SEP
SECRETARÍA DE EDUCACIÓN PÚBLICA

Subsecretaría de Educación Superior

Tecnológico Nacional de México

Instituto Tecnológico de Orizaba

"Año del Centenario de la Promulgación de la Constitución Política de los Estados Unidos Mexicanos"

DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN

OPCIÓN I.- TESIS

TRABAJO PROFESIONAL

"Determinación del diagnóstico situacional de las autopsias en el H.R.R.B aplicando algoritmos de aprendizaje automático para las tareas de minería de datos".

QUE PARA OBTENER EL GRADO DE:

Maestra en Sistemas Computacionales

PRESENTA:

I.S.C. Elayne Rubio Delgado

DIRECTOR DE TESIS:

Dra. Lisbeth Rodríguez Mazahua

CODIRECTOR DE TESIS:

M.C. Silvestre Gustavo Sergio Peláez Camarena

Ciencia - Técnica - Cultura 1957 - 2017 Aniversario

ORIZABA, VERACRUZ, MÉXICO. **SEPTIEMBRE 2017**

"Año del Centenario de la Promulgación de la Constitución Política de los Estados Unidos Mexicanos"

FECHA: 14/09/2017
DEPENDENCIA: POSGRADO
ASUNTO: Autorización de Impresión
OPCIÓN: I

C. **ELAYNE RUBIO DELGADO**
CANDIDATO A GRADO DE MAESTRO EN:
SISTEMAS COMPUTACIONALES

De acuerdo con el Reglamento de Titulación vigente de los Centros de Enseñanza Técnica Superior, dependiente de la Dirección General de Institutos Tecnológicos de la Secretaría de Educación Pública y habiendo cumplido con todas las indicaciones que la Comisión Revisora le hizo respecto a su Trabajo Profesional titulado:

"DETERMINACION DEL DIAGNOSTICO SITUACIONAL DE LAS AUTOPSIAS EN EL H.R.R.B APLICANDO ALGORITMOS DE APRENDIZAJE AUTOMATICO PARA LAS TAREAS DE MINERIA DE DATOS".

Comunico a Usted que este Departamento concede su autorización para que proceda a la impresión del mismo.

C.A. TITULACIÓN
SECRETARIA DE EDUCACIÓN PÚBLICA
INSTITUTO TECNOLÓGICO DE ORIZABA

A T E N T A M E N T E

M.C. MA. ELENA GARCÍA REYES
JEFE DE LA DIV. DE ESTUDIOS DE POSGRADO

ggc

"Año del Centenario de la Promulgación de la Constitución Política de los Estados Unidos Mexicanos"

FECHA : 04/09/2017

ASUNTO: Revisión de Trabajo Escrito

C. M.C. MA. ELENA GARCÍA REYES
JEFE DE LA DIVISION DE ESTUDIOS
DE POSGRADO E INVESTIGACION.
P R E S E N T E

Los que suscriben, miembros del jurado, han realizado la revisión de la Tesis del (la) C. :

ELAYNE RUBIO DELGADO

la cual lleva el título de:

"DETERMINACION DEL DIAGNOSTICO SITUACIONAL DE LAS AUTOPSIAS EN EL H.R.R.B APLICANDO ALGORITMOS DE APRENDIZAJE AUTOMATICO PARA LAS TAREAS DE MINERIA DE DATOS".

Y concluyen que se acepta.

A T E N T A M E N T E

PRESIDENTE : DRA. LISBETH RODRIGUEZ MAZAHUA — FIRMA

SECRETARIO : M.C. SILVESTRE GUSTAVO SERGIO PELAEZ — FIRMA

VOCAL : M.C. MARIA ANTONIETA ABUD FIGUEROA — FIRMA

VOCAL SUP. : M.S.C.ALMA IVONNE SANCHEZ GARCIA — FIRMA

EGRESADO(A) DE LA MAESTRIA EN SISTEMAS COMPUTACIONALES

OPCION: I Tesis

Agradecimientos

Considero un gran privilegio el haber tenido la oportunidad de pertenecer a este maravilloso sistema de posgrado. Sin lugar a dudas, fue una meta alcanzada gracias al apoyo incondicional de mi esposo, quien fungió como motor impulsor haciendo posible que hoy pudiese estar escribiendo estas palabras de agradecimiento. Le doy gracias a él, a mis padres y demás familiares, por permitirme volar alto y apoyarme incondicionalmente.

Agradezco al comité de la maestría, especialmente al maestro Camarena, por darme la oportunidad de formar parte del posgrado, a la maestra Celia por su paciencia, apoyo y comprensión durante el complejo proceso de permanencia en el programa por mi condición de extranjera, a la maestra Betty, como cariñosamente todos la llamamos, porque me apoyó mucho tanto en el plano académico como en el personal y por supuesto, al excelente claustro de profesores por su exigencia, compromiso y dedicación, ya que sin ellos esta experiencia no hubiese sido lo mismo.

Quiero agradecer especialmente a la Dra. Lisbeth por su distinguible manera de apoyarme durante la investigación, escuchando y respetando mis puntos de vista y ayudándome a encontrar siempre la mejor alternativa. De igual manera, al Dr. Palet por estar siempre accesible y dispuesto a colaborar. También tengo mucho que agradecer al Dr. Asdrúbal por su asesoría en la estancia, ya que los resultados de este período fueron muy importantes para toda la investigación.

Quiero finalizar, por ser muy importante, agradeciéndole a mis compañeros de clase, que desde el primer día fueron incondicionales, sin duda alguna son una de las mejores cosas que me llevo de este período.

Sin lugar a dudas han sido dos años que han enriquecido mi vida a un nivel que no imaginé, por lo que le agradezco a la vida el haberme permitido estar y formar parte de la generación 2015-2017 de la Maestría en Sistemas Computacionales del Instituto Tecnológico de Orizaba, Veracruz, México.

«El valor del producto se halla en la producción»

Albert Einstein

Índice General

Índice de Tablas

Índice de Figuras

Resumen

La minería de datos (MD) propone un conjunto de técnicas que permiten analizar datos y extraer conocimiento novedoso y útil. Ello favorece la toma de decisiones y la concepción de estrategias dentro de las empresas, cuyo fin supremo es la elevación de la calidad de sus procesos. En este sentido, el objetivo de la presente investigación es identificar, mediante la aplicación de algoritmos de aprendizaje automático para las tareas de MD, los factores que influyen en la disminución de la realización de autopsias en el Hospital Regional de Río Blanco (H.R.R.B.).

El conjunto de datos está conformado por las opiniones del personal médico de este centro de salud, recogidas a través de encuestas y codificadas en una serie de variables categóricas. La investigación se enfoca en realizar un análisis descriptivo para comprender los datos y, a partir de esto, determinar las relaciones existentes entre las variables categóricas que los conforman, lo que permite obtener un modelo basado en reglas mediante el que un experto es capaz de evaluar cuáles son los factores influyentes en el rechazo hacia la práctica de autopsias.

Las técnicas de minería de datos que se aplicaron en los conjuntos de datos formados a partir de las opiniones de los médicos del hospital fueron las reglas de asociación y las redes bayesianas. Para la exploración y extracción del conocimiento, se consideraron los algoritmos Apriori, FPGrowth, PredictiveApriori, Tertius, J48, NaiveBayes, SMO, MultilayerPerceptron y BayesNet, todos ellos provistos por la API (*Application Programming Interface*, Interfaz de Programación de Aplicaciones) de Weka. Para generar los modelos de minería y presentar el nuevo conocimiento en lenguaje natural, se desarrolló una aplicación web. Los resultados presentados en este estudio son los obtenidos de los algoritmos mejor evaluados, los cuales fueron validados por un especialista del campo de la patología.

El marco de trabajo de esta investigación se conformó de la siguiente manera: se siguió el método KDD (*Knowledge Discovery from Data*, Descubrimiento de Conocimiento a partir de los Datos) para guiar el proceso de minería de datos y se utilizó la API de Weka para generar e interpretar modelos de minería desde la aplicación web desarrollada en Java bajo la metodología ágil UWE (UML-Based Web Engineering, Ingeniería Web basada en UML). Además, para garantizar la persistencia de los datos se ocupó el Sistema Gestor de Bases de Datos (SGBD) PostgreSQL.

Abstract

Data mining (DM) proposes a set of techniques that allow analyzing data and extracting new and useful knowledge. This favors the decision-making and the conception of strategies within the companies, whose supreme aim is the increase of the quality of their processes. In this sense, the objective of the present research is to identify, through the application of machine learning algorithms for the tasks of DM, the factors that influence the reduction of autopsies in the Regional Hospital of Rio Blanco (H.R.R.B.).

The data set consists of the opinions of the medical staff of this health center, collected through surveys and codified in a series of categorical variables. The research focuses on performing a descriptive analysis to understand the data and, from this, determining the existing relationships between the categorical variables that make them, what allows obtaining a model based on rules by which an expert can evaluate what are the influential factors in the rejection towards the practice of autopsies.

The data mining techniques that were applied to data sets formed from the opinions of hospital physicians were the association rules and the Bayesian networks. For the exploration and extraction of knowledge, the algorithms of Apriori, FPGrowth, PredictiveApriori, Tertius, J48, NaiveBayes, SMO, MultilayerPerceptron and BayesNet were considered, all provided by the Weka API (Application Programming Interface). To generate the mining models and present the new knowledge in a natural language, a web application was developed. The results presented in this study are those obtained from the best evaluated algorithms, which were validated by a specialist in the field of pathology.

The framework of this research was shaped as follows: the KDD (Knowledge Discovery from Data) method was followed to guide the data mining process and the Weka API was used to generate and interpret mining models from the web application developed in Java under the agile methodology UWE (UML-Based Web Engineering). In addition, to ensure the persistence of the data, the Database Management System (DBMS) PostgreSQL was used.

Introducción

La MD es una de las herramientas más usadas hoy en día en sistemas computacionales con manejo de grandes cantidades de datos. Su uso permite a las empresas aprovechar el conocimiento para ajustar sus procesos en aras de ser más competitivas y eficientes, de ahí que sean muchas las áreas en las que se emplea.

La problemática que estudia esta investigación se relaciona con la alarmante disminución en la práctica de autopsias en los centros hospitalarios del país, hecho sobre el que se pretende esclarecer sus factores influyentes, mediante la aplicación de algoritmos de aprendizaje automático para las técnicas de MD.

Para esto último, se analizan las características de cada técnica de minería, lo que permite determinar la más adecuada para la solución demandada. De esta manera, los resultados servirán a las autoridades competentes para el diseño de estrategias y la toma de acciones que posibiliten un resurgir de la aceptación y realización de la citada práctica médica, que contribuye a la calidad de la Medicina.

Para una mayor comprensión de la investigación, sus objetivos, metodología, entre otros, el presente trabajo de investigación se estructura en cinco capítulos. En el primero de ellos se describe el problema a resolver, los objetivos perseguidos y las razones que justifican la necesidad de dar solución al problema. En el segundo se presenta un análisis comparativo de los trabajos existentes que se relacionan con los procesos implicados en la investigación descrita en este proyecto y la alternativa de solución al problema abordado donde se eligen las tecnologías y metodologías más adecuadas, tanto para el desarrollo del proyecto como para llevar a cabo el proceso de MD. En el capítulo cuatro se describen las actividades que dicta la metodología que apoya a la investigación y los resultados obtenidos. En el quinto y último capítulo se presentan las conclusiones y recomendaciones de este trabajo.

.

Capítulo 1. Antecedentes

Este primer capítulo define las categorías y subcategorías de análisis del estudio a partir de postulados de varios autores que reflexionaron sobre temas relacionados con el mismo, lo cual permite dar forma al cuerpo teórico que sustenta toda investigación científica.

Contiene también la fundamentación metodológica de la investigación, en la que se analizan las metodologías existentes para la minería de datos, sus tipos de análisis, sus técnicas y plataformas más utilizadas, de los cuales, algunos apoyarán y guiarán el proceso en aras de alcanzar y resolver el objeto de estudio.

De igual forma se explicitan aspectos metodológicos esenciales de toda obra investigativa como el problema de investigación, los objetivos generales y específicos del estudio, y una breve justificación sobre la pertinencia, relevancia y actualidad del tema central.

1.1 Marco teórico

1.1.1 Autopsia

La palabra autopsia define el método médico mediante el que un especialista en patología reconoce un cadáver para determinar y dictaminar la verdadera causa de la muerte. Este método también se practica con objetivos científicos, para estudios e investigaciones, ya que constituye el análisis más minucioso que se hace para comprobar los efectos de una enfermedad en la anatomía humana y los procesos biológicos, incluso a nivel celular. Mediante las autopsias los expertos evalúan cualitativamente los diagnósticos y tratamientos que se practicaron al paciente pre-mortem, razón por la que es universalmente aceptado que son prácticas que preservan la calidad en la medicina [1].

1.1.2 Conjunto de datos

Un conjunto de datos no es más que un grupo de información estructurada que responde a determinada cuestión de interés, también conocido como *"dataset"*. El conjunto se representa como matriz o tabla, donde cada columna responde a una variable, y cada fila representa un registro.

1.1.3 Aprendizaje Automático

El aprendizaje automático dentro del área computacional se enfoca en analizar cómo funciona el razonamiento lógico, ya que su objetivo principal es crear sistemas que infieran conocimiento de forma completamente automática o semi-automática y sean capaces de dar respuestas adecuadas. *"... cualquiera que sea el proceso de adquisición de conocimientos o habilidades, si lo realiza una máquina se denomina aprendizaje automático*" [2]. El aprendizaje automático se logra a través de diferentes ámbitos como son aprendizaje supervisado, cuando se pretende generalizar asociaciones a partir del conjunto de datos de entrada (etiquetados) y responder apropiadamente cuando reciban nuevas instancias. Cuando hay una parte de los datos iniciales cuya salida esperada se desconoce, se pone en práctica el aprendizaje semi-supervisado y el no supervisado se implementará cuando no se conozca *a priori* el etiquetado de las entradas.

1.1.4 Minería de datos

En las empresas o entidades de cualquier área, ya sea en la educativa, la política, la médica, por citar algunas, el recurso más importante es sin duda la información que se genera en grandes volúmenes, siendo así hasta en las pequeñas organizaciones. Proporcionalmente al crecimiento de los datos, aumenta la dificultad para analizarlos. La Minería de Datos (MD) se convirtió en una herramienta muy importante, explotada por años, para el análisis avanzado de grandes cantidades de datos. Es la manera de descubrir de forma automatizada y no trivial el conocimiento que se encuentra implícito en los datos almacenados y que es totalmente desconocido, lo que resulta potencialmente útil.

La MD está muy ligada a métodos matemáticos y a la Inteligencia Artificial y su origen surge de la necesidad de analizar y descubrir el conocimiento alojado en los diferentes repositorios de información existentes hoy en día. Es un proceso que se basa en técnicas para la exploración de los datos y la identificación de patrones para extraer el conocimiento nuevo y útil de forma comprensible, válido para determinar factores de influencia dentro de procesos, para apoyar la toma de decisiones y que permita predecir variables o comportamientos futuros y agrupe elementos por semejanzas. Muchas áreas, desde hace algunos años utilizan la minería

con algunos de estos fines, encontrando así, resultados satisfactorios, novedosos y necesarios para su desarrollo.

1.1.4.1 Metodología

"Ciencia del método", es el significado que le da la Real Academia Española [3] a la palabra metodología, se interpreta como el procedimiento a seguir en aras de solucionar una problemática dada.

La MD es una ciencia compleja que necesita la guía de una metodología para desarrollar productos de calidad, y es por eso que hoy existen muchas de éstas que dirigen la planificación y ejecución de los proyectos. De acuerdo con la comunidad KDnuggets, las más utilizadas son CRISP-DM (*Cross Industry Standard Process for Data Mining*, Proceso Estándar Industrial de Corte para Minería de Datos), SEMMA (*Sample, Explore, Modify, Model and Assess*, Probar, Explorar, Modificar, Modelar y Evaluar), KDD y Catalyst, en ese orden, cada una de ellas se analiza en el capítulo tres.

1.1.4.2 Tipos de Análisis

Las tareas de la Minería de Datos se clasifican en predictivas y descriptivas.

El tipo de análisis predictivo, mediante las tareas de Clasificación y Regresión, precisa de un conjunto de entrenamiento conformado por los atributos y una etiqueta de clase. En las tareas de predicción discreta, la variable de clase es categórica y en las de predicción continua, la etiqueta de clase es numérica.

El análisis descriptivo solo requiere del conjunto conformado por los atributos que se quieren analizar para encontrar similitudes o asociaciones mediante la aplicación de tareas de Agrupamiento y Asociación.

1.1.4.3 Técnicas

Son muchas las técnicas de minería de datos que hoy en día se utilizan y para identificar cuál de ellas aplicar a una determinada situación, se tienen que analizar los tipos de atributos e identificar el objetivo que se quiere alcanzar mediante la aplicación de la minería. Las técnicas

se agrupan en supervisadas, cuando se predicen datos desconocidos a partir de los que ya se conocen, y en no supervisadas, cuando lo que se busca es obtener patrones y tendencias en los datos.

A continuación se listan algunas de las técnicas más conocidas, para el análisis predictivo es posible ocupar Redes Neuronales, Árboles de Decisión, Máquinas de Soporte Vectorial, Métodos de Regresión y Métodos Bayesianos, en cambio, para el análisis descriptivo se utilizan los métodos de Agrupamiento Particional, Jerárquico, basado en Densidad y basado en Cuadrícula, así como Correlación y Reglas de Asociación [4].

Descripción de las tareas y técnicas más utilizadas:

- **Redes Neuronales:** Se trata de un sistema de interconexión entre los elementos en una red que colaboran para producir un resultado de salida.
- **Árboles de Decisión:** Es un modelo predictivo útil para representar condiciones y categorizarlas.
- **Clasificación:** Su objetivo es predecir a qué clase pertenece una nueva instancia de acuerdo a los datos de entrada, de los cuales es necesario conocer su etiqueta o clase.
- **Regresión:** Predice en una nueva instancia el valor de una variable cuantitativa mediante las variables conocidas.
- **Métodos Bayesianos:** Infieren la probabilidad de que una hipótesis sea cierta.
- **Agrupamiento:** Es la agrupación de una serie de vectores usando criterios de distancia donde estarán más cercanos los que más características en común tengan.
- **Correlaciones:** Muestran la semejanza entre variables cuantitativas.
- **Reglas de asociación:** Permiten reconocer las relaciones implícitas que existen entre variables categóricas, por lo que es posible descubrir hechos comunes dentro de un conjunto de datos. Dentro de estas reglas también se encuentran las secuenciales que permiten buscar relaciones temporales en los datos.

1.1.4.4 Plataformas

Para dar soporte y aplicar las técnicas de la minería de datos existen algunas herramientas disponibles. De acuerdo con [5] las que tienen mejor criterio entre los profesionales son WEKA (*Waikato Environment for Knowledge Analysis*, Entorno para el Análisis del Conocimiento de la Universidad de Waikato), R y RapidMiner. Otras de estas herramientas son *"SAS Enterprise Miner"* y *"SPSS Modeler"*.

1.1.5 Minería de texto

Así como la minería de datos extrae conocimiento útil y novedoso dentro de grandes colecciones de datos, la minería de texto (MT) [6] busca patrones y relaciones en grandes cantidades de documentos no estructurados hasta extraer el conocimiento implícito dentro de los textos de la colección. Las tres técnicas más importantes dentro de MT son: extracción de términos, extracción de información y análisis de enlaces. Los códigos computacionales generados para estos fines son generalmente complejos, además se requiere de expertos para extraer la información.

1.2 Situación tecnológica, económica y operativa de la empresa.

El Hospital Regional de Río Blanco, ubicado en el estado de Veracruz, es uno de los centros médicos más importantes de la región. Actualmente brinda atención médica a más de 735 mil pacientes pertenecientes a los 57 municipios más cercanos como son Río Blanco, Orizaba, Córdoba, Coscomatepec, Zongólica, Nogales y Cd. Mendoza, por mencionar algunos, e incluso de otros estados como Oaxaca y Puebla.

Este centro médico es considerado uno de los hospitales con mayor cantidad de atenciones a pacientes registradas en México, por lo que procura mantener actualizado su equipamiento técnico, procurando que sean equipos de última tecnología, para brindar un mejor servicio a los pacientes y elevar la calidad de los resultados en los exámenes médicos. Ejemplo de ello es la adquisición de un equipo de tomografía computarizada multicortes en el mes de Julio de 2015, el cual se gestionó debido a la demanda de este tipo especializado de diagnóstico.

1.3 Planteamiento del problema

La ausencia y/o práctica de las autopsias en general, refleja en una parte un estado de práctica de la medicina institucional moderna basada o sustentada en información obtenida a partir del uso y aplicación de la tecnología en todas sus manifestaciones y enfoques, incluyendo las informáticas, que se aplican en la medicina actual, totalmente alejado del principio de la observación *in situ* de la enfermedad que se manifiesta en los cambios anatómicos, histológicos, bioquímicos y fisiopatológicos, por decir los más importantes.

La consideración primaria acerca de que los estudios de autopsias son la parte más importante de la práctica médica en relación con la aportación de información vital de los hallazgos encontrados que enriquecen la práctica médica diaria, plantea la pregunta incuestionable acerca de su práctica, su aceptación o de su rechazo. La tendencia nacional y mundial va dirigida hacia la disminución en la práctica de autopsias.

La revisión de las referencias menciona frecuencias y estadísticas acerca de los números de autopsias, sin otras especificaciones, pero poco se aclara lo relacionado con las causas o motivos por las que no se hacen este tipo de estudios.

Existen pocos, no existen o no se conocen estudios, donde se abra una investigación directa y específica hacía los médicos y autoridades que informen o refieran por qué no se solicitan los estudios de las autopsias en los hospitales.

Se reconoce en general, que la tendencia en la práctica de los estudios posmortem es baja en la mayor parte de los servicios de patología de países de primer orden. También se conoce que la tendencia es similar en los centros hospitalarios de las ciudades más grandes del país, como son las ciudades de México, Guadalajara y Monterrey. Particularmente en el estado de Veracruz, la información acerca de esta situación es de menor cuantía o no existe por lo que se carece de información estatal acerca de la producción de autopsias en el sistema hospitalario de los Servicios de Salud de Veracruz.

Es por esto que se propone utilizar técnicas de minería de datos para determinar cuáles son las causas por las que los médicos no solicitan autopsias en el Hospital Regional de Río Blanco.

1.4 Objetivo general y específico

1.4.1 Objetivo general

Aplicar algoritmos de minería de datos para identificar las causas, motivos y circunstancias por las cuáles los médicos no solicitan autopsias en el Hospital Regional de Río Blanco (H.R.R.B).

1.4.2 Objetivos específicos

- Conocer la situación problemática a través de un experto en el tema para delimitar el alcance de la investigación.
- Conocer los factores coincidentes en la población diferencial del H.R.R.B., como parte de la opinión que se tiene acerca del problema para comprender cómo influyen estos en la realización de autopsias.
- Analizar los algoritmos de minería de datos aplicables a la información obtenida mediante las encuestas, para determinar cuáles permiten obtener mejores resultados.
- Seleccionar el marco de trabajo adecuado para los algoritmos funcionales con vistas a obtener una solución al problema planteado.
- Probar la funcionalidad de los algoritmos en el caso de estudio para determinar causas, motivos y circunstancias por las que no se efectúan autopsias.
- Clasificar las causas, motivos y circunstancias para que el experto en el tema tenga elementos que le permitan proponer posibles soluciones a la situación planteada.

1.5 Justificación

A pesar de la incuestionable importancia del estudio posmortem, se identificó una alarmante disminución en la realización de esta actividad a nivel mundial. En el H.R.R.B desde hace algún tiempo el personal médico dejó de proponer por alguna o algunas razones, hasta hoy desconocidas, la aplicación de esta práctica. Es necesario identificar las causas que llevaron a los especialistas del hospital a no proponer ni realizar autopsias.

Para determinar las causales que provocan que no se soliciten estos estudios, se aplicarán técnicas de minería de datos, como reglas de asociación y redes bayesianas, sobre los datos recogidos en el hospital, resultado de las encuestas realizadas al personal médico de este

centro de salud, ya que con estas técnicas se establecerán asociaciones, patrones y tendencias entre los factores influyentes en la disminución de esta práctica.

La información se procesará y preparará de acuerdo a lo que indique la metodología seleccionada para guiar el proceso de la minería de datos y concluirlo satisfactoriamente, con la obtención de un modelo de reglas que al evaluarlas resulten eficientes. A través de estas reglas se busca establecer las relaciones entre las variables que permitan identificar las razones del rechazo a la realización de autopsias.

El principal objetivo de esta investigación va dirigido a proporcionar los argumentos necesarios a las autoridades competentes, que les permitan tomar medidas y lanzar estrategias en función de retomar el uso de esta práctica en el H.R.R.B. De esta forma el hospital logrará aumentar su aporte a la medicina con los descubrimientos que se generarán en este campo, así como influir en otros centros hospitalarios para que aumenten sus números de autopsias y alcanzar mayor prestigio en la comunidad médica.

Capítulo 2. Estado de la práctica

Al inicio de la investigación, se revisó un gran número de artículos de los cuales se extrajo información elemental y se analizaron los trabajos más actuales y de vanguardia relacionados con el tema a desarrollar. El estudio de las investigaciones mostró el enfoque que hoy en día se persigue en la búsqueda de conocimientos dentro de bases de datos, dirigiendo el interés hacia el área médica. De alguna manera, cada trabajo consultado hizo su aporte al conocimiento que permitió llevar a cabo esta investigación. A continuación, se presenta una breve panorámica de los más significativos.

2.1 Trabajos relacionados

En esta sección se exponen las investigaciones más significativas de actualidad que se consultaron con el objetivo de comprender la problemática que aborda este trabajo, su importancia, necesidad y beneficios para la sociedad. También se presentan otros estudios que proponen esquemas de solución para el análisis de datos mediante tareas de MD que ilustraron a los autores sobre las herramientas, tareas y técnicas de MD más adecuadas para dar solución a la problemática planteada.

Los primeros trabajos que se estudiaron estaban enfocados en el problema de la disminución de autopsias. Fueron muchos los estudios que se encontraron relacionados con este tema lo que determinó que no era un problema trivial y que urge darle una rápida solución.

Suleiman [7] explicó los beneficios de las autopsias destacando su utilidad para determinar la causa de la muerte, la describió como un instrumento para dilucidar el espectro cambiante de las enfermedades; que permite la confirmación, la clasificación, y corrección de diagnósticos clínicos, así como la identificación de enfermedades nuevas y reemergentes, entre otras. Aunque la investigación se enfocó en un hospital de Nigeria, hace alusión a la disminución de las tasas de autopsias en el Reino Unido que descendió a menos del 10% en los hospitales universitarios y menos del 5% en otros lugares, y en los Estados Unidos y las Américas las tasas de autopsias oscilan entre el 20 y 45 porciento. En la investigación se expusieron varias razones que intervienen en la disminución de las autopsias, van desde las objeciones religiosas

y socioculturales a la versión pública del procedimiento, pero incluso dentro del hospital los médicos no están solicitando las autopsias y se especula que una de las causas sea la preocupación por las demandas por negligencia. La sugerencia que brinda el trabajo para fomentar el uso de la práctica médica es que sea obligatoria en los casos en los que no se obtiene un diagnóstico clínico antes de la muerte, o en los casos en que el deterioro repentino de la situación clínica no pueda explicarse completamente por el diagnóstico realizado. El trabajo cierra con las palabras de Giovanni Battista Morgagni, "Los médicos que han hecho o visto muchas autopsias han aprendido al menos a desconfiar de su diagnóstico; los que no se enfrentan a los hallazgos desalentadores de las autopsias, viven en las nubes con una vana ilusión "...

Se analizaron los datos del Registro de Cáncer de Zúrich en [8] desde el año 1980 hasta el 2010. Los registros contenían información de 102,434 pacientes registrados con cáncer y 89,933 fallecidos. Para la exploración de los datos los autores se basaron en la técnica estadística descriptiva. La investigación abordó la problemática de la disminución de autopsias y los resultados del caso de estudio mostraron la reducción de un 60% de las autopsias en 1980 a un 7% en 2010.

Fueron alarmantes los resultados que se obtuvieron del número de autopsias de algunos países. La investigación [9] se realizó en el año 2013 e identificó que la tasa de autopsias para Reino Unido era de un 0.69% en ese momento, 0.51% para Inglaterra, 2.13% en Escocia, 0.65% en Gales y 0.46% en Irlanda del Norte. Los datos demostraron que la práctica médica se encuentra al borde de la extinción en el Reino Unido.

Henshaw et al. [10] afirmaron que las técnicas modernas del siglo XXI influyeron en el decremento de las autopsias, sin embargo, no reducen el número de errores en los diagnósticos clínicos. Los resultados de la investigación demostraron un decremento en las autopsias de un 25% a un 0.5% entre los años 1983 y 2013 en el Reino Unido y de similar manera se perdió esta práctica en Europa y los EE.UU. Los investigadores identificaron algunas causas involucradas en la disminución de las autopsias como son: las pocas solicitudes de los médicos para su realización bajo el supuesto de que los familiares rara vez dan su consentimiento, excusa que cuestionan los autores porque sus análisis indicaron que en una ciudad como

Londres el 89% de las familias daría el consentimiento para el estudio si se ofrece de manera apropiada. Los investigadores proponen medidas como: sensibilizar a los médicos, a la sociedad, a los responsables de las políticas de atención de la salud y a los políticos acerca de los beneficios de la autopsia.

En [11], los investigadores analizaron las opiniones de los estudiantes de medicina respecto a las autopsias para justificar la inclusión continua de este método al plan de estudio. Para ello encuestaron a un total de 210 estudiantes. De los resultados obtenidos se constató que la mayoría de los estudiantes, 188 (89.52%), están de acuerdo respecto a la importancia de la autopsia en la educación médica y 158 (75.23%) sugirieron que los estudiantes de medicina deben observar y participar en más autopsias. Mientras que 193 (91.90%) estudiantes sintieron que la autopsia no se debe desechar del plan de estudio médico. A 50 (23.80%) de los encuestados no les importaría que se les realizara la autopsia tanto a sí mismos como a sus familiares. Sólo unos pocos estudiantes accedieron a especializarse en medicina forense y 19 sintieron que la medicina forense no es lucrativa. El trabajo concluyó que la autopsia es una importante herramienta de enseñanza que debe ser retomada y cuidadosamente ajustada en los programas de enseñanza médica.

A continuación, se hace referencia a estudios que consideraron en sus soluciones diferentes técnicas de MD.

Hoy en día es mucha la diversidad de opciones que proponen las aplicaciones en línea para cada perfil de usuario o cliente, provocando que el hecho de encontrar el producto, la opción o información deseada dentro del sistema, se convierta en una experiencia complicada y demorada a percepción del usuario. Es por ello que en [12] se presentó un estudio sobre minería de datos de uso web y del sistema de recomendación basado en el comportamiento del usuario actual a través de sus datos, con el fin de proporcionar información pertinente a la persona sin la necesidad de pedirla. Los autores entrenaron el método de clasificación *"K-Nearest-Neighbor (KNN)"* para utilizarlo en línea y tiempo real con el objetivo de identificar a tipos de clientes aprendiendo de su secuencia de clics. Esta técnica permitió agrupar usuarios con características similares y la recomendación de opciones de navegación adaptadas a sus necesidades y de igual manera a modo personal. Para llevar a cabo la solución propuesta los

datos se almacenaron en un *"data mart"* de manera limpia y agrupada en sesiones y se extrajeron de las *"Really Simple Syndication (RSS)"* de los usuarios en Internet. El resultado de esta investigación fue que el método clasificador KNN es consistente, sencillo, fácil de entender, con alta tendencia a poseer cualidades deseables y de fácil implementación, de lo que carecen otras técnicas de aprendizaje automático específicamente cuando hay poco o ningún conocimiento previo acerca de la distribución de datos.

En [13] se analizaron varios métodos de MD adaptables a varios dominios y aplicaciones, pero dándole una especial referencia a algunos de los métodos de detección de anomalías entre los supervisados, semi-supervisados y sin supervisión. La investigación se enfocó en el tema de las actividades anómalas, específicamente en las redes sociales por la tendencia creciente de las mismas en diferentes dominios. A la par del aumento de los beneficios que les brindan a los usuarios el uso abierto y libre de estas redes sociales, se detectó el uso indebido y fraudulento. Estas acciones inusuales suelen presentar diversos comportamientos por lo que en este estudio se analizaron los diferentes tipos de anomalías, su clasificación con base en características junto con los supuestos y las razones de su existencia, así como las técnicas para prevenir y detectarlas. La importancia de este estudio fue que generó una fuente de información elemental para llevar a cabo nuevas investigaciones dentro de cualquier área interesada en la utilización de técnicas para detección de atípicos como: los métodos basados en proximidad (KNN utilizando distancia o densidad), los basados en agrupamiento y los basados en clasificación (como el clasificador Bayesiano, las máquinas de soporte vectorial y las redes neuronales).

En los últimos tiempos se aplica con éxito la Minería de Datos en muchas áreas y dominios científicos como el médico y la bioinformática, solo por mencionar algunas, recientemente se introdujo en el campo de la construcción, debido a que esta actividad genera grandes cantidades de datos sobre el funcionamiento del sistema, el comportamiento de los ocupantes, el consumo de energía, las condiciones climáticas, entre otros. En [14] se describieron los avances que se llevaron a cabo a través de las tareas principales de MD, tanto predictivas como descriptivas, debido a su importancia y aporte. La investigación expuso los desafíos

encontrados y previstos, los posibles desarrollos futuros y recomendaciones para la aplicación de técnicas de minería de datos en el campo de la construcción.

Dentro del impulso por mejorar la eficiencia energética y las características de diseño ecológico sostenible en edificios nuevos y existentes, las tareas de MD dieron su aporte. Con el fin de proporcionar información útil para los diseñadores y planificadores de autoridad, dirigido a la identificación de las principales causas de los altos consumos de energía y los valores de referencia para impulsar un enfoque de diseño de sostenibilidad en un edificio, se genera una gran cantidad de datos durante las simulaciones de energía. Por esta razón, Capozzoli et al. [15] presentaron información útil que ayuda a reconocer los patrones que conducen a la evaluación de la eficiencia energética de edificios y describe en detalle la aplicación del algoritmo *"k-means",* que permitió dividir las muestras de consumo en grupos similares. El resultado de este trabajo fue el desarrollo de una herramienta que ayuda a los equipos de proyecto y las autoridades públicas a evaluar e identificar patrones útiles en grandes poblaciones en construcción. Para este propósito, utilizaron la combinación de diferentes técnicas de minería de datos del aprendizaje automático, reconocimiento de patrones y estadística para extraer automáticamente conceptos, relaciones y patrones de interés dentro de grandes conjuntos de datos.

Otra área en la que es crucial la MD es en el desarrollo de software. Uno de los principales beneficios aportados dentro de esta área fue el identificar los factores de éxito para las aplicaciones. Pero existe una necesidad de herramientas eficientes que utilizando datos de los repositorios de *"software"* disponibles predigan los módulos defectuosos de un nuevo proyecto con el fin de permitirle a los jefes de proyecto y equipos de calidad saber dónde invertir su tiempo en la prueba y la depuración. Por esta razón, el estudio realizado en [16] proporcionó una arquitectura de soluciones que mejora el desarrollo de aplicaciones basado en los datos de los repositorios de software y ofrece un punto de referencia que brinda un conjunto de modelos de MD referentes al problema de predicción de módulos defectuosos y la comparación de los resultados. Se demostró con este trabajo que los algoritmos de MD son eficaces cuando se utilizan para las relaciones entre los indicadores de calidad y los posibles módulos defectuosos, así como también que los mejores algoritmos de predicción fueron, en

primer lugar, el algoritmo *"Naive Bayes",* luego el de Red Neuronal y por último, los de Árboles de Decisión.

En [17] se evaluó la efectividad del aprendizaje del entorno colaborativo en línea en función de la información almacenada en los archivos de registro de los estudiantes, tema interesante, debido a la popularidad que alcanzó esta enseñanza, sustentada en los avances de la tecnología, que hace posible los intercambios a distancia. El estudio revela que un aprendiz en silencio es capaz de beneficiarse igualmente de este método educativo y los predictores generados mediante la utilización de la técnica de MD les permiten a los profesores comprender la influencia y logros de aprendizaje de los estudiantes que participan en este tipo de enseñanza. Esta investigación es una valiosa fuente de información para asesorar a los estudiantes en el aprendizaje colaborativo en línea de manera efectiva.

Siguiendo el tema de la educación, Harwati et al. [18] presentaron un estudio realizado en la Universidad Islámica de Indonesia para examinar el patrón de desempeño de los estudiantes mediante el uso de la técnica de agrupamiento *"K-means".* La información para esta investigación responde a los estudiantes del Departamento de Ingeniería Industrial desde la clase de 2009 a la 2011 y se conformó por el género, el origen nacional, el trabajo de los padres, promedio de calificaciones (por sus siglas en Inglés: GPA), optimización de valor y grado de Planificación y Control de la Producción. El resultado fue que cada grupo tiene sus respectivas categorías, que el grupo más grande se componía de los estudiantes más inteligentes y activos con un (45.74%), el segundo fue el grupo de alumnos con capacidad inferior a la media (33.33%) y el menor comprendía al restante 20.91% de los estudiantes. La técnica usada para las predicciones de clasificación es el principal aporte de esta investigación.

Hoy en día son muchos los fármacos que producen reacciones adversas y es un reto para los expertos la identificación eficiente de estas drogas. Los métodos de minería de datos de filtrado automático, se aplican en el cuidado de la seguridad de estos medicamentos. En la investigación [19], se propuso el uso de las reglas de asociación para buscar conjuntos de elementos frecuentes, relacionando medicamentos y reacciones adversas. Además, los investigadores establecieron una comparación mediante simulaciones, de las capacidades de las reglas de asociación con otros métodos, para detectar asociaciones y demostrar su

aplicación. Las reglas aplicadas detectaron asociaciones entre enfermedades y padecimientos consistentes de acuerdo a evidencia clínica y análisis de casos, lo que valida su uso para la detección de los efectos secundarios que pude causar un medicamento.

Para este siglo se espera un aumento del nivel del mar a consecuencia de una serie de factores como el desarrollo y los cambios ambientales, por citar algunos, por esta razón, Gutiérrez et al. [20] detallaron los datos que utilizaron para desarrollar y evaluar el rendimiento de una Red Bayesiana, que definió las relaciones entre las fuerzas motrices, las limitaciones geológicas y la respuesta costera en la costa atlántica de Estados Unidos. Una de las principales causas de erosión en esta zona que identificaron los investigadores fue las diferencias relativamente moderadas en las tasas de aumento del nivel del mar. Los resultados que se obtuvieron con la Red Bayesiana predijeron el cambio de este litoral a largo plazo y su aporte sirvió de apoyo a la administración costera de la zona, para trazar estrategias de respuesta ante diferentes escenarios futuros relacionados con cambios en los niveles del mar.

Se presupone que los resultados de la investigación descrita anteriormente no se aplican de forma genérica a otros sitios costeros. Por ello, es necesario explorar otros conjuntos de datos de este tipo con el fin de comprobar si estas relaciones entre la subida del nivel del mar y la erosión del litoral son fortuitas, se deben a las condiciones locales o si se aplican a muchas costas de todo el mundo. Otra de estas investigaciones que da su aporte al conocimiento es [21], la cual estudió el conjunto de datos costeros de la isla La Reunión a través del método *"Bayes Net (BN)".* El resultado fue una visión diferente de los procesos costeros respecto a los estudios anteriores. Este trabajo confirmó el considerable potencial de las redes bayesianas para explorar las bases de datos costeras, profundizar en estos procesos y seguir indagando en los factores que causan los cambios del litoral, incluyendo cambios del nivel del mar. Sin embargo, también identificaron varias dificultades para explorar datos de este tipo con BN. Entre las más relevantes señalaron que la calidad de la información en cuanto a niveles de precisión debe ser alta, si existen inconsistencias en el conjunto, hay que procesar los datos para eliminarlas, y como otras de las dificultades encontraron que a pesar de que BN permite operar con muy pocos datos, es importante contar con un número suficiente de muestras en la fase de aprendizaje.

La MD es la única tecnología que permite encontrar patrones, relaciones entre variables y comprender el contenido oculto de una base de datos. En [22], se propuso un conjunto de patrones extraídos del Registro Poblacional de Cáncer del Municipio de Pasto, en Colombia. Los datos que se utilizaron en el análisis fueron específicamente los que se registraron entre 1998 y 2007. El análisis de esta información dictaminó, mediante la tarea de Clasificación a través de la técnica de Árboles de Decisión que la supervivencia de las mujeres con este padecimiento es mayor a tres años, partiendo desde la fecha del diagnóstico de la enfermedad. Sin embargo, a través de la tarea de Asociación, los investigadores determinaron cuáles son los factores socioeconómicos y clínicos coligados a la supervivencia de las mujeres afectadas por esta dolencia. Los autores señalan que estos resultados apoyaron a entidades administrativas y privadas del sector de salud en la toma de decisiones, en el establecimiento de políticas, en la creación de programas de apoyo y protección dirigidos al sector afectado.

Por la motivación de apoyar el desarrollo en el sector de salud en las ciudades inteligentes, Oviedo et al. [23] describieron una serie de avances y tendencias de la minería de datos dentro de esta área. Los autores describieron algunas metodologías, herramientas y técnicas de la MD involucradas en aplicaciones médicas e identificaron a CRISP-DM como metodología más empleada, a Weka como herramienta más común para el análisis de datos de salud mediante técnicas de redes neuronales y árboles de decisión en análisis predictivo y para el descriptivo *"K-means".* El estudio también presentó como las principales tendencias y trabajos a futuro el análisis de datos no estructurados, así como las metodologías con etapas de pre-procesamiento e indexación para estos datos y la incorporación de herramientas con soporte a minería multimedia.

Durante el período en el que un paciente es atendido, se registra un gran número de información como: tratamiento médico, medicamentos y el perfil del enfermo, por mencionar algunos. Los datos registrados se convierten en una gran fuente de conocimiento dentro del marco de la medicina, que solo podrá descubrirse a través de la MD. El estudio en [24], demostró que, mediante la aplicación de reglas de asociación a un conjunto de datos de registros de pacientes, se obtuvieron correlaciones entre los diferentes atributos (exámenes, medicamentos, tratamientos y perfil del paciente). Los resultados también se analizaron de

acuerdo a diferentes niveles de abstracción, en la búsqueda de normas más específicas, por ello se realizó un *"drill-down"* a partir del subconjunto de normas de alto nivel. El resultado demostró eficacia en el descubrimiento de reglas interesantes de acuerdo a diferentes niveles.

La identificación de las cuestiones que ponen en peligro la vida de un paciente en unidades de cuidados intensivos (UCI) se hace mediante modelos de predicción de riesgo, los cuales están limitados a una cierta cantidad de condiciones y la mayoría de las escalas se enfocan hacia adultos. Sin embargo, las nuevas tecnologías en el campo médico, permiten registrar de cada paciente un amplio número de datos complejos, transitorios y en diferentes formatos, que no se analizan en su máxima expresión por los modelos convencionales, en función de extraer la mayor cantidad de conocimiento posible, oculto dentro de esa información. Es evidente entonces, que la necesidad de un modelo novedoso que permita adaptar nuevas características e incorporar modalidades temporales para una predicción de riesgo personalizado, se hace imperiosa. En [25], se presentó el nuevo sistema de predicción de riesgo ICU ARM-II (Minería de Reglas de Asociación para Unidades de Cuidados Intensivos). El conjunto de datos referentes a 4,975 pacientes se tomó de la Unidad de Cuidados Intensivos Pediátrica de Salud de Niños de Atlanta. La investigación identificó un grupo de reglas de asociación para predecir riesgos con base en todas las condiciones disponibles y un indicador que evalúa la fiabilidad de estas reglas.

