Análisis de la propagación y evolución de mutaciones de COVID-19 en Ecuador aplicando datos abiertos

Cesar Guevara, Dennys Coronel

Centro de Mecatrónica y Sistemas Interactivos – MIST, Universidad Tecnológica Indoamérica

***Resumen* — En la actualidad, el análisis y predicción de datos relacionados con el virus del COVID-19 extraídos de repositorios de información de pacientes recopilados por hospitales y organizaciones de salud ha sido de gran importancia. Ya que han contribuido significativamente al desarrollo de vacunas y a la formulación de técnicas de contingencia, proporcionando así herramientas esenciales para prevenir rebrotes y gestionar eficazmente la propagación de la enfermedad.**

En este contexto, la presente investigación se focaliza en el análisis de la información biológica del genoma del virus y los datos clínicos de pacientes afectados por COVID-19, utilizando datos de acceso público del Ecuador. Esto implica considerar variables como edad, género y ubicación geográfica para comprender la evolución de las mutaciones y su distribución en las provincias ecuatorianas. En este estudio se aplica la metodología Crisp DM, el cual es fundamental para el análisis de datos. Por otro lado, se emplearon diversas técnicas de preprocesamiento de datos y análisis estadístico, que incluyeron la correlación de Pearson, la prueba de chi-cuadrado y el análisis de varianza (Anova). Además, se utilizaron diagramas y gráficos estadísticos con el objetivo de facilitar una mejor visualización de los resultados.

Los resultados de esta investigación resaltan la diversidad genómica del virus y su relación con variables clínicas, proporcionando una comprensión profunda de la dinámica de la propagación del COVID-19 en Ecuador. Se identifican variables críticas que influyen en la vulnerabilidad de la población. Además, las conclusiones enfatizan la importancia del monitoreo de mutaciones y proponen la ampliación de la investigación a nivel global.

I. Introducción

Sin duda, el análisis de la propagación de enfermedades transmisibles se ha convertido en un área de investigación sumamente prometedora. Esto se debe a su contribución significativa en los últimos años al desarrollo de vacunas, técnicas de prevención y planes de respuesta sanitaria [1], [2]. En las últimas dos décadas, hemos sido testigos de la propagación de numerosas enfermedades transmisibles entre diferentes países. En estudios realizados recientemente como el de Prabu [3], detalla que estas enfermedades han sido de transmisión directa desde bacterias, virus y otros patógenos.

De acuerdo con la Federación Internacional de Sociedades de la Cruz Roja y de la Media Luna Roja (IFRC) y la Organización Mundial de la Salud (OMS), las enfermedades más contagiosas a nivel global son las enfermedades tropicales, la tuberculosis, la malaria, el coronavirus, el dengue, la hepatitis, el sarampión y el VIH/SIDA [4], [5]. Estas enfermedades representan desafíos significativos para la salud pública a nivel mundial debido a su capacidad de propagación y sus consecuencias en las poblaciones afectadas. Un ejemplo destacado de propagación de enfermedades transmisibles fue el virus del MERS-CoV causante de infecciones respiratorias graves, que afectó a más de 2.468 personas y provocó más de 851 muertes en 27 países desde 2012 [6].

Por otro lado, en el año 2009, apareció el virus de la gripe porcina (H1N1), que se propagó a 214 países y causó más de 18.449 muertes confirmadas por la OMS [7]. No obstante, no fue el único caso de esta naturaleza en el período mencionado. Entre los años 2003 y 2019, apareció el virus de la influenza aviar (H5N1), que infectó a 861 personas en el mundo [8].

Otro virus de gran propagación y relevancia fue la pandemia de SARS (Síndrome Respiratorio Agudo Severo), que tuvo su origen a finales de 2002 y se extendió a 29 países. El SARS provocó 8.096 casos de contagio y resultó en 774

fallecimientos.

Finalmente, el virus del COVID-19, el cual probablemente surgió en Wuhan-China en el año 2019, provocó una epidemia global que ha afectado gravemente a numerosos países [9]. Según la OMS [10], el COVID-19 afectó a 770 millones de personas y 6 millones de personas han perdido la vida debido a esta enfermedad. En el caso de Ecuador, más de un millón de personas han sido diagnosticadas con COVID-19 y más de 36 mil personas han fallecido [11]. Sin embargo, es importante destacar que se ha llevado a cabo una significativa campaña de vacunación en el país. Según datos del Ministerio de Salud del Ecuador [12], se han administrado cerca de 39 millones de dosis de vacunas a la población.

La investigación en este campo es de vital importancia, especialmente en el contexto de la enfermedad por coronavirus. La propagación de enfermedades es un área de suma importancia debido a la evolución y mutación de los virus de manera impredecible en cualquier parte del mundo [13]. Esto es aún más relevante en Ecuador, un país caracterizado por su diversidad climática única. Esta diversidad climática podría propiciar el desarrollo de diversas mutaciones.

Por esta razón, el objetivo central de la presente investigación es analizar tanto la información biológica del genoma del virus (cadena de proteínas) como la información clínica de los pacientes infectados con COVID-19 en Ecuador. Esto incluye variables como la edad, el género, la ubicación geográfica, entre otras. El propósito es identificar las variables más relevantes y comprender la evolución de las mutaciones del virus, así como su distribución geográfica en las diferentes provincias de Ecuador. Con todos estos datos, se buscará identificar los grupos de población más vulnerable en función de sus características clínicas, como la edad, el género y la ubicación geográfica, en relación con las diferentes mutaciones y variantes del COVID-19. Este enfoque permitirá una mejor comprensión de la dinámica de la enfermedad y una toma de decisiones más informada en términos de salud pública.

Este documento se estructura de la siguiente manera. En la sección 2, se describen los trabajos relacionados en el área de medicina, biología y análisis de datos sobre el COVID-19. En la sección 3, se detallan los métodos y materiales empleados para llevar a cabo la investigación. En la sección 4, se centra en las fases del preprocesamiento de datos y la aplicación de herramientas tecnológicas a la información. En la sección 5, se aborda el análisis de los datos obtenidos, que incluyen tanto la información relacionada con las proteínas y mutaciones del virus como los perfiles de los pacientes extraídos de la base de datos. Además, se enriquece la discusión al incorporar perspectivas y hallazgos de otros autores en el campo. Por último, en la sección 6, se exponen las conclusiones derivadas de la investigación y se plantean las posibles direcciones para futuros trabajos.

II. Trabajos Relacionados

En esta sección se realizó un análisis bibliográfico de artículos relacionados con mutaciones genómicas en el COVID-19. La búsqueda se limitó a artículos publicados en los últimos cinco años entre 2019 y 2023.

Uno de los trabajos de investigación relevantes en este campo es el artículo publicado por Rui Wang [14]. En este estudio se identificó y analizó las posiciones, frecuencias y proteínas codificadas de las mutaciones del SARS-CoV-2 a nivel global. El objetivo principal de esta investigación fue aislar el genoma del SARS-CoV-2 y cuantificar el número de mutaciones presentes mediante la técnica de genotipado. Los resultados de esta investigación fueron considerados satisfactorios, ya que identificó un total de 13.402 mutaciones únicas. Además, el estudio reveló que el 51.4% de las mutaciones del SARS-CoV-2 corresponden al tipo C→T.

