



Universidad Nacional

SAN LUIS GONZAGA



Reconocimiento-NoComercial-CompartirIgual 4.0 Internacional

Esta licencia permite a otras combinar, retocar, y crear a partir de su obra de forma no comercial, siempre y cuando den crédito y licencia a nuevas creaciones bajo los mismos términos.

<http://creativecommons.org/licenses/by-nc-sa/4.0>



EVALUACION DE ORIGINALIDAD

CONSTANCIA

El que suscribe, deja constancia que se ha realizado el análisis con el software de verificación de similitud de la **TESIS** cuyo título es:

"APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING EN EL TRATAMIENTO DE DATOS COVID-19 DEL HOSPITAL MARÍA REICHE NEUMAN DE NAZCA"

Presentado por:

CUCHO RODAS JESÚS FÉLIX

De la **MAESTRÍA EN INGENIERÍA DE SISTEMAS** mención **GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN**.

Que, se ha recibido del operador del programa informático evaluador de originalidad de la Escuela de Posgrado de la UNICA, el informe automatizado de originalidad, el mismo que concluye de la siguiente manera:

El documento de investigación APRUEBA los criterios de originalidad con un porcentaje de similitud de 1%.

Para dar fe, se adjunta al presente el reporte de similitud de las bases de datos de iThenticate. En Ica 11 de diciembre de 2024

Atentamente


UNIVERSIDAD NACIONAL "SAN LUIS GONZAGA"
ESCUELA DE POSGRADO
Dr. MARIO GUSTAVO REYES MEJÍA
DIRECTOR

UNIVERSIDAD NACIONAL “SAN LUIS GONZAGA”
VICERRECTORADO DE INVESTIGACIÓN
ESCUELA DE POSGRADO
MAESTRÍA: INGENIERÍA DE SISTEMAS
Mención: Gestión de Tecnologías de la Información



TESIS

**“APLICACIÓN DE ALGORITMOS DE MACHINE LEARNING EN EL
TRATAMIENTO DE DATOS COVID-19 DEL HOSPITAL MARÍA
REICHE NEUMAN DE NAZCA”**

Línea de investigación:

Ciencias naturales, ingeniería y tecnologías sostenibles

PRESENTADA POR:

Bach. Jesús Félix Cucho Rodas

GRADO ACADÉMICO A OBTENER: MAESTRO

ASESOR:

Dr. EDGAR LEONARDO PEÑA CASAS

Ica - Perú

2026

Dedicatoria

Dedico esta Tesis a mi madre, quien siempre me inculcó los valores que tengo actualmente y me quedarán por toda la vida, por ser artífice en la culminación de cada paso de mi carrera, siempre dándome desde el cielo, el impulso para salir adelante ante cualquier adversidad. A mi papá Victor, quien me enseñó la perseverancia en cada paso que doy. A mi esposa Susan, mis hijos Leia, Italy y Pariz, que sin ellos no sentiría la motivación que tengo de avanzar cada día un paso adelante y que me tomen como ejemplo para su vida futura, y para ti mi último hijito, pequeño Luke Antoine, que constituyes ese pedacito lleno de amor en mi vida.

Agradecimientos

En primer lugar, expresar mi agradecimiento a Dios, quien es el que guía el camino que voy trazando día a día. A mi padre Víctor, quien siempre está pendiente por ver que obtenga un logro más. A mi asesor de la Tesis Dr. Edgar Peña Casas, por la dedicación, apoyo, orientación, dirección y el rigor que impuso para el correcto desarrollo de la presente Tesis.

ÍNDICE DE CONTENIDOS

Dedicatoria	ii
Agradecimientos	iii
Índice de contenidos	iv
Índice de tablas	v
Índice de figuras	vi
Resumen	vii
Abstract	viii
I. INTRODUCCIÓN	1
II. ESTRATEGIA METODOLÓGICA	5
1. Establecer el problema	5
2. Adquisición de los datos	5
3. Preparación de los datos	10
4. Exploración y análisis de los datos	12
5. Identificación del modelo	14
6. Visualización y comunicación	14
III. RESULTADOS	
a. Análisis descriptivo de las variables	15
b. Estimación de las variables	20
3.3 Análisis de los componentes principales – PCA	22
3.4 Clúster jerárquico de las variables	27
3.5 Evaluación de las variables obtenidas	30
IV. DISCUSIÓN	33
V. CONCLUSIONES	34
VI. RECOMENDACIONES	35
VII. REFERENCIAS BIBLIOGRÁFICAS	36
VIII. ANEXOS	38
Interfaces del sistema ESSI	38

ÍNDICE DE TABLAS

TABLA I ÁREAS REORGANIZADAS POR CONSECUENCIA COVID-19	3
TABLA II NUEVO PERSONAL CONTRATADO	4
TABLA III CARACTERÍSTICAS O VARIABLES SELECCIONADAS	11
TABLA IV DIAGNOSTICO DE COVID-19	15
TABLA V DIAGNOSTICO VS SEXO	15
TABLA VI DIAGNOSTICO VS PRIORIDAD	16
TABLA VII DIAGNOSTICO VS RESULTADOS DE ATENCIÓN	16
TABLA VIII DIAGNOSTICO VS SOSPECHA DE COVID-19	16
TABLA IX FALLECIDOS POR DISTRITO Y SEXO	17
TABLA X POSITIVOS POR DISTRITO Y SEXO	17
TABLA XI POSITIVOS POR DISTRITO Y MÉTODO DE PRUEBA	18
TABLA XII RESUMEN ESTIMACIÓN DE LAS VARIABLES	20
TABLA XIII CALCULO DEL ERROR ESTÁNDAR	21
TABLA XIV ESTIMACIÓN DEL ESTADÍSTICO EDAD	22

ÍNDICE DE FIGURAS

Fig. 1. Información de los datos	7
Fig. 2. Detalle de los registros de los datos	7
Fig. 3. Modelo para la adquisición de los datos del sistema ESSI	8
Fig. 4. Datos de los fallecidos COVID-19 provincia de Nazca	9
Fig. 5. Datos de los positivos COVID-19 provincia de Nazca	9
Fig. 6. Selección de características	10
Fig. 7. Datos de las características seleccionadas	11
Fig. 8. Dataset final con las características y registros seleccionados	12
Fig. 9. Selección de las variables de los datos abiertos del MINSA-Fallecidos COVID-19	13
Fig. 10. Selección de las variables de los datos abiertos del MINSA-Positivos COVID-19	14
Fig. 11. Comportamiento gráfico de los diagnosticados con COVID-19, virus identificado HMRN	18
Fig. 12. Comportamiento de los positivos COVID- MINSA	19
Fig. 13. Comportamiento de los fallecidos COVID-19 – MINSA	19
Fig. 14. Obtención de la media de la variable edad	21
Fig. 17. Selección de variables o características y variable objetivo (target)	22
Fig. 18. Conversión de las variables categóricas en variables continuas	23
Fig. 19. Gráfico “test scree” para el análisis de PCA inicial	24
Fig. 20. Gráfico “test scree” para el análisis de PCA ajustado	24
Fig. 21 variables de los componentes principales (1)	25
Fig. 22. variables ordenadas del Componente PC1	25
Fig. 23. Modelo de análisis PCA	26
Fig. 24. Proyección lineal de los componentes PC1 y PC2 para el diagnóstico	27
Fig. 25 Modelo de Clustering jerárquico (Hierarchical Clustering)	27
Fig. 26. Clúster jerárquico	28
Fig. 27. Datos clasificados y agrupados por Clúster	29
Fig. 28. Clustering de los datos mediante un mapa de calor	30
Fig. 29. Test y score de los algoritmos de predicción para las 14 variables	31
Fig. 31. Modelo de evaluación de las variables	32

RESUMEN

Objetivo. Estimar la incidencia de COVID-19 en el hospital María Reiche Neuman de Nazca 1er semestre 2021, mediante algoritmos de Machine Learning. El algoritmo para Clúster Jerárquico y Análisis de componentes principales.

Método. Se aplicó el método de Ciencia de los Datos para los casos presentados en el periodo de enero a junio del 2021, utilizando los registros de los diagnosticados del Hospital María Reiche Neuman como el diagnostico, sospecha de COVID-19, sexo entre otras. Además, los datos abiertos del MINSA para conocer la incidencia de los casos en el periodo en estudio sobre positivos y fallecidos.

Resultados. En el HMRN el análisis descriptivo encontró 1517 casos diagnosticados con COVID-19, y 1491 casos no fueron sospechosos de COVID-19, y solo 26 casos si fueron sospechosos. 422 fueron del sexo femenino (64.68%) y 1095 (53.42%) del sexo masculino. Se encontró que aun 908 (62.23%) casos permanecían en servicio, con 67(69.79%) casos trasferidos para hospitalización, en el caso de la Edad se estimó un intervalo de edades entre 35.78 a 46.36 las edades de mayor prevalencia (Tabla XII). Según los datos de MINSA se reportaron 368 casos fallecidos por COVID-19 de los cuales 248 fueron masculinos y 120 femeninos. En el caso de los positivos se reportaron 397 casos, de los cuales 247 fueron masculinos y 150 femeninos. En el caso de los positivos se reporta un caso menos que los fallecidos. Para el Clúster Jerárquico el dendrograma clasifico diversos grupos de las variables en estudio, el análisis PCA redujo las características iniciales de 14 a solo 6 características. Las 6 características se evaluaron con diversos algoritmos de predicción logrando tener una mayor precisión que utilizando las 14 características.

Conclusión. El análisis descriptivo con Machine Learning, detecto algunos patrones importantes, mientras que los algoritmos de clasificación y agrupamiento Clúster Jerárquico y el análisis PCA permitió conocer las características o variables que son mejores predictoras para los algoritmos de predicción.

Palabras clave: COVID-19, Machine Learning, Clúster Jerárquico, análisis PCA.

ABSTRACT

Objective. To estimate the incidence of COVID-19 in the hospital María Reiche Neuman of Nazca 1st semester 2021, by means of Machine Learning algorithms. The algorithm for Hierarchical Clustering and Principal Component Analysis.

Method. The Data Science method was applied for the cases presented in the period from January to June 2021, using the records of those diagnosed from the María Reiche Neuman Hospital such as diagnosis, COVID-19 suspicion, sex among others. In addition, open data from MINSA to know the incidence of cases in the period under study on positives and deaths.

Results. In the HMRN the descriptive analysis found 1517 cases diagnosed with COVID-19, and 1491 cases were not suspected of COVID-19, and only 26 cases were suspected. 422 were female (64.68%) and 1095 (53.42%) were male. It was found that 908 (62.23%) cases remained in service, with 67 (69.79%) cases transferred for hospitalization, in the case of age, the age range was estimated to be between 35.78 and 46.36, the most prevalent ages (Table XII). According to MINSA data, 368 cases of COVID-19 deaths were reported, of which 248 were male and 120 female. In the case of positives, 397 cases were reported, of which 247 were male and 150 female. In the case of positives, one case less than deaths was reported. For the Hierarchical Cluster the dendrogram classified various groups of the variables under study, the PCA analysis reduced the initial characteristics from 14 to only 6 characteristics. The 6 features were evaluated with different prediction algorithms achieving a higher accuracy than using the 14 features.

Conclusion. The descriptive analysis with Machine Learning, detected some important patterns, while the classification and clustering algorithms Hierarchical Clustering and the PCA analysis allowed to know the features or variables that are better predictors for the prediction algorithms.

Keywords: COVID-19, Machine Learning, Hierarchical Clustering, PCA analysis.

I. INTRODUCCIÓN

El COVID-19 ha traído una serie de cambios abruptos en el país, desde su aparición en nuestro país en el mes de marzo del 2020. Según información de la Organización Mundial de la Salud (OMS) al 02 de junio 2021, se han reportado un total de casos de 1,961,087, y de 184,169 fallecidos con un total de pacientes recuperados de 1,914,169. Igualmente se han reportado un total de nuevos casos de 5,618, total de nuevos fallecidos 115,165 y con nuevos pacientes recuperados 8,736. Este COVID19 es una enfermedad infecciosa causada por un tipo de coronavirus tipo 2 del síndrome respiratorio agudo grave o SARS-CoV-2 [1].

La salud de los peruanos igualmente se ha visto afectada por esta pandemia, al externo de tener más de 180,000.00 decesos por ella. Los hospitales del país empezaron a colapsar, la falta de oxígeno, camas UCI, pruebas moleculares, las limitaciones de personal de salud para la gran cantidad de pacientes que tenían que ser atendidos en los diversos hospitales, de la cual no se escapa el Hospital María Reiche Newman de la provincia de Nazca, ha creado una serie de problemas en la prestación de salud a la que desde antes de la pandemia tenía algunas deficiencias.

En el contexto mundial, nuestro país ha reportado la mayor cantidad de muertes en el mundo. Las defunciones ocasionadas por la emergencia del Covid-19 ha creado gran alarma en los gobiernos en diferentes sectores, como el sector productivo, responsables del sector salud pública y el público en general [2]. Esta preocupación mundial ha generado diversas investigaciones al respecto. Investigaciones como el modelo que abarca la atención en el lugar (point-of- care) y herramientas de inteligencia artificial, con lo cual los países podrían conocer lo real de la enfermedad y determinar diagnósticos más rápidos [3]. Por otro lado, es de vital importancia la confianza en la ciencia y los expertos para conocer estas epidemias [4], al respecto igualmente [4] reporta lo importante de tener confianza en la ciencia como predictora en la prevención del COVID-19. Al mismo tiempo se sabe que la inteligencia artificial y el Big Data se conjugan para afrontar diversos

problemas concernientes con el análisis de datos masivos [5] como es lo ocurrido con la COVID-19 en todo el planeta.

Para el presente estudio la inteligencia artificial y su rama como es el aprendizaje de máquina o Machine Learning son de gran ayuda. Identificar y seleccionar las variables relevantes de un conjunto de datos, mediante técnicas de selección de variables y dimensionalidad a fin contar con aquellas que mejor se ajuste al modelo [6] . y esto es de suma importancia porque según el estudio de [6], la cantidad de datos requeridos para utilizar un modelo de aprendizaje automático que generalice correctamente aumenta exponencialmente con mayor cantidad de datos. Adicionalmente crece las necesidades de poder de cómputo. Por lo tanto, la búsqueda de características relevantes que permitan a la comunidad científica identificar factores que faciliten la detección [5] es de suma importancia y de allí el uso del análisis de componentes principales (PCA) y Clúster jerárquico para la clasificación de dichas características.

De un lado el algoritmo de Clúster jerárquico es de suma importancia en la tarea de agrupar o seleccionar datos. La agrupación jerárquica de datos epidemiológicos contribuye a la toma de decisiones [7]. Ya que la toma de decisiones es de vital importancia en este tipo de escenarios cambiantes. Igualmente, los resultados de este agrupamiento podrían ser útiles para responsables políticos, médicos y otros, como lo establece [7].

De otro lado el algoritmo de selección de características o variables con el PCA, ayuda a reducir la dimensionalidad de los datos, seleccionando los que más aporten a los modelos de predicción en la Inteligencia artificial.

Otros estudios igualmente muestran el uso de algoritmos de Inteligencia artificial como el desarrollado por [8] donde el modelo propuesto de población un modelo autorregresivo de medias móviles (ARIMA) para series de tiempo en el contagio de Covid-19. Al mismo tiempo estudios donde algoritmos de inteligencia artificial aplicados a los datos, permiten un monitoreo de la evolución de la epidemia, con mayor precisión para la toma de decisiones [1].

Igualmente, otros estudios con inteligencia artificial han logrado predecir la ubicación del próximo brote, lo cual es fundamental para controlar la pandemia [9].

El método de clustering jerárquico consiste inicialmente en encontrar el clúster inicial y luego crear uno nuevo o separar uno existente dando lugar a otros dos para con lo que se maximiza la medida de similitud, o se minimizar la distancia [7].

Debemos precisar, finalmente, la utilización de herramientas tecnológicas para inteligencia artificial, podrían ayudar significativamente la batalla contra la COVID-19.

En nuestro país se viene registrando en el sistema de salud del MINSA las atenciones que se dan en los distintos hospitales contando con ingente cantidad de datos de todas estas atenciones, sean de

COVID-19 o no. Existe una gran cantidad de datos relacionadas con el COVID-19 que no está siendo aprovechada, los datos pueden generar información valiosa que ayude a los directivos a mejorar o a prevenir sus sistemas de atención, sobre todo del comportamiento del COVID-19., vale decir ayudaría a tomar decisiones más certeras.

Ante esta situación y con la gran cantidad de datos que se han generado, y con la finalidad de poder aportar en nuestra región, y específicamente en la provincia de Nazca en el Hospital María Reiche Newman a cargo del sistema donde se registran las atenciones, obliga a tomar estos datos para poder mediante algoritmos de Inteligencia Artificial, como el Machine Learning, conocer el comportamiento del virus en nuestra provincia.

