Predicción Pronóstico Pacientes COVID19

Alejandro M. Sevillano Mantas

Comisión: 29825

Data Science:

CoderHouse

Tabla de Contenidos

Alejandro M. Sevillano Mantas
Comisión: 29825

1. Introducción.

La enfermedad de COVID-19 ha causado más de 1 millón de muertes en el mundo desde su aparición a finales del 2019. Debido a las graves consecuencias que produjo esta enfermedad tanto para la salud como para el proceder social, se ha convertido en una enfermedad de interés general, y por tanto su estudio requiere un mayor detalle; poder predecir el pronóstico favorable o desfavorable de un paciente es de suma importancia.

Según la Organización mundial de la salud, esta enfermedad puede clasificarse en función de su sintomatología: enfermos muy graves; aquellos que son hospitalizados en cuidados intensivos con intubación (grado 1), enfermos que no presentas síntomas graves y son dados de altas tras unos días en el hospital (grado 5). Indistintamente a su grado de intensidad y cuadro clínico, muchos de los enfermos por COVID19 acudieron al hospital como primera medida para poder obtener un tratamiento acorde con su sintomatología. Esto derivo que en 2020 se produjera una crisis en casi todos los sistemas de salud mundial, colapsando las salas de atención primaria y de urgencias.

Aunque la mayoría de los pacientes que acudieron al hospital fueron dados de alta a los pocos días, muchos otros sufrieron un curso de la enfermedad desfavorable, necesitando incluso, cuidados intensivos y/o intubación o sucumbiendo frente a la enfermedad. Por ello, es de suma importancia poder atender a los pacientes según el pronóstico de evolución de la enfermedad; poder derivar a aquellos pacientes con posible transcurso desfavorable hacia cuidados más concretos ayudaría a la pronta intervención de tratamientos específicos para cuadros agudos de COVID19.

2. Objetivos

Para aliviar la carga inicial de pacientes afectados por COVID19 y por tanto la posibilidad de colapsar las salas de atención primaria y de urgencias, se propone generar un modelo de predicción que ayude a clasificar a pacientes atendiendo a su cuadro clínico inicial (día de hospitalización, t_0) en dos grupos según el pronóstico de curso de la enfermedad: pronostico favorable – pronostico desfavorable.

Los algoritmos seleccionados son:

- DecisionTree
- RamdonForest
- Bagging estimator
- Boosting estimators (AdaBoosting, Gradient Boosting and XGB)

Los datos para la realización de este trabajo se han obtenido de una publicación del Cell-Reports donde se realiza un estudio proteómico de pacientes afectados por COVID19[i]

3. Limpieza y preparación de los datos

Los datos clínicos facilitados por la empresa Olink[i], fueron cargados en Jupiter Notebook y analizados brevemente con el objetivo de comprobar la salud del dataset. Se realizó el siguiente procedimiento:

1) **Comprobación de que los datos se han cargado correctamente**.

Para ello se inspeccionó las posibles anomalías tanto en el encabezado como la cola del dataset. A su vez se verificó que el archivo original *.txt* se cargó correctamente usando el parámetro *'sep=';"* de la función *pd.read_csv()*.

2) **Inspección de los tipos de datos que compone el dataset.**

El dataset se compone de un total de 384 sujetos, con un total de 44 variables numéricas.

3) **Detección y tratamientos de valores nulos (NaN).**

Se observó una gran cantidad de valores NaN en variables diferentes a las correspondientes tomadas a t_0. Esto es debido a que las variables se miden en intervalos de tiempo correspondientes a la estancia del paciente en el hospital, por lo que estas variables no se les puede otorgar otro valor como por ejemplo la mediana o media de la serie de su grupo. Se decide únicamente tener en cuenta los valores tomados a t_0.

4) **Creación de un nuevo dataset.**

Se observó alrededor de 100 pacientes COVID19 negativos, por lo que se descartaron para este estudio. Se utilizó para el siguiente apartado este dataset generado con pacientes COVID19 positivos.

4. Exploratory Data Analysis.

Desde la aparición de la enfermedad COVID19, se han identificado múltiples factores de riesgo asociados a la evolución desfavorable de la infección por SARS-COV2[ii].

Para poder entender mejor como las diferentes variables clínicas pueden derivar en un factor de riesgo, se estudió como dichas variables se comportan en el dataset.

En esta sección, se ha estudiado el comportamiento de las variables a lo largo del dataset, incluyendo tres secciones:

1) Creación de nuevas variables.

    a) *Pronóstico (‘prognosis’)*. Esta nueva variable va a ser la variable objetivo (target) para nuestro modelo. Surge a raíz del grado de enfermedad en la que se encuentra el paciente al último día de estudio (Acuity 28) ya que en esta variable se incluyen el grado de enfermedad de pacientes dados de alta con anterioridad o pacientes que no sobrevivieron a la enfermedad. Siendo pronóstico favorable aquellos sujetos con grado 4-5 (pacientes sin intubar, con síntomas leves o dados de alta) y no favorables aquellos que tienen grado 1-2 (situación crítica o fallecidos). Para los pacientes que al día 28 seguían en grado 3, se les asigno la etiqueta de indeterminados (*indeterminate*). Tras un análisis posterior, estos pacientes fueron descartados debido su baja representación en el dataset.
    b) Tiempo en el hospital (‘days_hosp)l. Esta variable se ha generado con el objetivo de conocer cuántos días el paciente está hospitalizado.

2) Análisis univariado.