El estudio realizado por Wang [15] identificó las mutaciones de rápido crecimiento en el dominio de unión al receptor (RBD) y analiza la tendencia evolutiva del SARS-CoV-2. El objetivo principal de esta investigación fue examinar un conjunto de datos genómicos del SARS-CoV-2 registrados en Mutation Tracker utilizando el método de aprendizaje profundo. Los resultados de este estudio son muy alentadores, ya que esta investigación identificó un total de 6.945 mutaciones únicas y 2 '194.305 mutaciones no únicas en el gen S del SARS-CoV-2 en todo el mundo. Además, los autores determinaron que la mayoría de las mutaciones en el SARS-CoV-2 corresponden a los tipos A→G, C→T y T→C. También, esta investigación destacó que aproximadamente el 70% de estas mutaciones debilitaría la eficacia de los anticuerpos conocidos.

En el estudio llevado a cabo por Thanh [16],

realiza un análisis exhaustivo de las mutaciones genómicas en las regiones codificantes del SARS-CoV-2 y explora la posible estructura secundaria de las proteínas resultantes. El objetivo central de esta investigación fue evaluar todas las mutaciones puntuales que se han registrado hasta la fecha en el SARS-CoV-2. Además, este estudio identifica los diferentes patrones de mutación mediante diferentes modelos de aprendizaje profundo. Los resultados de este estudio revelan que existen un total de 3.089 mutaciones en la proteína S del SARS-CoV-2.

Lucy Van Dorp [17] analiza las mutaciones relacionadas con la transmisión del virus SARS-CoV-2. El objetivo principal de este estudio fue cuantificar el número de descendientes que heredan un alelo específico en comparación con aquellos que no lo hicieron. Este estudio utilizó el índice filogenético para el análisis de los datos. Los resultados de este estudio revelaron un total de 12.706 mutaciones de tipo C→U. Sin embargo, este estudio concluyó que ninguna de estas mutaciones estaba asociada con un aumento significativo en la transmisión del virus.

Pachetti et al. [18] Realizaron un análisis y evaluación de la distribución de las mutaciones del SARS-CoV-2 en distintas áreas geográficas (Asia, Oceanía, Europa y América del Norte) mediante el método Clustal Omega. Este estudio se basó en datos recopilados de manera aleatoria de la base de datos GISAID. Los resultados de esta investigación arrojaron hallazgos significativos, ya que este estudio identificó un total de 14.408 mutaciones en la proteína P a L. Además, los autores demostraron que algunas de estas mutaciones podrían generar resistencia a ciertos medicamentos.

Rozhgar [19], identifica y analiza las mutaciones genómicas del SARS-CoV-2. El objetivo principal de este estudio fue determinar las mutaciones más comunes del SARS-CoV-2 mediante programas bioinformáticos. El estudio analizó 95 secuencias completas del genoma del SARS-CoV-2 disponibles en GenBank National MicrobiologyData Center (NMDC). Los resultados de este estudio determinaron que existen 116 mutaciones correspondientes al gen ORF1ab, ORF8 y al gen N.

Ahmad [20], analiza las mutaciones del genoma completo del SARS-CoV-2. El objetivo principal de este estudio fue determinar las posibles mutaciones y la evolución del COVID-19. Esta investigación utilizó un software llamado BioEdit para realizar sus alineamientos genómicos. Los resultados de esta investigación determinaron que existen 596 mutaciones en todos los genes.

Finalmente, el estudio realizado por Abdel-Rahman [21] analiza las mutaciones secuenciales presentes en el genoma del SARS-CoV-2 y determina los diversos patrones de mutación que se manifestaron en pacientes egipcios infectados. Esa investigación utilizó los métodos desclasificación de linaje de Pangolin y Nextstrain. El objetivo principal de esta investigación fue determinar la mejor clasificación de los genomas del SARS-CoV-2. Los resultados de esta investigación revelaron la existencia de un total de 1.115 mutaciones únicas. Además, se encontró que aproximadamente el 60.5% de estas mutaciones se localizan en la poliproteína ORF1ab.

III. Métodos y materiales

En esta sección, se presentan detalladamente los materiales como la base de datos utilizada y los métodos aplicados para el análisis de datos para el desarrollo de este estudio.

A. Materiales

1) Base de datos COVID-19 – GISAID Ecuador

La base de datos del COVID-19 en Ecuador fue recopilada a través de EpiFlu™, una iniciativa de GISAID [22]. Este conjunto de datos abarca un total de 8.992 registros y 13 atributos, un atributo es información genómica (secuencias proteína del virus) y 12 atributos de información médica (información relacionada con pacientes), como se presenta en la Tabla 1.

TABLA I
Variables de base de datos COVID-19 Ecuador

Tipo de Datos	NOMBRE	TIPO	PRECISIÓN	EJEMPLO
Proteínas	Cadena de proteínas	Texto	Caracteres	[--------accaaccaactctaa.....]
Paciente	Código del paciente	Texto	Caracteres	EPI_ISL_10137512
	Length	Numérico	Entero	29557
	Provincia	Texto	Caracteres	Imbabura
	Ciudad	Texto	Caracteres	Ibarra
	Latitud	Numérico	Entero con cinco decimales	0.35987
	Longitud	Numérico	Entero con cinco decimales	-78.12825
	Edad	Texto	Caracteres	Rango18

	Genero	Texto	Caracteres	Male
	Linajes	Texto	Caracteres	BA.1.1
	Código de virus	Texto	Caracteres	GRA
	Fecha	Texto	Caracteres	21/2/2022
	Variable	Texto	Caracteres	Ómicron

Una vez procesados los datos de la cadena de proteínas, se generó una nueva base de datos que incluye información sobre los 150 aminoácidos del genoma del COVID-19, así como el número de mutaciones asociadas a cada una de ellas.

B. *Métodos*

1) *Correlación de Pearson*

El método de Correlación de Pearson se utiliza para medir el grado de correlación lineal entre dos variables. El coeficiente de correlación en este método puede variar en un rango de valores que va desde más uno hasta menos uno (Ecuación 1). Si el valor total entre las dos variables se acerca a 1, indica una fuerte dependencia entre ellas. Mientras que si el valor se acerca a -1, indica una dependencia débil entre las variables [23].

$$P_{xy} = \frac{\Sigma Z_x Z_y}{N} \qquad \text{Eq. (1)}$$

En esta ecuación, X representa la variable uno, Y corresponde a la variable dos, ZX es la desviación estándar de la variable uno, ZY es la desviación estándar de la variable dos y N Es el número total de datos.

2) *Chi-cuadrado* X^2

El chi-cuadrado o $X^2 x^2$ es una prueba estadística que se utiliza para evaluar si hay una asociación significativa entre dos variables, como se lo presenta en la Ecuación 2 [24], [25].

$$X_c^2 = \sum \frac{(x_i - m_i)^2}{m_i} \qquad \text{Eq. (2)}$$

Donde cc son los grados de libertad, $x\,x$ representa los valores del conjunto de datos y mm los valores esperados.

3) *Método de ANOVA*

El análisis de varianza (ANOVA) es una técnica estadística que se utiliza para comparar las medias de tres o más grupos, con el fin de determinar si existen diferencias significativas entre ellos considerando el grado de libertad de cada parámetro [26]. La Ecuación 3 muestra el cálculo de ANOVA.

$$A = \frac{MSX}{MSY} \qquad \text{Eq. (3)}$$

Donde $MSXMSX$ representa la media de las sumas de cuadrados de la variable de XX y MSY MSY representa la media de las sumas de cuadrados de la variable de $Y\,Y$ [27].

