Modelo clasificación supervisado para predecir si un paciente hospitalizado por covid-19 requiere UCI

Cesar Vicuña H., Johan Callomamani B., Adolfo Ramon P., Paul Cusi H.,Jairo Pinedo T.
Universidad Nacional de Ingeniería (Maestría de Inteligencia Artificial, Fundamentos de la inteligencia artificial) 18-05-2024

***Resumen*.**
Un modelo de clasificación para determinar si un paciente requiere ingreso en la Unidad de Cuidados Intensivos (UCI) se basa en datos clínicos recopilados durante la hospitalización. Utiliza variables como edad, sexo, resultados de pruebas de laboratorio, datos demográficos. El modelo clasifica a los pacientes en categorías como "paciente va a ingresar a UCI" o " paciente no va a ingresar a UCI", proporcionando una herramienta para que los médicos tomen decisiones rápidas y basadas en evidencia. La precisión del modelo puede mejorar con datos actualizados y ajustes continuos, optimizando la asignación de recursos y la atención del paciente.

Abstract
Modelos de clasificación han emergido como herramientas fundamentales en la lucha contra COVID-19, facilitando la toma de decisiones clínicas y la gestión de recursos. Estos modelos, que utilizan algoritmos de aprendizaje automático y estadísticas avanzadas, procesan datos de pacientes como síntomas, resultados de pruebas y factores demográficos para predecir el riesgo de infección, gravedad de la enfermedad y necesidad de hospitalización. La capacidad de estos modelos para clasificar a los pacientes en categorías como "alto riesgo" o "bajo riesgo" permite a los profesionales de la salud priorizar tratamientos y recursos de manera más eficiente. Además, estos modelos han demostrado ser valiosos en la predicción de brotes y en la evaluación de la efectividad de intervenciones. Sin embargo, su rendimiento depende de la calidad de los datos y la constante actualización de los algoritmos, lo que resalta la necesidad de una vigilancia continua y de investigaciones adicionales para mejorar su precisión y aplicabilidad en la gestión de la pandemia.

Palabras clave: Sudoku, algoritmo de búsqueda en profundidad (DFS).

I. INTRODUCCIÓN

Desde el inicio de la pandemia de COVID-19, Perú ha enfrentado desafíos significativos en términos de salud pública y gestión de recursos. El impacto del virus se ha reflejado en un alto número de fallecidos, hospitalizados y vacunados a lo largo de la crisis sanitaria. Las cifras de fallecimientos han subrayado la gravedad de la enfermedad y la necesidad urgente de intervenciones efectivas. Al mismo tiempo, el sistema de salud ha tenido que gestionar una gran cantidad de pacientes hospitalizados, exacerbando la presión sobre los recursos médicos y de infraestructura. En respuesta, el gobierno y las autoridades sanitarias han implementado campañas de vacunación masiva para reducir la propagación del virus y proteger a la población. Estas campañas han sido cruciales para controlar la pandemia, pero también han presentado desafíos logísticos y de distribución. Analizar estos datos no solo ayuda a entender la magnitud del impacto de COVID-19 en Perú, sino que también proporciona información valiosa para la planificación de futuras estrategias de salud pública y respuesta a emergencias sanitarias.

II. IDENTIFICACIÓN DEL PROBLEMA

La pandemia de COVID-19 ha puesto de manifiesto una crítica problemática en el sistema de salud peruano: la insuficiencia de camas en las Unidades de Cuidados Intensivos (UCI). Durante los picos de contagio, la demanda de camas UCI superó con creces la capacidad disponible, lo que llevó a situaciones de colapso en numerosos hospitales a lo largo del país. Este déficit de camas UCI se debe a varios factores, entre ellos la limitada infraestructura hospitalaria, la escasez de equipos médicos especializados y la falta de personal capacitado para operar estas unidades.

La insuficiencia de camas UCI ha tenido consecuencias graves para los pacientes que requerían cuidados intensivos, no solo por COVID-19, sino también por otras enfermedades críticas. La necesidad de priorizar a los pacientes con mayor probabilidad de supervivencia ha generado dilemas éticos y ha puesto a los profesionales de la salud en situaciones extremadamente difíciles. Además, la saturación de las UCI ha impedido una atención adecuada y oportuna, incrementando la mortalidad y prolongando la recuperación de muchos pacientes.

Para abordar esta problemática, es fundamental una planificación estratégica a largo plazo que incluya la ampliación de la infraestructura de salud, la adquisición de equipos médicos y la formación de personal especializado. También es crucial la implementación de sistemas de predicción y gestión de recursos que permitan una respuesta más eficiente ante futuros brotes y emergencias sanitarias.