Un estudio realizado sobre el apoyo que brinda la MD enfocada a la inteligencia de negocios en el área de salud se describe en [26]. La intención de los autores fue determinar los factores de riesgo coligados a la Diabetes Mellitus Tipo 2 (DM2) aplicando reglas de asociación. Se ocuparon los registros de datos de los pacientes con DM2 atendidos por uno de los especialistas del Centro de Atención Médica en Colombia. Los resultados que se obtuvieron mediante las reglas de asociación fueron que se encontraron nuevos factores de riesgo que permiten identificar grupos propensos a contraer la enfermedad, facilitando así la atención a los mismos de forma preventiva. El aporte de este trabajo fue ayudar en la labor de los profesionales del sector médico en el manejo de la enfermedad de Diabetes Mellitus.

En muchos países de escasa atención médica la mayoría de las muertes ocurren en los hogares. A diferencia de las muertes que se producen en los hospitales, éstas no cuentan con un

estándar para su validación según se indica en [27]. Es por ello que estudios anteriores evidencian contradicciones en los resultados de los métodos automatizados para la clasificación de causas de muertes (COD). En esta investigación se realizó una comparación de los clasificadores Naive Bayes (NB), Tarificación de código abierto (OTM) e InterVA-4 en tres conjuntos de datos que comprenden alrededor de 21,000 registros de muertes de niños y adultos. El resultado del clasificador NB superó a los demás clasificadores, aunque se evidencia que ninguno de los clasificadores automatizados actuales es capaz de realizar adecuadamente la clasificación de COD individuales.

Sumalatha y Muniraj [28] plantearon que los expertos encuentran dificultades para determinar el grado de una enfermedad cuando carecen de pruebas suficientes para el diagnóstico médico y también en el caso contrario, cuando disponen de demasiadas pruebas. Por ello, los autores analizaron importantes investigaciones que tratan sobre la aplicación de algoritmos de aprendizaje automático para las tareas de minería de datos encaminadas a apoyar el diagnóstico de enfermedades del corazón, cáncer de mama y diabetes. El análisis de los autores estuvo dirigido a identificar los algoritmos de minería de datos que es posible utilizar en el campo de la predicción médica de manera eficiente. En ese sentido, reafirmó la importancia de diagnosticar estas enfermedades en sus primeras etapas y consagró la necesidad de un nuevo enfoque para reducir la tasa de falsas alarmas y aumentar la de detección de la enfermedad.

En [29], con el objetivo de identificar enfoques de clasificación y agrupamiento útiles para el desarrollo de sistemas de predicción, los autores hacen una revisión de fuentes que describen la aplicación de las diferentes técnicas de minería de datos en el campo médico. Asimismo, discuten sobre las herramientas disponibles para el procesamiento y clasificación de los datos y explican que, para el reconocimiento de patrones, la elección de las tareas de minería depende de las características de los datos. Por ello indican el uso de técnicas de agrupamiento cuando los datos no estén etiquetados y la clasificación para el caso contrario. El estudio resalta la importancia que tiene la exactitud en el diagnóstico de enfermedades que ponen en riesgo la vida, como el cáncer y las enfermedades cardíacas, y señala que es un factor que

requiere de un enfoque novedoso, que permita disminuir las falsas alarmas y mejorar el diagnóstico en las primeras etapas de la enfermedad.

Las técnicas de minería de datos se utilizan para mejorar la toma de decisiones en áreas como la gestión hospitalaria. También son muy útiles para sustituir el análisis manual de los datos del seguro médico. Con el incremento de personas que adscritas a algún plan, esa tarea, acudiendo solo al conocimiento limitado profesional, se torna cada vez más difícil e imposible de realizar de manera eficiente. En [30] se propuso una clasificación basada en tres criterios (precisión, estabilidad y complejidad), que permitiría analizar más eficazmente el volumen de datos en comparación con un análisis manual. El conjunto de datos utilizado para probar la eficacia de este enfoque comprende decenas de miles de pacientes de una ciudad y cientos de miles de registros de reembolsos médicos durante el lustro 2010-2015. Los resultados del experimento realizado a partir de los datos médicos analizados por el algoritmo FPGrowth demuestran que el enfoque propuesto mejora el modelo de decisión, por lo que la toma de decisiones gana en flexibilidad y eficiencia y supera a los otros esquemas en términos de la precisión para la clasificación.

Song et al. [31] realizaron un estudio, basado en algoritmos de aprendizaje automático para minería de datos, sobre las características de las enfermedades provocadas por el mosquito, como el dengue-1, dengue-4, fiebre amarilla, infección por el virus del Nilo occidental y filariasis. A pesar de que para algunas de las enfermedades mencionadas existe cura, los autores suponen que como estas afectan mayormente a zonas de mucha pobreza como son el continente africano y Asia Occidental, para la mayoría de las personas el tratamiento indicado se hace inaccesible. Por ello su objetivo es encontrar características semejantes en las secuencias de aminoácidos, que permitan crear una cura capaz de sanar al paciente de una sola vez. Los resultados del estudio mostraron que, aunque parezca que hay rasgos similares entre el virus del Dengue, el virus de la fiebre amarilla, WestNile y Brugia Malayimitochondrion, las diferencias entre estos son más fuertes que sus semejanzas. Por otro lado, descubrieron que el control de la Leucina contribuye en el desarrollo de una cura única eficaz para los casos de WestNile y Brugia Malayimitochondrion.

A partir de la alta mortalidad del cáncer, Cho et al. [32] investigaron las secuencias de diversas citoquinas mediante algoritmos como Apriori, árbol de decisión, y máquinas de soporte vectorial (SVM). Las citoquinas desempeñan un papel central en el sistema inmunológico, por lo que el estudio, de cumplirse su objetivo, contribuirá con otros en el hallazgo de nuevas reglas que permitan determinar si una citoquina tiene propiedades anti cancerígenas o no.

En [33], se compararon diferentes esquemas para la identificación de etiquetas de clases para un conjunto de datos determinado y demuestran como su propuesta IGFSS (Improved Global Feature Selection Scheme) es más eficiente que las clásicas. También se describió el uso de algoritmos comúnmente usados para la clasificación de textos como Naive Bayes(NB) y Support vector machine (SVM), con el objetivo de demostrar la eficacia de su propuesta.

Para predecir las causas de muertes relacionadas con el estándar de clasificación de enfermedades de la Organización Mundial de la Salud, Mujtaba et al. [34] aplicaron técnicas de aprendizaje automático en informes forenses de texto. A su vez, realizaron una comparativa de enfoques de extracción de características, enfoques de representación de valores de características y clasificadores de texto como SVM, Random Forest (RF) y NB para la clasificación de informes de autopsia forenses. El conjunto de datos utilizado fue el resultado de 400 reportes de autopsia forense de un hospital en Kuala Lumpur, Malasia, que comprendía ocho de las causas de muertes más comunes. Los resultados de los modelos de decisión para SVM superaron a los de RF y NB.

En [35], propusieron un sistema de clasificación automática (multi-clase) para predecir las causas de muerte a partir de modelos de decisión de clasificación automática de textos. Los datos analizados fueron 2200 registros de autopsias por accidentes del hospital Kuala Lumpur. Los investigadores evaluaron los algoritmos SVM, NB, k-nearest neighbor (KNN), decision tree (DT) y Random forest (RF) de acuerdo a las métricas *precision*, *recall*, *F-measure* y Área ROC (*Receiver Operating Characteristics*), desde la herramienta para minería de datos Weka. Random forest y J48 resultaron ser los modelos de decisión mejor evaluados.

El desarrollo de esquemas eficientes y robustos para la clasificación de texto es de gran importancia para la inteligencia empresarial y otras áreas. Por ello, en [36] realizaron un

análisis empírico sobre métodos estadísticos para la extracción de palabras claves utilizando las colecciones de documentos ACM y Reuters-21578. También describen el comportamiento predictivo de los algoritmos de clasificación y métodos de aprendizaje conjunto cuando se utilizan de las palabras claves para representar documentos de texto científicos, demostrando así que a medida que aumenta el número de palabras clave, el desempeño predictivo de los clasificadores tiende a aumentar.

La búsqueda de patrones frecuentes constituye hoy día una actividad costosa debido al gran volumen de las bases de datos que requieren ser exploradas. La generación de conjuntos candidatos es una tarea compleja y requiere de mayor espacio en memoria debido a la cantidad de iteraciones que se realizan para establecer las comparaciones necesarias que permiten determinar si un conjunto de elementos candidato es frecuente o no. En [37], discutieron las diferentes perspectivas para optimizar la poda de los subconjuntos candidatos que no son frecuentes en los algoritmos de reglas de asociación como *Apriori*, *FP-tree* y *Fuzzy FP-tree*.

Manimaran y Velmurugan [38] analizaron el algoritmo *Apriori*, ya que es uno de los algoritmos más ampliamente utilizado en diversos dominios, entre los que se destaca por su gran uso el área de asistencia médica. Sin embargo, el objetivo principal fue integrar el algoritmo *Apriori* en la minería de texto. Los autores demostraron que es una forma eficiente para encontrar patrones interesantes que son fácilmente interpretados por técnicas de visualización. Por tanto, es posible analizar la información desconocida disponible en datos de texto utilizando el algoritmo *Apriori*.

2.2 Análisis comparativo

Como se puede notar, en la descripción de los estudios que formulan sus soluciones con técnicas de MD, han sido muy diversos los esquemas propuestos debido a las peculiaridades y necesidades de los problemas plantados, así como de las diferentes características de los datos a tratar en cada caso. Para comprender la intención de cada esquema se muestra en la Tabla 2.1 el análisis comparativo de estos artículos.

Tabla 2.1 Análisis comparativo de los artículos relacionados

Artículo	Problema	Objetivo	Solución	Área	Resultados
Adeniyi et al. [12]	A la par de los beneficios que se derivan del aumento de las aplicaciones en línea y la disponibilidad de la información en Internet, surge como efecto colateral la dificultad para localizar la información requerida por el usuario.	Demostrar que mediante la técnica KNN es posible generar modelos eficientes para sistemas de recomendación en tiempo real.	**Tareas MD** Clasificación **Técnica MD** KNN (Distancia Euclidiana) **Algoritmo MD** KNN **Software** MATLAB	Internet: Aplicaciones en línea	Prototipo de sistema de recomendación en tiempo real basado en la técnica KNN, capaz de producir clasificaciones y recomendaciones útiles y precisas al cliente.
Ravneet y Sarbjeet [13]	Uso indebido y fraudulento en las redes sociales.	Identificar las técnicas de la MD que permiten la detección de atípicos.	Técnicas para detección de anomalías como: los métodos basados en proximidad (KNN utilizando distancia o densidad), los basados en agrupamiento y clasificación (como el clasificador Bayesiano, SVM y las redes neuronales).	Redes Sociales	El estudio generó una valiosa fuente de información que realza el potencial de las técnicas de la minería de datos en busca de fraude.

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
Yu et al. [14]	El campo de la construcción genera mucha información que necesita analizarse para mejorar la calidad de sus resultados.	Destacar los avances recientes de las tareas predictivas y descriptivas en el campo de la construcción.	**Tareas MD** Clasificación Agrupamiento Asociación **Técnicas MD** SVM. Red neuronal. Árboles de decisión Reglas de asociación.	Construcción	La investigación da una panorámica general de los trabajos recientes que estudian el aporte de las tareas predictivas y descriptivas en el área de la construcción en función de mejorar su rendimiento. Sin embargo, destaca que hay que garantizar calidad en los datos y seguir las tendencias de la MD, como los métodos de minería escalables, basados en restricciones y enfoque de gestión "Big Data".
Capozzoli et al. [15]	Se necesita extraer conocimiento de la información almacenada durante las simulaciones de energía, que permita identificar las causas de	Obtener patrones para evaluar la eficiencia energética de edificios mediante técnicas de clasificación.	**Tareas MD** Clasificación **Técnica MD** Árboles de decisión	Energía Eléctrica	Desarrollo de una herramienta que ayuda a los equipos de proyecto y las autoridades públicas a evaluar e identificar patrones útiles en grandes

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	alto consumo energético y los valores de referencia para impulsar un enfoque de diseño sostenible para nuevos edificios.		**Algoritmo MD** CART (Classification and Regression Tree) **Software** RapidMiner		poblaciones en construcción.
Yousef y Ahmed [16]	Inexistencia de herramientas eficientes que mediante datos de repositorios de software predigan los módulos defectuosos en un proyecto.	Demostrar cuáles son los atributos que predicen el estado defectuoso de módulos de software mediante el uso de algoritmos de MD.	**Tareas MD** Clasificación **Técnica MD** "Naive Bayes", Red Neuronal y Árboles de Decisión	Software	Propuso una arquitectura de soluciones para mejorar el desarrollo de aplicaciones.
Shukor et al. [17]	De acuerdo al avance tecnológico surgió una nueva modalidad de enseñanza, el aprendizaje del entorno colaborativo en línea. Cada día aumenta el número de usuarios que forman parte de ello. Se	Identificar mediante técnicas de MD los atributos que predicen los resultados del estudiante mediante el aprendizaje en línea.	**Tareas MD** Clasificación **Técnica MD** Árboles de decisión **Algoritmo MD**	Educativa: Aprendizaje en línea.	Los resultados dictaminaron que los estudiantes sí mejoraron su aprendizaje mediante la enseñanza colaborativa en línea.

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	hace imprescindible evaluar la efectividad del aprendizaje de los estudiantes en esta modalidad de estudio.		C4.5 **Software** Weka		
Harwati y Wula, [18]	Se necesitaba determinar el desempeño de los estudiantes en la Universidad Islámica de Indonesia.	Aplicar la MD mediante técnicas de agrupamiento "*K-means*" para determinar el desempeño de los estudiantes en el Departamento de Ingeniería Industrial en la Universidad Islámica de Indonesia.	**Tareas MD** Agrupamiento **Técnica MD** Agrupamiento basado en partición. **Algoritmo MD** *"K-means"*	Educativa	Se determinaron tres grupos de características diferentes: los estudiantes más inteligentes y activos (45.74%), los de capacidad sobre la media (3.33%) y los de bajo rendimiento (20.91%).
Chao Wang et al. [19]	Se necesita identificar eficientemente los fármacos que producen reacciones adversas y las reacciones asociadas a ellos. Debido a que existe un gran número de fármacos se hace una tarea	Obtener un modelo mediante las reglas de asociación que permita encontrar relaciones de medicamentos con reacciones adversas y conjuntos de elementos frecuentes que se	**Tareas MD** Asociación **Técnica MD** Reglas de asociación. **Software**	Farmacología	Las reglas aplicadas detectaron asociaciones entre enfermedades y padecimientos consistentes de acuerdo a evidencia clínica y análisis de casos, lo que valida su uso para la detección de

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	muy complicada para los expertos.	produzcan dadas estas relaciones.	*SAS Analytics*		los efectos secundarios que causa un medicamento.
Gutierrez et al. [20]	Las zonas costeras presentaron cambios en los niveles de erosión, se necesita determinar qué factores influyen en este proceso.	Evaluar el rendimiento de una Red Bayesiana con los datos registrados de la costa atlántica de Estados Unidos.	**Tareas MD** Clasificación **Técnica MD** Métodos Bayesianos **Algoritmo MD** Red Bayesiana **Software** Netica y MATLAB	Ecología	Identificó como uno de los factores que afectan a la erosión costera los cambios moderados en las tasas de aumento del nivel del mar, y predijo el cambio de este litoral a largo plazo, lo que permite que las entidades competentes logren trazar estrategias de respuesta.
Bulteau et al. [21]	Otro caso de estudio del tema tratado en [20].	Evaluar el rendimiento de una Red Bayesiana con datos costeros de la Isla La Reunión.	**Tareas MD** Clasificación **Técnica MD** Métodos Bayesianos **Algoritmo MD** Red Bayesiana	Ecología	Los resultados difieren del caso anterior. Este trabajo confirmó el potencial de las BN para explorar datos costeros, aunque identificó algunas dificultades. Es necesario que los datos tengan altos niveles de

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
					precisión, sean consistentes y que exista un número suficiente de muestras para la fase de aprendizaje.
Timarán y Yépez, [22].	Se necesita extraer la información inmersa en datos del Registro Poblacional de Cáncer del Municipio de Pasto, en Colombia, relacionada con la supervivencia al cáncer uterino.	Obtener modelos y reglas eficientes mediante la MD que permitan determinar los factores que influyen en la supervivencia de las mujeres que padecen de cáncer uterino.	**Tareas MD** Clasificación y Asociación **Técnica MD** Árboles de decisión y Minería de Reglas de Asociación **Algoritmo MD** J48 Apriori **Software** Weka	Salud	Mediante la técnica de Árboles de Decisión se dictaminó que la supervivencia a este padecimiento es mayor a tres años, desde el momento de su diagnóstico. A través de las tareas de Asociación, los investigadores determinaron los factores socioeconómicos y clínicos coligados a la supervivencia en los pacientes.
	Uno de los principales	Identificar los avances	**Tareas MD**	Salud	Presentó los aportes de la

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
Oviedo et al. [5]	propósitos de las ciudades inteligentes hoy en día es mejorar la calidad de vida de los ciudadanos, especialmente el servicio de salud. La minería de datos es capaz de hacer su aporte.	alcanzados y las tendencias a futuro de la minería de datos que serán explotadas en el área de salud de las ciudades inteligentes.	Clasificación y Agrupamiento **Técnica MD** Árboles de Decisión, Redes Neuronales, "*K-means*" **Software** Weka		MD en el campo de la salud, así como su tendencia al futuro. Demostró que dentro del área médica para análisis predictivos, se usan más las técnicas de redes neuronales y árboles de decisión, para el descriptivo *"K-means",* como metodología y herramientas para la minería CRISP-DM y Weka, respectivamente.
Antonelli et al. [24]	Durante el período en el que un paciente es atendido, se registra información como: tratamiento médico, medicamentos y el perfil del enfermo, por mencionar algunos. Se analizaron los datos de los	Obtener conocimiento útil del conjunto de datos de pacientes diabéticos de un NHC de Italia, mediante la aplicación de reglas de asociación.	**Tareas MD** Asociación **Técnica MD** Reglas de asociación. **Software** MeTA (Medical Treatment Analysis)	Salud	El resultado demostró la eficacia en el descubrimiento de reglas interesantes de acuerdo a diferentes niveles de abstracción. Los resultados se validaron por expertos del dominio clínico.

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	pacientes diabéticos del Centro de Salud Nacional (NHC) de una provincia italiana para encontrar nuevo conocimiento que aporte al desarrollo de la medicina.		Framework enfocado en el análisis de datos médicos que se centra en la caracterización de tratamientos en los diferentes niveles de granularidad.		
Chih-Wen et al. [25]	Los modelos de predicción de riesgo, que identifican las causas que ponen en peligro la vida de un paciente en unidades de cuidados intensivos (UCI), están limitados a una cierta cantidad de condiciones y la mayoría de las escalas se enfocan hacia adultos.	Obtener un modelo que se adapte a nuevas características e incorpore modalidades temporales para una predicción de riesgo personalizado.	**Tareas MD** Clasificación y Asociación **Algoritmo** CBA (Clasificación basada en Asociación) **Software** MATLAB	Salud	Los resultados del trabajo fueron: un nuevo grupo de reglas de asociación para predecir riesgos con base en todas las condiciones disponibles, un indicador que evalúa la fiabilidad de estas reglas y el sistema de predicción de riesgo, ICU ARM-II.
Franco y León [26]	Se necesitaba extraer el conocimiento de los datos de pacientes con DM2 registrados por un especialista del Centro de	Aplicar reglas de asociación para determinar los factores de riesgo coligados a la DM2.	**Tareas MD** Asociación **Algoritmo**	Salud	Se encontraron factores de riesgo que permiten identificar grupos propensos a contraer la enfermedad, facilitando

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	Atención Médica en Colombia.		A priori y FP-Growth		así la atención a los mismos de forma preventiva.
Miasnikof et al. [27]	En países con escasa atención médica muchas de las muertes ocurren en los hogares sin atención médica. Se necesita de una herramienta que permita la clasificación eficiente para las autopsias verbales.	Los resultados de los clasificadores automatizados actuales generan muchas contradicciones en la clasificación de causas de muertes individuales respecto a la clasificación dada por los médicos, por lo que se necesita identificar un clasificador eficiente para clasificar autopsias verbales de acuerdo a la clasificación de los médicos.	**Tareas MD** Clasificación **Técnica MD** Clasificación **Algoritmo MD** Naive Bayes Tarificación de código abierto (OTM) InterVA-4	Medicina	Se realizó una comparación entre diferentes clasificadores donde resultó más eficiente Naive Bayes para la clasificación de causas de muertes. Sin embargo, consideran que se necesitan realizar más investigaciones porque los clasificadores actuales tienen muchas contradicciones de acuerdo a las clasificaciones de los médicos.
Sumalatha y Muniraj [28]	Los expertos encuentran dificultades para determinar el grado de	Identificar los algoritmos de minería de datos que se logran utilizar en el	**Tareas MD** Clasificación	Medicina	El estudio revela la importancia de diagnosticar estas

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	una enfermedad cuando carecen de pruebas suficientes para el diagnóstico médico y en el caso contrario, demasiadas pruebas podrían complicarlo.	campo de la predicción médica de manera eficiente.	Agrupamiento **Algoritmo MD** SVM Decision Trees Naive Bayes Multilayer Perceptron Self Organizing Map (SOM) **Software** Weka Matlab		enfermedades en sus primeras etapas y la necesidad de proponer un nuevo enfoque para reducir la tasa de falsas alarmas y aumentar la tasa de detección de la enfermedad.
Sharma et. al [29]	Se hace necesario identificar patrones ocultos en los datos médicos para diagnosticar enfermedades que ponen en riesgo la vida.	Proponer un enfoque adecuado a partir de la revisión de literatura relacionada con la búsqueda de conocimiento mediante minería de datos para	**Tareas MD** Clasificación Agrupamiento **Técnica** Árboles de Decisión	Medicina	El estudio resalta la importancia que tiene la exactitud en el diagnóstico de enfermedades que ponen en riesgo la vida como el cáncer y las

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
		enfermedades cardíacas y cáncer.	Redes Neuronales **Algoritmo** ID3 C4.5 SVM CART Naïve Bayes **Sotware** RapidMiner Weka R-Programming Orange KNIME NLTK		enfermedades cardíacas, y señala que es un factor que requiere de un enfoque novedoso que permita disminuir las falsas alarmas y mejorar el diagnóstico de las mismas en las primeras etapas de la enfermedad.
Duan et al. [30]	El análisis manual de los datos del seguro médico, dado el	Proponer un enfoque de clasificación para la minería de datos que	**Tareas MD** Asociación	Medicina: Seguros médicos	El modelo de decisión propuesto mejora la toma de decisiones con

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	incremento de las personas que lo contratan, se torna en una tarea difícil e imposible de realizar de manera eficiente acudiendo solo al conocimiento limitado profesional.	mejore la eficacia de la toma de decisiones en los análisis de seguros médicos de los hospitales.	Clasificación **Técnicas** Arboles de decisión Reglas de asociación **Algoritmo MD** FPGrowth		mayor flexibilidad y eficiencia y supera a los otros esquemas en términos de la precisión para la clasificación.
Song et al. [31]	Enfermedades como el dengue-1, el dengue-4, la fiebre amarilla, la infección por el virus del nilo occidental y la filariasis, afectan mayormente a zonas de mucha pobreza como son el continente africano y Asia Occidental, donde las personas no son capaces	Buscar similitudes entre las secuencias de aminoácidos que permita producir una cura para estas enfermedades de una sola vez.	**Tareas MD** Clasificación Asociación **Algoritmo MD** Árboles de Decisión Apriori	Medicina	El experimento descrito en este trabajo mostró que son más fuertes las diferencias entre las características de estas enfermedades que sus semejanzas. Por otro lado, descubrieron que el control de la Leucina contribuye en el desarrollo de una cura única eficaz para los

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	de costear las curas.				casos de WestNile y Brugia Malayimitochondrion.
Cho et al. [32]	En la actualidad la tasa de mortalidad que está provocando el cáncer es realmente alta. Por lo tanto, se hace necesario conocer las características de las citoquinas para combatir esta enfermedad	Analizar las secuencias de diversas citoquinas mediante algoritmos de minería de datos para determinar si tienen propiedades anti cancerígenas o no.	**Tareas MD** Asociación Clasificación **Técnicas** Reglas de asociación Árboles de decisión **Algoritmo MD** Apriori SVM	Medicina	Se obtuvieron las reglas de las secuencias de las citoquinas anti cancerígenas, por tanto es posible usar las mismas para determinar si una nueva citoquina tendrá propiedades anti cancerígenas.
Mujtaba et al. [34]	En los estudios de autopsia forense los expertos generan informes que demoran de 30 a 45 días	Aplicar técnicas de aprendizaje automático en informes forenses de	**Tareas MD** Clasificación	Médico- Forense	Se identificaron las características más adecuadas en la comparación de los

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
	debido a que es una tarea muy laboriosa por lo que consume mucho tiempo.	texto para predecir las causas de muertes y comparar enfoques para determinar el más adecuado para la clasificación de informes de autopsia forenses.	**Técnica MD** Clasificación de texto **Algoritmo MD** Naive Bayes SVM Random Forest		diferentes enfoques para la clasificación de texto y el mejor modelo de decisión resultó ser el SVM con una precisión de 78.25%.
Mujtaba et al. [35]	La elaboración de un informe de autopsia es una tarea que toma mucho tiempo debido a que es un examen muy minucioso.	Desarrollar un sistema de clasificación para determinar las causas de muerte a partir de los reportes de autopsias por accidentes.	**Tareas MD** Clasificación Agrupamiento **Técnica MD** Clasificación de texto Árboles de decisión **Algoritmo MD** J48 Random Forest K-nearest neighbor	Medicina: Patología	Sistema factible y práctico que apoya a los patólogos para determinar con precisión y rapidez la causa de muerte a partir de informes de autopsias.

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
			Decision tree Naive Bayes SVM **Software** Weka		
Onan et al. [36]	La clasificación de texto es parte complementaria para muchos de los procesos automatizados, es por ellos que se hace necesario identificar un esquema eficiente para la extracción de palabras clave.	Realizar un análisis empírico sobre métodos estadísticos para la extracción de palabras clave y analizar su aporte en el comportamiento predictivo de los algoritmos de clasificación y métodos de aprendizaje conjunto cuando se utilizan las palabras clave para representar documentos de texto científico.	**Tareas MD** Clasificación **Algoritmo MD** *Naïve Bayes* SVM *Logistic Regression* *Random Forest*	Investigación científica	Presentan un esquema eficiente y robusto para la clasificación de texto y demuestran que a medida que aumenta el número de palabras clave, el desempeño predictivo de los clasificadores tiende a aumentar.

Tabla 2.1 Análisis comparativo de los artículos relacionados cont.

Artículo	Problema	Objetivo	Solución	Área	Resultados
Solanki y Patel [37]	Se necesita mejorar el rendimiento de los algoritmos en la búsqueda de conjuntos de elementos frecuentes.	Optimizar la poda de candidatos de elementos no frecuentes en los algoritmos de reglas de asociación.	**Algoritmo** Apriori, FP-tree y Fuzzy FP-tree.	Minería de Datos	Los autores exponen sus perspectivas para optimizar la poda de los subconjuntos candidatos que no son frecuentes en los algoritmos de reglas de asociación como Apriori, *FP-tree* y *Fuzzy FP-tree*.
Manimaran y Velmurugan [38]	Se busca realizar minería de texto de manera eficiente a través del algoritmo *Apriori*.	Integrar el algoritmo *Apriori* en la minería de texto.	**Técnica** Reglas de asociación **Algoritmo** *Apriori* **Software** Matlab	Minería de Datos y Minería de Texto	Los autores demostraron que de forma eficiente es posible utilizar el algoritmo *Apriori* para encontrar patrones interesantes en datos de texto.

Como se observa en la Tabla 2.1, los diferentes estudios analizados demuestran la utilidad de las técnicas de minería de datos para la solución de problemas en varias áreas, principalmente en el área médica. Sin embargo, después de revisar varios trabajos relacionados, no se encontró alguno donde se utilicen reglas de asociación y redes Bayesianas para analizar el decremento en el número de autopsias realizadas en un hospital, por lo que esto determina la importancia de esta investigación.

2.3 Propuesta de solución

Con el objetivo de establecer el marco de trabajo más acorde a la propuesta de solución de la investigación, se hizo un análisis bajo criterios de funcionalidad y pertinencia de las principales herramientas y metodologías existentes. Se investigaron plataformas y métodos para las tareas de MD, metodologías para el desarrollo de software, se identificó el lenguaje de programación más idóneo para el desarrollo de la aplicación, así como IDEs (*Integrated Development Environments,* Entornos de Desarrollo Integrado) y sistemas gestores de bases de datos.

La selección de dichas herramientas se fundamentó en la coherencia de éstas con el proceso, es decir, que fueran válidas para los objetivos de la investigación y beneficiosas para todo el proceso. La solución propuesta considera como herramienta para aplicar los algoritmos de MD a Weka, a UWE como metodología para el desarrollo de la aplicación, al lenguaje Java mediante el IDE NetBeans para el desarrollo de la aplicación y al SGBD PostgreSQL para la implementación del repositorio de la herramienta, tal como se ilustra en la Tabla 2.2.

Tabla 2.2 Alternativa de solución

Aspectos	**Propuesta**
Herramienta MD	Weka
Metodología MD	KDD
Lenguaje de programación	Java
IDE	NetBeans
SGBD	PostgreSQL

Tabla 2.2 Alternativa de solución cont.

Aspectos	**Propuesta**
Metodología de desarrollo	UWE
Frameworks	JavaServer Faces PrimeFaces

Respecto a las herramientas para aplicar los algoritmos de minería de datos se identificaron como las más prometedoras a Weka y RapidMiner, ambas son gratuitas y desarrolladas bajo plataforma Java, pero la seleccionada fue Weka porque se destaca por tener más amplio el abanico de algoritmos, como se muestra en la Tabla.2.3

Tabla 2.3 Comparativa de las herramientas de MD

Características	**Weka**	**Knime**	**RapidMiner**
Algoritmos nativos	168	9	34
Algoritmos importados desde Weka	0	102	101
Algoritmos implementados	168	111	135
Algoritmos nativos (%)	100%	8.1%	25.2%

Se seleccionó Java debido a que es el lenguaje nativo de Weka y facilitará la interpretación del modelo generado por esta herramienta y la aplicación. Además, Java es un lenguaje de propósito general, orientado a objetos, concurrente y hasta hoy, uno de los más populares.

En cuanto a los SGBD se selecciona PostgreSQL sobre Oracle, atendiendo a que es libre de paga, estas herramientas son muy similares y brindan prácticamente las mismas características y aunque Oracle cuenta con una versión gratuita, dicha versión tiene algunas limitaciones en cuanto a capacidad de información y características avanzadas de seguridad, entre otras.

Atendiendo a las características de los IDEs Eclipse y NetBeans se selecciona el segundo, ya que permite de manera predefinida tener el ambiente de trabajo con todas las características necesarias para el desarrollo de la aplicación. En cambio, Eclipse requiere de una preparación previa del marco de trabajo basado en la búsqueda e instalación de *plugins*.

Para desarrollar la minería de datos se seleccionó el método KDD por sus resultados probados para descubrir el conocimiento útil y novedoso implícito en los datos, como se evidencia en las referencias consultadas.

Como metodología para el desarrollo de la aplicación se seleccionó UWE. La decisión se fundamentó en que UWE se enfoca en el desarrollo de aplicaciones web y propone las herramientas necesarias para guiar y documentar adecuadamente la construcción de dichas aplicaciones.

JavaServer Faces (JSF) [39] es un *framework* que ofrece una estructura de componentes del lado del servidor para el desarrollo de aplicaciones web mediante lenguaje Java. Es una poderosa API que brinda componentes, gestión de estados y eventos, validaciones del lado del servidor y conversión de datos, definiciones de navegación en las páginas, internacionalización y accesibilidad. Además, proporciona biblioteca de etiquetas para añadir componentes a las páginas y las conexiones de dichos componentes con objetos que se encuentran en el lado del servidor. Se fundamenta en un modelo de programación bien estructurado en cuatro capas, lo cual hace en combinación a lo anteriormente expuesto que el desarrollo y el mantenimiento de las aplicaciones web se realice de manera eficiente, fácil y rápida.

PrimeFaces [40] es un *framework* de componentes visuales de código abierto para mejorar la experiencia y posibilidades de los componentes de JSF. Es una biblioteca que facilita mucho el trabajo a los desarrolladores de aplicaciones enriquecidas de la web. Actualmente, el *framework* está ampliamente difundido, tiene una amplia gama de interesantes componentes y temas de apariencia, tiene soporte para Ajax, es muy ligero, no requiere de configuraciones y cuenta con una poderosa documentación.

Capítulo 3. Aplicación de la metodología

De acuerdo a la problemática que se estudia en esta investigación y al objetivo trazado para su solución, aspectos descritos en el apartado 1.3 del capítulo uno, se necesita procesar la información obtenida a partir de la encuesta aplicada a los médicos del H.R.R.B y analizarla mediante técnicas de MD.

Por ello, en este estudio se propone el desarrollo de una aplicación web capaz de registrar la opinión médica a través de una encuesta en línea y representar los resultados generados por los algoritmos de aprendizaje automático de la minería de datos.

En la Figura 3.1 se presenta el esquema de solución que indica que es posible acceder a la aplicación desde un dispositivo cliente para insertar y consultar encuestas. Además, la aplicación permite generar modelos de minería de datos haciendo uso de la API de Weka y guardarlos físicamente en el servidor que la hospeda.

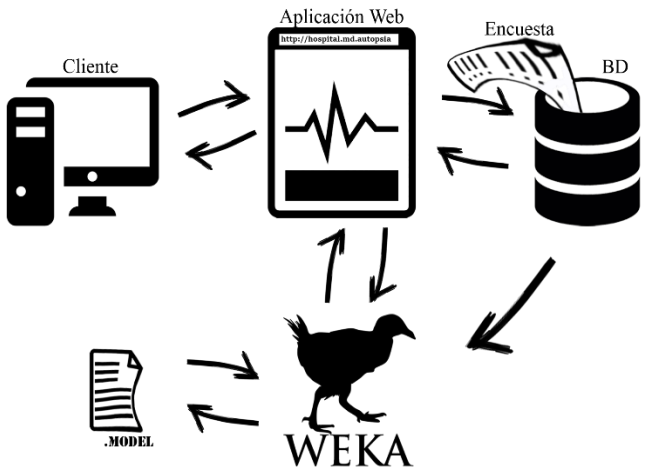


Figura 3.1: Esquema de solución.

La biblioteca de Weka permite asimismo consultar los modelos para interpretarlos y, mediante la aplicación, mostrar los resultados de manera gráfica y comprensible para los usuarios finales.

La API de Weka provee todas las funcionalidades a las que se accede desde la herramienta de escritorio Weka, desarrollada en lenguaje Java en la Universidad de Waikato, en Nueva Zelanda.

Para la investigación se emplean técnicas de minería como las de asociación, con el fin de encontrar relaciones entre los atributos y determinar patrones de comportamiento en los datos, y las de clasificación, para definir a qué clase pertenecen determinados elementos.

En la implementación de estas funcionalidades se hace necesario el uso de paquetes proveídos por la API de Weka. Los utilizados en este estudio fueron esencialmente dos:

weka.associations: Es el paquete que agrupa los algoritmos para el aprendizaje de reglas de asociación.

weka.classifiers: Es el paquete que contiene los algoritmos de clasificación discreta y los de predicción numérica. Incluye varios subpaquetes, muchos de los cuales se emplean en la investigación como *weka.classifiers.bayes*, *weka.classifiers.rules*, *weka.classifiers.lazy*, *weka.classifiers.trees*, *weka.classifiers.functions*, *weka.classifiers.meta*, entre otros.

A continuación, se describe la arquitectura de la aplicación, el proceso para su desarrollo guiado por la metodología UWE y las actividades de MD de acuerdo al método KDD.

3.1 Diseño de la aplicación

La fase de diseño en el desarrollo de software permite que todos los involucrados en el proceso a través de una descripción de todo el sistema conozcan su comportamiento antes de llevarlo a cabo, así como estimar tiempos y costos, detectar errores y hacer correcciones en etapas tempranas. La primera actividad dentro de esta fase es la definición de la arquitectura del sistema para identificar las necesidades reales a cubrir por el mismo. En la Figura 3.2 se representa la arquitectura para la aplicación web propuesta en este trabajo.

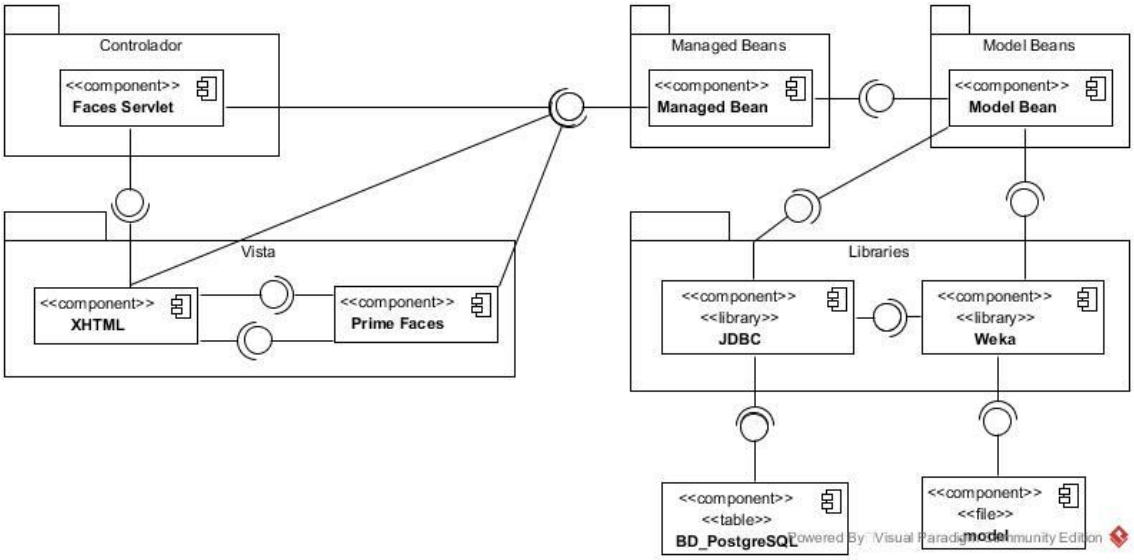


Figura 3.2 Arquitectura de la aplicación.

3.1.1 Arquitectura

Los patrones arquitectónicos son muy utilizados en el diseño de la arquitectura de software porque proporcionan soluciones a problemas comunes dentro del entorno y su efectividad se probó en problemas similares.

Estos patrones de alto nivel ayudan a definir la estructura y organización de los sistemas de software. Su aplicación genera grandes beneficios en el desarrollo de software como son: reducción en tiempos de desarrollo y mayor robustez, calidad y facilidad del mantenimiento.

La arquitectura propuesta para la aplicación se fundamentó sobre uno de los patrones más utilizados [41] para el desarrollo de aplicaciones web, conocido como Modelo-Vista-Controlador (MVC).

Patrón MVC

El MVC separa la lógica de negocio de la representación y la persistencia, y define tres capas dentro de la aplicación. Estas son:

***Modelo*:** Agrupa la lógica de la aplicación, que está representada en los *beans* administrados, con acceso a los componentes de la interfaz y capacidad para pasar información a los *beans* de modelo. A su vez, estos últimos representan las clases importantes del dominio; usan la API de Weka para generar los modelos de minería de datos de los que se extraerá el nuevo conocimiento y la de JDBC para controlar el acceso al gestor de base de datos PostgreSQL y manipular la información.

***Vista*:** Es la capa en la que, mediante archivos XHTLM, se representa el modelo y maneja la interacción con el usuario. Para este propósito se utilizan etiquetas propias de JSF y componentes de la biblioteca *PrimeFaces*.

***Controlador*:** Esta capa contiene al servlet de JSF, que funciona como vínculo entre el modelo y la vista. Se ocupa además de gestionar las peticiones de los recursos accediendo al modelo requerido en la petición del usuario y seleccionando la vista adecuada para representarlo.

Como se observa, la estructura y organización que propone el MVC proporcionan un buen acoplamiento entre los componentes, lo que permite que los cambios solo sean perceptibles para las partes directamente involucradas.

La arquitectura del software instrumentado en esta investigación se representó mediante el Lenguaje de Modelado Unificado (UML) por ser estándar, lo cual lo hace más fácil de entender y coherente con el lenguaje que propone UWE, la metodología a seguir en el desarrollo de la aplicación.

3.1.2 Análisis de requerimientos

Su objetivo es encontrar los requisitos funcionales de la aplicación web para representarlos como casos de uso en un diagrama.

El primer paso es conocer los diferentes actores de la aplicación y qué actividades realizan. La Tabla 3.1 muestra cada uno de los actores y su descripción.

Tabla 3.1 Actores de la aplicación web.

Actores	Descripción
Anónimo/Encuestado	Representa al usuario que no se ha autenticado en el sistema y solamente es capaz de responder encuestas y consultar resultados.
Especialista	Representa al usuario que se autentica en el sistema y tiene permisos para consultar y eliminar encuestas, generar modelos, consultar resultados y cerrar sesión.
Administrador	Representa al usuario que se autentica en el sistema y tiene permisos para insertar, modificar, consultar y eliminar usuarios, consultar resultados y cerrar sesión.

Después de identificar los actores y las funciones que realiza la aplicación, se representaron las relaciones entre éstos mediante un diagrama de casos de uso, mostrado en la Figura 3.3.

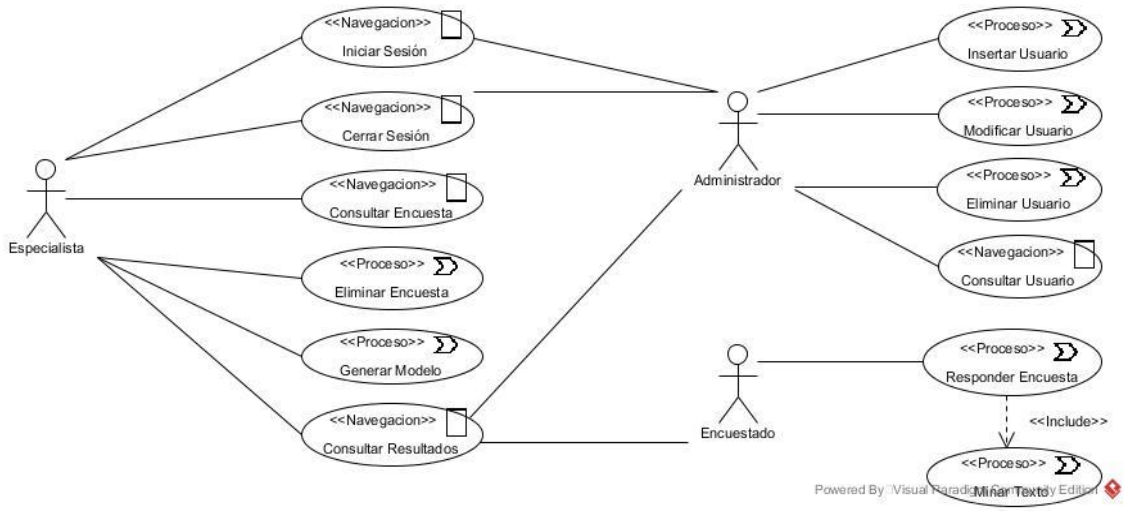


Figura 3.3 Diagrama de casos de uso de la aplicación web.

