UNIVERSIDAD CÉSAR VALLEJO

FACULTAD DE INGENIERÍA Y ARQUITECTURA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

Modelo de clasificación basado en minería de datos para la identificación de factores que influyen en las infecciones respiratorias agudas graves de pacientes

TESIS PARA OBTENER EL TÍTULO PROFESIONAL DE GRADO ACADÉMICO DE INGENIERA DE SISTEMAS

AUTOR:

Quijano Nayhua, Rossmery Jackeline (ORCID: 0000-0001-5226-9587)

ASESOR:

Dr.Saboya Rios, Nemias (0000-0002-7166-2197)

LÍNEA DE INVESTIGACIÓN:

Sistemas de Información y Comunicaciones

LIMA – PERÚ

2021

Dedicatoria

Dedico esta tesis a mi familia que pese a la distancia, siempre estuvieron dándome su apoyo incondicional, logrando que no baje la guardia por más dura que se ponga la situación.

Agradecimiento

A mi madre Julia Nayhua y a mi padre Florentino Quijano, pese a que no seremos la familia perfecta o que tenga dinero, siempre se encargaron de sacarme adelante y alentarme a ser una buena hija y estudiante. También a mi prima Edith Huaman, por los consejos, gracias a todos por guiarme en esta etapa de mi vida.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

Resumen

La presente tesis permitió, determinar el mejor algoritmo de acuerdo a las técnicas que ofrece la minería de datos para predecir el diagnóstico, a partir de la población de pacientes en un periodo determinado y que hayan sido diagnosticados con algún tipo de infección respiratoria en el Policlinico Cruz Verde. El diseño de la investigación fue pre-experimental, la técnica de recolección de datos fue la observación de las historias clínicas y el instrumento fue la ficha de registro de los datos físicos, los cuales fueron validados por los médicos a cargo del diagnóstico del paciente. La metodología que se empleó para la construcción del modelo de clasificación, fue CRISP-DM y para el análisis de los datos, se usó la herramienta Anaconda, Editor de códigos SPYDER con lenguaje de programación Python.

Finalmente, realizando una evaluación de los algoritmos empleados, el árbol de decisión tipo CART y clustering tipo K-MEANS, permitiendo una clasificación correcta de datos con una asertividad de 0.97 y 0.6, en el que permitió llegar a la conclusión que el mejor algoritmo es el árbol de decisión, ya que permite clasificar los datos en falso o verdadero para saber que sentencias se cumplen y de cuanto es su entropía.

Palabras claves: minería de datos, algoritmos de aprendizaje, crisp-dm.

Abstract

This thesis proposes to determine the best algorithm for data mining classification techniques to predict the correct diagnosis of severe acute respiratory infection diseases, from the population of patients with a diagnosis of some type of respiratory infection at the Polyclinic Cruz Green, for which the research design was pre-experimental, the data collection technique was the observation of the medical records and the instrument was the physical data record sheet, which were validated by the physicians at charge of the patient's diagnosis;. The methodology used for the construction of the classification model was CRISPDM and for the data analysis, the Anaconda tool, SPYDER code editor with Python programming language, was used.

Finally, it is shown that the decision tree classification techniques with the classification algorithm and clusters, fulfilled the objectives of the research, allowed a correct classification of data with a precision of 0.97 and 0.6, and in a very long period of time. short, coming to the conclusion that data mining tools can help to carry out a preliminary diagnosis on the part of the personnel.

Keywords: data mining, learning algorithms, diagnosis.

I. INTRODUCCIÓN

Según la Organización Mundial de la Salud (2020) plantea reconocer a todos los pacientes con IRAG en el primer contacto con el sistema sanitario para poder gestionar-administrar los tratamientos, de la misma manera, mejorar la atención clínica de estos pacientes y facilitar las orientaciones más actualizadas con la correcta prevención de las infecciones. Si bien es cierto, existen las consultas ambulatorias, las cuales pueden ayudar cuando la enfermedad está iniciando, por lo contrario, ya se requiere intervención de profesionales, ya que en la actualidad se encuentran como las primeras causas de mortalidad. Así mismo, se debe tomar en cuenta el diagnóstico clínico, ya que, se basa en un procedimiento donde los profesionales de la salud pueden identificar la enfermedad (o no enfermedad) de un paciente en función de los síntomas con la ayuda de diversas herramientas que pueden definir su situación clínica (Sánchez, S, 2019).

Tal es el caso, las infecciones respiratorias agudas graves, en los últimos años han sido consideradas como una de las enfermedades que afectan a la sociedad, ya que, se pueden presentar en los niños a partir de los 5 años y en las personas adultas de 18 años a más, si bien es cierto, existen las consultas ambulatorias, las cuales pueden ayudar cuando la enfermedad está iniciando, por lo contrario ya se requiere intervención de profesionales, ya que en la actualidad se encuentran como las primeras causas de mortalidad. Además, la Organización Mundial de la Salud, declara la clasificación de los pacientes con IRAG asociados a COVID-19 en la primera fase con el sistema sanitario, ya que debe valorarse como probable etiología de un cuadro respiratorio agudo en diferentes situaciones (2020, p.2).

Por otro lado, a pesar de que ya existen factores afectivos a cerca de las enfermedades infecciosas respiratorias, este problema de salud muestra una variedad de virus respiratorios, los cuales son causas frecuentes de las IRAG en niños menores o adultos de tercera edad, en la revisión sistemática se observó que la etiología viral fue de 50,4% de IRAG, variando entre 48,7% en neumonías y 66,3% en bronquiolitis (Becerra, et al., 2019).

Desafortunadamente, en el Centro de Salud Javier Loyola, a partir del año 2019 se observó un gran aumento de casos con IRAG, por lo que en Ecuador se puede denominar como una de las causas frecuentes de morbilidad, tanto en personas mayores de 50 años y entre 0 a más de 5 años, tal como se muestra en la Figura 1. Una de las enfermedades más comunes es la Rinofaringitis y le sigue la Faringo-amigdalitis.

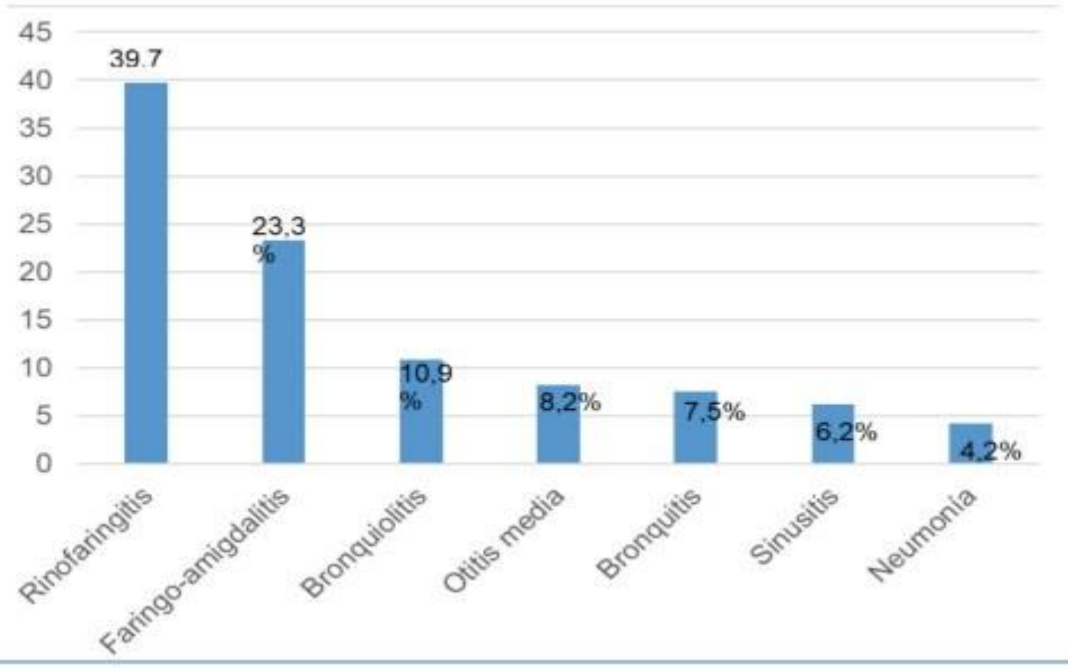


Figura 1. IRAG > 5 años del Centro de Salud de Javier Loyola.

Para el Ministerio de Salud del Perú, la semana de epidemiologia (SE), se notificaron 406 542 casos de infecciones respiratorias agudas, por lo que se puede estimar que fueron casos acumulados de 1443 episodios por cada 10 mil niños menores de 5 años, asimismo, se observó una reducción en la infección de tipo SOB/Asma del 9,1 %. El índice epidémico va sumando, por lo que la incidencia es alta, ya que si se ubicaría entre 0,76-1,24%, la incidencia estaría normal, ello se puede observar en la Tabla 1. (2019, p.4).

Tabla 1. *Indicadores de la Vigilancia IRAG, Perú 2017-2019*

Variables	2017	2018	2019
IRA < 5 años	434520	406260	406542
Incidencia Acumulada x 10 000	**1526.9**	**1435.0**	**1443.1**
SOB / ASMA < 5 años	24711	23091	20982
Incidencia Acumulada x 10 000	**86.8**	**81.6**	**74.5**
Neumonías < 5 años	4870	3411	3619
Incidencia Acumulada x 10 000	**17.1**	**12.0**	**12.8**
Hospitalizados < 5 años	1608	1196	1261
Tasa hospitalización x 100	**33.0**	**35.1**	**34.8**
Defunciones < 5 años	44	37	26
Letalidad x 100	**0.90**	**1.08**	**0.72**
Mortalidad x 100 000	**1.5**	**1.3**	**0.9**
Neumonías > 60 años	2861	3380	3813
Incidencia Acumulada x 10 000	**9.2**	**10.5**	**11.4**
Hospitalizados > 60 años	1015	1169	1391
Tasa hospitalización x 100	**35.48**	**34.59**	**36.48**
Defunciones > 60 años	242	311	286
Letalidad x 100	**8.5**	**9.2**	**7.5**
Mortalidad x 100 000	**7.8**	**9.6**	**8.5**

Fuente: Centro Nacional de Epidemiología, Prevención y Control de Enfermedades.

Además, como bien se sabe, dentro de las infecciones respiratorias agudas graves, hay diferentes tipos, dentro de está la neumonía, por ejemplo en la Figura 2. , se puede observar que la incidencia de casos es mayor en la región de la selva, por lo que se deduce que va ir variando de acuerdo a la localidad.

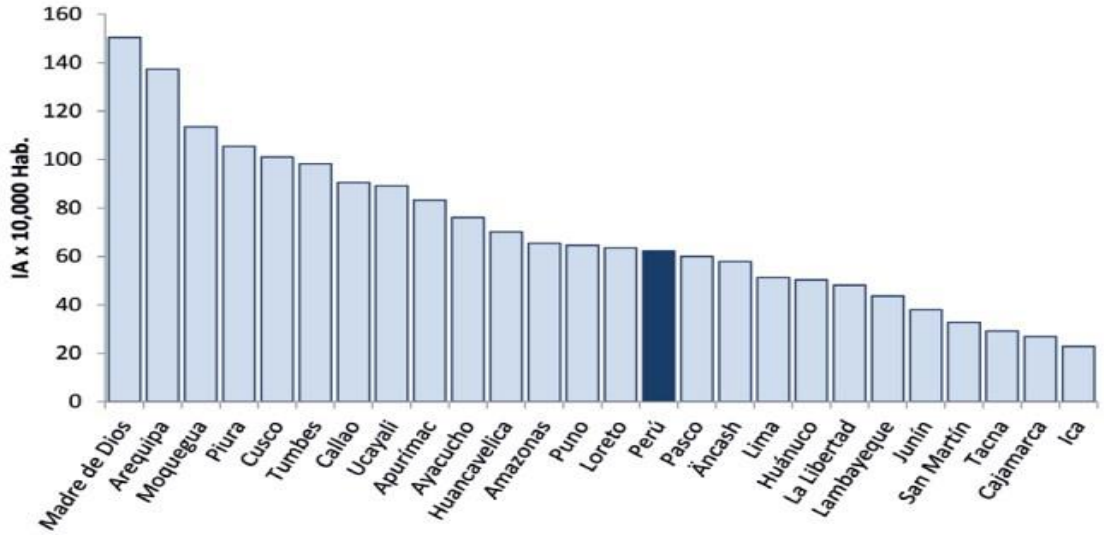


Figura 2. Incidencia acumulada de neumonías en mayores de 60 años según departamento, Perú 2019.

Así también, en particular relacionado a las empresas de salud, la minería de datos es más conocido como ciencia de datos, si bien es cierto es difícil la administración, clasificación y en ocasiones se presentan información altamente protegida, la cual confunde al controlador o procesador de los datos, ya que se debería de crear un modelado simplificado para minimizar el riesgo de una clasificación excesiva o insuficiente (Almagro, 2019).

Cabe mencionar que el Policlínico Cruz Verde, enfrenta una situación preocupante en la actualidad por el virus del COVID-19, ya que el tiempo de espera de cada paciente para poder ser atendidos, genera la aglomeración y ello puede causar nuevas enfermedades, ya que no todos van con los mismos síntomas, algunos van para pasar examen médico general o pruebas de laboratorio. Cabe mencionar, ello se genera, ya que, las enfermeras no pueden brindar un diagnóstico previo al paciente, ya que no cuentan con el conocimiento o actualización de información a detalle sobre el tema.

En consideración con la situación actual que enfrenta el Policlínico Cruz Verde, se plantea:

Problemática general:

¿De qué manera el modelo de clasificación basado en minería de datos identifica los factores que influyen en las infecciones respiratorias agudas graves de pacientes?

Problemáticas secundarias:

- ¿Qué algoritmos de aprendizaje se pueden usar para determinar los síntomas más relevantes en pacientes con infecciones respiratorias agudas graves?

- ¿Cuáles son las características de los pacientes que presentan infecciones respiratorias agudas graves según diagnóstico?

Por otro lado, en Perú, si bien es cierto, las enfermedades respiratorias agudas graves son preocupantes, ya que a través de estudios se ha verificado la

existencia de grandes disparidades en materia de morbilidad evitable, lo cual, se podría evitar si se tuviera información que estableciera alertas en tiempo real de acuerdo a los síntomas.

Tal es el caso, por lo que este trabajo de investigación se justifica de acuerdo al rubro de la tecnología, emplear la minería de datos para la construcción de un modelo de acuerdo a historias clínicas de pacientes con diagnóstico de algún tipo de infección respiratoria aguda graves para que sirva de apoyo en la toma de decisiones y contribuir en el diagnóstico previo de parte del personal. Cabe mencionar, esta tecnología, quiere lograr la clasificación de acuerdo a factores que influyen en las infecciones respiratorias agudas graves de pacientes en el Policlínico Cruz Verde, se observa, que la enfermedad común, dentro de, es la faringitis y bronquitis, por lo que, para detectar el tipo y saber el tratamiento adecuado, valerse de un modelo, el cual sirva de apoyo para el diagnóstico médico sería de gran ayuda para disminuir la incidencia de casos. En cuanto a la economía, dado a la pandemia del COVID-19, ha ocasionado el aumento de pacientes que sufren de infecciones respiratorias agudas graves, lo cual es notaria la concentración en los recursos costosos, por lo que se debe asignar prioridades en el área de admisión y se pueda brindar un diagnóstico previo.

Ante la investigación realizada, se plantean objetivos:

Objetivo general:

Desarrollar un modelo de clasificación basado en minería de datos para identificar los factores que influyen en las infecciones respiratorias agudas graves de pacientes.

Objetivos específicos:

- Identificar algoritmos de aprendizaje para determinar los síntomas más relevantes en pacientes con infecciones respiratorias agudas graves.

- Identificar las características de los pacientes que presentan infecciones respiratorias agudas graves según diagnóstico.

Estos objetivos, permiten plantear la hipótesis general y las hipótesis específicas:

Hipótesis general:

Un modelo de clasificación basado en minería de datos identifica los factores que influyen en las infecciones respiratorias agudas graves de pacientes.

Hipótesis específica:

- Un modelo de clasificación basado en minería de datos, permite reconocer los factores que influyen en las infecciones respiratorias agudas graves de pacientes del Policlínico Cruz Verde.

Se plantea la hipótesis, puesto que según Gamarra y Santos (2019). Un modelo de minería de datos permite aumentar la confiabilidad del diagnóstico, mediante la aplicación de algoritmos para predecir el cáncer de mama, para ello, se extrae los signos o síntomas de determinados pacientes para convertirlos en patrones que se puedan clasificar en subgrupos e incrementar el índice de incidencia, teniendo como objetivo la precisión, optimización y reducción de intervención médica para obtención de los datos mediante respuestas computacionales.

II. MARCO TEÓRICO

Borja y Castaño (2017) en su tesis de maestría titulada “Minería de datos de salud: investigación sobre factores individuales, familiares y de vivienda que afectan la diabetes e hipertensión basada en la encuesta de atención primaria de salud en el área metropolitana del valle de Aburra”, el cual tuvo como objetivo, aplicar técnicas de minería de datos en diferentes encuestas de atención primaria de salud para determinar los factores domésticos, personales y familiares afectados por enfermedades crónicas. Dicha investigación es de enfoque cuantitativo y tuvo como población a todos los pacientes del Área Metropolitana del Valle de Aburra. Como resultados se obtuvo que, los factores de investigación que tienen mayor impacto en el diagnóstico de diabetes son: desnutrición, resultado de la última mamografía, edad, incidencia, incidencia de EPOC y dislipidemia. Para la hipertensión, se encontraron las siguientes correlaciones: desnutrición, edad, peso y morbilidad Dislipidemia. Sobre esta base, se concluye que si el área metropolitana quiere seguir analizando desde APS, necesita contar con equipos mediante tecnología de minería de datos. Además, tener un servidor de datos puede reducir el tiempo de análisis porque la evidencia del proceso de desarrollo del trabajo indica que hay hasta 20 procesos, dependiendo de la cantidad de días para cargar los resultados (p.232).

Según un estudio realizado por Sánchez (2017), para optar su maestría de TIC, plantea como título “Minería de datos de la salud: análisis de los factores que influyen en la realización de cirugías estéticas”, el cual como objetivo era determinar el impacto de la decisión del paciente, acerca de la cirugía estética, teniendo como resultado que todo los datos personales encontrados, podrían determinar la probabilidad que el paciente se someta a una cirugía plástica, por lo que, definir los factores que influyen en las decisiones de cirugía estética, son importantes, puesto que es una estrategia para minimizar el tiempo de espera de la consulta para asignar la valoración. Asimismo, se concluyó que los factores más resaltantes; tenían que ver con el sentido del paciente, de acuerdo a una valoración dependiendo de su deseo e intención de realizarse un procedimiento específico y el rol del cirujano plástico es brindarle la confianza al paciente (p. 70).

