ANÁLISIS DE AGRUPAMIENTO DE DATOS PARA EL APOYO AL DIAGNÓSTICO DE LA TUBERCULOSIS

KARINA ALEJANDRA ORTIZ NEIRA

Trabajo Dirigido

Tutor
Ing Álvaro David Orjuela Cañon Ph.D

UNIVERSIDAD DEL ROSARIO
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
PROGRAMA DE INGENIERÍA BIOMÉDICA
BOGOTÁ D.C
2016

AGRADECIMIENTOS

En primer lugar expresó puntualmente mi agradecimiento con Dios, padre todo poderoso, por guiarme en cada paso de la vida, por brindarme la sabiduría en cada decisión tomada, además por permitirme adquirir una formación académica, adecuada y completa para que a futuro logre ejecutar las labores propias de ingenia biomédica a nivel profesional.

Seguidamente fijo mis agradecimientos al director de la presente tesis de pregrado, ingeniero Álvaro David Orjuela Cañón, quien con su dedicación, apoyo y confianza, forjó en mí el compromiso, la disciplina y las actitudes necesarias para culminar la misión encomendada, igualmente agradezco por su respeto, sus ideas y sus sugerencias en la ardua labor de esta tutoría, así mismo reitero una vez más mi gratificación por el tiempo que reservo para orientarme, guiarme e implementar en mi un desarrollo cognitivo a nivel profesional, por tal trabajo siempre quedare en deuda y agradeceré por su amabilidad y comprensión.

Finalmente agradezco a mi familia, padres y hermanos por su apoyo moral, económico y su amor incondicional, también por el esfuerzo, la lucha constante, la misión de guiarme e infundir en mí el ejemplo correcto y preciso para cumplir cada uno de los objetivos propuestos no solo en el ámbito profesional, sino en la familiar, social y religioso, además de su constante compañía en momentos de debilidad angustia o temor por desistir, sin antes dar la lucha para llegar a la meta con una actitud siempre triunfante.

Por último es de conocimiento público que la docencia desempeña un gran papel en la vida del ser humano, por ello reconozco un agradecimiento a todos y cada uno de los docentes de la universidad del Rosario y de la Escuela Colombiana de Ingeniería Julio Garavito del programa de ingeniería biomédica por su apoyo total y por brindarme semestre a semestre la cognición necesaria para desenvolverme en la realidad.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

La tuberculosis (TB) es una enfermedad infecciosa considerada de emergencia mundial por la Organización Mundial de la Salud (OMS) porque se encuentra entre las diez principales causas de muerte por infección cada año [1]. Esta enfermedad es causada por una bacteria llamada "*Mycobacterium tuberculosis*" que por lo general suele atacar los pulmones, aunque también puede afectar otras partes de cuerpo. La TB, se contagia principalmente por el aire, ya que cuando una persona con TB estornuda expulsa la bacteria [2].

Según reporte de la OMS, con fecha, del 26 de Septiembre del año 2018, la meta 3.3 de los objetivos de desarrollo sostenible (ODS) incluye el objetivo de poner fin a la epidemia de tuberculosis para la vigencia 2030. Esto, teniendo en cuenta que cada año la epidemia de la TB ha ido aumentando, mostrando registros de hasta 1,5 millones de muertes hasta el año. Al mismo tiempo, cabe mencionar que las personas que padecen trastornos que perjudican al sistema inmunitario, son las más propensas a desarrollar la enfermedad, con trastornos como: tabaquismo, abuso del alcohol, VIH y desnutrición, entre otros. De los 10 millones de personas que se enfermaron de TB en 2018, se atribuyen 2.3 millones a la desnutrición, 0.83 millones al abuso de alcohol y 0.86 millones al tabaquismo; así mismo la OMS, asocia más del 95% de los casos reportados a países en desarrollo, en la región de Asia Sudoriental se presentaron 44%, en África, 24% y finalmente en el Pacifico Occidental 18% de los nuevos casos reportados en este mismo año [3].

En el año 2018, el Ministerio de Salud de Colombia, reportó 14338 casos de TB en el país, alcanzando una tasa de incidencia de 26 casos por cada cien mil habitantes. El comportamiento demográfico y social también fue analizado, donde personas del sexo masculino aportaron el mayor número de casos al tener 9422 (65.7 %), con grupos de edades mayores de 65 años siendo los que tienen el mayor riesgo con 2877 casos (20.1 %), seguido de 25 a 29 años, 1627 (11.3 %), ubicados en cabecera municipal principalmente. Según este Ministerio, las personas con afiliación al régimen subsidiado aportaron 7611 casos. Finalmente, las personas en condiciones de vulnerabilidad con mayor enfermedad de tuberculosis presentan su mayor proporción en población privada de la libertad, seguida de habitante de calle, trabajador de la salud y población procedente del exterior [4].

Actualmente, el procedimiento para diagnóstico de TB se lleva a cabo haciendo uso de dos métodos, el primero es la baciloscopia de esputo, ese método se fundamenta en un examen microscópico de muestras de esputo por técnicos de laboratorio, para detectar la presencia de la bacteria. Sin embargo, la microscopía solo detecta la mitad de los casos siendo un problema para determinar si un cuadro clínico corresponde a dicha enfermedad en realidad [1]. El segundo método consiste en el cultivo de micobacterias, en el cual para que el desarrollo de la bacteria sea visible macroscópicamente sobre el medio de cultivo se requiere por lo menos 15 días y hasta ocho días de incubación [4]. Debido a esto, la OMS recomendó la prueba Xpert MTB/RIF, la cual detecta en un plazo de dos horas la enfermedad así como la resistencia a la rifampicina, el cual es el fármaco más importante contra la TB.

En Colombia, para el año 2015, el fondo financiero de proyectos de desarrollo (FONADE), seleccionó las ciudades de: Bogotá, Bucaramanga, Cúcuta y Medellín, para dotarlas con el

sistema GenExpert, cuyo principio es la implementación de una prueba molecular completamente automatizada que simultáneamente detecta la presencia de la bacteria y resistencia a rifampicina, conocido como prueba Xpert MTB/RIF [5]. A pesar de ello, es una prueba con un costo elevado y con poca cobertura para las regiones que reporta el Ministerio de Salud en Colombia con mayor tasa de casos con TB.

Conforme a dicho reporte, para el año 2018, en Colombia Los departamentos que reportaron mayor número de casos de Tuberculosis Pulmonar, fueron Antioquia con 1.910 casos, Valle 1.206, Santafé de Bogotá con 457, Santander 286, Quindío con 196, Caldas 169 y Putumayo 148. Por su parte los de mayor incidencia, donde se reportan más de 50 casos por 100.000 habitantes y se consideran de riesgo muy alto para la transmisión fueron en su orden: Amazonas, Vaupés, Guainía, Guajira, San Andrés y Putumayo. Ahora bien, en el Plan Obligatorio de Salud POS, se garantiza la atención integral a la población afiliada a través de acciones de promoción y prevención. La mejor estrategia preventiva para la enfermedad consiste en la búsqueda de personas con tos de más de 15 días de evolución, realizar la prueba de baciloscopia de esputo. Si sale positivo se inicia el Tratamiento Acortado Supervisado (TAS), en el cual los trabajadores de la salud o voluntarios entrenados supervisan a los pacientes a diario para que tomen cada dosis del tratamiento durante los dos primeros meses y luego 2 veces por semana durante los cuatro meses siguientes [4].

De acuerdo con lo expuesto, donde la tasa de detección es baja ya que el procedimiento para buscar la enfermedad basa su principio en la espera de presencia de tos por más de 15 días, lo cual genera el riesgo de que si una persona efectivamente tiene la enfermedad pueda llegar a propagarla. Además los métodos aportados por el gobierno nacional no son suficientes, ya que como se evidencia en el reporte del 2018, los departamentos que presentan mayor incidencia de TB se encuentran geográficamente retirados de las ciudades dotadas con los equipos Xpert dificultando su diagnóstico y a su vez generando una posible reproducción de la enfermedad. Es por esto, que se hace necesario plantear nuevas estrategias que ayuden en la tarea de detección de esta enfermedad, generando un apoyo y una inclusión mayor de todas las personas que padecen esta enfermedad en el territorio nacional.

En este trabajo se hará énfasis en las soluciones algorítmicas que se basan en algoritmos conocidos como redes neuronales, es decir, un sistema informático modelado a partir del cerebro humano. De esta forma, poder proporcionar mecanismos que cada vez más son usados en la práctica médica, pues cada capa de la red neuronal funciona de manera independiente pero coordinada, separando aspectos importantes que los profesionales no ven, y así integrar los resultados con los modelos implementados. Estas nuevas herramientas visuales prometen transformar el diagnóstico médico e incluso pueden buscar cáncer a nivel de células individuales [6].

1.1 ESTADO DEL ARTE

Desde la ingeniería han crecido los aportes en el campo del diagnóstico de enfermedades durante los últimos años. Es por esto que se ha venido incrementando el uso de la inteligencia artificial en este tipo de aplicaciones. El tema no es nuevo, existe desde 1956, y en los años 70 se popularizó un poco más con su primera experiencia en el sector salud, denominado como Mycin. Un sistema experto orientado a la detección de enfermedades infecciosas de la sangre, el cual se comunicaba en lenguaje natural con el usuario y recetaba medicaciones de forma personalizada a cada paciente [7]. La Inteligencia Artificial puede aplicarse de maneras diferentes, como lo son: soluciones algorítmicas, tratamiento de imágenes.

A lo largo de los años se han ido implementando técnicas de algoritmos basados en redes neuronales en pro del diagnóstico de la TB, por ejemplo en India, del hospital de Osmania se recolectó información de 700 pacientes diagnosticados con la enfermedad, cuya información constaba de variables como: tos crónica por semanas, pérdida de peso, fiebre, sudores nocturnos, dolor en el pecho, sida (VIH). Con esos datos clínicos se hace una comparación de las técnicas más usadas para éste fin, las cuales son: *lineal discriminant analysis* (LDA), *support vector machine* (SVM), C4.5, vecino más cercano (k-NN),*partial least Squares Regression* (PLS-DA), BLR, *multinomial logistic regression* (MLR), *the k-means algorithm*, *entropy based mean clustering* (EMC) algorithm,the Apriori algorithm, obteniendo como resultado que el PLS-DA es el mejor para el diagnóstico de TB [9].

En Europa, realizaron un estudio con anamnesis la base de datos fue proporcionada por el hospital Diyarbakir, en Turquía, en este estudio se utilizaron 150 muestras; de las cuales 50 corresponden a pacientes diagnosticadas con TB, las otras 100 personas libres de la enfermedad, las entradas que se usaron para el método máquinas de vectores de soporte SVM, fueron: dolor en el pecho, debilidad, molestias de tos, temperatura corporal, disnea por esfuerzo, hábito de fumar, sonajero en el pecho, presión en el pecho, esputo, sonido en el tracto respiratorio, leucocitos, eritrocitos, trombosis, hematocrito, hemoglobina, albúmina 2, álcalis fosfatasa 2 L, amilasa, aspartato aminotransferasa. Para llevar a cabo el estudio primero la base de datos se divide en grupos de entrenamiento que son usados para crear el modelo de entrenamiento, después de obtener el modelo predictivo se realiza la prueba para determinar diagnóstico, finalmente se concluye que al hacer uso de SVM se alcanza un rendimiento significativo del 96 % al compararlo con otros estudios que también usaron SVM [14].

Por otro lado, se realizó un estudio en el cual agrupan y clasifican la TB en dos categorías; la Tuberculosis Pulmonar (PTB) y la Tuberculosis Pulmonar Retroviral (RPTB), haciendo uso de información clínica de un hospital estatal de pacientes que padecen la TB. La información clínica incluía: edad, tos crónica (registrada en cantidad de semanas), pérdida de peso, fiebre intermitente (registrada en cantidad de días), sudores nocturnos, esputo, tos con sangre, dolor en el pecho, VIH, hallazgos radiográficos, empleando un proceso de agrupación de las variables para realizar el estudio mediante dos etapas. En la primera, fue usado el algoritmo de *k-means* y en la segunda etapa un SVM, que consiste en tomar un conjunto de datos de entrada y a partir de los mismos datos, genera unos planos de separación predice entre dos clases, para cada entrada dada, realizando una clasificación lineal binaria no probabilística. Posteriormente, en el mismo estudio, hacen una comparación con siete algoritmos existentes: árboles de decisión, vecinos más cercanos,

bosques aleatorios (*random forest*) y clasificador bayesiano (*Naïve Bayesian Classifier*). El algoritmo *k-means* en comparación con los existentes tuvo una precisión del 98.7% con SVM, en la agrupación adecuada de las variables obtenidas en la anamnesis [11].

En algunas ocasiones, se suele modificar el principio del algoritmo que se quiere implementar, haciéndole unas variaciones o uniendo dos algoritmos, un claro ejemplo de esto, es un estudio que se realizó, en el cual se usa una modificación del algoritmo *Artificial Immune Recognition System2* (AIRS2) para realizar una clasificación de variables, empleando modelos de *k-vecinos próximos*, y de esta forma realizar la clasificación de las variables a utilizar en el desarrollo del algoritmo. Sin embargo, en lugar de usar el algoritmo de *k-vecinos próximos* hacen uso de SVM, dándole nombre a SAIRS2. La información empleada para llevar acabo el estudio se recolecto del laboratorio *Pasteur de amol* en Irán con variables como: dolor de pecho, pérdida de peso, tos, sudoración nocturna, fiebre, falta de aliento, concentración de hemoglobina, recuento de plaquetas, recuento total de glóbulos blancos (WBC), recuento de neutrófilos, recuento de linfocitos, velocidad de sedimentación globular, nivel de alanina aminotransferasa, nivel de fosfatasa alcalina, concentración de lactato deshidrogenasa y el estado de la tuberculosis (positivo o negativo). Para evaluar el desempeño del algoritmo se hizo validación cruzada, sensibilidad y especificidad, obteniendo valores del 100% en las dos últimas. En comparación, otro estudio que se llevó a cabo con estos mismos datos fue FuzzyART obteniendo precisión del 99%, sensibilidad 87% y especificidad 86.12% [12].

En América Latina, la inteligencia artificial (IA) ha tenido varios desarrollos, por ejemplo en Brasil, usan algoritmos junto al *big data* para prevención y control de enfermedades en las zonas donde no se cuenta con el suficiente apoyo para su diagnóstico [8]. Por otro lado, en Argentina se han desarrollado algoritmos que predicen la malnutrición de niños, basando su funcionamiento en medidas antropométricas como su peso, altura e índice de masa corporal [8]. Con estos ejemplos se puede evidenciar la importancia de estos algoritmos para ayudar al profesional de la salud usando otro punto de vista que puede mejorar la forma en la que se realiza cada proceso actualmente [8]. Otros resultados del uso de la IA en Brasil, fueron reportados en el hospital universitario de Rio de Janeiro, se llevó a cabo un estudio que consistía en evaluar un método de predicción para diagnosticar tuberculosis con baciloscopia negativa (SNPT). La información con la que desarrollaron el modelo consistía en información de la historia clínica de un paciente e incluyó factores demográficos y de riesgo típicamente empleados en el diagnóstico de la TB. Se consideraron 26 variables clínicas para el desarrollo del modelo, entre las cuáles fueron incluidas: edad, tos, esputo, sudor, fiebre, pérdida de peso, dolor en el pecho, estremecimiento, disnea, diabetes, alcoholismo, entre otras. Como resultado se obtuvo que el modelo basado en una red neuronal logró un buen rendimiento de clasificación, exhibiendo sensibilidad del 71 al 82% y especificidad del 60 al 83% [10].

En Brasil, más específicamente en Rio de Janeiro, se realizó un estudio diagnóstico de TB pleural, dividido en dos fases de diagnóstico de TB pleural, en la primera fase se manejaron variables de anamnesis tales como: edad, género, VIH. Las cuales fueron complementadas en la segunda fase, que se incluyeron variables como: adenosina desaminasa (ADA), bacilo resistente al alcohol acidal (BAAR), prueba serológica (ELISA), reacción en cadena de la polimerasa (PCR) y cultivo de líquido pleural. El estudio contenía información de 135 pacientes del hospital de Geral Santa Casa da Misericordia, de éstos pacientes 96 presentaban la enfermedad y 36 no la tenían. Se realizó un entrenamiento con SOM y con

análisis múltiple de correspondencia (MCA), en el cual se obtuvo como resultado que el agrupamiento adecuado se generaba en tres grupos, asimilándolo a una tarea de triage, y efectivamente para la primera fase se obtuvo un porcentaje de 83.33% y en la segunda fase 84.38% en el diagnóstico de TB pleural [30].

