infección de forma asintomática, más bien, significa que al momento de la evaluación no presentaban síntomas, independientemente si su caso fue luego *confirmado, rechazado* o marcado como *caso probable*.

Tabla 3-1 Características clínicas: Se presenta un resumen de características clínicas relevantes que pueden servir para describir y conocer los datos presentes.

Variable	Frecuencia absoluta	Frecuencia relativa
Mortalidad		
Vivos	6.728.735	99,5%
Muertos	35.884	0,05%
Etapa clínica		
Descartada	5.139.390	75,97%
Confirmada	680.632	10,06%
Búsqueda Activa	587.146	8,68%
Sospecha	219.307	3,24%
Probable	138.144	2,04%
Hospitalización		
Si	72979	1,078%
No	6.691.639	98,922%
Paciente crítico	27.115	0,4%
Ventilación mecánica Invasiva	9.921	0,147%
Ventilación mecánica	8.748	0,13%
No invasiva		
Presentación clínica		
Sintomático	2.349.187	34,73%
Asintomático	4.306.860	63,67%

3.2.3 Biomarcadores

De acuerdo con la literatura, los biomarcadores pueden representar muy buenos predictores de resultados de la enfermedad. En algunos casos se utilizaron solo biomarcadores y en otros, luego de realizar la selección de características se encontró que algunos eran de gran relevancia para poder predecir variables de interés[11].

En la base de datos se encuentran registrados 25 biomarcadores tomados a 24 y 48 horas. Ahora bien, estos registros poseen valores perdidos por falta de registro, en especial los de 48 horas. En la Fig. 3-3 y Fig. 3-4 se resume la cantidad de registros presentes para cada uno de los biomarcadores. Se incluyeron en este análisis algunas de las características clínicas para comparar la densidad de datos.

Si bien la cantidad de registros presentes para los biomarcadores es baja, en relación a la cantidad de registros de características clínicas, no se debe pensar en eliminarlos ya que de acuerdo a la literatura pueden ser buenos predictores. También es importante considerar que el modelo final se entrenará con un sub conjunto de la base de datos original, procurando que para las variables de interés exista una buena densidad de datos. Otro punto a tener en cuenta es que por la naturaleza de los datos y de la enfermedad, no se puede esperar que una gran porción de pacientes tenga registros no nulos en las variables de biomarcadores, ya que su observación depende del estado del paciente. Por ejemplo, en un caso confirmado de baja complejidad, es probable que no exista registros de todos los biomarcadores, por otro lado, en un caso confirmado, hospitalizado y con requerimiento de ingreso en UCI o UTI, es más probable contar con registros de biomarcadores completos.

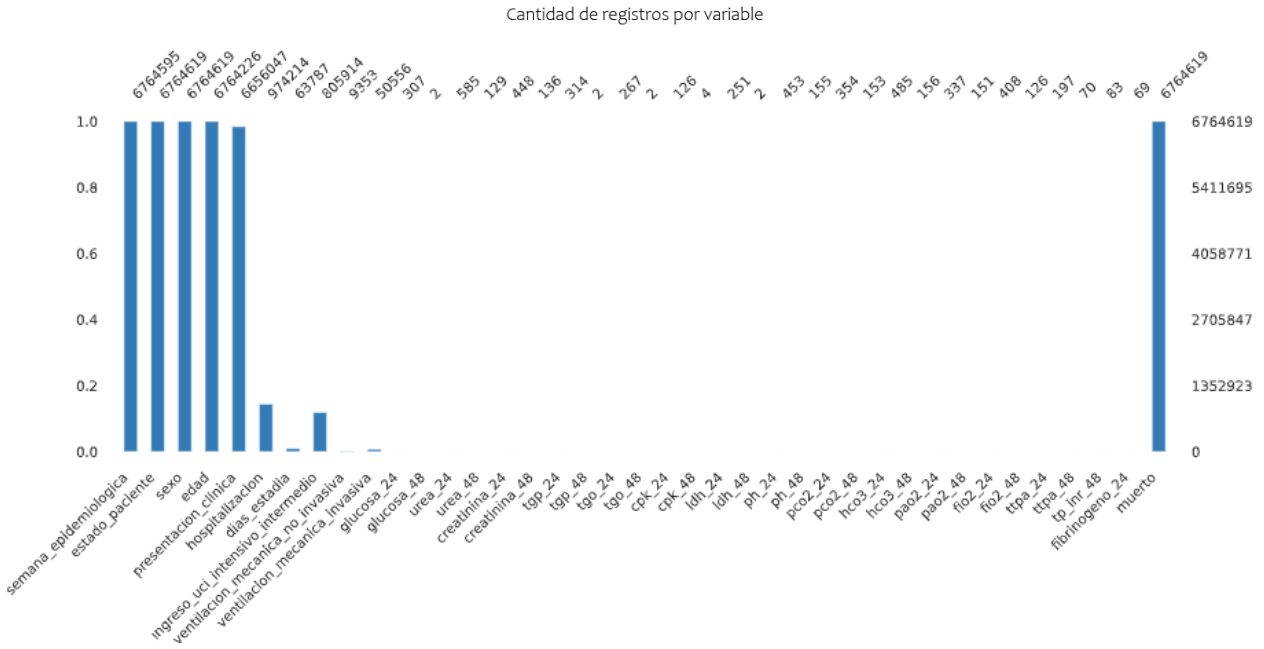


Fig. 3-3 Cantidad de registros no nulos para variables de salida y biomarcadores (I)

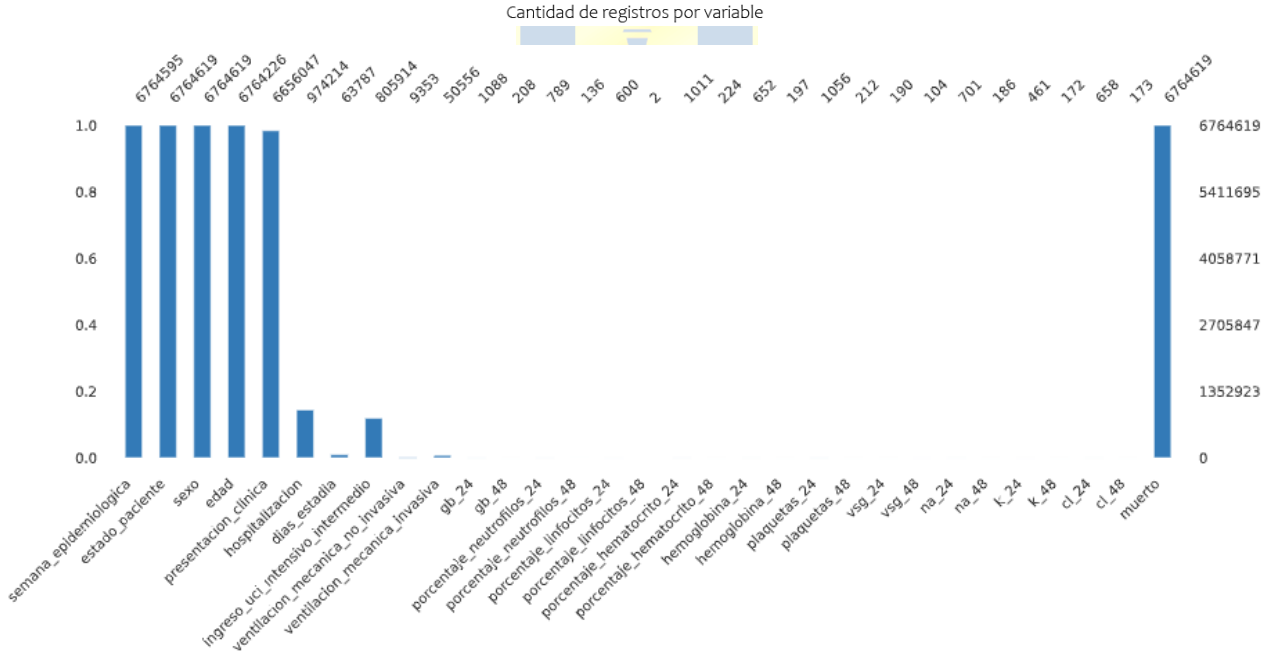


Fig. 3-4 Cantidad de registros no nulos para variables de salida y biomarcadores (II).

3.3 Preprocesamiento de la base de datos

Sobre el set de datos se realizó un pre procesado con el objetivo de limpiar, ordenar y estandarizar los datos presentes. Se revisó que, para cada variable continua las unidades de medida y escalas sean consistentes entre los pacientes. Para cada variable categórica se revisaron las categorías únicas, buscando duplicados, errores ortográficos o sinónimos que pudieran aumentar la cantidad de categorías presentes.

Para el manejo de valores faltantes se analizó el tipo de variable al que pertenece, así como la naturaleza de esta para evaluar si requiere imputación o si se reporta como valor faltante debido a que determinado paciente no presentó esa característica. Por ejemplo, pacientes con valores faltantes en la variable "ventilación mecánica invasiva" debido a que no la requirieron. Eliminar a los pacientes con valor nulo en dicha variable podría inducir riesgo de sesgo hacia pacientes de mayor gravedad, ya que son estos quienes necesitan este tipo de apoyo ventilatorio [19].

3.3.1 Variables categóricas

Dentro de las variables categóricas presentes se puede identificar dos grupos, variables categóricas dicotómicas y variables categóricas nominales. Para las variables categóricas dicotómicas se utilizaron los números 0 y 1 para etiquetar sus valores. También se revisó la consistencia de los datos, por ejemplo, para las variables hospitalización y ventilación mecánica invasiva existían variables que contenían las fechas de inicio y término. Para cada registro se analizó la existencia de registro en la variable dicotómica cuyo valor podía ser "True", "False" o "NaN", y se analizó la consistencia con los registros de fecha de inicio y término. Así se pudo corregir casos que aparecían como "False" aun teniendo fechas de inicio y/o término. También se corrigió el registro para pacientes que aparecían como ventilados, pero no como hospitalizados.

En el caso de las variables categóricas nominales como comorbilidades y síntomas, se utilizó procesamiento natural del lenguaje ya que, para cada registro, la cantidad de síntomas y comorbilidades es variable. Utilizando procesamiento natural del lenguaje se identificaron 14 síntomas de interés y 11 comorbilidades, cada uno de estos se transformó en una columna la cual se rellenó con 1 si estaba presente en el paciente y 0 si no fue observada.

3.3.2 Variables continuas

Edad, signos vitales y biomarcadores componen el grupo de variables continuas. Las variables de edad y signos vitales no necesitaron preprocesamiento previo a imputaciones y escalamiento, ya que todos sus registros presentaban consistencia, en cambio para cada biomarcador presente fue necesario revisar las escalas en las que se encontraban sus datos. Los resultados de laboratorio pueden ser entregados en distintas escalas, considerando distintas unidades de medida. Por ejemplo, la creatinina puede ser medida en mg/dL o mmol/L, y sus valores pueden ser transformados fácilmente usando un factor de conversión.

3.3.3 Reducción de la base de datos

Debido a la gran cantidad de datos faltantes y la baja densidad de registros no nulos en variables de interés, principalmente en biomarcadores, se redujo la cantidad de registros a analizar. Se consideró la cantidad de registros no nulos para el biomarcador de mayor densidad como umbral. Así, las filas de la base de datos fueron reducidas a 1.062. También se eliminaron todas las columnas que no correspondieran a las variables predictoras propuestas en la literatura y las de biomarcadores a 24 horas con una densidad de datos menor al 25% de los registros. De igual forma, se eliminaron los registros de biomarcadores a 48 horas por baja densidad de datos.

3.3.4 Imputación de valores faltantes

Los algoritmos de clasificación a utilizar requieren completitud de datos, es por esto que es necesario realizar imputación de datos faltantes para su ejecución.

Para imputar los valores faltantes se utilizó el paquete MICE en R, del inglés *Multiple Imputation by Cahined Equations*, en RStudio. MICE utiliza un método en el que cada una de las variables es considerada como objetivo y el resto se utiliza para predecir los valores faltantes. Además, el algoritmo de MICE es capaz de imputar mezclas de datos continuos, binarios y categóricos lo que resulta muy útil para la base de datos utilizada.

En la Fig. 3-5 se presenta la distribución de densidad para algunas de las variables imputadas. Con MICE se generaron 5 set de datos imputados los que luego fueron promediados para obtener solo uno. Cada uno de los sets imputados está representado con magenta, los datos originales se presentan en azul. Se observa que las distribuciones de los datos imputados son muy similares a las de los datos

reales. La distribución de densidad para todas las variables imputadas puede ser revisada en el Anexo B.

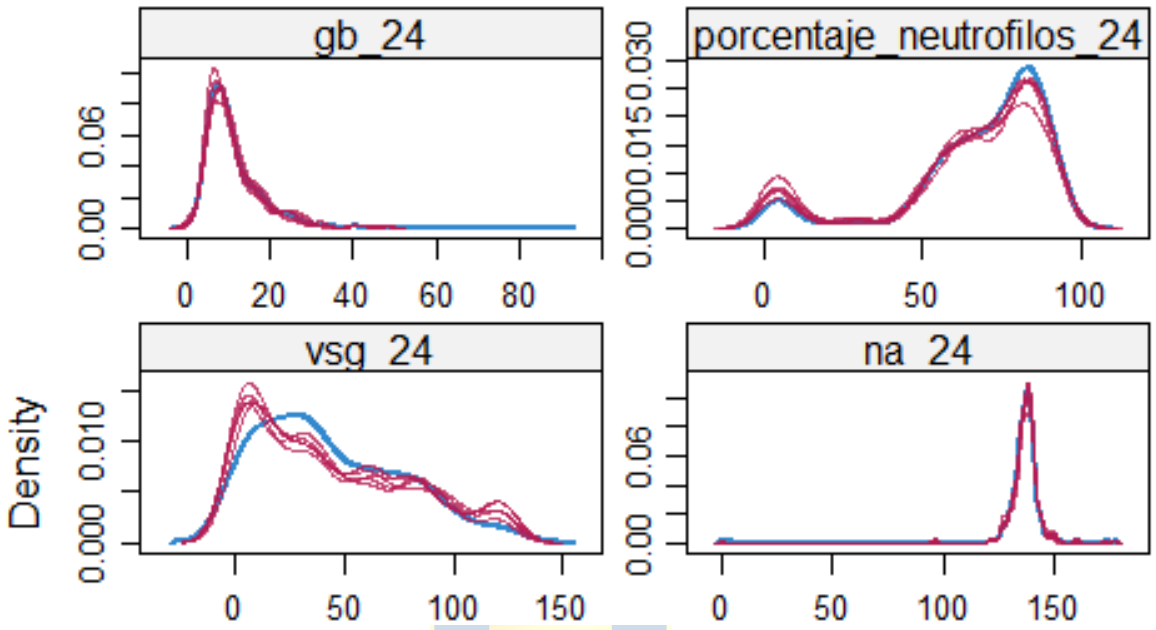


Fig. 3-5 Distribuciones de densidad para variables imputadas. Se observan en rojo las distribuciones de densidad obtenidas luego de imputación con MICE y en azul las distribuciones de densidad originales.

