

Clasificación de pacientes con COVID-19 con predisposición a cuidados intensivos usando técnicas de SVM y Random Forest

Juan Víctor Sanguinetti Valdivia
20153324@aloe.ulima.edu.pe
Universidad de Lima, Perú
doi: <https://doi.org/10.26439/ciis2021.5631>

El COVID-19 es una enfermedad respiratoria de alto nivel de contagio causada por el SARS-CoV-2. Los algoritmos predictivos permitirían la identificación de aquellas personas que podrían ser admitidas a cuidados intensivos. En este trabajo se siguió una metodología que consiste en la selección de un conjunto de datos que luego será procesado mediante el uso de técnicas como One Hot Encoding, MICE y LASSO. Luego se desarrollarán y evaluarán los modelos propuestos de Random Forest y Support Vector Machine mediante las métricas de sensibilidad, especificidad y AUC. Los resultados indican que el modelo de Random Forest obtiene un mejor rendimiento para la clasificación de los pacientes que son admitidos a una sala de cuidados intensivos.

CLASSIFICATION OF PATIENTS WITH COVID-19 PREDISPOSED TO INTENSIVE CARE USING SVM AND RANDOM FOREST TECHNIQUES

COVID-19 is a highly contagious respiratory disease caused by SARS-CoV-2. Predictive algorithms would allow the identification of those who could be admitted to intensive care. In this work, a methodology was followed that consists of selecting a data set, which will then be processed through techniques such as One Hot Encoding, MICE, and LASSO. Then the proposed Random Forest and Support Vector Machine models will be developed and evaluated using sensitivity, specificity, and AUC metrics. The results indicate that the Random Forest model obtains a better performance for the classification of patients who are admitted to an intensive care ward.

Clasificación de pacientes con COVID-19 con predisposición a cuidados intensivos usando técnicas de SVM y Random Forest

Juan Víctor Sanguinetti Valdivia
20153324@aloe.ulima.edu.pe

RESUMEN. El COVID-19 es una enfermedad respiratoria de alto nivel de contagio causada por el SARS-CoV-2. Los algoritmos predictivos permitirían la identificación de aquellas personas que podrían ser admitidas a cuidados intensivos. En este trabajo se siguió una metodología que consiste en la selección de un conjunto de datos que luego será procesado mediante el uso de técnicas como One Hot Encoding, MICE y LASSO. Luego se desarrollarán y evaluarán los modelos propuestos de Random Forest y Support Vector Machine mediante las métricas de sensibilidad, especificidad y AUC. Los resultados indican que el modelo de Random Forest obtiene un mejor rendimiento para la clasificación de los pacientes que son admitidos a una sala de cuidados intensivos.

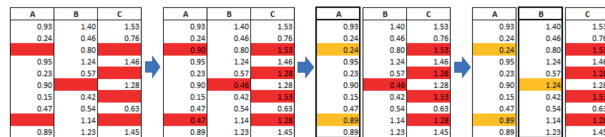
Introducción

La enfermedad del COVID-19 pone a prueba los sistemas de salud en todo el mundo, ya que se debe tener una capacidad máxima de hospitalización para tratar a los pacientes internados, y con mayor razón si es que la expansión del virus no es mitigada. La manera tradicional de analizar la evolución de los pacientes es de forma manual, lo cual a los encargados les toma tiempo. Los modelos predictivos permiten predecir la evolución de los pacientes que se encuentran en el hospital, con lo que se podría reducir la demanda pico, asegurando la priorización de recursos correctamente.



Metodología

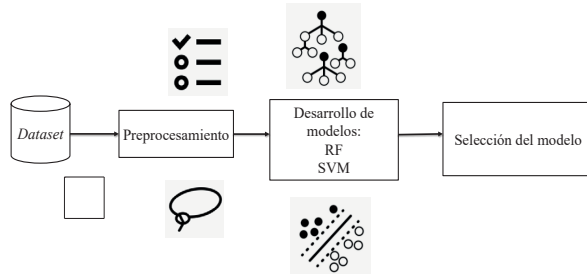
Se siguió una metodología que empieza con la exploración del conjunto de datos y la aplicación de técnicas de preprocesamiento como One Hot Encoding y MICE.



Representación de MICE

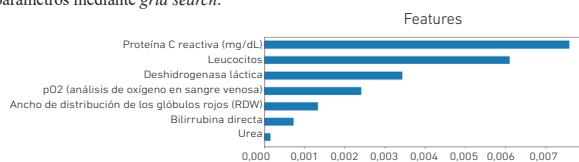
Fuente: Tomado de <https://cran.r-project.org/web/packages/miceRanger/vignettes/miceAlgorithm.html>

Para luego desarrollar los modelos de RF y SVM, los cuales serán posteriormente evaluados.



Experimentación

El conjunto de datos utilizado describía los resultados de pruebas de laboratorio tomadas al ingresar al hospital, además de información demográfica de los pacientes. Al aplicar LASSO se encontraron los atributos más relevantes para el rendimiento de los modelos. Adicionalmente, se hizo una selección de hiperparámetros mediante *grid search*.



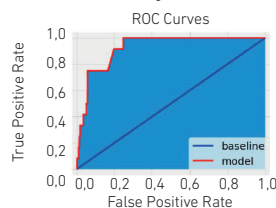
Resultados de la aplicación de LASSO. Elaboración propia

Modelo e hiperparámetro	Resultado
Random Forest	
Profundidad	5
Número de árboles	256
Support Vector Machine	
Regularización C	10
Kernel	rbf
Grado polinomial	-

Resultados de *grid search*. Elaboración propia

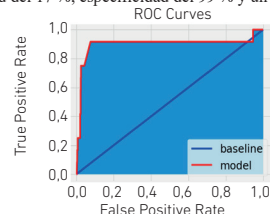
Resultados

Luego del desarrollo se siguió con la parte experimental, en la que Random Forest obtuvo una sensibilidad del 50 %, una especificidad del 95 % y un AUC del 90 %.



Curva ROC del Random Forest. Elaboración propia

En el caso de Support Vector Machine se observa que se tenía una sensibilidad del 17 %, especificidad del 99 % y un AUC del 90 %.



Curva ROC del SVM. Elaboración propia

Referencias

Burdick, H., Lam, C., Mataraso, S., Siefkas, A., Braden, C., Dellinger, R. P., McCoy, A., Vincent, J. L., Green-Saxena, A., Barnes, G., Hoffman, J., Gilbert, J., Pellegrini, E., y Das, R. (2020). Prediction of Respiratory Decompenation in COVID-19 Patients Using Machine Learning: The READY Trial. *Computers in Biology and Medicine*, 124, 103949. <https://doi.org/10.1016/j.compbiomed.2020.103949>

Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., Li, Y., Guan, W.,

Sang, L., Liu, J., Xu, Y., Chen, G., Guo, H., Guo, J., Chen, Z., Zhao, Y., Li, S., Zhang, N., Zhong, N., y He, J. (2020). Development and Validation of a Clinical Risk Score to Predict the Occurrence of Critical Illness in Hospitalized Patients with COVID-19. *JAMA Internal Medicine*, 160(6), 1061-1069. <https://doi.org/10.1001/jamainternmed.2020.2033>

Schwab, P., DuMont Schütte, A., Dietz, B., y Bauer, S. (2020). Clinical Predictive Models for COVID-19: Systematic Study. *Journal of Medical Internet Research*, 22(10), e21439. <https://doi.org/10.2196/21439>

Agradecimientos

Agradezco a los profesores Víctor Ayma y Juan Gutiérrez por el apoyo en la realización de este trabajo.