Considerando que se estima hasta una cuarta Ola hacia el 2022, la investigación se considera de mucha relevancia importancia y justifica el estudio de los datos del HMRN con estas herramientas de Machine Learning.

Con la Declaratoria de Emergencia Sanitaria en el Perú motivada por la declaratoria Sanitaria de Pandemia Mundial por parte de la OMS a raíz del incremento de casos confirmados de COVID-19, a nivel Institucional en concordancia con los lineamientos determinados por el Ministerio de Salud (MINSA) se suspendieron las actividades de Consulta Ambulatoria, Procedimientos, priorizando atenciones en áreas críticas y áreas diferenciadas para atención de pacientes sospechosos y confirmados por la COVID-19 sin descuidar y garantizando a su vez la atención en áreas críticas NO COVID del hospital María Reiche Newman (HMRN).

Por ello hubo a necesidad de reorganizarse en salvaguarda de la salud de los pacientes asegurados de este distrito, de tal manera que existen dos áreas:

TABLA I
ÁREAS REORGANIZADAS POR CONSECUENCIA COVID-19

ÁREA COVID-19	ÁREA NO COVID
Hospitalización con 16 camas	Hospitalización con 06 camas
Observación con 02 posiciones (Carpa)	Observación con 03 posiciones
	Shock Trauma con 02 Posiciones
Tópico Medico	Tópico Medico
Toma de Muestras de Laboratorio	Toma de Muestras de Laboratorio
Ambiente de Rayos X	Ambiente de Rayos X

Al mismo tiempo, se gestionó para minimizar la Brecha de Recurso Humano existente de los diferentes grupos ocupacionales mediante la contratación de personal mediante la modalidad de CAS COVID durante el periodo Mayo 2020 – Abril 2021.

TABLA II
NUEVO PERSONAL CONTRATADO

GRUPO OCUPACIONAL	<u>CONTRATACIÓN DE PERSONAL</u>		RENUNCIAS
	2020	2021	
MEDICO GENERAL	3	1	2
UC ENFERMERÍA	1	3	1
TÉCNICO DE ENFERMERÍA	0	4	0
T.M. LABORATORIO	1	0	0
CHOFER ASISTENCIAL	1	0	0

El HMRN se vio obligado a atender ahora a un nuevo tipo de paciente ocasionado por el contagio del virus del COVID-19, estos nuevos pacientes empezaron a generar nuevos datos; datos que a la se han incrementado, y a pesar de obtener algunos indicadores básicos de los datos, estos no están siendo aprovechado de manera que genere conocimiento valioso para los directivos del hospital y hacer mejorar en el servicio de atención de este tópico con información que describa el comportamiento del virus en el HMRN.

El objetivo planteado para la investigación fue “Estimar la incidencia de COVID-19 en el hospital María Reiche Neuman de Nazca 1er semestre 2021, mediante algoritmos de Machine Learning”, con objetivos específicos para la misma: 1. Describir la incidencia de COVID-19 en el hospital María Reiche Neuman de Nazca en el 1er semestre 2021 con lenguaje de análisis de datos Python, 2. Determinar la clasificación de datos mediante algoritmo de Clustering Jerárquico de Machine Learning en el hospital María Reiche Neuman de Nazca en el 1er semestre 2021, y 3. Determinar los componentes principales de los datos de COVID-19 con algoritmo PCA de Machine Learning para el análisis de componente principales del hospital María Reiche Neuman de Nazca 1er semestre 2021. Bajo estos objetivos se tiene como propósito aprovechar la data existente en el HMRN y utilizar herramientas de Machine Learning para seleccionar y agrupar los datos y reducir la dimensionalidad para tener datos que aporten a los algoritmos de predicción de datos.

En general la inteligencia artificial que es una disciplina bastante genérica cuyo objetivo es dotar a una máquina inteligencia similar a un ser humano. En la actualidad se ha convertido en un aliado en el estudio de diversas enfermedades en el campo de la salud., donde estudios como el de [1] en la que mediante algoritmos de clustering se trazó como objetivo identificar grupos poblacionales en la ciudad de Cali-

Colombia, cuya vulnerabilidad poblacional dependen de sus características sociodemográficas y socioeconómicas de los casos de COVID19. O estudios como el de [2] en la que por medio de algoritmos de clustering, se busca agrupar a los países en función de los casos activos por población y los casos activos por población y zona del COVID19, basándose en los datos epidemiológicos de Johns Hopkins. En ese mismo contexto relacionado con el COVID19, en la que nuestro estudio busca clasificar ar los datos del COVID19 que se han presentado en la provincia de Nazca, tomando como fuente los datos registrados en el sistema SERVICIO DE SALUD INTELIGENTE (SSI) del MINSA.

Como se muestra en los estudios mencionados, se trata de clasificar los datos mediante técnicas de clustering jerárquico para ver que patrones de comportamiento del virus se presentan en las zonas estudiadas, por lo que reviste vital importancia igualmente conocer el comportamiento en la provincia de Nazca en la cual como trabajador me veo en la necesidad de aportar en el conocimiento del virus.

II. ESTRATEGIA METODOLÓGICA

En el presente trabajo de investigación sobre el COVID-19 en el HMRN de la provincia de Nazca, se aprovechó el acceso a la información del hospital sobre las atenciones que viene siendo registradas en el sistema SERVICIO DE SALUD INTELIGENTE (SSI), extrayendo la información de dicho sistema para el periodo en estudio 9564 registros, esta información tenía todos los tipos de atenciones realizadas en el hospital, por lo que hubo la necesidad de filtrar la información con solo la información que de una u otra manera estaba relacionada con el COVID-19, en el atributo o variable diagnóstico. Igualmente se aprovechó de la información registrada en el MINSA sobre el COVID-19 que es de acceso libre, similarmente a datos utilizando en la investigación de [2] y [10], donde se utilizaron datos de acceso libre en sus respectivos países.

Los datos finales del HMRN que solo correspondían a el estudio fueron 2483 y 8 características, atributos o variables a estudiar (Fecha atención, Hora de atención, años, sexo, prioridad, diagnostico, resultado de atención y sospecha de covid)

El diseño aplicado a la investigación se basó en un diseño no experimental, no habiendo manipulación de ninguna variable; retrospectivo debido a que los datos utilizados fueron recogidos por el MINSA como parte de su labor diaria, extrayendo la información para el periodo en estudio Enero – junio 2021.

Se empleo para el procesamiento de los datos el software para Data minning Orange, que está basado en el lenguaje Python, por ser de uso libre pero además de ser una herramienta que no requiere de programación, con un empleo visual.

El Machine Learning es un subcampo de la Inteligencia artificial, que es la aplicación práctica de modelos matemáticos para dotar a un ordenador de inteligencia artificial, pero sobre todo que aprende de forma automática en base a datos históricos, por lo tanto, necesita datos históricos para aprender.

TABLA I

TIPOS DE PROBLEMAS EN MACHINE LEARNING.

Tipo de aprendizaje Tipo de variable	Variable continua	Variable categórica
	Regresión	Clasificación
Supervisado		
No supervisado	Reducción de la dimensionalidad	Clustering

Según se desprende de la Tabla I, de los diversos tipos de algoritmos de Machine Learning, debido al tipo de variable con que se contaba de la base de datos del MINSA, se eligió el algoritmo para el análisis de Machine Learning de Inteligencia artificial con Clustering. Como se aprecia en la Tabla, este tipo de aprendizaje es del tipo no Supervisado. El mismo que se usa para segmentar o agrupar individuos u observaciones, de tal forma que los individuos al interior de cada clúster sean los más cercanos posibles, pero además cada clúster sea lo más diferente posible a los otros clústers [2].

Aprendizaje no supervisado: En el aprendizaje no supervisado no se tiene una variable la que se desea buscar, no existe clase, ni nada solo existen los datos. Por lo cual en este tipo de problemas lo que se desea es estudiar los datos, buscar la estructura dentro de estos datos o la síntesis de estos y a partir de ello automáticamente producir salidas (outputs).

Clustering (agrupamiento): en esencia un clúster, básicamente es un grupo, este algoritmo está orientado a las siguientes posibilidades de procesamiento:

1. Segmentar un Dataset para comprender sus características intrínsecas y encontrar grupos de elementos similares (anexo 01).
2. Detectar elementos anómalos (esto es, elementos que no pertenecen a ningún grupo)
3. Simplificar Dataset al agrupar conjuntos de variables con valores similares.

Algoritmo K-Medias: es un algoritmo de clustering, para el agrupamiento que sigue los pasos siguientes:

1. Elegir K que es un hiperparametro que se le tiene que indicar a K-medias. Es decisión del investigador decidir en cuanto clúster se desea dividir los datos.

2. Asignar K centroides al azar. Centroides son los puntos que están al centro de cada clúster.
3. Asignar cada observación al clúster más cercano. Se puede usar cualquier método, sin embargo, una buena técnica es usar la distancia euclidiana, según la siguiente fórmula.

$$\sqrt{\sum_{i=1}^n |x_i - y_i|^2} \dots\dots\dots (1)$$

4. Calcular la inercia, la condición de parada del algoritmo Kmedias es la inercia de los clústeres, medida como:

$$\sum_{i=0}^n \min (\|x_j - u_i\|^2), \mu_j \in C \dots\dots\dots (2)$$

La sumatoria al cuadrado de las distancias de cada observación al centroide más cercano (el clúster C al que pertenecen). Si la inercia ha llegado a su mínimo o es menor que un valor especificado anteriormente, el algoritmo se detiene.

5. Calcular los nuevos centroides, los centroides se calculan ahora como las medias de los puntos a los que pertenecen a cada centroide (de allí el nombre de K-medias). De allí vuelve al paso 3.

para Clúster Jerárquico para el agrupamiento de los datos, y el Algoritmo de PCA para encontrar los componentes principales de las variables del Dataset del hospital. Se comparó diversos algoritmos de predicción como Random Forest (árboles de decisión), Tree (árboles), SVM (máquina de soporte vectorial) y Neural Network (redes neuronales) para ver el nivel predictivos de dichos algoritmos en base a los datos del HMRN.

La metodología utilizada se basa en la metodología de la ciencia de los datos [11], para Inteligencia artificial con el lenguaje de programación Python, del software Orange Data minning, cuyas fases describimos a continuación:

1. Establecer el problema

El problema de que se presenta en el hospital María Reich Neuman cuenta con importante información sobre las atenciones médicas que se realizan a lo largo del tiempo. Como caso especial desde el 2020 con la aparición del COVID-19. Muchas personas se han atendido para casos de esta pandemia, sin embargo, los datos registrados no son analizados en la medida de que con ellos se puede obtener información valiosa para el hospital, más aún en un contexto donde se está apreciando un crecimiento en la detección de nuevos casos de COVID-19 en el primer semestre del 2022 y no se conoce las características de este comportamiento del virus en la provincia de Nazca.

2. Adquisición de los datos

Los datos para analizar fueron obtenidos de dos fuentes:

- ✓ Información registrada en el sistema del MINSA que tiene implementado en el hospital María Reiche Neuman.
- ✓ Datos para análisis que fueron obtenidos del portal del MINSA en los datos abiertos que publican sobre los contagiados, fallecidos y vacunados.

2.1 Datos del Hospital María Reiche Neuman: los datos del hospital María Reiche Neuman se extrajeron para el periodo en estudio Enero -Junio 2021 (extrayendo 01 archivo por cada mes en formato Excel) del sistema SSI, los archivos tuvieron que ser concatenados para unirlos en un solo Dataset. Esta data extraída una vez concatenada presentó la siguiente información.

Data Info - Orange	
Data Set Name	047_20210101_20210131_PacAtendxTopico-047_20210101_20210131_PacAtendx
Data Set Size	Rows: 9564 Columns: 33
Features	Categorical: 14 Numeric: 13
Targets	None
Meta Attributes	Categorical: - Numeric: - Text: 6
Location	Data is stored in memory
Data Attributes	
9564	

Fig.1. Información de los datos

La Fig. 1 El Dataset cuenta con 9564 registros y 33 columnas (características o variables) para el periodo de enero-junio 2021. Estas características tienen cuentan con 14 características de tipo categóricas y 13 características de tipo numérica, y 6 meta atributos de tipo texto.

	SVL_PROFESIONAL	AUTOGENERADO	DIRE	PACIENTE	DIAGNOSTICO	ESC_DIAGNOSTICO
1	9893408	5105120CANVDS...	1-22067967	CARNERO DE V...	R04.0	EPISTAXIS
2	9893408	8803011AMD4EL...	1-49902803	MENDOZA HU...	S40.0	CONTUSION D...
3	21446130	3405151MIEA008	1-22061993	MEJA JIMENEZ...	S70.0	CONTUSION D...
4	21446130	6805181EOEAM...	1-29387080	ESCOBEDO ESC...	V48.5	OCUPANTE DE...
5	21446130	3405161MIEA008	1-22065893	MEJA JIMENEZ...	S70.0	CONTUSION D...
6	21446130	8101230EIVAB009	1-40781529	ESPILCO VELAS...	S70.0	CONTUSION D...
7	21448226	8409290FCRAJ005	1-44121937	FALCON PEÑA A...	Z35.9	SUPERVISION D...
8	21448226	1802120LACRV...	1-80841167	LOBATON CHIR...	S52.8	FRACTURA DE...
9	21448226	8012101LNCVE...	1-40731817	LLANOS CHAV...	J80	SINDROME DE...
10	21448226	0406091LEVGA...	1-78050430	LOPEZ VARGAS...	R50.9	FIEBRE, NO ESP...
11	21448226	3006260CARGA...	1-23090311	CAMARGO DE...	G44.1	CEREALEA VASC...
12	21448226	7803080BOHHC...	1-40233128	BOHORQUEZ H...	G43.9	MBGAA A, NO...
13	21448226	6510111AN8DH...	1-22102763	ARANA BARDA...	S90.2	CONTUSION D...
14	21448226	8905071RMD005	1-45696651	RAMIREZ MAR...	H11.3	HEMORRAGIA...
15	21448226	7703050CCHC...	1-21568875	CANCHO HUA...	K30	DISPEPSIA
16	21448226	0803091MRSCP...	1-72626286	MARROQUIN S...	Z00.0	EXAMEN MEDIC...
17	21448226	8112171VLOCJO...	1-41827530	VALLADOLID O...	M79.2	NEURALGIA Y...
18	21448226	6202021CAEIX009	1-22062969	CANALES ESPI...	M54.4	LUMBAGO CO...
19	21448226	0906191SOPEFO...	1-82541838	SOTO PEREYRA...	Z48.9	CUIDADO POST...
20	21448226	8306281NADSF...	1-43163788	NIZAMA DIOSE...	M54.8	DORSALGIA, N...
21	21448226	6601160CNFEN...	1-58533343	CORNEJO PINE...	K30	DISPEPSIA
22	21448226	591A100GENE...	1-22064840	GOMEZ CUADR...	M75.3	TENDINITIS CAL...
23	21448226	76110205ORPHZ...	1-31187207	SEGOBIA RICH...	K02.9	CARIES DENTAL...
24	21448226	8812201FNCDIO...	1-40124766	FERNANDEZ C...	M54.0	SINDROME DE...
25	21448226	7112280PEMDA...	1-22061777	RAMIREZ MEN...	I13.9	ENFERMEDAD...
26	21448226	7404300CVVEED...	1-22092716	CUEVAS VENE...	O20.0	AMENAZA DE...
27	21448226	8802160CTRRD...	1-44833599	CASTAÑA EDA R...	K35.0	APENDICITIS A...

Fig. 2. Detalle de los registros de los datos

En este Dataset de las 27 características (features), se tenía 6.4% de datos perdidos o faltantes, y de las 6 características de tipo meta texto, se tenía 13.9% de datos perdidos o faltantes. Los valores de estos datos son mostrados en la Fig. 2.