4) *Metodología CRISP DM*

La metodología CRISP-DM es una de las más utilizadas en la actualidad para el desarrollo de proyectos de minería de datos, ya que ofrece un enfoque cíclico en la gestión de proyectos. Este enfoque permite un ciclo de vida estructurado, lo que facilita la comprensión y gestión eficiente de cada etapa del proceso de minería de datos [28]. La Figura 1 ilustra cada una de las etapas, comenzando por la comprensión del problema, seguida por la comprensión de datos, la preparación de datos, el modelado, la evaluación del modelo y por último la implementación.

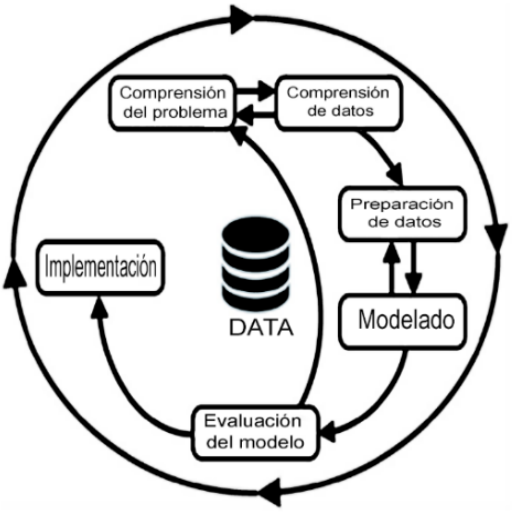

Fig. 1 Etapas de la metodología Crisp DM [29].

Esta metodología CRISP-DM consta de seis etapas fundamentales. La metodología Crisp DM, inicia con la comprensión del problema, donde se identifica la problemática, se definen los objetivos del proyecto y se evalúa el estado actual. Posteriormente, en la fase de comprensión de datos, se recopilan y exploran los datos para comprender su significado y propiedades. La etapa de preparación de datos se centra en la limpieza, transformación y creación de indicadores a partir de los datos existentes. Luego, en la etapa de modelado, se elige la técnica más apropiada, se

ajustan los parámetros del modelo y se evalúa su rendimiento. Después en la evaluación del modelo, se determina su calidad a través de métricas estadísticas y comparaciones con objetivos previamente establecidos, asegurando que cumpla con las expectativas del proyecto. Y finalmente luego de una evaluación satisfactoria del modelo entrenado, se procede a su implementación en la cual se configura una infraestructura específica para el procesamiento de datos [29].

IV. Procesamiento de datos

El preprocesamiento de datos es una fase crucial, ya que la eficiencia en el preprocesamiento influye en gran medida en la calidad de los resultados finales. El procesamiento de datos lo dividimos en cuatro fases distintas basadas en la metodología Crisp DM, donde se aplicaron una serie de técnicas y transformaciones a los datos con el objetivo de limpiar, organizar y prepararlos para su análisis posterior.

Para cada fase del procesamiento de datos hemos hecho uso del servidor del Instituto de Ciencias Matemáticas (ICMAT), llamado Clúster LOVELACE. Este servidor cuenta con 32 nodos de cómputo general, un nodo equipado con procesadores Xeon Phi, dos nodos con unidades de procesamiento gráfico (GPU Tesla) y tres nodos con una elevada capacidad de memoria RAM.

1) Fase 1- Integración y recopilación de datos

En esta fase, se procedió a descargar la base de datos de COVID-19 de Ecuador desde la plataforma GISAID [22]. Esta base de datos estaba dividida en dos partes, la primera en formato FASTA que contenía la información sobre la cadena de proteínas, la fecha y el código del paciente, entre otros detalles relevantes. La segunda parte comprendía datos médicos de los pacientes en archivos .CSV que contenían información epidemiológica como edad, género, identificación del paciente, ubicación, variante, entre otros.

2) Fase 2- Selección y limpieza de datos

Durante esta fase, se identificaron y se corrigieron los problemas de datos, con el objetivo de garantizar la precisión y confiabilidad de la información. Para este procedimiento, se desarrolló un algoritmo llamado “DataClean”, que fue creado en lenguaje de programación Python de alto nivel. Este algoritmo utiliza múltiples librerías de manejo de datos como Pandas, Numpy, Matplotib, entre otros. Este algoritmo elimina datos duplicados, valores atípicos y aquellos que contengan errores. Además, este algoritmo depura la información de cada variable eliminando saltos de línea y caracteres desconocidos. Adicional en esta fase, se eliminó variables que no son relevantes para el estudio, como son: host debido a que todos los infectados del virus son humanos, originating_lab información incoherente e incompleta e imprecisa, authors solo muestra quien recolectó la muestra, origenVariante por su información incompleta y no aporta datos a la investigación.

La tabla 2 muestra los coeficientes de correlación de Pearson entre las variables.

Tabla II
Correlación de Pearson

Variable	**Edad**	**Género**	**Variante**
Edad	1	-0.31	0.85
Género	-0.31	1	0.54
Variante	0.54	0.85	1

La tabla 3 presenta los valores de la estadística Chi-cuadrado entre las variables.

Tabla III
Prueba de Chi-Cuadrado

Variable	**Edad**	**Género**	**Variante**
Edad	-	6.84	185.08
Género	6.84	-	89.98
Variante	185.08	89.98	-

La tabla 4 expone los valores de la estadística de ANOVA, donde se representan estadísticas F, utilizada para analizar diferencias significativas entre las variables.

Tabla IV
Método ANOVA

Variable	Edad	Género	Variante
Edad	-	0.41	2.38
Género	0.41	-	1.92
Variante	2.38	1.92	-

3) Fase 3 - Transformación de datos

En esta fase, se preparó los datos para su posterior análisis, esto incluye la normalización de datos (edad y género). Por otro lado, se realizó la categorización de la variable edad, donde se presentan 20 grupos de datos comprendidos en periodos de 5 años, es decir, grupo 1 (0 – 5 años), grupo 2 (6 – 10 años), grupo 3 (11 – 15 años), grupo 4 (16 – 20 años), grupo 5 (21 – 25 años), grupo 6 (26 – 30 años), grupo 7 (31 – 35 años), grupo 8 (36 – 40 años), grupo 9 (41 – 45 años) , grupo 10 (46 – 50 años) , grupo 11 (51 – 55 años) , grupo 12 (56 – 60 años) , grupo 13 (61 – 65 años) ,

grupo 14 (66 – 70 años) , grupo 15 (71 – 75 años) , grupo 16 (76 – 80 años) , grupo 17 (81 – 85 años) , grupo 18 (86 – 90 años) , grupo 19 (91 – 95 años) , hasta grupo 20 (96 – mayor de 100 años). Durante esta fase de preparación de datos, se optó por categorizar la variable edad en grupos de 5 años con el objetivo de estructurar y organizar la información de manera más efectiva para análisis posteriores. Esta elección posibilita una representación detallada de la distribución de edades en la muestra, facilitando la identificación de patrones y tendencias específicas en diferentes segmentos de la población. Además, la categorización en intervalos pequeños proporciona una visión más refinada de cómo la edad puede influir en los resultados, permitiendo así una interpretación más precisa de los hallazgos. Por otro lado, se ha transformado la variable género en dos categorías numéricas (0 – masculino y 1 femenino), para poder aplicar las técnicas de selección de variables como Anova y Chi cuadrado, que no permiten analizar datos categóricos.

