

Minería de datos. Detección de patrones de análisis clínicos de laboratorio

Eje temático: Tecnologías de la Información y la Comunicación

Autor: Alcaraz Cabral, Enrique; enriquealc1989@gmail.com

Orientador: Ing. Ayala Díaz, Katia; ktiaayala@gmail.com

Co-Orientador: Ing. Arrúa Ginés, Jorge; jorgearrua@gmail.com

Facultad Politécnica

Universidad Nacional del Este

Resumen

Las instituciones de salud generan gran cantidad de datos, pero estos por lo general no se encuentran estructurados, son dispersos, de mala calidad y no se encuentran a disposición de manera precisa y oportuna, estos datos no constituyen información y menos aún son herramientas de utilidad para la toma de decisiones.

Uno de los servicios de suma importancia en las instituciones de salud son los laboratorios de análisis clínicos, perteneciente al grupo de las que se denominan comúnmente medios de diagnóstico. Los exámenes de laboratorio tienen como objetivos, ayudar a confirmar o descartar un diagnóstico, establecer un pronóstico, controlar la evolución de la enfermedad y los resultados del tratamiento entre otros. Por lo cual se plantea como objetivo del trabajo detectar patrones en análisis clínicos de laboratorio empleando técnicas de minería de datos.

En este trabajo se realizó la extracción de patrones mediante técnicas de minería de datos, utilizando los registros de análisis clínicos de un laboratorio correspondiente a un hospital de Ciudad del Este. Los algoritmos utilizados fueron el de a priori, Simple k-means y árbol de decisiones. Entre los resultados más significativos se obtuvieron árboles clasificando los análisis de colesterol y triglicéridos, cuyo resultado arrojó una probabilidad de 66,7% de que mujeres de entre 20 y 26 años tengan niveles altos de colesterol y una probabilidad del 50% de que mujeres de entre 22 a 27 años tengan niveles altos de triglicéridos.

Palabras clave: Técnicas de Minería de Datos, Análisis clínicos, Toma de decisión, ETL.

Introducción

Al mayor poder de procesamiento y transporte de información se suman los avances tecnológicos en materia de base de datos que soportan grandes volúmenes de información de diferentes orígenes y fuentes, pudiendo incluir texto, imágenes, video, etc. Estos avances en materia de gran capacidad de almacenamiento de información han dado lugar a la creación de nuevas tecnologías como la Minería de Datos [1].

La minería de datos es una herramienta que cobra mayor importancia en diversos entornos de la sociedad, desde el académico, pasando por el comercial, hasta el de investigación y desarrollo. Tratar de encontrar patrones, tendencias y anomalías es uno de los grandes retos de vida moderna. Código de barras, automatización de procesos en general, avances tecnológicos en almacenamiento de información y abaratamiento de precios en memoria, son algunos de los factores que han contribuido a la generación masiva de datos [2].

Objetivos

General:

Detectar patrones en análisis clínicos de laboratorio empleando técnicas de la Minería de datos.

Específicos:

- Definir los procesos de extracción, Transformación y limpieza (ETL) de datos que serán practicados en la etapa del análisis.
- Seleccionar las técnicas de minería de datos adecuadas a la ETL definida.
- Extraer datos resultantes de los modelados de MD aplicadas.
- Evaluar relaciones de patologías comunes en Paraguay entre los atributos seleccionados con expertos del área.

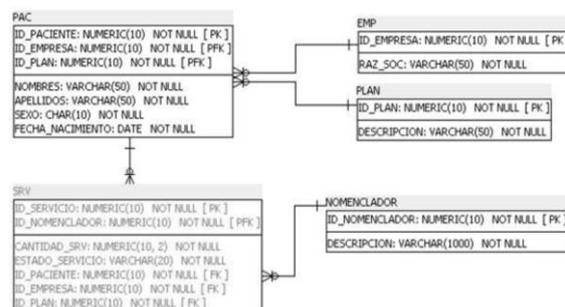
Materiales y Métodos

El presente trabajo considera las etapas del proceso KDD implementadas en base a la metodología CRISP-DM la cual permite

desarrollar un proyecto de Minería de Datos, aplicado a un sistema informático de un “Laboratorio de Análisis Clínico”, porque además de especificar detalladamente las tareas en cada fase, proporcionan guías sobre como ejecutarlas. Participantes: Se tiene la base de datos de un sanatorio privado de Ciudad Del Este, con módulos de consultas clínicas, urgencias, internaciones, laboratorio, farmacia, facturación, entre otros. Este proyecto se centró en el módulo del laboratorio de análisis clínicos. El mismo almacena datos sobre: el paciente, el análisis realizado, y a partir del segundo semestre del año 2016 almacena los resultados de dichos análisis.