Durante este tipo de análisis, se ha estudiado el comportamiento de todas las variables clínicas tomadas al tiempo inicial del estudio (t_0). A continuación, se muestran las dos figuras más representativas: Variación de la edad de los pacientes y niveles de proteína c- reactiva.

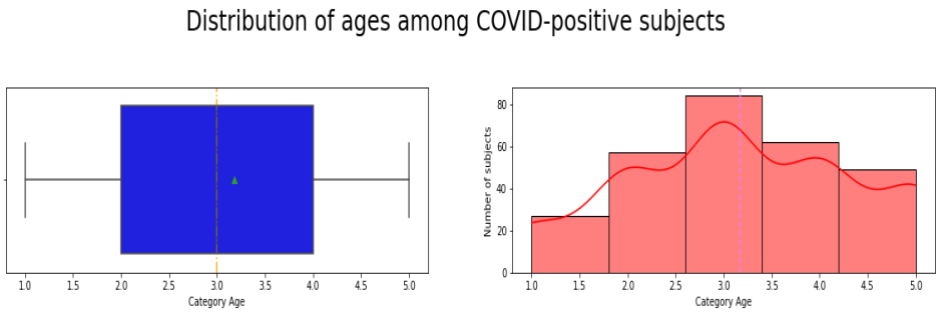

Figura 1. Distribución de la variable edad a lo largo del dataset. Análisis de la edad de los pacientes que ingresan en el hospital. La media viene representada por un triángulo verde en la figura de caja de bigotes y por unas rayas moradas en la figura de histograma de barras.

Como se observa en la figura 1, la variable edad (*age*) sigue una distribución normal con un ligero desplazamiento hacia la derecha de sus datos. El análisis presente indica que la mayoría de los individuos que llegan al hospital tienen una edad media de entre 50-64 años de edad.

La figura 2 por el contrario, muestra una clara desviación hacia la izquierda, por lo que se espera la presencia de outliers para niveles bajos de proteína C reactiva. Al encontrar los niveles desplazados se intuye que la mayoría de los pacientes que ingresan en el hospital tienen la proteína c reactiva elevada, sugiriendo el desarrollo de una afección inmune aguda.

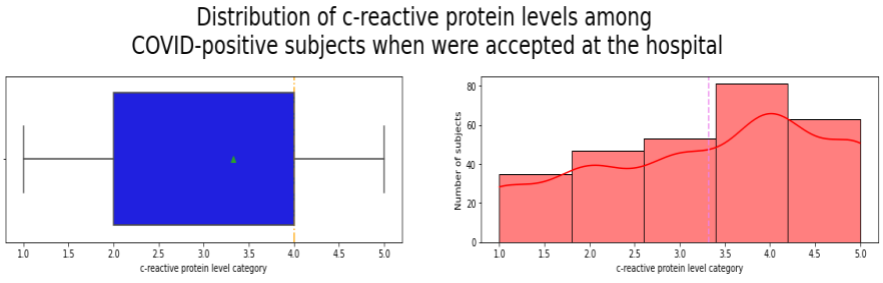

Figura 2. Análisis de la distribución de la proteína c reactiva. La proteína C-reactiva se encuentra en niveles altos el día t_0, indicando una afección inmune aguda por parte de los pacientes que ingresas en el hospital. La media viene representada por un triángulo verde en la figura de caja de bigotes y por unas rayas moradas en la figura de histograma de barras.

Alejandro M. Sevillano Mantas
Comisión: 29825

3) Análisis multivariado.

En este análisis, se ha estudiado la relación entre diferentes variables y la variable prognosis con el objetivo de conocer dicho comportamiento.

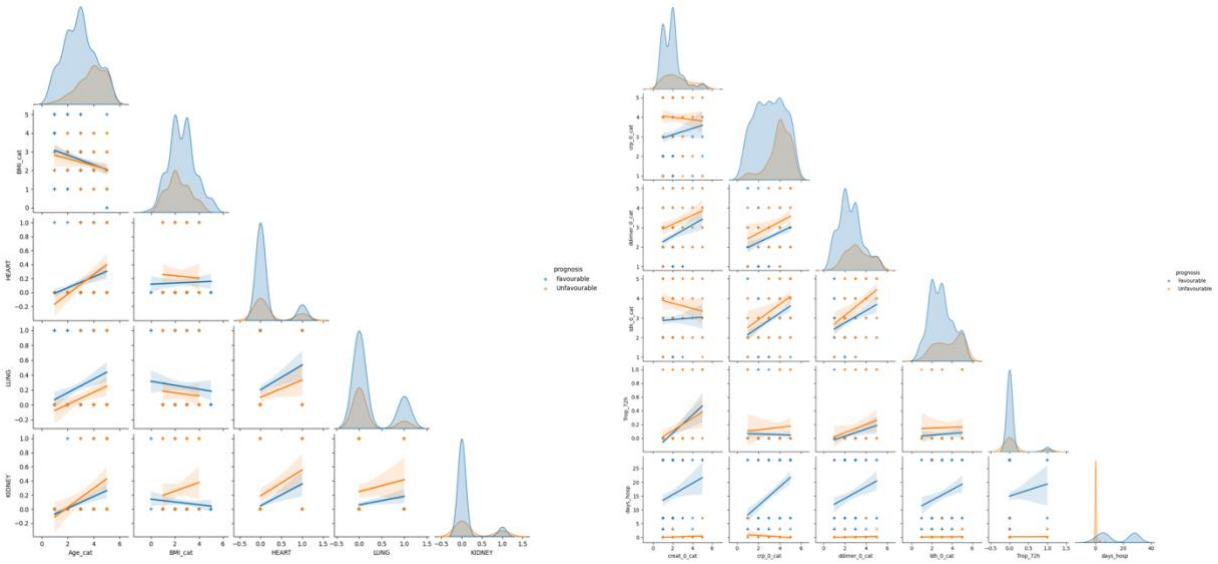

Figura 3. Análisis univariado de las variables más representativas. Se analizó de forma individual cada variable a lo largo del dataset con el fin de entender la variación de los valores clínicos de los pacientes. En este análisis solo se tienen en cuenta variables tomadas a t_0.