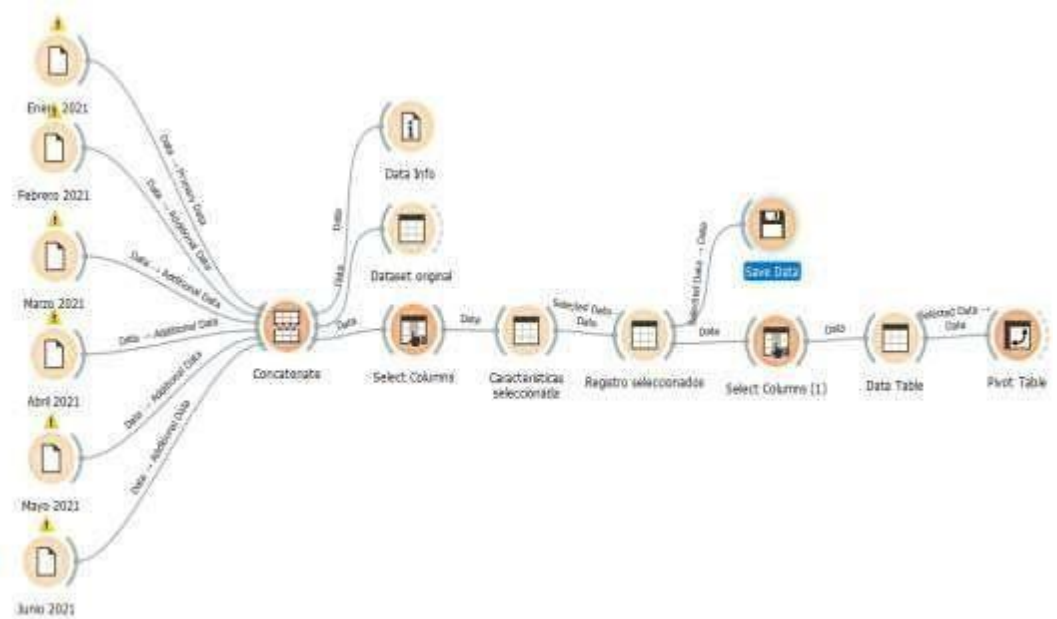


Fig. 3. Modelo para la adquisición de los datos del sistema ESSI

En la Fig. 3 se tienen 6 archivos de datos en bruto, uno por cada mes de enero – junio 2021, utilizando widgets para concatenar los datos y tener una sola base de datos, la misma que es guardada para utilización posterior.

2.2 Datos libres del portal MINSA

El MINSA pone a disposición en su portal y de acceso libre para la comunidad en general la información sobre los positivos COVID-19, los fallecidos del COVID-19 desde el inicio de la pandemia, pero que se tuvo que seleccionar solo los registros para el periodo en estudio.

En la Fig. 4 de los datos extraídos del portal del MINSA realizando el filtro para obtener solo los datos de los fallecidos de COVID-19 se han encontrado 368 registros, no se encontraron registros perdidos ni faltantes para las 10 variables reportadas para el periodo en estudio.

	CHA_FALLECIDO	FECHA_DECLARADA	SEXO	LABORATORIO	DEPARTAMENTO	PROVINCIA	DISTRITO	UBIGEO	UBIGEO	FECHA_CORTE
1	2021-06-20 00:00...	72	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.5151e+07	20220020
2	2021-06-20 00:00...	53	MASCULINO	Centro viral	ICA	NAZCA	MARCONA	11004	1.5705e+07	20220020
3	2021-06-28 00:00...	38	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	4.1738e+06	20220020
4	2021-06-26 00:00...	53	FEMENINO	Centro viral	ICA	NAZCA	EL INGENIO	11003	3.7004e+06	20220020
5	2021-06-26 00:00...	88	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.0845e+07	20220020
6	2021-06-26 00:00...	70	FEMENINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	1.3400e+07	20220020
7	2021-06-23 00:00...	70	MASCULINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	3.8715e+06	20220020
8	2021-06-23 00:00...	85	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.3491e+07	20220020
9	2021-06-22 00:00...	40	MASCULINO	Centro viral	ICA	NAZCA	MARCONA	11004	1.4035e+07	20220020
10	2021-06-22 00:00...	89	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.3927e+07	20220020
11	2021-06-20 00:00...	63	FEMENINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.5738e+07	20220020
12	2021-06-19 00:00...	44	FEMENINO	Centro viral	ICA	NAZCA	MARCONA	11004	1.8740e+07	20220020
13	2021-06-17 00:00...	45	MASCULINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	1.9858e+07	20220020
14	2021-06-17 00:00...	85	FEMENINO	Centro viral	ICA	NAZCA	NAZCA	11001	9.1243e+06	20220020
15	2021-06-16 00:00...	57	MASCULINO	Centro viral	ICA	NAZCA	MARCONA	11004	1.4821e+07	20220020
16	2021-06-15 00:00...	74	FEMENINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.5151e+07	20220020
17	2021-06-15 00:00...	79	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	3.7004e+06	20220020
18	2021-06-13 00:00...	59	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	2.8017e+06	20220020
19	2021-06-13 00:00...	60	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	5.3737e+06	20220020
20	2021-11-06 00:00...	86	FEMENINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	1.8914e+07	20220020
21	2021-10-06 00:00...	74	MASCULINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	2.7807e+06	20220020
22	2021-07-06 00:00...	87	FEMENINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.4760e+07	20220020
23	2021-06-06 00:00...	82	MASCULINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	5.7474e+07	20220020
24	2021-05-06 00:00...	71	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	9.9010e+06	20220020
25	2021-04-06 00:00...	91	FEMENINO	Centro viral	ICA	NAZCA	VISTA ALEGRE	11005	6.7707e+06	20220020
26	2021-04-06 00:00...	47	MASCULINO	Centro viral	ICA	NAZCA	NAZCA	11001	1.2872e+07	20220020

Fig. 4. Datos de los fallecidos COVID-19 provincia de Nazca

	FECHA_RESULTADO	FECHA_CORTE	DEPARTAMENTO	PROVINCIA	DISTRITO	METODOLOGIA	EDAD	SEXO
1	2021-06-30 00:00...	20220020	ICA	NAZCA	EL INGENIO	AG	64	MASCULINO
2	2021-06-28 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	27	MASCULINO
3	2021-06-28 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	21	MASCULINO
4	2021-06-23 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	PR	44	FEMENINO
5	2021-06-18 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	38	FEMENINO
6	2021-06-18 00:00...	20220020	ICA	NAZCA	MARCONA	AG	40	FEMENINO
7	2021-06-16 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	69	FEMENINO
8	2021-06-14 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	40	FEMENINO
9	2021-06-14 00:00...	20220020	ICA	NAZCA	MARCONA	AG	29	FEMENINO
10	2021-06-14 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	39	MASCULINO
11	2021-06-14 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	34	FEMENINO
12	2021-06-14 00:00...	20220020	ICA	NAZCA	MARCONA	AG	34	MASCULINO
13	2021-11-06 00:00...	20220020	ICA	NAZCA	NAZCA	PCR	55	MASCULINO
14	2021-09-06 00:00...	20220020	ICA	NAZCA	MARCONA	AG	57	MASCULINO
15	2021-07-06 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	40	FEMENINO
16	2021-04-06 00:00...	20220020	ICA	NAZCA	MARCONA	AG	49	FEMENINO
17	2021-03-06 00:00...	20220020	ICA	NAZCA	NAZCA	PCR	29	FEMENINO
18	2021-02-06 00:00...	20220020	ICA	NAZCA	MARCONA	PR	61	MASCULINO
19	2021-01-06 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	72	MASCULINO
20	2021-05-31 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	71	FEMENINO
21	2021-05-31 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	43	FEMENINO
22	2021-05-31 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	PR	38	FEMENINO
23	2021-05-28 00:00...	20220020	ICA	NAZCA	CHANGULLO	AG	42	MASCULINO
24	2021-05-27 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	40	FEMENINO
25	2021-05-27 00:00...	20220020	ICA	NAZCA	VISTA ALEGRE	AG	73	FEMENINO
26	2021-05-27 00:00...	20220020	ICA	NAZCA	EL INGENIO	AG	45	FEMENINO

Fig. 5. Datos de los positivos COVID-19 provincia de Nazca

En la Fig. 5 de los datos extraídos del portal del MINSA realizando el filtro para obtener solo los datos de los positivos de COVID-19 se han encontrado 368 registros, no se encontraron registros perdidos ni faltantes para las 8 variables reportadas para el periodo en estudio.

3. Preparación de los datos

3.1 Selección de las características o variables

El Dataset estaba conformado por 27 características, por lo que hubo primero la necesidad de seleccionar las características más relevantes para ser estudiadas, ya que no todas aportaban al estudio. Después de la selección de las características importantes para el estudio el Dataset se redujo a 8 características (features).

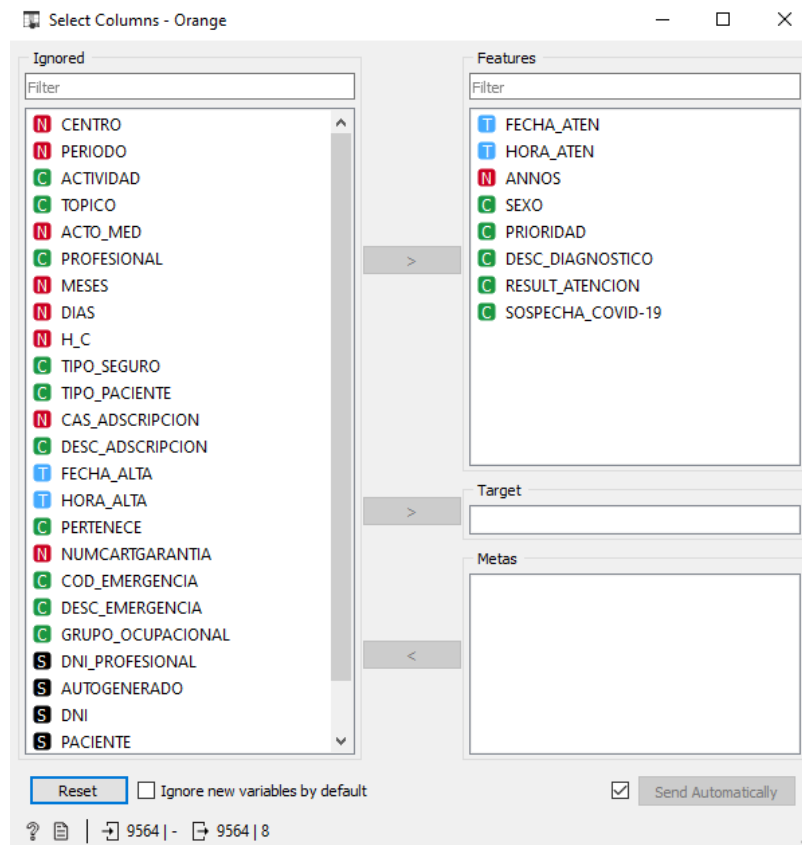


Fig. 6. Selección de características

La Fig. 6 muestra la selección de 8 características, aún se conservan cantidad inicial de registros. En la parte inferior se muestra los 9564 registros.

Los datos de las características seleccionadas se muestran en la Fig. 7

	FECHA_ATEN	HORA_ATEN	AÑOS	SEXO	PRIORIDAD	DESC_DIAGNOSTICO	RESULT_ATENCION	SUSPECHA_COVID
1	2021-01-21 00:00	08:22:00	89	F	PRIORIDAD III	EPISODIOS	RACIENTE CON	NO
2	2021-01-21 00:00	12:51:00	52	M	PRIORIDAD II	CONTUSION D.	ALTA MEDICA	NO
3	2021-01-28 00:00	13:54:00	86	M	PRIORIDAD II	CONTUSION D.	RACIENTE CON	NO
4	2021-01-20 00:00	20:20:00	52	M	PRIORIDAD I	OCURANTE DE	I	NO
5	2021-01-28 00:00	13:39:00	56	M	PRIORIDAD II	CONTUSION D.	REFERENCIA	NO
6	2021-01-08 00:00	22:02:00	39	F	PRIORIDAD II	CONTUSION D.	ALTA MEDICA	NO
7	2021-01-18 00:00	07:04:00	36	F	PRIORIDAD II	SUPERVISION D.	RACIENTE CON	NO
8	2021-01-23 00:00	15:41:00	2	F	PRIORIDAD I	FRACTURA DE	RACIENTE CON	NO
9	2021-01-24 00:00	16:36:00	40	M	PRIORIDAD II	SINDROME DE	RACIENTE CON	NO
10	2021-01-25 00:00	20:47:00	76	M	PRIORIDAD II	FIEBRE, NO ESP.	RACIENTE CON	NO
11	2021-01-23 00:00	12:10:00	76	F	PRIORIDAD III	CEFALEA VASC.	RACIENTE CON	NO
12	2021-01-18 00:00	06:57:00	41	F	PRIORIDAD III	MIGRAÑA, NO	RACIENTE CON	NO
13	2021-01-23 00:00	20:09:00	55	M	PRIORIDAD IV	CONTUSION D.	RACIENTE CON	NO
14	2021-01-06 00:00	20:21:00	31	M	PRIORIDAD IV	HEMORRAGA	RACIENTE CON	NO
15	2021-01-04 00:00	19:03:00	83	F	PRIORIDAD V	DISPEPSIA	ALTA MEDICA	NO
16	2021-01-04 00:00	10:11:00	34	M	PRIORIDAD V	EXAMEN MED.	ALTA MEDICA	NO
17	2021-01-05 00:00	15:42:00	39	M	PRIORIDAD V	REUBALGA Y	ALTA MEDICA	NO
18	2021-01-27 00:00	12:40:00	58	M	PRIORIDAD V	LUMBAGO COL.	ALTA MEDICA	NO
19	2021-01-23 00:00	10:16:00	11	M	PRIORIDAD V	CUIDADO POST.	ALTA MEDICA	NO
20	2021-01-23 00:00	10:22:00	35	M	PRIORIDAD IV	DORSALGA, N.	ALTA MEDICA	NO
21	2021-01-23 00:00	10:21:00	55	F	PRIORIDAD IV	DISPEPSIA	ALTA MEDICA	NO
22	2021-01-23 00:00	11:06:00	74	F	PRIORIDAD IV	TEODONTIS CA.	ALTA MEDICA	NO
23	2021-01-23 00:00	15:20:00	44	F	PRIORIDAD IV	CARIES DENTAL	ALTA MEDICA	NO
24	2021-01-23 00:00	18:42:00	31	M	PRIORIDAD IV	SINDROME DE	ALTA MEDICA	NO
25	2021-01-05 00:00	14:57:00	46	F	PRIORIDAD II	ENFERMEDAD	RACIENTE CON	NO
26	2021-01-05 00:00	12:46:00	46	F	PRIORIDAD II	AMENAZA DE	RACIENTE CON	NO
27	2021-01-04 00:00	13:59:00	32	F	PRIORIDAD II	APENDICITIS A.	RACIENTE CON	NO
28	2021-01-05 00:00	18:59:00	71	M	PRIORIDAD II	DIARREA Y GAS.	RACIENTE CON	NO
29	2021-01-11 00:00	13:03:00	3	F	SINDROME DE	ENTEROCOLITIS	RACIENTE CON	NO

Fig. 7. Datos de las características seleccionadas

En la Fig. 7 se muestra los datos de las características que han sido seleccionada, y el Dataset tiene ahora solo 1.2% de datos faltantes o perdidos. Estos datos si son relevantes para el análisis con algoritmos de Machine Learning. Las mismas que se detallan en la Tabla III.

TABLA III
CARACTERÍSTICAS O VARIABLES SELECCIONADAS

Característica o variable	Tipo de variable
Fecha de atención	Fecha
Hora de atención	Fecha
Años	Numérico
Sexo	Categorico
Prioridad	Categorico
Diagnóstico	Categorico
Resultado de atención	Categorico
Sospecha de COVID	Categorico

3.2 Selección de los registros

Para tener la información referente al COVID-19, se tuvo que realizar el filtrado de la información de la característica DESC_DIAGNOSTICO, en la que se tiene los diagnósticos de todos los pacientes atendidos en el hospital, sin embargo, para el análisis se tuvo que seleccionar solo los diagnósticos relacionados con el COVID-19. Los registros sobre el tema tenían como diagnóstico: COVID19, VIRUS IDENTIFICADO y COVID19, VIRUS

NO IDENTIFICADO, por lo que nuestro Dataset se redujo a 2483 registros los que se presentan en la Fig.5.

Para ir resguardando el Dataset de estudio, en este punto se tuvo que guardar los registros en un nuevo archivo independiente.

	FEO_H_ATEN	HORA_ATEN	AÑOS	SEXO	PRIORIDAD	DESC_DIAGNOSTICO	RESULT_ATENCION	SEXO
24	2021-01-29 00:00	20:16:00	41	M	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
25	2021-01-31 00:00	14:46:00	58	M	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
26	2021-01-31 00:00	16:14:00	33	M	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
27	2021-01-22 00:00	08:48:00	44	F	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
28	2021-01-23 00:00	17:47:00	57	M	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
29	2021-01-23 00:00	18:41:00	53	M	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
30	2021-01-26 00:00	18:52:00	39	F	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
31	2021-01-26 00:00	19:45:00	6	M	PRIORIDAD II	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
32	2021-01-14 00:00	19:30:00	21	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
33	2021-01-21 00:00	19:57:00	39	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
34	2021-01-22 00:00	21:35:00	38	F	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
35	2021-01-23 00:00	08:04:00	34	F	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
36	2021-01-23 00:00	01:18:00	41	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
37	2021-01-14 00:00	08:16:00	61	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
38	2021-01-13 00:00	15:58:00	69	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
39	2021-01-13 00:00	16:11:00	44	F	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
40	2021-01-18 00:00	17:38:00	38	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
41	2021-01-20 00:00	17:28:00	34	F	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
42	2021-01-20 00:00	20:09:00	38	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
43	2021-01-20 00:00	17:54:00	34	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
44	2021-01-21 00:00	15:29:00	38	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
45	2021-01-21 00:00	17:29:00	33	F	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	PACIENTE CONL.	NO
46	2021-01-11 00:00	11:11:00	49	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	ALTA MEDICA	NO
47	2021-01-13 00:00	12:50:00	23	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	ALTA MEDICA	NO
48	2021-01-16 00:00	19:35:00	41	M	PRIORIDAD V	COVID-19, VIRUS IDENTIFICADO	ALTA MEDICA	NO

Fig. 8. Dataset final con las características y registros seleccionados

En la Fig. 8 se muestra que el resultado del archivo final cuenta con 8 características, y 2483 registro de datos. Estos datos finales no tienen registro perdidos por lo cual no va a ser necesario la imputación de datos (técnica que se realiza para datos faltantes o perdidos).

