PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

FACULTAD DE CIENCIAS E INGENIERÍA

ANÁLISIS DE CLUSTERING DE SECUENCIAS GENÓMICAS DE SARS-COV-2 IDENTIFICADAS EN PERÚ

Tesis para obtener el título profesional de Ingeniera Informática

AUTORA:

Carolina Estefania Mejia Mujica

ASESOR:

Dr. Edwin Rafael Villanueva Talavera

Lima, agosto, 2023

Informe de Similitud

Yo, Edwin Rafael Villanueva Talavera,

docente de la Facultad de Ciencias e Ingeniería de la Pontificia

Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado:

ANÁLISIS DE CLUSTERING DE SECUENCIAS GENÓMICAS DE SARS-COV-2 IDENTIFICADAS EN PERÚ,

del/de la autor(a)/ de los(as) autores(as):

Carolina Estefania Mejia Mujica,

dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 24%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 14/08/2023.
- He revisado con detalle dicho reporte y la Tesis o Trabajo de Suficiencia Profesional, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 14 de Agosto de 2023

Apellidos y nombres del asesor / de la asesora: Edwin Rafael Villanueva Talavera	
DNI: 29714308	Firma 
ORCID: https://orcid.org/0000-0002-6540-1230	

Resumen

El presente proyecto de tesis tiene como objetivo el desarrollo de una herramienta analítica interactiva para representar visualmente la diversidad de las secuencias genómicas de SARS-CoV-2 en el Perú, que facilite el análisis de agrupamientos en el espacio y tiempo; y, que permita incorporar nuevas secuencias. Este trabajo pretende resolver la necesidad de realizar una analítica avanzada que incluya representaciones de agrupamiento en el espacio y tiempo de la diversidad de las secuencias genómicas de SARS-CoV-2 en el Perú con el fin de apoyar la vigilancia genómica.

Esta investigación surge debido a la pandemia y al virus que evoluciona aceleradamente presentando constantes variantes genómicas, por lo que, la comunidad académica y autoridades sanitarias están interesados en entender la diversidad de estas para así realizar más estudios sobre su propagación. En el caso del Perú, de acuerdo a la revisión bibliográfica, no se encontraron estudios publicados que investiguen la distribución de las variantes del virus SARS-CoV-2. Comprender esta dinámica es importante porque ayudará a conocer el impacto de la pandemia en el país, además de tener un mejor conocimiento de la propagación del virus SARS-CoV-2 para que las autoridades sanitarias tomen acciones informadas.

Por ello, los objetivos planteados son el desarrollo de un módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2; el desarrollo de un módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 con la capacidad de incorporar nuevas secuencias; y la implementación de vistas con capacidades interactivas en ambos módulos para que el usuario interactúe con ellos.

Finalmente, la herramienta desarrollada cumple con realizar un análisis de agrupamiento y una representación visual en el espacio-tiempo de la diversidad de secuencias genómicas SARS-CoV-2 para apoyar la vigilancia genómica en el Perú.

Dedicatoria

El presente proyecto de tesis se lo dedico a mi familia, principalmente a mi madre Karim y a mis abuelos Teresa y Ricardo, por su apoyo incondicional y motivación a lo largo de toda mi carrera universitaria y a lo largo de mi vida, y quienes siempre han confiado y creído en mí.

A mi asesor, Edwin Villanueva por guiarme y brindarme las herramientas para culminar satisfactoriamente esta tesis.

A mis amigos, que me motivaron y brindaron su apoyo para realizar con éxito este proyecto.

Tabla de Contenidos

Resumen i
Dedicatoria ii
Tabla de Contenidos iii
Índice de Figuras vi
Índice de Tablas x
Capítulo 1. Generalidades 1
1.1 Problemática 1
1.1.1 Árbol de Problemas 1
1.1.2 Descripción 2
1.1.3 Problema seleccionado 5
1.2 Objetivos 5
1.2.1 Objetivo general 5
1.2.2 Objetivos específicos 6
1.2.3 Resultados esperados 6
1.2.4 Mapeo de objetivos, resultados y verificación 7
1.3 Métodos y Procedimientos 9
1.3.1. Herramientas 11
1.3.2. Métodos 13
Capítulo 2. Marco Conceptual 18
2.1 Introducción 18
2.2 Conceptos de biología 18
2.2.1 Ácidos nucleicos 18
2.2.2 Ácidos ribonucleicos 18
2.2.3 Ácidos desoxirribonucleicos 19
2.3 Conceptos de genética 20
2.3.1 Secuencias genómicas 20
2.3.2 Variante genética 20
2.3.3 SARS-CoV-2 21
2.4 Conceptos de ciencias de la computación 22
2.4.1 Análisis de agrupamiento 22
2.4.2 Analítica visual 24
Capítulo 3. Estado del Arte 25
3.1 Introducción 25
3.2 Objetivos de revisión 25
3.3 Preguntas de revisión 26
3.4 Estrategia de búsqueda 26
3.4.1 Motores de búsqueda a usar 26
3.4.2 Cadenas de búsqueda a usar 26

3.4.3 Documentos encontrados 27
3.4.4 Criterios de inclusión/exclusión 27
3.4.5 Documentos revisados 29
3.5 Formulario de extracción de datos 31
3.6 Resultados de la revisión 32
3.6.1 Respuesta a pregunta: ¿Qué métodos de análisis no supervisados se han usado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2, qué resultados, ventajas y limitantes se han obtenido y cómo son evaluados? 32
3.6.2 Respuesta a pregunta: ¿Qué métricas de distancia son las más usadas en los métodos o algoritmos de agrupamiento para secuencias genómicas SARS-CoV-2? 35
3.6.3 Respuesta a pregunta: ¿Cuáles son las estrategias de representación genómica aplicadas en el análisis de secuencias genómicas SARS-CoV-2? 36
3.6.4 Respuesta a pregunta: ¿Qué paradigmas o formas de visualización de resultados de agrupamiento se han usado en el análisis de secuencias genómicas SARS-CoV-2? 38
3.7 Conclusiones 39
Capítulo 4. Módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú 40
4.1 Introducción 40
4.2 Resultados alcanzados 40
4.2.1 Estructura de la base de datos de los módulos de software a desarrollar. 40
4.2.2 Módulo de software para realizar el pre-procesamiento de los datos de las secuencias genómicas SARS-CoV-2. 47
4.2.3 Módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú. 54
4.3 Discusión 60
Capítulo 5. Módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 62
5.1 Introducción 62
5.2 Resultados alcanzados 62
5.2.1 Selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software. 62
5.2.2 Módulo de software que implementa los métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2. 64
5.2.3 Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2. 66
5.3 Discusión 74
Capítulo 6. Implementación de vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento 76
6.1 Introducción 76
6.2 Resultados alcanzados 76

6.2.1 Lista de requerimientos a considerar para realizar la interactividad de las representaciones visuales. 76
6.2.2 Interfaces del módulo para visualizar la representación espacio-temporal (RE.1.3) y del módulo para visualizar el análisis de agrupamiento (RE.2.3) con capacidades interactivas de acuerdo con los requerimientos especificados. 78
6.3 Discusión 86
Capítulo 7. Conclusiones y trabajos futuros 87
7.1 Conclusiones 87
7.2 Trabajos futuros 88
Referencias 90
Anexos 98
Anexo A: Plan de Proyecto 98
Anexo B: Script para la creación de las tablas de la base de datos 111
Anexo C: Acta de validación del modelo físico de base de datos 112
Anexo D: Catálogo de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas SARS-CoV-2 113
Anexo E: Reporte del resultado de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas de SARS-CoV-2 120
Anexo F: Acta de validación del cumplimiento de los requerimientos del módulo espacio-temporal 127
Anexo G: Reporte de apreciación del módulo espacio-temporal 128
Anexo H: Acta de validación de métodos de análisis de agrupamiento 130
Anexo I: Catálogo de pruebas del módulo de agrupamiento 131
Anexo J: Reporte del resultado de pruebas del módulo de agrupamiento 138
Anexo K: Acta de validación del cumplimiento de los requerimientos del módulo de agrupamiento 144
Anexo L: Reporte de apreciación del módulo de agrupamiento 145
Anexo M: Acta de validación de los requerimientos 146
Anexo N: Acta de validación de cumplimiento de los requerimientos de usabilidad 147

Índice de Figuras

Figura 1. Estructura del ácido ribonucleico (ARN) que presenta los cuatro nucleótidos con las bases nitrogenadas como adenina (A), citosina (C), guanina (G) y uracilo (U). 19
Figura 2. Estructura de doble hélice del ácido desoxirribonucleico. Las hélices rotan en el sentido de las agujas del reloj, cada una de estas tiene 3.4 mm de largo y cada base nitrogenada está a 0.34 mm de distancia (Blanco & Blanco, 2017). 20
Figura 3. Estructura del virus del SARS-CoV-2. 21
Figura 4. Agrupación de un conjunto de datos utilizando el método de particionamiento k-means. 22
Figura 5. Representación dendrograma para un conjunto de datos utilizando el método jerárquico. 23
Figura 6. Accesibilidad de densidad y conectividad de densidad en el método basados en densidad (Han et al., 2012a). 23
Figura 7. Representación jerárquica del método basado en cuadrícula STING. 24
Figura 8. Distribución de los métodos no supervisados usados en 9 artículos. 33
Figura 9. Modelo físico de la base de datos. 41
Figura 10. Flujo de obtención de datos de las secuencias genómicas SARS-CoV-2. 48
Figura 11. Grafica del acumulado de varianza explicada con 10 componentes. 50
Figura 12. Flujo del preprocesamiento de las secuencias genómicas SARS-CoV-2. 52
Figura 13. Filtros de rango de fechas y departamentos. 56
Figura 14. Gráfico del mapa del Perú y gráfico circular de la distribución de secuencias genómicas SARS-CoV-2. 57
Figura 15. Gráfico lineal que muestra la evolución de variantes en el tiempo. 57
Figura 16. Tabla con los datos de las secuencias genómicas SARS-CoV-2 utilizadas en los gráficos. 58
Figura 17. Detalle de los filtros de rango de fechas y departamentos. 58
Figura 18. Detalle de las herramientas de los gráficos del módulo espacio-temporal. 59
Figura 19. Detalle de la opción de descarga en la visualización de los datos de las secuencias genómicas SARS-CoV-2. 60
Figura 20. Filtros de rango de fechas y departamentos. 68
Figura 21. Captura de pantalla del módulo de agrupamiento de secuencias genómicas SARS-CoV-2 realizado con el algoritmo k-means. 69

Figura 22. Captura de pantalla del módulo de agrupamiento de secuencias genómicas SARS-CoV-2 realizado con el algoritmo jerárquico.....69
Figura 23. Tabla con los datos de las secuencias genómicas SARS-CoV-2 agrupadas con el algoritmo jerárquico.....70
Figura 24. Captura de pantalla del módulo de agrupamiento de secuencias genómicas SARS-CoV-2 realizado con el algoritmo DBSCAN.....71
Figura 25. Detalle de los filtros de rango de fechas, algoritmo y departamentos.....72
Figura 26. Detalle de las herramientas de los gráficos de dispersión del módulo de agrupamiento.....73
Figura 27. Detalle de la opción de descarga en la visualización de los datos de las secuencias genómicas SARS-CoV-2 agrupadas con los diferentes algoritmos.....74
Figura 28. Flujo del servicio para incorporar nuevas secuencias genómicas SARS-CoV-2 al análisis.....82
Figura 29. Captura de la pantalla con las acciones que se puede realizar en los gráficos. 83
Figura 30. Captura de la pantalla con la barra de selección de número de clústeres o valor de épsilon dependiendo del algoritmo seleccionado. 84
Figura 31. Captura de la pantalla para importar los datos de las secuencias genómicas SARS-CoV-2.....85
Figura 32. Captura de la pantalla para eliminar las secuencias genómicas SARS-CoV-2. 85
Figura A1. Estructura de descomposición del trabajo del proyecto 104
Figura D1. Formato del archivo de entrada .fasta.....115
Figura D2. Formato del archivo de entrada .tsv.....115
Figura D3. Resultado de la prueba del Preprocesamiento. 116
Figura D4. Datos de entrada para la prueba del Servicio que realiza el gráfico del mapa del Perú. 116
Figura D5. Resultado de la prueba del Servicio que realiza el gráfico del mapa del Perú. ...116
Figura D6. Datos de entrada para la prueba del Servicio que realiza el gráfico circular.....117
Figura D7. Resultado de la prueba del Servicio que realiza el gráfico circular.....117
Figura D8. Datos de entrada para la prueba del Servicio que realiza el gráfico de línea. 118
Figura D9. Resultado de la prueba del Servicio que realiza el gráfico de línea. 118
Figura D10. Datos de entrada para la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2. 119
Figura D11. Resultado de la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2. 119

Figura D12. Resultado de la prueba del Servicio que calcula la cantidad total de secuencias obtenidas y la cantidad de secuencias utilizadas en el análisis. 119
Figura E1. Formato del archivo de entrada .fasta. 120
Figura E2. Formato del archivo de entrada .tsv. 121
Figura E3. Resultado de la prueba del Preprocesamiento. 121
Figura E4. Datos de entrada para la prueba del Servicio que realiza el gráfico del mapa del Perú. 122
Figura E5. Resultado de la prueba del Servicio que realiza el gráfico del mapa del Perú. 122
Figura E6. Datos de entrada para la prueba del Servicio que realiza el gráfico circular. 123
Figura E7. Resultado de la prueba del Servicio que realiza el gráfico circular. 123
Figura E8. Datos de entrada para la prueba del Servicio que realiza el gráfico de línea. 124
Figura E9. Resultado de la prueba del Servicio que realiza el gráfico de línea. 124
Figura E10. Datos de entrada para la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2. 125
Figura E11. Resultado de la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2. 125
Figura E12. Resultado de la prueba del Servicio que calcula la cantidad total de secuencias obtenidas y la cantidad de secuencias utilizadas en el análisis. 126
Figura I1. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means. 133
Figura I2. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means. 134
Figura I3. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico. 134
Figura I4. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico. 135
Figura I5. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN. 136
Figura I6. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN. 136
Figura I7. Datos de entrada para la prueba del Servicio que realiza el dendrograma. 137
Figura I8. Resultado de la prueba del Servicio que realiza el dendrograma. 137
Figura J1. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means. 138

Figura J2. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means. 139
Figura J3. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico. 140
Figura J4. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico. 140
Figura J5. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN. 141
Figura J6. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN. 142
Figura J7. Datos de entrada para la prueba del Servicio que realiza el dendrograma. 142
Figura J8. Resultado de la prueba del Servicio que realiza el dendrograma. 143

Índice de Tablas

Tabla 1. Árbol de problemas....1
Tabla 2. Resultados esperados, medios de verificación e indicadores objetivamente verificables para el objetivo específico O1....7
Tabla 3. Resultados esperados, medios de verificación e indicadores objetivamente verificables para el objetivo específico O2....8
Tabla 4. Resultados esperados, medios de verificación e indicadores objetivamente verificables para el objetivo específico O3....9
Tabla 5. Herramientas, métodos y procedimientos a emplear para los resultados esperados del objetivo específico O1....10
Tabla 6. Herramientas, métodos y procedimientos a emplear para los resultados esperados del objetivo específico O2....10
Tabla 7. Herramientas, métodos y procedimientos a emplear para los resultados esperados del objetivo específico O3....11
Tabla 8. Documentos encontrados por motor de búsqueda....27
Tabla 9. Documentos obtenidos luego de aplicar los criterios de inclusión/exclusión en cada motor de búsqueda....28
Tabla 10. Documentos seleccionados para la revisión sistemática....29
Tabla 11. Campos del formulario de extracción de información....31
Tabla 12. Diccionario de datos de la tabla Departamentos....43
Tabla 13. Diccionario de datos de la tabla Secuencias....43
Tabla 14. Diccionario de datos de la tabla Variantes....44
Tabla 15. Diccionario de datos de la tabla Algoritmos....45
Tabla 16. Diccionario de datos de la tabla Archivos....46
Tabla 17. Diccionario de datos de la tabla Agrupamiento....46
Tabla 18. Tabla de prioridad....54
Tabla 19. Lista de requerimientos para el módulo espacio-temporal....55
Tabla 20. Tabla de prioridad....66
Tabla 21. Lista de requerimientos para el módulo de agrupamiento....66
Tabla 22. Tabla de prioridad....77
Tabla 23. Lista de requerimientos para realizar la interactividad de las representaciones visuales....77
Tabla A1. Riesgos del proyecto identificados....103

Tabla A2. Lista de tareas. 105
Tabla A3. Costeo del proyecto.....110
Tabla D1. Catálogo de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas de SARS-CoV-2. 113
Tabla I1. Catálogo de pruebas del módulo de agrupamiento.....131

Capítulo 1. Generalidades

1.1 Problemática

Dentro de este capítulo se desarrollará el problema central del trabajo, la descripción de cómo surgió en la situación actual y los efectos que puede generar. En las siguientes secciones, se planteará lo mencionado mediante un esquema de árbol de problemas, descripción del árbol y se concluirá en el problema seleccionado.

1.1.1 Árbol de Problemas

En la Tabla 1, se presenta el árbol de problemas, en el cual se detalla los problemas causa de la problemática, el problema central o principal del presente estudio de investigación y los problemas efectos de esta problemática.

Tabla 1. Árbol de problemas

Problemas efectos	Al no tener herramientas informáticas de apoyo de fácil uso para que las autoridades y los investigadores obtengan conocimiento sobre la dinámica espacio-temporal de la diversidad genómica del SARS-CoV-2 en Perú se dificulta anticipar acciones informadas a fin de contener la propagación del virus.	Dificultad en el análisis y vigilancia continua de la diversidad genómica de variantes del virus SARS-CoV-2 en Perú.	Las herramientas analíticas de diversidad genómica que ofrecen vistas estáticas y no permiten al usuario interactuar con el proceso y los resultados, limitan la interpretación de los mismos y un mejor entendimiento del fenómeno subyacente y la formulación de nuevas hipótesis.
Problema central	Necesidad de realizar una analítica avanzada que incluya representaciones visuales interactivas y analíticas de agrupamiento en el espacio y a lo largo del tiempo de la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú a fin de apoyar la vigilancia genómica		

Problemas causas	No se cuenta con estudios publicados que comprendan o investiguen el análisis espacio-temporal de la diversidad genómica de las muestras de virus SARS-CoV-2 secuenciadas en Perú.	No se cuenta con estudios publicados en el Perú sobre el análisis de agrupamiento de las secuencias genómicas de las variantes de SARS-COV-2, con la posibilidad de incorporar nuevas secuencias sin la necesidad de rehacer la representación y el análisis.	La mayoría de recursos y librerías para el análisis de diversidad genómica son de uso complejo y generan mayormente vistas estáticas, no permitiendo al usuario interactuar con las vistas y resultados.

Nota. Elaboración propia.

1.1.2 Descripción

En diciembre del 2019, se reportó un grupo de casos de neumonía de origen desconocido en Wuhan, China (Wenjie Tan et al., 2020). Unas semanas después, se detectaron casos de COVID-19 en diferentes países a nivel mundial; por ello, la Organización Mundial de la Salud (OMS) declaró, el 30 de enero de 2020, el brote del SARS-CoV-2, una emergencia de salud pública internacional (Cucinotta & Vanelli, 2020). El 11 de marzo de 2020, la OMS declaró el brote del COVID-19 una pandemia mundial (Cucinotta & Vanelli, 2020).

SARS-CoV-2 se transmite principalmente por vía respiratoria, siendo el período de incubación de aproximadamente 14 días y presenta características clínicas variables; la mayoría de personas presentan síntomas leves o son asintomáticos y otras personas requieren hospitalización o una terapia con ventilación mecánica, llegando en muchos casos a la muerte (Córdova-Aguilar et al., 2020).

Al inicio de la pandemia las personas mayores presentaban una mayor afección por el virus y una mayor tasa de mortalidad, al igual que los hombres; sin embargo, conforme se han ido detectando diferentes variantes alrededor del mundo, la tasa de mortalidad en la población joven ha ido aumentando (Hahn et al., 2021). Según la OMS, hasta el 11 de noviembre de 2021, se tiene más de 251 millones de casos confirmados de COVID-19 y más de 5 millones de muertes a nivel mundial (World Health Organization, 2021).

En el Perú, según el Instituto Nacional de Salud y Centro Nacional de Epidemiología, Prevención y Control de Enfermedades – MINSA, hasta el 11 de noviembre de 2021, se registraron 2 211 366 casos confirmados y 200 554 muertes con una letalidad del 9,07% (MINSA, 2021).

Dada la rápida propagación del SARS-CoV-2, la comunidad académica, empresas farmacéuticas y autoridades sanitarias están interesados en entender la diversidad de secuencias genómicas obtenidas de las personas contagiadas de este virus, para así realizar más estudios sobre la propagación de este, así como desarrollar posibles vacunas. Por ello, el intercambio internacional de información de secuencias genómicas del virus SARS-CoV-2 se ha tornado en algo fundamental en las investigaciones del SARS-CoV-2. Es así que han surgido repositorios centrales y de acceso público como la base de datos GISAID (*Global Initiative on Sharing All Influenza Data*) en donde diferentes centros de investigación depositan constantemente datos de secuencias genómicas del SARS-CoV-2 recolectadas a lo largo del mundo, incluyendo el INS del Perú (Elbe & Buckland-Merrett, 2017; Zhao Id et al., 2020). A la fecha (20/01/2022), GISAID contiene 8 509 522 de secuencias de SARS-CoV-2 recolectadas de hospederos humanos, de los cuales más de 17 395 muestras corresponden a secuencias recolectadas en Perú (GISAID, 2022).

Gracias al repositorio GISAID es que se ha posibilitado la investigación de la diversidad genética, la propagación espacial y la filodinámica del virus a nivel mundial (Hahn et al., 2021; Toyoshima et al., 2020; X. Yang et al., 2020). Por ejemplo, se han realizado estudios sobre la dinámica espacio-temporal de la transmisión del virus en Bangladesh (Islam et al., 2021), China (Huang et al., 2020), Irán (Pourghasemi et al., 2020) y Estados Unidos (Y. Wang et al., 2021). También se han realizado análisis filogenéticos para comprender las mutaciones, agrupando las mutaciones para así encontrar patrones de evolución del virus (Hozumi et al., 2021). A partir de estos estudios, se ha ayudado en la focalización de lugares donde se necesita

tomar mejores decisiones y medidas para enfrentar el virus (Islam et al., 2021). Además, algunos resultados de estos estudios han ayudado a la creación de vacunas efectivas dado que han generado conocimientos para entender la variabilidad genética del virus en determinadas regiones (Hahn et al., 2021).

En el caso del Perú, de acuerdo a nuestra revisión bibliográfica, no se encontraron estudios publicados que investiguen la distribución en el espacio y tiempo de las variantes del virus SARS-CoV-2. Comprender esta dinámica es importante porque ayudaría a conocer el impacto de la pandemia en el país, así como, a tener un mejor conocimiento de la propagación del virus SARS-CoV-2 para que las autoridades sanitarias puedan tomar acciones informadas (Islam et al., 2021).

A pesar de las posibilidades que nos ofrece un repositorio de secuencias como GISAID y los avances de los estudios derivados del mismo, se ha podido observar en la revisión de la literatura que la mayoría de estudios que realizan análisis de agrupamiento de las secuencias genómicas para entender la evolución y transmisión del virus SARS-CoV-2 plantean análisis de manera estática (Hozumi et al., 2021; Morais et al., 2020). Esto quiere decir que, si se quiere aumentar la cantidad de datos en un futuro, debido a que en el tiempo se presentan más secuencias genómicas, se tendría que realizar todo el análisis de nuevo. De acuerdo a la revisión sistemática realizada, se concluye que no se han explorado mayormente métodos que permitan hacer análisis de agrupamientos de forma online, en la cual se pueda agregar nuevas secuencias sin necesidad de tener que recrear la representación de todo el conjunto de datos y el análisis de agrupamiento o clustering. Esta capacidad es importante si se quiere tener una herramienta que apoye la vigilancia genómica en la cual se pueda mapear nuevas secuencias de variantes a medida que estas son disponibilizadas (Santos, 2021). Rehacer todo el análisis con cada nueva secuencia del virus SARS-CoV-2, además de ser ineficiente y demorado, puede generar diferentes representaciones cada vez que se ejecuta, lo que dificultaría el análisis y seguimiento.

Otra limitación que se ha podido observar en la mayoría de los estudios de la literatura, es que ellos están enfocados en realizar un análisis de agrupamiento estático, con el cual se obtienen y muestran resultados de la situación pasada de la diversidad genómica del virus SARS-CoV-2 (X. Yang et al., 2020). Generalmente no ofrecen herramientas con capacidad de interactuar con ellas. Ello dificulta replicar dichos estudios en nuevos lugares, como el Perú, y poder inspeccionar o avizorar si está emergiendo alguna nueva variante del virus, lo cual es importante ya que el aumento de casos por covid-19 se debe en parte a la propagación de nuevas variantes genómicas del virus (Serrano, 2021).

En conclusión, todos estos problemas evidencian la falta de una herramienta informática que apoye en la vigilancia genómica, donde se carguen los datos de las secuencias del virus SARS-CoV-2, recolectados en el Perú desde el inicio de la pandemia, para analizar, visualizar e interactuar con vistas que nos ayuden a comprender sobre la distribución en el espacio y el tiempo de la diversidad genómica del virus en Perú.

1.1.3 Problema seleccionado

De acuerdo a lo mencionado anteriormente, el problema central que el presente estudio de investigación pretende resolver, es la necesidad de realizar una analítica avanzada que incluya representaciones visuales interactivas y analíticas de agrupamiento en el espacio y a lo largo del tiempo de la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú a fin de apoyar la vigilancia genómica.

1.2 Objetivos

1.2.1 Objetivo general

En el presente proyecto, se ha definido como objetivo general desarrollar una herramienta analítica interactiva para representar visualmente la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú, que facilite la realización de análisis de agrupamientos en el espacio y a lo largo del tiempo; y, que permita incorporar nuevas secuencias en el análisis.

1.2.2 Objetivos específicos

O 1. Desarrollar un módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú.

O 2. Desarrollar un módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 y su visualización con capacidad de incorporar nuevas secuencias de forma en línea.

O 3. Implementar las vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento para que el usuario interactúe con los módulos de software.

1.2.3 Resultados esperados

O 1. Desarrollar un módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú.

RE.1.1. Estructura de la base de datos de los módulos de software a desarrollar.

RE.1.2. Módulo de software para realizar el pre-procesamiento de los datos de las secuencias genómicas SARS-CoV-2 y su representación en baja dimensión.

RE.1.3. Módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú.

O 2. Desarrollar un módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 y su visualización con capacidad de incorporar nuevas secuencias de forma en línea.

RE.2.1. Selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.

RE.2.2. Módulo de software que implementa los métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2.

RE.2.3. Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2.

O 3. Implementar las vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento para que el usuario interactúe con los módulos de software.

RE.3.1. Lista de requerimientos a considerar para realizar la interactividad de las representaciones visuales.

RE.3.2. Interfaces del módulo para visualizar la representación espacio-temporal (RE.1.3) y del módulo para visualizar el análisis de agrupamiento (RE.2.3) con capacidades interactivas de acuerdo con los requerimientos especificados.

1.2.4 Mapeo de objetivos, resultados y verificación

En la Tabla 2, Tabla 3 y Tabla 4, se presentan los objetivos con sus resultados esperados, medios de verificación e indicadores objetivamente verificables para la medición y validación de los resultados.

Tabla 2. Resultados esperados, medios de verificación e indicadores objetivamente verificables para el objetivo específico O1

Objetivo: O1. Desarrollar un módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú.		
Resultado esperado	**Medio de verificación**	**Indicador objetivamente verificable**
RE.1.1. Estructura de la base de datos de los módulos de software a desarrollar.	- Documento que describe el modelo físico de la base de datos. - Script de creación de la base de datos.	- Validación del documento por parte de un especialista en base de datos.
RE.1.2. Módulo de software para realizar el pre-procesamiento de los datos de las secuencias genómicas SARS-CoV-2 y su representación en baja dimensión.	- Código fuente del software. - Catálogo de pruebas.	- Resultado de pruebas unitarias aprobadas al 100%.
RE.1.3. Módulo para visualizar la representación espacio-	- Documento que contiene la lista de requerimientos	- Aprobación de que el software cumpla con el 100% de los requerimientos especificados,

temporal de las secuencias genómicas SARS-CoV-2 en el Perú.	para el desarrollo del módulo. - Código fuente y manual de usuario del módulo.	validados por un especialista en ingeniería informática. - Reporte de apreciación del módulo de visualización espacio-temporal por un especialista en bioinformática o biología molecular.

Nota. Elaboración propia.

Tabla 3. Resultados esperados, medios de verificación e indicadores objetivamente verificables para el objetivo específico O2

Objetivo: O2. Desarrollar un módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 y su visualización con capacidad de incorporar nuevas secuencias de forma en línea.		
Resultado esperado	**Medio de verificación**	**Indicador objetivamente verificable**
RE.2.1. Selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.	- Informe de selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.	- Validación del documento por parte de un especialista en inteligencia artificial o ciencia de datos.
RE.2.2. Módulo de software que implementa los métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2.	- Código fuente del software. - Catálogo de pruebas.	- Resultado de pruebas unitarias aprobadas al 100%. - Se debe implementar por lo menos un algoritmo de clustering particional y uno de clustering jerárquico.
RE.2.3. Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2.	- Documento que contiene la lista de requerimientos para el desarrollo del módulo. - Código fuente y manual de usuario del módulo.	- Aprobación de que el software cumpla con el 100% de los requerimientos especificados, validados por un especialista en ingeniería informática. - Reporte de apreciación del módulo de visualización del análisis de agrupamiento por un especialista en bioinformática o biología molecular.

Nota. Elaboración propia.

Tabla 4. Resultados esperados, medios de verificación e indicadores objetivamente verificables para el objetivo específico O3

Objetivo: O3. Implementar las vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento para que el usuario interactúe con los módulos de software.		
Resultado esperado	**Medio de verificación**	**Indicador objetivamente verificable**
RE.3.1. Lista de requerimientos a considerar para realizar la interactividad de las representaciones visuales.	- Documento que contiene la lista de requerimientos.	- Validación del 100% de los requerimientos por parte de un especialista en bioinformática o biología molecular.
RE.3.2. Interfaces del módulo para visualizar la representación espacio-temporal (RE.1.3) y del módulo para visualizar el análisis de agrupamiento (RE.2.3) con capacidades interactivas de acuerdo con los requerimientos especificados.	- Código fuente.	- Aprobación de que el software cumpla con el 100% de los requerimientos de usabilidad especificados, validados por un especialista en interacción humano computador (HCI).

Nota. Elaboración propia.

1.3 Métodos y Procedimientos

En esta sección, se presentan las herramientas, métodos y procedimientos a emplear para materializar cada resultado esperado. En la Tabla 5, Tabla 6 y Tabla 7, se presentan los objetivos con sus resultados esperados asociados a las herramientas, métodos y procedimientos a emplear.

Tabla 5. Herramientas, métodos y procedimientos a emplear para los resultados esperados del objetivo específico O1

Objetivo: O1. Desarrollar un módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú.	
Resultados esperados	**Herramientas, métodos y procedimientos**
RE.1.1. Estructura de la base de datos de los módulos de software a desarrollar.	**Herramientas**: PostgreSQL
RE.1.2. Módulo de software para realizar el pre-procesamiento de los datos de las secuencias genómicas SARS-CoV-2 y su representación en baja dimensión.	**Herramientas:** Python, Biopython, SciPy, Scikit Learn, NumPy, TensorFlow, FastApi, *Amazon Web Services* (AWS) **Métodos:** Alineamiento múltiple de secuencias (MSA, por sus siglas en inglés), Análisis de componentes principales (PCA, por sus siglas en inglés), Escalamiento multidimensional (MDS, por sus siglas en inglés), *Landmark* MDS (LMDS) y Perceptrón multicapa (MLP, por sus siglas en inglés).
RE.1.3. Módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú.	**Herramientas:** React, Bokeh, *Amazon Web Services* (AWS) **Métodos**: Visualización de datos (mapa regional)

Nota. Elaboración propia.

Tabla 6. Herramientas, métodos y procedimientos a emplear para los resultados esperados del objetivo específico O2

Objetivo: O2. Desarrollar un módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 y su visualización con capacidad de incorporar nuevas secuencias de forma en línea.	
Resultados esperados	**Herramientas, métodos y procedimientos**
RE.2.1. Selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.	

RE.2.2. Módulo de software que implementa los métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2.	**Herramientas:** Python, FastApi, SciPy, Scikit Learn, *Amazon Web Services* (AWS) **Métodos:** Algoritmo *k-means*, agrupación jerárquica y el algoritmo de agrupación espacial basado en densidad de aplicaciones con ruido (DBSCAN, por sus siglas en inglés).
RE.2.3. Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2.	**Herramientas:** React, Bokeh, *Amazon Web Services* (AWS) **Métodos:** Visualización de datos (gráfico de dispersión)

Nota. Elaboración propia.

Tabla 7. Herramientas, métodos y procedimientos a emplear para los resultados esperados del objetivo específico O3

Objetivo: O3. Implementar las vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento para que el usuario interactúe con los módulos de software.	
Resultados esperados	**Herramientas, métodos y procedimientos**
RE.3.1. Lista de requerimientos a considerar para realizar la interactividad de las representaciones visuales.	
RE.3.2. Interfaces del módulo para visualizar la representación espacio-temporal (RE.1.3) y del módulo para visualizar el análisis de agrupamiento (RE.2.3) con capacidades interactivas de acuerdo con los requerimientos especificados.	**Herramientas:** React, Bokeh, *Amazon Web Services* (AWS) **Métodos:** Visualización de datos (gráficos interactivos)

Nota. Elaboración propia.

1.3.1. Herramientas

- **PostgreSQL**: es un sistema de base de datos relacional orientado a objetos (PostgreSQL, 1996). Es muy potente y de código abierto, con más de 30 años de desarrollo activo se ha

ganado una sólida reputación por la integridad de datos, una gran rendimiento y funciones robustas (PostgreSQL, 1996).

- **Python**: es un lenguaje de programación orientado a objetos y de alto nivel que permite el desarrollo rápido de aplicaciones, posee una sintaxis sencilla y fácil de aprender, admite módulos y paquetes de tal manera que permite reutilizar el código (Python Software Foundation, 2001). Esta herramienta se utilizará para el preprocesamiento de las secuencias genómicas SARS-CoV-2, del desarrollo del módulo espacio-temporal y del módulo de agrupamiento.
- **Biopython**: es un conjunto de herramientas de Python para biología computacional que provee funciones para manejar las secuencias genómicas (Biopython, 2021). Esta herramienta será utilizada para el preprocesamiento de las secuencias genómicas SARS-CoV-2.
- **FastApi**: es un *framework* moderno, rápido y de alto rendimiento para el desarrollo de aplicaciones web en Python, está diseñado para el desarrollo de modelos de aprendizaje automático (FastAPI, s.f.). Esta herramienta será utilizada para el desarrollo del módulo espacio-temporal y del módulo de agrupamiento.
- **Scikit Learn**: es un módulo de Python que incorpora algoritmos de aprendizaje automático para problemas supervisados y no supervisados (Pedregosa et al., 2011). Esta herramienta será utilizada para implementar los métodos de agrupamiento.
- **SciPy**: es un software de código abierto conformado por herramientas matemáticas y de computación de datos para matemáticas, ciencias e ingeniería (SciPy, s.f.). Esta herramienta será utilizada para implementar los métodos de agrupamiento.
- **NumPy**: es un software de código abierto que permite la computación numérica con Python (NumPy, 2019). Esta herramienta será utilizada para el preprocesamiento de las secuencias genómicas SARS-CoV-2.