Es necesario que los actores *Administrador* y *Especialista* se autentiquen en el sistema para realizar sus operaciones y, por tanto, también son capaces de finalizar su sesión en el momento que así lo consideren.

El *Administrador* es el rol encargado de gestionar los usuarios registrados en la aplicación, por lo que tiene permisos para insertar uno nuevo y modificar, consultar o eliminar uno existente.

Por su parte, el rol *Especialista* es el encargado de gestionar la información de las encuestas dentro del sistema, por lo que es capaz de consultar y eliminar encuestas. Este actor también es el responsable de generar los modelos de minería.

El *Encuestado,* por su lado, no se autentica en la aplicación, ya que juega el papel de usuario anónimo que responde una encuesta, proceso que invoca a otro llamado *Minar Texto,* encargado de clasificar las respuestas dadas por los encuestados en las preguntas abiertas. Para el *Encuestado* y el resto de los actores la funcionalidad de consultar resultados está disponible.

Mediante el diagrama de casos de uso descrito es que se representan las interacciones de los actores con la aplicación, lo que permite identificar los requerimientos funcionales. Además, cada caso de uso, como lo propone UWE, se describe por un diagrama de *workflow* para identificar las actividades que en ellos intervienen.

El caso de uso *Iniciar Sesión* se representa en el diagrama de actividad de la Figura 3.4. Éste indica que el proceso se inicia mostrando en pantalla un formulario mediante el

cual el usuario proporciona sus credenciales de acceso para que el sistema inicie la sesión indicada de acuerdo a las mismas, lo cual da fin al referido caso de uso.

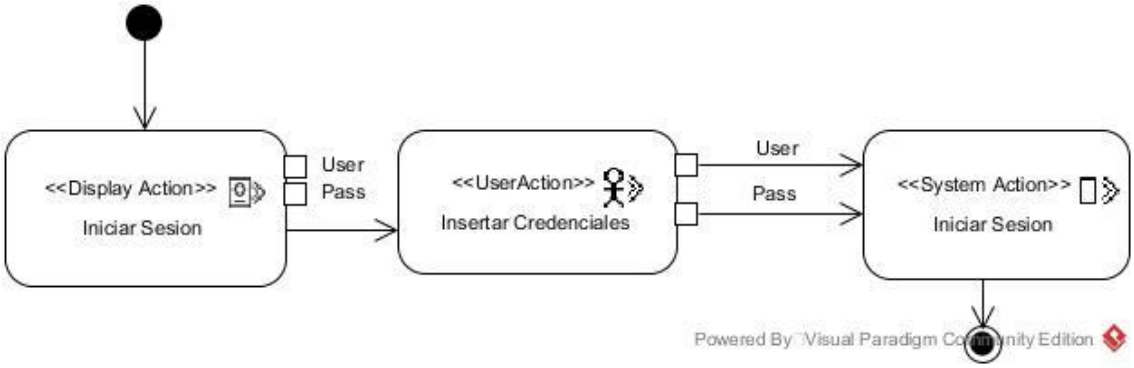


Figura 3.4 Diagrama de actividad del caso de uso *Iniciar Sesión*.

El diagrama de actividad del caso de uso contrapuesto, *Cerrar Sesión*, se muestra en la Figura 3.5. Éste indica que el caso inicia mostrando en pantalla un formulario mediante el cual el usuario selecciona la opción de cerrar sesión, para que el sistema cierre la sesión actual y finalice así el caso de uso.

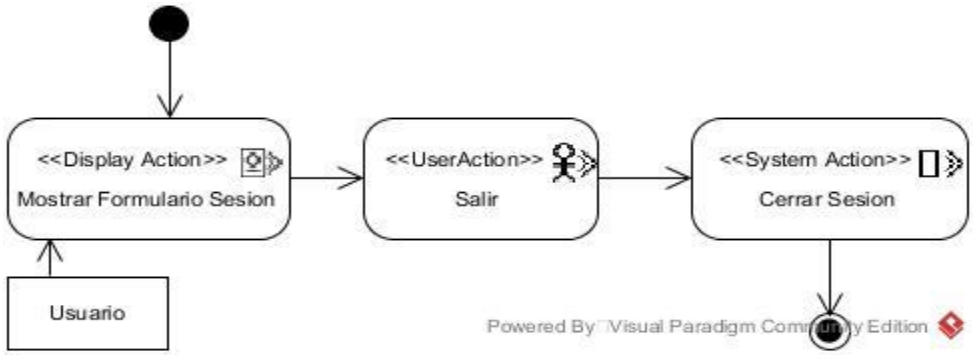


Figura 3.5 Diagrama de actividad del caso de uso *Cerrar Sesión*.

En la Figura 3.6 se muestra el diagrama de actividad para *Responder Encuesta*. Se indica que el caso de uso inicia mostrando en pantalla un formulario mediante el cual el usuario responde las preguntas de la encuesta, para que seguidamente el sistema guarde la nueva encuesta y lance el caso de uso *Minar Texto*. Cuando la encuesta queda registrada en la aplicación y fue objeto de la actividad de minería de texto, el caso de uso se termina.

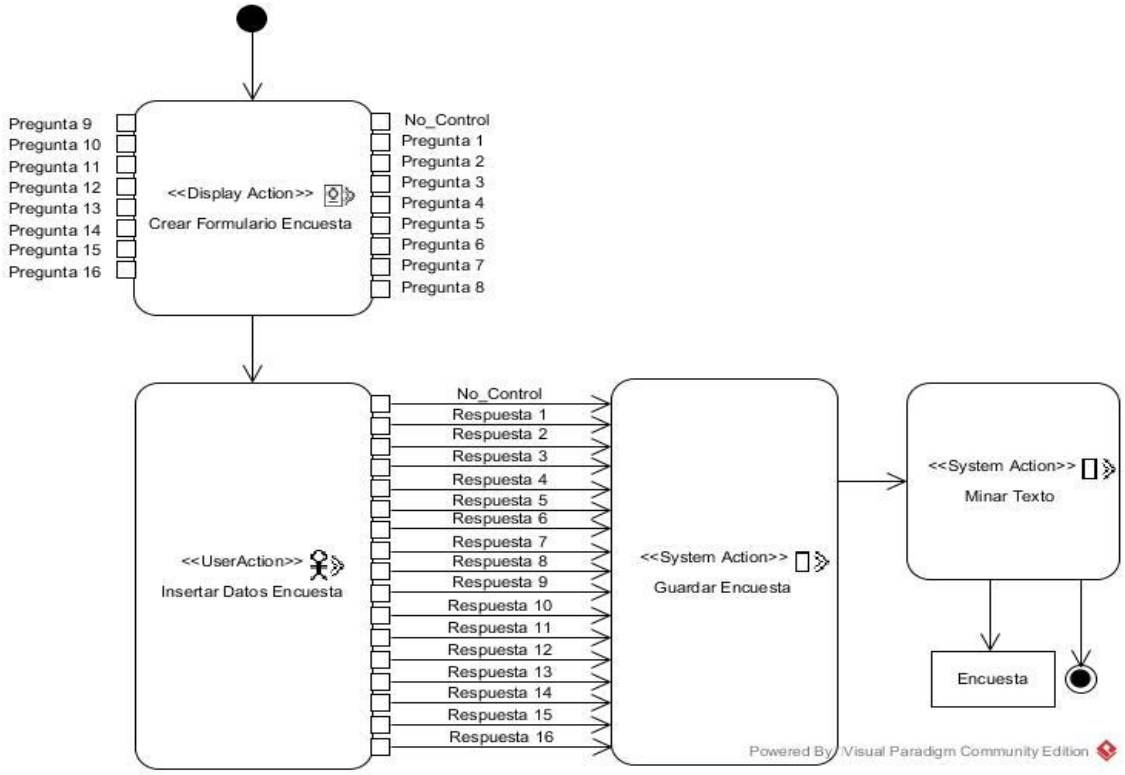


Figura 3.6 Diagrama de actividad del caso de uso *Responder Encuesta*.

El caso de uso *Consultar Encuesta*, representado por su diagrama de actividad en la Figura 3.7, inicia mostrando en pantalla el listado de las encuestas y un formulario mediante el que el usuario proporciona el número de la encuesta requerida, para que entonces el sistema proporcione los datos de la misma y finalice así el caso de uso.

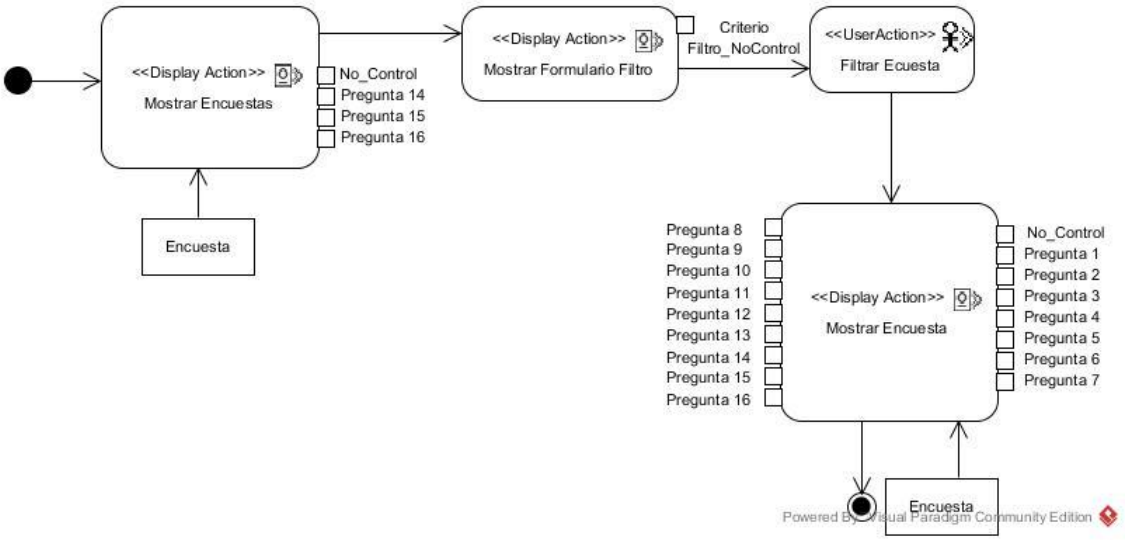


Figura 3.7 Diagrama de actividad del caso de uso *Consultar Encuesta*.

Por otro lado, para el caso de uso *Eliminar Encuesta*, representado en el diagrama de actividad mostrado en la Figura 3.8, el sistema muestra en pantalla el listado de las

encuestas y un formulario mediante el cual el usuario proporciona el número de la encuesta que se quiere eliminar. A esto el sistema responderá con la muestra de los datos de la misma, para que entonces el usuario indique si desea eliminar la encuesta y proceder a realizar la operación. El caso de uso termina con la eliminación exitosa de la encuesta.

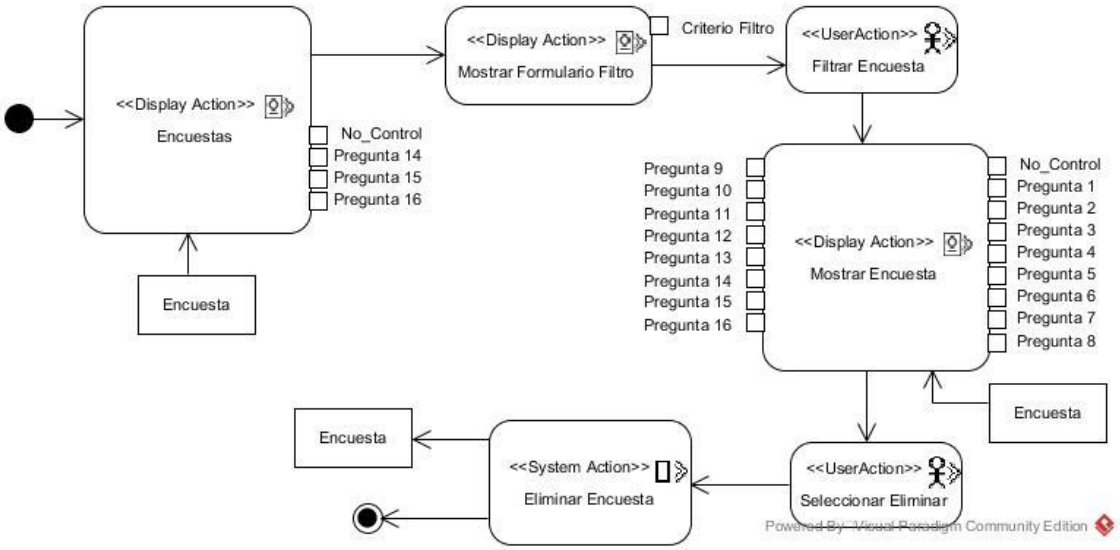


Figura 3.8 Diagrama de actividad del caso de uso *Eliminar Encuesta*.

En la Figura 3.9 se muestra el diagrama de actividad para el caso de uso *Generar Modelo*. Este inicia cuando se muestran en pantalla los resultados del modelo existente y el usuario selecciona la opción *Generar Modelo*. Cuando el nuevo modelo es generado, termina el caso de uso.

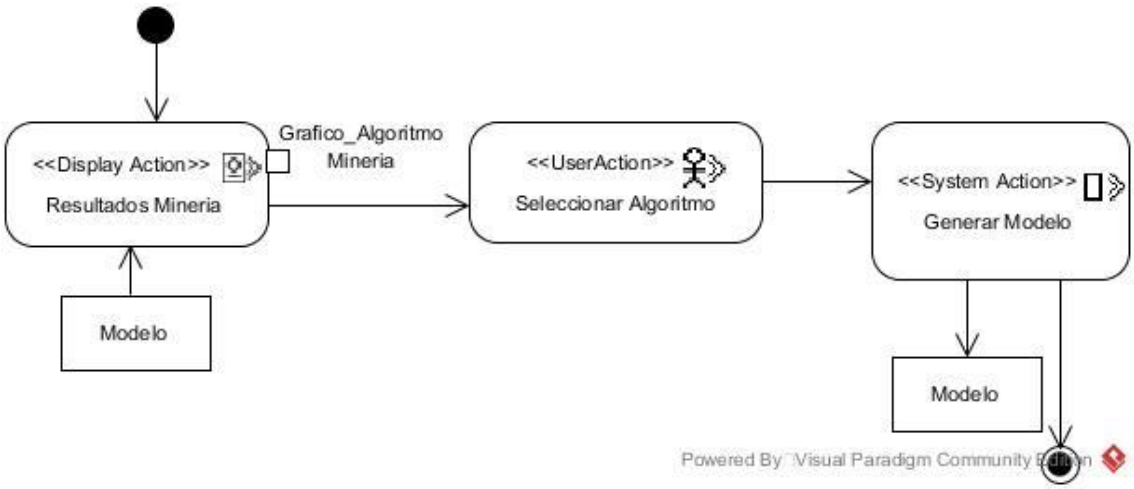


Figura 3.9 Diagrama de actividad del caso de uso *Generar Modelo*.

El caso de uso *Interpretar Resultados de Minería* está representado por su diagrama de actividad en la Figura 3.10, donde se indica que el caso inicia cuando el sistema hace una interpretación del modelo y termina una vez que los resultados son mostrados en pantalla.

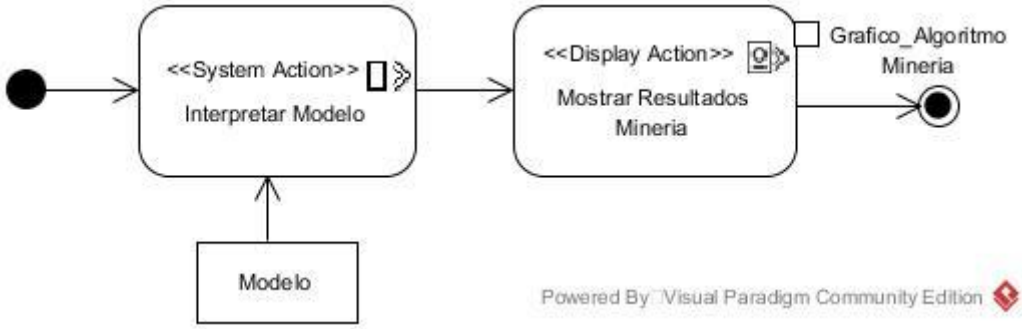


Figura 3.10 Diagrama de actividad del caso de uso *Interpretar Resultados de Minería*.

El diagrama de actividad para *Insertar Usuario* se muestra en la Figura 3.11. Ésta indica que el caso de uso inicia mostrando en pantalla un formulario mediante el cual el usuario autorizado proporciona los datos necesarios, para que seguidamente el sistema guarde el nuevo usuario y dé fin al caso.

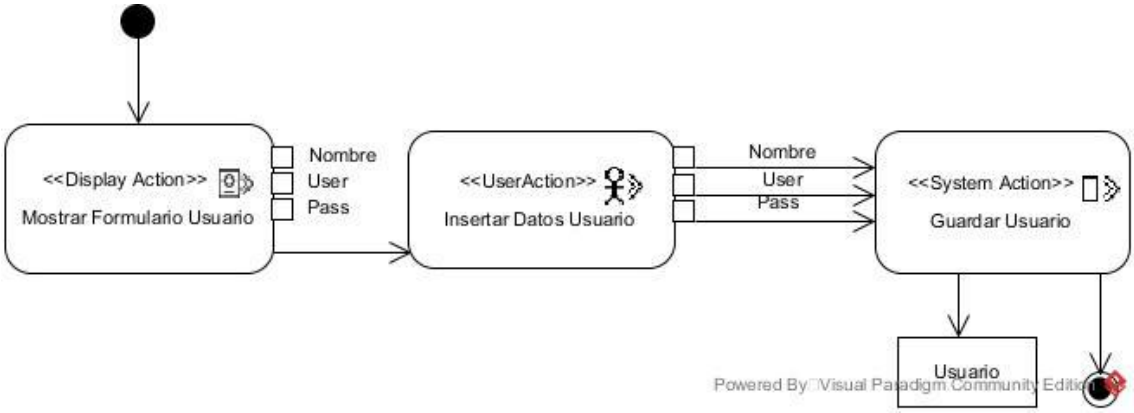


Figura 3.11 Diagrama de actividad del caso de uso *Insertar Usuario*.

El caso de uso *Consultar Usuario*, como explica su diagrama de actividad en la Figura 3.12, inicia con la muestra en pantalla del listado de usuarios registrados en la aplicación y un formulario mediante el cual el usuario autorizado filtra los resultados. Para ello proporciona el identificador del usuario que se quiere consultar, a lo que el sistema responderá consecuentemente y dará fin al caso de uso.

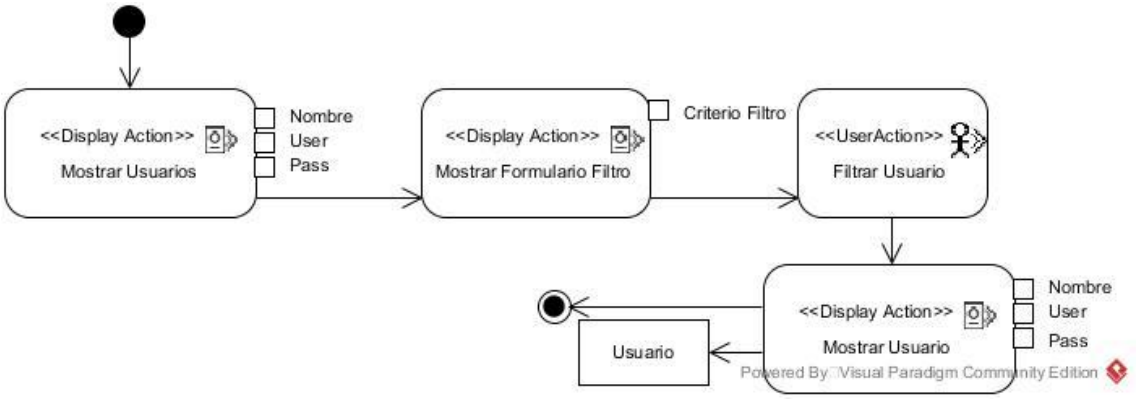


Figura 3.12 Diagrama de actividad del caso de uso *Consultar Usuario*.

En la Figura 3.13 se muestra el diagrama de actividad para el caso de uso *Actualizar Usuario*. Éste comienza mostrando en pantalla el listado de los usuarios registrados en la aplicación y un formulario mediante el cual el usuario autorizado proporciona el identificador del o los usuarios que desea actualizar. Luego el sistema proporciona los datos correspondientes y el usuario hace los cambios deseados para finalmente, previa corroboración al sistema de que desea actualizar la información, cumplimentar el caso de uso, cuyo término se materializa en el registro permanente de los cambios.

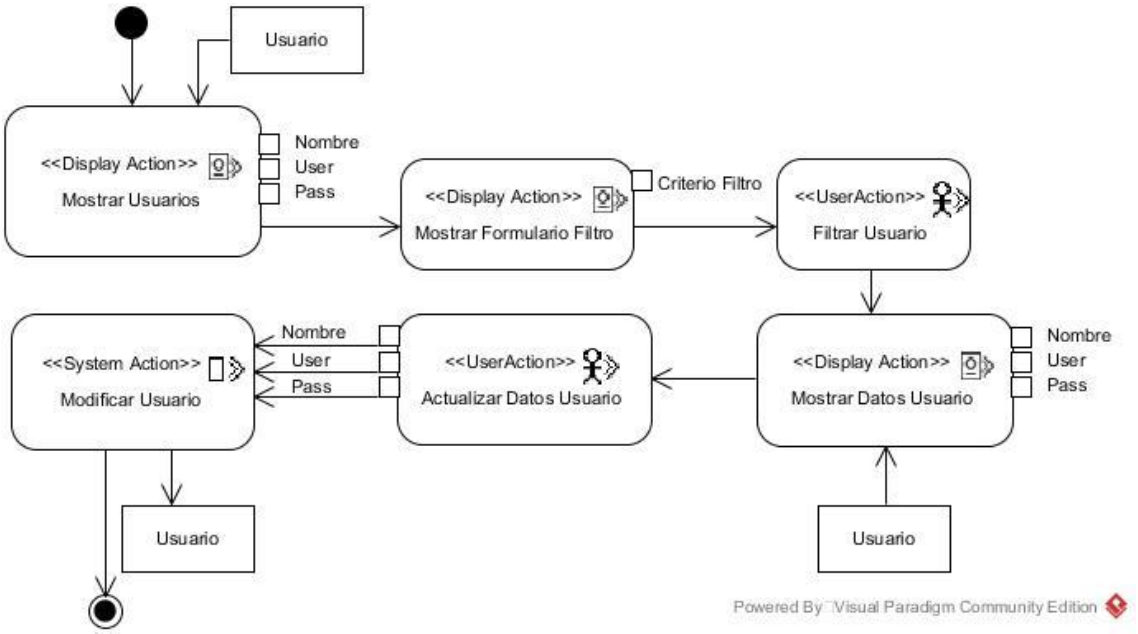


Figura 3.13 Diagrama de actividad del caso de uso *Actualizar Usuario*.

El diagrama de actividad para el caso de uso *Eliminar Usuario* se muestra en la Figura 3.14. Comienza mostrando en pantalla el listado de los usuarios registrados en la aplicación y un formulario mediante el cual el usuario autorizado proporciona el

identificador del usuario que desea eliminar. Hecho esto el sistema proporcionará los datos del mismo y en este punto el usuario indica que desea eliminar al usuario, para que el sistema proceda a realizar la operación. El caso de uso termina con la eliminación exitosa del usuario.

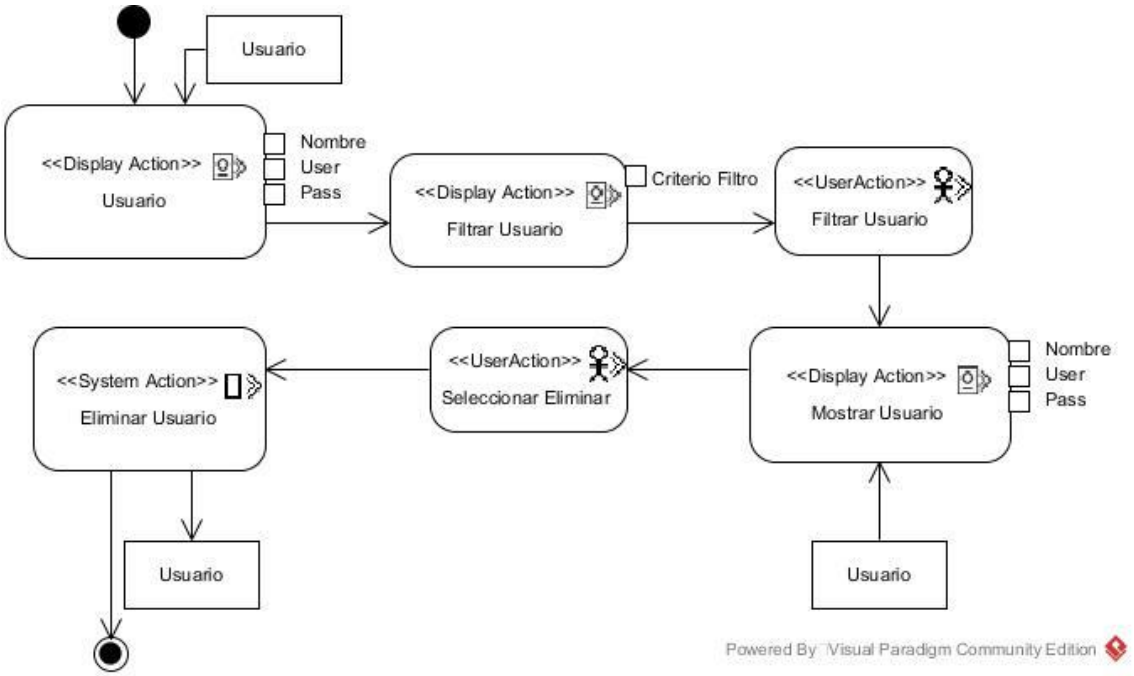


Figura 3.14 Diagrama de actividad del caso de uso *Eliminar Usuario*.

3.1.3 Modelo conceptual

El objetivo de este modelo es representar de una manera conceptual el dominio de la aplicación a partir de los requisitos. Propone diagramas como: conceptual, lógico y físico de la base de datos, Figuras 3.15, 3.16 y 3.17, respectivamente.

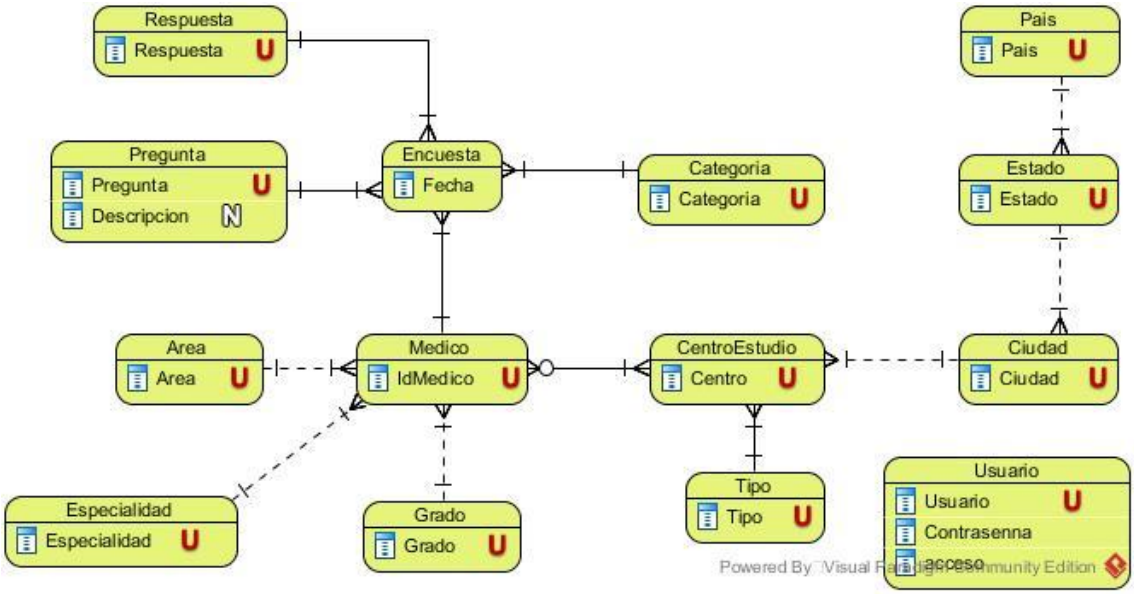


Figura 3.15 Diagrama conceptual de la aplicación.

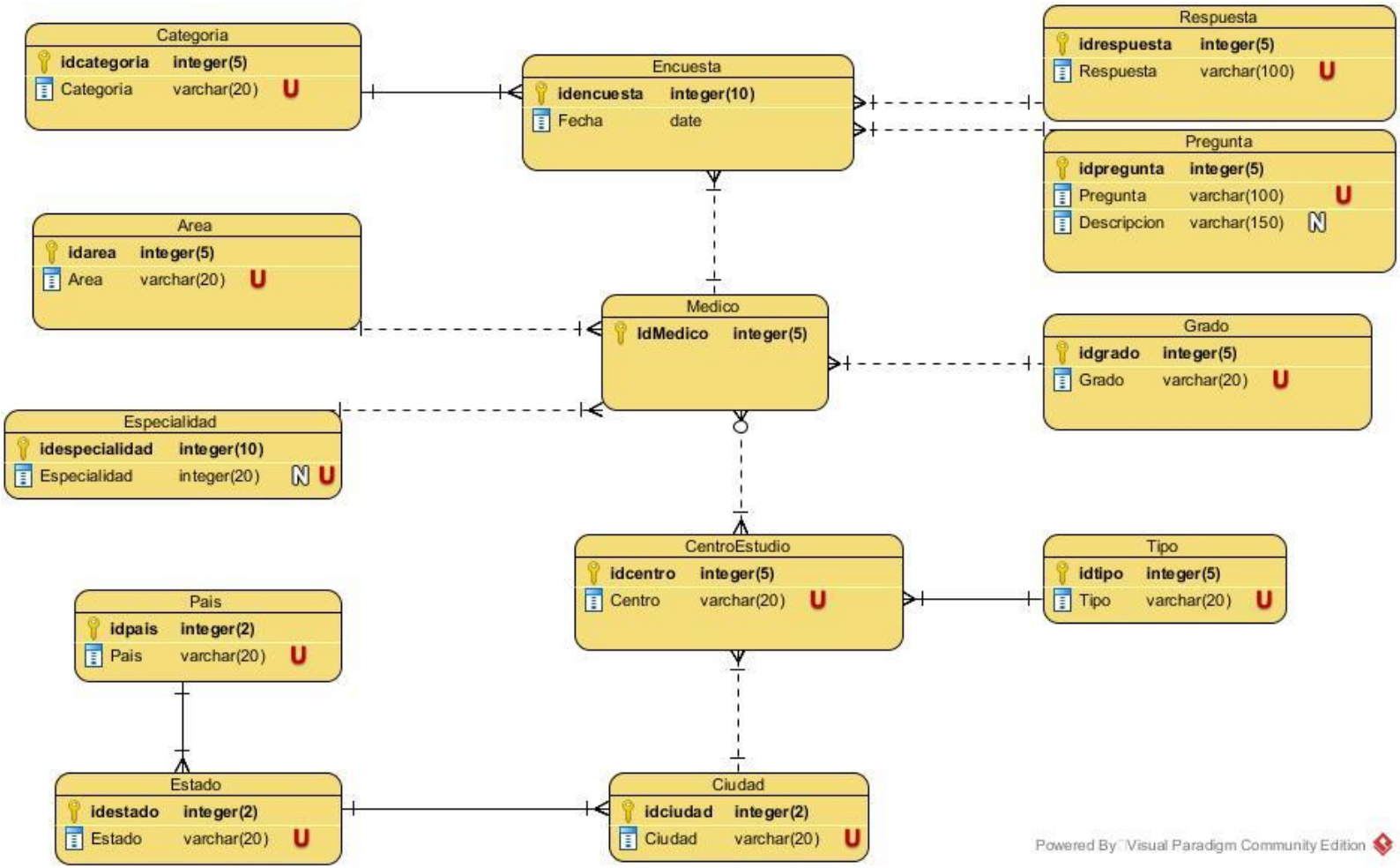


Figura 3.16 Diagrama lógico de la aplicación.

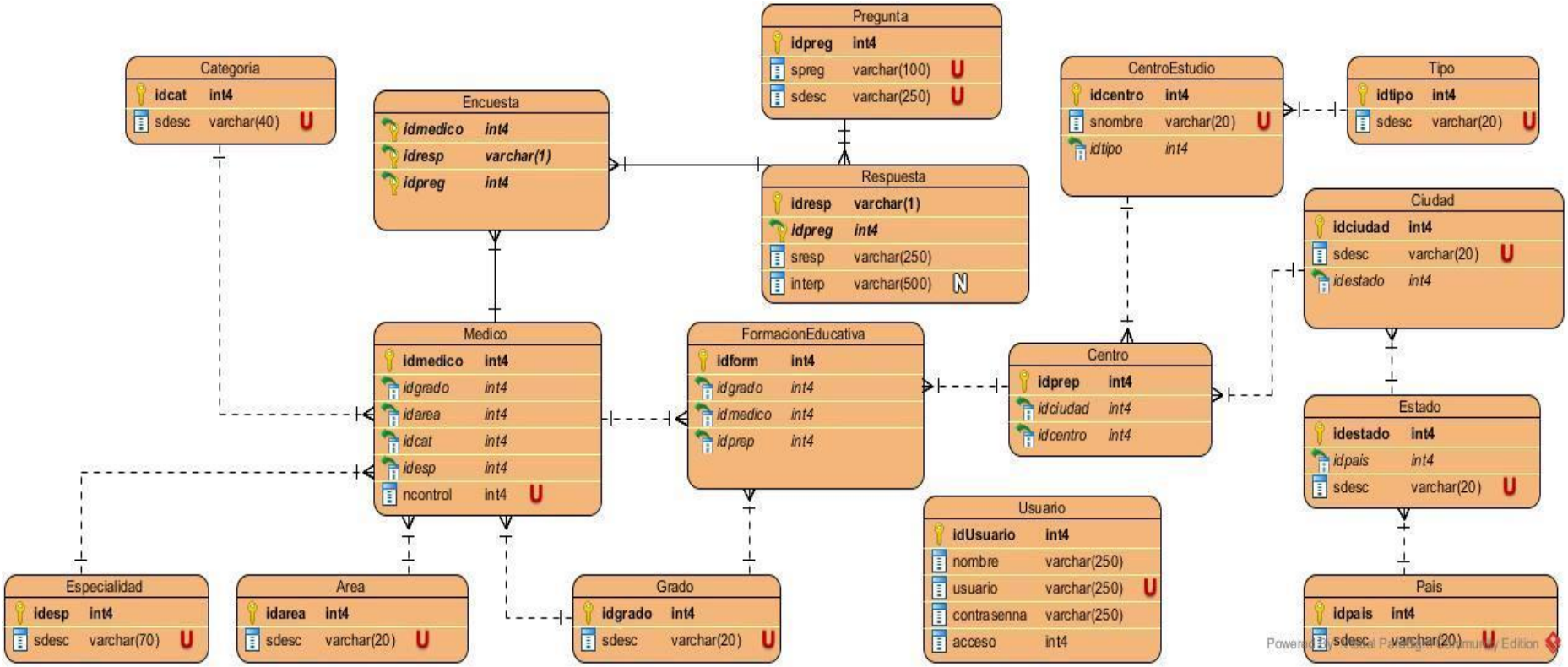


Figura 3.17 Diagrama físico de la aplicación.

Tablas de la base datos

***Categoría*:** Almacena la información referente a la forma en la que el encuestado accedió a responder la encuesta, por ejemplo, si fue mediante una "Invitación interna" o "Invitación accionada fuera del servicio".

Especialidad: Guarda el valor que permitirá indicar a qué campo pertenece un encuestado. Para este caso de estudio el valor que hace referencia al sector médico es "Medicina".

Área: Registra la información que señala en calidad de qué presta servicio el encuestado al hospital, por ejemplo, si es "Adscrito", "Residente" o "Interno".

Grado: Guarda el valor que permitirá indicar el nivel de especialización dentro del campo médico que tiene el encuestado, como "Práctica Universitaria", "Medicina General" o "Especialidad".

Centro Estudio: Contiene los nombres de centros que están involucrados en la enseñanza de medicina. Algunos ejemplos son: "Facultad Medicina", "Universidad Veracruzana", "Hospital Regional de Río Blanco", entre otros.

Tipo: La información contenida en esta tabla responde a la clasificación de los centros de preparación, como son "Escuela" u "Hospital".

Ciudad, Estado, País: Estas tablas y sus relaciones permiten darles una ubicación geográfica a los centros educacionales. Por ejemplo, es posible conocer si en un momento determinado se hace referencia a la "Universidad Veracruzana" del estado "Veracruz" y ciudad "Mendoza", o del estado de "Veracruz" y ciudad "Xalapa".

Pregunta: La tabla registra cada una de las preguntas que conforman la encuesta. Como ejemplo se mencionan las preguntas "Años de práctica" y "¿Quién considera que debe solicitar la autopsia?".

Respuesta: La tabla respuesta contiene todas las respuestas posibles para la encuesta y establece la relación de éstas con su pregunta correspondiente. Ejemplo de ello son casos como "Menos de 5" y "Más de 20", que son posibles respuestas de la pregunta "Años de práctica". La relación se establece mediante la llave foránea *idpreg* en la tabla *Respuesta,* que responde a la llave primaria *idpreg* de la tabla *Pregunta.*

Centro: De acuerdo a las relaciones establecidas con las tablas *Centro Estudio* y *Ciudad* se logra asignar a cada centro de preparación médica su ubicación geográfica.

Médico: Es la tabla que almacena las encuestas y relaciona el número de control de cada una con su categoría, especialidad, grado y área.

Formación Educativa: En esta tabla se establecen relaciones con las tablas *Médico, Grado y Centro*; de esta manera se registra la información de los centros donde se formó cada encuestado.

Encuesta: Las relaciones establecidas con las tablas *Médico, Pregunta* y *Respuesta* permiten registrar las respuestas para cada una de las preguntas por cada encuestado.

Clasificado: En esta tabla persisten las respuestas de las preguntas abiertas con su clasificación correspondiente, resultado del proceso de minería para la clasificación de texto.

Interpretación: Con el objetivo de dar una explicación lógica de los resultados que generan los algoritmos de minería aplicados en esta investigación, esta tabla guarda el significado y la descripción para cada atributo contenido en los conjuntos de datos analizados.

Matriz_binaria y Vista_minable: Estas tablas representan a toda la información de las encuestas insertadas en la base de datos y se crean automáticamente por funciones SQL.

Menú*:* La estructura para el menú de la aplicación está contenida en esta tabla y brinda la posibilidad que pueda ser personalizado de acuerdo a los diferentes usuarios.

3.1.4 Modelo de navegación

Este modelo propone la construcción de un mapa de navegación mediante un diagrama de clases estereotipado, que abarca todos los caminos posibles para la navegación de los usuarios. El modelo de navegación de la aplicación se muestra en la Figura 3.18.

En el mapa de navegación se ven todos los caminos posibles que indican cómo un usuario se mueve de un sitio a otro dentro de la aplicación. Partiendo de la página de inicio, desde el Menú Principal un usuario elige ir a consultar los resultados, insertar una encuesta o iniciar sesión. Si decide seguir por la última opción tiene cuatro opciones para continuar.

La primera de estas últimas muestra la posibilidad desde el Menú Acciones de navegar hasta la página de Usuarios, donde podría insertar nuevo usuario, seleccionar uno específico y a partir de este momento, de acuerdo al Menú Usuario, podría ver los datos del seleccionado, eliminarlo, modificarlo o insertar uno nuevo.

La segunda ofrece al usuario la opción de cerrar su sesión en la aplicación. En la tercera, desde el Menú Acciones un usuario navega hasta la página de Encuestas, punto en el que elige una específica y a partir de ese momento, de acuerdo al Menú Encuesta, ver los datos de la seleccionada o eliminarla.

Por último, la cuarta opción permite al usuario autorizado generar los modelos de minería.

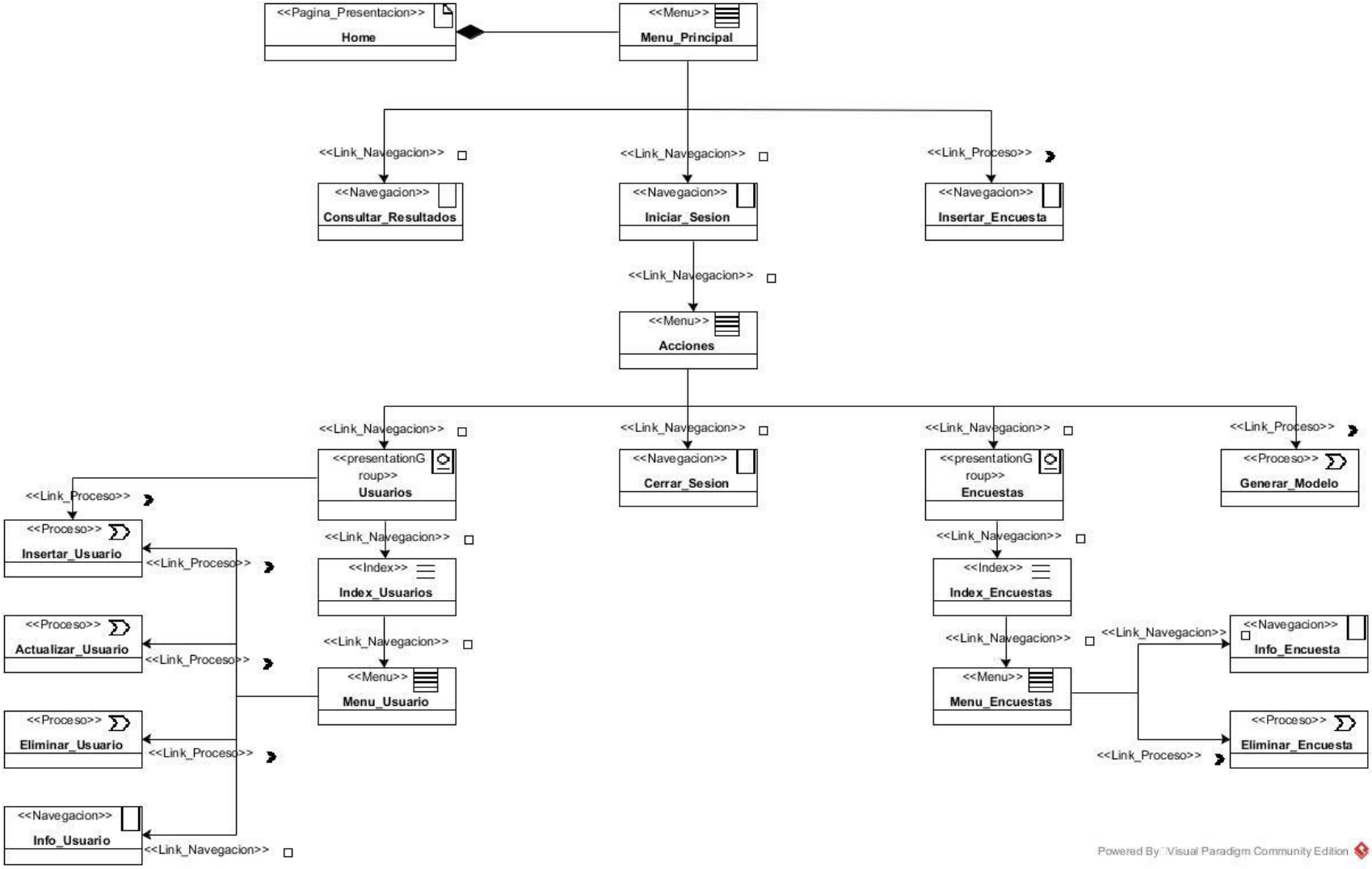


Figura 3.18 Diagrama de navegación de la aplicación web.

