# UNIVERSIDAD PÚBLICA DE EL ALTO

## INGENIERÍA DE SISTEMAS


**TESIS DE GRADO**

# "MODELO DE PREDICCIÓN SOBRE EL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA EN LAS MUJERES DE EDADES ENTRE 20 A 40 AÑOS DE LA CIUDAD DE LA PAZ, BASADO EN MINERÍA DE DATOS"

Para optar al título de Licenciatura en Ingeniería de Sistemas

**MENCIÓN:** Gestión y Producción


| | | |
|---|---|---|
| **POSTULANTE** | **:** | Miriam Apaza Ajnota |
| **TUTOR METODOLÓGICO** | **:** | Ing. Marisol Arguedas Balladares |
| **TUTOR ESPECIALISTA** | **:** | Ing. Enrique Flores Baltazar |
| **TUTOR REVISOR** | **:** | Ing. Julio Mamani Luque |



La Paz – Bolivia

2020

## RESUMEN

En la actualidad el cáncer de mama es uno de los problemas que más impacto tiene en la salud de las mujeres de la ciudad de La Paz y el índice de crecimiento va aumentando ya que no existe buena información sobre la enfermedad, y a medida que transcurre el tiempo se va incrementando la tasa de mortalidad y perdida de seno por lo tanto se hace difícil controlar el problema.

El presente trabajo de investigación tiene como finalidad realizar un Modelo de Predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos, para la toma de decisiones por un intervalo de 5 años con el propósito de evitar futuras muertes.

Por medio de la presente investigación se realizara un Modelo Predictivo, con la aplicación de técnicas y algoritmos de Minería de Datos, además se utilizaran las herramientas de software libre como ser: Windows 10, Lenguaje de Programación Java, NetBeans, Interfaz gráfica Weka, Sublime Text y las metodologías CRISP-DM (Cross Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Databases), ASD (Desarrollo de Software Adaptable) y el método científico.



**Palabras Claves:** Modelo Predictivo, Cáncer de mama, Minería de Datos, Proceso KDD y CRISP-DM.

## ABSTRACT

Currently, breast cancer is one of the problems that has the most impact on the health of women in the city of La Paz and the growth rate is increasing since there is no good information on the disease, and as the Over time the mortality and breast loss rate increases, therefore it is difficult to control the problem.

The purpose of this research work is to carry out a prediction model of the growth rate of breast cancer in women between the ages of 20 and 40 in the city of La Paz, based on Data Mining, for decision-making by an interval of 5 years in order to avoid future deaths.

Through this research, a Predictive Model will be carried out, with the application of Data Mining techniques and algorithms, and free software tools will be used, such as: Windows 10, Java Programming Language, NetBeans, Weka graphical interface, Sublime Text and CRISP-DM (Cross Industry Standard Process for Data Mining), KDD (Knowledge Discovery in Databases), ASD (Adaptive Software Development) and the scientific method.



**Key Words:** Predictive Model, Breast Cancer, Data Mining, KDD Process and CRISP-DM.

## DEDICATORIA

*El presente Tesis de Grado dedico principalmente a Dios, por ser el inspirador y darme fuerzas.*

*A mis padres Guillermo y Cristina, por su amor, cariño y sacrificio en todos estos años que me han brindado su apoyo incondicionalmente.*

*A mis hermanos Susy, Edwin y Elmer, que siempre me brindaron su apoyo.*

*Y especialmente a mi querido hijo Yahir por darme la fortaleza para seguir adelante en la culminación de mis estudios.*

## AGRADECIMIENTOS

*Primeramente agradezco a Dios por darme fortaleza y sabiduría.*

*Con mucho cariño a mis padres Guillermo y Cristina por el apoyo incondicional que me brindaron para seguir adelante con culminación de mis estudios.*

*Al Ing. Marisol Arguedas Balladares como Tutor Metodológico por guiarme y sobre todo su paciencia y colaboración.*

*Al Ing. Enrique Flores Baltazar como Tutor Especialista por su colaboración y tiempo, que me oriento al desarrollo de mi Tesis de Grado.*

*Al Ing. Julio Mamani Luque como Tutor Revisor por su tiempo y colaboración en la revisión de mi Tesis de Grado.*

*Y finalmente a la Universidad Pública de El Alto, por haberme cobijado y brindado una formación académica.*

# ÍNDICE

# ÍNDICE DE TABLAS

## ÍNDICE DE FIGURAS

# Capítulo I: Marco Preliminar

**Resumen**

En el presente capitulo se plantea el trabajo de investigación "Modelo Predictivo del índice de crecimiento del cáncer de mama de mujeres entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos", planteando el problema principal , los objetivos y mostrando las metodologías, técnicas y herramientas que se aplicaran.