Por otro lado, Medrano (2016), en su proyecto de tesis que lleva como título "Modelo de minería de datos de aprendizaje automático para identificar síntomas y patrones de enfermedades respiratorias en registros médicos para mejorar el diagnóstico de pacientes en Trujillo en 2016", lleva como objetivo, progresar el diagnóstico de pacientes en la ciudad de Trujillo, aplicando el desarrollo de un modelo de minería de datos, el cual se armó con reconocimiento de patrones de síntomas y enfermedades respiratorias de los pacientes, tal es el caso que su resultado se profundizó de acuerdo al descubrimiento de los patrones, asociaciones, cambios, anomalías y estructuras con gran contenido de datos, los cuales se almacenan en base de datos, repositorios de información, la nube u otros. Para lo cual, se suele usar técnicas, programas o páginas que puedan ayudar al análisis de la información logrando sacar estadísticas para cumplir con el proceso.

Además, Sosa (2017), en su proyecto de "Trazabilidad de datos en Salud", el cual fue ejecutado en Medellín, destaca como objetivo, basarse en identificar la trazabilidad de los datos en salud, ya que se podrá determinar los baches de información, tipos y la frecuencia con la que se presentan, para ello se puede usar unidades de análisis, tales como generadora de datos (UGD), compiladora de datos (UCD) y la analizadora de datos (UAD) que hoy en día son existente en el sector salud. Tal es caso, en el que demuestra que los actores del sistema de salud, se ven involucrados, dependiendo de la consolidación de análisis y de los flujos de información administrativa, ya que son el pilar para cumplir todas las fases de la trazabilidad de los datos.

De la misma manera, Rojas y Aguilar (2017), en su tesis que se titula "Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá y Colombia" para optar el título como ingenieros de sistemas, indican que su principal objetivo, era descubrir los patrones de enfermedades respiratorias en pacientes de la ciudad de Bogotá, dando uso a las técnicas de la minería de datos. De tal manera, lograron como resultados dar a conocer los comportamientos y síntomas asociados a las enfermedades respiratorias para que de esa manera la secretaria de salud pueda plantear nuevas estrategias de

prevención para disminuir la tasa de mortalidad y proponen el uso de un método de investigación para profundizar los análisis que se realizan, tomando como ejemplo, los boletines epidemiológicos, en las que se presentan estadísticas por edad, enfermedad y año (p.17).

De acuerdo a los conceptos establecidos, para poder analizar más a detalle la investigación se tomó en cuenta algunas teorías relacionadas. Tal es el caso de las fases de la minería de datos, como principal fase, es la regla de descubrimiento, que mostrará nueva relación entre variables o excepciones dependiendo de la empresa que va usar el proceso. Asimismo, sucede que si hay reglas establecidas, no pueden ser cambiadas, en caso no sea para mejorar su desempeño. Una vez que se descubren reglas importantes, pueden ser utilizadas para estimar algunas variables de salida para que se complementen las estadísticas tradicionales y los de inteligencia artificial. Cabe mencionar que las fases, están distribuidas en dos tipos de tareas.

Las tareas predictivas, se encargan de la previsión, análisis de secuencia, análisis de desviaciones y los análisis de similitud en series temporales, es de gran importancia que se cumplan todas ellas, ya que permiten procesar las variables cualitativas-cuantitativas y la predicción neuronal, también la búsqueda de los patrones de acuerdo a secuencias o transacciones en una colección de series temporales para lograr la comparación-análisis de todos los datos obtenidos. A la vez, si se cumple con todo se podrá demostrar la regresión de manera estadística y la predicción será numérica para que pueda asignar valores reales (Lejarza, I, 2020, p.6).

Por otro lado, las tareas descriptivas, se presentan más que todo en el agrupamiento o clustering para obtener datos naturales con la debida clasificación según clases o segmentos, las reglas de asociación, también están dentro de, ya que permiten identificar las relaciones no explicitas y las versiones cualitativas, es decir, las correlaciones no implican una relación de causa-efecto sino de concomitancia (frecuente), estas se presentan en tiempos subsecuentes. Todas ellas, pueden ser clasificadas en arboles de decisión (Lejarza, I, 2020, p.6).

A la vez, están las técnicas de la minería de datos, las cuales se originan de paradigmas metodológicos, entre ellas hay métodos que dependen de algoritmos-variantes para determinar su efectividad, ya sea mayor o menor para lograr sacar estadísticas con el adecuado análisis jerárquico o no jerárquico, a la vez, va permitir resolver problemas de agrupamiento. A más detalle, se observa la clasificación:

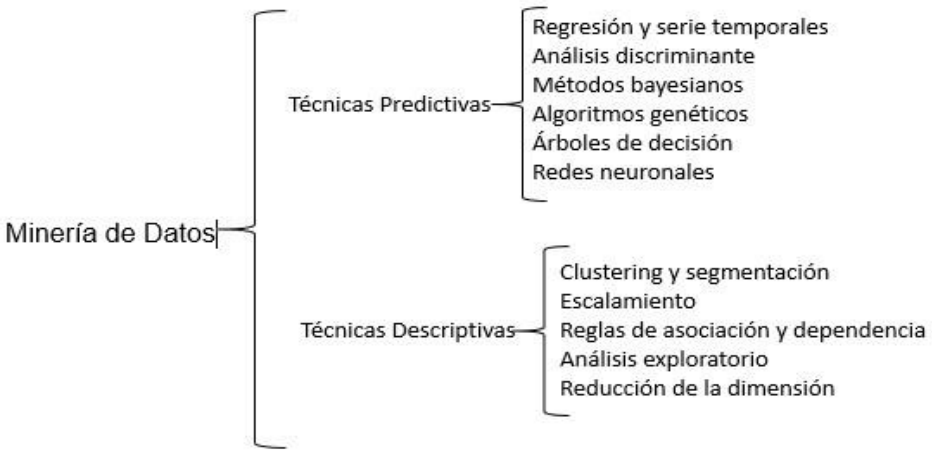


Figura 3. Principales técnicas de minería de datos. Fuente: Universidad tecnológica nacional

Se distribuye en 2 técnicas, las técnicas predictivas que se utilizan para obtener un modelo que servirá para aplicarlo a datos futuros, principalmente para predecir comportamientos. En inteligencia artificial, se les llama modelos de aprendizaje supervisado. Las variables utilizadas pueden ser categóricas y numéricas. Dentro de ellas, están considerados, la regresión y serie temporales, las cuales, se conocen como la predicción o estimación para que se pueda determinar una función real, logrando minimizar el error entre el valor determinado y el valor predictivo. Asimismo, se debe tomar en cuenta el tiempo para poder distribuir los recogidos en un orden determinado (Zamora, 2018, p.25).

Como bien se sabe hay diferente tipos de redes, por ejemplo la de MLP que sirve para entrenar con un gran conjunto de datos, por ejemplo en la Figura 4. , se muestra una clasificación de síntomas en una empresa de salud. La cual se elaboró por capas, la capa de entrada con nodos para cada variable

clasificadora, la de salida como tipo de seguro y 2 capas ocultas que predominan el proceso ciego.

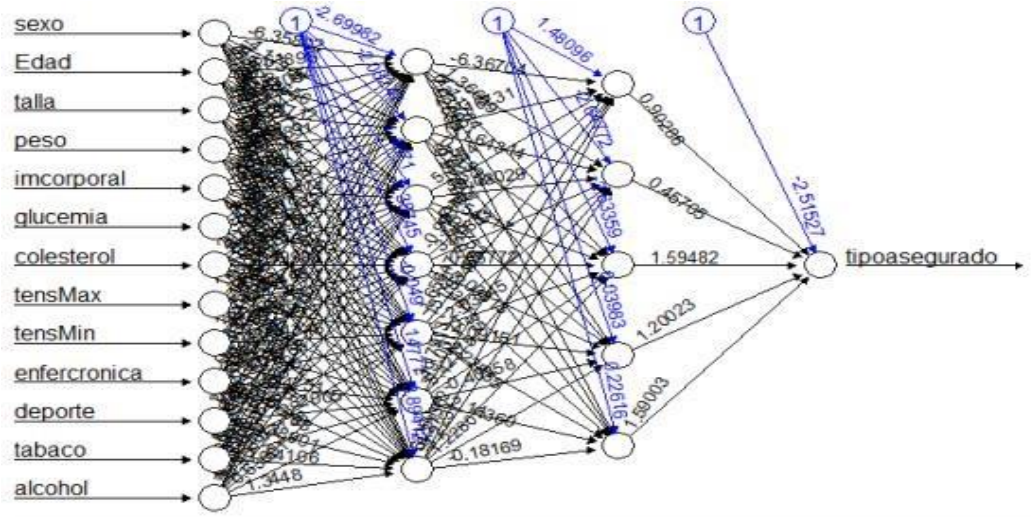


Figura 4. Clasificación de síntomas.

Otra de las técnicas, son las descriptivas, están orientadas a definir grandes conjuntos de datos, teniendo como objetivo encontrar la descripción de los datos. Según Romero, 2019. Los algoritmos más utilizados son los de K-Means que se basan en aprendizaje de máquinas, de acuerdo al uso de vectores de entrada en los conjuntos de datos, sin necesidad que dependan de resultados etiquetados. Otro de los algoritmos es el de vecino K más cercano, se encarga de la clasificación en el aprendizaje por máquina y tiene dominio de aprendizaje supervisado.

De igual importancia, son las metodologías en minería de datos, primero mencionamos al KDD, metodología en el que se usa la minería de datos para extraer el conocimiento, dando uso a una base de datos junto a un pre procesamiento, sub muestreo y transformación requerida en la base de datos, esta cuenta con cinco etapas, tal cual se muestra en la Figura 5.:

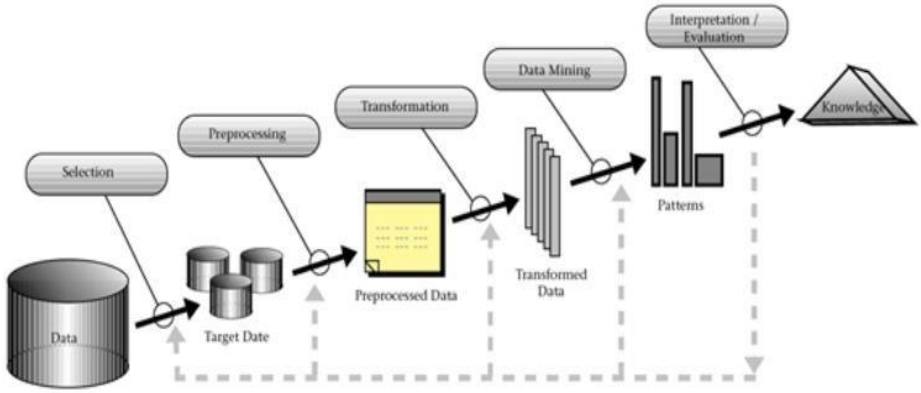


Figura 5. Etapas de KDD.

De acuerdo a la Figura 5., se empieza con la selección que consiste en realizar el descubrimiento de las variables o muestra de datos que se usarán más adelante, es decir, se escoge la información que va resultar útil (Grández, 2017). Antes de finalizar, se emplea la minería de datos que consiste en la búsqueda de los patrones y la elección del método a emplear (Galán, 2016). Finalmente, la interpretación y evaluación de los patrones encontrados en la etapa anterior para comprender los resultados obtenidos (Grández, 2017).

Otra metodología SEMMA, en la que prima la representación de los datos y se aplican estadísticas de visualización, se transforman las variables y se evalúa con exactitud el modelo, esta cuenta con cinco fases, tal cual se muestra en la Figura 6.:

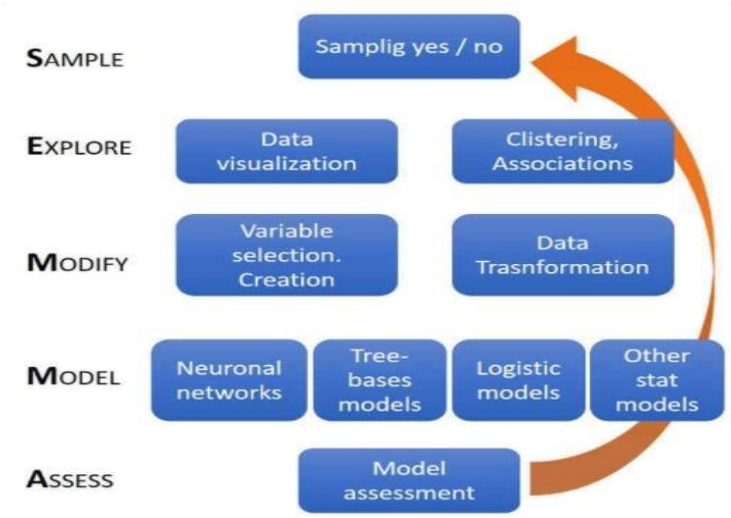


Figura 6. Fases de SEMMA por Romero Jorge 2019.

De acuerdo a la Figura 6., la primera fase es el muestreo (SAMPLE) que se realiza una muestra significativa para reducir el tiempo de procesamiento, seguidamente del entrenamiento para ajustar el modelo (Romero, 2019). Antes de finalizar, se emplea el modelado (MODEL) para la predicción de resultados esperados de acuerdo a los datos obtenidos, ya que engloba a las redes neuronales (Romero, 2019). Por último, la evaluación (ASSESS), en esta fase, se evalúa la fiabilidad y eficiencia de los modelos, ello se realiza con la frecuencia de los datos y la predicción aceptable (Grández, 2017).

Por último, está la metodología CRISP-DM que tiene todas las fases de un proyecto, las tareas primarias y las relaciones entre ellas. La secuencia de fases que ofrece la metodología no es rígida, por lo que todo depende de los resultados de cada fase, cuenta con seis fases, tal cual se muestra en la Figura

7. :

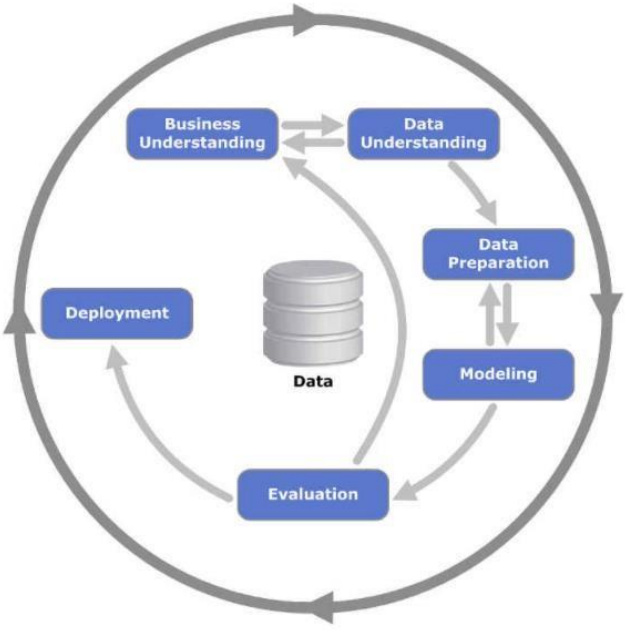


Figura 7. Fases de CRISP-DM por Kenneth Jensen.

Primero se realiza el análisis del problema, el cual, se centra en el análisis de objetivos y requisitos para poder desarrollar un plan preliminar para que demuestre los logros de objetivos (Romero, 2019).Luego se emplea el análisis de los datos, en esta fase se realiza la obtención de los datos, identificación de los problemas de calidad, detectar subconjuntos para formular hipótesis de

información que es desconocida (Romero, 2019). A la vez, se realiza la preparación de los datos para construir el conjunto de los datos finales que están sin procesar desde la obtención de ellos (Romero, 2019). Antes de finalizar, se realiza el modelado, selección y aplicación de modelos con parámetros óptimos, en caso que los valores no cuenten con las características, se vuelve a la preparación de los datos (Romero, 2019). Luego la evaluación para verificar el rendimiento-integridad de los pasos y se verifica si se han cumplido debidamente todos los objetivos del negocio (Romero, 2019). Finalmente la implementación, se revisa los cumplimientos de todos los modelado, en caso no sea así, se aumenta el conocimiento de los datos, se organiza y se presenta de tal manera que el cliente pueda usarlo (Grández, 2017).

Otro aspecto a tomar en cuenta en la investigación es el proceso de atención de salud, según la Organización Mundial de la Salud (2020) plantea reconocer a todos los pacientes con IRAG en el primer contacto con el sistema sanitario para poder gestionar-administrar los tratamientos y mejorar la atención clínica de estos pacientes, facilitando las orientaciones más actualizadas con la correcta prevención de las infecciones. Si bien es cierto, existen las consultas ambulatorias, las cuales pueden ayudar cuando la enfermedad está iniciando, por lo contrario, ya se requiere intervención de profesionales, ya que en la actualidad se encuentran como las primeras causas de mortalidad. Así como también, dentro de se evalúa el diagnóstico, el cual, se basa en un procedimiento para que los profesionales de la salud identifiquen la enfermedad (o no enfermedad) de un paciente en función de los síntomas con la ayuda de diversas herramientas que pueden definir su situación clínica (Sánchez, S, 2019).

Para Molina y García (2018, p. 120) “La clasificación es el proceso de dividir un conjunto de datos en grupos mutuamente excluyentes, de tal forma que cada miembro de un grupo esté lo más cerca posible de otros y los grupos diferentes estén lo más lejos posible de otros, en donde la distancia se va a medir con respecto a las variables especificadas que se pretende predecir”. Es decir, el árbol de decisión es una forma gráfica y analítica que por medio de distintos caminos posibles van a efectuar la clasificación de los datos, los caminos posibles representan las ramificaciones del árbol, estas ramificaciones forman

nodos u hojas, cada una de los nodos del árbol representa los diferentes atributos presentes en los datos.

De los diversos algoritmos de árboles de decisión, los más reconocidos en el mundo de la minería de datos son el IDE3 (induction of decisión trees), el C4.5 ambos desarrollados por J.R. Quinlan y CART (Classification and regression Trees). El algoritmo CART elabora particiones binarias con una habilidad de poda fundamentada en un criterio de coste-complejidad, según Romeu (2019) las particiones binarias son el resultado de evaluar una condición que tiene dos únicas respuestas, si Zl es un atributo categórico con valores {Z_1, Z_2, Z_3… Zn}, se pregunta si Zl está entre un subconjunto de los valores categóricos, siendo la respuesta sí o no. Si Zl es un atributo continuo, Q incluye preguntas del tipo ¿Zl <= v?, siendo v un valor real cualquiera. Para reducir, CART captura v como el punto medio entre dos valores consecutivos de Zl.

De la misma manera, está el algoritmo IDE3 planteado en el año 1986 por Quinlan, es un algoritmo simple, pero a la vez robusto, poderoso que va permitir diseñar un árbol de decisión con atributos discretos y continuos, tiene la capacidad de trabajar con ruido en los datos y también es capaz de adquirir conocimientos de una disyunción de expresiones, para construir el árbol, el algoritmo aplica el análisis de la entropía como función de impurezas. La entropía es la incertidumbre o aleatoriedad que existe en un sistema, en otras palabras, ante una determinada situación la probabilidad que suceda cada uno de los probables resultados (Romeu, 2018, p.51).