Para la gran mayoría de las variables estudiadas, no se aprecia una clara diferenciación entre prognosis favorable y desfavorable. Únicamente se puede apreciar dicha diferenciación cuando se combinan las variables “días en el hospital (*days_hosp*)” con las variables: ‘*creat_0_cat', 'crp_0_cat', 'ddimer_0_cat', 'ldh_0_cat'* y *'Trop_72h*’. Por lo que sería importante tener en cuenta estas variables para la modelización.

Alejandro M. Sevillano Mantas
Comisión: 29825

- **Variación de la prognosis con la edad.**

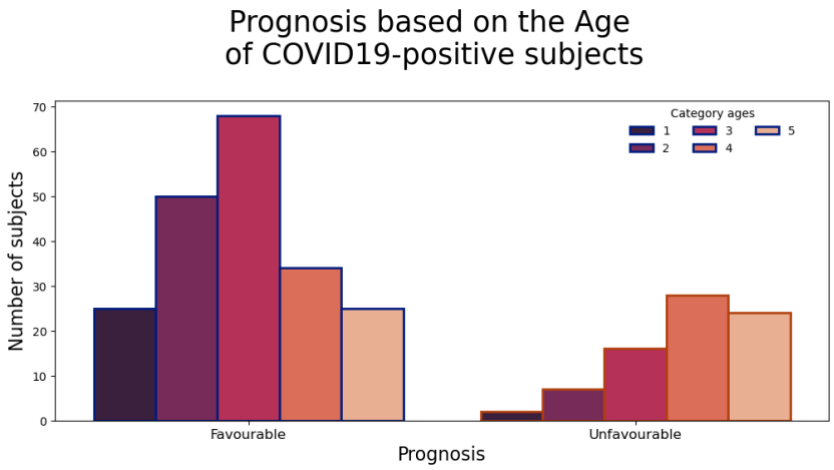

Figura 4. Variación de la prognosis con la edad. Grafica de barras representando el número de pacientes y su edad (categoría) y su relación con su prognosis. Age Category: 1 = 20-34 2 = 36-49 3 = 50-64 4 = 65-79 5 = 80+

La figura 4 representa la cantidad de pacientes con una prognosis determinada y la categoría de edad en la que se encuentran. Como se observa, hay una mayor cantidad de pacientes mayores de 65 años (categoría 4 y 5) en el grupo con prognosis desfavorable que en el grupo con prognosis favorable. Esto indica que la edad es un factor de riesgo importante y que personas mayores de 65 pueden tener más probabilidad de desarrollar un cuadro desfavorable.

- **Variación de la prognosis con la hipertensión.**

Otra gráfica importante es la mostrada en la figura 5, donde se representa el número de pacientes con diferente prognosis y la existencia de hipertensión a la hora de su hospitalización. Como se observa, la relación entre sujetos con hipertensión y sin hipertensión es mayor en el grupo de pacientes con una prognosis desfavorable, indicando que la hipertensión es otro factor de riesgo en pacientes con COVID19.

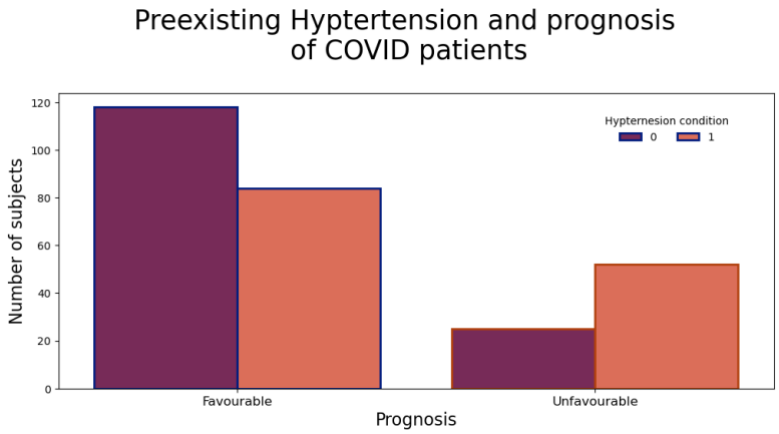

Figura 5. Variación de la prognosis con la hipertensión. Grafica de barras representando el número de pacientes y la presencia de hipertensión (0:no, 1:sí) y su relación con su prognosis.

