BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Análisis de Factores que inciden en Infecciones de Transmisión Sexual en el Estado de Puebla

TESIS

QUE PARA OBTENER EL GRADO DE LICENCIATURA EN INGENIERÍA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA

SELENE VEGA ROMERO

ASESOR

DRA. MARÍA JOSEFA SOMODEVILLA GARCÍA

PUEBLA, Diciembre 2015

Agradecimientos

Quiero agradecer principalmente a Dios, por todos los buenos momentos que me ha dado en la vida, y por los tropiezos que me ha puesto en los que he descubierto, que soy una mujer luchona y valiente, ya que día a día hay metas que lograr y gracias a él las hago con sabiduría.

A mis hermosos padres Edith y Albino, que me brindaron la vida con todo su amor, que sé que aunque se enojen, nunca me dejarán en ningún instante de la vida, y siempre compartiremos todos los momentos ya sea para bien o para mal, como una verdadera familia.

A mis hermanos Luis y Guadalupe, que sé que definitivamente no sería lo mismo sin ellos, y que en cada momento están ahí para mí y yo para ellos.

A mi pequeño motor, mi hijo Gabriel… Sin él yo no sería la mujer valiente que soy ahora, ese niño que es mi vida y alegría entera, Te Amo chaparro.

Y a mí profesora y asesora la Dra. María Josefa Somodevilla, que confió en mí en cada momento, a quien le tengo todo mi respeto y admiración por ser la persona que me enseñó tanto, y quien me motivó a finalizar en este gran paso de mi vida profesional.

Muchas gracias por sus apoyos, motivaciones y principalmente amor en todo momento.

Resumen

Las enfermedades de Transmisión Sexual (ETS), son un grupo heterogéneo de padecimientos que se adquieren a través de las relaciones sexuales coitales sin responsabilidad, generalmente como una consecuencia de una falta de educación, y conocimientos sobre la sexualidad humana por parte de los padres y de los educadores. Se tiene un estimado que al día se presentan 685,000 infecciones y 250 millones al año.

Las ETS son un problema de salud pública en todo el mundo; sin embargo, son los países en desarrollo los que se ven principalmente afectados. La negativa de los gobiernos para admitir la existencia de las altas prevalencias de las ETS, la falta de servicios con experiencia para atender estos padecimientos, la limitación para realizar exámenes de detección de ETS a las personas que buscan atención para otros problemas de salud, son condiciones que prevalecen.

Hasta la fecha no existe en el horizonte alguna perspectiva de disminución de las ETS como problemas de salud. La diversidad son agentes infecciosos, la dificultad para obtener vacunas, la existencia de problemas de tratamiento y especialmente por lo resistentes que pueden llegar a ser las bacterias, virus etc.

Todas las enfermedades infecciosas son el resultado de una interacción entre un agente causal al que se atribuye el desencadenamiento de la enfermedad y una persona vulnerable. No todas las enfermedades infecciosas son contagiosas, en realidad son provocadas por microorganismos que pueden transmitirse del enfermo o incluso de un portador no enfermo al individuo sano, siempre que este sea susceptible de padecer la enfermedad.

La importancia de esta tesis recae en describir las condiciones de cada tipo de ETS, en las cuales se desarrolla cada una de ellas y, cada lugar donde se encuentran más personas infectadas para así poder proponer soluciones factibles desde el punto de vista de las tecnologías de la información. La información fue colectada en centros de salud, datos online y el Hospital General de Agua Santa. Con estos datos se seguirá una metodología, la cual comenzará por la selección y limpieza de los datos, seguido del diseño de la base de datos. Posteriormente, se inicia el análisis con la asistencia del software de distribución libre WEKA que contribuye a la visualización y análisis de datos a

partir de algoritmos de aprendizaje que tienen como resultado modelos descriptivos y predictivos.

Índice General

Índice de Figuras

Índice de Tablas

CAPÍTULO 1

1. Introducción

A lo largo de muchas décadas, se ha buscado combatir las ETS (Enfermedades de Transmisión Sexual) y a la fecha hay muchas que no tienen cura o solo se encuentra el tratamiento para poder controlarlas como el VIH/SIDA. Se sabe que la cantidad de personas infectadas ha crecido conforme pasa el tiempo y gracias a las bases de datos que se tienen en los centros de salud y hospitales, se cuenta con los informes emitidos por organizaciones de Salud gubernamentales.

Partiendo de lo anterior, en la actualidad se utilizan técnicas y arquitecturas las cuales nos permiten conocer con exactitud los problemas más comunes que afronta la sociedad respecto a las ETS. El procesamiento de grandes volúmenes de datos y la extracción de conocimiento tiene el propósito de ayudar en la toma de decisiones estratégicas que influyen directamente en el mejoramiento del desempeño de los centros de salud y hospitales. En particular se trabajó con datos emanados del Estado de Puebla que fueron procesados con Weka.

El ***Data Mining*** o extracción de información útil y no evidente de grandes bases de datos es una tecnología con un gran potencial para ayudar en este caso a los hospitales a enfocar su información alrededor de la información importante contenida en las bases de datos [1].

WEKA es una poderosa herramienta y es muy fácil de utilizar. Nos permite tener conocimiento y aprender de esta con la minería de datos pues maneja un gran volumen de estos para la toma de decisiones y aquí el depósito de cada dato es confiable con resultados de calidad. Las aplicaciones de dichas herramientas facilitarán el problema de acceso a la información y en consecuencia, se acelerará el análisis y consulta y será mucho menor el tiempo en que se tendrá uso de la información además de que al mismo tiempo será más entendible.

En este proyecto se aborda el aprendizaje de datos que contiene un historial clínico. Estos datos son reales y corresponden a pacientes los cuales han sido atendidos en los centros de salud y hospitales del estado de Puebla. Entre otros datos se especifica la vía de adquisición de la ETS, es decir, mediante la relación sexual, vía sanguínea o vía postnatal.

Para tener información más clara y precisa será necesario pasar por una serie de fases los cuales serán los siguientes y en el primer capítulo serán explicados a detalle [2].

- Filtrado de Datos
- Selección de Variables
- Extracción de conocimiento
- Interpretación y Evaluación

1.1 Planeamiento de la Investigación

En esta sección se precisa el problema de la investigación a resolver, se definen los objetivos del proyecto tanto generales como particulares, al mismo tiempo se plantea la propuesta de solución y por último se describe la organización de la tesis para así obtener una propuesta de solución.

1.1.1 Problema a Resolver

El problema que se aborda en este proyecto, tiene sus orígenes en datos obtenidos por el HOSPITAL GENERAL DE PUEBLA AGUASANTA e información relevante de hace algunos años que se encuentra disponible en INTERNET (sus ligas se muestran en la bibliografía). Las fuentes de datos tienen la intención de dar a conocer los programas de salud y a las personas principalmente, cifras actualizadas acerca del problema que se tiene con las ETS, los datos obtenidos se agruparán de acuerdo a la región, las edades, el sexo y los síntomas.

Debido al gran volumen de información, se aplicará un proceso de extracción de conocimiento con el objetivo de sentar bases para disminuir el impacto con que las ETS están atacando desde hace ya varios años especialmente en jóvenes menores de 30 años y al mismo tiempo contribuir en la disminución del riesgo de contagio en un futuro.

1.1.2 Objetivos de la Investigación

El objetivo general de la siguiente tesis es la siguiente:

Aplicar técnicas de Minerías de Datos para implementar un sistema para el análisis enfermedades de Transmisión Sexual en el estado de Puebla.

Los objetivos particulares se especifican a continuación:

- Investigar el estado del arte relativa a ETS y a Técnicas de minería de Datos.
- Separar la información sobresaliente para describir las condiciones en que se desarrollan las ETS.

- Realizar un análisis de datos de las ETS más comunes considerando una división por localidades del Estado de Puebla.

1.1.3 Justificación de la Investigación

A continuación se presentan las razones que justifican este trabajo de tesis:

1.-Las ETS son un problema que afectan a cada país en el mundo, no importa el nivel socioeconómico de las personas, si son menores o mayores de edad, cada ETS ataca a las personas por el tipo de vida sexual que llevan sin ninguna protección e incluso puede atacar a los hijos de las mujeres infectadas que está embarazadas o amamantando. Como se mencionó anteriormente, estas enfermedades venéreas atacan a la población en general.

2.-En México son consideradas un problema de salud pública desde mediados de los años 80 y a partir de entonces ha ido incrementando día a día el índice de personas infectadas.

3.-Se identificarán las causas que provocan las altas tasas de ETS principalmente en jóvenes menores hasta adultos jóvenes. Esto con el fin de describir las causas que provocan las ETS a partir del comienzo de su vida sexual y teniendo en cuenta el número de parejas sexuales que se han tenido.

4.-Ante la duda se saber realmente el impacto que tienen las ETS respecto a las personas que están infectadas, se investiga la aplicación de nuevas herramientas para el tratamiento de información (bases de datos, Data Mining), que mejoren la calidad, cantidad y eficiencia de los datos, así como el análisis, procesamiento y comunicación de los mismos.

5.-Se puede mantener al tanto del número de personas infectadas, con morbilidad y mortalidad a las personas que no están infectadas para que tomen conciencia de su vida sexual activa en la actualidad y para un futuro.

6.-Con estas nuevas herramientas podríamos ofrecer una proyección a futuro de las ETS. Las instituciones que lo deseen puedan extraer dichos datos, información y el conocimiento que necesitan para identificar las causas que provocan las ETS y así establecer alguna estrategia y enfrentar retos para disminuir epidemias futuras de mayor envergadura.

1.2 Presentación de la Solución

A continuación se presenta la propuesta de la solución al problema definido en la sección 1.1.1, donde se describen las herramientas con las cuales se cumplirá el propósito de esta tesis.

1.2.1 Propuesta de Solución

De acuerdo al problema ya expuesto, se presenta la siguiente solución:

Se propone el diseño, construcción e implementación de una base de datos para identificar patrones para la toma de decisiones con datos específicos que den como resultado, un análisis experto mediante técnicas adecuadas de Minerías de Datos.

El sistema se basa fundamentalmente en la creación de una base de datos siguiendo la metodología KDD[1], la cual inicia con la selección, pre procesamiento y transformación de datos. También se requiere la ayuda de la herramienta WEKA[2] 3.7, el cual es un software de experimentación en el análisis de datos que ofrece un aprendizaje automático para el descubrimiento de patrones.

[1]Knowledge Discovery in Databases

[2]Waikato Environment for Knowledge Analysis

1.3 Aportaciones a la investigación

Uno de los objetivos principales de esta tesis es describir la información que se obtiene con el uso de técnicas de minería de datos como consecuencia de los tipos más comunes de ETS, y que aporte información para apoyar a los procesos de toma de decisiones en el área de salud.

1.4 Organización de la Tesis

Este trabajo está dividido en 6 capítulos, los cuales estarán presentados de la siguiente manera:

1. **Capítulo 1. Introducción:** En esta parte se describe el problema y se plantea la forma de resolución, los objetivos que se desean alcanzar, la propuesta de solución al problema, la justificación del proyecto y los resultados obtenidos.

2. **Capítulo 2. Estado del Arte:** Aquí se muestran detalladamente los trabajos realizados en torno al tema de las ETS en el Estado de Puebla, respecto a la Minería de Datos.

3. **Capítulo 3. Marco Teórico:** En este apartado se presentan conceptos relacionados con Minería de Datos, así como información básica de las herramientas que se utilizarán.

4. **Capítulo 4. Análisis y Diseño:** En este capítulo se muestra el preprocesamiento de los datos, el proceso de selección y limpieza de datos. También se presenta el diseño de la base de datos detallando cada atributo siendo utilizado.

5. **Capítulo 5. Resultados:** Aquí se muestran los resultados obtenidos con una visión general del desarrollo al aplicar Minería de Datos y se proponen futuras investigaciones para mejorar los resultados.

6. **Capítulo 6. Conclusiones y Trabajos Futuros:** Finalmente se presenta el conocimiento generado en base a los resultados obtenidos, el cual sería un punto de partida para ofrecer una propuesta de solución para mitigar las ETS.

Finalmente se muestran las referencias que aportan las bases teóricas y prácticas a este trabajo.

1.5 Conclusiones

Cabe destacar que se cuenta un gran número de personas que no son detectadas a tiempo con ETS, y también pacientes infectados en centros de salud y hospitales. Las ETS son temas muy importantes para la población mexicana. Como ya mostraron los datos reales, diariamente el número de personas infectadas va en incremento debido a la poca información que se tiene. La falta de información es un factor muy importante para incrementar las tasas de ETS, su morbilidad y mortalidad y no existe información adecuada dirigida a los jóvenes para combatirlas a tiempo.

Debemos entonces concientizar a la población para poder disminuir las ETS. Se cuentan muchos tipos de ETS pero 7 son las que más predominan en la actualidad, todas pueden ser mortales si no se detectan o se tratan a tiempo. Para el VIH (Virus de Inmunodeficiencia Humana) por ejemplo, no se ha encontrado cura alguna aunque se puede mejorar la salud de la persona infectada y prolongar su tiempo de vida con antiretrovirales. Las tecnologías aquí propuestas servirán para mejorar la información relativa a este tema, como lo es la minería de datos, sobre todo ayudará a reconocer patrones relacionados con las ETS, y de esta forma anticipar soluciones.

CAPÍTULO 2

2. Estado del Arte

Desde épocas coloniales, se relacionó la prostitución con enfermedades venéreas, en ese periodo se creó el primer hospital especial para atender los casos de las ETS y a partir de entonces, se realizaron y aplicaron las primeras medidas preventivas. Organismos internacionales como la OMS dan cifras alarmantes al plantear que anualmente se producen más de 250 millones de nuevos casos con ETS. Una de cada 20 personas padece alguna enfermedad sexual anualmente [2].

Hipócrates fue el primero en describir las ITS en el año 460 A.C. describiendo lesiones genitales duras, suaves, secundarias al contacto sexual. Las enfermedades venéreas en el pueblo azteca eran conocidas como *Ciuatlaueliloc* y utilizaban la raíz de una hierba que era útil para el dolor de pecho, curar la fiebre y para personas con sangrados y pus en la orina.

Infecciones de Transmisión Sexual =! Enfermedades de Transmisión Sexual
La Infección es diferente de la Enfermedad, ya que esta es la invasión del organismo mediante el microorganismo patógeno, el cual puede producir o no daño al humano mientras que la enfermedad se presenta una vez que el humano ya ha sido infectado.

Ej. Una infección puede estar presente sin que haya síntoma como el VIH.

La Organización Mundial de la Salud (OMS) recomienda como estrategia fundamental para la prevención de VIH/SIDA el diagnóstico oportuno, tratamiento e información de las ETS. Las ITS son causadas por bacterias, virus y parásitos. Según informes de la OMS en el 2011, cada año se producen 448 millones de nuevos casos curables de ITS y la mitad de estas son en personas de entre 15 y 29 años, en la actualidad se estima un total de 34 millones de personas infectadas por VIH/SIDA de las cuales en 2011 se infectaron unos 2.5 millones de personas [3]. En México, las ITS ocupan uno de los cinco primeros lugares de demanda de consulta y se ubica entre las primeras diez causas de morbilidad general [4].

La infección anogenital por el virus del papiloma humano (VPH), es la infección de transmisión sexual más frecuente en todo el mundo. Se han identificado alrededor de 100 genotipos de los cuales unos 40 infectan la región anogenital. 15 genotipos son causantes necesarios del cáncer cervicouterino y han sido aplicados como causantes carcinogénicos de la vulva, vagina, pene, ano y región orofaríngea, los virus de bajo riesgo solo llegan a causar verrugas genitales (condilomas). Se calcula que unas 300 millones de mujeres en el mundo portan el VPH.

2.1 Proyectos relacionados con ETS e ITS

En esta sección se darán a conocer proyectos Web, las cuales dan información a las personas que desean saber acerca y más de cada una de las enfermedades, también se muestran foros donde las personas resuelven sus dudas y dan características que ellos mismos tienen acerca de las ETS. La ciencia está tan avanzada que también ya se cuentan con tecnologías para detectar ETS y en este proyecto de tesis serán explicados.

2.1.1 Proyecto STD Triage

STD Triage es una aplicación la cual permite tomar una foto de una erupción en los genitales por ejemplo, y a cambio de cierta cantidad de dinero poder enviarla y recibir en un día el probable diagnóstico de un médico. Esta aplicación examina ETS que tengan síntomas típicos, como sífilis, herpes y verrugas genitales.

Otra herramienta tecnológica contra las ETS ayuda a entablar conversaciones difíciles con sus antiguos amantes. Los estudios demuestran que el 23% de las parejas de los pacientes diagnosticados con una ETS, nunca reciben la advertencia de que también podrían estar en riesgo y para esto se creó la aplicación en la cual, los usuarios notifiquen a sus antiguas parejas sus propios diagnósticos de clamidias, gonorrea y tricomoniasis.

En la actualidad los usuarios pueden compartir los resultados confirmados de sus pruebas de detección de clamidia, VIH, gonorrea y sífilis pero por el momento solo está disponible en Estados Unidos [5].