Algoritmo C4.5 creado por Quinlan como extensión del IDE3, en este algoritmo se detectaron algunos inconvenientes al presentar problemas de no clasificar los atributos numéricos de forma adecuada, y la tendencia de favorecer a los atributos de numerosos valores.

III. METODOLOGÍA

3.1 Tipo y Diseño de Investigación

El presente proyecto de investigación será de tipo aplicado porque busca generar conocimiento mediante la aplicación de soluciones para los problemas que surgen en la realidad de la población. Del mismo modo, la investigación aplicada, según Schwarz (2017) es considerada aquella que se plantea como objetivo principal resolver un determinado problema o un método específico (p. 11).

Asimismo, el diseño de investigación será pre-experimental porque el grado de control fue mínimo, es decir no existió mayores posibilidades de comparación de grupos, ya que no hubo más variables que intervinieran. Asimismo, Guevara, Verdesoto y Castro (2020) Declara que este método incluye la manipulación por parte del investigador de una o más variables para controlar sus cambios y efectos (p. 168).

El alcance del estudio será explicativo, según Ramos (2020), el alcance busca esclarecer situaciones específicas y menciona que se debe formular una hipótesis para encontrar la causa y efecto del fenómeno. Además, el estudio utiliza un método cuantitativo, Sánchez (2019) menciona que el método utiliza técnicas estadísticas para analizar los datos recogidos mediante fenómenos medibles (página 104).

3.2 Variables y Operacionalización

Variable independiente: Minería de datos

Tabla 2. *Variables y Operacionalización.*

Variables	Definición Conceptual	Definición Operacional	Dimensiones	Indicadores	Instrumento	Escala de Medición
Modelo de clasificación basado en Minería de datos	El modelo de clasificación, permite la predicción de acuerdo a los datos recolectados, basándose en la minería de datos se realiza una búsqueda de relaciones, correlaciones, dependencias, asociaciones, segmentos, los cuales se obtienen de grandes juegos de datos (Akbari, S, 2016).	Los algoritmos de aprendizajes, si bien es cierto, son procesos capaces de seleccionar funciones que no se repitan, ya que se encargan de que sus valores o respuestas sean auténticas, teniendo como finalidad resolver el problema planteado.	Resultado de algoritmo de aprendizaje	Precisión del algoritmo de aprendizaje	Registro de datos físicos	Razón
				Clasificación de instancias correctas		

Fuente: Elaboración propia.

3.3. Población, Muestra y Muestreo

La población es el total del objeto a estudiar en donde, se tiene que tomar en cuenta que deben poseer una característica en común para que se establezcan criterios para determinar los pacientes, en este caso, son todos los pacientes con diagnóstico médico de tipos de infección respiratoria aguda grave del Policlínico Cruz Verde en la provincia de Carhuaz, en el que se recolectó información de un total de 300 pacientes.

- Criterios de selección:
  - Pacientes con diagnóstico de Faringitis, Amigdalitis aguda, Bronquitis aguda, Covid 19, etc. en el Policlínico Cruz Verde, en el periodo Agosto-Setiembre 2019 y Agosto-Setiembre 2020 que son los meses en el que se presentaron la mayor incidencia de diagnóstico de infecciones respiratorias.

  - Historia clínica presente.

- Unidad de análisis

  Se toma en cuenta a cada uno de los 300 pacientes con diagnóstico de infección respiratoria aguda grave común en el Policlínico Cruz Verde.

Dividir las necesidades del grupo de investigación seleccionando una muestra que se denomina subconjunto del todo o parte representativa del todo, se compone de unidades que son elementos de la investigación. El muestreo es una herramienta que apoya la investigación científica, ya que sirve para determinar la parte de la población a estudiar (Hernández, 2019).

3.4. Técnicas e instrumentos de recolección de datos

Para esta investigación se usará como técnica de recolección de datos la observación y como instrumento la ficha de registro. Teniendo en cuenta que ello serán las historias clínicas, ya que son usadas por doctores que

evalúan directamente al paciente, en el cual se podrá observar sus síntomas, diagnóstico, edad, el tratamiento, temperatura, presión arterial y su procedencia. Cabe destacar, ello servirá como fuente de apoyo para la variable independiente minería de datos, ya que se va evaluar por indicadores. Asimismo, la dimensión, los indicadores, la técnica e instrumentos de la variable, se podrán observar en la Tabla 2.

Tabla 3*. Recolección de datos.*

DIMENSIÓN	INDICADOR	TÉCNICA	INSTRUMENTO
Resultado de algoritmo de aprendizaje	Precisión del algoritmo de aprendizaje	Observación	Ficha de registro de datos físicos
	Clasificación de instancias correctas		

Fuente: Elaboración propia.

3.5 Procedimientos

En esta investigación se evaluó una de las problemáticas recurrentes en el Policlínico Cruz Verde que es el tiempo de espera, lo cual genera aglomeración de las personas y en la actualidad, no debería suceder ello, ya que, si hay pacientes con algún tipo de infección respiratoria aguda grave, estarían exponiendo a los pacientes que van por alguna enfermedad pasajera. Luego, se pasa a revisar tesis en las cuales se hayan expuesto la variable independiente. Tomando en cuenta, las fases de la metodología a usar de la minería de datos, se toma la decisión que a futuro en el Policlínico, se pueda implementar un modelo de minería de datos para la clasificación de factores de las infecciones respiratorias agudas graves

Cabe destacar que, para ello, primero será de suma importancia concluir con la recolección de datos, siguiendo determinados pasos:

- Ordenar las historias clínicas, dependiendo del año.

- Se procede a la revisión y dictamen de las historias clínicas con los expertos en las enfermedades el Dr. Julio Vilca y la Dra. Johana Narciso.
- Se digita los datos en el programa informático Excel de Microsoft si cumple con los campos de recolección de datos, los cuales son diagnóstico, síntomas, edad, tratamiento, sexo, temperatura, PA (presión arterial), SPS (Saturación), FC (Frecuencia cardiaca) y procedencia.

Al finalizar la recolección de los datos, se procede a seguir las fases de la metodología a emplear en este caso será CRISP-DM, asimismo, se establecen las herramientas y librerías que se emplean para la construcción del modelo de clasificación.

3.6. Método de análisis de datos

El método de análisis de datos que se ha utilizado en el proyecto se basa en tablas estadísticas para ello, se empleó la herramienta anaconda y el desarrollador de códigos spyder con el lenguaje Python para generar las gráficas de acuerdo a las conjugaciones que se podía realizar con los campos.

Se opta por el lenguaje Python, ya que es bastante flexible y cuenta con librerías para implicar nuevos datos en las plataformas o sistemas más conocidos, de la misma manera, con la data importada, se puede empezar a construir el modelo, tal es el caso, en el que ya se procede a escoger el tipo de algoritmo de aprendizaje a emplear, de acuerdo a las gráficas, ya que primero se mide la relevancia de los datos.

Además, se aplicará hipótesis para validar la confiabilidad de la investigación.

HE1: Un modelo de clasificación basado en minería de datos identifica los factores que influyen en las infecciones respiratorias agudas graves de pacientes.

Ha: Un modelo de clasificación de minería de datos, permite reconocer los factores que influyen en las infecciones respiratorias agudas graves de pacientes del Policlínico Cruz Verde.

H0: Un modelo de clasificación de minería de datos, no permite reconocer los factores que influyen en las infecciones respiratorias agudas graves de pacientes del Policlínico Cruz Verde.

3.7 Aspectos éticos

La presente investigación está sujeta a lo que dicta la ética de un investigador, en el ámbito de la salud, esta promueve el respeto, la protección de los derechos de seres humanos y la confiabilidad de su información. Por cuestiones legales, en el proyecto de investigación, se restringió el uso de datos personales, es decir sus nombres e identificación que involucre al paciente, por lo que, se decidió para cumplir con lo establecido, solo la recolección de información de síntomas, tratamiento, el diagnóstico y los rangos de edades. Además, se centrará en respetar todos los derechos de propiedad intelectual que puedan citarse en esta investigación, citando correctamente a cada autor, asegurando la alta calidad del proyecto y autenticidad de su información, para que los futuros investigadores puedan utilizar dicha información.

IV. RESULTADOS

Se muestra el flujo de trabajo de las fases de la Metodología CRISP-DM, la cual es empleada para desarrollar el modelo de clasificación basado en minería de datos, a continuación, se detalla por cada fase en la Figura 8.

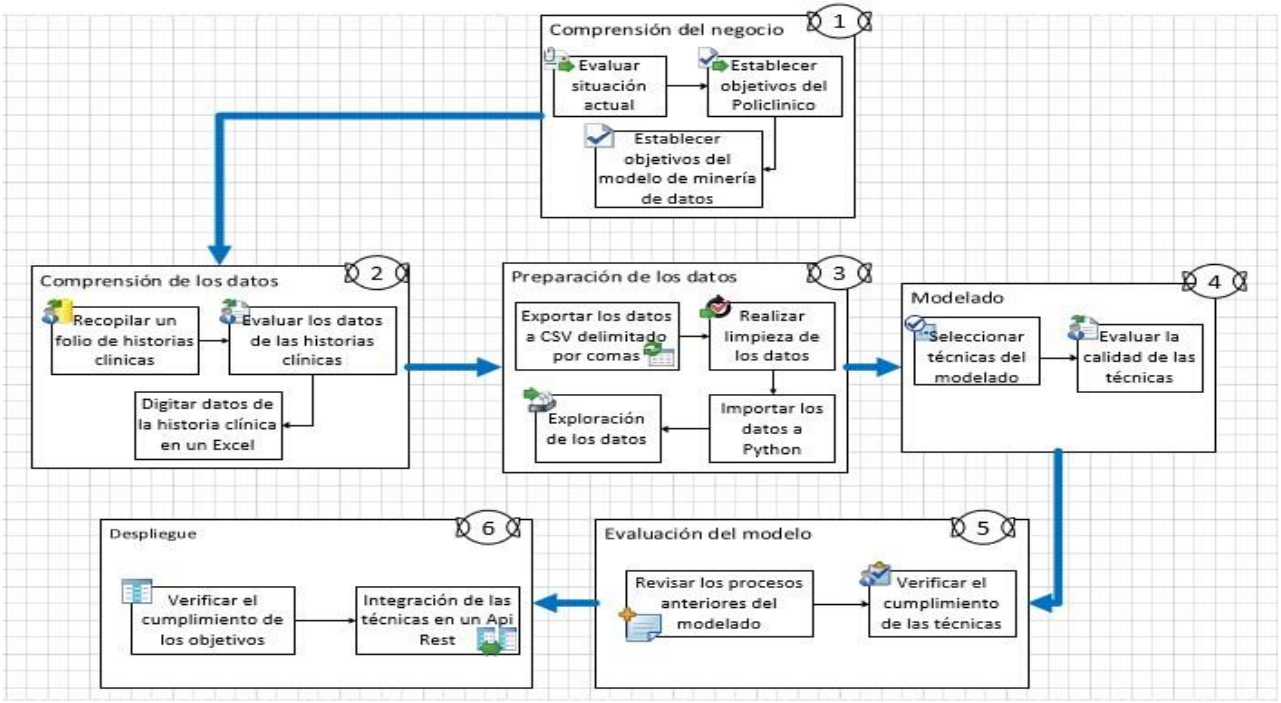


Figura 8. Flujo de trabajo de las fases de CRISP-DM.

4.1. Comprensión del negocio

4.1.1 Breve descripción de la empresa

El Policlínico Cruz Verde, es una organización nivel I-III, es decir, cuenta con mayores unidades productoras de servicios de salud. Se encuentra ubicado en la esquina de Jr. Comercio y 2 de Mayo-Carhuaz, tiene 12 años de antigüedad, se encarga de la atención primaria de Salud, su misión es brindar servicios de calidad, con personal altamente calificado y humanizado, que garantiza una atención personalizada a sus usuarios; antes, durante y después del desarrollo de su enfermedad, como visión quieren logran ser una institución acreditada y reconocido a nivel local, regional que cuenta con diversas especialidades y servicios para brindar una atención de salud integral de calidad, oportuna y accesible, centrado en las necesidades de cada usuario y que brinda apoyo en la atención de pacientes de bajos recursos. Cabe mencionar, en la provincia de Carhuaz el sector salud, es poco explotado actualmente. A pesar de ello, el policlínico cuenta con una barrera alta de sobrepasar y es la fidelidad de los pacientes, la credibilidad de los médicos que laboran en la institución y los proveedores que se encargan del servicio terceriazado, los cuales se manejan a base de convenios con la cláusula de exclusividad de atención en la institución.

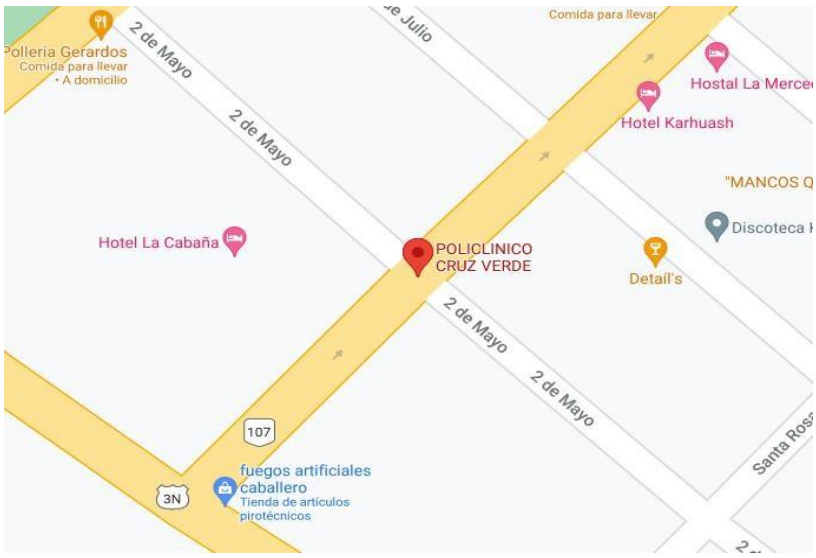


Figura 9 Mapa de la ubicación del Policlinico Cruz Verde.

4.1.2 Objetivo de la empresa

Los objetivos estratégicos del Policlinico, se enfocan en:

- ❖ Brindar atención médica segura y de alta calidad.
- ❖ Apoyar y desarrollar los servicios médicos
- ❖ Garantizar los servicios de atención médica a todos, en particular a los grupos desfavorecidos.

4.1.3 Objetivo del modelo

Este proyecto de clasificación basado en minería de datos tiene como propósito determinar un algoritmo de aprendizaje para apoyo en el diagnóstico previo por parte del personal que se encarga de realizar el triaje para cada paciente en el Policlinico Cruz Verde, puesto que en la actualidad la pandemia del COVID 19, ha generado la aglomeración de personas y nuevos diagnósticos de infecciones respiratorias. Por lo que, la técnica de minería y la elección del algoritmo de aprendizaje, debe aumentar el grado de acierto del diagnóstico, disminuir la incertidumbre del personal y lograr la clasificación de instancia. Cabe mencionar que la meta de mi proyecto de investigación, es lograr un modelo a partir del mejor algoritmo de aprendizaje que genere ciertos patrones o sentencias con datos históricos obtenidos de los pacientes con el diagnóstico de infecciones respiratorias agudas graves.

4.1.4 Finalidad del modelo

El modelo generado debe ser comprensible y dinámico, así el usuario final (enfermera) pueda entenderlo e interpretarlo para tomar una decisión.

4.2. Recolección de los datos

Para el comienzo o ejecución del proyecto la fase 2 inició con la recolección de los datos de historias clínicas de los pacientes diagnosticados con algún tipo de infección respiratoria aguda grave en el Policlínico Cruz Verde. Cabe destacar, que para recibir o me puedan entregar un folio de las historias clínicas, se contó con un documento en el que el Dr. Vilca Begazo Julio, acepta la construcción del modelo de clasificación basado en minería de datos, teniendo en cuenta que se hará uso de las historias clínicas de los pacientes con diagnóstico de IRAG, ello se podrá observar en el ANEXO 2. Luego de ello, se procedió a digitar datos de las historias clínicas, para lo cual, se usó un Excel, el cual contaba con las siguientes columnas la fecha de atención, los síntomas, diagnóstico, PA, SPS, Temperatura, la edad, sexo, FC y procedencia que presentan los pacientes. La especificación, exploración y revisión de estos atributos o variables, se realizó con la ayuda de los médicos responsables del diagnóstico y control de los pacientes, en este caso el Dr. Julio Vilca y la Dra. Johana Narciso. Una vez concluida la recolección de información para crear el modelo de mi proyecto de investigación, se consolida 300 datos, para lo cual se tomó en consideración los criterios de selección de la población, es decir, los pacientes tenían que haber sido diagnosticados en el periodo agosto-setiembre del 2019 y agostosetiembre del 2020. Finalmente, el formato para proceder exportar el archivo, será .csv delimitado por comas, dado que Excel permite guardar en esta extensión, luego se procede a verificar que los campos no estén vacíos, ello se realizar en un editor de texto. Luego, se empleó la herramienta anaconda y el desarrollador de códigos spyder con el lenguaje Python para generar las gráficas de acuerdo a las conjugaciones que se podía realizar con los campos.

A continuación se presenta una lista de los campos que se emplearon en el Excel, con su respectiva abreviación y descripción:

Tabla 4. *Abreviación y Descripción de los campos.*

N°	NOMBRE	ABREVIATURA	DESCRIPCIÓN

1	Fecha de atención	FA	Fecha que el paciente ha ido a pasar consulta el paciente.
2	Diagnóstico	DIAG	Tipo de infección respiratoria diagnosticada.
3	Síntomas	SINT	Síntomas de la infección respiratoria.
4	Edad	ED	Edad del paciente.
5	Tratamiento	TRAT	Medicamentos que se le recomienda que debería tomar para recuperarse.
6	Sexo	S	Genero del paciente (F o M)
7	Temperatura	T	Nivel de fiebre de acuerdo a su tipo de infección
8	Presión Arterial	PA	Presión Arterial del paciente.
9	Saturación	SPS	Saturación de paciente oxigenado.
10	Frecuencia Cardiaca	FC	Pulso por minuto del paciente.
11	Procedencia	PROC	Lugar de vivienda del paciente.
12	Antecedentes	ANT	Prueba de Covid 19 (positivo o negativo).

Fuente: Elaboración propia.