## 1.1 Introducción

El cáncer de mama es uno de los problemas que más impacto tiene en la salud de las mujeres, a pesar de avances tecnológicos su propósito sigue dependiendo principalmente del estadio de la enfermedad en el momento de la detección.

De acuerdo a datos del Registro Nacional de Cáncer de Base Poblacional del Ministerio de Salud el número de casos nuevos de cáncer en mujeres registrados en los últimos tres años (2016 al 2018) asciende al 18.118 siendo, la tasa cruda de incidencia de cáncer es de 351.35 por cada 100.000 habitantes en el sexo femenino y la tasa cruda por cáncer de mama es de 159.70 por cada 100.000 habitantes, se afirma que el segundo caso de cáncer más frecuentes en mujeres es el cáncer de mama representando el 14% del total de los canceres en este sexo. (Ministerio de Salud, 2018).

En el Departamento de La Paz en las últimas tres gestiones se registraron 1.025 casos de cáncer de mama, para la gestión 2016 se registraron 384 casos, 2017 se

registraron 369 y 2018 se registraron 272 casos nuevos de cáncer de mama. (Ministerio de Salud, 2018).

Actualmente en la ciudad de La Paz el cáncer de mama su índice de crecimiento va aumentando ya que no existe buena información sobre la enfermedad, y a medida que transcurre el tiempo se va incrementando la tasa de mortalidad y perdida de seno por lo tanto se hace difícil controlar el problema (Periódico Pagina Siete, 19 de octubre de 2015).

Por medio de la presente investigación se proporcionará un Modelo para la predicción del índice de crecimiento del cáncer de mama de las mujeres de edades 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos, para la toma de decisiones y prevenir muertes futuras.

Se utilizara las herramientas de Windows 10, Lenguaje de Programación Java, NetBeans 8.2, Interfaz gráfica Weka, Sublime Text y la metodología CRISP-DM), metodología ASD y el método científico.

## 1.2 Antecedentes

### *1.2.1 Local*

- En la Universidad Pública de El Alto por: Mamani Mamani Víctor de la gestión 2012 presento el tema "MODELO PREDICTIVO BASADO EN REDES NEURONALES ARTIFICIALES PARA EL CRECIMIENTO VEGETATIVO DE ESTUDIANTES DE LA UPEA". Presenta un modelo que pueda pronosticar el crecimiento vegetativo estudiantil para el año 2020.con la ayuda de redes neuronales artificiales la cantidad de estudiantes.

- En la Universidad Pública de El Alto por: Hidalgo de la gestión 2012 presento el tema "APLICACIÓN DE MINERIA DE DATOS PARA EL ANALISIS DE RENDIMIENTO ACADEMICO PARA ESTUDIANTES DE SECUNDARIA" CASO DISTRITO 2 DE EL ALTO. Con el objetivo de presentar un prototipo llamado RendAC, la cual extrae información útil a partir de almacenamiento de grandes cantidades de datos estudiantiles.

### *1.2.2 Nacional*

- En la Universidad Mayor de San Andrés por: Apaza Quevedo Boris Rodrigo de la gestión 2009 presento el tema: "MODELO DE PREDICCIÓN DE PREVALENCIA DE ENFERMEDADES PARA CENTROS DE SALUD BASADO EN MINERÍA DE DATOS" La presente investigación centra su atención en la información almacenada en bases de datos de centros de salud en Bolivia a los que la ONG Medicus Mundi presta servicio. Información de pacientes que acuden al centro de salud, síntomas y tratamiento de enfermedades, estos datos no se utilizan para identificar patrones o información novedosa.
- En la Universidad Mayor de San Andrés por : Quispe Condori Rubén de la gestión 2011 presento el tema : "PREDICCIÓN DEL CONSUMO DEL AGUA POTABLE DE LA CIUDAD DE LA PAZ APLICANDO REDES NEURONALES Y LÓGICA DIFUSA" La aplicación de redes neuronales y lógica difusa al consumo de agua potable , permitirá predecir la demanda del agua en los siguientes años , con una efectividad del 60%.