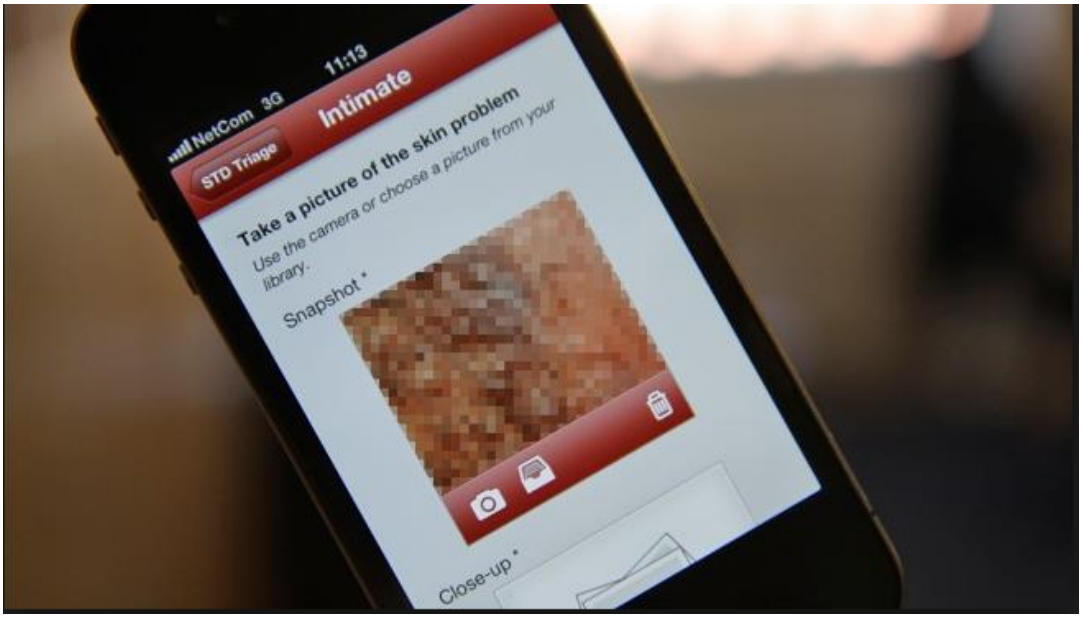

Figura 2. 1 Aplicación STD Triage

Figura 2. 2 Logo STD Triage

2.1.2 Proyecto IK4

Se trata de un dispositivo creado por la empresa *Globionic Technology* en el 2012 en la India, es un dispositivo creado para la detección de la Clamidia, el VPH y otras ETS en menos de media hora. Procura especialmente la Clamidia, la cual es una enfermedad silenciosa en el 90% de los casos, transmitida por cada contacto sexual, que repercute únicamente en las mujeres. Con alta incidencia también en Europa, la Clamidia afecta el cerebro, provoca ceguera y produce abortos, partos prematuros o rotura prematura de las membranas y es factor clave en el elevado grado de morbilidad infantil o la infertilidad.

El dispositivo de diagnóstico rápido y de bajo costo, no requiere la infraestructura de un laboratorio, lleva a cabo complejas técnicas de detección a partir de muestras clínicas y es más eficaz y rentable que otras técnicas rápidas actuales [6].

Figura 2. 3 Clamidia atacando a estómago

Figura 2. 4 Bacteria de Clamidia

2.1.3 Proyecto RCmultimedios

Se tiene otro proyecto en mente en Reino Unido, la cual consta de lo siguiente:

La detección de las ETS, es una aplicación creada por Microchip y se utiliza exclusivamente en teléfonos móviles y permite checar las ITS de forma barata y discreta. Para esto la aplicación consiste en que la persona que sospecha de una ETS coloque 2 gotas de orina o saliva sobre el chip y lo inserte en el móvil, inmediatamente la muestra será enviada al UK *Clinical Research Collaboration* y en un par de minutos será regresado el diagnóstico.

La idea es vender los dispositivos en máquinas dispensadoras donde compran condones los usuarios, haciendo la autoexaminación un asunto barato y discreto, motivando a las personas que generalmente les costaría mucho ir al doctor, y al parecer el chip funcionaría a través de micro USB [7].

Figura 2. 5 Aplicación RCMultimedios

2.2 Datos OMS

La OMS estima una cantidad de 1 millón de personas infectadas diariamente debido a las ITS. Anualmente existen unos 500 millones de personas que contraen alguna de las siguientes enfermedades: clamidiasis, gonorrea, sífilis o tricomoniasis. Más de 530 millones de personas son portadoras del virus que provoca el herpes genital tipo 2 (HSV2). Más de 290 millones de mujeres están infectadas con el VPH. Considerando las cifras y datos emanados de la OMS nos hacemos los siguientes cuestionamientos:

¿Desde qué épocas las ETS representan un problema de salud para México?
¿Existía algún control en épocas antiguas?
¿Cuáles son los tratamientos utilizados en las ETS?
¿Cuáles podrían ser las consecuencias si no son tratadas?
¿Cómo se puede prevenir las ITS?

Las Enfermedades de Transmisión Sexual se vienen incrementando a lo largo de los años, y la oferta en salud es deficiente, sobre todo en los países en vías de desarrollos. La ONU ha publicado algunos puntos importantes con relación a las ETS [8]:

Existe un vínculo muy estrecho entre las ETS y la Transmisión Sexual de la Infección por el VIH, aumentando su factor hasta por 10.

A menudo las ETS no presentan síntomas, y en las mujeres la mayor parte de las infecciones gonocócicas y clamidiales son asintomáticas.

En los países en desarrollo, las ETS y sus complicaciones incluso no contabilizando los casos de infecciones por el VIH-SIDA se encuentran entre las 5 primeras categorías de enfermedades para las que los adultos solicitan asistencia.

El objetivo de la prevención y atención de las ETS es reducir la prevalencia de esas enfermedades a través de la prevención primaria y del tratamiento de casos eficaces.

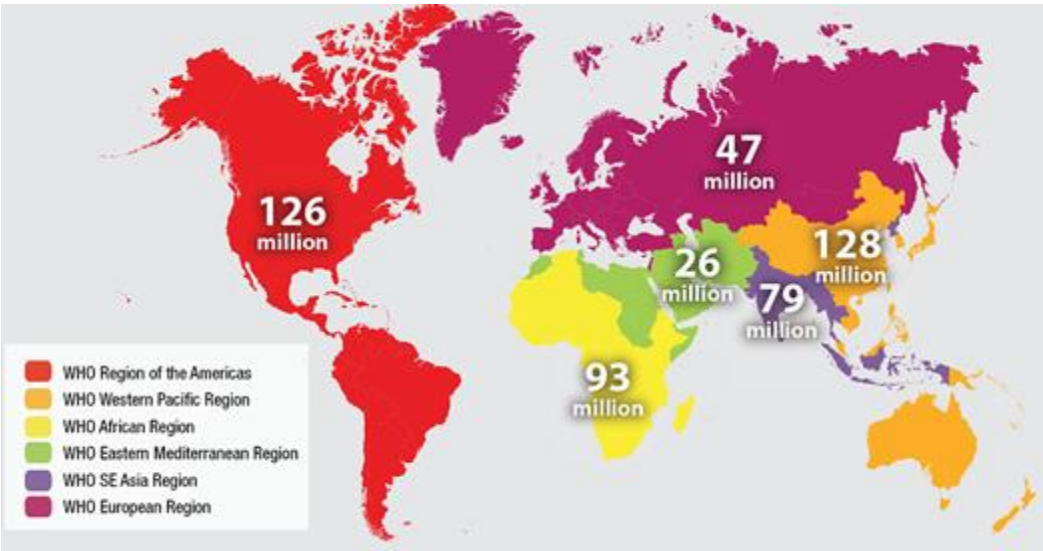

Figura 2. 6 Número de personas infectadas

2.2.1 Datos y estadísticas

Prevalencia

El VPH es una de las ITS más comunes en el mundo que los estudios coordinados por la *International Agency for Reasearch on Cancer* (IARC) en más de 18 000 mujeres mayores de 15 años en 113 países han mostrado tasas de prevalencia de ADN de VPH que van desde el 1.6% en España y Hanoi (Vietnam) hasta el 64% en Estados Unidos. La prevalencia es mayor en mujeres menores de 25 años y va disminuyendo paulatinamente hasta llegar a los niveles de la cuarta o quinta década de vida.

Incidencia

La adquisición de la infección es muy común, sobre todo en los adolescentes y adultos jóvenes. Se estima que por lo menos 75% de personas sexualmente activas puede adquirir una infección durante su vida. Se han realizado estudios que demuestran que del 50% al 60% de los jóvenes de entre 15 y 19 años y a los 4 años de haber iniciado su vida sexual adquieren el VPH. 49 de 60 mujeres de los 14 a 17 años fueron positivas con VPH lo que indica un 82% de incidencia.

Transmisión

La gran mayoría del VPH se transmite durante el acto sexual con o sin penetración. También se puede transmitir de la madre al niño durante el parto vaginal lo que causa una alta morbilidad en el niño. El factor de riesgo más alto de adquisición es el número de

compañeros sexuales, otro factor de infección es la iniciación temprana de relaciones sexuales y el consumo del cigarrillo [9].

Cerca del 30% de los mexicanos entre 18 y 30 años de edad han padecido alguna enfermedad de transmisión sexual. Diariamente son detectados y atendidos, cerca del 15% de pacientes con este diagnóstico. De acuerdo con cifras proporcionadas por el Instituto Mexicano del Seguro Social (IMSS), las enfermedades con mayor número de infecciones son la candidiasis, tricomoniasis, vulvovaginitis, sífilis, gonorrea y clamidia.

En el mundo, la OMS ha calculado una incidencia de 340 millones de casos anuales de ITS incurables. En México, se estima que la mayor tasa de incidencia en infecciones es por el Virus del Papiloma Humano (VPH), con una tasa de 23.3 casos por cada 100 mil habitantes.

Los grupos de mayor riesgo para la transmisión y adquisición de las ITS, son las mujeres sexoservidoras y los homosexuales. Otros padecimientos que han reaparecido son el SIDA, la gonorrea y el VPH, debido a la creciente apertura sexual y a la falta de información para prevenir contagios [9].

2.3 ETS y la toma de decisiones

Hasta el momento se han presentado proyectos que se han llevado a cabo en diferentes partes del mundo, y también se han mostrado números de personas que viven con morbilidad debido a las ETS en diferentes países, todo esto gracias a los datos que la OMS ha mostrado y los cuales son significativos para el tema referente a las ETS e ITS, cada uno busca formas y estrategias diferentes de disminuir las ETS, pero aún falta mucha tecnología para poder combatirlas principalmente (en este caso) en México.

Cada uno de los proyectos anteriores tiene el mismo objetivo en común… disminución de muertes, morbilidad, incrementos de las ETS etc., también se buscan cifras sobresalientes que muestre la situación en la que se encuentra la población actualmente.

El alcance de esta tesis corresponde con los datos sobresalientes de la ciudad de Puebla y sus alrededores, y se propone para su tratamiento la minería de datos. Esta herramienta se usará para estudiar detalladamente la información obtenida y posteriormente explicar el comportamiento de la población ante las ETS, y de este modo contribuir a mitigar las ITS.

2.4 Conclusiones

Varios proyectos ya están listos para la detección de las ETS, de esta forma las personas sabrán si es que están infectadas y deberían comenzar a cuidarse, la mayoría de estos

proyectos son muy económicos y finalmente lo que se tiene en común en (por lo menos) estos 3 mencionados anteriormente, es la disminución de las ITS y mejorar la calidad de vida para el caso de las personas que vivan con morbilidad, sus parejas e hijos infectados.

No cabe duda que la tecnología también va desarrollándose para combatir ETS, ampliar horizontes y mentalidades en todo el mundo, y así lograr en un futuro acabar estas enfermedades con las que personas de todas las razas enfrentan.

CAPÍTULO 3

3. Marco Teórico

Para el presente proyecto de investigación, a continuación se expondrán definiciones así como la aplicación de estos conceptos en las herramientas utilizadas. Se parte del modelo KDD que inicia con la discriminación de las fuentes de información a utilizar, y se seleccionarán, limpiarán y transformarán sus datos. Una vez preprocesados los datos, serán aplicadas las técnicas apropiadas de minería de datos, para finalmente mostrar y difundir el nuevo conocimiento.

No hay que olvidar mencionar que este modelo KDD es muy amplio, y por lo tanto es la parte más extensa de este trabajo de tesis, ya que se trata de la forma de describir el funcionamiento y la conveniencia de la aplicación de la minería de datos.

3.1 Análisis de grandes volúmenes de datos

Hoy en día los almacenes de datos, son el centro de atención para grandes empresas, ya que estos constituyen uno de los soportes fundamentales para el proceso de la toma de decisiones, de aquí que la información guardada sea confiable y con calidad. Uno de los procesos más importantes de la toma de decisiones es la limpieza de datos.

Los datos son resultados de estudios o información recabada en el pasado la cual puede provenir de diferentes fuentes. Los grandes volúmenes de información son útiles para la búsqueda de patrones de comportamiento y la toma de decisiones.

La idea de la minería de datos no es nueva, y esta era manejada como *Data Fishing, Data Mining o Data Archeology* con la idea de encontrar correlaciones sin una hipótesis previa de base de datos con ruido.

La minería de datos forma parte de la metodología KDD, la cual sugiere una serie da etapas para lograr con éxito los objetivos planteados. La minería de datos y la metodología KDD tienen metas en común como: procesar grandes cantidades de datos, identificar lo más sobresaliente y así poder presentarlos para que puedan ser utilizados en beneficio de quien lo solicite.

3.2 ¿Que es la Minería de Datos?

La estadística es la primera ciencia que históricamente extrae información de los datos básicos. Cuando se empezaron a utilizar las computadoras, surgió el concepto de *Machine Learning* traducido como aprendizaje automático. Posteriormente con el incremento de tamaño y la estructuración de los datos, es cuando se empieza a hablar de la minería de datos [10].

La minería de datos, se refiere al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utilizar los métodos de la inteligencia artificial, aprendizaje automático, estadísticas y sistema de bases de datos. Su objetivo principal consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior.

Actualmente existen más de 100 empresas en el mundo que ofrecen alrededor de 300 soluciones. La minería de datos requiere de varias fases que son necesarias y las cuales son:

- Comprensión del problema a resolver
- Determinación, obtención y limpieza de los datos a utilizar
- Aplicación de la minería de datos necesaria para los datos

3.2.1 Propósito de la Minería de Datos

Explorar los datos que se encuentra en las profundidades de las bases de datos, como los almacenes de datos, que algunas veces contienen información almacenada durante varios años.

1. En algunos casos, los datos se consolidan en un almacén de datos y en mercados de datos, en otros se mantienen en servidores de internet e intranet.

2. Las herramientas de minería de datos, ayuda a extraer el mineral de la información enterrado en archivos corporativos, o en registros públicos archivados.

3. Clasificar un dato dentro de una de las clases categóricas predefinidas, preguntas tales como: ¿Cuál es el riesgo de tener relaciones sexuales a corta edad?

4. Agrupar registros, observaciones o casos en clases de objetos similares.

5. Generar reglas en referencia al descubrimiento de relaciones de asociación.

3.2.2 Generalidades de la Minería de Datos

La minería de datos es una etapa, si bien la más importante, de lo que se ha venido llamando el proceso de extracción de conocimiento a partir de datos. Este proceso consta de varias fases e incorpora e incorpora diferentes técnicas de los campos del aprendizaje automático, la estadística, las bases de datos, los sistemas de toma de decisión, la inteligencia artificial y otras áreas de la informática y de la gestión de información.

La minería de datos o exploración de datos, es un campo referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos, se deriva de la metodología KDD y se define como el proceso que tiene el propósito de extraer y descubrir información almacenada durante varios años. Su tarea es el análisis automático o semiautomático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos.

Figura 3. 1 Ciclo de Vida de la minería de datos

La minería de datos es ya un concepto muy evolucionado que necesita ser llevado a cabo por etapas. Dentro de las etapas (Figura 3.1) que se sugieren para cumplir con éxito la extracción de información se encuentran [11]:

1. **Análisis de problema:** Incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

2. **Análisis de datos:** Recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

3. **Preparación de los datos:** Para que los datos puedan ser tratados por las técnicas del modelado. Incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato. Está muy relacionada con la fase del modelado, puesto que en función de la técnica del modelado que vaya a ser utilizada los datos necesitan ser procesados en diferentes formas.

4. **Modelado:** Se seleccionan las técnicas del modelado más apropiadas para el proyecto de minería de datos específico. Antes de procedes a esta etapa se debe establecer un diseño del método de evaluación de los modelos.

5. **Evolución:** En esta fase se evalúa el modelo, desde el cumplimiento de los criterios de éxito del problema.

6. **Explotación:** Se deben documentar y presentar los resultados de manera comprensible en orden a incrementar el conocimiento. Además se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados [12].

El objetivo de la minería de datos es encontrar información oculta la cual debe contar con ciertas características.

- **Nueva:**
  No se buscan conceptos ya existentes, deben ser desconocidos.

- **Correcta:**
  La selección de los datos debe ser acorde a los resultados que se esperan obtener sino, tendrá como consecuencia información errónea.

- **Significativa:**
  La información obtenida debe tener algún significado y deber ser muy clara para el usuario.

- **Aplicable:**
  La información obtenida debe obtener un propósito o dar soluciones según el teme elegido.

Existen dos modelos que se usan para el manejo de datos, el modelo descriptivo y el modelo predictivo que se utilizan de acuerdo a las características de los datos y las necesidades del estudio. A continuación se describe cada uno de ellos [13]:

Modelo predictivo o supervisado: Predicen el valor de un atributo (etiqueta) de un conjunto de datos, conocidos otros atributos. A partir de los datos cuya etiqueta se conoce se induce una relación entre dicha etiqueta y otra serie de atributos. Esas relaciones sirven para realizar la predicción en datos cuya etiqueta es desconocida. Esta forma de trabajar también es conocida como aprendizaje supervisado y se desarrolla en dos fases:

Entrenamiento: Construcción de un modelo usando un subconjunto de datos con etiqueta conocida.

Prueba: Prueba del modelo sobre el resto de los datos.