- **TensorFlow**: es una librería de código abierto que permite desarrollar y/o entrenar modelos de aprendizaje automático (TensorFlow, s.f.). Esta herramienta será utilizada para el preprocesamiento de las secuencias genómicas SARS-CoV-2.
- ***Amazon Web Services*** **(AWS)**: es una plataforma tecnológica en la nube que ofrece servicios web seguros y a bajo costo (Amazon Web Services, 2021). Proporciona una plataforma de infraestructura global escalable, accesible y flexible lo que permite innovar con mayor rapidez (Amazon Web Services, 2021). Esta herramienta será utilizada para alojar la base de datos y desplegar la herramienta en la nube.
- **React**: es una librería de JavaScript de software libre desarrollada por *Facebook* para el desarrollo de interfaces de usuario interactivas de forma sencilla (React, s.f.). Esta herramienta será utilizada debido a la gran flexibilidad que ofrece para desarrollar las interfaces gráficas del módulo espacio-temporal y del módulo de agrupamiento.
- **Bokeh**: es una librería de visualización interactiva de alto rendimiento en conjunto de datos muy grandes, permite crear gráficos versátiles e interactivos (Bokeh, s.f.). Esta herramienta será utilizada para el desarrollo de interfaces del módulo espacio-temporal y del módulo de agrupamiento.

1.3.2. Métodos

- **Alineamiento múltiple de secuencias (MSA, por sus siglas en inglés)**: es el alineamiento de más de dos secuencias, ya que si la alineación es de dos secuencias sería una alineación de pares (Bawono & Heringa, 2014). Para realizar este alineamiento se necesita que todas las secuencias tengan la misma longitud (Bawono & Heringa, 2014). Este alineamiento se utiliza para conocer los sitios variables, donde en una misma posición para todas las secuencias se tiene diferentes aminoácidos o nucleótidos dentro de la secuencia, proporcionando información valiosa respecto a las relaciones evolutivas y funcionales de la secuencia (Bawono & Heringa, 2014). El alineamiento múltiple de

secuencias tiene como objetivo representar de manera óptima las relaciones evolutivas entre las secuencias (Bawono & Heringa, 2014). Este alineamiento será utilizado como parte de la estrategia de reducción de dimensionalidad de las secuencias genómicas SARS-CoV-2.

- **Análisis de componentes principales (PCA, por sus siglas en inglés)**: es una técnica de reducción de dimensionalidad de aprendizaje no supervisado, que permite la transformación de los datos a otro espacio de menor dimensión mediante la identificación de ejes ortogonales, componentes principales, y no correlacionadas entre sí, los cuales deben explicar la varianza máxima del conjunto de datos (Tharwat, 2016). Este análisis de componentes principales será utilizado como estrategia de reducción de dimensionalidad de las secuencias genómicas SARS-CoV-2.
- **Escalamiento multidimensional (MDS, por sus siglas en inglés):** es una técnica para el análisis de similitud de un conjunto de datos en un espacio de baja dimensión (Borg & Groenen, 2005). Se forma un espacio geométrico con las distancias entre puntos del conjunto de datos, esto se realiza porque se quiere tener una representación gráfica de la estructura del conjunto de datos donde se muestre información esencial (Borg & Groenen, 2005). Existen variedades de MDS dependiendo del tipo de geometría que se quiere para representar a los datos. Esta técnica de escalamiento multidimensional será utilizada como estrategia de reducción de dimensionalidad de las secuencias genómicas SARS-CoV-2.
- ***Landmark* MDS (LMDS)**: es un método que utiliza las distancias entre unos puntos de referencia y todos los demás puntos en un conjunto de datos para así determinar dónde deben ir todos los demás puntos (De Silva & Tenenbaum, 2004). Este método soporta la introducción de nuevos puntos de datos, ya que como los puntos de referencia son fijos, se utiliza estos para calcular la posición de los nuevos puntos (De Silva & Tenenbaum, 2004). La elección de los puntos de referencia suele ser de manera aleatoria pero también se puede

utilizar otras técnicas determinísticas para elegir los mejores puntos de referencia (De Silva & Tenenbaum, 2004). Este método será utilizado como parte del preprocesamiento de las secuencias genómicas SARS-CoV-2.

- **Perceptrón multicapa (MLP, por sus siglas en inglés)**: es una estructura de red neuronal que se aproxima a cualquier función continua y consta de tres tipos de capas: la capa de entrada, las capas ocultas o intermedias y la capa de salida (Abirami, S., & Chitra, P., 2020). El modelo en cada neurona de la red tiene una función de activación no lineal (Sainlez, M., & Heyen, G., 2011). Además, los MLP pueden resolver problemas de clasificación de conjunto de datos que no son linealmente separables (Abirami, S., & Chitra, P., 2020). Esta red neuronal será utilizada como parte del preprocesamiento de las secuencias genómicas SARS-CoV-2.
- **Algoritmo *K-means***: es un método de agrupamiento que agrupa el conjunto de datos en k clústeres o grupos, siendo estos clústeres representados por su centroide(Georgieva et al., 2013). Se asigna cada punto de datos al clúster cuyo centroide está más próximo, este es un procedimiento iterativo, donde en cada iteración se calcula nuevos centroides y se vuelven a formar los clústeres, esto se repite hasta que los centroides no cambien de posición (Georgieva et al., 2013). Este método será utilizado en la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.
- **Agrupación jerárquica**: es un método de agrupamiento que proporciona información sobre la estructura o jerarquía de los clústeres que se pueden formar con el conjunto de datos (Georgieva et al., 2013). Esta jerarquía se visualiza mediante el dendrograma (Georgieva et al., 2013). Los algoritmos de agrupación jerárquica pueden ser aglomerativos o divisivos (Georgieva et al., 2013). El agrupamiento jerárquico aglomerativo comienza a formar grupos o clústeres desde abajo hacia arriba, combinando

uno a uno los clústeres hasta llegar a la raíz (Georgieva et al., 2013). En cambio, el agrupamiento jerárquico divisivo comienza con todo el conjunto de datos como un clúster o grupo y va haciendo divisiones consecutivas hasta llegar abajo, obteniendo finalmente varios grupos (Georgieva et al., 2013). Este método será utilizado en la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.

- **Algoritmo de agrupación espacial basado en densidad de aplicaciones con ruido (DBSCAN, por sus siglas en inglés)**: Este algoritmo está diseñado para descubrir agrupaciones de forma arbitraria y el ruido en el espacio (Ester, Kriegel, Sander, & Xu, 1996). Este algoritmo es eficiente para grandes bases de datos espaciales y es escalable porque realiza un único recorrido al conjunto de datos (Ester, Kriegel, Sander, & Xu, 1996). Este método será utilizado en la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.
- **Visualización de datos**: Cuando se lee datos numéricos se procesa esta información abstracta y se trata de convertir en una visión más profunda e intuitiva. Por ello, se debe representar la información de manera visual porque es más compacto y accesible para todos, se logra un mejor entendimiento de la información de manera llamativa (Grant, 2019). La visualización de la información no solo es útil para comunicar mensajes sino también para que especialistas y analistas puedan comprender los datos en profundidad (Grant, 2019).

  Respecto a los gráficos interactivos, se sabe que hay muchas formas en las que el usuario puede interactuar con una visualización (Grant, 2019). A continuación, se presentan los enfoques que más se utilizan en la actualidad; siendo el primero, el desplazamiento del mouse para que aparezca un cuadro flotante con información extra, hacer un clic para mostrar más información, hacer un clic para reemplazar el contenido con más detalle, hacer

clic y arrastrar el mouse para hacer zoom, mover controles para cambiar el aspecto de la visualización, alternar entre mostrar y ocultar diferente información, se puede desplazar o hacer zoom en un mapa, etc. (Grant, 2019). Estos enfoques y técnicas serán empleadas en este proyecto para presentar de manera visual e interactiva la información relacionada a las secuencias genómicas SARS-CoV-2.

Capítulo 2. Marco Conceptual

2.1 Introducción

En este capítulo, se presenta una breve definición de los conceptos necesarios para este estudio. Se mencionan definiciones sobre conceptos biológicos, genéticos y computacionales.

2.2 Conceptos de biología

2.2.1 Ácidos nucleicos

Los ácidos nucleicos son macromoléculas que se encuentran en todas las células y virus, formados por cadenas de nucleótidos (Blanco & Blanco, 2017). Estos ácidos tienen como función almacenar información genética en sistemas biológicos; y, se clasifican en ácidos ribonucleicos (ARN) y ácidos desoxirribonucleicos (ADN), la diferencia de estos está en la pentosa que presentan, ribosa en el ARN y desoxirribosa en el ADN, en las bases nitrogenadas que contienen y la estructura de las cadenas (Blanco & Blanco, 2017). La información almacenada en la secuencia de nucleótidos de los ácidos nucleicos es lo que determina la singularidad funcional de cada ser vivo (Blanco & Blanco, 2017).

2.2.2 Ácidos ribonucleicos

El ácido ribonucleico (ARN) es un polinucleótido que tiene una sola cadena, está constituido por ribosa (azúcar de cinco carbonos), fosfato y bases nitrogenadas como adenina (A), citosina (C), guanina (G) y uracilo (U) (Blanco & Blanco, 2017; Minchin & Lodge, 2019). Existen diferentes tipos de ARN en la célula, entre ellos están: el ácido ribonucleico mensajero (ARNm), el ácido ribonucleico ribosómico (ARNr) y el ácido ribonucleico de transferencia (ARNt), ARN nuclear pequeño, microARN, entre otros (Blanco & Blanco, 2017).

- El **ácido ribonucleico mensajero** (ARNm) es una cadena simple que transfiere información genética desde el ADN nuclear hacia el citoplasma donde se realiza la síntesis de proteínas, comprende aproximadamente el 5% del ARN (Blanco & Blanco, 2017).

- El **ácido ribonucleico ribosómico** (ARNr) se encarga del proceso de síntesis de proteínas, representa alrededor del 80% del ARN, constituye más del 55% del peso de un ribosoma (Blanco & Blanco, 2017).
- El **ácido ribonucleico de transferencia** (ARNt) es una molécula pequeña del ARN que participa en la síntesis de proteínas; asimismo, asegura la ubicación exacta de cada aminoácido (Blanco & Blanco, 2017).

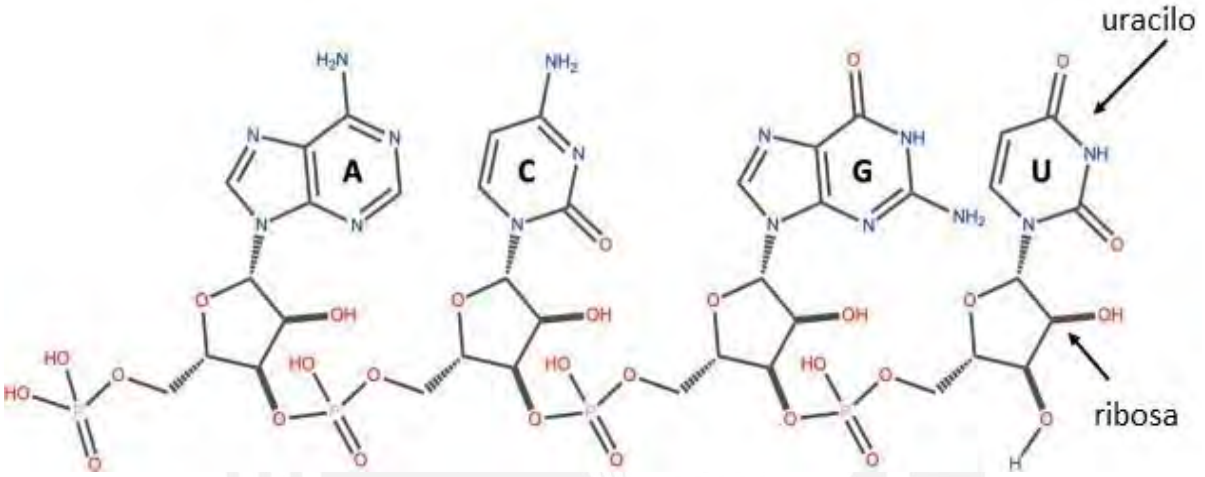


Figura 1. Estructura del ácido ribonucleico (ARN) que presenta los cuatro nucleótidos con las bases nitrogenadas como adenina (A), citosina (C), guanina (G) y uracilo (U).

Adaptado de (Minchin & Lodge, 2019).

2.2.3 Ácidos desoxirribonucleicos

El ácido desoxirribonucleico (ADN) es una molécula que codifica la información que la célula necesita para producir proteínas (Blanco & Blanco, 2017). El ADN se compone de dos cadenas de polinucleótidos que se enrollan entre ellas en un mismo eje formando una estructura de doble hélice (Blanco & Blanco, 2017), estas dos cadenas corren en direcciones opuestas (Blanco & Blanco, 2017; Minchin & Lodge, 2019). Está constituida por bases nitrogenadas como adenina (A), citosina (C), guanina (G) y timina (T); siendo la secuencia de estas las que codifican la información genética de cada ser vivo (Blanco & Blanco, 2017).

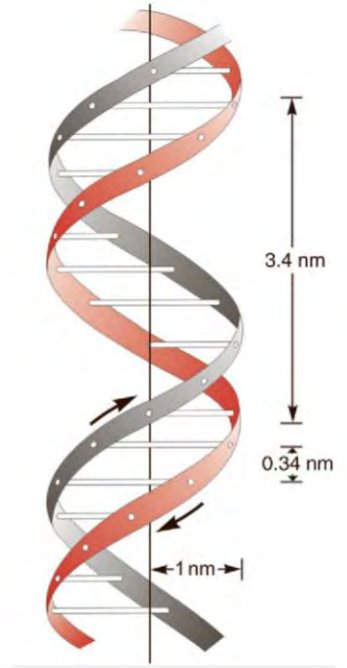


Figura 2. Estructura de doble hélice del ácido desoxirribonucleico. Las hélices rotan en el sentido de las agujas del reloj, cada una de estas tiene 3.4 mm de largo y cada base nitrogenada está a 0.34 mm de distancia (Blanco & Blanco, 2017).

2.3 Conceptos de genética

2.3.1 Secuencias genómicas

Las secuencias genómicas se refieren a las secuencias del genoma completo de un organismo (Saraswathy & Ramalingam, 2011), que proporcionan una gran cantidad de datos los cuales ayudan en el conocimiento de nuevas enfermedades o para la realización de medicinas (Hood & Rowen, 2013). En estas secuencias se observa el orden de los nucleótidos o bases nitrogenadas del ADN de un genoma, los cuales pueden ser adenina (A), citosina (C), guanina (G) y timina (T) (Saraswathy & Ramalingam, 2011).

2.3.2 Variante genética

La variante genética es el cambio en la secuencia genómica (Lauring & Hodcroft, 2021), ésta a su vez, puede comprometer una o varias mutaciones. Una mutación puede ser una sustitución, inserción o eliminación de un nucleótido en la secuencia genómica (Lauring & Hodcroft, 2021). Por otro lado, se puede definir a la variante genética como una cepa que

presenta un fenotipo comprobablemente diferente (Lauring & Hodcroft, 2021). La interpretación de variantes puede variar en los diferentes laboratorios, de acuerdo al criterio profesional, a pesar de los esfuerzos para estandarizar esta interpretación (Zhang et al., 2020).

2.3.3 SARS-CoV-2

El SARS-CoV-2 es un virus cuyas siglas en inglés significan Síndrome Respiratorio Agudo Severo CoronaVirus 2, el cual es el agente causante de la enfermedad por coronavirus 2019 (COVID-19) (Ludwig & Zarbock, 2020). El SARS-CoV-2 se propaga de persona a persona a través de secreciones nasales, gotas de saliva o aerosoles (Ludwig & Zarbock, 2020).

El coronavirus (CoV) es un virus envuelto con una membrana lípida derivada de la célula huésped, donde se incrustan proteínas de la superficie viral (Ludwig & Zarbock, 2020). Estas proteínas que sobresalen le dan la forma de un halo; lo que da lugar al nombre corona (Ludwig & Zarbock, 2020). El genoma del coronavirus está formado por una cadena de ácido ribonucleico (ARN) monocatenario con polaridad positiva, esta cadena es similar a la estructura de un ácido ribonucleico mensajero (ARNm) (Ludwig & Zarbock, 2020; Soto, 2020). Además, posee 6 proteínas codificadas por sus genes ORF3a, ORF6, ORF7a, ORF7b y ORF8 (Khailany et al., 2020).

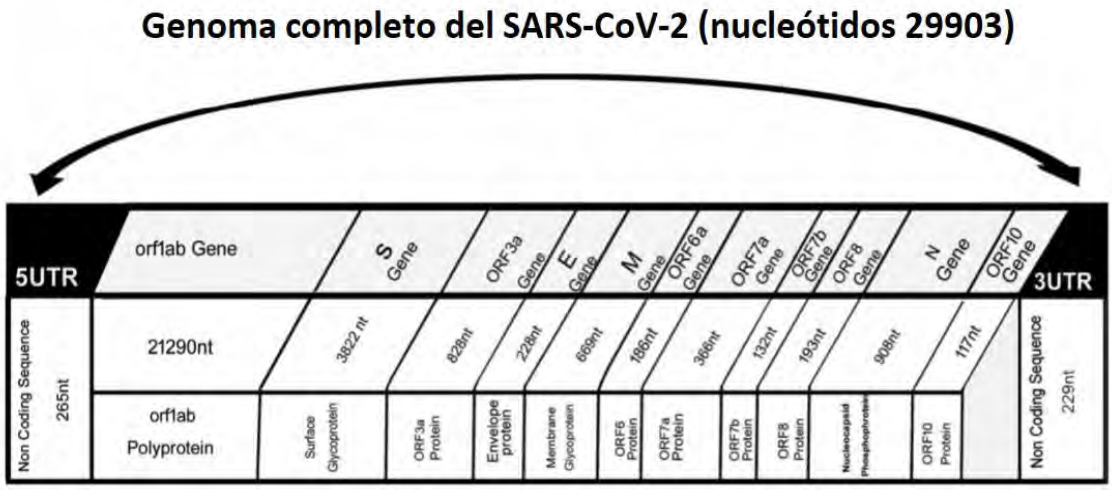


Figura 3. Estructura del virus del SARS-CoV-2.

Adaptado de (Khailany et al., 2020).

2.4 Conceptos de ciencias de la computación

2.4.1 Análisis de agrupamiento

El análisis de agrupamiento es una técnica en la cual se divide el conjunto de datos en subconjuntos mediante el agrupamiento de estos, usando medidas de semejanza (Diday & Simon, 1976; Han et al., 2012a). Las técnicas de agrupación representan los datos en grupos, de tal modo, que los datos que pertenecen a un mismo grupo son similares entre sí, y diferentes a los datos de otros grupos (Han et al., 2012b). Las similitudes y diferencias se evalúan, en la mayoría de casos, según una métrica de distancia pre-definida (Han et al., 2012b).

Las técnicas de agrupamiento se pueden clasificar en métodos de particionamiento, jerárquicos, basados en densidad y basados en cuadrículas (Han et al., 2012a).

- Los **métodos de particionamiento** construyen k grupos de particiones del conjunto de datos (Han et al., 2012a). Se divide el conjunto de datos de tal manera que cada dato pertenece exactamente a un grupo y cada grupo tiene como mínimo un dato (Han et al., 2012a).

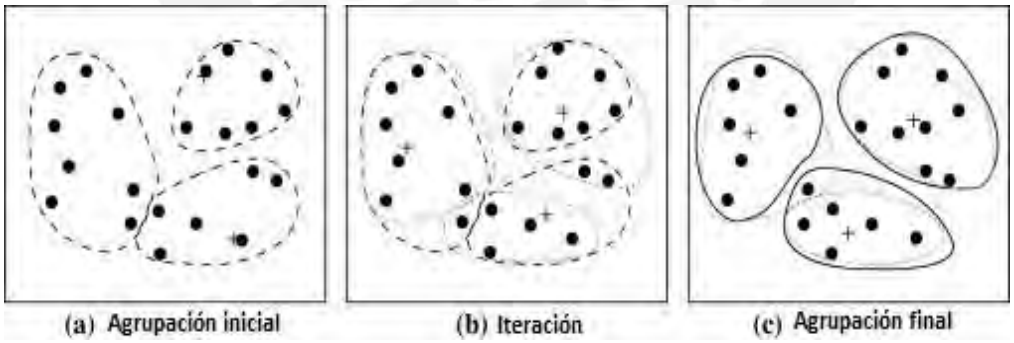


Figura 4. Agrupación de un conjunto de datos utilizando el método de particionamiento k-means. Adaptado de (Han, Kamber, & Pei, 2012a).

- Los **métodos jerárquicos** crean una descomposición jerárquica del conjunto de datos (Han et al., 2012a). Estos métodos se apoyan en la distancia o en la densidad y la continuidad; y, se caracterizan en que una vez realizado un nuevo grupo o separado alguno, no se puede deshacer ese cambio (Han et al., 2012a).

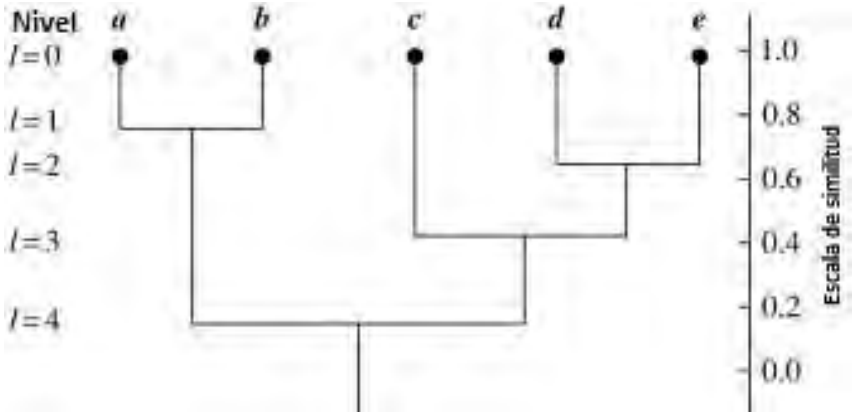


Figura 5. Representación dendrograma para un conjunto de datos utilizando el método jerárquico. Adaptado de (Han et al., 2012a).

- Los **métodos basados en densidad** consisten en la detención de zonas en donde no exista un mínimo de datos alrededor de otros grupos de datos (Han et al., 2012a). Este método sirve mucho para filtrar los valores atípicos (Han et al., 2012a).

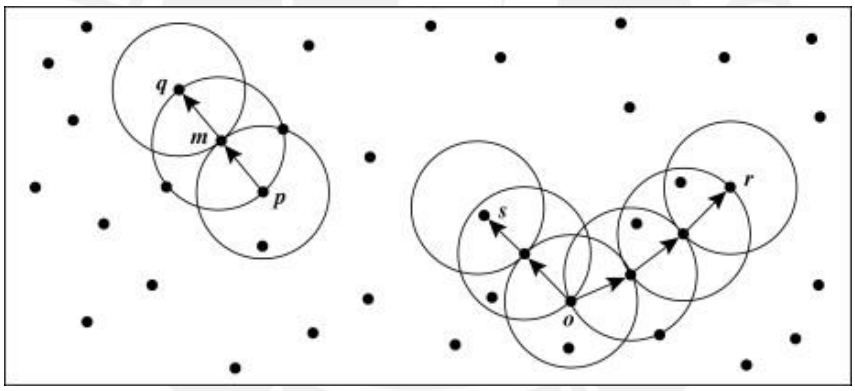


Figura 6. Accesibilidad de densidad y conectividad de densidad en el método basados en densidad (Han et al., 2012a).

- Los **métodos basados en cuadrículas** forman una estructura de cuadrícula, un espacio cuantificado (Han et al., 2012a). El tiempo de procesamiento al usar este método es rápido; asimismo, ayuda a muchos problemas relacionados con el análisis espacial (Han et al., 2012a).

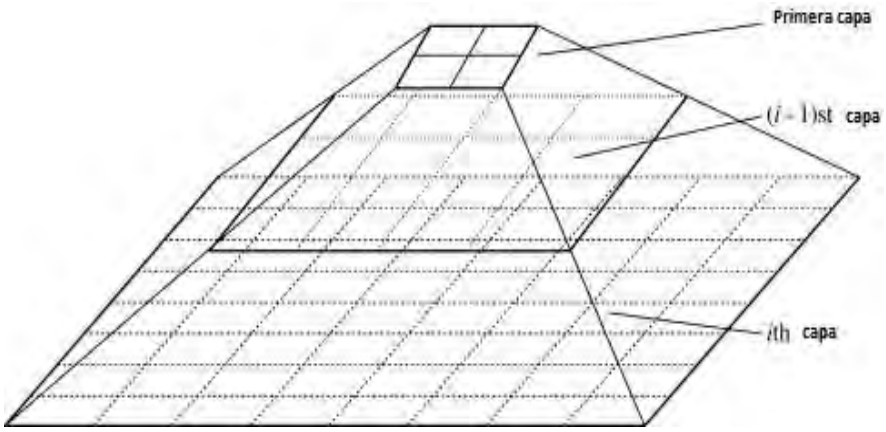


Figura 7. Representación jerárquica del método basado en cuadrícula STING.

Adaptado de (Han et al., 2012a).

2.4.2 Analítica visual

La analítica visual es la ciencia que facilita el razonamiento analítico mediante interfaces visuales interactivas (Thomas & Cook, 2005). El objetivo es proporcionar este proceso de razonamiento mediante una herramienta de software, que ayude a maximizar la capacidad de comprender y razonar sobre información y datos complejos de manera dinámica (Thomas & Cook, 2005). Esta a su vez, presenta las siguientes áreas de enfoque: técnicas de razonamiento analítico, representaciones visuales y técnicas de interacción, representaciones y transformaciones de datos complejos en formas de visualización; y, técnicas que sirven para apoyar la producción, presentación y divulgación de los resultados (Thomas & Cook, 2005).

Capítulo 3. Estado del Arte

3.1 Introducción

En este capítulo se presenta el estado del arte, el cual tiene como fin realizar una revisión sistemática mediante los objetivos de revisión, el establecimiento de las preguntas de revisión para dar a conocer cómo y qué métodos de aprendizaje no supervisado son los más utilizados, para realizar análisis de agrupamiento de las variantes de secuencias genómicas del SARS-Cov-2 en otros países.

3.2 Objetivos de revisión

El objetivo general de la presente revisión sistemática es obtener conocimiento sobre los métodos y técnicas no supervisadas que se vienen usando para el análisis de secuencias genómicas de variantes del nuevo coronavirus SARS-CoV-2. Dada la gran cantidad y variedad de artículos e investigaciones existentes relacionadas al SARS-CoV-2 se ha restringido la búsqueda a estudios enfocados en este microorganismo. El tipo de revisión a realizar es del estado del arte. Los objetivos específicos de esta revisión son:

- Conocer los principales métodos o algoritmos no supervisados aplicados en el análisis de secuencias genómicas SARS-CoV-2.
- Conocer los resultados alcanzados, las ventajas y limitaciones de los métodos no supervisados aplicados en el análisis de secuencias SARS-CoV-2.
- Conocer cómo se validan o evalúan los métodos de agrupamiento en el análisis de secuencias genómicas SARS-CoV-2.
- Conocer las métricas de distancia más usadas para realizar el respectivo agrupamiento de secuencias genómicas SARS-CoV-2.
- Conocer qué estrategias de representación de las secuencias genómicas son usadas para realizar el análisis de agrupamiento de variantes SARS-CoV-2.

- Conocer cómo se muestra el resultado de los algoritmos de agrupamiento en el análisis de secuencias genómicas SARS-CoV-2.

3.3 Preguntas de revisión

- P1. ¿Qué métodos de análisis no supervisados se han usado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2, qué resultados, ventajas y limitantes se han obtenido y cómo son evaluados?
- P2. ¿Qué métricas de distancia son las más usadas en los métodos o algoritmos de agrupamiento para secuencias genómicas SARS-CoV-2?
- P3. ¿Cuáles son las estrategias de representación genómica aplicadas en el análisis de secuencias genómicas SARS-CoV-2?
- P4. ¿Qué paradigmas o formas de visualización de resultados de agrupamiento se han usado en el análisis de secuencias genómicas SARS-CoV-2?

3.4 Estrategia de búsqueda

3.4.1 Motores de búsqueda a usar

En esta investigación se utilizan los motores de búsqueda de Scopus y Web of Science, debido a que son bases de datos que indexan material científico publicado en revistas y congresos internacionales de reconocida calidad.

3.4.2 Cadenas de búsqueda a usar

En la construcción de la cadena de búsqueda primero se identifican los términos o palabras claves para la revisión sistemática. Estas se dividen de la siguiente manera:

- Términos claves relacionados a la secuencia genómica de estudio: Sars-Cov-2, SARSCOV2, Covid-19, Covid19, Covid y Sars-Cov2.
- Términos asociados a métodos no supervisados: *clustering*, *cluster* y *unsupervised learning*.

- Términos relacionados a biología molecular: *genomics*, *genomic sequences*, *genome*, *genotype* y *variants*.

Para responder las preguntas de revisión anteriormente mencionadas, se utilizará la siguiente cadena conformada por las palabras claves descritas buscadas en el tópico del artículo (título, resumen y palabras clave):

Cadena en Scopus: TITLE-ABS-KEY(("Sars-Cov-2" OR "SARSCOV2" OR "Covid-19" OR "Covid19" OR "Covid" OR “Sars-Cov2”) AND ("clustering" OR "cluster" OR "unsupervised learning") AND ("genomics" OR "genomic sequences" OR “genome” OR "genotype" OR “variants”))

Cadena en Web of Science: TS=("Sars-Cov-2" OR "SARSCOV2" OR "Covid-19" OR "Covid19" OR "Covid" OR “Sars-Cov2”) AND TS=("clustering" OR "cluster" OR "unsupervised learning") AND TS=("genomics" OR "genomic sequences" OR “genome” OR "genotype" OR "variants")

3.4.3 Documentos encontrados

Luego de realizar la búsqueda por cadena, en ambos motores de búsqueda se obtuvieron artículos e investigaciones relevantes para el proyecto. En la Tabla 8, se muestra la cantidad de documentos obtenidos en cada motor de búsqueda.

Tabla 8. Documentos encontrados por motor de búsqueda

Scopus	Web of Science	Total
198	71	269

Nota. Elaboración propia.

3.4.4 Criterios de inclusión/exclusión

Los artículos encontrados en la revisión sistemática pasaron por un análisis de inclusión y exclusión con el fin de considerar aquellos artículos que son de utilidad para responder las preguntas de revisión.

Los criterios de inclusión utilizados para la revisión sistemática fueron los siguientes:

1. El estudio detalla los métodos o algoritmos de agrupamiento en el análisis de secuencias genómicas SARS-CoV-2.
2. El estudio pertenece a las áreas: ciencias de la computación, biología computacional, matemática, ingeniería, aplicaciones interdisciplinarias de la informática, ciencia de la decisión, inteligencia artificial informática, genética y ciencias multidisciplinarias. Estas áreas son consideradas porque están relacionadas y contienen los temas necesarios para la revisión sistemática.
3. El estudio está escrito en español o inglés. Siendo el inglés el idioma predominante en el que están escritos la mayoría de artículos e investigaciones.

Además, no se considera un límite en cuanto a la fecha de publicación de los estudios debido a que los estudios relacionados al SARS-CoV-2 fueron publicados a partir del año 2020.

Finalmente, los estudios que se excluyeron cumplían con el siguiente criterio:

1. El estudio no utiliza métodos no supervisados o análisis de agrupamientos para analizar secuencias genómicas de SARS-CoV-2. Dado que no estaría relacionado al tema de estudio.

En la Tabla 9, se presenta la cantidad de artículos obtenidos luego de aplicar los criterios de inclusión y exclusión en cada motor de búsqueda.

Tabla 9. Documentos obtenidos luego de aplicar los criterios de inclusión/exclusión en cada motor de búsqueda

Scopus	Web of Science
11	8

Nota. Elaboración propia.

Finalmente, se obtuvieron 14 artículos sin incluir repeticiones.

3.4.5 Documentos revisados

Luego de aplicar los criterios de inclusión y exclusión, se obtuvieron 14 artículos para la revisión sistemática. En la Tabla 10, se presenta la lista de los 14 artículos revisados con su referencia APA.

Tabla 10. Documentos seleccionados para la revisión sistemática

N° de Artículo	Título Artículo	Referencia
1°	*Non-standard bioinformatics characterization of SARS-CoV-2*	Bielińska-Wąż, D., & Wąż, P. (2021). Non-standard bioinformatics characterization of SARS-CoV-2. Computers in Biology and Medicine, 131, 104247. https://doi.org/10.1016/j.compbiomed.2021.104247
2°	*Identification and computational analysis of mutations in SARS-CoV-2*	Dey, T., Chatterjee, S., Manna, S., Nandy, A., & Basak, S. C. (2021). Identification and computational analysis of mutations in SARS-CoV-2. Computers in Biology and Medicine, 129, 104166. https://doi.org/10.1016/j.compbiomed.2020.104166
3°	*Genomic characterization and phylogenetic analysis of the first SARS-CoV-2 variants introduced in Lebanon*	Feghali, R., Merhi, G., Kwasiborski, A., Hourdel, V., Ghosn, N., & Tokajian, S. (2021). Genomic characterization and phylogenetic analysis of the first SARS-CoV-2 variants introduced in Lebanon. PeerJ, 9, 19. https://doi.org/10.7717/peerj.11015
4°	*Unsupervised cluster analysis of SARS-CoV-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of SARS-CoV-2 virus*	Hahn, G., Lee, S., Weiss, S. T., & Lange, C. (2021). Unsupervised cluster analysis of SARS-CoV-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of SARS-CoV-2 virus. https://doi.org/10.1002/gepi.22373
5°	*UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets*	Hozumi, Y., Wang, R., Yin, C., & Wei, G. W. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. Computers in Biology and Medicine, 131, 104264. https://doi.org/10.1016/j.compbiomed.2021.104264
6°	*Geospatial dynamics of COVID-19 clusters and hotspots in Bangladesh*	Islam, A., Sayeed, M. A., Rahman, M. K., Ferdous, J., Shariful Islam, |, Mohammad, |, & Hassan, M. (2021). Geospatial dynamics of COVID-19 clusters

		and hotspots in Bangladesh. Transbound Emerg Dis, 00, 1–15. https://doi.org/10.1111/tbed.13973
7°	*The global population of SARS-CoV-2 is composed of six major subtypes*	Morais, I. J., Polveiro, R. C., Medeiros Souza, G., Bortolin, D. I., Sassaki, T., Talis, A., & Lima, M. (2020). The global population of SARS-CoV-2 is composed of six major subtypes. Scientific Reports. https://doi.org/10.1038/s41598-020-74050-8
8°	*In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh*	Shishir, T. A., Bin, I., Id, N., & Faruque, S. M. (2021). In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. PLOS ONE. https://doi.org/10.1371/journal.pone.0245584
9°	*SARS-CoV-2 genomic variations associated with mortality rate of COVID-19*	Toyoshima, Y., Nemoto, • Kensaku, Matsumoto, S., Nakamura, Y., & Kiyotani, K. (2020). SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. Journal of Human Genetics, 65, 1075–1082. https://doi.org/10.1038/s10038-020-0808-9
10°	*Emergence of novel SARS-CoV-2 variants in the Netherlands*	Urhan, A., & Abeel, T. (2021). Emergence of novel SARS-CoV-2 variants in the Netherlands. Scientific Reports |, 11, 6625. https://doi.org/10.1038/s41598-021-85363-7
11°	*Principal Component Analysis Applications in COVID-19 Genome Sequence Studies*	Wang, B., & Jiang, L. (2021). Principal Component Analysis Applications in COVID-19 Genome Sequence Studies. Cognitive Computation, 1, 3. https://doi.org/10.1007/s12559-020-09790-w
12°	*Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants*	Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., & Wei, G.-W. (2021). Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. COMMUNICATIONS BIOLOGY, 4. https://doi.org/10.1038/s42003-021-01754-6
13°	*Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations*	Yang, H.-C., Chen, C.-H., Wang, J.-H., Liao, H.-C., Yang, C.-T., Chen, C.-W., … Liao, J. C. (2020). Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.2007840117/-/DCSupplemental

14°	*Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization*	Zhao Id, Z., Sokhansanj Id, B. A., Malhotra Id, C., Zheng, K., & Rosenid, G. L. (2020). Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. Plos Computational Biology. https://doi.org/10.1371/journal.pcbi.1008269

Nota. Elaboración propia.