En cuanto a los datos de los fallecidos y positivos del portal libre del MINSA, los datos se tuvieron que filtrar para el periodo en estudio, vale decir de enero a junio del 2021.

4. Exploración y análisis de los datos

4.1 Exploración datos del hospital María Reiche Neuman

Con los datos se realizó una serie de análisis de las características, mediante técnica de pivoteo de datos, que se presentan en el capítulo de resultados.

4.2 Exploración datos del MINSA sobre positivos y fallecidos

Los datos obtenidos del portal del MINSA, para el caso de los positivos y los fallecidos, se seleccionó las variables adecuadas para el análisis como se muestra en las Figs. 9 y 10.

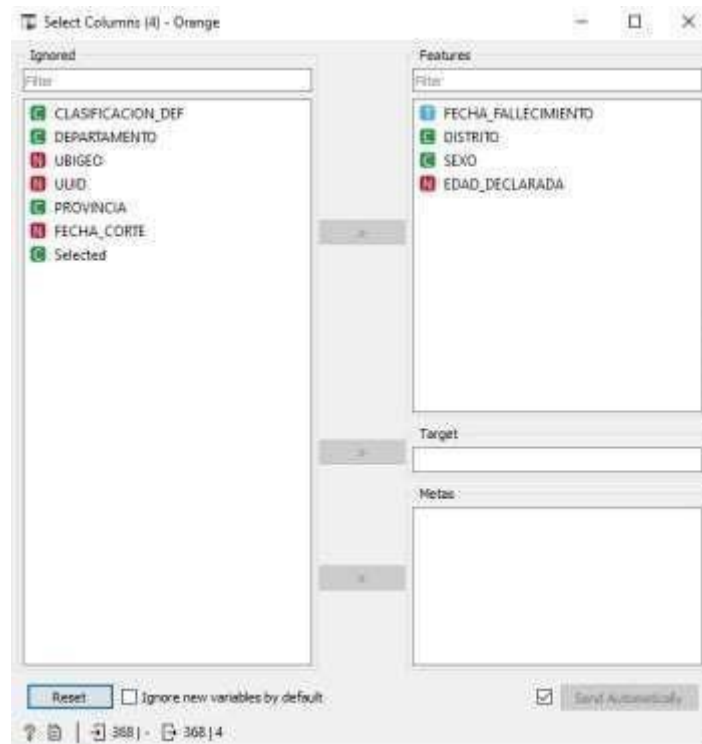


Fig. 9. Selección de las variables de los datos abiertos del MINSA-Fallecidos COVID-19

En la Fig. 9 los datos seleccionados para el análisis se considera la Fecha para poder utilizarla como serie de tiempo y ver el comportamiento de los fallecidos COVID-19.

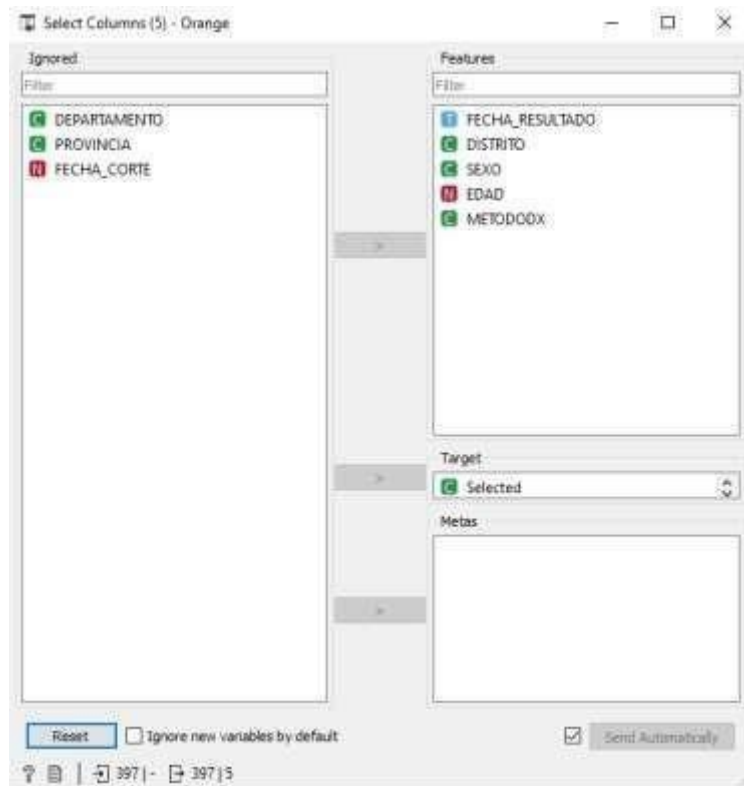


Fig. 10. Selección de las variables de los datos abiertos del MINSA-Positivos COVID-19

En la Fig. 10 al igual que la Fig. 9, los datos seleccionados para el análisis se considera la Fecha para poder utilizarla como serie de tiempo y ver el comportamiento de los positivos con COVID-19.

5. Identificación del modelo

La identificación de los modelos de Machine Learning, se enmarca en dos tipos, por un lado, el PCA (Principal Component Analysis/Análisis de Componentes Principales) para la identificación de los componentes principales de la data del hospital. Hay que destacar que PCA es una de las herramientas para selección de variables pudiendo también utilizar como complemento otros algoritmos como la matriz de correlación, el diagrama de tamiz, gráficos de dispersión y gráficos de proyección lineal.

y el algoritmo de Clustering jerárquico especificado en los objetivos del estudio. Pero que pueden ser utilizados otros algoritmos como el K-means, KNN, CN2.

6. Visualización y comunicación

Con la herramienta de procesamiento Orange Datamining, se obtendrá los resultados gráficos de los para los algoritmos en estudio.

III. RESULTADOS

3.1 Análisis descriptivo de las variables

A. Hospital María Reiche Neuman

Son datos de tipo categórico, se presenta una serie de tablas resultantes de la combinación de estas por medio del widget Pivot table.

TABLA IV
DIAGNOSTICO DE COVID-19

DESC_DIAGNOSTICO	DESC_DIAGNOSTICO		
	Count COVID-19, VIRUS IDENTIFICADO	COVID-19, VIRUS NO IDENTIFICADO	Total
COVID-19, VIRUS IDENTIFICADO	1517.0	0.0	1517.0
COVID-19, VIRUS NO IDENTIFICADO	0.0	966.0	966.0
Total	1517.0	966.0	2483.0

En la Tabla IV, se muestra que para el periodo en estudio se han identificado 1517 pacientes con diagnóstico COVID-19, virus identificado y 966 pacientes COVID-19, virus no identificado.

TABLA V
DIAGNOSTICO VS SEXO

SEXO	DESC_DIAGNOSTICO		
	Count COVID-19, VIRUS IDENTIFICADO	COVID-19, VIRUS NO IDENTIFICADO	Total
F	422.0	368.0	790.0
M	1095.0	598.0	1693.0
Total	1517.0	966.0	2483.0

En la Tabla V, para el diagnóstico COVID-19, virus identificado se encontraron 422 pacientes del sexo femenino y 1095 pacientes del sexo masculino. De la misma manera para el diagnóstico COVID-19, virus no identificado se encontraron 368 pacientes del sexo femenino y 598 pacientes del sexo masculino.

TABLA VI
DIAGNOSTICO VS PRIORIDAD

		DESC_DIAGNOSTICO	
		Count COVID-19, VIRUS IDENTIFICADO	COVID-19, VIRUS NO IDENTIFICADO
PRIORIDAD	PRIORIDAD II - EMERGENCIA	53.0	37.0
	PRIORIDAD III - URGENCIA MAYOR	489.0	137.0
	PRIORIDAD IV - URGENCIA MENOR	810.0	628.0
	PRIORIDAD V - CONSULTA	165.0	161.0
	Total	1517.0	966.0
		Total	2483.0

En la Tabla VI, El diagnóstico COVID-19, virus identificado, se encontraron 810 pacientes en la prioridad IV de emergencia menor, 489 pacientes en prioridad III urgencia mayor, en emergencia se encontró solo 53 casos y 165 casos en consulta. Igualmente, en cuanto al diagnóstico COVID-19, virus no identificado se diagnosticaron en urgencia menor 628 pacientes

TABLA VII
DIAGNOSTICO VS RESULTADOS DE ATENCIÓN

		DESC_DIAGNOSTICO	
		Count COVID-19, VIRUS IDENTIFICADO	COVID-19, VIRUS NO IDENTIFICADO
RESULT_ATENCION	ALTA MEDICA	528.0	378.0
	PACIENTE CONTINUA EN SERVICIO	908.0	551.0
	REFERENCIA	10.0	7.0
	RETIRO VOLUNTARIO	4.0	1.0
	TRANSFERENCIA HOSPITALIZACION	67.0	29.0
		Total	1517.0
		Total	966.0
		Total	2483.0

La Tabla VII, para el diagnóstico COVID-19, virus identificado se encontraron 908 pacientes que continuaban en servicio, 528 pacientes con alta médica, 67 que tuvieron transferencia para hospitalización con solo 10 pacientes referenciados y 4 pacientes que se retiraron voluntariamente.

TABLA VIII
DIAGNOSTICO VS SOSPECHA DE COVID-19

		DESC_DIAGNOSTICO	
		Count COVID-19, VIRUS IDENTIFICADO	COVID-19, VIRUS NO IDENTIFICADO
SOSPECHA_COVID_19	NO	1491.0	945.0
	SI	26.0	21.0
	Total	1517.0	966.0
		Total	2483.0

En la Tabla VIII, según el diagnóstico COVID-19, virus identificado, se encontraron 1491 pacientes que no tenían sospecha del COVID-19, y solo 26 pacientes sospechosos, vale decir se

ingresaron estos pacientes sin sospecha de COVID-19 pero finalmente se les identificó el virus. En cuanto al diagnóstico COVID-19., virus no identificado se tuvo la sospecha de 21 pacientes que posiblemente tenían COVID-19.

B. Datos abiertos MINSA

TABLA IX
FALLECIDOS POR DISTRITO Y SEXO

		DISTRITO					
	Count	CHANGUILLO	EL INGENIO	MARCONA	NAZCA	VISTA ALEGRE	Total
SEXO	FEMENINO	1.0	6.0	33.0	62.0	18.0	120.0
	MASCULINO	4.0	5.0	80.0	116.0	43.0	248.0
	Total	5.0	11.0	113.0	178.0	61.0	368.0

Según la Tabla IX, los fallecidos en el periodo en investigación según el sexo muestran 120 casos del sexo femenino y 248 del sexo masculino. Habiéndose diagnosticado 422 casos del sexo Femenino y 1095 diagnosticados del sexo masculino según la Tabla II.

TABLA X
POSITIVOS POR DISTRITO Y SEXO

		DISTRITO					
	Count	CHANGUILLO	EL INGENIO	MARCONA	NAZCA VISTA ALEGRE	Total	
SEXO	FEMENINO	11.0	7.0	63.0	1.0	68.0	150.0
	MASCULINO	17.0	14.0	149.0	5.0	62.0	247.0
	Total	28.0	21.0	212.0	6.0	130.0	397.0

En la Tabla X, de los casos positivos de COVID-19 determinados por el MINSA, 150 corresponden al sexo femenino y 247 al sexo masculino, teniendo una mayor cantidad de contagios en el distrito de Vista Alegre con 130 casos, seguido de Changuillo con 28 casos.

TABLA XI
POSITIVOS POR DISTRITO Y MÉTODO DE PRUEBA

		DISTRITO					
	Count	CHANGUILLO	EL INGENIO	MARCONA	NAZCA	VISTA ALEGRE	Total
METODOXX	AG	28.0	11.0	115.0	0.0	100.0	254.0
	PCR	0.0	0.0	42.0	6.0	3.0	51.0
	PR	0.0	10.0	55.0	0.0	27.0	92.0
	Total	28.0	21.0	212.0	6.0	130.0	397.0

En la Tabla XI, se muestra los casos positivos reportados por el MINSA, en la que la prueba AG tuvo la mayor cantidad de detectados con 254 seguido de la prueba PCR con 51 y 92 para la prueba PR.

Análisis gráfico de serie de tiempo de los diagnósticos del HMRN, positivos y fallecidos reportados por el MINSA

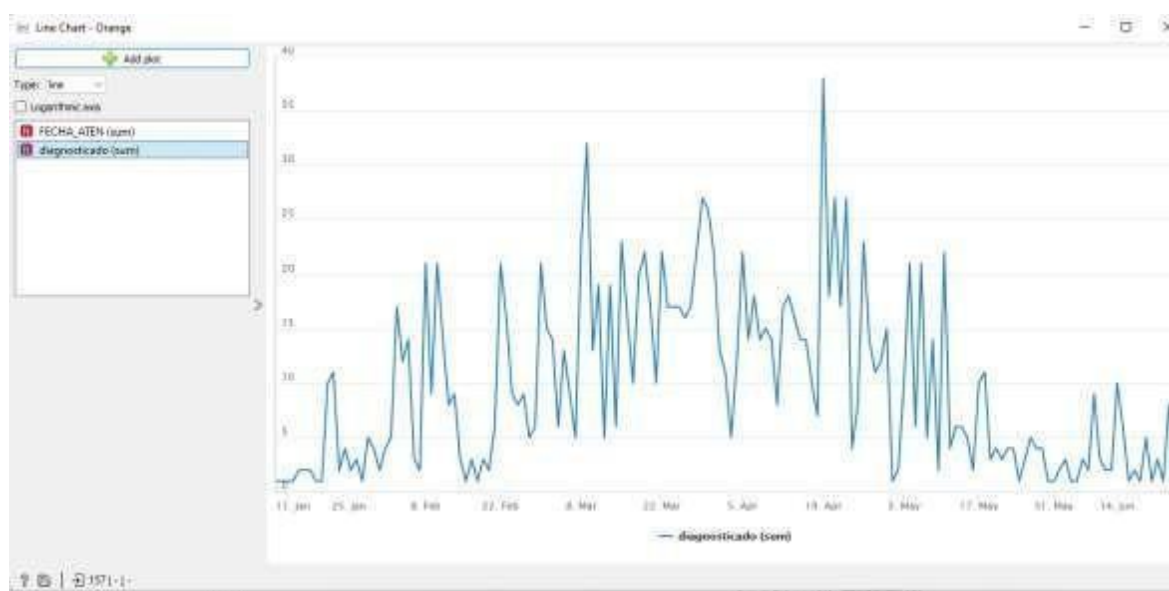


Fig. 11. Comportamiento gráfico de los diagnosticados con COVID-19, virus identificado HMRN.

La Fig. 11, muestra el comportamiento mediante series de tiempo de los diagnosticados con COVID-19, virus identificado, vale decir que el virus fue identificado en el diagnóstico.

La gráfica muestra un nivel alto de diagnosticados y virus identificado en los meses de febrero a mediados de mayo.