En otro proyecto de investigación se ha desarrollado un sistema neuro-difuso para el diagnóstico de TB. Los síntomas de la enfermedad se utilizaron como entrada del sistema, estos síntomas fueron investigados por expertos, proponiendo entradas como: tos persistente de dos semanas o más, tos con sangre, pérdida de peso, cansancio, fiebre, sudor nocturno, dolor en el pecho, dificultad para respirar, pérdida de apetito y agrandamiento de ganglios linfáticos. Es mencionada también la ventaja de la lógica difusa en procesar procesos no lineales, este estudio se desarrolló mediante el software Matlab [15].

En Colombia, la universidad Antonio Nariño apoyó un estudio sobre el desempeño de las redes neuronales ART Y FuzzyART, obteniendo como resultado una sensibilidad de 96.87%, lo cual indica que el diagnóstico de TB pleural usando éstos algoritmos es bastante efectiva [31].

El presente trabajo pretende hacer uso de algunas de las técnicas de agrupamiento mencionadas anteriormente basadas en inteligencia computacional, en este caso redes neuronales artificiales, para crear herramientas de apoyo al diagnóstico de la TB. Mediante, algoritmos de agrupamiento evaluando el índice de Davies Bouldin de cada algoritmo empleado, de esta manera generar información útil para el profesional de la salud. Las ayudas médicas y los métodos complementarios para el pre diagnóstico de TB se implementan en los centros de salud, el propósito del presente trabajo es analizar un conjunto de datos proporcionados por la Unidad de Servicios de Salud Santa Clara, adscrita a la Subred Integrada de Servicios de Salud Centro Oriente E.S.E.

1.2. MARCO TEÓRICO

1.2.1 TUBERCULOSIS PULMONAR

Las bacterias que provocan la tuberculosis se transmiten de una persona a otra por el aire. Esto ocurre cuando una persona enferma de tuberculosis tose, estornuda, habla o canta. Las personas que se encuentran cerca pueden inhalar las bacterias e infectarse. Hay dos tipos de afecciones de la TB: la infección de TB latente y la enfermedad de TB.

La infección de TB latente significa que las bacterias de TB pueden vivir en el cuerpo sin que la persona que las posea se enferme, en la mayoría de personas que inhalan las bacterias, el cuerpo puede combatir éstas evitando que se multipliquen y posteriormente evitando la enfermedad [26].

Las personas con la infección latente presentan una serie de características las cuales son:

- No tienen ningún síntoma.
- No pueden transmitir las bacterias de la tuberculosis a los demás.
- Por lo general, tienen una reacción positiva en la prueba cutánea de la tuberculina o un resultado positivo en el examen de sangre para detectar la tuberculosis.
- Pueden presentar enfermedad de tuberculosis si no reciben tratamiento para la infección de tuberculosis latente [26].

Si estas bacterias se activan en el cuerpo y se multiplican, la persona pasará de tener la infección de TB latente a tener la enfermedad de TB. Las personas que tienen la enfermedad de tuberculosis por lo general presentan síntomas y pueden transmitir las bacterias de la tuberculosis a los demás.

Las bacterias de la tuberculosis se multiplican con más frecuencia en los pulmones y pueden causar síntomas como los siguientes [26]:

- Una tos intensa que dura 3 semanas o más.
- Dolor en el pecho.
- Tos con sangre o esputo (flema que sale del fondo de los pulmones).
- Debilidad o cansancio.
- Pérdida de peso.
- Falta de apetito.
- Escalofríos.
- Sudor en la noche
- Fiebre.

1.2.2 FACTORES DE RIESGO DE LA TUBERCULOSIS

Cualquier persona puede contraer la enfermedad, sin embargo existen dos categorías de personas que presentan alto riesgo, las cuales son:

- Personas que hayan sido infectadas recientemente por las bacterias de la tuberculosis.
- Personas con afecciones que debilitan el sistema inmunitario.

Dentro de las categorías mencionadas anteriormente, se encuentran todas aquellas personas que presenten alguna de las siguientes características [26]:

- Aquellas personas que tengan contacto con una persona que tiene la enfermedad de tuberculosis.
- Es originario de un país donde la TB es muy común o lo ha visitado.
- Vive o trabaja en lugares donde la tuberculosis es más común, como un refugio para desamparados, una prisión o cárcel o establecimientos de cuidados a largo plazo.
- Es un trabajador de atención médica que atiende a clientes o pacientes con un alto riesgo de la enfermedad de tuberculosis.
- Habitantes de calle.
- Tiene la infección por el VIH.
- Es un niño menor de 5 años.
- Se infectó con la bacteria de tuberculosis en los últimos dos años.
- Tiene otros problemas de salud que dificultan que su cuerpo combata la enfermedad.
- Fuma cigarrillos o abusa del alcohol o las drogas.
- No le trataron adecuadamente la infección de tuberculosis latente o la enfermedad de tuberculosis en el pasado.

1.2.3 PRUEBAS DE DIAGNÓSTICO

Para el diagnóstico de tuberculosis pulmonar hay tres técnicas en la medicina para ayudar a los pacientes con síntomas de tuberculosis pulmonar como la baciloscopia, cultivo micobacteriano y el Xpert MTB/RIF.

1.2.3.1 BACILOSCOPIA

La baciloscopia consiste en una prueba de tres días consecutivos, donde se toma una muestra de esputo (catarro), para observar qué bacteria se encuentra presente. Esta prueba se utiliza para dictar el diagnóstico bacteriológico de la tuberculosis. Además de que es una técnica que permite identificar al 50-80% de los casos pulmonares positivos. Esta prueba se hace en ayunas y sin cepillarse.

El diagnóstico bacteriológico de la tuberculosis no está exento de errores: la baciloscopia, a través del examen directo de la muestra y coloración con la técnica de Ziehl-Neelsen, no es 100% confiable, pues el bacilo no siempre es detectado en las muestras clínicas examinadas. La microscopía, utilizando la técnica de coloración de Ziehl-Neelsen, es rápida, económica y sencilla. La sensibilidad deja mucho que desear, varía dependiendo del tipo de muestra y la micobacteria involucrada, ya que como regla deben existir entre cinco mil a diez mil bacilos por ml. de expectoración para que tengan un 50% de posibilidades de ser detectados al microscopio; sólo cuando el número de bacilos alcanza a más de 100.000 por ml. de expectoración, se puede esperar que el resultado sea positivo

[26]. Haciendo uso de la tinción fluorescente, este número puede ser tan bajo como 1.000 bacilos por ml [26]. Además, la presencia de bacilos ácido alcohol resistentes en el examen directo de muestras clínicas, no siempre garantiza que se trate de un bacilo tuberculoso, pues puede tratarse de una micobacteria atípica o de otro microorganismo que comparta la característica de ácido alcohol resistencia (M. leprae, Actinomyces, Nocardia). Esto puede ocasionar graves problemas diagnósticos y terapéuticos. El rango de sensibilidad de la baciloscopia, oscila entre 50 – 80% y la especificidad es virtualmente del 100% [26].

1.2.3.2 CULTIVO MICOBACTERIANO

Acompañando a la baciloscopia, el diagnóstico bacteriológico de la tuberculosis, debe complementarse con el cultivo. El cultivo es una técnica que tiene mayor sensibilidad (70-90%), ya que solo necesita que existan más de 10 bacilos, en muestras digeridas y concentradas, para que sea positivo. Realizando una pequeña comparación cabe recordar que la baciloscopia sólo utiliza 0,01 ml. de la muestra, es decir, de 10.000 campos microscópicos, en el mejor de los casos, se llegan a leer 100 a 200 campos; en cambio, en el cultivo se procesan 0,1 ml. de expectoración. El aislamiento de las micobacterias por cultivo es entorpecido por su lento crecimiento, ya que se necesita como mínimo un periodo de ocho días de incubación en medios convencionales para que pueda ser detectado. Los métodos de identificación convencionales después del cultivo incluyen determinación de la velocidad de crecimiento, crecimiento a diferentes temperaturas, morfología de la colonia, producción de pigmentos y susceptibilidad a los agentes. De tal manera que en casos en los que se requiera una toma de decisiones rápida para instaurar un tratamiento efectivo su valor es limitado. Se han hecho muchos intentos para mejorar los métodos de cultivo del bacilo tuberculoso, de modo que se pueda disponer de sus resultados en plazos más breves. Las técnicas más útiles a este respecto parecen ser las radiométricas, que permiten hacer el diagnóstico de muchas infecciones bacterianas en pocas horas y de la tuberculosis en pocos días. Además, tienen una mayor sensibilidad que los métodos bacteriológicos tradicionales [4].

Se han descrito métodos de cultivo más rápidos como: El método radiométrico BACTEC, el sistema ESP Myco, el sistema MB/BacT Alert, el sistema de tubo MGIT, el sistema Septi-Chek AFB, la detección de microcolonias en medios sólidos. Desafortunadamente tales pruebas son costosas en su inicio, requieren de personal altamente calificado y su sensibilidad y especificidad son muy variables de uno a otro laboratorio. Por lo tanto no están disponibles en todos los países subdesarrollados [4].

Por lo mencionado anteriormente, las técnicas bacteriológicas tradicionales de laboratorio (baciloscopia y cultivo), a pesar que tienen una buena sensibilidad y especificidad, no son lo suficientemente útiles cuando se requiere de un diagnóstico precoz y específico [4].

1.2.3.3 PRUEBA XPERT MTB/RIF

El método Xpert MTB/RIF es una prueba de amplificación del ácido nucleico totalmente automatizada que emplea un cartucho para diagnosticar la tuberculosis y la resistencia a la rifampicina, la cual permite diagnosticar TB en no más de dos horas con empleo de tiempo mínimo por parte de personal técnico.

Para iniciar el procedimiento se obtiene una muestra de esputo del paciente y se inactiva con un reactivo especialmente formulado para matar la bacteria, licuarla y estabilizar los componentes. La manipulación de la muestra es mínima, basta con agitar la mezcla, dejarla reposar quince minutos y agitarla de nuevo. De esa mezcla, se retira un pequeño volumen, de 2 a 3 mililitros, y se inserta en un cartucho en el que se encuentran todos los reactivos necesarios para realizar el análisis. A partir de ahí todo lo hace la máquina, que después de una hora y cuarenta y cinco minutos facilita el resultado [5].

1.2.4 REDES NEURONALES ARTIFICIALES

Al interior de la IA existen diferentes subáreas de estudio, una de ellas son las redes neuronales, las cuales están inspiradas en el funcionamiento del cerebro humano. Son definidas como una estructura de procesamiento paralelo masivo constituida por unas unidades sencillas (denominadas neuronas), que tienen la capacidad de almacenar conocimiento experimental y ponerla a disposición para su uso.

Las redes neuronales artificiales se asemejan a las redes neuronales biológicas en varios aspectos: el primero es que las neuronas son elementos simples y altamente interconectados, el segundo las conexiones llamadas pesos sinápticos que van entre las neuronas determinan la función de la red, en las cuales se almacena el conocimiento adquirido y por último el conocimiento se adquiere por medio del entorno gracias al aprendizaje [27].

Las redes neuronales se dividen en dos áreas principales: aprendizaje supervisado y aprendizaje no supervisado.

1.2.4.1 APRENDIZAJE SUPERVISADO

En el aprendizaje supervisado, los algoritmos trabajan con variables que están etiquetadas previamente, intentado encontrar una función que, dadas las variables de entrada, les asigne la etiqueta de salida adecuada. El algoritmo se entrena a partir de datos de entrenamiento, los cuales consisten de pares de objetos, una componente del objeto son los datos de entrada y la otra los resultados esperados, es decir, predice el valor de salida, este tipo de aprendizaje se suele usar para problemas de clasificación y problemas de regresión.

1.2.4.2 APRENDIZAJE NO SUPERVISADO

El aprendizaje no supervisado tiene lugar cuando no se dispone de datos etiquetados para el entrenamiento. Sólo conocemos los datos de entrada, pero no existen datos de salida que correspondan a un determinado *input.* Por tanto, sólo podemos describir la estructura de los datos, para intentar encontrar algún tipo de organización que simplifique el análisis, se usa para problemas de agrupamiento, algunos de los más usados son: los mapas autoorganizados de kohonen (SOM), vecino más cercano (KNN) y el k-means [8].

MAPAS AUTOORGANIZADOS (SOM):

Los mapas autoorganizados de Kohonen (SOM, del inglés *Self Organizing Maps*), fueron presentados en el año 1982 por el profesor finlandés Teuvo Kohonen. Es un algoritmo evolutivo de aprendizaje competitivo no supervisado, muy útil para problemas en los que hay búsqueda e identificación de estructuras (patrones o jerarquías entre los datos) o la reducción de la dimensionalidad sin perder información, dado su proceso de medir distancias entre los datos formando grupos o conglomerados (clúster) a través del encuentro de un representante o centroide por cada grupo [16].

Su arquitectura tiene características tales como: cada neurona es conectada a las demás neuronas pero es localmente conectada solo a sus vecinas, cada neurona tiene un vector de pesos *W* de entrada asociado y la neurona con el peso más cercano a la entrada se activará, las neuronas compiten entre sí para cumplir con la tarea propuesta, ésta competencia se da en todas las capas de la red, para finalmente generar la cantidad de agrupamiento o clúster según lo considere el mapa [16].

Un modelo SOM está compuesto por dos capas de neuronas. La capa de entrada la cual está formada por N neuronas, una por cada variable de los datos, se encarga de recibir y transmitir a la capa de salida la información procedente del exterior, por otra parte la capa de salida formada por M neuronas, es la encargada de llevar a cabo el procesamiento de la información y de generar el mapa de rasgos [16].

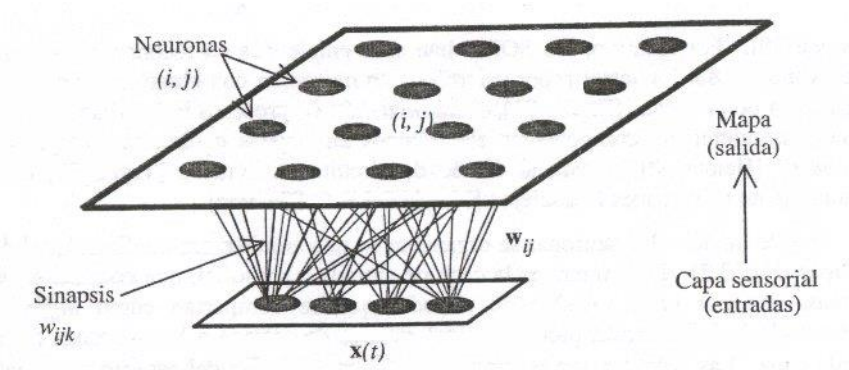

Ilustración 1 Arquitectura Mapas de Kohonen [17]

Como se observa en la ilustración 1, ésta es la manera de representar de manera intuitiva una red SOM.

Ahora bien, de manera descriptiva, las neuronas de las capas se encuentran conectadas por sus pesos *Wij* respectivamente. De esta forma, las neuronas de salida tienen asociado un vector de pesos *Wj* llamado vector de referencia, debido a que constituye el vector prototipo (o promedio) de la categoría representada por la neurona de salida j. Así, el SOM define una proyección desde un espacio de datos en alta dimensión a un mapa bidimensional de neuronas [17].

El proceso de aprendizaje está conformado por tres etapas: competitivo, cooperativa y adaptativa. En aprendizaje competitivo, se calcula la distancia euclidiana de los pesos de

cada entrada a todas las unidades o neuronas, aquel con el peso más similar a la entrada es definido como la mejor unidad de coincidencia (BMU). En el aprendizaje cooperativo el proceso se da alrededor de BMU, y las unidades cercanas a él se actualizan basado en la función de vecindad. Finalmente, el proceso adaptativo cambia los pesos de BMU de acuerdo con la entrada, esto se realiza mediante la siguiente ecuación [22].

$$W_V(S+1) = \mathrm{Wv}(\mathrm{s}+1) = \mathrm{Wv}(\mathrm{s}) + \theta(\mathrm{u},\mathrm{v},\mathrm{s})\alpha(\mathrm{s})(\mathrm{D}(\mathrm{t}) - \mathrm{Wv})$$

Dónde

- s, es el índice del paso.
- t, es el índice dentro del conjunto a entrenar.
- u, es el índice de la unidad de mejor correspondencia.
- D (t), es el vector de entrada.
- α(s), función entre conjuntos ordenados que preserva o invierte el orden dado decreciente de aprendizaje.
- θ (u, v, s), corresponde a la función de vecindad, la cual depende de la distancia entre la unidad de mejor correspondencia y la neurona v, la cual puede obtener un valor de 1 o 0 según la cercanía con la BMU o elegir una función gaussiana.

Este proceso se repite para cada vector de entrada un número λ (ciclos), generalmente grande. Por último, calculando la distancia euclidiana entre el vector de entrada y los de pesos se descubrirá la única neurona ganadora, que es aquella que se encuentra más cerca del vector de entrada [22].