3.5 Formulario de extracción de datos

En la Tabla 11, se presentan los campos del formulario de extracción de información para los artículos encontrados en la revisión sistemática.

Tabla 11. Campos del formulario de extracción de información

Campo	Descripción	Pregunta
N° de Artículo	Número identificador del artículo	General
Año de publicación	Año de publicación del artículo	General
Título Artículo	Título del artículo	General
Autores	Autor o autores del artículo	General
Nombre de la Fuente	Nombre de la revista, congreso, libro o tesis	General
Resumen	Breve resumen del artículo	General
Tipo de análisis	Tipo de análisis realizado	General
Algoritmo(s) usado(s)	Qué métodos de análisis no supervisados se han usado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2	P1
Métricas de distancia usada	Qué métricas de distancia son usadas en los métodos o algoritmos de agrupamiento para secuencias genómicas SARS-CoV-2	P2
Estrategia de reducción de dimensionalidad	Cuáles son las estrategias de representación genómica aplicadas en el análisis de secuencias genómicas SARS-CoV-2	P3

Datos analizados	Los datos usados para el análisis del modelo	General
País de análisis	País donde fue hecho el análisis	General
Fuente de datos	Fuente de donde se obtuvieron los datos para el respectivo análisis	General
Resultados obtenidos	Resultados obtenidos al utilizar el método de análisis no supervisado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2	P1
Limitaciones del modelo	Limitantes que presenta el método no supervisado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2	P1
Ventajas del modelo	Ventajas de emplear el método de análisis no supervisado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2	P1
Paradigmas o formas de visualización de resultados	Paradigmas o formas en las que se visualiza el resultado de los algoritmos de agrupamiento que se han usado en el análisis de secuencias genómicas SARS-CoV-2	P4
¿Cómo se validan o evalúan los métodos?	Se detalla cómo se validan o evalúan los métodos de análisis no supervisados	P1
¿Tiene código disponible?	Pregunta cerrada para saber si el artículo presenta o tiene código disponible	General
Comentarios	Comentarios personales sobre el artículo revisado	General

Nota. Elaboración propia.

3.6 Resultados de la revisión

3.6.1 Respuesta a pregunta: ¿Qué métodos de análisis no supervisados se han usado para realizar el agrupamiento de secuencias genómicas SARS-CoV-2, qué resultados, ventajas y limitantes se han obtenido y cómo son evaluados?

De la revisión sistemática realizada, se puede observar que los estudios utilizan distintos métodos de análisis no supervisados para realizar el agrupamiento de secuencias genómicas SARS-CoV-2. Entre estos se encuentran el agrupamiento por ISM (marcadores informativos

de subtipos) (Zhao Id et al., 2020) *K-means* (Hozumi et al., 2021; R. Wang et al., 2021) agrupación jerárquica (Toyoshima et al., 2020; H.-C. Yang et al., 2020), agrupamiento sin modelo que compara las secuencias genómicas a nivel de todo el genoma (Hahn et al., 2021) agrupamiento espacial mediante las estadísticas de Local Moran's I y el agrupamiento mediante las estadísticas de exploración de espacio-tiempo mediante un modelo de Poisson discreto (Islam et al., 2021).

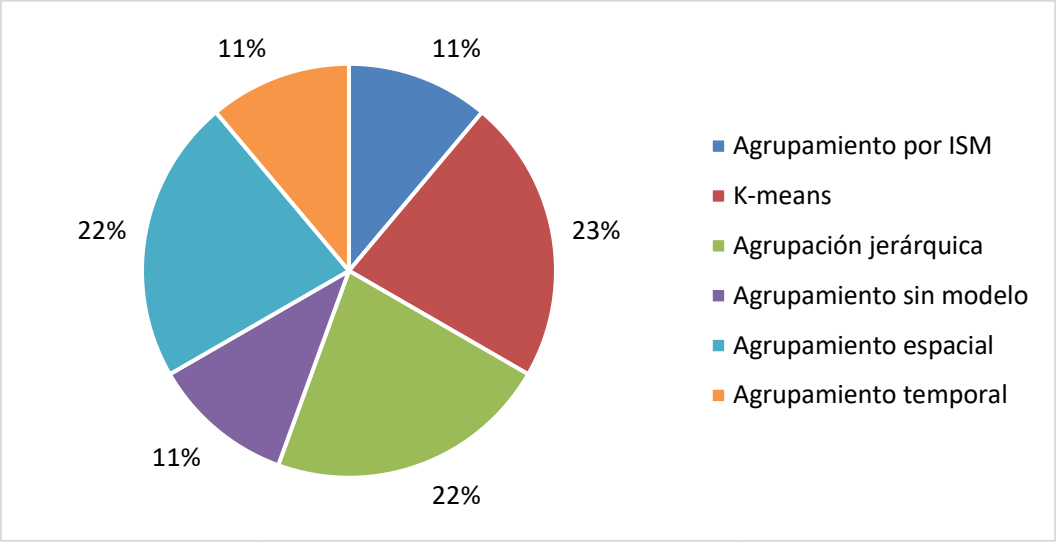


Figura 8. Distribución de los métodos no supervisados usados en 9 artículos.

En la Figura 8, se presenta la distribución de los métodos no supervisados utilizados en los estudios revisados. Se observa que en un 67% (6 artículos) se utilizaron 3 métodos, *k-means*, agrupación jerárquica y agrupamiento espacial, con un 23%, 22%, 22% respectivamente.

En cuanto al método *k-means*, el cual es uno de los métodos de análisis no supervisados más populares (Hozumi et al., 2021). Para evaluar cuantitativamente el rendimiento y la precisión del agrupamiento se modifican los problemas de clasificación con etiquetas, en problemas de agrupación de *k-means*, estableciendo el número de grupos (k) igual al número de categoría reales; para finalmente, calcular la precisión de *k-means* para todo el conjunto de datos (Hozumi et al., 2021); además, el rendimiento de *k-means* depende de la selección de la métrica de distancia (R. Wang et al., 2021).

Con respecto a los resultados y ventajas de *k-means*, se tiene como principal ventaja la obtención de mejores resultados al tener una mayor cantidad de datos (Hozumi et al., 2021). Sin embargo, la limitante es que *k-means* calcula la distancia entre los centroides de cada grupo hacía cada muestra, y al tener una gran cantidad de secuencias genómicas SARS-CoV-2 el cálculo se hace muy costoso al necesitar una gran cantidad de memoria para realizar el agrupamiento (Hozumi et al., 2021). Como resultado se obtiene que *k-means* presenta una mejor agrupación y rendimiento cuando se utiliza algoritmos de reducción de dimensionalidad (Hozumi et al., 2021).

Seguidamente se menciona el método agrupamiento jerárquico, se puede apreciar que es uno de los métodos más usados para identificar y agrupar las secuencias genómicas SARS-CoV-2. Con respecto a los resultados, el agrupamiento jerárquico clasificó a 28 países en tres grupos (Toyoshima et al., 2020). En el primer grupo se incluye a la mayoría de países asiáticos, mientras que en el segundo grupo se encuentran países europeos y sudamericanos; y, el tercer grupo incluye países de Europa, Norteamérica, Oceanía, África y algunos países de Asia (Toyoshima et al., 2020).

Por otro lado, el agrupamiento espacial mediante las estadísticas de Local Moran's I se usó para identificar la autocorrelación espacial de las secuencias genómicas SARS-CoV-2 (Islam et al., 2021). Asimismo, se realizó el agrupamiento espacial mediante las estadísticas de exploración espacio-tiempo mediante un modelo de Poisson discreto (Islam et al., 2021). Como resultado, se obtuvo una fuerte autocorrelación espacial de las secuencias genómicas en Bangladesh durante los primeros 5 meses, de marzo a julio de 2020 (Islam et al., 2021).

Los métodos de agrupamiento por ISM, agrupamiento sin modelo y agrupamiento temporal fueron menos usados en los estudios revisados. Se conoce que el ISM define un conjunto de sitios de nucleótidos que representan las posiciones más variables en las secuencias genómicas (Zhao Id et al., 2020); por ello, se realiza un agrupamiento por ISM. Este método

tiene como ventaja que se puede usar para la visualización espacio-temporal de la variación genética del SARS-CoV-2; así como también, es más eficiente computacionalmente y escalable a medida que se tiene una mayor cantidad de secuencias genómicas SARS-CoV-2 (Zhao Id et al., 2020).

En el agrupamiento sin modelo se aplica el análisis de componentes principales (PCA, por sus siglas en inglés) a una matriz de similaridad en la cual se compara todos los pares de las secuencias genómicas SARS-CoV-2 utilizando el índice de Jaccard (Hahn et al., 2021). Con dicho método se identificó cuatro subgrupos genéticos diferentes en Europa y Estados Unidos (Hahn et al., 2021). La principal limitante de este método es que no permite secuencias genómicas con información faltante (Hahn et al., 2021).

El agrupamiento temporal de las variantes del virus se realiza mediante estadísticas de exploración de espacio-tiempo usando un modelo de Poisson discreto (Islam et al., 2021). Las estadísticas espacio-temporales son métodos de suma importancia para identificar la dinámica de transmisión del SARS-CoV-2 (Islam et al., 2021). Con este método se identificó que los casos en Bangladesh incrementaron gradualmente de abril a julio de 2020; así como también se identificaron doce grupos estadísticamente significativos (Islam et al., 2021).

En resumen, no todos los métodos de análisis no supervisados descritos en los artículos revisados describen sus ventajas, limitantes o cómo son validados o evaluados. Sin embargo, en todos los artículos revisados se pudo apreciar resultados relevantes al aplicar métodos de agrupamientos en las secuencias genómicas SARS-CoV-2 analizadas.

3.6.2 Respuesta a pregunta: ¿Qué métricas de distancia son las más usadas en los métodos o algoritmos de agrupamiento para secuencias genómicas SARS-CoV-2?

Entre las métricas de distancia más usadas en los estudios revisados se encuentran, la distancia de Hamming (Hahn et al., 2021; Zhao Id et al., 2020), el índice global de Moran (Islam et al., 2021) y la distancia Jaccard (Hozumi et al., 2021; R. Wang et al., 2021).

En cuanto a la distancia de Hamming, es una métrica de distancia que mide la divergencia de cada secuencia genómica a una secuencia de referencia (Hahn et al., 2021; Zhao Id et al., 2020).

En lo que se refiere al índice global de Moran, es una medida estadística que analiza e indica la existencia de una autocorrelación espacial, los valores van de -1 a +1, donde -1 significa una agrupación de valores diferentes, +1 agrupación de valores similares y 0 significa que no hay autocorrelación espacial (Islam et al., 2021).

Para finalizar, la distancia Jaccard mide la diferencia entre las secuencias del genoma SARS-CoV-2 ya que ofrece una diferencia filogenética o topológica entre las muestras (Hozumi et al., 2021). La distancia Jaccard se calcula restando el índice Jaccard de 1 (Hozumi et al., 2021). Al emplear esta métrica de distancia en los métodos de agrupamiento, ayuda a mejorar el rendimiento de la agrupación (Hozumi et al., 2021).

3.6.3 Respuesta a pregunta: ¿Cuáles son las estrategias de representación genómica aplicadas en el análisis de secuencias genómicas SARS-CoV-2?

En los estudios se observan diversas estrategias de representación de secuencias genómicas; como, el análisis de componentes principales (PCA, por sus siglas en inglés) (Hahn et al., 2021; Hozumi et al., 2021; B. Wang & Jiang, 2021; Zhao Id et al., 2020), alineamiento de secuencias genómicas múltiples (Dey et al., 2021; Feghali et al., 2021; Hozumi et al., 2021; Shishir et al., 2021; Toyoshima et al., 2020; Urhan & Abeel, 2021; B. Wang & Jiang, 2021; R. Wang et al., 2021; Zhao Id et al., 2020), algoritmo de corrección de errores (Zhao Id et al., 2020), UMAP (*Uniform Manifold Approximation and Projection*) (Hozumi et al., 2021), Incrustación de vecinos estocástica distribuida (t-SNE, por sus siglas en inglés) (Hozumi et al., 2021) y el análisis de variación de secuencias (Morais et al., 2020).

El análisis de componentes principales (PCA, por sus siglas en inglés) es una de las técnicas más populares para la reducción de la dimensionalidad (Hozumi et al., 2021; B. Wang

& Jiang, 2021), el cual encuentra las direcciones que maximizan la varianza del conjunto de datos, siendo éstas proyectadas en un nuevo espacio para así obtener una transformación del conjunto de datos con menor dimensionalidad, pero manteniendo la varianza (Hozumi et al., 2021). La ventaja de este método es que puede trabajar con grandes números de características, como es el caso de las secuencias genómicas SARS-CoV-2 (B. Wang & Jiang, 2021). Una de sus limitantes es que, si bien logra cubrir la máxima varianza entre las características de los datos, se puede perder información si se elige una cantidad equivocada de componentes principales (Hozumi et al., 2021).

Otra estrategia de representación genómica, es el alineamiento de secuencias genómicas múltiples con una secuencia de referencia en donde se ven los nucleótidos con similitud o variación, así como también, sirve para identificar los puntos de mutación (Dey et al., 2021). La secuencia de referencia usada en algunos de los estudios es la secuencia SARS-CoV-2_Wuhan-Hu-1 (Feghali et al., 2021; Toyoshima et al., 2020; Urhan & Abeel, 2021).

El algoritmo de corrección de errores es un método para resolver y corregir la ambigüedad que presentan algunos nucleótidos en cierta posición (Zhao Id et al., 2020). Este método reemplaza los marcadores informativos de subtipos (ISM, por sus siglas en inglés) con una base ambigua en una posición por otro ISM que no presente error en esa posición (Zhao Id et al., 2020).

UMAP (*Uniform Manifold Approximation and Projection*) es un método de reducción dimensional no lineal eficiente y adecuado, que ayuda al algoritmo *k-means* en el agrupamiento de grandes cantidades de datos (Hozumi et al., 2021). UMAP está basado en gráficos, tiene como objetivo crear una representación gráfica ponderada k-dimensional ya definida de cada dato de alta dimensión (Hozumi et al., 2021).

El método de incrustación de vecinos estocástica distribuida (t-SNE, por sus siglas en inglés), es utilizado para reducir la alta dimensión de los datos a un espacio bidimensional o

tridimensional (Hozumi et al., 2021). En este método, primero se calcula la distribución de probabilidad por pares de datos, asignando una alta probabilidad a los pares de datos cercanos; luego, se define una distribución de probabilidad en el espacio incrustado similar al espacio original de datos, con el objetivo de reducir la divergencia de Kullback-Leibler (KL) entre los datos (Hozumi et al., 2021). Finalmente, el análisis de variación de secuencias permite detectar regiones genómicas con mayor varianza genética de posiciones que albergan dos o más bases de nucleótidos distintos (Morais et al., 2020).

3.6.4 Respuesta a pregunta: ¿Qué paradigmas o formas de visualización de resultados de agrupamiento se han usado en el análisis de secuencias genómicas SARS-CoV-2?

Entre los principales paradigmas o formas de visualización de resultados están: los mapas de densidad de puntos (Islam et al., 2021), los gráficos temporales (Dey et al., 2021), el gráfico polar 2D (Dey et al., 2021), el gráfico de puntuación de PCA (B. Wang & Jiang, 2021), los gráficos de mapas de similitud en espacios 2D y 3D (Bielińska-Wąż & Wąż, 2021); y, la gráfica de asociación generalizada (GAP, por sus siglas en inglés) (H.-C. Yang et al., 2020).

Los mapas de densidad de puntos se utilizan para visualizar como cambia la distribución espacial en el tiempo (Islam et al., 2021). Con respecto a los gráficos temporales, estos se han usado para entender cómo se ha ido propagando el SARS-CoV-2 (Dey et al., 2021). También, se ha utilizado la gráfica polar 2D para representar secuencias en un sistema de coordenadas polares bidimensionales y evidenciar diferentes grupos mutados en las proteínas del SARS-CoV-2 (Dey et al., 2021).

En lo que se refiere al gráfico de puntuación de PCA (análisis de componentes principales), este se aplicó para separar las muestras humanas de las muestras animales (B. Wang & Jiang, 2021).

Los gráficos de mapas de similitud en espacios 2D y 3D se usaron para mostrar las similitudes o diferencias de las secuencias genómicas del SARS-CoV-2, agrupando puntos que representan grupos particulares de secuencias (Bielińska-Wąż & Wąż, 2021).

En el gráfico de asociación generalizada (GAP, por sus siglas en inglés) se visualiza el agrupamiento y variación de las secuencias genómicas; así como también, identifica los valores atípicos en las variaciones (H.-C. Yang et al., 2020).

3.7 Conclusiones

Según los estudios encontrados en esta revisión sistemática, se puede concluir que se utilizan varios métodos para el agrupamiento de grandes cantidades de datos, como es el caso de las secuencias genómicas del SARS-CoV-2; por ello, al usar estos métodos se obtienen mejores resultados (Hozumi et al., 2021; Islam et al., 2021; Toyoshima et al., 2020; R. Wang et al., 2021; H.-C. Yang et al., 2020; Zhao Id et al., 2020). Así como también, se ha encontrado que otros estudios utilizan herramientas empaquetadas, donde no se obtuvieron resultados precisos sobre la propagación del virus en los diferentes países de estudio.

Con respecto a las estrategias de representación genómica, la mayoría de los estudios utilizan uno o más estrategias para reducir la dimensión de las secuencias genómicas (Dey et al., 2021; Feghali et al., 2021; Hahn et al., 2021; Hozumi et al., 2021; Shishir et al., 2021; Toyoshima et al., 2020; Urhan & Abeel, 2021; B. Wang & Jiang, 2021; R. Wang et al., 2021; Zhao Id et al., 2020), y otros reemplazan algunos nucleótidos faltantes o que representen alguna ambigüedad (Zhao Id et al., 2020). Por último, se utilizan diferentes formas de visualización de los resultados.

Capítulo 4. Módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú

4.1 Introducción

En el presente capítulo, se presentan los resultados alcanzados para lograr el cumplimento del primer objetivo específico, el cual es "Desarrollar un módulo de software que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2 recolectadas en Perú". A continuación, se detalla los resultados alcanzados los cuales son: la estructura de la base de datos de los módulos de software, el módulo de software para realizar el pre-procesamiento y la reducción dimensional de los datos de las secuencias genómicas SARS-CoV-2; y, el módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú. Además, se presentan los medios de verificación y los indicadores objetivamente verificables con los cuales se va a validar el cumplimiento de cada resultado alcanzado.

4.2 Resultados alcanzados

4.2.1 Estructura de la base de datos de los módulos de software a desarrollar.

En este resultado se aprecia la estructura de la base de datos para la herramienta analítica interactiva a desarrollar en el presente proyecto de investigación. Para lograr esto, se propone como medios de verificación, un documento donde se describa el modelo físico de la base de datos y el script de creación de la misma.

A continuación, se detalla el modelo físico de la base de datos para la herramienta analítica interactiva a desarrollar en el presente proyecto de investigación.

Para realizar la estructura de la base de datos se ha utilizado la herramienta de base de datos PostgreSQL. En la Figura 9, se presenta el modelo físico de base de datos realizado, en el que se incluyen las relaciones y restricciones para almacenar y acceder a los datos; y, la especificación de llaves primarias y foráneas.

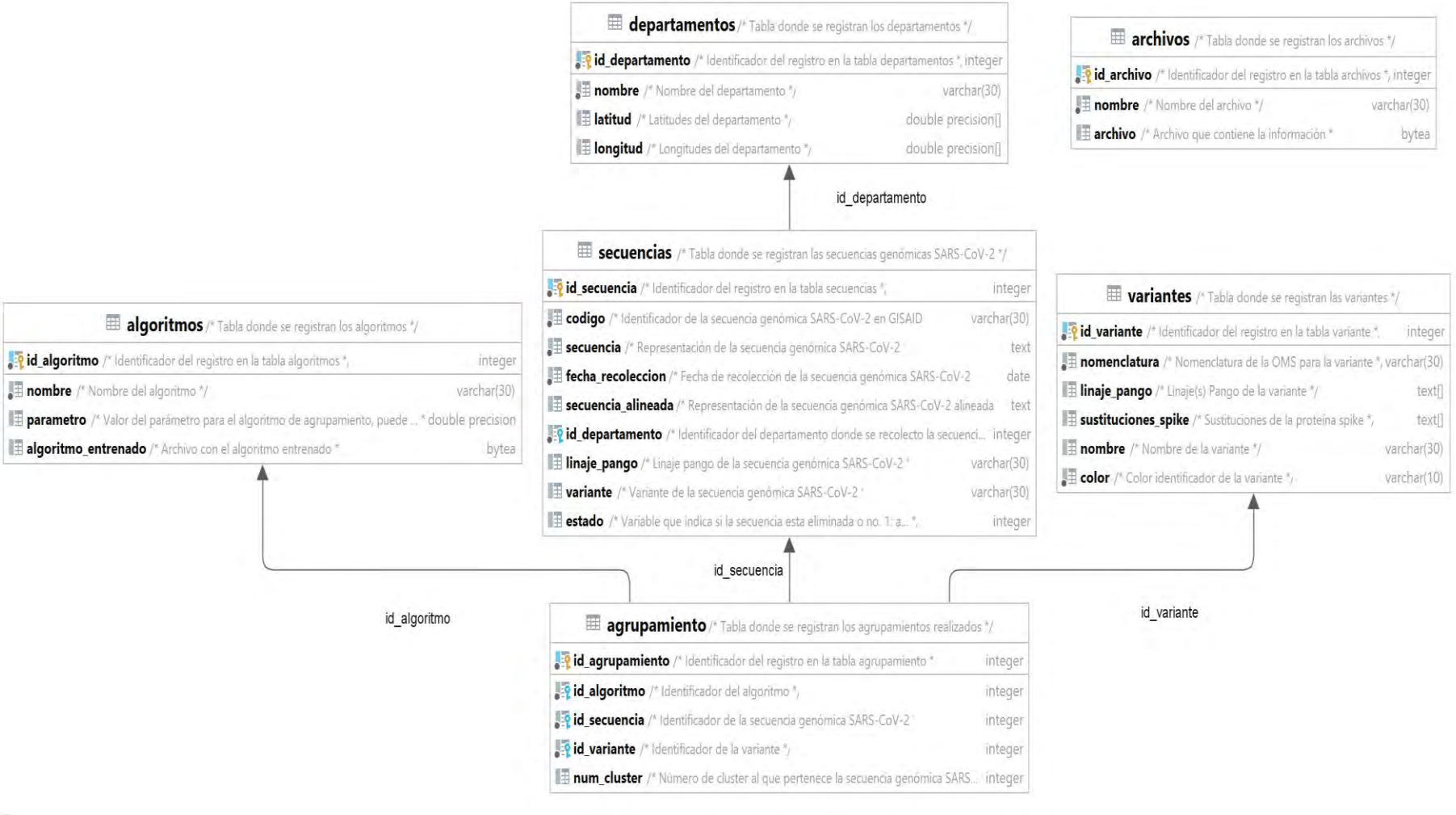


Figura 9. Modelo físico de la base de datos.

Para definir la estructura se consideró el propósito de la base de datos, la cual es organizar y almacenar la información, proporcionándole de una forma más sencilla para la herramienta a desarrollar. Asimismo, se tomó de referencia la información que muestra la herramienta analítica Nextstrain, ya que esta herramienta ayuda a la comprensión epidemiológica del virus SARS-CoV-2 a nivel mundial (Nexstrain, s.f.).

Teniendo en cuenta lo anterior, se identificó las siguientes tablas: Departamentos, Secuencias, Variantes, Algoritmos, Archivos y Agrupamiento. Se creó, la tabla Departamentos, para almacenar los nombres y las coordenadas geográficas de los departamentos del Perú; la tabla Secuencias, donde se almacena la información de las secuencias genómicas SARS-CoV-2 como su identificador en GISAID, su linaje pango, la fecha en que se recolectó y la representación de la secuencia; la tabla Variantes, para las variantes identificadas por la OMS (Organización Mundial de la Salud); la tabla Algoritmos, donde se almacena la información de los algoritmos de agrupamiento a utilizar en el módulo de agrupamiento; y, la tabla Archivos, para guardar los archivos del análisis de escalamiento multidimensional (MDS), el análisis de componentes principales (PCA) y la matriz de distancia de las secuencias genómicas SARS-CoV-2. Por último, se define la tabla Agrupamiento, la cual contiene la relación muchos a muchos de las tablas secuencias, variantes y algoritmos; así como también, contiene el número de clúster al que pertenece la secuencias genómica SARS-CoV-2 luego de realizado el modelo de agrupamiento.

A continuación, se presenta el diccionario de datos del modelo físico de la base de datos, para un mejor entendimiento de lo descrito anteriormente.

Tabla 12. Diccionario de datos de la tabla Departamentos.

Nombre de la tabla		**Departamentos**			
Descripción		Tabla donde se registran los departamentos			
Nombre del campo	**Tipo de dato**	**No Nulo**	**PK (Llave primaria)**	**FK (Llave foránea)**	**Descripción**
id_departamento	*integer*	Sí	Sí	No	Identificador del registro en la tabla departamentos
nombre	*varchar*(30)	Sí	No	No	Nombre del departamento
latitud	*double precision*[]	No	No	No	Latitudes del departamento
longitud	*double precision*[]	No	No	No	Longitudes del departamento

Nota. Elaboración propia.

Tabla 13. Diccionario de datos de la tabla Secuencias.

Nombre de la tabla		**Secuencias**			
Descripción		Tabla donde se registran las secuencias genómicas SARS-CoV-2			
Nombre del campo	**Tipo de dato**	**No Nulo**	**PK (Llave primaria)**	**FK (Llave foránea)**	**Descripción**
id_secuencia	*integer*	Sí	Sí	No	Identificador del registro en la tabla secuencias
codigo	*varchar*(30)	Sí	No	No	Identificador de la secuencia genómica SARS-CoV-2 en GISAID

secuencia	*text*	Sí	No	No	Representación de la secuencia genómica SARS-CoV-2
fecha_recoleccion	*date*	Sí	No	No	Fecha de recolección de la secuencia genómica SARS-CoV-2
secuencia_alineada	*text*	No	No	No	Representación de la secuencia genómica SARS-CoV-2 alineada
id_departamento	*integer*	Sí	No	Sí	Identificador del departamento donde se recolectó la secuencia genómica SARS-CoV-2
linaje_pango	*varchar*(30)	No	No	No	Linaje pango de la secuencia genómica SARS-CoV-2
variante	*varchar*(30)	No	No	No	Variante de la secuencia genómica SARS-CoV-2
estado	*integer*	No	No	No	Variable que indica si la secuencia esta eliminada o no. 1: activo 0: eliminado

Nota. Elaboración propia.

Tabla 14. Diccionario de datos de la tabla Variantes.

Nombre de la tabla		**Variantes**			
Descripción		Tabla donde se registran las variantes			
Nombre del campo	**Tipo de dato**	**No Nulo**	**PK (Llave primaria)**	**FK (Llave foránea)**	**Descripción**
id_variante	*integer*	Sí	Sí	No	Identificador del registro en la tabla variante

Nomenclatura	*varchar*(30)	Sí	No	No	Nomenclatura de la OMS para la variante
linaje_pango	*text*[]	No	No	No	Linaje(s) Pango de la variante
sustituciones_spike	*text*[]	No	No	No	Sustituciones de la proteína spike
Nombre	*varchar*(30)	No	No	No	Nombre de la variante
Color	*varchar*(10)	Sí	No	No	Color identificador de la variante

Nota. Elaboración propia.

Tabla 15. Diccionario de datos de la tabla Algoritmos.

Nombre de la tabla		**Algoritmos**			
Descripción		Tabla donde se registran los algoritmos			
Nombre del campo	**Tipo de dato**	**No Nulo**	**PK (Llave primaria)**	**FK (Llave foránea)**	**Descripción**
id_algoritmo	*integer*	Sí	Sí	No	Identificador del registro en la tabla algoritmos
nombre	*varchar*(30)	Sí	No	No	Nombre del algoritmo
parametro	*double precision*	No	No	No	Valor del parámetro para el algoritmo de agrupamiento, puede ser k, épsilon o número de clúster
algoritmo_ entrenado	*bytea*	No	No	No	Archivo con el algoritmo entrenado

Nota. Elaboración propia.

Tabla 16. Diccionario de datos de la tabla Archivos.

Nombre de la tabla		**Archivos**			
Descripción		Tabla donde se registran los archivos			
Nombre del campo	**Tipo de dato**	**No Nulo**	**PK (Llave primaria)**	**FK (Llave foránea)**	**Descripción**
id_archivo	*integer*	Sí	Sí	No	Identificador del registro en la tabla archivos
nombre	*varchar*(30)	Sí	No	No	Nombre del archivo
archivo	*Bytea*	No	No	No	Archivo que contiene la información

Nota. Elaboración propia.

Tabla 17. Diccionario de datos de la tabla Agrupamiento.

Nombre de la tabla		**Agrupamiento**			
Descripción		***Tabla donde se registran los agrupamientos realizados***			
Nombre del campo	**Tipo de dato**	**No Nulo**	**PK (Llave primaria)**	**FK (Llave foránea)**	**Descripción**
id_agrupamiento	*integer*	Sí	Sí	No	Identificador del registro en la tabla agrupamiento
id_algoritmo	*integer*	Sí	No	Sí	Identificador del algoritmo
id_secuencia	*integer*	Sí	No	Sí	Identificador de la secuencia genómica SARS-CoV-2
id_variante	*integer*	Sí	No	Sí	Identificador de la variante

num_cluster	*integer*	No	No	No	Número de clúster al que pertenece la secuencia genómica SARS-CoV-2

Nota. Elaboración propia.

El script que contiene los comandos para la creación de las tablas de la base de datos se encuentra en el Anexo B.

Por otro lado, el indicador objetivamente verificable para este resultado contiene la validación del documento por parte de un especialista en base de datos. Este documento se encuentra en el Anexo C.

4.2.2 Módulo de software para realizar el pre-procesamiento de los datos de las secuencias genómicas SARS-CoV-2.

En este resultado se realiza el pre-procesamiento de las secuencias genómicas SARS-CoV-2 donde se realiza su representación en baja dimensión; así como también, se realiza el desarrollo del backend del módulo espacio-temporal. Para lograr esto, se propone como medios de verificación, el código fuente del pre-procesamiento y módulo espacio-temporal, y el catálogo de pruebas.

Para el primer medio de verificación, se tiene en primer lugar, la descripción del flujo a seguir para obtener los datos de las secuencias genómicas SARS-CoV-2. En la Figura 10, se presenta el flujo de manera gráfica correspondiente al proceso descrito a continuación.

1. En primer lugar, se accede a la base de datos GISAID y se selecciona la pestaña de EpiCoV – *Search*.
2. Se filtra la búsqueda de las secuencias genómicas SARS-CoV-2 por "*Location*", seleccionando la opción "*South America / Peru*" y en "*Host*" se selecciona "*Human*".
3. Se selecciona la opción "*complete*" para solo obtener las secuencias genómicas completas.
4. Luego, se seleccionan todas las secuencias genómicas obtenidas con los filtros realizados.

5. Se procede a descargar todas las secuencias seleccionadas haciendo clic en el botón "*Download*", seguidamente se selecciona el formato de descarga "*Sequences* (FASTA)" y la opción "*Replace spaces with underscores in* FASTA *header*" y finalmente, se dará clic en el botón "*Download*".
6. Por último, se vuelve a dar clic en el botón "*Download*", se selecciona el formato de descarga "*Sequencing technology metadata*" y se da clic en el botón "*Download*".

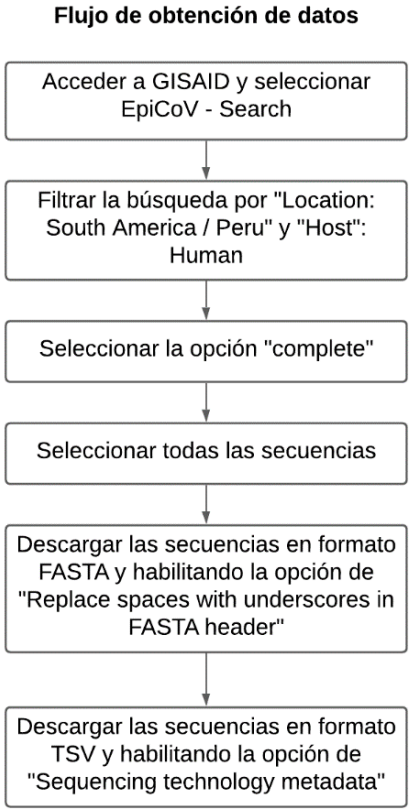


Figura 10. Flujo de obtención de datos de las secuencias genómicas SARS-CoV-2.

Luego de la obtención de los datos se procede a realizar el preprocesamiento de los datos de las secuencias genómicas SARS-CoV-2. El procedimiento realizado para el preprocesamiento es el siguiente:

1. Una vez descargadas las secuencias genómicas SARS-CoV-2, es necesario convertir los datos a un formato más sencillo de trabajar, dado que uno de los archivos es de extensión FASTA y el otro de extensión TSV. Por ello, se leen estos archivos para obtener los datos.

2. Se obtiene el identificador de acceso, la fecha de recolección y el lugar de donde se obtuvo cada secuencia genómica SARS-CoV-2. Se consideran solo las secuencias que pertenecen o hacen mención a un departamento del Perú.
3. Se elimina las secuencias con errores de lectura, definido como letras distintas de A, C, G y T.
4. Se obtiene el linaje pango de las secuencias genómicas SARS-CoV-2.
5. Luego se realiza el alineamiento múltiple de las secuencias genómicas SARS-CoV-2 para conocer los lugares en donde, en una misma posición para todas las secuencias, se tiene diferentes nucleótidos. Para realizar este alineamiento se utiliza la función *MultipleSeqAlignment* de la librería Biopython.
6. Se calcula la matriz de distancia Hamming de las secuencias, estas distancias son la representación de la desigualdad de cada nucleótido de una secuencia con respecto a otra. Esta distancia se calcula de cada secuencia con respecto a las demás y se representa en el porcentaje de los nucleótidos que divergen entre las secuencias.
7. Luego se realiza el escalamiento multidimensional (MDS) para transformar las distancias entre las secuencias genómicas a un espacio de 10 dimensiones.
8. Se aplica el análisis de componentes principales (PCA) para transformar los datos a un espacio de dos dimensiones y así poder realizar las visualizaciones de agrupamiento en dos dimensiones. El porcentaje de varianza explicada con los dos primeros componentes principales es del 32,04%. En la Figura 11, se presenta el gráfico del acumulado de varianza explicada en las nuevas dimensiones.

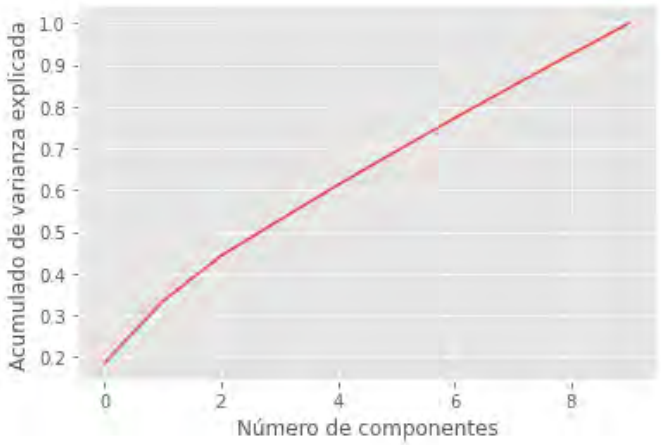


Figura 11. Grafica del acumulado de varianza explicada con 10 componentes.