3.3.5 Estandarización

Los datos fueron estandarizados usando el valor z de acuerdo a la ecuación 3-1

$$Z = \frac{x - \mu}{\sigma} \qquad (3\text{-}1)$$

donde:

x : valor de la variable

μ : media de la variable

σ : desviación estándar de la variable.

Al usar esta estandarización, las distintas variables pueden ser comparadas entre sí por los algoritmos sin entregarle mayor importancia a alguna solo porque su orden de magnitud sea mayor.

3.3.6 Set de datos a utilizar en la creación del modelo

Luego del preprocesamiento, el set de datos a utilizar quedó definido como un *dataframe* de 1.062 filas y 55 columnas, de las cuales 3 corresponden a las variables objetivo de hospitalización, ingreso a UCI y ventilación mecánica invasiva.

La distribución de casos positivos y negativos para cada una de las variables de salida se observa en la Fig. 3-6. De acuerdo a lo mostrado, para todas las variables objetivo las clases se encuentran desbalanceadas, siendo el caso de la variable ventilación mecánica invasiva el con mayor desbalance, contando con solo un 15% de casos positivos.

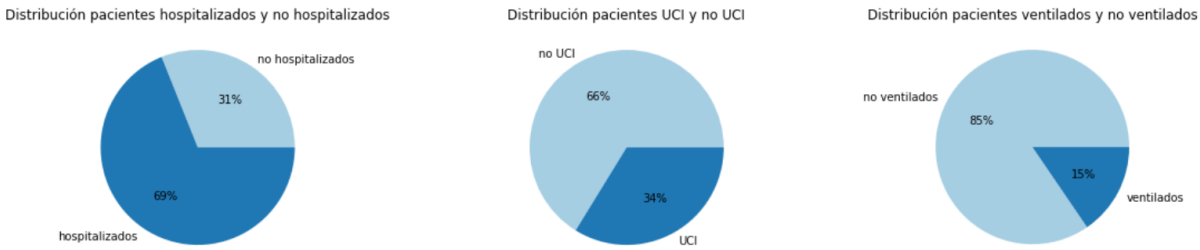


Fig. 3-6 Distribución de casos positivos y negativos para variables de salida. Se observa desbalance de clases para todas las salidas.

3.4 Selección de características

Para realizar selección de características se utilizó regresión logística. Considerando las sugerencias de la literatura, se incorporaron 52 variables candidatas a variable predictoras, de las cuales 5 corresponden a signos vitales, 20 a biomarcadores, 11 a comorbilidades, 14 a síntomas y las dos restantes edad y sexo.

Para tomar en consideración las posibles interacciones no lineales entre las variables predictoras y las variables de salida, se utilizaron transformaciones de splines cúbicos restringidos en cada una de las variables continuas [20].

Los datos se dividieron en set de entrenamiento y prueba, dejando un 30% de los datos para prueba. Se utilizó validación cruzada con *repeated 3-Fold Cross Validation*, con numero de

repeticiones igual a 100. Este procedimiento se hizo de forma individual para cada una de las variables de salida.

Utilizando el estadístico Wald chi-cuadrado se analizó la importancia de las variables predictoras. Wald chi-cuadrado se usa para probar la hipótesis de que al menos uno de los predictores tiene coeficiente no nulo. Este test se realizó utilizando el paquete caret en R.

Para la variable de salida hospitalización, las variables más importantes de acuerdo con el estadístico Wald chi-cuadrado fueron: edad, frecuencia respiratoria, frecuencia cardiaca, presión sistólica, cefalea, fiebre, asma, inmunocomprometido, enfermedad cardiovascular, enfermedad pulmonar crónica, enfermedad renal crónica, glóbulos blancos, porcentaje de hematocrito, creatinina, alanina-aminotransferasa (TGP), potencial de hidrógeno (pH), presión parcial de oxígeno (PaO2), fracción inspirada de oxígeno (FiO2) e hipertensión.

Para variable objetivo de ventilación mecánica, los atributos seleccionados fueron: Edad, presión sistólica, cefalea, disnea, fiebre, anosmia, mialgia, ageusia, diarrea, dolor abdominal, taquipnea, dolor torácico, asma, cardiopatía, enfermedad cardiovascular, enfermedad pulmonar crónica, enfermedad renal crónica, enfermedad neurológica crónica, enfermedad hepática crónica, porcentaje de neutrófilos, porcentaje de linfocitos, porcentaje de hematocrito, velocidad de segmentación globular (VSG), sodio, cloro, glucosa, TGP, aspartato aminotransferasa (TGO), creatinina fosfoquinasa (CPK), pH, PaO2, tiempo parcial de tromboplastina activada (TTPa), potasio, diabetes, hipertensión , tos y postración.

Para el resultado de muerte las características más importantes fueron: Edad, sexo, frecuencia cardiaca, presión diastólica, cefalea, anosmia, ageusia, mialgia, dolor abdominal, taquipnea, dolor torácico, obesidad, inmunocomprometido, enfermedad cardiovascular, enfermedad renal crónica, enfermedad hepática crónica, porcentaje de neutrófilos, hemoglobina, VSG, cloro, glucosa, urea, TGO, lactato deshidrogenasa (LDH), PaO2, TTPa, hipertensión y diabetes.

En el Anexo C se encuentran los valores del estadístico Wald chi- cuadrado para cada variable considerando interacciones lineales y no lineales según corresponda, para cada una de las tres variables objetivo.

Capítulo 4 Modelos de Aprendizaje Supervisado

4.1 Introducción

Para predecir las variables objetivo, se realizaron 3 modelos de aprendizaje supervisado para cada una de las variables objetivo, utilizando distintos algoritmos de aprendizaje automático, a saber, regresión logística, *random forest* y redes neuronales.

Los modelos fueron diseñados de acuerdo con el algoritmo presentado en la Fig. 4-1. Para todos los casos se dividieron los datos en sets de entrenamiento y prueba, considerando el 30% de estos como set de prueba utilizando la función de scikit-learn train_test_split. Para el proceso de ajuste de parámetros se incluyó validación cruzada de k-iteraciones estratificadas repetidas con k = 3 y 100 repeticiones.

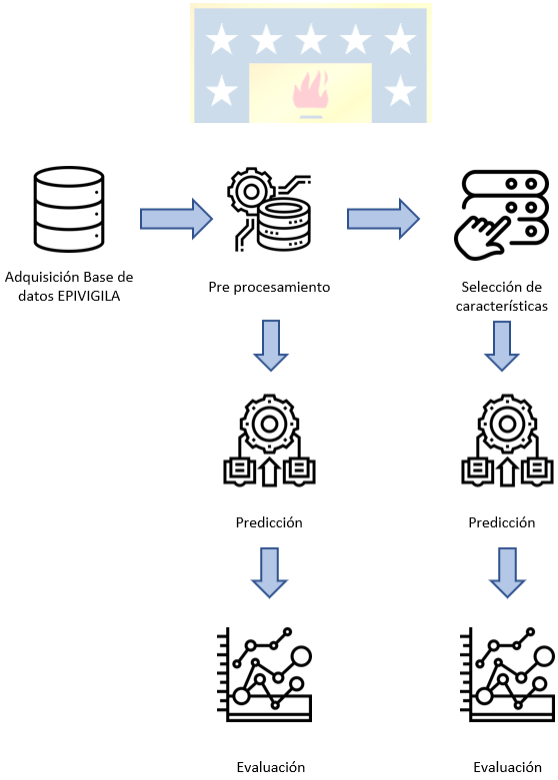


Fig. 4-1 Algoritmo de predicción.

4.2 Regresión logística

El modelo de regresión logística fue construido utilizando el paquete scikit-learn de Python como se muestra en la Fig. 4-1. En primera instancia se llevó a cabo el ajuste de los parámetros utilizando el set de entrenamiento. Se ajustaron los "C", correspondiente al inverso del parámetro de regularización λ, "*solver*" y "*penalty*" o regularización, la que podía ser de LASSO (L1) o RIDGE (L2); para esto se utilizó la función *GridSearchCV* y se encontraron los parámetros listados en la Tabla 4-1.

Tabla 4-1 Parámetros ajustados para regresión logística.

Salida	C	Solver	Penalty
Hospitalización	0,01	Newton-cg	L2
Ingreso UCI	0,1	liblinear	L2
Ventilación mecánica.	10	liblinear	L1

Una vez ajustados los parámetros, se entrenó el modelo con todo el set de entrenamiento y las 52 variables predictoras. Luego se entrenó considerando solo las variables seleccionadas en 3.4.

4.3 Random Forest

Al igual que los modelos de regresión logística, los de RF fueron construidos usando scikit-learn, siguiendo los pasos mencionados en 6.1. Los parámetros a ajustar fueron "criterion", correspondiente a el criterio para evaluar la calidad de la división en cada nodo, "max_depth", correspondiente a la profundidad máxima para cada árbol y "class_weight", parámetro que indica el manejo de pesos para las muestras de cada clase. Los parámetros ajustados se muestran en la Tabla 4-2.

Tabla 4-2 Parámetros ajustados para Random Forest.

Salida	Criterion	max_depth	class_weight
Hospitalización	entropía	11	balanced_subsample
Ingreso UCI	entropía	5	balanced_subsample
Ventilación mecánica.	gini	1	balanced_subsample

4.4 Redes Neuronales

Para construir los modelos de redes neuronales se utilizó la plataforma TensorFlow y su API Keras en Python. Los modelos definidos fueron perceptrones multicapa (PMC), compuestos por una capa de entrada, una capa oculta y una capa de salida.

Se probaron distintas configuraciones y se ajustaron los parámetros de forma manual. Se usó validación interna, destinando un 30% del set de entrenamiento para validación. Las configuraciones seleccionadas se resumen en la Tabla 4-3.

Tabla 4-3 Configuraciones Redes Neuronales.

Salida	Capas ocultas	Función de activación	Optimizador	Nodos
Hospitalización	1	Elu	Nadam	34
Ingreso UCI	1	Elu	Nadam	34
Ventilación mecánica.	1	Relu	RMSprop	34

En el caso de los modelos de *deep learning*, una vez ajustados no se volvieron a entrenar con las características seleccionas con el método de selección de características, ya que este tipo de modelos es muy dependiente de la cantidad de entradas, influyendo tanto en la cantidad de capas, neuronas, funciones de activación y funciones optimizadoras necesarias, por lo cual habría que ajustar nuevamente dichos parámetros obteniendo modelos distintos. Luego, la comparación entre estos modelos y los iniciales no tendría elementos constantes.

4.5 Evaluación

El rendimiento de cada uno de los modelos fue evaluado utilizando diversas métricas derivadas de la matriz de confusión. Al tratarse de clasificaciones en datos desbalanceados y con pocos ejemplos disponibles, es necesario analizar los resultados considerando diversas métricas.

Las métricas seleccionadas fueron exactitud, precisión, sensibilidad y especificidad. Además, se graficaron las curvas de Precisión-Sensibilidad y ROC para realizar comparaciones entre modelos.

4.5.1 Matriz de confusión

La matriz de confusión (ver Fig. 4-2) es una herramienta útil para resumir el comportamiento de un clasificador. En ella se identifican cuatro casos posibles para las etiquetas predichas, respecto de su valor real.

- Verdaderos positivos (VP): Los casos verdaderos positivos son los que tienen etiqueta predicha positiva y etiqueta real positiva.
- Falsos positivos (FP): Los casos falsos corresponden a los que tienen etiqueta predicha positiva y etiqueta real negativa.
- Verdaderos negativos (VN): Los casos verdaderos negativos corresponden a los que tienen etiqueta predicha negativa y etiqueta real negativa.
- Falsos negativos (FN): Los casos falsos negativos corresponden a los que tienen etiqueta predicha negativa y etiqueta real negativa.

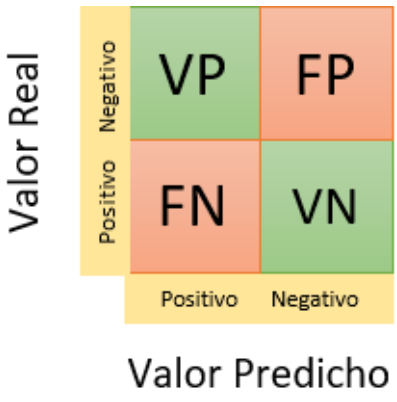


Fig. 4-2 Matriz de confusión.

4.5.2 Métricas derivadas de la matriz de confusión

A partir de los cuatro casos posibles definidos por la matriz de confusión se derivan diversas métricas que ayudan a evaluar el rendimiento de los modelos de clasificación supervisada. En este trabajo se utilizarán las métricas exactitud, precisión, sensibilidad y especificidad.