Por otro lado, se realizó la estandarización de la cadena de proteínas, donde se efectuó un alineamiento de secuencia aplicando el sistema MAFFT (Multiple Alignment Fast Fourier Transform) versión 7 mediante el método progresivo de Transformada Rápida de Fourier y el método de refinamiento iterativo (FFT-NS-2). El propósito central de esta fase fue alinear las secuencias de proteínas de cada paciente, asegurando que todas tengan una longitud uniforme (29904 nucleótidos). Este proceso se ha realizado por medio de un algoritmo desarrollado en Python. Posteriormente, el conjunto de datos de proteínas alineados con el mismo tamaño se transforma en cadena de aminoácidos, aplicando las librerías de EigenMS y LibMUSCLE.

Finalmente, con la cadena de aminoácidos resultante se identificó el número de mutaciones presentes en cada registro y comparadas con la muestra de COVID-19 del paciente en las primeras etapas de contagio. Esta tarea se desarrolló con un algoritmo en el lenguaje de Python aplicando la librería de Biopython y Pandas. El propósito central de este algoritmo fue identificar las distintas mutaciones presentes en cada uno de los aminoácidos. Para detectar estas mutaciones, se analizó los cambios en los aminoácidos en posiciones específicas de la secuencia en comparación con otras secuencias. Como resultado, se generó una base de datos que detalla el número de mutaciones que afectan a cada aminoácido.

4) *Fase 4 - Integración de datos*

En esta fase final, se realizó un algoritmo de integración de la base de datos epidemiológicos del paciente y la base de datos de aminoácidos y la cantidad de mutaciones detectadas del virus en la fase 3. En esta integración de datos se desarrolló un algoritmo en Python para fusionar los dos conjuntos de datos por medio del Id del paciente. Mediante esta fusión de datos se estableció un conjunto de datos coherente que facilitó el análisis de forma conjunta y unificada.

V. Análisis de datos de base de datos COVID-19 y Discusión

En esta sección, se presenta el análisis de datos relacionados con pacientes de COVID-19 en Ecuador, posteriormente al preprocesamiento de datos. Para llevar a cabo este análisis se utilizó Python y las librerías de Matplotlib, Geopandas y Seaborn para la visualización de la información procesada. Estas bibliotecas permitieron generar diagramas estadísticos, correlación de variables, lo que facilitó el análisis de cada una de las gráficas.

La Tabla 5 muestra la relación entre las provincias más afectadas por las diferentes variantes del COVID-19, junto con los porcentajes de infección asociados. Estos datos evidencian que las provincias de Pichincha, Guayas y en menor medida Chimborazo, han sido las provincias más afectadas en términos de mayor porcentaje de infección relacionadas con las diferentes variantes del COVID-19.

Tabla V
Porcentaje de infección de cada variante por provincia.

Variante	Provincia	Porcentaje de Infección
Variante Alfa	Chimborazo	15.19%
	Guayas	10.13%
	Pichincha	39.24%
Variante Delta	Guayas	18.64%
	Manabí	15.21%
	Pichincha	16.21%
Variante Lambda	El Oro	18.90%
	Guayas	20.06%
	Pichincha	15.41%
Variante Mu Gh	Chimborazo	13.99%
	Guayas	14.63%

	Pichincha	31.83%
Variante Gamma	El Oro	6.59%
	Guayas	20.16%
	Pichincha	29.32%
Variante Ómicron	Guayas	19.10%
	Manabí	7.66%
	Pichincha	34.89%

En la Figura 2, se presenta la distribución de las diferentes variantes de COVID-19 en las provincias de Ecuador. Las provincias que presentan mayor índice de contagio son: Pichincha, Guayas, Manabí, Chimborazo, Azuay y Cotopaxi. Además, se observa que la variante predominante en cada una de ellas es Ómicron, seguida por Delta, Mu Gh, Gamma, Lambda y finalmente la variante Alpha (Tabla 6).

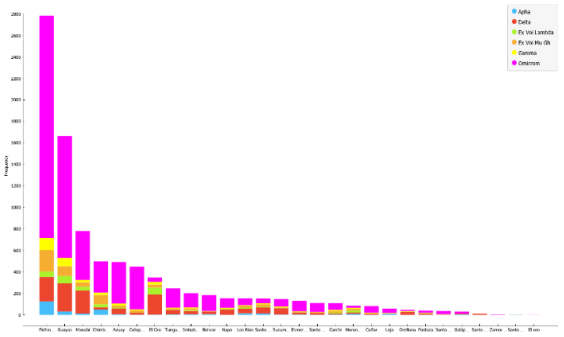

Fig. 2 Variantes de Covid-19 por provincias del Ecuador.

Tabla VI
Provincias con el mayor índice de contagio según variante de COVID-19.

Provincia	Variantes					
	Apha	Delta	Lambda	Mu Gh	Gamma	Omicron
Pichincha	4.46 %	8.16%	1.91%	7.12%	4.03%	74.34%
Guayas	1.93 %	15.70 %	4.15%	5.48%	4.63%	68.11%
Manabí	1.67 %	27.34 %	4.62%	5.26%	2.82%	58.28%
Chimborazo	9.66 %	4.63%	5.63%	17.51 %	4.23%	58.35%
Azuay	1.22 %	10.79 %	1.83%	3.46%	4.28%	78.41%
Cotopaxi	0.45 %	3.58%	0.67%	4.70%	2.24%	88.37%

Por otro lado, la Figura 3 muestra un diagrama de la cronología del contagio del COVID-19 por variantes, desde enero de 2021 hasta octubre de 2023. En estos datos se puede apreciar que la variante Alfa fue predominante hasta julio de 2021 y la variante Delta desde julio de 2021 hasta enero de 2021. Pero la variante que se ha mantenido como predominante en el Ecuador ha sido Omicron, la cual inició su propagación desde diciembre de 2021 hasta octubre de 2023.

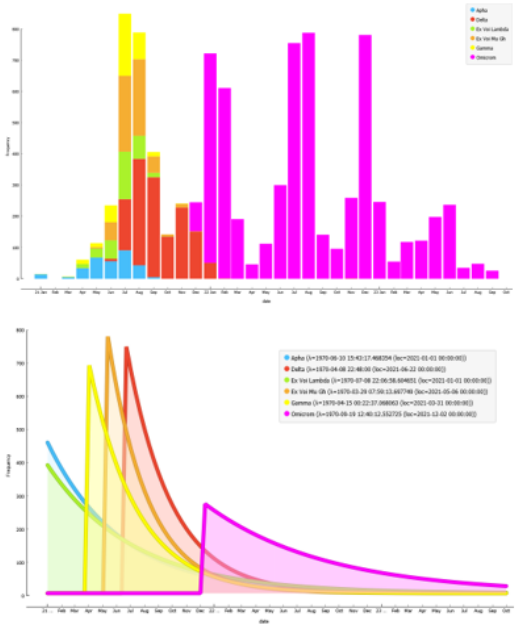

Fig. 3 Cronología de propagación de variantes de COVID-19.