- **Variación de la prognosis con los niveles de proteína c-reactiva**

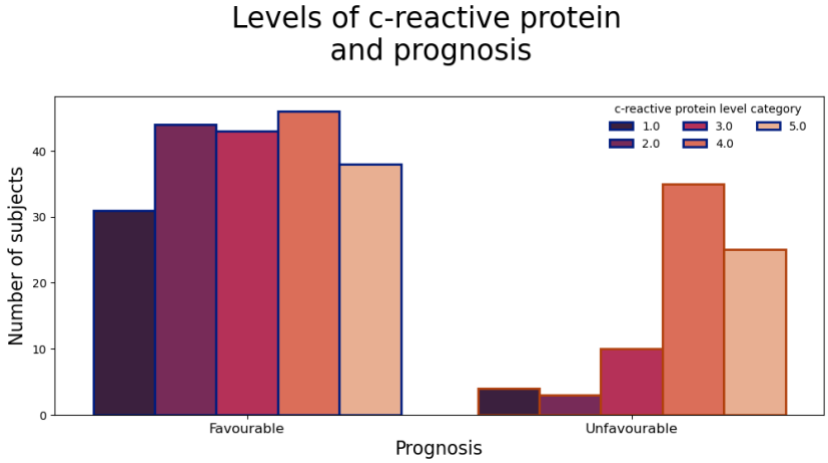

Figura 6. Variación de la prognosis con los niveles de la proteína c-reactiva. Grafica de barras representando el número de pacientes y los niveles de proteína c-reactiva y su relación con su prognosis. Leyenda: 1 = 0-19.9 2 = 20-59.0 3 = 60-99.9 4 = 100-179 5 = 180+ units/ml

La proteína c-reactiva es un biomarcador que se encuentra en la sangre e indica la presencia de una respuesta inmune aguda. También sirve como indicador de fallo hepático en respuesta a la activación abrupta de la respuesta inmune frente a una infección. Como se observa en la figura 6, la mayoría de los pacientes que derivan en una prognosis desfavorable poseen aumentada la cantidad de proteína c-reactiva al llegar al hospital, por lo que este parámetro es otro factor de riesgo.

5. Preparación de datos para modelización.

Para la realización del modelo, se ha tenido en cuenta únicamente las variables correspondientes al tiempo inicial del estudio(t_0), ya que el objetivo principal es realizar un modelo de predicción de la prognosis de los pacientes con los datos clínicos disponibles al primer día de ingreso hospitalario. Con este fin, primero se han determinado las variables dependientes o target (Y) y las variables independientes dependientes (X)

A. **Selección de X, Y**.

a. X→ Se asignaron las variables independientes: *'Age_cat', 'BMI_cat',' HEART', 'LUNG', 'KIDNEY', 'DIABETES', 'HTN', 'IMMUNO', 'Resp_Symp', 'Fever_Sympt', 'GI_Symp', 'Acuity_0', 'abs_neut_0_cat', 'abs_lymph_0_cat', 'abs_mono_0_cat', 'creat_0_cat', 'crp_0_cat', 'ddimer_0_cat', 'ldh_0_cat', 'Trop_72h'* y *'days_hosp'*.

b. Y→ Se asignó la variable target a: *'prognosis'*.

B. **Reducción de dimensiones**. Con el objetivo de poder reducir la dimensionalidad de los datos, se realizó el análisis de los componentes principales o PCA. Se encontró que únicamente dos variables podrían explicar la variación del 30% de los datos, mientras que el resto de las variables poseen un peso similar entre ellas.

6. Estudio modelo clasificación supervisada-I

Para la realización del modelo de predicción se utilizaron los algoritmos mencionados en el apartado 2.

El criterio de evaluación para los diferentes modelos es el siguiente:

- Si el modelo predice una buena prognosis o pronóstico positivo (favorable) cuando el paciente en realidad tiene una prognosis negativa (desfavorable) (**Falso Negativo**), podría ocurrir que un paciente derive en una situación crítica debido a la asignación *a posteriori* de un tratamiento erróneo.

- Si el modelo predice una prognosis desfavorable cuando el paciente en realidad tiene una buena prognosis, puede ocurrir que el paciente obtenga un tratamiento más agresivo y por lo tanto poder derivar en otras patologías diferentes a las producidas por el COVID, o bien podría darse de alta con mayor antelación (**Falso Positivos**).

Aunque en este caso los falsos positivos son importantes, una predicción cuyo peso se centre en reducir los falsos positivos, sacrificando la detección de falsos negativos podría derivar en un mayor número de pacientes con peor prognosis y tratamientos erróneos, llevando al individuo a una prognosis desfavorable. Por tanto, se considera como factor más importante la optimización de la sensibilidad o reducción de falsos positivos.

Previamente, las variables X e Y se dividieron en grupos de entrenamiento (*train*) y testeo (*test*).

En todos los modelos probados, se realizó una optimización de el parámetro *recall* o sensibilidad (*hypertuning*) con el objetivo de generar un modelo que detecte el menor número de falsos positivos tanto para el set de entrenamiento como el set de testeo.

Tabla 1. Métricas de los diferentes modelos. Métricas obtenidas para los diferentes modelos usando como base el set de entrenamiento.

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Adaboost Classifier	Adaboost Tuned	Gradient Boost	Gradient Boost Tuned	XGB	XGB Tuned	Stacking Classifier	Bagging Classifier	Bagging Classifier Tuned
Accuracy	1.00	0.65	1.00	0.97	1.00	0.84	1.00	1.00	1.00	0.97	0.98	1.00	0.97
Recall	1.00	0.57	1.00	1.00	1.00	0.84	1.00	1.00	1.00	1.00	1.00	1.00	0.99
Precision	1.00	0.91	1.00	0.96	1.00	0.93	1.00	1.00	1.00	0.96	0.97	1.00	0.98
F1	1.00	0.70	1.00	0.98	1.00	0.88	1.00	1.00	1.00	0.98	0.99	1.00	0.98

Tabla 2. Métricas de los diferentes modelos. Métricas obtenidas de los diferentes modelos usando como base el set de testeo.