4.5.2.1 Exactitud

La métrica de exactitud entrega el porcentaje de casos que el modelo ha etiquetado correctamente. La exactitud está definida por la ecuación 4-1.

$$\text{Exactitud} = \frac{(\text{VP} + \text{VN})}{(\text{VP}+\text{VN}+\text{FP}+\text{FN})} \times 100 \quad \textbf{(4-1)}$$

4.5.2.2 Precisión

La métrica precisión, también llamada valor predictivo positivo, indica que porcentaje de los casos etiquetados como positivos corresponde realmente a casos y está definida por la ecuación 4-2.

$$\text{Precisión} = \frac{\text{VP}}{(\text{VP}+\text{FP})} \times 100 \quad \textbf{(4-2)}$$

4.5.2.3 Sensibilidad

La métrica sensibilidad, también conocida como exhaustividad, entrega el porcentaje de casos positivos que el modelo de aprendizaje automático es capaz de identificar y está definida por la ecuación 4-3.

$$\text{Sensibilidad} = \frac{\text{VP}}{(\text{VP}+\text{FN})} \times 100 \quad \textbf{(4-3)}$$

4.5.2.4 Especificidad

La especificidad indica la capacidad de un clasificador para identificar correctamente los casos negativos.

$$\text{Especificidad} = \frac{\text{VN}}{\text{VN} + \text{FP}} \times 100 \quad \textbf{(4-4)}$$

4.6 Curva ROC

La curva ROC, del inglés *reciever operating characteristic curve,* es una herramienta estadística que puede ser usada para analizar la capacidad discriminante de un modelo de clasificación. En ella se grafican sensibilidad versus 1- especificidad para distintos valores de umbral. La curva ROC resume las compensaciones entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos.

Una de sus ventajas es que toma en consideración la distribución de clases, dándole más peso a las clasificaciones correctas de la clase minoritaria.

Para poder cuantificar el performance global del clasificador es necesario calcular el área bajo la curva (AUC). Este valor representa el rendimiento del clasificador sobre todos los escenarios posibles de costo de clasificación.

4.7 Curva Precisión-Sensibilidad

La curva de precisión-sensibilidad (PRC) resume las compensaciones entre sensibilidad y precisión de un modelo clasificador, usando distintos umbrales de probabilidad.

Una de sus ventajas es que, en el caso de clasificación con clases desbalanceadas, en particular con clase positiva minoritaria, ofrece una imagen más real que la curva ROC del rendimiento de un clasificador ya que, tanto para precisión como para sensibilidad, no se toman en consideración los verdaderos negativos y se centra en la evaluación de la correcta identificación de la clase positiva minoritaria.

Capítulo 5 Análisis de resultados

5.1 Introducción

El rendimiento de cada uno de los modelos fue evaluado utilizando diversas métricas derivadas de la matriz de confusión. Al tratarse de clasificaciones en datos desbalanceados y con pocos ejemplos disponibles, es necesario analizar los resultados considerando distintos puntos de vista.

La evaluación de los modelos se basará en la comparación de desempeño de los modelos diseñados además del efecto del método de selección de características empleado en la variación de las métricas consideradas.

Las métricas seleccionadas fueron exactitud, precisión, sensibilidad y especificidad. Además, se graficaron las curvas de Precisión-Sensibilidad y ROC para realizar comparaciones entre modelos.

5.2 Resultado clínico: Hospitalización

Las métricas obtenidas para la variable de salida hospitalización se resumen en la Tabla 5-1 y Tabla 5-2.

Tabla 5-1 Métricas resultado clínico Hospitalización, todas las características.

Modelo	Exactitud	Sensibilidad	Precisión
RL	73,7	76,2	85,3
RF	82,4	94	83,5
PMC	81,9	90,9	84,8

Tabla 5-2 Métricas resultado clínico Hospitalización, características seleccionadas

Modelo	Exactitud	Sensibilidad	Precisión
RL	75,9	76,6	88,1
RF	83,3	90,5	86,7

Tanto para regresión logística como RF se observa una mejora en exactitud y precisión al considerar las características seleccionadas con el método de splines, ver Tabla 5-2. En RF la sensibilidad cae 3,5 puntos, pero la precisión aumenta en 3,2. Para regresión logística mejoran todas las métricas. En particular, el aumento de la precisión indica una mejora en la capacidad de discriminación de los clasificadores, de forma estadística, aumenta la probabilidad de que un caso clasificado como positivo (requiere hospitalización), realmente lo sea. Sin embargo, la disminución en la sensibilidad de RF, revela un deterioro en la capacidad de identificar los verdaderos positivos por parte del clasificador, en este caso, para identificar los pacientes que realmente requerirán hospitalización.

De acuerdo con la Tabla 5-1, RF presenta mayor exactitud y sensibilidad que RL y PMC, sin embargo, ambos modelos superan a RF en precisión.

A pesar de las diferencias en las métricas, todos los modelos presentan un rendimiento aceptable. Esto puede deberse a que, en este caso, la clase predominante es la positiva, sin embargo, es evidente que todos los modelos presentan cierta dificultad para identificar entre clases, esto se observa en las compensaciones evidenciadas al reducir la dimensión del espacio de características. Si bien aumenta la precisión y exactitud, disminuye la sensibilidad.

En general, el ajuste entre precisión y sensibilidad es un fenómeno esperado y al analizar las matrices de confusión de la Fig. 5-1 se evidencian ciertos *intercambios* entre ambas métricas.

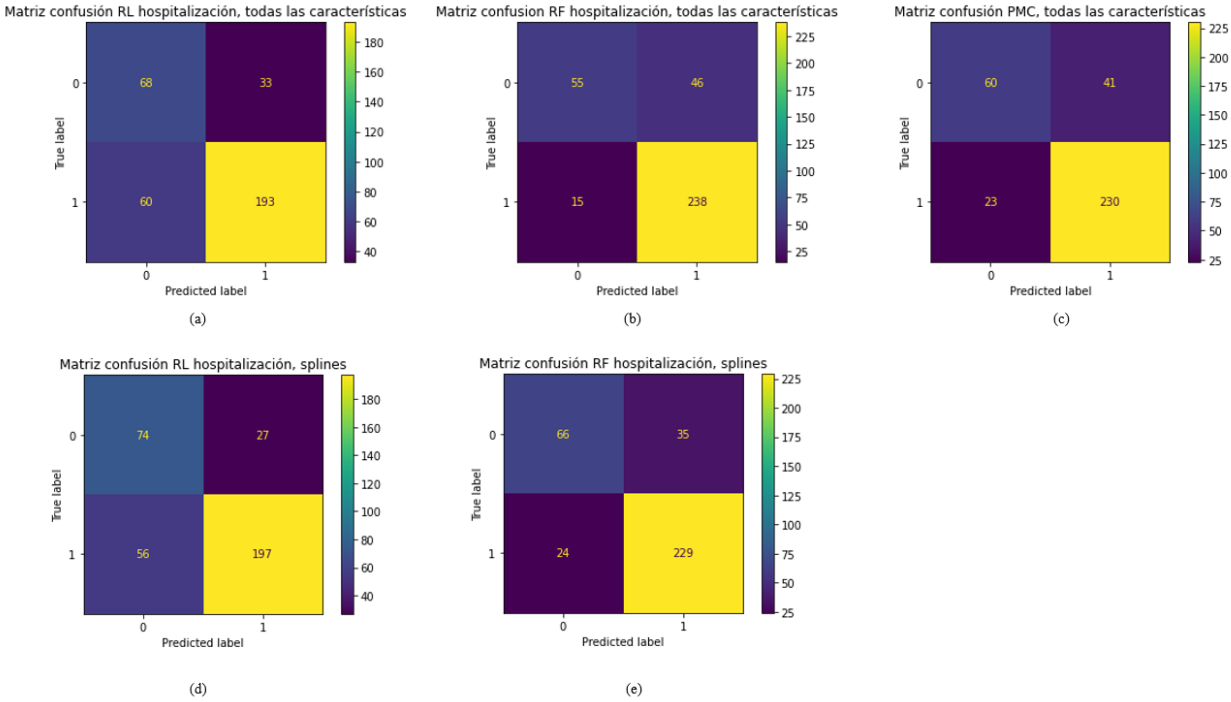


Fig. 5-1 Matrices de confusión para salida de Hospitalización, todas las características y dimensión reducida.

Como se puede ver en la Fig. 5-1, *con Random Forest* se obtienen más verdaderos positivos en todos los casos. A pesar de esto, también se puede apreciar que tanto para el caso con todas las características como para el que tiene espacio de características reducido, la regresión logística es capaz de identificar de mejor forma la clase negativa que RF, obteniendo así menos falsos positivos (FP) en ambos casos.

De (a), (b) y (c) se aprecia que PMC presenta un rendimiento que podría calificarse de intermedio entre el de los otros dos clasificadores. Vemos que, es capaz de obtener un 19% más de verdaderos positivos que la regresión logística con aproximadamente el 38% de sus falsos negativos. Respecto de *Random Forest*, la red neuronal obtiene un 3,5% menos de verdaderos positivos y consigue una reducción de aproximadamente un 5% en los falsos positivos.

Al observar la especificidad (Tabla 5-3) de los modelos, podemos ver que, para ambos espacios de características, regresión logística es el modelo que presenta mejor valor en esta métrica. Esto indica que regresión logística es el modelo que mejor identifica la clase negativa.

Tabla 5-3 Especificidad Hospitalización, todos los modelos.

Modelo	Especificidad
RL, todas las características	0,67
RF, todas las características	0,54
PMC, todas las características	0,59
RL, splines	0,73
RF, splines	0,65

Para poder integrar los análisis anteriores y presentar un resumen de forma gráfica que ayude a comprar e interpretar de forma global el rendimiento de los distintos clasificadores, se calcularon las curvas ROC y PRC, desplegadas en la Fig. 5-2 y Fig. 5-3.

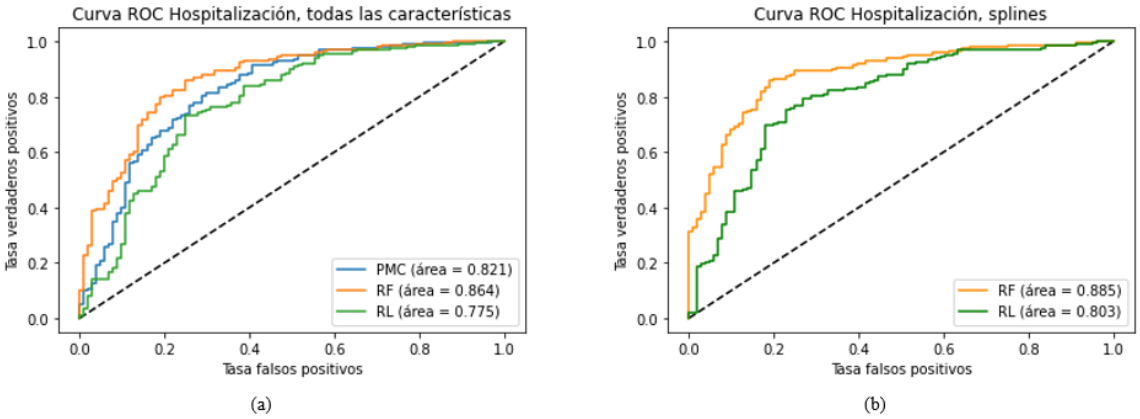


Fig. 5-2 Curvas ROC, resultado clínico Hospitalización.

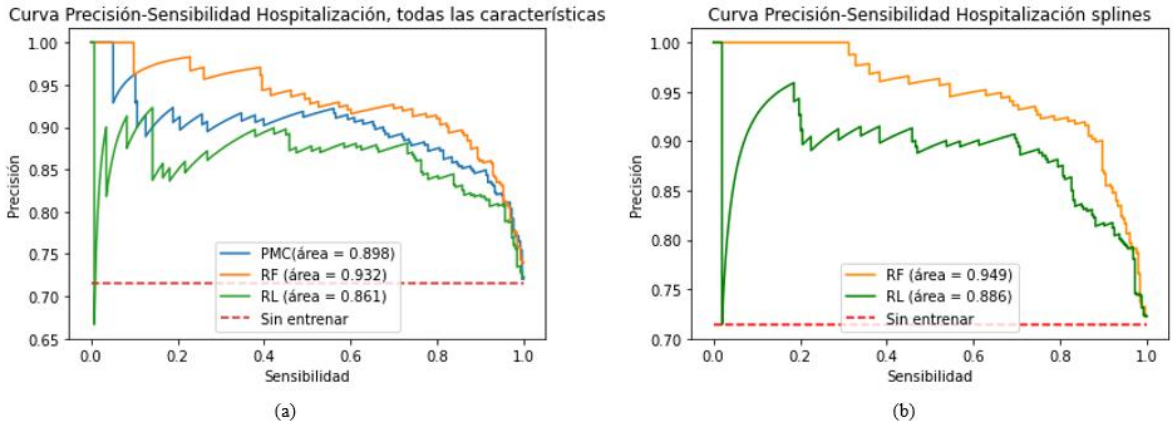


Fig. 5-3 Curvas PRC, resultado clínico Hospitalización.

Al comparar los puntajes AUC para todas las curvas, es posible apreciar que RF es el modelo que presenta mejor rendimiento. Además, al comparar los puntajes para RF y RL antes y después de la reducción del espacio de características, se evidencia una mejora en los clasificadores.

Se considera necesario analizar ambas curvas para evaluar el desempeño de los clasificadores ya que cada una entrega información importante respecto del poder de clasificación de los modelos. Las curvas ROC, al entregar la relación entre las tasas de verdaderos positivos y falsos positivos nos dan una idea de la calidad de la clasificación general, considerando ambas clases, mientras la curva PRC se concentra exclusivamente en la clase positiva, excluyendo los VN del análisis de rendimiento. La precisión es también llamada valor predictivo positivo, que puede ser interpretado, en este caso, como la probabilidad de que un paciente que fue etiquetado como con necesidad de hospitalización, realmente lo requiera.

5.3 Resultado clínico: Ingreso UCI

Las métricas obtenidas para la variable de salida ingreso UCI se resumen en la Tabla 5-4 y Tabla 5-5.

Tabla 5-4 Métricas resultado clínico ingreso UCI, todas las características.

Modelo	Exactitud	Sensibilidad	Precisión
RL	70,3	73,7	55,2
RF	75,4	57,3	66,6
PMC	83,0	70,4	78,1

Tabla 5-5 Métricas resultado clínico ingreso UCI, características seleccionadas.

