Articulo

Diseño de un Modelo de Aprendizaje Automático para la Predicción de Casos de Infección por SARS-CoV-2 a Partir de Parámetros Clínicos de Laboratorio

Diana Carolina Prada Robles[1,*]

[1]Facultad de Ingeniería y Ciencias Básicas, Fundación Universitaria Los Libertadores
[*]Autor de correspondencia: dcpradar@libertadores.edu.co

Resumen
La tuberculosis (TB) es la décimo tercera causa de muerte en el mundo, en Colombia el país ha adoptado la "Estrategia Mundial denominada Fin a la TB 2016-2035" a fin de mitigar el contagio y muertes por la enfermedad. (abecé-tuberculosis. Minsalud). En el análisis de la TB se han definido variables tales como la edad, sexo, las condiciones de salubridad y residencia del paciente. Variables que están asociadas al desarrollo más temprano de esta enfermedad, no obstante, no se ha confirmado que necesariamente sean estas quienes determinen una condición mortal en el paciente. Es por ello que a través de un modelo machine learning se determinan las características más importantes que están relacionadas con la evolución de la enfermedad TB y caracterizar los perfiles de pacientes con TB, de acuerdo a la información de las bases de datos de la plataforma de notificación de eventos en salud pública Sivigila y así poder estimar el porcentaje de mortalidad que puede llegar a tener un paciente de TB. Realizando la implementación se pudo mejorar el modelo base del modelo basado en reglas siendo el Quadratic Discriminant Analysis el mejor por sus métricas las cuales no son muy buenas pero tienen una tendencia de superar el modelo base.

Palabras clave: Tuberculosis, variables, salud pública, modelo de aprendizaje.

Facultad de Ingeniería y Ciencias Básicas

APUNTES DE CIENCIA E INGENIERÍA

Recibido: 14 de mayo de 2022
Aceptado: 5 de agosto de 2022
Publicado: 22 de noviembre de 2023

Como citar este articulo

Prada-Robles, D.C., "Diseño de un Modelo de Aprendizaje Automático para la Predicción de Casos de Infección por SARS-CoV-2 a Partir de Parámetros Clínicos de Laboratorio", *Revista Apuntes de Ciencia e ingeniería*, 1, 1, nov, pag 34-47. 2023. doi: 10.37511/apuntesci.v1n1a4

1. Introducción

La pandemia ocasionada por la infección del virus SARS-CoV-2, diariamente genera nuevos retos al actual sistema de salud, lo que ha mantenido en alerta sanitaria a todos los países del mundo por la evidente

capacidad del microorganismo para modificarse y generar nuevas cepas que resultaron ser altamente contagiosas (Cabezas, 2020).

Aunque ya se han implementado planes para inmunizar a la población colombiana desde febrero de 2017(Ministerio, 2021), la enfermedad sigue cobrando vidas, lo que sugiere realizar avances o planes de mejora en la identificación de la infección, con el fin de captar y tratar oportunamente los nuevos casos evitando desarrollos de estadíos graves de la enfermedad o decesos (Gao et al., 2020).

Actualmente se cuenta con pruebas diagnósticas altamente específicas y sensibles que pueden realizar la identificación de la infección por SARS-CoV-2, como la Reacción en Cadena de la Polimerasa en Tiempo Real (RT-PCR), que es una prueba molecular que amplifica una porción específica del virus logrando así su reconocimiento, en muestras de hisopado nasal y saliva. Así mismo se cuenta con pruebas como cuantificación de anticuerpos IgM e IgG específicos para COVID que se realizan en suero, como pruebas cualitativas de cromatografía o pruebas rápidas de cassette, que se realizan también con hisopados nasofaríngeos (Mazariegos-Herrera et al., 2020) (Yamayoshi et al., 2020).

No obstante, el costo del uso de estas tecnologías es muy elevado y en algunos casos los insumos se ven afectados por el alto consumo que la pandemia actualmente lo requiere, por tanto, realizar un diagnóstico a tiempo puede significar la diferencia entre identificar pacientes positivos con alto riesgo a desarrollar la enfermedad de forma grave y aquellos positivos que sólo requieren cuidados básicos (Kukar et al., 2021).

Estudios posteriores han identificado parámetros entre los exámenes de laboratorio clínico que pueden estar presentes de manera significativa en el contagio por SARS-CoV-2, permitiendo ampliar la vigilancia de estos pacientes sospechosos, al tiempo que se va procesando la prueba diagnóstica, manteniendo en cercana observación a dichos pacientes de manera que se puedan controlar etapas graves de la enfermedad (Sánchez et al., 2021) (Wang et al., 2020).

La presencia de esta infección puede confirmarse mediante la prueba gold estándar, mientras que una ronda de parámetros clínicos de detección inicial, puede proporcionar una indicación probabilística de la presencia de la enfermedad. Es muy difícil para un médico extraer información completa de diferentes tipos de análisis de sangre. Sin embargo, los modelos de aprendizaje automático pueden diferenciar varios patrones obtenidos a partir de parámetros sanguíneos (Chadaga et al., 2022). Por lo tanto, muchos investigadores como Schwab et al. (2020) y entusiastas de ML han explorado en su revisión sistemática, el desarrollo de modelos de ML que pueden diagnosticar COVID-19.

El objetivo principal de este estudio se centra en diseñar un modelo machine learning que permita pronosticar casos por infección de SARS-CoV-2 a partir de parámetros clínicos de laboratorio.

2. Antecedentes

La enfermedad respiratoria COVID-19, es causada por el virus SARS-CoV-2, identificado inicialmente en la provincia de Wuhan (China), en noviembre de 2019 (Xu et al., 2020). SARS-CoV-2 pertenece al linaje Betacoronavirus, subgénero Sarbecovirus. Los síntomas de este virus incluyen fiebre, tos, dificultad para respirar, leucopenia y neumonía en ambos pulmones. Tiene una mayor afectación en aquellos pacientes de edad avanzada (>50 años) y que presentan comorbilidades significativas. Originalmente, se pensó que estaba asociado principalmente con los adultos mayores, pero ahora el virus está afectando ampliamente a personas más jóvenes: incluso niños. Los pacientes con el desarrollo de la enfermedad en estado grave requieren de cuidados intensivos y tienen una alta tasa de muerte (Zhang et al., 2020).

La OMS (2020) estableció que la mayoría de las personas infectadas por el virus experimentaron una enfermedad respiratoria de leve a moderada y se recuperarán sin requerir un tratamiento especial. Sin embargo, algunas enfermaron gravemente y requirieron atención médica. Las personas mayores y aquellos que padecen enfermedades subyacentes, como enfermedades cardiovasculares, diabetes, enfermedades respiratorias crónicas o cáncer, tienen más probabilidades de desarrollar la enfermedad de forma grave.