III. SOLUCIÓN

Para determinar si un paciente requiere una cama uci vamos a realizar un modelamiento de clasificación, con lo cual vamos a poder determinar si el paciente requiere o no una cama UCI, los modelos que se van a emplear en este trabajo son:

1) Árboles de decisión.
2) Regresión logística
3) Random forest
4) Clasificador XGBoost

IV. DESCRIPCIÓN DEL CONJUNTO DE DATOS

La información utilizada en el presente estudio abarca datos detallados sobre personas fallecidas, hospitalizadas y vacunadas a causa de la COVID-19 en el territorio peruano durante el año 2021. Estos datos han sido meticulosamente recopilados y descargados del portal de Datos Abiertos del

Gobierno de Perú, el cual proporciona acceso a información pública con el fin de fomentar la transparencia y la investigación científica.

https://www.datosabiertos.gob.pe/dataset/fallecidos-hospitalizados-y-vacunados-por-covid-19

Variables categóricas:

sexo (object): Indica el género del paciente. Puede tomar valores como "Masculino" o "Femenino".

criterio_fallecido (object): Especifica si el paciente ha fallecido. Puede tomar valores como "Sí" o "No".

ubigeo_cdc (object): Código de ubicación geográfica del paciente según el Centro para el Control de Enfermedades (CDC) en Perú.

cdc_positividad (object): Indica el resultado positivo de la prueba COVID-19 del paciente. Puede tomar valores como "Positivo" o "Negativo".

flag_hospitalizado (object): Indica si el paciente ha sido hospitalizado. Puede tomar valores como "Sí" o "No".

eess_renaes (object): Código del establecimiento de salud donde se atendió al paciente, según el Registro Nacional de Establecimientos de Salud (RENAES).

evolucion_hosp_ultimo (object): Describe la última evolución hospitalaria del paciente. Puede incluir categorías como "Mejorado", "Estable", "Grave", etc.

ubigeo_inei_domicilio (object): Código de ubicación geográfica del domicilio del paciente según el Instituto Nacional de Estadística e Informática (INEI) de Perú.

flag_vacuna (object): Indica si el paciente ha sido vacunado contra COVID-19. Puede tomar valores como "Sí" o "No".

flag_uci (object): Indica si el paciente ingreso o no a UCI. Puede tomar valores como "Sí" o "No".

Variables numéricas:

edad (int64): Edad del paciente en años.

fecha_fallecimiento(int64): Fecha en la cual el paciente fallece.

fecha_ingreso_uci(int64): Fecha en la cual el paciente ingresa a uci.

fecha_ingreso_ucin(int64): Fecha en la cual el paciente sale de uci.

fecha_ingreso_hosp(int64): Fecha en la cual el paciente ingresa a hospitalización.

Tratamiento de datos antes de realizar el modelamiento:

Se trabajaron los datos de fecha y estas se transformaron convenientemente en las siguiente variables numéricas:

dias_hospitalizado (int64): Número de días que el paciente ha permanecido hospitalizado.
dias_seguimiento_hosp (int64): Número de días de seguimiento hospitalario del paciente.
dias_uci (int64): Número de días que el paciente ha permanecido en la Unidad de Cuidados Intensivos (UCI).
dias_ucin (int64): Número de días que el paciente ha permanecido en la Unidad de Cuidados Intensivos Neonatal (UCIN), si aplica.

También dividimos nuestra data en 2 dataframes para poder estudiar su comportamiento de forma independiente.
data numerica conformada por:

edad	int64
dias_hospitalizado	int64
dias_seguimiento_hosp	int64
dias_uci	int64
dias_ucin	int64

data categorica conformada por:

sexo	object
criterio_fallecido	object
ubigeo_cdc	object
cdc_positividad	object
flag_hospitalizado	object
eess_renaes	object
evolucion_hosp_ultimo	object
ubigeo_inei_domicilio	object
flag_vacuna	object

variable objetivo o dependiente:

flag_uci	object

se analiza las datas numéricas.