3.1.5 Modelo de presentación

El modelo de presentación define las páginas de la aplicación, la estructura y los componentes de las mismas. La aplicación tendrá seis páginas, que se describen a continuación.

La página *Inicio,* ver Figura 3.19, da la bienvenida a los usuarios que acceden a la aplicación. La compone un menú principal que despliega opciones de acuerdo a los permisos de los usuarios, un formulario mediante el cual los usuarios registrados abren y cierran sesión, y un texto de bienvenida.

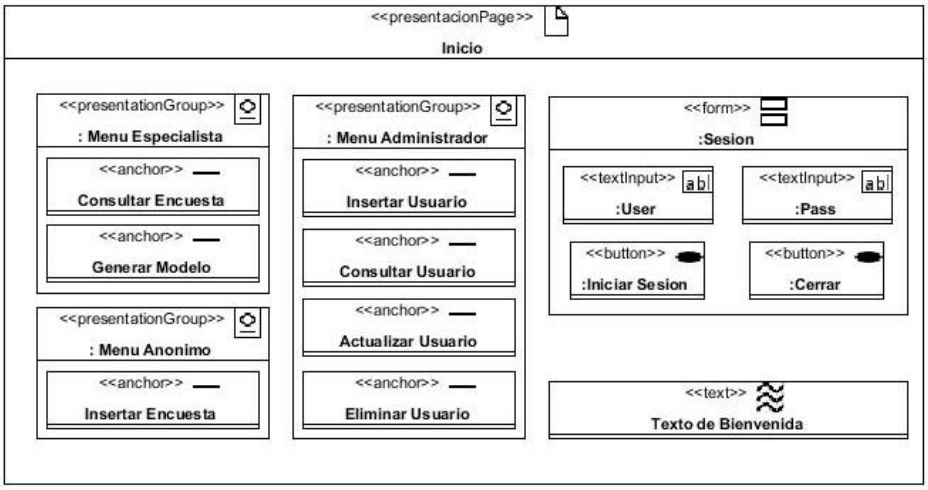


Figura 3.19 Página *Inicio* del modelo de presentación.

Usuario, ver Figura 3.20, es la página desde la cual un administrador gestiona las cuentas de los usuarios registrados en la aplicación. Se muestra una tabla con los datos de los usuarios y las operaciones de actualizar y eliminar vinculadas a cada uno de ellos. Mediante el formulario *Filtro Usuario* se especifica un determinado usuario. Si se desea modificar uno, la alternativa *Gestionar Usuario* mostrará controles de captura para el ítem especificado. La opción de crear uno nuevo también está presente mediante un botón que, al ser seleccionado, muestra la alternativa *Nuevo Usuario* con un formulario para la captura de los datos necesarios para hacer el nuevo registro. *Mensajes* es otra alternativa que indicará mensajes de errores de validación o mensajes de confirmación, según sea el caso.

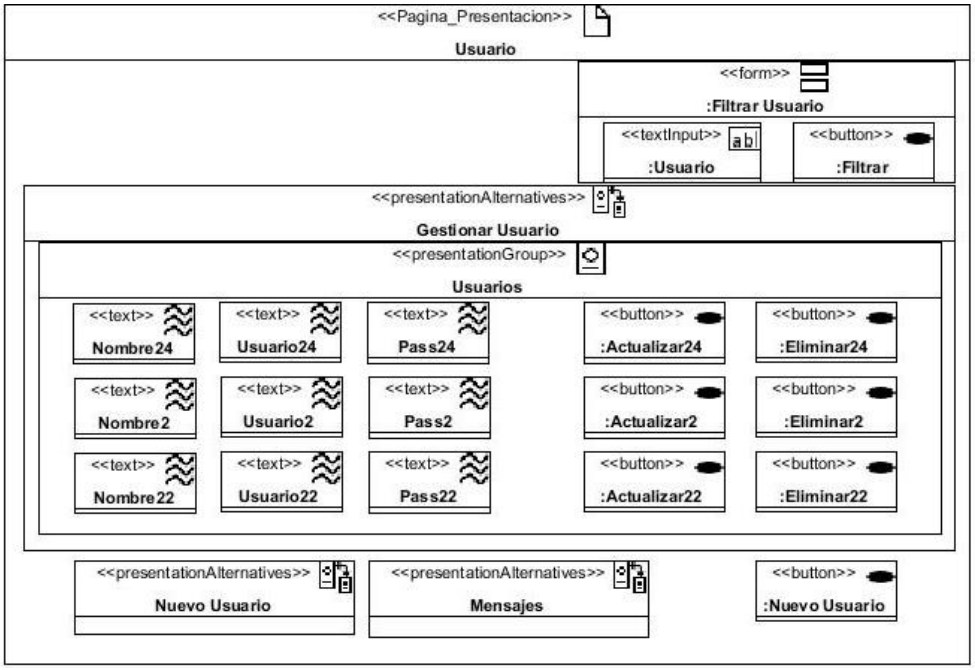


Figura 3.20 Página *Usuario* del modelo de presentación.

La página *Gestionar Encuesta*, ver Figura 3.21, permite al especialista realizar el proceso indicado. Para ello se muestra una tabla con algunos datos de las encuestas y las operaciones de seleccionar y eliminar, vinculadas a cada una de ellas. Mediante el formulario *Filtrar Encuesta* se especifica cuál se desea gestionar. Si se desean ver todas las respuestas de una se elige la operación *Seleccionar* y esta hará una llamada a la página *Encuesta*. Mediante la alternativa de *Mensajes* se indicarán mensajes de errores de validación o mensajes de confirmación, según corresponda.

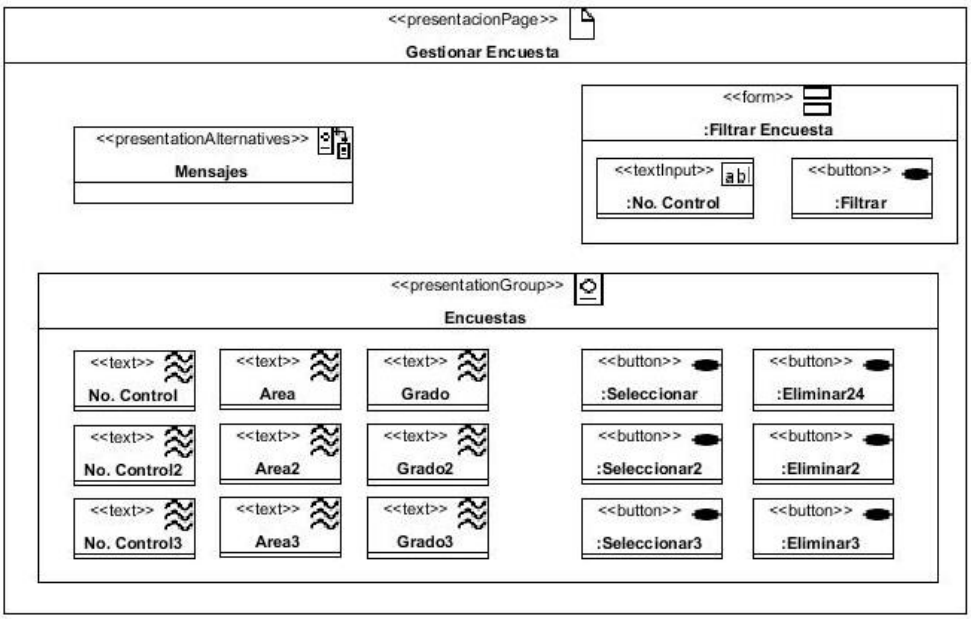


Figura 3.21 Página *Gestionar Encuesta* del modelo de presentación.

Encuesta, ver Figura 3.22, es la página que proporciona un formulario para responder una nueva encuesta o editar una existente, de acuerdo a la operación que haya hecho la llamada a esta página. Los mensajes de error y validación en esta página se manejan de la misma manera que en las anteriores.

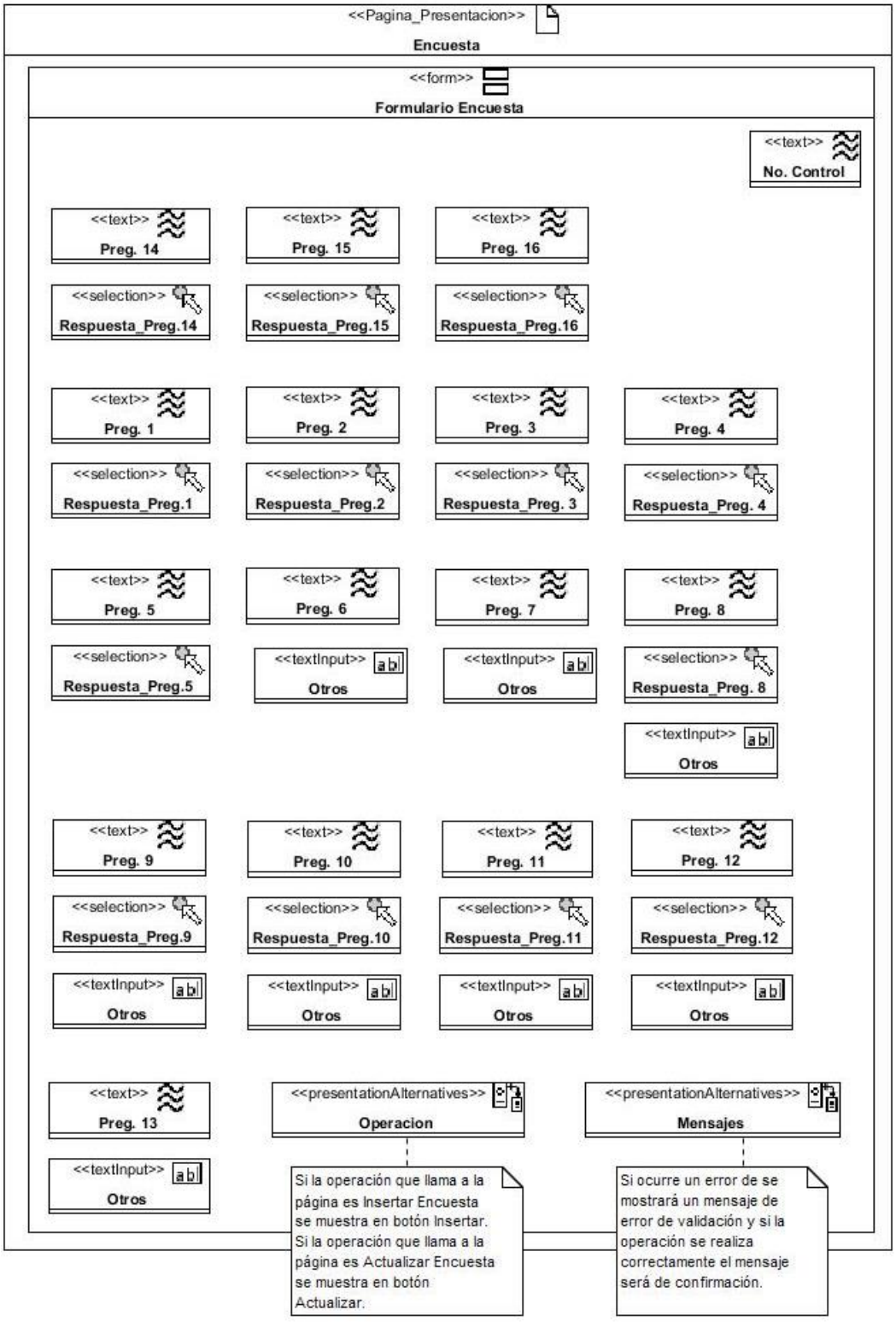


Figura 3.22 Página *Encuesta* del modelo de presentación.

El especialista es capaz de generar los modelos desde la página *Genera Modelos*, ver Figura 3.23, donde selecciona el conjunto de datos que quiere minar, el algoritmo que quiere aplicar y especifica sus parámetros. Al generar el modelo requerido por el especialista, en el área resultados de la página se mostrará la interpretación del modelo generado.

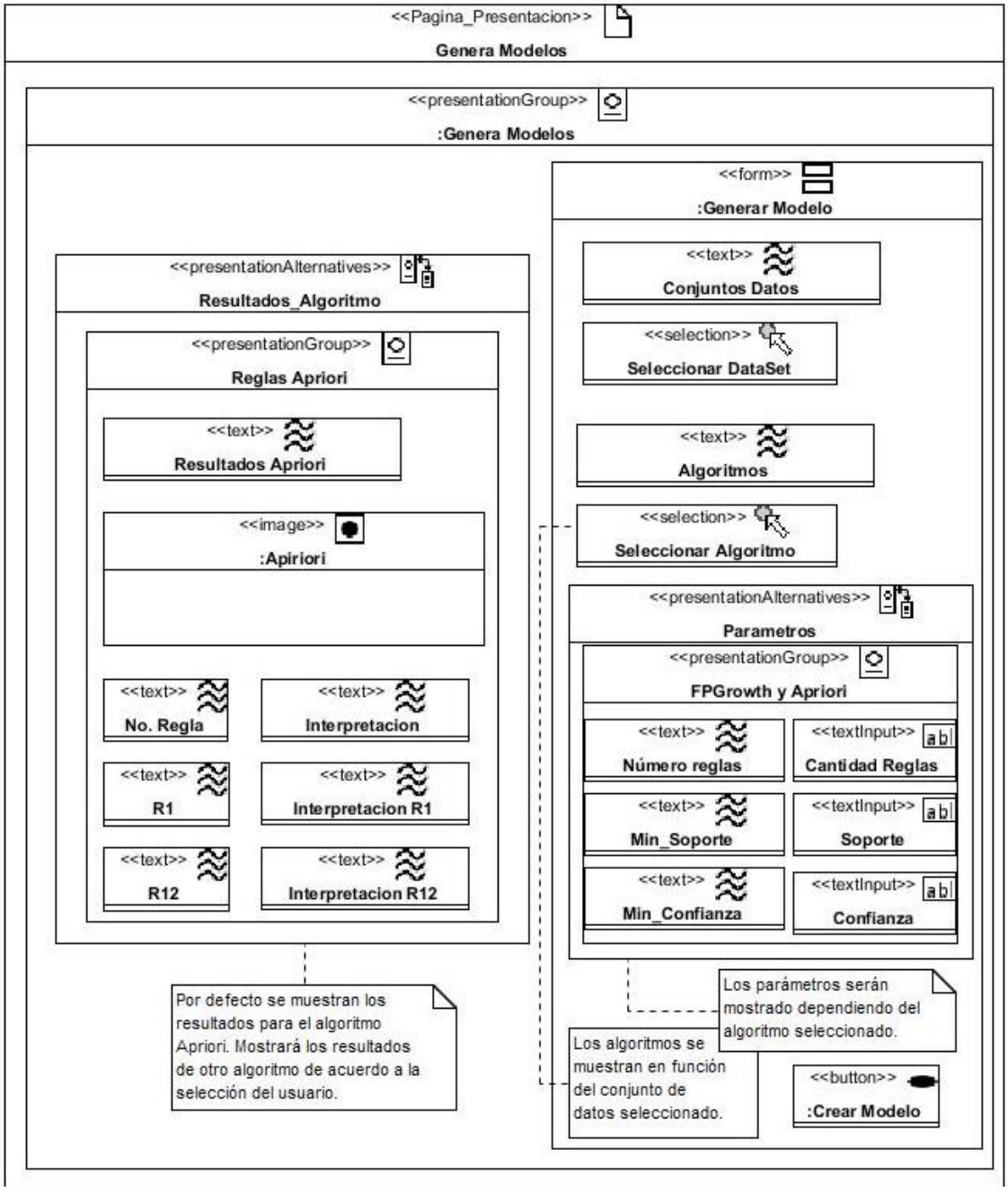


Figura 3.23 Página *Genera modelo* del modelo de presentación.

La página *Resultados*, ver Figura 3.24, muestra la interpretación de los modelos previamente generados. Todos los actores de la aplicación son capaces de llegar a esta página y seleccionar un modelo de interés para ver sus resultados.

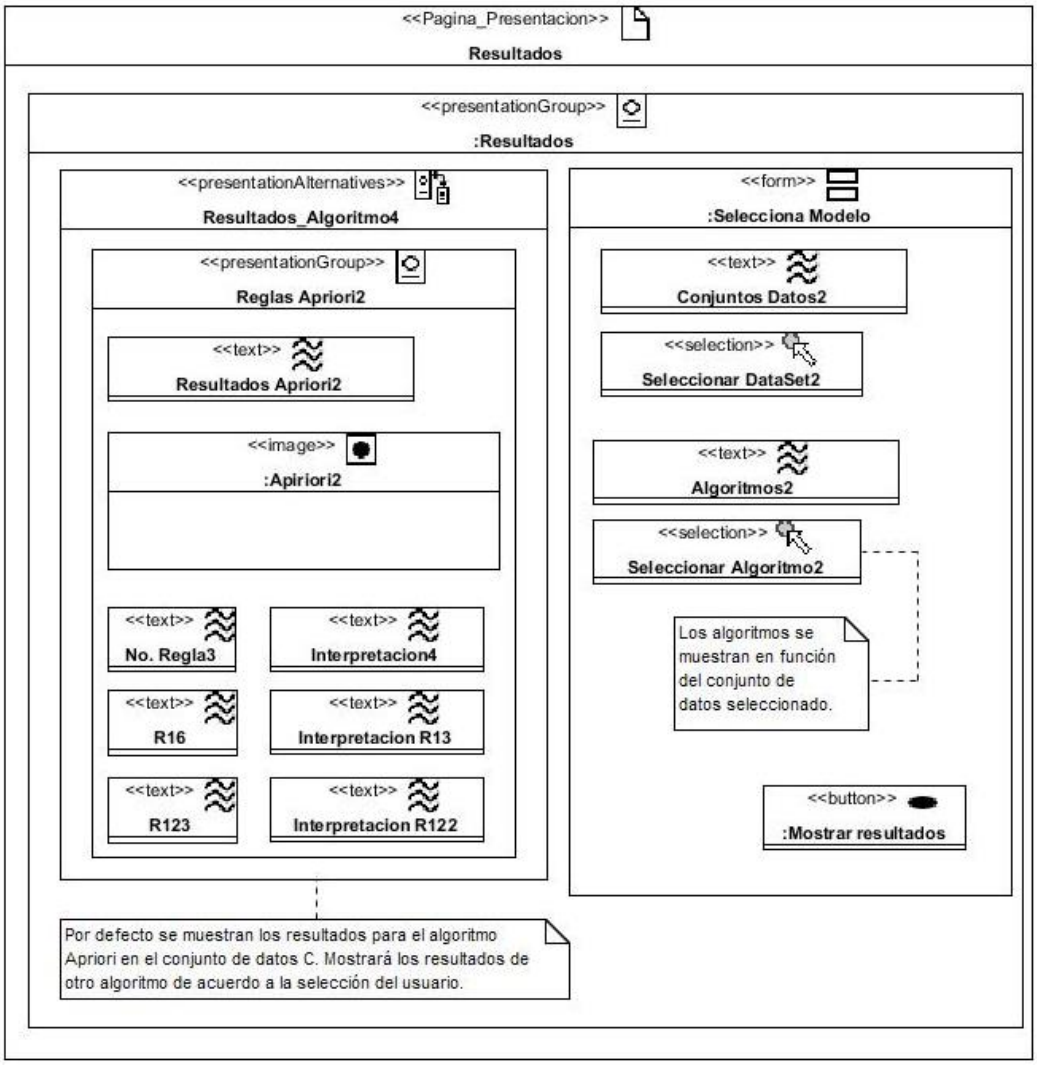


Figura 3.24 Página *Resultados* del modelo de presentación.

A continuación, se describen las áreas alternativas del modelo de presentación. Estas zonas muestran contenido que varía en función de las elecciones de los usuarios al interactuar con la aplicación.

Operación, ver Figura 3.25, mostrará y ocultará los botones de Insertar y Actualizar de acuerdo a la operación que haya hecho la llamada. Para el caso que un encuestado haya solicitado a la aplicación responder una encuesta, se activará el botón Insertar; sin embargo, cuando un administrador haya pedido actualizar los datos de un usuario se activará el botón Actualizar.

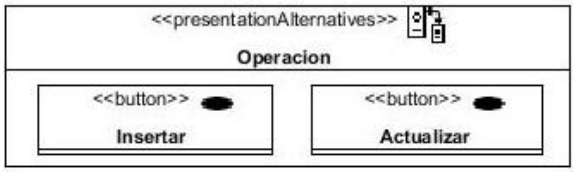


Figura 3.25 Área alternativa “Operación”.

Mensajes, ver Figura 3.26, es el área que permitirá mostrar en las páginas mensajes de errores de validación o mensajes de confirmación.

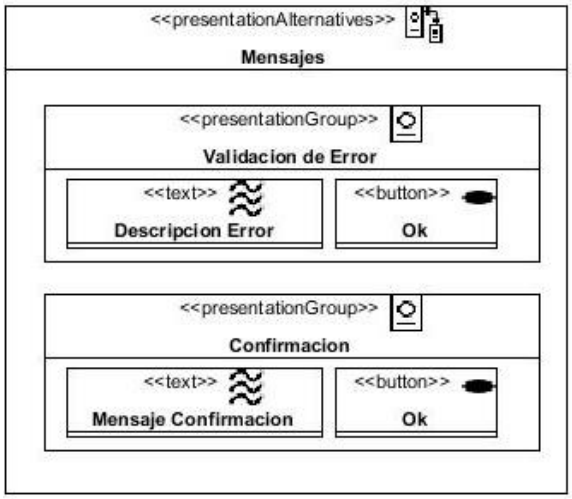


Figura 3.26 Área alternativa “Mensajes”.

Nuevo Usuario, ver Figura 3.27, por defecto oculta su contenido y éste solo será mostrado si un administrador selecciona la opción de registrar un usuario en la aplicación.

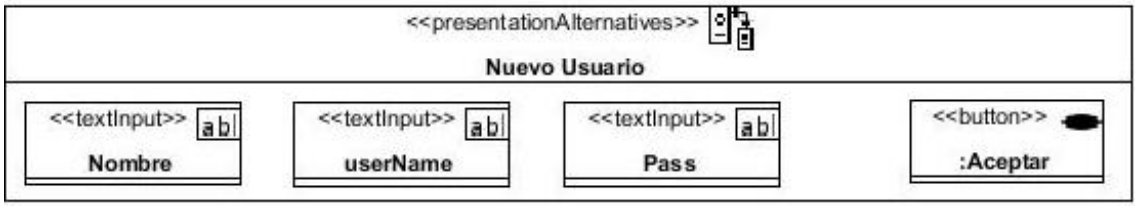


Figura 3.27 Área alternativa “Nuevo Usuario”.

Gestionar Usuario, ver Figura 3.28, muestra por defecto información de los usuarios mediante componentes de solo lectura, pero cuando un administrador quiere actualizar un usuario, los controles que muestra ese objeto en específico cambian a componentes de entrada de texto que permiten capturar los cambios.

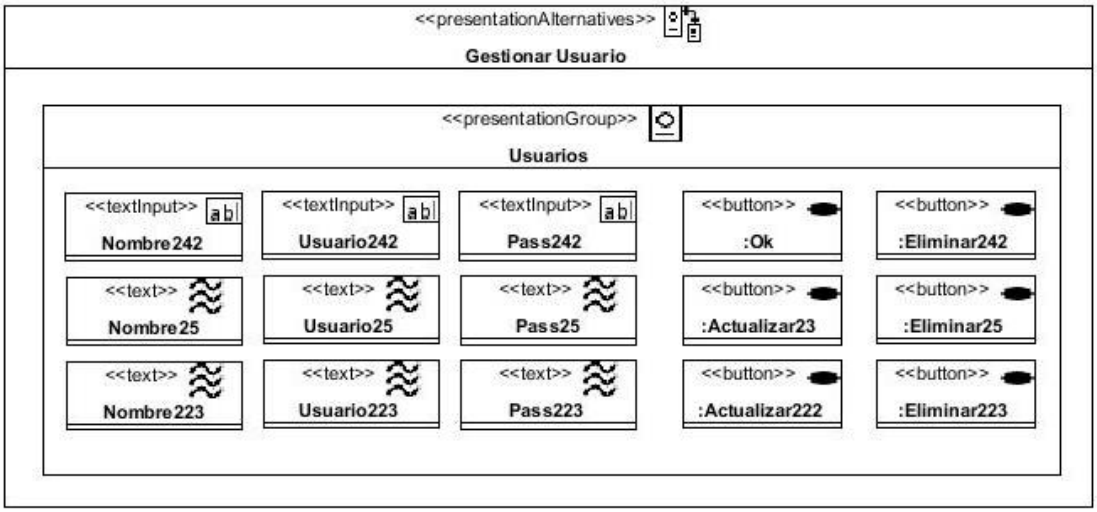


Figura 3.28 Área alternativa "Gestionar Usuario".

Parámetros, ver Figura 3.29, muestra por defecto los parámetros para los algoritmos de *Apriori* y *FPGrowth*. Al igual que en la alternativa anterior, esta información varía de acuerdo al algoritmo elegido por el usuario.

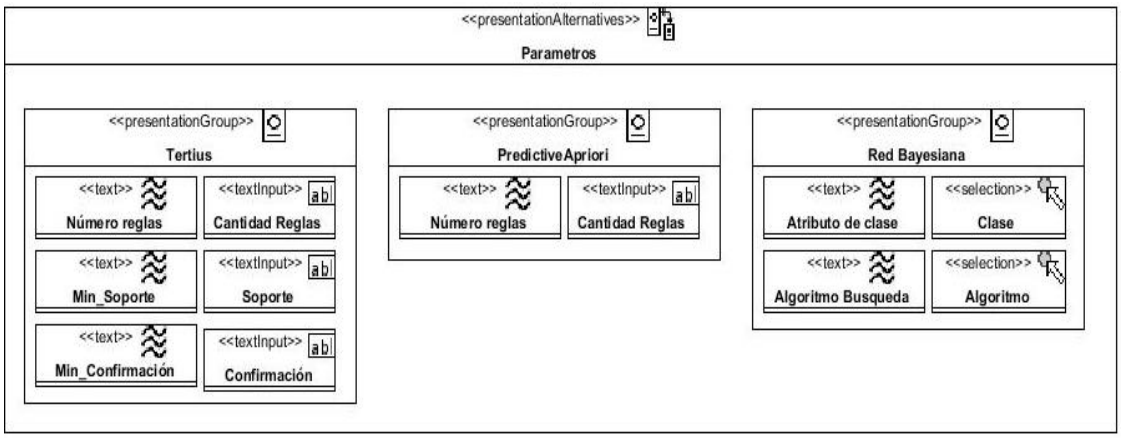


Figura 3.29 Área alternativa "*Parámetros*".

Resultados Algoritmos, ver Figura 3.30, muestra por defecto la interpretación para el modelo de *Apriori* para el conjunto "C". Esta información varía de acuerdo al modelo elegido por el usuario.

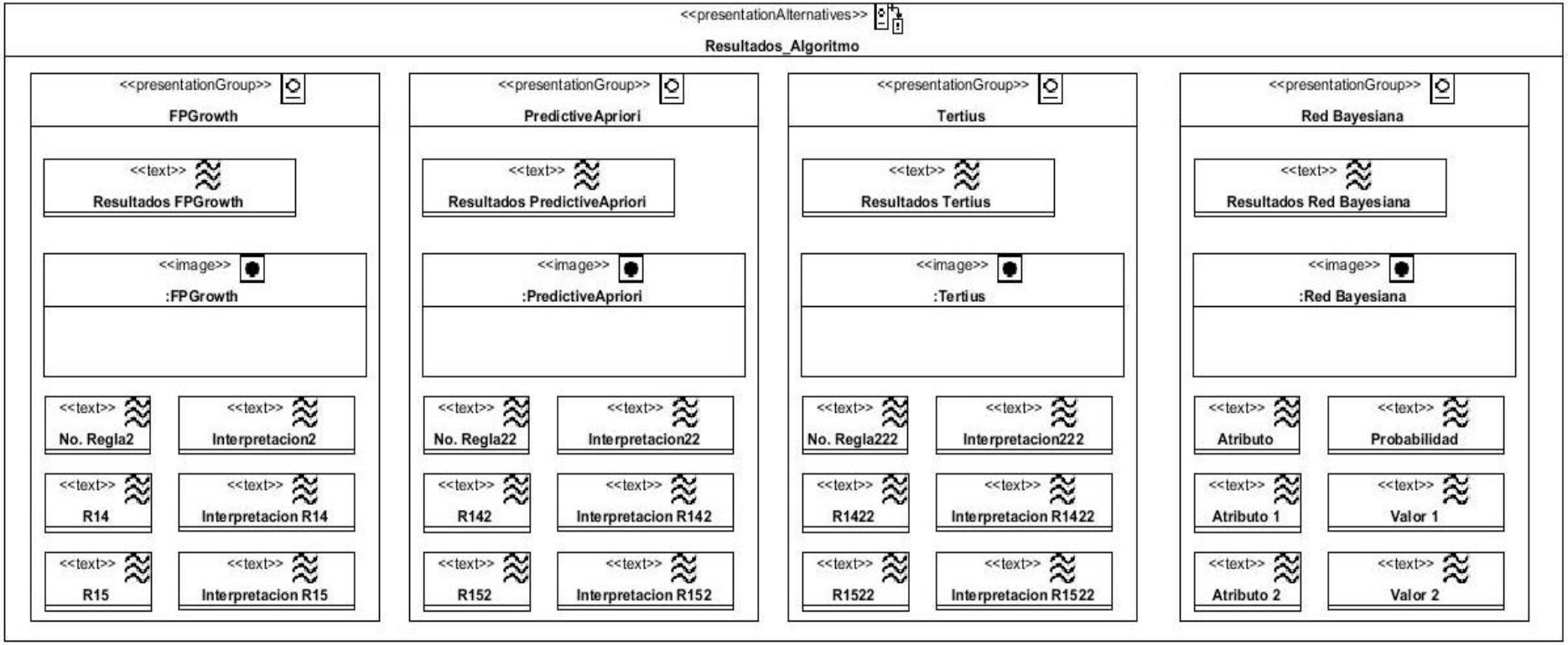


Figura 3.30 Área alternativa "*Resultados_Algoritmo*".

.

3.1.6 Modelo de procesos

En este modelo se detallan los casos de uso mediante diagramas de actividades estereotipados. A continuación, se presentan los diagramas de cada uno de los casos de uso que se identificaron para implementarse en la aplicación.

Un usuario es capaz de autenticarse en la aplicación mediante un formulario que le solicitará sus credenciales. El sistema procederá a validar los datos, si estos son incorrectos se le mostrará al usuario un mensaje de error y regresará al formulario para que corrija la información. Cuando los datos proporcionados sean correctos el sistema iniciará la sesión solicitada. Así se describe el proceso en la Figura 3.31.

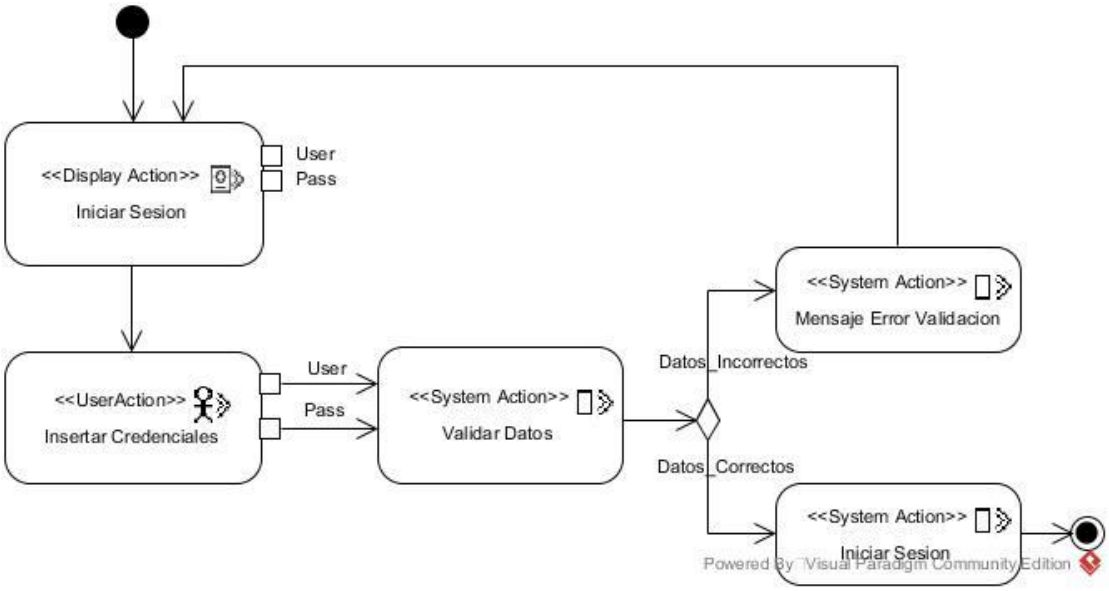


Figura 3.31 Diagrama de proceso del caso de uso *Iniciar Sesión*.

El diagrama de proceso para el caso de uso *Cerrar Sesión* se muestra en la Figura 3.32. Una vez que un usuario se autenticó en el sistema es posible que quiera cerrar su sesión en cualquier momento. Esto lo realiza a partir del formulario que le permite seleccionar la opción de cerrar sesión. El sistema pedirá confirmación al usuario para proceder con la operación; en caso de que el usuario quiera cancelarla, el sistema volverá al estado donde comienza el caso de uso, si confirma, entonces el sistema cerrará la sesión.

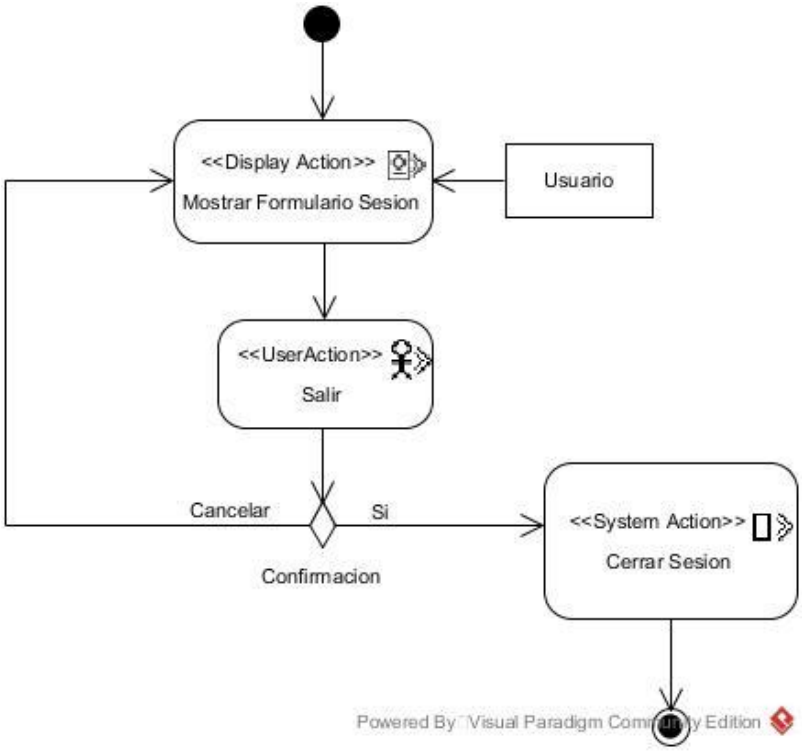


Figura 3.32 Diagrama de actividad del caso de uso *Cerrar Sesión*.

En la Figura 3.33 se describe el proceso de *Responder Encuesta*. Se indica que el caso de uso inicia mostrando en pantalla un formulario mediante el cual el usuario responde las preguntas de la encuesta, seguidamente el sistema pasa a validar los datos. Cuando los datos son incorrectos se le indica al usuario mediante un mensaje de error y se muestra nuevamente el formulario para que corrija su información, cuando los datos son validados correctamente el sistema guarda la nueva encuesta y lanza el caso de uso *Minar Texto*, encargado de clasificar las respuestas de las preguntas abiertas insertadas por el usuario.

El proceso para *Consultar Encuesta* se describe en la Figura 3.34. Comienza mostrando en pantalla el listado de las encuestas y un formulario mediante el cual el usuario proporciona el número de control de la encuesta requerida. Cuando no se obtienen resultados para el valor proporcionado por el usuario, el sistema mostrará un mensaje indicando la inexistencia de esa encuesta, el usuario debe confirmar y será enviado nuevamente al listado donde aparecen todas las encuestas. Si el filtro arroja resultados, el sistema mostrará todos los datos de la encuesta especificada.

Para eliminar una encuesta el sistema muestra en pantalla el listado de las encuestas existentes. Mediante un filtro el usuario proporciona el número de control de la encuesta que desea eliminar, si no existe la encuesta solicitada el sistema lanza un

mensaje indicando la ausencia de ésta, el usuario confirma y regresa al listado de las encuestas existentes. Cuando la encuesta es identificada, se le muestra al usuario toda su información. A partir de este momento el usuario seleccionará que desea eliminar la encuesta y el sistema le pedirá confirmación para proceder con la operación. Si el usuario cancela la operación es regresado al listado de las encuestas existentes y en caso contrario el sistema eliminará de forma permanente la encuesta. Este proceso se muestra en la Figura 3.35.

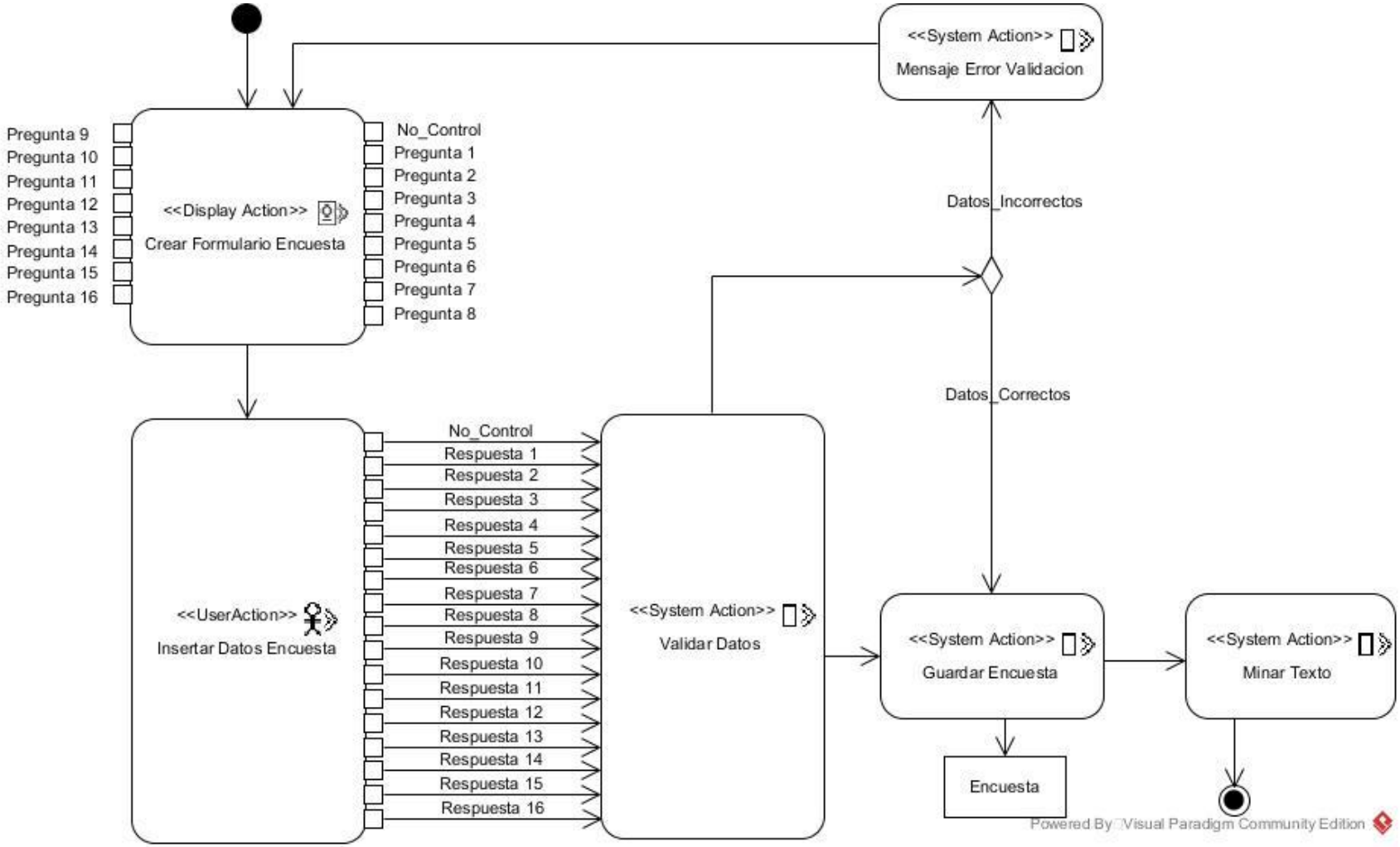


Figura 3.33 Diagrama de actividad del caso de uso *Responder Encuesta*.