El fichero principal de datos, llamado en este trabajo como conjunto principal de datos, contiene muestras correspondientes a 42.684 análisis de laboratorios realizados a partir del segundo semestre del año 2016 (01-07-16) hasta el segundo semestre del año 2017 (30-06-17). Los atributos que contienen cada registro son: nombre del estudio (análisis), fecha del estudio, identificador del paciente, sexo, fecha de nacimiento del paciente, identificador del seguro médico e identificador del plan de seguro médico. Fueron integrados la información de las siguientes tablas: pac (pacientes), srv (servicios), emp (empresa de seguro médico), plan (plan del seguro médico) y nomenclador (ver figura 1). De tal manera a que todos los datos sean contenidos en una única tabla, y puedan ser interpretados por Weka.

Figura 1: Diagrama entidad-relación con los datos seleccionados



Además del conjunto principal de datos, también se analizaron tres subconjuntos de datos, que además de las tablas y sus atributos correspondientes ya citados más arriba, contienen las siguientes tablas del sistema informático: turno lab (tabla que contiene referencias a los datos del paciente, fecha del estudio, todos los análisis realizados en cada turno, y los resultados de cada estudio), nomenclador lab (distinto a la tabla nomenclador, ya que contiene detalles específicos referente a cada análisis laboratorial) y resultado lab (resultado del análisis laboratorial). Estas tablas nos permiten obtener el atributo resultado lab (atributo del tipo numérico). Cabe mencionar que la tabla srv (servicios) no tiene referencia directa a la tabla resultado lab. Los tres subconjuntos de datos corresponden a los estudios de colesterol total (694 registros) y triglicéridos (669 registros).

La metodología utilizada se basa en el proceso KDD (Knowledge Discoverd in Database, descubrimiento de conocimiento en base de datos), es el proceso más ampliamente utilizado en MD, que dispone de tecnología de recolección, almacenamiento y gerenciamiento para grandes bases de datos estructuradas.

Selección de datos: Los datos fueron extraídos mediante consulta SQL, y almacenados en un archivo de Excel.

Limpieza de los datos: Se tiene un total de 290 tipos de análisis de laboratorio realizados, correspondientes al año 2016, y un total de 271 tipos de análisis de laboratorio realizados correspondientes al año 2017. De estos análisis se elaboró un ranking con los estudios más realizados en ambos periodos, y con más relevancia, según la bioquímica encargada del laboratorio, atendiendo las necesidades de recursos e insumos que estos utilizan. Se pudo encontrar análisis duplicados, es decir, un mismo tipo de análisis con nombres diferentes.

Transformación de los datos: Debido a incompatibilidades con el formato del tipo

fecha en el Weka, se optó por realizar transformaciones en dichos atributos: En el atributo denominado Fecha Servicio se extrajeron el año y el mes de la columna mencionada, utilizando formulas del software Microsoft Excel, y simplificando la nomenclatura del mes a sus tres letras iniciales.

Figura 2. Fórmula para atributo año

	C	D	E	F	G	H	I
1	ANALISIS	FECHA	ANHO	MES	HORA	INTERNADO_ID	CONSULTA
2	HGY	01-07-16	2016	JUL	6		14081

Instrumentos: Se utilizó el software Weka, programa se distribuido como software de libre distribución (licencia gnu-gpl desarrollado en Java, dispone de tres entornos de trabajo gráficos y un entorno en modo consola, permitiendo la implementación de algoritmos para preprocesamiento de datos, clasificación, regresión, clustering, selección de atributos, reglas de asociación, etc [5].

Procedimiento: En este trabajo se realizó la extracción de patrones mediante técnicas de minería de datos, utilizando los registros de análisis clínicos de un laboratorio correspondiente a un hospital de la región (Ciudad del Este - Paraguay). Se abordaron los siguientes modelos: Respecto al aprendizaje no supervisado, correspondientes a la parte de segmentación.