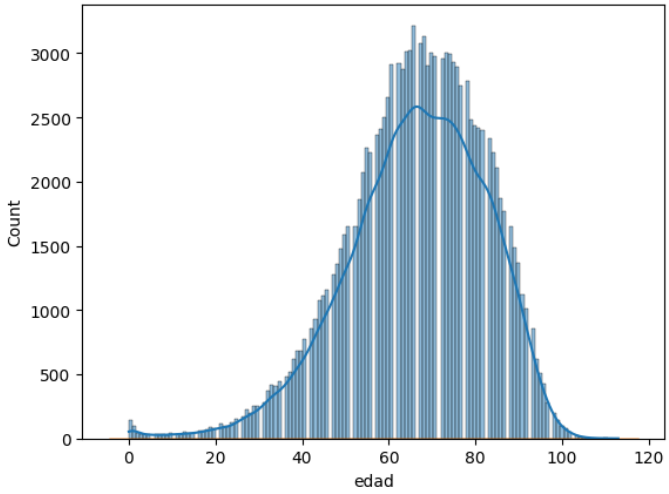


Figura 1. Histograma de la edad de los pacientes

Se realiza la limpieza de outlayers de la data numérica.
Realizamos la matriz de correlación entre nuestras variables númericas.

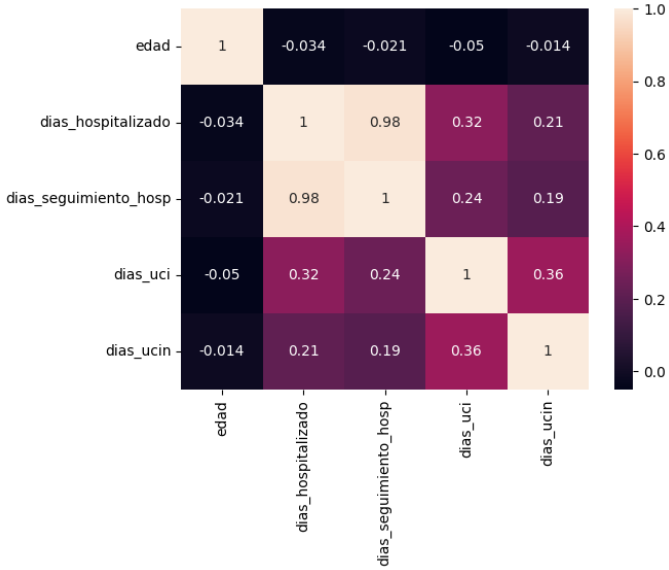


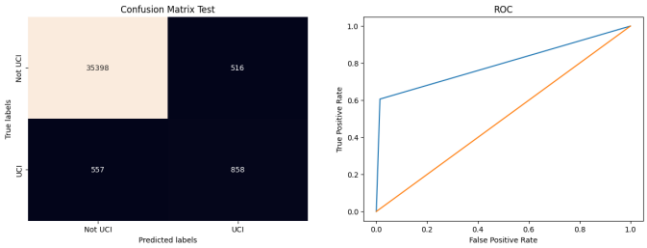
Figura 2. Matriz de correlación entre variables numéricas.

Para el caso de las variables categóricas, tenemos que transformarlas de forma que se pueda procesar para lo cual tratamos para obtener las variables dummies correspondientes.

V. DESCRIPCIÓN DE TÉCNICAS UTILIZADAS

De las técnicas que se van a emplear se tienen las siguientes matrices de confusión:

1. Árbol de decisión, con una precisión de 0.98 para no UCI, sin embargo para UCI 0.62.

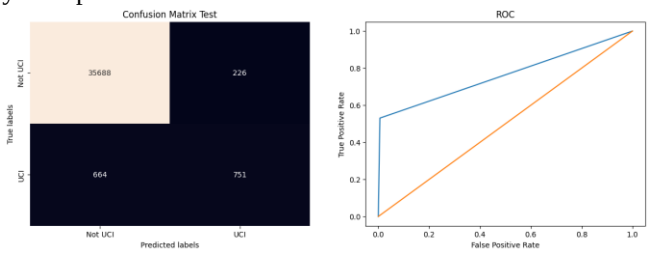


	precision	recall	f1-score	support
0	0.98	0.99	0.99	35914
1	0.62	0.61	0.62	1415
accuracy			0.97	37329
macro avg	0.80	0.80	0.80	37329
weighted avg	0.97	0.97	0.97	37329

AUC ROC : 0.7959963839804983

Figura 3. Matriz de confusión Árbol de decisión.