Cualquier persona, de cualquier edad, puede contraer COVID-19, enfermar gravemente e incluso fallecer.

A nivel mundial, se han reportado alrededor de 465,9 millones de casos de coronavirus, con un total aproximado de muertes de 6 millones de personas, encabezando el continente americano con 2,7 millones de decesos reportados en marzo 2022 (Statista, 2022). En Colombia el Ministerio de Salud y Protección Social, en marzo del 2020, declara a nivel nacional la Emergencia Sanitaria frente a COVID-19. Implementando las principales medidas de aislamiento y contingencia para mitigar la propagación de la infección del nuevo coronavirus (Ministerio, 2020).

Así mismo el Ministerio de Salud y Protección Social (2020), fue claro al mencionar que la crisis desencadenada por el SARS-CoV-2 ha desnudado la precariedad de algunos de muchos sistemas de salud del mundo incluso en países económicamente poderosos, haciendo énfasis que la infraestructura de muchas instituciones de salud no estaban preparadas para el abordaje de la cantidad de pacientes contagiados.

La prueba Gold Standard, que permite identificar el virus SARS-CoV-2, es la Reacción en Cadena de la Polimerasa con Transcripción Inversa en tiempo real (RT-PCR), es una técnica de biología molecular utilizada para estudiar la expresión génica a nivel de transcripción. Implica los siguientes pasos: aislamiento de ARN de muestras y síntesis de ADNc utilizando un kit de transcripción inversa. (Jalandra et al., 2020).

Chu et al. (2020), diseñaron dos ensayos de PCR de transcriptasa inversa cuantitativa en tiempo real de 1 paso que detectan dos regiones diferentes: ORF1b (marco abierto de lectura) y N (proteína del núcleo), ya que estas son las secuencias altamente conservadas en los coronavirus. Se analizaron muestras de esputo y frotis de garganta de dos pacientes junto con controles positivos y negativos, y los resultados son altamente específicos y sensibles. El gen N funciona como una herramienta de detección, y ORF1b se utiliza como prueba de confirmación. El 14 de enero de 2020, se publicó en el sitio web de la OMS el protocolo de RT-PCR para la detección de SARS-CoV-2 para su uso en la identificación de COVID-19 (Headquarters, 2020).

En Colombia la prueba comenzó a implementarse en marzo de 2020 a través del Instituto Nacional de Salud, y 22 laboratorios satélite a nivel nacional quienes validaron y obtuvieron el aval para comenzar el testeo en las regiones. Santander fue uno de los primeros departamentos en apoyar con el procesamiento de pruebas diagnósticas para COVID-19 (INS, 2020).

También se han desarrollado pruebas de detección rápida de antígenos, para captar la infección activa, aunque por aumento de casos por las variantes emergentes, algunas veces se dispone de un número limitado de dichas pruebas. Sin embargo, en comparación con la RT-PCR, las pruebas de detección rápida de antígenos carecen de sensibilidad y debido al mayor riesgo de resultados falsos negativos, por lo tanto, se han considerado un complemento de las pruebas de biología molecular. Las pruebas de anticuerpos pueden tener un papel complementario a las pruebas de RT-PCR en el diagnóstico de COVID-19, aproximadamente 10 días o más después del inicio de los síntomas, en la evaluación de infecciones pasadas y/o en las infecciones presentes (Vandenberg et al, 2021).

No obstante, en periodos de nuevas cepas emergentes, los picos de contagio se elevan, consumiendo los insumos para el diagnóstico de la infección, ocasionando retrasos en la atención de los usuarios, al tratar de clasificar los pacientes altamente sugerentes a contagio por SARS-CoV-2, de aquellos pacientes con otro tipo de anomalías respiratorias.

Los reportes de laboratorio pueden ser pieza clave al momento de pronosticar casos activos de COVID-19, ya varios estudios se han dedicado a agrupar los analitos más representativos durante el curso de la enfermedad. En el estudio de Ferrari et al. (2020), observaron diferencias estadísticamente significativas para conteo de glóbulos blancos (WBC), proteína c reactiva (CRP), aspartato aminotransferasa (AST), alanina aminotransferasa (ALT) y lactato deshidrogenasa (LDH). El punto de corte empírico para AST y LDH permitió la identificación del 70 % de los pacientes con COVID-19 positivo o negativo sobre la base de los resultados de los análisis de sangre de rutina.

López & Mazzuco (2020) determinaron en su estudio que las alteraciones de laboratorio informadas al

Facultad de Ingeniería y Ciencias Básicas
Imagina, diseña y construye: Revista Apuntes de Ciencia e Ingeniería

momento del diagnóstico y durante la hospitalización son variadas y que van a depender del desenlace inmunológico de cada hospedador frente a la infección y de cómo sea la presentación del cuadro clínico. Marcadores como la ferritina (síndrome de activación macrofágica), interleucina 6 (IL-6) y proteína C reactiva (respuesta inflamatoria), dímero D (coagulopatía), LDH (daño de órgano), troponinas (infarto agudo de miocardio), el recuento linfocitario (respuesta inmune) y ALT/ AST (daño hepático) son claves y deben ser medidos tanto al ingreso como en el seguimiento de los pacientes contagiados con SARS-CoV-2.

El aprendizaje automático puede ser una herramienta que permita facilitar la clasificación de pacientes con sospecha alta COVID-19 positivos de aquellos negativos, a través de algunos parámetros clave de laboratorio. Parsons et al.(2021) identificaron en su modelo que la combinación de los resultados del recuento de glóbulos blancos, los linfocitos y la ferritina en una prueba de sangre combinada COVID (CCBT) tuvo un área bajo la curva de 0,79. Al analizar esto en comparación con una revisión retrospectiva adicional de 181 pacientes sospechosos de COVID-19, utilizando el mismo umbral CCBT, se obtuvo una sensibilidad de 0,73 y una especificidad de 0,75. Donde la sensibilidad de su estudio fue comparable a la RT PCR del SARS-CoV-2.

Los hallazgos de Kukar et al. (2021) establecen la selección del punto ROC operativo con una sensibilidad del 81,9 % y una especificidad del 97,9 %. El área bajo la curva (AUC) con validación cruzada fue 0,97. Donde los cinco parámetros sanguíneos de rutina más útiles para el diagnóstico de COVID-19 según la puntuación de importancia de características del algoritmo XGBoost fueron: Concentración de hemoglobina por glóbulo rojo (MCHC), recuento de eosinófilos, albúmina, International Normalized Ratio (INR) y porcentaje de actividad de protrombina (PT).