La Figura 4 muestra la distribución cuantitativa de los datos entre variantes de COVID-19 y género de paciente, donde se observa que los pacientes de género femenino han sido más afectados por la variante Ómicron, representando un 69.17% de los casos, seguido por la variante Delta, que afectó al 14.88%. Por otro lado, los pacientes de género masculino han experimentado una afectación mayor por la variante Ómicron, que representa el 61.72% de los casos, seguida por la variante Delta, que afectó al 16.46%. Estos datos resaltan diferencias significativas en la afectación de las variantes del COVID-19 entre géneros.

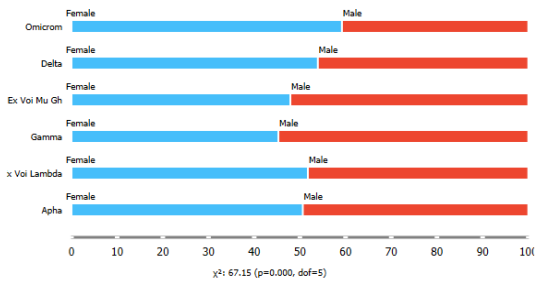

Fig. 4 Variantes de COVID-19 por género de pacientes.

La Figura 5, proporciona una distribución cuantitativa de los datos relacionados con las personas infectadas por COVID-19, desglosadas por género y provincia, donde se muestra que los pacientes de género femenino han sido más afectados por el COVID-19 en las provincias de Pichincha con un 30.97%, Guayas con un 19.36%, y Manabí con un 8.75%. Asimismo, los pacientes de género masculino han experimentado una alta incidencia en las provincias de Pichincha con un 30.91%, Guayas con un 17.36%, y Manabí con un 8.55%. Estos datos revelan que las provincias de Pichincha, Guayas y Manabí han sido las más afectadas por la pandemia de COVID-19, tanto en pacientes de género femenino como de género masculino.

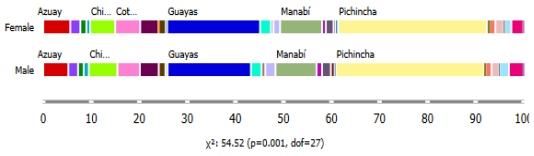

Fig. 5 Personas infectadas por género y provincia.

La Figura 6 presenta la distribución de las edades de los pacientes por cada variante de COVID-19 en Ecuador, donde se evidencia que los pacientes comprendidos entre 31 y 35 años de edad son el grupo de mayor contagio, seguidos por pacientes de edades entre 26 y 30 años y en tercer lugar, se localizan los pacientes con edades entre 36 y 40 años.

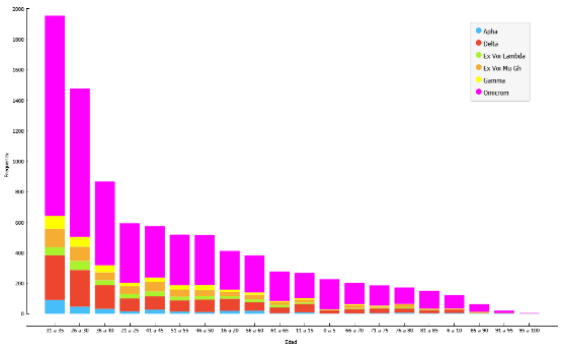

Fig. 6 Pacientes infectados por edad y variante de COVID-19.

La Figura 7 presenta la mayor distribución de pacientes infectados por edad y variante de COVID-19. Por otro lado, la Tabla 7 detalla la distribución de edades que se han contagiado de cada variante de COVID-19, donde la variante Ómicron en las edades de 31 a 35 años tiene un porcentaje de 22.17%, la variante Delta afectó un 20.93%, la variante Mu Gh afectó un 19.45% y la variante Gamma 21.73%. Por otro lado, la variante Lambda afectó a más pacientes de 26 a 30 años con un 17.15% y la variante Alpha afectó un 22.17 %.

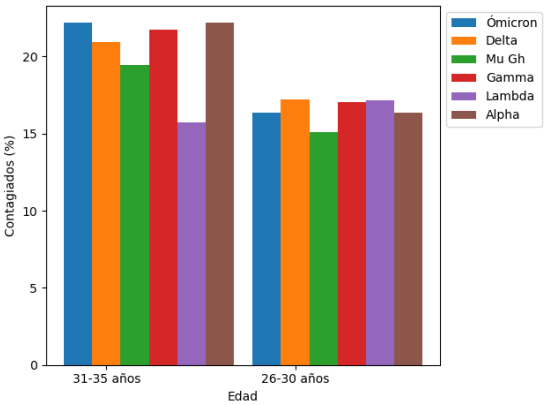

Fig. 7 Mayor porcentaje pacientes infectados por edad y variante.

Tabla VII
Porcentaje de pacientes infectados por edad y variante.

Variante	26-30 años (%)	31-35 años (%)
Ómicron	16.38	22.17
Delta	17.21	20.93
Mu Gh	15.11	19.45
Gamma	17.02	21.73
Lambda	17.15	15.7
Alpha	16.38	22.17

El análisis de información biológica del virus se ha centrado específicamente en el Spike, la cual en anteriores estudios como el de Wang [15] y Thanh [16], han tenido la mayor cantidad de mutaciones. La Tabla 8 muestra que la mayoría de aminoácidos con dos mutaciones son de la cadena de proteína Spike (S). Las mutaciones en esta cadena de proteínas pueden tener un impacto en la transmisibilidad, la capacidad de evadir el sistema inmunológico y la eficacia de las vacunas. El 98% de aminoácidos que generaron dos mutaciones están relacionadas con la proteína Spike(S) y el 2 % no tiene cadena proteínica definida. La columna "Mutaciones" describe el tipo de cambio genético que ocurrió en una ubicación específica de la secuencia genómica. La columna "Proteína" identifica la proteína relacionada con dicha mutación. La columna "Posición" indica la ubicación precisa dentro de la secuencia genómica donde se registró la mutación. En la columna

"Secuencia Original" se muestra la secuencia genética inicial en la ubicación afectada antes de la mutación, mientras que en la columna "Secuencia Mutada" se detalla la secuencia resultante después de la alteración genética. En el caso de las secuencias donde no se ha asignado una proteína específica, no se dispone de información sobre la secuencia original ni sobre la secuencia a la que se mutará.