4.3. Preparación de los datos

Para este proyecto de investigación, se recolecto 300 datos de pacientes con diagnóstico de algún tipo de infección respiratoria aguda grave de Policlinico Cruz Verde, de los cuales fueron diagnosticados con Amigdalitis aguda, Bronconeumonía, Bronquitis, Bronquitis aguda, Bronquitis aguda con COVID 19, Bronquitis asmática, Bronquitis crónica, COVID 19, Faringitis, Faringitis aguda, Faringitis amigdalitis aguda, Fibrosis pulmonar, Neumonía, Rinitis aguda. Cabe mencionar, que para identificar los tipos de enfermedades, se realizó reuniones con los doctores especializados para decidir si todos los síntomas de cada enfermedad era necesario para identificar la enfermedad, por lo que se decidió mantener todos los síntomas, datos del paciente, en cuanto a su temperatura, presión, frecuencia cardiaca, edad y de acuerdo a la relevancia e incidencia de los datos, se podría establecer el más preeminente, dependiendo de las conjugaciones que se realice. Como bien se mencionó para verificar los datos, se realizaron reuniones con los especialistas, las evidencias de ello, se podrá observar en el ANEXO 3

Seguidamente, primero se realiza la limpieza de los datos, es decir, que no haya datos nulos o vacíos, que todo esté en su determinada columna, ello se puede verificar luego de exportar al formato .csv, visualizándolo en un editor de texto, cada campo que pertenece a una columna debe de estar en comillas, ya que al importar a Python, sino cumple con esa condición, todo va figurar en una columna y las demás columnas estarán vacías, tal como se observa en la Figura 10 y la manera correcta

que los datos estén digitalizados es como se observa en la Figura 11.

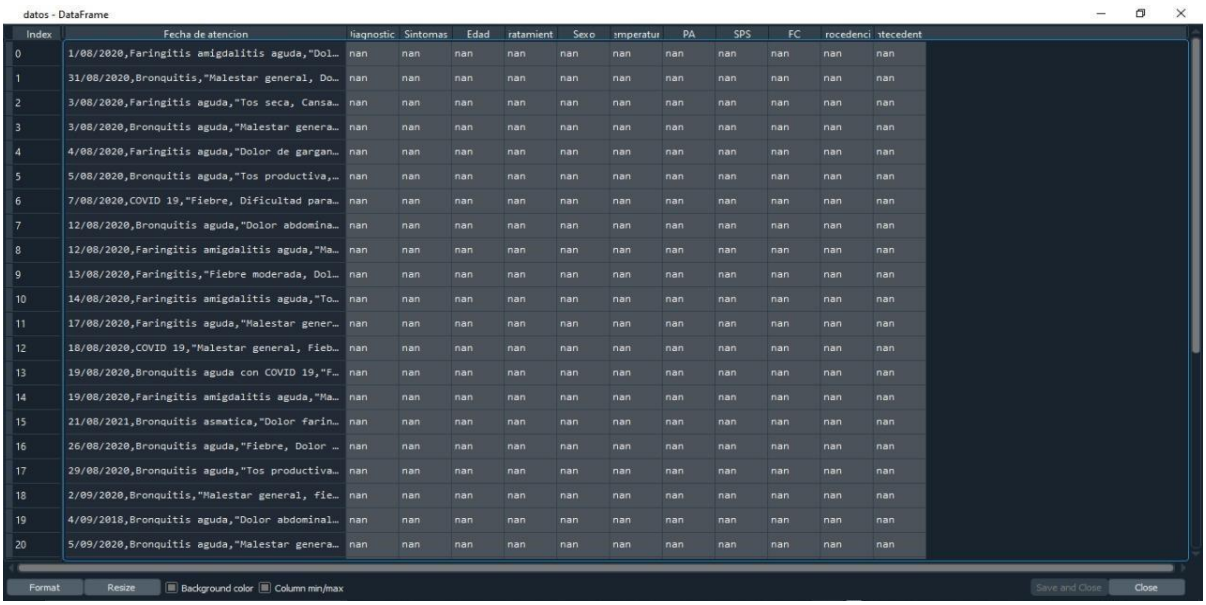
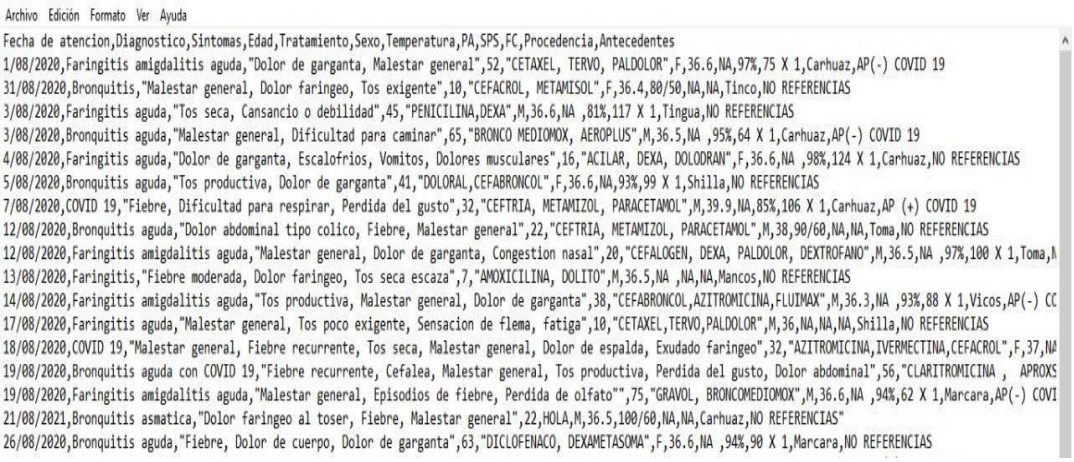


Figura 10. Tabla de datos importados.

En la Figura 11, se edita las líneas de datos para que se delimiten por comillas, de tal manera, lograr que todos los datos estén en su correspondiente columna, ya que Python tiene la librería Pandas, la cual, trabaja con consultas por nombre de columnas y filas de la tabla. Cabe mencionar, los datos que se delimitan por comillas son los SINTOMAS y TRATAMIENTO, ya que cuentan con diferentes datos, es decir es más que 1, la información que se procesa es de los pacientes diagnosticados con algún tipo de infección respiratoria aguda grave. A continuación se muestra el editor de texto.



```
Archivo  Edición  Formato  Ver  Ayuda
Fecha de atencion,Diagnostico,Sintomas,Edad,Tratamiento,Sexo,Temperatura,PA,SPS,FC,Procedencia,Antecedentes
1/08/2020,Faringitis amigdalitis aguda,"Dolor de garganta, Malestar general",52,"CETAXEL, TERVO, PALDOLOR",F,36.6,NA,97%,75 X 1,Carhuaz,AP(-) COVID 19
31/08/2020,Bronquitis,"Malestar general, Dolor faringeo, Tos exigente",10,"CEFACROL, METAMISOL",F,36.4,80/50,NA,NA,Tinco,NO REFERENCIAS
3/08/2020,Faringitis aguda,"Tos seca, Cansancio o debilidad",45,"PENICILINA,DEXA",M,36.6,NA ,81%,117 X 1,Tingua,NO REFERENCIAS
3/08/2020,Bronquitis aguda,"Malestar general, Dificultad para caminar",65,"BRONCO MEDIOMOX, AEROPLUS",M,36.5,NA ,95%,64 X 1,Carhuaz,AP(-) COVID 19
4/08/2020,Faringitis aguda,"Dolor de garganta, Escalofrios, Vomitos, Dolores musculares",16,"ACILAR, DEXA, DOLODRAN",F,36.6,NA ,98%,124 X 1,Carhuaz,NO REFERENCIAS
5/08/2020,Bronquitis aguda,"Tos productiva, Dolor de garganta",41,"DOLORAL,CEFABRONCOL",F,36.6,NA,93%,99 X 1,Shilla,NO REFERENCIAS
7/08/2020,COVID 19,"Fiebre, Dificultad para respirar, Perdida del gusto",32,"CEFTRIA, METAMIZOL, PARACETAMOL",M,39.9,NA,85%,106 X 1,Carhuaz,AP (+) COVID 19
12/08/2020,Bronquitis aguda,"Dolor abdominal tipo colico, Fiebre, Malestar general",22,"CEFTRIA, METAMIZOL, PARACETAMOL",M,38,90/60,NA,NA,Toma,NO REFERENCIAS
12/08/2020,Faringitis amigdalitis aguda,"Malestar general, Dolor de garganta, Congestion nasal",20,"CEFALOGEN, DEXA, PALDOLOR, DEXTROFANO",M,36.5,NA ,97%,100 X 1,Toma,N
13/08/2020,Faringitis,"Fiebre moderada, Dolor faringeo, Tos seca escaza",7,"AMOXICILINA, DOLITO",M,36.5,NA ,NA,NA,Mancos,NO REFERENCIAS
14/08/2020,Faringitis amigdalitis aguda,"Tos productiva, Malestar general, Dolor de garganta",38,"CEFABRONCOL,AZITROMICINA,FLUIMAX",M,36.3,NA ,93%,88 X 1,Vicos,AP(-) CO
17/08/2020,Faringitis aguda,"Malestar general, Tos poco exigente, Sensacion de flema, fatiga",10,"CETAXEL,TERVO,PALDOLOR",M,36,NA,NA,NA,Shilla,NO REFERENCIAS
18/08/2020,COVID 19,"Malestar general, Fiebre recurrente, Tos seca, Malestar general, Dolor de espalda, Exudado faringeo",32,"AZITROMICINA,IVERMECTINA,CEFACROL",F,37,NA
19/08/2020,Bronquitis aguda con COVID 19,"Fiebre recurrente, Cefalea, Malestar general, Tos productiva, Perdida del gusto, Dolor abdominal",56,"CLARITROMICINA ,  APROXS
19/08/2020,Faringitis amigdalitis aguda,"Malestar general, Episodios de fiebre, Perdida de olfato"",75,"GRAVOL, BRONCOMEDIOMOX",M,36.6,NA ,94%,62 X 1,Marcara,AP(-) COVI
21/08/2021,Bronquitis asmatica,"Dolor faringeo al toser, Fiebre, Malestar general",22,HOLA,M,36.5,100/60,NA,NA,Carhuaz,NO REFERENCIAS"
26/08/2020,Bronquitis aguda,"Fiebre, Dolor de cuerpo, Dolor de garganta",63,"DICLOFENACO, DEXAMETASOMA",F,36.6,NA ,94%,90 X 1,Marcara,NO REFERENCIAS
```

Figura 11. Editor de texto de los datos a usar.

A parte, luego de modificar en el editor de texto, se sube en editor de códigos Spyder, en el que se va llenando las columnas que figuran en NAN con el número 0 para que de esa manera se pueda hacer consultas de todas las filas x columnas, para ello, se usó la siguiente sentencia que se muestra en la Figura 12.

```
n=df.fillna(0)
print(n)
```

Figura 12. Código para cambiar datos nulos.

Seguidamente, para continuar con la preparación de los datos, se cambia la cabecera de la tabla con la abreviatura de las columnas ya mencionadas anteriormente en la fase 2 de recolección, ello se realiza con el comando cabecera=["FA","DIAG","SINT","ED","TRAT","S","T","PA","SPS","FC","PRO C","ANT"] y df.columns=cabecera, el cambio se puede visualizar en la Figura 13.

```
cabecera=["FA","DIAG","SINT","ED","TRAT","S","T","PA","SPS","FC","PROC","ANT"]
df.columns=cabecera
           FA                          DIAG  ...     PROC              ANT
0    1/08/2020  Faringitis amigdalitis aguda  ...  Carhuaz  AP(-) COVID 19
1   31/08/2020                    Bronquitis  ...    Tinco  NO REFERENCIAS
2    3/08/2020            Faringitis aguda  ...   Tingua  NO REFERENCIAS
3    3/08/2020            Bronquitis aguda  ...  Carhuaz  AP(-) COVID 19
4    4/08/2020            Faringitis aguda  ...  Carhuaz  NO REFERENCIAS
```

*Figura 13.*Sentencia y lista de las primeras 5 filas al cambiar la cabecera.

Seguidamente, se procedió a eliminar columnas con las que no se iban a conjugar, tales como el tratamiento, presión arterial, frecuencia cardiaca, antecedentes (positivo o negativo a COVID 19) y saturación de paciente (SPS), ya que, son datos que van a variar de acuerdo a la edad, análisis que se haya realizado el paciente y si tiene alguna otra enfermedad cardiaca. Para ejecutar ello, se usó la siguiente sentencia que se puede ver en la Figura 14.

```
v=df.drop(columns=["TRAT","PA","FC","ANT","SPS"])
print(v)
```

Figura 14. Sentencia para eliminar columnas.

Luego se digita un comando para saber el tipo de dato que se maneja por columnas, tal como se observa en la Figura 15. Cabe mencionar que ello, es de gran ayuda para tener en cuenta que columnas necesitan ser cambiadas a números binarios o enteros para que se pueda realizar las correctas conjugaciones y lograr la construcción de los modelos.

```
u=df.dtypes
print(u)
FA        object
DIAG      object
SINT      object
ED         int64
TRAT      object
S         object
T        float64
PA        object
SPS       object
FC        object
PROC      object
ANT       object
dtype: object
```

Figura 15 Tipo de datos de las variables.

Luego de visualizar el tipo de datos de las variables, sigue las conjugaciones de las variables, primero se realiza una sentencia para saber los tipos de enfermedades respiratorias únicas hay y la cantidad de pacientes por cada una de ellas, en la Figura 16, se muestra en gráfico de barras y su leyenda de la características con las que cuenta la enfermedad o diagnóstico.

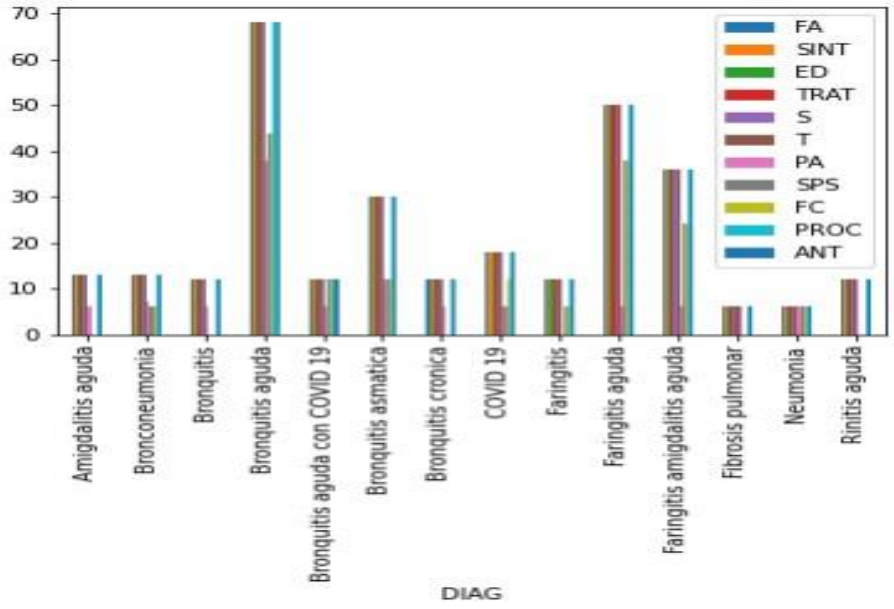


Figura 16. Gráfico de barras de Diagnósticos de infecciones respiratorias agudas graves.

En la figura anterior, se muestran los diferentes tipos de diagnóstico con su cantidad de paciente, a pesar de ello, no se llega a los datos exactos, por lo que se procede a crear otra sentencia, la cual contabilice el diagnóstico y saber la cantidad de tipo de diagnósticos que se presentan y el número de pacientes por enfermedad. En la Figura 17, se muestra la sentencia y el resultado.

```
sint=df.groupby('DIAG').count()
print(sint)
```

```
In [71]: runfile('C:/Users/HP/.spyder-py3/temp.py', wdir='
                          FA  SINT  ED  TRAT   S  ...
DIAG                                               ...
Amigdalitis aguda         13    13  13    13  13  ...
Bronconeumonia            13    13  13    13  13  ...
Bronquitis                12    12  12    12  12  ...
Bronquitis aguda          68    68  68    68  68  ...
Bronquitis aguda con COVID 19  12  12  12  12  12  ...
Bronquitis asmatica       30    30  30    30  30  ...
Bronquitis cronica        12    12  12    12  12  ...
COVID 19                  18    18  18    18  18  ...
Faringitis                12    12  12    12  12  ...
Faringitis aguda          50    50  50    50  50  ...
Faringitis amigdalitis aguda  36  36  36  36  36  ...
Fibrosis pulmonar          6     6   6     6   6  ...
Neumonia                   6     6   6     6   6  ...
Rinitis aguda             12    12  12    12  12  ...

[14 rows x 11 columns]
```

*Figura 17.*Cantidad de datos por tipo de Diagnostico.

Seguidamente, se tiene que realizar la conversión de los síntomas a booleanos, es decir presenta el síntoma o no. Cabe mencionar, para ello se requiere primero saber la cantidad de síntomas que se repiten para lo cual,

se utiliza el siguiente comando que se muestra en la Figura 18:

```
u=df['SINT'].value_counts()
print (u)
```

```
In [121]: runfile('C:/Users/HP/.spyder-py3/temp.py', wdir='C:/Users/HP/.spyder-py3')
Dolor abdominal tipo colico,Fiebre moderada,Malestar general
3
Fiebre recurrente,Irritabilidad al pasar agua
2
Dolor de garganta,Malestar general,Perdida del olfato
2
Malestar general,Fiebre moderada,Perdida de olfato
2
Dolor de garganta,Malestar general
1
Escalofrios,Malestar general,Perdida del olfato
1
Tos exigente,Sensacion de falta de aire,Malestar general
1
Malestar general,Dolor de garganta,Congestion nasal,Perdida del olfato
1
```

Figura 18. Datos únicos de los síntomas.

Luego de identificar que filas se repiten, se procedió a pasar los síntomas a una tabla para otorgarles un valor a cada uno de ello y luego proceder con la conversión de ellos a número enteros para poder realizar las conjugaciones. Ello se podrá observar en la Tabla 5.

Tabla 5. *Cuadro de síntomas con sus valores.*

VALOR	SINTOMA	VALOR	SINTOMA
1	Malestar general	20	Dolor de espalda
2	Dolor de garganta	21	Exudado faríngeo
3	Dolor faríngeo	22	Cefalea
4	Tos exigente	23	Perdida del olfato
5	Tos seca	24	Fiebre esporádica
6	Dificultad para caminar	25	Fiebre intensa
7	Escalofríos	26	Ronquido en el pecho
8	Vómitos	27	Sensación de olor térmica
9	Dolores musculares	28	Irritabilidad al pasar agua

10	Fiebre	29	Dolor torácico
11	Dificultad para respirar	30	Sensación de falta de aire
12	Perdida del gusto	31	Secuelas en los ojos
13	Dolor abdominal	32	Producción de mucosidad
14	Fiebre moderada	33	Sibilancias
15	Congestión nasal	34	Amígdalas rojas
16	Tos productiva	35	Voz rasposa
17	Sensación de flema	36	Tos con vomito
18	Fatiga	37	Dificultad para tragar
19	Fiebre recurrente	38	Dolor al tragar

Fuente: Elaboración propia.