3. Referentes teóricos

Conservando los objetivos propuestos del presente proyecto, se han presentado diversos documentos que describen el uso y aplicación de machine learning y los temas del riesgo de crédito. Una publicación muy amplia en su descripción ha sido “Propuesta de Modelo para evaluación de Riesgo de Crédito utilizando algoritmos de predicción para la Cooperativa de Ahorro y Crédito LA” (Cuenca et, 2019). Manifiesta que la regresión logística se desempeña como mejor modelo dado su rendimiento y sus métricas para la predicción del riesgo, su modelo presentó un accuracy de 0.6634, sensibilidad de 0.6448, especificidad 0.6821 y precisión de 0.6698. Con herramientas como la curva de características operativas receptoras (ROC), su área bajo la curva (AUC) y favorece el descarte de modelaciones econométricas, que han sido usados convencionalmente. Asimismo, hace reconocimiento a las posibles dificultades que puede presentar la modelación, entre ellas se encuentran la calidad de los datos, el filtro de las variables y aplicación de algoritmos.

La regresión logística cuenta con ciertas limitaciones a la hora de aplicar el modelo, afectaciones subyacentes como la apropiada vinculación de las variables, que podrían afectarse por multicolinealidad. Para ellos surgen nuevas herramientas como el random forest, cuyas métricas lo catapultan como superior por sus estimaciones. (Kruppa et al., 2013). Una revisión sistemática de la literatura, lleva a definir que no existe consenso alguno en cuanto al rendimiento de los modelos estadísticos y machine learning para determinar la calificación crediticia. (Dastile et al., 2020). Esto, porque pueden presentar dificultades en la explicación de las predicciones y datos desequilibrados.

4. Metodología

Se analizaron los datos de RT-PCR del servicio de Urgencia, tomados del Sistema de Información del Laboratorio Clínico Higuera Escalante (SILHE) ubicados en el Nororiente Colombiano, recopilados durante los meses enero y octubre del año 2021, y los parámetros de laboratorio clínico que se obtuvieron de rutina en la misma fecha, como valoración inicial de la sintomatología. Al revisar las frecuencias de los parámetros del laboratorio, se eliminaron exámenes duplicados para cuantificar el tamaño real de la muestra de los datos.

Así mismo se revisó el récord de análisis de laboratorio clínico realizados durante el mismo rango de tiempo, para la misma área hospitalaria (urgencia), esto con el fin de realizar los estadísticos descriptivos iniciales y caracterizar los parámetros de laboratorio que llegaron a ser parte del modelo de aprendizaje automático.

5. Análisis, validación y tratamiento de datos

Para el tiempo estimado del análisis comprendido entre enero y octubre de 2021, se obtuvieron 3048 observaciones o registros de pruebas PCR-RT para COVID-19, por parte de la institución se obtuvo las bases de datos de los parámetros clínicos más relevantes para ese periodo de tiempo y se realizaron las respectivas asociaciones con la base de datos de registros pruebas COVID, se encontraron 44 variables identificadas. Se evidenció que existen observaciones donde no se les realizaron paraclínicos o pruebas anexas, por tanto, esos registros no se tuvieron en cuenta, obteniendo una base de 2126 registros.

Con una nueva revisión de la data, 20 de las variables venían con un porcentaje mayor a 30 % de datos nulos, estas no se tuvieron en cuenta para el análisis debido a que los valores biológicos o datos de biomarcadores no se pueden imputar, ya que se le asignaría un valor no existente a un parámetro clínico, afectando la sensibilidad del modelo; de igual manera, se procede con no contar con las observaciones cuyos datos se encontraron incompletos para las variables potenciales, de esta nueva revisión se obtiene un registro final de 1164 registros de pacientes con toda la información de biomarcadores clínicos completos. Al realizar estadísticos inferenciales, se detectaron 4 variables que no eran estadísticamente significativas para el análisis, por tanto, tampoco se tuvieron en cuenta (ver Figura 1).

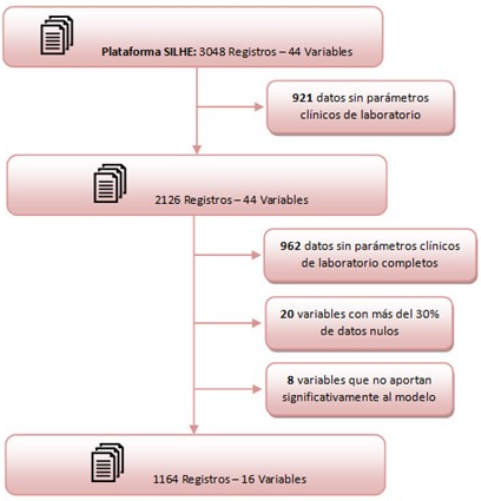

Figura 1: Definición y transformación de variables.

5.1. Herramientas y librerías implementadas

Los análisis de bases de datos y estadísticos se realizaron utilizando el software de Google Collaboratory así como para el código PYTHON.

Para los estadísticos descriptivos, se estimaron las frecuencias de las variables y la caracterización sociodemográfica de los datos a implementar. Para los estadísticos inferenciales, se realizó mediante la prueba U de Mann-Whitney para las variables continuas, el umbral de significación se fijó en 0,05, teniendo en cuenta que es el valor globalmente adoptado para probar la significancia estadística. De esta forma los datos de este documento pueden ser corroborados por la literatura existente (Altini et al., 2021).

Los análisis exploratorios de los datos (EDA) se elaboraron con la librería sweetviz V.2.1.3 para realizar un análisis que permitió relacionar las variables seleccionadas con la variable de salida (reporte RT- PCR).

Otra de las librerías que se emplearon fue PyCaret, utilizada para llevar a cabo todas las transformaciones de los datos de forma ágil y eficiente. También esta librería cuenta con una función en la que ejecuta 18 modelos de regresión distintos, esto permite explorar las distintas opciones y de acuerdo a las métricas seleccionar los modelos que mejor desempeño tienen para el conjunto de datos bajo estudio.

Por último, se utilizó la librería Shap que permite explicar la predicción de una variable objetivo, calculando la contribución de cada característica a la predicción.