- Asociativo: A priori.
- Cluster: Simple K-means

Respecto a los modelos predictivos de aprendizaje supervisados, se aplicó el siguiente modelo:

- Arboles de decisión: J48

Resultados y Discusión

Área de estudio 1: Identificar una relación en función al tipo de seguro médico.

Asociativo:

Para la utilización de los datos en el

algoritmo a priori se procedió a realizar un pre-proceso en la herramienta Weka, que consistió en discretizar los datos, que no sean nominales. Para la ejecución se tomó el valor de soporte mínimo igual a 70%, debido a que se verificó que con valores mayores al mismo no se obtiene ninguna regla de asociación. Para este modelado se obtuvieron ocho reglas (ver figura 3).

1. HORA='(6.9-9.2)' SEXO=FEMENINO EMPRESA=5 7588 ==> RANGO=ADULTO 6193
2. HORA='(6.9-9.2)' SEXO=FEMENINO 12811 ==> RANGO=ADULTO 10147
3. HORA='(6.9-9.2)' EMPRESA=5 12071 ==> RANGO=ADULTO 9368
4. SEXO=FEMENINO EMPRESA=5 12523 ==> RANGO=ADULTO 9603
5. HORA='(6.9-9.2)' 21209 ==> RANGO=ADULTO 16053
6. SEXO=FEMENINO 21807 ==> RANGO=ADULTO 15967
7. EMPRESA=5 20527 ==> RANGO=ADULTO 14686
8. HORA='(6.9-9.2)' SEXO=MASCULINO 8398 ==> RANGO=ADULTO 5906

Figura 3. Reglas obtenidas con a priori

Para la realización de este modelado se han considerado los siguientes atributos análisis, mes, sexo, hora, empresa, rango, y la conjuración utilizada es la siguiente: Con la opción \-U" se indica el límite superior de cobertura requerido para aceptar un conjunto de ítems. Si no se encuentran conjuntos de ítems suficientes para generar las reglas requeridas se va disminuyendo el límite hasta llegar al límite inferior (opción \-M"). Con la opción \-C" se indica la confianza mínima para mostrar una regla de asociación; Y con la \-N" se indica el número de regla que se desean que aparezcan en pantalla.

Se verifica que todas las reglas cuentan con una confiabilidad (conf) con un mínimo de 70 %. Además, el indicador lift presenta en todas las reglas, valores mayores a la unidad, lo que indica que ese conjunto aparece una cantidad de veces superior a lo esperado bajo condiciones de independencia.

Realizando un análisis de las reglas para filtrarlas, y tomando en cuenta la confianza y soporte de estas, se verifico la predominancia de las reglas 1 y 3. La regla 1 presenta una frecuencia elevada de estudios clínicos realizados entre las 6 am a 9 am, de pacientes adultos del sexo femenino que poseen seguro médico de la

empresa 5, siendo esta empresa, inclusive la única presentada en la mayoría de las reglas, esta combinación aparece en un 82% de la base de datos. Además, en la regla 3 se observa nuevamente que en el mismo horario los adultos (sin discriminar por sexo) con el mismo seguro médico (empresa 5) son los que hacen uso del análisis laboratorial en el hospital, esta información con un porcentaje de confianza elevado, equivalente a un 78 %.

Clustering:

Para este modelado se tuvo en cuenta que el valor del clúster sea el valor 3, por lo cual se obtiene la siguiente salida:

```
Number of iterations: 4
Within cluster sum of squared errors: 131199.0

Initial starting points (random):

Cluster 0: CRE,ENE,'(9.2-11.5)\'',MASCULINO,'(9.5-19)\'',NINHO,5
Cluster 1: ORI,NOV,'(6.9-9.2)\'',FEMENINO,'(19-28.5)\'',ADULTO,5
Cluster 2: HEM,ENE,'(6.9-9.2)\'',FEMENINO,'(19-28.5)\'',ADULTO,1

Missing values globally replaced with mean/mode

Final cluster centroids:

Attribute      Full Data      Cluster#      1      2
                (37132.0)     (16657.0)     (14316.0)     (6159.0)

ANALISIS      HEM      HEM      ORI      HEM
MES            ENE      ENE      NOV      ENE
HORA           '(6.9-9.2)' '(6.9-9.2)' '(6.9-9.2)' '(6.9-9.2)'
SEXO          FEMENINO MASCULINO FEMENINO FEMENINO
EDAD          '(28.5-38]' '(-inf-9.5]' '(28.5-38]' '(19-28.5]'
RANGO        ADULTO   ADULTO   ADULTO   ADULTO
EMPRESA       5        5        5        1
```