9. Luego, se eligió los puntos de referencia (*landmark*), los cuales fueron seleccionados por el algoritmo de agrupamiento *k-means*, seleccionando los puntos más cercanos a los centroides obtenidos al realizar el agrupamiento. Luego de realizar pruebas con diferentes cantidades de puntos de referencia, siendo estas 10, 20, 40, 50 y 70, donde se obtuvo que con 40 puntos de referencia se logra representar adecuadamente los datos.
10. Teniendo los puntos de referencia, se coloca en una matriz X de tamaño n x m la distancia de cada punto con respecto a cada punto de referencia, donde n es la cantidad de secuencias genómicas SARS-CoV-2 a preprocesar y m la cantidad de puntos de referencia.
11. Se asocia un *target* de las posiciones a obtener de los datos, es decir, lo que va a predecir el modelo. Este *target* es de tamaño n x 2, donde n es la cantidad de secuencias genómicas SARS-CoV-2 a preprocesar y 2 es la cantidad de valores que el modelo de mapeamiento debe predecir, las coordenadas en dos dimensiones de las secuencias genómicas SARS-CoV-2. Este *target* tiene la representación de las secuencias genómicas SARS-CoV-2 en dos dimensiones para su visualización.
12. Posteriormente, se procede a crear una red neuronal perceptrón multicapa (MLP, por sus siglas en inglés), con 4 capas. La cantidad de entradas en la red neuronal es la cantidad de puntos de referencia (*landmark*), entonces en la capa de entrada se tiene 40 neuronas. Se tiene 2 capas ocultas, la primera capa oculta con 20 neuronas y la segunda capa oculta con

10 neuronas. La salida de la red es igual a la cantidad de coordenadas del *target*, en este caso dos dimensiones. Todas estas capas tienen como función de activación la tangente hiperbólica (tanh) y como inicializador, el inicializador normal de Glorot (GlorotNormal). Luego, de definir las capas se compila el modelo con un optimizador, el algoritmo Nadam, el cual tiene como parámetro la tasa de aprendizaje de 0,000 09.

13. Por último, se procede a entrenar el modelo con los datos de las secuencias genómicas SARS-CoV-2 con 15000 épocas (*epoch*), que representa el número de veces que se ejecutará el modelo, donde en cada época se ajustan los pesos del modelo de red neuronal con todos los datos de entrenamiento.

En la Figura 12, se presenta el flujo de manera gráfica del procedimiento anteriormente mencionado.

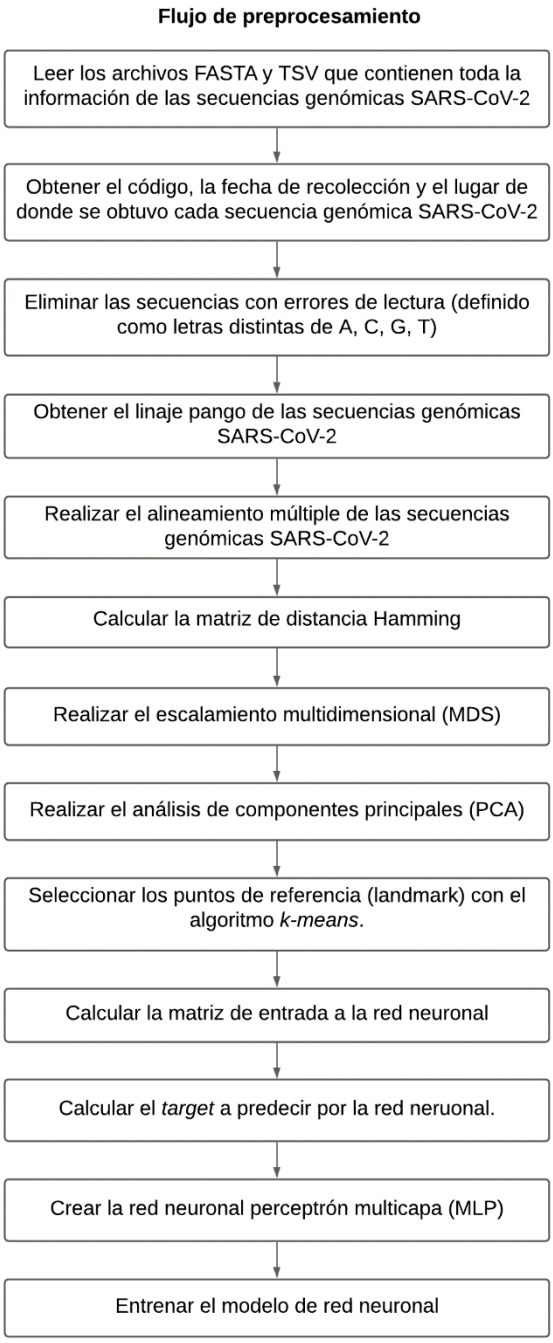


Figura 12. Flujo del preprocesamiento de las secuencias genómicas SARS-CoV-2.

El código fuente del preprocesamiento realizado se encuentra en el siguiente repositorio: https://github.com/CarolinaMejiaMujica/Preprocesamiento

Por último, para la programación de los servicios del backend, del módulo espacio-temporal, se ha usado el *framework* de Python, FastApi. Los servicios requeridos para el presente módulo son los siguientes:

- ❖ Servicio que realice el gráfico del mapa del Perú a nivel regional de las variantes identificadas en cada departamento.
- ❖ Servicio que realice el gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo.
- ❖ Servicio que realice el gráfico de línea con la cantidad de secuencias genómicas SARS-CoV-2 pertenecientes a las variantes identificadas en el tiempo.
- ❖ Servicio que lista los datos de las secuencias genómicas SARS-CoV-2. Estos datos serán los siguientes: Departamento, ID de acceso de las secuencias genómicas (identificador en la base de datos de GISAID), Fecha de recolección, Nomenclatura según la OMS de la variante identificada y Nombre de la variante identificada.

Cada uno de estos servicios recibe como parámetros el rango de fechas, fecha inicio y fecha fin, y los departamentos seleccionados por el usuario, para así poder realizar el filtro de los datos. Esta fecha corresponde a la fecha de recolección de las secuencias genómicas SARS-CoV-2. Así mismo, también se recibe como parámetro el departamento o departamentos para realizar el filtrado de datos.

- ❖ Servicio que calcula la cantidad de secuencias genómicas SARS-CoV-2 obtenidas de GISAID y la cantidad de secuencias genómicas SARS-CoV-2 utilizadas en el análisis.

Este último servicio no recibe ningún dato de entrada solo retorna las cantidades mencionadas anteriormente.

El código fuente del backend del módulo del presente objetivo específico se encuentra en el siguiente repositorio: https://github.com/CarolinaMejiaMujica/Modulo-Espacio-Temporal

Como segundo medio de verificación, se tiene el catálogo de pruebas del módulo espacio-temporal y el preprocesamiento, el cual consta de cinco pruebas unitarias sobre un conjunto de 9 196 datos de secuencias genómicas SARS-CoV-2. La primera prueba unitaria está

relacionada al preprocesamiento, la segunda, al servicio que realiza el gráfico del mapa del Perú, la tercera, al servicio que realiza el gráfico circular, la cuarta, al servicio que realiza el gráfico de línea, y la quinta, al servicio que lista los datos de las secuencias genómicas SARS-CoV-2.

En el Anexo D, se detalla cada prueba unitaria realizada y los resultados obtenidos. Por otro lado, el indicador objetivamente verificable para este resultado contiene el resultado de las pruebas unitarias aprobadas al 100%, esto indica que las pruebas se han completado de forma satisfactoria cumpliendo con los resultados esperados. Este documento se encuentra en el Anexo E.

4.2.3 Módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú.

Este resultado, hace uso del resultado descrito en el punto 4.2.2 para así mostrar visualmente la representación en el espacio y tiempo de las secuencias genómicas SARS-CoV-2 recolectadas en el Perú, esto con el fin de que los especialistas y analistas puedan ver la evolución del virus a nivel regional del Perú, así como también, en el tiempo desde que inició la pandemia hasta la actualidad (20/01/2022) para que así puedan realizar estudios o tomar decisiones con la información brindada.

Para lograr esto, se propone como primer medio de verificación identificar todos los requerimientos necesarios para el desarrollo del módulo espacio-temporal. En la Tabla 18, se presenta el nivel de prioridad y su descripción de lo que significa cada nivel. Los requerimientos que tengan un nivel de prioridad “Alta” serán realizados primero, posteriormente con los de prioridad “Media”; y, por último, con los de prioridad “Baja”.

Tabla 18. Tabla de prioridad

Nivel de Prioridad	Descripción
1	Alta

2	Media
3	Baja

Nota. Elaboración propia.

En la Tabla 19, se presenta la lista de requerimientos con su tipo de requerimiento, el nivel prioridad y si es exigible o deseable.

Tabla 19. Lista de requerimientos para el módulo espacio-temporal

N°	Requerimiento	Tipo	Nivel de Prioridad	Exigible / Deseable
1	El módulo permitirá seleccionar el rango de fechas (fecha inicio y fin).	Funcional	1	Exigible
2	El módulo permitirá filtrar por departamentos.	Funcional	1	Exigible
3	El módulo mostrará en un mapa del Perú a nivel regional las variantes identificadas en cada departamento.	Funcional	1	Exigible
4	El módulo permitirá visualizar en cada departamento el detalle con la siguiente información: Departamento, Total de secuencias genómicas, Variantes identificadas y la Variante predominante.	Funcional	2	Exigible
5	El módulo permitirá visualizar cada departamento de un color, el cual corresponderá a la variante predominante en ese departamento.	Funcional	3	Exigible
6	El módulo mostrará un gráfico de línea con la cantidad de secuencias genómicas SARS-CoV-2 pertenecientes a las variantes identificadas en el tiempo.	Funcional	1	Exigible

7	El módulo mostrará un gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo.	Funcional	2	Exigible
8	El módulo permitirá visualizar los datos de las secuencias genómicas de acuerdo con los filtros. Estos datos serán los siguientes: Departamento, ID de acceso de las secuencias genómicas (identificador en la base de datos de GISAID), Fecha de recolección, Nomenclatura según la OMS de la variante identificada y Nombre de la variante identificada.	Funcional	1	Exigible
9	El módulo permitirá descargar los datos de las secuencias genómicas SARS-CoV-2.	Funcional	2	Exigible

Nota. Elaboración propia.

Como segundo medio de verificación, se tiene el código fuente desarrollado para el frontend del módulo espacio-temporal, para ello, se utilizó el *framework* React.

Se cuenta con los filtros de rango de fechas, fecha inicio y fecha fin, y el filtro por departamento(s); una vez seleccionado estos filtros se le da clic al botón “Generar”, con el cual se actualiza la información mostrada en los gráficos y la tabla. En la Figura 13, se muestran los filtros anteriormente descritos.

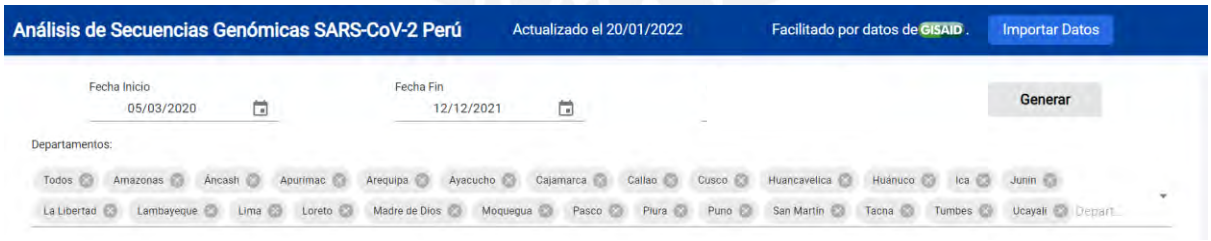


Figura 13. Filtros de rango de fechas y departamentos.

En la Figura 14, se muestra el gráfico del mapa del Perú a nivel regional donde se aprecia la variante predominante de cada departamento por su respectivo color, la cantidad de secuencias genómicas SARS-CoV-2 obtenidas de GISAID, la cantidad de secuencias

genómicas SARS-CoV-2 utilizadas en el análisis; y, el gráfico circular donde se aprecia la distribución de secuencias genómicas SARS-CoV-2 por variantes.

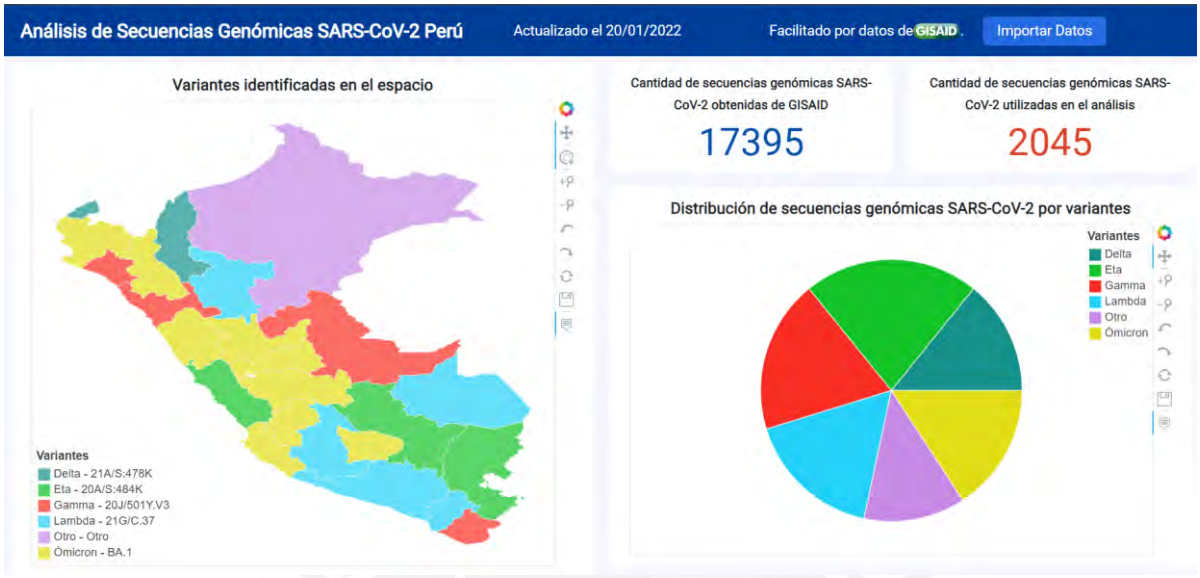


Figura 14. Gráfico del mapa del Perú y gráfico circular de la distribución de secuencias genómicas SARS-CoV-2.

En la Figura 15, se muestra el gráfico de línea de las variantes identificadas en el tiempo, donde se muestra una leyenda de cada variante con su respectivo color correspondiente.

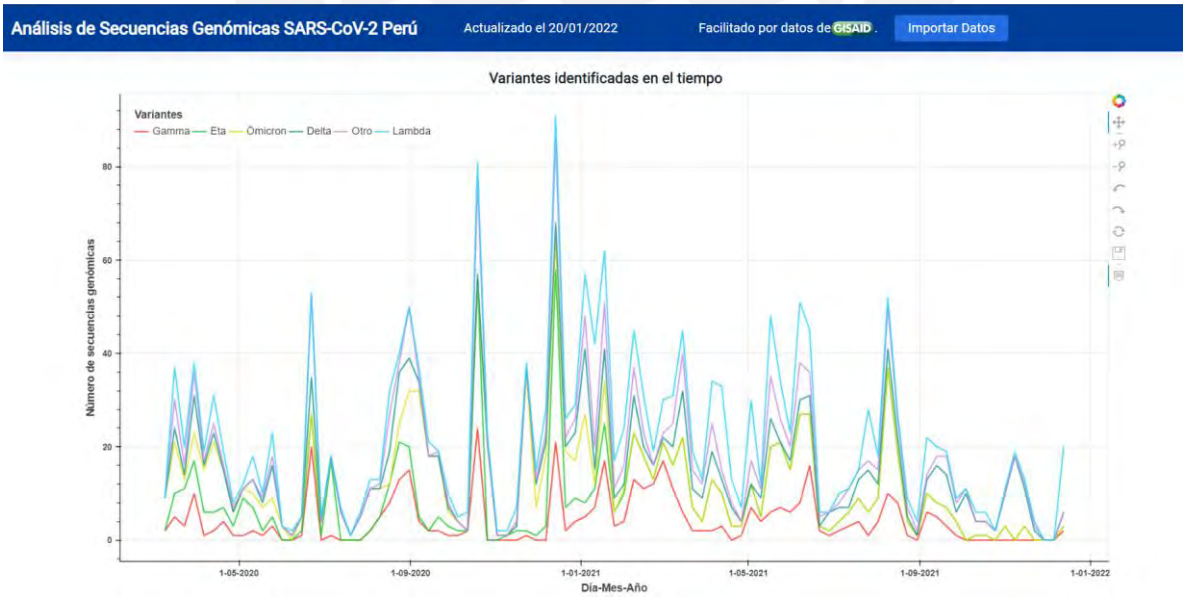


Figura 15. Gráfico lineal que muestra la evolución de variantes en el tiempo.

En la Figura 16, se muestra una tabla con los datos de las secuencias genómicas SARS-CoV-2 utilizados en los gráficos anteriores, se muestra los datos a mayor detalle, como el

departamento de donde se obtuvo la secuencia, el identificador de acceso de la secuencia en la base de datos GISAID, la fecha de recolección, la nomenclatura según la OMS y el nombre de la variante identificada a la que pertenece la secuencia genómica SARS-CoV-2.

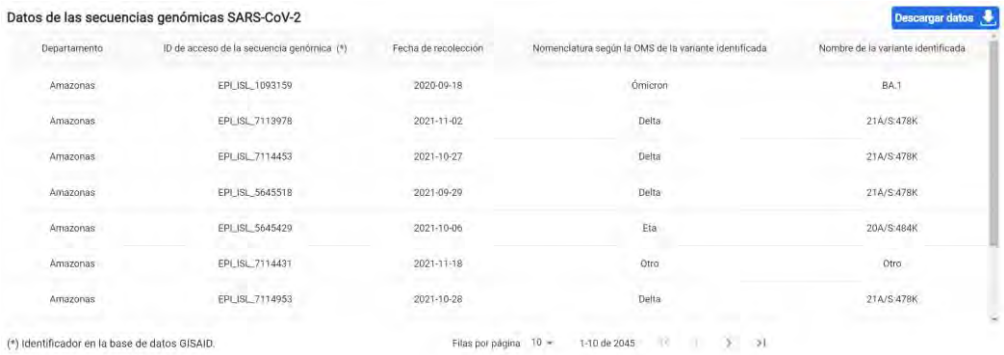


Figura 16. Tabla con los datos de las secuencias genómicas SARS-CoV-2 utilizadas en los gráficos.

El código fuente del frontend del módulo del presente objetivo específico se encuentra en el siguiente repositorio: https://github.com/CarolinaMejiaMujica/Modulo-Espacio-Temporal-Front

Adicionalmente, como segundo medio de verificación se tiene el manual de usuario del módulo espacio temporal. En este manual se detalla la función de cada herramienta u opciones que existen.

1. Primero se tiene la sección de filtros donde se encuentran los filtros de fecha inicio, fecha fin y departamentos. Luego de seleccionar los filtros, se da clic al botón Generar para actualizar los datos de las secuencias genómicas SARS-CoV-2 y los gráficos del módulo espacio-temporal. En la Figura 17, se explica la sección de filtros.

Figura 17. Detalle de los filtros de rango de fechas y departamentos.

2. Luego, se presentan los gráficos del módulo espacio-temporal, el gráfico del mapa del Perú, el gráfico circular donde se aprecia el porcentaje de variantes identificadas en el tiempo y el gráfico de línea de las variantes identificadas en el tiempo. Estos gráficos tienen herramientas, de las cuales se explica su funcionalidad en la Figura 18.

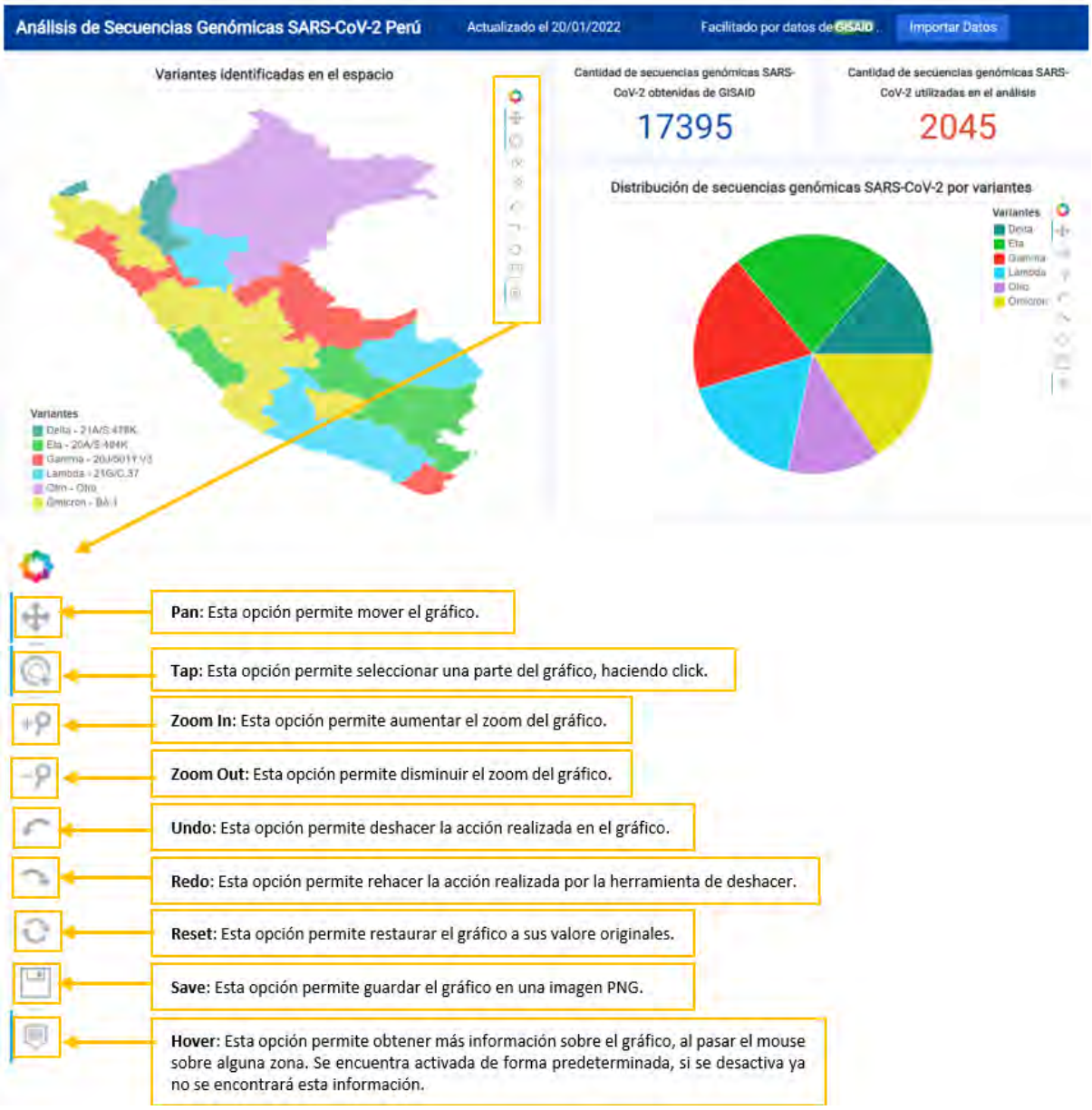


Figura 18. Detalle de las herramientas de los gráficos del módulo espacio-temporal.

3. Por último, se tiene la visualización de los datos de las secuencias genómicas SARS-CoV-2, donde se tiene la opción de descargar los datos observados en la tabla, como se muestra en la Figura 19.

Figura 19. Detalle de la opción de descarga en la visualización de los datos de las secuencias genómicas SARS-CoV-2.

Por otro lado, el primer indicador objetivamente verificable para este resultado contiene la validación de que el software cumpla con el 100% de los requerimientos especificados por parte de un especialista en ingeniería informática. Este documento se encuentra en el Anexo F.

El segundo indicador objetivamente verificable es el reporte de apreciación del módulo de visualización espacio-temporal por un especialista en bioinformática o biología molecular. Este documento se encuentra en el Anexo G.

4.3 Discusión

Para solucionar el problema que no se cuenta con estudios publicados que comprendan o investiguen el análisis espacio-temporal de la diversidad genómica de las muestras de virus SARS-CoV-2 secuenciadas en Perú se obtuvieron tres resultados.

En primer lugar, se realizó la estructura de la base de datos de toda la herramienta analítica. En esta estructura se consideró los dos módulos a desarrollar, el módulo espacio temporal y el de agrupamiento. Asimismo, se tomó de referencia la información que muestra la herramienta analítica Nextstrain, ya que esta herramienta ayuda a la comprensión epidemiológica del virus SARS-CoV-2 a nivel mundial (Nexstrain, s.f.).

En segundo lugar, se realizó el preprocesamiento de los datos de las secuencias genómicas SARS-CoV-2 la cual incluye la representación en baja dimensión de las mismas para una mejor representación de las secuencias visualmente, esto se logró con el alineamiento múltiple de secuencias, el escalamiento multidimensional (MDS) y el análisis de componentes principales (PCA). Adicional a ello, como parte del resultado se implementó los servicios a utilizar en el módulo espacio-temporal, estos incluyen el gráfico del mapa del Perú de las variantes identificadas a nivel regional, el gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo, el gráfico de línea con la cantidad de secuencias genómicas SARS-CoV-2 pertenecientes a las variantes identificadas en el tiempo y la visualización de los datos de las secuencias genómicas.

En tercer lugar, se definieron los requerimientos necesarios para el desarrollo del módulo espacio-temporal. Una vez definidos, y con los servicios ya implementados en el anterior resultado, se procedió a generar las representaciones visuales de los servicios. Con estos gráficos se comprende mejor el análisis espacio-temporal de la diversidad genómica del virus SARS-CoV-2 en el Perú.

La principal limitante de los resultados son los datos de las secuencias genómicas, ya que muchas de estas presentan errores de lectura por lo que estas secuencias son descartadas debido a que estos errores de lectura dificultan el preprocesamiento de las mismas. Se sugiere como trabajos futuros investigar formas de tratar estas secuencias genómicas, como por ejemplo técnicas de imputación, a fin de no descartar secuencias.

Capítulo 5. Módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2

5.1 Introducción

En el presente capítulo, se presenta los resultados alcanzados para lograr el cumplimento del segundo objetivo específico, el cual es "Desarrollar un módulo de software que permita realizar un análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 y su visualización con capacidad de incorporar nuevas secuencias de forma en línea". A continuación, se detalla los resultados alcanzados los cuales son: la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2, el módulo de software que implementa estos métodos de análisis de agrupamiento; y, el módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 en el Perú. Además, se presentan los medios de verificación y los indicadores objetivamente verificables con los cuales se va a validar el cumplimiento de cada resultado alcanzado.

5.2 Resultados alcanzados

5.2.1 Selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.

En este resultado se realiza la selección de los métodos de análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 a ser implementados. Para lograr esto, se propone como medio de verificación, un informe de selección de métodos de análisis de agrupamiento.

Los métodos de análisis de agrupamiento a implementar son el método de agrupamiento *k-means*, el método jerárquico y el método de agrupamiento espacial DBSCAN. A continuación, se detalla el motivo de la elección de estos métodos.

El método *k-means* es uno de los métodos de análisis no supervisados más populares (Hozumi et al., 2021) y utilizados para realizar el agrupamiento de secuencias genómicas SARS-CoV-2. La principal ventaja de utilizar *k-means* es la obtención de mejores resultados

al tener una mayor cantidad de datos (Hozumi et al., 2021). Además, *k-means* presenta una mejor agrupación y rendimiento cuando se utiliza en conjunto con algoritmos de reducción de dimensionalidad (Hozumi et al., 2021).

Por otro lado, el método agrupamiento jerárquico es otro de los métodos más usados para identificar y agrupar las secuencias genómicas SARS-CoV-2. En algunos estudios se utiliza el agrupamiento jerárquico con una matriz de similitud de coincidencia de las secuencias genómicas SARS-CoV-2 con una secuencia de referencia, como la secuencias Wuhan-Hu-1 (H.-C. Yang et al., 2020). Estudios con el agrupamiento jerárquico ha ayudado a identificar diferentes tipos de cepas de SARS-CoV-2 (H.-C. Yang et al., 2020).

Por último, el algoritmo de agrupación espacial basada en densidad de aplicaciones con ruido (DBSCAN) tiene como ventaja determinar la cantidad de grupos automáticamente (Wisesty & Mengko, 2021). Además, DBSCAN detecta datos atípicos, cuando estos datos están muy lejos de los grupos existentes son considerados ruido, en lugar de forzar a los datos atípicos a pertenecer a un clúster específico (Wisesty & Mengko, 2021). En un estudio se comparó este algoritmo con otros, como *k-means*, agrupamiento jerárquico aglomerativo, modelo de mezcla gaussiana (GMM) y algoritmo de cambio medio, utilizando el análisis de siluetas, en donde resultó que el algoritmo DBSCAN con PCA obtuvo mejores resultados en cuanto al agrupamiento realizado, con una puntuación de silueta de 0,8 133 (Wisesty & Mengko, 2021).

Por lo anteriormente mencionado, estos métodos son los seleccionados debido a que son los más utilizados y han presentado mejores resultados al realizar el agrupamiento de las secuencias genómicas SARS-CoV-2. Además, porque ofrecen distintas formas de encontrar los grupos, permitiendo así que el usuario utilice cualquiera de estos tres algoritmos para un mejor análisis de agrupamiento.

Por otro lado, el indicador objetivamente verificable para este resultado contiene la validación del documento por parte de un especialista en inteligencia artificial o ciencia de datos. Este documento se encuentra en el Anexo H.

5.2.2 Módulo de software que implementa los métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2.

En este resultado se realiza el desarrollo del backend del módulo de agrupamiento. Para lograr esto, se propone como medios de verificación, el código fuente del módulo de agrupamiento, y el catálogo de pruebas.

En el primer medio de verificación se tiene la programación de los servicios del backend, del módulo de agrupamiento, para el cual se ha usado el *framework* de Python, FastApi. Los servicios requeridos para el presente módulo son los siguientes:

- Servicio que realice el gráfico de dispersión que represente el agrupamiento realizado con el algoritmo *k-means* para mostrar las variantes identificadas en el Perú.
- Servicio que realice el gráfico de dispersión que represente el agrupamiento realizado con el algoritmo jerárquico para mostrar las variantes identificadas en el Perú.
- Servicio que realice el gráfico de dispersión que represente el agrupamiento realizado con el algoritmo DBSCAN para mostrar las variantes identificadas en el Perú.
- Servicio que realice el dendrograma para visualizar la estructura o jerarquía de los clústeres que se pueden formar.
- Servicio que lista los datos de las secuencias genómicas SARS-CoV-2 utilizadas en este módulo para el agrupamiento. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Cada uno de estos servicios recibe como parámetros el rango de fechas, fecha inicio y fecha fin, los departamentos seleccionados por el usuario, el algoritmo de agrupamiento y el

valor del parámetro del algoritmo, para así poder realizar el filtro de los datos. Esta fecha corresponde a la fecha de recolección de las secuencias genómicas SARS-CoV-2. Así mismo, también se recibe como parámetro el departamento o departamentos para realizar el filtrado de datos.

El código fuente del backend del módulo del presente objetivo específico se encuentra en el siguiente repositorio:

https://github.com/CarolinaMejiaMujica/Modulo-Agrupamiento

Como segundo medio de verificación, se tiene el catálogo de pruebas del módulo de agrupamiento, el cual consta de cinco pruebas unitarias sobre un conjunto de 1 365 datos de secuencias genómicas SARS-CoV-2. La primera prueba unitaria está relacionado al servicio que realiza el gráfico de agrupamiento con el algoritmo *k-means*, la segunda, al servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico, la tercera, al servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN, la cuarta, al servicio que realiza el dendrograma, y la quinta, al servicio que lista los datos de las secuencias genómicas SARS-CoV-2 utilizadas en este módulo para el agrupamiento.

En el Anexo I, se detalla cada prueba unitaria realizada y los resultados obtenidos.

Por otro lado, el primer indicador objetivamente verificable para este resultado contiene el resultado de las pruebas unitarias aprobadas al 100%, esto indica que las pruebas se han completado de forma satisfactoria cumpliendo con los resultados esperados. Este documento se encuentra en el Anexo J.

El segundo indicador objetivamente verificable se cumplió al implementar el algoritmo de *k-means*, que es un algoritmo de clustering particional, y el algoritmo jerárquico aglomerativo, el cual es un algoritmo de clustering jerárquico; esto se comprueba en el código fuente.

5.2.3 Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2.

Este resultado, hace uso del resultado descrito en el punto 5.2.2 para así visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 recolectadas, esto con el fin de que los especialistas y analistas puedan ver la diversidad de secuencias genómicas SARS-CoV-2 en el Perú, así como también, comprender las posibles variantes que se encuentren presentes y encontrar patrones en la evolución del virus.

Para lograr esto, se propone como primer medio de verificación identificar todos los requerimientos necesarios para el desarrollo del módulo de agrupamiento. En la Tabla 20, se presenta el nivel de prioridad y su descripción de lo que significa cada nivel. Los requerimientos que tengan un nivel de prioridad "Alta" serán realizados primero, posteriormente con los de prioridad "Media"; y, por último, con los de prioridad "Baja".

Tabla 20. Tabla de prioridad

Nivel de Prioridad	Descripción
1	Alta
2	Media
3	Baja

Nota. Elaboración propia.

En la Tabla 21, se presenta la lista de requerimiento con su tipo de requerimiento, el nivel de prioridad y si es exigible o deseable.

Tabla 21. Lista de requerimientos para el módulo de agrupamiento

N°	Requerimiento	Tipo	Nivel de Prioridad	Exigible / Deseable
1	El módulo permitirá seleccionar el rango de fechas (fecha inicio y fin), los departamentos y el	Funcional	1	Exigible

	algoritmo de agrupamiento (*k-means*, jerárquico y DBSCAN).			
2	El módulo permitirá visualizar el agrupamiento realizado de las secuencias genómicas SARS-CoV-2 identificadas en el Perú.	Funcional	1	Exigible
3	El módulo mostrará en un gráfico de dispersión las variantes identificadas en el Perú para los algoritmos *k-means*, jerárquico y DBSCAN.	Funcional	1	Exigible
4	El módulo mostrará un dendrograma para el algoritmo jerárquico.	Funcional	2	Exigible
5	El módulo permitirá visualizar el detalle de cada secuencia genómica con la siguiente información: Departamento, ID de acceso (identificador en la base de datos de GISAID), Fecha de recolección, Variante de la secuencia y Variante predominante del grupo (correspondiente al grupo al que pertenece).	Funcional	1	Exigible
6	El módulo mostrará los datos de las secuencias genómicas por clústeres obtenidos de acuerdo al algoritmo seleccionado. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.	Funcional	1	Exigible
7	El módulo permitirá descargar los datos de las secuencias genómicas SARS-CoV-2.	Funcional	2	Exigible
8	El módulo de agrupamiento permitirá filtrar por clústeres en los algoritmos *k-means* y jerárquico.	Funcional	1	Exigible
9	El módulo de agrupamiento permitirá filtrar por el valor de épsilon en el algoritmo DBSCAN.	Funcional	2	Exigible

Nota. Elaboración propia.

Como segundo medio de verificación, se tiene el código fuente desarrollado para el frontend del módulo de agrupamiento, para ello, se utilizó el *framework* React.

Se cuenta con los filtros de rango de fechas, fecha inicio y fecha fin, el filtro por departamento(s), y, el filtro de algoritmo de agrupamiento; una vez seleccionado estos filtros se le da clic al botón "Generar", con el cual se actualiza la información mostrada en los gráficos y la tabla. En la Figura 20, se muestran los filtros anteriormente descritos.

Figura 20. Filtros de rango de fechas y departamentos.

En la Figura 21, se muestra el agrupamiento de las secuencias genómicas SARS-CoV-2 con el algoritmo de *k-means*, donde se aprecia al lado izquierdo en un gráfico de dispersión las variantes identificadas en el Perú con el algoritmo *k-means* y al lado derecho se muestra una tabla con los datos de las secuencias genómicas agrupadas con el algoritmo *k-means*. En la tabla se muestra los datos a mayor detalle, como el departamento de donde se obtuvo la secuencia, el identificador de acceso de la secuencia en la base de datos GISAID, la fecha de recolección, el número de clúster, y la nomenclatura variante identificada según la OMS.