### *1.2.3 Internacional*

- En la Universidad Mayor de San Andrés por : Ayala Jiménez Luis de la gestión 2009 presento el tema : "MODELO DE REDES NEURONALES PARA LA PREDICCIÓN DE LA VARIACIÓN DEL VALOR DE LA ACCIÓN DE FIRST SOLAR" Es posible obtener rendimientos superiores al mercado, transando con acciones volátiles de manera diaria y con un modelo de predicción basado en redes neuronales. Este supuesto se basa en la idea de que el modelo de redes es capaz de captar patrones conductuales que otorga el mercado al realizar transacciones con dichos instrumentos.
- En la Universidad Mayor de San Andrés por : Garduño García Gabriela de la gestión 2011 presento el tema : "METODOLOGÍA PARA CALCULAR EL PRONÓSTICO DE VENTAS Y UNA MEDICIÓN DE SU PRECISIÓN EN UNA EMPRESA FARMACÉUTICA: CASO DE ESTUDIO" En el presente trabajo se propone una metodología para el cálculo del Pronóstico de Ventas de una empresa de Sector Farmacéutico que contribuya a disminuir el abastecimiento excesivo que generalmente existe en la compañía y con esto minimizar los costos, como el de almacenamiento de productos terminados y materiales.

## 1.3 Planteamiento del Problema

Actualmente el índice de crecimiento sobre el cáncer de mama es uno de los problemas fundamentales en la ciudad de La Paz, que causa la muerte en las mujeres; por esta razón, existe una amplia evidencia de que la detección temprana juega un papel importante en la reducción de la mortalidad de este cáncer en la ciudad de La Paz.

Actualmente las instituciones de salud de la ciudad de La Paz no cuentan con un Modelo Predictivo que pueda predecir el índice de crecimiento del cáncer de mama de las mujeres, ya que se ha visto un gran cantidad de personas de diferentes edades que padecen este cáncer, por lo que se considera realizar este modelo principalmente de la mujeres de edades entre 20 a 40 años, con el fin de realizar una buena toma de decisiones para la prevención de las muertes.

### *1.3.1 Problema Principal.*

El problema principal es que actualmente no existe un Modelo Predictivo que pronostique el Índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años en las instituciones de salud de la ciudad de La Paz, para la prevención de futuras muertes.

### *1.3.2 Problemas Secundarios*

- La falta de aplicación de nuevas técnicas y diseño que puedan predecir el índice de crecimiento del cáncer de mama, para prevenir muertes futuras.
- Existencia de datos desorganizados del índice de crecimiento del cáncer de mama en diferentes instituciones de salud de la ciudad de La Paz, lo que no permite tomar decisiones óptimas.
- Datos no representados o interpretados en diferentes instituciones de salud sobre el factor de riesgo del cáncer de mama.
- Falta de aplicación de nuevos TIC que constituyen de herramientas y programas que administran, transmitan y compartan la información mediante soporte tecnológico

**¿De qué manera ayudaría un Modelo de Predicción sobre el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años en la Ciudad de La Paz, basado en Minería de Datos, para la toma de decisiones para prevenir las muertes?**

## 1.4 Objetivos

### *1.4.1 Objetivo General*

Aplicar un Modelo de Predicción sobre el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos, para la toma de decisiones para prevenir las muertes por esta patología.

### *1.4.2 Objetivos Específicos*

- Investigar y analizar el índice de crecimiento actual sobre el cáncer de mama en las mujeres de edades entre 20 a 40 años en la ciudad de La Paz para prevención de muertes.
- Determinar cuál es el factor principal del índice de crecimiento del cáncer de mama mediante el estudio de la investigación.
- Diseñar un Modelo de predicción del índice de crecimiento del cáncer de mama aplicando pruebas de Algoritmo de Minería de Datos.
- Seleccionar las herramientas y software libres mediante el estudio de las alternativas en el campo del modelado.
- Aplicar técnicas para el diseño del Modelo.
- Realizar las estrategias del modelado para la predicción.

## 1.5 Hipótesis

El Modelo Predictivo con la ayuda de Minería de Datos demuestra que existe un alto índice de crecimiento del cáncer de mama en mujeres de edades entre 20 a 40 años en la ciudad de La Paz con una eficiencia de 95% por un intervalo de 5 años.

Con la presente tesis se demostrara la veracidad de la tesis Nula.