Figura 4. Modelado de Clustering

Es importante recordar que el clustering no es un problema de clasificación sino de agrupación de datos. En el caso del K-means, se agruparon los datos mediante su representación por tres centroides (uno por cada clase ya que $k = 3$). Después de 4 iteraciones del algoritmo de ajuste de los centroides, se muestra cada uno de los puntos mediante sus dimensiones y la agrupación que ha descubierto. Se puede visualizar los clustering descubiertos. Otro dato interesante para resaltar es el tiempo de respuesta aplicando este algoritmo con un tiempo aproximado de 410 msec, mencionando que se inició con 3 clusters seleccionados aleatoriamente.

Análisis de resultados: Se verifica que el análisis HEM (hemograma completo) es realizado en mayor proporción en el mes de enero en el horario comprendido entre las

6,9 a 9,2 horas, teniendo una relación aproximada de 50% más de mujeres que realizan este análisis que los hombres, siendo en su mayor adulto. La misma en función a la edad puede ser clasificada de la siguiente manera:

Mujeres:

19 ≤ Edad ≤ 28, de la empresa de seguro médico 1, en menor proporción.

29 ≤ Edad ≤ 38, de la empresa de seguro médico 5, en mayor proporción.

Hombres:

0 < Edad ≤ 38, de la empresa de seguro médico 5, en una proporción importante, pero menor a la de las mujeres adultas.

El análisis ORI (orina simple) es importante que el laboratorio clínico disponga de insumos y especialistas para realizarlo en el mes de noviembre en el mismo horario mencionado arriba, para pacientes adultos, del sexo femenino.

Área de estudio 2: Buscar una relación de los análisis realizados por periodo de tiempo, ande facilitar la planificación de insumos y personal necesarios.

Asociativo: Para el siguiente modelado se tuvo en cuenta la siguiente configuración: *weka.associations.Apriori* -N 10 -T 0 -C 0.7 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1.

Utilizando los siguientes atributos: análisis, mes, hora, sexo y rango de edad, para este modelado se ha obviado el atributo empresa. De esta manera se ha generado las siguientes reglas:

1. HORA=(6.9-9.2)' SEXO=FEMENINO 12811 ==> RANGO=ADULTO 10147 <conf:(0.79)> lift:(1.15)
2. HORA=(6.9-9.2)' 21209 ==> RANGO=ADULTO 16053 <conf:(0.76)> lift:(1.1)
3. SEXO=FEMENINO 21807 ==> RANGO=ADULTO 15967 <conf:(0.73)> lift:(1.06)
4. HORA=(6.9-9.2)' SEXO=MASCULINO 8398 ==> RANGO=ADULTO 5906 <conf:(0.7)> lift:(1.02)

Figura 5. Modelado Clustering

Realizando un análisis de las reglas inicialmente se verifica la confianza mínima correspondiente a un 70 %, todas con un valor lift superior a la unidad. Una vez verificada la calidad de las reglas, se seleccionan para su presentación de análisis

las reglas 1, 2 y 4 por la información relevante adquirida en las mismas.

El resultado analizado indica, y de esta manera corrobora el horario predominante existente desde las 6.9 horas a las 9.2 horas. En la regla 1 se indica nuevamente que los pacientes del sexo femenino son las que más realizan diversos tipos de análisis, mientras que en la regla 4 se mencionan (pero en aproximadamente el 50% menos) al sexo masculino, indicando en todos los casos independientemente del sexo la predominancia de análisis en adultos.

Clustering: SimpleKmeans:

Para esta segunda área de estudio se tomó solamente los atributos que tengan relación, y aporten información útil, los cuales se citan a continuación: análisis, mes, hora, sexo, edad, rango.