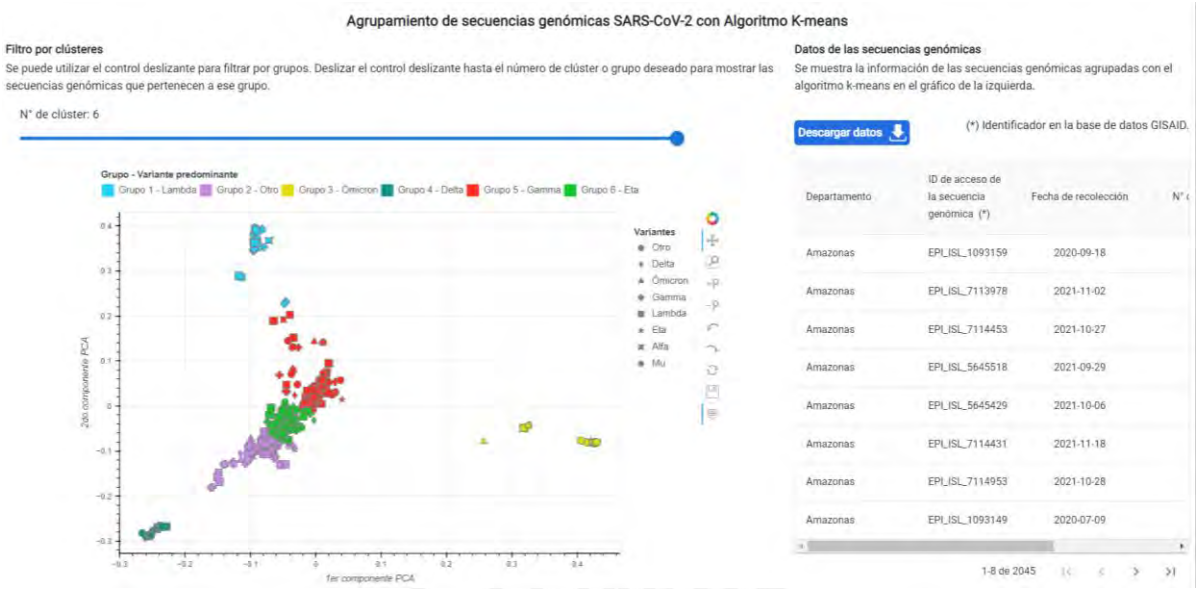


Figura 21. Captura de pantalla del módulo de agrupamiento de secuencias genómicas SARS-CoV-2 realizado con el algoritmo k-means.

En la Figura 22, se muestra el agrupamiento de las secuencias genómicas SARS-CoV-2 con el algoritmo jerárquico, donde se aprecia al lado izquierdo el dendrograma de las secuencias genómicas SARS-CoV-2 y al lado derecho un gráfico de dispersión donde se visualiza las variantes identificadas en el Perú con el algoritmo jerárquico.

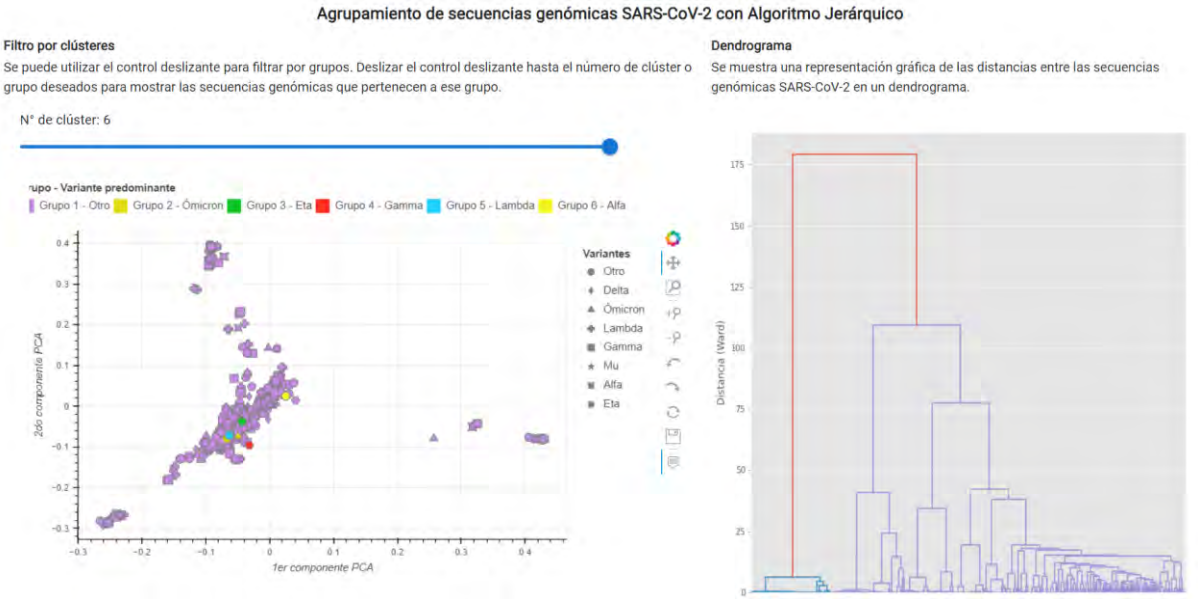


Figura 22. Captura de pantalla del módulo de agrupamiento de secuencias genómicas SARS-CoV-2 realizado con el algoritmo jerárquico.

En la Figura 23, se muestra una tabla con los datos de las secuencias genómicas SARS-CoV-2 agrupadas con el algoritmo jerárquico, se muestra los datos a mayor detalle, como el departamento de donde se obtuvo la secuencia, el identificador de acceso de la secuencia en la base de datos GISAID, la fecha de recolección, el número de clúster, y la nomenclatura variante identificada según la OMS.

Datos de las secuencias genómicas SARS-CoV-2

Descargar datos

Se muestra la información de las secuencias genómicas agrupadas con el algoritmo jerárquico en el gráfico superior.

Departamento	ID de acceso de la secuencia genómica (*)	Fecha de recolección	N° de clúster	Nomenclatura según la OMS de la variante identificada
Amazonas	EPI_ISL_1093159	2020-09-18	1	Otro
Amazonas	EPI_ISL_7113978	2021-11-02	1	Otro
Amazonas	EPI_ISL_7114453	2021-10-27	1	Otro
Amazonas	EPI_ISL_5645518	2021-09-29	1	Otro
Amazonas	EPI_ISL_5645429	2021-10-06	1	Otro
Amazonas	EPI_ISL_7114431	2021-11-18	1	Otro
Amazonas	EPI_ISL_7114953	2021-10-28	1	Otro

(*) Identificador en la base de datos GISAID.

Filas por página 10 1-10 de 2045

Figura 23. Tabla con los datos de las secuencias genómicas SARS-CoV-2 agrupadas con el algoritmo jerárquico.

En la Figura 24, se muestra el agrupamiento de las secuencias genómicas SARS-CoV-2 con el algoritmo DBSCAN, donde se aprecia al lado izquierdo en un gráfico de dispersión las variantes identificadas en el Perú con el algoritmo DBSCAN y al lado derecho se muestra una tabla con los datos de las secuencias genómicas agrupadas con el algoritmo DBSCAN. En la tabla se muestra los datos a mayor detalle, como el departamento de donde se obtuvo la secuencia, el identificador de acceso de la secuencia en la base de datos GISAID, la fecha de recolección, el número de clúster, y la nomenclatura variante identificada según la OMS.

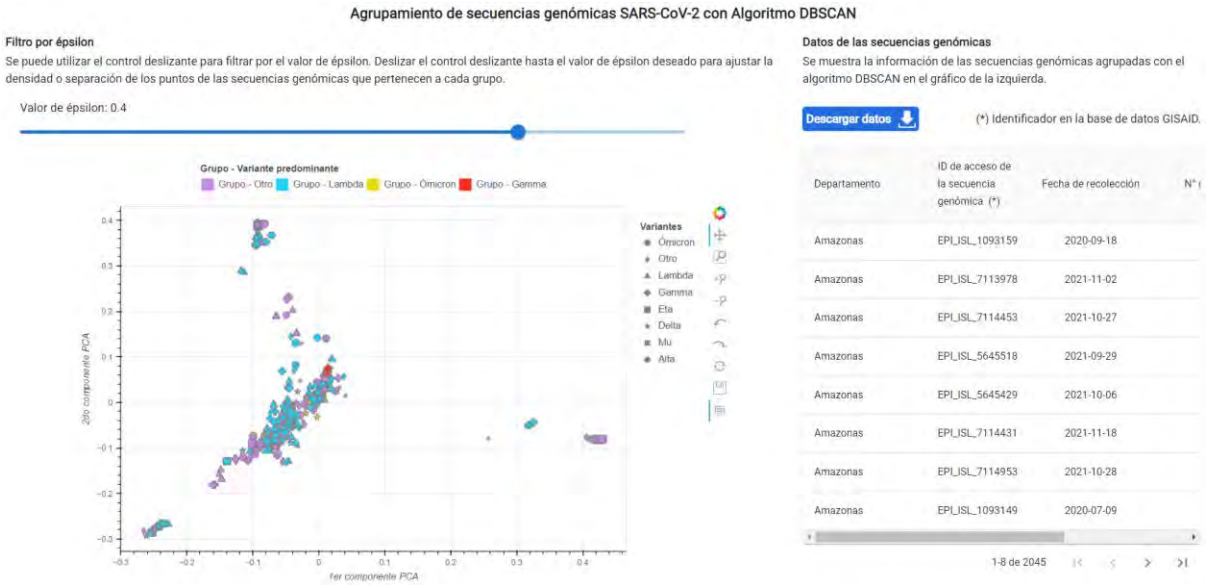


Figura 24. Captura de pantalla del módulo de agrupamiento de secuencias genómicas SARS-CoV-2 realizado con el algoritmo DBSCAN.

El código fuente del frontend del módulo del presente objetivo específico se encuentra en el siguiente repositorio: https://github.com/CarolinaMejiaMujica/Modulo-Agrupamiento-Front

Adicionalmente, como segundo medio de verificación se tiene el manual de usuario del módulo de agrupamiento. En este manual se detalla la función de cada herramienta u opciones que existen.

1. Primero se tiene la sección de filtros donde se encuentran los filtros de fecha inicio, fecha fin, algoritmo de agrupamiento y departamentos. Luego de seleccionar los filtros, se da clic al botón Generar para actualizar los datos de las secuencias genómicas SARS-CoV-2 y los gráficos del módulo de agrupamiento. En la Figura 25, se detalla las opciones mencionadas.

Figura 25. Detalle de los filtros de rango de fechas, algoritmo y departamentos.

2. Luego, se presentan los gráficos de dispersión del módulo de agrupamiento, los cuales son: el gráfico de dispersión del agrupamiento realizado con el algoritmo *k-means*, jerárquico y DBSCAN. Estos gráficos tienen herramientas, de las cuales se explica su funcionalidad en la Figura 26.

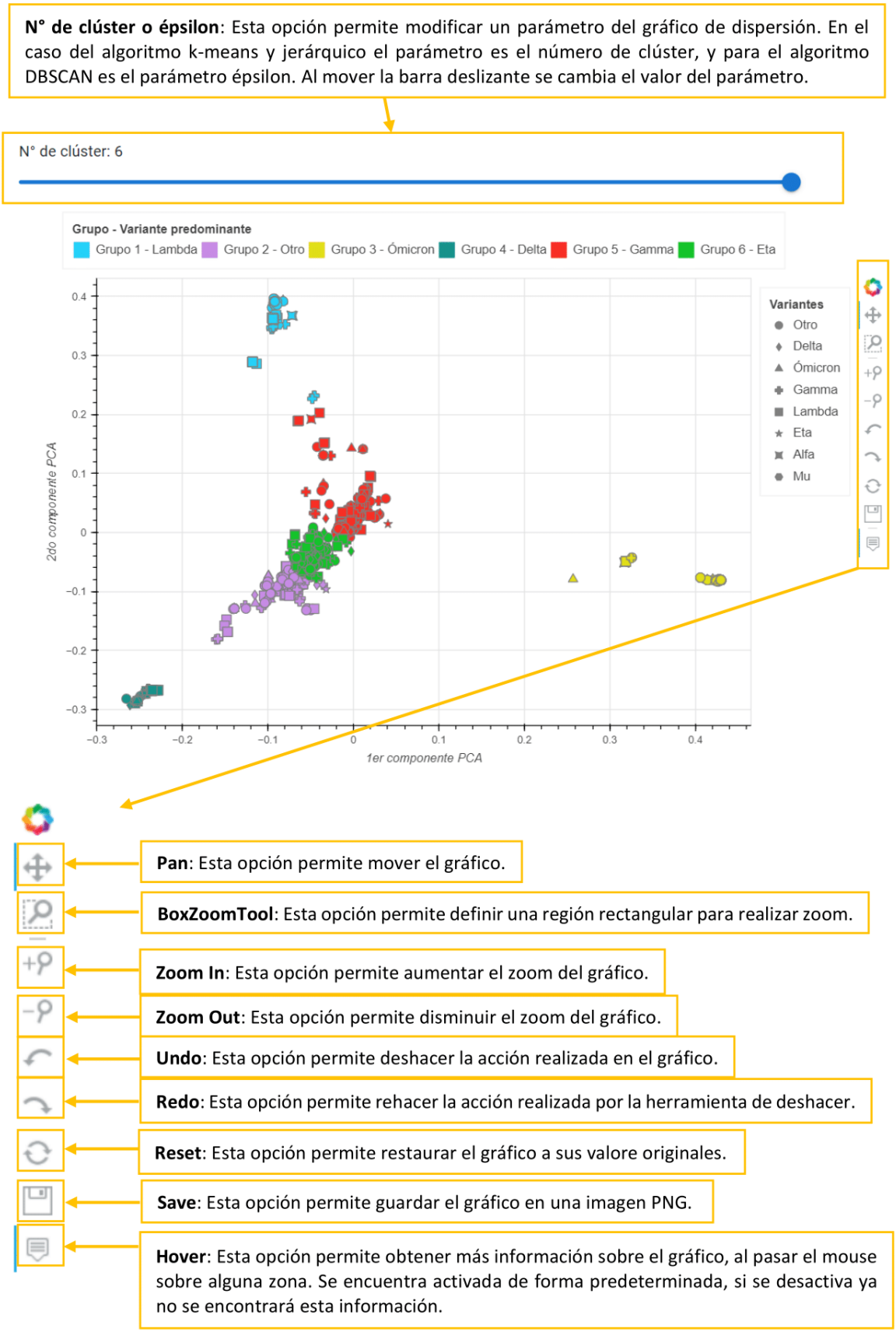


Figura 26. Detalle de las herramientas de los gráficos de dispersión del módulo de agrupamiento.

3. Por último, se tiene la visualización de los datos de las secuencias genómicas SARS-CoV-2 agrupadas con los algoritmos *k-means*, jerárquico y DBSCAN, donde se tiene la opción de descargar los datos observados en la tabla.

Figura 27. Detalle de la opción de descarga en la visualización de los datos de las secuencias genómicas SARS-CoV-2 agrupadas con los diferentes algoritmos.

Por otro lado, el primer indicador objetivamente verificable para este resultado contiene la validación de que el software cumpla con el 100% de los requerimientos especificados por parte de un especialista en ingeniería informática. Este documento se encuentra en el Anexo K.

El segundo indicador objetivamente verificable es el reporte de apreciación del módulo de visualización del análisis de agrupamiento por un especialista en bioinformática o biología molecular. Este documento se encuentra en el Anexo L.

5.3 Discusión

Para solucionar el problema que no se cuenta con estudios publicados en el Perú sobre el análisis de agrupamiento de las secuencias genómicas de las variantes de SARS-COV-2, con la posibilidad de incorporar nuevas secuencias sin la necesidad de rehacer la representación y el análisis, se obtuvieron tres resultados.

En primer lugar, se realizó la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados. Estos métodos de análisis de agrupamiento fueron el método de agrupamiento *k-means*, el método jerárquico y el método de agrupamiento espacial DBSCAN, los cuales se seleccionaron debido a que son los más utilizados y han presentado mejores resultados al realizar el agrupamiento de las secuencias genómicas SARS-CoV-2. Además, porque ofrecen distintas formas de encontrar los grupos, permitiendo así que el usuario utilice cualquiera de estos tres algoritmos para un mejor análisis de agrupamiento.

En segundo lugar, se implementó los métodos de análisis de agrupamiento y los servicios a utilizar en el módulo de agrupamiento de las secuencias genómicas SARS-CoV-2, estos incluyen el gráfico de dispersión del agrupamiento con el algoritmo *k-means*, el algoritmo jerárquico y el algoritmo DBSCAN para mostrar las variantes identificadas en el Perú, el gráfico del dendrograma para visualizar la estructura o jerarquía de los clústeres que se pueden formar y la visualización de los datos de las secuencias genómicas.

En tercer lugar, se definieron los requerimientos necesarios para el desarrollo del módulo de agrupamiento. Una vez definidos, y con los servicios ya implementados en el anterior resultado, se procedió a generar las representaciones visuales de los servicios. Con estos gráficos se comprende mejor el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 en el Perú.

Capítulo 6. Implementación de vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento

6.1 Introducción

En el presente capítulo, se presentan los resultados alcanzados para lograr el cumplimento del tercer objetivo específico, el cual es “Implementar las vistas con capacidades interactivas del módulo espacio-temporal y del módulo de agrupamiento para que el usuario interactúe con los módulos de software”. A continuación, se detalla los resultados alcanzados los cuales son: la lista de requerimientos para realizar la interactividad de las representaciones visuales y las interfaces del módulo para visualizar la representación espacio-temporal y el análisis de agrupamiento con capacidades interactivas. Además, se presentan los medios de verificación y los indicadores objetivamente verificables con los cuales se va a validar el cumplimiento de cada resultado alcanzado.

6.2 Resultados alcanzados

6.2.1 Lista de requerimientos a considerar para realizar la interactividad de las representaciones visuales.

En este resultado se identifica todos los requerimientos relacionados a la interactividad de las representaciones visuales de la herramienta analítica interactiva, con el fin de que los especialistas y analistas puedan analizar las secuencias genómicas de SARS-CoV-2 de manera sencilla e interactiva y realizar un análisis con mayor profundidad y entendimiento.

En la Tabla 22, se presenta el nivel de prioridad y su descripción de lo que significa cada nivel. Los requerimientos que tengan un nivel de prioridad “Alta” serán realizados primero, posteriormente con los de prioridad “Media”; y, por último, con los de prioridad “Baja”.

Tabla 22. Tabla de prioridad

Nivel de Prioridad	Descripción
1	Alta
2	Media
3	Baja

Nota. Elaboración propia.

En la Tabla 23, se presenta la lista de requerimientos con su tipo de requerimiento, el nivel prioridad y si es exigible o deseable.

Tabla 23. Lista de requerimientos para realizar la interactividad de las representaciones visuales.

N°	Requerimiento	Tipo	Nivel de Prioridad	Exigible / Deseable
1	La herramienta permitirá mover y realizar zoom a los gráficos de cada módulo.	Funcional	1	Exigible
2	La herramienta permitirá guardar o descargar los gráficos de cada módulo.	Funcional	1	Exigible
3	La herramienta permitirá recargar el gráfico, para así volver a la visualización inicial en caso de hacerle zoom o algún movimiento a los gráficos.	Funcional	2	Deseable
4	La herramienta permitirá activar o desactivar la opción de mostrar el detalle en los gráficos de cada módulo.	Funcional	2	Deseable
5	La herramienta permitirá deshacer o rehacer los cambios realizados en la interacción con los gráficos de cada módulo.	Funcional	1	Exigible
6	El módulo de agrupamiento mostrará una barra de selección de número de clústeres para filtrar por clústeres en los algoritmos *k*-	Funcional	2	Deseable

	means y jerárquico; y en el caso del algoritmo DBSCAN se filtrará por el valor de épsilon.			
7	La herramienta permitirá incorporar nuevas secuencias genómicas SARS-CoV-2 sin rehacer todo el análisis.	Funcional	2	Exigible
8	La herramienta permitirá al usuario administrador eliminar las secuencias genómicas SARS-CoV-2.	Funcional	2	Deseable
9	La herramienta permitirá al usuario administrador subir las secuencias genómicas en un archivo FASTA. y .TSV.	No funcional	1	Exigible
10	La herramienta muestra mensajes de error entendibles expresados en lenguaje plano sin códigos de error e indicando claramente el problema.	No funcional	1	Exigible
11	El diseño de la interfaz de la herramienta emplea íconos representativos que se asocian correctamente con las acciones que puede realizar el usuario.	No funcional	1	Exigible

Nota. Elaboración propia.

Por otro lado, el indicador objetivamente verificable para este resultado contiene la validación del 100% de los requerimientos por parte de un especialista en bioinformática o biología molecular. Este documento se encuentra en el Anexo M.

6.2.2 Interfaces del módulo para visualizar la representación espacio-temporal (RE.1.3) y del módulo para visualizar el análisis de agrupamiento (RE.2.3) con capacidades interactivas de acuerdo con los requerimientos especificados.

En este resultado se realiza el desarrollo de las interfaces de los módulos espacio-temporal y agrupamiento con capacidades interactivas de acuerdo a los requerimientos

especificados. Para lograr esto, se propone como medio de verificación, el código fuente de la interactividad realizada en estos módulos.

En cuanto al medio de verificación, primero se tiene la programación de los servicios del backend, para el cual se ha usado el *framework* de Python, FastApi. La mayoría de estos servicios son los desarrollados en los módulos espacio-temporal y agrupamiento, explicados en el capítulo 4 y 5 respectivamente, con la diferencia que se ha agregado funciones de la librería Bokeh para lograr la interactividad. Estas funciones consisten en poder realizar zoom a los gráficos, mover las vistas, guardar o descargar, recargar el gráfico, deshacer o rehacer los cambios realizados en los gráficos, y activar o desactivar la opción de mostrar el detalle en los gráficos. Asimismo, para los gráficos de los algoritmos *k-means* y jerárquico se realiza la opción de seleccionar el número de clúster, y así poder filtrar por clúster en cada algoritmo; y para el algoritmo DBSCAN, se realiza la opción de filtrar por el valor del parámetro épsilon.

Los otros servicios requeridos para realizar la interactividad son los siguientes:

- Servicio que realice la incorporación de nuevas secuencias genómicas SARS-CoV-2, brindando la opción de realizar todo el agrupamiento nuevamente con estas secuencias incorporadas o de lo contrario, solo incorporar las secuencias sin rehacer todo el agrupamiento.
- Servicio que liste todas las secuencias genómicas SARS-CoV-2.
- Servicio que realice la eliminación de secuencias genómicas SARS-CoV-2.

El primer servicio recibe como parámetros un archivo .FASTA y un archivo .TSV y un valor que indica si se realiza el agrupamiento con las nuevas secuencias genómicas SARS-CoV-2 o no. El segundo servicio no recibe parámetros solo devuelve la lista de todas las secuencias genómicas SARS-CoV-2 registradas en la base de datos. Por último, el tercer servicio recibe como parámetros los identificadores de acceso de las secuencias genómicas (identificador en la base de datos de GISAID) a eliminar.

Con respecto al primer servicio, el proceso para incorporar nuevas secuencias genómicas SARS-CoV-2 sin rehacer todo el análisis, es el siguiente:

1. En primer lugar, se recuperan los archivos del preprocesamiento guardados en la base de datos. Estos archivos son: la matriz de las secuencias pre procesadas anteriormente, la matriz de distancia de las secuencias genómicas SARS-CoV-2, el arreglo con los puntos de referencia; y, el modelo de la red neuronal entrenado.
2. Se leen los nuevos archivos, el de extensión .FASTA y el de extensión .TSV para obtener los datos de las nuevas secuencias genómicas SARS-CoV-2 incorporadas.
3. Se obtiene el identificador de acceso, la fecha de recolección y el lugar de donde se obtuvo cada secuencia genómica SARS-CoV-2. Se consideran solo las secuencias que pertenecen o hacen mención a un departamento del Perú.
4. Se eliminan las secuencias con errores de lectura, definido como letras distintas de A, C, G y T.
5. Se obtiene el linaje pango de las nuevas secuencias genómicas SARS-CoV-2.
6. Luego se realiza el alineamiento múltiple de las nuevas secuencias genómicas SARS-CoV-2 para conocer los lugares en donde, en una misma posición para todas las secuencias, se tiene diferentes nucleótidos. Para realizar este alineamiento se utiliza la función *MultipleSeqAlignment* de la librería Biopython. Para este alineamiento de las nuevas secuencias genómicas se utiliza matriz de las secuencias pre procesadas anteriormente.
7. Se calcula la matriz de distancia Hamming de todas las secuencias genómicas SARS-CoV-2. Para ello, se utiliza la matriz de secuencias anterior y la de las nuevas secuencias, para así calcular la matriz de distancias de cada secuencia con respecto a las demás.
8. Luego se calcula la distancia de cada punto, de las nuevas secuencias genómicas SARS-CoV-2, con respecto a cada punto de referencia (*Landmark*).

9. Con el modelo de la red neuronal recuperado se predice los valores de la representación de las secuencias genómicas SARS-CoV-2 en dos dimensiones para su visualización.
10. Dependiendo de la opción elegida por el usuario se realiza el agrupamiento con todas las secuencias genómicas o solo se incorporan sin rehacer el agrupamiento. En el caso de elegir la opción de rehacer todo el agrupamiento, se procede a eliminar el agrupamiento realizado anteriormente y seguidamente, se realiza el agrupamiento con todas las secuencias genómicas SARS-CoV-2 para todos los algoritmos (*k-means*, jerárquico y DBSCAN). Si se elige la opción de solo incorporar las secuencias, se guardan los datos de la secuencia con la variante identificada en la base de datos GISAID, sin realizar ningún agrupamiento.

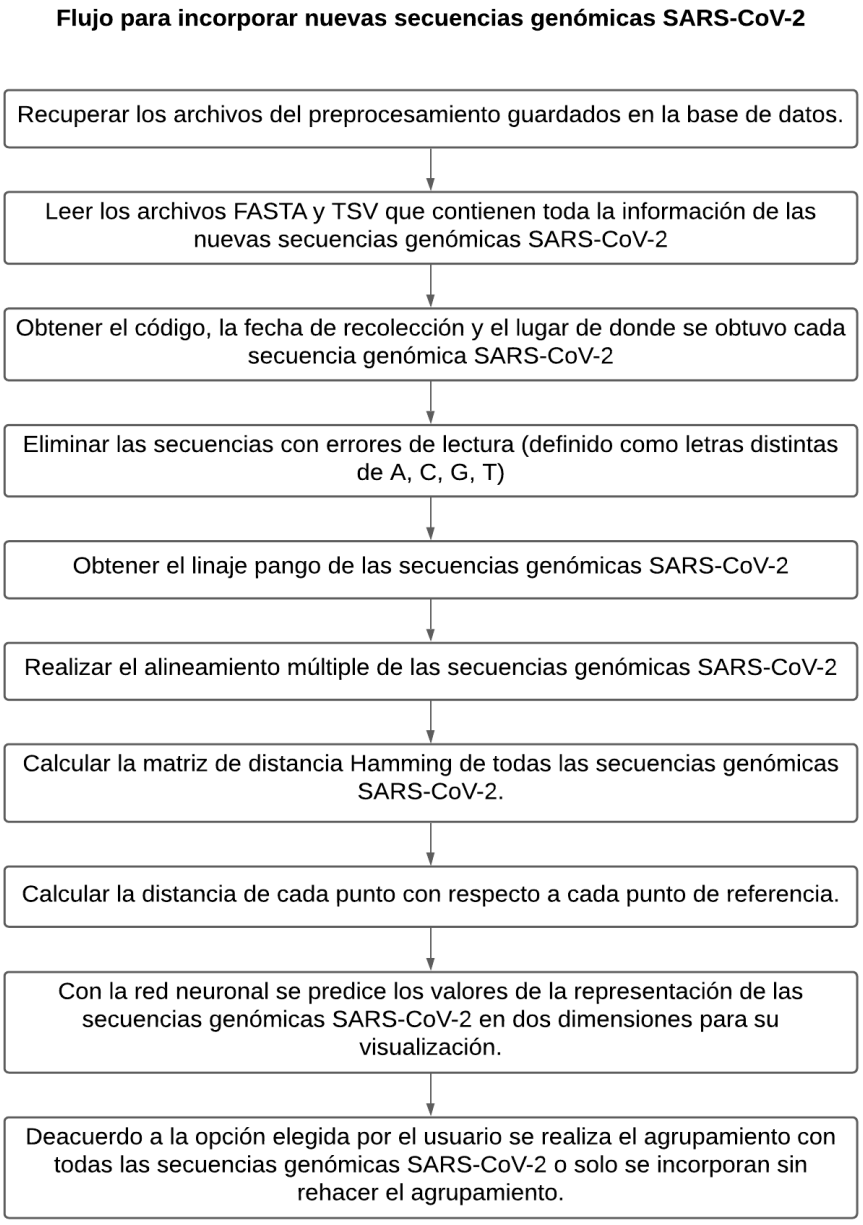


Figura 28. Flujo del servicio para incorporar nuevas secuencias genómicas SARS-CoV-2 al análisis.

El código fuente del backend del módulo del presente objetivo específico se encuentra en el siguiente repositorio:

https://github.com/CarolinaMejiaMujica/Interactividad-Backend

Con respecto se tiene el código fuente desarrollado para el frontend, para ello, se utilizó el *framework* React.

En la Figura 29, se muestra las acciones de realizar zoom a los gráficos, mover las vistas, guardar o descargar, recargar el gráfico, deshacer o rehacer los cambios realizados en los gráficos, y activar o desactivar la opción de mostrar el detalle en los gráficos.

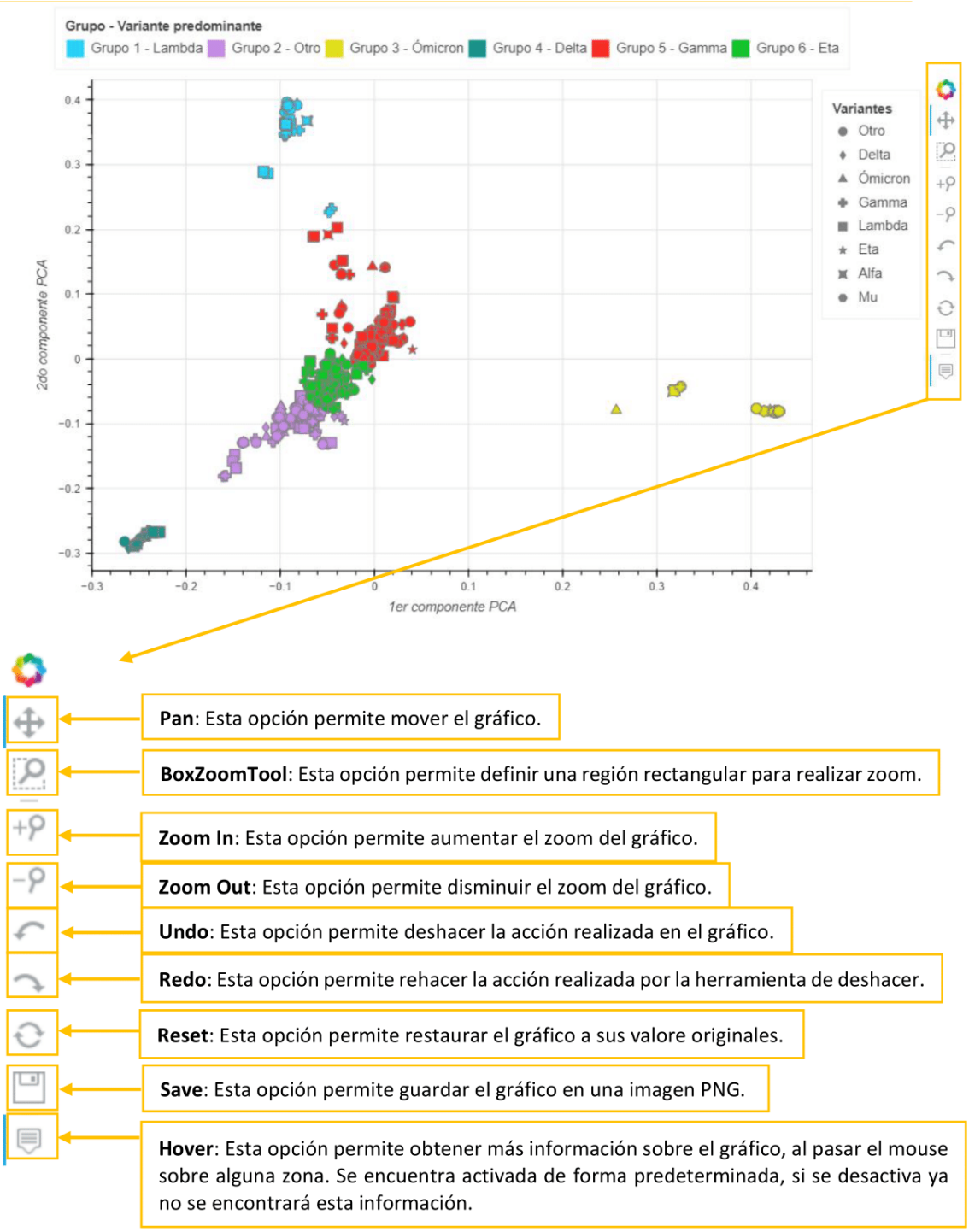


Figura 29. Captura de la pantalla con las acciones que se puede realizar en los gráficos.

En la Figura 30, se muestra la barra de selección de número de clústeres para filtrar por clúster en los algoritmos *k-means* y jerárquico; y en el caso del algoritmo DBSCAN se filtrará por el valor de épsilon.

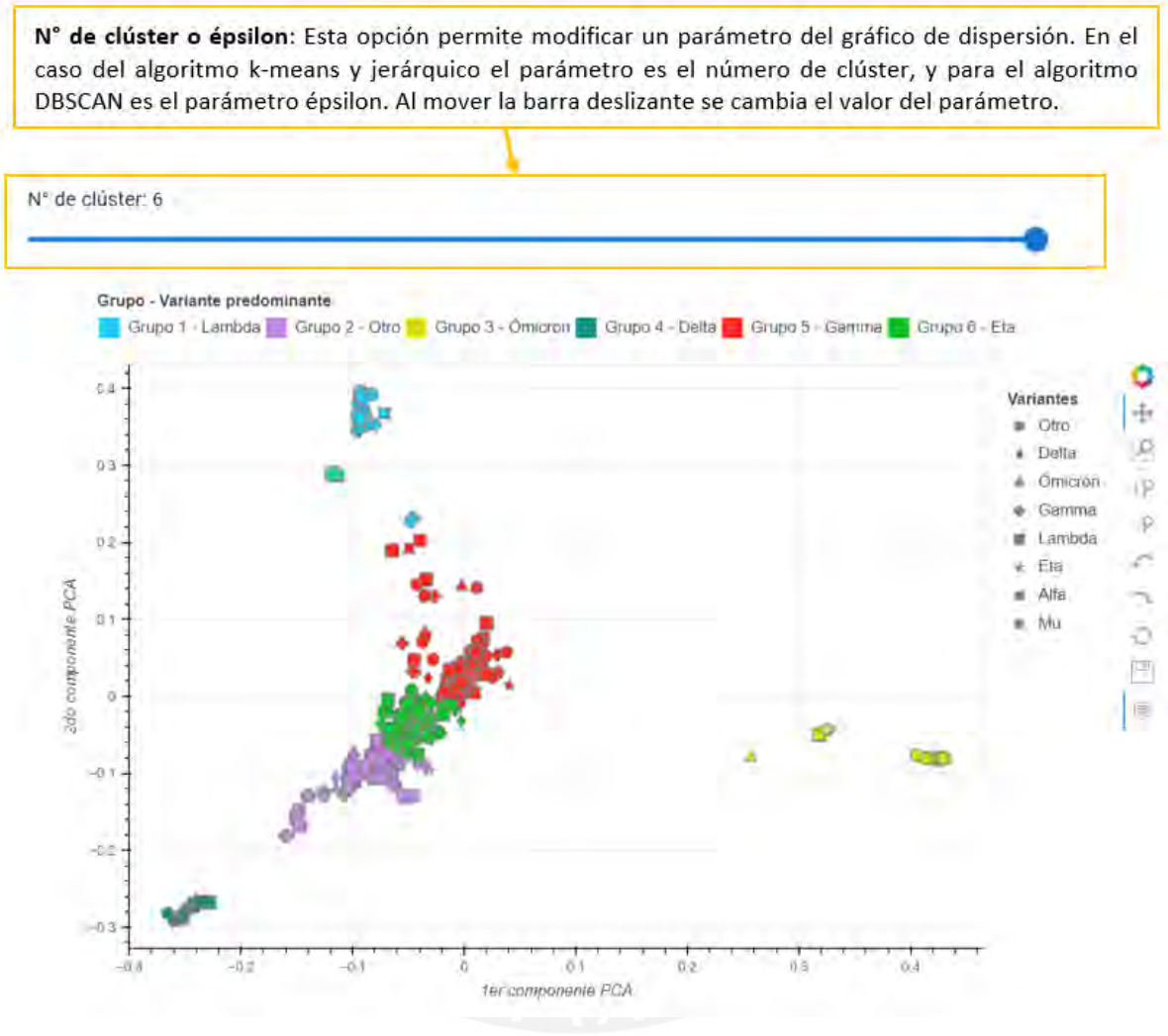


Figura 30. Captura de la pantalla con la barra de selección de número de clústeres o valor de épsilon dependiendo del algoritmo seleccionado.