El objetivo de la red es descubrir rasgos comunes, correlaciones o categorías en los datos de entrada, e incorporarlos a su estructura interna de conexiones. Es por esto, que se dice que las neuronas deben auto-organizarse en función de los estímulos, (datos) procedentes del exterior [17].

La funcionalidad de los mapas auto-organizados de Kohonen o red SOM está en descubrir la estructura subyacente al conjunto de datos introducidos en la red con fines de estudio. A continuación de forma descriptiva se presenta la estructura de una red SOM:

- Matriz de neuronas: las neuronas de la red se distribuyen sobre un campo o rejilla de dos dimensiones, esta puede ser rectangular, hexagonal entre otras, de tal manera en la que se van identificando los grupos o clústeres.
- Espacio de entrada: el conjunto de datos de entrada se identifican como un vector de *N* componentes, por cada atributo o correlación, determinado así la dimensión del vector de pesos de inicialización.
- Espacio de salida: corresponde al conjunto de datos de salida de la red neuronal SOM en un mapa de menor dimensión al conjunto de entrada, suele ser bidimensional.
  Relación topológica entre neuronas: entre las neuronas existe una relación de vecindad, la cual es subyacente al conjunto de datos debido a su estructura topológica natural. Lo más importante de esta relación es definir la regla de asociación en la inicialización y aprendizaje de la Red Neuronal SOM [22].

ADAPTIVE RESONANCE THEORY (ART)

La red de resonancia adaptiva (ART), fue desarrollado por Steven Grossberg y Gair Carpenter en 1987, ésta se comporta como un clasificador de vectores, el cual dependiendo de la entrada se encarga de organizar la información dependiendo de las categorías de los datos almacenados en la que más se parezca [18].

Para lograr su funcionamiento las redes ART tienen una organización en fases: la primera fase comúnmente llamada fase de reconocimiento, en la cual el vector de entrada se compara con la clasificación ya existente en el nodo de salida, para tomar un valor de "1" si coincide con la clasificación impuesta o por el contrario "0" si no coincide. La segunda fase es la fase de comparación, en la cual se realiza otra comparación entre el vector de entrada y el vector de la capa de comparación, en esta fase la condición principal es que la similitud entre los dos vectores sea menor que el parámetro de vigilancia. Por último, se encuentra la fase de búsqueda, para la cual si existe un restablecimiento y la coincidencia no es tan buena el proceso se repite hasta lograr una adecuada similitud, por el contrario si la coincidencia es buena se finaliza el entrenamiento [19]

REDES FUZZY-ART

La red Fuzzy-Art es una de los tipos de las redes (ART), de carácter binario a carácter analógico, a través del operador difuso *and*, en lugar de la intersección booleana. Sin embargo sigue conservando las características generales de la arquitectura ART, como lo son: la constante comparación entre las entradas y la información ya almacenada en el sistema, la búsqueda paralela y la autoorganización. Además posee ventajas:

- La lógica difusa permite que se puedan dar entradas analógicas.
- En el proceso de aprendizaje los pesos sinápticos solo decrecen con el tiempo, lo que permite asegurar la estabilidad del sistema.
- Se realiza un preprocesamiento de las entradas mediante el uso de código complementario, que asegura la normalización (evitando la saturación) y la conservación de la amplitud relativa de las componentes de entrada [20].

Como en el resto de arquitecturas ART, Fuzzy ART es una red modular, que incluye una capa de entrada de nodos que almacenan el vector de entrada, una capa de elección que contiene las categorías activas y una capa de emparejamiento que recibe las conexiones [20].

1.2.5 PROCESO DE AGRUPAMIENTO

El proceso de agrupamiento consiste en distribuir los datos logrando que los objetos de un mismo clúster (grupo) tengan una similitud alta, y baja con elementos de otros clúster. Ésta medida está basada en las características que describen los objetos. Por lo general se utilizan distancias como: distancia euclidiana, de Manhatan, de Mahalanobis, entre otras.
El agrupamiento es una técnica de aprendizaje automático no supervisado. Desde un punto de vista práctico, el agrupamiento juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la

información y minería de texto, aplicaciones web, marketing, diagnóstico médico, análisis de ADN en biología computacional, y muchas otras [6].

Los métodos de agrupamiento se dividen en:

- Particionales: se busca una distribución en cierto número de clases, que cumplan con el objetivo planteado.
- Jerárquicos: en la cual se realizan distribuciones anidadas, donde cada nivel de la jerarquía es una misma distribución, obtenida por el grupo de la jerarquía del nivel inferior [6].

En este caso, se empleó un método de agrupamiento particional llamado *Kmeans* como método base para comparar con los agrupamientos generados con el empleo de las redes neuronales mencionadas.

KMEANS

Creado por MacQueen en 1967, es un algoritmo de aprendizaje no supervisado, el cual agrupa una cierta cantidad de objetos en "*k*" grupos teniendo en cuenta sus características, el principio del algoritmo es basado en la disminución de las distancias entre el objeto y el centroide del grupo al que se va asociar, para este algoritmo se puede implementar la distancia de la preferencia, sin embargo se suele hacer uso de la euclidiana o de la cuadrática [21].

Se le llama *Kmeans* porque cada uno de los clúster es representado por la media (o media ponderada) de sus puntos, es decir, por su centroide. La representación por centroides presenta la ventaja de tener un significado gráfico y estadístico inmediato. Se basa en la minimización de la distancia interna dentro de su grupo. En realidad, se minimiza la suma de las distancias al cuadrado de cada patrón al centroide de su grupo.

El algoritmo del *K-means* se realiza en 4 etapas:

- Etapa 1: Se eligen aleatoriamente *"k"* objetos formando así los *"k" gr*upos iniciales. Para cada grupo, el valor inicial del centro es el valor de la entrada asociada.
- Etapa 2: Cada entrada es reasignada al grupo de centroide más próximo a ella según una medida de distancia, la cual suele ser la medida euclidiana.
- Etapa 3: Una vez que todas las entradas son acomodadas, recalcular los centros de "*k"* grupos(los baricentros).
- Etapa 4: Se repiten las etapas 2 y 3 hasta que no hayan más reasignaciones [21].

2. OBJETIVOS

2.1. General

Analizar información clínica de la Unidad de Servicios de Salud Santa Clara, empleando algoritmos de agrupamiento para generar grupos de riesgo en sujetos bajo sospecha de tener Tuberculosis pulmonar como herramienta de apoyo a su diagnóstico.

2.2. Específicos

a) Establecer una base de datos con la información de sujetos bajo sospecha de tener Tuberculosis pulmonar, codificando sus variables para poder ser empleadas en algoritmos de agrupamiento.
b) Comparar tres algoritmos de agrupamiento basados en redes neuronales para establecer grupos de riesgo en el diagnóstico de la Tuberculosis pulmonar.
c) Validar la información obtenida a través de la interacción con profesionales de la salud, estableciendo hallazgos que puedan ser tenidos en cuenta en el diagnóstico de la Tuberculosis pulmonar.

3. METODOLOGÍA

El trabajo descrito en este documento propone su desarrollo en tres grandes etapas, descritas en los numerales de la presente sección.

3.1. ACONDICIONAMIENTO DE DATOS

En primer lugar, los datos con los que se desarrolló el presente trabajo corresponden al proyecto de investigación titulado “Generación de modelos alternativos basados en inteligencia computacional para tamización y diagnóstico de Tuberculosis pulmonar”. Los cuales hacen parte de la Unidad de Servicios Santa Clara, que a su vez está vinculada a la Subred Integrada de Servicios de Salud Centro Oriente E.S.E.

La unidad de servicios Santa Clara, suministró dos bases de datos, una correspondiente al año 2017, la cual contenía 62 registros y otra correspondiente al año 2018 con 89 registros, cada uno de estos registros fueron guardados por el hospital cada vez que ingresaba un registro al programa de Tuberculosis de dicha institución, debido a que presentaba los síntomas de la enfermedad. Allí, se encuentra información bastante general de los pacientes debido a que es un registro propio de la institución realizado por profesionales en enfermería. Variables como el sexo, la edad, el tipo de tuberculosis, pertenencia étnica, grupo poblacional, régimen de afiliación, barrio y localidad de domicilio son almacenados. Seguido de información específica como la prueba y resultado de la afección de sida (VIH) y condición de egreso. En las tablas 1, 2 y 3 se muestran algunos registros brindados por el hospital.

Trimestre del año	Ingresa a tratamiento	Edad	Sexo	Pertenencia étnica	Grupo poblacional	Barrio	Comuna localidad
I	No	52	M	Otro	Habitante de calle	Sin dato	Sin dato
I	Si	57	M	Otro	Habitante de calle	San Bernardo	Sin dato
I	Si	53	M	Otro	Habitante de calle	Sin dato	Sin dato
I	Si	53	M	Otro	Otros		Usaquén
I	si	34	M	Otro	Habitante de calle		

Tabla 1 Registros originales de pacientes con TB

Afiliación	Tipo TB	Prueba molecular	Se realizó prueba	Resultado prueba	Prueba norma
s-subsidiado	Pulmonar	Detectado	Vih+previo	Vih+previo	Vih+previo
n-no asegurado	Pulmonar	Detectado	Si	Negativo	Negativo
n-no asegurado	pulmonar		Pte no acepta	Pte no acepta	Pte no acepta
c-contributivo	Pulmonar	No detectado	Vih+previo	Vih+previo	Vih+previo
n-no asegurado	Pulmonar	detectado	Si	Negativo	Negativo

Tabla 2. Registros originales de pacientes con TB.

Tipo de farmacorresistencia	Condición de egreso
Ninguna	Fallecido durante el tratamiento
Ninguna	Fallecido durante el tratamiento
Ninguna	Perdida en el seguimiento
Ninguna	Tratamiento terminado

Tabla 3. Registros originales de pacientes con TB.

Las bases de datos suministradas por el hospital, difieren en que la del año 2018 no cuenta con la variable de condición de egreso. Con los datos mostrados anteriormente se prosigue a plantear 3 escenarios para el presente trabajo: *i)* información únicamente del año 2017, *ii)* información únicamente del año 2018, y *iii)* unión de las dos bases de datos (2017 y 2018) sin tener en cuenta la variable de condición de egreso, logrando que los datos tengan concordancia debido a que los datos del año 2018 no la tienen. A continuación se presenta la definición de cada variable:

- Trimestre del año: variable cualitativa, sus categorías eran (I, II, III, IV), ésta variable no se usó para el desarrollo del trabajo.
- Sexo: variable cualitativa define si el individuo es femenino o masculino
- Edad: variable cuantitativa, específica los años de cada individuo los cuales están en el rango de 20 a 91.
- Grupo poblacional: Variable cualitativa, con dos categorías que hacían referencia si el paciente era persona en condición de indigencia o persona del común.
- Barrio y localidad de domicilio: variable cualitativa, cuyas categorías se encontraban los barrios y localidades de Bogotá.
- El régimen de afiliación: variable cualitativa, cuyas categorías eran (Subsidiado, no asegurado, contributivo)
- Prueba molecular, resultado y prueba con parámetros según la norma para infección VIH: éstas 3 variables cualitativas, tenían las mismas categorías, las cuales eran: (Paciente no acepta, negativo, positivo, VIH+previo)
- Tipo de farmacorresistencia: variable cualitativa, cuyas categorías eran: (mono r, ninguna).

- Condición de egreso, variable cualitativa, cuyas categorías eran (fallecido durante el tratamiento, paciente no reportado, perdida en el seguimiento, recuperado, tratamiento terminado, sin dato), ésta variable no fue empleada durante el trabajo.

3.1.1 ANÁLISIS DE VARIABLES

Para el correcto desarrollo del trabajo, es necesario realizar un análisis descriptivo que nos permita visualizar de una manera visual la información con la que se va a trabajar.

VARIABLE CUANTITATIVA

En este grupo, se encuentra la variable edad, con la que se realiza un análisis de medidas de tendencia central para cada uno de los tres escenarios planteados anteriormente, mostrado en la tabla 4.

BASE DE DATOS 2017			
Moda	**Mediana**	**Media**	**Medidas de dispersión**
La edad con más frecuencia es 45 años y corresponde a un 8% de los datos	La mayoría de la población contagiada de TB está por encima de 40 años y la minoría por debajo de este valor	El promedio de la edad de personas con TB es de 42 años	El promedio es de 42 años con una desviación de $\pm$ 15.58
BASE DE DATOS 2018			
Moda	**Mediana**	**Media**	**Medidas de dispersión**
La edad con más frecuencia es 29 años y corresponde a un 6.7% de los datos	La mayoría de la población contagiada de TB está por encima de 40 años y la minoría por debajo de este valor	El promedio de la edad de personas con TB es de 43 años	El promedio es de 42 años con una desviación de $\pm$ 18.09
BASE DE DATOS AÑOS 2017 Y 2018			
Moda	**Mediana**	**Media**	**Medidas de dispersión**
La edad con más frecuencia es 29 años y corresponde a un 6.6% de los datos	La mayoría de la población contagiada de TB está por encima de 40 años y la minoría por debajo de este valor	El promedio de la edad de personas con TB es de 42 años	El promedio es de 42 años con una desviación de $\pm$ 17.05

Tabla 4. Medidas de tendencia central

En las figuras 2,3 y 4, se evidencia la información sobre la edad más frecuente de los pacientes que acuden al hospital por diagnóstico de TB, confirmando al análisis estadístico realizado previamente.

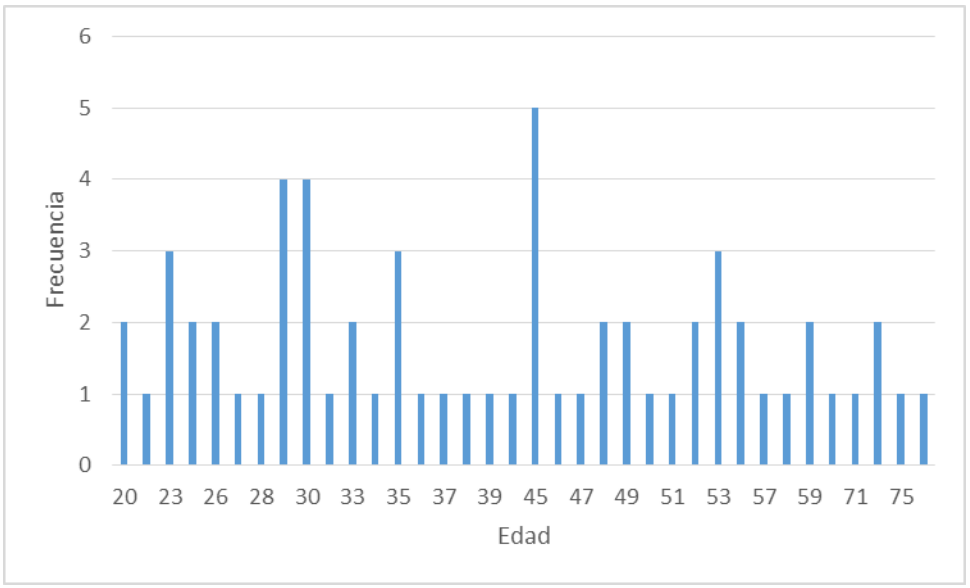

Ilustración 2 Histograma de la edad base de datos 2017

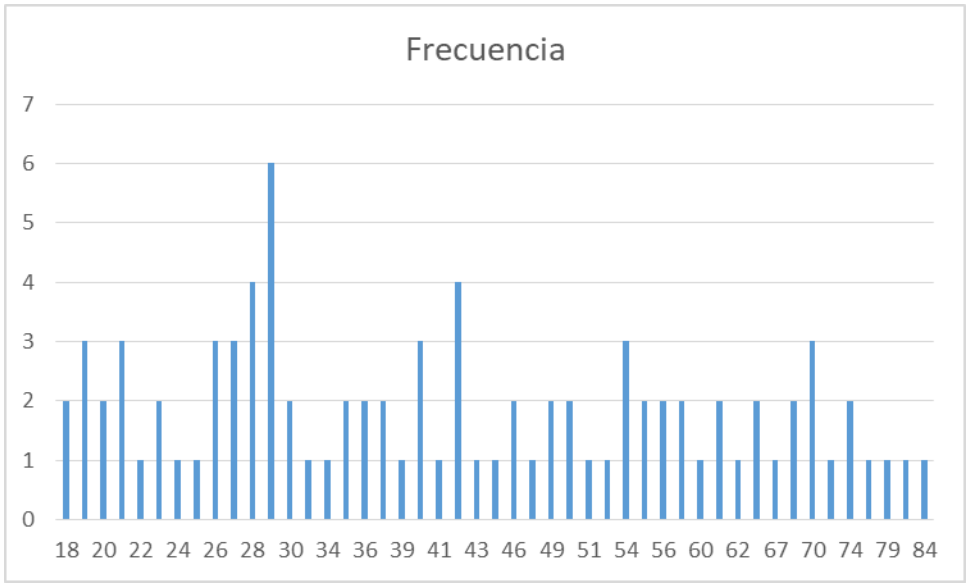

Ilustración 3 Histograma base de datos 2018

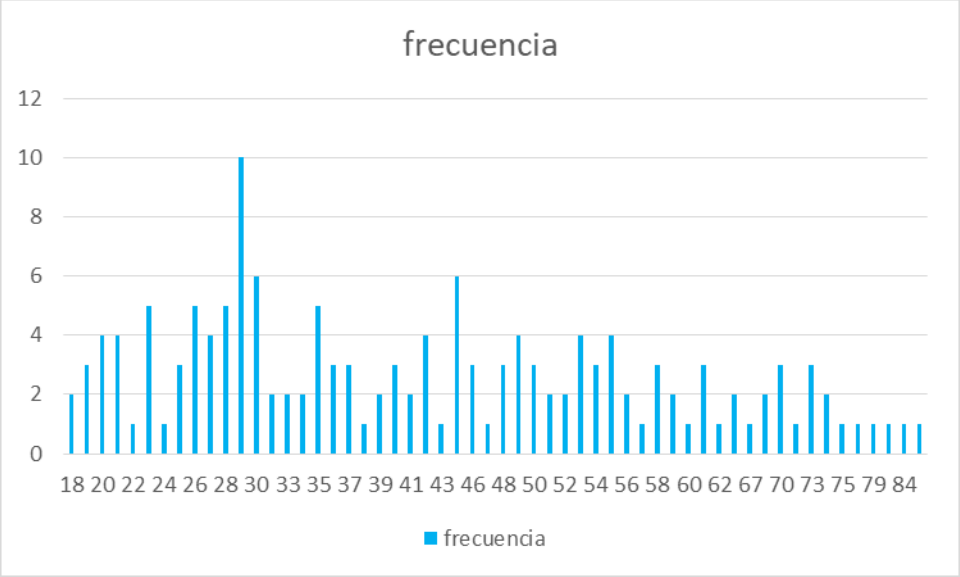

Ilustración 4 Histograma de edad de base de datos 2017 y 2018

VARIABLES CUALITATIVAS

En este grupo se cuenta con 9 variables cualitativas, a las cuales se les hizo un estudio descriptivo para cada escenario planteado.