Para empezar con la conversión, se determinó que son 38 síntomas únicos, los cuales se van a distribuir por columnas para verificar que enfermedad tienen el síntoma y cuál no, de esa manera se podrá determinar los síntomas que si o si debe presentar el paciente para saber el tipo de infección respiratoria aguda grave, para ello primero se deben digitar los 38 síntomas y si se cumple posicionar un número tal como se muestra en la Figura 19:

```
sint_={'Malestar general':1,'Dolor de garganta':2,'Dolor faringeo':3,'Tos exigente':4, 'Tos seca':5,
       'Dificultad para caminar':6, 'Escalofrios':7,'Vomitos':8,'Dolores musculares':9, 'Fiebre':10,
       'Dificultad para respirar':11, 'Perdida del gusto':12, 'Dolor abdominal':13, 'Fiebre moderada':14,
       'Congestion nasal':15, 'Tos productiva':16, 'Sensación de flema':17, 'Fatiga':18, 'Fiebre recurrent
       'Dolor de espalda':20,'Exudado faringeo':21,'Cefalea':22, 'Perdida del olfato':23, 'Fiebre esporadi
       'Fiebre intensa':25, 'Ronquido en el pecho':26, 'Sensacion de olor termica':27, 'Irritabilidad al p
       'Dolor toraxico':29, 'Tos con vomito':30, 'Sensacion de falta de aire':31, 'Secuelas en los ojos':3
       'Produccion de mucosidad':33, 'Sibilancias':34, 'Amigdalas rojas':35, 'Voz rasposa':36,'Dificultad
       'Dolor al tragar':38}
datos['SINT1']=datos['SINT1'].map(sint_)
datos['SINT2']=datos['SINT2'].map(sint_)
datos['SINT3']=datos['SINT3'].map(sint_)
datos['SINT4']=datos['SINT4'].map(sint_)
datos['SINT5']=datos['SINT5'].map(sint_)
datos['SINT6']=datos['SINT6'].map(sint_)
```

Figura 19. Conversión de síntomas.

En la Figura 20, se visualiza la conversión de los síntomas a números enteros. Los cuales, se han distribuido en columnas, dando un total de 6 para que se pueda conjugar con todos los registros que se presentan.

SINT1	SINT2	SINT3	SINT4	SINT5	SINT6
2	1	0	0	0	0
1	3	4	0	0	0
5	18	0	0	0	0
1	6	0	0	0	0
2	7	8	9	0	0
16	2	0	0	0	0
10	11	12	0	0	0
0	14	1	0	0	0
1	2	15	0	0	0
14	3	5	0	0	0
16	0	2	0	0	0
1	5	0	18	0	0
1	19	5	20	21	0
19	22	1	16	12	13
1	14	0	0	0	0
3	10	1	0	0	0

Figura 20.*Cuadro de columnas.*

Por otro lado, también se realiza la conversión del diagnóstico de la enfermedades, ya que son datos String y deben ser numéricos, para ello, se utilizó la siguiente sentencia, tal como se muestra en la Figura 21. Tal sentencia, permite la distribución por columnas de todos los valores únicos, por lo que, se valida entre 0 y 1, es decir True or False.

```
df=pd.get_dummies(df, columns=["DIAG"])
```

DIAG_Bronquitis	DIAG_Bronquitis aguda	DIAG_Bronquitis aguda con COVID 19	DIAG_Bronquitis asmatica	DIAG_COVID 19	DIAG_Faringitis	DIAG_Faringitis aguda
0	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	1
0	1	0	0	0	0	0
0	0	0	0	0	0	1
0	1	0	0	0	0	0
0	0	0	0	1	0	0
0	1	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	0
0	0	0	0	0	0	1
0	0	0	0	1	0	0
0	0	1	0	0	0	0
0	0	0	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	0

Figura 21. Imagen de conversión de columna diagnóstico.

Otras conjugaciones que se realiza es el Diagnostico con Temperatura para saber en qué enfermedad o enfermedades tiene un nivel alto de fiebre y la temperatura en la que se presenta mayormente el tipo de infección respiratoria, ello se observa en la Figura 22.

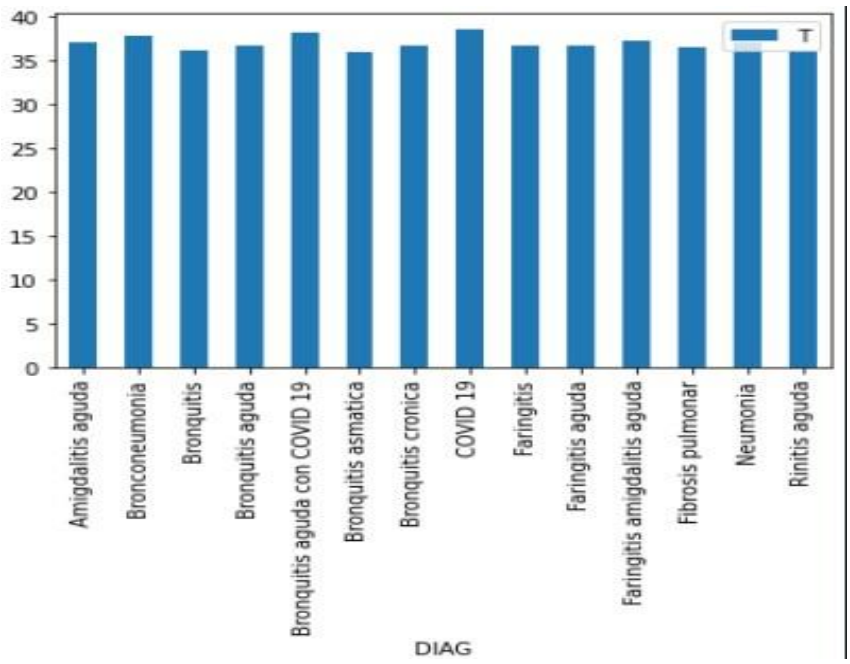


Figura 22. Gráfico de barras de la temperatura según diagnóstico.

En la Figura 22, el gráfico en el que se muestra el rango de temperaturas por diagnóstico, se puede deducir que es un síntoma relevante, ya que a diferencia de otras enfermedades respiratorias, en Bronconeumonía, Bronquitis aguda con Covid-19, Covid 19 y Neumonía, es notorio por que se presenta a partir de 38 °.

Por otro lado, se presenta la conjugación de Edad y Diagnóstico, con el cual se podrá ver el rango de edades frecuentes en el que se presentan las enfermedades. Es decir, hay enfermedades que acusan a las personas de tercera edad, tal es el caso de Faringitis aguda, Bronquitis aguda, COVID_19, Faringitis amigdalitis aguda, las cuales se presentan en personas con edades entre los 50 a 70 años de edad. Tal como se muestra en la Figura 23.

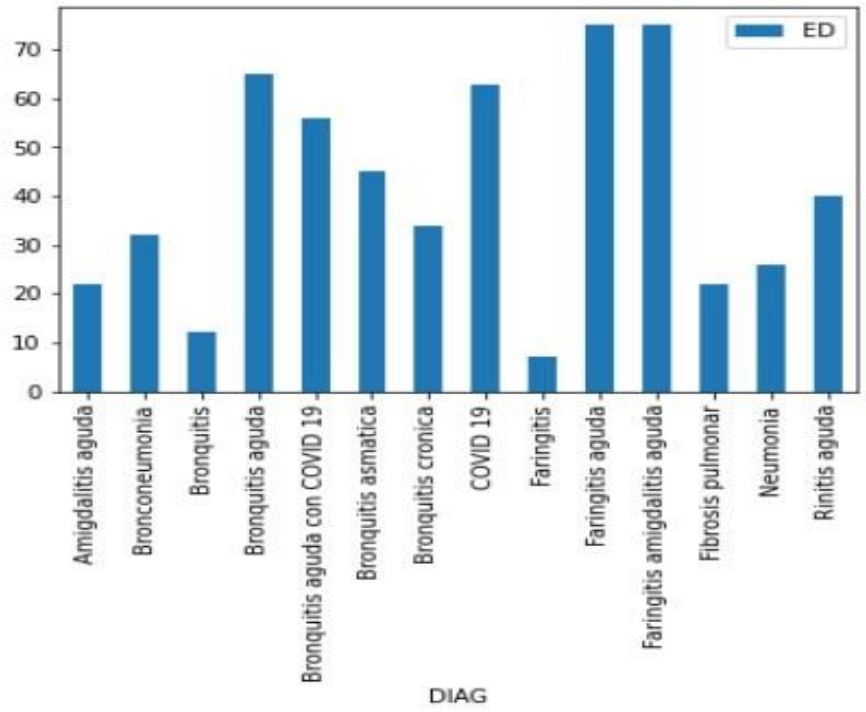


Figura 23. Gráfico de barras de rango de edades por diagnóstico.

Otro grafico estadístico que se puede mostrar es el de localidades, para lo cual, primero se contabiliza la cantidad de pacientes por localidad, tal como se muestra en la Figura 24. Luego para generar el gráfico, se agrega un plot, ello se podrá observar en la Figura 25.

```
f=df['PROC'].value_counts()
print(f)

In [176]: runfile('C:/Users/HP/.spyder-py3/t
Shilla      68
Toma        60
Carhuaz     48
Tingua      32
Marcara     32
Tinco       12
Mancos      12
Vicos       12
Malpaso      6
Maya         6
Anta         6
Nunocoto     6
Name: PROC, dtype: int64
```

Figura 24. Localidad de pacientes.

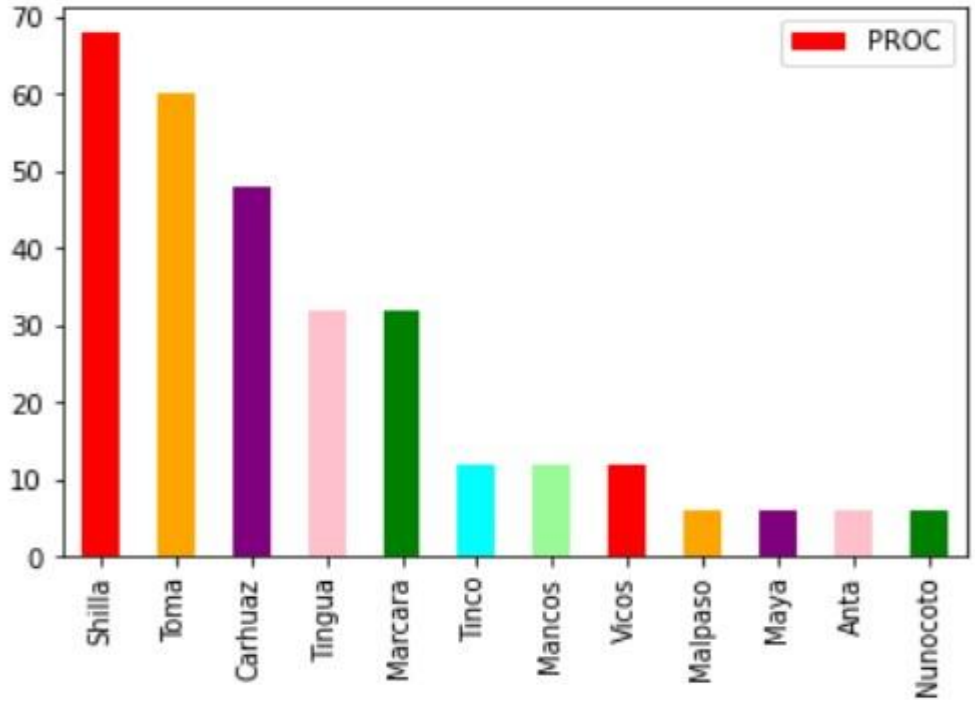


Figura 25. Gráfico de barras de localidad para saber cantidad de pacientes.

Posteriormente, luego de tener en cuenta cuales son los diferentes tipos de diagnósticos de infecciones respiratorias y localidades, se pasa a realizar conjugaciones de acuerdo al nombre del diagnóstico y su respectiva localidad, en la primera sentencia se contabiliza los datos. Ello se muestra en la Figura 26.

```
diag= list(zip(df['DIAG']=='Bronquitis aguda', df['DIAG']=='Faringitis aguda',
               df['PROC']=='Marcara',df['PROC']=='Shilla'))
a= pd.Series(diag).value_counts()
print(a)
```

```
In [178]: runfile('C:/Users/HP/.spyder-py3/tem
HP/.spyder-py3')
(False, False, False, False)    138
(True, False, False, False)      38
(False, False, False, True)      32
(False, True, False, False)      24
(True, False, False, True)       24
(False, True, True, False)       14
(False, True, False, True)       12
(False, False, True, False)      12
(True, False, True, False)        6
dtype: int64
```

Figura 26. Conjugación de datos.

En los resultados de la Figura 27, Se muestra True y False que significa si cumple o no cumple dicha sentencia, son 138 datos que no cumplen, 38 datos que cumplen con el Diagnostico de Bronquitis aguda, 32 que son de localidad de Shilla, 24 que tienen Faringitis aguda, 24 que cumplen con Diagnostico de Bronquitis aguda y con de la localidad de Shilla, 14 que provienen de Marcará, 12 que tienen Faringitis aguda y son de Shilla, 6 con Diagnostico de Bronquitis aguda y son de Marcará. Gráficamente, se muestra en la Figura 28.

```
diag= list(zip(df['DIAG']=='Faringitis', df['DIAG']=='Amigdalitis aguda',
               df['PROC']=='Marcara',df['PROC']=='Toma'))
a= pd.Series(diag).value_counts()
print(a)
In [181]: runfile('C:/Users/HP/.spyder-py3/temp.
py3')
(False, False, False, False)    189
(False, False, False, True)      54
(False, False, True, False)      32
(True, False, False, False)      12
(False, True, False, False)       7
(False, True, False, True)        6
dtype: int64
```

Figura 27. Agrupamiento de datos.

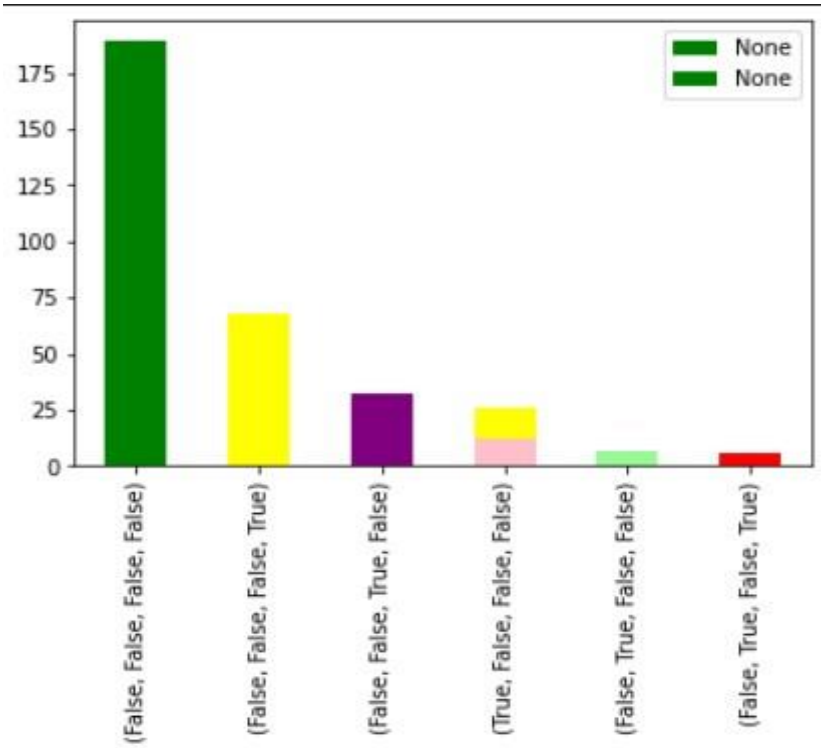


Figura 28. Gráfico de barras de cumplimiento.

En otro aspecto, se relaciona con un comando los datos numéricos para saber el máximo de los datos, tal como se muestra en Figura 29, la edad máxima de los pacientes es de 75 años y la temperatura máxima es de 38°, tal es el caso que en las Figuras anteriores, se pudo ver a detalle los rangos de cada columna.

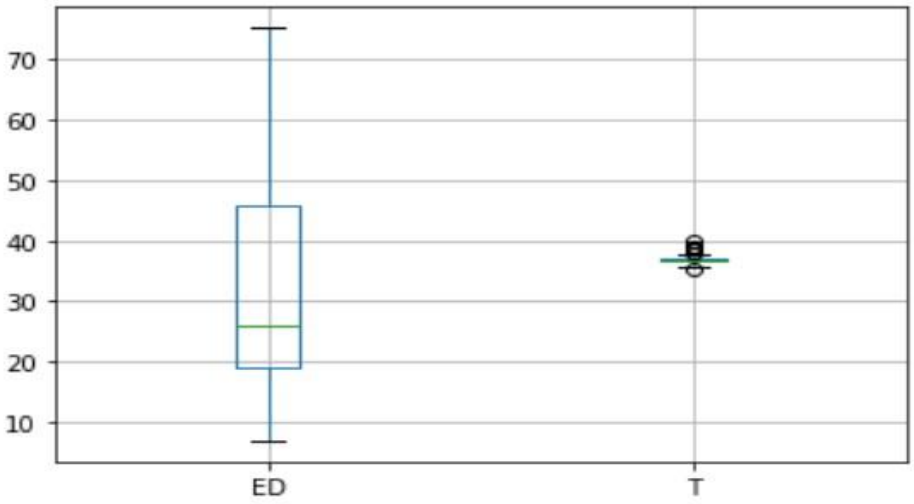


Figura 29. Máxima Edad y temperatura de los pacientes.

Luego de realizar las conjugaciones, se puede determinar que la temperatura y la edad son de importancia para saber qué tipo de enfermedad puede llegar a tener el paciente. De la misma manera, es primordial, usar los síntomas para que se pueda realizar un diagnóstico previo, es decir, se tienen que evaluar que síntomas son relevantes en las enfermedades, a la vez, evaluar la procedencia del paciente para saber si hay algún otro factor para una enfermedad determinada.

Cabe mencionar que para comenzar con el entrenamiento de los algoritmos que se van usar para el modelado, primero se tienen que pasar los datos que cuentan con dos respuestas a números binarios. Tal es el caso del sexo del paciente que M=Masculino y F= Femenino, para ello, se usa el siguiente comando que se muestra en la Figura 30. :

```
sex_={'M':1, 'F':0}
datos['S']=datos['S'].map(sex_)
print(datos)
```

Figura 30. Sentencia de conversión.