Modelo descriptivo, no supervisado o de descubrimiento del conocimiento: Cuando una aplicación no es lo suficientemente madura y no tiene el potencial necesario para una solución predictiva, entonces se recurre a este método, la que trata de descubrir patrones y tendencias en los datos actuales. El descubrimiento de esa información sirve para llevar a cabo acciones y obtener un beneficio de ellas.

Tabla 3.1 Clasificación de las técnicas de minería de datos

Supervisados	**No supervisados**
Árboles de decisión	Detección de desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento ("clustering")
Series temporales	Reglas de asociación
	Patrones secuenciales

3.2.3 Disciplinas aplicadas a la minería de datos

La minería de datos es una actividad de la metodología KDD, y está a medio camino de la informática, la estadística y la documentación, y que se ha estado utilizando en números disciplinas para el análisis de grandes cantidades de datos.

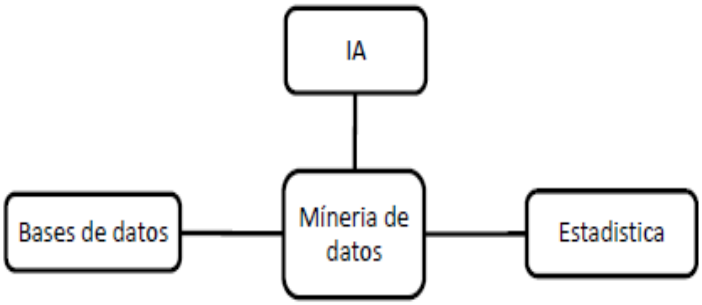

Figura 3. 2 Disciplinas de la minería de datos

La inteligencia artificial implica elementos de aprendizaje, evolución y lógica que tienen como resultados aprendizaje automatizado el cual se alcanza al aplicar algoritmos inteligentes como las redes neuronales, algoritmos genéticos, etc. [14].

La estadística analiza y recolecciona información para su posterior experimento donde se busca describir fenómenos, tal como en la minería de datos. Algunos conceptos utilizados en la estadística son medias centralizadas (media aritmética, mediana y moda), medidas de posición (cuartiles, deciles y porcentiles), medidas de dispersión (rango, desviación, media, varianza y desviación típica).

Y finalmente las bases de datos, aquí es donde se almacenan las grandes cantidades de datos de manera ordenada, y proporcionan de manera fácil el acceso a información.

3.2.4 Áreas de aplicación

Existen varias áreas en las que se puede aplicar la minería de datos, tales áreas están expuestas a continuación con su respectiva información [15]:

- Sistemas empresariales
    - Cambios de usuarios a otras compañías
    - Ventas de más productos en empresas
- Medicina :
    - Algún medicamento que sea el más solicitado
    - Asociación de síntomas y clasificación de diferentes patologías
    - Estudio de factores de riesgo para la salud en distintas patologías
    - Estudios epidemiológicos, prevención, vacunas.

- Banca
    - Detección de tarjetas clonadas (fraudes)

- Identificación de clientes leales
- Determinar gastos con tarjetas de crédito

- Procesos industriales
  - Modelos de calidad
  - Predicción de fallos
  - Extracción de modelos de producción
  - Extracción de modelos de coste

3.2.5 Metodología KDD

La metodología KDD fue propuesta por Fayyad en 1996. Esta metodología consta de 5 fases selección, preprocesamiento, transformación, minería de datos y evaluación e implementación, como se muestra en la figura 3.3.

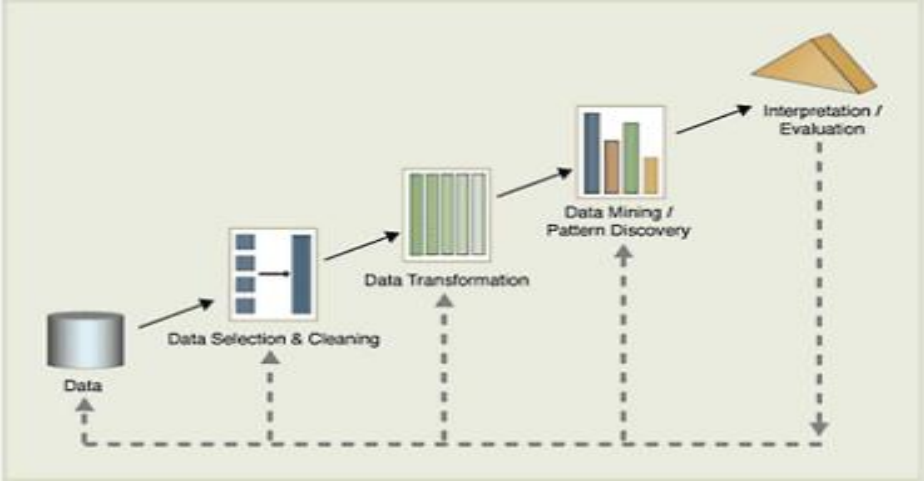

Figura 3. 3 Metodología KDD [16]

1. Desarrollar un entendimiento de la aplicación de dominio y los conocimientos previos y la identificación de la meta del proceso de KDD desde el punto de vista del cliente.

2. Crear un conjunto (o en su defecto obtener un gran almacén de datos) de donde el descubrimiento se llevará a cabo.
3. Limpieza y preprocesamiento de datos. Incluir las operaciones básicas que incluyen la eliminación del ruido, campos de datos vacíos etc.

4. Reducción de datos y la proyección: la búsqueda de características útiles para representar los datos en función del objetivo de la tarea (reducción de datos).
5. Comenzar con el paso número 1 de la metodología KDD, se comienza el objetivo de este método a un método de minería de datos.
6. Análisis exploratorio y de hipótesis y el modelo de selección: la elección del algoritmo de minería de datos se utilizará para la búsqueda de patrones de datos.
7. Aplicar la minería de datos: la búsqueda de patrones de interés en una determinada forma de una o de un conjunto de representación.
8. Interpretación de los patrones minados, se puede regresar en los pasos del 1 al 7 para tener mejor iteración. Aquí se puede aplicar la visualización de los patrones y modelos extraídos o visualización de los datos que figuran modelos.
9. Finalmente aquí es donde actúa el conocimiento descubierto. Este proceso incluye la comprobación y solución de posibles conflictos en los conocimientos [17].

3.2.6 Minería de Datos y metodología KDD

Entre la minería de datos y la metodología KDD existe una pequeña diferencia, ya que los dos aplican al conocimiento de la base de datos.

KDD es el descubrimiento de conocimientos en bases de datos y la DM es la minería de datos. El proceso completo del KDD es, la extracción no trivial de conocimiento implícito, previamente desconocido y potencialmente útil, a partir de una base de datos.

Mientras que en la Minería de datos se tiene una etapa de descubrimiento en el proceso de KDD, el cual consiste, en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados. Dos tipos de metas que se puede conseguir a consecuencia de la DM y el proceso KDD son los siguientes:

- Verificación: El sistema está limitado a la hipótesis del usuario
- Descubrimiento: El sistema automáticamente, encuentra nuevos patrones.

Pueden ser para:

- Predicción: Cuando el sistema encuentra patrones por predicción del comportamiento de alguna entidad.
- Descripción: Donde el sistema presenta los datos de forma inteligible para los humanos [18].

3.3 Tareas y técnicas de minerías de datos

La minería de datos, no es más que un caso especial de aprendizaje computacional inductivo. Es la identificación de patrones, de regularidades, existentes en la evidencia, también se puede ver como la predicción de observaciones futuras con plausibilidad.

Existen las taxonomías de Técnicas del DM tales como [19]:

- **Predictivo**
- Interpolación y Predicción secuencial.
    - Datos continuos
    - Regresión lineal
    - Regresión no lineal

- Datos discretos

- **Descriptivo**
- Aprendizaje supervisado
    - Clasificación
    - Categorización

- Dependiendo del número y tipo de clases
    - Clase discreta
    - Clase continua

3.3.1 Tareas de minerías de datos

La tarea de minería de datos real, es el análisis automático o semiautomático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, tales como grupos de registros de datos, registros poco usuales y dependencias. Esto generalmente implica el uso de técnicas de bases de datos como los índices especiales.

Como ya se ha mencionado, las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son ms que los algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener resultados.

3.3.2 Tareas del modelo predictivo

El modelo predictivo es una rama de la minería de datos que tiene relación con la predicción de las probabilidades y tendencias futuras, permite extraer conclusiones confiables sobre eventos futuros, a través de la aplicación de métodos estadísticos, matemáticos y de reconocimientos de patrones.

También se le conoce como a un proceso utilizado en el análisis para crear un modelo estadístico de comportamiento futuro. El análisis predictivo es el área de minería de datos

en cuestión, con probabilidades de pronósticos y tendencias. Para cumplir los objetivos del modelo predictivo, son dos los retos de la minería de datos los cuales son:

1. Trabajar con grandes volúmenes de datos, que proceden generalmente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos, etc.).

2. Usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos, la utilidad del conocimiento extraído está relacionado con la comprensibilidad del modelo inferido.

El elemento central del análisis predictivo es el predictor, una variable que puede ser medida para una entidad individual o de oro tipo para predecir el comportamiento futuro. Por ejemplo:

En el Hospital General del Sur es probable que se tenga en cuenta los posibles predictores de seguridad, tales como la edad, el género y un registro de ETS, al momento de tomar una decisión para llevar a cabo un tratamiento necesario para cada tipo de paciente

Para desarrollar un modelo predictivo, primero se necesita ensamblar la base de datos que será utilizada para capacitación. Para esto, un conjunto de campos de entrada que represente al cliente, por ejemplo, se ensambla junto en un registro. Este registro puede contener características como edad, género, ETS, si es mujer y está embarazada, tiempo de detección y si es que ya se encuentra en tratamiento. Cuando todos los registros de paciente son ensamblado juntos, se convierten en un conjunto de datos que puede contener millones de registros.

Las técnicas de modelo predictivo permiten el desarrollo de modelos predictivos precisos, siempre y cuando existan datos suficientes y la calidad de los datos no sea un problema. Los malos datos otorgan un mal modelo, sin importar lo bueno que sea la técnica predictiva. Y por esto existe la máxima: “Basura que entra, basura que sale” [20].

3.3.3 Tareas del modelo descriptivo

En el modelo descriptivo se identifican patrones que describen los datos mediante tareas, por ejemplo, en agrupamiento y técnicas de asociación. Cabe destacar que mediante este modelo, se identifican patrones que explican o resumen el conjunto de datos, siendo estos útiles para explorar las propiedades de los datos examinados [21].

Estos modelos siguen un tipo un tipo de aprendizaje no supervisado, que consiste en adquirir conocimiento desde los datos disponibles, sin requerir influencia externa que indique un comportamiento deseado al sistema.

- **Agrupamiento**

En esta tarea se evalúan similitudes entre los datos para construir modelos descriptivos, analizar correlaciones entre las variables o representar un conjunto de datos en un pequeño número de regiones. El agrupamiento es considerado como la tarea de dividir una población heterogénea en un número de subgrupos homogéneos de acuerdo a las similitudes de sus registros. Dentro de esta manera existen dos tipos principales de agrupamiento como: el jerárquico que se caracteriza por el desarrollo recursivo de una estructura en forma de árbol, y el particional que organiza los registros dentro de k grupos.

Los métodos particionales tienen ventajas en aplicaciones que involucran gran cantidad de datos para los cuales la construcción de un árbol resulta complicada. Una característica de este tipo de agrupamiento es el de establecer a priori el número de grupos de entrada (k), por lo que el en la práctica es necesario repetir la prueba estableciendo diferentes números de grupos, eligiendo la solución que mejor se adapte al objetivo del problema.

Un método sugerido por Milligan (1985) y Hair (1995) para determinar el número de grupos de entrada (k) es usar el resultado obtenido por algún algoritmo jerárquico, mediante el cual se obtiene el número deseado de grupos, posteriormente se aplica algún algoritmo particional.

- **Técnicas de asociación**

Con este modelo, se identifican afinidades entre la colección de los registros examinados, buscando relaciones o asociaciones entre ellos. Las afinidades son expresadas como reglas de la forma: “si X entonces Y” donde X y Y son los registros de una transacción.
El interés por esta tarea se debe principalmente a que, las reglas proporcionan una forma concisa de declarar la información potencialmente útil. Las reglas se evalúan usando dos parámetros: precisión y cobertura.

La cobertura es el número de instancias o datos hallados correctamente, mientras que la precisión, es el porcentaje de instancias halladas correctamente. Las ventajas más frecuentes en las reglas de asociación son el descubrimiento de asociación y de frecuencia.

El descubrimiento de asociación encuentra relaciones, que aparecen conjuntamente a un acontecimiento y la consecuencia la asocia al tiempo.

- **Correlaciones**

Las correlaciones, son una tarea descriptiva que se usan para determinar el grado de similitud de los valores de dos variables numéricas. Un mecanismo estándar para medir la correlación es el coeficiente de correlación, para este caso llamado S, el cual es un valor real comprendido entre -1 y 1. Así, si S es 1, las variables están totalmente correlacionadas; si S es -1, las variables están correlacionadas negativamente; y si S es 0, entonces no existe correlación.

Por consiguiente, cuando S es positivo, las variables tienen un comportamiento similar y cuando S es negativo una variable crece y la otra decrece.

3.3.4 Relación entre tareas y métodos

La minería de datos, es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utilizan análisis matemáticos para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos, ya que las relaciones son demasiado complejas, o porque hay demasiados datos.

En seguida se describen cada una de las técnicas y de los métodos que se pueden utilizar para llegar a una minería de datos:

- **Técnicas basadas en arboles de decisión:**

Es un algoritmo híbrido que incorpora distintos métodos para crear un árbol y admite varias tareas de análisis, incluyendo la regresión, la clasificación y la asociación. Representan funciones lógicas (if-then). Por medio de Aprendizaje Automático se refiere un Árbol de decisión a partir de un conjunto de instancias o ejemplos. El algoritmo J48 de Weka, utiliza un método heurístico para inferir el árbol, donde se realiza la selección del atributo en cada nivel del árbol en función de la calidad de la división que produce [22].

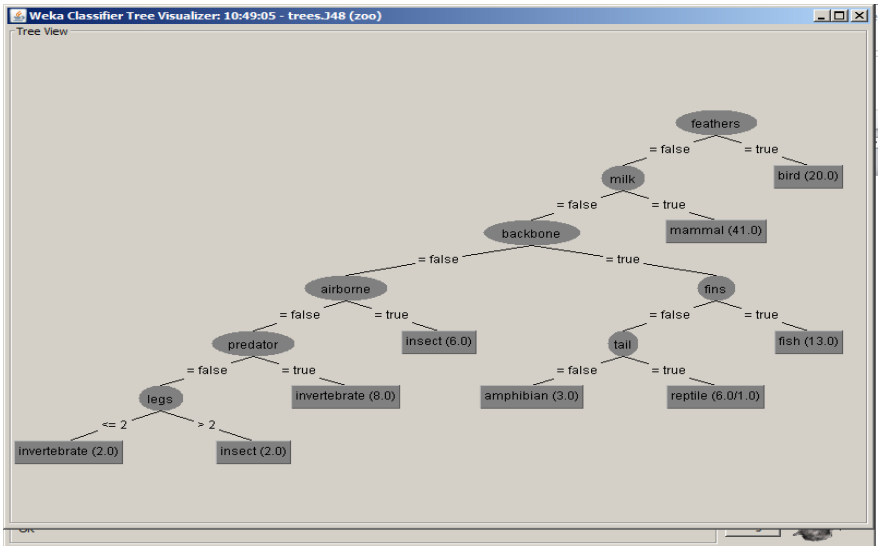

Figura 3. 4 Árbol de decisión

- **Técnicas basadas en redes neuronales:**
  Son ampliamente utilizadas para tareas relacionadas con reconocimiento de patrones y clasificación. Aunque son clasificadores muy precisos, no son comúnmente muy utilizados para *Data Mining* porque producen modelos de

aprendizaje inexplicables. Se representa con un árbol de decisión que aproxima a la red [23].

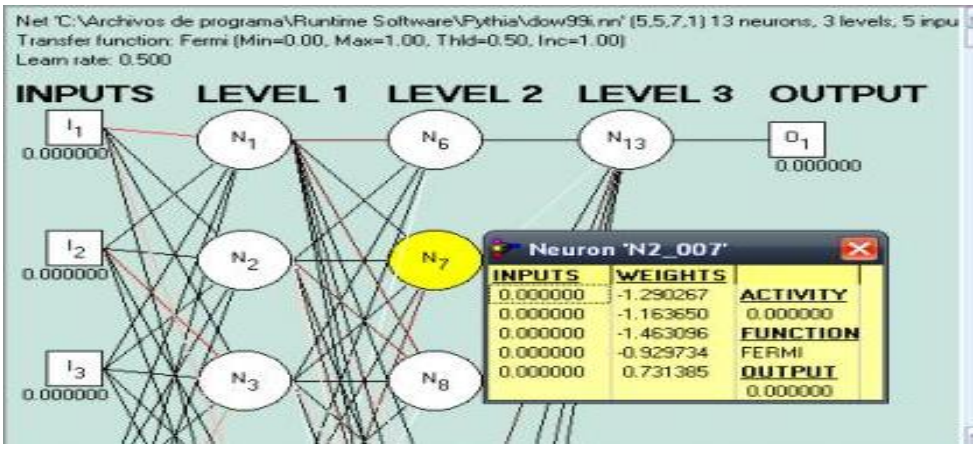

Figura 3. 5 Redes neuronales

- **Técnicas de regresión:**
  Un modelo logístico de árbol consiste básicamente en una estructura de árbol de decisión estándar con funciones de regresión logística en las hojas [24].