Modelo	Exactitud	Sensibilidad	Precisión
RL	72,3	81,1	56,8
RF	77,9	81,9	64,1

Analizando las métricas contenidas en la Tabla 5-4 y Tabla 5-5 se observa que PMC es el modelo con el que se obtiene mejor exactitud y precisión, siendo superado solo en sensibilidad por RL. RL presenta un valor de 55,2 en precisión, esto significa que de los casos etiquetados como positivos, solo el 55,2% realmente lo era, a pesar de poder recuperar el 73,7% de los casos positivos de forma correcta. Una vez más se evidencian los *intercambios* entre precisión y sensibilidad. Si bien es complejo obtener buenos valores para ambas métricas, es necesario establecer ciertos umbrales de acuerdo a la utilidad que se le dará a los modelos. En el caso de RF, la sensibilidad de 57,3 indica que solo dicho porcentaje de casos positivos fue correctamente identificado por el clasificador. Esto significa que el modelo tiene bajo rendimiento la hora de recuperar la clase positiva.

En la Tabla 5-5 se puede observar el efecto de la reducción del espacio de características para regresión logística y *Random Forest*. Es evidente la mejora en todas las métricas, sin embargo, para obtener altas sensibilidades, se sacrifica la precisión y a pesar de poder identificar en ambos casos aproximadamente el 81% de los casos positivos de forma correcta; del total de casos etiquetados como con necesidad de ingreso en UCI, hay un porcentaje importante que no lo necesitaba. Para poder observar de mejor forma estas relaciones, en la Fig. 5-4 se presentan las matrices de confusión para la salida de ingreso UCI.

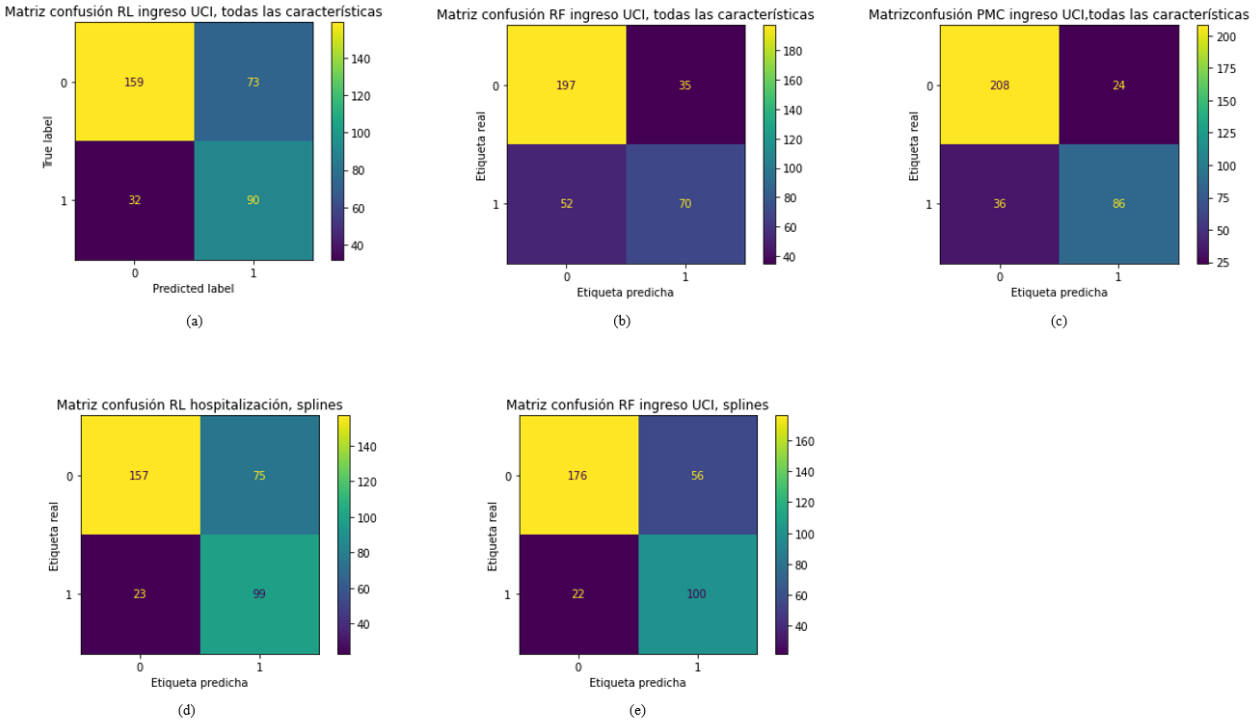


Fig. 5-4 Matrices de confusión para resultado clínico ingreso UCI, todas las características y dimensión reducida.

Para el caso de clasificación con todas las características, RL es el modelo que entrega mayor cantidad de VP, sin embargo, es también el que entrega más FP. En la Fig 5-4 (b) es posible observar la baja sensibilidad de RF, siendo capaz de recuperar solo 70 VP y obteniendo la mayor cantidad de FN con 52 entradas en dicha categoría. PMC obtiene aprox. un 95% de los VP de RL, no obstante, es capaz de reducir la cantidad de FP en un 67,12%, transformándose en el clasificador con mejor rendimiento para el resultado clínico de ingreso UCI al utilizar todas las características.

Tabla 5-6 Especificidad resultado clínico ingreso UCI, todos los modelos.

Modelo	Especificidad
RL, todas las características	0,685
RF, todas las características	0,849
PMC, todas las características	0,896
RL, splines	0,676
RF, splines	0,758

Con la Tabla 5-6, podemos comparar el rendimiento de los distintos modelos respecto de su capacidad para identificar correctamente los casos negativos. Debemos recordar que ingreso UCI presenta un desbalance de clases similar al de hospitalización (ver Fig. 3-6), solo que, en este caso, la clase positiva es la minoritaria. Al comparar las especificidades obtenidas para el resultado clínico de hospitalización, con las obtenidas para el resultado clínico de ingreso UCI, se observa una mejor capacidad de identificación de para el resultado clínico de ingreso UCI. Esto puede atribuirse al desbalance de clases. En el caso de ingreso UCI, hay muchos más ejemplos de clase negativa con los cuales alimentar a los modelos, mejorando el nivel de aprendizaje en esta categoría.

Para resumir y evaluar el rendimiento global de los modelos se presentan las gráficas de las curvas ROC y PRC para la salida de ingreso UCI en la Fig. 5-5 y Fig. 5-6, respectivamente.

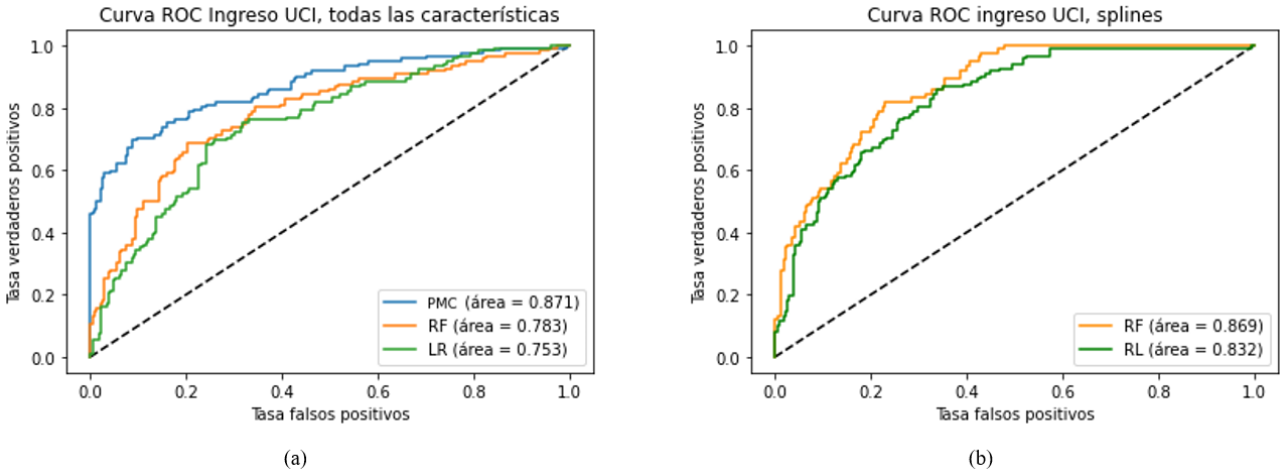


Fig. 5-5 Curvas ROC resultado clínico ingreso UCI

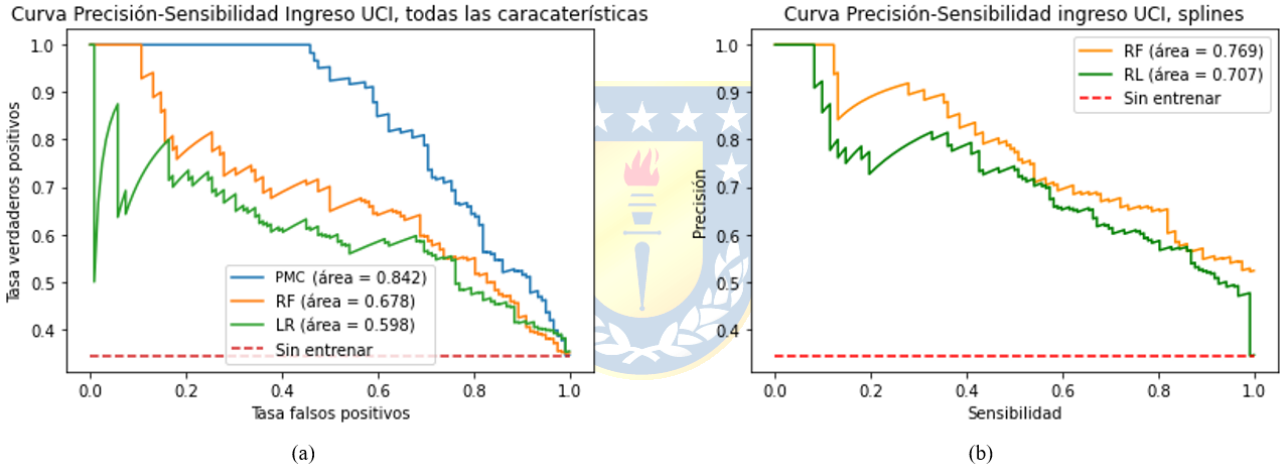


Fig. 5-6 Curvas PRC resultado clínico ingreso UCI.

Al comparar las curvas ROC y PRC de las figuras anteriores, se observa el "optimismo" de las curvas ROC frente a clases desbalanceadas que se menciona en [21]. Si solo se utilizara la curva ROC, se podría concluir que el rendimiento de los clasificadores es mucho mejor que el real e incluso concluir que la reducción del espacio de características tuvo un efecto muy positivo en este, mas, este optimismo se diluye al observar las curvas PRC.

Para el caso con todas las características es evidente que el PMC es el modelo con mejor rendimiento; esto puede observarse tanto en la curva ROC como en la PRC, sin embargo, al analizar esta última, podemos notar una caída sostenida y con pendiente considerable de la precisión a medida que aumenta la sensibilidad. La precisión se mantiene constante en 100% hasta un umbral de

sensibilidad de aproximadamente 45%, punto en el cual comienza a decaer la gráfica con una pendiente que podría considerarse constante.

En el caso con selección de características, la curva PRC presenta una disminución en la pendiente respecto de su par con todas las características, aumentando el área bajo la curva en un 13,42% para RF y en un 18,22% para RL, no obstante, debemos recordar que el aumento de la sensibilidad en este caso, viene aparejada con un aumento no despreciable de los falsos positivos.

5.4 Resultado clínico: Ventilación mecánica invasiva

Las métricas obtenidas para la variable de salida ventilación mecánica invasiva se muestran en la Tabla 5-7 y Tabla 5-8.

Tabla 5-7 Métricas resultado clínico Ventilación mecánica invasiva, todas las características.

Modelo	Exactitud	Sensibilidad	Precisión
RL	72	69	31,6
RF	73,4	56,3	30,6
PMC	92,6	69	80,8

Tabla 5-8 Métricas resultado clínico Ventilación mecánica invasiva, características seleccionadas.

Modelo	Exactitud	Sensibilidad	Precisión
RL	72,3	78,1	33,3
RF	72,3	67,2	31,6

De la Tabla 5-7 se concluye que, al considerar las 52 características, el mejor rendimiento se obtiene con el PMC. RL presenta sensibilidad y exactitud aceptables, pero el valor de precisión obtenido es muy bajo. RF tiene sensibilidad de tan solo 56,3 indicando que la cantidad de casos positivos recuperado de forma exitosa por este clasificador, apenas sobrepasa la mitad del total. Tanto para regresión logística como *Random Forest* se obtienen precisiones que no sobrepasan el 32%. Esto indica que, para lograr las sensibilidades mostradas, se genera una gran cantidad de falsos positivos.

Con la Tabla 5-8 podemos apreciar el efecto del método de selección de características en el desempeño de RL y RF. En ambos casos la sensibilidad aumenta en aproximadamente 10 puntos porcentuales, sin embargo, el aumento en la precisión es marginal. En este caso podemos notar que la reducción del espacio de características contribuyó en una mejora de la capacidad de los clasificadores para identificar la clase positiva de forma correcta, a pesar de ello, no tuvo una influencia importante en el valor predictivo positivo de los clasificadores.