En la Figura 31, se muestra la pantalla donde se importa los datos de las secuencias genómicas SARS-CoV-2, se debe importar dos archivos, un archivo .fasta y otro archivo .tsv, estos archivos se obtienen de la base de datos GISAID. Además, se debe seleccionar una opción para el agrupamiento, una de ellas es realizar el agrupamiento de todas las secuencias genómicas SARS-CoV-2 nuevamente, es decir, agrupar las nuevas secuencias agregadas con

las anteriores registradas en el sistema; y la otra opción es no realizar el agrupamiento con las nuevas secuencias agregadas, solo ubicarlas visualmente en los gráficos para una mayor comprensión, es decir, que estas nuevas secuencias no pertenecerán a ningún clúster en específico.

Figura 31. Captura de la pantalla para importar los datos de las secuencias genómicas SARS-CoV-2.

En la Figura 32, se presenta la pantalla para eliminar las secuencias genómicas SARS-CoV-2 donde se tiene la opción de seleccionar las secuencias genómicas que se desean eliminar.

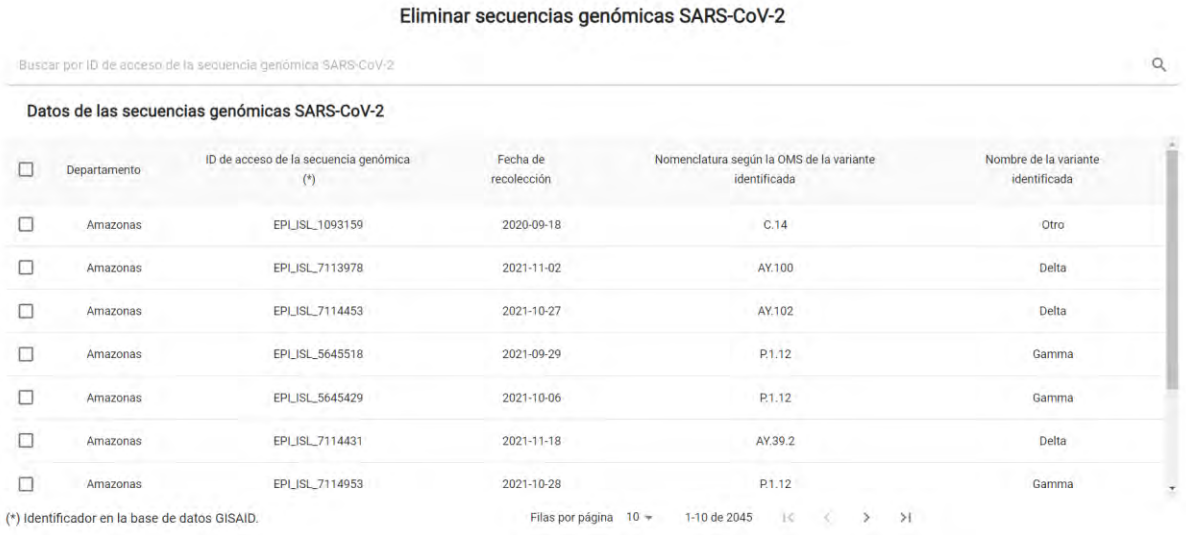


Figura 32. Captura de la pantalla para eliminar las secuencias genómicas SARS-CoV-2.

El código fuente del frontend del presente objetivo específico se encuentra en el siguiente repositorio: https://github.com/CarolinaMejiaMujica/Frontend-Tesis

Por otro lado, el indicador objetivamente verificable para este resultado contiene la validación de que se ha implementado los requerimientos de usabilidad especificados por parte de un especialista en interacción humano computador (HCI). Este documento se encuentra en el Anexo N.

6.3 Discusión

Para solucionar el problema que la mayoría de recursos y librerías para el análisis de diversidad genómica son de uso complejo y generan mayormente vistas estáticas, no permitiendo al usuario interactuar con las vistas y resultados, se obtuvieron tres resultados.

En primer lugar, se definieron los requerimientos necesarios para realizar la interactividad de las representaciones visuales. Con estos requerimientos definidos se procede a realizar la implementación de las mismas para así lograr la interactividad en la herramienta para que así ayude en el análisis de las secuencias genómicas SARS-CoV-2.

En segundo lugar, se realizó las interfaces del módulo para visualizar la representación espacio-temporal y del módulo para visualizar el análisis de agrupamiento con capacidades interactivas de acuerdo con los requerimientos especificados. Además, en este resultado se implementaron dos servicios, el primer servicio es sobre la incorporación de nuevas secuencias genómicas SARS-CoV-2, y el otro servicio realiza la eliminación de secuencias genómicas SARS-CoV-2.

Capítulo 7. Conclusiones y trabajos futuros

7.1 Conclusiones

Dado que actualmente (20/01/2022) la salud pública mundial se ve afectada por la pandemia que se vive a consecuencia del Covid-19, y que este virus evoluciona aceleradamente presentando constantes variantes genómicas a raíz de su alta capacidad de transmisión; la comunidad académica, empresas farmacéuticas y autoridades sanitarias están interesados en entender la diversidad de secuencias genómicas obtenidas de las personas contagiadas de este virus, para así tener un mejor conocimiento de la propagación del virus SARS-CoV-2.

Por ello, teniendo en consideración este contexto, se llega a identificar la problemática la cual es la necesidad de realizar una analítica avanzada que incluya representaciones visuales interactivas y analíticas de agrupamiento en el espacio y a lo largo del tiempo de la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú a fin de apoyar la vigilancia genómica.

En base a esta problemática y a fin de brindar una solución, en el presente proyecto de investigación se desarrolló una herramienta analítica interactiva para representar visualmente la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú. Para lograr esto se planteó tres objetivos, en primer lugar, se desarrolló un módulo que permita realizar una representación visual espacio-temporal de las secuencias genómicas SARS-CoV-2; en segundo lugar, se desarrolló otro módulo que permita realizar un análisis de agrupamiento con capacidad de incorporar nuevas secuencias de forma en línea; y, por último, se implementó las vistas con capacidades interactivas del módulo espacio-temporal y de agrupamiento.

En el primer objetivo, se elaboró la estructura de la base de datos de toda la herramienta, luego se implementó el módulo para realizar el preprocesamiento de los datos de las secuencias genómicas SARS-CoV-2, en el cual se realizó la reducción dimensional de estas secuencias para así tener una mejor representación de los datos en una menor dimensión. Por último, se

desarrolló el módulo de la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú. En base a ello, se concluye que se alcanzó el primer objetivo específico.

Para el segundo objetivo, fue necesario realizar la selección de métodos de análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 a implementar, para esta selección se basó en los métodos utilizados y los resultados alcanzados en diversos artículos obtenidos en la revisión sistemática. Luego de la selección, se implementaron los métodos de agrupamiento seleccionados, que fueron el algoritmo *k-means*, jerárquico y DBSCAN.

Para el tercer objetivo, se realizó la lista de requerimientos a considerar al momento de realizar la interactividad de las representaciones visuales de los dos módulos anteriormente mencionados, en estos requerimientos se consideraron aspectos de interactividad que no se encontraron en los estudios de la revisión sistemática y que sirven para una mejor compresión de la evolución del virus SARS-CoV-2.

7.2 Trabajos futuros

En primer lugar, dado que se dispone de información sobre las secuencias genómicas de SARS-CoV-2 en diferentes repositorios de base de datos es viable considerar como un trabajo futuro el integrar la herramienta con estas bases de datos. De forma que sea posible ir actualizando los datos de las secuencias genómicas conforme se vaya actualizando esta información en la base de datos.

En segundo lugar, también es posible considerar como trabajo futuro el realizar el alineamiento de las secuencias genómicas SARS-CoV-2 a partir de los genes codificantes de la proteína spike de cada secuencia, ya que los especialistas en virología comparan la proteína spike en todas las secuencias genómicas SARS-CoV-2 para así detectar en que aminoácidos se produce los cambios entre las secuencias, identificando así las variantes del virus SARS-CoV-2.

Por otro lado, con respecto al preprocesamiento de las secuencias, es viable aplicar otras técnicas de reducción dimensional y técnicas proyectivas. De esta manera, se puede obtener una diferente representación de las secuencias genómicas SARS-CoV-2 en una menor dimensionalidad, para así realizar el agrupamiento y la visualización de las mismas. Además, se puede comparar los resultados obtenidos con otras dimensionalidades intermedias aparte de 10 dimensiones, para así visualizar los diferentes resultados obtenidos en el preprocesamiento y analizar cuál sería la mejor solución que capture la mayor información de los datos en una menor dimensión.

Finalmente, esta herramienta fue desarrollada para el análisis de la evolución del virus SARS-CoV-2 en el Perú, sin embargo, este análisis es necesario realizar en diferentes países a nivel mundial afectados por este virus para así entender la evolución y transmisión del virus SARS-CoV-2. Por ello, esta herramienta puede soportar los datos de secuencias genómicas SARS-CoV-2 de otros países, de esta forma se puede ampliar el análisis a secuencias recolectadas en otros países.

Referencias

Abirami, S., & Chitra, P. (2020). *Energy-efficient edge based real-time healthcare support system*. The Digital Twin Paradigm for Smarter Systems and Environments: The Industry Use Cases, 339–368. doi:10.1016/bs.adcom.2019.09.007

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1). https://doi.org/10.1088/1742-6596/1142/1/012012

Amazon Web Services (2021). *About AWS*. Recuperado el 9 de junio de https://aws.amazon.com/about-aws/

Bawono, P., & Heringa, J. (2014). Phylogenetic Analyses. In *Comprehensive Biomedical Physics* (Vol. 6, pp. 93–110). Elsevier. https://doi.org/10.1016/B978-0-444-53632-7.01108-4

Bielińska-Wąż, D., & Wąż, P. (2021). Non-standard bioinformatics characterization of SARS-CoV-2. *Computers in Biology and Medicine*, *131*, 104247. https://doi.org/10.1016/j.compbiomed.2021.104247

Biopython (2021). *Biopython - Python Tools for Computational Molecular Biology*. Recuperado el 9 de junio de 2021 de https://biopython.org/

Blanco, A., & Blanco, G. (2017). Nucleic Acids. In *Medical Biochemistry* (pp. 121–140). Elsevier. https://doi.org/10.1016/B978-0-12-803550-4.00006-9

Bokeh (s.f.). *The Bokeh Visualization Library*. Recuperado el 9 de junio de 2021 de https://bokeh.org/

Borg, Ingwer., & Groenen, P. J. F. (2005). *Modern multidimensional scaling : theory and applications*. Springer.

Córdova-Aguilar, A., Rossani, G. A., & Revisión, A. de. (2020). COVID-19: REVISIÓN DE LA LITERATURA Y SU IMPACTO EN LA REALIDAD SANITARIA PERUANA.

Revista de La Facultad de Medicina Humana URP, *20*(3), 471–477. https://doi.org/10.25176/RFMH.v20i3.2984

Cucinotta, D., & Vanelli, M. (2020). WHO Declares COVID-19 a Pandemic. *Acta Biomed*, *91*, 157–160. https://doi.org/10.23750/abm.v91i1.9397

De Silva, V., & Tenenbaum, J. B. (2004). *Sparse multidimensional scaling using landmark points*. Technicalreport, Technical report, Stanford University.

Dey, T., Chatterjee, S., Manna, S., Nandy, A., & Basak, S. C. (2021). Identification and computational analysis of mutations in SARS-CoV-2. *Computers in Biology and Medicine*, *129*, 104166. https://doi.org/10.1016/j.compbiomed.2020.104166

Diday, E., & Simon, J. C. (1976). *Clustering Analysis*. https://doi.org/https://doi.org/10.1007/978-3-642-96303-2_3

Elbe, S., & Buckland-Merrett, G. (2017). Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges*, *1*(1), 33–46. https://doi.org/10.1002/gch2.1018

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Retrieved from www.aaai.org

FastAPI (s.f.). *FastAPI*. Recuperado el 9 de junio de 2021 de https://fastapi.tiangolo.com/

Feghali, R., Merhi, G., Kwasiborski, A., Hourdel, V., Ghosn, N., & Tokajian, S. (2021). Genomic characterization and phylogenetic analysis of the first SARS-CoV-2 variants introduced in Lebanon. *PeerJ*, *9*, 19. https://doi.org/10.7717/peerj.11015

Georgieva, Petia., Mihaylova, Lyudmila., Jain, L. C., & (Eds.). (2013). *Advances in Intelligent Signal Processing and Data Mining* (P. Georgieva, L. Mihaylova, & L. C. Jain, Eds.; Vol. 410). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-28696-4

GISAID (2022). *GISAID*. Recuperado el 20 de enero de 20222 de https://www.gisaid.org/

Grant, Robert. (2019). *Data Visualization Charts, Maps and Interactive Graphics*. https://www.crcpress.com/go/asacrc

Hahn, G., Lee, S., Weiss, S. T., & Lange, C. (2021). Unsupervised cluster analysis of SARS-CoV-2 genomes reflects its geographic progression and identifies distinct genetic subgroups of SARS-CoV-2 virus. *Genetic Epidemiology*. https://doi.org/10.1002/gepi.22373

Han, J., Kamber, M., & Pei, J. (2012a). Cluster Analysis: Basic Concepts and Methods. In *Data Mining* (pp. 443–495). Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00010-1

Han, J., Kamber, M., & Pei, J. (2012b). Data Preprocessing. In *Data Mining* (pp. 83–124). Elsevier. https://doi.org/10.1016/b978-0-12-381479-1.00003-4

Han, J., Kamber, M., & Pei, J. (2012c). Data Preprocessing. In *Data Mining* (pp. 83–124). Elsevier. https://doi.org/10.1016/b978-0-12-381479-1.00003-4

Hood, L., & Rowen, L. (2013). The human genome project: Big science transforms biology and medicine. *Genome Medicine*, *5*(9). https://doi.org/10.1186/gm483

Hozumi, Y., Wang, R., Yin, C., & Wei, G. W. (2021). UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets. *Computers in Biology and Medicine*, *131*, 104264. https://doi.org/10.1016/j.compbiomed.2021.104264

Huang, R., Liu, M., & Ding, Y. (2020). Coronavirus Pandemic Spatial-temporal distribution of COVID-19 in China and its prediction: A data-driven modeling analysis. *The Journal of Infection in Developing Countries - JIDC*, *14*(3), 246–253. https://doi.org/10.3855/jidc.12585

Instituto Nacional de Salud y Centro Nacional de Epidemiología, Prevención y Control de Enfermedades – MINSA (2021). *Sala situacional COVID-19 Perú*. Recuperado el 11 de noviembre de 2021 de https://covid19.minsa.gob.pe/sala_situacional.asp

Islam, A., Sayeed, M. A., Rahman, M. K., Ferdous, J., Shariful Islam, |, Mohammad, |, & Hassan, M. (2021). Geospatial dynamics of COVID-19 clusters and hotspots in Bangladesh. *Transbound Emerg Dis*, *00*, 1–15. https://doi.org/10.1111/tbed.13973

Khailany, R. A., Safdar, M., & Ozaslan, M. (2020). Genomic characterization of a novel SARS-CoV-2. *Gene Reports*, *19*. https://doi.org/10.1016/j.genrep.2020.100682

Lauring AS, Hodcroft EB. (2021). Genetic Variants of SARS-CoV-2—What Do They Mean?. *JAMA.* 2021;325(6):529–531. https://doi.org/10.1001/jama.2020.27124

Leaflet (2010). *Leaflet - an open-source JavaScript library for mobile-friendly interactive maps*. Recuperado el 9 de junio de 2021 de https://leafletjs.com/index.html

Loayza Alarico, M. J., & de La Cruz Vargas, J. A. (2021). Effect of SARS-CoV-2 variants on the transmission of COVID-19 in Peru. *Revista de La Facultad de Medicina Humana*, *21*(1), 10–11. https://doi.org/10.25176/rfmh.v21i1.3606

Ludwig, S. P., & Zarbock, A. M. (2020). *Coronaviruses and SARS-CoV-2 A Brief Overview*. https://doi.org/10.1213/ANE.0000000000004845

Meneses Escobar, C. A., Vanessa, L., Murillo, R., & Soto, J. F. (2011). Tecnologías bioinformáticas para el análisis de secuencias de ADN. *Scientia et Technica Año XVI*, *49*.

Minchin, S., & Lodge, J. (2019). Understanding biochemistry: structure and function of nucleic acids. *Essays in Biochemistry*, *63*, 433–456. https://doi.org/10.1042/EBC20180038

Morais, I. J., Polveiro, R. C., Medeiros Souza, G., Bortolin, D. I., Sassaki, T., Talis, A., & Lima, M. (2020). The global population of SARS-CoV-2 is composed of six major subtypes. *Scientific Reports*. https://doi.org/10.1038/s41598-020-74050-8

Nexstrain (s.f.). *Genomic epidemiology of novel coronavirus - Global subsampling*. Recuperado el 4 de setiembre de 2021 de https://nextstrain.org/ncov/gisaid/global

NumPy (2019). *ABOUT US*. Recuperado el 9 de junio de 2021 de https://numpy.org/about/

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in PythonJournal of Machine Learning Research, 12, 2825–2830.

PostgreSQL (1996). *PostgreSQL: The World's Most Advanced Open Source Relational Database*. Recuperado el 9 de junio de 2021 de https://www.postgresql.org/about/

Pourghasemi, H. R., Pouyan, S., Heidari, B., Farajzadeh, Z., Fallah Shamsi, S. R., Babaei, S., Khosravi, R., Etemadi, M., Ghanbarian, G., Farhadi, A., Safaeian, R., Heidari, Z., Tarazkar, M. H., Tiefenbacher, J. P., Azmi, A., & Sadeghian, F. (2020). Spatial modeling, risk mapping, change detection, and outbreak trend analysis of coronavirus (COVID-19) in Iran (days between February 19 and June 14, 2020). *International Journal of Infectious Diseases*, *98*, 90–108. https://doi.org/10.1016/j.ijid.2020.06.058

Python Software Foundation (2001). *What is Python? Executive Summary*. Recuperado el 9 de junio de 2021 de https://www.python.org/doc/essays/blurb/

React (s.f.). *React - A JavaScript library for building user interfaces*. Recuperado el 9 de junio de 2021 de https://reactjs.org/

Sainlez, M., & Heyen, G. (2011). *Recurrent neural network prediction of steam production in a Kraft recovery boiler.* 21st European Symposium on Computer Aided Process Engineering, 1784–1788. doi:10.1016/b978-0-444-54298-4.50135-5

Santos, G. (2021). “Un escenario de riesgo para nuevas variantes es tener a mucha gente contagiada a la vez”. *Ojo Público*. Recuperado de https://ojo-publico.com/2665/mucha-gente-contagiada-la-vez-trae-riesgos-para-nuevas-variantes

Saraswathy, N., & Ramalingam, P. (2011). Genome sequencing methods. In *Concepts and Techniques in Genomics and Proteomics* (pp. 95–107). Elsevier. https://doi.org/10.1533/9781908818058.95

SciPy (s.f.). *SciPy.org*. Recuperado el 9 de junio de 2021 de https://www.scipy.org/

Serrano, C. (2021). "Variantes del coronavirus: por qué la escasa vigilancia del virus en América Latina puede convertirse en un problema global". *BBC News Mundo*. Recuperado de https://www.bbc.com/mundo/noticias-56580755

Shishir, T. A., Bin, I., Id, N., & Faruque, S. M. (2021). In silico comparative genomics of SARS-CoV-2 to determine the source and diversity of the pathogen in Bangladesh. *PLOS ONE*. https://doi.org/10.1371/journal.pone.0245584

Shu, Y., & Mccauley, J. (2017). GISAID: Global initiative on sharing all influenza data-from vision to reality. *Euro Surveill*, 1. https://doi.org/10.2807/1560-7917.ES.2017.22.13.30494

Soto, G. P. (2020). Bases Genéticas y Moleculares del COVID-19 (SARS-CoV-2). Mecanismos de Patogénesis y de Respuesta Inmune. In *Int. J. Odontostomat* (Vol. 14, Issue 3). https://doi.org/http://dx.doi.org/10.4067/S0718-381X2020000300331

TensorFlow (s.f.). *Una plataforma de aprendizaje automático de código abierto de extremo a extremo*. Recuperado el 5 de noviembre de 2021 de https://www.tensorflow.org/

Tharwat, A. (2016). Principal component analysis-a tutorial. In *Int. J. Applied Pattern Recognition* (Vol. 3, Issue 3).

Thomas, J. J., & Cook, K. A. (2005). *Illuminating the Path: The Research and Development Agenda for Visual Analytics*.

Toyoshima, Y., Nemoto, • Kensaku, Matsumoto, S., Nakamura, Y., & Kiyotani, K. (2020). SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. *Journal of Human Genetics*, *65*, 1075–1082. https://doi.org/10.1038/s10038-020-0808-9

Urhan, A., & Abeel, T. (2021). Emergence of novel SARS-CoV-2 variants in the Netherlands. *Scientific Reports* |, *11*, 6625. https://doi.org/10.1038/s41598-021-85363-7

Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. In *Journal of Machine Learning Research* (Vol. 9).

Vergara, Í. (2021). "Científicos rastrean posible variante peruana del covid-19 en el país". *Andina*. Recuperado de https://andina.pe/agencia/noticia-cientificos-rastrean-posible-variante-peruana-del-covid19-el-pais-837522.aspx

Wang, B., & Jiang, L. (2021). Principal Component Analysis Applications in COVID-19 Genome Sequence Studies. *Cognitive Computation*, *1*, 3. https://doi.org/10.1007/s12559-020-09790-w

Wang, B., & Kennedy, M. A. (2014). Principal components analysis of protein sequence clusters. *J Struct Funct Genomics 15*, 1–11. https://doi.org/10.1007/s10969-014-9173-2

Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., & Wei, G.-W. (2021). Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. *COMMUNICATIONS BIOLOGY*, *4*. https://doi.org/10.1038/s42003-021-01754-6

Wang, Y., Liu, Y., Struthers, J., & Lian, M. (2021). Spatiotemporal Characteristics of the COVID-19 Epidemic in the United States. *Clinical Infectious Diseases : An Official Publication of the Infectious Diseases Society of America*, *72*(4), 643–651. https://doi.org/10.1093/cid/ciaa934

Wenjie Tan, Xiang Zhao, Xuejun Ma, Wenling Wang, Peihua Niu, Wenbo Xu, George F. Gao, & Guizhen Wu. (2020). A Novel Coronavirus Genome Identified in a Cluster ofPneumonia Cases — Wuhan, China 2019−2020. *China CDC Weekly*, *2*(4), 61–62. https://doi.org/10.46234/ccdcw2020.017

Wisesty, U. N., & Mengko, T. R. (2021). Comparison of dimensionality reduction and clustering methods for sars-cov-2 genome. *Bulletin of Electrical Engineering and Informatics*, *10*(4), 2170–2180. https://doi.org/10.11591/EEI.V10I4.2803

World Health Organization (2021). *WHO Coronavirus (COVID-19) Dashboard*. Recuperado el 11 de noviembre de 2021 de https://covid19.who.int/

Yang, H.-C., Chen, C.-H., Wang, J.-H., Liao, H.-C., Yang, C.-T., Chen, C.-W., Lin, Y.-C., Kao, C.-H., Lu, M.-Y. J., & Liao, J. C. (2020). Analysis of genomic distributions of SARS-CoV-2 reveals a dominant strain type with strong allelic associations. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.2007840117/-/DCSupplemental

Yang, X., Dong, N., Wai-Chi Chan, E., & Chen, S. (2020). *Genetic cluster analysis of SARS-CoV-2 and the identification of those responsible for the major outbreaks in various countries*. https://doi.org/10.1080/22221751.2020.1773745

Zhang, J., Yao, Y., He, H., & Shen, J. (2020). Clinical Interpretation of Sequence Variants. *Current Protocols in Human Genetics*, *106*(1). https://doi.org/10.1002/cphg.98

Zhao Id, Z., Sokhansanj Id, B. A., Malhotra Id, C., Zheng, K., & Rosenid, G. L. (2020). Genetic grouping of SARS-CoV-2 coronavirus sequences using informative subtype markers for pandemic spread visualization. *Plos Computational Biology*. https://doi.org/10.1371/journal.pcbi.1008269

Anexos

Anexo A: Plan de Proyecto

En este anexo se detalla el plan de proyecto del presente trabajo de investigación. Se describe la justificación, viabilidad, alcance, limitaciones y riesgos del proyecto. Asimismo, se detallará la lista de tareas y se realizará un cronograma para la ejecución de estas. También, se establecerá la estructura de descomposición del trabajo, la lista de recursos y finalmente, el costo total del proyecto.

- **Justificación**

Actualmente (20/01/2021), la salud pública mundial se ve afectada por la pandemia que se vive a consecuencia del SARS-CoV-2, llevando ya más de un año en esta situación. Este virus evoluciona aceleradamente presentando constantes variantes genómicas a raíz de su alta capacidad de transmisión, las variantes tienen un alto nivel de contagio y un alto grado de letalidad (Loayza Alarico & de La Cruz Vargas, 2021). Dada esta situación, la comunidad académica, empresas farmacéuticas y autoridades sanitarias están interesados en entender la diversidad de secuencias genómicas obtenidas de las personas contagiadas de este virus, para así conocer el impacto de la pandemia en regiones geográficas, y tener un mejor conocimiento de la propagación del virus SARS-CoV-2 para que las autoridades sanitarias puedan tomar acciones informadas (Islam et al., 2021). En Perú, los esfuerzos por hacer analítica avanzada de los datos de las secuencias genómicas recolectadas en el país han sido incipientes. Por ello, en el presente trabajo se plantea el desarrollo de una herramienta que brinde soporte a estudios sobre la propagación del virus en el espacio y tiempo.

Asimismo, se dispone de información sobre las secuencias genómicas de SARS-CoV-2 en diferentes repositorios de base de datos, como GISAID (GISAID, 2022), para realizar diferentes análisis.

Existe una gran necesidad de realizar análisis de agrupamiento para comprender las posibles variantes que se encuentren presentes y encontrar patrones en la evolución del virus. A raíz de esta necesidad, se han desarrollado diferentes herramientas que apoyan a los estudios sobre la evolución del virus, así como de la diversidad de secuencias genómicas en diferentes países. Muchas de estas herramientas realizan análisis de agrupamientos de las secuencias genómicas de manera estática (Hozumi et al., 2021; Morais et al., 2020), es decir, no permite incorporar nuevas secuencias sin necesidad de rehacer todo el análisis de agrupamiento, lo que puede generar diferentes representaciones cada vez que se ejecuta, dificultando el análisis y seguimiento. Por ello, se necesita contar con métodos que permitan realizar un análisis de agrupamientos de forma online, en la cual se pueda agregar nuevas secuencias sin necesidad de tener que recrear la representación de todo el conjunto de datos y el análisis de agrupamiento.

En el caso del Perú, no se ha encontrado alguna publicación en la que se indique que se cuenta con una herramienta que investigue la distribución en el espacio y tiempo de las variantes del virus SARS-CoV-2.

Por tal motivo, este proyecto de investigación tiene como finalidad proporcionar una herramienta analítica interactiva que permita visualizar y analizar el agrupamiento en el espacio y a lo largo del tiempo de la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú, donde se pueda incorporar nuevas secuencias en el análisis. De esta manera, la herramienta apoyará en la investigación de diversidad genómica del virus, evolución del virus y en la realización de un análisis descriptivo de los grupos de las diferentes variantes de SARS-CoV-2 que circulan en el Perú. Asimismo, permitirá analizar las secuencias genómicas de SARS-CoV-2 de manera sencilla e interactiva, para que especialistas y analistas puedan realizar un análisis con mayor profundidad y entendimiento.

En consecuencia, se espera que luego de contar con la herramienta, esta sea de gran utilidad en el sector de salud, para que se pueda realizar análisis sobre la diversidad genómica del SARS-CoV-2 del Perú en el espacio y en el tiempo de manera interactiva, donde se pueda seleccionar o configurar los parámetros para realizar el análisis.

Además, esta herramienta podría ser adaptada en la visualización y análisis del agrupamiento, en el espacio y a lo largo del tiempo, para otros países.

- **Viabilidad**

✓ **Viabilidad técnica**: El proyecto de investigación requiere de herramientas que son de libre acceso, por lo que no requieren de un costo económico adicional. Además, se cuenta con el equipo físico que está acondicionado de acuerdo con las necesidades para el desarrollo del proyecto. Asimismo, se cuenta con apoyo del asesor del proyecto, quien tiene conocimiento y experiencia en los métodos bioinformáticos y de clustering a utilizar en el desarrollo del proyecto. Por otro lado, se cuenta con repositorios de acceso público como la base de datos GISAID de donde se obtendrán los datos de las secuencias genómicas SARS-CoV-2.

✓ **Viabilidad operativa**: El proyecto de investigación requiere de la participación de un experto en inteligencia artificial o ciencia de datos que será el encargado de validar y aprobar los resultados relacionados a esta área. En tal sentido, el experto será contactado por el asesor ya que pertenece al grupo de investigación de inteligencia artificial.

✓ **Viabilidad temporal:** El proyecto de investigación tiene una estimación aproximada de desarrollo de 4 meses; por ello, se cuenta con un cronograma que permite cumplir con los objetivos en el plazo establecido.

✓ **Viabilidad económica**: El proyecto de investigación no requiere una inversión económica adicional, ya que las herramientas a utilizar son de acceso libre y se cuenta

con el equipo físico necesario para desarrollar el proyecto. A ello se suma, que se cuenta con un crédito educacional en *Amazon Web Services* que brinda la universidad.

- **Alcance**

El presente proyecto pertenece a la rama de ciencias de la computación dentro de la especialidad de Ingeniería Informática, dado que se emplearán métodos de aprendizaje automático no supervisado. Este proyecto tiene como finalidad desarrollar una herramienta que permita visualizar y analizar el agrupamiento en el espacio y a lo largo del tiempo de la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú. Además, esta herramienta contará con una interfaz con funcionalidades interactivas para que el usuario pueda comprender los datos en profundidad.

Por otro lado, solo se trabajará con las secuencias genómicas de SARS-CoV-2 del Perú obtenidas de la base de datos GISAID. Asimismo, no se realizará un análisis adicional de las secuencias genómicas, se va a tomar como correctos los datos obtenidos de la base de datos GISAID.

En cuanto a la herramienta, esta tendrá tres funcionalidades; una de ellas es visualizar la diversidad de las secuencias genómicas de SARS-CoV-2 en un mapa del Perú en el espacio y tiempo, desde que se registraron los datos en GISAID hasta el presente. Se trabajará con datos a nivel regional del Perú porque no se cuenta con información detallada a nivel de distritos en la base de datos. Otra funcionalidad es visualizar a través de un gráfico de dispersión el agrupamiento de las secuencias genómicas de SARS-CoV-2, donde se pueda incorporar nuevas secuencias de forma online. Por último, se brindará la opción de elegir algunos parámetros de los algoritmos de agrupamiento, brindando así la interactividad a los módulos a desarrollar.

Además, la herramienta será desplegada en la nube y se brindará el servicio como una aplicación web, utilizando los servicios brindados por *Amazon Web Services*.

- **Limitaciones**

En esta sección se identifica las limitantes para el desarrollo del presente proyecto, estas son las siguientes:

Dado que el proyecto va a utilizar información de la base de datos de GISAID, este va a estar limitado por la cantidad de datos que se tenga disponible de secuencias genómicas de SARS-CoV-2 en el Perú. Solo se va a utilizar las secuencias genómicas que estén disponibles en GISAID.

En vista de que no se cuenta con servidores dedicados para desplegar el sistema, se utilizará los créditos brindados por la universidad en *Amazon Web Services Educate*. Por ello, se tendrá en cuenta el uso de servicios de bajo costo para así poder utilizar por un mayor tiempo dichos servicios brindados, lo cual implica una limitante en el poder computacional de la herramienta.

- **Identificación de los riesgos del proyecto**

En la Tabla A1, se presenta la lista de riesgos identificados que pueden afectar el desarrollo del presente proyecto; así como también, se definirá para cada riesgo, las situaciones que permite que el riesgo ocurra (síntomas), la probabilidad de ocurrencia, el impacto en caso se presente el riesgo, la severidad, las acciones que se realizará para prevenir el riesgo (mitigación) y las acciones a realizar en caso se presente el riesgo (contingencia).

Tabla A1. Riesgos del proyecto identificados

Riesgo	Síntomas	Probabilidad	Impacto	Severidad	Mitigación	Contingencia
Poca disponibilidad de los especialistas o expertos.	Postergación de las reuniones por parte de los especialistas.	Bajo	Medio	Medio	Planificar con anticipación las fechas de reuniones junto a los especialistas.	Solicitar el apoyo de otros especialistas en el tema.
Pérdida del avance del proyecto debido a una avería en el computador	- El computador va lento. - El computador se reinicia de manera habitual. - En el computador aparece la pantalla azul.	Bajo	Alto	Medio	- Almacenar en la nube y en un disco duro externo el avance del proyecto. - Realizar un mantenimiento periódico del computador.	Hacer uso de un computador alternativo.

Nota. Elaboración propia.

- **Estructura de descomposición del trabajo (EDT)**

En la Figura A1, se presenta la estructura de descomposición del trabajo del proyecto de investigación.