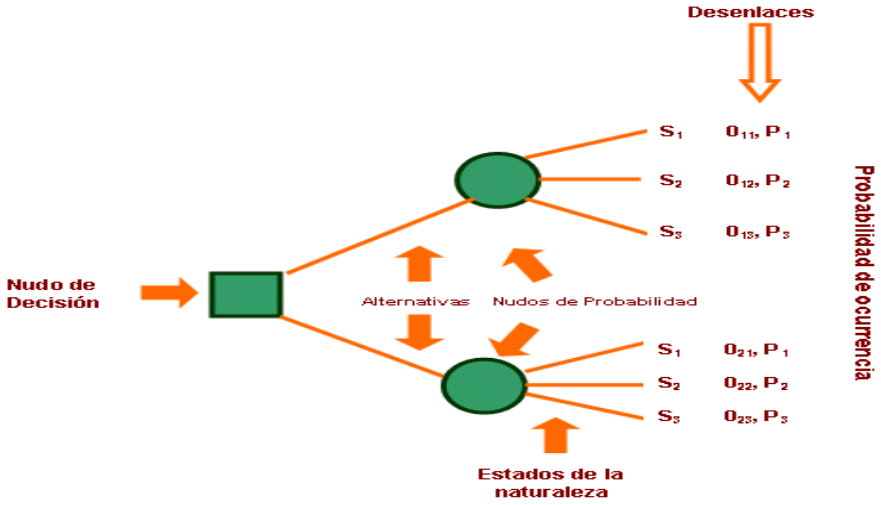

Figura 3. 6 Técnicas de regresión

- **Técnicas de series de tiempo o temporales:**
  El propósito de las series de tiempo significa que extendemos los valores históricos a un futuro, donde no hay mediciones disponibles. El propósito se realiza generalmente para optimizar áreas como los niveles de inventario, la capacidad de producción o los niveles de personal [25].

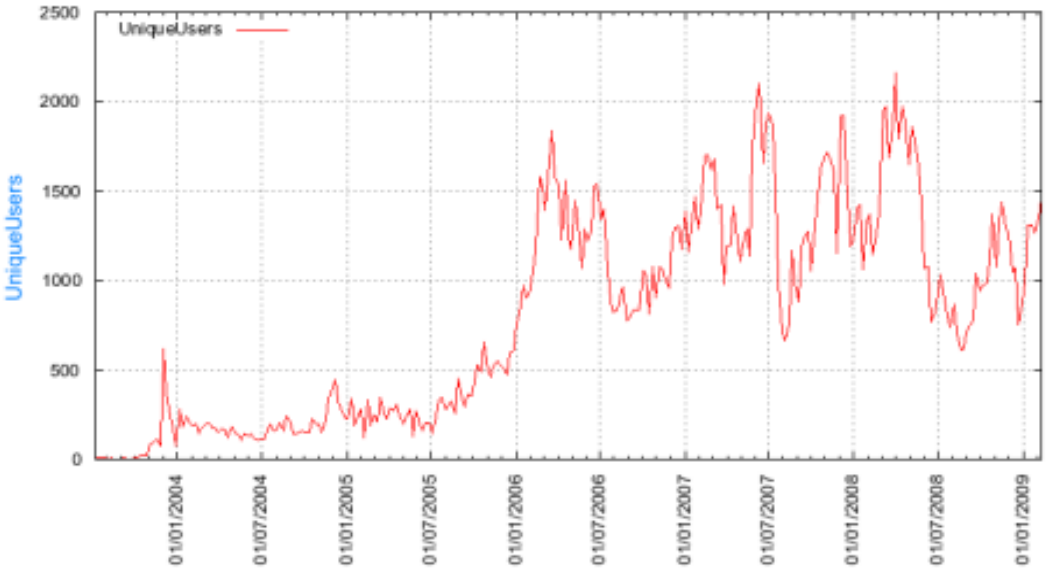

Figura 3. 7 Técnicas de seres de tiempo o temporales

- **Técnicas de segmentación:**
  Técnica que consiste en la agrupación de los datos con características similares, por ejemplo (Delegaciones/embarazadas, etc.). Esta es una importante herramienta que permite detectar ETS acordes a diferentes tipos de comportamiento [26].

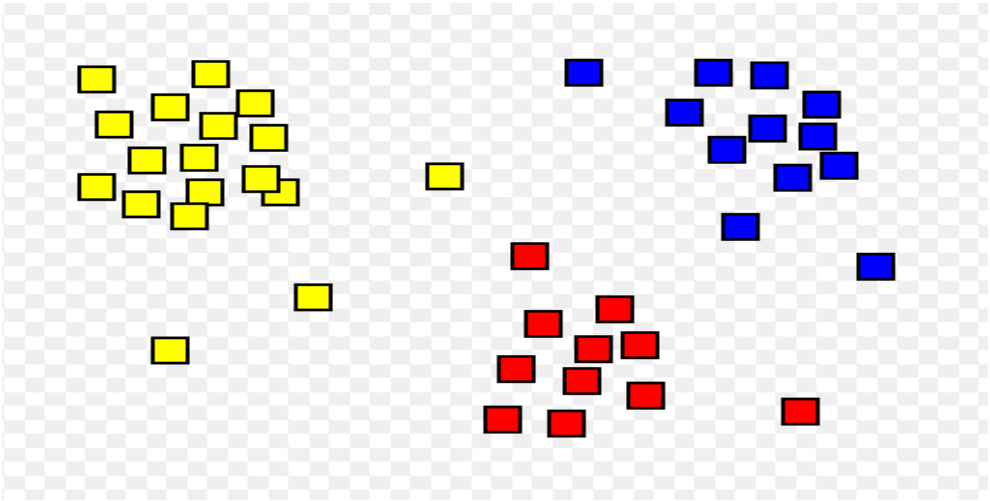

Figura 3. 8 Técnicas de segmentación

Los métodos anteriormente planteados corresponden con varias de las tareas más significativas de la minería de datos. Es importante conocer características funcionales de cada técnica y seleccionar el o los métodos ideales para alcanzar los objetivos según sea el caso de estudio.

3.4 Conclusiones

Es muy preciso saber qué tipo de tareas y métodos son los necesarios para aplicar a la minería de datos, por lo tanto también deben de ser conocidos cada uno para así poder aplicarlos. Por esta razón, en este capítulo fueron expuestos las tareas y métodos más comúnmente usados en la mayoría de los sistemas para la toma de decisiones, así como también la metodología KDD.

CAPÍTULO 4

4. Metodología

En el presente capítulo se describe la metodología a utilizar para el análisis y diseño del proyecto, desde construir el conjunto de datos de aprendizaje hasta el descubrimiento de los patrones en el conjunto anterior.

Conceptos tratados en el capítulo anterior, serán implementados de manera práctica sobre los datos seleccionados, así como también serán implementadas las técnicas de minerías de datos.

4.1 Planteamiento y requerimientos de análisis

Como se planteó anteriormente en la sección 1.1.1, se consideran los objetivos que se desean alcanzar los cuales fueron especificados en la sección 1.1.2. Se siguen los pasos de la metodología KDD, que se indican en la sección 3.2.4 como se presentan a continuación:

1. **Preparación de los datos:**
   - Se recolectan las Bases de Datos de distintas fuentes de información.

   - Preparación de datos:
     - Decidir qué datos serán utilizados para el análisis.
     - Limpieza de datos, lo cual implica la selección de los subconjuntos de datos limpios.
     - Construir datos, esta tarea incluye la construcción de operaciones de preparación de datos, tales como atributos derivados o el ingreso de nuevos registros.
     - Integración de datos, éste es el método donde la combinación es integrada de múltiples tablas o registros para crear nuevos registros o valores.
     - Formatear datos, se refiere a modificaciones principales sintácticas hechas a los datos que no cambian su significado, pero podría ser requerido por la herramienta de modelado.

2. **Aplicación de las técnicas de Minerías de Datos:**

Se aplicarán las técnicas de minerías de datos ya mencionadas en la sección 3.3.3 (agrupamiento, clasificación, etc.) y se elegirá el método que mejores resultados entregue.

3. **Evaluación e interpretación:**

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

4. **Difusión y uso:**

Incorporar el conocimiento descubierto al sistema, lo cual puede incluir resolver conflictos existentes. El conocimiento se obtiene para realizar acciones o la toma de decisiones [27].

4.2 Integración y recopilación de datos

En la fase de integración y recopilación de datos, se determinan las fuentes de información que pueden ser útiles y donde conseguirlas. A continuación, se transforman todos los datos a un formato común, frecuentemente mediante un almacén de datos que consiga unificar de manera operativa toda la información recogida, detectando y resolviendo las inconsistencias. Este almacén de datos facilita enormemente la navegación y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.

Dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes. Estas situaciones se tratan en la fase de selección, limpieza y transformación, en la que se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos.

Se proyectan los datos para considerar únicamente aquellas variables o atributos que van a ser relevantes, con el objetivo de hacer más fácil la tarea propia de minería, y para que los resultados de la misma sean más útiles.

En la siguiente sección se presentará la recolección de información proporcionada por el Hospital General del Sur de Puebla. Este sistema actualiza información nacional pública desde 1981, los atributos correspondientes a los datos de ETS se describen a continuación.

4.3 Selección, limpieza y transformación

La calidad del conocimiento descubierto, no solo depende del algoritmo de minería utilizado, sino también de la calidad de los datos minados. Por ello, después de la recopilación, el siguiente paso en el proceso de la extracción del conocimiento es seleccionar y preparar el subconjunto de datos que se va a minar, los cuales constituyen lo que se conoce como vista minable. Este paso es necesario, ya que algunos datos coleccionados en la fase anterior, son irrelevantes o innecesarios para la tarea de minería que se desea realizar.

La recopilación de datos, debe ir acompañada de una limpieza e integración de los mismos, para que éstos estén en condiciones para su análisis. Los beneficios del análisis y de la extracción de conocimiento a partir de datos, dependen de una gran medida, de la calidad de los datos recopilados. Además, generalmente, debido a las características propias de las técnicas de minería de datos, es necesario realizar una transformación de los datos, para obtener una "materia prima" que sea adecuada para el propósito concreto, y las técnicas que se quieren emplear.

Todas las acciones descritas anteriormente se efectuaron con la finalidad de mejorar la eficiencia de la herramienta de minería de datos, y de esta manera mejorar la calidad del conocimiento obtenido. Existen maneras diversas para llevar a cabo este proceso. Algunas de ellas son hacerlo manualmente registro a registro y también es posible hacer uso de herramientas automatizadas.

Esta etapa del proyecto es muy importante, ya que si la información obtenida no es selecciona adecuadamente, la búsqueda de conocimiento con la aplicación de técnicas de minería de datos, puede resultar errónea.

La calidad del conocimiento obtenido, depende mucho y en gran parte de la calidad de los datos minados, es por ellos que después de haber seleccionado los datos, el siguiente paso en la metodología KDD es preparar el subconjunto de datos que serán minados, los cuales van a constituir lo que se conoce como vista minable.

En el transcurso del proceso ya mencionado, se eliminaron valores outliers (valores que no se ajustaron al comportamiento general de los datos). Posteriormente se derivaron algunos atributos que eran relevantes, se rellenaron algunos valores faltantes o simplemente se cambiaron por *, también se cambiaron algunos datos de tipo numérico a su representación porcentual.

Todas las acciones descritas anteriormente se efectuaron con la finalidad de mejorar la eficiencia de la herramienta de minería de datos, y de esta manera mejorar la calidad del conocimiento obtenido. Existen maneras diversas para llevar a cabo este proceso. Algunas de ellas son hacerlo manualmente registro a registro y también es posible hacer uso de herramientas automatizadas.

4.3.1 Tratamiento de datos

En esta sección, se dará a conocer todo el preprocesamiento de los datos, hasta obtener la información que se usó para las pruebas.

Con las modificaciones que se hicieron a los datos, se encuentra la conversión de ciertos atributos, separación de columnas, y creación de nuevos campos, así como también la eliminación de datos que no eran de utilidad.

4.3.2 Fase de Minería de Datos

Cualquier proyecto de minería de datos, independiente de su enfoque y de las técnicas de extracción utilizadas al transcurso del proceso, debe atravesar por una serie de fases que hace que el proceso sea exitoso desde que inicia hasta que culmina. Dando así un análisis completo y efectivo para tomar una decisión correcta.

- Filtrado de datos
    - El objetivo de esta fase, es filtrar los datos de tal motivo que se eliminen todos los valores incorrectos, valores no válidos y valores desconocidos o simplemente valores que no son necesarios.

- Selección de variables
    - Para reducir el tamaño de los datos elegidos, se deben establecer características correspondientes y necesarias para ser aplicadas a la selección correcta de los datos.

- Extracción de conocimiento
    - Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema.

- Interpretación y evaluación
    - Una vez obtenido el modelo final, se deben validar las conclusiones obtenidas al finalizar el proceso de extracción. Se debe comprobar que las conclusiones arrojadas son válidas, suficientes y satisfactorias.

Tabla 4.1 Datos almacenados de VIH

Viewer

Relation: BDD CAPASITS

No.	1: INDIGENA Nominal	2: SEXO Nominal	3: EDAD String	4: VIHEMBARAZADA Nominal	5: ULTIMACONSULTA Nominal	6: PRIVLIBERTAD Nominal	7: NOTIFICACION Nominal	8: INGRESO Nominal	9: JURISDICCION Nominal	10: ESTADO Nominal	11: DELMUNICIPIO Nominal	12: LOCALIDAD Nominal	13: FECHAALTA Nominal	14: TTO DIAS String	15: TTO MES String	16: TTO Strin
1	NO	M	53	NO APLICA	2014-11-27	N	2005-06-20	2005-06-24	PUEBLA	PUEBLA	PUEBLA		2006-12-31 00...	3447	114.9	9.58
2	NO	F	33	NO	2014-12-10	N	2001-02-07	2001-04-19	ACATLáN	PUEBLA	PETLALCINGO	TEPEJILLO	2006-12-31 00...	5054	168.5	14.04
3	NO	F	21	NO	2014-12-09	N	2004-06-06	2004-08-11	PUEBLA	PUEBLA	PUEBLA		2006-12-31 00...	3838	127.9	10.66
4	NO	F	38	NO	2015-01-19	N	2005-12-01	2006-12-01	EL SECO	PUEBLA	LIBRES	NUEVO MéXIC...	2006-12-31 00...	3336	111.2	9.27
5	NO	M	55	NO APLICA	2009-04-20	N	2004-03-04	2004-03-05	PUEBLA	PUEBLA	PUEBLA		2006-12-31 00...	1873	62.4	5.20
6	SI	F	36	NO	2014-10-02	N	2008-06-12	2006-12-01	HUAUCHINANGO	PUEBLA	HUAUCHINANGO	HUAUCHINANGO	2006-12-31 00...	2303	76.8	6.40
7	NO	M	44	NO APLICA	2014-12-12	N	1999-12-12	2000-02-24	HUEJOTZINGO	PUEBLA	SAN ANDRéS CHO...	SAN ANDRéS ...	2006-12-31 00...	5479	182.6	15.22
8	NO	M	46	NO APLICA	2014-11-26	N	1999-10-11	2000-01-06	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	5525	184.2	15.35
9	NO	M	84	NO APLICA	2013-10-01	N	2005-08-01	2005-10-01		MéXICO	NAUCALPAN DE J...		2006-12-31 00...	2983	99.4	8.29
10	NO	M	36	NO APLICA	2014-11-20	N	2004-09-28	2004-11-01	TEPEXI DE RODR...	PUEBLA	TEPEACA	SAN PABLO AC...	2006-12-31 00...	3705	123.5	10.29
11	NO	M	45	NO APLICA	2014-07-10	N	2000-07-20	2000-11-14	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	5103	170.1	14.18
12	SI	M	50	NO APLICA	2014-11-06	N	2006-01-25	2006-01-30	TEHUACáN	PUEBLA	TEHUACáN	TEHUACáN	2006-12-31 00...	3207	106.9	8.91
13	SI	M	47	NO APLICA	2013-05-07	N	2003-09-17	2003-10-28	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3520	117.3	9.78
14	NO	M	33	NO APLICA	2014-11-10	N	2005-11-08	2005-05-13	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3289	109.6	9.14
15	NO	M	36	NO APLICA	2014-07-10	N	2004-11-12	2004-12-07	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3527	117.6	9.80
16	NO	M	46	NO APLICA	2013-03-15	N	2003-07-12	2004-02-11	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3534	117.8	9.82
17	NO	M	42	NO APLICA	2014-11-19	N	2004-09-14	2004-10-18	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3718	123.9	10.33
18	NO	F	34	NO	2014-10-08	N	2004-04-02	2004-10-05	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3841	128.0	10.67
19	NO	M	49	NO APLICA	2013-07-29	N	2001-11-22	2002-01-16	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	4267	142.2	11.85
20	NO	M	21	NO APLICA	2012-03-14	N	2000-06-11	1999-05-31	NO APLICA	DISTRITO ...	CUAUHTéMOC	CUAUHTéMOC	2006-12-31 00...	4294	143.1	11.93
21	NO	M	14	NO APLICA	2015-01-12	N	2003-11-24	2003-10-09	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	4067	135.6	11.30
22	NO	F	13	NO APLICA	2014-12-19	N	2005-05-06	2005-04-26	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3514	117.1	9.76
23	NO	M	54	NO APLICA	2014-10-17	N	2006-10-04	2006-09-13	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	2935	97.8	8.15
24	NO	M	39	NO APLICA	2014-04-29	N	1996-10-16	2006-05-02	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	6404	213.5	17.79
25	NO	M	20	NO APLICA	2014-02-25	N	2006-04-03	2011-06-08	HUEJOTZINGO	PUEBLA	SAN SALVADOR E...	ANALCO DE P...	2006-12-31 00...	2885	96.2	8.01
26	NO	M	70	NO APLICA	2011-03-08	N	2005-05-11	2006-01-27	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	2127	70.9	5.91
27	NO	M	36	NO APLICA	2014-09-12	N	2005-05-06	2005-06-20	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3416	113.9	9.49
28	NO	M	39	NO APLICA	2014-11-24	N	2003-04-03	2003-09-06	EL SECO	PUEBLA	RAFAEL LARA GR...	MáXIMO SERDáN	2006-12-31 00...	4253	141.8	11.81
29	NO	M	48	NO APLICA	2014-10-30	N	2004-08-12	2004-09-02	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3731	124.4	10.36
30	NO	M	47	NO APLICA	2014-12-04	N	2002-02-25	2002-12-18	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	4665	155.5	12.96
31	NO	M	58	NO APLICA	2006-01-23	N	2005-10-14	2005-08-14	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	101	3.4	0.28
32	NO	F	42	NO APLICA	2007-02-12	N	2004-07-08	2004-09-02	PUEBLA	PUEBLA	SAN GABRIEL CHI...	SAN GABRIEL ...	2006-12-31 00...	949	31.6	2.64
33	NO	M	46	NO APLICA			2004-07-01	2004-07-08	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	#iVALOR!	#iVALOR!	#iVALOF
34	NO	M	53	NO APLICA	2014-06-13	N	2004-10-28	2005-02-07	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3515	117.2	9.76
35	NO	M	48	NO APLICA	2014-12-03	N	2005-04-01	2005-04-28	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3533	117.8	9.81
36	NO	M	68	NO APLICA	2014-12-05	N	2004-03-18	2004-05-12	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	3914	130.5	10.87
37	NO	F	54	NO	2011-01-12	N	1999-01-30	1999-09-30	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	4365	145.5	12.13
38	NO	M	41	NO APLICA	2014-12-17	N	2002-07-19	2002-12-16	PUEBLA	PUEBLA	PUEBLA	HERóICA PUEB...	2006-12-31 00...	4534	151.1	12.59

La tabla 4.1 muestra los datos que se utilizan para esta tesis, son datos exclusivos de personas infectadas y detectadas con VIH, no se eliminó ni se modificó ninguna información obtenida, para que esto pudiera ser posible solo se guardó la información con la extensión .CSV, dicha extensión también puede ser utilizada mediante la herramienta Weka.