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Adaboost Classifier	Adaboost Tuned	Gradient Boost	Gradient Boost Tuned	XGB	XGB Tuned	Stacking Classifier	Bagging Classifier	Bagging Classifier Tuned
Accuracy	0.71	0.55	0.85	0.82	0.80	0.75	0.81	0.82	0.80	0.79	0.75	0.76	0.82
Recall	0.75	0.52	0.97	0.95	0.89	0.82	0.87	0.90	0.87	0.92	0.87	0.82	0.93
Precision	0.84	0.78	0.84	0.83	0.84	0.83	0.87	0.86	0.85	0.81	0.80	0.85	0.84
F1	0.79	0.63	0.90	0.89	0.86	0.83	0.87	0.88	0.86	0.86	0.83	0.83	0.88

Atendiendo a los parámetros obtenidos en los diferentes algoritmos de clasificación supervisada podemos concluir lo siguiente:

- En general los modelos poseen *overfitting* debido principalmente a que el peso de la variable-target creada: 'prognosis', posee macho peso una de sus observaciones, se recomienda realizar un *oversampling* para generar mayores observaciones para *'unfavourable'*.

- La métrica más importante para este caso de clasificación es la sensibilidad o *recall*, ya que la optimización de la predicción del número de falsos positivos (falsos pacientes con buena prognosis) podría llevar a pacientes con una predicción peor derivar a una situación crítica debido a un tratamiento equivocado.

- Teniendo en cuenta la métrica de *recall*, los modelos que menor *overfitting* y mejor rendimiento ofrecen son:

- Set de entrenamiento: AdaBoost Classifier y Bagging Classifier.

- Set de testeo: Random Forest (pre- and after- tuning) y Bagging Classifier.

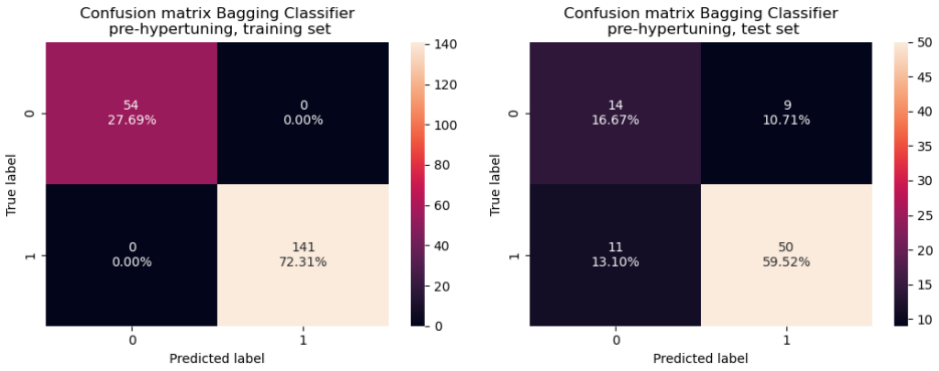

Figura 7. Matriz de confusión. Representación grafica de las métricas del algoritmo Bagging Classifier, donde la etiqueta *True label* indica los valores verdaderos mientras que la etiqueta *Predicted label* muestra la predicción basada en el algoritmo

En la figura 7 se muestra las matrices de confusión (*Confusion Matrix*) donde se reflejan el número y porcentaje de los diferentes pacientes captados por el algoritmo 'Bagging Classifier' previo a la optimización de la métrica sensibilidad. Este algoritmo muestra una gran diferencia entre el número de verdaderos positivos (1,1), verdaderos negativos (0,0), falsos negativos (1,0) y falsos positivos (0,1) entre el set de entrenamiento y el set de testeo, indicando que el este modelo, a pesar de las metricas obtenidas, no posee un buen rendimiento a la hora de clasificar en el set de testeo.

7. Estudio modelo clasificacion supervisada-II

Debido al *oversampling* en las muestras, seguidamente al estudio descrito en el apartado anterior, se realizó el estudio de varios modelos con un nuevo set de datos derivados de la generación sintética de nuevas observaciones mediante el algoritmo SMOTE.

Tras la aplicación de dicho algoritmo, el número de observaciones es el que sigue:

```
Antes OverSampling, número de etiquetas '1': 141
Antes OverSampling, número de etiquetas '0': 54

Después OverSampling, número de etiquetas '1': 141
Después OverSampling, número de etiquetas '0': 126

Después OverSampling, the shape of train_X: (267, 20)
Después OverSampling, the shape of train_y: (267,)
```

Tabla 4. Métricas de los diferentes modelos para el set de entrenamiento. Métricas obtenidas de los diferentes modelos usando como base el set de *training_oversampled*.

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Adaboost Classifier	Adaboost Tuned	Gradient Boost	Gradient Boost Tuned	XGB	XGB Tuned	Stacking Classifier	Bagging Classifier	Bagging Classifier Tuned
Accuracy	1.00	0.76	1.00	0.90	0.86	0.78	0.98	1.00	1.00	0.93	0.96	0.99	0.97
Recall	1.00	0.87	1.00	0.91	0.88	0.85	0.99	1.00	1.00	0.94	0.97	0.99	0.96
Precision	1.00	0.81	1.00	0.91	0.86	0.75	0.97	1.00	1.00	0.92	0.95	0.99	0.98
F1	1.00	0.84	1.00	0.91	0.87	0.80	0.98	1.00	1.00	0.93	0.96	0.99	0.97

Tabla 5. Métricas de los diferentes modelos para el set de testeo. Métricas obtenidas de los diferentes modelos usando como base el set de *training_oversampled*.