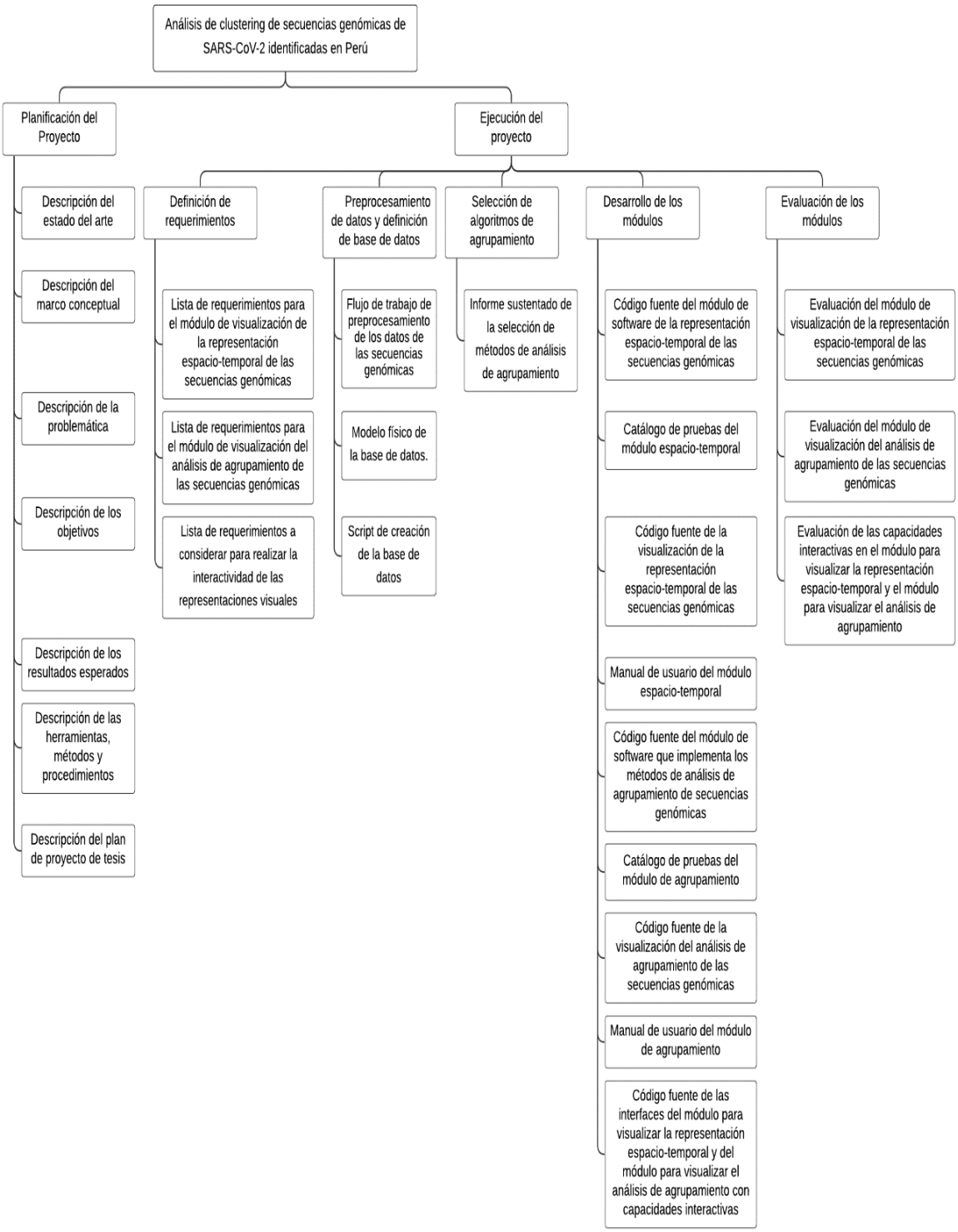


Figura A1. Estructura de descomposición del trabajo del proyecto

- **Lista de tareas**

En la Tabla A2, se presenta la lista de tareas de la planificación y ejecución del proyecto de investigación, en donde se detalla para cada tarea la duración estimada, el esfuerzo asociado y el costo estimado.

Tabla A2. Lista de tareas.

Tarea	Duración estimada (días)	Esfuerzo asociado (horas - persona)	Costo estimado (soles)
Planificación del proyecto			
Elaborar el entregable parcial 1.1: Ficha de registro de idea de tesis y asesor	4	4 (tesista)	80
Elaborar el entregable parcial 1.2: Protocolo de revisión. Diseño de Formulario de extracción.	4	4 (tesista)	80
Elaborar el entregable parcial 1.3: Reporte de ejecución de la revisión. Formulario de extracción.	4	4 (tesista)	80
Elaborar el entregable parcial 1.4: Reporte de ejecución de la revisión. Formulario de extracción.	4	4 (tesista)	80
Elaborar el entregable parcial 1.5: Marco conceptual.	4	4 (tesista)	80
Elaborar el entregable 1: Problemática. Marco conceptual/Marco teórico/Marco legal. Estado del Arte.	7	7 (tesista)	140
Elaborar el entregable parcial 2.1: Árbol de objetivos. Objetivo general. Objetivos específicos.	7	7 (tesista)	140
Elaborar el entregable 2: Objetivo general. Objetivos específicos. Resultados Esperados. Medios de verificación.	16	16 (tesista)	320
Elaborar el entregable 3: Resultados esperados. Herramientas, métodos y procedimientos. Alcance y Limitaciones.	14	14 (tesista)	280

Elaborar el entregable 4: Proyecto de fin de carrera completo incluyendo: todas las correcciones y el Anexo de Plan de Tesis.	7	7 (tesista)	140
Reuniones con asesor del proyecto	12	12 (tesista) 12 (asesor)	960
Ejecución del proyecto			
Realizar el modelo físico de la base de datos	1	6 (tesista)	120
Creación de scripts para la base de datos	1	6 (tesista)	120
Realizar el documento que describe el modelo físico de la base de datos	1	5 (tesista)	100
Generar el flujo de trabajo de preprocesamiento de los datos de las secuencias genómicas (reducción de dimensionalidad)	2	16 (tesista)	320
Programar backend del módulo espacio-temporal	6	36 (tesista)	720
Realizar el catálogo de pruebas de módulo espacio-temporal	1	8 (tesista)	160
Redactar lista de requerimientos del módulo espacio-temporal	2	10 (tesista)	200
Programar frontend del módulo espacio-temporal	6	36 (tesista)	720
Redactar el manual de usuario del módulo	2	12 (tesista)	240
Definir lista de los métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2	1	5 (tesista)	100
Realizar el informe sustentado de selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software	1	7 (tesista)	140
Programar backend del módulo de agrupamiento	6	42 (tesista)	840

Realizar el catálogo de pruebas de módulo de agrupamiento	1	8 (tesista)	160
Redactar lista de requerimientos del módulo agrupamiento	1	5 (tesista)	100
Programar frontend del módulo de agrupamiento	6	36 (tesista)	720
Redactar el manual de usuario del módulo	2	12 (tesista)	240
Redactar lista de requerimientos de la vista interactiva	1	6 (tesista)	120
Programar backend de la vista interactiva del módulo espacio-temporal y del módulo de agrupamiento	6	30 (tesista)	600
Programar frontend de la vista interactiva del módulo espacio-temporal y del módulo de agrupamiento	6	42 (tesista)	840
Reunión con asesor del proyecto	16	16 (tesista) 16 (asesor)	1280
Reunión para la validación por parte de los expertos	5	10 (tesista) 10 (asesor) 2 (por cada experto)	1400

Nota. Elaboración propia.

- **Cronograma del proyecto**

A continuación, se adjunta el documento con el cronograma del proyecto de investigación. Este documento contiene la descripción de las tareas, la duración, el rango de fechas y la dependencia que existe entre las tareas de la fase de planificación y ejecución del proyecto. Asimismo, se presenta el diagrama de Gantt con las tareas de la fase de ejecución del proyecto. Se puede acceder al documento mediante los siguientes enlaces:

- ✓ En formato PDF:

https://drive.google.com/file/d/1-zRHKDXNtefCyZeceBmidTgUr4uAITMY/view?usp=sharing

- ✓ En formato Excel:

https://docs.google.com/spreadsheets/d/1_N_eLHE3UYnmQWMgT9mU-AXAp1WksK7T/edit#gid=1478261005

- **Lista de recursos**

En esta sección, se presenta la lista de recursos con los que se cuenta para el proyecto de investigación.

- o **Personas involucradas**

- ✓ Asesor de tesis: El Dr. Edwin Villanueva será el asesor de tesis y jefe de proyecto. También, será quien guíe y oriente a la tesista durante la planificación y ejecución del proyecto de investigación.
- ✓ Tesista: La tesista Carolina Mejía será quién desarrolle los módulos espacio-temporal y de agrupamiento para obtener la herramienta analítica interactiva para representar visualmente la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú.
- ✓ Especialista en base de datos: Profesional con experiencia en base de datos que ayudará a validar uno de los resultados esperados relacionado a la estructura de la base de datos.
- ✓ Especialista en Ingeniería Informática: Profesional que ayudará a validar los resultados esperados.
- ✓ Especialista en bioinformática o biología molecular: Profesional con experiencia en el área de bioinformática o biología molecular que ayudará a validar los resultados esperados relacionados al módulo espacio-temporal y el módulo de agrupamiento; así como de la capacidad interactiva que deben tener estos módulos.

- ✓ Especialista en inteligencia artificial o ciencia de datos: Profesional con experiencia en el área de inteligencia artificial o ciencia de datos que ayudará a validar uno de los resultados esperados relacionado con la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2.
- ✓ Especialista en interacción humano computador (HCI): Profesional con experiencia en el área de usabilidad que ayudará a validar uno de los resultados esperados relacionado con la visualización y capacidad interactiva de la herramienta.

- o **Materiales requeridos para el proyecto**

En el presente proyecto de investigación, este recurso no aplica.

- o **Estándares utilizados en el proyecto**

En el presente proyecto de investigación, este recurso no aplica.

- o **Equipamiento requerido**

Computador: se utilizará una laptop para el desarrollo del módulo espacio-temporal y del módulo de agrupamiento, entre otras actividades del proyecto de investigación.

Servicios en la nube: se utilizará un servidor en la nube de AWS donde estará alojada la base de datos y se desplegará la herramienta del proyecto de investigación.

- o **Herramientas requeridas**

A continuación, se presentan las herramientas necesarias para el desarrollo del proyecto de investigación. La descripción correspondiente a cada una de estas herramientas se encuentra en la sección Lista de Herramientas del presente proyecto.

- ✓ PostgreSQL
- ✓ Python
- ✓ Biopython
- ✓ FastApi
- ✓ Scikit Learn

- ✓ SciPy
- ✓ NumPy
- ✓ *Amazon Web Services* (AWS)
- ✓ React
- ✓ Bokeh

- **Costeo del Proyecto**

En la Tabla A3, se presenta de manera detallada los costos estimados del proyecto de investigación para lograr establecer el presupuesto.

Tabla A3. Costeo del proyecto.

Ítem	Descripción	Unidad	Cantidad	Valor Unitario (S/.)	Monto Total (S/.)	Monto Acumulado (S/.)
0	**Costo total del proyecto**	-	-	-	-	15660
1	**Personas involucradas en el proyecto**	-	-	-	-	11620
1.1	Tesista	Horas	437	20	8740	
1.2	Asesor de tesis	Horas	38	60	2280	
1.3	Experto en Base de datos	Horas	2	60	120	
1.4	Experto en Inteligencia artificial	Horas	2	60	120	
1.5	Experto en Ingeniería informática	Horas	2	60	120	
1.6	Experto en Bioinformática	Horas	2	60	120	
1.7	Experto en HCI	Horas	2	60	120	
2	**Bienes y equipos**	-	-	-	-	3500
2.1	Laptop	Equipo	1	3500	3500	

3	Licencias de software	-	-	-	-	540
3.1	Servicios AWS	Créditos	150	3.6	540	

Nota. Elaboración propia.

Anexo B: Script para la creación de las tablas de la base de datos

Este anexo presenta el script que contiene los comandos para la creación de las tablas de la base de datos en el archivo 01CreateDDL.sql, el cual puede ser accedido por medio del siguiente enlace: https://github.com/CarolinaMejiaMujica/Script

Anexo C: Acta de validación del modelo físico de base de datos

Este anexo contiene el acta de validación del documento por parte de un especialista en base de datos. El cual se presenta a continuación.

Acta de validación de resultados

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Estructura de la base de datos de los módulos de software a desarrollar.

Descripción del resultado: Se describe el modelo físico de la base de datos a emplear para almacenar los datos a utilizar por la herramienta analítica interactiva para representar visualmente la diversidad de las secuencias genómicas de SARS-CoV-2 recolectadas en Perú.

Evaluador: Ing. Juan Carlos Tovar Galarreta

Veredicto: Aceptado

Comentario adicional:
De acuerdo a la revisión realizada, se valida el modelo físico de la base de datos con las siguientes sugerencias de ser necesarias:

- Si formará parte de una aplicación, seguir con el siguiente estándar:
  <idAplicacion>_<idObjeto>_<NombreObjeto>
  Identificador de Aplicación. De 3 a 6 caracteres
  Identificador de Objeto. Tabla o vista.
    TM. Tabla Maestra
    TD. Tabla Detalle
    TH. Tabla Histórica
  Nombre de Objeto. Para nombres compuestos separar con '_'.
- Para las columnas de más de 4000 caracteres, se recomienda utilizar datos tipos CLOB o su equivalente.
- Agregar check constraints para asegurar la integridad de los datos más relevantes.
- Considerar campos de auditoría, para las tablas de mayor relevancia.

Firma

Lima, 06 de septiembre de 2021

Anexo D: Catálogo de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas SARS-CoV-2

Este anexo contiene el catálogo de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas SARS-CoV-2, en donde se detalla las pruebas unitarias del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas de SARS-CoV-2. Estas pruebas unitarias se realizaron sobre un conjunto de 9 196 datos de secuencias genómicas SARS-CoV-2. En la Tabla D1, se presenta el catálogo de pruebas, en el cual se detalla la función o servicio a probar, los datos para realizar la prueba, una descripción y el resultado esperado.

Tabla D1. Catálogo de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas de SARS-CoV-2.

Prueba	Datos	Descripción	Resultado Esperado
Preprocesamiento	❖ Archivos .fasta y .tsv con los datos de las secuencias genómicas SARS-CoV-2.	En este proceso se realiza el preprocesamiento de las secuencias genómicas SARS-CoV-2.	Datos de las secuencias genómicas preprocesadas, en una menor dimensión.
Servicio que realiza el gráfico del mapa del Perú	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta.	Este servicio realiza el gráfico del mapa del Perú a nivel regional de las variantes identificadas en cada departamento.	Gráfico del mapa del Perú con las variantes identificadas en cada departamento en formato JSON.
Servicio que realiza el gráfico circular	❖ Rango de fechas: - Fecha Inicio - Fecha Fin	Este servicio realiza el gráfico circular con el porcentaje de aparición de cada variante del	Gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo en formato JSON.

	❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta.	SARS-CoV-2 en el tiempo.	
Servicio que realiza el gráfico de línea	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta.	Este servicio realiza el gráfico de línea con la cantidad de secuencias genómicas SARS-CoV-2 pertenecientes a las variantes identificadas en el tiempo.	Gráfico de línea con las variantes identificadas en el tiempo en formato JSON.
Servicio que lista los datos de las secuencias genómicas SARS-CoV-2	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta.	Este servicio devuelve la lista de los datos de las secuencias genómicas SARS-CoV-2.	Lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de las secuencias genómicas (identificador en la base de datos de GISAID), Fecha de recolección, Nomenclatura según la OMS de la variante identificada y Nombre de la variante identificada.
Servicio que calcula la cantidad de secuencias genómicas SARS-CoV-2 obtenidas de GISAID y la cantidad de secuencias utilizadas en el análisis.	❖ No recibe datos de entrada	Este servicio devuelve la cantidad de secuencias genómicas SARS-CoV-2 obtenidas de GISAID y la cantidad de secuencias genómicas SARS-CoV-2 utilizadas en el análisis.	Cantidad total de las secuencias genómicas obtenidas de GISAID y la cantidad de secuencias genómicas utilizadas en el análisis en formato JSON.

Nota. Elaboración propia.

1. Prueba del Preprocesamiento

Esta prueba tiene como objetivo preprocesar y reducir la dimensión de los datos de las secuencias genómicas SARS-CoV-2. Se tiene como datos de entrada los archivos .fasta y .tsv con la información de las secuencias genómicas SARS-CoV-2, los cuales se obtienen descargando las secuencias de la base de datos GISAID, el proceso para descargar los datos se detalla en el presente proyecto de investigación. En la Figura D1, se presenta el formato que debe tener el archivo de entrada .fasta.

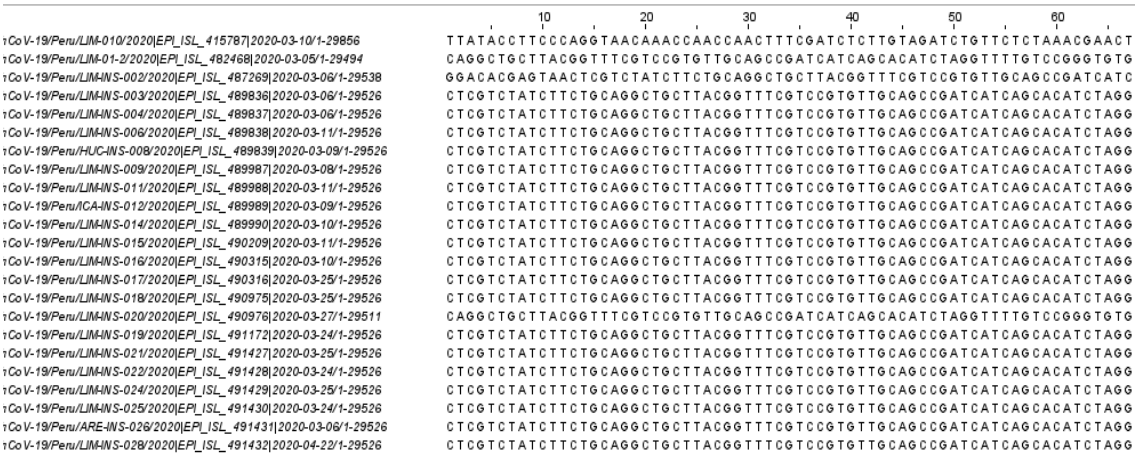


Figura D1. Formato del archivo de entrada .fasta.

En la Figura D2, se presenta el formato que debe tener el archivo de entrada .tsv

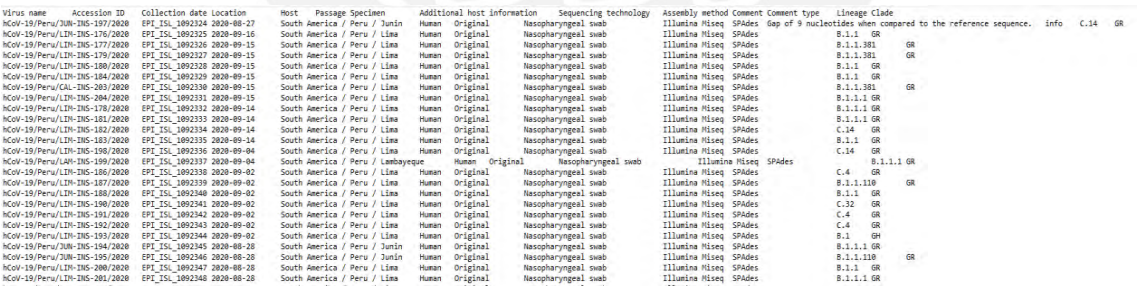


Figura D2. Formato del archivo de entrada .tsv.

En la Figura D3, se presenta el resultado esperado luego de ejecutar la prueba.

```
array([[-0.0657005 , -0.06535234],
       [-0.01780591, -0.01308275],
       [-0.07733525, -0.07017369],
       ...,
       [-0.18357708,  0.37995702],
       [-0.07294369, -0.10820307],
       [-0.07607677, -0.11662673]], dtype=float32)
```

Figura D3. Resultado de la prueba del Preprocesamiento.

Para poder ejecutar las siguientes pruebas se debe tener registrado los datos de las secuencias genómicas SARS-CoV-2 en la base de datos.

2. **Prueba del Servicio que realiza el gráfico del mapa del Perú**

Esta prueba tiene como objetivo verificar que se obtenga el gráfico del mapa del Perú a nivel regional de las variantes identificadas en cada departamento en formato JSON. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura D4, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

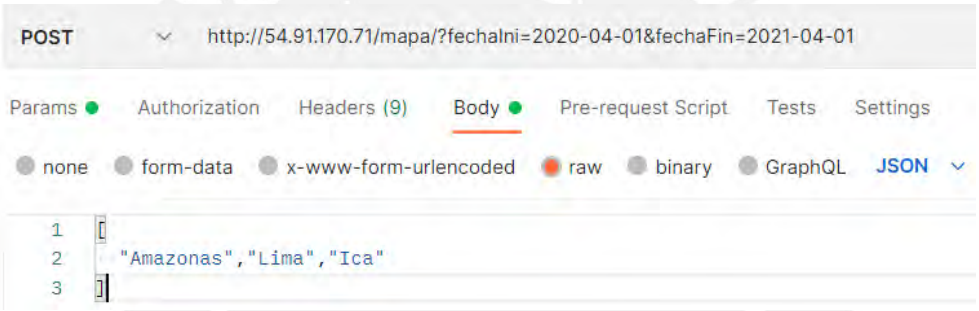


Figura D4. Datos de entrada para la prueba del Servicio que realiza el gráfico del mapa del Perú.

En la Figura D5, se presenta el resultado esperado luego de ejecutar la prueba.

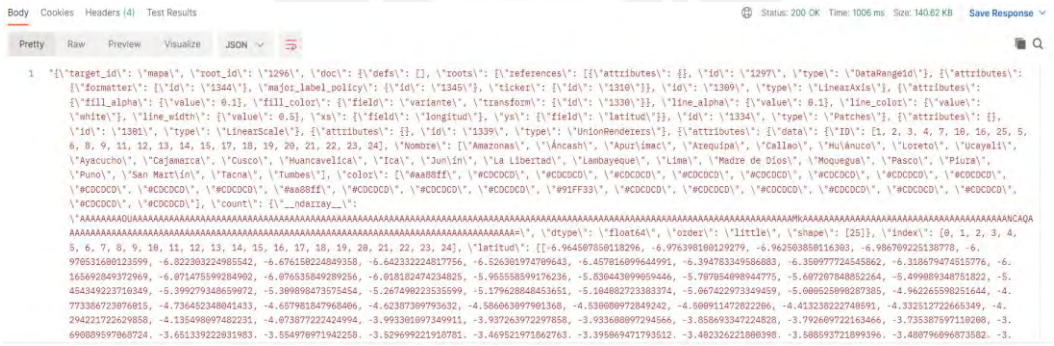


Figura D5. Resultado de la prueba del Servicio que realiza el gráfico del mapa del Perú.

3. **Prueba del Servicio que realiza el gráfico circular**

Esta prueba tiene como objetivo verificar que se obtenga el gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo en formato JSON. Se tiene como

datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura D6, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

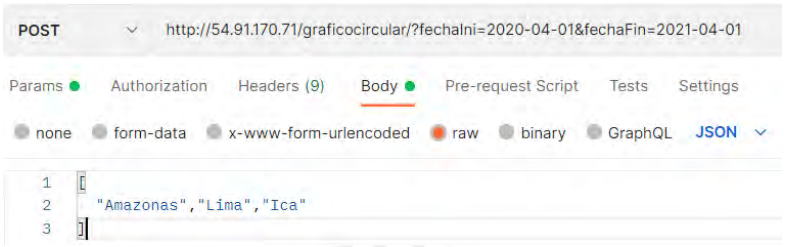


Figura D6. Datos de entrada para la prueba del Servicio que realiza el gráfico circular.

En la Figura D7, se presenta el resultado esperado luego de ejecutar la prueba.

Figura D7. Resultado de la prueba del Servicio que realiza el gráfico circular.

4. Prueba del Servicio que realiza el gráfico de línea

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de línea con la cantidad de secuencias genómicas SARS-CoV-2 pertenecientes a las variantes identificadas en el tiempo en formato JSON. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura D8, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

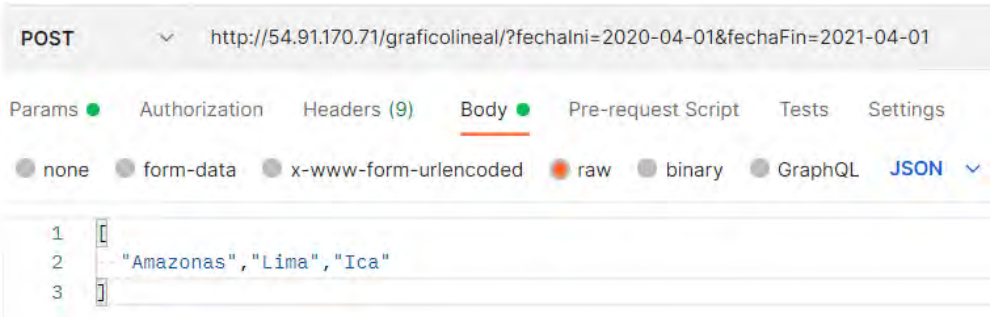


Figura D8. Datos de entrada para la prueba del Servicio que realiza el gráfico de línea.

En la Figura D9, se presenta el resultado esperado luego de ejecutar la prueba.

Figura D9. Resultado de la prueba del Servicio que realiza el gráfico de línea.

5. Prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2

Esta prueba tiene como objetivo verificar que se liste los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos deben ser los siguientes: Departamento, ID de acceso de las secuencias genómicas (identificador en la base de datos de GISAID), Fecha de recolección, Nomenclatura según la OMS de la variante identificada y Nombre de la variante identificada. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura D10, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

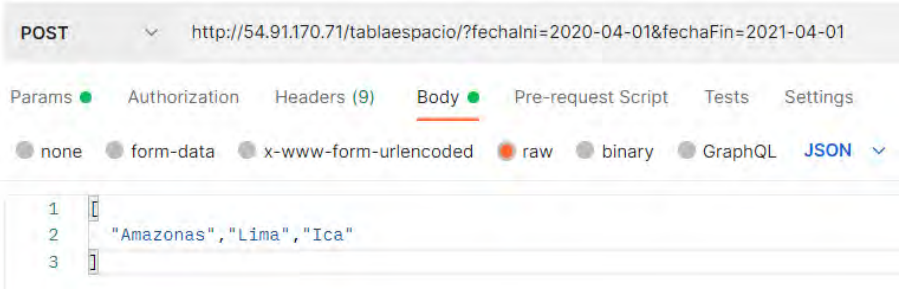


Figura D10. Datos de entrada para la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2.

En la Figura D11, se presenta el resultado esperado luego de ejecutar la prueba.

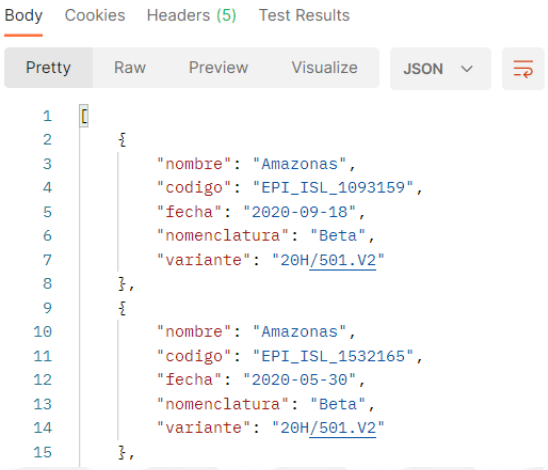


Figura D11. Resultado de la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2.

6. Prueba del Servicio que calcula la cantidad de secuencias genómicas SARS-CoV-2 obtenidas de GISAID y la cantidad de secuencias utilizadas en el análisis.

Esta prueba tiene como objetivo verificar que se devuelva la cantidad total de las secuencias genómicas obtenidas de GISAID y la cantidad de secuencias genómicas utilizadas en el análisis en formato JSON. En la Figura D12, se presenta el resultado esperado luego de ejecutar la prueba.

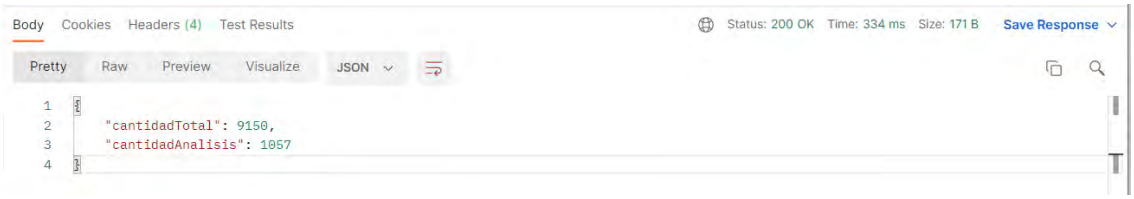


Figura D12. Resultado de la prueba del Servicio que calcula la cantidad total de secuencias obtenidas y la cantidad de secuencias utilizadas en el análisis.

Anexo E: Reporte del resultado de pruebas del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas de SARS-CoV-2

Este anexo contiene el resultado de las pruebas unitarias aprobadas al 100%, esto indica que las pruebas se han completado de forma satisfactoria cumpliendo con los resultados esperados. A continuación, se detalla los resultados de las pruebas unitarias del módulo espacio-temporal y el preprocesamiento de las secuencias genómicas de SARS-CoV-2. Estas pruebas unitarias se realizaron sobre un conjunto de 9 196 datos de secuencias genómicas SARS-CoV-2. A continuación, se detalla cada prueba realizada y el resultado alcanzado.

1. **Prueba del Preprocesamiento**

Esta prueba tiene como objetivo preprocesar y reducir la dimensión de los datos de las secuencias genómicas SARS-CoV-2. Se tiene como datos de entrada los archivos .fasta y .tsv con la información de las secuencias genómicas SARS-CoV-2, los cuales se obtienen descargando las secuencias de la base de datos GISAID, el proceso para descargar los datos se detalla en el presente proyecto de investigación. En la Figura E1, se presenta el formato que debe tener el archivo de entrada .fasta.

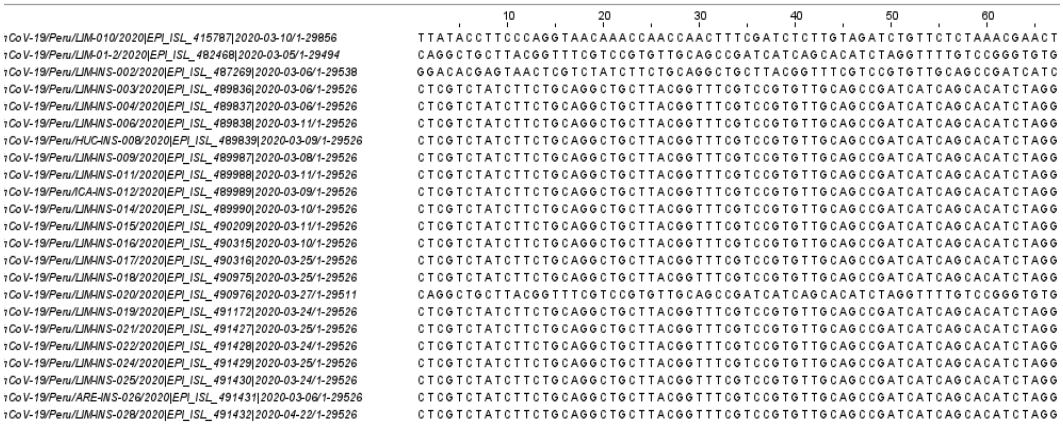


Figura E1. Formato del archivo de entrada .fasta.

En la Figura E2, se presenta el formato que debe tener el archivo de entrada .tsv

Virus name	Accession ID	Collection date	Location	Host	Passage	Specimen	Additional host information	Sequencing technology	Assembly method	Comment	Comment type	Lineage	Clade			
hCoV-19/Peru/JUN-INS-197/2020	EPI_ISL_1092324	2020-08-27	South America / Peru / Junin	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes	Gap of 9 nucleotides when compared to the reference sequence.	info	C.14	GR			
hCoV-19/Peru/LIM-INS-176/2020	EPI_ISL_1092325	2020-09-16	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1	GR			
hCoV-19/Peru/LIM-INS-177/2020	EPI_ISL_1092326	2020-09-15	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.381	GR			
hCoV-19/Peru/LIM-INS-179/2020	EPI_ISL_1092327	2020-09-15	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.381	GR			
hCoV-19/Peru/LIM-INS-180/2020	EPI_ISL_1092328	2020-09-15	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1	GR			
hCoV-19/Peru/LIM-INS-184/2020	EPI_ISL_1092329	2020-09-15	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1	GR			
hCoV-19/Peru/CAL-INS-203/2020	EPI_ISL_1092330	2020-09-15	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.381	GR			
hCoV-19/Peru/LIM-INS-204/2020	EPI_ISL_1092331	2020-09-15	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.1	GR			
hCoV-19/Peru/LIM-INS-178/2020	EPI_ISL_1092332	2020-09-14	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.1	GR			
hCoV-19/Peru/LIM-INS-181/2020	EPI_ISL_1092333	2020-09-14	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.1	GR			
hCoV-19/Peru/LIM-INS-182/2020	EPI_ISL_1092334	2020-09-14	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			C.14	GR			
hCoV-19/Peru/LIM-INS-183/2020	EPI_ISL_1092335	2020-09-14	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1	GR			
hCoV-19/Peru/LIM-INS-198/2020	EPI_ISL_1092336	2020-09-04	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			C.14	GR			
hCoV-19/Peru/LAM-INS-199/2020	EPI_ISL_1092337	2020-09-04	South America / Peru / Lambayeque	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.1	GR			
hCoV-19/Peru/LIM-INS-186/2020	EPI_ISL_1092338	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			C.4	GR			
hCoV-19/Peru/LIM-INS-187/2020	EPI_ISL_1092339	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.110	GR			
hCoV-19/Peru/LIM-INS-188/2020	EPI_ISL_1092340	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1	GR			
hCoV-19/Peru/LIM-INS-190/2020	EPI_ISL_1092341	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			C.32	GR			
hCoV-19/Peru/LIM-INS-191/2020	EPI_ISL_1092342	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			C.4	GR			
hCoV-19/Peru/LIM-INS-192/2020	EPI_ISL_1092343	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			C.4	GR			
hCoV-19/Peru/LIM-INS-193/2020	EPI_ISL_1092344	2020-09-02	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1	GH			
hCoV-19/Peru/JUN-INS-194/2020	EPI_ISL_1092345	2020-08-28	South America / Peru / Junin	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.1	GR			
hCoV-19/Peru/JUN-INS-195/2020	EPI_ISL_1092346	2020-08-28	South America / Peru / Junin	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.110	GR			
hCoV-19/Peru/LIM-INS-200/2020	EPI_ISL_1092347	2020-08-28	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1	GR			
hCoV-19/Peru/LIM-INS-201/2020	EPI_ISL_1092348	2020-08-28	South America / Peru / Lima	Human	Original		Nasopharyngeal swab	Illumina Miseq	SPAdes			B.1.1.1	GR			

Figura E2. Formato del archivo de entrada .tsv.

El resultado esperado es los datos de las secuencias genómicas preprocesadas, en una menor dimensión. En la Figura E3, se presenta el resultado obtenido luego de ejecutar la prueba.

```
array([[-0.0657005 , -0.06535234],
       [-0.01780591, -0.01308275],
       [-0.07733525, -0.07017369],
       ...,
       [-0.18357708,  0.37995702],
       [-0.07294369, -0.10820307],
       [-0.07607677, -0.11662673]], dtype=float32)
```

Figura E3. Resultado de la prueba del Preprocesamiento.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

2. Prueba del Servicio que realiza el gráfico del mapa del Perú

Esta prueba tiene como objetivo verificar que se obtenga el gráfico del mapa del Perú a nivel regional de las variantes identificadas en cada departamento en formato JSON. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura E4, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

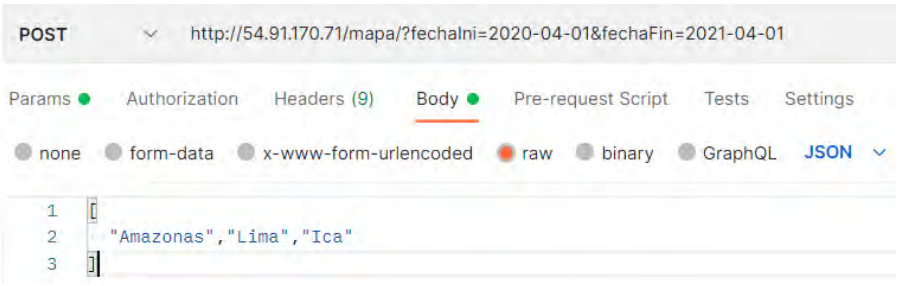


Figura E4. Datos de entrada para la prueba del Servicio que realiza el gráfico del mapa del Perú.

El resultado esperado es el gráfico del mapa del Perú con las variantes identificadas en cada departamento en formato JSON. En la Figura E5, se presenta el resultado obtenido luego de ejecutar la prueba.