### *1.5.1 Operación de Variables*

**Tabla 1**

*Operación de Variables*

| **Operación de variables** | |
|---|---|
| **Variable Dependiente** | Índice de crecimiento. |
| **Variable Independiente** | Modelo Predictivo<br>Cáncer de mama en mujeres de edades entre 20 a 40 años en la ciudad de La Paz. |
| **Variable Interrelacionado** | Predicción<br>Minería de Datos |

Fuente: (Elaboración Propia)

## 1.6 Justificación

### *1.6.1 Justificación Científica*

Para la presente investigación se analizará el comportamiento sobre el índice de crecimiento del cáncer de mama en las mujeres de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años, por lo que obtendremos datos predichos, para así realizar una toma de decisiones para la prevención de las muertes.

La propuesta de construir un modelo de predicción del cáncer de mama es importante en el sentido de que se lo construye aplicando la técnica en la minería de datos.

### *1.6.2 Justificación Técnica*

El trabajo se justifica técnicamente por la utilización de herramientas de software libre para el diseño del Modelo Predictivo como: Lenguaje de Programación Java, Netbeans, Weka, Sublime Text 3 y la Metodología CRISP-DM (Cross Industry Standard Process for Data Mining), Metodología de KDD (Knowledge Discovery in Databases), Metodología ASD (Knowledge Discovery in Databases), el Método Científico.

### *1.6.3 Justificación Económica*

La presente investigación se justifica económicamente al proponer un Modelo de Minería de Datos como producto final para poder predecir el índice de crecimiento del cáncer de mama de las mujeres de 20 a 30 años de la ciudad de La Paz, basado en Minería de Datos, por lo que no se tendrá ningún costo de las herramientas de software libres que se utilizaran para el diseño del Modelo de Predicción.

Así mismo las mujeres de edades entre 20 a 40 años de la ciudad de La Paz podrán conocer si corren el riesgo de padecer el cáncer de mama a tiempo y así lograr menores gastos por tratamiento a tiempo.

### *1.6.4 Justificación Social*

El modelo de predicción sobre el índice de crecimiento del cáncer de mama de las mujeres de la ciudad de La Paz, colaborará para una mejor prevención, motivará a futuros estudios sobre esta enfermedad y beneficiara en la prevención de mujeres de edades entre 20 a 40 años.

## 1.7 Metodología

Para realizar el Modelo de predicción se utilizarán las siguientes metodologías:

- Método Científico
- Metodología (CRISP-DM)
- Metodología ASD (Software de Desarrollo Adaptable)
- Métrica de calidad ISO/IEC 9126
- Modelo Cocomo II

### *1.7.1 Método Científico*

El presente trabajo a desarrollar utiliza el método científico bajo el enfoque cualitativo estableciendo con las siguientes características.

- **Observación y Elección del problema a investigar**

La minería de datos ayudaría a predecir el índice de crecimiento del cáncer de mama de las mujeres de la ciudad de La Paz donde se puede observar un nivel elevado de cáncer de mama y aun no dan solución a este problema.

- **Planteamiento del problema**

Actualmente el incremento del cáncer de mama en la ciudad de La Paz es uno de los problemas fundamentales, esto debido a la falta de información y predicción sobre el cáncer de mama en mujeres de edades entre 20 a 40 años de la ciudad de La Paz.

- **Recopilación de datos**

La recolección de los datos se obtuvo de diferentes fuentes como ser del Hospital de la Mujer, Ministerio de Salud, Fundación sobre el cáncer de mama, Oncológico La Paz.

- **Planteo de hipótesis**

Con el modelo de predicción se pretende demostrar que existe un crecimiento severo del cáncer de mama de las mujeres del Municipio de La Paz.

Con la ayuda de la Minería de Datos, se podrá obtener un Modelo Predictivo con una eficiencia del 95% para la predicción del cáncer de mama de las mujeres de La Paz, basado en Minería de Datos.

- **Diseño de la Aplicación o Experimentación**

Para la experimentación del Modelo Predictivo del cáncer de mama de las mujeres de La Paz, se utilizaran varias herramientas libres para el modelado en base a la Minería de Datos, donde se podrá predecir el índice de crecimiento del cáncer de mama de la mujer de edades entre 20 a 40 años.

- **Análisis de Resultados.**

Se probará el Modelo Predictivo del índice de crecimiento del cáncer de mama de las mujeres de la ciudad de La Paz mediante las herramientas y técnicas.

- **Elaboración del reporte de investigación**

Se elabora el reporte final que se presentara, en base a los resultados del modelo predictivo con minería de datos.

### *1.7.2 La Metodología de Cross Industry Standard Process for Data Mining (CRISP-DM)*

Es un método probado para orientar sus trabajos de minería de datos. Como metodología, incluye descripciones de las fases normales de un proyecto, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas.