4.4 Weka

Se trata de un entorno para hacer análisis de un conocimiento y tiene una plataforma de software para el aprendizaje automático y una minería de datos en Java los cuales fueron en la universidad de Waikato.

El paquete Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelo predictivo, para 1997 se desarrolla con el fin de ser utilizada en distintas áreas, particularmente con finalidades docentes y de investigación.

Weka soporta preprocesamiento de datos, Clustering, Clasificación, Regresión, Visualización y Selección de atributos. También proporciona acceso a Bases de datos vía SQL mediante la conexión JDBC (Java Database Connectivity) y puede procesar el resultado devuelto por una consulta hecha a la base de datos. No puede realizar minería de datos multi-relacional, pero existen aplicaciones que pueden convertir una colección de tablas relacionadas de una base de datos en una única tabla que podrá ser procesada con Weka.

4.5 Conclusiones

En este capítulo se describe la metodología que se utiliza para el análisis y diseño del presente proyecto, la cual incluye: la preparación de los datos que fueron recopilados de las distintas fuentes, la integración de datos, y finalmente las fases de la minería de datos.

También se describen las características del software Weka el cual es una máquina de aprendizaje automático y la herramienta utilizada en este proyecto para descubrir patrones.

CAPÍTULO 5

5. Aplicación de Técnicas de Minería de Datos

El análisis de datos, para cumplir con los objetivos de este proyecto, se llevó a cabo utilizando la herramienta Weka v3.7. Esta herramienta permitió realizar entrenamiento y pruebas mediante la aplicación de múltiples algoritmos para seleccionar la información más sobresaliente sobre el tema de ETS en una población seleccionada del estado de Puebla. Los resultados variaron según los algoritmos aplicados, estos resultados se dan a conocer en esta tesis para que en un futuro, puedan servir de soporte para la toma de decisiones en el tratamiento de las ETS.

El conjunto de datos de aprendizaje, el cual se corresponde en particular con enfermos de VIH, cuenta con los atributos que se muestran en la tabla 5.1. Los pacientes considerados en esta muestra comprenden a aquellos que fueron notificados de padecer la enfermedad a partir del año 2008.

Tabla 5.1 Descripción Conjunto de Datos de Aprendizaje

ATRIBUTO	DESCRIPCIÓN
indígena	0.004 % del total de la muestra
sexo	24% mujeres, 76% hombres
edad	En el rango [1, 92años]. 6% [1,13años]. 2% [60,92años] media 35 años
embarazada	0.5% del total de la muestra
última consulta	Enero 2015
privado de la libertad	0.018 % del total de la muestra
notificación	31% en 2001, 61 en 2008%
ingreso	42 pacientes entre [1998-2001] 426 entre [2013-2014]
jurisdicción	15 jurisdicciones sanitarias en el estado de Puebla
estado	Puebla, DF, México, Tabasco, Chiapas, Tlaxcala , Veracruz
delegación/municipio	171 Municipios de los estados antes mencionados
localidad	520 localidades 40% Puebla de Zaragoza
fecha de alta	2015

A partir de los atributos de fechas de notificación, ingreso y alta, mostrados en la tabla 5.1 se derivaron atributos de año de notificación, año de ingreso y año de alta. Estos atributos nos dan una medida más acertada acerca de la adherencia al tratamiento de los enfermos. En la figura 5.1 se presenta la vista preliminar de los datos de aprendizaje la cual se compone de 17 atributos y 3830 ejemplos.

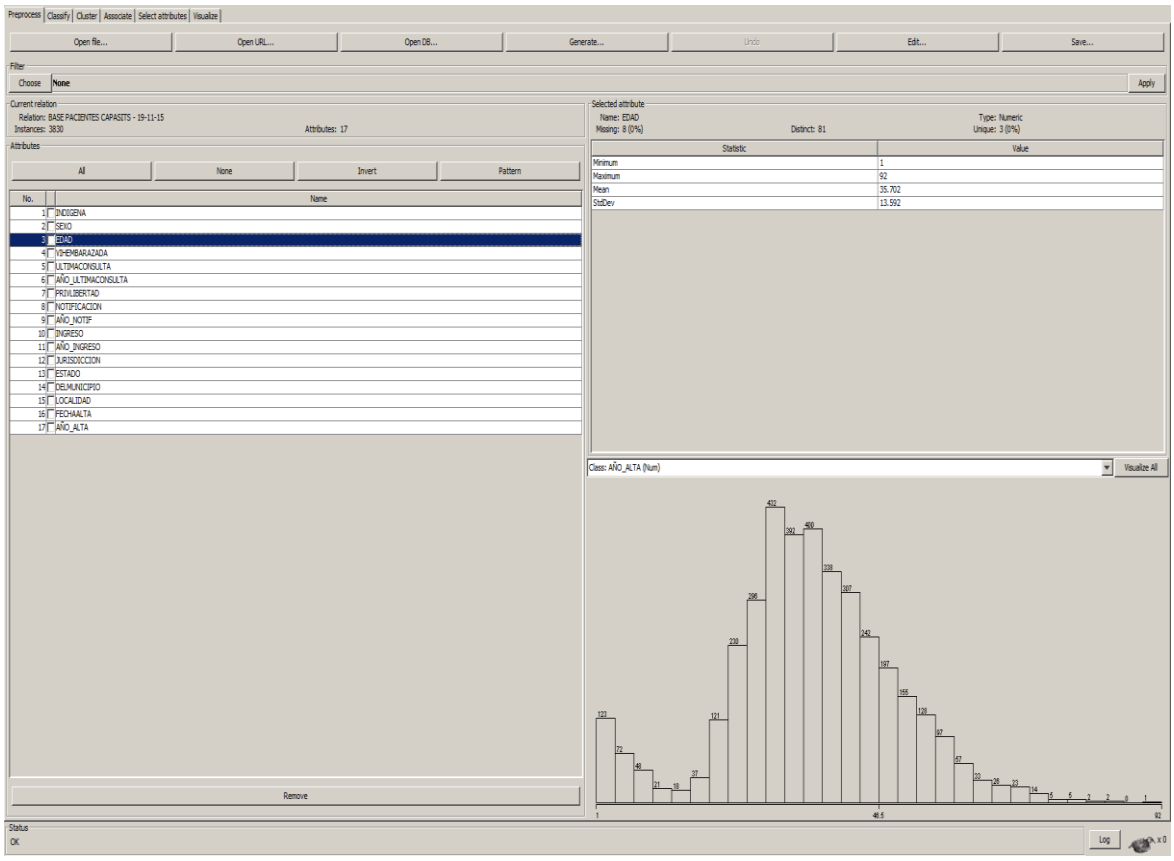

Figura 5. 1 Conjunto inicial de datos de aprendizaje

5.1 Aplicación de filtros

Los filtros proveen un mecanismo para procesar datasets e instancias y realizar transformaciones sobre ello. Soporta distintos formatos, conectividad a base de datos y filtrados. Weka cuenta con más de 100 métodos de clasificación. En la selección de filtros, se pueden transformar los datos de continuos a nominales y se pueden eliminar los atributos irrelevantes o redundantes [28].

Teniendo en cuenta la descripción de los datos presentada en la tabla 5.1 se aplicaron una serie de filtros para lograr un conjunto de datos de aprendizaje bien dimensionado y balanceado. Un conjunto de aprendizaje con las características anteriormente mencionadas representa una vista minable de datos. Una vista minable de datos ofrece

las características mínimas necesarias para la correcta aplicación de técnicas de minería de datos.

5.1.1 Filtros de eliminación de atributos

En primera instancia, se aplicó filtro *remove* para remover aquellos atributos que no son significativos. Los atributos removidos son *indígena, embarazadaconVIH, privación de la libertad y estado.* Como se muestra en la tabla 5.1 los valores para pacientes indígenas, embarazadas y con privación de la libertad representan menos del 1% de la muestra. Por otra parte el 99% de los pacientes viven en el estado de Puebla por lo tanto no se consideró significativo el número de pacientes provenientes de otros estados. El resultado de la aplicación del filtro *remove* se muestra en la figura 5.2, donde la muestra se redujo a 13 atributos.

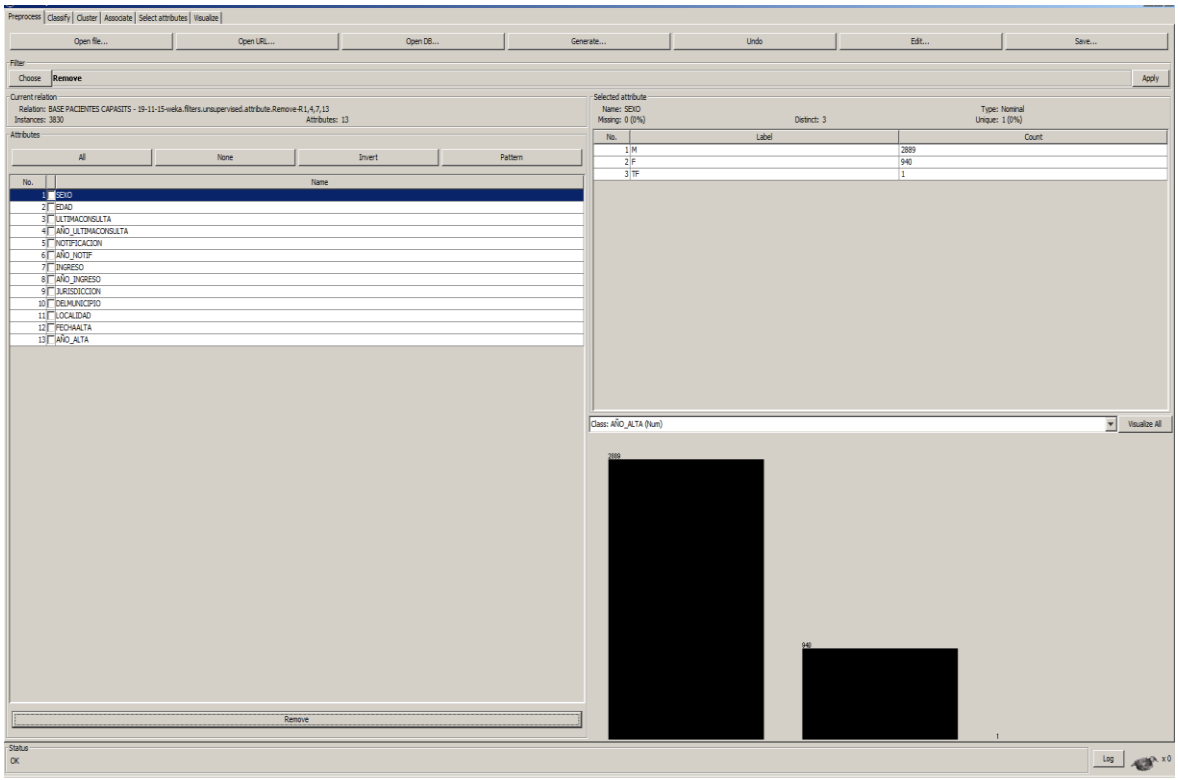

Figura 5. 2 Aplicación del filtro *remove*

5.1.2 Filtros para añadir expresiones

Se consideró interesante añadir un atributo que indicara el tiempo en el cual un paciente ha estado en tratamiento. El atributo *tratamientoenaños* indica la diferencia entre la fecha de ingreso y la fecha de alta. La figura 5.3 muestra cómo se deriva este atributo.

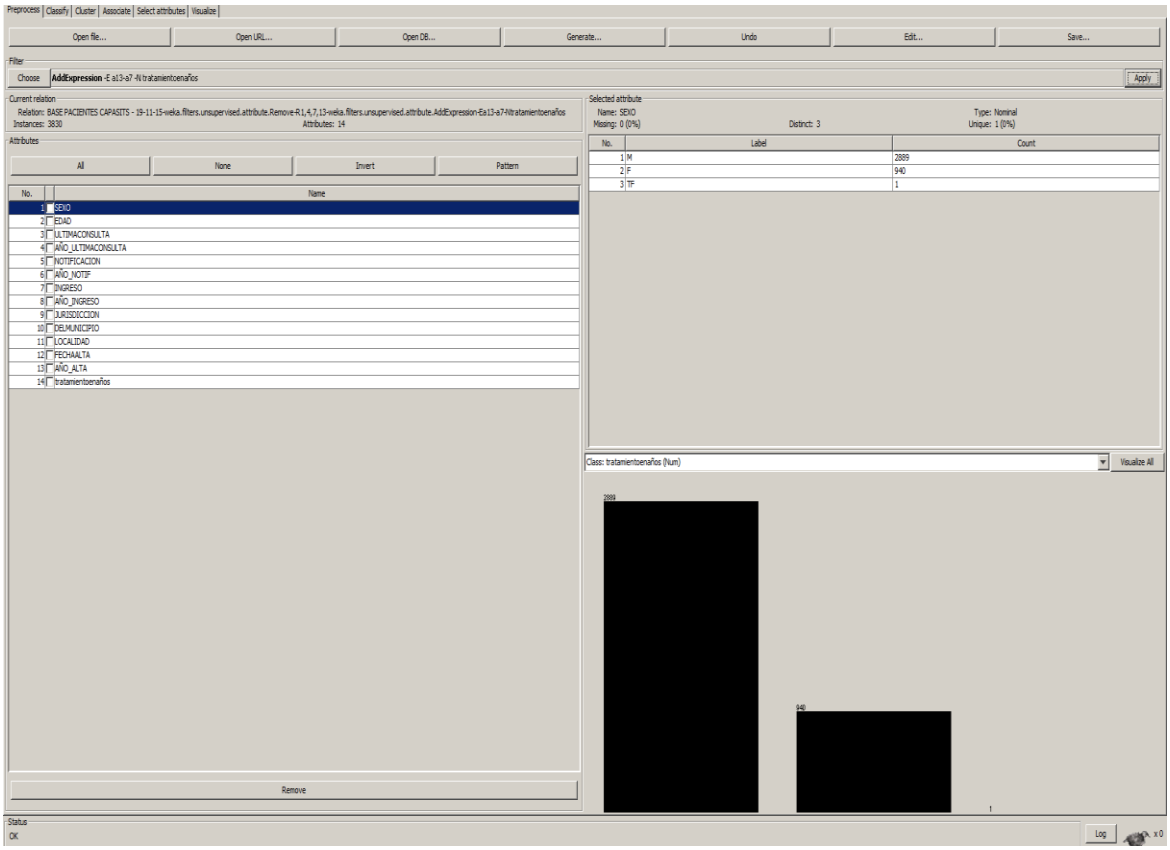

Figura 5. 3 Derivación del atributo *tratamiento en años*

5.2 Técnicas de Agrupamiento

Se aplicó primero el método de agrupamiento de máxima expectación EM con número de clusters -1, con el objetivo de que este método nos sugiera un número adecuado de clusters. Este número sugerido de clusters lo utilizaremos posteriormente como el parámetro *k* del método Kmeans.