Tabla VIII
Aminoácidos afectados con dos mutaciones

Mutación	Proteína	Posición	Secuencia Original	Secuencia Mutada
DEL144/144	No Asignada	144	-	-
DEL157/158		157/158	-	-
DEL241/243		241/243	-	-
DEL25/27		25/27	-	-
DEL3675/3677		3675/3677	-	-
DEL69/70		69/70	-	-
E1264D		1264	Glu (E)	Asp (D)
G662S	Spike (S)	662	Gly (G)	Ser (S)
I1566V		1566	Ile (I)	Val (V)
I2230T		2230	Ile (I)	Thr (T)
K3353R		3353	Lys (K)	Arg (R)
P1000L		1000	Pro (P)	Leu (L)
P2046L		2046	Pro (P)	Leu (L)
P2287S		2287	Pro (P)	Ser (S)
P3395H		3395	Pro (P)	His (H)
S1188L		1188	Ser (S)	Leu (L)
T1001I		1001	Thr (T)	Ile (I)
T265I		265	Thr (T)	Ile (I)
T3255I		3255	Thr (T)	Ile (I)
T3646A		3646	Thr (T)	Ala (A)
V2930L		2930	Val (V)	Leu (L)
A1306S		1306	Ala (A)	Ser (S)
A1708D		1708	Ala (A)	Asp (D)
A1918V		1918	Ala (A)	Val (V)

La Tabla 9 muestra los aminoácidos más afectados con una mutación, proporcionando información clave sobre a qué cadena de proteína pertenece, en qué posición está ubicada, la secuencia de aminoácidos original y la secuencia de aminoácidos resultante tras la mutación. Estos resultados muestran que la mayor cantidad de aminoácidos con una mutación se centran en el Spike (S) con el 68.29%, 14.63% corresponden a la proteína Nucleocápside (N), el 14.63% están vinculados a la proteína no estructural (nsp2) y un 2.45% en la proteína no estructural (nsp1).

Tabla IX
Aminoácidos afectados con una mutación.

Mutación	Proteína	Posición	Secuencia Original	Secuencia Mutada
D614G	Spike (S)	614	Asp (D)	Gly (G)
S84L		84	Ser (S)	Leu (L)
P681H		681	Pro (P)	His (H)
H655Y		655	His (H)	Tyr (Y)
N679K		679	Asn (N)	Lys (K)
H69X		69	His (H)	Insertado X
D796Y		796	Asp (D)	Tyr (Y)
V70X		70	Val (V)	Insertado X
T478K		478	Thr (T)	Lys (K)
A63T		63	Ala (A)	Thr (T)
N501Y		501	Asn (N)	Tyr (Y)
S375F		375	Ser (S)	Phe (F)
S373P		373	Ser (S)	Pro (P)
G339D		339	Gly (G)	Asp (D)
T223I		223	Thr (T)	Ile (I)
S413R		413	Ser (S)	Arg (R)
Y505H		505	Tyr (Y)	His (H)
Q498R		498	Gln (Q)	Arg (R)
E484A		484	Glu (E)	Ala (A)
S477N		477	Ser (S)	Asn (N)
T376A		376	Thr (T)	Ala (A)
K417N		417	Lys (K)	Asn (N)
S371F		371	Ser (S)	Phe (F)
D405N		405	Asp (D)	Asn (N)
DEL31/33		31/33	Deleción	Deleción
L452R		452	Leu (L)	Arg (R)
R408S		408	Arg (R)	Ser (S)
N440K		440	Asn (N)	Lys (K)
R203K	Nucleocápside (N)	203	Arg (R)	Lys (K)
G204R		204	Gly (G)	Arg (R)
N969K		969	Asn (N)	Lys (K)
Q954H		954	Gln (Q)	His (H)
N764K		764	Asn (N)	Lys (K)
T9I	No estructural (nsp2)	9	Thr (T)	Ile (I)
Q19E		19	Gln (Q)	Glu (E)
G142D		142	Gly (G)	Asp (D)

T19I		19	Thr (T)	Ile (I)
Y144X		144	Tyr (Y)	Insertado X
T95I		95	Thr (T)	Ile (I)
D3N		3	As	-
P13L	No estructural (nsp1)	13	Pro (P)	Leu (L)

La Tabla 10 presenta los aminoácidos que han experimentado dos mutaciones en diversas variantes. Los aminoácidos más relevantes son E484A y P681R. El aminoácido E484A se ha asociado con una reducción en la capacidad de los anticuerpos. Además, la ubicación del aminoácido P681R en la proteína espiga del virus podría influir en su capacidad para fusionarse con las células humanas.

Tabla X
Aminoácidos afectados por dos mutaciones que se repiten en diferentes variantes.

Variante	**Variantes**	**Proteína**
T265I	Ómicron, Delta, Lambda, Mu GH, Gamma, Alpha	Spike (S)
P3395H	Ómicron, Delta, Lambda, Mu GH, Alpha	Spike (S)
S1188L	Delta, Lambda, Mu GH, Gamma, Alpha	Spike (S)
A1306S	Ómicron, Lambda, Mu GH, Alpha	Spike (S)
A1708D	Ómicron, Lambda, Mu GH, Alpha	Spike (S)
A1918V	Ómicron, Delta, Mu GH, Alpha	Spike (S)
DEL144/144	Ómicron, Delta, Mu GH, Alpha	Spike (S)
DEL157/158	Ómicron, Delta, Mu GH, Alpha	Spike (S)
DEL241/243	Ómicron, Delta, Mu GH, Alpha	Spike (S)
DEL25/27	Ómicron, Delta, Mu GH, Alpha	Spike (S)
DEL69/70	Ómicron, Delta, Mu GH, Alpha	Spike (S)
E1264D	Ómicron, Delta, Mu GH, Alpha	Spike (S)
G662S	Ómicron, Mu GH, Alpha	Spike (S)
I1566V	Ómicron, Delta, Mu GH, Alpha	Spike (S)
I2230T	Ómicron, Lambda, Mu GH, Alpha	No Estructurales (nsp2)
K3353R	Ómicron, Lambda, Mu GH, Alpha	No Estructurales (nsp2)
P1000L	Ómicron, Delta, Mu GH, Alpha	Spike (S)
P2287S	Ómicron, Lambda, Mu GH, Alpha	No Estructurales (nsp2)
T1001I	Lambda, Mu GH, Gamma, Alpha	Spike (S)
T3255I	Ómicron, Delta, Lambda, Mu GH	No Estructurales (nsp2)
T3646A	Ómicron, Delta, Lambda, Mu GH, Alpha	No Estructurales (nsp2)
V2930L	Ómicron, Lambda, Mu GH, Alpha	No Estructurales (nsp2)
L24S	Mu GH, Alpha	c
P26S	Mu GH, Alpha	No Estructurales (nsp1)
P681R	Mu GH, Alpha	Spike (S)
T19R	Mu GH, Alpha	Spike (S)
E484A	Mu GH	Nucleocápside (N)
K1655N	Gamma	Nucleocápside (N)
K1795Q	Gamma	Nucleocápside (N)
L18F	Gamma	Spike (S)
R346T	Mu GH	Nucleocápside (N)
T19I	Mu GH	No Estructurales (nsp1)
T20N	Gamma	Spike (S)
Y144X	Mu GH	Spike (S)

La Tabla 11 proporciona una visión detallada de los aminoácidos que han experimentado una mutación en diversas variantes y a que cadena proteínica pertenecen, los aminoácidos más relevantes son D614G, N501Y y P681H. El aminoácido D614G ha sido asociado con una mayor capacidad de transmisión del virus. Además, el aminoácido N501Y se ha relacionado con un posible aumento en la unión del virus a las células huésped. Finalmente, el aminoácido P681H similar a la mutación P681R afecta la infectividad del virus al influir en su capacidad para ingresar a las células humanas. Estos aminoácidos señalados en la tabla 8 son de especial interés debido a su potencial influencia en la dinámica de transmisión y la interacción del virus con las células huésped.

Tabla XI
Aminoácidos afectados por una mutación que se repite en diferentes variantes.