Variable	**Descripción**	**Unidades**	**Transformación**	**Puntos de corte asociados a posible contagio**
Resultado	Efecto que resulta del proceso de la RT-PCR. Variable objetivo.	Variable con dos únicas opciones: 1. Negativo 2. Positivo	Recategorización en 0 = Negativo 1 = Positivo	Un resultado POSITIVO es determinante a contagio por SARS-Cov-2
Creatinina (CREA)	Producto de desecho generado por los músculos como parte de la actividad diaria.	mg/dL	Se toma el logaritmo en base 10 (CREA_log10)	>1.2 mg/dL asociado a una lesión o enfermedad renal.
Lactato Deshidrogenasa (LDH)	Enzima catalizadora que se encuentra en muchos tejidos del cuerpo.	U/L	Se toma el logaritmo en base 10 (LDH_log10)	>150 U/L asociado a algún daño de tipo tisular.
Nitrógeno Uréico en Sangre (BUN)	Sustancia formada por la descomposición de proteínas en el hígado	mg/dL	Se toma el logaritmo en base 10 (BUN_log10)	>20 mg/dL asociado a una lesión o enfermedad renal.
Proteína C Reactiva (PCR)	Marcador de inflamación	mg/dL	Se toma el logaritmo en base 10 (PCR_log10)	>0.3 asociado a infección viral severa o sepsis viral
Conteo de Glóbulos Rojos (RBC)	Recuento de glóbulos rojos por milímetro cúbico de sangre.	mm3	No Aplica	>5.8M/mm3 asociado a hemoconcentración viral
Hemoglobina (HB)	Hemoproteína con función de transportar O2 desde los pulmones a los tejidos.	g/dL	No Aplica	>15 g/dL asociado a hemoconcentración viral
Hematocrito (HTO)	Volumen de glóbulos rojo con relación al total de la sangre.	Porcentaje (%)	No Aplica	>45 % asociado a hemoconcentración viral
Volumen Corpuscular Medio (VCM)	Se refiere a la media del volumen individual de los eritrocitos	Femtolitro (fL)	No Aplica	>93 fL asociado a hemoconcentración viral
Conteo de Glóbulos Blancos (WBC)	Recuento de glóbulos blancos por milímetro cúbico de sangre.	mm3	No Aplica	>10000 asociado a respuesta inmune viral
Neutrófilos (NEU)	Primeras célula inmunitarias que reaccionan cuando entran al cuerpo microorganismos, comobacterias o virus.	Porcentaje (%)	No Aplica	>70 % asociado a respuesta inmune viral
Linfocitos (LIMP)	Segundas células inmunitarias que reaccionan cuando entran al cuerpo microorganismos, como bacterias o virus	Porcentaje (%)	No Aplica	<20 % asociado a la disminución de la respuesta inmune viral
Monocitos (MONO)	Son un tipo de glóbulo blanco. Ayudan a combatir bacterias, virus y otras infecciones en tu cuerpo.	Porcentaje (%)	No Aplica	>8 % asociado a respuesta inmune viral
Eosinófilos (EOS)	Tipo de célula inmunitaria que tiene gránulos (partículas pequeñas) con enzimas que se liberan durante las infecciones, las reacciones alérgicas y el asma.	Porcentaje (%)	No Aplica	>4 % asociado a respuesta inmune viral
Basófilos (BASO)	Tipo de célula inmunitaria que tiene gránulos (partículas pequeñas) con enzimas que se liberan durante las reacciones alérgicas y el asma.	Porcentaje (%)	No Aplica	No Aplica
Plaquetas (PLAQ)	Células sanguíneas asociadas en gran parte a los procesos de coagulación.	mm3	No Aplica	<150000 mm3 se asocia a coagulopatía por consumo

Tabla 1: Variables asociadas.

Para el desarrollo del modelo de aprendizaje automático, se empleó la metodología de base que se aplica en proyectos de aprendizaje automático (ver Figura 2).

Como parte de la metodología de elaboración de modelos predictivos, se decide dividir la base de datos en grupo de entrenamiento y en grupo de testeo, para realizar las respectivas ejecuciones de los diferentes modelos, se utilizó la librería PyCaret para realizar las diferentes comparaciones con cada

modelo dispuesto en la librería.

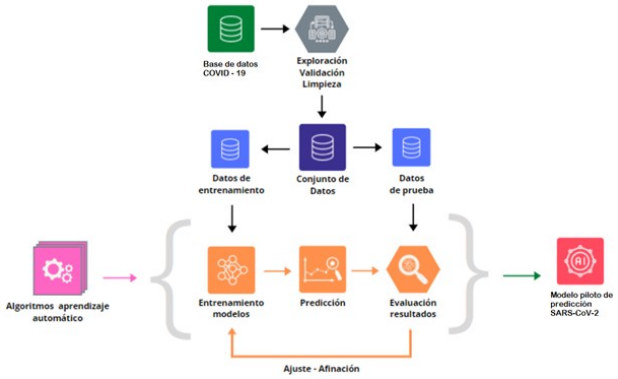

Figura 2: Evaluación del modelo.

La matriz de confusión, permite evaluar el modelo y sus métricas, mostrando de una forma explícita cuándo un grupo es confundido con otro, facilitando trabajar de forma separada con los diferentes tipos de error.

6. Resultados

6.1. Estadísticos descriptivos

Durante los meses de enero a octubre del 2021, se procesaron 370567 prestaciones o biomarcadores, asociados al diagnóstico de COVID-19. Se destacan los parámetros clínicos como el hemograma o cuadro hemático (16.94 %), creatinina (10,85 %), potasio (9.04 %), nitrógeno ureico (8,77 %) entre otros que se presentan en la Figura 3.

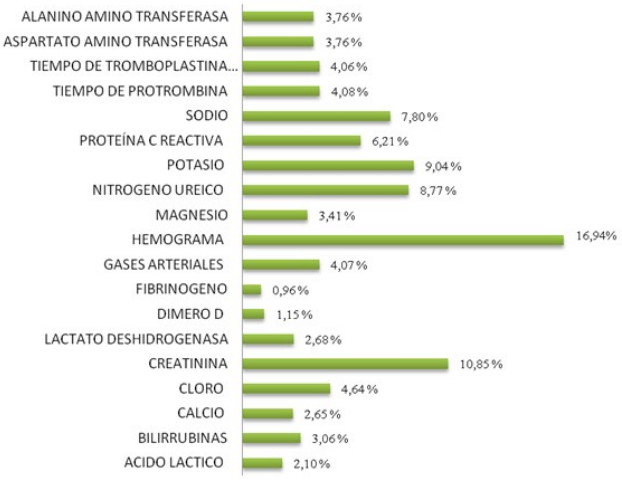

Figura 3: Biomarcadores implementados para el diagnóstico COVID-19.