Inicialmente se utilizaron 6 cluster montados aleatoriamente, la cual sirvió para finalmente adquirir los clusters centroides finales.

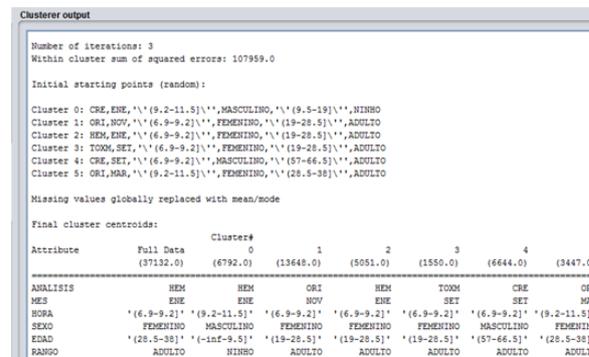


Figura 6. Modelado

En la misma se presentan los siguientes resultados:

Se corrobora los resultados referidos a mujeres presentadas en el modelado 1.2, con respecto al análisis HEM (hemograma completo), mientras que para el masculino el rango de edad se puntualizó a un rango más acotado siendo (0; 9,5 años) perteneciente a un rango de edad infantil. Para el análisis de ORI, se presentan 2 cluster (Cluster 1 y 5), pudiendo ser resumidos de la siguiente manera: Dependiendo del mes, existe un rango de

edades en que se presentan pacientes que realizan este análisis:

- En el mes de noviembre, mujeres con rango de edad adultos-jóvenes comprendidos en una edad entre 19 a 28 años.
- Marzo, mujeres con rango de edad adultos, comprendidos entre los 29 a 38 años.
- En el mes de Setiembre, dependiendo del sexo:
 - Femenino: el estudio TOXM (toxoplasmosis igm), con rango de edad comprendido entre 19 y 28 años.
 - Masculino: el estudio CRE (creatinina), con rango de edad comprendido entre los 57 a 66 años.

Área de estudio 3: Encontrar relaciones relevantes, referentes a niveles elevados de ciertos indicadores de potenciales patológicas.

Para esta _área de estudio tomando en consideración el tipo de variable, y principalmente las recomendaciones de los especialistas se seleccionaron estudios laboratoriales relevantes para el Paraguay que son:

- Colesterol.
- Triglicéridos.

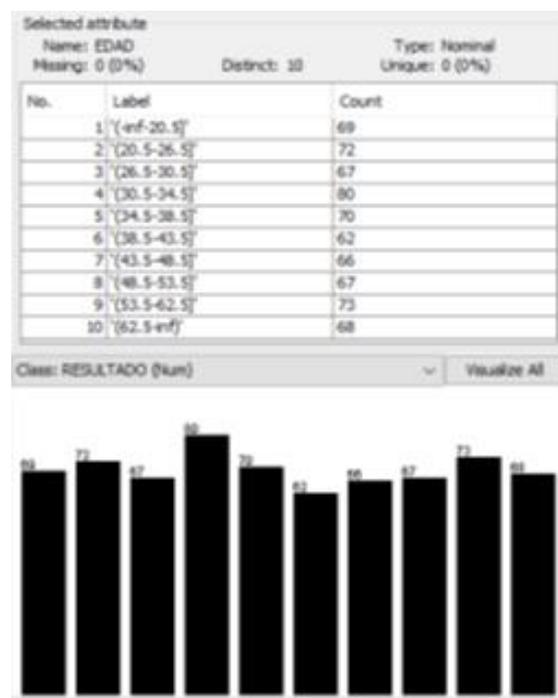
A fin de disponibilizar un arbol que presenten las probabilidades de padecer estos en función a datos como edad y sexo.

Para este estudio se utilizó el algoritmo J48 debido a que posibilito la clasificación de los niveles numéricos de estos estudios, con relación a diversos rangos etarios.

Clasificación de niveles de colesterol total (A) por Árbol J48.

Para el modelado (colesterol total), se tuvieron en cuenta los siguientes atributos: sexo (nominal), edad (numérico), resultado (numérico). Se aplicó el filtro de discretización (filters->unsupervised->attribute->discretize) para el atributo edad, con 10 intervalos equitativos (equal frequency).