Número	**Aminoácido**	**Variantes**	**Proteínas**
1	D614G	Alpha, Mu GH, Lambda, Delta, Ómicron	Spike (S)

2	N501Y	Alpha, Mu GH, Ómicron	
3	P681H	Alpha, Mu GH, Ómicron	
4	DEL31/33	Alpha, Mu GH, Lambda, Delta	
5	S84L	Alpha, Mu GH, Lambda, Delta, Ómicron	
6	R203K	Alpha, Mu GH, Lambda, Ómicron	
7	G204R	Alpha, Mu GH, Lambda, Ómicron	
8	T40I	Alpha, Lambda, Delta, Ómicron	No definidas
9	V70X	Alpha, Lambda	

Los resultados expuestos del análisis del COVID-19 en Ecuador, muestran similitudes con los hallazgos de Wang [15] y Thanh [16], ya que indican que la mayoría de las mutaciones están relacionadas con el Spike (S). Por otro lado, el estudio publicado por Rozhgar [19] coincide con nuestros hallazgos al determinar que la cadena proteica afectada incluye la Nucleocápside (N). Además, el artículo de Pachetti et al. [17] al igual que nuestro estudio identificó una serie de aminoácidos afectados, específicamente aquellos ubicados en la secuencia de la proteína desde 'P' hasta 'L'. Dentro de esta secuencia, se destacaron las mutaciones P1000L y P2046L, presentes en un alto porcentaje de pacientes infectados en las variantes Ómicron, Delta, Mu GH y Alpha. Estos aminoácidos específicos (P1000L y P2046L) experimentaron dos mutaciones relevantes. Esta observación resalta su implicación en las mencionadas variantes del virus y su presencia recurrente en un número significativo de pacientes infectados.

La revisión de la literatura no encontró trabajos comparables con los datos epidemiológicos de COVID-19 del Ecuador.

VI. Conclusiones y Trabajos futuros

En esta investigación, se realizó un análisis extenso de datos de COVID-19 de Ecuador. Esto abarcó tanto la información biológica del genoma del virus (la secuencia de proteínas) como los datos epidemiológicos relacionados con los pacientes, los cuales se obtuvieron a través de la iniciativa GISAID. En el transcurso de esta investigación, se emplearon diversas técnicas de preprocesamiento de datos y análisis estadístico, que incluyeron la correlación de Pearson, la prueba de chi-cuadrado y el análisis de varianza (Anova). Además, se utilizaron diagramas y gráficos estadísticos con el objetivo de facilitar una mejor visualización de los resultados.

En conclusión, este estudio ha generado descubrimientos significativos al analizar la distribución geográfica de variables relacionadas con el COVID-19 en las diversas provincias de Ecuador. Se observó que la variante Ómicron, tuvo mayor prevalencia en gran parte del territorio ecuatoriano, seguida de cerca por la variante Delta, mientras que la variante Lambda tuvo presencia en algunas regiones específicas.

En esta línea, se identificó que las variantes Ómicron, Delta y Lambda afectaron a más del 50% de pacientes de género femenino, mientras que las variantes Mu GH y Gamma impactaron a más del 50% de pacientes de género masculino. Además, se observó una mayor incidencia del COVID-19 en pacientes de género femenino en las provincias de Pichincha con un 30.97%, Guayas con un 19.36%, y Manabí con un 8.75%. Por otro lado, los pacientes de género masculino han experimentado una alta incidencia en las provincias de Pichincha con un 30.91%, Guayas con un 17.36%, y Manabí con un 8.55%. Estos datos revelan que las provincias de Pichincha, Guayas y Manabí han sido las más afectadas por la enfermedad.

Por otro lado, durante el análisis se ha descubierto que el grupo de edad de 26 a 35 años ha sido el más afectado por todas las variantes, siendo la variante Delta más notoria en pacientes de 31 a 35 años, mientras que la variante Mu GH tuvo un impacto relevante en pacientes de 31 a 35 años y de 41 a 45 años. También se constató que, desde enero de 2021 hasta finales de noviembre de ese mismo año, las variantes predominantes en las infecciones fueron Mu GH, Gamma y Delta. No obstante, a partir de principios de diciembre de 2021 hasta 2023, la variante que ha prevalecido mayormente ha sido Ómicron.

Con este enfoque de información epidemiológica relacionado con los pacientes, las instituciones de salud pública de Ecuador podrán mejorar la comprensión de la dinámica de la enfermedad, comprender la importancia de la monitorización de enfermedades y definir políticas de seguridad

sanitaria para prevenir futuras enfermedades de variantes peligrosas del COVID-19.

Además, se descubrió información genómica de gran relevancia que destaca la relación entre las variantes y los aminoácidos. Se identificó que los aminoácidos asociados con las proteínas S, Nucleocápside (N), y las cadenas de proteínas no estructurales (nsp1) y (nsp2) fueron las más afectadas por las mutaciones. Específicamente, se observó que los aminoácidos D614G, N501Y, P681H, E484A y P681R tienen efectos significativos en la transmisibilidad del virus, su capacidad de unión a las células huésped y su habilidad para evadir la respuesta inmunológica. Estas conclusiones resaltan la importancia de comprender y monitorear estas mutaciones para desarrollar estrategias efectivas tanto en el tratamiento como en la prevención de la infección por el virus.

Para futuras investigaciones se propone incorporar mayor información de diferentes bases de datos libres que tengan validez científica para mejorar el análisis y realizar comparaciones con datos existentes, como integrar información genómica y epidemiológica de diferentes países para identificar patrones globales o específicos. Además, se sugiere considerar la inclusión de variables adicionales, como factores ambientales y las condiciones médicas subyacentes de los pacientes, con el objetivo de lograr una comprensión más profunda y holística de la infección.

Otro aspecto relevante sería realizar un análisis detallado de las mutaciones identificadas en los aminoácidos clave. Esto permitiría evaluar su impacto individual en la interacción del virus con las células huésped, su capacidad replicativa y su influencia en la respuesta inmunológica.

Asimismo, se propone la utilización de modelos predictivos o algoritmos de aprendizaje automático para prever la evolución de las variantes y su impacto, lo que podría proporcionar una visión anticipada y precisa de los posibles escenarios de la enfermedad.

Referencias

[1] G. L. Gilbert, C. Degeling, and J. Johnson, "Communicable Disease Surveillance Ethics in the Age of Big Data and New Technology," Asian Bioeth Rev, vol. 11, no. 2, pp. 173–187, Jun. 2019, doi: 10.1007/S41649-019-00087-1/METRICS.

[2] Z. S. Y. Wong, J. Zhou, and Q. Zhang, "Artificial Intelligence for infectious disease Big Data Analytics," Infect Dis Health, vol. 24, no. 1, pp. 44–48, Feb. 2019, doi: 10.1016/J.IDH.2018.10.002.

[3] S. R. Prabhu, "Infectious and Communicable Diseases: An Overview," Textbook of General Pathology for Dental Students, pp. 63–72, 2023, doi: 10.1007/978-3-031-31244-1_9.