Como datos sociodemográficos para las 1164 observaciones contenidas en el data set dispuesto para elaborar el modelo de aprendizaje automático, tenemos que la edad media de los pacientes se encontró en 64 años (DE 19.9), la distribución por sexo fue de 51 % para hombres y 49 % para mujeres. Respecto al comportamiento del resultado total de la prueba RT-PCR comprendida entre enero y octubre del 2021, el 65 % fue negativo y el 35 % positivo. Y la distribución de genes amplificados y detectados en pacientes positivos fue de 34,1 % para el gen del núcleo (Gen N), 0,2 % para el gen que codifica dos proteínas específicas del virus (Gen ORF1ab), y un 65,7 % para la amplificación de ambos genes (Gen N y Gen ORF1ab).

6.2. Estadísticos inferenciales

Para establecer asociaciones estadísticamente significativas entre la variable objetivo (Resultado RT-PCR) y las posibles variables explicativas, se implementó la prueba U de Mann-Whitney, obteniendo significancia estadística (p = <0.05) solo para 15 de las variables potenciales ver Figuras 4 y 5.

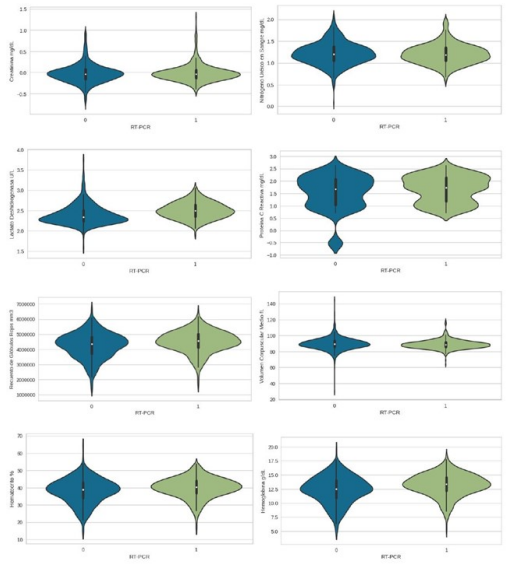

Figura 4: Gráficos de violín de la distribución de las características de laboratorio seleccionadas considerando la prueba de RT-PCR como resultado.

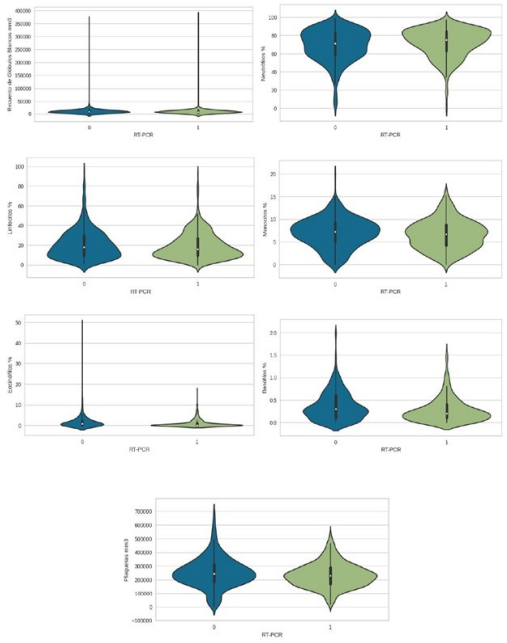

Figura 5: Gráficos de violín de la distribución de las características de laboratorio seleccionadas considerando la prueba de RT-PCR como resultado.

Las variables que se seleccionaron para elaborar el modelo de predicción fueron: creatinina, lactato deshidrogenasa, nitrógeno úrico, proteína c reactiva, recuento de glóbulos rojos, hemoglobina, hematocrito, volumen corpuscular medio, recuento de glóbulos blancos, neutrófilos, linfocitos, monocitos, eosinófilos, basófilos y recuento de plaquetas.

Teniendo en cuenta la distribución de los datos de cada variable se realizaron transformaciones logarítmicas a los datos de la creatinina, nitrógeno ureico, lactato deshidrogenasa y proteína c reactiva que se encontraban sesgadas a la derecha y podrían inferir en el resultado del modelo.

6.3. Modelos predictivos

Los resultados de cada algoritmo en ejecución se pueden observar en la Figura 6, relacionando las mejores métricas para cada modelo.

	Model	Accuracy	AUC	Recall	Prec.	F1
gbc	Gradient Boosting Classifier	0.6685	0.7328	0.6706	0.6606	0.6637
rf	Random Forest Classifier	0.6629	0.7091	0.6152	0.6741	0.6414
et	Extra Trees Classifier	0.6616	0.7208	0.6649	0.6578	0.6589
lightgbm	Light Gradient Boosting Machine	0.6575	0.7035	0.6427	0.6561	0.6484
ada	Ada Boost Classifier	0.6466	0.6813	0.6653	0.6363	0.6482
lda	Linear Discriminant Analysis	0.6398	0.6867	0.6344	0.6353	0.6338
ridge	Ridge Classifier	0.6357	0.0	0.6316	0.6305	0.6302
qda	Quadratic Discriminant Analysis	0.5976	0.6743	0.7701	0.5693	0.6511
dt	Decision Tree Classifier	0.5606	0.5606	0.5568	0.5515	0.553
lr	Logistic Regression	0.5511	0.5839	0.5899	0.5392	0.5611
nb	Naive Bayes	0.536	0.5821	0.6755	0.4766	0.5565
dummy	Dummy Classifier	0.5075	0.5	0.0	0.0	0.0
svm	SVM - Linear Kernel	0.5034	0.0	0.525	0.2994	0.3697
knn	K Neighbors Classifier	0.498	0.533	0.4902	0.4895	0.489

Figura 6: Desempeño de los modelos desarrollados.

De acuerdo a la tabla comparativa obtenida previamente, el mejor modelo con las métricas de mayor ajuste, es el Gradient Boosting Classifier o “gbc”, que son un grupo de algoritmos de aprendizaje automático que combinan muchos modelos débiles para crear un modelo predictivo sólido. Sin embargo, al potencializar este poder de predicción del modelo “gbc”, las métricas fueron superadas por el modelo Random Forest Classifier o “rf”.

El modelo Random Forest Classifier, ajusta una serie de clasificadores de árboles de decisión en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste. Los datos del área bajo la curva (AUC) para el clasificador, se pueden apreciar en ela Figura 7.