Después de realizar la conversión, se realiza consultas para saber que síntomas se cumplen por cada diagnóstico, para ello, se usó la sentencia

que se muestra en la Figura 31.

```
diag= list(zip(n['DIAG_Bronquitis'], n['PROC_Carhuaz'],n['SINT1']==2,n['SINT2']==3))
a= pd.Series(diag).value_counts()
print(a)
```

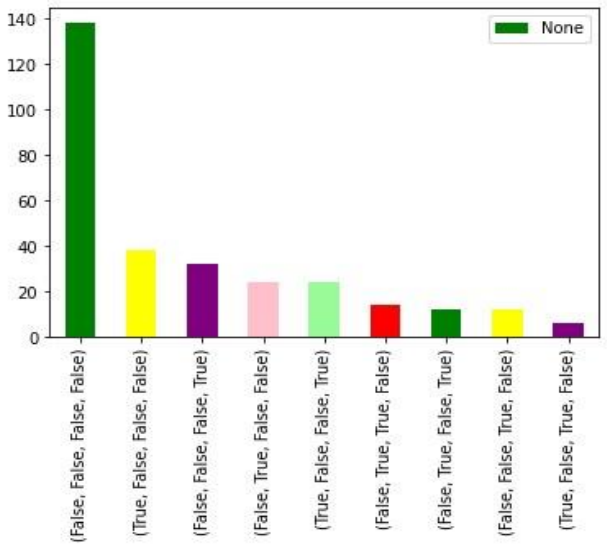


Figura 31. Gráfico de barras de True o False.

Continuando con las consultas para saber que síntomas se cumplen por cada diagnóstico, se siguieron con la sentencia, claro que modificando el valor del síntoma y el tipo de diagnóstico que se quería que cumpla, tal como se observa en la Figura 32.

```
diag= list(zip(n['DIAG_Bronquitis aguda'],n['SINT1']==2,n['SIN2']==8, n['SINT3']==10))
a= pd.Series(diag).value_counts().plot(kind='bar', legend='Reverse')
print(a)
```

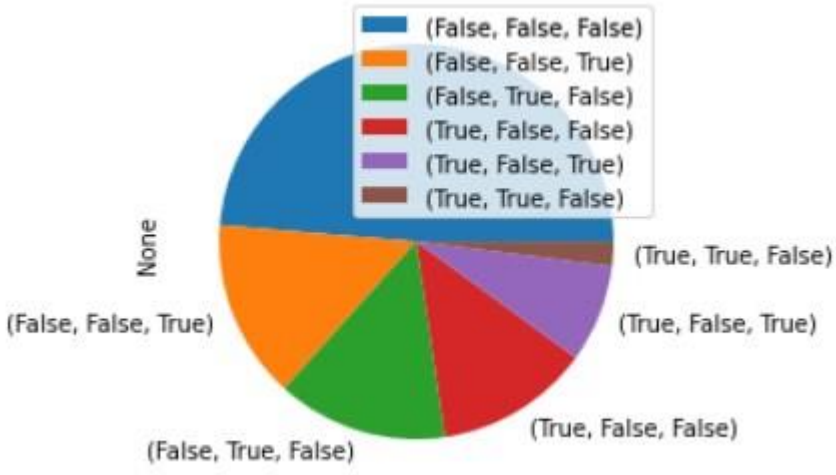


Figura 32. Grafico circular de True or False, de acuerdo a valores.

Viendo los anteriores gráficos con las sentencias, de esa manera se puede ir conjugando con las diferentes enfermedades y valores que tiene el registro del paciente. Otro ejemplo más en la Figura 33.

```
diag= list(zip(n['DIAG_Faringitis aguda'],n['SIN4']==5, n['SINT6']==15))
a= pd.Series(diag).value_counts().plot(kind='bar', legend='Reverse')
print(a)
```

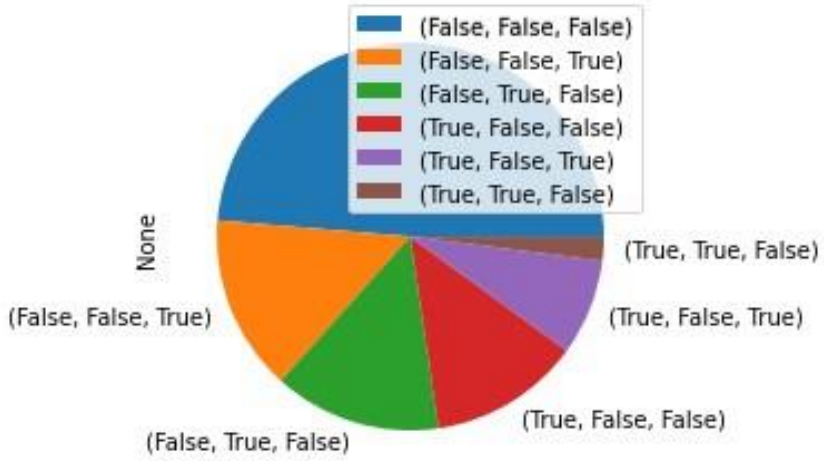


Figura 33. Gráfico circular de True or False para saber si se cumple la enfermedad de Faringitis.

Otro ejemplo es la conjugación del tipo de diagnóstico con el Sexo que está distribuido en 0 y 1, en donde 0 es Mujer y 1 Varón, ello se observa en la Figura 34, en el que se relaciona con diagnóstico de faringitis aguda y da un total de 10 personas que son mujeres y 10 varones.

```
diag= list(zip(n['DIAG_Faringitis aguda'],n['S']==0, n['S']==1))
a= pd.Series(diag).value_counts().plot(kind='bar', legend='Reverse')
print(a)
```

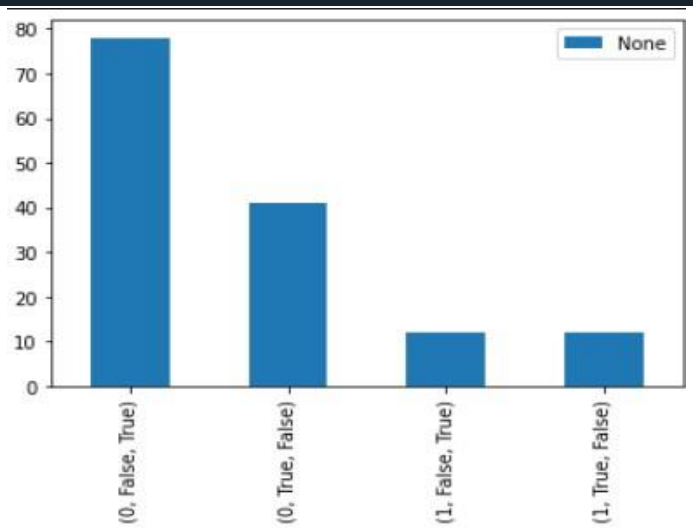


Figura 34. Sentencia y gráfico de barras de conjugación diagnóstico y sexo.

RESULTADOS DE LA METODOLOGÍA

4.4. Modelo

Para la fase 4 de nuestro proyecto, se consideró los algoritmos de árboles de decisión y clustering. La salida es la evaluación de los modelos en si donde se resumen los resultados de esta tarea, se enlista las calidades de los modelos generados y se clasifica su calidad en relación con cada otro en la tarea de construcción del modelo.

EXPERIMENTO I: MODELO USANDO ARBOLES DE DECISIÓN TIPO CART

A continuación, vamos a aplicar el algoritmo de árbol de decisión basado en clases, las cuales abarcan los síntomas y diagnóstico, es decir, los diagnósticos tiene valor de 0 y 1, los síntomas que están clasificados numéricamente desde el 1 al 38. La primera sentencia a usar, consiste en poner el intervalo de columnas que se va tener en cuenta para la evaluación, en este caso primero se realiza la consulta con el DIAG_Bronquitis, con las clases 13 y 4 que son los síntomas, al realizar la consulta, se va graficar un árbol de decisión para saber si se cumple o no uno de los valores para determinar que el paciente tiene Bronquitis. En la Figura 35, se muestra las sentencias del algoritmo.

```
x=n.iloc[:,0:24].values
y=n.iloc[:,299].values

x_train, x_test, y_train, y_test=train_test_split(x,y,test_size=0.4,random_state=0)
classifier=DecisionTreeClassifier(criterion='entropy',max_depth=4,random_state=0)
classifier.fit(x_train,y_train)
print(classifier)
y_pred=classifier.predict(x_test)
print(y_pred)
cm=confusion_matrix(y_test, y_pred)
print(cm)
print("Accuracy:", metrics.accuracy_score(y_test,y_pred))
print("El Score:", metrics.f1_score(y_test, y_pred, average='weighted'))
print("ROC:", metrics.roc_auc_score(y_test,y_pred))
print("Recall:", metrics.recall_score(y_test, y_pred, average='weighted'))
tree.export_graphviz(classifier,out_file='tree_social.dot')
dot_n=tree.export_graphviz(classifier,out_file=None, class_names=['13','4']
                           ,feature_names=list(n.drop(['DIAG_Bronquitis'],axis=1)),filled=True)
u=graph=pydotplus.graph_from_dot_data(dot_n)
graph.write_png('LI 7.png')

print(u)
```

Figura 35. Código de construcción del modelo.

En la Figura 36, se muestra la clasificación, en el que se lee de la siguiente manera, si el SINT 4 es menor a 8, en la columna SINT2, se observa que cumple con el valor 4, es decir tiene el síntoma Tos exigente y no se cumple el valor o clase 13, por lo que se determina que Bronquitis no tiene el síntoma de Dolor abdominal.

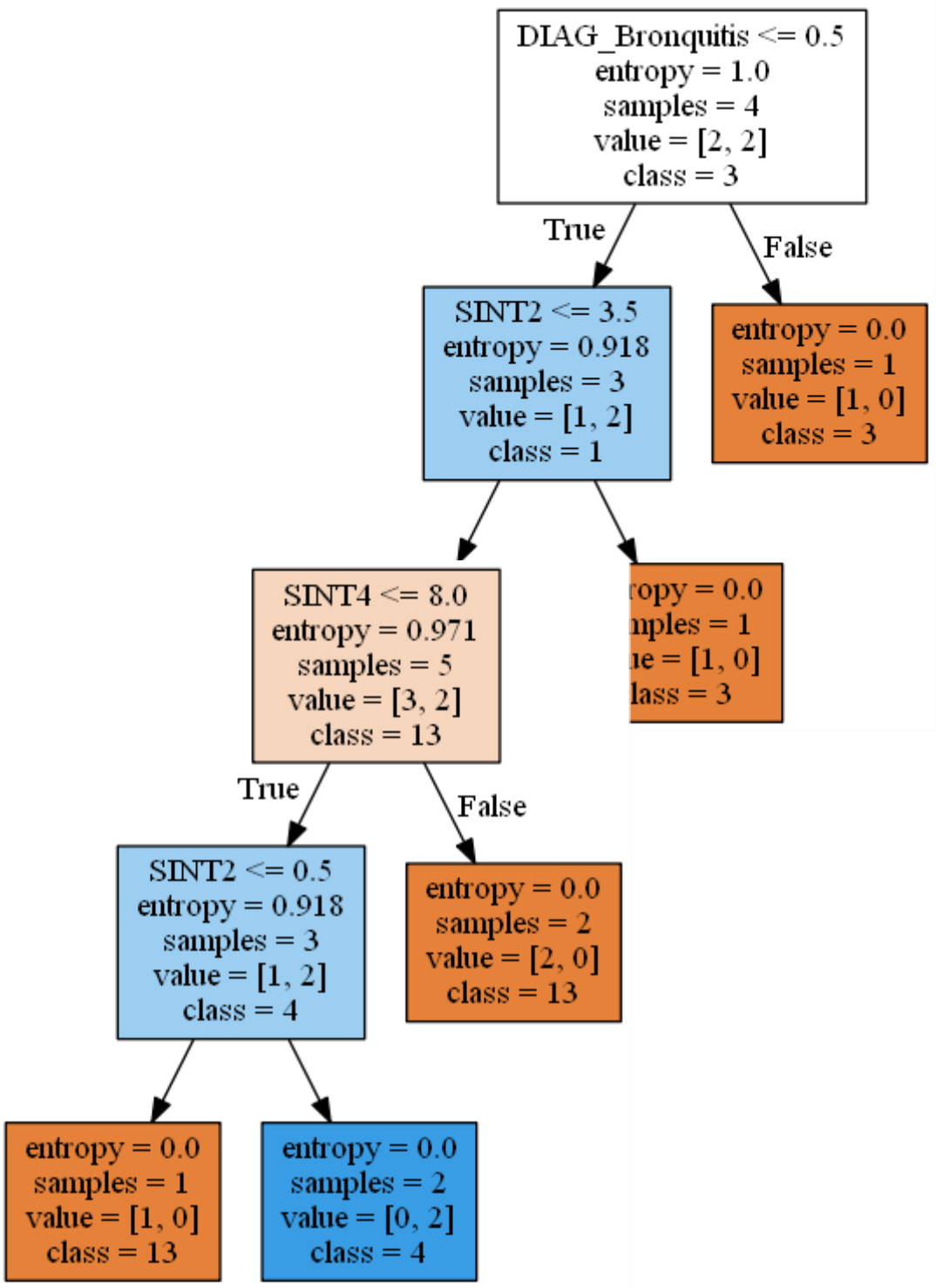


Figura 36. Árbol de decisión de Bronquitis.

Por consiguiente, estableciendo diferentes sentencias, es decir modificando la clase y el drop, el cual permite clasificar de manera adecuada el algoritmo, se va estableciendo según las conjugaciones que el usuario desee, es decir, puede poner el diagnostico con intervalo de edades para saber en qué edad se presenta dicha enfermedad. Ello, podremos observarlo en la Figura 37,

en la que se estableció como drop DIAG_ Bronquitis aguda con COVID 19 y las clases o valores de 18 y 20.

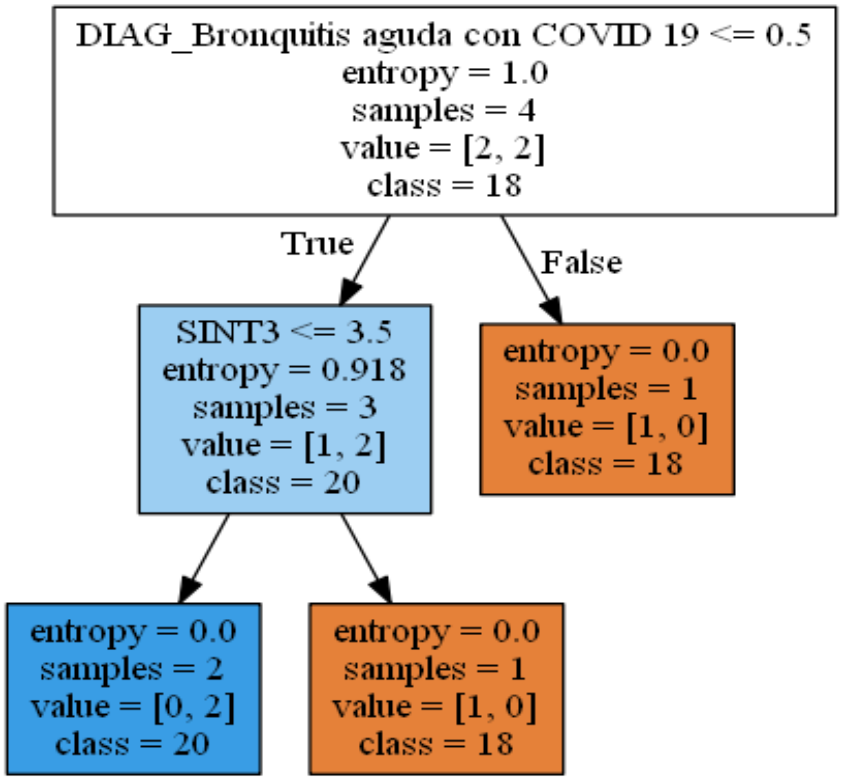


Figura 37. Árbol de decisión de Bronquitis aguda con COVID19, factor edades.

Por otro lado, evaluando el mismo diagnóstico, se toma en cuenta posibles síntomas, en el que se establece clases 1, 2, 3 para que proceda a evaluar, por lo que el árbol de decisión, muestra que en la columna SINT3, debes ser <=2.5 y si se cumple la clase 2, tiene el diagnóstico de Bronquitis aguda con COVID 19, ya que su entropía está al 0.918, es decir es una respuesta favorable.

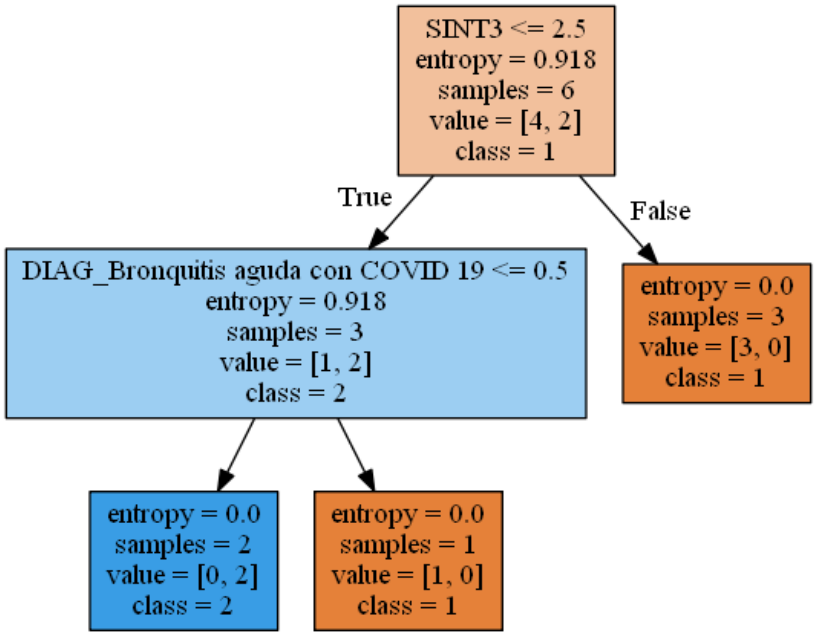


Figura 38. Árbol de decisión Bronquitis aguda con COVID 19, factor síntomas.

Otra conjugación que también se realizó, es el de Faringitis amigdalitis aguda con síntomas posibles, tal como se visualiza en la Figura 39, en el que se cumple el síntoma 13 Fiebre moderada con una entropía del 0.871 que es favorable para la asertividad del árbol de decisión.

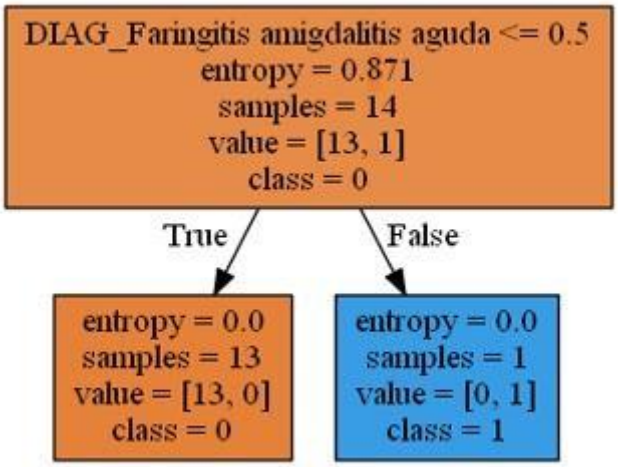


Figura 39. Árbol de decisión de Faringitis amigdalitis aguda.