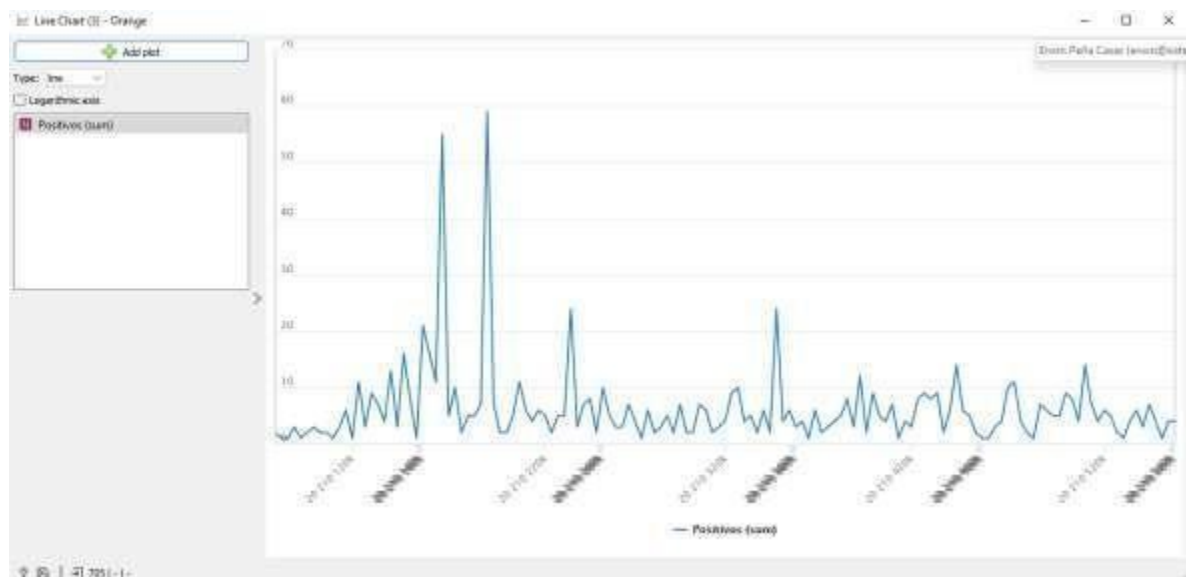


Fig. 12. Comportamiento de los positivos COVID- MINSA

En la Fig. 12 se muestra el comportamiento de los casos positivos publicados por el MINSA, con picos de positivos en el mes de febrero para luego regularizarse en los siguientes meses, con algunos picos en marzo y abril.

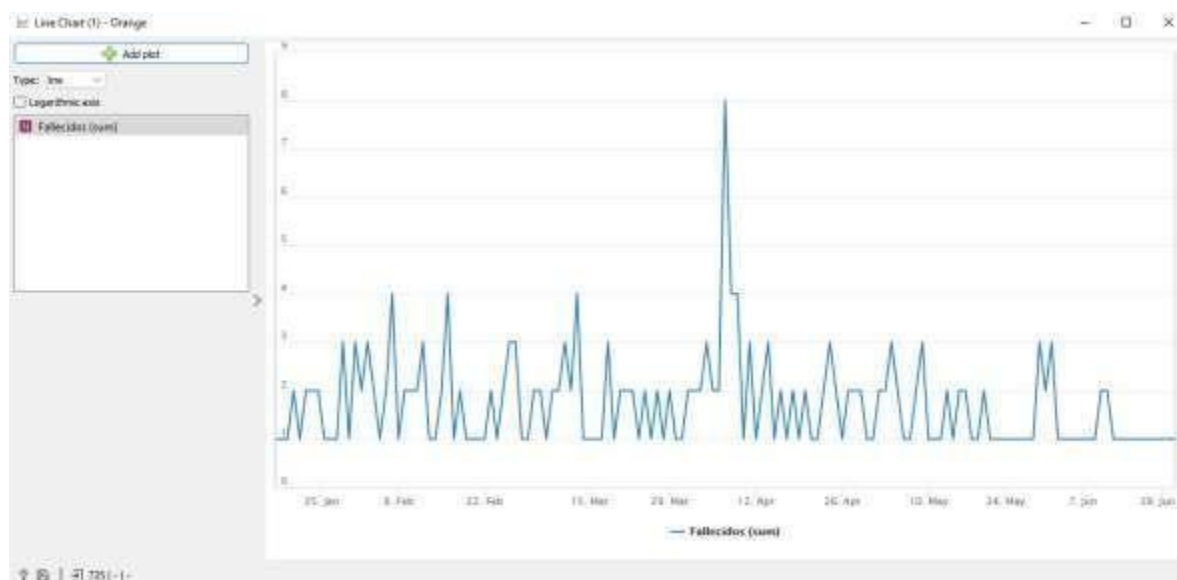


Fig. 13. Comportamiento de los fallecidos COVID-19 - MINSA

En la Fig. 13, se muestra el comportamiento de los fallecidos registrados por el MINSA, en la que se presenta comportamiento regular de enero a marzo, con un pico de fallecidos en el mes de abril, para luego regularizarse entre fines de abril hasta junio del 2021.

3.2 Estimación de las variables

En el presente apartado y con la finalidad de cumplir con el objetivo de la investigación “Estimar la incidencia de COVID-19 en el hospital María Reiche Neuman de Nazca 1er semestre 2021, mediante algoritmos de Machine Learning”, de los datos procesados con los algoritmos de Machine Learning se resumen en la Tabla XII los resultados de la estimación mediante el análisis estadístico para el intervalo de confianza al 95% cuyo valor representativo de la curva normal es 1,96 para el cálculo de los límites inferior y superior.

TABLA XII
RESUMEN ESTIMACIÓN DE LAS VARIABLES

RESUMEN ESTIMACION DE LAS VARIABLES				
ESTADÍSTICO	Prevalencia	Error Estándar	Intervalo de confianza 95%	
			Límite Inferior	Límite Superior
SEXO				
MASCULINO	0.6468	0.0096	0.6280	0.6656
FEMENINO	0.5342	0.0100	0.5146	0.5538
PRIORIDAD				
PRIORIDAD II - EMERGENCIA	0.5889	0.0099	0.5695	0.6083
PRIORIDAD III - URGENCIA MAYOR	0.7812	0.0083	0.7649	0.7975
PRIORIDAD IV - URGENCIA MENOR	0.5629	0.0100	0.5434	0.5824
PRIORIDAD V - CONSULTA	0.5030	0.0100	0.4833	0.5227
RESULTADO DE ATENCIÓN				
ALTA MEDICA	0.5828	0.0099	0.5634	0.6022
PACIENTE CONTINUA EN SERVICIO	0.6223	0.0097	0.6032	0.6414
REFERENCIA	0.5882	0.0099	0.5688	0.6076
RETIRO VOLUNTARIO	0.8000	0.0080	0.7843	0.8157
TRANSFERENCIA HOSPITALIZACIÓN	0.6979	0.0092	0.6798	0.7160
SOSPECHA DE COVID-19				
SI	0.5532	0.0100	0.5336	0.5728
NO	0.6121	0.0098	0.5929	0.6313

Para la Tabla XII

$$\text{Límite inferior} = \text{Prevalencia} - (1,96 * \text{Error Estándar}) \quad (1)$$

$$\text{Límite superior} = \text{Prevalencia} + (1,96 * \text{Error Estándar}) \quad (2)$$

TABLA XIII
CALCULO DEL ERROR ESTÁNDAR

ESTADÍSTICO	Muestra	Prevalencia	1-Prevalencia	Error Estándar
	n	p	q	$\sqrt{\frac{p \cdot q}{n}}$
SEXO				
MASCULINO	2483	0.6468	0.3532	0.0096
FEMENINO	2483	0.5342	0.4658	0.0100
PRIORIDAD				
PRIORIDAD II - EMERGENCIA	2483	0.5889	0.4111	0.0099
PRIORIDAD III - URGENCIA MAYOR	2483	0.7812	0.2188	0.0083
PRIORIDAD IV - URGENCIA MENOR	2483	0.5629	0.4371	0.0100
PRIORIDAD V - CONSULTA	2483	0.5030	0.4970	0.0100
RESULTADO DE ATENCIÓN				
ALTA MEDICA	2483	0.5828	0.4172	0.0099
PACIENTE CONTINUA EN SERVICIO	2483	0.6223	0.3777	0.0097
REFERENCIA	2483	0.5882	0.4118	0.0099
RETIRO VOLUNTARIO	2483	0.8000	0.2000	0.0080
TRANSFERENCIA HOSPITALIZACIÓN	2483	0.6979	0.3021	0.0092
SOSPECHA DE COVID-19				
SI	2483	0.5532	0.4468	0.0100
NO	2483	0.6121	0.3879	0.0098

La Tabla XIII, se utiliza para poder calcular el error estándar, base a la prevalencia como se detalla en los títulos de la Tabla.

En el caso de la variable numérica de la edad, se construye el siguiente cuadro, para ello previo se debe calcular el estadístico de la media, la cual la obtenemos con el widget de Box plot como se muestra en la Fig. 14.

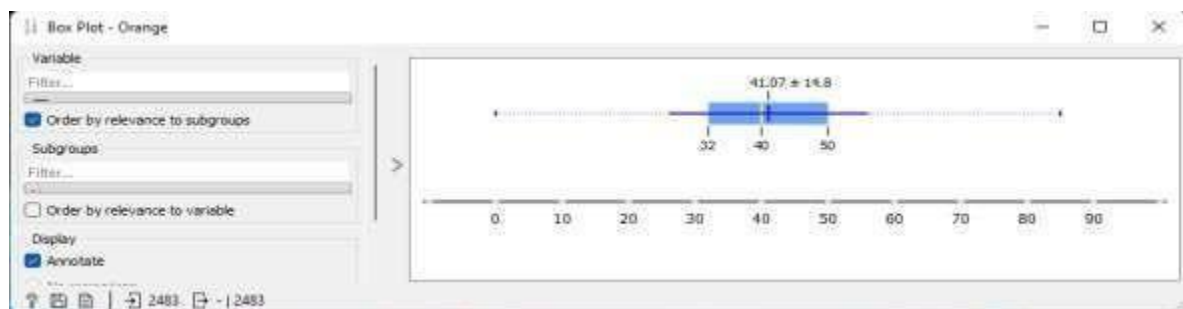


Fig. 14. Obtención de la media de la variable edad

TABLA XIV
ESTIMACIÓN DEL ESTADÍSTICO EDAD

	Media	Error estándar	Intervalo de confianza 95%	
			Límite Inferior	Límite superior
Edad	41.07	2.6989	35.78	46.36

3.3 Análisis de los componentes principales - PCA

La reducción de la dimensionalidad de las variables es un punto de suma importancia para los modelos de Machine Learning, ya que se pueden estar utilizando variables que no aportan a los modelos de predicción. Para realizar esta reducción se aplica el algoritmo de PCA como se describe a continuación.

Como primer paso se debe seleccionar las variables que aportan al modelo (features) y además de debe seleccionar la variable objetivo (target).

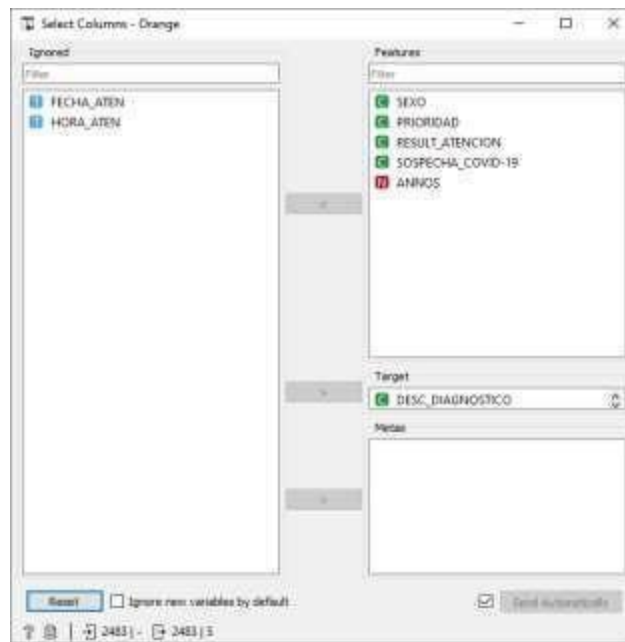


Fig. 17. Selección de variables o características y variable objetivo (target)

En la Fig. 17, se han seleccionado 05 características o variables y 01 variable objetivo (target), no se considera la Fecha ni la hora debido a que estas variables no aportan para el modelo, pero si para series de tiempo.

Todas las variables de tipo categórica deben se convertirse en variables continuas para lo cual debe aplicarse el método de one hot encoder, que se aplica en el análisis PCA. El resultado se muestra en la Fig. 18.

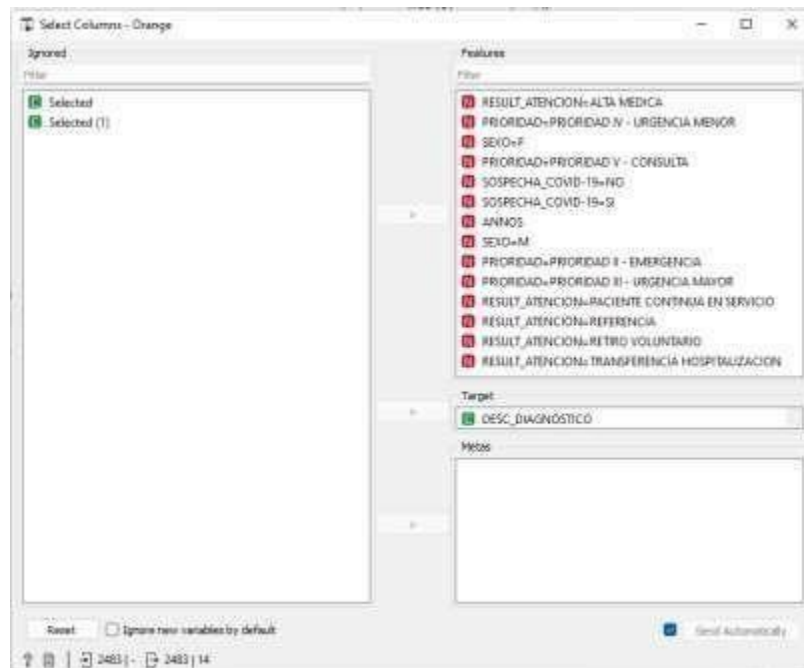


Fig. 18. Conversión de las variables categóricas en variables continuas

Según la Fig. 18, la conversión de las variables categóricas en variables continuas con el widget **continuize** que aplica el método del **one hot encoder** da como resultado 14 variables, además de la variable objetivo.

Aplicando el widget de PCA, el resultado muestra una relación de los componentes principales y la proporción de las variables con tres componentes, y un porcentaje resultante de solo es del 45.6%. El análisis de componentes principales requiere de al menos 75% en su porcentaje de variabilidad, el mismo que se expresa en la gráfica del “test scree” o gráfico de sedimentación.

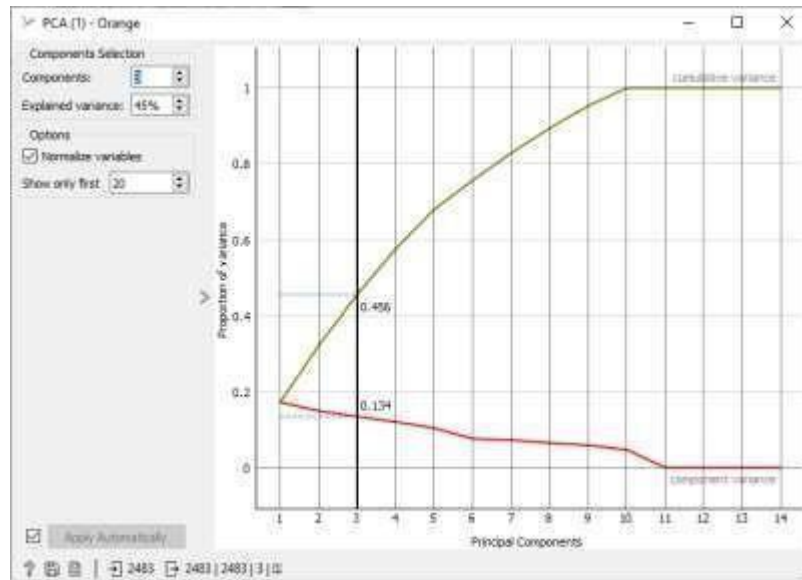


Fig. 19. Gráfico “test scree” para el análisis de PCA inicial

El gráfico de “test scree” de la Fig. 19 muestra los componentes principales, que por estándar se asignan 03 componentes principales, pero estos tres componentes solo representan el 45.6% no siendo el adecuado para este análisis.

Este porcentaje no representa los componentes ideales, ya que el porcentaje debe estar por encima de los 70%, por lo que se hace necesario ir configurando hasta que el porcentaje llegue al 75% como mínimo.

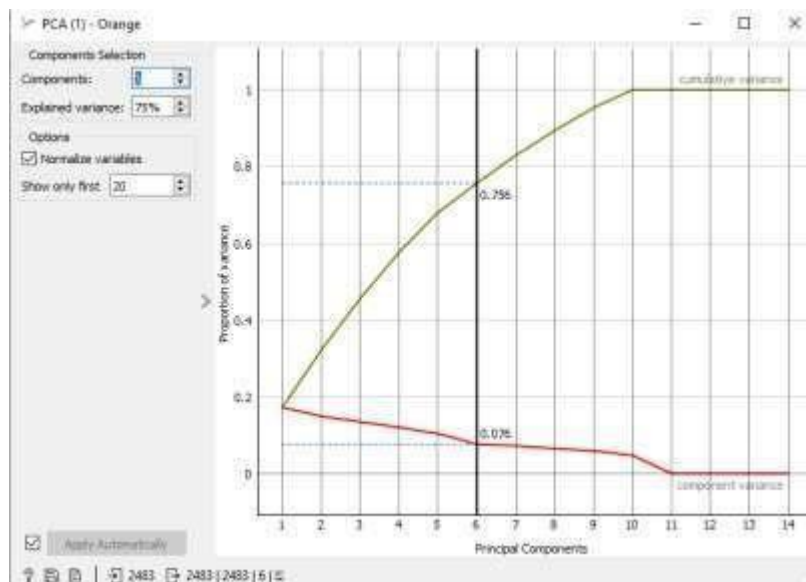


Fig. 20. Gráfico “test scree” para el análisis de PCA ajustado

En la Fig. 20, se puede apreciar que la gráfica se ha ajustado la relación de los componentes a 6, con lo que se consigue un 75.6% mayor al 75% necesario.