	Decision Tree	Decision Tree Tuned	Random Forest	Random Forest Tuned	Adaboost Classifier	Adaboost Tuned	Gradient Boost	Gradient Boost Tuned	XGB	XGB Tuned	Stacking Classifier	Bagging Classifier	Bagging Classifier Tuned
Accuracy	0.63	0.70	0.70	0.76	0.76	0.73	0.71	0.70	0.69	0.73	0.70	0.65	0.97
Recall	0.69	0.89	0.87	0.87	0.80	0.87	0.79	0.79	0.79	0.82	0.85	0.69	0.96
Precision	0.78	0.75	0.76	0.82	0.86	0.78	0.81	0.80	0.79	0.81	0.76	0.81	0.98
F1	0.73	0.81	0.81	0.84	0.83	0.82	0.80	0.79	0.79	0.81	0.81	0.74	0.97

Atendiendo a los parámetros obtenidos en los diferentes algoritmos de clasificación supervisada podemos concluir lo siguiente:

- En general los modelos han mejorado bastante el rendimiento, evitando el *overfitting*.

- Teniendo en cuenta la métrica de *recall*, los modelos que mejor rendimiento ofrecen son:

  - Set de entrenamiento: Los mejores modelos son Decision-Tree tuned, AdaBooster y AdaBooster-tuned

  - Set de testeo: Tras la optimización de las métricas (*hypertuning*) todos los modelos mejoran su rendimiento notablemente evitando en casi todos el *overfitting*.

Teniendo en cuenta la similitud entre las medidas en el set de entrenamiento y el set de testeo, el modelo de Adaboost posee mayor similitud entre ambos grupos se puede considerar un buen modelo para la predicción de pacientes con una prognosis desfavorable.

La figura 8 muestra la matriz de confusión del algoritmo '*AdaBoost*' después de la optimización de la métrica sensibilidad. Como se observa, tras la aplicación de SMAT en el set de entrenamiento hay un mayor número de casos negativos, esto permite al modelo reconocer de forma más eficiente el número de positivos y negativos verdaderos y, por tanto, el número de falsos positivos.

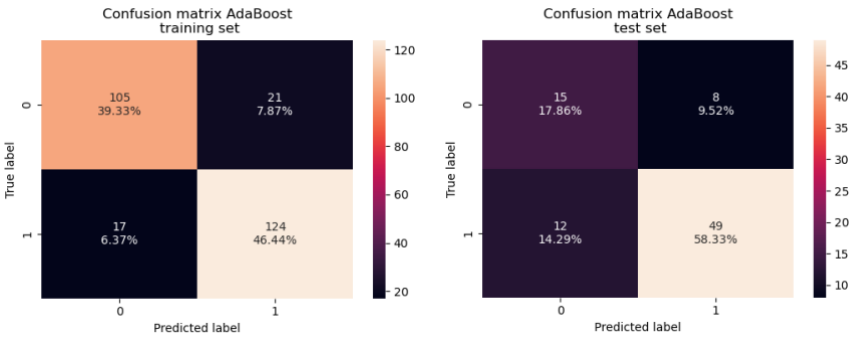

Figura 8. Matriz de Confusión del modelo *AdaBoost*. Para este modelo se aplicó un set de entrenamiento con observaciones sintéticas donde los valores de la categoría 'prognosis' se balancearon usando la técnica de SMAT. La etiqueta *True label* indica los valores verdaderos mientras que la etiqueta *Predicted label* muestra la predicción basada en el algoritmo

8. Modelos seleccionados

Tras el estudio de varios modelos de clasificación, el algoritmo de AdaBoost Classifier es el que mejor rendimiento ofrece tras la adición de datos sintéticos mediante SMAT.

En la figura 9, se observa las variables con mayor importancia relativa para el modelo de clasificación. Valores de neutrófilos y monocitos se encontraron relativamente bajos en la mayoría de los pacientes al día t_0. Por otro lado, durante la fase exploratoria, se pudo observar que los niveles de la proteína c reactiva y la edad, eran factores de riesgo, debido a sus altos valores en pacientes con prognosis desfavorable.

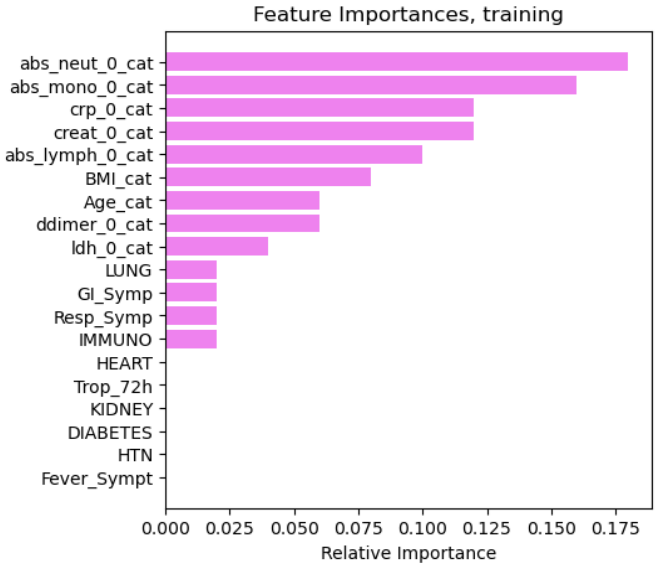

Figura 9. Gráfico de barras representando la importancia relativa de las diferentes variables.