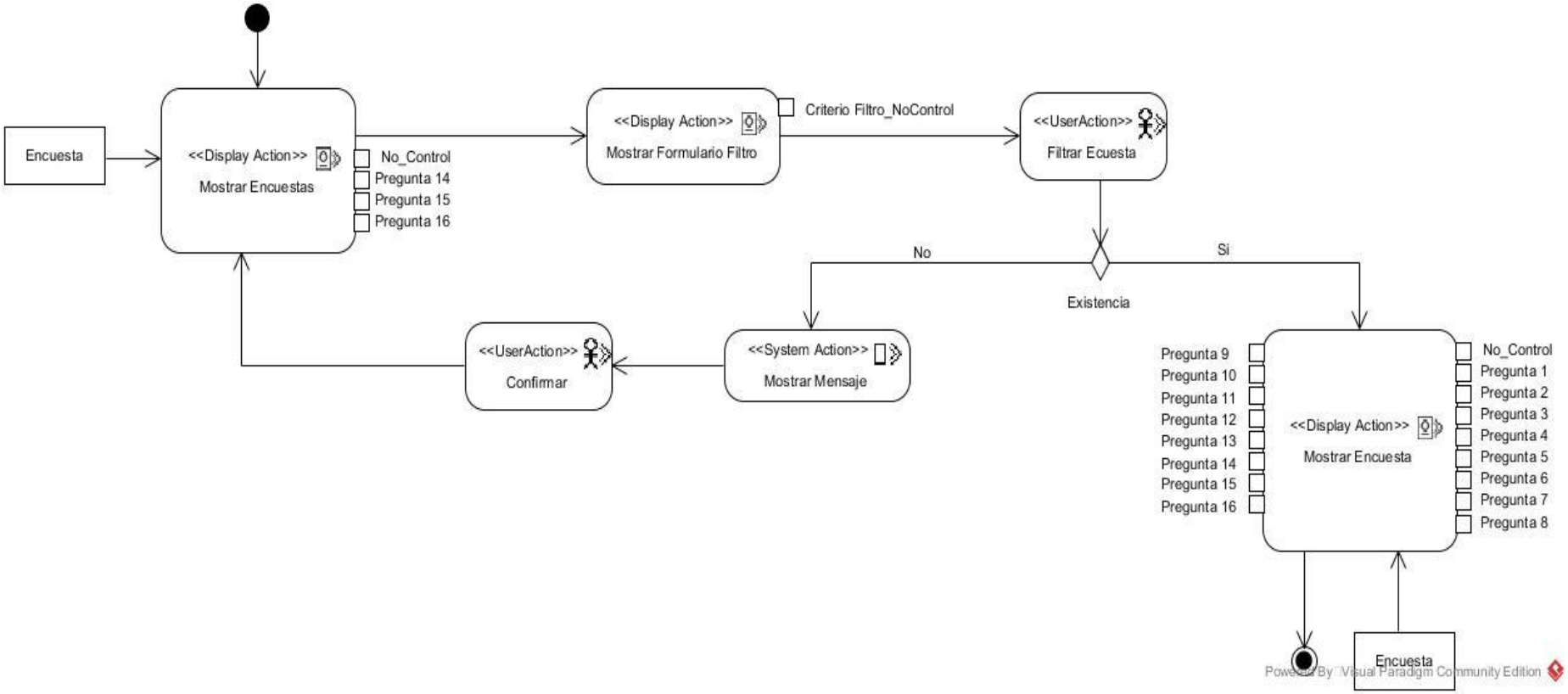


Figura 3.34 Diagrama de actividad del caso de uso *Consultar Encuesta*.

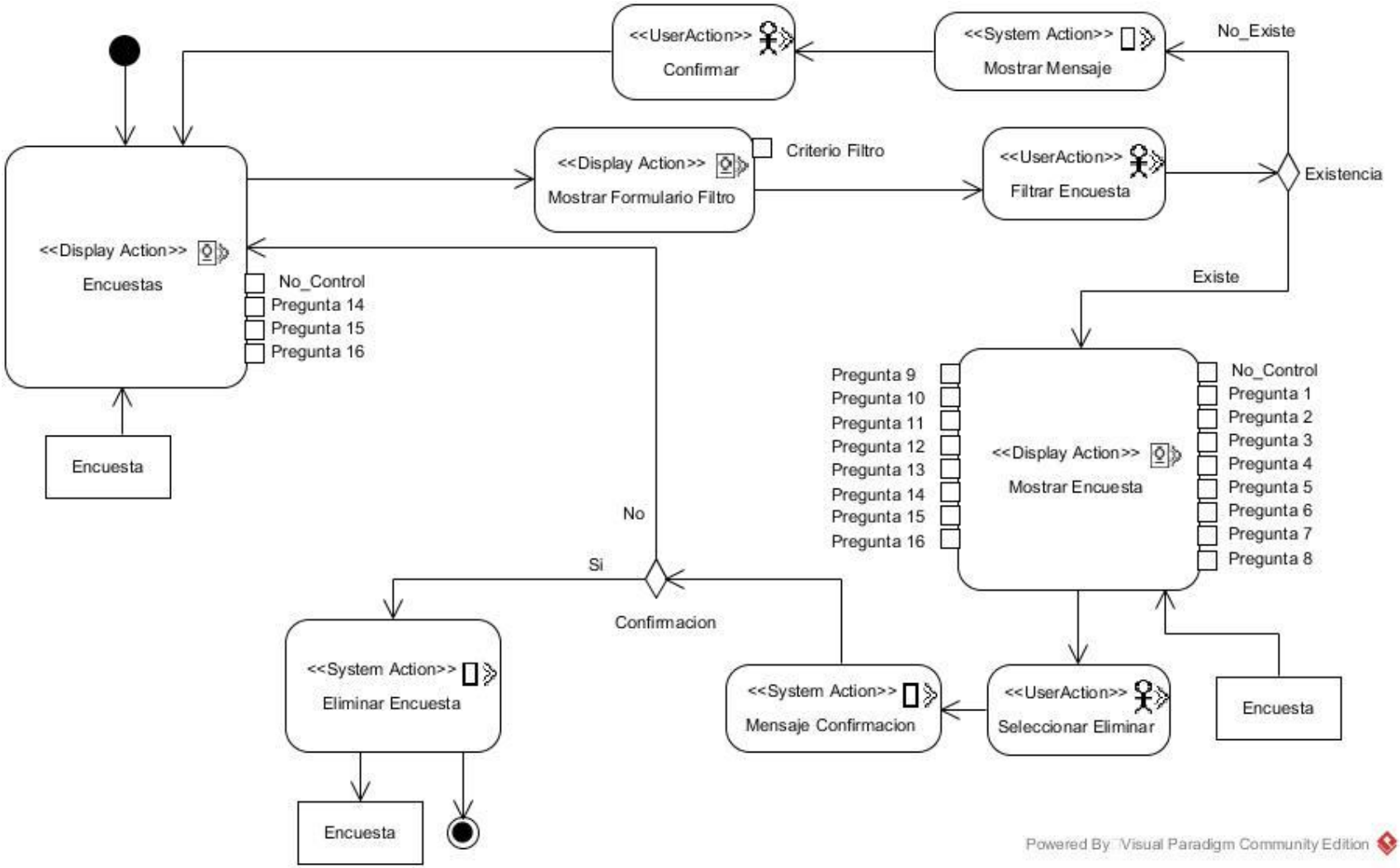


Figura 3.35 Diagrama de actividad del caso de uso *Eliminar Encuesta.*

En la Figura 3.36 se muestra el proceso para el caso de uso *Generar Modelo*. Éste inicia cuando se muestran al especialista los resultados del modelo existente. El especialista selecciona el algoritmo de minería que desea aplicar y el sistema generará el modelo correspondiente. En caso de fallo en la creación del modelo el sistema enviará un mensaje de error y regresará al escenario donde comienza el caso de uso, sin embargo, si la operación termina con éxito, el modelo es guardado físicamente en la aplicación.

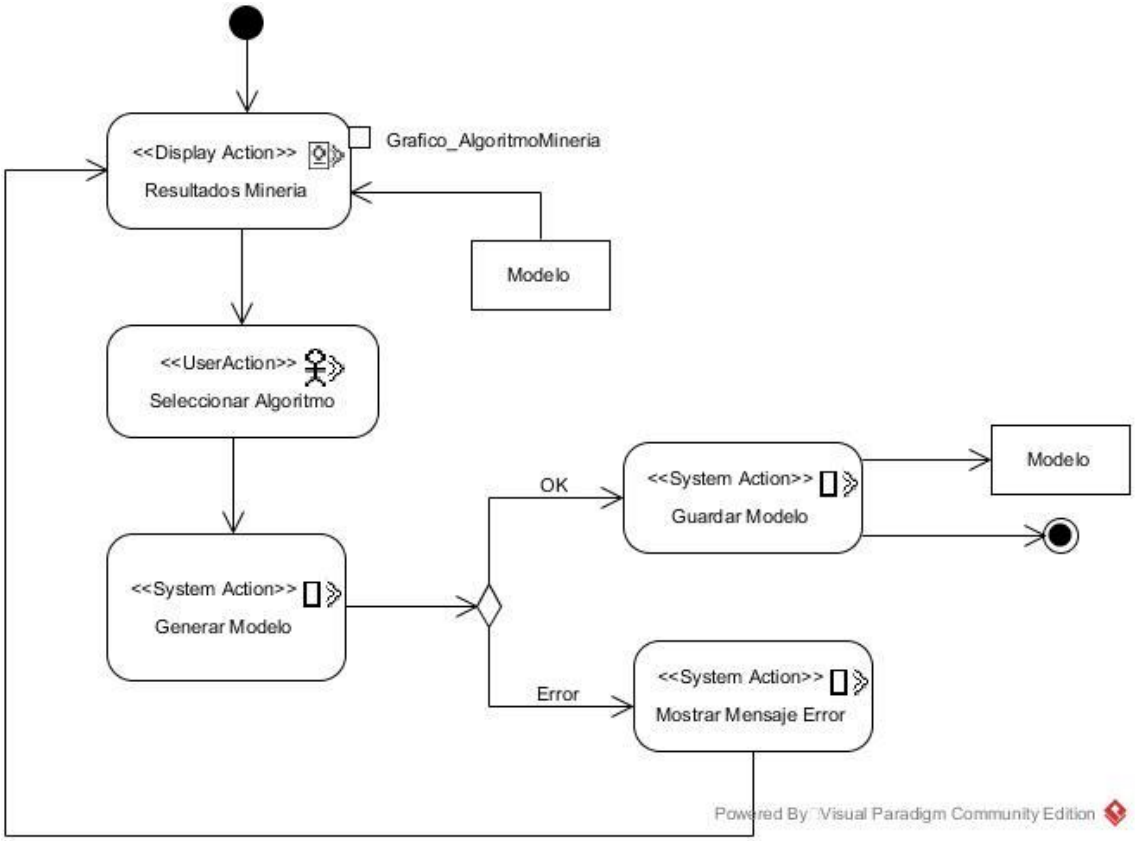


Figura 3.36 Diagrama de actividad del caso de uso *Generar Modelo*.

El proceso para interpretar los resultados de los modelos se describe en la Figura 3.37, donde se indica que inicia cuando el sistema comienza a interpretar el modelo. En caso de error el sistema notifica al usuario y regresa a interpretar el modelo una vez más. Cuando lo interpreta exitosamente muestra los resultados en pantalla.

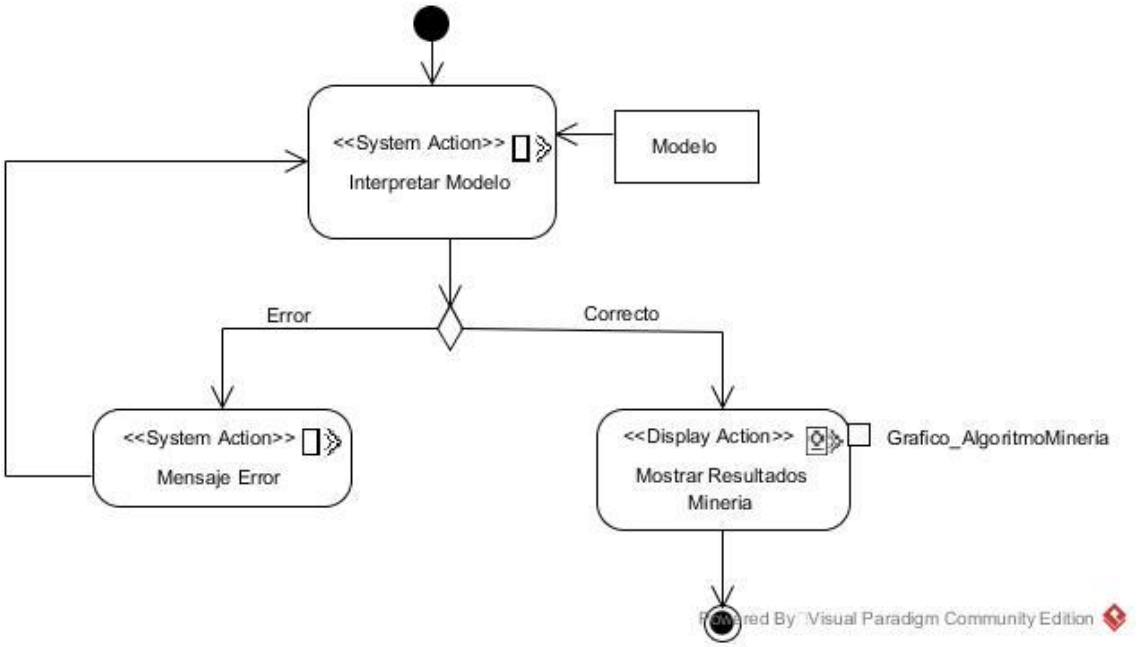


Figura 3.37 Diagrama de actividad del caso de uso *Interpretar Resultados de Minería*.

En la Figura 3.38 se describe el proceso para insertar un nuevo usuario. Comienza mostrando en pantalla un formulario mediante el cual el administrador proporciona la información del nuevo usuario. El sistema validará los datos y si detecta algún error en estos enviará un mensaje de error y regresará al administrador al formulario para que corrija la información. Cuando los datos son correctamente validados, el sistema registra al nuevo usuario en la aplicación.

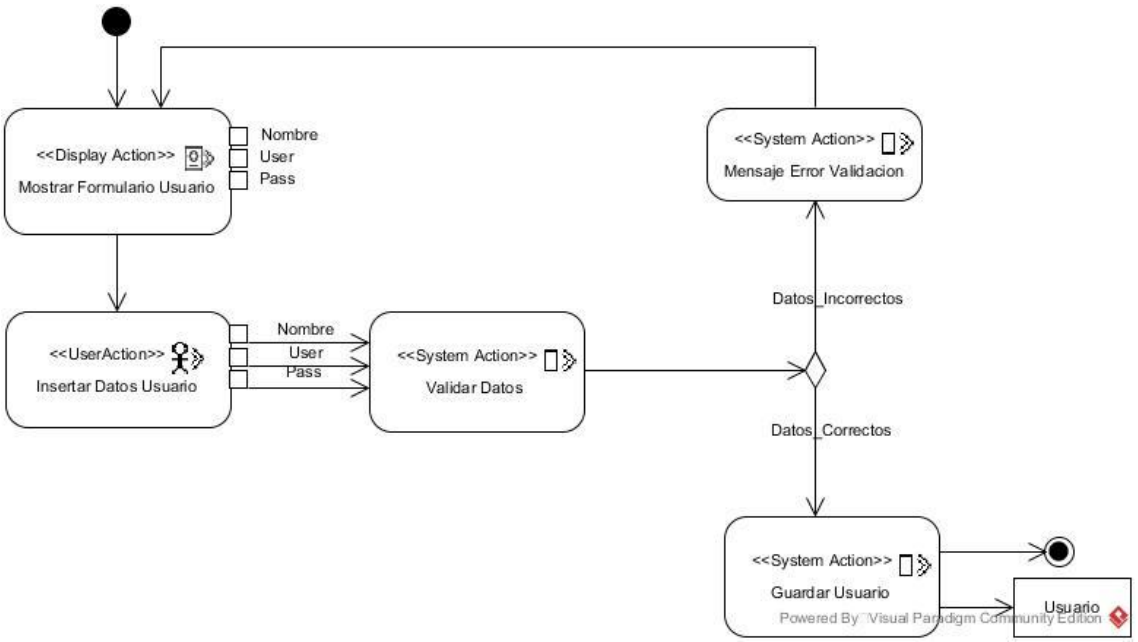


Figura 3.38 Diagrama de actividad del caso de uso *Insertar Usuario*.

El procedimiento que sigue un administrador para consultar un usuario se describe en la Figura 3.39. Se parte desde un listado de usuarios registrados en la aplicación. El administrador, mediante la opción de filtrar, especifica el usuario que quiere consultar. Si no existe, el sistema lanza un mensaje de error; el usuario confirma entonces y es dirigido al listado de los usuarios. En caso de encontrar con éxito al requerido, el sistema muestra todos sus datos.

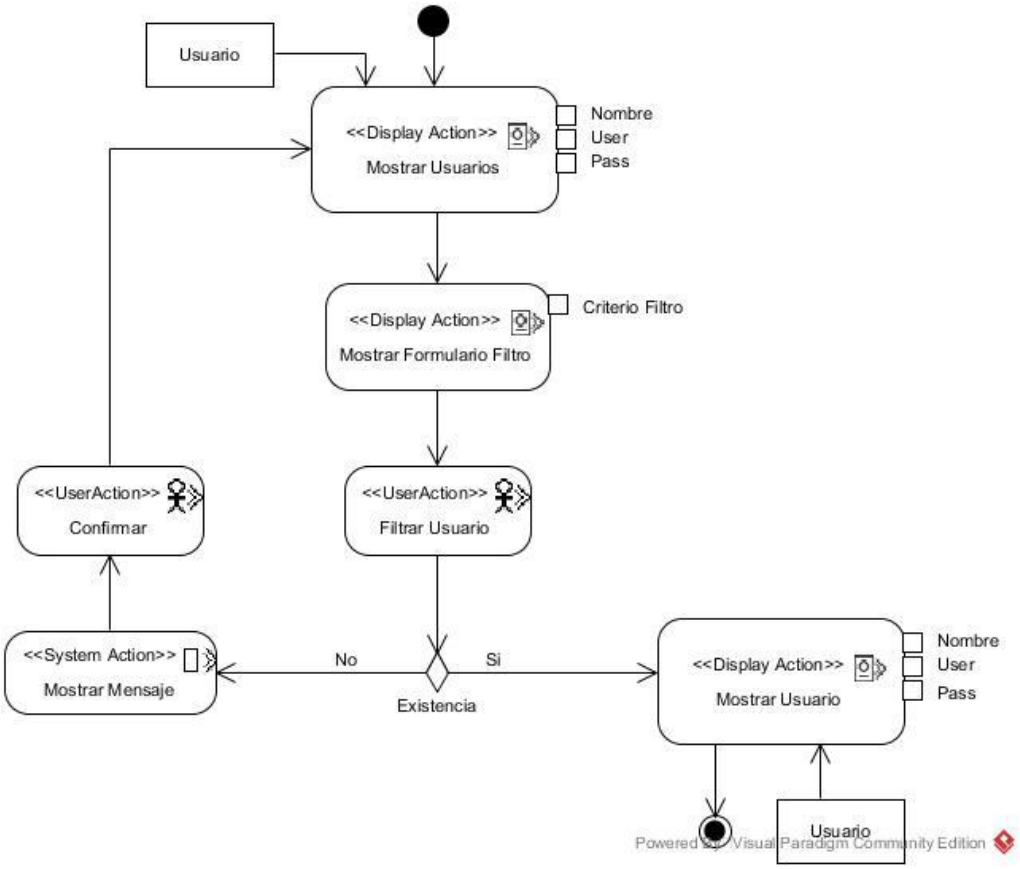


Figura 3.39 Diagrama de actividad del caso de uso *Consultar Usuario*.

En la Figura 3.40 se muestra el procedimiento para actualizar los datos de un usuario. Inicia mostrando en pantalla un listado de los usuarios registrados en la aplicación. El administrador, mediante la opción de filtrar, especifica el usuario que quiere actualizar. Si no existe, el sistema lanza un mensaje de error para que el administrador confirme y sea redirigido al listado de los usuarios. En caso de encontrar con éxito al usuario requerido, el sistema muestra todos sus datos y permite al administrador hacer las modificaciones pertinentes para luego validarlas. De generarse algún error de validación, se regresa al formulario del usuario. En caso contrario, se actualiza el usuario con la nueva información.

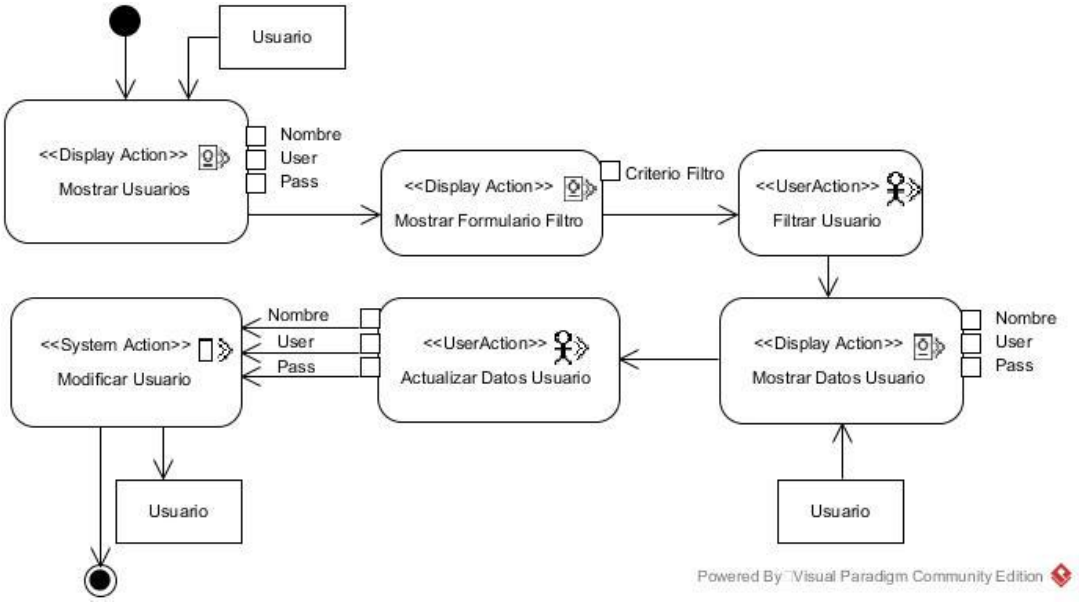


Figura 3.40 Diagrama de actividad del caso de uso *Actualizar Usuario*.

El proceso para eliminar un usuario inicia al presentar en pantalla un listado de los usuarios registrados en la aplicación. El administrador, mediante la opción de filtrar, especifica el usuario que quiere eliminar. Si no existe, el sistema lanza un mensaje indicando la inexistencia del objeto buscado. Seguidamente el usuario confirma y es dirigido al listado de los usuarios. En caso de encontrar con éxito al objeto demandado, el sistema muestra todos sus datos y el administrador selecciona la opción de eliminar usuario. El sistema pide confirmación para continuar con la operación; si se cancela, el sistema regresa a mostrar los datos del usuario y si se confirma, elimina permanentemente al usuario seleccionado. Este proceso se muestra en la Figura 3.41.

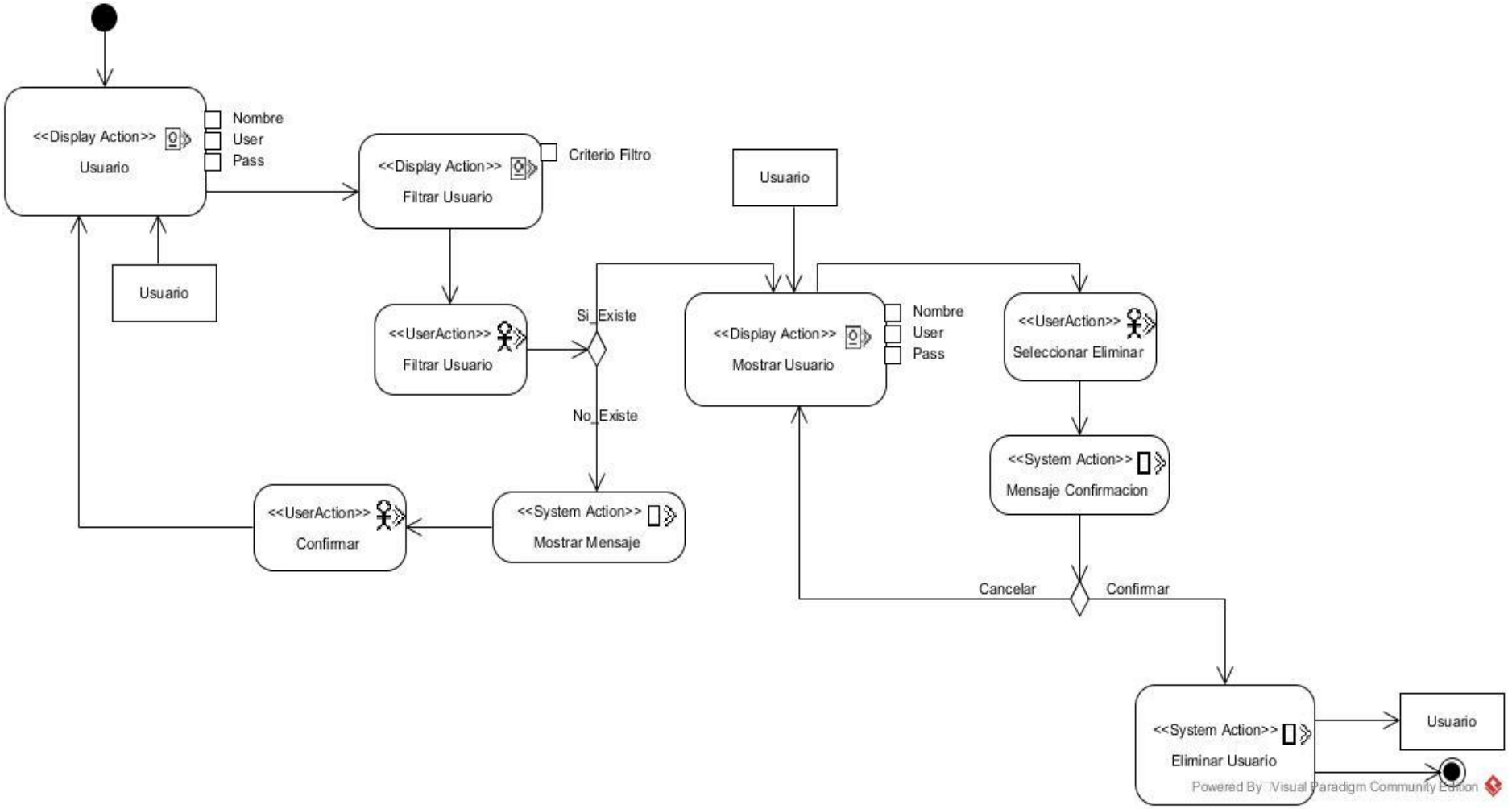


Figura 3.41 Diagrama de actividad del caso de uso *Eliminar Usuario*.

3.2. Desarrollo de los modelos de minería de datos utilizando KDD

En este apartado se describe cómo se aplicó la minería de datos con el objetivo de encontrar conocimiento útil, que sirviera de base para identificar los motivos que conducen a que en el H.R.R.B no se realicen autopsias. Este proceso fue guiado por la metodología KDD, los pasos que define están ilustrados en la Figura 3.42.

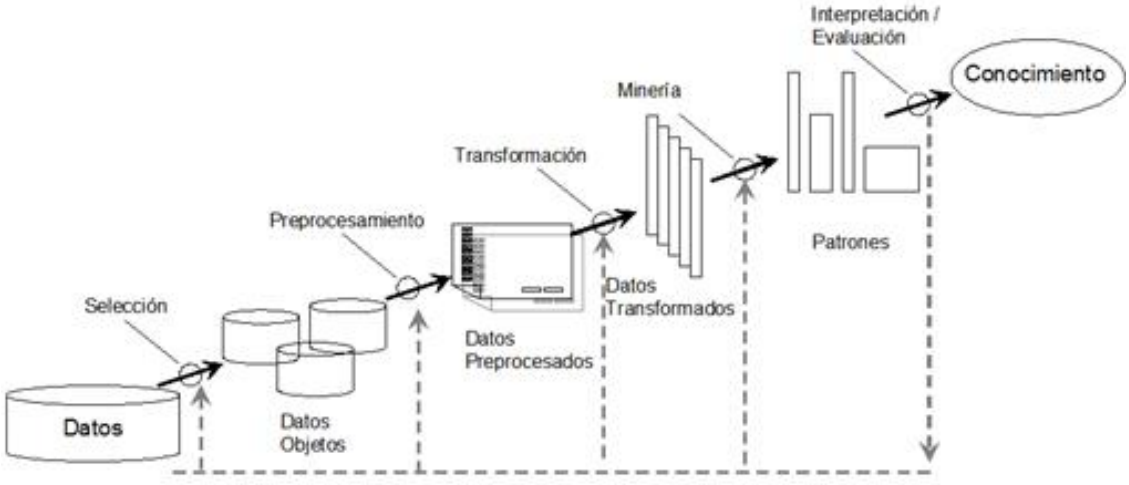


Figura 3.42 Proceso de descubrimiento del conocimiento.

3.2.1. Selección de Datos

Cuando se identificó la problemática que aborda esta investigación, se hizo evidente que no existía una fuente de información que permitiera investigar de forma directa y específica los motivos por los que no se solicitan las autopsias en el hospital de Río Blanco. Por ello, se consideró utilizar una encuesta como técnica de investigación social cuyo objetivo era fundamentalmente indagar sobre la opinión, actitudes o comportamientos de los médicos ante la práctica de autopsias, así como los valores, creencias o motivos que los caracterizan. El cuestionario contiene un total de dieciséis preguntas, divididas en tres de tipo abierta y trece de tipo cerrada, de las cuales cinco incluyen una sección para especificar otras respuestas consideradas por los encuestados.

A continuación, se presenta un resumen de la información identificada de acuerdo al instrumento:

• 27 categorías relacionadas con factores que los médicos consideran negativos para la realización de autopsias y 26 más para los factores positivos.

• Nueve motivos por los que el familiar no solicita el estudio y ocho posibles razones por las que no se realizan suficientes autopsias en el hospital.

• Respecto a la opinión de los médicos sobre el procedimiento para solicitar una autopsia, se consideraron 14 métodos eficientes y seis opciones sobre personal adecuado para solicitarla.

• Las respuestas de comentarios generales que dieron los médicos se consideraron en 25 categorías.

• Tres posibles respuestas para el área y el grado del médico y cinco para cada una de las tres preguntas relacionadas con la opinión de los médicos sobre los hallazgos encontrados en las autopsias.

La Tabla 3.2 muestra un resumen de la encuesta aplicada y la cantidad de categorías generadas por respuesta.

Tabla 3.2 Resumen de la encuesta aplicada al personal médico

Aspectos	Preguntas	Tipo de Pregunta	Categorías generadas
Formación del médico	Área	Cerrada	3
	Grado	Cerrada	3
	Centro Formación Medicina General	Cerrada	47
	Centro Formación Especialidad	Cerrada	47
Experiencia del médico	Años de práctica	Cerrada	5
	Casos de autopsias en los que ha participado	Cerrada	5
Opinión del médico sobre los hallazgos de autopsias	Originan discrepancias con los diagnósticos clínicos	Cerrada	5
	Originan casos de demandas	Cerrada	5
	Originan casos de arbitraje	Cerrada	5
Opinión del encuestado ante la solicitud de autopsias	Motivos para que el médico la solicite	Abierta	26
	Motivos para que el médico no la solicite	Abierta	27
	Motivos para que el familiar no la solicite	Cerrada	9
	Motivos por los que en el hospital no se realizan suficientes autopsias	Cerrada	8
Opinión del médico sobre el procedimiento para solicitar una autopsia	Personal adecuado para solicitar autopsias	Cerrada	6
	Método eficiente para solicitar autopsias	Cerrada	14
Aspecto General	Comentarios	Abierta	25

En la Figura 3.43 se presenta el porcentaje de preguntas que corresponde a cada uno de los aspectos que explora la encuesta aplicada.

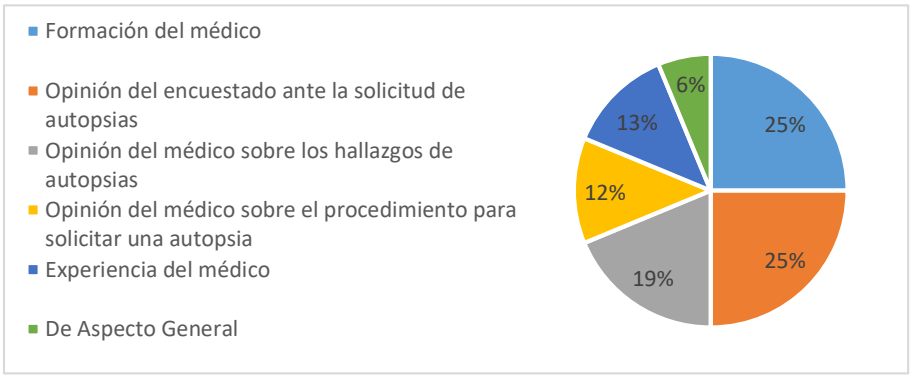


Figura 3.43 Principales áreas exploradas por la encuesta.

Algunos de los aspectos de interés para los objetivos de la investigación se muestran de forma gráfica en las Figuras 3.44 y 3.45. Estos se encuentran relacionados específicamente con la aplicación de la encuesta y la información obtenida mediante ella.

Los resultados del instrumento detallan que la mayor parte de los médicos integrantes de la muestra del estudio, en la que la mitad de los galenos son especialistas y menos del 20 por ciento se clasifican como practicantes, han realizado menos de cinco autopsias. Incluso, más del 30 por ciento de los encuestados no ha realizado ninguna. Ello es evidencia de la problemática que se aborda en este estudio.

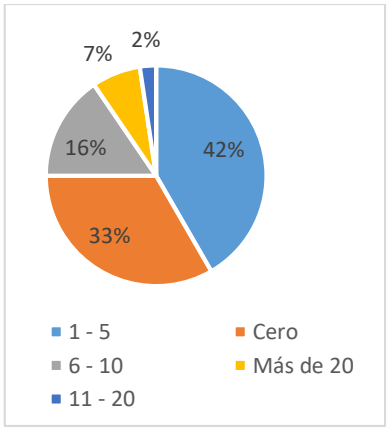


Figura 3.44 Intervenciones de los médicos en casos de autopsia.

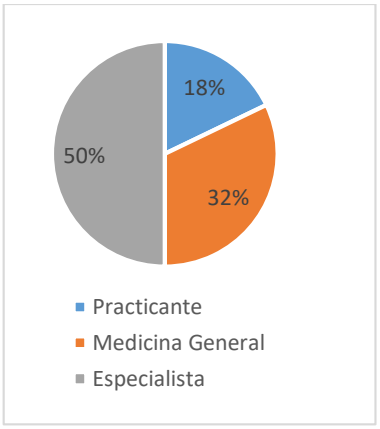


Figura 3.45 Nivel de escolaridad de los médicos.

La encuesta se desarrolló por uno de los especialistas patólogos del hospital, el Dr. José Antonio Palet, y fue posible aplicarla a 86 médicos del hospital. La información obtenida se registró en una base de datos relacional (ver Figura 3.17) para garantizar la

persistencia de estos datos y poder utilizarlos en posteriores análisis. Esta base de datos es la única fuente de datos con la que cuenta la investigación.

3.2.2. Pre-procesamiento de Datos

Las respuestas de los médicos se transformaron en dos representaciones adecuadas (matriz_binaria y vista_minable) para aplicar técnicas de minería de datos, como ilustra la Figura 3.46. Estas estructuras se crearon mediante funciones SQL (Structured Query Language, Lenguaje de Consulta Estructurado) y de esta manera se conformaron dos conjuntos de datos distintos a partir de los mismos datos. En este trabajo se nombrará a la matriz-binaria como ‘C’ y a la vista-minable como ‘D’.

También se crearon los conjuntos de datos *mcc_aut*, *mcc_no_aut* y *com_sug_op* que se utilizarán para la clasificación de las respuestas de los encuestados de las preguntas abiertas como son: motivos para solicitar autopsias, motivos para no solicitarlas y comentarios, respectivamente. Estos conjuntos contienen las respuestas de los encuestados a las preguntas anteriormente mencionadas y se consideraron como valores para el atributo de clase las categorías seleccionadas *a priori* de conjunto con el experto.

Vista_minable: Tabla que se genera dinámicamente en dependencia de los datos registrados mediante una función SQL. La función construye una matriz donde las filas significan las combinaciones de respuestas para las encuestas y las columnas representan las respuestas. El valor de cada columna responde a la intersección que se lee como un par <pregunta, respuesta>.

	area character var	hall_disc character var	hall_arb character var	hall_dem character var	mcc_aut character var	mcc_no_aut character var	rechazo_fam character var	no_hosp character var	...
1	a1	7b	8b	11b	14z	16z	17a	18b	19
2	a1	7b	8b	11b	14z	16z	17a	18b	19
3	a1	7b	8b	11b	14z	16z	17a	18c	19
4	a1	7b	8b	11b	14z	16z	17a	18c	19
5	a3	7b	8b	11b	14z	16z	17c	18c	19
6	a1	7c	8c	11d	14z	16z	17d	18a	19
7	a3	7b	8b	11a	14z	16z	17d	18b	19
8	a3	7b	8b	11a	14z	16z	17d	18b	19
9	a3	7b	8b	11a	14z	16z	17d	18b	19
10	a3	7b	8b	11a	14z	16z	17d	18b	19
11	a3	7b	8b	11a	14z	16z	17d	18d	19
12	a3	7b	8b	11a	14z	16z	17d	18d	19
13	a3	7b	8b	11a	14z	16z	17d	18d	19
14	a3	7b	8b	11a	14z	16z	17d	18d	19
15	a3	7b	8b	11a	14z	16z			
16	a3	7b	8b	11a	14z	16z			
17	a3	7b	8b	11a	14z	16z			
18	a3	7b	8b	11a	14z	16z			
19	a3	7b	8b	11a	14z	16z			
20	a3	7b	8b	11a	14z	16z			
21	a3	7b	8b	11a	14z	16z			
22	a3	7b	8b	11a	14z	16z			
23	a3	7b	8b	11a	14z	16z			
24	a3	7b	8b	11a	14z	16z			
...	a3	7b	8b	11a	14z	16z			

Matriz_binaria: Tabla que se genera dinámicamente en dependencia de los datos registrados mediante una función SQL. La función construye una matriz binaria, en la que cada fila representa un encuestado y las columnas representan las respuestas. El valor de cada columna responde a la intersección que se lee como un par <respuesta, valor>, siendo valor igual a 'S' si se respondió dicha respuesta y '' en caso contrario.

	p1_1 character var	p2_1 character var	p3_2 character var	p11_1 character var	p12_1 character var	p14_8 character var	p15_1 character var	p16_8 character var	..
1	S	S	S	S	S	S	S	S	S
2			S					S	
3	S			S	S	S	S		S
4			S	S	S			S	
5				S				S	
6	S			S		S	S	S	S
7		S	S	S				S	
8				S	S				
9				S	S			S	
10			S					S	
11		S	S	S	S				
12			S	S					
13		S		S	S			S	
14		S		S					
15	S	S		S		S	S	S	S
16	S	S		S		S	S	S	S
17	S	S	S			S	S	S	S
18				S					
19	S		S	S					S
20		S	S		S			S	
21				S	S				
22			S	S	S			S	
23				S				S	
24				S					
...	S			S		S	S		S

Figura 3.46 Tablas *matriz_binaria* y *vista_minable.*

Tabla 3.3 Características de los conjuntos de datos C y D.

Características	Conjunto C	Conjunto D
Atributos	166	18
Objetos	4	7859
Tipo de datos	Nominal-Binarios- Asimétricos	Nominal
Descripción	Matriz binaria <respuesta, valor>	Matriz representada por <pregunta, respuesta>
Valores faltantes	Si	No
Valores fuera de rango	No	No
Valores inconsistentes	No	No

3.2.3. Minería de Datos

En la fase de transformación se analizaron las características de los conjuntos de datos y se determinó que no era necesario realizar ninguna transformación, por lo que se pasó directamente a la fase de minería de datos.

De acuerdo a los datos y al objetivo de este trabajo se consideraron dos tareas de minería para atacar el problema. Se pensó en primer lugar realizar un análisis de asociación para determinar las relaciones entre los atributos, y por otro lado, utilizar las redes Bayesianas para reconocer las dependencias relevantes entre los atributos, de acuerdo a la probabilidad y la estadística.

3.2.3.1. Análisis de asociación

Se estudiaron cada uno de los algoritmos que propone Weka y se consideró la posibilidad de su aplicación en los conjuntos *C* y *D* (ver Tabla 3.4). A continuación, se detallan los aspectos de cada uno de estos.

Apriori [42]: Es un algoritmo clásico para el aprendizaje de reglas de asociación, el cual genera reglas mediante un proceso incremental que realiza búsquedas de relaciones frecuentes entre atributos acotado por una confianza mínima. El algoritmo es configurado para que se ejecute bajo ciertos criterios como son límite superior e inferior de cobertura para aceptar conjuntos de elementos que cumplan con la restricción, la confianza mínima, un criterio de

ordenamiento para mostrar las reglas, así como un parámetro para indicar la cantidad específica de reglas que se quiere obtener.

FPgrowth [43]**:** Se basa en *Apriori* para realizar la primera exploración de los datos donde identifica los conjuntos de ítems frecuentes y su soporte, valor que permite organizar los conjuntos de manera descendente. El método propone una buena selectividad y reduce sustancialmente el coste de la búsqueda, ya que inicia buscando los patrones frecuentes más cortos para luego concatenarlos con los menos frecuentes (sufijos), y así identificar los patrones frecuentes más largos. Se demostró que es aproximadamente un orden de magnitud más rápido que el algoritmo *Apriori*.

PredictiveApriori [44]: El algoritmo alcanza un favorable rendimiento computacional debido a su técnica de poda dinámica que utiliza el límite superior de todas las reglas de los superconjuntos de un conjunto de elementos dado. Además, mediante un sesgo hacia atrás de las reglas, consigue eliminar aquellas redundantes que se derivan de las más generales. Para este algoritmo solo se necesita especificar el número de reglas que se requieren.

Tertius [45]**:** Realiza una búsqueda óptima basada en encontrar las *k* hipótesis más confirmadas (interesantes) haciendo uso de un operador de refinamiento no redundante para eliminar resultados duplicados. El algoritmo cuenta con una serie de parámetros de configuración que posibilitan su aplicación a múltiples dominios.

Tabla 3.4 Aplicación de los algoritmos en los conjuntos de datos.

Conjunto de datos	Apriori	FPGrowth	PredictiveApriori	Tertius
C	√	√	√	√
D	√	✕	✕	√

Seguidamente se describen los resultados del experimento realizado para determinar la eficiencia de estos algoritmos en cada conjunto de datos, en las Tablas 3.5 y 3.6. Cada prueba fue ejecutada 100 veces para estimar valores promedios en cuanto a tiempo, soporte y confianza.

Tabla 3.5 Resultados de las pruebas para *Apriori* y *FPGrowth.*

Algoritmos	**Conf / Sop**	**Reglas**	**Tiempo**	**Confianza**	**Soporte**
Apriori	0.9 / 0.5	9	11	0.93	0.57
FPGrowth		9	4	0.93	0.57
Apriori	0.9 / 0.5	11	9	0.92	0.58
FPGrowth		11	5	0.92	0.58
Apriori	0.8 / 0.6	2	9	0.88	0.76
FPGrowth		2	3	0.88	0.76
Apriori	0.9 / 0.4	25	15	0.93	0.49
FPGrowth		24	8	0.93	0.49
Apriori	0.9 / 0.3	87	21	0.95	0.38
FPGrowth		78	12	0.94	0.39
Apriori	0.8 / 0.3	167	20	0.90	0.37
FPGrowth		140	12	0.91	0.39
Apriori	0.8 / 0.4	48	13	0.89	0.47
FPGrowth		47	8	0.90	0.47
Apriori	0.9 / 0.2	568	45	0.95	0.25
FPGrowth		528	28	0.96	0.26
Apriori	0.8 / 0.2	985	45	0.91	0.25
FPGrowth		961	26	0.91	0.25
Apriori	0.9 / 0.1	16463	285	0.97	0.12
FPGrowth		9349	108	0.97	0.13
Apriori	0.7 / 0.6	2	9	0.88	0.76
FPGrowth		2	4	0.88	0.76
Apriori	0.7 / 0.5	12	11	0.90	0.58
FPGrowth		12	4	0.90	0.58

Tabla 3.6 Resultados de las pruebas de *PredictiveApriori* y *Tertius*.