BASE DE DATOS 2017

En el primer escenario, de los sesenta y dos registros existentes, se tiene que cincuenta y cinco pacientes ingresaron a tratamiento, mientras que siete no lo hicieron, como se muestra en la tabla 4.

- **Ingresa a tratamiento**

Tratamiento	No Personas
Si	55
No	7

Tabla 5 Distribución de frecuencias variable ingresa a tratamiento

Para el año 2017, el hospital Santa Clara atendió una mayor cantidad de hombres que mujeres, cuya información se muestra en la tabla 5.

- **Sexo**

Sexo	No personas
F	14
M	48

Tabla 6 Distribución de frecuencias sexo

Para este mismo año, del total de los pacientes que asistieron al hospital, se encontró que uno, que pertenecía de la población étnica, como se evidencia en la tabla 7.

- **Pertenencia étnica**

Pertenencia Étnica	No personas
Otro	61
Indígena	1

Tabla 7 Distribución de frecuencias variable pertenencia étnica

Por otro lado, en la tabla 8 e ilustración 5, se observa que del total de los registros, se brindó atención médica a veinte habitantes de calle, a cuarenta y una personas del común, y a un paciente del cual no se conoce su grupo poblacional.

- **Grupo poblacional**

Grupo Poblacional	No personas
Habitante calle	20
otros	41
sin dato	1

Tabla 8 Distribución de frecuencias variable grupo poblacional

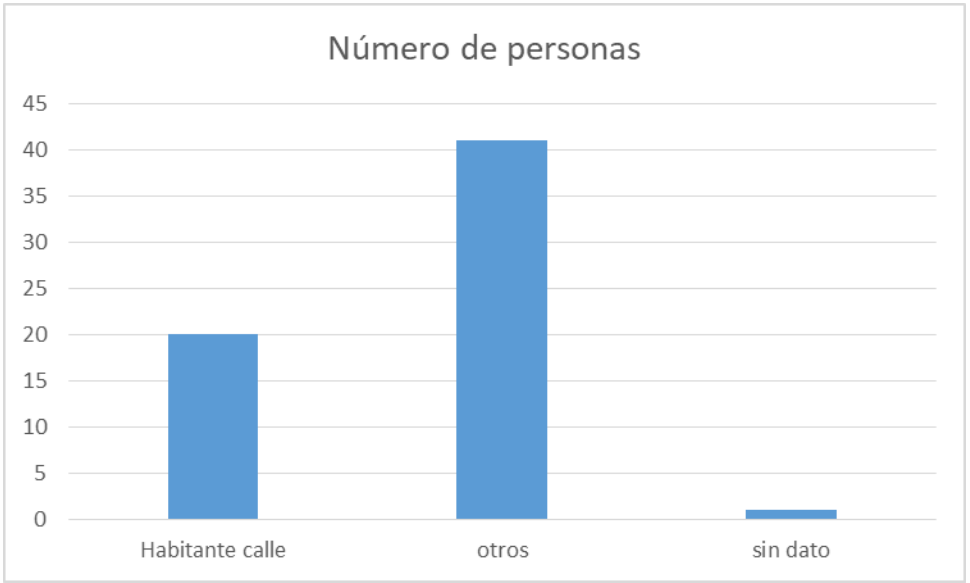

Ilustración 5 Histograma variable grupo poblacional base de datos 2017

El hospital Santa Clara está ubicado en una zona vulnerable de la ciudad, por ende como se observa en la tabla 9 e ilustración 6,la mayor cantidad de pacientes corresponde a un régimen subsidiado, seguida de los pacientes que no cuentan con sistema de salud, finalmente los pacientes con sistema de salud contributivo.

- **Régimen de afiliación**

Afiliación	No personas
S - SUBSIDIADO	44
C - CONTRIBUTIVO	2
N - NO ASEGURADO	16
E - ESPECIAL	0

Tabla 9 Distribución de frecuencias variable régimen de afiliación

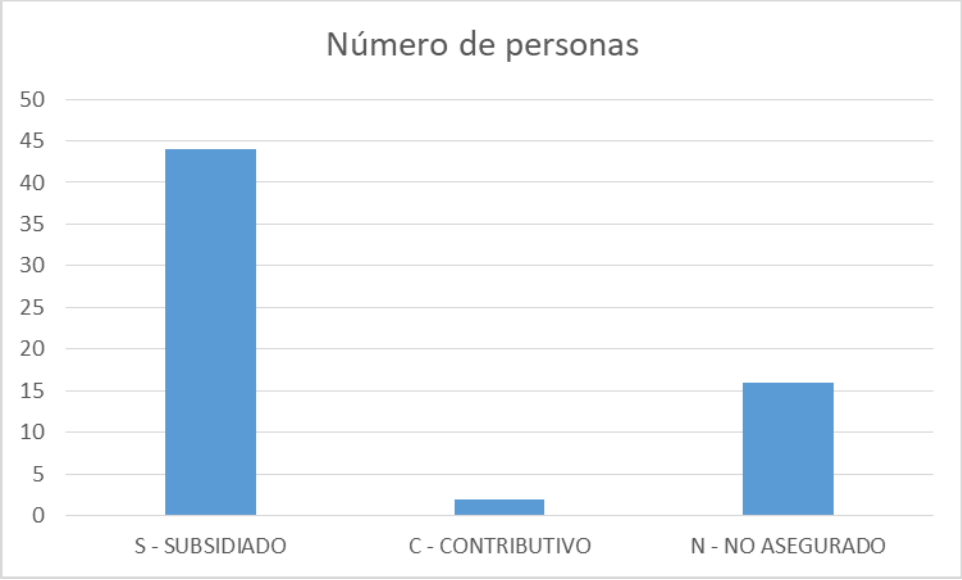

Ilustración 6 Histograma variable régimen de afiliación base de datos 2017

En la tabla 10 e ilustración 7, se muestra que del total de los registros con enfermedad de TB pulmonar, en veintisiete gracias a la prueba molecular se les detectó VIH, seguidos de los registros que no se tienen información, y por último a trece no se les detectó.

- **Prueba molecular VIH**

Prueba molecular	Número de personas
Detectado	27
No detectado	13
Sin dato	22

Tabla 10 Distribución de frecuencias variable prueba molecular

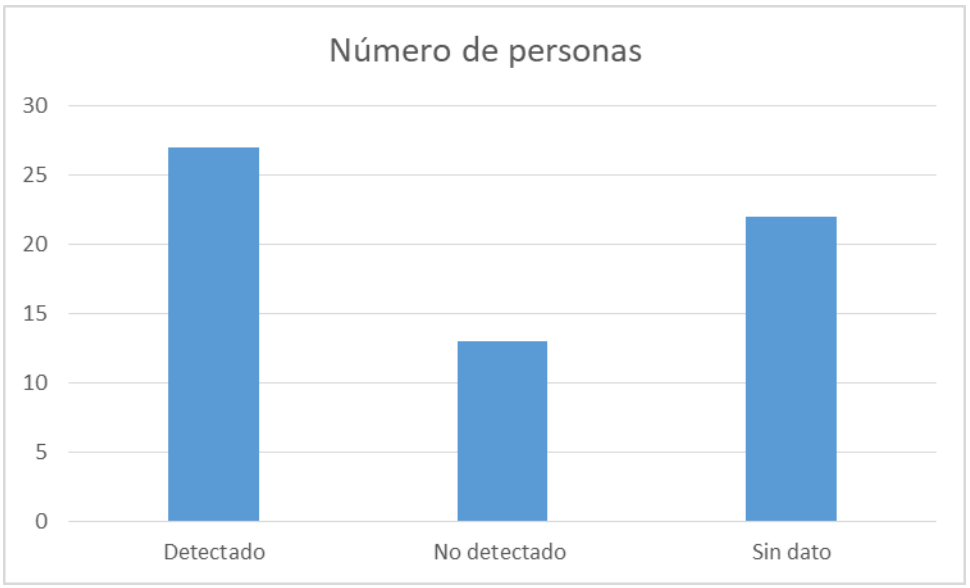

Ilustración 7 Histograma variable prueba molecular base de datos 2017

El hospital realizó la prueba VIH, en donde treinta y seis pacientes que aceptaron, a dos no se les realizo, a veintidós se les realizó VIH+previo, finalmente hubo un paciente que no cuenta con ésta información y un paciente que no acepta realizarse la prueba. Como se evidencia en la tabla 11 e ilustración 8.

- **Se realizó prueba VIH**

Se realizó prueba	Número de personas
VIH+ PREVIO	22
NO	2
SI	36
PTE NO ACEPTA	1
SIN DATO	1

Tabla 11 Distribución de frecuencias variable prueba VIH

Ilustración 8 Histograma variable realizó prueba VIH base de datos 2017

En la tabla 12 e ilustración 9, se muestran los resultados de la prueba VIH realizada por el hospital, en la cual se tienen veinticinco pacientes con resultado negativo, once con resultado positivo, veintidós con VIH+previo y cuatro (paciente que no aceptó y aquellos que no tienen información).

- **Resultado prueba VIH**

Resultado Prueba	Número de personas
NEGATIVO	25
POSITIVO	11
VIH+PREVIO	22
PTE NO ACEPTA	1
SIN DATO	3

Tabla 12 Distribución de frecuencias variable resultado prueba VIH

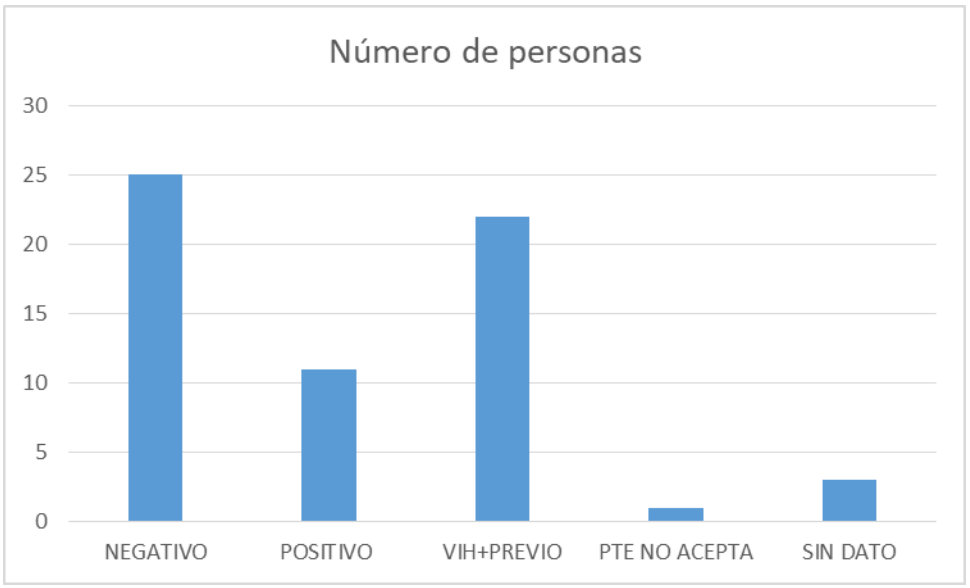

Ilustración 9 Histograma variable resultado prueba VIH base de datos 2017

Finalmente para el primer escenario, del total de los registros, cuarenta y uno no tienen ningún tipo de farmacorresistencia, uno posee mono-r y existen veinte registros de los que no se tiene esta información, como se observa en la tabla 13.

- **Tipo de farmacorresistencia**

Farmacorresistencia	Número de personas
NINGUNA	41
MONO R	1
SIN DATO	20

Tabla 14 Distribución de frecuencias variable farmacorresistencia

- **BASE DE DATOS 2018**

En el segundo escenario, de los ochenta y nueve registros existentes, se tiene que ochenta y tres pacientes ingresaron a tratamiento, mientras que siete no lo hicieron, como se muestra en la tabla 14.

- **Ingresa a tratamiento**

Tratamiento	Número de Personas
SI	83
NO	6

Tabla 14 Distribución de frecuencias variable tratamiento

Para el año 2018, el hospital Santa Clara atendió una cantidad mayor de hombres, específicamente sesenta y seis, en cambio fueron atendidas veintitrés mujeres, cuya información se muestra en la tabla 15.

- **Sexo**

Sexo	Sexo
M	66
F	23

Tabla 15 Distribución de frecuencias variable sexo

Para este mismo año, de los pacientes que asistieron al hospital, se encontró que uno del total de los registros, hacía parte de la población étnica indígena, como se evidencia en la tabla 16.

- **Pertenencia étnica**

Pertenencia Étnica	Número de personas
Otro	88
indígena	1

Tabla 16 Distribución de frecuencias variable pertenencia étnica

Por otro lado, en la tabla 17, se observa que del total de los registros, se brindó atención médica a un habitante de calle, a ochenta y ocho personas del común, y a un paciente del cual no se conoce su grupo poblacional.

- **Grupo poblacional**

Grupo poblacional	Número de personas
HABITANTE DE CALLE	1
OTROS	88

Tabla 17. Distribución de frecuencias variable grupo poblacional

El hospital Santa Clara está ubicado en una zona vulnerable de la ciudad, por ende como se observa en la tabla 18 e ilustración 10, la mayor cantidad de pacientes corresponde a un régimen subsidiado, seguida de los pacientes que no cuentan con sistema de salud, finalmente los pacientes con sistema de salud contributivo.

- **Régimen de afiliación**

RÉGIMEN DE AFILIACIÓN	Número de personas
S – SUBSIDIADO	54
N - NO ASEGURADO	33
C - CONTRIBUTIVO	2

Tabla 18.Distribución de frecuencias variable régimen de afiliación

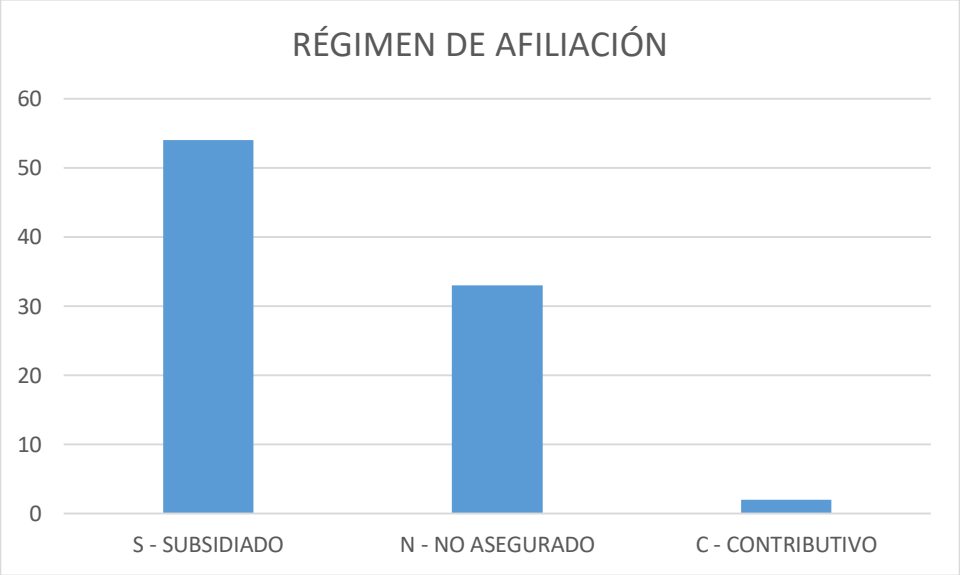

Ilustración 10 Histograma variable régimen de afiliación, base de datos 2018

En la tabla 19 e ilustración 11, se muestra que del total de los registros con enfermedad de TB pulmonar, en treinta y siete gracias a la prueba molecular se les detectó VIH, seguidos de los registros de los cuales no se tiene información, y por último a siete a no se les detectó.