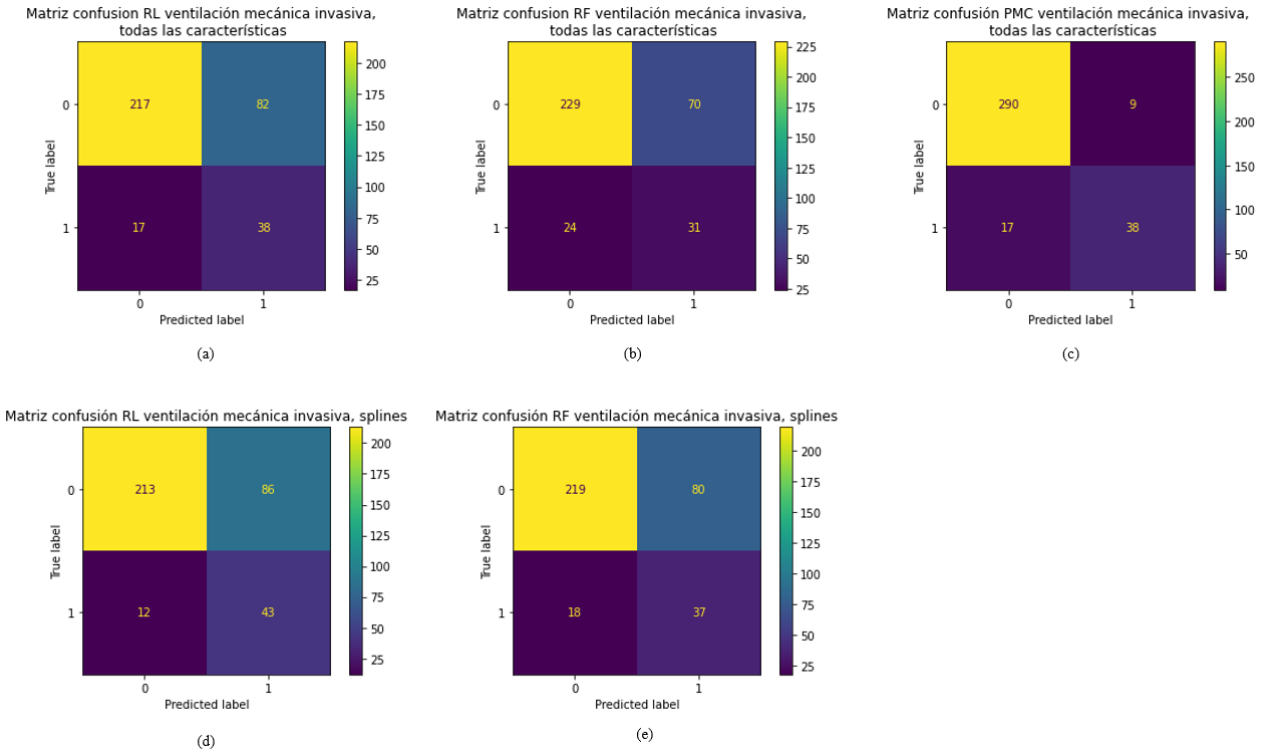


Fig. 5-7 Matrices de confusión resultado clínico Ventilación mecánica invasiva, todas las características y dimensión reducida.

De la Fig. 5-7 podemos apreciar que PMC es el modelo que posee mejor balance entre todas las componentes de la matriz de confusión. De (a) y (c) podemos observar que RL y PMC poseen igual cantidad de VP, sin embargo, con PMC se logra una reducción en los FP de un 89% respecto de los obtenidos con regresión logística. Con (a) y (b) se puede ver que con *Random Forest* se obtienen menos FP que con el modelo de RL, no obstante, se produce un aumento en los FN evidenciando

nuevamente los comentados *intercambios o compensaciones* entre las componentes de la matriz de confusión y su consecuente efecto en las métricas que de ella de derivan. Este fenómeno vuelve a quedar claro en (d) y (e), donde se observa que, al considerar las características identificadas con el método de selección de características, se produce un aumento en la cantidad de VP para RL y RF, que, aunque sutil en cantidad, se ve reflejado como un aumento porcentual de la sensibilidad de aproximadamente 10 puntos para cada caso. Debemos recordar que la variable de salida ventilación mecánica invasiva es la que presenta mayor desbalance, contando con solo un 15% de casos positivos.

Tabla 5-9 Especificidad resultado clínico Ventilación mecánica invasiva, todos los modelos,

Modelo	Especificidad
RL, todas las características	0,725
RF, todas las características	0,765
PMC, todas las características	0,969
RL, splines	0,712
RF, splines	0,732

Como se muestra en la Tabla 5-9, todos los modelos presentan buena especificidad y evidentemente, con PMC se obtiene un puntaje sobresaliente. Estos buenos resultados no deberían ser muy sorprendentes ya que como se comentó anteriormente, en problemas de clasificación desbalanceada con clase positiva minoritaria, los clasificadores presentan facilidad para identificar la clase negativa mayoritaria. A pesar de ello, la alta especificidad de PMC está dada no solo por su gran capacidad para identificar la clase negativa, sino también por un muy buen desempeño en la clase positiva.

El buen rendimiento del clasificador PMC se evidencia también en sus curvas ROC y PRC, así como en sus puntajes de área bajo la curva para ambas gráficas. En la Fig. 5-8 (a) se distingue el buen rendimiento de PMC frente a RL y RF, los cuales nuevamente evidencian el efecto optimista que se comentó anteriormente. Al revisar (b) este efecto es aún más evidente, ya que tanto RL como

RF ambos clasificadores aumentan su área bajo la curva con lo que se puede concluir una mejora en el rendimiento global de los modelos al considerar solo las variables seleccionadas con el método de selección de características empleado. Sin embargo, en este punto es importante recordar que al analizar las matrices de confusión de la Fig. 5-7 si bien se observó un aumento de los VP respecto del caso con todas las características, también se produjo un aumento en los FP, con lo que se observa una vez más el fenómeno de compensaciones ya comentado.

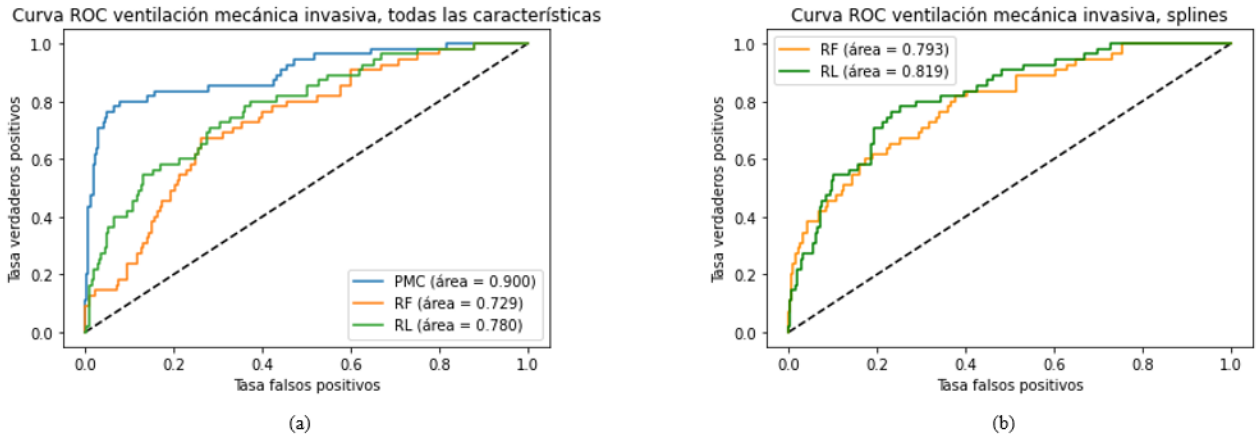


Fig. 5-8 Curvas ROC resultado clínico Ventilación mecánica invasiva.

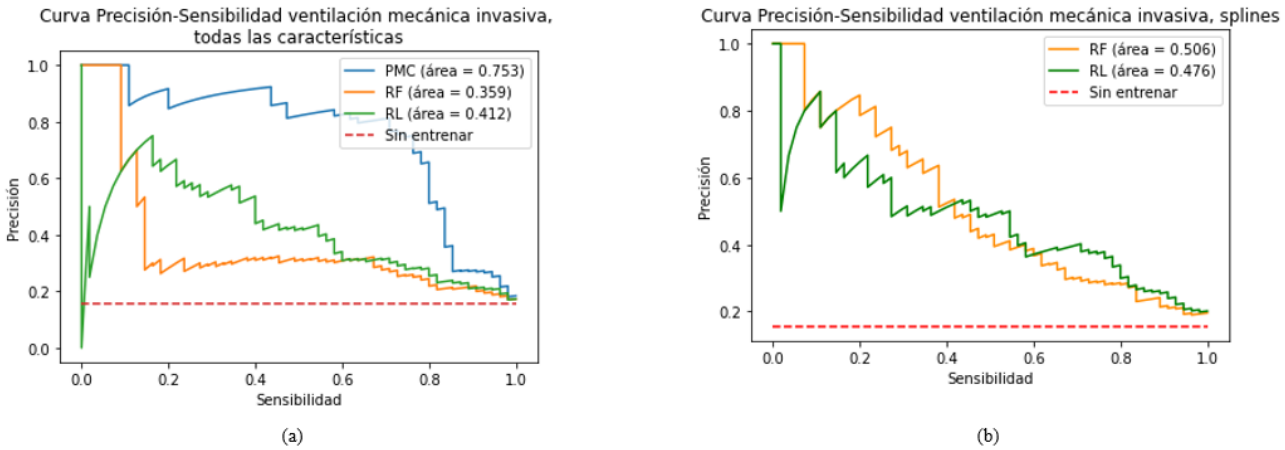


Fig. 5-9 Curvas PRC resultado clínico Ventilación mecánica invasiva.

En la Fig. 5-9 (a) se puede observar la diferencia de rendimiento en la clase positiva de PMC frente a los otros dos modelos. La curva PRC de PMC presenta una forma muy distinta a las de los otros clasificadores, con un área bajo la curva mucho mayor y con un cambio de pendiente muy pronunciado. Podemos ver que el clasificador logra alcanzar aproximadamente 80% de precisión y sensibilidad, para luego comenzar a decaer en ambas métricas de forma muy pronunciada. Para RL y RF las gráficas poseen una forma muy distinta a la de PMC. No solo no alcanzan una buena aproximación a la esquina superior derecha, si no que presentan pendientes que podrían considerarse constantes. La curva PRC de RF es casi horizontal, manteniendo un valor de precisión de aproximadamente 0,3 hasta una sensibilidad de 0,64 donde comienza a descender. La curva de RL desciende de forma constante y se corta con la de RF en el punto que podría considerarse de mejor rendimiento para ambos clasificadores.

En la Fig. 5-9 (b) se observa una leve mejora en las curvas PRC y sus puntajes AUC luego de reducir el espacio de características. RF tiene ahora una pendiente similar a la de RL. El área bajo la curva de RF aumenta mucho más que para RL.

Capítulo 6 Discusión y conclusiones

6.1 Discusión

El objetivo principal de este trabajo era el de poder predecir el riesgo de tres resultados clínicos para pacientes con COVID-19. Dentro de los principales desafíos encontrados en su desarrollo el primero fue el de la baja densidad de datos en variables consideradas de importancia según la literatura, en especial en los biomarcadores. Este hecho constituyó una limitación importante en el trabajo ya que, a pesar de contar con más de 6 millones de registros, finalmente este universo se redujo a 1.062, dentro de los cuales aún existían biomarcadores con una cantidad no despreciable de elementos a imputar. El hecho de imputar datos crea nuevas distribuciones en la base de datos, que si bien de acuerdo con la Fig. 3-5, no difieren mucho de las distribuciones originales observadas en los

datos, con los métodos de imputación se pueden crear ciertas dependencias entre variables que tal vez, no existían en los datos reales.

En general dentro de la bioinformática es común que los problemas de clasificación presenten desbalances de clases considerables, principalmente con una clase positiva minoritaria. Los efectos de los desbalances de clase muy pronunciados pueden ser aminorados al contar con altos volúmenes de datos, para que, con esto, los modelos puedan ser alimentados con una cantidad de ejemplos importante que les permita aprender de mejor forma sobre la clase minoritaria.

En nuestro caso, el set de datos utilizado para la creación de los modelos no solo presentaba importantes desbalances, sino que también era pequeño y al dividirlo en los conjuntos de entrenamiento y prueba, a pesar de hacerlo de forma estratificada, la cantidad de ejemplos disponibles para las clases positivas se veía aún más reducida. Esto también ocurrió en las validaciones cruzadas para ajuste de parámetros.

La clasificación desbalanceada presenta un desafío en sí misma y para abordarla se sugieren diversas técnicas que van desde métodos de muestreo hasta el uso y análisis eficiente de las diversas métricas de evaluación existentes [22]. En nuestro caso se optó por realizar un análisis con diversas métricas por sobre la utilización de técnicas de muestreo.

Al analizar de manera holística diversas métricas de la matriz de confusión y estudiar el comportamiento de esta última de forma gráfica, es posible obtener una idea más completa y real del rendimiento de los clasificadores. Esto permite identificar qué se sacrifica en cada caso para obtener mejores resultados.

El uso de las curvas ROC en clasificación desbalanceada puede llevar a un falso optimismo en el rendimiento de los modelos. Este fenómeno se pudo apreciar en las salidas de ventilación mecánica invasiva e ingreso UCI, las cuales presentaban desbalance con clase positiva minoritaria. A pesar de obtener buenos puntajes AUC, los modelos de regresión logística y *Random Forest* presentaban *compensaciones* muy marcadas entre verdaderos y falsos positivos, efecto que podía evidenciarse al analizar las gráficas de sus curvas PRC.

Estos son precisamente los fenómenos por los cuales se establece una preferencia de este tipo de curvas por sobre las ROC para evaluar y comparar el rendimiento de modelos que trabajan con clases desbalanceadas[23]. Se considera que las curvas PRC son mucho más informativas ya que se centran solo en la calidad de predicción de la clase positiva, en cambio con las curvas ROC se toman en consideración los verdaderos negativos que, como ya se ha comentado, suelen ser bien identificados al representar a la clase mayoritaria.

Sin embargo, los modelos que presentaron mejor rendimiento para todos los casos, tenían también los puntajes AUC más altos para ambas curvas. De esto se puede comentar que la idoneidad de una curva por sobre otra a la hora de evaluar y comprar los modelos, no está dada por una diferencia en el modelo que se va a considerar como de mejor rendimiento, si no que ambas curvas se complementan para entregar una caracterización más acabada del modelo y permiten no quedarse con la falsa sensación de un muy buen desempeño que podría darse al observar solo las curvas ROC.

Es muy importante el poder analizar los resultados con distintos puntos de vista, sin olvidar el objetivo que se persigue al construir los modelos. En nuestro caso y en general, en los problemas de predicción de resultados clínicos o apoyo diagnóstico es muy importante obtener altas sensibilidades, ya que se espera poder recuperar la mayor cantidad de casos positivos posible. Modelos con bajo rendimiento en esta métrica no son de ayuda, ya que en salud es preferible algunos falsos positivos que dejar muchos verdaderos positivos, por ejemplo, sin diagnóstico o tratamiento. No obstante, es necesario analizar esta métrica en conjunto con otras, como la precisión.