[4] La Federación Internacional de Sociedades de la Cruz Roja y de la Media Luna Roja (IFRC), "Enfermedades transmisibles | IFRC." Accessed: Sep. 20, 2023. [Online]. Available: https://www.ifrc.org/es/nuestro-trabajo/salud-y-cuidado/salud-comunitaria/enfermedades-transmisibles

[5] Organización Mundial de la Salud and Organización Panamericana de la Salud, "Enfermedades transmitidas por vectores que ponen en riesgo a la población de las Américas." Accessed: Sep. 20, 2023. [Online]. Available: https://www.paho.org/es/noticias/7-4-2014-diez-enfermedades-transmitidas-por-vectores-que-ponen-riesgo-poblacion-americas

[6] T. P. Sheahan et al., "Comparative therapeutic efficacy of remdesivir and combination lopinavir, ritonavir, and interferon beta against MERS-CoV," Nature Communications 2020 11:1, vol. 11, no. 1, pp. 1–14, Jan. 2020, doi: 10.1038/s41467-019-13940-6.

[7] C. Taylor and J. Kidgell, "Flu-like pandemics and metaphor pre-covid: A corpus investigation," Discourse, Context & Media, vol. 41, p. 100503, Jun. 2021, doi: 10.1016/J.DCM.2021.100503.

[8] S. Chowdhury et al., "The Pattern of Highly Pathogenic Avian Influenza H5N1 Outbreaks in South Asia," Tropical Medicine and Infectious Disease 2019, Vol. 4, Page 138, vol. 4, no. 4, p. 138, Nov. 2019, doi: 10.3390/TROPICALMED4040138.

[9] C. J. C. Nicomedes and R. M. A. Avila, "An analysis on the panic during COVID-19 pandemic through an online form," J Affect Disord, vol. 276, pp. 14–22, Nov. 2020, doi: 10.1016/J.JAD.2020.06.046.

[10] OMS, "Panel de control del coronavirus (COVID-19) de la OMS | Panel de control del coronavirus (COVID-19) de la OMS con datos de vacunación." Accessed: Sep. 19, 2023. [Online]. Available: https://covid19.who.int/

[11] E. Mathieu et al., "Coronavirus Pandemic (COVID-19)," Our World in Data, Mar. 2020, Accessed: Sep. 19, 2023. [Online]. Available: https://ourworldindata.org/coronavirus

[12] Ministerio de Salud Pública de Ecuador, "Vacunómetro Covid-19." Accessed: Sep. 19, 2023. [Online]. Available: https://app.powerbi.com/view?r=eyJrIjoiYTkzNTFkMmUtZmUzNi00NDcwLTg0MDEtNjFkNzhhZTg5ZWYyIiwidCI6IjcwNjIyMGRiLTliMjktNGU5MS1hODI1LTI1NmIwNmQyNjlmMyJ9&pageName=ReportSection

[13] A. Ruiz-Bravo, M. Jiménez-Valera, A. Ruiz-Bravo, and M. Jiménez-Valera, "SARS-CoV-2 y pandemia de síndrome respiratorio agudo (COVID-19)," Ars Pharmaceutica (Internet), vol. 61, no. 2, pp. 63–79, 2020, doi: 10.30827/ARS.V61I2.15177.

[14] R. Wang, Y. Hozumi, C. Yin, and G. W. Wei, "Mutations on COVID-19 diagnostic targets," Genomics, vol. 112, no. 6, pp. 5204–5213, Nov. 2020, doi: 10.1016/J.YGENO.2020.09.028.

[15] R. Wang, J. Chen, K. Gao, and G. W. Wei, "Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries," Genomics, vol. 113, no. 4, pp. 2158–2170, Jul. 2021, doi: 10.1016/J.YGENO.2021.05.006.

[16] T. T. Nguyen et al., "Genomic mutations and changes in protein secondary structure and solvent accessibility of SARS-CoV-2 (COVID-19 virus)," Scientific Reports 2021

11:1, vol. 11, no. 1, pp. 1–16, Feb. 2021, doi: 10.1038/s41598-021-83105-3.
[17] L. van Dorp, D. Richard, C. C. S. Tan, L. P. Shaw, M. Acman, and F. Balloux, "No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2," Nature Communications 2020 11:1, vol. 11, no. 1, pp. 1–8, Nov. 2020, doi: 10.1038/s41467-020-19818-2.
[18] M. Pachetti et al., "Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant," J Transl Med, vol. 18, no. 1, pp. 1–9, Apr. 2020, doi: 10.1186/S12967-020-02344-6/FIGURES/4.
[19] R. A. Khailany, M. Safdar, and M. Ozaslan, "Genomic characterization of a novel SARS-CoV-2," Gene Rep, vol. 19, p. 100682, Jun. 2020, doi: 10.1016/J.GENREP.2020.100682.
[20] S. U. Ahmad et al., "A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries," J Infect Public Health, vol. 15, no. 8, pp. 878–891, Aug. 2022, doi: 10.1016/J.JIPH.2022.07.002.
[21] A. R. N. Zekri et al., "Characterization of the SARS-CoV-2 genomes in Egypt in first and second waves of infection," Scientific Reports 2021 11:1, vol. 11, no. 1, pp. 1–11, Nov. 2021, doi: 10.1038/s41598-021-99014-4.
[22] Federal Ministry of food and Agriculture and Max plack institut informatic, "Iniciativa GISAID." Accessed: Oct. 15, 2023. [Online]. Available: https://www.epicov.org/epi3/frontend#28bd1e
[23] S. Peng, W. Han, and G. Jia, "Pearson correlation and transfer entropy in the Chinese stock market with time delay," Data Science and Management, vol. 5, no. 3, pp. 117–123, Sep. 2022, doi: 10.1016/J.DSM.2022.08.001.
[24] C. Guevara and M. S. Penas, "Surveillance Routing of COVID-19 Infection Spread Using an Intelligent Infectious Diseases Algorithm," IEEE Access, vol. 8, pp. 201925–201936, 2020, doi: 10.1109/ACCESS.2020.3036347.
[25] S. Bahassine, A. Madani, M. Al-Sarem, and M. Kissi, "Feature selection using an improved Chi-square for Arabic text classification," Journal of King Saud University - Computer and Information Sciences, vol. 32, no. 2, pp. 225–231, Feb. 2020, doi: 10.1016/J.JKSUCI.2018.05.010.
[26] V. Sharma and R. K. Sharma, "Application of Taguchi Method and ANOVA in Parameters Optimization for Fluidization Characteristic of Pine Needles in Fluidized Bed," Lecture Notes in Mechanical Engineering, pp. 869–878, 2021, doi: 10.1007/978-981-16-0159-0_77/COVER.
[27] F. Wang, G. H. Huang, Y. Fan, and Y. P. Li, "Robust Subsampling ANOVA Methods for Sensitivity Analysis of Water Resource and Environmental Models," Water Resources Management, vol. 34, no. 10, pp. 3199–3217, Aug. 2020, doi: 10.1007/S11269-020-02608-2/METRICS.
[28] J. A. Solano, D. J. Lancheros Cuesta, S. F. Umaña Ibáñez, and J. R. Coronado-Hernández, "Predictive models assessment based on CRISP-DM methodology for students performance in Colombia - Saber 11 Test," Procedia Comput Sci, vol. 198, pp. 512–517, Jan. 2022, doi: 10.1016/J.PROCS.2021.12.278.
[29] S. Huber, H. Wiemer, D. Schneider, and S. Ihlenfeldt, "DMME: Data mining methodology for engineering applications – a holistic extension to the CRISP-DM model," Procedia CIRP, vol. 79, pp. 403–408, Jan. 2019, doi: 10.1016/J.PROCIR.2019.02.106.