- **Prueba molecular**

PRUEBA MOLECULAR	Número de personas
Detectado	37
NO Detectado	7
No Reportado	45

Tabla 19 Distribución de frecuencias variable prueba molecular

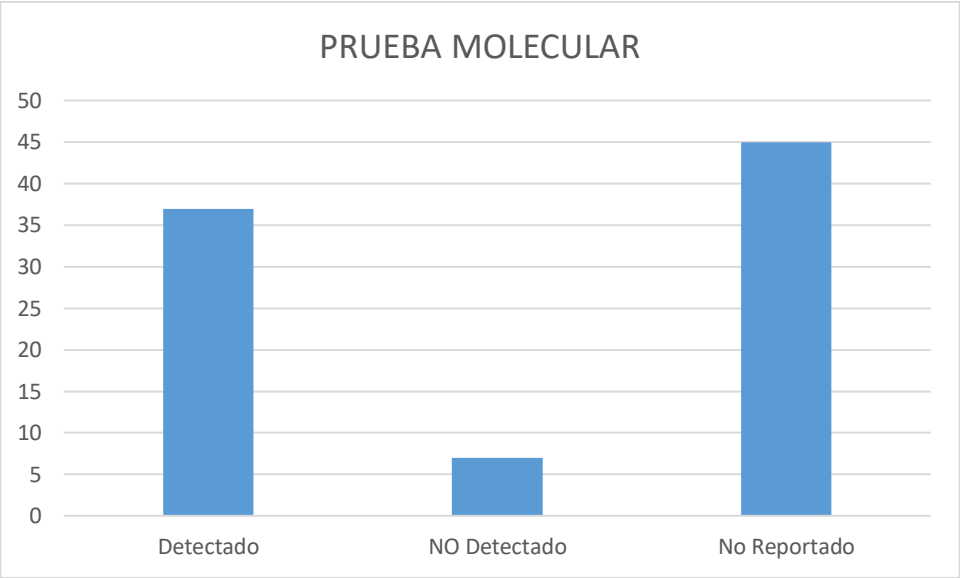

Ilustración 11 Histograma variable prueba molecular, base de datos 2018

El hospital realizó la prueba VIH, en la cual cuarenta y cinco pacientes aceptan, a diecisiete no se les realiza, a veintisiete se les realizó VIH+previo. Como se evidencia en la tabla 20 e ilustración 12.

- **Se realizó prueba VIH**

SE REALIZÓ PRUEBA	SE REALIZÓ PRUEBA
VIH + Previo	27
SI	45
NO	17

Tabla 20 Distribución de frecuencias variable prueba VIH

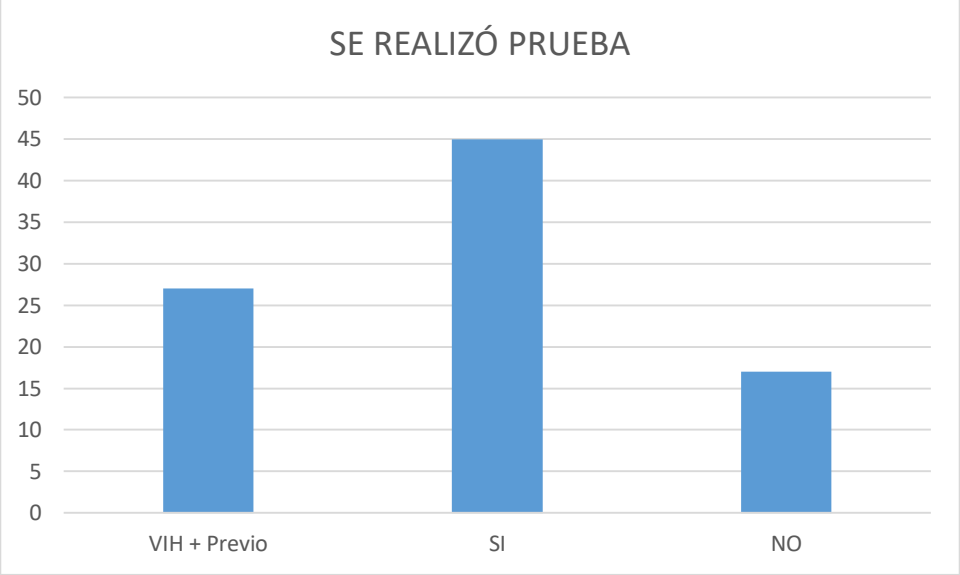

Ilustración 12 Histograma variable prueba VIH, base de datos 2018

En la tabla 21 e ilustración 13, se muestran los resultados de la prueba VIH realizada por el hospital, en la cual se tienen treinta y cuatro pacientes con resultado negativo, once con resultado positivo, veintiséis con VIH+previo y dieciocho entre el paciente que no aceptó y de los que no se tiene información.

- **Resultado prueba**

RESULTADO PRUEBA	RESULTADO PRUEBA
VIH + Previo	26
Negativo	34
Paciente no Acepta	1
Positivo	11
Nreportado	17

Tabla 21 Distribución de frecuencias variable resultado prueba

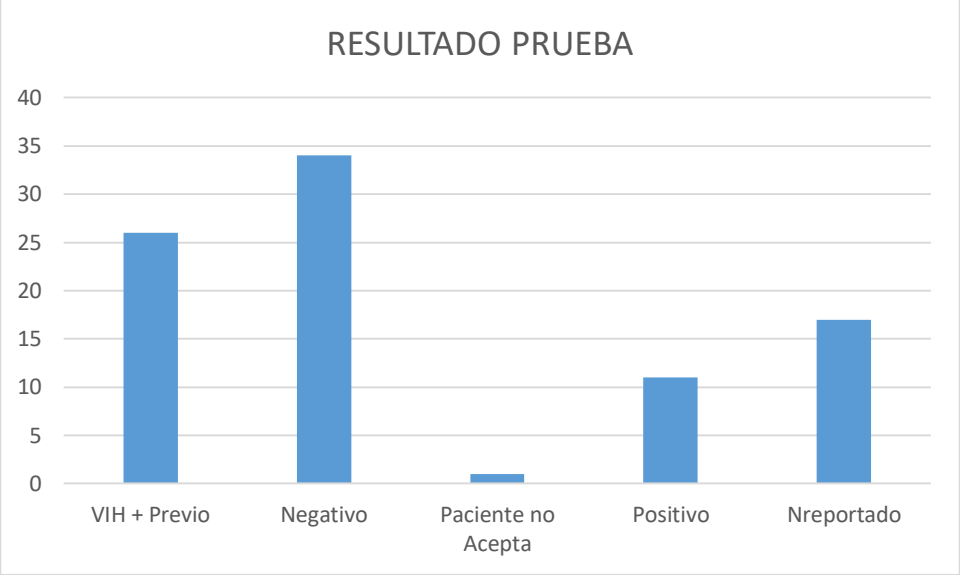

Ilustración 13 Histograma variable resultado prueba, base de datos 2018

Finalmente para el segundo escenario, del total de los registros, sesenta y cinco no tienen ningún tipo de farmacorresistencia, tres poseen mono-r y existen veintiún registros de los que no se tiene esta información, como se observa en la tabla 22.

- **Tipo de farmacorresistencia**

TIPO DE FARMACORRESISTENCIA	TIPO DE FARMACORRESISTENCIA
Mono R	3
No Reportado	65
Ninguna	21

Tabla 22 Distribución de frecuencias variable farmacorresistencia

- **BASE DE DATOS AÑOS 2017 Y 2018**

En el tercer escenario planteado, de los ciento cincuenta y un registros existentes, se tiene que ciento treinta y ocho pacientes ingresaron a tratamiento, mientras que trece no lo hicieron, como se muestra en la tabla 23.

- **Ingresa a tratamiento:**

INGRESA A TRATAMIENTO	INGRESA A TRATAMIENTO
SI	138
NO	13

Tabla 23 Distribución de frecuencia variable tratamiento

En el periodo comprendido entre el año 2017 y 2018, el hospital Santa Clara atendió una cantidad mayor de hombres, específicamente ciento catorce, en cambio fueron atendidas treinta y siete mujeres, cuya información se muestra en la tabla 24.

- **Sexo:**

SEXO	SEXO
M	114
F	37

Tabla 24 Distribución de frecuencia variable sexo

Para este mismo periodo, de los pacientes que asistieron al hospital, se encontró que dos del total de los registros, hacen parte de la población étnica, como se evidencia en la tabla 25.

- **Pertenencia étnica indígena**

PERTENENCIA ÉTNICA	PERTENENCIA ÉTNICA
Otro	149
Indígena	2

Tabla 25 Distribución de frecuencia variable pertenencia étnica

Por otro lado, en la tabla 26, se observa que del total de los registros, se brindó atención médica a veintiuno habitantes de calle, a ciento veintinueve personas del común, y a un paciente del cual no se conoce su grupo poblacional.

- **Grupo poblacional**

GRUPO POBLACIONAL	GRUPO POBLACIONAL
Habitante de calle	21
Otros	129
No Reportado	1

Tabla 26 Distribución de frecuencia variable grupo poblacional

El hospital Santa Clara está ubicado en una zona vulnerable de la ciudad, por ende como se observa en la tabla 27 e ilustración 14, la mayor cantidad de pacientes corresponde a un régimen subsidiado, seguida de los pacientes que no cuentan con sistema de salud, finalmente los pacientes con sistema de salud contributivo.

- **Régimen de afiliación**

RÉGIMEN DE AFILIACIÓN	RÉGIMEN DE AFILIACIÓN
S - SUBSIDIADO	98
N - NO ASEGURADO	49
C - CONTRIBUTIVO	4

Tabla 27 Distribución de frecuencias variable régimen de afiliación

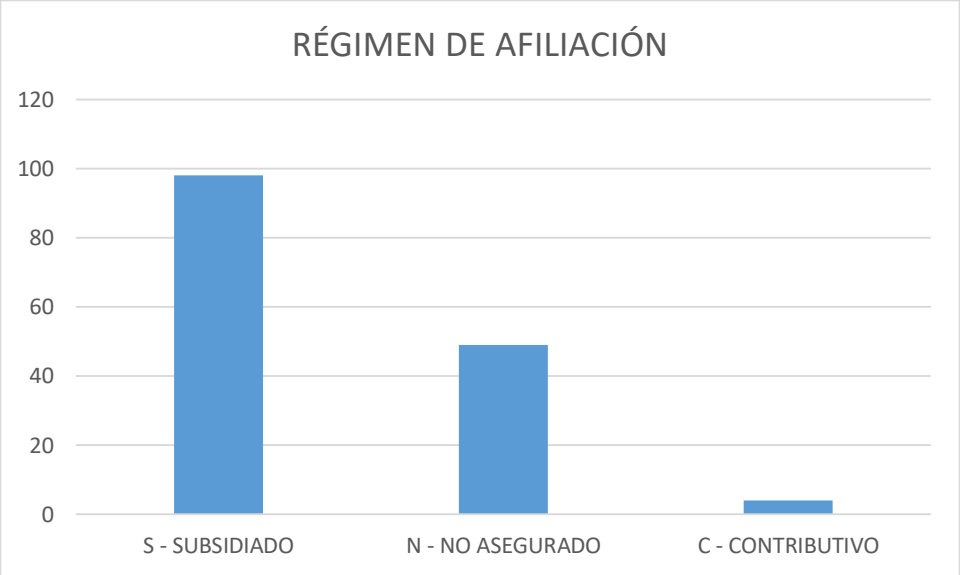

Ilustración 14 Histograma variable régimen de afiliación, base de datos unida

En la tabla 28 e ilustración 15, se muestra que del total de los registros con enfermedad de TB pulmonar, en sesenta y cuatro (64) gracias a la prueba molecular, se detectó VIH al 100%, seguidos de los registros de los cuales no se tiene información, y por último a siete a quienes no se les detectó

- **Prueba molecular VIH**

PRUEBA MOLECULAR	PRUEBA MOLECULAR
Detectado	64
No Detectado	20
No Reportado	67

Tabla 28 Distribución de frecuencias variable prueba molecular

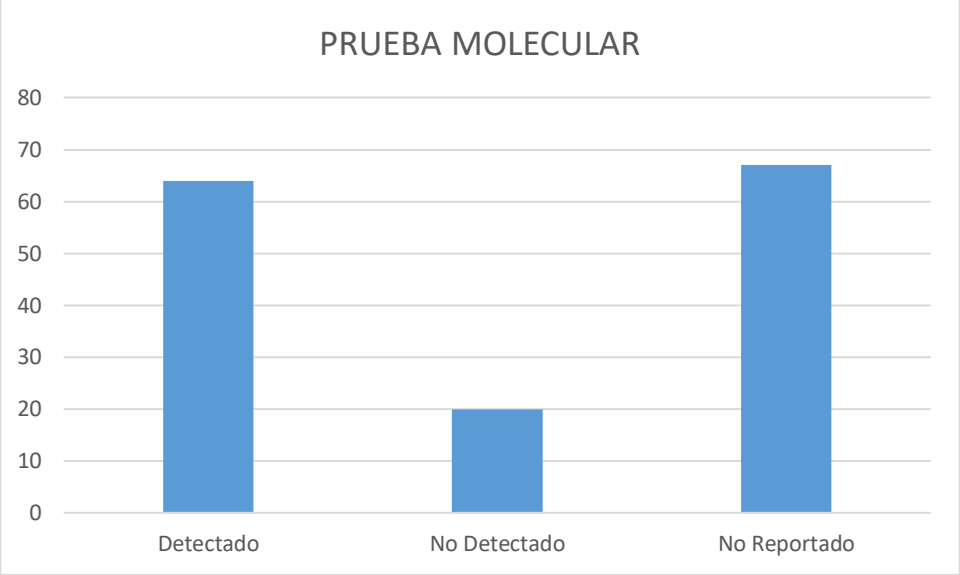

Ilustración 15 Histograma variable prueba molecular, base de datos unida

El hospital realizó la prueba VIH, en la cual ochenta y un pacientes aceptan, a diecinueve no se les realiza, a cuarenta y nueve se les realizó VIH+previo. Como se evidencia en la tabla 29 e ilustración 16.

- **Realizó prueba**

SE REALIZÓ PRUEBA	SE REALIZÓ PRUEBA
VIH + Previo	49
SI	81
Pte No Acepta	1
NO	19
No Reportado	1

Tabla 29 Distribución de frecuencias variable realizó prueba

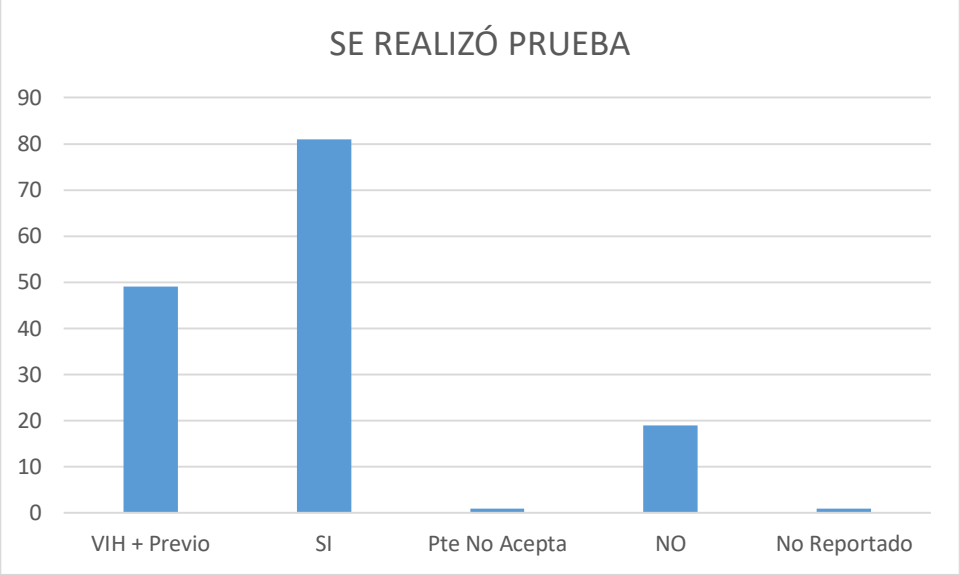

Ilustración 16 Histograma variable realizó prueba, base de datos unida

En la tabla 30 e ilustración 17, se muestran los resultados de la prueba VIH realizada por el hospital, en la cual se tienen cincuenta y nueve pacientes con resultado negativo, veinte dos con resultado positivo, cuarenta y ocho con VIH+previo y veintidós entre el paciente que no aceptó y aquellos que no se tiene información.