2. Regresión logística, con una precisión de 0.98 para No UCI y 0.77 para UCI.

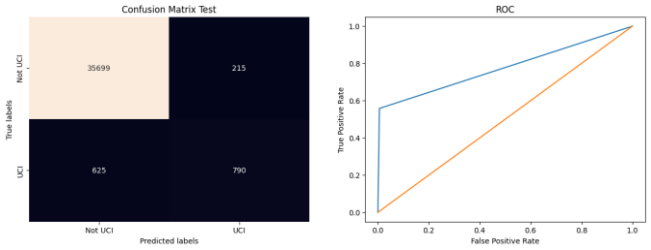


	precision	recall	f1-score	support
0	0.98	0.99	0.99	35914
1	0.77	0.53	0.63	1415
accuracy			0.98	37329
macro avg	0.88	0.76	0.81	37329
weighted avg	0.97	0.98	0.97	37329

AUC ROC : 0.7622246194334287

Figura 4. Matriz de confusión Regresión logística

3.Random forest, con una precisión de 0.98 para No UCI y 0.79 para UCI

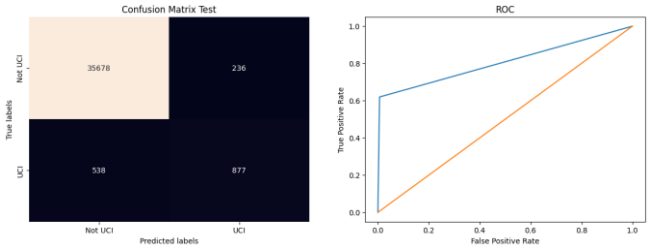


	precision	recall	f1-score	support
0	0.98	0.99	0.99	35914
1	0.79	0.56	0.65	1415
accuracy			0.98	37329
macro avg	0.88	0.78	0.82	37329
weighted avg	0.98	0.98	0.98	37329

AUC ROC : 0.7761586817822159

Figura 5. Matriz de confusión Random forest

4.XG Boost, con una precisión de 0.99 para No UCI y 0.79 para UCI.



	precision	recall	f1-score	support
0	0.99	0.99	0.99	35914
1	0.79	0.62	0.69	1415
accuracy			0.98	37329
macro avg	0.89	0.81	0.84	37329
weighted avg	0.98	0.98	0.98	37329

AUC ROC : 0.8066083661577884

Figura 6. Matriz de confusión Random forest

Podemos apreciar que al tener muchos mas datos con No UCI este resultado se entrega de mejor manera.

Revisando la data

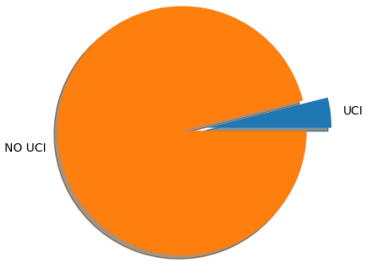
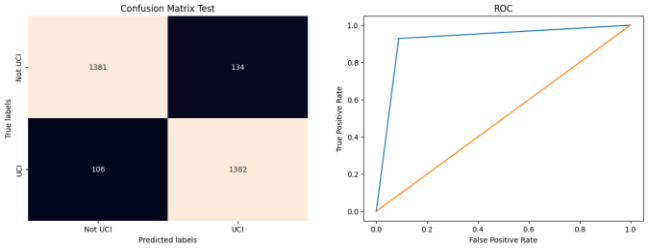


Figura 7. Pio de información de la variable dependiente UCI y NO UCI

Aplicamos una técnica de resample, donde vamos a igual la cantidad de datos a fin de mejorar el performance del modelo. Con los siguientes resultados.

1. Árbol de decisión, con una precisión de 0.92 en promedio.

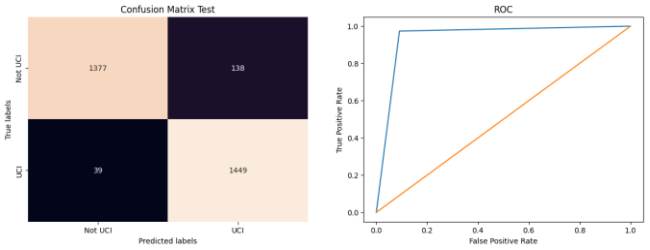


	precision	recall	f1-score	support
0	0.93	0.91	0.92	1515
1	0.91	0.93	0.92	1488
accuracy			0.92	3003
macro avg	0.92	0.92	0.92	3003
weighted avg	0.92	0.92	0.92	3003

AUC ROC : 0.9201572979878633

Figura 7. Matriz de confusión Árbol de decisión.

2. Regresión logística, con una precisión de 0.94 para el modelo en general.

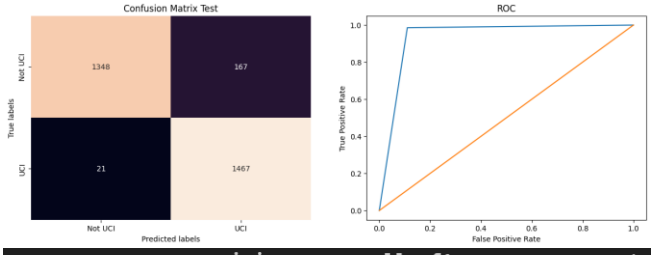


	precision	recall	f1-score	support
0	0.97	0.91	0.94	1515
1	0.91	0.97	0.94	1488
accuracy			0.94	3003
macro avg	0.94	0.94	0.94	3003
weighted avg	0.94	0.94	0.94	3003

AUC ROC : 0.941350606834877

Figura 8. Matriz de confusión Regresión logística

3.Random forest, con una precisión de 0.93 para el modelo en general.