En el análisis PCA, los componentes se ordenan en forma decreciente de importancia, por lo que el primer componente recoge la mayor variabilidad de los datos. Para un mejor entendimiento de los componentes y sus variables se realiza la transposición de los resultados

Feature name	PC1 No 0.172181	PC2 No 0.149251	PC3 No 0.134415	PC4 No 0.120299	PC5 No 0.104153	PC6 No 0.0761743
1 SEXO=F	0.226809	0.406532	-0.529477	0.020551	-0.0111487	0.0265758
2 SEXO=M	-0.226809	-0.406532	0.529477	-0.020551	0.0111487	-0.0265758
3 PRIORIDAD=PR...	-0.0668236	0.067545	-0.00751956	0.311666	0.527152	-0.221879
4 PRIORIDAD=PR...	-0.447121	0.00756488	-0.123945	0.231025	-0.271621	0.192581
5 PRIORIDAD=PR...	0.275566	0.00850528	0.096182	-0.629958	0.281119	0.0661357
6 PRIORIDAD=PR...	0.208598	-0.0593034	0.0229002	0.498882	-0.35243	-0.220867
7 RESULT_ATENCL...	0.544762	-0.158292	0.14203	0.139819	-0.148264	0.0532542
8 RESULT_ATENCL...	-0.51733	0.139042	-0.154523	-0.281485	-0.0739716	-0.19249
9 RESULT_ATENCL...	-0.00404827	0.0438223	0.0201636	0.183872	0.355713	-0.541555
10 RESULT_ATENCL...	-0.0212126	-1.79141e-06	-0.0236732	0.0620707	0.00283315	0.510077
11 RESULT_ATENCL...	-0.0326644	0.0193994	0.0367879	0.276532	0.406288	0.466537
12 SOSPECHA_CO...	0.00336073	-0.553528	-0.428767	-0.0300883	0.0880498	-0.0181604
13 SOSPECHA_CO...	-0.00336073	0.553528	0.428767	0.0300883	-0.0880498	0.0181604
14 ANOS	-0.0856239	0.039265	0.00784619	0.186206	0.332085	0.217452

Fig. 21 variables de los componentes principales (1)

Las Fig. 21 muestran todas las variables del modelo (14 variables), para cada uno de los componentes, sin embargo, se debe seleccionar 6 variables que serían las mejores predictoras para cualquier modelo de Machine Learning, y se toma del primer componente (PC1), ya que este recoge la mayor variabilidad de los datos.

Para seleccionar dichas variables es necesario hacer un ordenamiento de las variables del componente PC1 como se muestra en la Fig. x

Feature name	PC1 No 0.172181
7 RESULT_ATENCION=ALTA MEDICA	0.544762
5 PRIORIDAD=PRIORIDAD IV - URGENCIA MENOR	0.275566
1 SEXO=F	0.226809
6 PRIORIDAD=PRIORIDAD V - CONSULTA	0.208598
12 SOSPECHA_COVID-19=NO	0.00336073
13 SOSPECHA_COVID-19=SI	-0.00336073
9 RESULT_ATENCION=REFERENCIA	-0.00404827
10 RESULT_ATENCION=RETRO VOLUNTARIO	-0.0212126
11 RESULT_ATENCION=TRANSFERENCIA HOSPITALIZACION	-0.0326644
3 PRIORIDAD=PRIORIDAD II - EMERGENCIA	-0.0668236
14 ANOS	-0.0856239
2 SEXO=M	-0.226809
4 PRIORIDAD=PRIORIDAD III - URGENCIA MAYOR	-0.447121

Fig. 22. variables ordenadas del Componente PC1

Según la Fig. 22, las variables que tienen un mayor nivel de predicción para los modelos de Machine Learning, son las 6 primeras variables: RESULT_ATENCION=ALTA MEDICA, PRIORIDAD=PRIORIDAD IV - URGENCIA MENOR, SEXO =F, PRIORIDAD=PRIORIDAD V – CONSULTA, SOSPECHA DE COVID-19=NO, SOSPECHA DE COVID-19=SI.

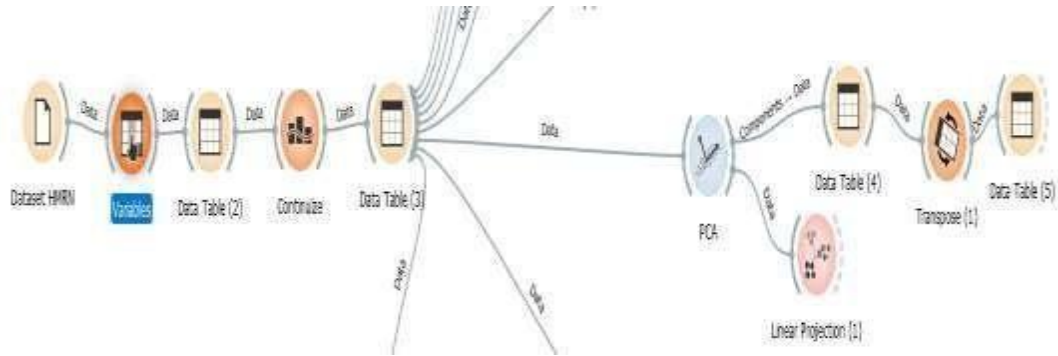


Fig. 23. Modelo de análisis PCA

La Fig. 23, muestra el modelo utilizado para el análisis PCA. El widget Contingency transforma las variables categóricas en variable numéricas para poder realizar el análisis PCA. El widget PCA crea la gráfica “test scree” para la selección de los componentes principales, los mismos que luego son exportados a una tabla para una mayor comprensión de estos y conocer cuáles serían las variables a utilizar en los modelos de predicción.

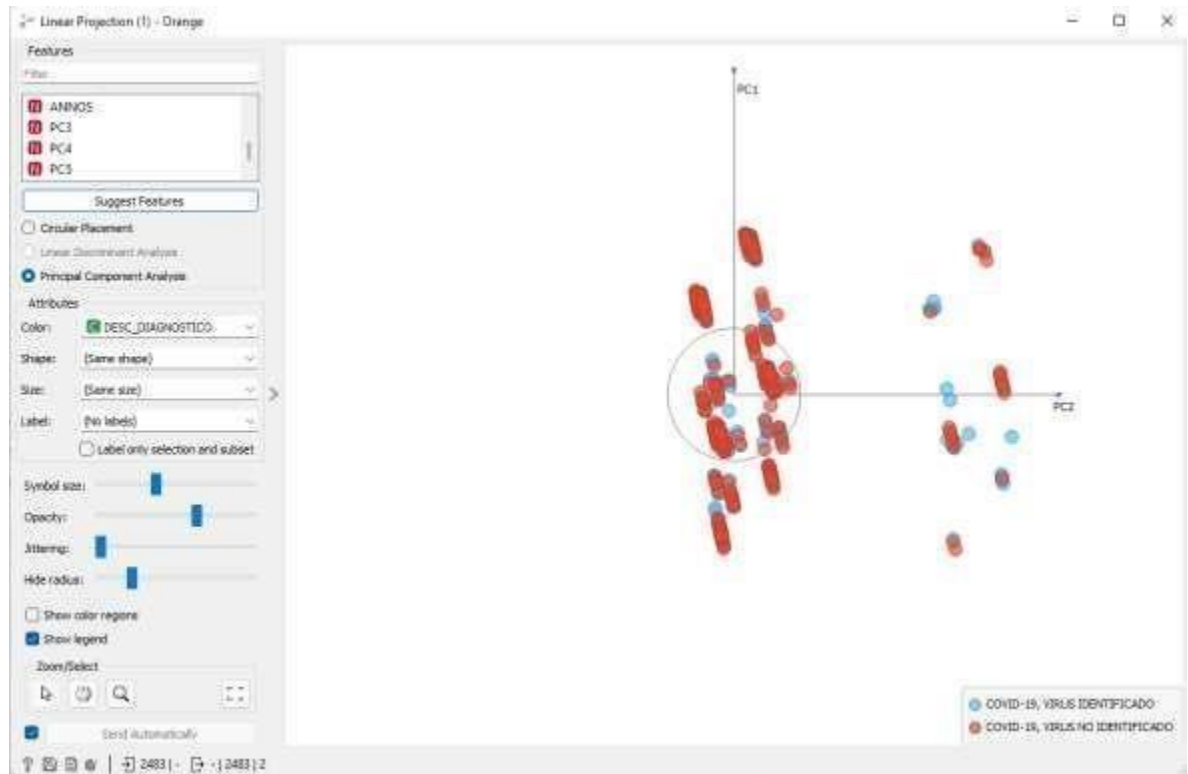


Fig. 24. Proyección lineal de los componentes PC1 y PC2 para el diagnóstico

3.4 Clúster jerárquico de las variables

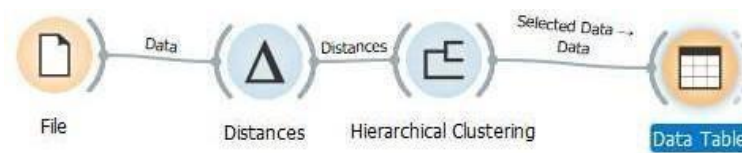


Fig. 25 Modelo de Clustering jerárquico (Hierarchical Clustering)

En el modelo de la Fig. 25, se puede apreciar los widgets para la evaluación de las variables previamente convertidas de categóricas a numéricas, previo al widget debe calcularse las distancias de los Clúster con el widget Distances.

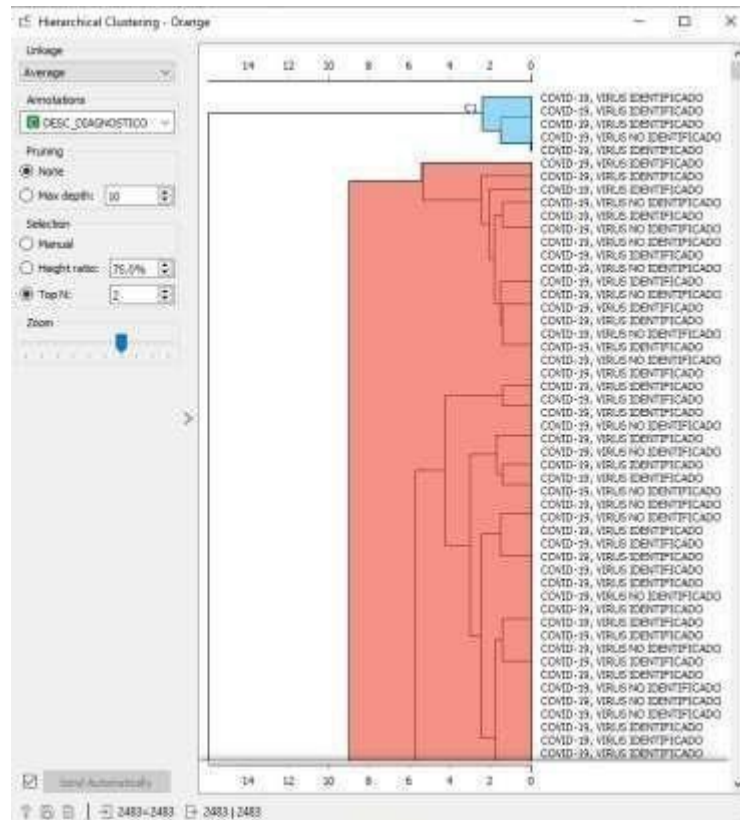


Fig. 26. Clúster jerárquico

En la Fig. 26, el modelo de clúster jerárquico, se configuro para dos categorías que se corresponden con el diagnóstico, el modelo de clasificación tiene algunas imprecisiones como se puede apreciar en la gráfica.

Para la clasificación de cuenta con una serie de algoritmos como son el K-meas, KNN, CN2 y Clúster jerárquico con K-means que es el aplicado en la presente investigación.

ID	DESC_DIAGNOSTICO	Cluster	ANOS	SEXO=M	DORIDAD III - UF	DORIDAD IV - UF	I=PRIORIDAD Y	PACIENTE CX
1500	COVID-19, VIRUS IDENTIFICADO	C2	37	1	0	0	0	1
1501	COVID-19, VIRUS IDENTIFICADO	C3	68	1	0	0	0	1
1502	COVID-19, VIRUS IDENTIFICADO	C2	52	0	0	0	0	1
1503	COVID-19, VIRUS IDENTIFICADO	C2	45	1	0	0	0	1
1504	COVID-19, VIRUS IDENTIFICADO	C2	32	1	0	0	0	1
1505	COVID-19, VIRUS IDENTIFICADO	C2	42	1	0	0	0	1
1506	COVID-19, VIRUS IDENTIFICADO	C2	39	0	0	0	0	1
1507	COVID-19, VIRUS IDENTIFICADO	C2	63	1	0	0	0	1
1508	COVID-19, VIRUS IDENTIFICADO	C2	34	1	0	0	0	1
1509	COVID-19, VIRUS IDENTIFICADO	C2	40	1	0	0	0	1
1510	COVID-19, VIRUS IDENTIFICADO	C3	53	1	0	0	0	1
1511	COVID-19, VIRUS IDENTIFICADO	C2	44	0	0	0	0	1
1512	COVID-19, VIRUS IDENTIFICADO	C2	25	1	0	0	0	1
1513	COVID-19, VIRUS IDENTIFICADO	C2	41	0	0	0	0	1
1514	COVID-19, VIRUS IDENTIFICADO	C2	41	1	0	0	0	1
1515	COVID-19, VIRUS IDENTIFICADO	C2	40	1	0	0	0	1
1516	COVID-19, VIRUS IDENTIFICADO	C2	20	1	0	0	0	1
1517	COVID-19, VIRUS IDENTIFICADO	C2	31	1	0	0	0	0
1518	COVID-19, VIRUS NO IDENTIFICADO	C2	33	0	1	0	0	0
1519	COVID-19, VIRUS NO IDENTIFICADO	C2	33	0	0	0	0	1
1520	COVID-19, VIRUS NO IDENTIFICADO	C2	66	1	1	0	0	0
1521	COVID-19, VIRUS NO IDENTIFICADO	C3	44	0	0	0	0	1
1522	COVID-19, VIRUS NO IDENTIFICADO	C2	66	1	0	1	0	0
1523	COVID-19, VIRUS NO IDENTIFICADO	C2	33	1	0	1	0	0
1524	COVID-19, VIRUS NO IDENTIFICADO	C2	67	1	0	0	0	0
1525	COVID-19, VIRUS NO IDENTIFICADO	C2	57	1	0	0	0	1
1526	COVID-19, VIRUS NO IDENTIFICADO	C3	81	1	1	0	0	0
1527	COVID-19, VIRUS NO IDENTIFICADO	C2	44	0	1	0	0	0

Fig. 27. Datos clasificados y agrupados por Clúster

En la Fig. 27, para la clasificación o agrupamiento donde se puede apreciar el clúster asignado a los datos (C2), donde se puede comprobar que existen algunas clasificaciones o agrupaciones no correctas para los clústeres.

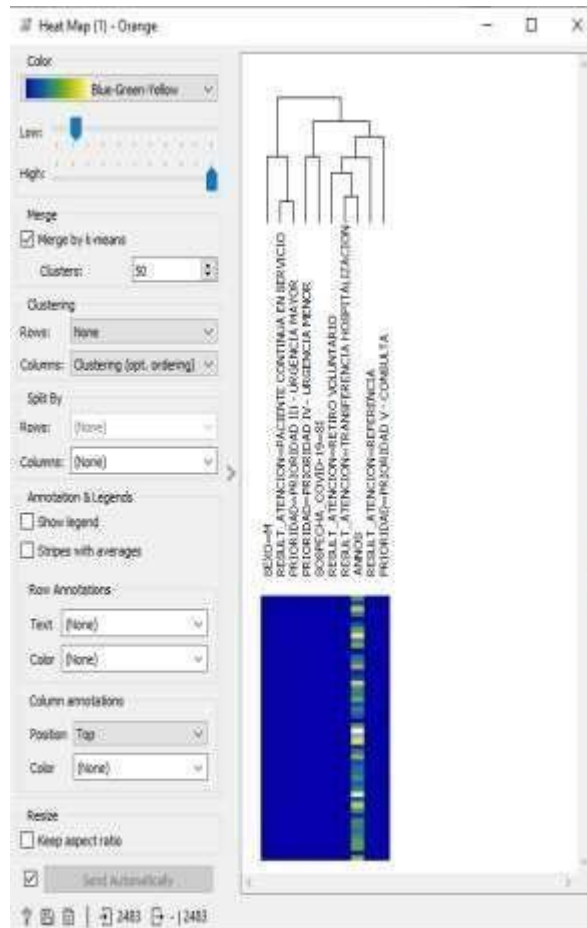


Fig. 28. Clustering de los datos mediante un mapa de calor

En la Fig. 28, se muestra por medio del mapa de calor de las variables, donde igualmente se configura para clustering, se nota el agrupamiento de las variables que tienen un nivel de predicción. Variables como el sexo, resultado (paciente continua en servicio) y prioridad (urgencia mayor) para un primer clúster y el segundo clúster las variables prioridad (urgencia menor), sospecha de covid-19 (si), hasta la prioridad (consulta).