La revisión de los modelos debe estar siempre enfocada al uso final de estos. Por ejemplo, si se quisiera usar un modelo para mejorar la gestión de los recursos hospitalarios siempre escasos, sobre todo en el contexto de la pandemia vivida durante los últimos dos años, los modelos de *Random Forest* y regresión logística de 5.4 no serían de mucha ayuda, ya que, a pesar de poder identificar una buena cantidad de casos positivos, las precisiones rondan el 31%. Esto no solo no ayudaría a gestionar los recursos de forma más eficiente, sino que empeoraría la visión del escenario real, retratando una situación mucho más grave que la que se tiene.

En todos los casos vistos, los modelos generados se comportaron mejor que la aleatoriedad, esto se puede evidenciar en las curvas ROC y PRC, donde todas las gráficas se encuentran por sobre los umbrales correspondientes a modelos sin entrenar.

En el caso de la variable de salida de hospitalización, la cual presentaba un desbalance favoreciendo a la clase positiva, los modelos de aprendizaje automático más sencillos como RL y RF tuvieron buenos rendimientos e incluso, este último superó los resultados obtenidos con la red neuronal.

Para los casos de ventilación mecánica invasiva e ingreso UCI, en los cuales el desbalance perjudicaba la clase positiva, tanto RF como RL evidenciaron dificultades para identificar entre clases. Dichas dificultades se pueden apreciar en las marcadas compensaciones entre verdaderos y falsos positivos y su consecuente efecto en precisión y sensibilidad y curvas PRC. Este fenómeno se puede explicar por la existencia de bordes de decisión poco marcados, lo cual podría ser producto de la baja

cantidad de ejemplos disponibles en las clases positivas, en la naturaleza de los datos y en las diferencias en la manifestación de la enfermedad entre pacientes.

El uso de *deep learning* fue clave para estas salidas, ya que incluso con su configuración más básica se pudo obtener una mejora importante en la clasificación.

El método de selección de características utilizado tuvo efectos que se evidenciaron en ciertos comportamientos extendidos para todos los resultados clínicos. En general su uso ayudó a mejorar la identificación de la clase minoritaria, independientemente de su etiqueta. En el caso de la hospitalización esto se tradujo como un aumento de la precisión, con disminución en la sensibilidad; y tanto para ingreso UCI como para ventilación mecánica invasiva, el mayor aumento se dio en la sensibilidad, pero sacrificando un poco la precisión obtenida por clasificadores. Al analizar los puntajes AUC de las curvas PRC, se puede comentar que el uso del método de selección de características ayudó a mejorar el rendimiento global de los clasificadores. la reducción en el espacio de características mejoró el rendimiento de los clasificadores.

Es importante comentar lo trabajoso que significa el ajuste de parámetros de las redes neuronales. Si bien existe consenso sobre esto, no hay reglas claras a la hora de elegir la cantidad de capas ocultas y nodos.

Distintos autores presentan distintas fórmulas para responder a esta interrogante, pero también son enfáticos al decir que estas recomendaciones no son “reglas de oro” y que pueden servir como puntos de partida para comparar los rendimientos obtenidos al realizar los ajustes de parámetros. Pese a lo anterior, el uso de estos modelos fue de utilidad para poder obtener mejores clasificaciones en los casos donde los modelos más clásicos no lograron desempeñarse de buena forma.

6.2 Conclusiones

En este trabajo se crearon distintos modelos de aprendizaje automático para predecir el riesgo de tres resultados clínicos en pacientes con COVID-19 de nuestro país.

Para el resultado clínico de hospitalización, el modelo de *Random Forest* fue capaz de identificar el riesgo de un paciente a dicho desenlace con un 94% de sensibilidad y un 83,5% de precisión.

Para el resultado clínico de ingreso UCI, el modelo de *deep learning* de perceptrón multicapa identificó correctamente el riesgo de ingresar a dicha unidad con un 70,4% de sensibilidad y un 78,1% de precisión.

Finalmente, para el resultado clínico de ventilación mecánica invasiva, el perceptrón multicapa identificó de forma correcta el riesgo de los pacientes a requerir de este tipo de apoyo ventilatorio con un 69% de sensibilidad y un 80,8% de precisión.

Todos los modelos presentados, y en particular los seleccionados como mejores modelos para cada variable de salida, suponen una mejora considerable respecto de la aleatoriedad.

A pesar de los desafíos propios de la clasificación con clases desbalanceadas y las limitaciones dadas por lo datos, fue posible obtener modelos de predicción con buenos rendimientos, recordando siempre analizar este punto con diversas métricas para poder recabar una imagen completa del rendimiento obtenido con los clasificadores.

Bibliografía

[1] Z. Wu y J. M. McGoogan, "Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention", *JAMA*, vol. 323, nº 13, pp. 1239–1242, abr. 2020, doi: 10.1001/jama.2020.2648.

[2] Amber L. Mueller, Maeve S.McNamara, y David A. Sinclair, "Why does COVID-19 disproportionately affect older people?", *Aging (Albany. NY).*, vol. 12, nº 10, pp. 9959–9981, 2020.

[3] M. Biswas, S. Rahaman, T. K. Biswas, Z. Haque, y B. Ibrahim, "Association of Sex, Age, and Comorbidities with Mortality in COVID-19 Patients: A Systematic Review and Meta-Analysis", *Intervirology*, vol. 64, nº 1, pp. 36–47, 2021, doi: 10.1159/000512592.

[4] A. Sanyaolu *et al.*, "Comorbidity and its Impact on Patients with COVID-19", *SN Compr. Clin. Med.*, vol. 2, nº 8, pp. 1069–1076, 2020, doi: 10.1007/s42399-020-00363-4.

[5] A. K. Singh, R. Gupta, A. Ghosh, y A. Misra, "Diabetes in COVID-19: Prevalence, pathophysiology, prognosis and practical considerations", *Diabetes Metab. Syndr. Clin. Res. Rev.*, vol. 14, nº 4, pp. 303–310, 2020, doi: 10.1016/j.dsx.2020.04.004.

[6] H. Ejaz *et al.*, "COVID-19 and comorbidities: Deltererious impact on infected patients", *J. Infect. Public Health*, vol. 13, nº January, pp. 1833–1839, 2020.

[7] S. Elezkurtaj *et al.*, "Causes of death and comorbidities in hospitalized patients with COVID-19", *Sci. Rep.*, vol. 11, nº 1, pp. 1–9, 2021, doi: 10.1038/s41598-021-82862-5.

[8] T. Chang, J. Wu, y L. Chang, "Clinical Course and outcomes of critically ill patients with COVID19 in Wuhan China", *Lancet Respir Med.*, vol. 8, nº January, pp. 475–81, 2020.

[9] F. Zhou *et al.*, "Clinical course and risk factors for mortality of adult inpatients with COVID-19 in Wuhan, China: a retrospective cohort study", *Lancet*, vol. 395, nº 10229, pp. 1054–1062, 2020, doi: 10.1016/S0140-6736(20)30566-3.

[10] M. Ludwig, J. Jacob, F. Basedow, F. Andersohn, y J. Walker, "Clinical outcomes and characteristics of patients hospitalized for Influenza or COVID-19 in Germany", *Int. J. Infect. Dis.*, vol. 103, pp. 316–322, 2021, doi: 10.1016/j.ijid.2020.11.204.

[11] L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal", *Liesbet Henckaerts*, vol. 15, p. 26, doi: 10.1136/bmj.m1328.

[12] D. van Klaveren *et al.*, "COVID Outcome Prediction in the Emergency Department (COPE): Development and validation of a model for predicting death and need for intensive care in COVID-19 patients", *medRxiv*, p. 2020.12.30.20249023, 2021, [En línea]. Disponible en: http://medrxiv.org/content/early/2021/02/08/2020.12.30.20249023.abstract.

[13] W. Khan, A. Hussain, S. A. Khan, M. Al-Jumailey, R. Nawaz, y P. Liatsis, "Analysing the impact of global demographic characteristics over the COVID-19 spread using class rule mining and pattern matching", *R. Soc. Open Sci.*, vol. 8, nº 1, 2021, doi: 10.1098/rsos.201823.

[14] W. H. Ng *et al.*, "Comorbidities in SARS-CoV-2 patients: A systematic review and meta-analysis", *MBio*, vol. 12, nº 1, pp. 1–12, 2021, doi: 10.1128/mBio.03647-20.

[15] S. C. Lee, K. J. Son, C. H. Han, J. Y. Jung, y S. C. Park, "Impact of comorbid asthma on severity of coronavirus disease (COVID-19)", *Sci. Rep.*, vol. 10, nº 1, pp. 1–9, 2020, doi: 10.1038/s41598-020-77791-8.

[16] X. Guan *et al.*, "Clinical and inflammatory features based machine learning model for fatal risk prediction of hospitalized COVID-19 patients: results from a retrospective cohort study", *Ann. Med.*, vol. 53, nº 1, pp. 257–266, 2021, doi: 10.1080/07853890.2020.1868564.

[17] P. Parchure *et al.*, "Development and validation of a machine learning-based prediction model for near-term in-hospital mortality among patients with COVID-19", *BMJ Support. Palliat. Care*, pp. 1–8, 2020, doi: 10.1136/bmjspcare-2020-002602.

[18] V. K. Gupta, A. Gupta, D. Kumar, y A. Sardana, "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model", *Big Data Min. Anal.*, vol. 4, nº 2, pp. 116–123, 2021, doi: 10.26599/BDMA.2020.9020016.

[19] M. E. Shipe, S. A. Deppen, F. Farjah, y E. L. Grogan, "Developing prediction models for clinical use using logistic regression : an overview", vol. 11, nº Suppl 4, pp. 574–584, 2019, doi: 10.21037/jtd.2019.01.25.

[20] J. Gauthier, Q. V. Wu, y T. A. Gooley, "Cubic splines to model relationships between continuous variables and outcomes: a guide for clinicians", *Bone Marrow Transplant.*, vol. 55, nº 4, pp. 675–680, 2020, doi: 10.1038/s41409-019-0679-x.

[21] J. Davis y M. Goadrich, "The relationship between Precision-Recall and ROC curves", *ICML 2006 - Proc. 23rd Int. Conf. Mach. Learn.*, vol. 2006, pp. 233–240, 2006.

[22] P. Branco, L. Torgo, y R. Ribeiro, "A Survey of Predictive Modelling under Imbalanced Distributions", pp. 1–48, 2015, [En línea]. Disponible en: http://arxiv.org/abs/1505.01658.

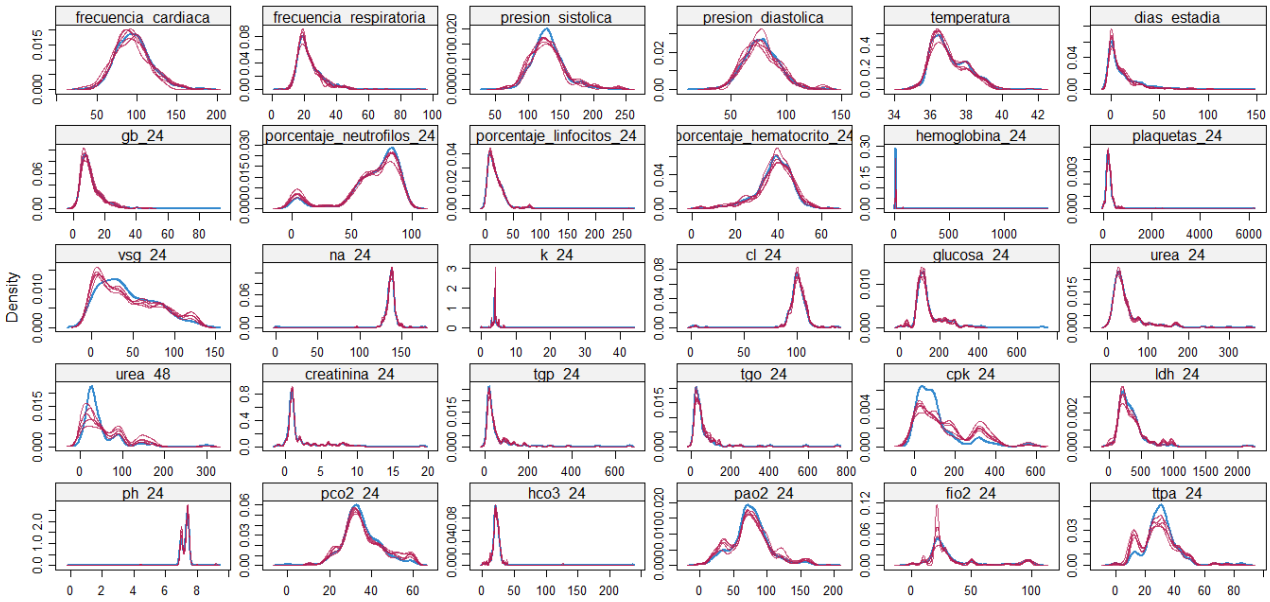
[23] T. Saito y M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets", *PLoS One*, vol. 10, nº 3, pp. 1–21, 2015, doi: 10.1371/journal.pone.0118432.