5.2.1 Agrupamiento utilizando EM

Antes de aplicar EM se aplicó el filtro *resample* para reducir la muestra original de 3830 instancias a un 10%. EM no puede manejar una muestra tan grande de datos. El resultado de *resample*, con 383 instancias, se muestra en la figura 5.4

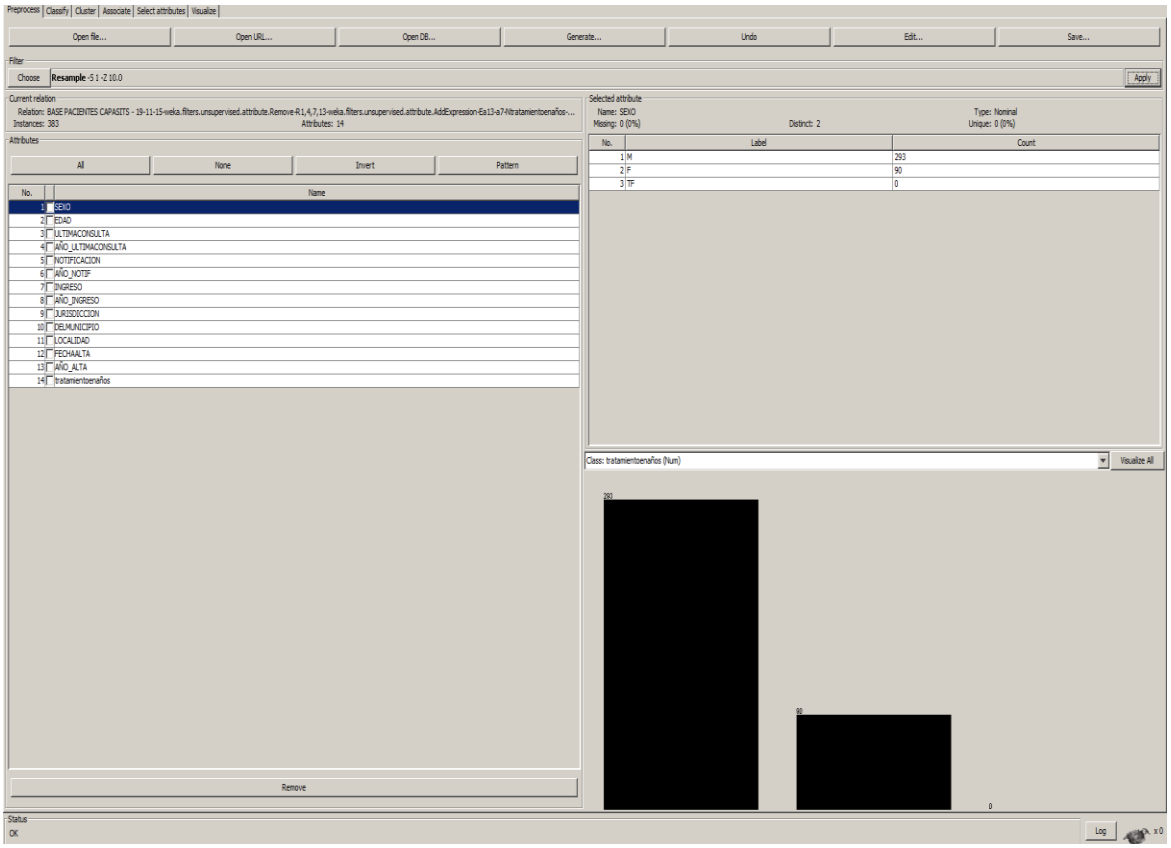

Figura 5. 4 Reducción de la muestra original en un 10%

Se aplicó el método EM en la muestra reducida y el resultado se presenta en la figura 5.5. EM propone 4 clusters los cuales se tomaron como referencia para la aplicación del método Kmeans

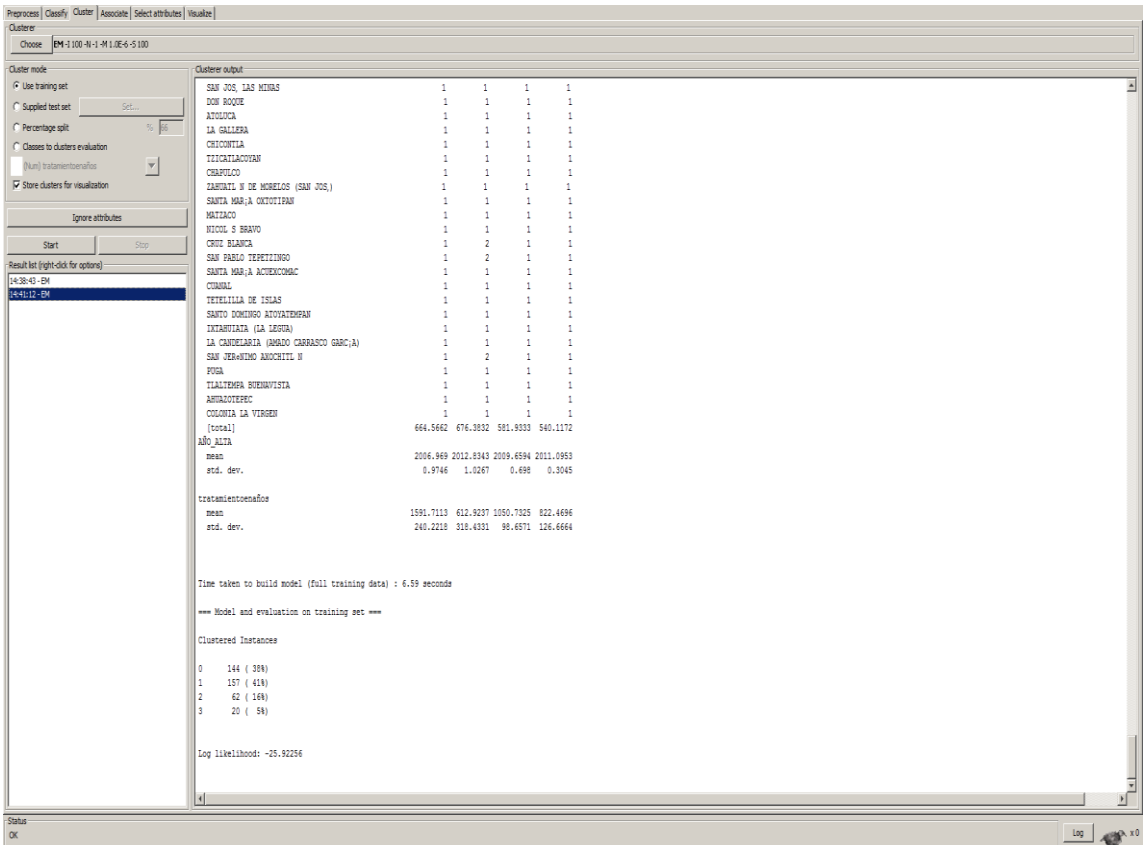

Figura 5. 5 Aplicación del método EM con número de clusters -1

5.2.2 Agrupamiento utilizando Kmeans

Para enriquecer la información suministrada por el agrupamiento se derivó un nuevo atributo: *morbilidad.* Este atributo es resultado de restar la fecha de notificación de la fecha de la última consulta. Consideramos que este atributo es importante ya que en el conjunto de datos original no existe información exacta del tiempo de vida del paciente.

Se realizaron varias ejecuciones del método *Kmeans* con diferentes números de clusters k. Se eliminaron las fechas de notificación, ingreso y alta así como también, el municipio para contar con variables que explicaran cada cluster de manera más significativa. En la figura 5.6 se muestra la aplicación de Kmeans con k=3, con un error resultante de 2263. Se nota que la población es mayoritariamente masculina radicando en las jurisdicciones de Puebla y Huejotzingo. Se observan en Puebla los mayores índices para años en tratamiento y morbilidad.

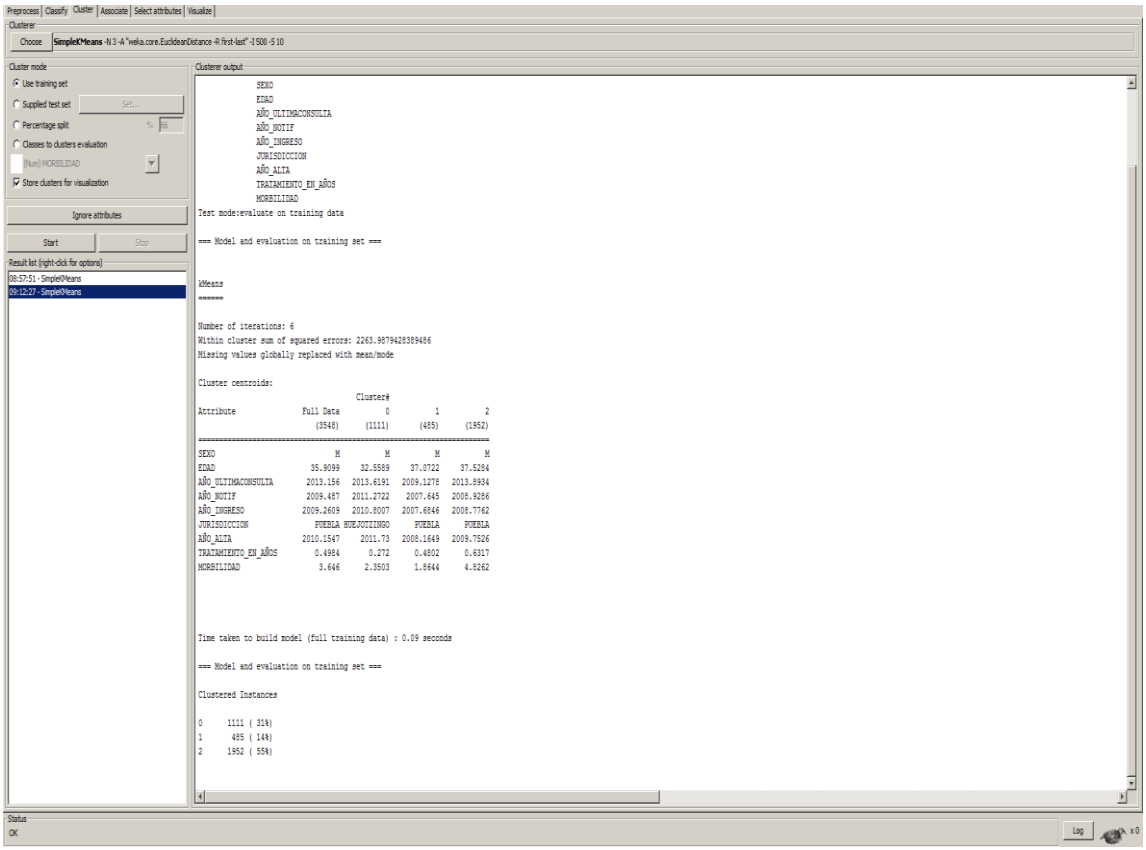

Figura 5. 6 Aplicación de Kmeans con k=3

En la figura 5.7 se grafican las jurisdicciones sanitarias de Puebla contra la morbilidad en años. Se puede apreciar a la extrema izquierda el cluster de Puebla exhibiendo una morbilidad menor a los 5 años.

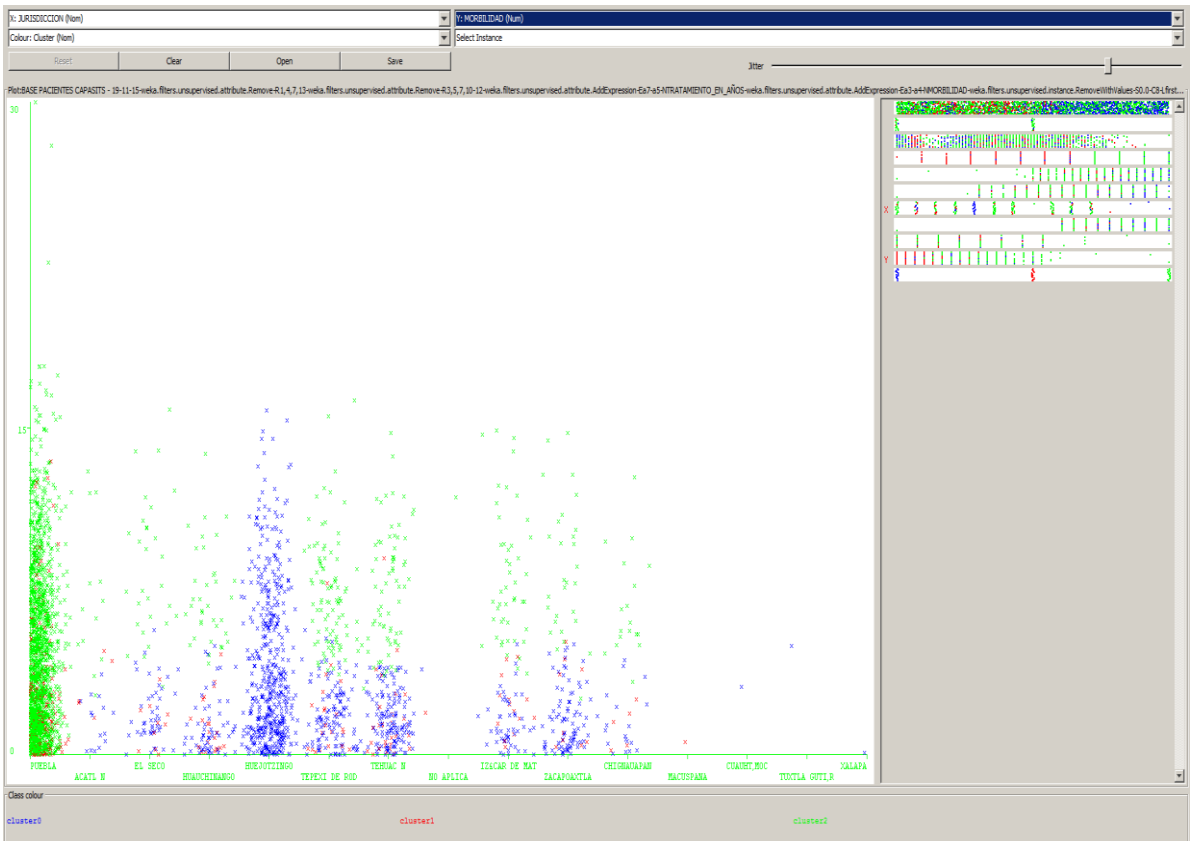

Figura 5. 7 Jurisdicción vs Morbilidad

En la figura 5.8 se muestra la aplicación de Kmeans con k=4, con un error resultante de 2214. La disminución del error con respecto a 3 clusters no es significativa. Se reportan 2 clusters para Puebla y otros 2 para Huejotzingo. Se podría destacar como una diferencia con respecto a la ejecución anterior, que la media de edad en un cluster de Huejotzingo se eleva a 43 años y por lo tanto la morbilidad asciende a más de 8 años.

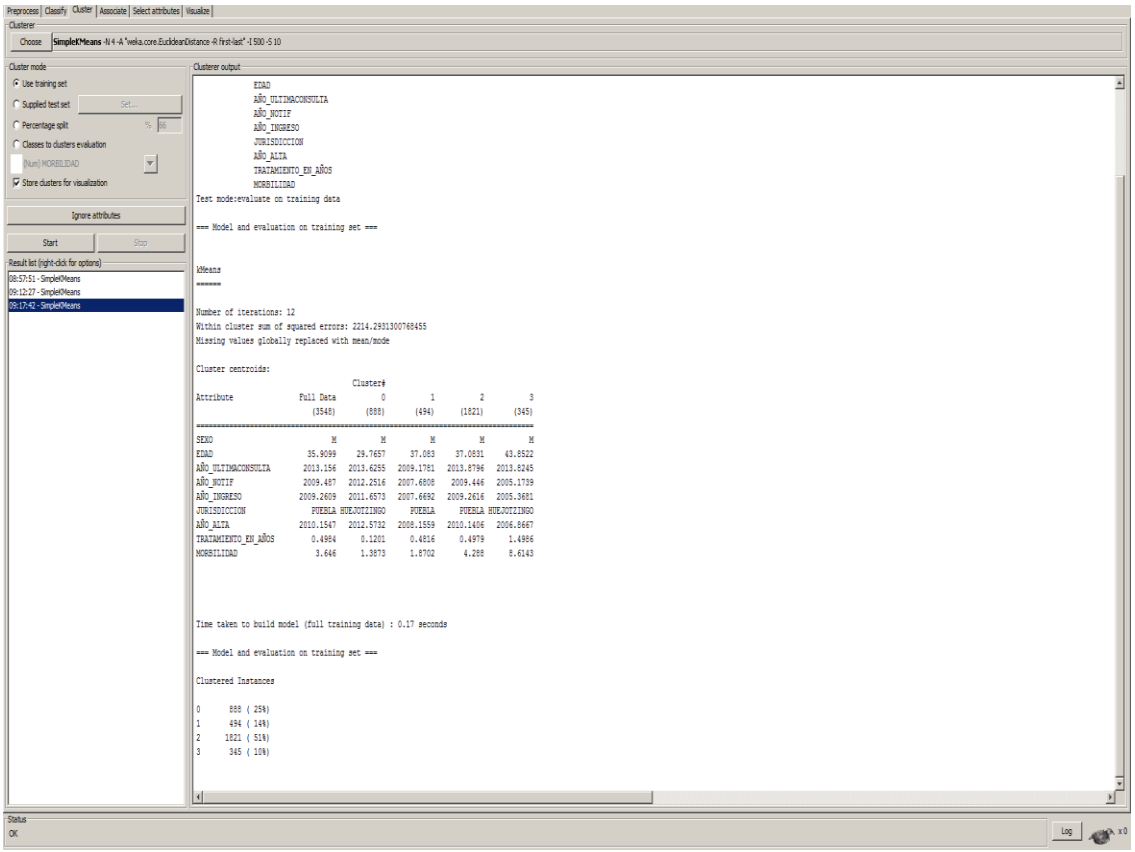

Figura 5. 8 Aplicación de Kmeans con k=4

En la figura 5.9 se grafican las jurisdicciones sanitarias de Puebla contra edad de los pacientes. Clusters en verde representan a Puebla y clusters en azul a Huejotzingo. Se aprecia una concentración de pacierntes en el custer de Huejotzingo alrededor de los 43 años.