Algoritmos	Soporte	Reglas	Tiempo	Confianza	Soporte
PredictiveApriori	0.5	9	4050	0.78	0.29
Tertius		9	97	-	0.25
PredictiveApriori	0.5	11	4136	0.82	0.28
Tertius		11	100	-	0.26
PredictiveApriori	0.6	2	2794	1	0.33
Tertius		2	83	-	0.49
PredictiveApriori	0.6	2	2877	1	0.33
Tertius		2	99	-	0.22
PredictiveApriori	0.5	12	4109	0.83	0.27
Tertius		12	97	-	0.27

Las Figuras 3.47, 3.48, 3.49 y 3.50, que aparecen a continuación, representan de forma gráfica estos resultados.

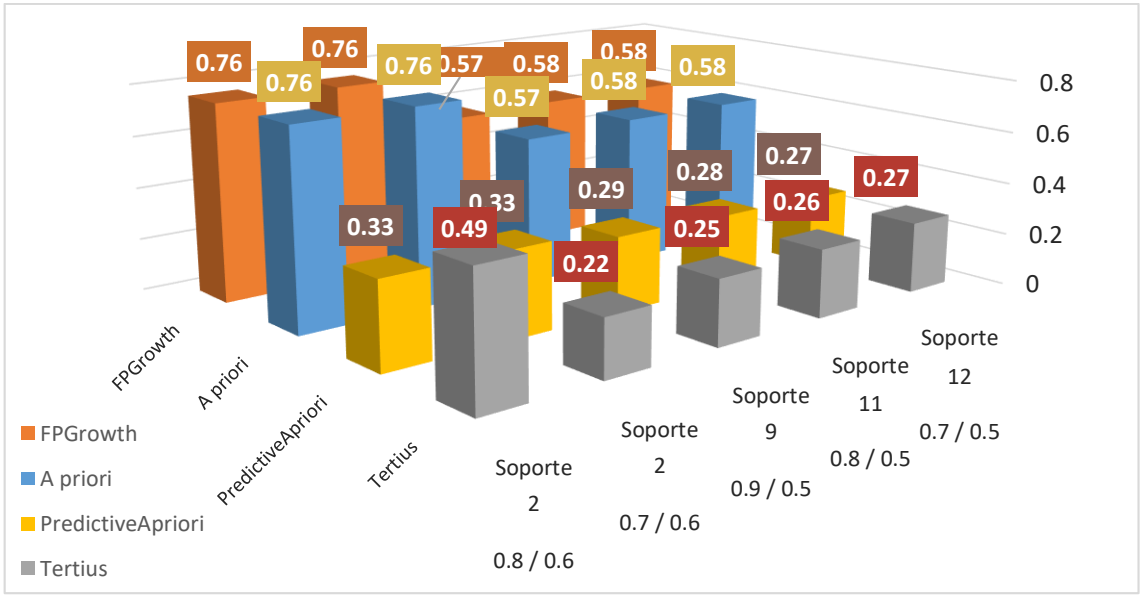


Figura 3.47 Comparación de los algoritmos Apriori, FPGrowth, PredictiveApriori y Tertius en cuanto a soporte.

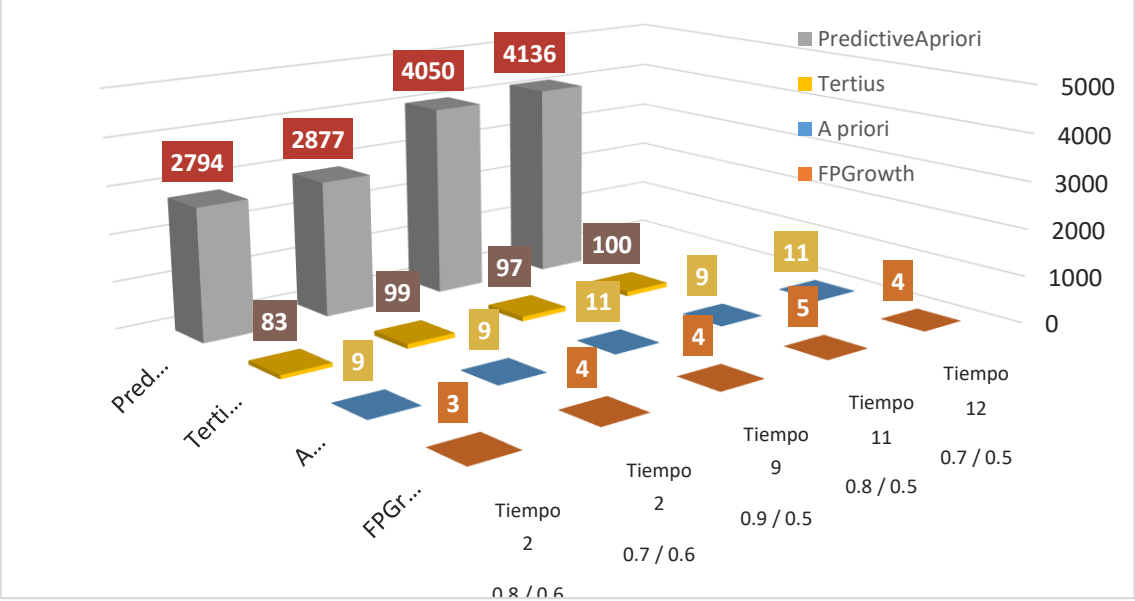


Figura 3.48 Comparación de los algoritmos Apriori, FPGrowth, PredictiveApriori y Tertius en cuanto a tiempo.

Observaciones: Los resultados indican que los algoritmos que generan reglas con mejor frecuencia dentro del conjunto de datos son Apriori y FPGrowth, también se evidencia que son computacionalmente más rápidos.

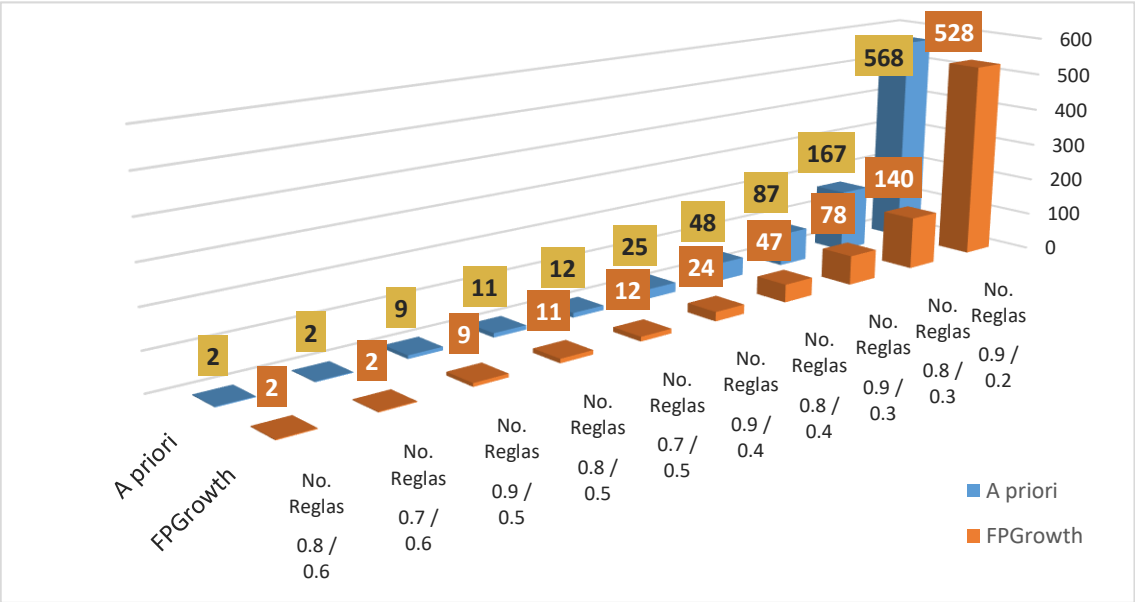


Figura 3.49 Comparación de los algoritmos Apriori y FPGrowth en cuanto a número de reglas.

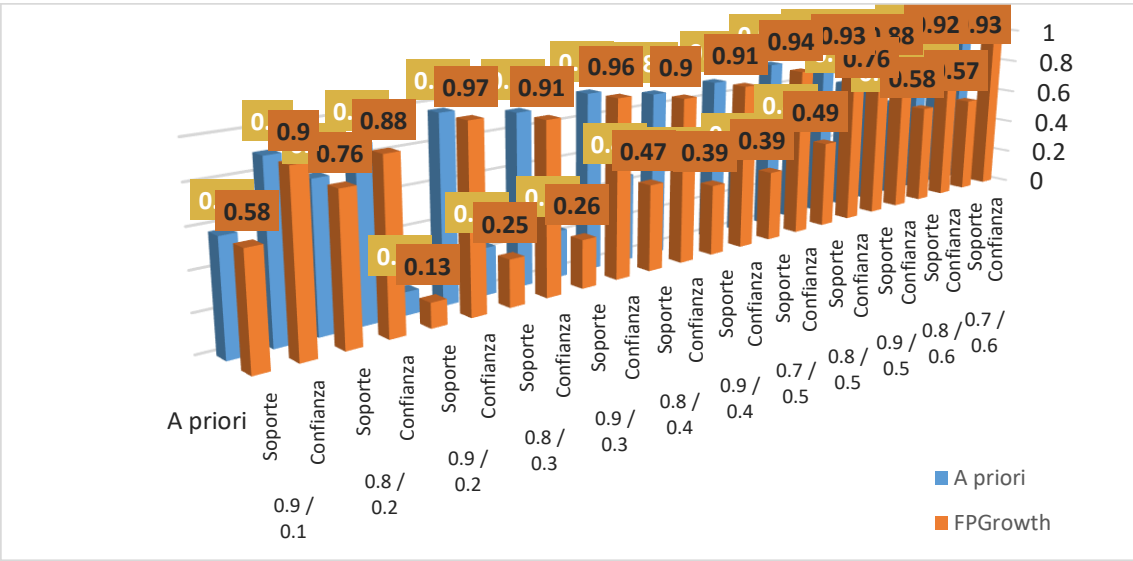


Figura 3.50 Comparación de los algoritmos Apriori y FPGrowth en cuanto a soporte y confianza.

Observaciones: Se evidencia gran similitud entre los resultados de ambos algoritmos, aunque cabe destacar que *Apriori* genera mayor cantidad de reglas que *FPGrowth* y que este último supera a *Apriori* en cuestiones de tiempo, así como en los valores promedios de confianza y soporte.

Conjunto de datos D

La Tabla 3.7 contiene los resultados de *Tertius* y *Apriori* para el conjunto de datos 'D' y las Figuras 3.51, 3.52 y 3.53.

Tabla 3.7 Resultados de las pruebas para *Apriori* y *Tertius*.

Algoritmos	Conf / Sop	Reglas	Tiempo	Soporte
Apriori	0.8 / 0.6	5	112	0.65
Tertius		5	9514	0.41
Apriori	0.7 / 0.6	6	118	0.64
Tertius		6	7467	0.42
Apriori	0.9 / 0.5	10	149	0.56
Tertius		10	9529	0.33
Apriori	0.8 / 0.5	28	144	0.56
Tertius		28	9700	0.42
Apriori	0.7 / 0.5	36	147	0.56
Tertius		36	9471	0.42
Apriori	0.9 / 0.4	82	198	0.45
Tertius		82	10042	0.39

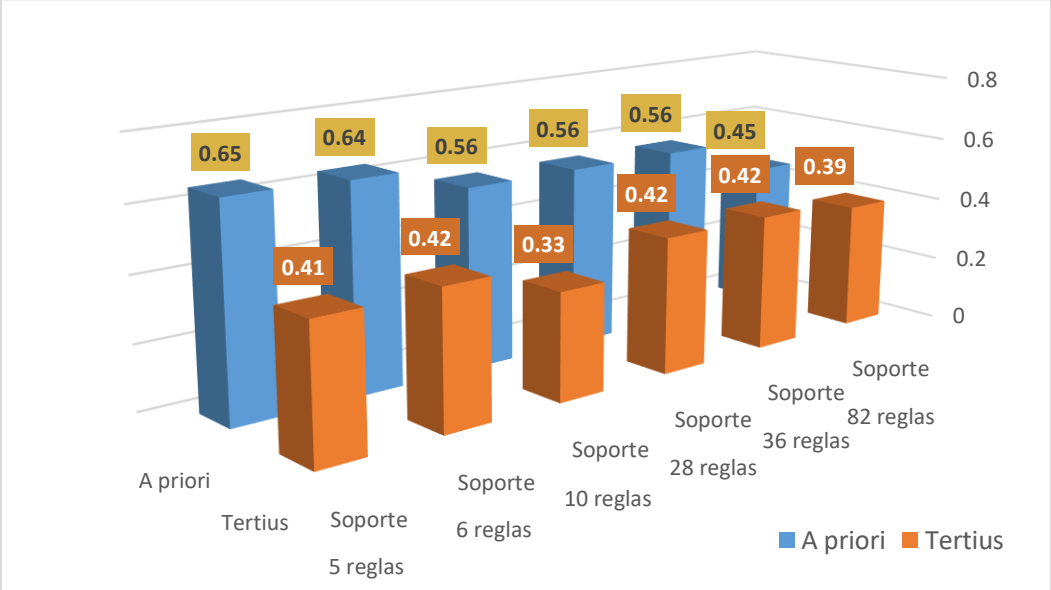


Figura 3.51 Comparación de los algoritmos *Apriori* y *Tertius* en cuanto a soporte.

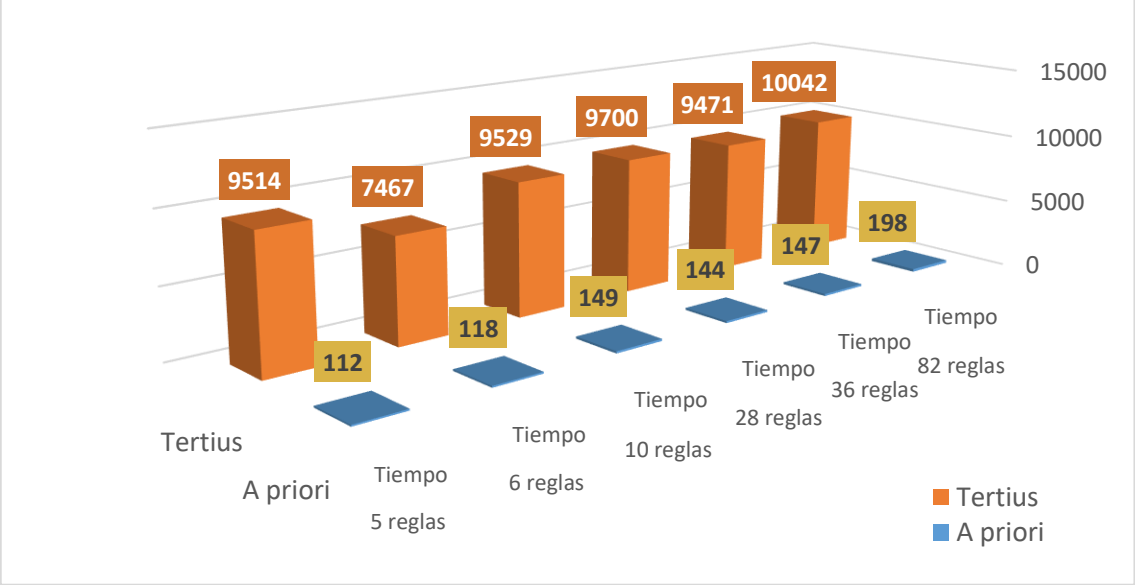


Figura 3.52 Comparación de los algoritmos *Apriori* y *Tertius* en cuanto a tiempo.

Observaciones: *Apriori* reporta mejores resultados que *Tertius*, fundamento que se sustenta porque el tiempo que gasta en resolver el mismo número de reglas es considerablemente más corto y las reglas que identifica tienen mejor soporte y confianza promedio.

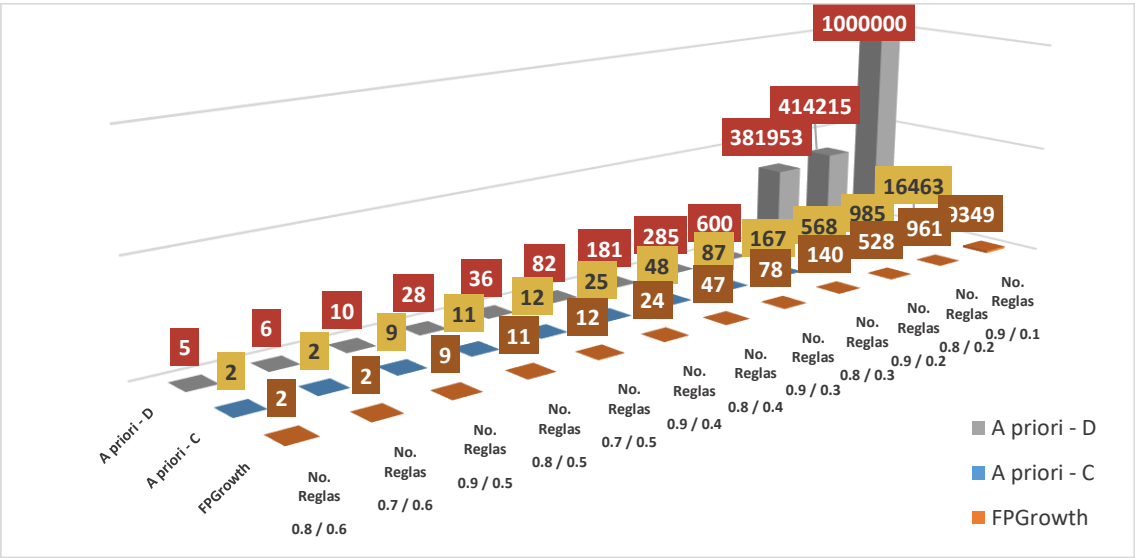


Figura 3.53 Comparación de los algoritmos Apriori en los conjuntos de datos C y D y Tertius para el conjunto C respecto a cantidad de reglas.

Observaciones: Se considera más adecuado para este trabajo el algoritmo *Apriori* dado que genera más reglas que *FPGrowth* y además permite analizar los dos conjuntos de datos concebidos en esta investigación.

3.2.3.2. Análisis de clasificación

Los algoritmos de clasificación también fueron evaluados. Las redes Bayesianas se consideraron para analizar los datos de las encuestas, en cambio, J48, Redes Neuronales, Naive Bayes y SMO se estudiaron considerando su aplicación en el proceso de clasificar las respuestas abiertas que proporcionan los usuarios.

***Redes bayesianas*:** Determinan las relaciones de dependencia e independencia probabilística entre todas las variables de un conjunto de datos, conformando así la estructura de la red Bayesiana presentada por un grafo acíclico donde los nodos son las variables y los arcos las dependencias probabilísticas entre los atributos enlazados [46], [47].

***J48*:** Construye un árbol de decisión binario para modelar el proceso de clasificación. Este algoritmo ignora los valores faltantes o predice estos en función a los valores conocidos del atributo en los demás registros [48], [49].

***Redes Neuronales*:** Son procedimientos matemáticos basados en la explotación del procesamiento local paralelo y las propiedades de la representación distribuida que imitan la estructura del sistema nervioso y se interpreta como la forma de obtener conocimiento a partir de la experiencia [50].

***Naive Bayes*:** Es un clasificador probabilístico que calcula las probabilidades en función de las combinaciones y frecuencias de ocurrencia de los datos en un conjunto de datos determinado [48] [49].

Optimización mínima secuencial (SMO): Implementa el algoritmo para entrenar máquinas de soporte vectorial (SVM) y resolver los problemas que éstas presuponen de programación cuadrática [51].

Se analizó la posibilidad de aplicar las redes Bayesianas en los conjuntos ‘C’ y ‘D’, la Tabla 3.8 muestra los resultados, y por el origen de las características de estos *“datasets”* solo es posible aplicar estas redes en el conjunto ‘D’.

Tabla 3.8 Aplicación de la red Bayesiana en los conjuntos de datos C y D.

Conjunto de Datos	Red Bayesianas
C	✗
D	√

Para evaluar las redes Bayesianas se consideraron los 18 atributos del conjunto de datos 'D' como clases. Cada prueba se ejecutó 100 veces para estimar el tiempo promedio de construcción de la red y se consideraron además los valores de precisión y Área ROC. La Tabla 3.9 muestra los resultados de las pruebas realizadas tomando como clase el grado del encuestado (*ult_grado*), de la misma manera se registraron los resultados de las 17 clases restantes, (ver anexos), se evidencia que los mejores resultados, ver Tabla 3.14, fueron obtenidos con los algoritmos de búsqueda Tan para 14 de las clases y HillClimber para las 4 restantes.

Tabla 3.9 Resultados de las redes Bayesianas para la clase *ult_grado.*

Clase: ult_grado	Precisión NetBeans	Área ROC NetBeans	Tiempo
K2	0,995	0,999	59
Tan	0,999	0.999	55411
TabuSearch	0,999	1	21571
RepeatedHillClimber	0,999	1	117975
LAGDHillClimber	0,993	0,999	310
HillClimber	0,999	1	10215

Observaciones: HillClimber para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Luego de estudiar las características de los conjuntos de datos *mcc_aut*, *mcc_no_aut* y *com_sug_op* y de los algoritmos J48, Redes Neuronales, NaiveBayes y SMO se determinó que es posible aplicar los cuatro algoritmos a los tres conjuntos como se ve en la Tabla 3.10.

Tabla 3.10 Aplicación de los algoritmos en los conjuntos de datos *mcc_aut, mcc_no_aut* y *com_sug_op*.

Conjunto de Datos	J48	Naive Bayes	Redes Neuronales	SMO
mcc_aut	√	√	√	√
mcc_no_aut	√	√	√	√
com_sug_op	√	√	√	√

La evaluación de los algoritmos J48, Redes Neuronales, NaiveBayes y SMO, contemplados para la clasificación de texto, se describe en las Tablas 3.11, 3.12 y 3.13 para los conjuntos de datos *mcc_aut*, *mcc_no_aut* y *com_sug_op*, respectivamente. Cada prueba fue ejecutada 100 veces para estimar valores promedios en cuanto a tiempo y además se consideraron otras métricas como precisión, área ROC, kappa e instancias sin clasificar.

Tabla 3.115 Resultados de las pruebas para la clase *mcc_aut.*.

Clase: *mcc_aut*		Precisión	Área ROC	Kappa	Instancias sin clasificar	Tiempo
J48	GainRatio	0,86	0,96	0,79	0	998
	InfoGain	0,86	0,97	0,80	0	1273
Naive Bayes	GainRatio	0,75	0,93	0,67	0	625
	InfoGain	0,75	0,93	0,67	0	647
Redes Neuronales	GainRatio	0,86	0,97	0,81	0	44611
	InfoGain	0,86	0,97	0,82	0	56820
SMO	GainRatio	0,83	0,91	0,75	0	1168
	InfoGain	0,83	0,91	0,75	0	560

Tabla 3.12 Resultados de las pruebas para la clase *mcc_no_aut.*

Clase: *mcc_no_aut*		Precisión	Área ROC	Kappa	Instancias sin clasificar	Tiempo
J48	GainRatio	0,78	0,97	0,74	0	1098
	InfoGain	0,78	0,97	0,74	0	1180
Naive Bayes	GainRatio	0,76	0,95	0,65	0	1279
	InfoGain	0,76	0,95	0,65	0	1244
Redes Neuronales	GainRatio	0,84	0,97	0,76	0	33674
	InfoGain	0,85	0,96	0,76	0	15904
SMO	GainRatio	0,81	0,94	0,73	0	995
	InfoGain	0,81	0,94	0,73	0	696

Tabla 3.13 Resultados de las pruebas para la clase *com_sug_op*.

Clase: *com_sug_op*		**Precisión**	**Área ROC**	**Kappa**	**Instancias sin clasificar**	**Tiempo**
J48	GainRatio	0,84	0,97	0,85	0	266
	InfoGain	0,84	0,97	0,85	0	326
Naive Bayes	GainRatio	0,84	0,97	0,76	0	332
	InfoGain	0,84	0,97	0,76	0	315
Redes Neuronales	GainRatio	0,90	0,97	0,88	0	14040
	InfoGain	0,92	0,98	0,90	0	25796
SMO	GainRatio	0,83	0,96	0,92	0	1473
	InfoGain	0,84	0,96	0,83	0	1208

En la Tabla 3.14 se presentan los mejores casos para cada uno de los algoritmos analizados de acuerdo a los diferentes conjuntos de datos. Esta información resulta útil para orientar al especialista sobre los parámetros con los que debe generar los modelos y así obtener resultados más precisos. Cabe destacar que solo se intenta proponer la mejor configuración para los algoritmos, pero independientemente de ello el especialista tiene la opción de configurarlos de acuerdo a sus intereses.

Tabla 3.14 Presentación de los mejores resultados de las evaluaciones de cada algoritmo.

Análisis de asociación		
Conjunto de Datos	**Algoritmo**	**Parámetros**
C	Apriori	Confianza = 0.9 Soporte = 0.4 Reglas = 15
	FPGrowth	Confianza = 0.9 Soporte = 0.4 Reglas = 8
	PredictiveApriori	Soporte = 0.5 Reglas = 12
	Tertius	Soporte = 0.5 Reglas = 20
D	Apriori	Confianza = 0.9 Soporte = 0.5 Reglas = 10
	Tertius	Soporte = 0.5 Reglas =10

Tabla 3.14 Presentación de los mejores resultados de las evaluaciones de cada algoritmo cont.

Análisis de clasificación			
Conjunto de Datos	**Algoritmo**	**Clase**	**Algoritmo Búsqueda**
D	Redes bayesianas	mcc_no_aut	Tan
		mcc_aut	HillClimber
		rechazo_fam	Tan
		anios_prac	Tan
		area	Tan
		casos	Tan
		categoria	Tan
		com_sug_op	Tan
		esc_esp	HillClimber
		esc_med_gral	Tan
		fmr_sol_aut	HillClimber
		hall_arb	Tan
		hall_dem	Tan
		hall_dis	Tan
		med_aut	Tan
		no_hosp	Tan
		per_sol_aut	Tan
		ult_grado	HillClimber
Conjunto de Datos	**Algoritmo**	**Medidas de selección**	
mcc_aut	Redes neuronales	InfoGain	
mcc_no_aut	Redes neuronales	InfoGain	
com_sug_op	Redes neuronales	InfoGain	

Capítulo 4. Resultados

Como parte de la investigación se desarrolló una aplicación web desde la cual los médicos contestarán la encuesta, lo que permite procesar la información de manera automatizada y mantener los datos actualizados. Esta aplicación además brinda un conjunto de operaciones de MD utilizando la API de Weka que permiten analizar los datos y extraer de ellos conocimiento útil y novedoso. La información generada por esos algoritmos es el resultado de las correlaciones entre las variables de interés manejadas en la encuesta y la probabilidad de que determinados patrones ocurran a partir del comportamiento de los datos entrenados. Este conocimiento generado es muy importante para el especialista porque está validado de manera objetiva mediante algoritmos estadísticos y probabilísticos y se complementa con su evaluación subjetiva sustentada por su experiencia y conocimientos del área de patología.

La aplicación permite analizar los datos mediante técnicas de asociación y de clasificación. Para el primer caso es posible seleccionar entre los algoritmos de Apriori, FPGrowth, Predictive Apriori y Tertius y para el segundo caso las Redes Bayesianas. Además, el sistema utiliza las Redes Neuronales para realizar una clasificación de texto para las preguntas abiertas en el momento que contestan una nueva encuesta. El especialista es capaz de generar cada uno de estos modelos y guardar de forma permanente el que considere más preciso. Los resultados que se muestran a los usuarios que visitan la aplicación son los leídos de los modelos guardados por el especialista. Los resultados de cada uno de estos algoritmos son interpretados y mostrados a los usuarios en un lenguaje natural, lo cual permite que puedan estudiar y analizar este conocimiento generado sin requerir de la intervención de un experto en MD.

En este capítulo se describe el funcionamiento de cada una de las operaciones de la aplicación, con base en el caso estudio del H.R.R.B. Así se ilustra cómo desde la aplicación un especialista es capaz de generar y guardar un modelo, así como eliminar y consultar encuestas. Por otro lado, se describe cómo el administrador para gestionar usuarios inserta nuevos, los edita, elimina y consulta. Por último, se describe cómo la aplicación muestra los resultados de los modelos de minería a todos los usuarios de la aplicación como son: administrador, especialista y anónimo.

4.1. Presentando sistema y caso de estudio

El acceso al sistema se realiza a través de su página de bienvenida, a la cual se llega en un primer acceso mediante una sesión de usuario anónimo, ver Figura 4.1, esta página solamente contiene información referente a la investigación, como el título, situación problemática y objetivos. Los enlaces de accesos son personalizados por los permisos de usuarios.

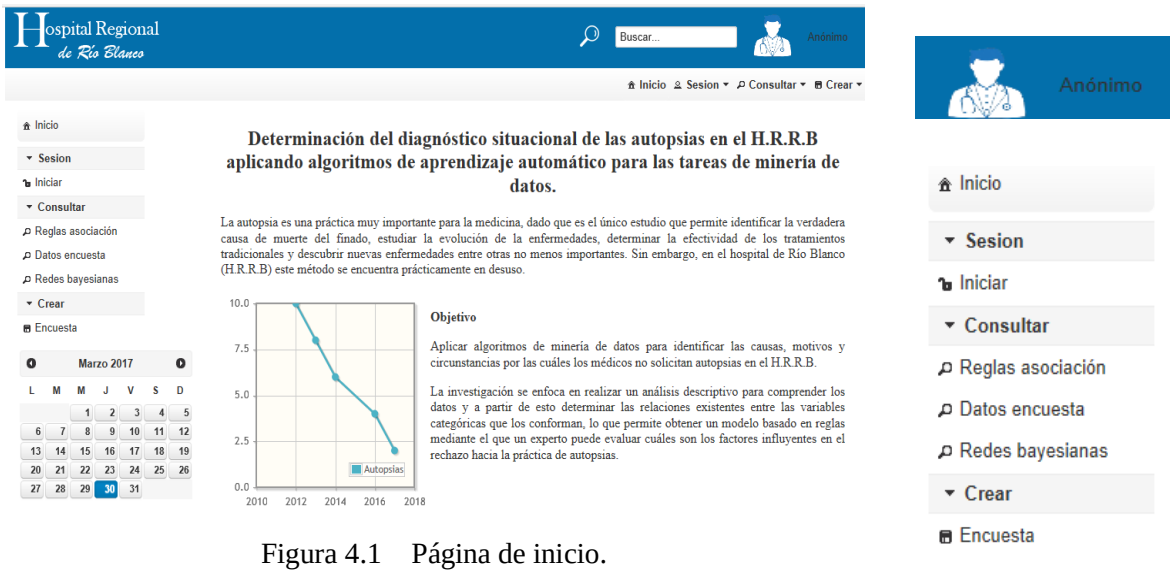


Figura 4.1 Página de inicio.

Un usuario anónimo es capaz de contestar una encuesta, para ello, debe seleccionar "*Encuesta*" en la sesión del menú "*Crear*" y navegará hacia la página para rellenar el formulario, ver Figura 4.2.

Figura 4.2 Formulario de la encuesta.

Una vez que el usuario contesta las preguntas, envía el formulario, y de inmediato se lanza el proceso de clasificación de texto para las preguntas abiertas, y una vez concluido éste, se guardará la encuesta de manera persistente en la base de datos. El resultado de clasificación para las preguntas abiertas se ve físicamente en la tabla "clasificados" de la base de datos donde se registra la respuesta de la pregunta, la que representa el texto que se necesita clasificar y su clase correspondiente resultante de dicho proceso, ver Figura 4.3.

Edit Data - PostgreSQL 9.2 (localhost:5432) - encuesta - clasificado

File Edit View Tools Help

No limit

	id_medico integer	id_preg integer	texto character varying	clasificado character va	idclasificado [PK] integer
1	1	14	Corroborar el diagnóstico de defunción	a	1
2	1	16	Que no exista el servicio	a	2
3	86	22	No hace	a	3
4	1	22	No hace	a	4
5	2	14	La práctica de autopsias se reduce a fines legales y polic:	b	5
6	2	14	El uso médico o de investigación no se hace con la frecuen(	c	6
7	2	16	Miedo a la demanda	b	7
8	2	16	Negativa de los familiares	c	8
9	2	16	Por el propio desconocimiento de médico	d	9
10	2	22	No hace	a	10
11	3	14	Para corroborar la causa de muerte	a	11
12	3	16	Temor a descubrir que la causa de muerte sea por negligenc:	b	12
13	3	22	No hace	a	13
14	4	14	Diagnóstico incierto	a	14
15	4	14	Duda diagnóstica	a	15

Figura 4.3 Respuestas clasificadas.

Desde una sesión anónima el usuario es capaz de firmarse en el sistema para interactuar con éste de acuerdo a los permisos del usuario con el que se haya ingresado. En esta aplicación existen dos roles: administrador y especialista. A continuación, se muestra el papel que juega un especialista dentro de la aplicación, ver Figura. 4.4.

Figura 4.4 Inicio de sesión de usuario *'especialista'*.

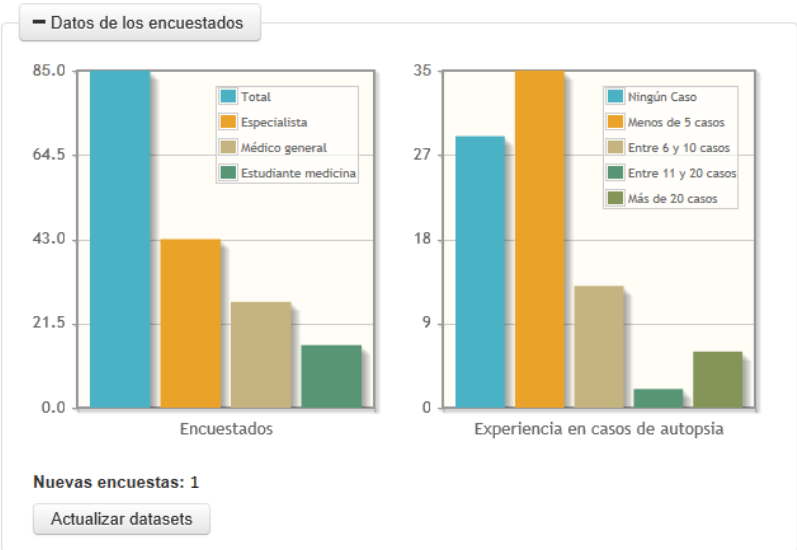


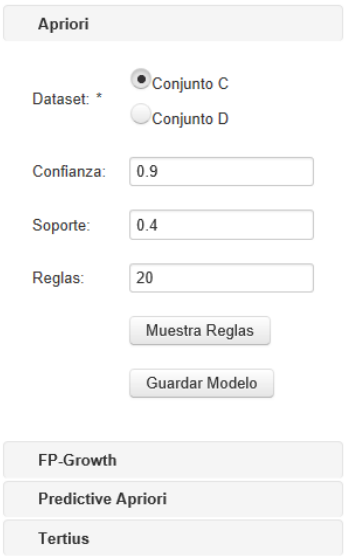
Figura 4.5 Actualizar *datasets*.

En la Figura 4.5 se observa que hay dos gráficas mostrando algunos detalles relacionados con la información de las encuestas que hasta el momento se registraron en el sistema. En la primera se ve que hay un total de 85 encuestas respondidas, de las cuales 43 fueron contestadas por especialistas, 26 por médicos generales y 16 por estudiantes de medicina. En la segunda gráfica se muestran detalles respecto a la experiencia de los médicos en la realización de autopsias, en este caso la cantidad de autopsias en las que intervinieron.

Además, se indica que fue insertada una nueva encuesta, por tanto es necesario actualizar los conjuntos de datos.

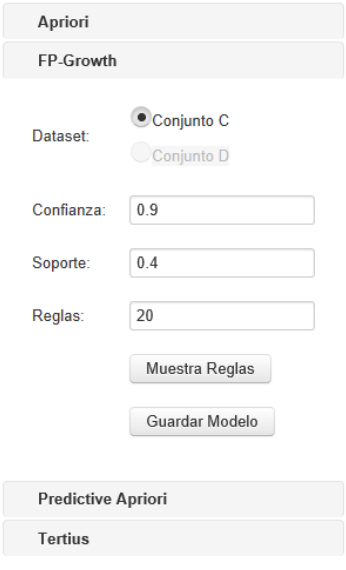
Una vez que el usuario hace clic en el botón “Actualizar *datasets*”, el sistema actualizará los conjuntos de datos “C y D” y además se actualizan los modelos de las redes Bayesianas que se utilizan para la clasificación de las preguntas abiertas. Una vez que el especialista ejecute esta operación, el sistema queda adecuadamente preparado para generar los modelos de minería.

Desde la sesión de “*Crear*” el especialista selecciona “*Reglas de asociación*” o “*Redes Bayesianas*”. Para el primer caso, la navegación llega hasta la página donde se configura cada uno de los algoritmos de acuerdo a los intereses del especialista, se proponen los valores que en esta investigación resultaron ser los más eficientes en el capítulo tres. Cuando un especialista selecciona un algoritmo, revisa sus resultados seleccionando el botón “*Mostrar reglas*”, y el sistema mostrará las reglas propuestas por dicho algoritmo, en las Figuras 4.6, 4.7, 4.8, 4.9, 4.10 y 4.11 se muestran los resultados de los modelos que se utilizaron en este caso de estudio, es decir, se presentan los modelos cuyos resultados fueron los analizados por los expertos para esta investigación.



No.	Regla	Confianza	Soporte
1	p9_4=S ==> p10_1=S	96%	58%
2	p11_1=S, p9_4=S ==> p10_1=S	96%	55%
3	p11_5=S ==> p10_1=S	95%	47%
4	p11_1=S, p15_3=S ==> p10_1=S	95%	47%
5	p6_3=S ==> p10_1=S	95%	42%
6	p6_1=S, p11_1=S ==> p10_1=S	95%	41%
7	p10_1=S, p8_6=S ==> p11_1=S	95%	41%
8	p11_1=S, p8_6=S ==> p10_1=S	95%	41%
9	p11_1=S, p14_1=S ==> p10_1=S	95%	41%
10	p2_2=S ==> p10_1=S	94%	40%
11	p9_4=S ==> p11_1=S	94%	57%
12	p10_1=S, p9_4=S ==> p11_1=S	94%	55%
13	p6_1=S ==> p10_1=S	93%	50%
14	p8_1=S ==> p10_1=S	93%	49%
15	p14_1=S ==> p10_1=S	93%	47%
16	p14_1=S ==> p15_3=S	93%	47%
17	p8_6=S ==> p11_1=S	93%	43%
18	p8_6=S ==> p10_1=S	93%	43%
19	p14_1=S, p15_3=S ==> p10_1=S	93%	43%
20	p10_1=S, p14_1=S ==> p15_3=S	93%	43%

Figura 4.6 Modelo de Apriori para el *dataset* C.

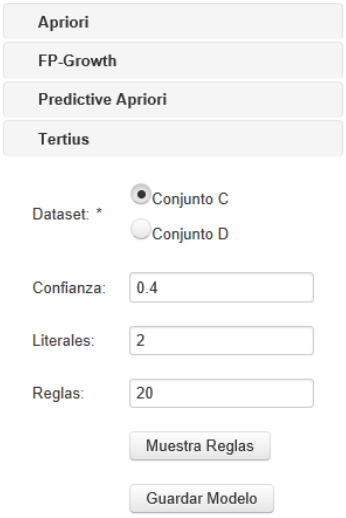


No.	Regla	Confianza	Soporte
1	p9_4=S ==> p10_1=S	96%	58%
2	p11_1=S, p9_4=S ==> p10_1=S	96%	55%
3	p11_5=S ==> p10_1=S	95%	47%
4	p11_1=S, p15_3=S ==> p10_1=S	95%	47%
5	p6_3=S ==> p10_1=S	95%	42%
6	p11_1=S, p6_1=S ==> p10_1=S	95%	41%
7	p11_1=S, p14_1=S ==> p10_1=S	95%	41%
8	p10_1=S, p8_6=S ==> p11_1=S	95%	41%
9	p11_1=S, p8_6=S ==> p10_1=S	95%	41%
10	p9_4=S ==> p11_1=S	94%	57%
11	p10_1=S, p9_4=S ==> p11_1=S	94%	55%
12	p6_1=S ==> p10_1=S	93%	50%
13	p8_1=S ==> p10_1=S	93%	49%
14	p14_1=S ==> p10_1=S	93%	47%
15	p14_1=S ==> p15_3=S	93%	47%
16	p8_6=S ==> p10_1=S	93%	43%
17	p8_6=S ==> p11_1=S	93%	43%
18	p10_1=S, p14_1=S ==> p15_3=S	93%	43%
19	p15_3=S, p14_1=S ==> p10_1=S	93%	43%
20	p11_1=S, p8_1=S ==> p10_1=S	92%	41%
21	p15_3=S ==> p10_1=S	92%	53%
22	p9_4=S ==> p10_1=S, p11_1=S	90%	55%
23	p11_1=S ==> p10_1=S	90%	76%

Figura 4.7 Modelo de FPGrowth para el *dataset* C.

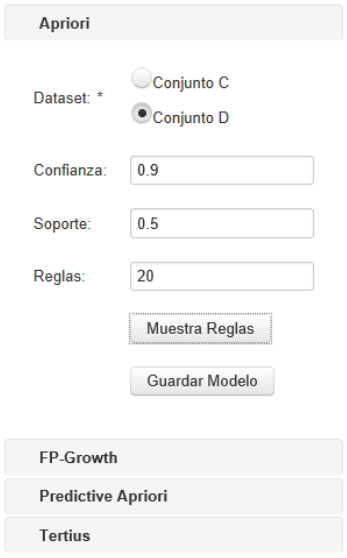
No.	Regla	Confianza	Soporte
1	p9_4=S, p15_3=S ==> p10_1=S	97%	38%
2	p6_1=S, p9_4=S ==> p10_1=S	97%	35%
3	p6_3=S, p15_3=S ==> p10_1=S	96%	30%
4	p12_3=S, p14_1=S ==> p15_3=S	96%	29%
5	p8_4=S ==> p10_1=S	96%	27%
6	p1_1=S, p11_1=S ==> p17_23=S	96%	27%
7	p7_5=S ==> p10_1=S	96%	24%
8	p13_1=S, p12_1=S ==> p10_1=S	95%	23%
9	p6_1=S, p11_1=S, p14_1=S ==> p10_1=S	95%	22%
10	p9_4=S ==> p10_1=S	94%	58%
11	p14_8=S ==> p1_1=S, p15_1=S	94%	19%
12	p14_8=S ==> p1_1=S, p17_23=S	94%	19%
13	p15_1=S ==> p1_1=S, p14_8=S	94%	19%
14	p15_1=S ==> p1_1=S, p17_23=S	94%	19%
15	p9_2=S ==> p10_1=S	94%	19%
16	p1_1=S, p6_1=S ==> p17_23=S	94%	17%
17	p1_1=S, p4_2=S ==> p17_23=S	94%	17%
18	p12_1=S ==> p10_1=S	94%	35%
19	p9_7=S ==> p11_1=S	94%	16%
20	p12_3=S, p10_1=S, p1_5=S ==> p15_3=S	93%	15%

Figura 4.8 Modelo de Predictive Apriori para el *dataset* C.