Anexo A Columnas de la Base de Datos original

Columna	Variable	Columna	Variable	Columna	Variable	Columna	Variable	Columna	Variable	Columna	Variable	Columna	Variable	Columna	Variable
1	id_formulario_eno	21	semana_gestacion	41	presion_diastolica	61	hospitalizacion	81	oxigeno_suplementario	101	tipos_antibioticos	121	caso_cerrado	141	resultado_otro_cultivo_1
2	numero_folio	22	prevision	42	temperatura	62	fecha_ingreso_hospital	82	fecha_inicio_oxigeno_suplementario	102	fecha_inicio_antibioticos	122	pais_contagio	142	resultado_otro_cultivo_detalle_1
3	id_enfermedad_eno	23	nacionalidad	43	comorbilidad	63	fecha_egreso_hospital	83	fecha_termino_oxigeno_suplementario	103	fecha_termino_antibioticos	123	fecha_ingreso_sistema	143	lab_2
4	enfermedad_notificada	24	pueblo_indigena	44	consumo_alcohol	64	dias_estadia	84	ingreso_uci_intensivo_intermedio	104	esteroides	124	vigente_no_eliminado	144	id_formulario_eno_2
5	semana_epidemiologica	25	otro_pueblo_indigena	45	consumo_tabaco	65	id_institucion_hospitalizacion	85	fecha_inicio_ingreso_uci_intensivo_intermedio	105	tipos_esteroides	125	lab_1	145	tipo_paciente_2
6	fecha_notificacion	26	region_residencia	46	uso_vapeadores	66	institucion_hospitalizacion	86	fecha_termino_ingreso_uci_intensivo_intermedio	106	fecha_inicio_esteroides	126	id_formulario_eno_1	146	id_tipo_muestra_2
7	etapa_clinica	27	comuna_residencia	47	antecedente_viaje_internacional	67	tipo_egreso	87	ventilacion_mecanica_no_invasiva	107	fecha_termino_esteroides	127	tipo_paciente_1	147	nombre_tipo_muestra_2
8	establecimiento_salud	28	descripcion_trabajo	48	contacto_caso_sospechoso	68	viaje_extranjero	88	fecha_inicio_ventilacion_mecanica_no_invasiva	108	otros_medicamentos	128	id_tipo_muestra_1	148	detalle_tipo_muestra_2
9	region	29	rubro_trabajo	49	nombre_caso_sospechoso	69	paises_viaje	89	fecha_termino_ventilacion_mecanica_no_invasiva	109	tipos_otros_medicamentos	129	nombre_tipo_muestra_1	149	fecha_toma_muestra_2
10	seremi	30	actividad_laboral_declarada	50	motivo_examen	70	fecha_ida_extranjero	90	ventilacion_mecanica_invasiva	110	fecha_inicio_otros_medicamentos	130	detalle_tipo_muestra_1	150	fecha_envio_muestra_2
11	dv	31	nombre_instruccion	51	detalle_otro_motivo_examen	71	fecha_vuelta_extranjero	91	fecha_inicio_ventilacion_mecanica_invasiva	111	fecha_termino_otros_medicamentos	131	fecha_toma_muestra_1	151	id_laboratorio_2
12	estado_paciente	32	anio_aprobado	52	uso_medicamentos_antipireticos	72	ciudades_viaje	92	fecha_termino_ventilacion_mecanica_invasiva	112	cie_10_diagnostico	132	fecha_envio_muestra_1	152	nombre_laboratorio_2
13	causa_basica_muerte	33	presentacion_clinica	53	fecha_inicio_toma_antipireticos	73	contacto_enfermos	93	drogas_vasoactivas	113	diagnostico	133	id_laboratorio_1	153	resultado_ifi_2
14	sexo	34	fecha_primeros_sintomas	54	uso_medicamentos_antibioticos	74	lugar_contacto_enfermos	94	fecha_inicio_drogas_vasoactivas	114	fecha_diagnostico	134	nombre_laboratorio_1	154	det_resultado_ifi_2
15	fecha_nacimiento	35	semana_epidemiologica_primeros_sintomas	55	fecha_inicio_toma_antibioticos	75	visita_mercados	95	fecha_termino_drogas_vasoactivas	115	confirmacion_biopsia	135	resultado_ifi_1	155	resultado_pcr_2
16	fecha_fallecimiento	36	sintomas	56	uso_medicamentos_antivirales	76	lugar_visita_mercados	96	antivirales	116	confirmacion_epidemiologo	136	det_resultado_ifi_1	156	det_resultado_pcr_2
17	edad	37	otro_sintoma	57	fecha_inicio_toma_antivirales	77	contacto_animales	97	tipos_antivirales	117	confirmacion_autopsia	137	resultado_pcr_1	157	resultado_hemocultivo_2
18	meses	38	frecuencia_cardiaca	58	fecha_primera_consulta	78	lugar_contacto_animales	98	fecha_inicio_antivirales	118	confirmacion_laboratorio	138	det_resultado_pcr_1	158	resultado_hemocultivo_detalle_2
19	dias	39	frecuencia_respiratoria	59	id_institucion_primera_consulta	79	trabajador_salud	99	fecha_termino_antivirales	119	confirmacion_clinico	139	resultado_hemocultivo_1	159	resultado_otro_cultivo_2
20	embarazo	40	presion_sistolica	60	institucion_primera_consulta	80	institucion_trabajo	100	antibioticos	120	estado_caso	140	resultado_hemocultivo_detalle_1	160	resultado_otro_cultivo_detalle_2

Columna	Variable	Columna	Variable	Columna	Variable	Columna	Variable
161	lab_3	191	hemoglobina_24	221	hco3_24	251	estado_ord
162	id_formulario_eno_3	192	hemoglobina_48	222	hco3_48	252	region_num
163	tipo_paciente_3	193	plaquetas_24	223	pao2_24	253	reg
164	id_tipo_muestra_3	194	plaquetas_48	224	pao2_48	254	reg_ser
165	nombre_tipo_muestra_3	195	vsg_24	225	fio2_24	255	ano_not
166	detalle_tipo_muestra_3	196	vsg_48	226	fio2_48	256	mes_not
167	fecha_toma_muestra_3	197	na_24	227	ttpa_24	257	dia_not
168	fecha_envio_muestra_3	198	na_48	228	ttpa_48	258	dif_fecha
169	id_laboratorio_3	199	k_24	229	tp_inr_24	259	muerto
170	nombre_laboratorio_3	200	k_48	230	tp_inr_48	260	actuales
171	resultado_ifi_3	201	cl_24	231	fibrinogeno_24	261	activos
172	det_resultado_ifi_3	202	cl_48	232	imagenologia	262	epivigila_pcr
173	resultado_pcr_3	203	glucosa_24	233	otra_comorbilidad	263	duplicado
174	det_resultado_pcr_3	204	glucosa_48	234	fecha_resultado_ifi_1		
175	resultado_hemocultivo_3	205	urea_24	235	fecha_resultado_pcr_1		
176	resultado_hemocultivo_detalle_3	206	urea_48	236	fecha_resultado_hemocultivo_1		
177	resultado_otro_cultivo_3	207	creatinina_24	237	fecha_resultado_otro_cultivo_1		
178	resultado_otro_cultivo_detalle_3	208	creatinina_48	238	fecha_resultado_ifi_2		
179	tipo_caso_busqueda	209	tgp_24	239	fecha_resultado_pcr_2		
180	lugar_busqueda_activa	210	tgp_48	240	fecha_resultado_hemocultivo_2		
181	detalle_lugar_busqueda	211	tgo_24	241	fecha_resultado_otro_cultivo_2		
182	lugar_reposo	212	tgo_48	242	fecha_resultado_ifi_3		
183	gb_24	213	cpk_24	243	fecha_resultado_pcr_3		
184	gb_48	214	cpk_48	244	fecha_resultado_hemocultivo_3		
185	porcentaje_neutrofilos_24	215	ldh_24	245	fecha_resultado_otro_cultivo_3		
186	porcentaje_neutrofilos_48	216	ldh_48	246	factor_riesgo_caso_confirmado		
187	porcentaje_linfocitos_24	217	ph_24	247	factor_riesgo_caso_sospechoso		
188	porcentaje_linfocitos_48	218	ph_48	248	factor_riesgo_viaje_internacional		
189	porcentaje_hematocrito_24	219	pco2_24	249	factor_riesgo_viaje_nacional		
190	porcentaje_hematocrito_48	220	pco2_48	250	ningun_factor_riesgo		

Anexo B Distribución de densidad variables imputadas



Anexo C Wald chi-cuadrado para atributos

Variable de salida hospitalización

Atributo	Df Chisq	Pr(>Chisq)	Significancia
`rcs(edad, 3)edad`	1 7.5865	0.0058806	**
`rcs(edad, 3)edad'`	1 3.1208	0.077301	.
sexo	1 1.5206	0.2175277	
`rcs(frecuencia_respiratoria, 3)frecuencia_respiratoria`	1 10.4751	0.0012099	**
`rcs(frecuencia_respiratoria, 3)frecuencia_respiratoria'`	1 3.2283	0.0723741	.
`rcs(frecuencia_cardiaca, 3)frecuencia_cardiaca`	1 6.2403	0.0124876	*
`rcs(frecuencia_cardiaca, 3)frecuencia_cardiaca'`	1 0.8969	0.3436202	
`rcs(presion_sistolica, 3)presion_sistolica`	1 5.8518	0.0155609	*
`rcs(presion_sistolica, 3)presion_sistolica'`	1 2.0473	0.1524767	
`rcs(presion_diastolica, 3)presion_diastolica`	1 0.0004	0.9833838	
`rcs(presion_diastolica, 3)presion_diastolica'`	1 0.2671	0.6052598	
`rcs(temperatura, 3)temperatura`	1 0.1198	0.7292033	
`rcs(temperatura, 3)temperatura'`	1 0.2971	0.5856792	
cefalea	1 3.9865	0.0458663	*
disnea	1 1.0842	0.2977582	
odinofagia	1 1.4373	0.2305733	
fiebre	1 3.6470	0.0561688	.
anosmia	1 0.1560	0.6928771	
ageusia	1 1.2398	0.2655033	
mialgia	1 0.0006	0.9802823	
diarrea	1 0.1602	0.6889547	
dolor_abdominal	1 1.4905	0.2221417	
taquipnea	1 0.0852	0.7704352	
dolor_toracico	1 1.6613	0.1974282	
asma	1 8.0114	0.0046484	**
obesidad	1 0.1652	0.6844568	
cardiopatia_cronica	1 0.1505	0.6980913	
inmunocomprometido	1 3.4865	0.0618709	.

enfermedad_cardiovascular	1 10.9068	0.0009581	***
enfermedad_pulmonar_cronica	1 5.7291	0.0166864	*
enfermedad_renal_cronica	1 4.4059	0.0358159	*
enfermedad_neurologica_cronica	1 0.0079	0.9291072	
enfermedad_hepatica_cronica	1 0.3473	0.5556486	
`rcs(gb_24, 3)gb_24`	1 5.3936	0.0202106	*
`rcs(gb_24, 3)gb_24'`	1 2.8629	0.0906441	.
`rcs(porcentaje_neutrofilos_24, 3)porcentaje_neutrofilos_24`	1 0.9978	0.317847	
`rcs(porcentaje_neutrofilos_24, 3)porcentaje_neutrofilos_24'`	1 0.4324	0.5108243	
`rcs(porcentaje_linfocitos_24, 3)porcentaje_linfocitos_24`	1 0.0444	0.8330423	
`rcs(porcentaje_linfocitos_24, 3)porcentaje_linfocitos_24'`	1 0.0591	0.8079308	
`rcs(porcentaje_hematocrito_24, 3)porcentaje_hematocrito_24`	1 8.5175	0.0035174	**
`rcs(porcentaje_hematocrito_24, 3)porcentaje_hematocrito_24'`	1 1.2598	0.2616817	
`rcs(hemoglobina_24, 3)hemoglobina_24`	1 0.1252	0.7234261	
`rcs(hemoglobina_24, 3)hemoglobina_24'`	1 0.1875	0.6649807	
`rcs(plaquetas_24, 3)plaquetas_24`	1 0.5876	0.4433609	
`rcs(plaquetas_24, 3)plaquetas_24'`	1 1.3640	0.2428423	
`rcs(vsg_24, 3)vsg_24`	1 0.1075	0.7429562	
`rcs(vsg_24, 3)vsg_24'`	1 0.0293	0.8640295	
`rcs(na_24, 3)na_24`	1 0.4437	0.5053197	
`rcs(na_24, 3)na_24'`	1 0.4353	0.5093859	
`rcs(cl_24, 3)cl_24`	1 0.3411	0.5591917	
`rcs(cl_24, 3)cl_24'`	1 0.4879	0.4848546	
`rcs(glucosa_24, 3)glucosa_24`	1 0.6960	0.4041212	
`rcs(glucosa_24, 3)glucosa_24'`	1 0.5554	0.4561227	
`rcs(urea_24, 3)urea_24`	1 0.0401	0.8412071	
`rcs(urea_24, 3)urea_24'`	1 0.8056	0.3694212	
`rcs(creatinina_24, 3)creatinina_24`	1 6.6510	0.0099097	**
`rcs(creatinina_24, 3)creatinina_24'`	1 0.8347	0.36091	
`rcs(tgp_24, 3)tgp_24`	1 12.1626	0.0004876	***
`rcs(tgp_24, 3)tgp_24'`	1 9.6801	0.0018628	**

`rcs(tgo_24, 3)tgo_24`	1 0.2092	0.647358	
`rcs(tgo_24, 3)tgo_24'`	1 0.2957	0.5865671	
`rcs(cpk_24, 3)cpk_24`	1 0.5517	0.4576387	
`rcs(cpk_24, 3)cpk_24'`	1 0.0071	0.9329931	
`rcs(ldh_24, 3)ldh_24`	1 0.1090	0.7412961	
`rcs(ldh_24, 3)ldh_24'`	1 0.2803	0.5964824	
`rcs(ph_24, 3)ph_24`	1 10.4173	0.0012484	**
`rcs(ph_24, 3)ph_24'`	1 0.2674	0.6051008	
`rcs(pco2_24, 3)pco2_24`	1 0.5914	0.4418756	
`rcs(pco2_24, 3)pco2_24'`	1 1.8488	0.1739275	
`rcs(hco3_24, 3)hco3_24`	1 2.2860	0.1305485	
`rcs(hco3_24, 3)hco3_24'`	1 2.5709	0.1088451	
`rcs(pao2_24, 3)pao2_24`	1 53.2507	2.94E-13	***
`rcs(pao2_24, 3)pao2_24'`	1 21.2244	4.09E-06	***
`rcs(fio2_24, 3)fio2_24`	1 7.1913	0.0073258	**
`rcs(fio2_24, 3)fio2_24'`	1 10.1537	0.0014402	**
`rcs(ttpa_24, 3)ttpa_24`	1 0.5153	0.4728505	
`rcs(ttpa_24, 3)ttpa_24'`	1 1.8679	0.1717146	
`rcs(k_24, 3)k_24`	1 1.8362	0.1753999	
`rcs(k_24, 3)k_24'`	1 1.1670	0.2800137	
diabetes	1 0.4592	0.4980162	
hipertension_arterial	1 9.6285	0.0019158	**
tos	1 0.4259	0.5140243	
cianosis	1 0.8280	0.3628571	
postracion	1 0.0406	0.840281	
Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '**		1	