Los algoritmos de Machine Learning empleados para la selección y agrupamiento de las variables en estudio, son de suma importancia para posteriormente aplicar los algoritmos de predicción y determinar el nivel de predicción de los diversos algoritmos.

3.5 Evaluación de las variables obtenidas

Para comprobar la efectividad de estas técnicas que permiten seleccionar las variables que más aportan al modelo de Machine Learning, se van a realizar la predicción con diversos algoritmos con todas las variables (14) y otra predicción con los mismos algoritmos para las variables seleccionadas (6).

Los algoritmos utilizados para evaluar las variables son: regresión logística, redes neuronales, árboles de decisión, árboles y SVM.

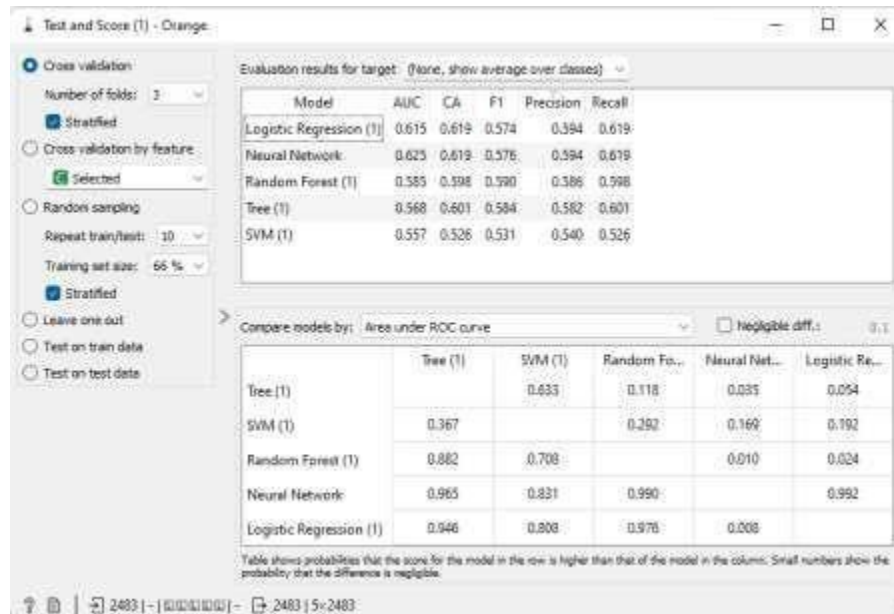


Fig. 29. Test y score de los algoritmos de predicción para las 14 variables

En la Fig. 29 la precisión de los modelos en la que la regresión logística y las redes neuronales tiene una mayor precisión con 59.40% con todas las características o variables del Dataset.

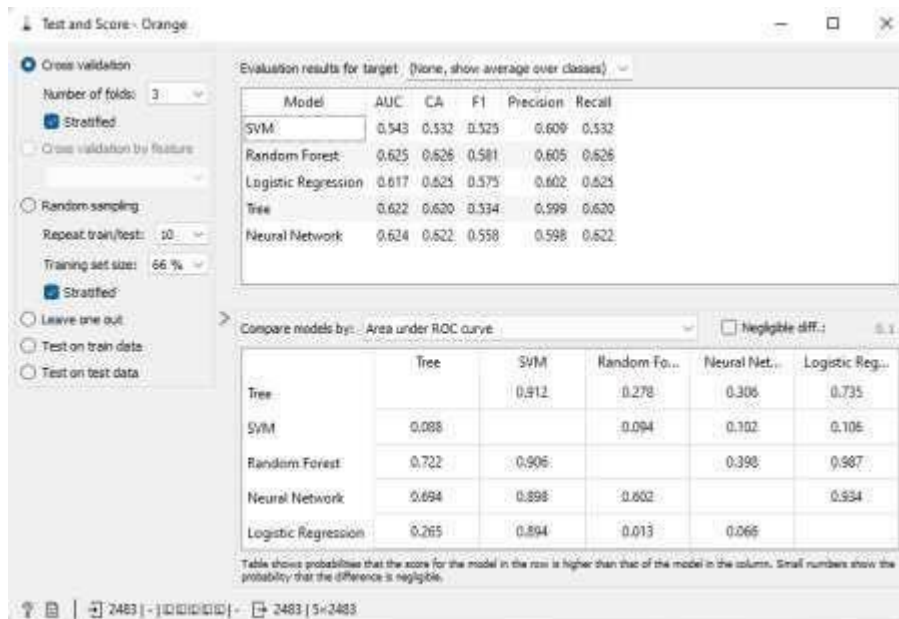


Fig. 30. Test y score de los algoritmos de predicción con 6 variables seleccionadas

En la Fig. 30, la precisión de los modelos de todos los modelos utilizados con las 6 variables seleccionadas por el PCA, tienen una mejor precisión entre 59.8% - 60.9%, que la obtenida con las 14 variables cuya mayor precisión fue 59.4%.

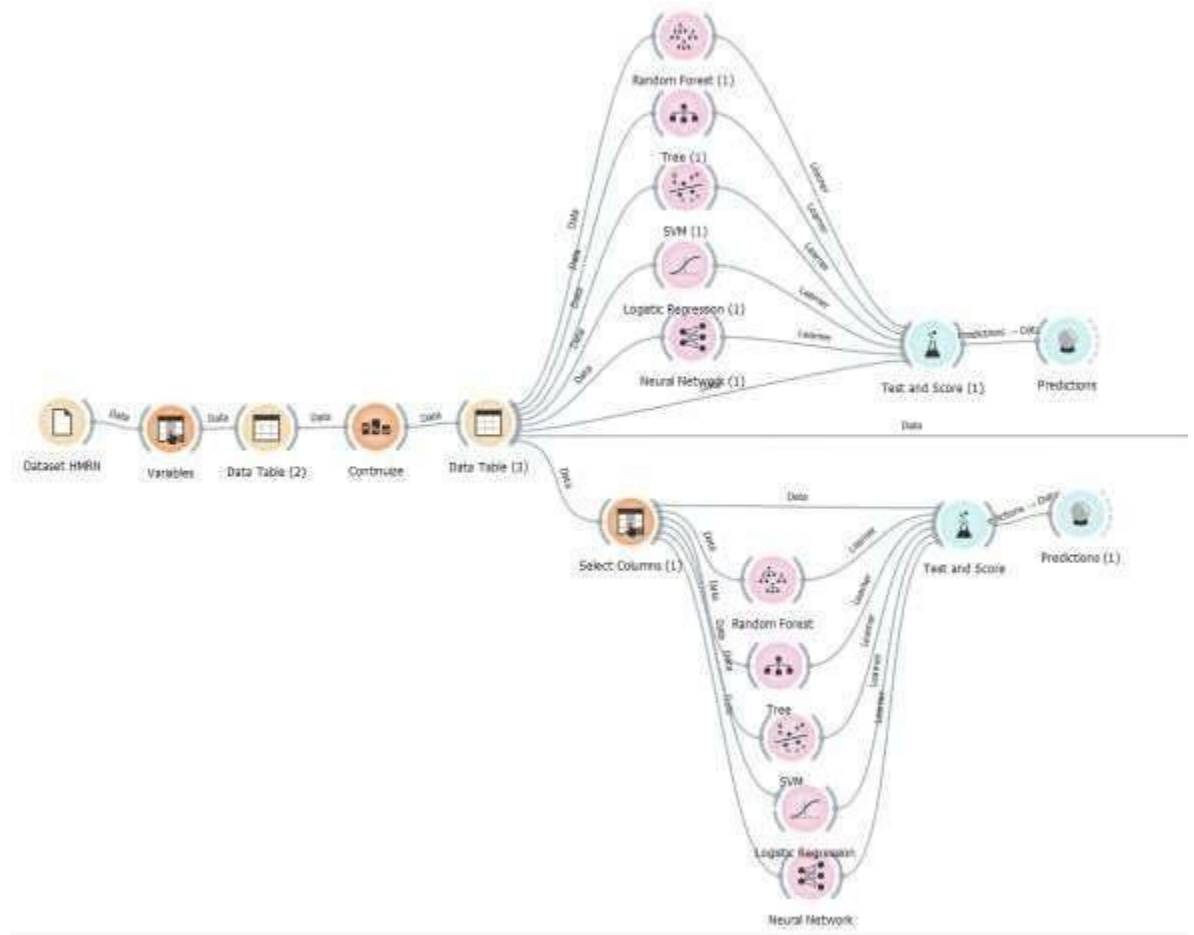


Fig. 31. Modelo de evaluación de las variables

En la Fig. 31 se puede apreciar la evaluación (Test and score) para los dos casos Data Table (3) con todas las variables y Select Column (1) para las variables con PCA y Clúster Jerárquico.

IV. DISCUSIÓN

Con la aplicación de modelos de Machine Learning para el manejo de datos relacionados con la COVID-19, El Hospital María Reiche Neuman y el MINSA pueden conocer las características de la pandemia y tomar mejores decisiones, como el caso colombiano cuyo estudio se podrían conocer el impacto epidémico real de la enfermedad y efectuar diagnósticos más rápidos [3]. Clasificar o agrupar los datos con algoritmos de inteligencia artificial como el Clúster Jerárquico permite conocer las relaciones entre las variables o características. Esta clasificación o agrupación será la base para generar comparaciones del comportamiento de la enfermedad [2]. De otro lado reducir la dimensionalidad de los datos con algoritmos de PCA ayuda a mejorar las predicciones, reduce los tiempos de procesamiento y reduce los costos de cómputo. Valorando la situación actual se pueden realizar estudios predictivos en bien de la sociedad [6], por lo cual el conocimiento obtenido es importante para la población, de tal forma que es momento de valorizar ese conocimiento en benéfico de la población [4]. Al respecto con el uso de algoritmos de inteligencia artificial, podemos estar mejor preparados para escenarios como el ocasionado por la COVID-19. Incluso se pueden minimizar los riesgos al formular tratamientos [5].

V. CONCLUSIONES

1. El impacto ocasionado por la COVID-19 ha ocasionado cambios en todas las actividades humanas, de tal manera que existe un esfuerzo por parte de los investigadores incluida la academia en la exploración del impacto de esta con los datos disponibles y aprovechables.
2. El preprocesamiento de los datos brutos es de suma importancia para que el modelo pueda tener los resultados esperados, a pesar de tener un importante consumo de tiempo 80-20, 80% del tiempo para el preprocesamiento (preparación de los datos) y un 20% del tiempo para aplicación de los modelos.
3. El análisis descriptivo permitió conocer el comportamiento de los datos, en forma de tablas o gráficas para una mejor comprensión de estos, en la que se encontraron algunos patrones de interés.
4. En la investigación se desarrolla modelos de Machine Learning con programación visual Python y la herramienta Orange para minería de datos, cuyos resultados para el algoritmo de clustering, específicamente Clúster Jerárquico y análisis PCA son consistentes y razonables para la data disponible.
5. El Clúster Jerárquico nos permite una clasificación de los datos agrupándolos y creando grupos de interés denominados Clúster. El análisis de componentes principales PCA permitió reducir la dimensionalidad de las variables y obtener las mejores variables que pueden ser utilizadas en los modelos de predicción, ayudando a tener un procesamiento más ligero, con menos variables en predicción, pero resultados más precisos.

VI. RECOMENDACIONES

La posibilidad de contar con los datos es la principal fuente para poder realizar una analítica de esos datos con predicciones con inteligencia artificial, como es el caso del Hospital María Reiche Neuman, sin embargo, los datos extraídos a pesar de tener una gran dimensionalidad con 33 características o variables, la mayoría de estas no son relevantes. Solo fueron seleccionados 6 características para el tratamiento y 02 de fecha y hora solo para el análisis descriptivo y ver el comportamiento de los datos. Por otro lado, los datos de acceso libre del MINSA son limitados en cuanto a características o variables.

Sugerir al MINSA, fomentar el empleo de la Inteligencia artificial para en el manejo de los datos médicos de todos los casos para mejorar la comprensión de las enfermedades, no solo para el COVID-19.

En base a lo anterior, con la finalidad de que se realicen investigaciones más precisas que apoyen a la toma de decisiones de los directivos del hospital, y del MINSA, estos deben poner a disposición de la comunidad científica los datos completos de las atenciones (eg. no se tiene acceso a datos como los tomados en triaje y otros del diagnóstico médico) que aportarían mucho a los modelos de Inteligencia artificial de acceso libre.

Queda en el tintero la realización de otras investigaciones de la aplicación del clustering, en diversas enfermedades que son de mucha frecuencia y que se presentan con frecuencia en el hospital.

VII. REFERENCIAS BIBLIOGRÁFICAS

- [1] G. Ortigoza, ; Lorandi, and Alberto, “*Simulación Numérica y Modelación Matemática de la propagación del Covid 19 en el estado de Veracruz Artículo Original Numerical Simulation and Mathematical Modeling of the spread of Covid 19 in the state of Veracruz*” *Rev Mex Med Forense*, vol. 5(3), no. 3, pp. 21–37, 2020.
- [2] Y. J. Alonso C., Julio César Ariza and C. C. Hoyos, “*Metodología para la agrupación de departamentos de Colombia para la comparación de indicadores de COVID-19 . Metodología para la agrupación de departamentos de Colombia para la comparación de indicadores de COVID-19*” *Cent. Investig. ICESI*, pp. 1–4, 2021, [Online]. Available: https://www.icesi.edu.co/covid-al-dia/images/excesos-muertes/Metodologia_clustering.pdf.
- [3] A. K. Bermúdez Rojas, Á. J. Romero Beltrán, J. D. Guacho Bonilla, and J. G. Méndez Reyes, “*Uso de tecnologías en el lugar de atención para el manejo de la pandemia por COVID-19*” *Reciamuc*, vol. 5, no. 4, pp. 206–216, 2021, doi: 10.26820/reciamuc/5.(4).noviembre.2021.206-216.
- [4] W. Cabanillas Rojas, “*Conducta y propagación del covid-19 en el Perú: Marco de referencia para el diseño de intervenciones conductuales de salud pública*” *Scielo*, no. June, p. 1, 2020, [Online]. Available: <https://preprints.scielo.org/index.php/scielo/preprint/view/868>.
- [5] J. Márquez Díaz, “*Artificial Intelligence and Big Data as solutions to COVID-19 Intel · ligència Artificial i Big Data per fer front a la COVID-19*” *Rev. Bioética y Derecho Perspect. Bioéticas*, no. c, pp. 315–331, 2020, [Online]. Available: https://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1886-58872020000300019.
- [6] V. Caínzos López, “*ANÁLISIS PREDICTIVO DE DATOS DE LA PANDEMIA COVID-19 UTILIZANDO APRENDIZAJE AUTOMÁTICO MASTER*” *Universidade Da Coruña*, 2021.
- [7] M. Y. Leyva Vasquez and G. B. Montero Castillo, “*COVID-19 time series analysis model based on hierarchical clustering*” *Universidad Politécnica Salesiana-Ecuador*, 2021.
- [8] Y. Martínez González and D. A. García Cortés, “*Modelo inspirado en un proceso hidrológico para el pronóstico de escenarios de contagio de la COVID-19*” *An. la Acad.Ciencias Cuba*, vol. 11, no. 1, p. 869, 2021, [Online]. Available: <http://revistaccuba.sld.cu/index.php/revacc/article/download/869/969>.

- [9] C. Quispe, P. Vela, M. Meza, and V. Moquillaza, “COVID-19: Una pandemia en la era de la salud digital” *Unidad Informática Biomédica en Salud Glob.*, pp. 1–19, 2020.
- [10] L.-M. A. Ortigoza-Capetillo GM, “Mathematical Projections for COVID-19 Cases in México” *Rev Mex Med Forense*, vol. 6(2), no. April, pp. 115–129, 2021.
- [11] Simplilearn, “Data Science In 5 Minutes” *Simplilearn*, 2018.
<https://www.youtube.com/watch?v=X3paOmcrTjQ> (accessed May 10, 2022).