9. Conclusión.

En el presente estudio se ha realizado el análisis de un conjunto de datos clínicos con el fin de determinar un modelo de predicción que ayude a clasificar pacientes según su sintomatología/valores clínicos correspondientes al primer día de ingreso. Una vez clasificados los pacientes con predicción favorable o predicción desfavorable, deberían ser sometidos al tratamiento correspondiente atendiendo a su gravedad, consiguiendo así una disminución de la carga hospitalaria por parte de pacientes cuya prognosis es positiva en el tiempo. Los modelos presentados, reflejan un claro *overfitting* donde predomina el valor 'favorable' de la variable prognosis, puesto que cerca del 72% de los pacientes transcurren la enfermedad de forma favorable. Por ello, se generaron valores

sintéticos de '*unfavourable'* de la variable prognosis, consiguiendo una mejora en la predicción de la mayoría de los modelos. Los modelos de 'Decision Tree-tuned','AdaBoost Classifier' y 'AdaBoost Classifier – tuned' son los que mejor rendimiento ofrecen en cuestión de sensibilidad de exactitud (*Accuracy*), siendo el algoritmo de AdaBoost Classifier el que posee métricas similares entre el set de entrenamiento y el set de testeo.

Para poder clasificar pacientes con un posible buen pronóstico y pacientes con una posible evolución desfavorable, se debe atender en el momento de hospitalización a los niveles altos de neutrófilos, monocitos, proteína c-reactiva y creatinina. A su vez, sujetos mayores de 65 deberían tratarse como pacientes con mayor riesgo de sufrir una evolución negativa.

10. Direcciones futuras.

El modelos propuestos en este estudio es el de AdaBoost Classifier Que poseen una sensibilidad (capacidad de reducir el número de falsos positivos) de 80% pudiendo predecir cerca del 76% de los casos. Para poder mejorar estas predicciones se propone la implementación de las siguientes acciones:

I. Adición de nuevas variables como: carga viral inicial, historia oncológica, actividad física, historial fumador y coinfección con otros virus[iii] entre otras.
II. Incrementar el número de pacientes. Con el objetivo de aumentar el volumen de observables para conseguir modelos de predicción mejores.
III. Inclusión de datos de otros hospitales. Se considera necesario la participación de otros hospitales para conseguir aumentar la base de datos.
IV. Implementación de nuevas medidas de acción para pacientes con pronóstico desfavorable. Teniendo en cuenta los factores de riesgo mostrados en los modelos anteriores, aquellos pacientes con altos niveles de factores de riesgo deben ser derivados a las salas correspondientes para un tratamiento específico.
V. Estudiar la evolución de los niveles de las variables: neutrófilos, monocitos, proteína c-reactiva y creatinina en pacientes a lo largo del estudio.

11. Anexo I. Diccionario de Variables.

Variable	Description
subject_id	Subject ID
COVID	COVID status (tested positive prior to enrollment or during hospitalization) 0 = negative 1= positive
Age cat	Age category 1 = 20-34 2 = 36-49 3 = 50-64 4 = 65-79 5 = 80+
BMI cat	Body mass index: 0 = <18.5 (underweight) 1 = 18.5-24.9 (normal) 2 = 25.0-29.9 (overweight) 3 = 30.0-39.9 (obese) 4 = >=40 (severely obese) 5 = Unknown
HEART	Pre-existing heart disease – HEART - (coronary artery disease, congestive heart failure, valvular disease) 0 = No 1 = Yes
LUNG	Pre-existing lung disease – LUNG - (asthma, COPD, requiring home O2, any chronic lung condition) 0 = No 1 = Yes
KIDNEY	Pre-existing kidney disease – KIDNEY - (chronic kidney disease, baseline creatinine >1.5, ESRD) 0 = No 1 = Yes
DIABETES	Pre-existing diabetes – DIABETES - (pre-diabetes, insulin and non-insulin dependent diabetes) 0 = No 1 = Yes
HTN	Pre-existing hypertension - HTN 0 = No 1 = Yes
IMMUNO	Pre-existing immunocompromised condition – IMMUNO (active cancer, chemotherapy, transplant, immunosuppressant agents, aspenic) 0 = No 1 = Yes
Resp_Symp	Respiratory symptoms – Symp_Resp (sore throat, congestion, productive or dry cough, shortness of breath or hypoxia, or chest pain) 0 = No 1 = Yes
Fever_Sympt	Febrile symptom