La distribución etaria generada para este modelo se observa en la figura, donde para el atributo edad se generaron intervalos etarios con frecuencias equitativas con relación a la cantidad de pacientes.



Figurado 7. Distribución etaria para modelado

Fue necesario estipular parámetros: el tamaño de hojas: 2; el grado de confianza=0,25.

Este estudio se centró en los resultados que señalan un nivel de colesterol superior a los 200 mg=dL, debido a que se consideran niveles altos a los superiores a dicho valor. Describiendo el resultado del árbol generado, de manera ordenada en relación con el rango etario.

Niveles	Rangos (mg/dL)
Aconsejable	< 200
Riesgo moderado	200 - 239
Alto	geq 240

Tabla 1. Valores de referencia para Colesterol Total

En la siguiente figura se observan algunos resultados del árbol.

Edad comprendida entre: 0 a 20,5 años.

- 4 de cada 6 personas del sexo femenino poseen colesterol entre 201 y 202 mg/dL.
- 1 de cada 2 personas del sexo masculino poseen colesterol entre 226 y 227 mg/dL.

Edad comprendida entre: 20,5 y 26,5 años.

- 10 de cada 14 mujeres poseen colesterol entre 209 y 213 mg/dL.
- 2 de cada 4 mujeres poseen colesterol entre 240 y 245 mg/dL.
- 1 de cada 2 mujeres poseen colesterol entre 258 y 261 mg/dL.
- 2 de cada 3 mujeres poseen colesterol entre 273 y 275 mg/dL.

Edad comprendida entre: 26,5 a 30,5 años.

- 1 de cada 3 mujeres poseen colesterol entre 245 y 248 mg/dL.
- 1 de cada 2 mujeres poseen colesterol entre 261 y 264 mg/dL.
- 1 de cada 2 hombres poseen colesterol entre 205 y 226 mg/dL.
- 1 de cada 2 hombres poseen colesterol entre 271 y 306 mg/dL.

Edad comprendida entre: 34,5 a 38,5 años.

- 3 de cada 5 personas poseen colesterol entre 200 y 201 mg/dL.

Figura 8. Figura Resultados de modelo

Rango edad (años)	Valores (mg/dL)	Porcentaje (%)
Mujeres 0 a 20,5	201 - 202	66,6
Hombres 0 a 20,5	226 - 227	50
Mujeres 20,5 a 26,5	273 - 275	66,7
Hombres 20,5 a 26,5	niveles normales	-
Mujeres 26,5 a 30,5	261 - 264	50
Hombres 26,5 a 30,5	271 - 306	50
Personas 30,5 a 34,5	niveles normales	-
Mujeres 34,5 a 38,5	233 - 234	75
Hombres 34,5 a 38,5	294 - 306	50
Mujeres 38,5 a 43,5	niveles normales	-
Hombres 38,5 a 43,5	233 - 235	60
Mujeres 43,5 a 48,5	265 - 273	28,6
Hombres 43,5 a 48,5	271 - 294	66,7
Mujeres 48,5 a 53,5	275 - 282	71,4
Hombres 48,5 a 53,5	202 - 205	50
Mujeres 53,5 a 62,5	248 - 250	50
Hombres 53,5 a 62,5	252 - 257	40
Mujeres > 62,5	275 - 282	50
Hombres > 62,5	niveles normales	-

Clasificación de niveles de colesterol total (B) por Árbol J48

En el modelado colesterol total, se realizó la prueba similar a la anterior, pero con intervalos de edad con frecuencias no equitativas.

Se enfocó en los resultados que señalan un nivel de colesterol superior a los 200 mg=dL.

Describiendo el resultado del árbol generado, de manera ordenada en relación con el rango etario. La distribución etaria generada para este modelo se observa en la figura, donde para el atributo edad se generaron intervalos etarios con anchura

constante.