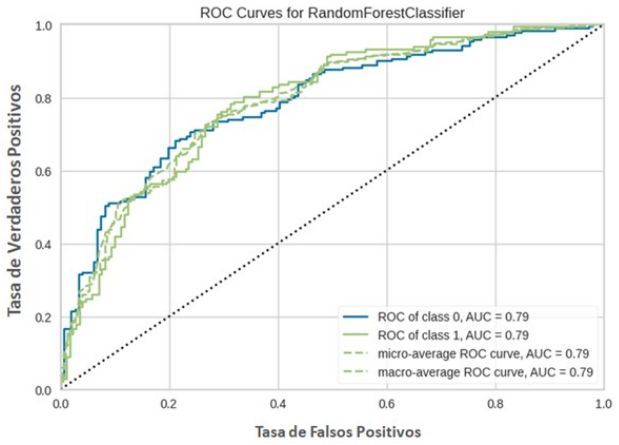

Figura 7: Curva ROC del Random Forest Classifier, para la predicción de COVID-19.

Teniendo en cuenta la curva del clasificador Random Forest, este presenta un área bajo la curva del 0.79. Esta curva se calcula variando el umbral de decisión, obteniendo tasas de verdaderos positivos y falsos positivos para cada uno de ellos. Cuanto más cerca esté el área de 1, mayor será la capacidad de discriminación del modelo en la prueba diagnóstica, en este caso el poder de diferenciar pacientes positivos o negativos para COVID-19.

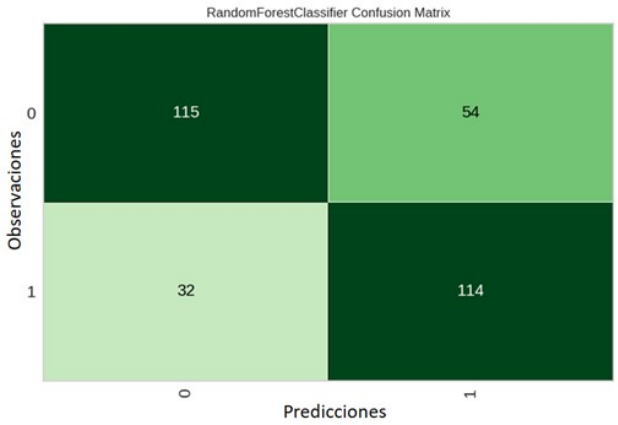

Figura 8: Matriz de confusión del Random Forest Classifier, para la predicción de COVID-19.

Para nuestra matriz tenemos que el valor 0 está indicado a un resultado Negativo de SARS-CoV-2, y el valor 1 a un resultado Positivo.

Los valores de la diagonal principal 115 y 114 corresponden con los valores estimados de forma correcta por el modelo, tanto los verdaderos positivos (TP), como los verdaderos negativos (TN). La otra diagonal, por tanto, representa los casos en los que el modelo "se ha equivocado" (32 falsos negativos (FN), 54 falsos positivos (FP)).

Las métricas de la anterior matriz indican que el modelo Random Forest tiene un Accuracy del 72 % indicando este porcentaje de predicciones correctas frente al total. La predicción de positivos de manera correcta se evidencia en una precisión de 67 %. Una sensibilidad del 78 % a casos positivos detectados y una especificidad de casos negativos detectados en un 68 %.

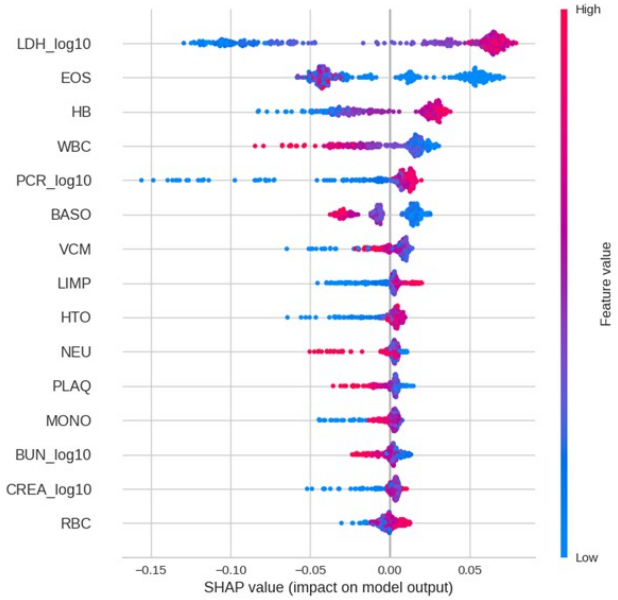

Figura 9: Gráfica SHAP para explicar los resultados del modelo.

La gráfica de SHAP-value, permite mostrar cuánto contribuye cada predictor, así como la relación positiva o negativa frente a la variable objetivo. Para el modelo Random Forest, la variable Lactato deshidrogenasa (LDH) tiene aporte significativo para el clasificador, un resultado elevado de la variable, es indicador de pronóstico positivo para SARS-CoV-2. De igual manera ocurre con los valores aumentados o concentrados de Hemoglobina (HB), y de Proteína C Reactiva (PCR). No obstante, los eosinófilos (EOS) cuando están disminuidos o ausentes, pueden ser indicador de pronóstico positivo para SARS-CoV-2.

7. Análisis de resultados

La detección inicial de la enfermedad de covid-19 es crucial para el tratamiento oportuno y para prevenir la propagación de la enfermedad. Las muestras de análisis de sangre han demostrado ser eficaces para el diagnóstico precoz de esta enfermedad. El modelo de aprendizaje automático que mejores métricas presenta frente a los predictores elegidos fue el clasificador Random Forest (AUROC 0,79, Accuracy 72 %, Precisión 67 %, Recall 78 %, Specificity 68 %), las mediciones son comparables con el estudio de Cabitza et al (2020) donde el mismo modelo contó con una Precisión 76 %, Recall 70 % y Specificity 82 %. Así mismo el mismo clasificador utilizado por Brinati et al (2020) mostró una sensibilidad muy alta (90 %) pero, en comparación, una especificidad limitada de solo el 65 %. Aunque el modelo es bajo en especificidad y en precisión, el porcentaje de detección de casos positivos es considerable por lo tanto es una herramienta más disponible entre aquellas que, siendo mucho más rápidas y económicas que las actuales pruebas diagnósticas de referencia, pueden utilizarse para el tamizaje de poblaciones enteras.