- **Resultado prueba**

RESULTADO PRUEBA	RESULTADO PRUEBA
VIH + Previo	48
Negativo	59
PACIENTE NO ACEPTA	2
Positivo	22
Nreportado	20

Tabla 30 Distribución de frecuencias variable resultado prueba

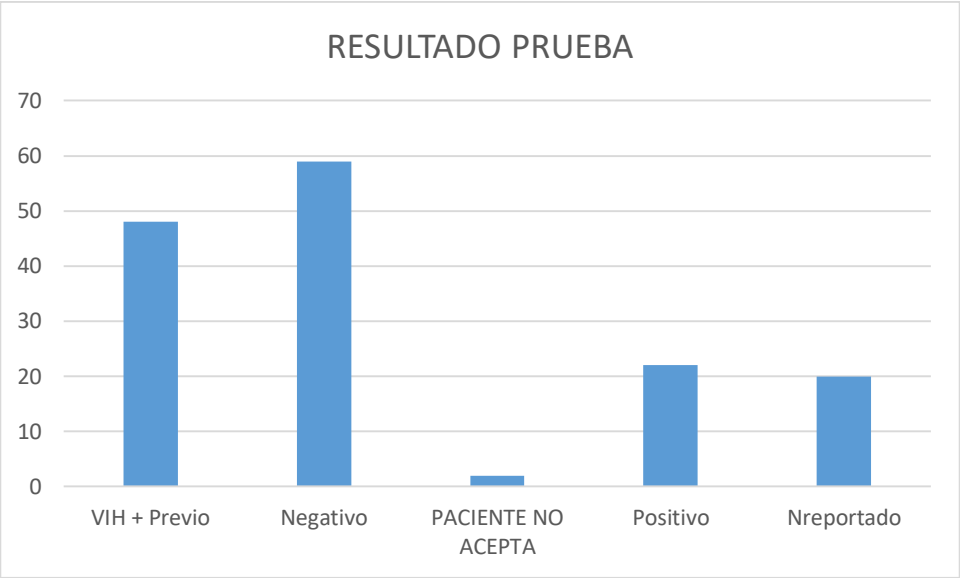

Ilustración 17 Histograma variable resultado prueba, base de datos unida

Finalmente para el tercer escenario, del total de los registros, ochenta y cinco no tienen ningún tipo de farmacorresistencia, cuatro poseen mono-r y existen sesenta y dos registros de los que no se tiene esta información, como se observa en la tabla 31.

- **Tipo de farmacorresistencia**

TIPO DE FARMACORRESISTENCIA	TIPO DE FARMACORRESISTENCIA
Mono R	4
No Reportado	85
Ninguna	62

Tabla 31 Distribución de frecuencia variable tipo de farmacorresistencia

3.1.2 CATEGORIZACIÓN DE VARIABLES

Realizada la primera validación de datos se encontró que cada variable alberga ciertas categorías, por este motivo se agrupan las opciones similares, para organizarlos en conjuntos más pequeños y realizar la codificación que necesitamos para desarrollar los experimentos empleando los algoritmos de agrupamiento.

En las bases de datos hay casillas marcadas como sin dato, o simplemente vacías, para realizar la codificación se decidió unificar esas dos posibles categorías en una sola llamada no reportado.

3.1.2 BINARIZACIÓN DE LAS VARIABLES

Para aplicar la información a las redes neuronales, fue necesario realizar una codificación previa para obtener la información apropiada, conocida como datos de entrada a las redes. Se tomaron las variables mencionadas anteriormente y se codificaron utilizando bits estos

de acuerdo a la cantidad de posibles valores de la variable que se tengan por cada una de estas, las cuales estará en 1 y los demás en 0. Esto, haciendo uso del método de binarización codificación *One-Hot*. A continuación se presenta la binarización de algunas variables.

Tratamiento	**SI**	**NO**
SI	1	0
NO	0	1
Sexo	**M**	**F**
M	1	0
F	0	1
Pertenencia Étnica	**Otro**	**indígena**
otro	1	0
indígena	0	1

Tabla 32 Binarización de las variables

Realizo Prueba	**vih+previo**	**SI**	**Pte no acepta**	**NO**	**Nreportado**
vih+previo	1	0	0	0	0
SI	0	1	0	0	0
Pte no acepta	0	0	1	0	0
NO	0	0	0	1	0
Nreportado	0	0	0	0	1

Tabla 33 Binarización de las variables

Al finalizar la binarización, se procede a organizar la información para los tres escenarios planteados, para lo cual cada variable mencionada anteriormente queda determinada por una cantidad de columnas, mostradas en las tablas 34, 35 y 36.

- **Base de datos 2017**

Variable	**Número de columnas**
Ingresa a tratamiento	2
Sexo	2
Pertenencia étnica	2
Grupo poblacional	3
Barrio	25

Localidad	17
Régimen de afiliación	3
Prueba molecular	3
Realizó prueba	5
Resultado prueba	5
Tipo de farmacorresistencia	3
Condición de egreso	7

Tabla 34 Descripción de cantidad de columnas a cada variable

- **Base de datos 2018**

Variable	Número de columnas
Ingresa a tratamiento	2
sexo	2
Pertenencia étnica	2
Grupo poblacional	2
Barrio	42
Localidad	9
Régimen de afiliación	3
Prueba molecular	3
Realizó prueba	3
Resultado prueba	5
Tipo de farmacorresistencia	3

Tabla 35 Descripción de cantidad de columnas a cada variable

- **Base de datos años 2017 y 2018**

Variable	Número de columnas
Ingresa a tratamiento	2
Sexo	2
Pertenencia étnica	2
Grupo poblacional	3
Barrio	58
Localidad	20
Régimen de afiliación	3
Prueba molecular	3
Realizó prueba	5
Resultado prueba	5
Tipo de farmacorresistencia	3

Tabla 36 Descripción de cantidad de columnas a cada variable

3.2 ALGORITMOS EMPLEADOS EN EL AGRUPAMIENTO

3.2.1 MAPAS AUTOORGANIZADOS (SOM)

Teniendo las bases de datos listas, se procede a realizar el entrenamiento de las redes neuronales. Para este paso, se inicializan las neuronas de la red SOM y se procede a realizar su entrenamiento, previa indicación del tamaño de cada uno de los mapas.

Para obtener el tamaño correcto de cada mapa o matriz se debe tener en cuenta la cantidad de datos con los que se va a entrenar la red neuronal, para la base de datos del 2017, eran 62 registros por 15 variables, entonces si se multiplica el 62 por la raíz cuadrada de 15 se obtiene un valor de 240, es decir, un aproximado del mapa sería 12 neuronas de ancho y 20 de alto. Para la base de datos del 2018, 17 neuronas de ancho y 20 de alto, finalmente para la de los dos años 25 neuronas de ancho por 20 de alto. En cuanto a la forma del mapa se prefiere neuronas de tipo hexagonal debido a que la distancia entre los centros de cada una de ellas es el mismo.
Para el desarrollo de este trabajo, la red SOM se combinará con Kmeans, de esta manera primero la red neuronal lleva a cabo el entrenamiento, y posteriormente esas neuronas resultantes del entrenamiento son las que se agrupan por medio de Kmeans.

3.2.2 REDES ART

Para realizar el entrenamiento de la red, es necesario establecer algunos parámetros: parámetro de vigilancia dentro del rango de (0 1], un valor cercano a uno implica una clasificación muy exigente para la cual dos medidas deben ser muy parecidas para pertenecer a un mismo clúster, por el contrario si el parámetro es cercano a cero permite un agrupamiento con características poco parecidas. El radio que se va a adecuar a las neuronas, la tasa de aprendizaje la cual si es más cercano a cero, presenta un funcionamiento más adecuado del entrenamiento.

La variación del radio y el parámetro de vigilancia fueron desde 0.1 hasta 1, aumentando cada 0.01, se hizo de esta manera ya que al inicio se planteó aumentando cada 0.1, pero se evidenció que el índice de validación de agrupamiento generaba cambios abruptos.

3.2.3 REDES Fuzzy-ART

Para el entrenamiento de la red Fuzzy-ART el parámetro relacionado con el radio de vigilancia varia cada 0.001, siguiendo la misma metodología de las redes ART. También es necesario establecer el parámetro de elección (alfa), el cual generalmente es mayor a cero. Este proceso permite encontrar el mejor radio para que se generen grupos de datos diferentes a uno, en el caso de que el radio sea muy grande, o diferentes al número de elementos a agrupar.

3.3 INDICE DE VALIDACIÓN DEL AGRUPAMIENTO

Es de importancia evaluar el resultado de los algoritmos de agrupamiento. Sin embargo, es difícil definir cuando el resultado de un agrupamiento es aceptable, debido a esta razón existen técnicas e índices para la validación de un agrupamiento realizado. En el presente trabajo, usamos el índice de Davies Bouldin (DB), el cual tiene como objetivo cuantificar la distancia intra-grupos, normalizando con la distancia entre-grupos. De esta forma, el valor del índice estará más cerca a cero para la mejor la agrupación, pues indica que las entradas se encuentran unidas en sus grupos y a su vez alejadas de los otros grupos.
La ecuación que rige el índice de Davies Bouldin, está dada por:

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max(j \neq i) \frac{\sigma_i + \sigma_j}{d(c_i, c_j}$$

Donde:

- n es el número de grupos
- c es el centroide del grupo x
- σ es la distancia media de todos los elementos en grupo al centroide x
- $d(ci, cj)$ es la distancia entre centroides.

Por otro lado si el índice DB nos da un valor mayor a cero punto siete (0.7), indica que los grupos se encuentran dispersos, mientras si es un valor menor a éste los grupos se encuentran compactos [24].

4. RESULTADOS

4.1 Primera Configuración Experimental

Al realizar el entrenamiento con los algoritmos de agrupamiento, se lleva a cabo la comparación del número de agrupamiento vs el índice de validación, de cada una de los 3 escenarios planteados. Los cuales se pueden observar a continuación.

Con el uso de SOM+kmeans, se puede observar que para el primer escenario, según los datos sumininistrados, se obtiene un número mayor a diez como el mejor agrupamiento, esto teniendo en cuenta que el índice de validación Davies Bouldin (DB) para diez grupos corresponde a un valor de 1.269, y como se observa en la figura 18 (línea azul) éste va de manera descendiente, sabiendo que entre menor sea el valor de DB, es mejor el agrupamiento. El comportamiento anterior se puede asociar a que la red neuronal apoyó su agrupamiento en aquellas variables con más cantidad de categorías posibles como correspondía a la información de domicilio de los registros.

Al aplicar la red neuronal *Kmeans,* se observa en la figura 18 (línea roja), que el índice DB, alcanza su valor mínimo en 1.5 para siete grupos, ahora bien en comparación con SOM+Kmeans, se obtiene un valor mayor pero una cantidad de grupos menor, es decir se observa un comportamiento inversamente proporcional. Por el contrario ésta red neuronal basó su agrupamiento en la variable condición de egreso la cual presentaba siete categorías.

Al implementar ART, se obtiene un índice de 1.937 con tres grupos, como se observa en la figura 26. Mientras que la implementación de FuzzyART cuatro grupos para un índice de 1.867, información mostrada en la figura 18 (línea morada). Con relación a los resultados anteriores se observa que la variación de éstas dos redes neuronales es menor a la variación de *Kmeans* implementado solo versus SOM+Kmeans.

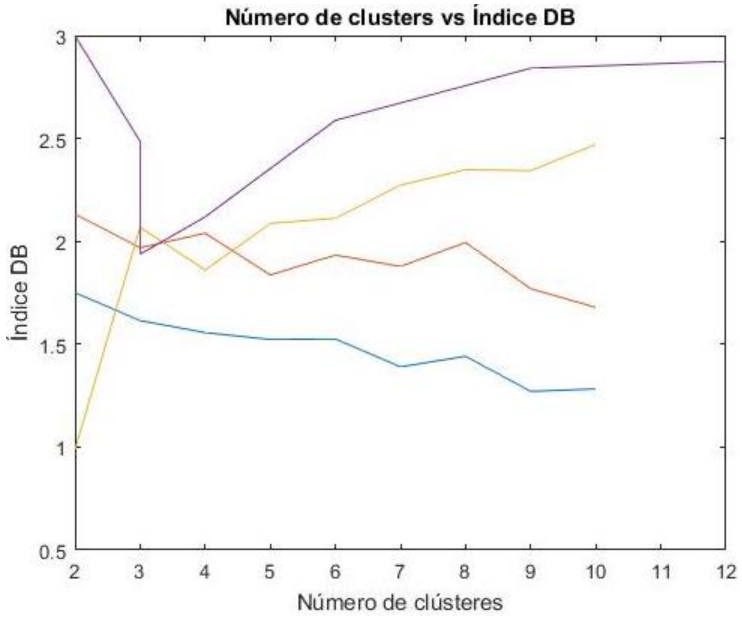

Ilustración 18 Número de grupos primera configuración experimental

4.2 Segunda Configuración Experimental:

Para el segundo escenario planteado, y como se evidencia en la figura 19 (línea morada), la implementación de la red SOM+Kmeans, arroja un índice DB de 1.1 para un total de diez grupos. En este caso sucede lo mismo que para el primer escenario el índice va de manera descendente lo cual nos dice que el número óptimo de grupos es mayor a diez.

Por otro lado con *Kmeans* se obtiene un valor de 1.52 con ocho grupos (línea naranja). Para la información anterior se observa un cambio bastante significativo entre la implementación de ésta red neuronal con respecto a los datos del año 2017 y año 2018, esto se puede deber a la cantidad de registros que contenía cada base de datos, es decir entre mayor registros menor cantidad de grupos y similar índice DB.

Con la red neuronal ART se obtienen cinco grupos con un índice DB de 2.424. Por otro lado para FuzzyART un índice de 2.138 con nueve grupos (línea roja), según lo anterior el cambio de estas dos redes es mayor con respecto al primer escenario, ya que para éste escenario en particular tuvo menos variación la implementación de som+Kmeans y *Kmeans,* lo cual deja una duda sobre cuál de los algoritmos planteados es el correcto y más adecuado para el agrupamiento de pacientes con enfermedad TB.

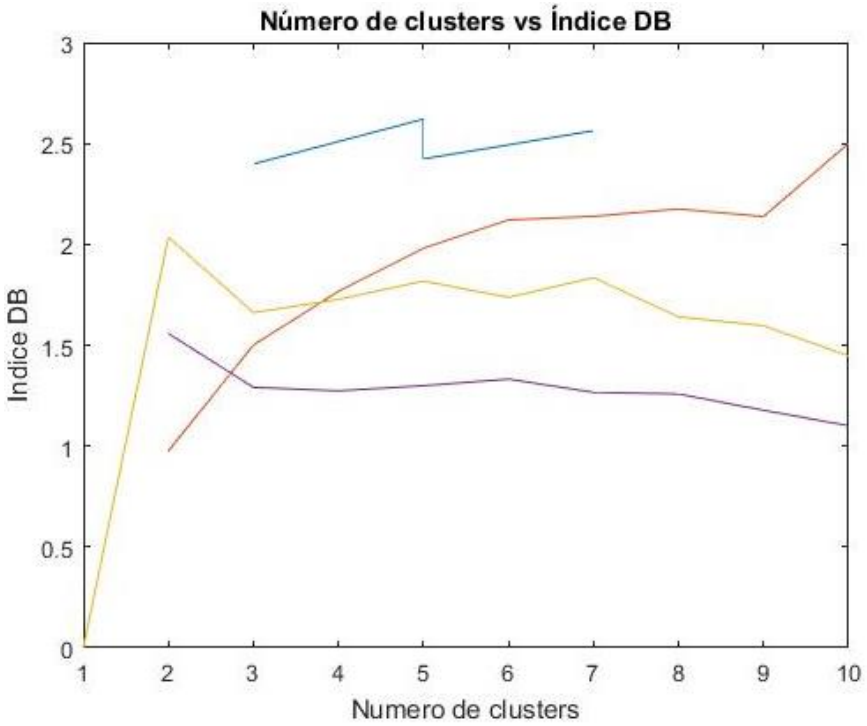

Ilustración 19 Número de grupos segunda configuración

4.3 Tercera Configuración Experimental:

Con relación al último escenario planteado se observa que para la red neuronal combinada SOM+Kmeans se tiene una cantidad de grupos que pueden ir de seis a siete con un índice de 1.23, éste caso en particular nos muestra una gráfica distinta a las dos anteriores de éste algoritmo, ya que como se evidencia en la figura 20 (línea morada), inicia descendiendo, posterior a eso presenta un ascenso seguido de otro descenso y finalmente otro ascenso.