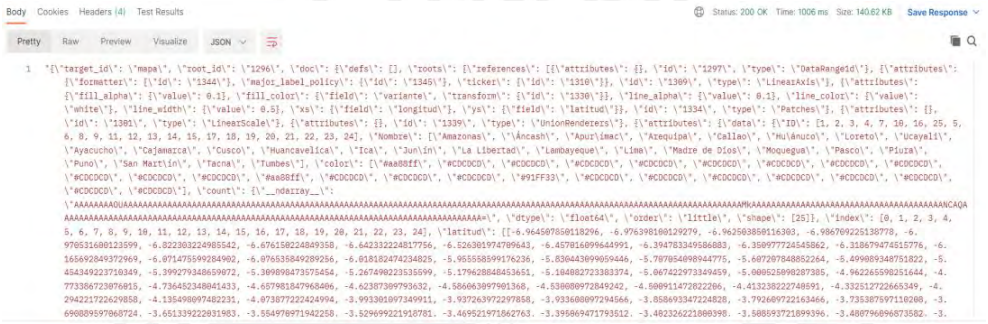


Figura E5. Resultado de la prueba del Servicio que realiza el gráfico del mapa del Perú.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

3. Prueba del Servicio que realiza el gráfico circular

Esta prueba tiene como objetivo verificar que se obtenga el gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo en formato JSON. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura E6, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

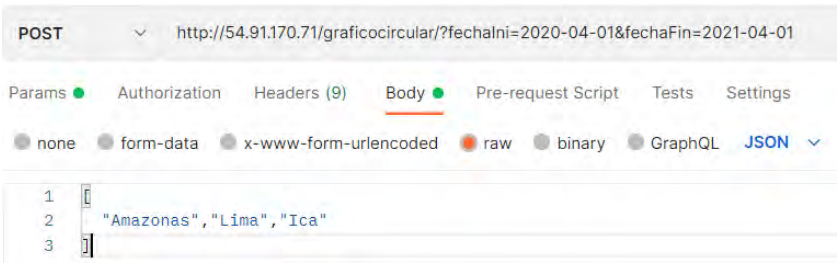


Figura E6. Datos de entrada para la prueba del Servicio que realiza el gráfico circular.

El resultado esperado es el gráfico circular con el porcentaje de aparición de cada variante del SARS-CoV-2 en el tiempo en formato JSON. En la Figura E7, se presenta el resultado obtenido luego de ejecutar la prueba.

Figura E7. Resultado de la prueba del Servicio que realiza el gráfico circular.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

4. Prueba del Servicio que realiza el gráfico de línea

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de línea con la cantidad de secuencias genómicas SARS-CoV-2 pertenecientes a las variantes identificadas en el tiempo en formato JSON. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura E8, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

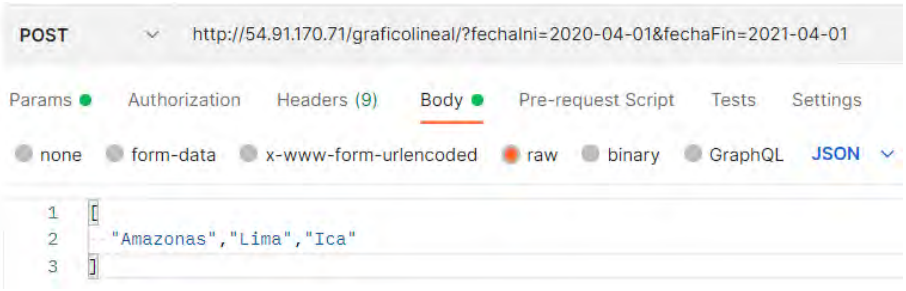


Figura E8. Datos de entrada para la prueba del Servicio que realiza el gráfico de línea.

El resultado esperado es el gráfico de línea con las variantes identificadas en el tiempo en formato JSON. En la Figura E9, se presenta el resultado obtenido luego de ejecutar la prueba.

Figura E9. Resultado de la prueba del Servicio que realiza el gráfico de línea.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

5. Prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2

Esta prueba tiene como objetivo verificar que se liste los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura E10, se muestra el formato de los datos de entrada, la fecha inicio y fin se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

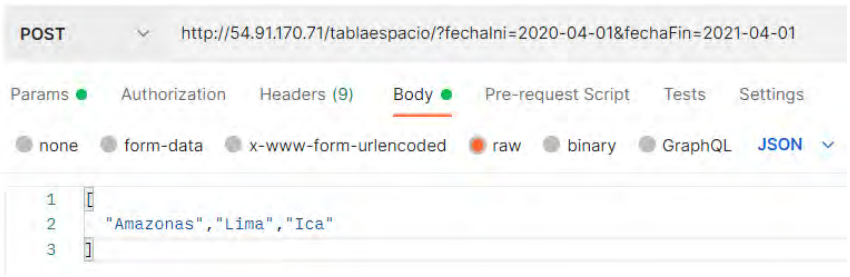


Figura E10. Datos de entrada para la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2.

El resultado esperado es la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos deben ser los siguientes: Departamento, ID de acceso de las secuencias genómicas (identificador en la base de datos de GISAID), Fecha de recolección, Nomenclatura según la OMS de la variante identificada y Nombre de la variante identificada. En la Figura E11, se presenta el resultado obtenido luego de ejecutar la prueba.

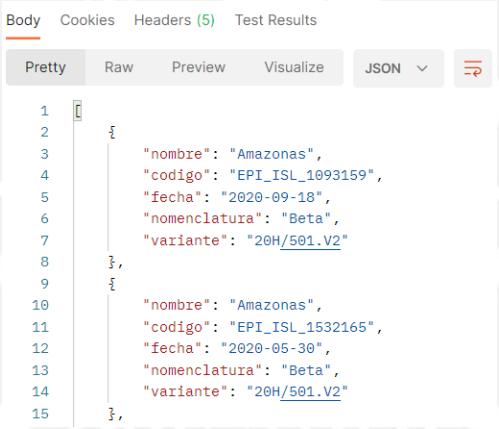


Figura E11. Resultado de la prueba del Servicio que lista los datos de las secuencias genómicas SARS-CoV-2.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

6. **Prueba del Servicio que calcula la cantidad de secuencias genómicas SARS-CoV-2 obtenidas de GISAID y la cantidad de secuencias utilizadas en el análisis.**

Esta prueba tiene como objetivo verificar que se devuelva la cantidad total de las secuencias genómicas obtenidas de GISAID y la cantidad de secuencias genómicas utilizadas en el análisis

en formato JSON. En la Figura E12, se presenta el resultado obtenido luego de ejecutar la prueba.

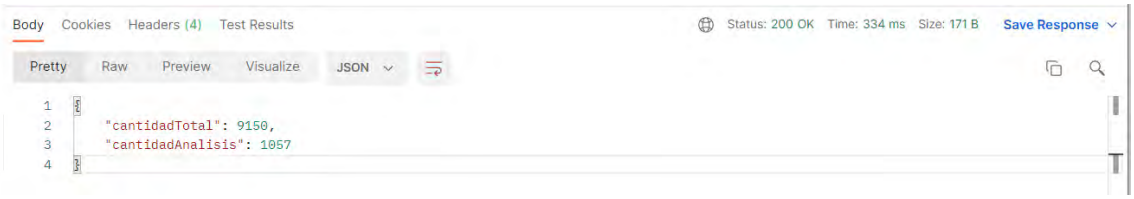


Figura E12. Resultado de la prueba del Servicio que calcula la cantidad total de secuencias obtenidas y la cantidad de secuencias utilizadas en el análisis.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

En conclusión, a través de este reporte se verifica que los resultados de las pruebas han sido aprobados al 100%, cumpliendo con el resultado esperado en cada una de ellas.

Anexo F: Acta de validación del cumplimiento de los requerimientos del módulo espacio-temporal

Este anexo contiene el acta de validación del cumplimiento de los requerimientos del módulo espacio-temporal por parte de un especialista en ingeniería informática. El cual se presenta a continuación.

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Acta de validación del cumplimiento de los requerimientos del módulo espacio-temporal

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

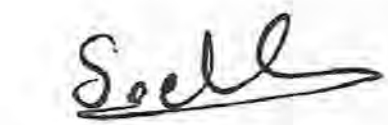
Nombre del resultado obtenido: Módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú.

Descripción del resultado: Se han implementado los requerimientos relacionados al módulo espacio-temporal para así mostrar visualmente la representación en el espacio y tiempo de las secuencias genómicas SARS-CoV-2 recolectadas en el Perú, esto con el fin de que los especialistas y analistas puedan ver la evolución del virus a nivel regional del Perú, así como también, en el tiempo desde que inició la pandemia hasta la actualidad para que así puedan realizar estudios o tomar decisiones con la información brindada.

Evaluador: Dra. Soledad Espezua Llerena

Veredicto: Aceptado

Comentario adicional: Con respecto al segundo gráfico, Porcentaje de variantes identificadas en el tiempo, creo que el titulo no condice con una relación temporal, más bien con una relación de distribución, que se usa para indicar proporciones numéricas en sectores.



Firma

Lima, 05 de noviembre de 2021

Anexo G: Reporte de apreciación del módulo espacio-temporal

Este anexo contiene el reporte de apreciación del módulo de visualización espacio-temporal por un especialista en bioinformática o biología molecular. El cual se presenta a continuación.

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Reporte de apreciación del módulo espacio-temporal

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

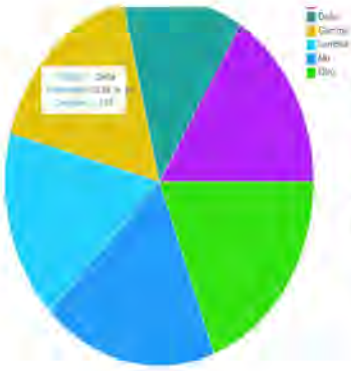
Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Módulo para visualizar la representación espacio-temporal de las secuencias genómicas SARS-CoV-2 en el Perú.

Descripción del resultado: Se han implementado los requerimientos relacionados al módulo espacio-temporal para así mostrar visualmente la representación en el espacio y tiempo de las secuencias genómicas SARS-CoV-2 recolectadas en el Perú, esto con el fin de que los especialistas y analistas puedan ver la evolución del virus a nivel regional del Perú, así como también, en el tiempo desde que inició la pandemia hasta la actualidad para que así puedan realizar estudios o tomar decisiones con la información brindada.

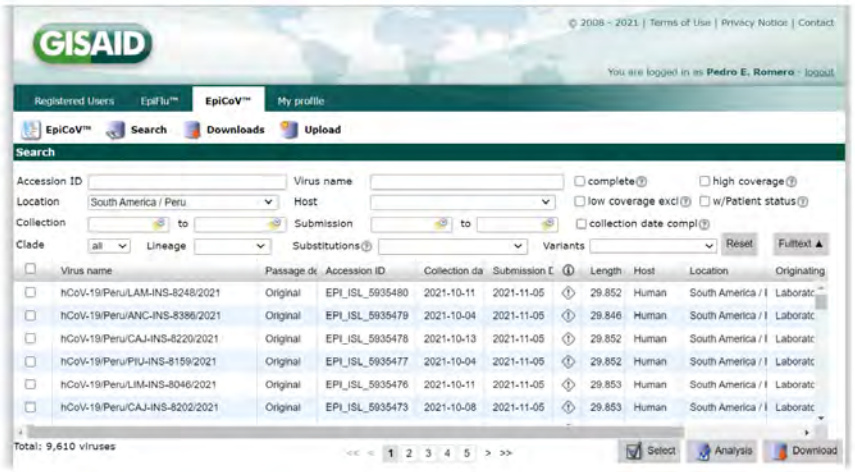
Evaluador: Dr. Pedro Eduardo Romero Condori

Comentarios de apreciación sobre el módulo:



Personalmente, no soy afín a los pie-charts. No me queda claro los números totales pues en el periodo 2020 – 2021 hay muchas más muestras que las indicadas en la figura.

Verificar estos números pues son necesarios para alcanzar el objetivo que se propone en la **Descripción del resultado**.



Firma

Lima, 05 de noviembre de 2021

Anexo H: Acta de validación de métodos de análisis de agrupamiento

Este anexo contiene el acta de validación del informe de selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 por parte de un especialista en inteligencia artificial o ciencia de datos. El cual se presenta a continuación.

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Acta de validación de métodos de análisis de agrupamiento

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en el módulo de software.

Descripción del resultado: Se ha revisado un informe (Anexo G) sobre la selección de métodos de análisis de agrupamiento de secuencias genómicas SARS-CoV-2 a ser implementados en la solución que se desarrolla en el marco de la tesis de la alumna.

Evaluador: Dr. Edwin Rafael Villanueva Talavera

Veredicto: Aceptado

Comentario adicional: La selección de los métodos ha sido adecuadamente sustentada. Los métodos son complementarios y ofrecen distintas formas de agrupamiento de las secuencias virales, por lo que se espera que vengan a enriquecer el aplicativo que se desarrolla

Firma

Lima, 28 de septiembrede2021

Anexo I: Catálogo de pruebas del módulo de agrupamiento

Este anexo contiene el catálogo de pruebas del módulo de agrupamiento, en donde se detalla las pruebas unitarias del módulo de agrupamiento. Estas pruebas unitarias se realizaron sobre un conjunto de 1365 datos de secuencias genómicas SARS-CoV-2. En la Tabla I1, se presenta el catálogo de pruebas, en el cual se detalla la función o servicio a probar, los datos para realizar la prueba, una descripción y el resultado esperado.

Tabla I1. Catálogo de pruebas del módulo de agrupamiento.

Prueba	Datos	Descripción	Resultado Esperado
Servicio que realiza el gráfico de agrupamiento con el algoritmo *k-means*	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. ❖ Valor del parámetro del algoritmo de agrupamiento.	Este servicio realiza el gráfico de dispersión que representa el agrupamiento realizado con el algoritmo *k-means* para mostrar las variantes identificadas en el Perú. Así como también, devuelve la lista los datos de las secuencias genómicas SARS-CoV-2 utilizados para realizar el gráfico.	Gráfico de agrupamiento con el algoritmo *k-means* donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.
Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. ❖ Valor del parámetro del	Este servicio realiza el gráfico de dispersión que representa el agrupamiento realizado con el algoritmo jerárquico para mostrar las variantes identificadas en el Perú. Así como también, devuelve la lista los datos de las secuencias genómicas SARS-CoV-2	Gráfico de agrupamiento con el algoritmo jerárquico donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de

	algoritmo de agrupamiento.	utilizados para realizar el gráfico.	GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.
Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. ❖ Valor del parámetro del algoritmo de agrupamiento.	Este servicio realiza el gráfico de dispersión que representa el agrupamiento realizado con el algoritmo DBSCAN para mostrar las variantes identificadas en el Perú. Así como también, devuelve la lista los datos de las secuencias genómicas SARS-CoV-2 utilizados para realizar el gráfico.	Gráfico de agrupamiento con el algoritmo DBSCAN donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.
Servicio que realiza el dendrograma	❖ Rango de fechas: - Fecha Inicio - Fecha Fin ❖ Arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta.	Este servicio realiza el dendrograma para visualizar la estructura o jerarquía de los clústeres que se pueden formar.	Gráfico del dendrograma donde se visualice la estructura o jerarquía de los clústeres que se pueden formar en formato HTML.

Nota. Elaboración propia.

Nota:

- ✓ El parámetro depende del algoritmo, si es el algoritmo es *k-means* o jerárquico entonces el parámetro a ingresar es el valor de K (cantidad de clústeres a formar), y en el caso del algoritmo DBSCAN el parámetro es épsilon.

- ✓ Para poder ejecutar las funciones, se debe tener registrado los datos de las secuencias genómicas SARS-CoV-2 en la base de datos.

1. Prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo *k-means*

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de agrupamiento con el algoritmo *k-means* donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; un arreglo de los nombres de los departamentos del Perú y el valor del parámetro del algoritmo de agrupamiento con los que se quiere realizar la consulta. En la Figura I1, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

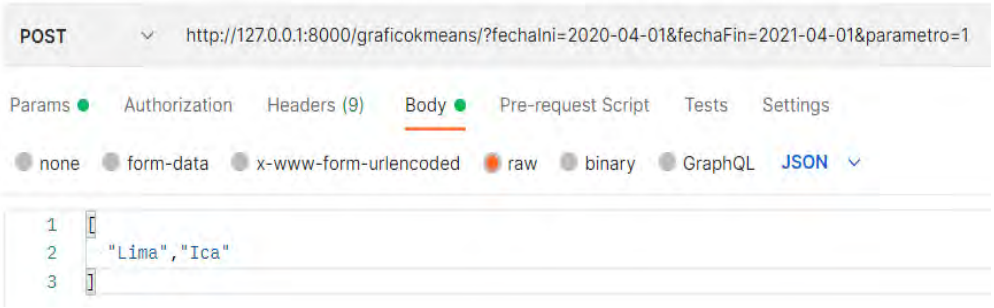


Figura I1. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means.

En la Figura I2, se presenta el resultado esperado luego de ejecutar la prueba. El primer elemento del arreglo obtenido es el gráfico y el segundo arreglo contiene la lista de los datos de las secuencias genómicas SARS-CoV-2.

Figura I2. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means.

2. Prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de agrupamiento con el algoritmo jerárquico donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; un arreglo de los nombres de los departamentos del Perú y el valor del parámetro del algoritmo de agrupamiento con los que se quiere realizar la consulta. En la Figura I3, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

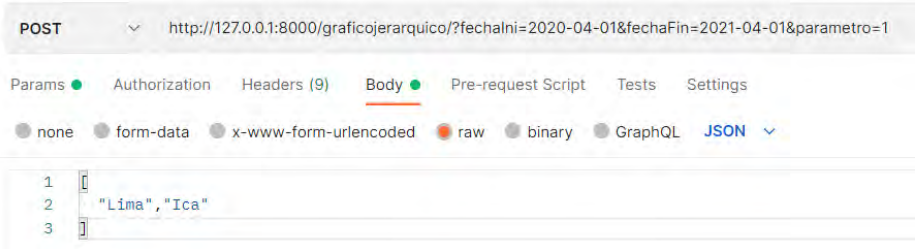


Figura I3. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico.

En la Figura I4, se presenta el resultado esperado luego de ejecutar la prueba. El primer elemento del arreglo obtenido es el gráfico y el segundo arreglo contiene la lista de los datos de las secuencias genómicas SARS-CoV-2.

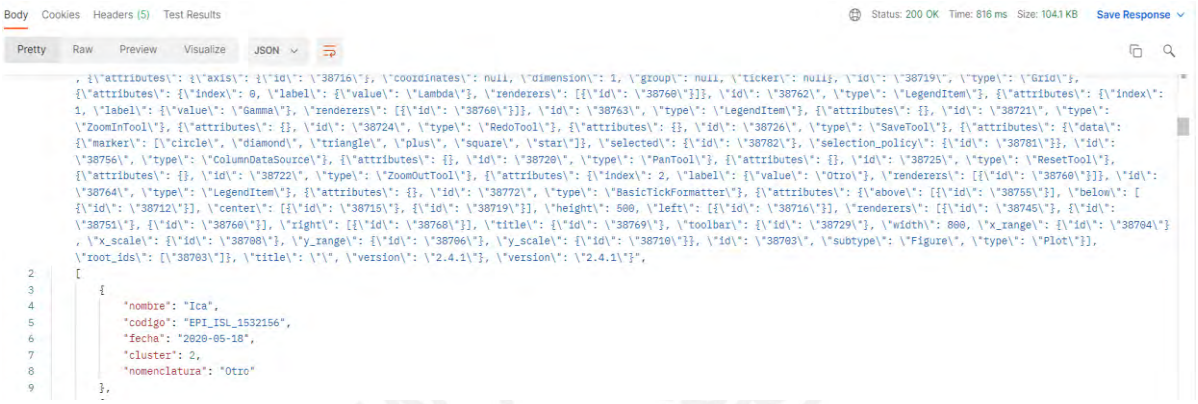


Figura I4. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico.

3. Prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de agrupamiento con el algoritmo DBSCAN donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; un arreglo de los nombres de los departamentos del Perú y el valor del parámetro del algoritmo de agrupamiento con los que se quiere realizar la consulta. En la Figura I5, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

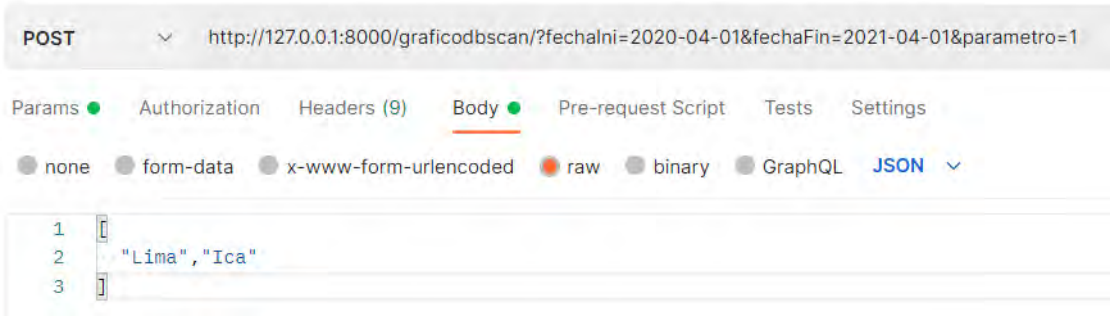


Figura I5. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN.

En la Figura I6, se presenta el resultado esperado luego de ejecutar la prueba. El primer elemento del arreglo obtenido es el gráfico y el segundo arreglo contiene la lista de los datos de las secuencias genómicas SARS-CoV-2.

Figura I6. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN.

4. Prueba del Servicio que realiza el dendrograma

Esta prueba tiene como objetivo verificar que se obtenga el gráfico del dendrograma donde se visualice la estructura o jerarquía de los clústeres que se pueden formar en formato HTML. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura I7, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

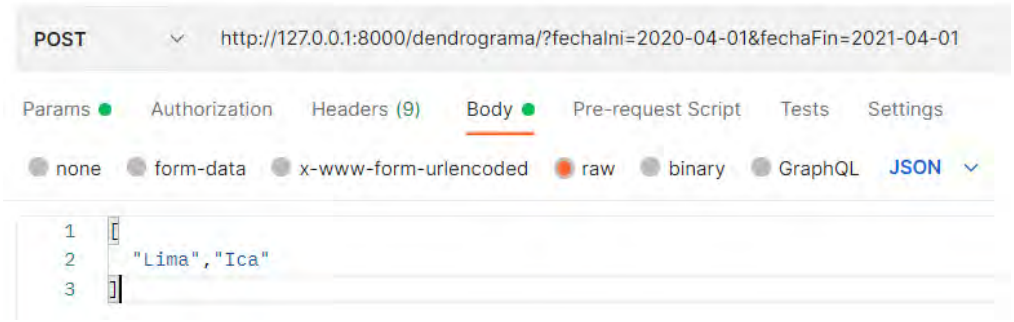


Figura I7. Datos de entrada para la prueba del Servicio que realiza el dendrograma.

En la Figura I8, se presenta el resultado esperado luego de ejecutar la prueba.

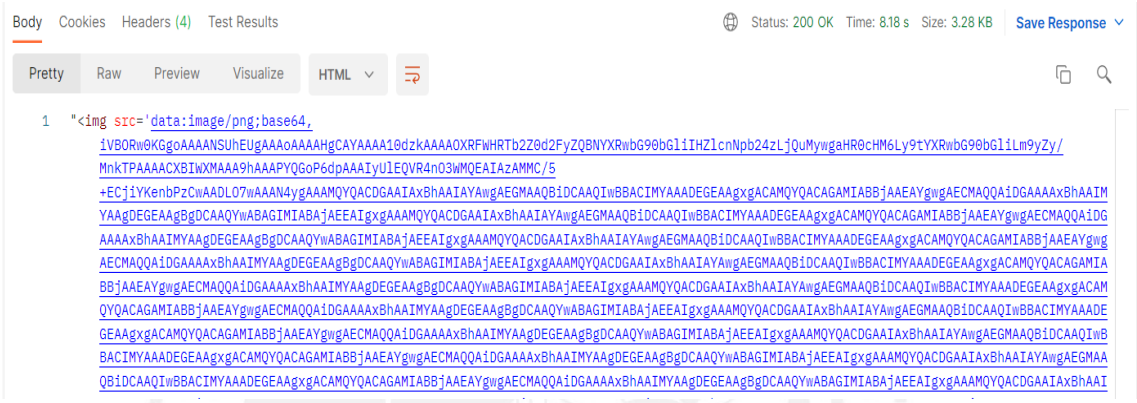


Figura I8. Resultado de la prueba del Servicio que realiza el dendrograma.

Anexo J: Reporte del resultado de pruebas del módulo de agrupamiento

Este anexo contiene el resultado de las pruebas unitarias aprobadas al 100%, esto indica que las pruebas se han completado de forma satisfactoria cumpliendo con los resultados esperados. A continuación, se detalla los resultados de las pruebas unitarias del módulo de agrupamiento. Estas pruebas unitarias se realizaron sobre un conjunto de 1365 datos de secuencias genómicas SARS-CoV-2. A continuación, se detalla cada prueba realizada y el resultado alcanzado.

1. **Prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo *k-means***

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de agrupamiento con el algoritmo *k-means* donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; un arreglo de los nombres de los departamentos del Perú y el valor del parámetro del algoritmo de agrupamiento con los que se quiere realizar la consulta. En la Figura J1, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

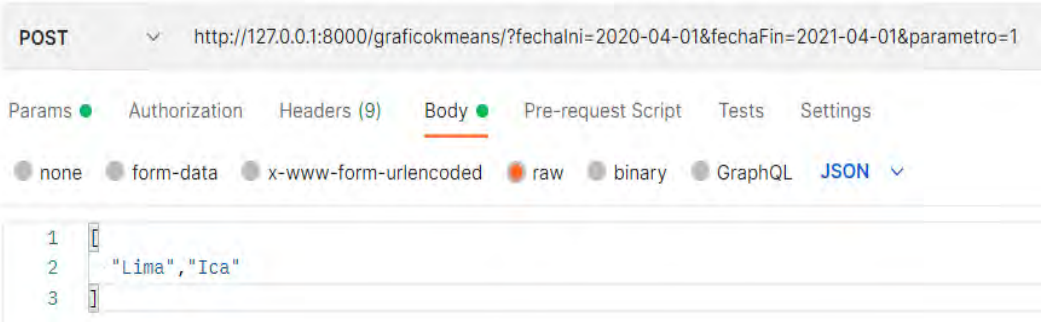


Figura J1. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means.

El resultado esperado es el gráfico de agrupamiento con el algoritmo *k-means* donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. En la Figura J2, se presenta el resultado obtenido luego de ejecutar la prueba. El primer elemento del arreglo obtenido es el gráfico y el segundo arreglo contiene la lista de los datos de las secuencias genómicas SARS-CoV-2.

Figura J2. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo k-means.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

2. **Prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico**

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de agrupamiento con el algoritmo jerárquico donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; un arreglo de los nombres de los departamentos del Perú y el valor del parámetro del algoritmo de agrupamiento con los que se quiere realizar la consulta. En la Figura J3, se muestra el formato de los datos

de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

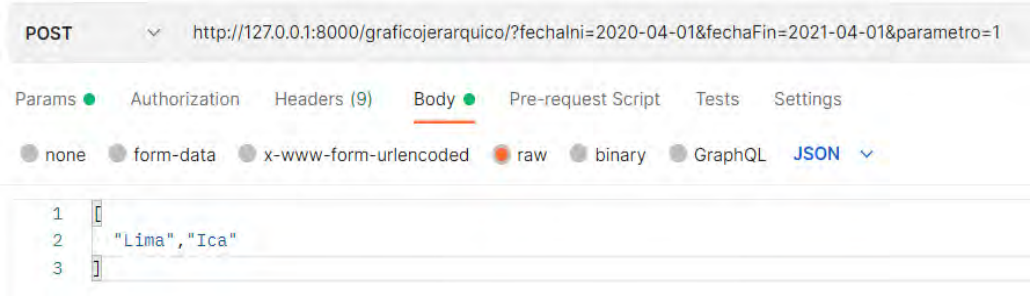


Figura J3. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico.

El resultado esperado es el gráfico de agrupamiento con el algoritmo jerárquico donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. En la Figura J4, se presenta el resultado obtenido luego de ejecutar la prueba. El primer elemento del arreglo obtenido es el gráfico y el segundo arreglo contiene la lista de los datos de las secuencias genómicas SARS-CoV-2.

Figura J4. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo jerárquico.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

3. **Prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN**

Esta prueba tiene como objetivo verificar que se obtenga el gráfico de agrupamiento con el algoritmo DBSCAN donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. Estos datos serán los siguientes: Departamento, ID de acceso de la secuencia genómica (identificador en la base de datos de GISAID), Fecha de recolección, N° de clúster y Nomenclatura según la OMS de la variante identificada.

Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; un arreglo de los nombres de los departamentos del Perú y el valor del parámetro del algoritmo de agrupamiento con los que se quiere realizar la consulta. En la Figura J5, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

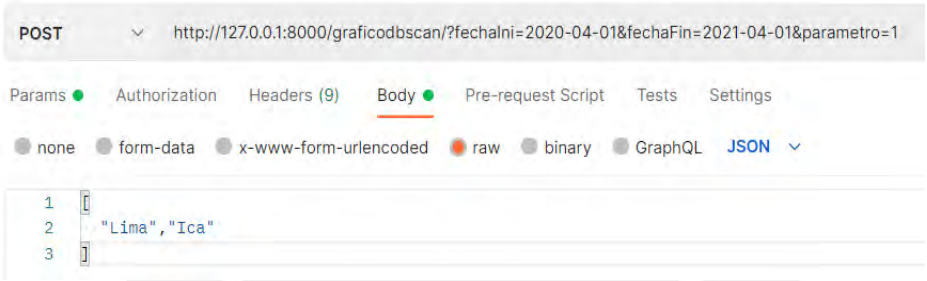


Figura J5. Datos de entrada para la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN.

El resultado esperado es el gráfico de agrupamiento con el algoritmo DBSCAN donde se muestre las variantes identificadas en el Perú y la lista con los datos de las secuencias genómicas SARS-CoV-2 en formato JSON. En la Figura J6, se presenta el resultado obtenido luego de ejecutar la prueba. El primer elemento del arreglo obtenido es el gráfico y el segundo arreglo contiene la lista de los datos de las secuencias genómicas SARS-CoV-2.

Figura J6. Resultado de la prueba del Servicio que realiza el gráfico de agrupamiento con el algoritmo DBSCAN.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

4. Prueba del Servicio que realiza el dendrograma

Esta prueba tiene como objetivo verificar que se obtenga el gráfico del dendrograma donde se visualice la estructura o jerarquía de los clústeres que se pueden formar en formato HTML. Se tiene como datos de entrada el rango de fechas, fecha inicio y fecha fin; y, un arreglo de los nombres de los departamentos del Perú con los que se quiere realizar la consulta. En la Figura J7, se muestra el formato de los datos de entrada, la fecha inicio, la fecha fin, el nombre del algoritmo y el valor del parámetro se colocan en la URL; y, en el body va el arreglo con los nombres de los departamentos.

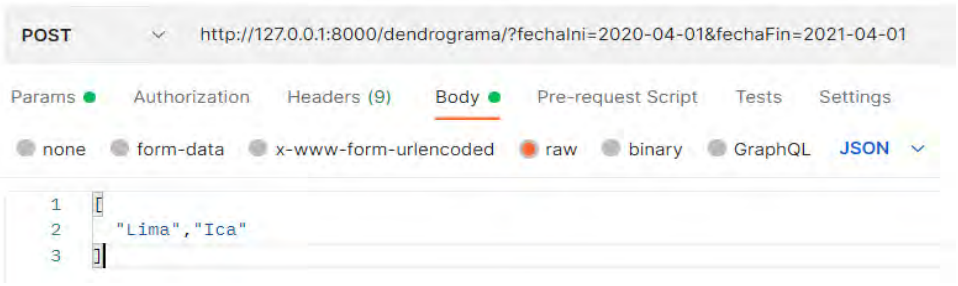


Figura J7. Datos de entrada para la prueba del Servicio que realiza el dendrograma.

El resultado esperado es el gráfico del dendrograma donde se visualice la estructura o jerarquía de los clústeres que se pueden formar en formato HTML. En la Figura J8, se presenta el resultado obtenido luego de ejecutar la prueba.

Figura J8. Resultado de la prueba del Servicio que realiza el dendrograma.

El resultado obtenido cumple con lo esperado como resultado de esta prueba, de esta manera, se verifica que la prueba ha sido aprobada satisfactoriamente.

En conclusión, a través de este reporte se verifica que los resultados de las pruebas han sido aprobados al 100%, cumpliendo con el resultado esperado en cada una de ellas.

Anexo K: Acta de validación del cumplimiento de los requerimientos del módulo de agrupamiento

Este anexo contiene el acta de validación del cumplimiento de los requerimientos del módulo de agrupamiento por parte de un especialista en ingeniería informática. El cual se presenta a continuación.

Acta de validación del cumplimiento de los requerimientos del módulo de agrupamiento

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2.

Descripción del resultado: Se han implementado los requerimientos relacionados al módulo de agrupamiento para así visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 recolectadas, esto con el fin de que los especialistas y analistas puedan ver la diversidad de secuencias genómicas SARS-CoV-2 en el Perú, así como también, comprender las posibles variantes que se encuentren presentes y encontrar patrones en la evolución del virus.

Evaluador: Dra. Soledad Espezua Llerena

Veredicto: Aceptado

Comentario adicional: Excelente trabajo, sobre clustering en secuencias genómicas de SARS-CoV-2. Sugiero reducir el tamaño de los gráficos, y mostrar los valores estadísticos en las representaciones visuales.

Firma

Lima, 05 de noviembre de 2021

Anexo L: Reporte de apreciación del módulo de agrupamiento

Este anexo contiene el Reporte de apreciación del módulo de visualización del análisis de agrupamiento por un especialista en bioinformática o biología molecular. El cual se presenta a continuación.

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Reporte de apreciación del módulo de agrupamiento

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Módulo para visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2.

Descripción del resultado: Se han implementado los requerimientos relacionados al módulo de agrupamiento para así visualizar el análisis de agrupamiento de las secuencias genómicas SARS-CoV-2 recolectadas, esto con el fin de que los especialistas y analistas puedan ver la diversidad de secuencias genómicas SARS-CoV-2 en el Perú, así como también, comprender las posibles variantes que se encuentren presentes y encontrar patrones en la evolución del virus.

Evaluador: Dr. Pedro Eduardo Romero Condori

Comentarios de apreciación sobre el módulo:

Se ve bastante bien. Es un poco complicado entender gráficamente cual es la variante predominante en cada grupo.

Firma

Lima, 05 de noviembre de 2021

Anexo M: Acta de validación de los requerimientos

Este anexo contiene el acta de validación del 100% de los requerimientos por parte de un especialista en bioinformática o biología molecular. El cual se presenta a continuación.

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Acta de validación de los requerimientos

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Lista de requerimientos a considerar para realizar la interactividad de las representaciones visuales.

Descripción del resultado: Se han identificado los requerimientos relacionados a la interactividad de las representaciones visuales de la herramienta analítica interactiva, con el fin de que los especialistas y analistas puedan analizar las secuencias genómicas de SARS-CoV-2 de manera sencilla e interactiva y realizar un análisis con mayor profundidad y entendimiento.

Evaluador: Dr. Pedro Eduardo Romero Condori

Veredicto: Aceptado

Comentario adicional: En los gráficos del final del Anexo D, se podría colocar los ejemplos de todos los requerimientos listados (once) y no solo de los primeros ocho.

Firma

Lima, 06 de septiembre de 2021

Anexo N: Acta de validación de cumplimiento de los requerimientos de usabilidad

Este anexo contiene el acta de validación del cumplimiento de los requerimientos de usabilidad por parte de un especialista en interacción humano computador (HCI). El cual se presenta a continuación.

PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Acta de validación de cumplimiento de los requerimientos de usabilidad

Título de Tesis: "Análisis de clustering de secuencias genómicas de SARS-CoV-2 identificadas en Perú"

Tesista: Carolina Estefanía Mejía Mujica

Nombre del resultado obtenido: Interfaces del módulo para visualizar la representación espacio-temporal y del módulo para visualizar el análisis de agrupamiento con capacidades interactivas de acuerdo con los requerimientos especificados.

Descripción del resultado: Se han implementado los requerimientos relacionados a la interactividad de las representaciones visuales de la herramienta analítica interactiva, con el fin de que los especialistas y analistas puedan analizar las secuencias genómicas de SARS-CoV-2 de manera sencilla e interactiva y realizar un análisis con mayor profundidad y entendimiento.

Evaluador: Dr. Freddy Alberto Paz Espinoza

Veredicto: Aceptado

Comentario adicional: Comentarios adjuntos a través de correo electrónico.

Firma

Lima, 08 de noviembre de 2021