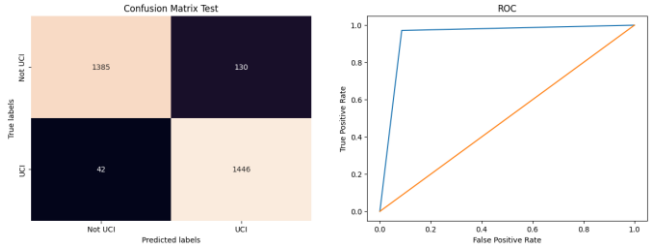


	precision	recall	f1-score	support
0	0.98	0.89	0.93	1515
1	0.90	0.99	0.94	1488
accuracy			0.94	3003
macro avg	0.94	0.94	0.94	3003
weighted avg	0.94	0.94	0.94	3003

AUC ROC : 0.9378280368359416

Figura 9. Matriz de confusión Random forest

4.XG Boost, con una precisión de 0.94 en general.



	precision	recall	f1-score	support
0	0.97	0.91	0.94	1515
1	0.92	0.97	0.94	1488
accuracy			0.94	3003
macro avg	0.94	0.94	0.94	3003
weighted avg	0.94	0.94	0.94	3003

AUC ROC : 0.9429828063451506

Figura 10. Matriz de confusión XG Boost

VI. CONCLUSIONES

La implementación de un modelo de machine learning para determinar el uso de camas UCI ha demostrado ser una herramienta valiosa para optimizar la gestión de recursos hospitalarios durante la pandemia de COVID-19. Se evaluaron varios modelos, incluyendo árboles de decisión, regresión logística, random forest y clasificador XGBoost, todos con una precisión inicial de 0.8 como máximo, luego e realizado un resample obtenemos una precisión de 0.94 lo cual es una mejora considerable en el modelo, esto debido a que se tenía mayoritariamente mayor data para valores o casos donde el paciente no entraba a UCI.A partir de esta evaluación, se pueden extraer las siguientes conclusiones:

1. Podemos concluir para este tipo de casos el modelo de XG Boost tiene mejor precisión alcanzando una precisión de 0.94.
2. El resample mejora el modelo de un 0.8 de precisión a 0.94 lo cual es considerable.
3. El modelo se puede mejorar con mayor cantidad de datos clínicos, como pueden ser enfermedades de los pacientes.

En resumen, aunque todos los modelos evaluados muestran una precisión aceptable después del resample realizado, cada uno tiene sus propias ventajas y desventajas en términos de interpretabilidad, manejo de complejidad y eficiencia. La elección del modelo óptimo es indiferente por la cercanía que tienen entre ellos. El uso de estos modelos puede mejorar el uso de camas UCI. Además, la implementación exitosa de estos modelos requerirá un flujo continuo de datos actualizados y la colaboración entre expertos en salud y en ciencia de datos

para garantizar su efectividad y aplicabilidad en situaciones reales.

VII. RECOMENDACIONES

-Se recomienda usar otros modelos para poder verificar y comprar los resultados.

-Se recomienda aumentar la cantidad de datos, sobre todo actuales, ya que la condición gracias a las vacunas afecta de cierta forma la aplicación de estos modelos en la actualidad.

-Se recomienda aumentar las variables ya que tener información clínica de ciertas enfermedades podría agregar valor al modelo.

VIII. REFERENCIAS

[1] "Machine Learning Applied to COVID-19: A Review of the Initial Pandemic Period" https://link.springer.com/article/10.1007/s44196-023-00236-3

[2] "Coronavirus Disease (COVID-19) Cases Analysis Using Machine-Learning Applications" https://link.springer.com/article/10.1007/s13204-021-01868-7

[3] "The Predictive Power of Data: Machine Learning Analysis for COVID-19 Mortality Based on Personal, Clinical, Preclinical, and Laboratory Variables" https://bmcinfectdis.biomedcentral.com/articles/10.1186/s12879-024-09298-w

[4] "Machine Learning-Based Prediction of COVID-19 Mortality Using Immunological and Metabolic Biomarkers" https://bmcdigitalhealth.biomedcentral.com/articles/10.1186/s44247-022-00001-0