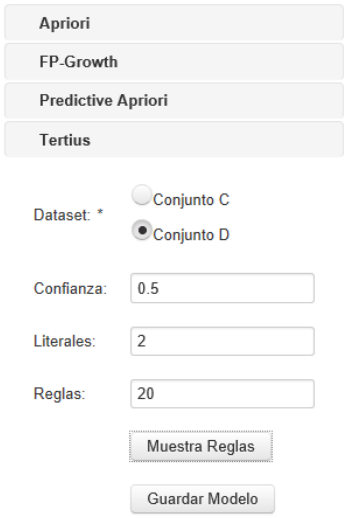
No.	Regla	Confianza	Soporte
1	p14_1=S ==> p15_3=S	70%	3%
2	p15_3=S ==> p14_1=S	70%	12%
3	p17_23=S ==> p1_1=S	66%	16%
4	p17_23=S ==> p14_2=S	61%	19%
5	p15_3=S ==> p1_5=S	46%	31%
6	p8_6=S ==> p12_4=S	42%	27%
7	p17_23=S ==> p14_8=S	41%	30%
8	p17_23=S ==> p15_1=S	41%	30%
9	p14_1=S ==> p1_5=S	39%	27%
10	p8_6=S ==> p15_3=S	37%	10%
11	p15_3=S ==> p8_6=S	37%	22%
12	p17_23=S ==> p15_2=S	31%	30%
13	p4_2=S ==> p3_2=S	30%	29%
14	p17_23=S ==> p5_3=S	28%	33%
15	p11_5=S ==> p8_1=S	28%	16%
16	p8_1=S ==> p11_5=S	28%	20%
17	p5_2=S ==> p4_2=S	28%	12%
18	p4_2=S ==> p5_2=S	28%	26%
19	p9_4=S ==> p8_6=S	28%	26%
20	p14_1=S ==> p5_1=S	28%	34%
21	p11_5=S ==> p9_2=S	26%	34%

Figura 4.9 Modelo de Tertius para el *dataset* C.



No.	Regla	Confianza	Soporte
1	ult_grado=g1, esc_med_gral=c1 ==> area=a3	100%	57%
2	ult_grado=g1, esc_med_gral=c1, per_sol_aut=19a ==> area=a3	100%	53%
3	ult_grado=g1, per_sol_aut=19a, com_sug_op=22a ==> area=a3	99%	53%
4	categoria=c4, ult_grado=g1, per_sol_aut=19a, com_sug_op=22a ==> area=a3	99%	51%
5	ult_grado=g1, per_sol_aut=19a ==> area=a3	99%	61%
6	categoria=c4, ult_grado=g1, per_sol_aut=19a ==> area=a3	99%	55%
7	ult_grado=g1, com_sug_op=22a ==> area=a3	98%	57%
8	categoria=c4, ult_grado=g1, com_sug_op=22a ==> area=a3	98%	52%
9	ult_grado=g1 ==> area=a3	98%	67%
10	categoria=c4, esc_med_gral=c1, com_sug_op=22a ==> area=a3	98%	50%
11	categoria=c4, ult_grado=g1 ==> area=a3	98%	57%
12	area=a3, categoria=c4, com_sug_op=22a ==> per_sol_aut=19a	98%	55%
13	area=a3, categoria=c4, ult_grado=g1, com_sug_op=22a ==> per_sol_aut=19a	98%	51%
14	area=a3, categoria=c4, ult_grado=g1 ==> per_sol_aut=19a	97%	55%
15	area=a3, per_sol_aut=19a, com_sug_op=22a ==> categoria=c4	97%	55%
16	ult_grado=g1, per_sol_aut=19a, com_sug_op=22a ==> categoria=c4	97%	51%
17	area=a3, ult_grado=g1, per_sol_aut=19a, com_sug_op=22a ==> categoria=c4	97%	51%
18	categoria=c4, ult_grado=g1, com_sug_op=22a ==> per_sol_aut=19a	97%	51%
19	area=a3, categoria=c4, esc_med_gral=c1 ==> per_sol_aut=19a	97%	51%
20	esc_med_gral=c1, com_sug_op=22a ==> area=a3	97%	53%

Figura 4.10 Modelo de Apriori para el *dataset* D.



No.	Regla	Confianza	Soporte
1	ult_grado=g1 ==> anios_prac=3e	55%	27%
2	hall_disc=7a ==> anios_prac=3e	46%	26%
3	hall_disc=7a ==> casos=4d	42%	38%
4	hall_disc=7a ==> esc_esp=c20	42%	38%
5	hall_disc=7a ==> hall_arb=8c	41%	35%
6	esc_med_gral=c1 ==> anios_prac=3e	40%	34%
7	ult_grado=g1 ==> casos=4d	38%	44%
8	ult_grado=g1 ==> esc_esp=c20	38%	44%
9	area=a3 ==> anios_prac=3e	38%	40%
10	esc_med_gral=c1 ==> hall_dem=11d	36%	39%
11	esc_med_gral=c1 ==> casos=4d	35%	48%
12	esc_med_gral=c1 ==> esc_esp=c20	34%	49%
13	esc_med_gral=c1 ==> hall_arb=8c	34%	45%
14	com_sug_op=22a ==> esc_esp=c20	33%	50%
15	com_sug_op=22a ==> casos=4d	32%	50%
16	categoria=c4 ==> hall_arb=8c	30%	50%
17	categoria=c4 ==> esc_esp=c20	30%	54%
18	ult_grado=g1 ==> hall_dem=11a	29%	48%
19	categoria=c4 ==> casos=4d	29%	54%
20	hall_disc=7a ==> hall_arb=8a	28%	39%

Figura 4.11 Modelo de Tertius para el *dataset* C.

Cuando el especialista determina que un modelo es adecuado, lo guarda seleccionando el botón "*Guardar modelo*", y así queda establecido el modelo que se leerá para mostrar los resultados a los usuarios que accedan a la aplicación.

En cambio, si la selección de usuario es "*Redes Bayesianas*", el sistema mostrará la página donde el especialista selecciona la clase, es decir, una pregunta de la encuesta para a partir de ésta generar el modelo. Para el caso de estudio se generaron dos redes Bayesianas considerando como clases a los atributos: *motivos por los que no se solicitan autopsias* y *motivos por los que se solicitan autopsias*, del conjunto de datos 'D', ver Figuras 4.12 y 4.13, donde se muestran los grafos generados.

A partir del grafo, el especialista consulta cada nodo para observar los resultados generados de acuerdo a las relaciones de probabilidad condicional entre los distintos atributos, el ejemplo mostrado en la Figura 4.14 representa los valores de probabilidad condicional para el nodo "*casos*" de la red Bayesiana que se generó a partir de la clase *motivos para no solicitar autopsias*.

Como se observa en la Figura 4.14, se generaron muchísimas salidas, se calcula la cantidad aproximada multiplicando cantidad de páginas por número de elementos en ella contenidos (234x12), lo que para este ejemplo dará una cantidad de 2808 resultados.

Es imposible revisar cada una de estas salidas y además no todas son de interés, por ello se implementó un filtro que da la posibilidad de reducir los resultados de acuerdo a los valores de interés del especialista, así como la opción de ordenarlos atendiendo a los valores de probabilidad de ocurrencia entre los atributos.

Esto se demuestra con un ejemplo en concreto. Se tiene interés de conocer la opinión de los médicos que se especializaron en el Instituto Nacional de Nutrición y participaron entre 11 y 20 casos de autopsias. Pues bien, hasta ahora se generó el grafo para la clase que responde a los motivos para no solicitar autopsias y se seleccionó el nodo "*casos*". En este punto lo que se hace es filtrar por los valores específicos de interés, como por ejemplo seleccionar en *Valor Nodo* la opción referida a 11 y 20 casos (4d), en *Nodo Padre* escoger la opción de la escuela donde el médico hizo su especialidad (*esc_espe*) y especificar en *Valor Padre* la escuela de interés, que sería para este ejemplo el Instituto Nacional de Nutrición (c20). Como se muestra en la Figura 4.15, el número de resultados se redujo a 24 salidas, se organizó de manera descendente de acuerdo a los valores de probabilidad y a su vez se eliminaron los resultados que carecían de interés.

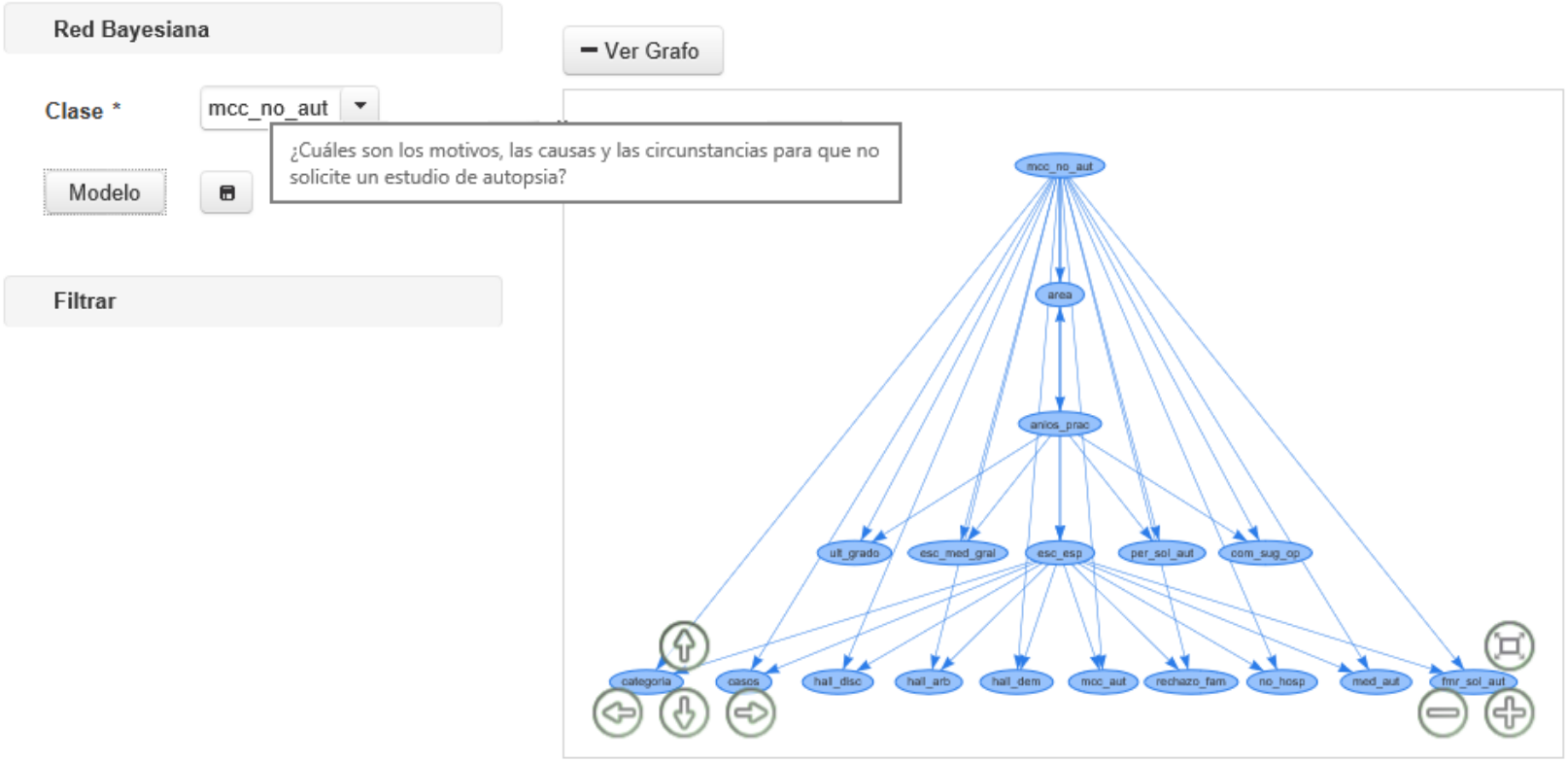


Figura 4.12 Grafo de la red bayesiana formado a partir de la clase "*mcc_no_aut*".

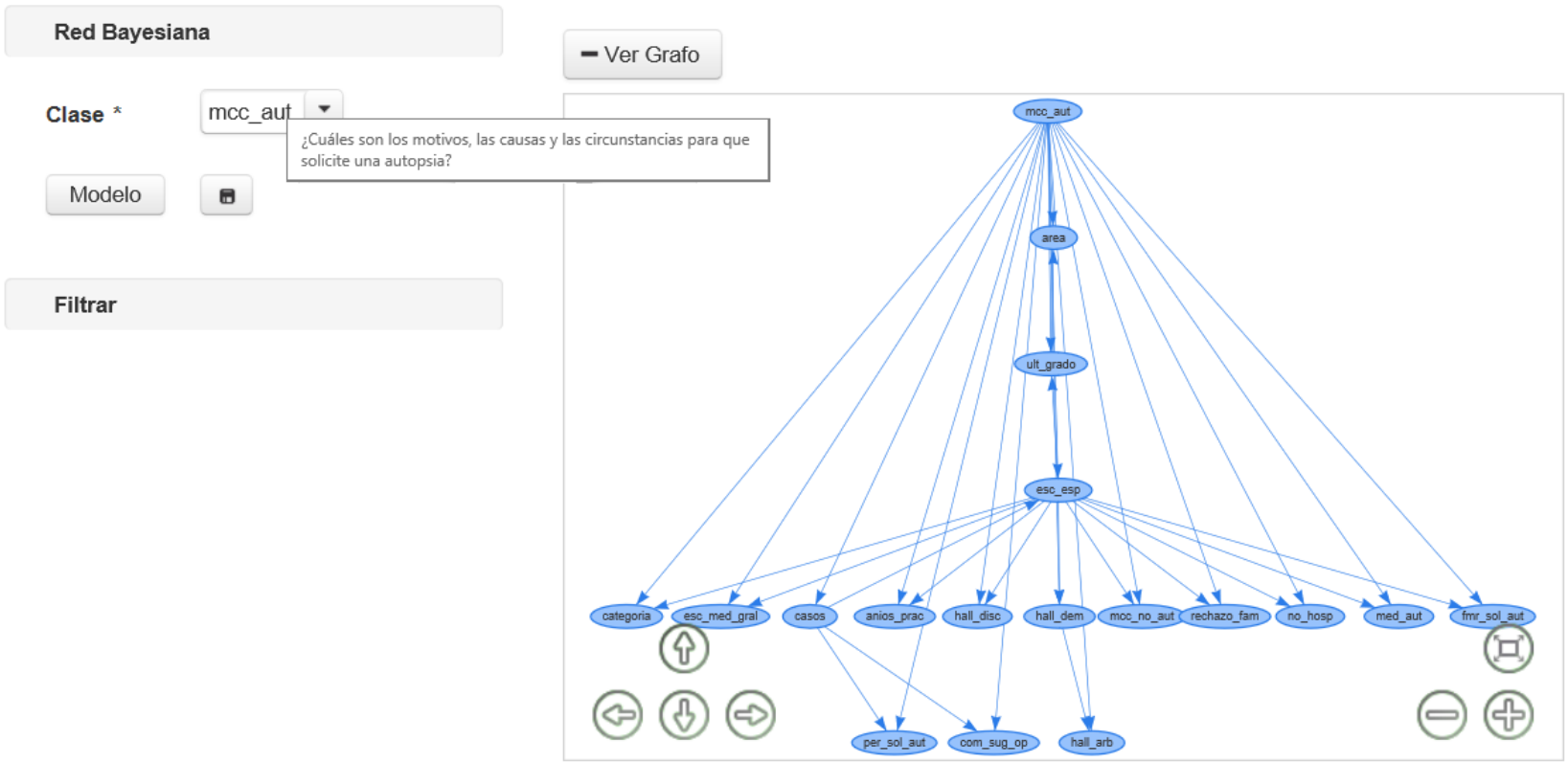


Figura 4.13 Grafo de la red bayesiana formado a partir de la clase "*mcc_aut*".

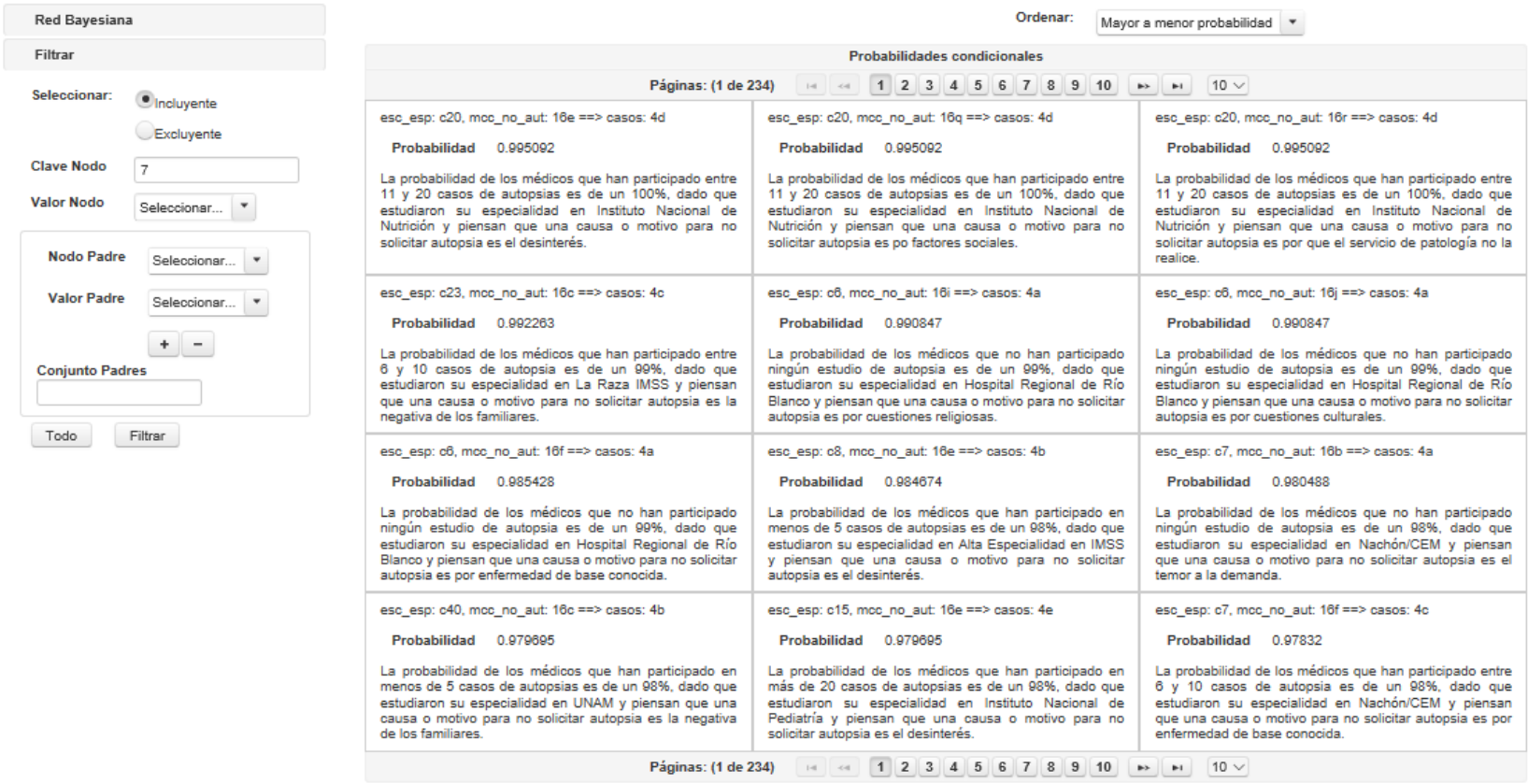


Figura 4.14 Probabilidades del nodo *casos*.

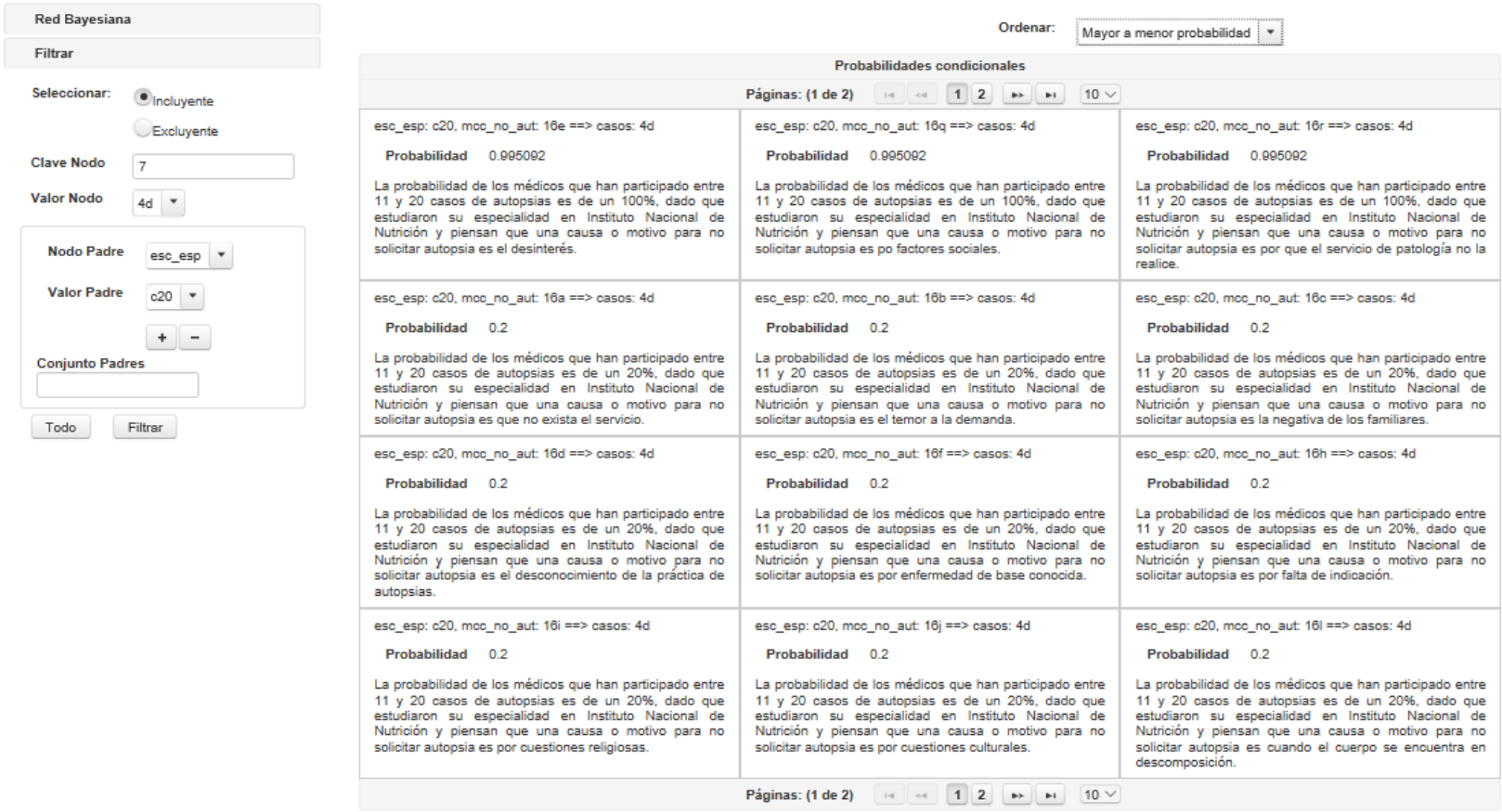


Figura 4.15 Filtro aplicado a los resultados del nodo *casos*.

Una vez que el especialista determina cuál es el modelo de interés, lo guarda seleccionando el botón "*Guardar modelo*", ver Figura 4.16, y así queda establecido el modelo que se leerá para mostrar los resultados a los usuarios que accedan a la aplicación.

Figura 4.16 Operación "*Guardar modelo*".

Un especialista también es capaz de seleccionar "*Encuesta*" desde las sesiones del menú "*Eliminar"* y "*Consultar*" si requiere realizar alguna de estas operaciones y el sistema mostrará una página como la que se representa en la Figura 4.17.

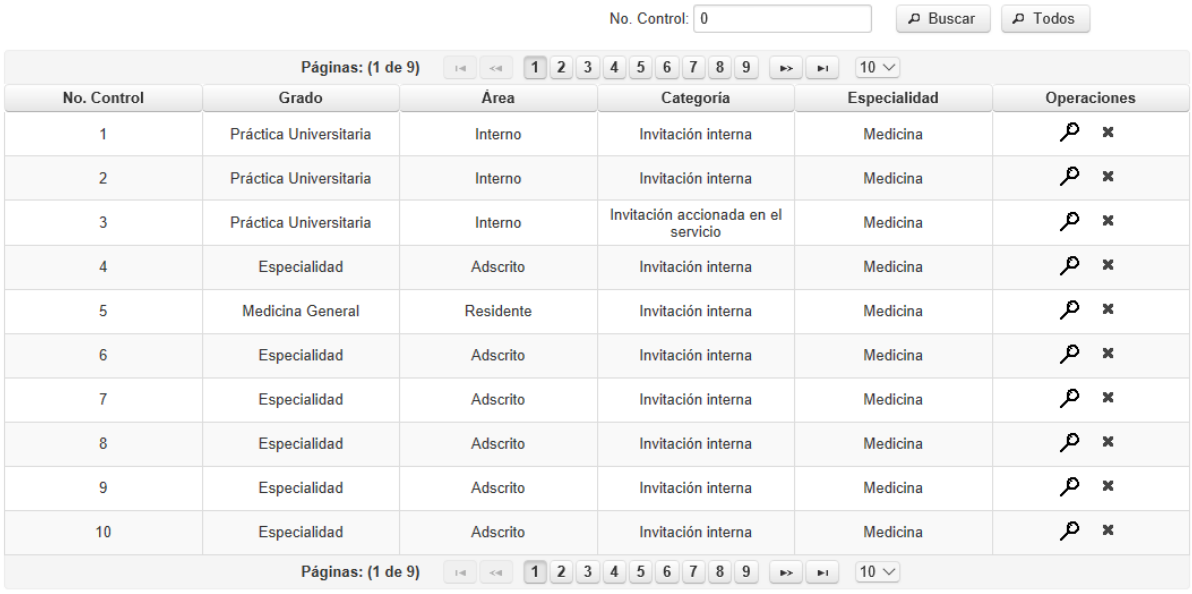


Figura 4.17 Gestionar encuesta.

El filtro de esta página facilita la ubicación de la encuesta requerida mediante el número de control cuyo valor es único y es asignado a cada encuesta en el momento de su registro en la base de datos, de esta manera y sin lugar a equívocos, el resultado será un único registro que responde al número de control especificado, ver Figura 4.18.

Figura 4.18 Filtrar encuesta.

Desde la opción de "*Operaciones*" se consultan las respuestas de la encuesta seleccionada (ver Figura 4.19) o eliminarla directamente (ver Figura 4.20).

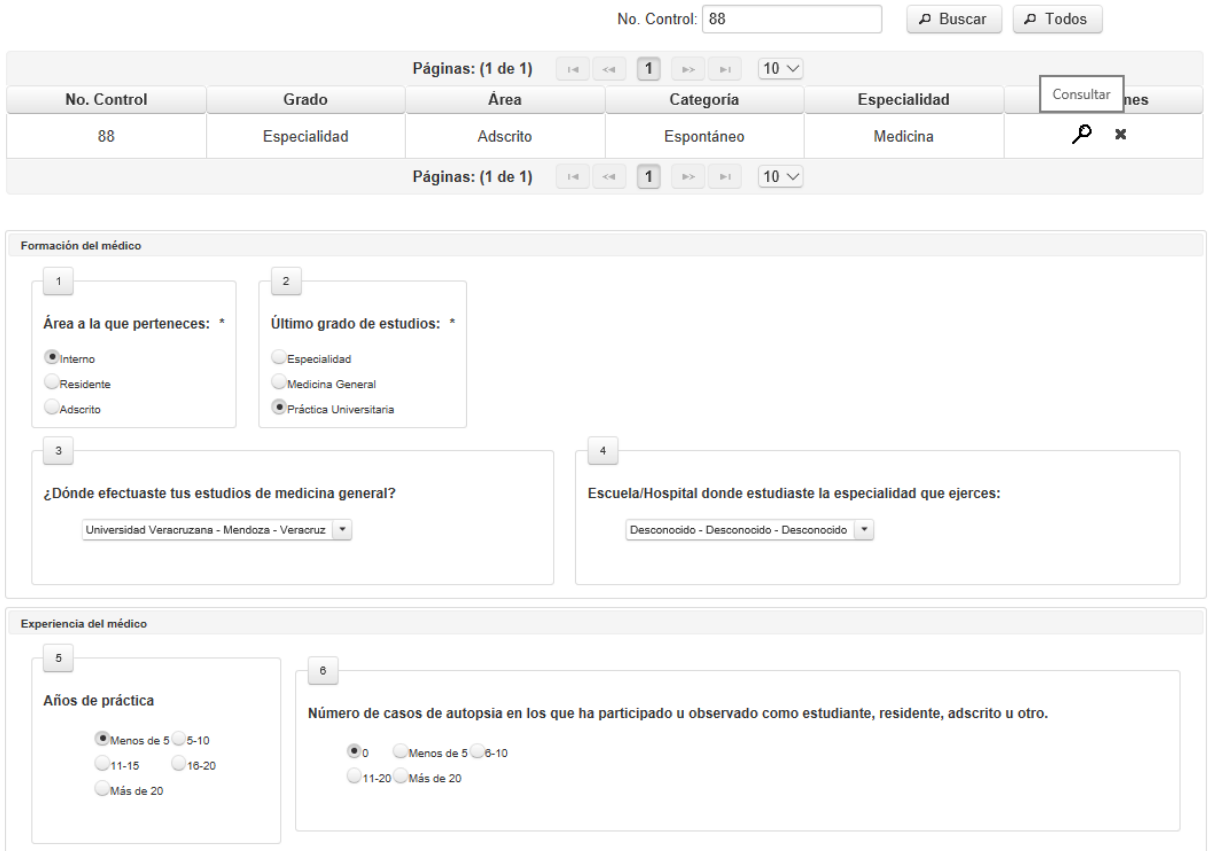


Figura 4.19 Consultar encuesta.

Figura 4.20 Eliminar encuesta.

El especialista cierra su sesión en el momento que lo desee, para ello selecciona la opción "*Cerrar*" en la sección "*Sesión*" del menú, el sistema pide confirmación para proceder con la operación (ver Figura 4.21) cierra la sesión especialista y genera una sesión para usuario anónimo.

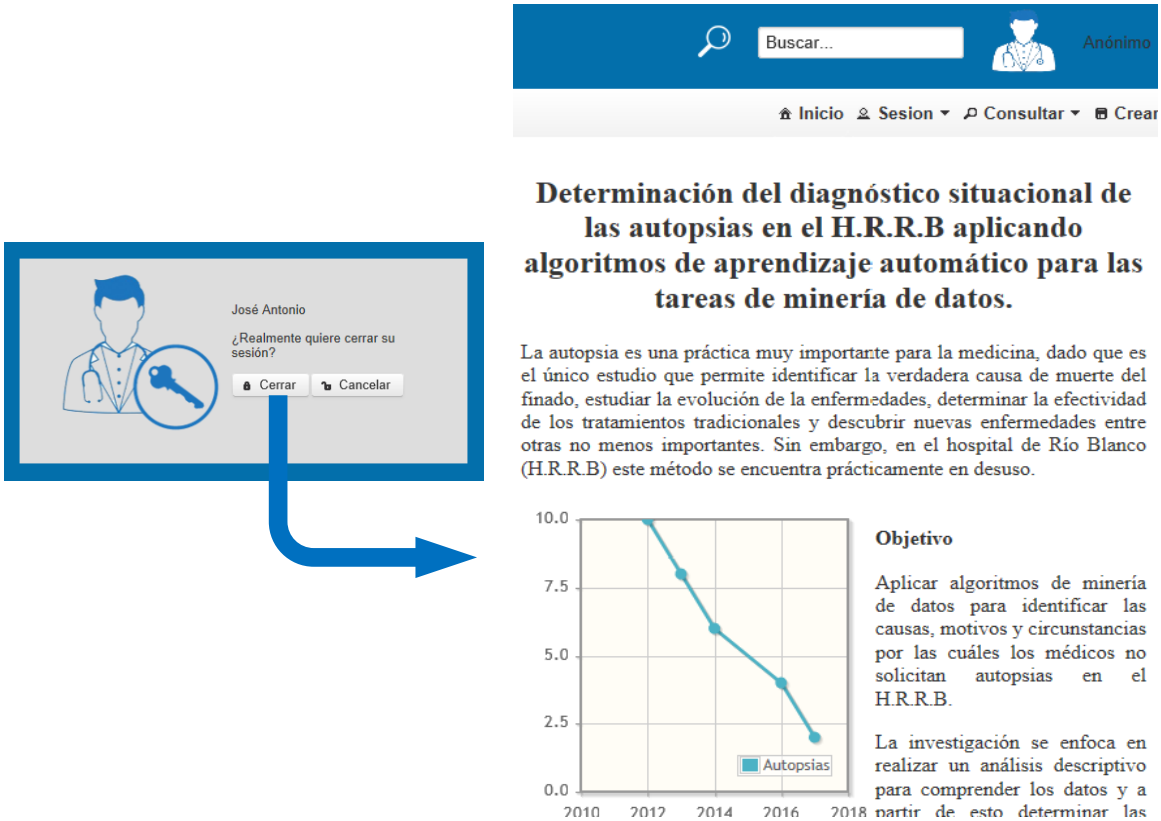


Figura 4.21 Cerrar sesión.

Un administrador tiene permisos para gestionar los usuarios de la aplicación, es decir, es quién es capaz de insertar, modificar, consultar y eliminar usuarios. El administrador debe firmarse en el sistema (ver Figura 4.22) y éste le mostrará la página desde la cual accederá a sus operaciones (ver figura 4.23).

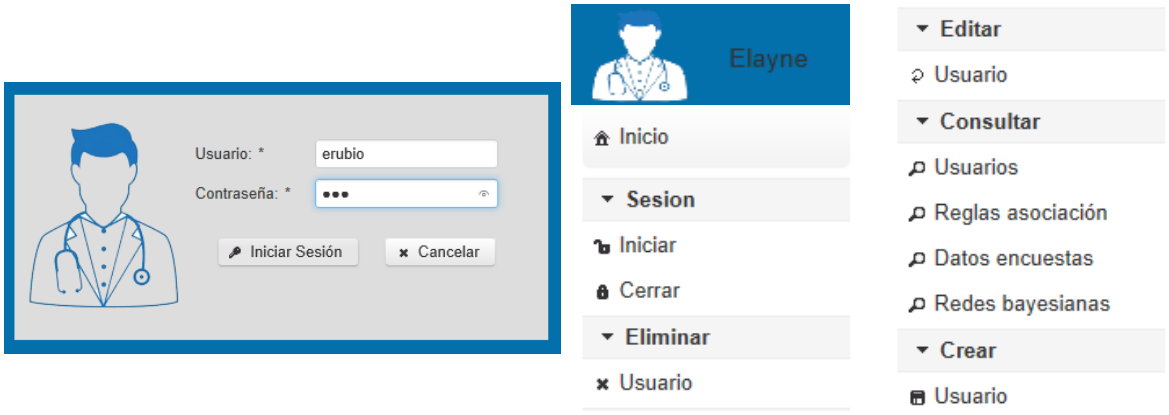


Figura 4.22 Iniciar sesión como administrador.

Figura 4.23 Página para gestionar usuarios.

Para crear un nuevo usuario, el administrador llena tanto los datos personales como los de la cuenta especificando el tipo de usuario (Administrador o Especialista). Después de llenar toda la información, se selecciona registrar y el nuevo usuario se guarda de manera persistente en la base de datos y quedan listas sus credenciales para que pueda firmarse en la aplicación (ver figura 4.24).

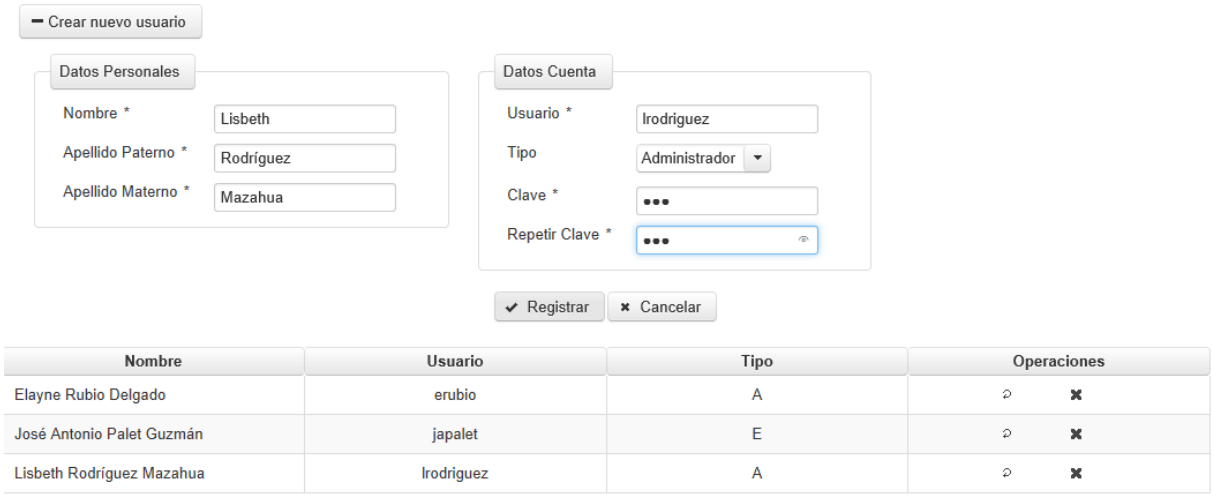


Figura 4.24 Insertar nuevo usuario.

Los usuarios son modificados o eliminados accediendo a estas operaciones mediante los botones que aparecen en la columna *Operaciones* en la tabla que contiene a los usuarios registrados de la aplicación. Al seleccionar la operación de actualizar un usuario, la aplicación muestra la información para ese usuario, el administrador realizará los cambios necesarios y confirmará el cambio haciendo clic en el botón "*Actualizar*" (ver Figura 4.25).

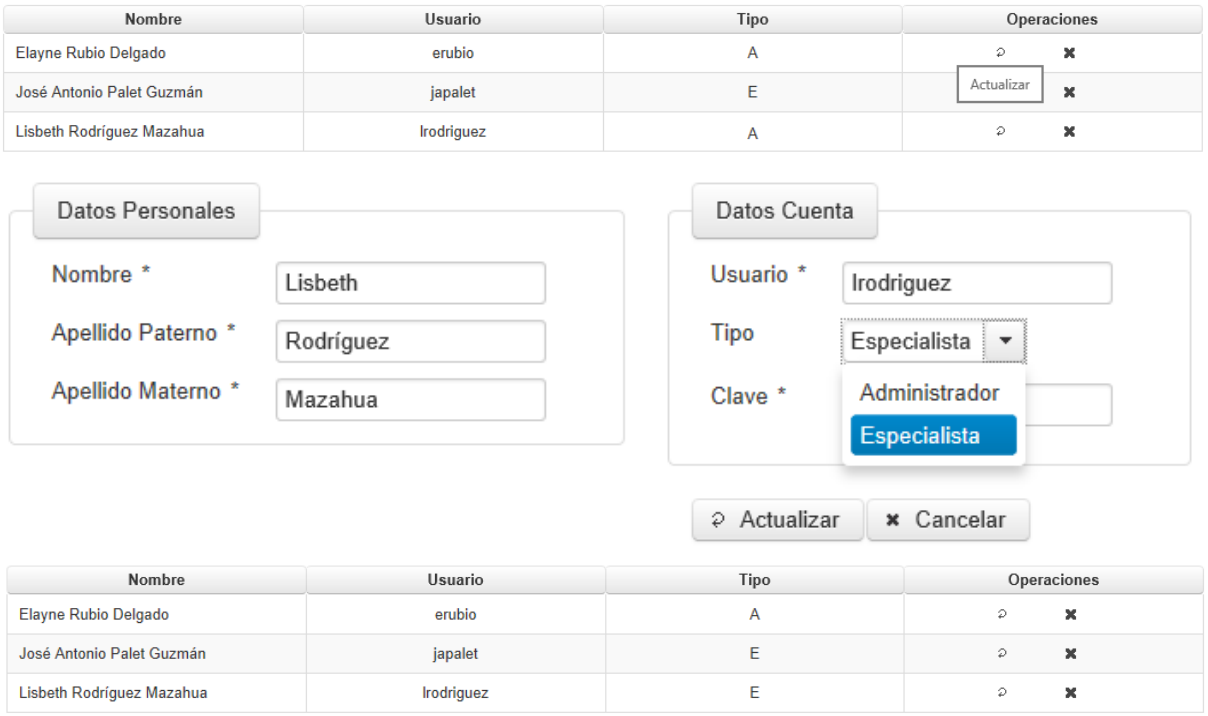


Figura 4.25 Actualizar usuario

Para eliminar un usuario, solo es necesario seleccionar la opción de "*Eliminar*" en la columna "*Operaciones*" y el usuario especificado será eliminado definitivamente de la aplicación (ver figura 4.26).

Figura 4.26 Eliminar usuario.

No existe ningún permiso o restricción para consultar los resultados de los modelos que fueron guardados previamente. Es por ello que cualquier usuario que acceda a la aplicación desde la sección "*Consultar*" es capaz de seleccionar las opciones para revisar los resultados de los modelos. Para facilitar la comprensión de los mismos, el sistema muestra de manera automática una explicación en lenguaje natural de los resultados.

Para ilustrar lo expuesto en el párrafo anterior, las descripciones de las reglas obtenidas por el modelo *Apriori* para el conjunto 'D' se muestran en la Figura 4.27. Por su parte, la Figura 4.28 muestra la interpretación de las relaciones de probabilidad entre los atributos del nodo '*med_aut*' de la red Bayesiana generada a partir de la clase '*mcc_no_aut*'.