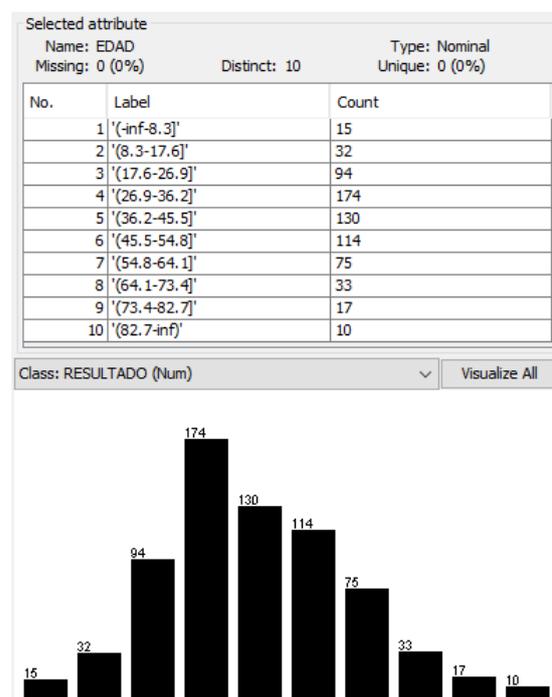


Figura 10. Distribución etaria para modelado

Tabla 5.3: Resultados resaltantes del modelado 3.2

Ambos modelos tienen resultados distintos debido a los parámetros utilizados al momento de establecer intervalos de edades, sin embargo, se pueden observar diversas similitudes que apuntan a niveles considerados altos de colesterol total:

- Se puede ver una probabilidad de 66,7% de colesterol alto en mujeres de 20,5 a 26,5 años. Mientras que en el siguiente modelado se obtuvo una probabilidad de 71,4% de colesterol moderado en mujeres de 17,6 a 26,6 años.
- También se observa una probabilidad de 50% de colesterol alto en mujeres de 26,5 a 30,5. Mientras que en el modelado 3.2 se obtuvo una probabilidad igual (50 %) de colesterol alto en mujeres de 26,9 a 36,2 años.
- Además, una probabilidad de 50% de colesterol alto en hombres de 26,5 a 30,5 años y también la misma probabilidad de colesterol alto para

hombres entre el rango de edades de entre 34,5 a 38,5 años.

- Otro resultado importante se obtuvo una probabilidad de 55,5% de colesterol alto en hombres de entre 26,6 y 36,2 años, complementando y validando los resultados.

Conclusión

Teniendo en cuenta el uso de los algoritmos, se presentan las siguientes conclusiones: Utilizando el algoritmo a priori, se obtuvo una rápida ejecución, resultando como mejor valor de confianza un 82 %, y valores de lift mayores a la unidad. Con Simple k-means, se obtuvieron resultados que prestaron relación con mayor cantidad de atributos, los cuales complementan en mayor proporción la información encontrada en el algoritmo a priori, lo cual es lógico por la naturaleza de ambos algoritmos.

Los arboles generados por el algoritmo J48 demostraron que, si bien presentaban tiempo de ejecución más elevado, los mismos cuentan con varias ramificaciones, que posibilitan tener relaciones más específicas para la toma de decisiones. Además, se concluye que el uso de MD implica la obtención de varios beneficios y ventajas para los interesados.

Dependiendo de la información que se pretenda adquirir resultara como material de apoyo para la toma de decisión tanto en

la planificación del cronograma del RRHH, ayudando a precisar la especialidad necesaria a cubrir en mayor proporción en periodos definidos de tiempo bien específicos. Asimismo, beneficiará en la planificación de adquisición insumos/materiales en base a los exámenes laboratoriales, y ofrecerá información sobre patologías en función a atributos como por ejemplo rangos etarios y sexo del paciente siendo verificada y avalada la información extraída por especialistas del área.

Bibliografía

- [1] Flores, H. Detección de Patrones de Daños y Averías en la Industria Automotriz (Doctoral dissertation, Tesis de Maestría en Ingeniería en Sistemas de Información. Facultad Regional Buenos Aires. Universidad Tecnológica Nacional). 2009.
- [2] Villalva, G., and Vinicio, M. Las herramientas de Business Intelligence para la Gestión del Recurso Humano en el Honorable Gobierno Provincial de Tungurahua (Master's thesis, Universidad Técnica de Ambato. Facultad de Ingeniería en Sistemas, Electrónica e Industrial. Maestría en Gestión de Bases de Datos. Dirección de Posgrado). 2012.
- [3] Dávila Hernández, F., and Sánchez Corales, Y. Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. Revista Cubana de Informática Médica, 4(2), 174-183. 2012.