Si profundizamos sobre los biomarcadores o parámetros clínicos utilizados para la elaboración del modelo de aprendizaje, es evidente que cada institución de salud, establezca los parámetros más significativos para el diagnóstico inicial, para el estudio, aquellos analitos sanguíneos de mayor importancia fueron el hemograma, creatinina, nitrógeno ureico en sangre, lactato deshidrogenasa y proteína c reactiva. Biomarcadores similares implementaron Alves et al. (2021) en los diferentes modelos en su estudio donde el aporte de cada predictor a la variable objetivo, obtuvo comportamientos similares a los de este estudio (Shape value). Por lo tanto, un valor bajo del número de conteo de glóbulos blancos (WBC), así como del número de plaquetas (PLT), visto en azul, tiende a impactar positivamente en la salida positiva de COVID-19. Esta tendencia también se observa para la cantidad de eosinófilos EOS y la eosinopenia, caracterizada por niveles bajos de EOS, parece estar relacionada con la gravedad de la enfermedad. En el caso de la proteína c reactiva (CRP), los valores más altos de este marcador tienden a impactar positivamente en un resultado reactivo de COVID-19.

Al revisar los resultados medibles del modelo diseñado, se observa que este modelo es mucho más sensible que específico. Esta situación es interesante, cuando se tiene en cuenta que las "falsas alarmas (positivas)" no son de mayor preocupación, y lo que se quiere evitar son los falsos negativos, ya que es de mayor relevancia una mayor sensibilidad.

No es de gran impacto que el modelo clasifique un falso positivo en COVID-19, ya que, indudablemente la prueba de RT-PCR se realizará al paciente. Sin embargo, es más importante que una persona en desarrollo de la enfermedad no diagnosticada, no acceda rápidamente al tratamiento adecuado debido a un falso negativo, ya que puede desarrollar estados graves de la infección. Por tanto y evaluando las salidas que tiene el modelo, se puede inferir que es mejor tener este modelo a no tener ninguno.

8. Conclusiones

El modelo de aprendizaje automático que obtuvo mejores métricas de clasificación fue el Random Forest (RF) con un AUROC del 79 %.

Los parámetros que contribuyeron de forma significativa al clasificador RF, fueron el hemograma, creatinina, nitrógeno ureico en sangre, lactato deshidrogenasa y proteína c reactiva.

De acuerdo a los resultados de la matriz de confusión para predecir SARS-CoV-2, el clasificador RF es más sensible que específico, permitiendo clasificar en un 78 % de casos positivos detectados y un 68 % de casos negativos detectados.

Las fortalezas del estudio se tiene el acceso a la información por parte del Laboratorio clínico y Banco de Sangre Higuera Escalante S.A.S, tanto para los datos de resultados de la PCR- RT, como los resultados de los biomarcadores utilizados. Como debilidades identificadas, se percibe que incluir una variable de comorbilidades, podría mejorar la especificidad del modelo, estos datos no se pudieron obtener ya que

contaba como datos sensibles al tratar de extraerlos de la historia clínica de los pacientes, lo cual requeriría reevaluar el estudio por el comité de ética de investigaciones de la clínica.

9. Agradecimientos

Al Laboratorio Clínico y Banco de Sangre Higuera Escalante, quien nos facilitó la base de datos anonimizada no solo para SARS-CoV-2, sino también para los biomarcadores clínicos que se implementaron en este estudio.

Referencias

[1] 20 nuevos laboratorios se alistan para iniciar el diagnóstico de covid-19 en el país. https://www.ins.gov.co/Noticias/Paginas/20-nuevos-laboratorios-se-alistan-para-iniciar-diagn%C3%B3stico-de-COVID-19-en-el-pa%C3%ADs.aspx. Accessed: 2022-03-21.

[2] Colombia comenzará la vacunación contra el covid-19 el 20 de febrero. https://www.minsalud.gov.co/Paginas/Colombia-comenzara-la-vacunacion-contra-el-covid-19-el-20-de-febrero-.aspx. Accessed: 2022-03-17.

[3] Coronavirus. https://www.who.int/es/health-topics/coronavirus#tab=tab_1. Accessed: 2022-06-18.

[4] Covid-19: número acumulado de casos en el mundo 2020–2022. https://es.statista.com/estadisticas/1104227/numero-acumulado-de-casos-de-coronavirus-covid-19-en-el-mundo-enero-mar Accessed: 2022-03-18.

[5] Parámetros de laboratorio clínico en pacientes con la covid-19. http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0138-65572021000200018&lng=es&tlng=es. Accessed: 2022-03-17.

[6] Presidente duque declara emergencia sanitaria frente a covid-19. https://www.minsalud.gov.co/Paginas/Presidente-Duque-declara-Emergencia-Sanitaria-frente-a-COVID-19.aspx. Accessed: 2022-03-21.

[7] Altini, N., Brunetti, A., Mazzoleni, S., Moncelli, F., Zagaria, I., Prencipe, B., Lorusso, E., Buonamico, E., Carpagnano, G. E., Bavaro, D. F., Poliseno, M., Saracino, A., Schirinzi, A., Laterza, R., Serio, F. D., D'Introno, A., Pesce, F., and Bevilacqua, V. (2021). Predictive machine learning models and survival analysis for COVID-19 prognosis based on hematochemical parameters. *Sensors*, 21(24):8503, DOI: 10.3390/s21248503, https://doi.org/10.3390/s21248503.

[8] Alves, M. A., Castro, G. Z., Oliveira, B. A. S., Ferreira, L. A., Ramírez, J. A., Silva, R., and Guimarães, F. G. (2021). Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. *Computers in Biology and Medicine*, 132:104335, DOI: 10.1016/j.compbiomed.2021.104335, https://doi.org/10.1016/j.compbiomed.2021.104335.

[9] Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., and Cabitza, F. (2020). Detection of COVID-19 infection from routine blood exams with machine learning: A feasibility study. *Journal of Medical Systems*, 44(8), DOI: 10.1007/s10916-020-01597-4, https://doi.org/10.1007/s10916-020-01597-4.

[10] Cabezas, C. (2020). Pandemia de la COVID-19: Tormentas y retos. *Revista Peruana de Medicina Experimental y Salud Pública*, 37(4):603–4, DOI: 10.17843/rpmesp.2020.374.6866, https://doi.org/10.17843/rpmesp.2020.374.6866.

[11] Cabitza, F., Campagner, A., Ferrari, D., Resta, C. D., Ceriotti, D., Sabetta, E., Colombini, A., Vecchi, E. D., Banfi, G., Locatelli, M., and Carobene, A. (2020). Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 59(2):421–431, DOI: 10.1515/cclm-2020-1294, https://doi.org/10.1515/cclm-2020-1294.