Regla
El 100% de los encuestados que son especialistas y estudiaron medicina general en Universidad Veracruzana, también son adscritos. Esta regla aparece con una frecuencia del 57%.
El 100% de los encuestados que son especialistas, estudiaron medicina general en Universidad Veracruzana y alegan que es el médico el personal adecuado para ordenar autopsia, también son adscritos. Esta regla aparece con una frecuencia del 53%.
El 99% de los encuestados que son especialistas, alegan que es el médico el personal adecuado para ordenar autopsia y no comenta, también son adscritos. Esta regla aparece con una frecuencia del 53%.
El 99% de los encuestados que respondieron la encuesta por una invitación accionada fuera del servicio, son especialistas, alegan que es el médico el personal adecuado para ordenar autopsia y no comenta, también son adscritos. Esta regla aparece con una frecuencia del 51%.
El 99% de los encuestados que son especialistas y alegan que es el médico el personal adecuado para ordenar autopsia, también son adscritos. Esta regla aparece con una frecuencia del 61%.
El 99% de los encuestados que respondieron la encuesta por una invitación accionada fuera del servicio, son especialistas y alegan que es el médico el personal adecuado para ordenar autopsia, también son adscritos. Esta regla aparece con una frecuencia del 55%.
El 98% de los encuestados que son especialistas y no comenta, también son adscritos. Esta regla aparece con una frecuencia del 57%.
El 98% de los encuestados que respondieron la encuesta por una invitación accionada fuera del servicio, son especialistas y no comenta, también son adscritos. Esta regla aparece con una frecuencia del 52%.
El 98% de los encuestados que son especialistas, también son adscritos. Esta regla aparece con una frecuencia del 67%.
El 98% de los encuestados que respondieron la encuesta por una invitación accionada fuera del servicio, estudiaron medicina general en Universidad Veracruzana y no comenta, también son adscritos. Esta regla aparece con una frecuencia del 50%.
El 98% de los encuestados que respondieron la encuesta por una invitación accionada fuera del servicio y son especialistas, también son adscritos. Esta regla aparece con una frecuencia del 57%.
El 98% de los encuestados que son adscritos, respondieron la encuesta por una invitación accionada fuera del servicio y no comenta, también alegan que es el médico el personal adecuado para ordenar autopsia. Esta regla aparece con una frecuencia del 55%.
El 98% de los encuestados que son adscritos, respondieron la encuesta por una invitación accionada fuera del servicio, son especialistas y no comenta, también alegan que es el médico el personal adecuado para ordenar autopsia. Esta regla aparece con una frecuencia del 51%.
El 97% de los encuestados que son adscritos, respondieron la encuesta por una invitación accionada fuera del servicio y son especialistas, también alegan que es el médico el personal adecuado para ordenar autopsia. Esta regla aparece con una frecuencia del 55%.
El 97% de los encuestados que son adscritos, alegan que es el médico el personal adecuado para ordenar autopsia y no comenta, también respondieron la encuesta por una invitación accionada fuera del servicio. Esta regla aparece con una frecuencia del 55%.
El 97% de los encuestados que son especialistas, alegan que es el médico el personal adecuado para ordenar autopsia y no comenta, también respondieron la encuesta por una invitación accionada fuera del servicio. Esta regla aparece con una frecuencia del 51%.
El 97% de los encuestados que son adscritos, son especialistas, alegan que es el médico el personal adecuado para ordenar autopsia y no comenta, también respondieron la encuesta por una invitación accionada fuera del servicio. Esta regla aparece con una frecuencia del 51%.
El 97% de los encuestados que respondieron la encuesta por una invitación accionada fuera del servicio, son especialistas y no comenta, también alegan que es el médico el personal adecuado para ordenar autopsia. Esta regla aparece con una frecuencia del 51%.
El 97% de los encuestados que son adscritos, respondieron la encuesta por una invitación accionada fuera del servicio y estudiaron medicina general en Universidad Veracruzana, también alegan que es el médico el personal adecuado para ordenar autopsia. Esta regla aparece con una frecuencia del 51%.
El 97% de los encuestados que estudiaron medicina general en Universidad Veracruzana y no comenta, también son adscritos. Esta regla aparece con una frecuencia del 53%.

Figura 4.27 Resultados de Apriori – 'D'.

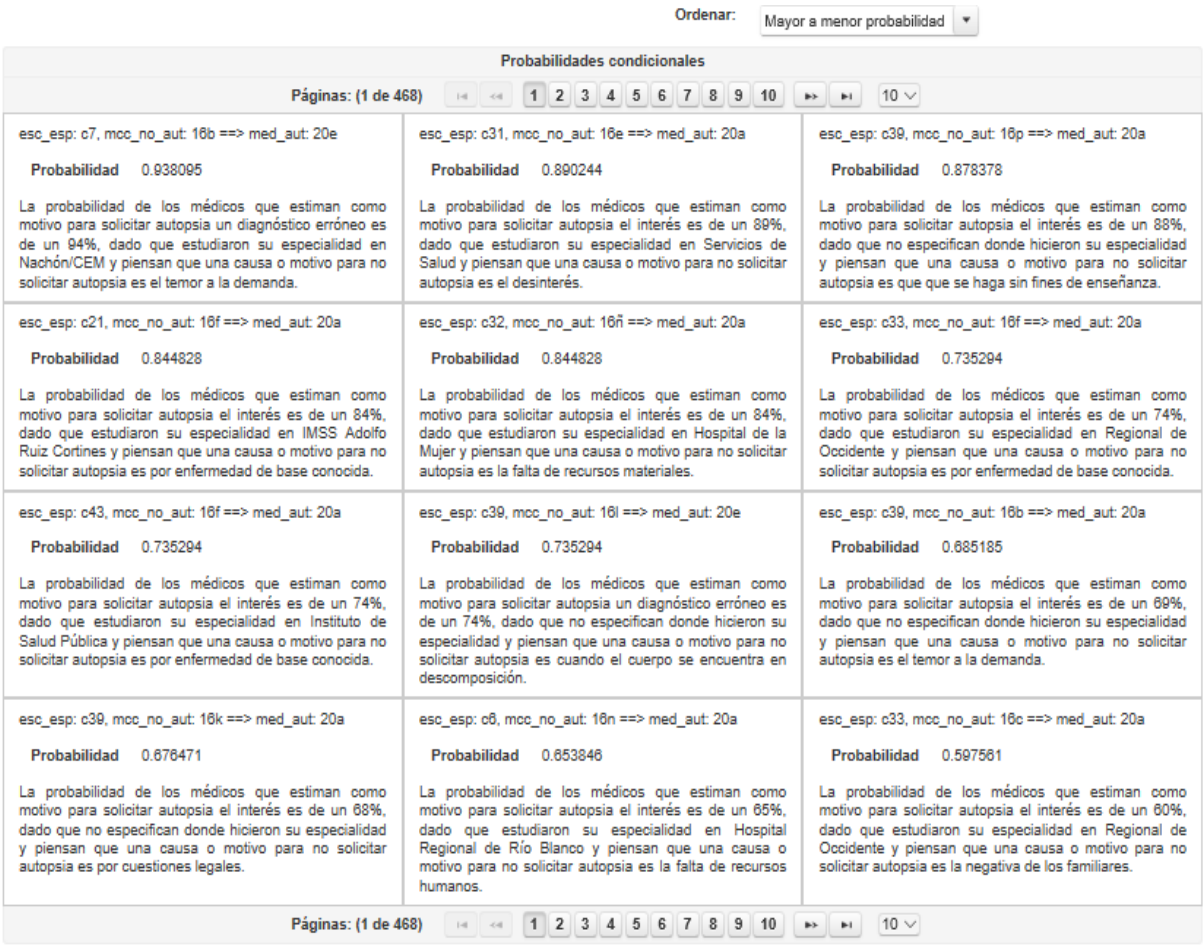


Figura 4.28 Resultados de las relaciones del nodo: *razones del médico para solicitar autopsias*.

4.2. Evaluación de resultados

Una vez generados los modelos se analiza el conocimiento extraído de los datos, es decir, las respuestas de las encuestas. Esta actividad se lleva a cabo por los expertos en el área que está siendo objeto de investigación, en este caso patología. Por ello, se consideró necesario que el sistema proporcione una explicación entendible de las reglas y de las relaciones de probabilidad condicional entre los atributos para apoyar al experto en el proceso de evaluación. De esta manera, el especialista es capaz de evaluar los resultados analizando de manera subjetiva la información extraída por los modelos con base en su experiencia y conocimientos sin la dependencia total de un especialista de minería de datos.

4.2.1. Reglas de asociación

Los resultados comprendían un total 120 reglas, 20 por modelo, de las cuales solo 100 pasaron a ser evaluadas por el especialista. Se tomó la decisión de no incluir las reglas de *FPGrowth* para evitar repeticiones en los resultados porque la mayoría aparecen también en *Apriori*.
Después de un análisis minucioso de las reglas los resultados concluyentes fueron: para un total de 100 reglas, quedaron aprobadas por el experto 75. Se descartaron ocho reglas en cada modelo de Apriori, siete en Predictive Apriori y dos en el modelo de Tertius con el conjunto 'D'. El algoritmo de mayor aceptación resultó ser Tertius con un 90% de reglas aprobadas con el conjunto 'D' y 100% con el conjunto 'C' (ver Tabla 4.1). De manera general se concluye en el análisis de asociación que los resultados tuvieron un 75% de aprobación.

Tabla 4.19 Evaluación de los resultados de asociación.

Algoritmo	Conjuntos	Aceptadas	Descartadas	Aceptación
Apriori	C	12	8	60%
	D	12	8	60%
Predictive Apriori	C	13	7	65%
Tertius	C	20	0	100%
	D	18	2	90%

4.2.2. Redes Bayesianas

Resulta complejo analizar los datos generados por los modelos Bayesianos debido al gran volumen de relaciones de probabilidad que se extrajeron de estas redes. Es por ello, que para esta investigación el experto delimitó el análisis los resultados con probabilidad mayor de 50% que relacionen *años de práctica*, *casos en los que ha intervenido el médico*, *hallazgos discrepantes, hallazgos de demanda, causas de rechazo a las autopsias, por qué no se solicitan en el hospital* y *causas por las que el médico no solicita autopsias*. De esta manera se sometieron a evaluación un total de 352 probabilidades condicionales, 186 fueron extraídas de la red generada a partir de la clase *mcc_aut* y 210 para la de *mcc_no_aut* (ver Tabla 4.2).

Tabla 4.2 Resultados de las redes Bayesianas.

BayesNet	Resultados
mcc_aut	168
mcc_no_aut	184
Total	352

Después de un análisis minucioso de las relaciones de probabilidad condicional los resultados concluyentes fueron: para un total de 352 probabilidades condicionales, quedaron aprobadas por el experto 347. Se descartó 1 probabilidad en la red Bayesiana para *mcc_aut* y 4 probabilidades en la red Bayesiana para *mcc_no_aut.* De manera general, se concluye que las Redes Bayesianas tuvieron un 98.6% de aprobación (ver Tabla 4.3).

Tabla 4.3 Evaluación de los resultados de redes Bayesianas.

Resultados	Aceptados	Descartados	Aceptación
mcc_aut	167	1	99.4%
mcc_no_aut	180	4	97.8%
Total	347	5	98.6%

Las redes permitieron además establecer un diagnóstico situacional de las autopsias en el H. R. R. B., el cual es detallado en las Figuras 4.29 y 4.30, referentes a las causantes y motivos para solicitar y no solicitar la realización de autopsias, respectivamente.

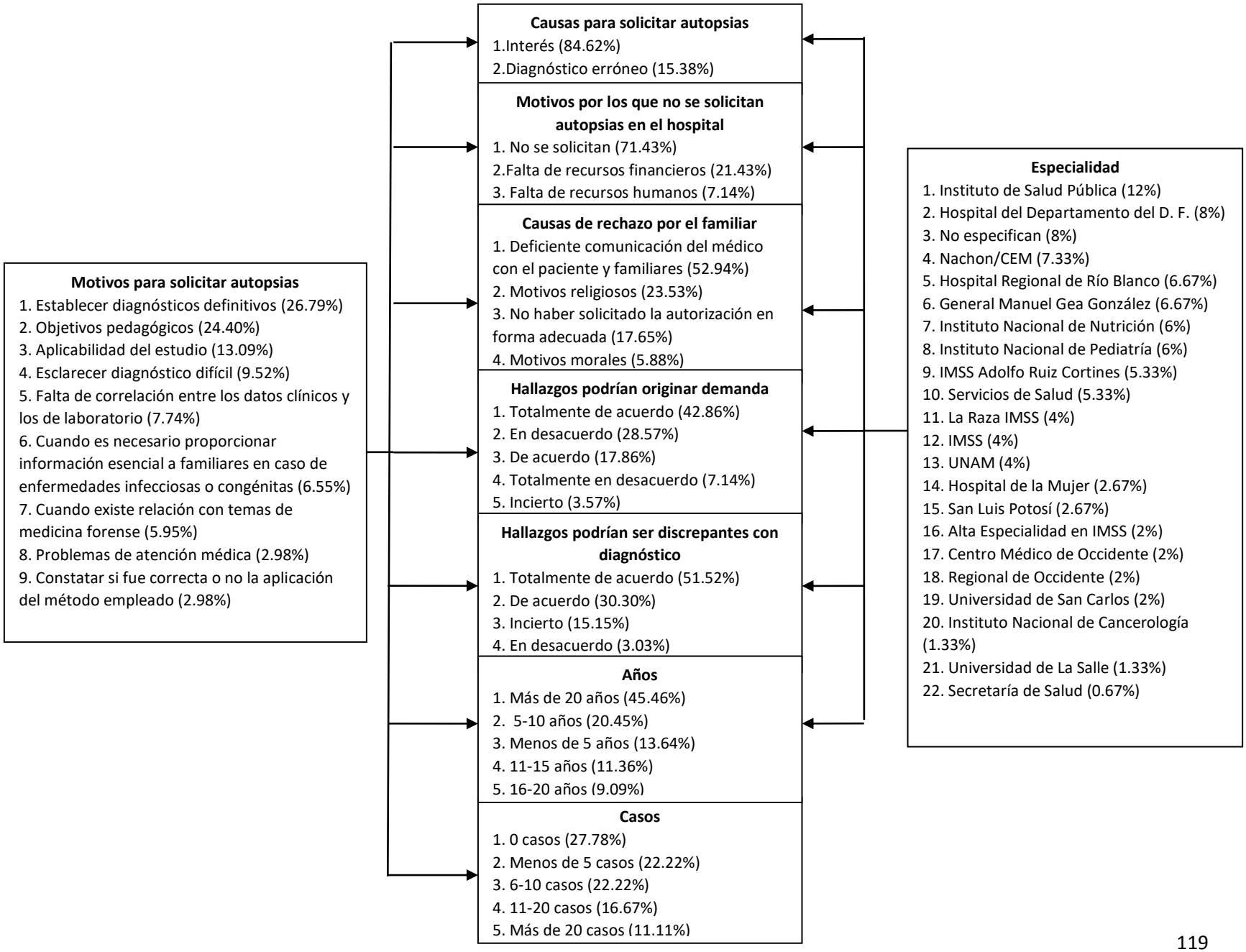


Figura 4.29 Diagnóstico situacional de las autopsias en el H. R. R. B. sobre los motivos para solicitar autopsias.

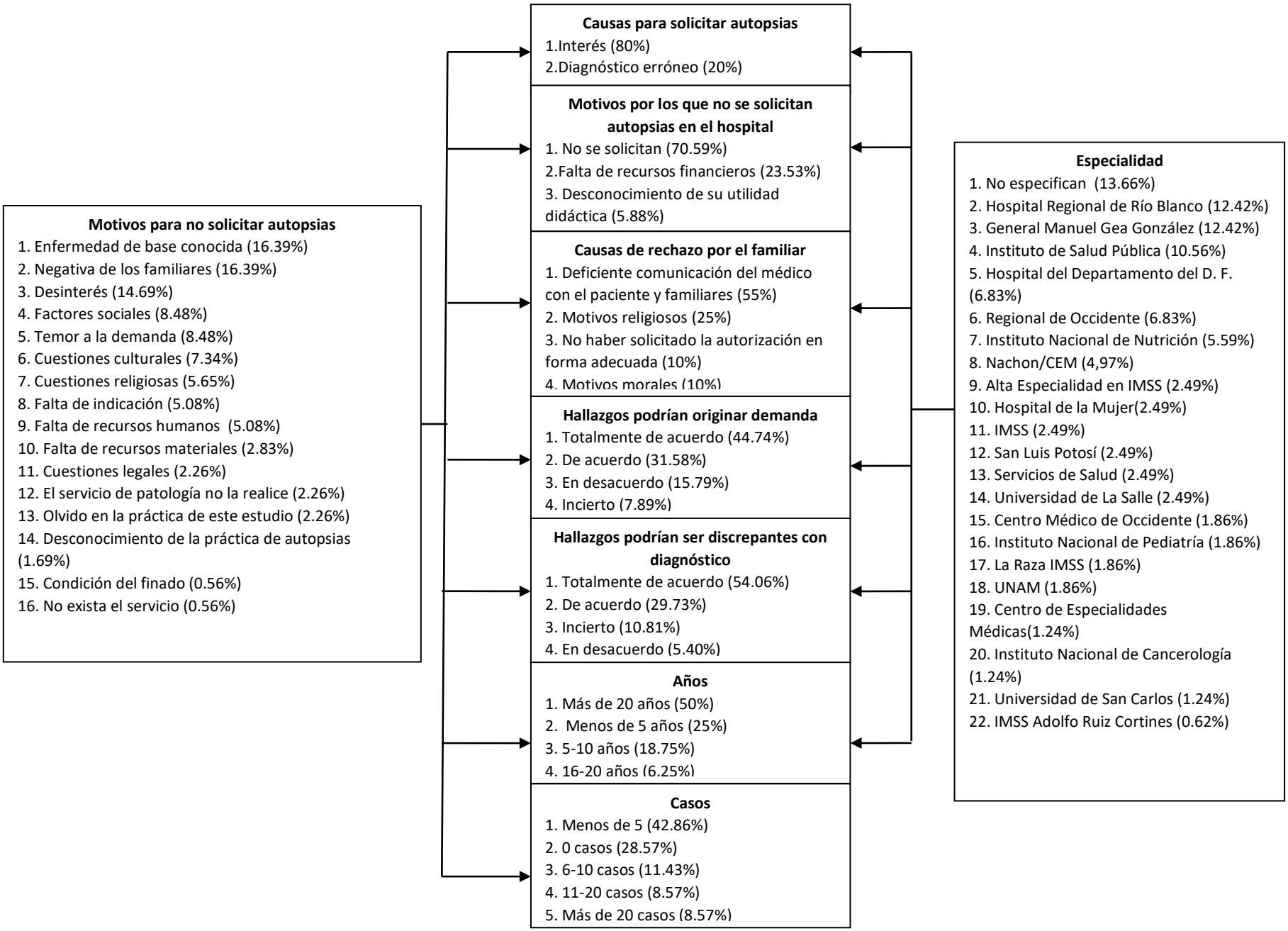


Figura 4.30 Diagnóstico situacional de las autopsias en el H. R. R. B. sobre los motivos para no solicitar autopsias.

Capítulo 5. Conclusiones y recomendaciones

5.1. Conclusiones

Dada la importancia que presentan los estudios de autopsias para la práctica médica, y teniendo en cuenta la problemática planteada, en esta investigación se busca identificar los elementos que en ella intervienen utilizando técnicas de MD que, como se comentó a lo largo de este trabajo, es una de las herramientas más poderosas para el manejo de grandes volúmenes de información. La MD posibilita extraer el conocimiento oculto en los datos, ese que las limitadas capacidades manuales no alcanzan a descubrir, de forma no trivial.

El propósito de este trabajo fue analizar las posibles causas de la reducción de las autopsias en el Hospital Regional de Río Blanco de Veracruz mediante la minería de reglas de asociación y redes Bayesianas a partir de los datos que pertenecen a las opiniones médicas sobre tal práctica médica. Asimismo, la investigación se apropia del concepto de aprendizaje automático para construir un sistema capaz de aprender o inferir conocimiento, a partir de un conjunto de datos de entrada, y generar respuestas adecuadas.

Los datos analizados fueron recogidos a través de una encuesta que se aplicó a los médicos del hospital. La encuesta se centró en las opiniones médicas sobre las causas o motivos por las que las autopsias no se realizaron, el nivel de estudio de los especialistas, sus años de experiencia, los casos de autopsias en los que estuvieron involucrados, entre otros.

El uso de técnicas de minería de reglas de asociación y redes Bayesianas permitió realizar un análisis descriptivo de la situación problemática y encontrar las correlaciones entre los atributos categóricos del conjunto de datos, que formaron la información obtenida del personal médico. Todo esto, a través de una aplicación web o sistema desarrollado especialmente para el caso. El sistema proporciona una explicación en lenguaje natural de los resultados de los modelos de minería, de modo que el patólogo los entienda. De esta manera, el especialista evalúa los resultados analizando subjetivamente, a partir de su experiencia y conocimiento, la información extraída por los modelos.

Atendiendo a los resultados del entrenamiento de los datos, es decir, realizando una valoración objetiva del rendimiento de los algoritmos, se demostró lo siguiente:

En cuanto a los Algoritmos de *Apriori*, *FPGrowth*, *Predictive Apriori* y *Tertius* para el conjunto 'C' resultó más eficiente *FPGrowth* por obtener reglas con mejor frecuencia y confianza, en segundo lugar quedó *Apriori*, ya que estos dos algoritmos generan casi las mismas reglas. En el *dataset* 'D' resultó ser mejor *Apriori* que *Tertius* porque generó reglas con mejor frecuencia de ocurrencia en los datos y por ser más rápido.

En la evaluación objetiva de las redes Bayesianas, se consideraron algoritmos de búsqueda como: *K2, Tan, TabuSearch, RepeatedHillClimber, LAGDHillClimber, HillClimber* con las 18 clases del conjunto de datos 'D'. Los mejores resultados para 14 de las clases fueron obtenidos con *Tan* y para las 4 restantes con *HillClimber*. Por este motivo, la generación de redes Bayesianas en la aplicación se implementó utilizando el algoritmo de búsqueda *Tan.*

Las reglas generadas por los modelos de asociación instrumentados para la investigación tuvieron un 75% de aprobación por parte del especialista cuando estas éstas se sometieron al análisis subjetivo por parte del doctor basado en su experiencia y conocimientos. En cuanto a los algoritmos, *Tertius* resultó ser el más preciso, porque el especialista aprobó el 90% de las reglas en el conjunto 'C' y 100% en el 'D'.

Las redes Bayesianas *mcc_aut* y *mcc_no_aut* se generaron con un 52% y 73% de precisión, respectivamente, atendiendo a la evaluación objetiva. Sin embargo, de acuerdo con el análisis de los resultados el doctor aprobó el 99.4% de los resultados para la primera mencionada y para la segunda el 97.8%.

El análisis de datos realizado en esta investigación mediante técnicas de MD, así como el conocimiento revelado por sus resultados, permiten demostrar cuán complejo es el problema abordado y reafirman la importancia de utilizar esquemas de extracción de conocimiento como el propuesto para formular soluciones reales. En tal sentido, los autores resaltan la eficiencia de las redes Bayesianas por la forma en la que representan la complejidad del tema tratado y por cómo muestran la información oculta en los datos de manera compresible y clara. En

resumen, la investigación constata las enormes posibilidades que brindan las redes Bayesianas para estudiar problemas complejos, así como la aplicabilidad de la MD en general.

5.2. Recomendaciones

Aun cuando esta investigación cumplió su objetivo y tuvo su logro, sería muy interesante profundizar en la problemática estudiada dándole una continuidad a este trabajo. Sería realmente beneficioso que se registraran un número mayor de encuestas para lograr mejor precisión en los modelos.

Además, como trabajo a futuro se sugiere estudiar datos de los registros clínicos de los pacientes fallecidos en el hospital, analizar con datos reales la tendencia de las causas que llevan a realizar autopsias en algunos pacientes y no en otros. Esto confirmará la veracidad de los resultados de esta investigación.

Como se dejó ver, en la revisión de trabajos relacionados, la disminución de autopsias está siendo tendencia no solo en México, muchos países se están viendo de igual manera afectados. Por ello, se propone utilizar el esquema presentado para investigar las causas en diferentes regiones y establecer comparativas para identificar si las opiniones médicas y las consecuencias del rechazo de las autopsias difieren según la región.

Dada la importancia de esta práctica médica, es necesario entender las causales que están haciendo que a nuestros días esté prácticamente en desuso. Se necesita de investigaciones como éstas que sirvan de herramientas para que las autoridades competentes comprendan los orígenes de la problemática, y en consecuencia establezcan medidas y lancen estrategias en aras de retomar el uso de las autopsias, examen vital para mantener y elevar la calidad de la medicina tradicional.

Productos Académicos

Elayne Rubio Delgado, Lisbeth Rodríguez Mazahua, Silvestre Gustavo Peláez Camarena, María Antonieta Abud Figueroa, José Antonio Palet Guzmán, Asdrúbal López Chau, Giner Alor Hernández.
Preliminary results of an analysis using association rules to find relations between medical opinions about the non-realization of autopsies in a Mexican hospital.
Research in Computer Science (ISSN 1870-4069).
Estado: *Publicado*

Elayne Rubio Delgado, Lisbeth Rodríguez Mazahua, Silvestre Gustavo Peláez Camarena, José Antonio Palet Guzmán, Asdrúbal López Chau.
Association analysis of medical opinions about the non-realization of autopsies in a Mexican hospital.
New Perspectives on Applied Industrial Tools and Techniques.
Estado: *Publicado*

Elayne Rubio Delgado, Lisbeth Rodríguez Mazahua, Silvestre Gustavo Peláez Camarena, José Antonio Palet Guzmán, Jair Cervantes, Asdrúbal López Chau.
Analysis of Medical Opinions about the Non-Realization of Autopsies in a Mexican Hospital Using Association Rules and Bayesian Networks.
Scientific Programming.
Estado: *Enviado*

Referencias

[1] J. Hurtado de Mendoza Amat, Autopsia. Garantía de calidad en la Medicina, La Habana: Editorial Ciencias Médicas, 2009, p. 200.

[2] J. D. D. Sampedro, *Estudio y aplicación de técnicas de aprendizaje automático orientadas al ámbito médico: estimación y explicación de predicciones individuales,* Madrid: Universidad Autónoma de Madrid, 2012.

[3] «Real Academia,» [En línea]. Available: http://dle.rae.es/. [Último acceso: Enero 2016].

[4] F. J. G. González, *Aplicación de técnicas de Minería de Datos a datos obtenidos por el Centro Andaluz de Medio Ambiente (CEAMA),* Granada, 2013.

[5] E. A. Oviedo Carrascal, A. I. Oviedo y G. L. Vélez Saldarriaga, «Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes.,» *Revista Politécnica ISSN 1900-2351(Impreso), ISSN 2256-5353 (En línea), Volumen 11, Número 20, 2015. pp. 111-120,* vol. 11, nº 20, pp. 111-120, 2015.

[6] C. C. Aggarwal y C. Zhai, Edits., Mining Text Data, Springer, 2012.

[7] D. Eneyamire Suleiman, «Reviving hospital autopsy in Nigeria: An urgent call for action,» *Annals of Nigerians Medicine,* vol. 19, nº 2, pp. 39-40, 2015.

[8] U. Bieri, H. Moch, S. Dehler, D. Korol y S. Rohrmann, «Changes in autopsy rates among cancer patients and their impact on cancer statistics from a public health point of view: a longitudinal study from 1980 to 2010 with data from Cancer Registry Zurich,» Springer, Berling, 2015.

[9] A. Turnbull, M. Osborn y N. Nicholas, «Hospital autopsy: Endangered or extinct?,» J Clin Pathol, 2015.

[10] H. Henshaw, L. Sharkey, D. Crowe y M. Ferguson, «The death of autopsy?,» *NIHR Nottingham Hearing Biomedical Research Unit,* vol. 386, 2015.

[11] A. Pervez Qasim, K. . U. Rehman Hashmi, M. Ahmad y K. Naheed, *THE VALUE OF AUTOPSY IN MEDICAL EDUCATION: STUDENT'S ATTITUDES & OPINION,* vol. 6, Punjab Medical College, 2015, pp. 17-25.

[12] D. Adeniyi, Z. Wei y Y. Yongquan, «Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method.,» *Original Research Article Applied Computing and Informatics,* vol. 12, pp. 90-108, 2016.

[13] R. Kaur y S. Singh, «A survey of data mining and social network analysis based anomaly detection techniques,» *Egyptian Informatics Journal, In Press, Corrected Proof,* nº 1-18, p. 2015.

[14] Z. J. Yu, F. Haghighat y C. M. F. Benjamin, «Advances and challenges in building engineering and data mining applications for energy-efficient communities.,» *Sustainable Cities and Society, In Press, Corrected Proof,* pp. 2-6, 2015.

[15] A. Capozzoli, D. Grassi y M. Savino Pi, «Discovering Knowledge from a Residential Building Stock through Data Mining Analysis for Engineering Sustainability.,» *Energy Procedia,* vol. 83, pp. 370-379, 2015.

[16] A. . H. Yousef., «Extracting software static defect models using data mining.,» *Ain Shams Engineering Journal,* vol. 6, pp. 133-144, 2015.

[17] N. . A. Shukor, Z. Tasir y H. Van der , «An Examination of Online Learning Effectiveness Using Data Mining.,» *Procedia - Social and Behavioral Sciences,* vol. 172, pp. 555-562, 2015.

[18] Harwati, Ardita Permata Alfiani y . F. Ayu Wula, «Mapping Student's Performance Based on Data Mining Approach (A Case Study).,» *Agriculture and Agricultural Science Procedia,* vol. 3, pp. 173-177, 2015.

[19] X.-J. G. J.-F. X. C. W. Y.-L. S. X.-F. Y. W. Q. X.-Q. M. W.-M. D. J. H. Chao Wang, «Exploration of the Association Rules Mining Technique for the Signal Detection of Adverse Drug Events in Spontaneous Reporting Systems,» *PLoS ONE,* vol. 7, nº 7, p. 6, 2012.

[20] B. Gutierrez, . N. Plant y E. Thieler., «A Bayesian network to predict coastal vulnerability to sea level rise.,» *J. Geophys. Res. Earth Surf.,* vol. 116, p. 15, 2011.

[21] T. Bulteau, A. Baills, L. Petitjean, M. Garcin, H., «Gaining insight into regional coastal changes on La Réunion island through a Bayesian data mining approach.,» *Geomorphology,* vol. 228, pp. 134-146, 2015.

[22] R. Timarán Pereira y M. C. Yépez Chamor, «La minería de datos aplicada al descubrimiento de patrones de supervivencia en mujeres con cáncer invasivo de cuello uterino.,» *Revista Universidad y Salud.,* vol. 14, pp. 117 - 129, 2012.

[23] E. A. Oviedo Carrascal, A. I. Oviedo y G. L. Vélez Saldarriaga, «Minería de datos: aportes y tendencias en el servicio de salud de ciudades inteligentes. ISSN 2256-5353 (En líne,» *Revista Politécnica ISSN 1900-2351(Impreso), ISSN 2256-5353 (En línea), Volumen 11, Número 20, 2015. pp. 111-120,* vol. 11, nº 20, pp. 111-120, 2015.

[24] D. Antonelli, . E. Baralis, . G. Bruno, L. Cagliero, T. Cerquitelli, . S. Chiusan, P. Garza y N. A. Mahoto., «MeTA: Characterization of Medical Treatments at Different Abstraction Levels.,» *ACM Trans. Intell. Syst. Technol. ,* nº 57, p. 25, 2015.

[25] N. C. K. M. M. D. W. Chih-Wen Cheng, «icuARM-II: improving the reliability of personalized risk prediction in pediatric intensive care units.,» *In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '14). ACM, New York, NY, USA,,* pp. 211-219, 2014.

[26] A. M. Franco Pérez y . E. León Guzmán, «An approach to the risk analysis of diabetes mellitus type 2 in a health care provider entity of Colombia using business intelligence.,» *In Proceedings of the 6th Euro American Conference on Telematics and Information Systems (EATIS '12), Rogerio Patricio Chagas do Nascimento (Ed.). ACM, New York, NY, USA,,* pp. 49-56, 2012.

[27] P. Miasnikof, V. Giannakeas, M. Gomes, L. Aleksandrowicz, A. Y. Shestopaloff, D. Alam, S. Tollman, A. Samarikhalaj y P. Jha, «- Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths,» *BMC Medicine,* vol. 13, nº 286, pp. 2-9, 2015.

[28] G.Sumalatha y D. Muniraj, «Survey on Medical Diagnosis Using Data Mining Techniques,» de *Proceedings of International Conference on Optical Imaging Sensor and Security*, Coimbatore, 2013.

[29] R. Sharma, S. Narayan y S. Khatri, «Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey,» de *Second International Conference on Computational Intelligence & Communication Technology*, 2016.

[30] G. Duan, D. Ding, Y. Tian y X. You, «An Improved Medical Decision Model Based on Decision Tree Algorithms,» de *IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)*, 2016.

[31] S. H. Song, Y. Choi y a. T. Yoon, «Comparison of episodes of mosquito borne disease: Dengue, Yellow Fever, West Nile, and Filariasis with Decision tree, Apriori Algorithm,» de *International Conference on Advanced Communications Technology (ICACT)*, 2016.

[32] Y. CHO, Y. AHN, S. YOON, J. KWON y T. YOON, «Analysis of Anti-cancer Cytokines by Apriori Algorithm, Decision Tree, and SVM,» de *BigComp*, 2015.

[33] A. K. Uysal, «An improved global feature selection scheme for text classification.,» *Expert*

Systems With Applications, vol. 43, pp. 82-92, 2015.

[34] G. Mujtaba, R. G. Raj, L. Shuib, R. Rajandram y K. Shaikh, «Automatic Text Classification of ICD-10 Related CoD from Complex and Free Text Forensic Autopsy Reports,» de *15th IEEE International Conference on Machine Learning and Applications*, 2016.

[35] G. Mujtaba, L. Shuib, R. G. Raj, R. Rajandram, K. Shaikh y M. A. Al-Garadi, «Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection.,» *PLoS ONE ,* vol. 12, nº 2, pp. 1-27, 2017.

[36] A. Onan, S. Korukoglu y H. Bulut, «Ensemble of keyword extraction methods and classifiers in text classification,» *Expert Systems With Applications,* nº 57, p. 232–247, 2016.

[37] S. K. Solanki y J. . T. Patel, *A Survey on Association Rule Mining,* Fifth International Conference on Advanced Computing & Communication Technologies ed., IEEE, 2015.

[38] J. Manimaran y T. Velmurugan, *A Survey of Association Rule Mining in Text applications,* IEEE , 2013.

[39] «JavaServer Faces.org,» [En línea]. Available: http://www.javaserverfaces.org/. [Último acceso: Septiembre 2016].

[40] C. Civicy, «Prime Faces User Guide 5.3,» [En línea]. Available: http://www.primefaces.org/docs/guide/primefaces_user_guide_5_3.pdf.

[41] J. T. Mora, «Arquitectura de software para aplicaciones web,» Ciudad México, 2011.

[42] R. Agrawal, T. Imielinski y A. Swami, «Mining association rules between sets of items in large databases,» Washington, DC, Proceedings of the ACM SIGMOD International Conference on Management of Data, 1993, pp. 207-216.

[43] J. Han, M. Kamber y J. Pei, Data Mining concepts and techniques, Elsevier, 2012.

[44] T. Scheffer, «Finding Association Rules that Trade Support Optimally Against Confidence,» pp. 424-435, 2001.

[45] P. A. FLACH y N. LACHICHE, «Confirmation-Guided Discovery of First-Order Rules with Tertius,» *Machine Learning,* nº 42, p. 61–95, 2001.

[46] P. Felgaer, «Optimización de Redes Bayesianas basado en Técnicas de Aprendizaje por Inducción,» *Reportes Técnicos en Ingeniería del Software,* vol. 6, nº 2, pp. 64-69, 2004.

[47] S. I. MARIÑO y P. L. ALFONZO, «SIMULACIÓN DEL RAZONAMIENTO EN EL PROCESO DE IDENTIFICACIÓN BOTÁNICA BASADO EN REDES BAYESIANAS,» *INVESTIGACION OPERATIVA,* vol. 24, nº 39, pp. 55-72, 2016.

[48] T. R. Patil y M. S. S. Sherekar, «Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification,» *International Journal Of Computer Science And Applications,* vol. 6, nº 2, pp. 256-261, 2013.

[49] H. Ibrahim, W. Yasin, N. I. Udzir y N. A. W. Abdul Hamid, «INTELLIGENT COOPERATIVE WEB CACHING POLICIES FOR MEDIA OBJECTS BASED ON J48 DECISION TREE AND NAÏVE BAYES SUPERVISED MACHINE LEARNING ALGORITHMS IN STRUCTURED PEER-TO-PEER SYSTEMS.,» *Journal of Information & Communication Technology,* vol. 15, nº 2, pp. 85-116, 2016.

[50] D. López, J. Hernández y E. Rivas, «Algorithm and Software Based on Multilayer Perceptron Neural Networks for Estimating Channel Use in the Spectral Decision Stage in Cognitive Radio Networks,» *International Journal of Computer, Electrical, Automation, Control and Information Engineering,* vol. 10, nº 12, pp. 1973-1977, 2016.

[51] D. D. Castillo, R. M. Pérez, L. H. Pérez, R. O. Morález y J. L. Ginori, «Algoritmos de aprendizaje automático para la clasificación de neuronas piramidales afectadas por el envejecimiento,» *Revista Cubana de Informática Médica,* vol. 8, nº 3, pp. 559-571, 2016.

Apéndice

Tabla A.1. Resultados de las redes Bayesianas para la clase *mcc_no_aut*

Clase: mcc_no_aut	Precisión	Área ROC	Tiempo
K2	0,644	0,929	360
Tan	0,729	0,958	299022
TabuSearch	0,679	0,933	92721
RepeatedHillClimber	0,679	0,933	640467
LAGDHillClimber	0,621	0,921	500
HillClimber	0,679	0,933	50533

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.2. Resultados de las redes Bayesianas para la clase *mcc_aut*

Clase: *mcc_aut*	Precisión	Área ROC	Tiempo
K2	0,533	0,779	133
Tan	0,515	0,788	129051
TabuSearch	0,534	0,782	31836
RepeatedHillClimber	0,534	0,782	158380
LAGDHillClimber	0,424	0,689	462
HillClimber	0,534	0,782	14666

Observaciones: *HillClimber* para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.3. Resultados de las redes Bayesianas para la clase *rechazo_fam*

Clase: *rechazo_fam*	Precisión	Área ROC	Tiempo
K2	0,526	0,718	97
Tan	0,601	0,748	91767
TabuSearch	0,506	0,724	38399
RepeatedHillClimber	0,506	0,724	238282
LAGDHillClimber	0,408	0,670	425
HillClimber	0,506	0,724	26329

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.4. Resultados de las redes Bayesianas para la clase *anios_prac*

Clase*: anios_prac*	Precisión	Área ROC	Tiempo
K2	0,956	0,996	72
Tan	0,998	0,999	70855
TabuSearch	0,964	0,998	30017
RepeatedHillClimber	0,964	0,998	121788
LAGDHillClimber	0,949	0,992	352
HillClimber	0,964	0,998	13304

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.5. Resultados de las redes Bayesianas para la clase *area*

Clase: *area*	Precisión	Área ROC	Tiempo
K2	0,982	0,998	52
Tan	0,999	1	48769
TabuSearch	0,989	0,999	165308
RepeatedHillClimber	0,989	0,999	57253
LAGDHillClimber	0,721	0,865	372
HillClimber	0,989	0,999	5913

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.6. Resultados de las redes Bayesianas para la clase *casos*

Clase: casos	Precisión	Área ROC	Tiempo
K2	0,944	0,997	76
Tan	0,994	0,999	72886
TabuSearch	0,961	0,998	36084
RepeatedHillClimber	0,961	0,998	164573
LAGDHillClimber	0,913	0,992	303
HillClimber	0,961	0,998	16567

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.7. Resultados de las redes Bayesianas para la clase *categoria*

Clase: *categoria*	Precisión	Área ROC	Tiempo
K2	0,972	0,995	70
Tan	0,996	0,999	68143
TabuSearch	0,984	0,996	30796
RepeatedHillClimber	0,984	0,996	120521
LAGDHillClimber	0,888	0,901	367
HillClimber	0,984	0,996	12384

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.8. Resultados de las redes Bayesianas para la clase *com_sug_op*

Clase: *com_sug_op*	Precisión	Área ROC	Tiempo
K2	0,960	0,995	250
Tan	0,970	0,999	303301
TabuSearch	0,969	0,996	100937
RepeatedHillClimber	0,969	0,996	626635
LAGDHillClimber	0,954	0,993	404
HillClimber	0,969	0,996	55224

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.9. Resultados de las redes Bayesianas para la clase *esc_esp*

Clase: *esc_esp*	Precisión	Área ROC	Tiempo
K2	0,997	0,999	478
Tan	0,999	0,999	550533
TabuSearch	0,999	0,999	135571
RepeatedHillClimber	0,999	0,999	441493
LAGDHillClimber	0,997	0,999	453
HillClimber	0,999	0,999	51733

Observaciones: *HillClimber* para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.10. Resultados de las redes Bayesianas para la clase *esc_med_gral*

Clase: *esc_med_gral*	Precisión	Área ROC	Tiempo
K2	0,978	0,999	466
Tan	0,997	0,999	452131
TabuSearch	0,983	0,999	134640
RepeatedHillClimber	0,983	0,999	826818
LAGDHillClimber	0,937	0,992	568
HillClimber	0,983	0,999	65817

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.11. Resultados de las redes Bayesianas para la clase *fmr_sol_aut*

Clase: fmr_sol_aut	Precisión	Área ROC	Tiempo
K2	0,442	0,812	109
Tan	0,450	0,843	105580
TabuSearch	0,475	0,817	21597
RepeatedHillClimber	0,475	0,817	120410
LAGDHillClimber	0,402	0,729	348
HillClimber	0,475	0,817	12637

Observaciones: *HillClimber* para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.12. Resultados de las redes Bayesianas para la clase *hall_arb*

Clase: *hall_arb*	Precisión	Área ROC	Tiempo
K2	0,963	0,997	89
Tan	0,996	0,999	65493
TabuSearch	0,980	0,998	38364
RepeatedHillClimber	0,980	0,998	205574
LAGDHillClimber	0,939	0,992	307
HillClimber	0,980	0,998	21312

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.13. Resultados de las redes Bayesianas para la clase *hall_dem*

Clase: *hall_dem*	Precisión	Área ROC	Tiempo
K2	0,970	0,998	113
Tan	0,995	0,999	77853
TabuSearch	0,975	0,999	19794
RepeatedHillClimber	0,975	0,999	68691
LAGDHillClimber	0,947	0,996	293
HillClimber	0,975	0,999	7450

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.14. Resultados de las redes Bayesianas para la clase *hall_dis*

Clase: hall_dis	Precisión	Área ROC	Tiempo
K2	0,898	0,981	87
Tan	0,990	0,999	66431
TabuSearch	0,959	0,992	33905
RepeatedHillClimber	0,959	0,992	206065
LAGDHillClimber	0,797	0,847	324
HillClimber	0,959	0,992	22511

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.15. Resultados de las redes Bayesianas para la clase *med_aut*

Clase: med_aut	Precisión	Área ROC	Tiempo
K2	0,498	0,755	155
Tan	0,570	0,775	288933
TabuSearch	0,514	0,758	50563
RepeatedHillClimber	0,514	0,758	321545
LAGDHillClimber	0,417	0,717	377
HillClimber	0,514	0,758	34482

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.16. Resultados de las redes bayesianas para la clase *no_hosp*

Clase: *no_hosp*	Precisión	Área ROC	Tiempo
K2	0,449	0,767	113
Tan	0,493	0.804	105658
TabuSearch	0,447	0,772	48647
RepeatedHillClimber	0,447	0,772	577330
LAGDHillClimber	0,333	0,729	362
HillClimber	0,447	0,772	38101

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.17. Resultados de las redes bayesianas para la clase *per_sol_aut*

Clase: *per_sol_aut*	Precisión	Área ROC	Tiempo
K2	0,877	0,891	83
Tan	0,887	0,915	165852
TabuSearch	0,861	0,897	46498
RepeatedHillClimber	0,861	0,897	716458
LAGDHillClimber	0,767	0,829	351
HillClimber	0,861	0,897	56120

Observaciones: Tan para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.

Tabla A.18. Resultados de las redes Bayesianas para la clase *ult_grado*

Clase: ult_grado	Precisión	Área ROC	Tiempo
K2	0,995	0,999	59
Tan	0,999	0.999	55411
TabuSearch	0,999	1	21571
RepeatedHillClimber	0,999	1	117975
LAGDHillClimber	0,993	0,999	310
HillClimber	0,999	1	10215

Observaciones: *HillClimber* para esta clase es el algoritmo seleccionado por presentar la mejor combinación de resultados.