Para la red neuronal *Kmeans* se tienen ocho grupos con un índice de 1.845, como se observa en la figura 20 (línea roja), de nuevo se observa un cambio significativo entre estas dos redes neuronales, para lo cual finalmente se concluye que para éste trabajo en particular estas dos redes pueden no ser las más apropiadas por los cambios observados durante la realización.

Finalmente para ART, como se evidencia en la figura 20 (línea azul), se obtuvo un índice de 2.5 para tres grupos, mientras que FuzzyART arrojo cinco grupos con un índice de 1.66, información existente en la figura 32 (línea naranja).

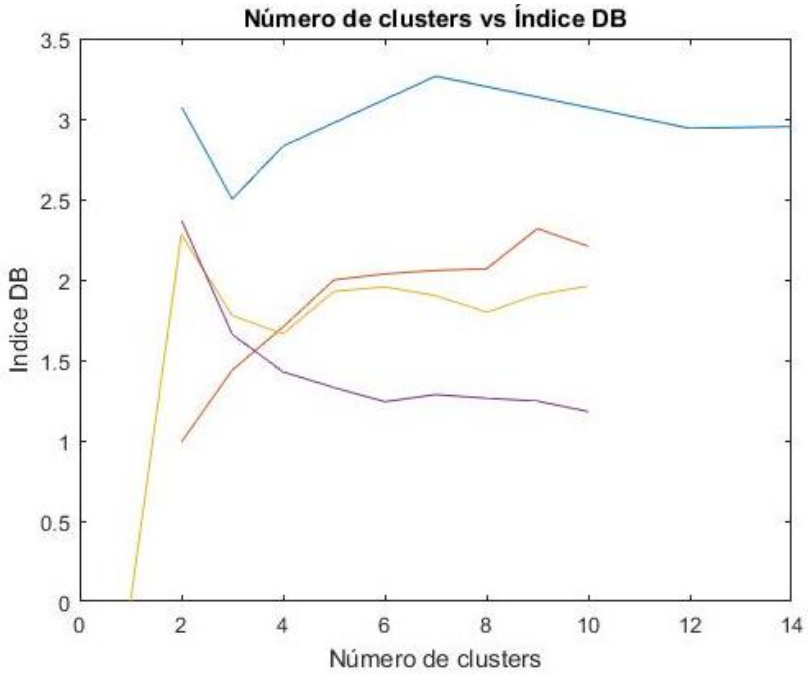

Ilustración 20 Número de grupos tercer escenario

En la tabla 37 se hace el resumen del agrupamiento desarrollado, en el cual se evidencia que el algoritmo con mejor agrupación dentro de los escenarios planteados es la red neuronal ART, el cual presenta menor variación dentro de los tres escenarios. Además como se mencionó anteriormente éste arrojo un resultado de tres grupos para dos de los tres escenarios planteados, indicando que tres es la mejor opción para la agrupación ya que se puede asociar a una clasificación mediante triage, en la cual se clasifican los registros suministrados por el hospital Santa Clara en tres grupos de riesgo, permitiéndole al personal médico centrar su atención en aquellos pacientes con mayor riesgo.

Datos	Som+kmeans	Índice DB	Kmeans	Índice DB	Art	Índice DB	FuzzyArt	Índice DB
2017	10	1.269	7	1.5	3	1.937	4	1.867
2018	10	1.1	4	1.52	5	2.424	9	2.138
2017 y 2018	6	1.23	8	1.845	3	2.5	5	1.99

Tabla 37 Número de grupos vs índice de DB, para cada escenario

Partiendo de tres grupos de riesgo y teniendo en cuenta que el primer escenario planteado contenía información sobre la condición de egreso de los pacientes, se procedió a generar el mapa de la red SOM, para visualizar la posición de las neuronas, con respecto a la variable condición de egreso. La identificación de los tres grupos de riesgo: alto, medio y bajo, se puede observar en la tabla 38.

	Zona 1.Alto riesgo
	Zona 2.Medio riesgo
	Zona 3.Bajo riesgo

Tabla 38 Identificación de grupos

Zona1. Pacientes que fallecen a causa de la enfermedad TB

Zona2. Pacientes con riesgo medio a causa de TB

Zona3. Pacientes que logran terminar el tratamiento de TB

En la figura 21 se observa la U-matriz, en la cual se puede evidenciar en donde se encuentran los tres niveles de riesgo.

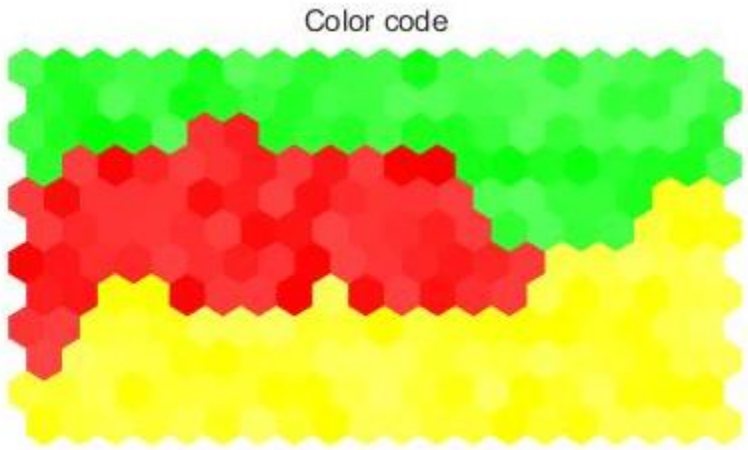

Ilustración 21 mapa agrupado con respecto a la tabla de agrupamiento

Posterior a esto, se obtiene la proyección del mapa para cada categoría de la variable condición egreso, logrando que el personal de la salud al observar la imagen pueda asociar a que zona pertenece el paciente y así prestar mayor atención a cuyos pacientes estén en mayor riesgo de morir, con la posibilidad de contrarrestar la causa fatal de convivir con ésta enfermedad.

En la figura 22 se tiene la categoría de fallecido durante el tratamiento, como se observa la zona 1 presenta mayor activación de las neuronas para un total de seis, lo cual confirma que ésta corresponde a aquellos pacientes que fallecen sin vencer la TB. Cabe resaltar que las otras neuronas activas dentro del mapa corresponden a falsos positivos, ya que se está evaluando la categoría de fallecimiento.

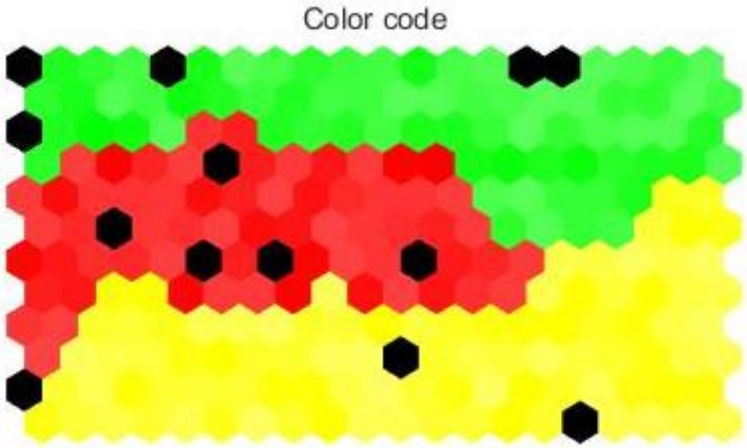

Ilustración 10 Categoría fallecido durante el tratamiento

En la figura 23, se encuentra la categoría de los registros de cuyos pacientes no se conoce la condición de su salida del hospital Santa Clara, al observar la proyección del mapa se evidencia que la región con mayor activación se encuentra ubicada en el centro del mapa entre las zonas 1 y 3.

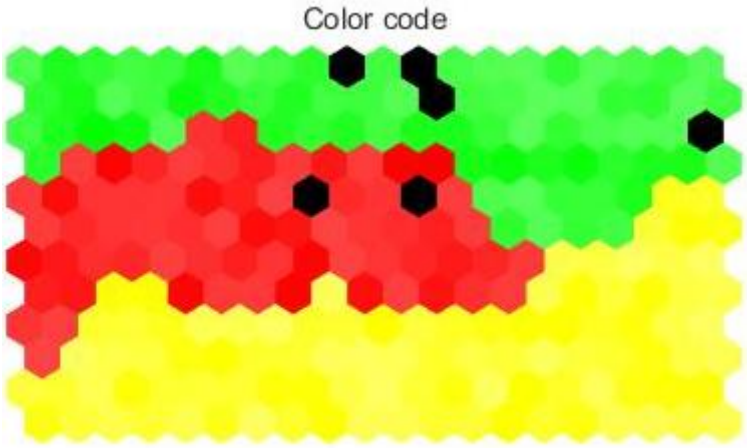

Ilustración 11 Categoría no reportado

En la figura 24 se tiene la categoría pérdida en el seguimiento, la cual corresponde a todos los pacientes con los que no fue posible continuar con el tratamiento, ésta región se encuentra distribuida por todas las zonas del mapa.

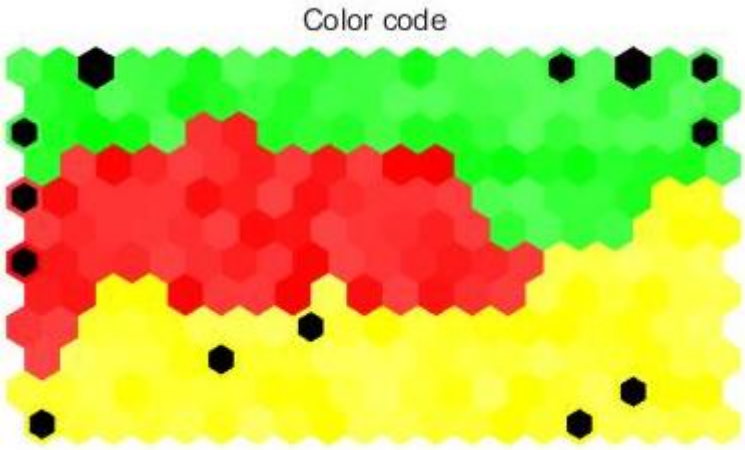

Ilustración 12 Categoría pérdida en el seguimiento

En la figura 25 se evidencia que la zona 2 con doce neuronas activas, corresponde a aquellos pacientes que logran culminar su tratamiento y superan la enfermedad TB.

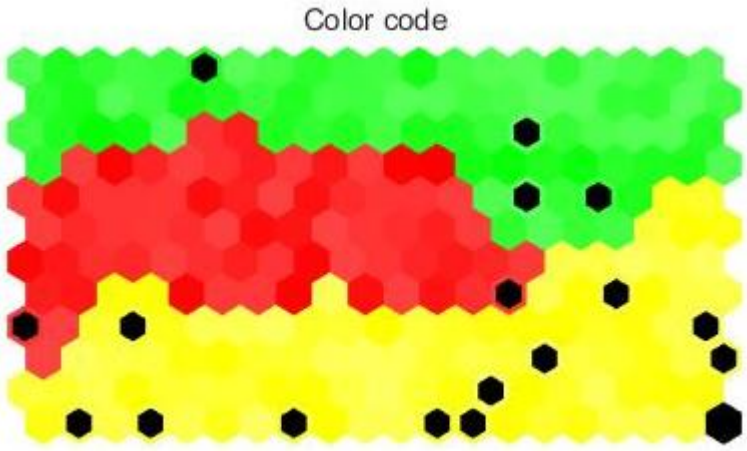

Ilustración 13 Categoría tratamiento terminado

En la figura 26, se evidencia que la región que hace parte a los pacientes curados de TB, corresponde a una parte derecha del mapa y están distribuidos entre la zona dos y tres.

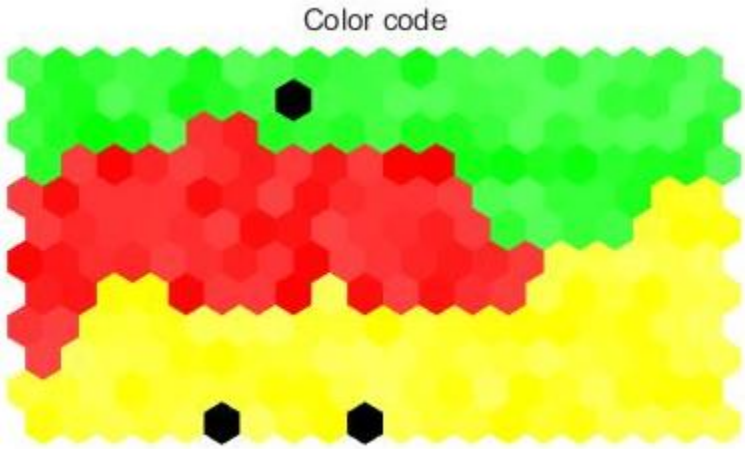

Ilustración 14 Categoría pacientes curados

En la figura 27, se muestra que la zona tres corresponde a aquellos pacientes que por un motivo u otro fueron apartados del tratamiento.

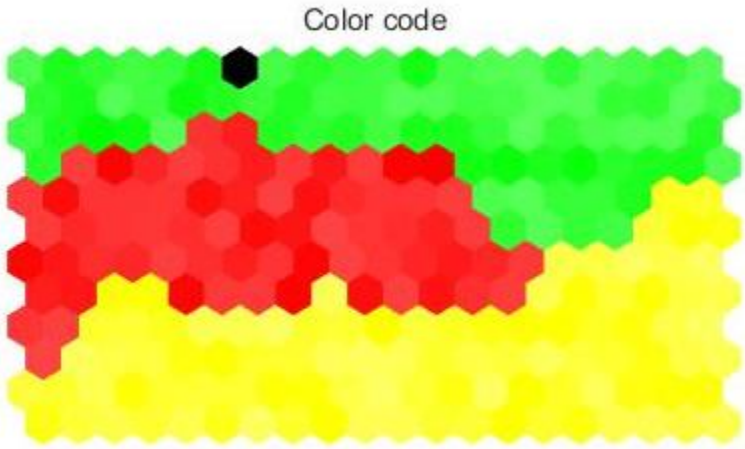

Ilustración 15 Categoría excluido por la cohorte

Por último en la figura 28, se observa que los registros de los pacientes que no fueron evaluados por el hospital, activan algunas neuronas de la parte izquierda del mapa de las zonas 1 y 2.

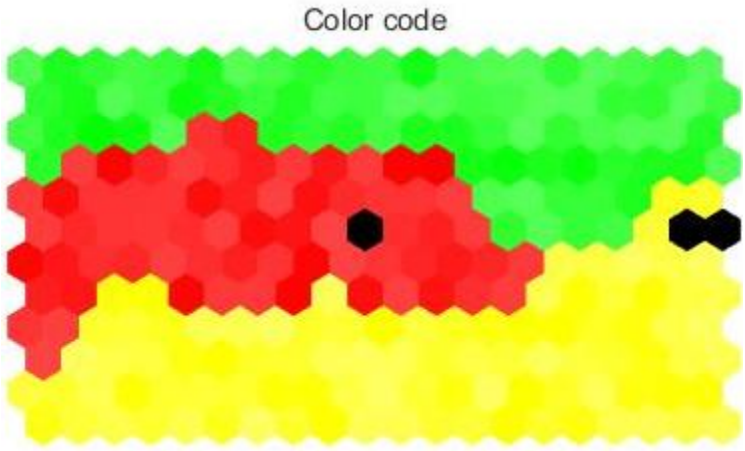

Ilustración 28 Categoría paciente no evaluado

5. DISCUSIÓN

Con la información anterior del cómo se activan las neuronas dentro del mapa de la red neuronal SOM, con respecto a la agrupación dentro de los tres grupos de riesgo y de la condición de egreso del hospital. Se procede a usar dicha información para observar el comportamiento de los dos escenarios restantes los cuales como se mencionó anteriormente no contaban con ésta variable.

Esto con el fin de brindarle una herramienta al personal de la salud, en la cual se pueda evidenciar los tres grupos de riesgo, para facilitar el proceso de atención médica a aquellas personas que sufren de ésta enfermedad, esto sin importar a que año pertenezcan los datos que serán usados para el entrenamiento.