[12] Chadaga, K., Prabhu, S., Bhat, K. V., Umakanth, S., and Sampathila, N. (2022). Medical diagnosis of COVID-19 using blood tests and machine learning. *Journal of Physics: Conference Series*, 2161(1):012017, DOI: 10.1088/1742-6596/2161/1/012017, https://doi.org/10.1088/1742-6596/2161/1/012017.

[13] Ferrari, D., Motta, A., Strollo, M., Banfi, G., and Locatelli, M. (2020). Routine blood tests as a potential diagnostic tool for COVID-19. *Clinical Chemistry and Laboratory Medicine (CCLM)*, 58(7):1095–1099, DOI: 10.1515/cclm-2020-0398, https://doi.org/10.1515/cclm-2020-0398.

[14] Gao, Y., Cai, G.-Y., Fang, W., Li, H.-Y., Wang, S.-Y., Chen, L., Yu, Y., Liu, D., Xu, S., Cui, P.-F., Zeng, S.-Q., Feng, X.-X., Yu, R.-D., Wang, Y., Yuan, Y., Jiao, X.-F., Chi, J.-H., Liu, J.-H., Li, R.-Y., Zheng, X., Song, C.-Y., Jin, N., Gong, W.-J., Liu, X.-Y., Huang, L., Tian, X., Li, L., Xing, H., Ma, D., Li, C.-R., Ye, F., and Gao, Q.-L. (2020). Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nature Communications*, 11(1), DOI: 10.1038/s41467-020-18684-2, https://doi.org/10.1038/s41467-020-18684-2.

[15] Gestoso-Pecellín, L., García-Flores, Y., González-Quintana, P., and Marrero-Arencibia, J. L. (2021). Recomendaciones y uso de los diferentes tipos de test para detección de infección por SARS-COV-2. *Enfermería Clínica*, 31:S40–S48, DOI: 10.1016/j.enfcli.2020.10.001, https://doi.org/10.1016/j.enfcli.2020.10.001.

[16] Jalandra, R., Yadav, A. K., Verma, D., Dalal, N., Sharma, M., Singh, R., Kumar, A., and Solanki, P. R. (2020). Strategies and perspectives to develop SARS-CoV-2 detection methods and diagnostics. *Biomedicine and Pharmacotherapy*, 129:110446, DOI: 10.1016/j.biopha.2020.110446, https://doi.org/10.1016/j.biopha.2020.110446.

[17] Kukar, M., Gunčar, G., Vovko, T., Podnar, S., Černelč, P., Brvar, M., Zalaznik, M., Notar, M., Moškon, S., and Notar, M. (2021). COVID-19 diagnosis by routine blood tests using machine learning. *Scientific Reports*, 11(1), DOI: 10.1038/s41598-021-90265-9, https://doi.org/10.1038/s41598-021-90265-9.

[18] Lai, C. K. C. and Lam, W. (2021). Laboratory testing for the diagnosis of COVID-19. *Biochemical and Biophysical Research Communications*, 538:226–230, DOI: 10.1016/j.bbrc.2020.10.069, https://doi.org/10.1016/j.bbrc.2020.10.069.

[19] Mazariegos-Herrera, C. J., Ozaeta-Gordillo, C. M., Menéndez-Veras, R. A., and Conde-Pereira, C. R. (2020). El papel de las pruebas diagnósticas en el manejo de la pandemia COVID-19: un enfoque desde el laboratorio clínico. *Ciencia, Tecnología y Salud*, 7(3):461–476, DOI: 10.36829/63cts.v7i3.990, https://doi.org/10.36829/63cts.v7i3.990.

[20] Parsons, I., Parsons, A., Balme, E., Hazell, G., Gifford, R., Stacey, M., Woods, D., and Russell-Jones, D. (2021). The use of routine blood tests to assist the diagnosis of COVID-19 in symptomatic hospitalized patients. *Annals of Clinical Biochemistry: International Journal of Laboratory Medicine*, page 000456322199907, DOI: 10.1177/0004563221999076, https://doi.org/10.1177/0004563221999076.

[21] Sanchez Vera, N., Saavedra Hernandez, D., Hidalgo Mesa, C. J., Aguila Lopez, M., Abreu Gutierrez, G., Herrera Gonzalez, V., and Rodriguez Garcia, I. (2021). ParÃ¡metros de laboratorio clÃnico en pacientes con la COVID-19. *Revista Cubana de Medicina Militar*, 50, ISSN: 0138-6557, http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0138-65572021000200018&nrm=iso.

[22] Schwab, P., Schütte, A. D., Dietz, B., and Bauer, S. (2020). Clinical predictive models for COVID-19: Systematic study. *Journal of Medical Internet Research*, 22(10):e21439, DOI: 10.2196/21439, https://doi.org/10.2196/21439.

[23] Vandenberg, O., Martiny, D., Rochas, O., van Belkum, A., and Kozlakidis, Z. (2020). Considerations for diagnostic COVID-19 tests. *Nature Reviews Microbiology*, 19(3):171–183, DOI: 10.1038/s41579-020-00461-z, https://doi.org/10.1038/s41579-020-00461-z.

[24] Wang, C., Deng, R., Gou, L., Fu, Z., Zhang, X., Shao, F., Wang, G., Fu, W., Xiao, J., Ding, X., Li, T., Xiao, X., and Li, C. (2020). Preliminary study to identify severe from moderate cases of COVID-19 using combined hematology parameters. *Annals of Translational Medicine*, 8(9):593–593, DOI: 10.21037/atm-20-3391, https://doi.org/10.21037/atm-20-3391.

[25] Yamayoshi, S., Sakai-Tagawa, Y., Koga, M., Akasaka, O., Nakachi, I., Koh, H., Maeda, K., Adachi, E., Saito, M., Nagai, H., Ikeuchi, K., Ogura, T., Baba, R., Fujita, K., Fukui, T., Ito, F., ichiro Hattori, S., Yamamoto, K., Nakamoto, T., Furusawa, Y., Yasuhara, A., Ujie, M., Yamada, S., Ito, M., Mitsuya, H., Omagari, N., Yotsuyanagi, H., Iwatsuki-Horimoto, K., Imai, M., and Kawaoka, Y. (2020). Comparison of rapid antigen tests for COVID-19. *Viruses*, 12(12):1420, DOI: 10.3390/v12121420, https://doi.org/10.3390/v12121420.

[26] Zhang, L.-P., Wang, M., Wang, Y., Zhu, J., and Zhang, N. (2020). Focus on the 2019 novel coronavirus (SARS-CoV-2). *Future Microbiology*, 15(10):905–918, DOI: 10.2217/fmb-2020-0063, https://doi.org/10.2217/fmb-2020-0063.