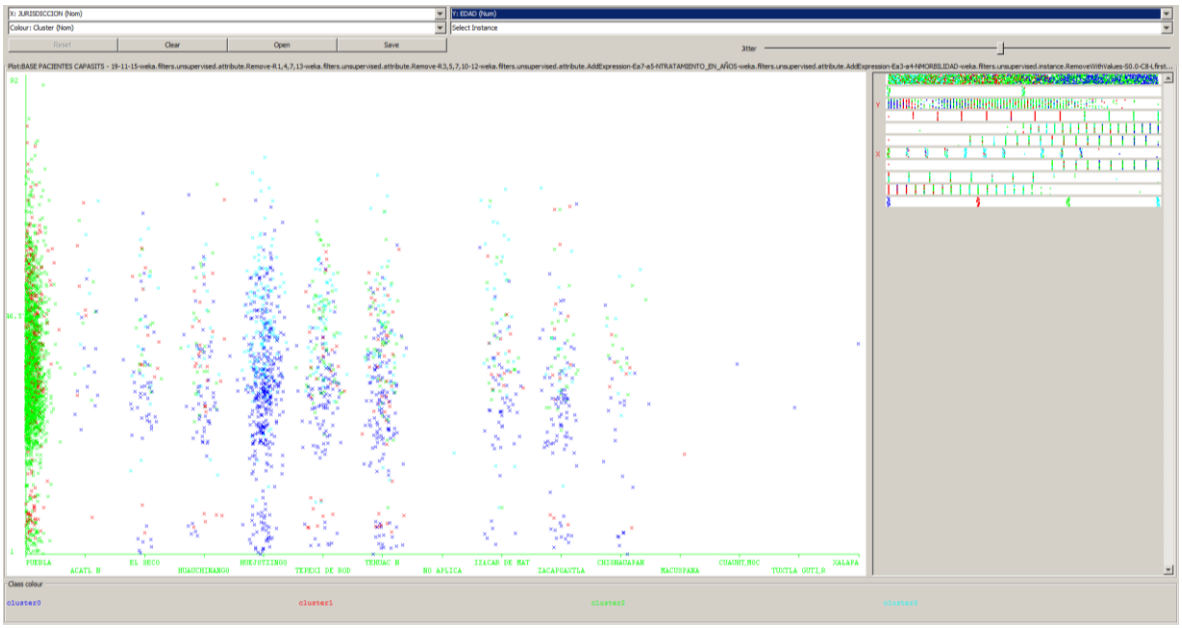

Figura 5. 9 Jurisdicción vs Edad

5.3 Técnicas de Clasificación

En esta sección se aplicarán técnicas de predicción en particular de clasificación basadas en árboles. Es requerido para la clasificación que la variable a predecir (de clase) sea nominal. Con este fin se aplicarán filtros de discretización.

5.3.1 Filtros de discretización

Estos filtros son muy útiles cuando se trabaja con atributos numéricos, puesto que muchas herramientas de análisis requieren datos simbólicos, y por lo tanto se necesita aplicar esta transformación. En la figura 5.10 se presenta la discretización del atributo *tratamientoenaños.* En particular se utilizó el filtro *numerictonominal,* es decir se conservaron los catorce valores numéricos correspondientes de 0 a 13 años de tratamiento pero ahora son valores discretos.

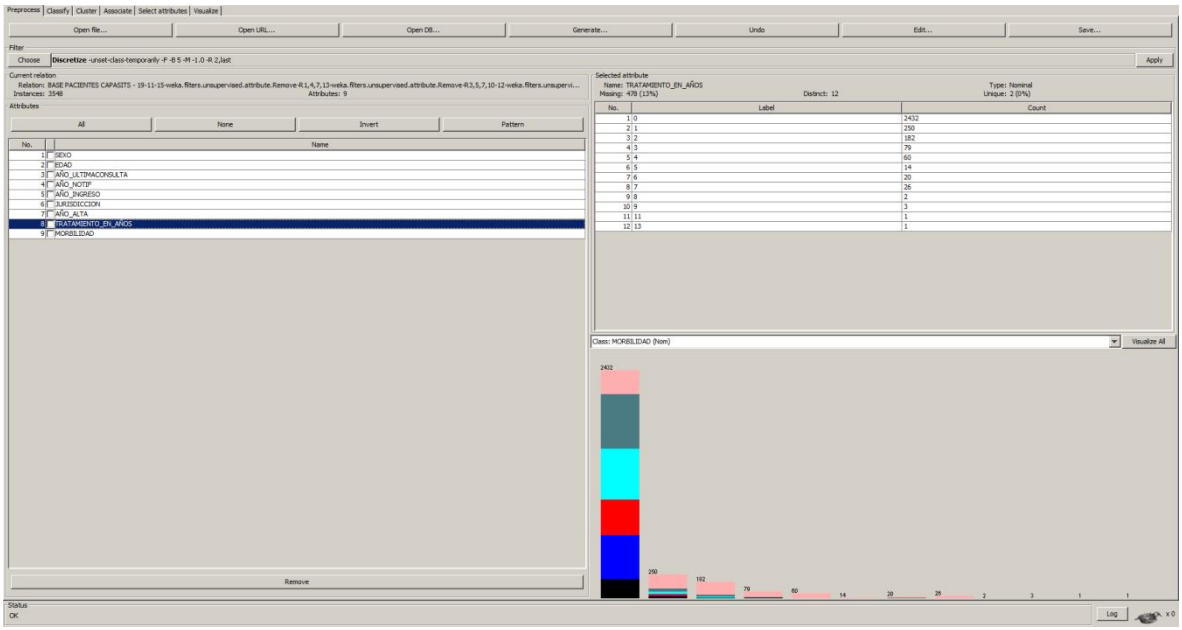

Figura 5. 10 Aplicación *NumerictoNominal* a *tratamientoenaños*

La figura 5.11 muestra la discretización del atributo edad en 5 bins de igual frecuencia pues se consideró que 5 grupos etáreos podrían facilitar la obtención y explicación de los patrones por J48.

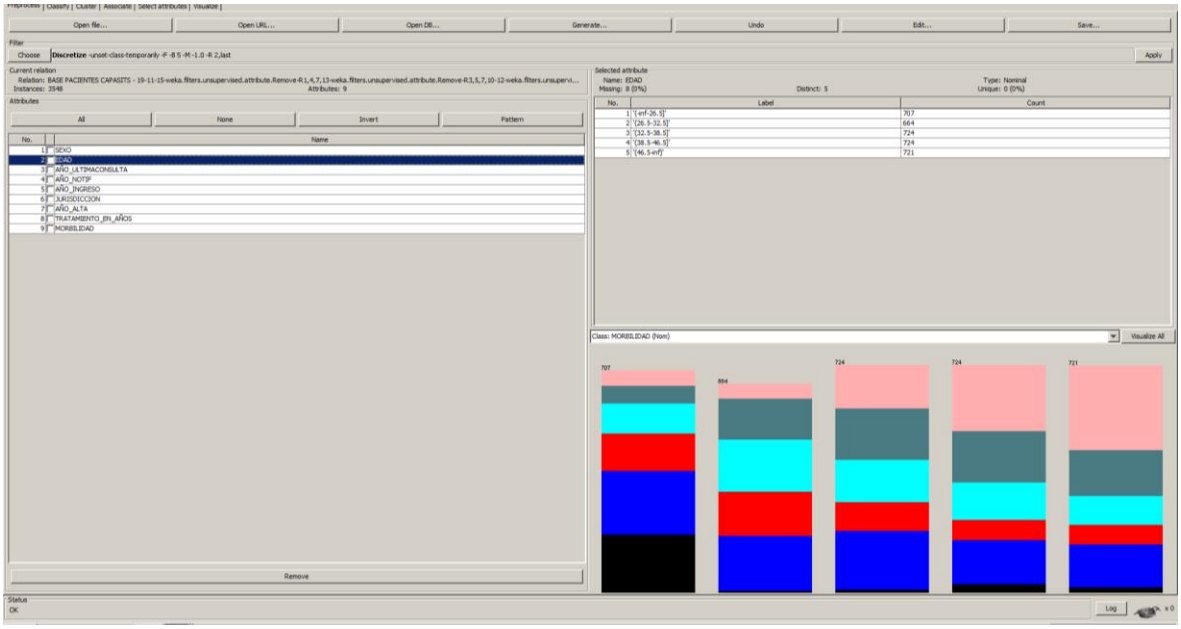

Figura 5. 11 Aplicación *Discretize* a *edad*

La figura 5.12 se muestra la discretización del atributo morbilidad en 5 bins de igual frecuencia también con el objetivo de reducir las clases a ser predichas por J48.

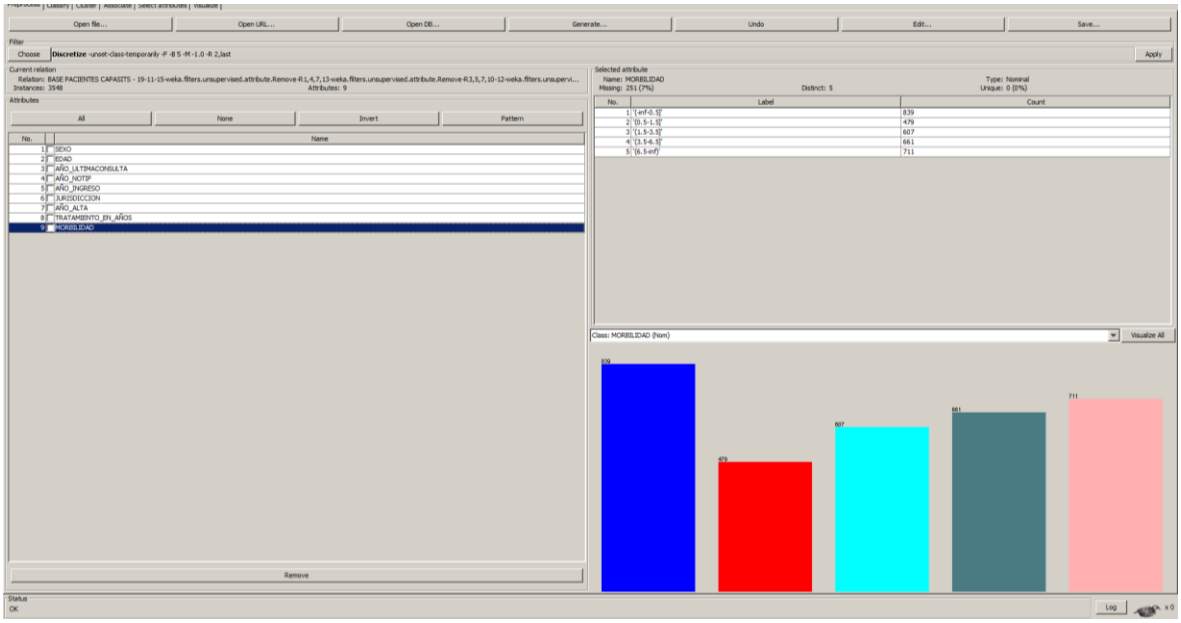

Figura 5. 12 Aplicación *Discretize* a morbilidad

Finalmente se removieron los atributos que se usaron para derivar morbilidad, estos son: años de notificación y fecha de la última consulta. En la sección 5.3.2 se realizará la predicción de morbilidad utilizando J48.

5.3.2 Predicción de Morbilidad

Como se mencionó en la sección 5.2.2, el atributo morbilidad es derivado de la fecha de última consulta y de la fecha de notificación para indicar el periodo de tiempo en que el paciente ha vivido con conocimiento de su enfermedad. En esta sección se hará la predicción de morbilidad usando J48. En la figura 5.13 se observan los resultados de la clasificación con un 68% de precisión para las 5 clases de morbilidad consideradas: (-inf,0.5), (0.5,1.5), (1.5,3.5), (3.5,6.5), (6.5,-inf).

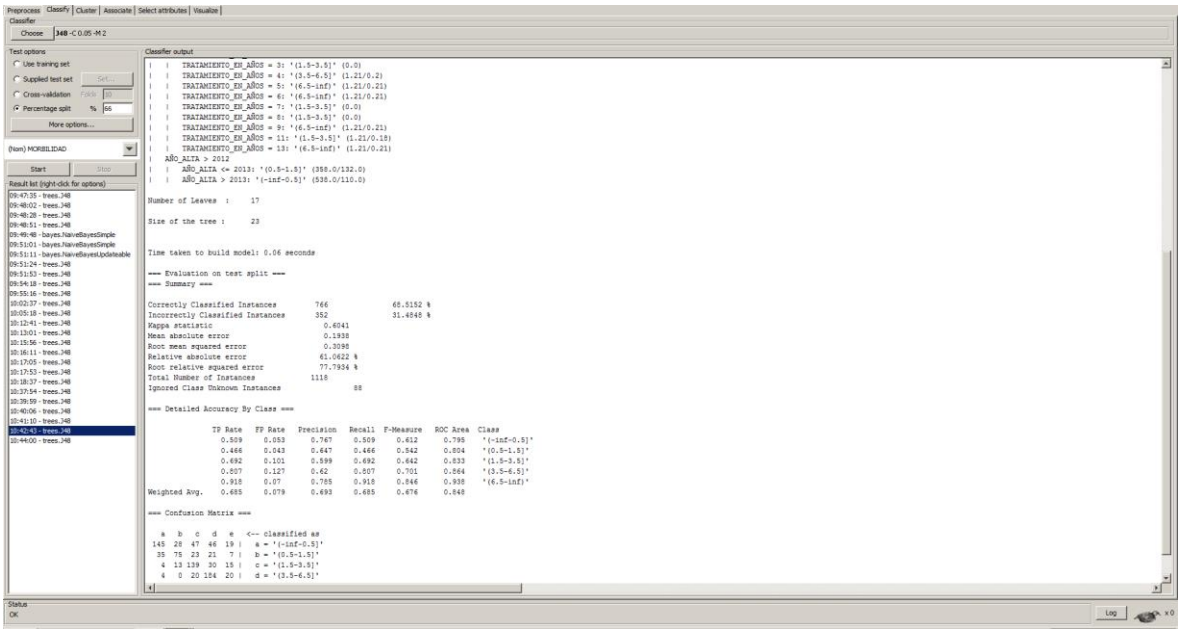

Figura 5. 13 Predicción de morbilidad por J48

En la figura 5.14 se muestra el árbol de clasificación el cual se genera en base a los atributos año de alta, ingreso y años de tratamiento. Las mejores predicciones se realizaron para la morbilidad menor a 6 meses y mayor a 6.5 años.

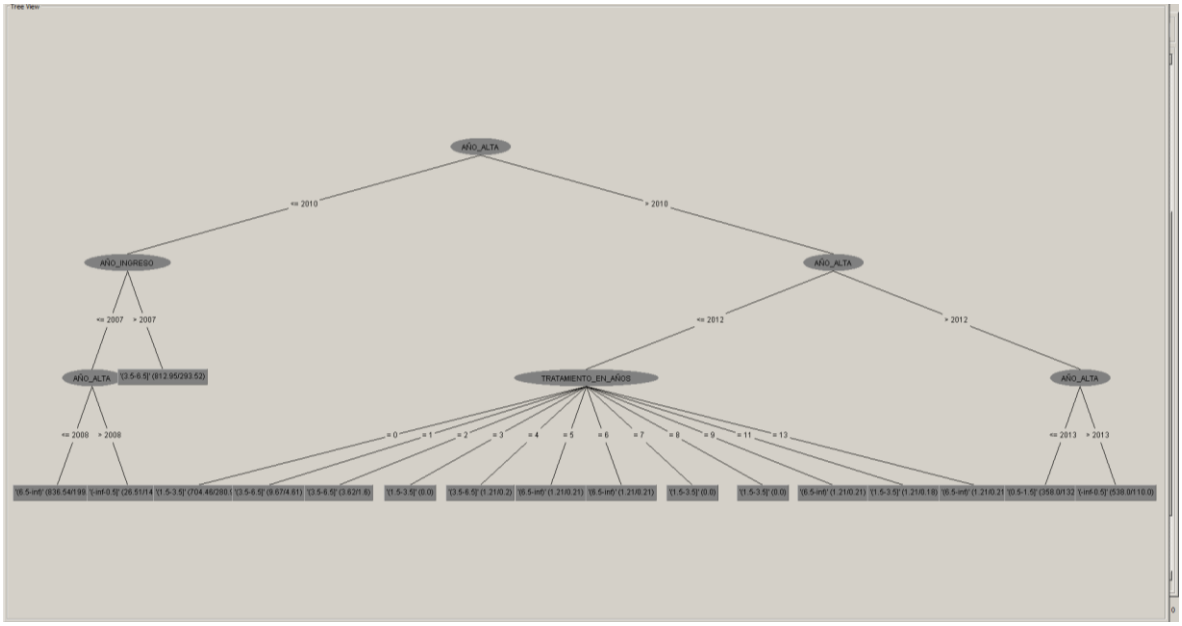

Figura 5. 14 Árbol de clasificación J48 de morbilidad

5.3.3 Predicción de tratamiento años

Se removieron los atributos que se usaron para derivar *tratamientoenaños*, estos son: año de ingreso y fecha de la fecha de alta. En la figura 5.15 se presentan los resultados del clasificador en la predicción del atributo *tratamientoenaños* con un 87% de precisión para las 14 clases de tratamientoenaños consideradas.