- **Base de datos 2018**

Para el segundo escenario planteado durante este trabajo, se observa en la figura 29, que aun sin tener las etiquetas de la condición de egreso del hospital, el comportamiento del mapa, sigue siendo igual a lo anteriormente dicho, ya que como se observa para la información del año 2018, hubo una mayor cantidad de neuronas activas, lo cual indica que el porcentaje de personas que pueden fallecer durante el tratamiento fue mayor con respecto al del año 2017. Es importante mencionar que el comportamiento de estos datos fue más compacto que en el año anterior, ya que no aparecen aquellas neuronas correspondientes a falsos positivos.

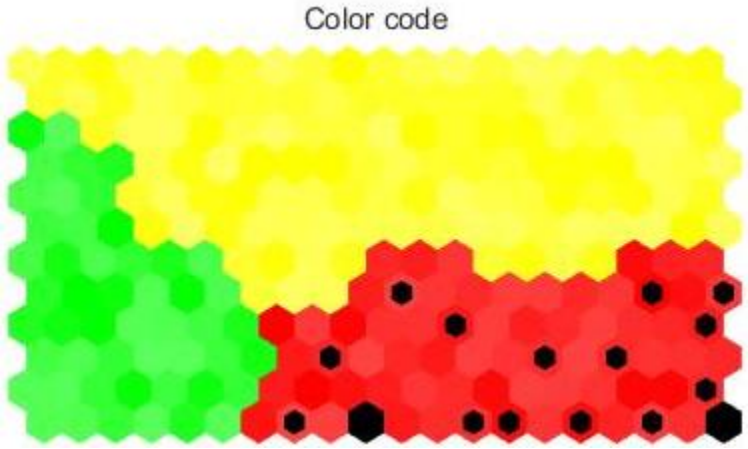

Ilustración 16 Categoría fallecido durante el tratamiento año 2018

Ahora bien, en la figura 30, se evidencia las personas que durante el año 2018, asistieron al hospital y fueron diagnosticados con TB, pero culminaron su tratamiento sin consecuencias fatales como el fallecimiento. Como bien se evidencia éstos pacientes caen en la zona de bajo riesgo, lo cual indica que el personal de salud puede estar más tranquilo después de visualizar esta información ya que estos registros no harían parte de los casos fatales por TB.

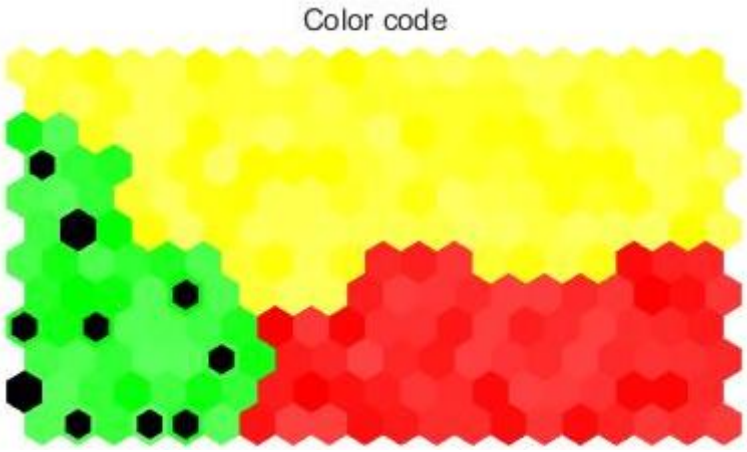

Ilustración 17 Categoría tratamiento terminado año 2018

Finalmente es importante observar el comportamiento de los registros del año 2018, que corresponden al porcentaje de pacientes curados, como se evidencia en la figura 31, las neuronas que se activan corresponden a la zona dos.

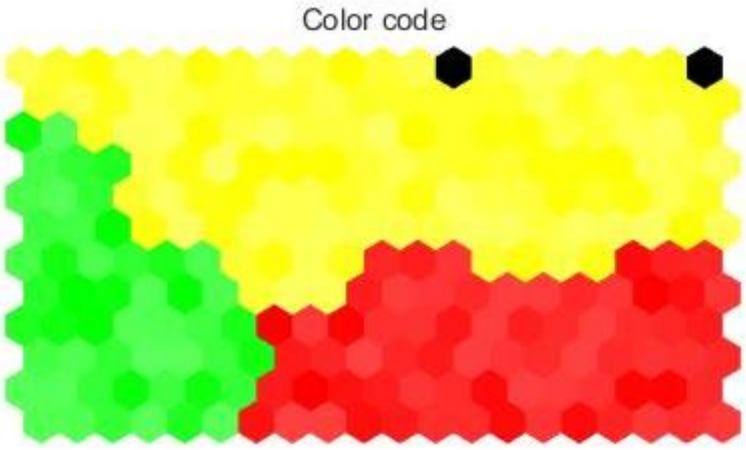

Ilustración 18 Categoría pacientes curados año 2018

- **Base de datos años 2017 y 2018**

En la figura 32, se observa que las personas que pueden fallecer durante el tratamiento, activan de nuevo la zona 1 correspondiente a alto riesgo.

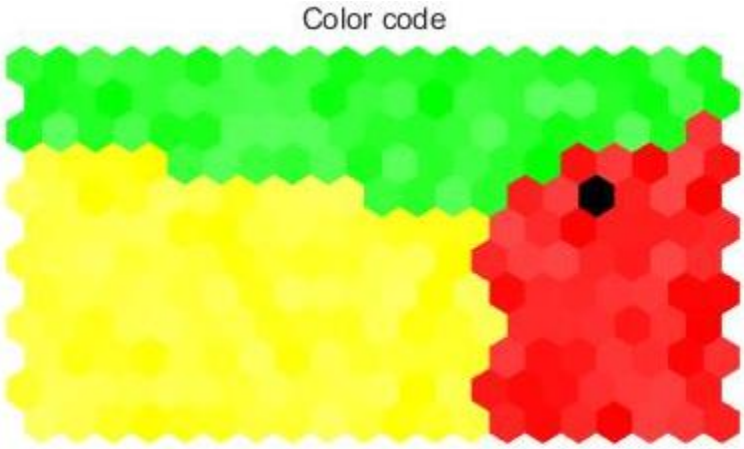

Ilustración 19 Categoría fallecido año 2017 y 2018

En la figura 33, se observa que para el tercer escenario, las neuronas que se activan para aquellos pacientes que tienen posibilidad de terminar el tratamiento caen dentro de la zona 2, es decir pueden presentar un estado de salud complicado, más no con causa fatal.

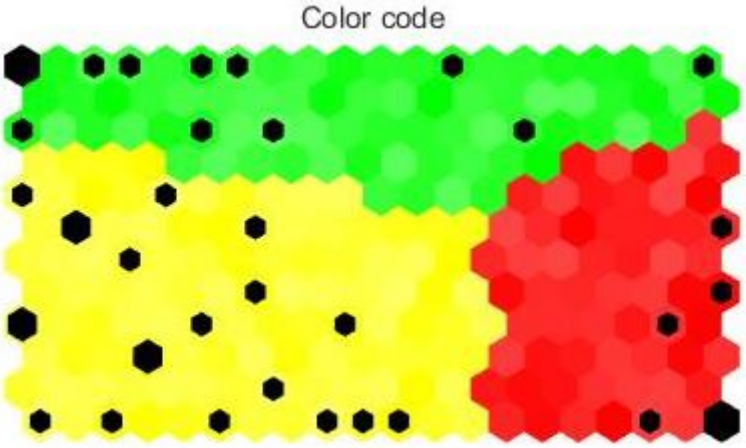

Ilustración 20 Categoría tratamiento terminado año 2017 y 2018

Finalmente, en la figura 34 se observa que los registros de aquellos pacientes que hacen parte de bajo riesgo, es decir que se curan de TB, siguen recayendo dentro de la zona 3.

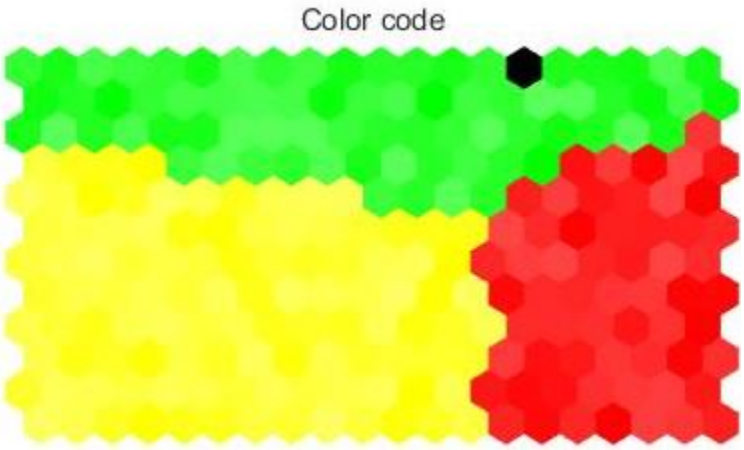

Ilustración 21 Categoría pacientes curados años 2017 y 2018

Comparando el trabajo desarrollado muestra similitudes, con estudios revisados en el estado del arte, como lo son el realizado en Rio de Janeiro y el apoyado por la Universidad Antonio Nariño. Como similitud se tiene que en los tres trabajos el agrupamiento es él mismo después de realizar el entrenamiento de las redes neuronales. Aunque los estudios planteados tienen como objetivo diagnosticar la TB pleural, éste en particular se enfoca en generar una herramienta para que el personal de la salud pueda observar dentro de todos sus pacientes a cual se le puede brindar una atención médica efectiva y oportuna evitando así que la curva exponencial de muertes a causa de TB siga creciendo. Además esta visualización de sus pacientes generada por el mapa agrupado, permite disminuir el tiempo invertido por el profesional de la salud, evitando que éste tenga que hacer un análisis exhaustivo sobre todos los registros de los pacientes, enfocándose sólo en aquellos que presenten alto riesgo de fallecimiento, es decir que se encuentren en la zona 1 del mapa.

6. CONCLUSIONES

Con el desarrollo del presente trabajo se logra apoyar al profesional de la salud, brindándole una herramienta en la cual pueda obtener de manera gráfica, la información de aquellos pacientes con TB que necesiten mayor atención o cuidado de parte del profesional, para así enfocar la atención y evitar más muertes por esta enfermedad.

Las redes neuronales artificiales pueden ayudar a encontrar soluciones a diversos problemas para los cuales no existen algoritmos que den soluciones eficientes en tiempos razonables, son útiles en diversos tipos de problemas, en este caso se evidenció que se pueden aprovechar para el diagnóstico de la tuberculosis pulmonar ya que actualmente es la segunda enfermedad con morbilidad a nivel mundial, pero se puede contrarrestar la curva gracias al agrupamiento realizado en este trabajo.

El presente trabajo muestra como las redes neuronales artificiales de aprendizaje no supervisado, donde no se tienen etiquetas de los datos, proporcionan una herramienta visual que puede apoyar al profesional en salud a realizar un diagnóstico con poca información. En el escenario ART se pudo evidenciar esa ayuda, debido a que se realizó un agrupamiento que puede representar un triage de tres grupos de riesgo.

En el desarrollo del trabajo se evidenció que mientras más información se le agregue al entrenamiento de las redes neuronales, éstas pueden llegar hacer más precisas, permitiendo obtener mejores resultados, pero se debe tener especial cuidado a la hora tanto de seleccionar la cantidad de variables como de la binarización de éstas, ya que la información puede ser excesiva y puede generar una confusión a la hora del agrupamiento.

7. REFERENCIAS

[1] World Health Organization WHO, “Global tuberculosis report 2019”

[2]Tuberculosis. Disponible en:https://medlineplus.gov/spanish/tuberculosis.html

[3] Tuberculosis. Disponible en: https://www.who.int/es/news-room/fact-sheets/detail/tuberculosis

[4] I, N,S. Boletín epidemiológico semanal. Disponible en: https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2019%20Bolet%C3%ADn%20epidemiol%C3%B3gico%20semana%2011.pdf

[5] Metrosalud. Genexpert, una herramienta para el diagnóstico de la tuberculosis ahora en Metrosalud. Disponible en: http://www.metrosalud.gov.co/actualidad/noticias/genexpert-una-herramienta-para-el-diagnostico-de-la-tuberculosis-ahora-en-metrosalud

[6] Instituto de ingeniería del conocimiento. Disponible en: https://www.iic.uam.es/lasalud/realidad-inteligencia-artificial-salud/

[7] La IA predice un futuro más saludable para América Latina. Disponible en: https://www.scidev.net/america-latina/gobernanza/especial/la-ia-predice-un-futuro-mas-saludable-para-america-latina.com

[8] Machine learning. Disponible en: https://empresas.blogthinkbig.com/que-algoritmo-elegir-en-ml-aprendizaje/

[9] K.R.Lakshmi, M.Veera Krishna, S.Prem Kumar: Utilization of Data Mining Techniques for Prediction and Diagnosis of Tuberculosis Disease Survivability. In: Modern Education and Computer Science Press (MECS Press), 2013. Pp. (8-17).

[10] Ciampi , A,Lechevallier, Y, Brazil: Statistical Models and Artificial Neural Networks: Supervised Classification and Prediction Via Soft Trees.pp(275-283).

[11] Asha.T, S. Natarajan, and K.N.B. Murthy: A Data Mining Approach to the Diagnosis of Tuberculosis by Cascading Clustering and Classification.

[12] Mahmoud, Reza S, Shahaboddin S, Shahram Golzari H, Teh Ying W, Saeed A, Mohamad Amin P, T Olariu. Diagnosing Tuberculosis With a Novel Support Vector Machine-Based Artificial Immune Recognition System. 2015

[13] Navneet,Walia, Harsukpreet,Singh Sharad, Kumar Tiwari,Anurag Sharma: A Decision Support System For Tuberculosis Diagnosability.In: International Journal on Soft Computing (IJSC). Vol.6, No. 3, August 2015

[14] Rusdah, Winarko, E: Review on Data Mining Methods for Tuberculosis Diagnosis. In: Review on Data Mining Methods for Tuberculosis Diagnosis.

[15] Jerome, Gumpy,M, Ibrahim G, Mohammed I. Neuro-Fuzzy Approach For Diagnosing And Control Of Tuberculosis. In: The International Journal of Computational Science, Information Technology and Control Engineering (IJCSITCE) Vol.5, No.1, 2018.

[16] Guénaël, C,Younès B. Learning the number of clusters in Self Organizing Map. Disponible en:https://lipn.univ-paris13.fr/~bennani/PUBLICATIONS/Cabanes/Cabanes_InTech10.pdf

[17] Los mapas autoorganizados de Kohonen. Disponible en: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema5dm.pdf

[18] L.G, Heinz. Adaptive Resonance Theory. Disponible en: https://scholarsmine.mst.edu/cgi/viewcontent.cgi?article=1173&context=comsci_facwork

[19] Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. Disponible en: https://dl.acm.org/doi/abs/10.1016/j.neunet.2012.09.017

[20] Non, T, H. Duy, B.An Improved Learning Rule for Fuzzy ART. In: JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 30, 713-726 ,2014.

[21] Michael,J,G..understanding kmeans clustering in Machine Learning. Disponible en: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

[22] Daniel, O, F. desarrollo de arquitecturas especializadas para sistemas de conucción inteligente. (Septiembre del 2017). Universidad de Alicante. Disponible en: https://rua.ua.es/dspace/bitstream/10045/70314/1/TFG-Daniel-Ortega.pdf

[23] Optimización de resultados mediante algoritmos de selección de variables. Disponible en:https://www.tdx.cat/bitstream/handle/10803/8451/6_CAPITULO2%28Cristhian%29.pdf?sequence=8&isAllowed=y

[24] Elizabeth, L, G.Métricas para la validaión de clustering. Disponible en: https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

[25] Elizabeth, L, G.Métricas para la validaión de clustering. Disponible en: https://disi.unal.edu.co/~eleonguz/cursos/mda/presentaciones/validacion_Clustering.pdf

[26]Tuberculosis.Disponible en: https://www.cdc.gov/tb/esp/default.htm

[27] Redes neuronales. Disponible en: http://avellano.fis.usal.es/~lalonso/RNA/index.htm

[28] Ivonne,G,G,(25 Septiembre 2013).algoritmo SVM para problemas sobre big data. Universidad Autónoma de Madrid.

[29] Ligdi, G. Naive bayes.Disponible en: https:ligdigonzalez.com/naive-bayes-teoria-machine.learning

[30] Orjuela-Cañon, A.D., de Seixas, J.M., Trajman, A.: SOM Neural Networks as a Tool in Pleural Tuberculosis Diagnostic. In: Braga, A. de P. and Bastos Filho, C.J.A. (eds.) Annals of the 11th Brazilian Congress on Computational Intelligence. pp. 1–5. SBIC, Porto de Galinhas, PE (2013).

[31] Orjuela-Cañon, A.D., de Seixas, J.: Fuzzy-ART neural networks for triage in pleural tuberculosis. In: Health Care Exchanges (PAHCE), 2013 Pan American. pp. 1–4 (2013)