Siguiendo conjugando diagnósticos y síntomas para que se construya el árbol de decisión, en la Figura se realiza la comparación de síntomas y se ve que la entropía es de 0.811, es decir es un buen árbol de clasificación.

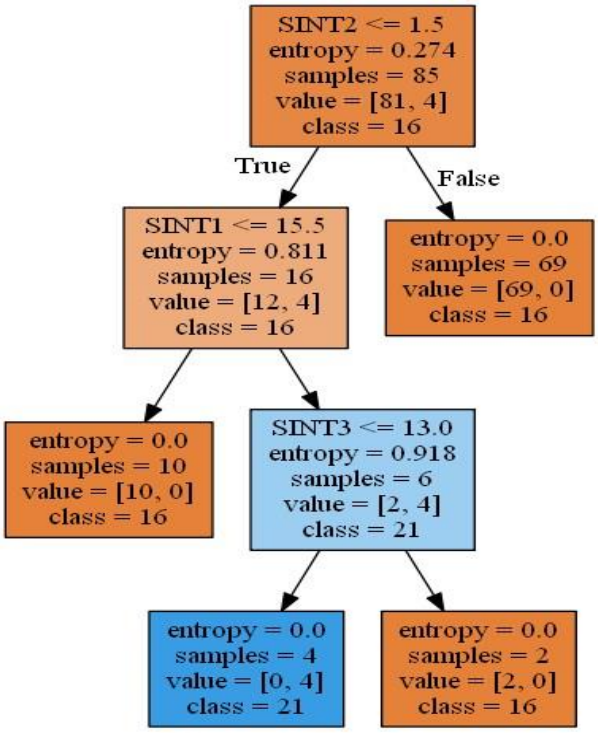


Figura 40. Árbol de decisión de Faringitis.

De acuerdo a los arboles, se puede ir determinando que sintomas debe cumplir dicha enfermedad para saber si el paciente tiene algun tipo de infección respiratoria aguda grave, por ejemplo con los arboles de la Figura 37 y Figura 38, en ellos, se enfatiza el Diagnostico de bronquitis y Bronquitis aguda con COVID 19, en el que nos explica lo siguiente:

Si una persona presenta tos exigente <=0.5 y tiene Malestar general<=3.5 Entonces el paciente tiene como diagnostico Bronquitis.

Si una persona presenta dolor de garganta, dolores musculares, tiene 20 años y temperatura 38 Entonces el paciente tiene como diagnostico Bronquitis aguda con COVID 19.

Si una persona presenta Dolor faríngeo <=0.3, tiene Dolor de garganta<=3.5 y Fiebre moderada Entonces el paciente tiene como diagnostico Faringitis amigdalitis aguda.

Si una persona presenta Dolor abdominal, perdida del gusto, Dificultad al tragar y temperatura 38 Entonces el paciente tiene como diagnostico

Faringitis.

EXPERIMENTO 2: CLASIFICACIÓN DE MODELOS USANDO CLUSTERS CON EL METODO K-MEANS

A continuación, vamos a aplicar el algoritmo de Clúster basado en clases, las cuales abarcan el análisis de todas las columnas. La primera sentencia a usar, consiste en poner el intervalo de columnas que se va tener en cuenta para la evaluación, luego de ello, se empieza a poner sentencias para crear el modelo de clustering, tal como se muestra en la Figura 32, luego se muestra el codo Jambú para saber el número de clúster, ello se ve en la Figura 39.

```
 x=n.iloc[:,0:24].values
 y=n.iloc[:,299].values
x_train, x_test, y_train, y_test=train_test_split(x,y,test_size=0.40,random_state=0)

print(x)
print(y)
wcss=[]
for i in range(0,20):
    kmeans=KMeans(n_clusters=i, max_iter=300)
    kmeans.fit(x,y)
    wcss.append(kmeans.inertia_)
plt.plot(range(0,20),wcss)
plt.title('Codo de Jambú')
plt.xlabel('Número de clusters')
plt.ylabel('wcss')
plt.show()

clustering=KMeans(n_clusters=4, max_iter=300)
clustering.fit(x,y)
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300, n_clusters=4, n_init=9,
       n_jobs=None, precompute_distances='auto', random_state=None, tol=0.0001, verbose=0)
n['KMeans_Clusters']=clustering.labels_
print(n.head())

fig=plt.figure(figsize=(6,6))
ax= fig.add_subplot(1,1,1)
ax.set_xlabel('Componente 1', fontsize=15)
ax.set_ylabel('Componente 2', fontsize=15)
ax.set_title('Componentes Principales',fontsize=25)
color_theme=np.array(['blue','yellow','green'])
ax.scatter(x=pca_nombres_diag.Componente_1, y=pca_nombres_diag.Componente_2,
           c=color_theme[pca_nombres_diag.KMeans_Clusters],s = 60)
plt.show()
print("Accuracy:", metrics.accuracy_score(y_test,y_pred))
print("El Score:", metrics.f1_score(y_test, y_pred, average='weighted'))
print("ROC:", metrics.roc_auc_score(y_test,y_pred))
print("Recall:", metrics.recall_score(y_test, y_pred, average='weighted'))
```

Figura 41. Sentencia para clustering.

El método K-MEANS, si bien es cierto, se encarga de agrupar las observaciones en K clusters distintos, los cuales se pueden observar que llegaron a un total de 22 K, de los cuales se cumplen 13, ello se puede determinar a la varianza que se da al crear el Codo de Jambú. Es decir, se

trata por lo tanto de un problema de optimización, en el que se reparten las observaciones en K clusters de forma que la suma de las varianzas internas de todos ellos sea lo menor posible. Para poder solucionar este problema es necesario definir un modo de cuantificar la varianza interna. Por lo que, se muestra el grafico, en el que se evaluó por componentes, el cual se muestra en la Figura 40. En ese caso, se evalúa por tipo de diagnóstico, es decir cada circulito equivale a los diferentes diagnósticos y cuanto estén más cerca, el grado de varianza aumenta.

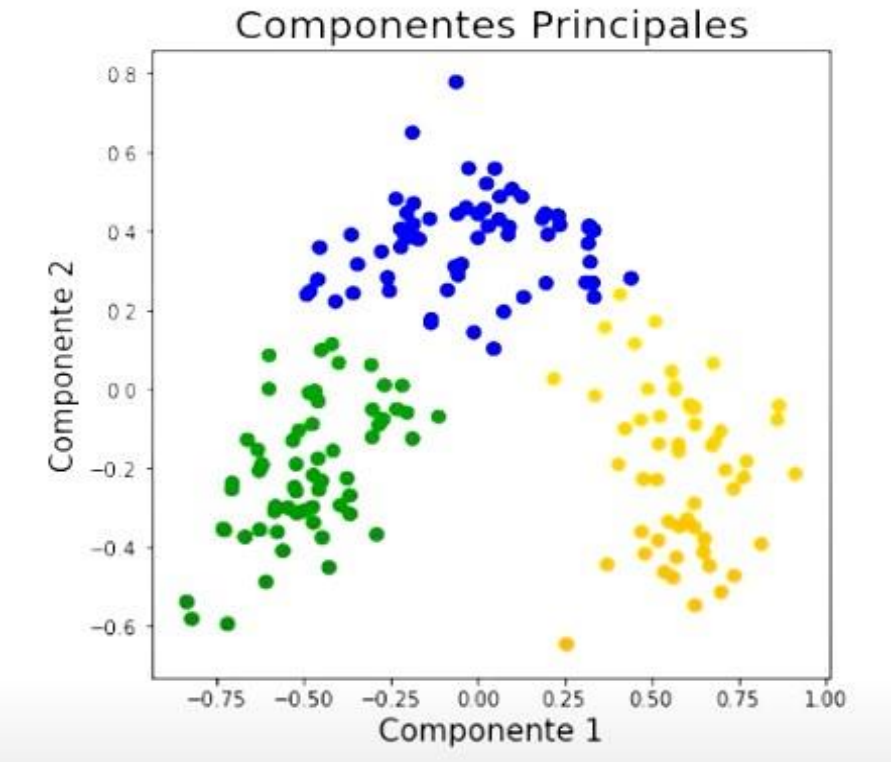


Figura 42. Modelo de clustering.

4.5. Evaluación del modelo

En relación con los árboles de decisión, se realiza la evaluación de los modelos para saber su grado de confiabilidad por cada uno de ellos, para lo cual se ejecuta una sentencia. Es la que se observa a continuación:

```
print("Accuracy:", metrics.accuracy_score(y_test,y_pred))
print("El Score:", metrics.f1_score(y_test, y_pred, average='weighted'))
print("ROC:", metrics.roc_auc_score(y_test,y_pred))
print("Recall:", metrics.recall_score(y_test, y_pred, average='weighted'))
```

```
DecisionTreeClassifier(criterion='entropy', max_depth=4,
random_state=0)
[0 1 0 1 1 0 0 0 1 1]
[[5 0]
 [0 5]]
Accuracy: 1.0
El Score: 1.0
ROC: 1.0
Recall: 1.0
<pydotplus.graphviz.Dot object at 0x0000023B8E6D5D60>
```

La confiabilidad se basa a las conjugaciones, por decir, en algunos casos, tendrá una obtención del 0.98 o 0.80, es decir se va mostrar así como se especifica en la Tabla 6.

Tabla 6. *Evaluación del modelo.*

ALGORITMO	*ACCURACY*	*ROC*	*SCORE*
ARBOL DE DECISIÓN	*Si es Faringitis amigdalitis aguda 0.811*	*0.811*	*0.811*
	Si es Bronquitis aguda con COVID 19 0.918	*0.918*	*0.918*

Fuente: Elaboración propia.

Se preguntarán porqué la varianza de los resultados, de acuerdo al estudio que se realizó, depende de la cantidad total de datos que se analicen, es decir cuantos más datos para realizar la predicción, será mayor el acierto.

El Codo de Jambú, muestra las altas y bajas de acuerdo a los clústeres, es decir, esté gráfico puede medir la acertibilidad del modelo de clústeres. En este caso si se llegan a cumplir todos los clústeres, el grado de acertibilidad

es mayor. A parte de ello, también, se puede generar la sentencia con el que se evaluó el modelo de árbol de decisión, ello se podrá ver en la Figura 43.

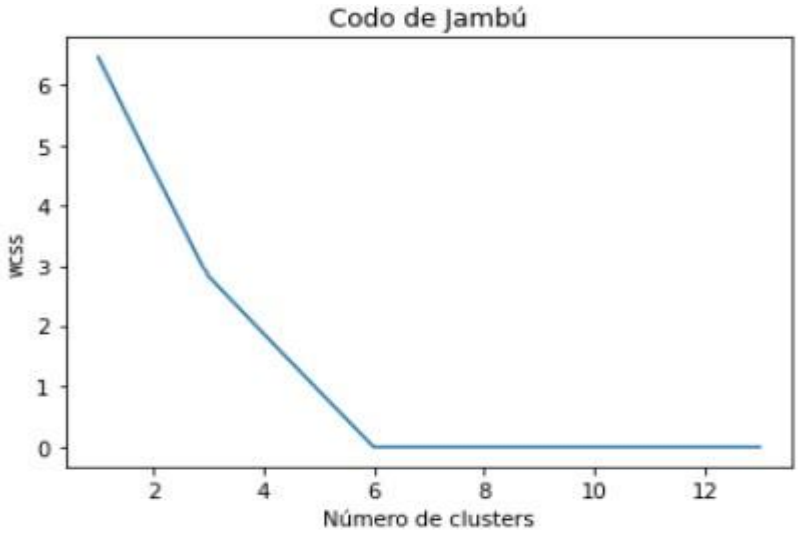


Figura 43. Codo de Jambú de clústeres.

En la Figura 44, se muestra la evaluación del modelo de clústeres, como se podrá observar el grado de acierto, es menor al de árbol de decisión, ya que tiene un Accuracy de 0.6, Score de 0,61, ROC de 0.61 y Recall 0.6, por lo que podemos ir concluyendo que el mejor algoritmo de aprendizaje es el Árbol de decisión.

```
n_clusters (20). Possibly due to duplicate points in X.
  kmeans.fit(x,y)
C:/Users/HP/.spyder-py3/temp.py:67: ConvergenceWarning:
Number of distinct clusters (15) found smaller than
n_clusters (21). Possibly due to duplicate points in X.
  kmeans.fit(x,y)
C:/Users/HP/.spyder-py3/temp.py:67: ConvergenceWarning:
Number of distinct clusters (15) found smaller than
n_clusters (22). Possibly due to duplicate points in X.
  kmeans.fit(x,y)
   ED  S     T  SINT1  ...  PROC_Tingua  PROC_Toma
PROC_Vicos  KMeans_Clusters
0  52  0  36.6    2.0  ...            0          0
0                2
1  10  0  36.4    1.0  ...            0          0
0                3
2  45  1  36.6    5.0  ...            1          0
0                1
3  65  1  36.5    1.0  ...            0          0
0                0
4  16  0  36.6    2.0  ...            0          0
0                1


Accuracy: 0.6
El Score: 0.6166666666666666
ROC: 0.619047619047619
Recall: 0.6
```

Figura 44. Evaluación del modelo clustering.

Para determinar con exactitud el grado de asertividad del algoritmo de aprendizaje árbol de decisión tipo CART, se realiza una matriz de confusión,

el cuál va permitir saber los negativos y positivos, dependiendo de la predicción que se realice, en la Figura 45, se puede observar la matriz que se crea entre 0 y 1, es decir se puede determinar cuántos cumplen con la predicción y cuantos no (positivo y negativo). Para ello, se realiza otras sentencias y sus resultados se pueden observar en la Figura 46. La matriz de confusión, también va de acuerdo a las predicciones que se usen, por ejemplo acá se empleó de acuerdo al árbol de decisión de la figura 40.

```
DecisionTreeClassifier(criterion='entropy', max_depth=4, random_state=0)
[0 0 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 1 0 1
 0 0 0 1 0 1 1 1 1 0 0 0 0 0 1 0 1 0 0 0 0]
```

Figura 45. Matriz de 0 y 1

Classes	0	1
Population	12	12
P: Condition positive	3	3
N: Condition negative	9	9
Test outcome positive	5	2
Test outcome negative	7	10
TP: True Positive	3	1
TN: True Negative	7	8
FP: False Positive	2	1
FN: False Negative	0	2
TPR: (Sensitivity, hit rate, recall)	1	0.3333333
TNR=SPC: (Specificity)	0.7777778	0.8888889
PPV: Pos Pred Value (Precision)	0.6	0.5
NPV: Neg Pred Value	1	0.8
FPR: False-out	0.2222222	0.1111111
FDR: False Discovery Rate	0.4	0.5
FNR: Miss Rate	0	0.6666667
ACC: Accuracy	0.8333333	0.75

Figura 46. Valores de la matriz de confusión.

4.6. Despliegue

De acuerdo a los modelos de algoritmo de aprendizaje, tanto del árbol de decisión y de clústeres, con las predicciones realizadas también con gráficas estadísticas de True or False, se puede construir un Api Rest, en el que se demuestre las conjugaciones que se realizó de manera dinámica, es decir primero empezamos con la creación de una interfaz en FIREBASE, luego se le puede ir alimentando desde la API o llenar los datos desde donde se van ir almacenando.

En la Figura 40, se muestra el primer paso a seguir para la construcción del API, primero se digita el código fuente, con el cuál el editor de códigos va poder interactuar con el FIREBASE.

```
export const environment = {
  production: false,
  firebaseConfig : {
    apiKey: "AIzaSyAop5-hzS0Jg4VOJ2xei8Xjs0eQ1hy2FlU",
    authDomain: "diagnostico-cfe06.firebaseapp.com",
    databaseURL: "https://diagnostico-cfe06-default-rtdb.firebaseio.com",
    projectId: "diagnostico-cfe06",
    storageBucket: "diagnostico-cfe06.appspot.com",
    messagingSenderId: "956517969934",
    appId: "1:956517969934:web:d8b603d0c9089f6070cc5f"
  }
};
```

Figura 47. Código de integración.

Seguidamente, establecemos los datos que van a ser ingresados desde el API. Es decir ponemos todos los campos posibles que pueden tener los pacientes y establecemos su tipo de dato. Ello se observa en la Figura 41.

```
export class Diagnostico {
    $keyRegistro: string;
    num_identificacion: string;
    diagnostico: string;
    sintoma1: string;
    sintoma2: string;
    sintoma3: string;
    sintoma4: string;
    sintoma5: string;
    sintoma6: string;
    temperatura: DoubleRange;
    edad: Int16Array;
    localidad: string;
    sexo: string;
}
```

Figura 48. Campos de predicción.

Seguido de ello, creamos una tabla, la cual será nuestra interfaz para que interactúe el usuario, es decir, se podrá realizar la búsqueda por síntomas, ahí también se establece los botones y actividades que podrá ejecutar. Ello se podrá observar en la Figura 42.

```
<br><br>
<div class="btn-group">
  <input type="text" class="form-control" id="buscar" placeholder="buscar diagnostico" [(ngModel)]='buscar'>
  <button class="btn btn-primary">consultar</button>
</div>

<br><br>
<table class="table table-bordered">

  <tr>
    <td> num_identificacion</td>
    <td>diagnostico</td>
    <td>sintoma1</td>
    <td>sintoma2</td>
    <td>sintoma3</td>
    <td>temperatura</td>
    <td>edad</td>
    <td>localidad</td>
    <td>sexo</td>
  </tr>
  <tr *ngFor="let diagnosticolista of diagnosticoList">
    <td>{{diagnosticolista.num_identificacion}}</td>
    <td>{{diagnosticolista.diagnostico}}</td>
    <td>{{diagnosticolista.sintoma1}}</td>
    <td>{{diagnosticolista.sintoma2}}</td>
    <td>{{diagnosticolista.sintoma3}}</td>
    <td>{{diagnosticolista.temperatura}}</td>
    <td>{{diagnosticolista.edad}}</td>
    <td>{{diagnosticolista.localidad}}</td>
    <td>{{diagnosticolista.sexo}}</td>
    <td>
      <button type="button" (click)="onEdit(diagnosticolista)" class="btn btn-outline-danger btn-sm">
        Edit
      </button>
    </td>
    <td>
      <button type="button" (click)="onDelete(diagnosticolista.$keyRegistro)" class="btn btn-outline-danger btn-sm">
        Delete
      </button>
    </td>
  </tr>

</table>
```

Figura 49. Creación de tabla.