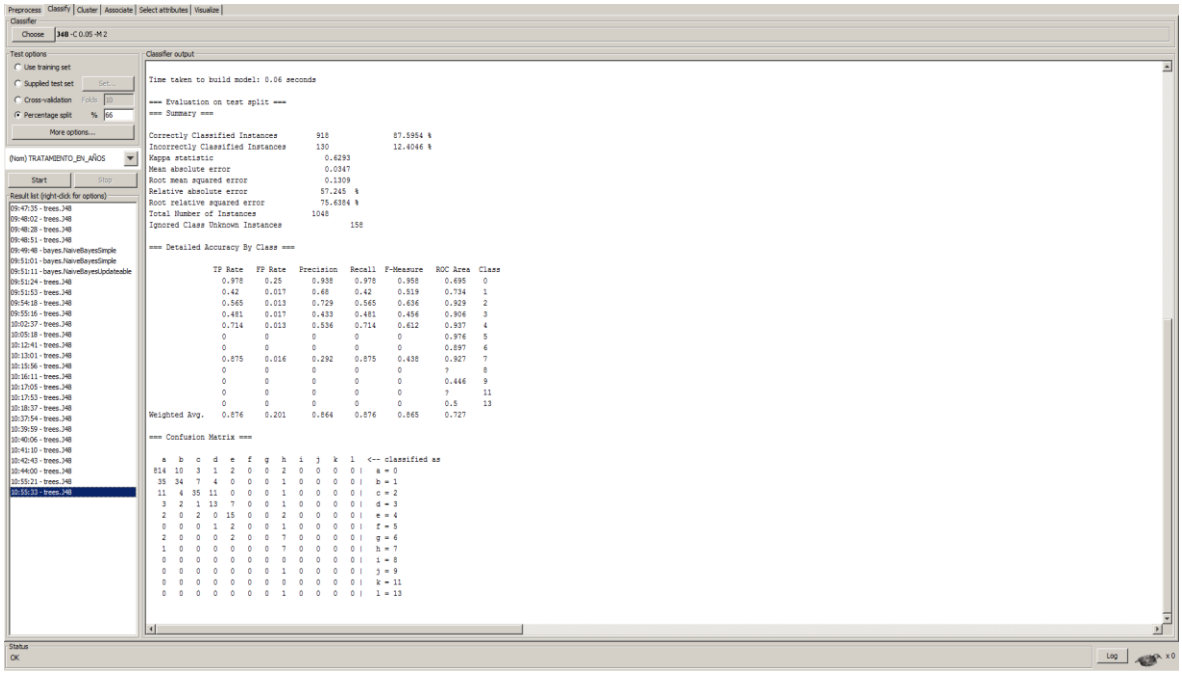

Figura 5. 15 Predicción de tratamientoenaños por J48

En la figura 5.16 se muestra el árbol de clasificación el cual se genera en base a los atributos año de notificación y años edad. Las mejores predicciones se realizaron para tratamientos menores a un año.

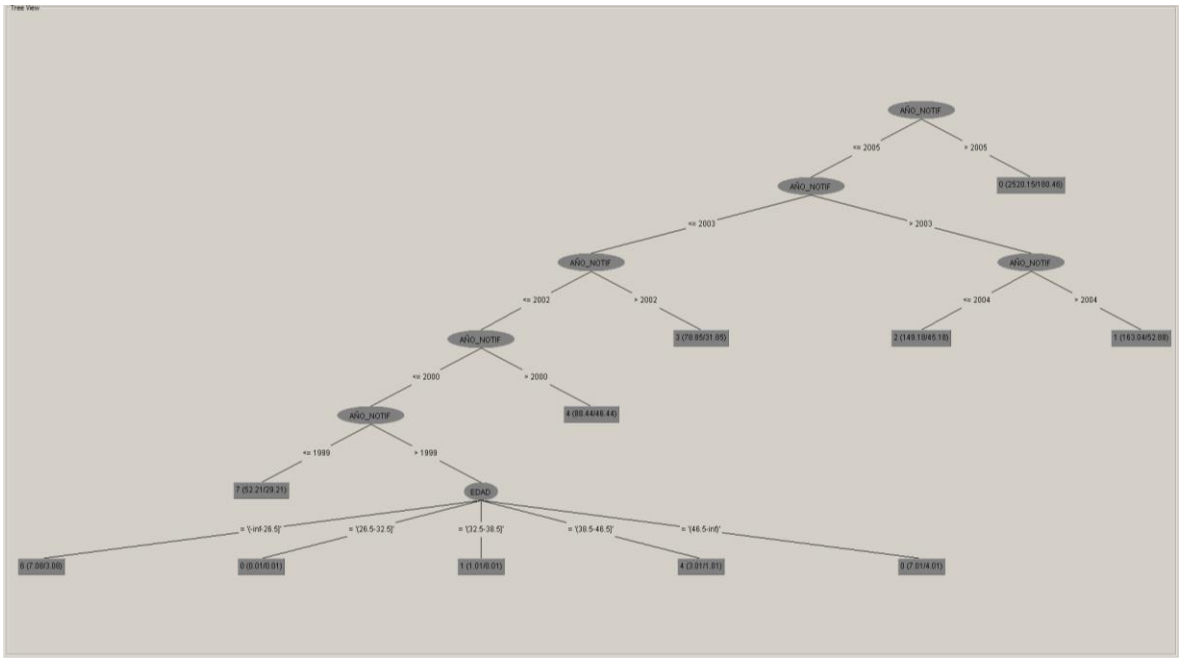

Figura 5. 16 Árbol de clasificación J48 de tratamiento en años

5.4 Conclusiones

En este capítulo se presentó la aplicación de técnicas de minería de datos. En primera instancia las técnicas de agrupamiento EM y Kmeans. Un agrupamiento de 3 clusters mostró que las jurisdicciones de Puebla y Huejotzingo presentan la mayor incidencia de VIH en población masculina de 35 años de edad. Posteriormente se aplicó el árbol de clasificación J48 donde se predijeron tratamientos menores de un año y una morbilidad de años

CAPÍTULO 6

6. Conclusiones y trabajo a futuro

Como último capítulo, se presentan las conclusiones finales de la investigación y aportaciones que se han alcanzado a lo largo de la tesis presente, así como también las mejoras que se pueden lograr en un futuro y para el caso de esta misma, le reducción de datos que se tienen a la fecha obtenidos.

6.1 Aportaciones del proyecto

Las Enfermedades de Transmisión Sexual son una gran amenaza para la salud, sobre todo para mujeres y niños en el estado de Puebla. El VIH es una de las infecciones más detectadas y con mayor índice de mortalidad si no es tratado oportunamente en las personas que lo contrajeron y que al mismo tiempo se vuelven más vulnerables.

Una de las principales aportaciones de esta tesis consiste en el conocimiento de datos reales de personas infectadas con VIH/SIDA del estado de Puebla, en relación a grupos etáreos y género. Se encontró el mayor índice de afectados en masculinos entre los entre los 28 y los 35 años de edad radicando en la ciudad de Puebla seguidos por Huejotzingo.

Se demostró con los datos tratados, que las ETS no son recientes, y que poco a poco han ido apareciendo distintas enfermedades. Estudios revelan que en especial, el VIH/SIDA hasta la fecha no tiene cura y es capaz de atacar a cualquier clase y raza de persona.

Se encontró que la muestra original de datos es inadecuada para aplicar técnicas de minería de datos porque está mal dimensionada y totalmente desbalanceada. Los datos cubren mayormente a personas de la capital del estado de Puebla los cuales tienen un mayor acceso a los servicios de salud. Es necesario que la muestra tenga una mayor presencia de pacientes que habitan en zonas rurales, de mujeres y de pacientes privados de la libertad. Por otra parte sería fundamental incluir la fecha de deceso de los pacientes de VIH ya que dicho atributo podría darnos una medida más certera acerca de la efectividad de los tratamientos que siguen los pacientes.

6.2 Trabajo a futuro

Algo fundamental en este trabajo realizado es que en la actualidad y también en el futuro, la investigación se siga enriqueciendo día a día para por lo menos, intentar combatir y/o

disminuir las ETS / VIH/SIDA, por lo que es necesario desarrollar alguna herramienta que reúna nuevos datos pero también los vaya procesando conforme vaya entrando cada uno de estos.

Como ya se demostró con los datos aportados por el Hospital General de Agua Santa, es de suma importancia que la gente conozca y aprenda de la información que se tiene acerca de cualquier ETS, en especial del VIH/SIDA por aquello de que es una enfermedad tratable pero que a la fecha no tiene una cura y para poder por lo menos disminuirla, es conociendo acerca de esta misma.

Con la información obtenida, los especialistas de estas ETS pueden definir acciones que pueden ayudar a combatir o disminuir estas Infecciones en los grupos más vulnerables, como por ejemplo los exámenes que son aplicados en distintas zonas de la ciudad y el estado de Puebla, dar la información necesaria y conferencias a personas que no estén tan informadas de estas ETS, aumentar los cuidados durante una relación sexual y disminuir el número de personas contagiadas en Puebla.

6.3 Conclusiones finales

En la tesis presente, se realizó el análisis, diseño e implementación de técnicas de minerías de datos para obtener modelos que soporten la información e infraestructura de los datos obtenidos acerca del conteo de las personas infectadas con el VIH/SIDA en la ciudad y estado de Puebla.

La experiencia obtenida al hacer esta tesis en conjunto con el área de CAPASITS del Hospital General Agua Santa del estado de Puebla, me permite concluir con las siguientes observaciones:

- Existe una cuantiosa información almacenada en distintos hospitales la cual no está siendo aprovechada para aprender más acerca de estos datos (no solo datos referentes con VIH/SIDA, ETS).

- Existe la herramienta de Minería da Datos con distribución libre y completamente gratuita, la cual se facilita para el aprendizaje con mucha información también totalmente gratis en internet.

- El software de esta tesis, puede ser utilizado por personas ajenas al ámbito informático.

- WEKA es una herramienta extraordinariamente amplia y completa, la cual puede ser de muchísima ayuda para aplicar búsquedas, filtros de datos, predicciones etc.

Respecto a las preguntas cuestionadas en el capítulo 2:

¿Desde qué épocas las ETS representan un problema de salud para México?
1980 México
Mediados de 1980 en Puebla

¿Existía algún control en épocas antiguas?

- Tripa de la res → condón
- Cianuro → morían

¿Cuáles son los tratamientos utilizados en las ETS?
VIH/SIDA → Retrovirales
En algunos ETS → Quemar la zona afectada

¿Cuáles podrían ser las consecuencias si no son tratadas?

- Esterilidad
- Ceguera en fetos
- Muerte
- Llagas en los genitales

¿Cómo se puede prevenir las ITS?

- Teniendo una sola pareja sexual
- Utilizando condón durante la relación sexual

7. Referencias

[1] Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J. & Zanasi, A.. (2005). Aplicaciones Empresariales en Data Mining. Barcelona: Book & Cd edition.

[2] Ana Teresa Fariñas Reinoso. (2001). Revista Cubana de Medicina General Integral. ISSN 1561-3038, 1, 5.

[3] Centro de Prensa. (2013). Infecciones de Transmisión Sexual. Septiembre 2014, de Organización Mundial de la Salud Sitio web: http://www.who.int/mediacentre/factsheets/fs110/es/

[4] Óscar Palacios. (2014). ITS. 12/09/14, de Norma Oficial de Salud Sitio web: http://prezi.com/appm4rrzqv3h/its-estado-del-arte/

[5] John Bonifield. (2013). Tecnologías para detectar ETS. 5/10/14, de CNN México Sitio web: http://mexico.cnn.com/tecnologia/2013/06/06/la-tecnologia-ayuda-a-saber-si-tienes-una-enfermedad-de-transmision-sexual

[6] Ana Erostarbe. (2010). Un dispositivo desarrollado por IK4 para detección de ETS. 11/10/2014, de Gaiker IK4 researche aliance Sitio web: http://www.gaiker.es/cas/noticias/un-dispositivo-desarrollado-por-ik4-para-detectar-enfermedades-de-transmision-sexual-se-comercializara-en-la-india.aspx?id=3803cdfe-4fa9-4d56-8a82-88df7fc4b2b7&pagina=8&origen=noticias

[7] RCMultimedios. (2011). Nueva aplicación para detección de ETS. 11/10/2014, de RCMultimedios Sitio web: http://rcmultimedios.mx/tecnologia/18700/nueva-aplicacion-para-detectar-enfermedades-de-transmision-sexual-

[8] Luis Susanpibar. (2014). ETS: panorama según la ONU. 28-SEP-2014, de Clínica de Urología Sitio web: http://drsusanibar.blogspot.mx/2011/08/ets-panorama-segun-la-onu.html

[9] invent_mariana. (2010). 30% de los Mexicanos ha padecido una ETS. Septiembre-2014, de Salud 180 Sitio web: http://www.salud180.com/jovenes/30-de-los-mexicanos-ha-padecido-una-ets

[10] TOMAS ALUJA. (2014). LA MINER´ DE DATOS, ENTRE LA ESTAD´ISTICA Y LA INTELIGENCIA ARTIFICIAL. 25/11/2015, de Universitat Politecnica de Catalunya Sitio web: file:///C:/Users/toshiba1/Downloads/27009-26933-1-PB.pdf

[11] Carlos Alba González-Fanjul. (2003). Proyectos Software. Estimación del Coste. 25/10/2015, de monografias.com Sitio web: http://www.monografias.com/trabajos27/estimacion-coste/estimacion-coste.shtml

[12] JC. Cantera. (2014). Míneria de datos sobre Ontologías. 2014, de Ibermática Sitio web: http://rtdibermatica.com/?p=376

[13] S/I. (2014). Métodos predictivos y Descriptivos - MINERÍA DE DATOS. 25/10/2014, de LinkedIn SlideShare Sitio web: http://es.slideshare.net/lalopg/mtodos-predictivos-y-descriptivos-minera-de-datos

[14] Fernando Berzal. (2003). Introducción al Data Introducción al Data Mining. 10/10/2014, de DECSAI Sitio web: http://elvex.ugr.es/decsai/intelligent/slides/dm/D1%20Data%20Mining.pdf

[15] N/A. (2004). Minería de datos. Noviembre-2014, de Monografias.com Sitio web: http://www.monografias.com/trabajos56/mineria-de-datos-venezuela/mineria-de-datos-venezuela2.shtml

[16] Elizabeth León Guzman. (1996). Mínería de Datos. Nov-2014, de Universidad Nacional de Colombia Sitio web: http://www.disi.unal.edu.co/profesores/eleonguz/cursos/md/presentaciones/Sesion5_Metodologias.pdf

[17] IVYthemes. (2013). From Data Mining to Knowledge Discovery in Databases. Noviembre 2014, de Arte dos Datos: True Sitio web: http://artedosdados.blogspot.mx/2015/04/from-data-mining-to-knowledge-discovery.html

[18] Juan Pedro Febles. (2005). KDD y MD. Nov-2014, de CITMA Sitio web: http://www.bibliociencias.cu/gsdl/collect/eventos/index/assoc/HASH018e.dir/doc.pdf

[19]Guillermo Molero Castillo. (2008). Técnicas de Minerías de Datos. Noviembre del 2014, de UNAM Sitio web: http://www.geologia-feflow.unam.mx/documentos/tesis%20mineria%20de%20datos.pdf

[20] Alex Guazzelli. (2012). Predicciones sobre el futuro. 05-Nov-2014, de Developer Works Sitio web: http://www.ibm.com/developerworks/ssa/industry/library/ba-predictive-analytics2/

[21]Guillermo Molero Castillo. (2008). Técnicas de Minerías de Datos. Noviembre del 2014, de UNAM Sitio web: http://www.geologia-feflow.unam.mx/documentos/tesis%20mineria%20de%20datos.pdf

[22] Jorge Iván Pincay Ponce. (2013). Dataminig con Weka. Caso: Árboles de decisión. 14/10/2015, In SlideShare Sitio web: http://es.slideshare.net/jpincay/weka-19112783

[23] Romina Laura Bot. (2005). Data Mining utilizado Redes Neuronales. 14/10/2015, de Universidad de Buenos Aires Sitio web: http://materias.fi.uba.ar/7500/bot-tesisdegradoingenieriainformatica.pdf

[24] J. C. González, M. Castellón y M. J. Castejón. (2007). 77 TÉCNICAS DE CLASIFICACIÓN EN EL ENTORNO DE WEKA. 14/10/2015, de Fundación Instituto Euro mediterráneo del Agua Sitio web: http://www.aet.org.es/congresos/xiii/cal20.pdf

[25] Joannès Vermorel. (2015). Definición de series de tiempo. 14/10/2015, de LAKAD Sitio web: http://www.lokad.com/es/que-es-el-pronostico-de-series-de-tiempo

[26] José Manuel Molina López Jesús García Herrero. (2006). TÉCNICAS DE ANÁLISIS DE DATOS . 14/10/2015, de Universidad Carlos III de Madrid Sitio web: http://www.giaa.inf.uc3m.es/docencia/II/ADatos/apuntesAD.pdf

[27] Scielo. (N/I). Esquema de las etapas del proceso del descubrimiento. 15/10/2015, de Revista Interameriacana Sitio web: http://www.google.com.mx/url?sa=i&rct=j&q=&esrc=s&source=images&cd=&ved=0CAcQjRxqFQoTCJXsvfvRxMgCFUGfgAodS5QL4Q&url=http%3A%2F%2Fwww.scielo.org.co%2Fscielo.php%3Fpid%3DS0120-09762012000100009%26script%3Dsci_arttext&psig=AFQjCNHsiyFHfJEtQ6oRRVH3l-g3YJScmQ&ust=1445002483143088

[28] Blanca Vargas. (2012). Taller Weka. 25/10/2015, in SlideShare Sitio web: http://es.slideshare.net/blancavg/taller-weka