GI_Symp	Any GI related symptoms at presentation (abdominal pain, nausea, vomiting, diarrhea)
D0_draw	Study draw on Day 0 0 = No 1 = Yes
D3_draw	Study draw in Day 3 window (study day 2 to study day 4) 0 = No 1 = Yes
D7_draw	Study draw in Day 7 window (study day 5 to study day 9) 0 = No 1 = Yes
DE_draw	Study draw for an event - event can include decompensation such as ICU admit or intubation, or extubation 0 = No 1 = Yes
Acuity 0	Acuity score maximum for day 0 study window - highest Acuity within Day 0 window (enrollment plus 24 hours) : 1 = Death 2 = Intubated / ventilated 3 = Hospitalized, supplementary O2 required 4 = Hospitalized, no supplementary O2 required 5 = Discharged / Not hospitalized
Acuity 3	Acuity score maximum for day 3 study window (study day 2 to study day 4): 1 = Death 2 = Intubated / ventilated 3 = Hospitalized, supplementary O2 required 4 = Hospitalized, no supplementary O2 required 5 = Discharged / Not hospitalized
Acuity 7	Acuity score maximum for day 7 study window (study day 5 to study day 9): 1 = Death 2 = Intubated / ventilated 3 = Hospitalized, supplementary O2 required 4 = Hospitalized, no supplementary O2 required 5 = Discharged / Not hospitalized
Acuity 28	Acuity score on day 28: 1 = Death 2 = Intubated / ventilated 3 = Hospitalized, supplementary O2 required 4 = Hospitalized, no supplementary O2 required 5 = Discharged / Not hospitalized
Acuity max	Acuity max is the highest Acuity level between Day 0 -28 1 = Death within 28 days 2 = Intubated / ventilated, survived to 28 days 3 = Hospitalized, supplementary O2 required, survived to 28 days 4 = Hospitalized, no supplementary O2 required, survived to 28 days 5 = Discharged, was not admitted to hospital within 28 day window, survived
abs_neut_0_cat	Absulute neutrophil count day 0 category: 1 = 0-0.99 2 = 1.0-3.99 3 = 4.0-7.99 4 = 8.0-11.99 5 = 12+
abs_lymph_0_cat	Absolute lymphocyte count day 0 category: 1 = 0-0.49 2 = 0.50-0.99 3 = 1.00-1.49 4 = 1.50-1.99 5 = 2+

abs_mono_0_cat	Absolute monocyte day 0 category 1 = 0-0.24 2 = 0.25-0.49 3 = 0.50-0.74 4 = 0.75-0.99 5 = 1.0+
creat_0_cat	Creatinine day 0 category 1 = 0-0.79 2 = 0.80-1.19 3 = 1.20-1.79 4 = 1.80-2.99 5 = 3+
crp_0_cat	c-reactive protein day 0 category: 1 = 0-19.9 2 = 20-59.0 3 = 60-99.9 4 = 100-179 5 = 180+
ddimer_0_cat	D-dimer day 0 category: 1 = 0-499 2 = 500-999 3 = 1000-1999 4 = 2000-3999 5 = 4000+
ldh_0_cat	Lactate dehydrogenase day 0 category: 1 = 0-200 2 = 200-299 3 = 300-399 4 = 400-499 5 = 500+
Trop_72h	Cardiac event – Trop_72h - (hs-cTn =>100 within first 72 hours of presentation) 0 = No 1 = Yes
abs_neut_3_cat	Absulute neutrophil count day 3 category: 1 = 0-0.99 2 = 1.0-3.99 3 = 4.0-7.99 4 = 8.0-11.99 5 = 12+
abs_lymph_3_cat	Absolute lymphocyte count day 3 category: 1 = 0-0.49 2 = 0.50-0.99 3 = 1.00-1.49 4 = 1.50-1.99 5 = 2+
abs_mono_3_cat	Absolute monocyte count day 3 category: 1 = 0-0.24 2 = 0.25-0.49 3 = 0.50-0.74 4 = 0.75-0.99 5 = 1.0+
creat_3_cat	Creatinine day 3 category 1 = 0-0.79 2 = 0.80-1.19 3 = 1.20-1.79 4 = 1.80-2.99 5 = 3+
crp_3_cat	c-reactive protein day 3 category: 1 = 0-19.9 2 = 20-59.0 3 = 60-99.9 4 = 100-179 5 = 180+
ddimer_3_cat	D-dimer day 3 category: 1 = 0-499 2 = 500-999 3 = 1000-1999 4 = 2000-3999 5 = 4000+
ldh_3_cat	Lactate dehydrogenase day 3 category: 1 = 0-200 2 = 200-299 3 = 300-399 4 = 400-499 5 = 500+
abs_neut_7_cat	Absulute neutrophil count day 7 category: 1 = 0-0.99 2 = 1.0-3.99 3 = 4.0-7.99 4 = 8.0-11.99 5 = 12+

abs_lymph_7_cat	Absolute lymphocyte count day 7 category: 1 = 0-0.49 2 = 0.50-0.99 3 = 1.00-1.49 4 = 1.50-1.99 5 = 2+
abs_mono_7_cat	Absolute monocyte count day 7 category: 1 = 0-0.24 2 = 0.25-0.49 3 = 0.50-0.74 4 = 0.75-0.99 5 = 1.0+
creat_7_cat	Creatinine day 7 category 1 = 0-0.79 2 = 0.80-1.19 3 = 1.20-1.79 4 = 1.80-2.99 5 = 3+
crp_7_cat	c-reactive protein day 7 category: 1 = 0-19.9 2 = 20-59.0 3 = 60-99.9 4 = 100-179 5 = 180+
ddimer_7_cat	D-dimer day 3 category: 1 = 0-499 2 = 500-999 3 = 1000-1999 4 = 2000-3999 5 = 4000+
ldh_7_cat	Lactate dehydrogenase day 7 category: 1 = 0-200 2 = 200-299 3 = 300-399 4 = 400-499 5 = 500+

12. Anexo II. Referencias.

[i] https://www.cell.com/cell-reports-medicine/pdfExtended/S2666-3791(21)00115-4

[ii] https://www.frontiersin.org/articles/10.3389/fmed.2022.1036556/full#:~:text=Among%20non%2Dhospitalized%20patients%20with,and%20the%20severity%20of%20symptoms

[iii] https://www.cdc.gov/coronavirus/2019-ncov/hcp/clinical-care/underlyingconditions.html.