Variable de salida ventilación mecánica invasiva

Atributo	Df Chisq	Pr(>Chisq)	Significancia
`rcs(edad, 3)edad`	1 0.5699	0.4502822	
`rcs(edad, 3)edad'`	1 21.9810	2.75E-06	***
sexo	1 0.6980	0.4034721	
`rcs(frecuencia_respiratoria, 3)frecuencia_respiratoria`	1 2.0653	0.1506872	
`rcs(frecuencia_respiratoria, 3)frecuencia_respiratoria'`	1 1.5390	0.2147614	
`rcs(frecuencia_cardiaca, 3)frecuencia_cardiaca`	1 2.2543	0.1332459	

`rcs(frecuencia_cardiaca, 3)frecuencia_cardiaca'`	1	0.3488	0.5548032	
`rcs(presion_sistolica, 3)presion_sistolica`	1	6.5518	0.010478	*
`rcs(presion_sistolica, 3)presion_sistolica'`	1	1.1154	0.2909076	
`rcs(presion_diastolica, 3)presion_diastolica`	1	0.4831	0.4870205	
`rcs(presion_diastolica, 3)presion_diastolica'`	1	0.1778	0.6732453	
`rcs(temperatura, 3)temperatura`	1	1.0067	0.3156941	
`rcs(temperatura, 3)temperatura'`	1	0.3868	0.5339739	
cefalea	1	4.4996	0.0339038	*
disnea	1	5.6670	0.0172874	*
odinofagia	1	1.0478	0.3060031	
fiebre	1	8.5553	0.0034451	**
anosmia	1	33.6455	6.61E-09	***
ageusia	1	1.0207	0.3123501	
mialgia	1	2.8941	0.0889071	.
diarrea	1	9.2255	0.0023867	**
dolor_abdominal	1	5.6505	0.0174506	*
taquipnea	1	4.2738	0.038704	*
dolor_toracico	1	7.2529	0.0070785	**
asma	1	12.0040	0.0005309	***
obesidad	1	0.7066	0.400583	
cardiopatia_cronica	1	3.9808	0.046022	*
inmunocomprometido	1	0.1834	0.668489	
enfermedad_cardiovascular	1	31.5475	1.95E-08	***
enfermedad_pulmonar_cronica	1	3.4894	0.0617637	.
enfermedad_renal_cronica	1	19.5514	9.79E-06	***
enfermedad_neurologica_cronica	1	30.5020	3.34E-08	***
enfermedad_hepatica_cronica	1	7.3065	0.0068704	**
`rcs(gb_24, 3)gb_24`	1	0.1280	0.7205167	
`rcs(gb_24, 3)gb_24'`	1	0.0192	0.8898944	
`rcs(porcentaje_neutrofilos_24, 3)porcentaje_neutrofilos_24`	1	9.3948	0.002176	**
`rcs(porcentaje_neutrofilos_24, 3)porcentaje_neutrofilos_24'`	1	0.5244	0.4689659	
`rcs(porcentaje_linfocitos_24, 3)porcentaje_linfocitos_24`	1	15.6742	7.52E-05	***
`rcs(porcentaje_linfocitos_24, 3)porcentaje_linfocitos_24'`	1	1.9872	0.1586305	
`rcs(porcentaje_hematocrito_24, 3)porcentaje_hematocrito_24`	1	13.6019	0.000226	***
`rcs(porcentaje_hematocrito_24, 3)porcentaje_hematocrito_24'`	1	0.1613	0.6879768	
`rcs(hemoglobina_24, 3)hemoglobina_24`	1	1.3315	0.2485391	
`rcs(hemoglobina_24, 3)hemoglobina_24'`	1	1.9229	0.1655321	
`rcs(plaquetas_24, 3)plaquetas_24`	1	0.4674	0.4941927	
`rcs(plaquetas_24, 3)plaquetas_24'`	1	0.4727	0.4917419	
`rcs(vsg_24, 3)vsg_24`	1	35.4466	2.62E-09	***
`rcs(vsg_24, 3)vsg_24'`	1	1.4922	0.2218762	
`rcs(na_24, 3)na_24`	1	5.7639	0.0163584	*
`rcs(na_24, 3)na_24'`	1	2.8014	0.0941792	.
`rcs(cl_24, 3)cl_24`	1	9.0623	0.0026094	**

`rcs(cl_24, 3)cl_24'`	1	0.1334	0.7148839	
`rcs(glucosa_24, 3)glucosa_24`	1	15.6965	7.44E-05	***
`rcs(glucosa_24, 3)glucosa_24'`	1	0.4307	0.5116238	
`rcs(urea_24, 3)urea_24`	1	2.0176	0.1554883	
`rcs(urea_24, 3)urea_24'`	1	1.0793	0.2988552	
`rcs(creatinina_24, 3)creatinina_24`	1	0.2159	0.6422052	
`rcs(creatinina_24, 3)creatinina_24'`	1	0.2380	0.6256265	
`rcs(tgp_24, 3)tgp_24`	1	5.7484	0.0165035	*
`rcs(tgp_24, 3)tgp_24'`	1	4.4481	0.0349401	*
`rcs(tgo_24, 3)tgo_24`	1	11.2263	0.0008065	***
`rcs(tgo_24, 3)tgo_24'`	1	2.4624	0.1166037	
`rcs(cpk_24, 3)cpk_24`	1	27.9624	1.24E-07	***
`rcs(cpk_24, 3)cpk_24'`	1	3.2035	0.073481	.
`rcs(ldh_24, 3)ldh_24`	1	0.6323	0.4265034	
`rcs(ldh_24, 3)ldh_24'`	1	0.2797	0.596872	
`rcs(ph_24, 3)ph_24`	1	7.4270	0.0064253	**
`rcs(ph_24, 3)ph_24'`	1	0.4033	0.5253792	
`rcs(pco2_24, 3)pco2_24`	1	0.0026	0.9592895	
`rcs(pco2_24, 3)pco2_24'`	1	0.0898	0.7644663	
`rcs(hco3_24, 3)hco3_24`	1	2.4066	0.1208218	
`rcs(hco3_24, 3)hco3_24'`	1	1.6110	0.2043458	
`rcs(pao2_24, 3)pao2_24`	1	8.4920	0.003567	**
`rcs(pao2_24, 3)pao2_24'`	1	5.5005	0.0190112	*
`rcs(fio2_24, 3)fio2_24`	1	0.1003	0.7514715	
`rcs(fio2_24, 3)fio2_24'`	1	0.0203	0.8867187	
`rcs(ttpa_24, 3)ttpa_24`	1	6.8511	0.0088588	**
`rcs(ttpa_24, 3)ttpa_24'`	1	6.1376	0.0132334	*
`rcs(k_24, 3)k_24`	1	13.4137	0.0002498	***
`rcs(k_24, 3)k_24'`	1	0.0220	0.882157	
diabetes	1	32.8374	1.00E-08	***
hipertension_arterial	1	13.2911	0.0002667	***
tos	1	45.7942	1.31E-11	***
cianosis	1	2.3355	0.1264563	
postracion	1	16.3470	5.27E-05	***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				

Variable de salida muerte

Atributo	**Df**	**Chisq**	**Pr(>Chisq)**	**Significancia**
`rcs(edad, 3)edad`	1	40.4009	2.07E-10	***
`rcs(edad, 3)edad'`	1	24.3877	7.88E-07	***

sexo	1	25.4323	4.58E-07	***
`rcs(frecuencia_respiratoria, 3)frecuencia_respiratoria`	1	1.6849	0.1942789	
`rcs(frecuencia_respiratoria, 3)frecuencia_respiratoria'`	1	0.5297	0.4667338	
`rcs(frecuencia_cardiaca, 3)frecuencia_cardiaca`	1	3.4143	0.0646339	.
`rcs(frecuencia_cardiaca, 3)frecuencia_cardiaca'`	1	0.4196	0.5171296	
`rcs(presion_sistolica, 3)presion_sistolica`	1	1.9324	0.1644994	
`rcs(presion_sistolica, 3)presion_sistolica'`	1	0.8975	0.3434415	
`rcs(presion_diastolica, 3)presion_diastolica`	1	4.9639	0.0258813	*
`rcs(presion_diastolica, 3)presion_diastolica'`	1	0.5566	0.455633	
`rcs(temperatura, 3)temperatura`	1	0.0888	0.7656549	
`rcs(temperatura, 3)temperatura'`	1	0.1362	0.7121007	
cefalea	1	4.0473	0.0442413	*
disnea	1	0.4616	0.4968878	
odinofagia	1	1.3726	0.2413652	
fiebre	1	0.0301	0.8623322	
anosmia	1	5.8331	0.0157272	*
ageusia	1	3.5221	0.060557	.
mialgia	1	18.6248	1.59E-05	***
diarrea	1	0.0738	0.7859244	
dolor_abdominal	1	9.1795	0.0024474	**
taquipnea	1	12.9224	0.0003247	***
dolor_toracico	1	2.3233	0.1274488	
asma	1	11.2900	0.0007792	***
obesidad	1	8.5405	0.0034733	**
cardiopatia_cronica	1	0.9179	0.3380387	
inmunocomprometido	1	2.9075	0.0881692	.
enfermedad_cardiovascular	1	9.2178	0.0023967	**
enfermedad_pulmonar_cronica	1	0.9392	0.3324951	
enfermedad_renal_cronica	1	13.3736	0.0002552	***
enfermedad_neurologica_cronica	1	1.6026	0.205535	
enfermedad_hepatica_cronica	1	6.6986	0.0096488	**
`rcs(gb_24, 3)gb_24`	1	2.4781	0.1154405	
`rcs(gb_24, 3)gb_24'`	1	0.0106	0.9178495	
`rcs(porcentaje_neutrofilos_24, 3)porcentaje_neutrofilos_24`	1	15.5155	8.18E-05	***
`rcs(porcentaje_neutrofilos_24, 3)porcentaje_neutrofilos_24'`	1	0.5097	0.4752523	
`rcs(porcentaje_linfocitos_24, 3)porcentaje_linfocitos_24`	1	0.9912	0.319457	
`rcs(porcentaje_linfocitos_24, 3)porcentaje_linfocitos_24'`	1	1.8634	0.1722293	
`rcs(porcentaje_hematocrito_24, 3)porcentaje_hematocrito_24`	1	1.2640	0.2609034	
`rcs(porcentaje_hematocrito_24, 3)porcentaje_hematocrito_24'`	1	0.9932	0.3189573	
`rcs(hemoglobina_24, 3)hemoglobina_24`	1	20.4140	6.24E-06	***
`rcs(hemoglobina_24, 3)hemoglobina_24'`	1	19.5786	9.66E-06	***
`rcs(plaquetas_24, 3)plaquetas_24`	1	1.7266	0.1888425	
`rcs(plaquetas_24, 3)plaquetas_24'`	1	2.4109	0.120495	
`rcs(vsg_24, 3)vsg_24`	1	21.7213	3.15E-06	***

`rcs(vsg_24, 3)vsg_24'`	1	1.1343	0.2868538	
`rcs(na_24, 3)na_24`	1	1.3874	0.2388502	
`rcs(na_24, 3)na_24'`	1	0.0010	0.9749264	
`rcs(cl_24, 3)cl_24`	1	0.2568	0.6123601	
`rcs(cl_24, 3)cl_24'`	1	4.0999	0.042886	*
`rcs(glucosa_24, 3)glucosa_24`	1	18.3081	1.88E-05	***
`rcs(glucosa_24, 3)glucosa_24'`	1	3.4485	0.0633106	.
`rcs(urea_24, 3)urea_24`	1	6.1141	0.0134111	*
`rcs(urea_24, 3)urea_24'`	1	2.0618	0.1510333	
`rcs(creatinina_24, 3)creatinina_24`	1	1.1967	0.2739834	
`rcs(creatinina_24, 3)creatinina_24'`	1	0.5592	0.4545774	
`rcs(tgp_24, 3)tgp_24`	1	2.4542	0.1172086	
`rcs(tgp_24, 3)tgp_24'`	1	0.4343	0.5098719	
`rcs(tgo_24, 3)tgo_24`	1	7.9378	0.0048413	**
`rcs(tgo_24, 3)tgo_24'`	1	0.1067	0.7439294	
`rcs(cpk_24, 3)cpk_24`	1	0.1136	0.7360572	
`rcs(cpk_24, 3)cpk_24'`	1	1.8311	0.1760012	
`rcs(ldh_24, 3)ldh_24`	1	16.0241	6.25E-05	***
`rcs(ldh_24, 3)ldh_24'`	1	6.6477	0.0099282	**
`rcs(ph_24, 3)ph_24`	1	0.2668	0.6054984	
`rcs(ph_24, 3)ph_24'`	1	0.3022	0.5825041	
`rcs(pco2_24, 3)pco2_24`	1	0.0183	0.8923502	
`rcs(pco2_24, 3)pco2_24'`	1	0.1528	0.695841	
`rcs(hco3_24, 3)hco3_24`	1	0.5592	0.4545907	
`rcs(hco3_24, 3)hco3_24'`	1	0.0138	0.9063558	
`rcs(pao2_24, 3)pao2_24`	1	18.2343	1.95E-05	***
`rcs(pao2_24, 3)pao2_24'`	1	6.5890	0.0102612	*
`rcs(fio2_24, 3)fio2_24`	1	1.3485	0.2455361	
`rcs(fio2_24, 3)fio2_24'`	1	0.2025	0.6527257	
`rcs(ttpa_24, 3)ttpa_24`	1	22.3440	2.28E-06	***
`rcs(ttpa_24, 3)ttpa_24'`	1	7.8920	0.0049655	**
`rcs(k_24, 3)k_24`	1	2.7790	0.0955104	.
`rcs(k_24, 3)k_24'`	1	0.9477	0.330294	
diabetes	1	16.7007	4.38E-05	***
hipertension_arterial	1	42.3959	7.45E-11	***
tos	1	1.3324	0.2483792	
cianosis	1	0.1073	0.7432737	
postracion	1	0.4044	0.5248432	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’	1			