VIII. ANEXOS

Anexo 01: Clasificación de clustering

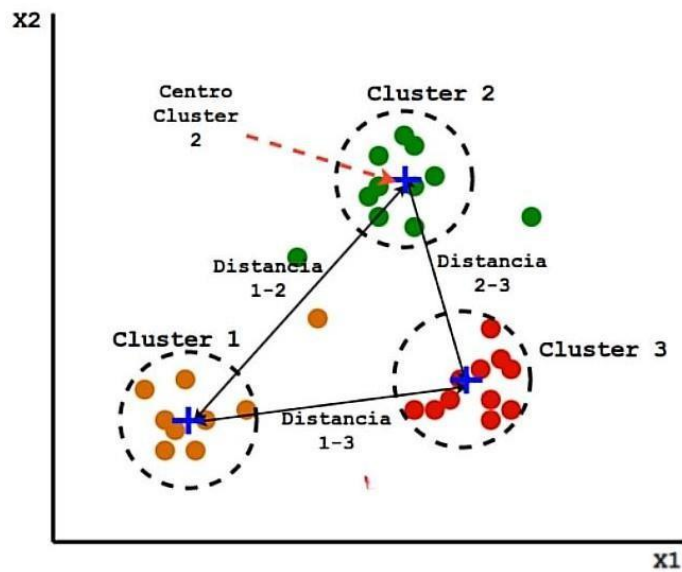


Fig. 32. Modelo de aplicación de Clustering

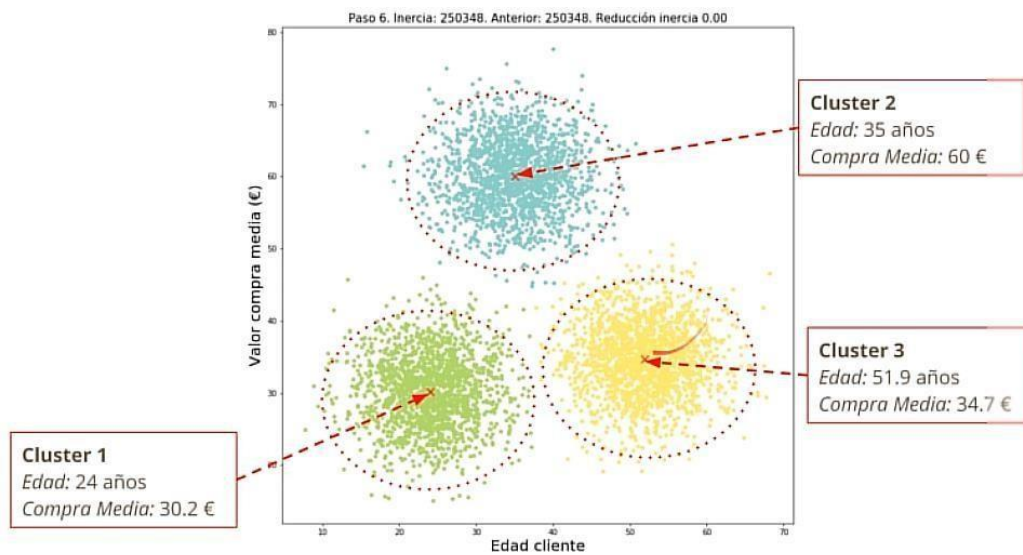


Fig. 33. Clustering K-medias de segmentación de clientes.

Anexo 02: Interfaces del Sistema ESSI



Fig. 32 Menú Principal

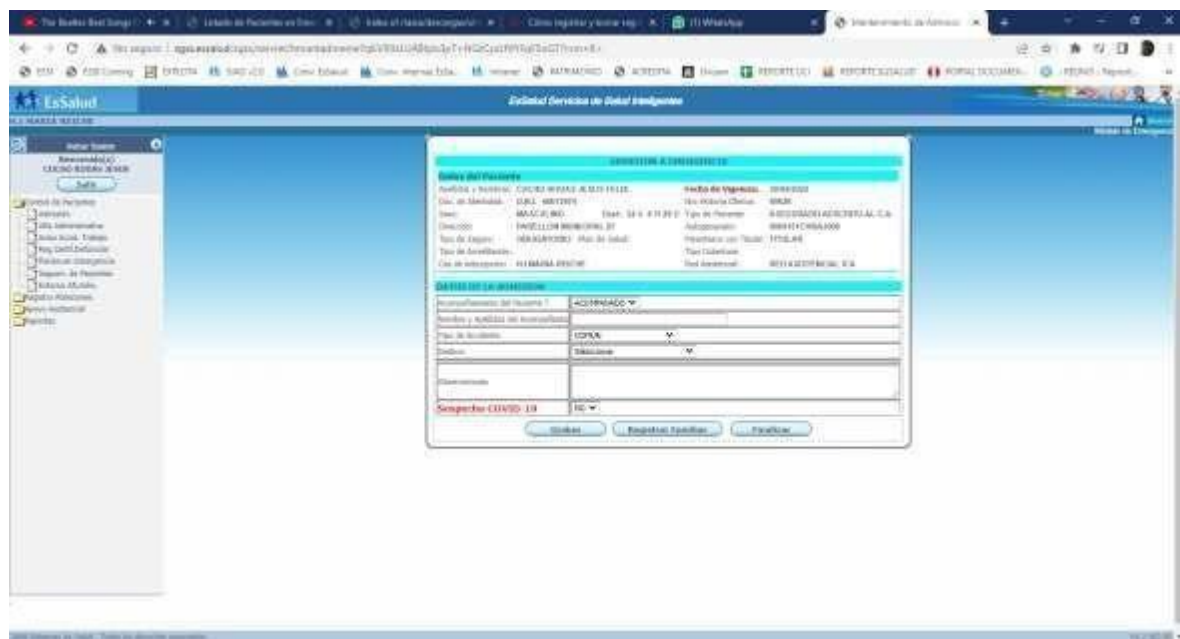


Fig. 33 Admisión de pacientes



Fig. 34 Interfaz de atención de pacientes



Fig. 35 Interfaz de pacientes atendidos según tópico

Anexo 03: Procedimiento de obtención de los datos

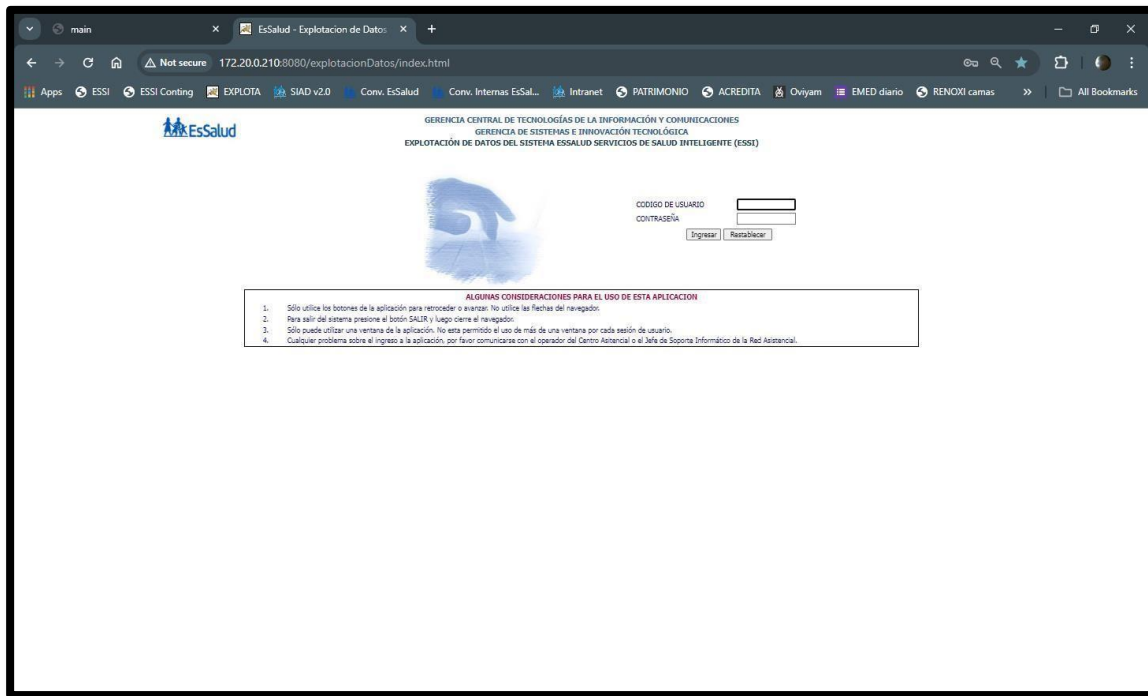


Fig. 36. Aplicativo EXPLOTA (tiene acceso a la BD del Sistema ESSI), es exclusivo para obtención de reportes.

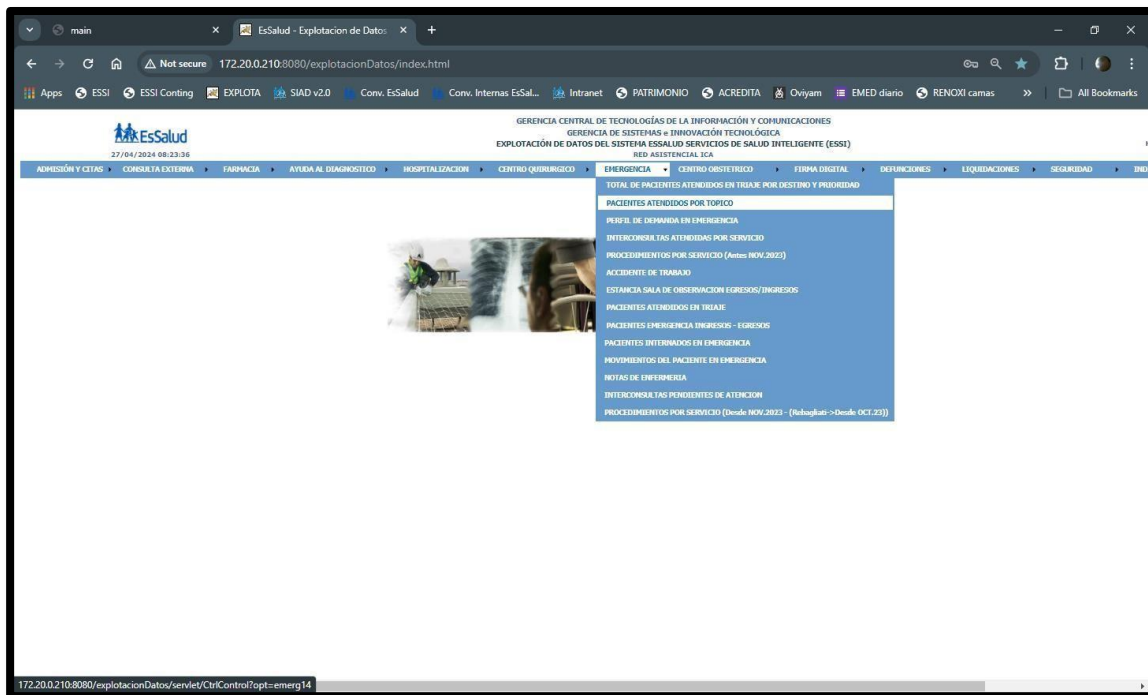


Fig. 37. Selección del Reporte de Atenciones de Pacientes de Emergencia

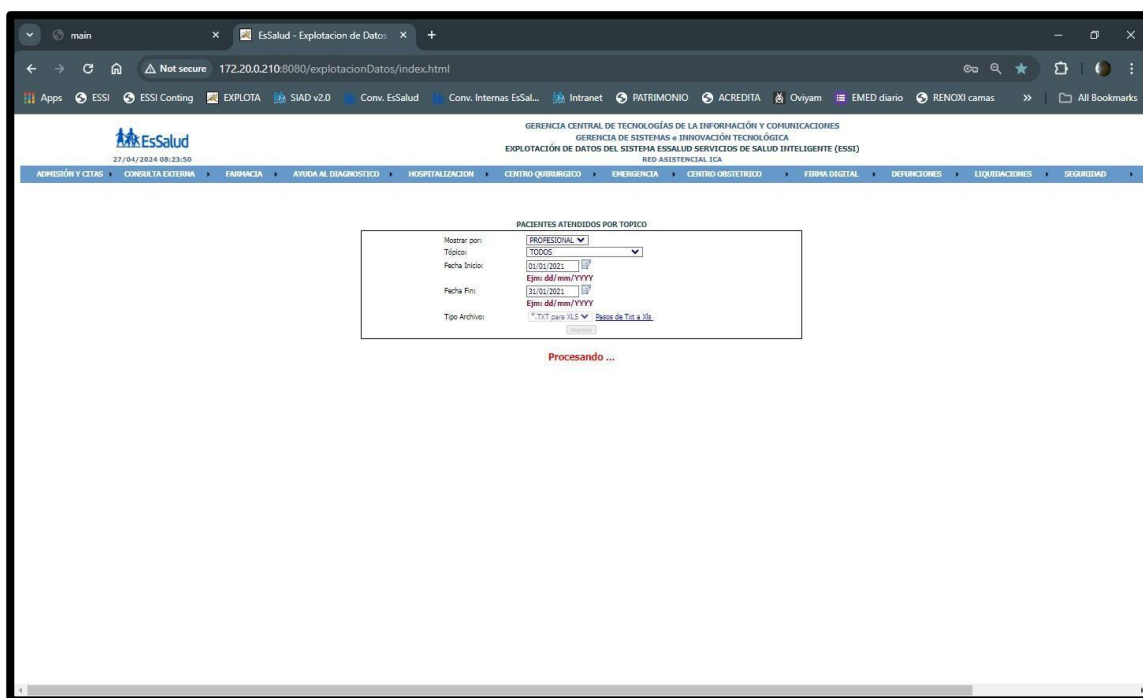


Fig. 38. Selección del periodo a consultar.

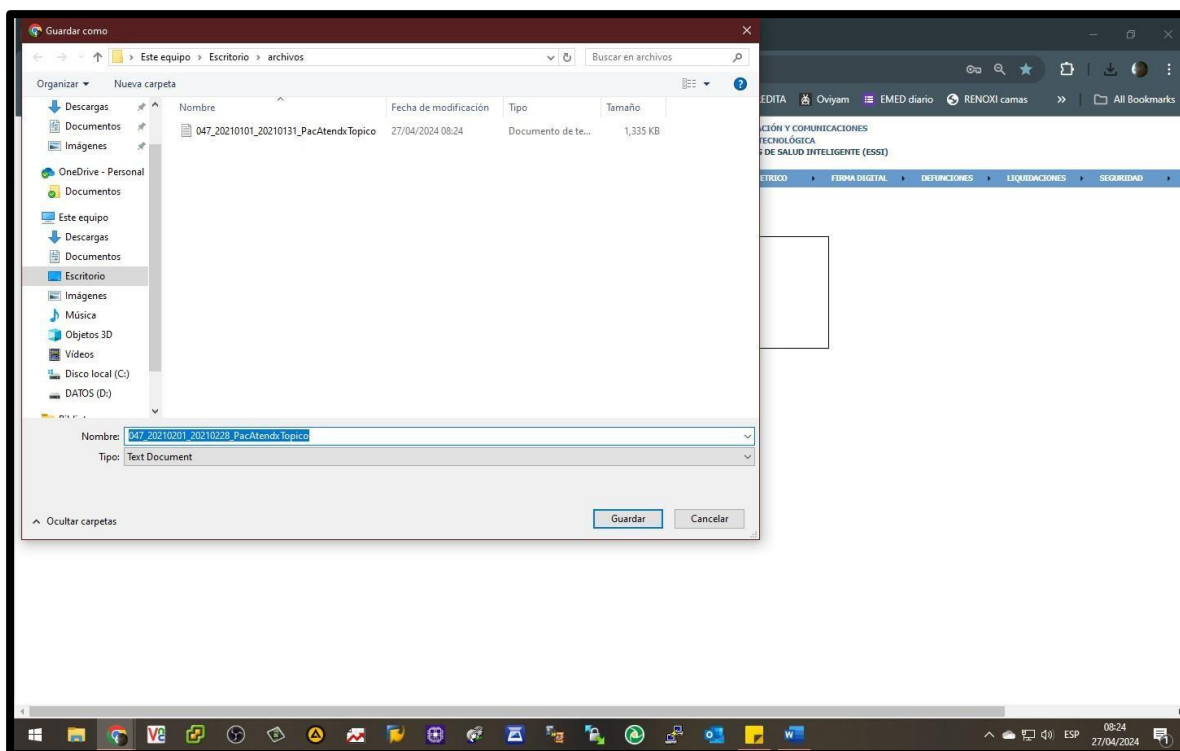


Fig. 39. Se descarga el reporte seleccionado