Luego de ello, se crea el formulario con el que se va interactuar, es decir, se va posicionando que datos van ingresar. Mira la Figura 43.

```
<br>
<form #diagnosticoForm="ngForm">

  <input type="hidden" name="$keyRegistro" #$keyRegistro="ngModel"
    [(ngModel)]="diagnosticoService.selectedDiagnostico.$keyRegistro">

  <label for="num_identificacion"><b>num_identificacion*</b></label>

  <input class="form-control" id="num_identificacion" name="num_identificacion" #num_identificacion="ngModel"
    [(ngModel)]="diagnosticoService.selectedDiagnostico.num_identificacion"placeholder="num_identificacion">

  <br>
  <label for="nombre"><b>nombre*</b></label>
  <input type="text" class="form-control" name="nombre" id="nombre" #nombre="ngModel"
    [(ngModel)]="diagnosticoService.selectedDiagnostico.nombre">
  <br>
```

Figura 50. Creación de formulario.

Al finalizar con toda la programación, se puede ir almacenando datos, es decir vamos alimentando a nuestro API, tal como se muestra en la Figura 44.

```
"Diagnosticos": {
    "001":{
        "Diagnostico":"Bronquitis","sintoma1":"tos","sintoma2":"Dolor de garganta","sintoma3":"Dolor faringeo",
        "temperatura":"38","edad":"25 > 50","localidad":"Shilla","sexo":"M"
    },
    "002":{
        "Diagnostico:":"Faringitis","sintoma1":"Malestar general","sintoma2":"Fiebre moderada","sintoma3":"Dolor faringeo",
        "sintoma4":"Tos seca","temperatura":"37","edad":"30>70","localidad":"Shilla","sexo":"F"
    },
    "003":{
        "Diagnostico:":"Faringitis amigdalitis aguda","sintoma1":"tos","sintoma2":"Dolor de garganta","sintoma3":"Dolor farin
        "temperatura":"38", "edad":"5>35","localidad":"Carhuaz","sexo":"M"
    },
    "004":{
        "Diagnostico":"Bronquitis aguda:","sintoma1":"Malestar general","sintoma2":"Tos productiva","sintoma3":"Dolor de garg
        "temperatura":"37","edad":"18<30","localidad":"Carhuaz","sexo":"M"
    },"005":{
    "Diagnostico":"COVID-19:", "sintoma1":"Fiebre", "sintoma2":"Dificultad para respirar", "sintoma3":"Perdida del gusto"
```


Figura 51. Ingresando datos.

Para interactuar con el API, el diseño de la interfaz es la siguiente, tal como se muestra en la Figura, se podrá editar, eliminar y realizar la búsqueda por síntomas, temperatura, edad, sexo para que de esa manera solo muestre el diagnostico que da con esos factores, en caso no se cumpla con ninguno de los campos, se mostrará toda la lista, en este caso, vamos interactuar con un total de 15 enfermedades y 38 síntomas que son los valores únicos que se han podido recopilar.

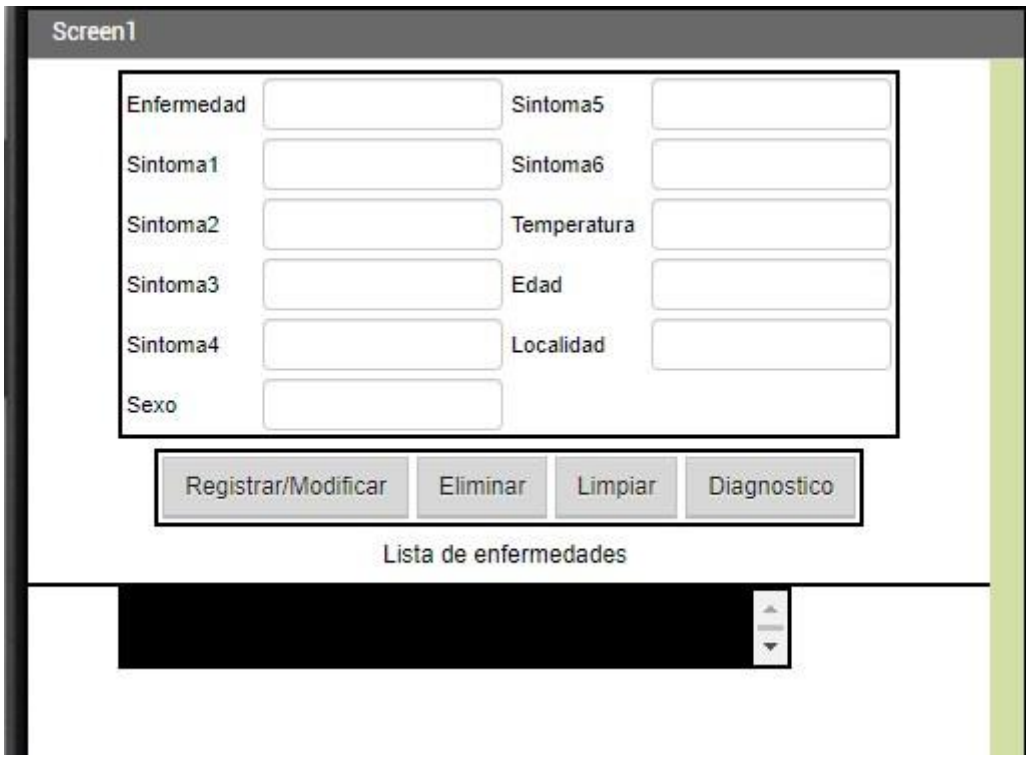


Figura 52. Interfaz del API.

V. DISCUSIÓN

Govindasamya y Velmuruganb (2017), en su trabajo “A Study on Classification and Clustering Data Mining Algorithms based on Students Academic

Performance Prediction”, se centraron, en el uso de técnicas con dominio educativo, en el que el objeto de análisis fueron los estudiantes de titulación UG y PG, utilizando algunos de los algoritmos de clasificación y agrupamiento en minería de datos. Asimismo, recopilaron de 4 universidades las características de la información de todos los estudiantes en sus exámenes de fin de semestre. Para su estudio, evaluaron algoritmos de predicción de rendimiento académico, tales como la maximización de expectativas(EM) y el algoritmo K-Means, en el que su precisión del segundo algoritmo es alta, ya el tiempo que tarda en construir el modelo es de 0,1 segundos, con una asertividad del 83 %.

Kumar (2018), en su artículo de investigación “Data Mining: A prediction for performance improvement using classification” desarrolló un modelo, el cual contó con datos ocultos que le ayudaron a mejorar el desempeño de los estudiantes, ya que se centró en técnicas de clasificación, con el fin de medir e identificar el rendimiento de los estudiantes con modalidad superior de la India. Para ello, usó una base de datos, que contaba con información de los estudiantes, los cuales fueron procesados en términos de completar los valores perdidos o nulos, transformar los valores, a tal caso que se seleccione por atributos y reconocer las variables relevantes. Cabe recalcar que para la construcción del modelo, empleó el método de clasificación tipo árbol, en el que si los valores daban mayor a 0.50 se les establecía como un alto factor de probabilidad para identificar el nivel del estudiante.

Torres y Diaz (2018), en su trabajo “Técnicas de inferencias, predicción y minería de datos” se enfocaron en el uso de la técnica de clasificación tipo árbol de decisión, en el que se evaluó desde su porcentaje de error que llegó a tener un 9.13% de acuerdo a las pruebas realizadas, es decir, llegaron a obtener un gran resultado en cuanto a la elaboración del modelo, ya que se implementaron técnicas, las cuales ayudaron a identificar patrones que se encarguen de la

predicción académica de los estudiantes para saber si aprobaron o desaprobaron.

Flores (2021), en su trabajo de investigación “Generación de árboles de decisión usando un algoritmo inspirado en la Física” desarrolló un modelo basado en arboles de decisión para predicción y clasificación, con el tipo de algoritmo centro evolutivo (ECA) [44] para inducir nodos internos que se distribuyeron linealmente para identificar sus valores internos como medida de su valor de aptitud y obtenga arboles cercanos al optimo ECA. El mejor árbol de decisión en la población final será refinado, reemplazando algunos nodos hoja con sub árboles para mejorar el desempeño del árbol, y posteriormente podado con el objetivo de reducir su tamaño. Los resultados obtenidos en esta investigación han resultado ser competitivos. La mayoría de los conjuntos de datos que se han utilizado para inducirlos, mejoran tanto en el tamaño, como en el porcentaje de clasificación, pues se han comparado con el método tradicional que se tienen como referente en la literatura para la inducción j48

[79]. Así mismo, ha mejorado algunos resultados obtenidos con el método EDADTSP V.

VI. CONCLUSIONES

De acuerdo a los resultados obtenidos durante el desarrollo de la investigación se llegó a las siguientes conclusiones, los cuales se determinaron en función de los objetivos específicos propuestos:

Objetivo Especifico 1: Identificar algoritmos de aprendizaje para determinar los síntomas más relevantes en pacientes con infecciones respiratorias agudas graves.

Conclusión: En esta investigación, se evalúo la aplicación de la minería de datos para el diagnóstico previo de las infecciones respiratorias en el Policlinico Cruz Verde, se aplicó algoritmos de clasificación de Arboles de decisión tipo CART y clustering tipo K-MEANS para resolver los objetivos trazados, el algoritmos de árboles de decisión fue el que tuvo mayor asertividad en los requerimientos del negocio, ya que permite clasificar los datos en falso o verdadero para saber que sentencias se cumplen y de cuanto es su entropía. Con el algoritmo de árboles de decisión, se analizó y se identificó el patrón de comportamiento del paciente con diagnóstico de algún tipo de infección respiratoria.

Objetivo Específico 2: Identificar las características de los pacientes que presentan infecciones respiratorias agudas graves según diagnóstico.

Conclusión: Se logró identificar las características al momento de construir y evaluar los modelos, luego de ello, se pudo construir un API REST, en el que digitas los síntomas, su temperatura, su edad y su sexo, me daría como diagnostico la posible enfermedad que tiene el paciente.

VII. RECOMENDACIONES

Se recomienda seguir contribuyendo e inculcando la investigación en cuanto al tema de minería de datos, ya que, es un campo muy importante de estudio que va tomando mucha importancia, en especial en el ámbito de la salud por lo que es una fuente en la que se puede conseguir, obtener o construir una gran base de datos de información. Cabe mencionar, el análisis de los algoritmos utilizados o empleados para construcción de modelos resultan muy interesantes en la etapa de construcción y evaluación, se preguntarán el motivo, puesto que, los algoritmo que se manejan permiten crear grandes aplicativos con inteligencia artificial, es decir, su resultado es a futuro. En el año 2017, se logró catalogar por la organización mundial de aduanas (WCO-World Customs Organization), que sea el año del análisis de datos, es decir su comienzo. Por otro lado, por un estudio de IBM Big Data & analyticz, en ese mismo año, hicieron que el Big Data y el análisis empezaría lleguen a ser tendencia.

Ante lo mencionado, como segunda recomendación, se debería de aplicar la investigación de Data Mining en el campo de Salud, siendo una de las formas o maneras que puedan contribuir en la detección o diagnóstico de nuevas enfermedad a base de síntomas de pacientes.

Asimismo, se debería considerar el diseño e implementación del modelo de clasificación, contando con interfaces más didácticas para que al personal, se le sea más sencillo el uso de la herramienta, es decir se puede distribuir por módulos e ir alimentando la base de datos con todas las variables de análisis para el Policlinico Cruz Verde, especialmente para los casos nuevos de infecciones respiratorias agudas graves ir agregando.

VIII. REFERENCIAS

Data Mining in Medicine. Zhang, L. 4 de Junio 2021. Journal of Healthcare Engineering.

Disponible en: https://www.hindawi.com/journals/jhe/si/875812/?utm_source=google&utm_medium=cpc&utm_capaign=HDW_MRKT_GBL_SUB_ADWO_PAI_DYNA_JOUR_X&gclid=Cj0KCQjwk4yGBhDQARIsAGfAeuDxeM7Soztk6HkuhTpsGdbYFdlIB-baBcJhu4y726vQzLjzQGAidcaAkMKEALw_wcB

Aplicación de minería de datos a información de pacientes pre diabéticos. 2017. Hernande, H. Universidad Juárez Autónoma de Tabasco.

Disponible en: https://www.researchgate.net/publication/ Aplicacion _de_mineria_de_datos_a_informacion_de_pacientes_prediabeticos

Medina,R y Gomez, C. Funcionalidades de la minería de datos. 31 de marzo 2016. Colombia.

Disponible en: https://www.researchgate.net/320219995_Funcionalidades _de_la_mineria_de_datos

Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia. Rojas, E y Aguilar, J. 2017 (pp.80). Disponible en: https://repository.ucatolica.edu.co/bitstream/10983/15329/1/Trabajo%20de%20grado.pdf

Minería de datos para mejorar el diagnóstico de la tuberculosis pulmonar en un hospital. Huerta, J. 2016 [Fecha de consulta: 28 de abril 2021].

Disponible en: https://repositorio.ucv.edu.pe/bitstream/handle/20.500.12692/34401/Huert a_CJ.pdf?sequence=1&isAllowe

Minería de datos. Belinchón, Y. Universidad Carlos III de Madrid: España, 2017. Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/10-11/15mem.pdf

Workshop on Data Mining for Medicine and Healthcare.Hilton, M. Publicado en Mayo 7 del 2016.
Disponible en: http://www.dmmh.org/sdm16

Data Mining and Medical Research Studies. Khajehei, M. Westmead. Hospital.2017.
Disponible en: https://www.researchgate.net/publication/232652822_Data_Mining_and_Medical_Research_Studies

Mineria de datos en medicina: Z. Yang, et.al.2019 [Fecha de consulta: 18 de Mayo 2021].

Disponible en: https://www.semanticscholar.org/paper/Software-Defined-WideArea-Network-(MineriaDatos)%3A-and-Yang-Cui/372c09f076334849284cbf739d6662deb9d470cd

Implementations Data Mining: Godeychik,S , Kolegov, D. 2018 [Fecha de consulta: 20 de Abril 2021].

Disponible en: https://www.researchgate.net/publication/336846762_Practical_Implement_Mining

How to make Data Mining in Medicine: Wood Richard, 2017.Francia.
Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1353485?via%3Dihub

Application of data mining techniques in service primary health (phc) G. Mine, J. Hai y L. Jin y Z. Huiying. 2020 [Fecha de consulta: 20 de Abril 2021].
Disponible en: https://ieeexplore.ieee.org/document/9240699930

Modelos de datos en salud: clasificación por arquitectura y usabilidad. Pedrera, M y Serrano, P. 2020 [Fecha de consulta: 16 de Mayo 2021]

Disponible en: https://www.researchgate.net/publication/342888010_Modelos_de_datos_en_salud_clasificacion_por_arquitectura_y_usabilidad

Minería de datos de la salud: Sistema de votación de técnicas analíticas para identificar los factores que influyen en la realización de cirugías estéticas. 2017.

Revista Politécnica. Oviedo, A.
Disponible en: https://www.researchgate.net/publication/331254746_Mineria_de_datos_de_la_salud_Sistema_de_votacion_de_tecnicas_analiticas_para_identificar_los_factores_que_influyen_en_la_realizacion_de_cirugias_esteticas

Sosa, M. Trazabilidad de datos en salud. Medellin 2016. Universidad Pontificia Bolivariana. Publicado en Abril 2017.

Disponible en: https://www.researchgate.net/publication/328006477_Trazabilidad_de_datos_en_salud_Medellin_2016

Manejo clínico de la infección respiratoria aguda grave (IRAG) en caso de sospecha de COVID-19. Organización Mundial de Salud. 13 de Marzo 2020.
Disponible en: https://apps.who.int/iris/bitstream/handle/10665/331660/WHO2019-nCoV-clinical-2020.4-spa.pdf

Etiología viral de las infecciones respiratorias agudas graves en una unidad de cuidados intensivos pediátricos. Rev Peru Med Exp Salud Publica 36 (2)

Disponible en: https://scielosp.org/article/rpmesp/2019.v36n2/231-238

Trazabilidad de datos en salud. Medellín. 2016.
Disponible en: https://www.researchgate.net/publication/328006477_Trazabilidad_de_datos_en_salud_Medellin_2016

Data mining in medicine. Italia. Ramírez, A y Álvarez, R. 2020
Disponible en: https://repository.upb.edu.co/bitstream/handle/20.500.11912 /3299/MINER%C3%8DA%20DE%20DATOS%20DE%20LA%20SLUD.pdf?sequence=1

Fases de la metodología de minería de datos. Universidad la católica.2020

Disponible en: https://repository.ucatolica.edu.co/bitstream/10983/15329/1/Trabajo%20de%20grado.pdf

Mineria de datos en medicina: Galan, V, et.al.2019 [Fecha de consulta: 25 de junio 2021].

Disponible en: https://e-archivo.uc3m.es/bitstream/handle/10016/22198/PFC_Victor_Galan_Cortina.pdf

Application of data mining techniques in service primary health (phc) G. Mine, J. Hai y L. Jin y Z. Huiying. 2020 [Fecha de consulta: 20 de Abril 2021].
Disponible en: https://ieeexplore.ieee.org/document/9240699930

Data Mining in Medicine. Uvidia, A. 24 de Junio 2021. Journal of Healthcare Engineering.

Disponible en: Dialnet-MineriaDeDatosParaLaTomaDeDecisionesEnLaUnidadDeNi-6836545.pdf

Mineria de datos. Belinhón, Y, 2019.
Disponible en: http://www.it.uc3m.es/jvillena/irc/practicas/10-11/15mem.pdf

Implementations Data Mining: Godeychik,S , Kolegov, D. 2018 [Fecha de consulta: 20 de Abril 2021].

Disponible en: https://www.researchgate.net/publication/336846762_Practical_Implement_Mining

ANEXOS

ANEXO 1. Carta de aceptación

POLICLÍNICO CRUZ VERDE

CARTA N° 002-2021-P*CV/GA*

De: Dr. Julio Vilca Bagazo Gerente
General del Policlinico Cruz Verde

A: Rossmery Quijano Nayhua
Estudiante de EP Ingeniería de Sistemas - Lima Norte Universidad César Vallejo SAC

ASUNTO: Aceptación para su Proyecto de Tesis

FECHA: 23 de Setiembre del 2021

Por medio de la presente se le hace llegar nuestro saludo cordial, y a la vez comunicarle que como Empresa de Servicios Médicos, aceptamos participar en su proyecto de "Modelo de clasificación basado en minería de datos para la identificación de factores que influyen en las infecciones respiratorias agudas graves de pacientes ", nos encargaremos de brindarle los documentos que requiera, con la finalidad que nos ayude en la actualización de conocimientos de nuestro personal.

Para ello le hacemos llegar la presente CARTA DE ACEPTACIÓN en conformidad.

Atentamente;

POLICLINICO
Julio E. Vilca Bagazo
GERENTE

ANEXO 2. Historia clínica de paciente

ANEXO 3. Evidencias de recolección de información