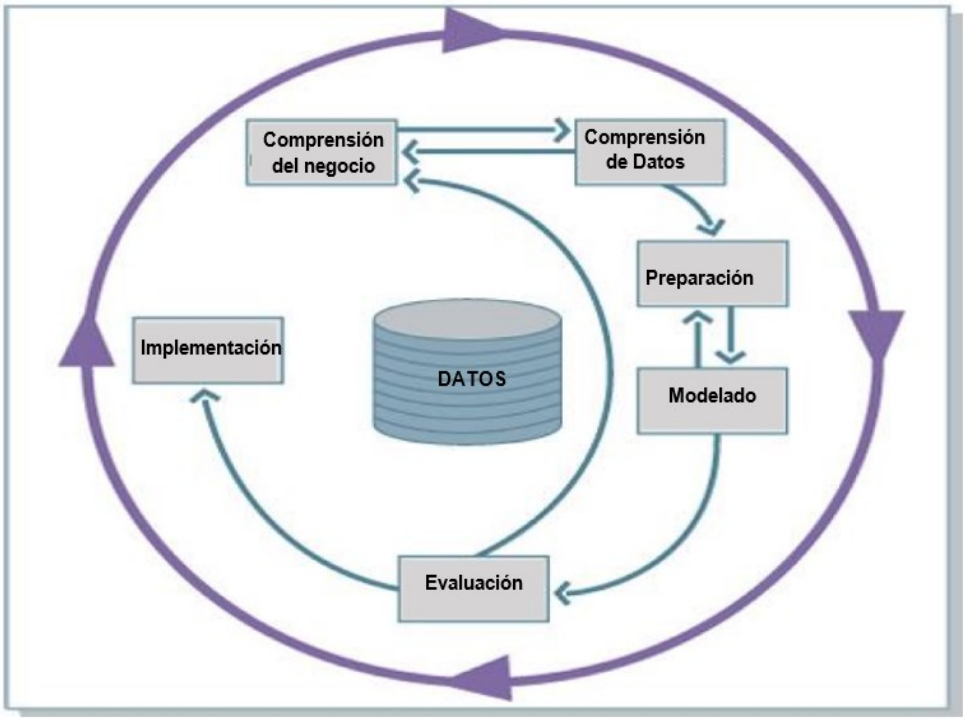
### 1.7.2.1 Fases de la Metodología CRISP-DM

El ciclo vital del modelo contiene seis fases con flechas que indican las dependencias más importantes y frecuentes entre fases. La secuencia de las fases no es estricta. De hecho, la mayoría de los proyectos avanzan y retroceden entre fases si es necesario.

CRISP-DM es la fase central del modelo e implica decidir tareas y algoritmos se divide el proceso de Minería de Datos en seis fases principales.

**Figura 1**

*Fases de la Metodología CRISP-DM*




*Fuente: (CRISP-DM, 2014)*

**a) Comprensión del Negocio**

(Objetivos y requerimientos desde una perspectiva no técnica)

- Establecimiento de los objetivos del negocio (Contexto inicial, objetivos, criterios de éxito)

- Evaluación de la situación (Inventario de recursos, requerimientos, supuestos, terminologías propias del negocio).
- Establecimiento de los objetivos de la minería de datos (objetivos y criterios de éxito)
- Generación del plan del proyecto (plan, herramientas, equipo y técnicas)

**b) Comprensión de Datos**

(Familiarizarse con los datos teniendo presente los objetivos del negocio)

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos
- Verificación de calidad de datos

**c) Preparación de datos**

(Obtener la vista minable o dataset)

- Selección de los datos
- Limpieza de datos
- Construcción de datos
- Integración de datos
- Formateo de datos

**d) Modelado**

(Aplicar las técnicas de minería de datos a los dataset)

- Selección de la técnica de modelado
- Diseño de la evaluación
- Construcción del modelo

- Evaluación del modelo

**e) Evaluación**

(De los modelos de la fase anteriores para determinar si son útiles a las necesidades del negocio)

- Evaluación de resultados
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones

**f) Implementación**

(Explotar utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización)

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Generación de informe final
- Revisión del proyecto

(Goicochea, 2009)

### *1.7.3 Metodología Ágil ASD (Adaptive Software Development)*

La técnica de Adaptive software Development fue desarrollada por Jim Highsmith y Sam Bayer a comienzos de 1990. Esta metodología se adapta al cambio en lugar de luchar contra él. Se basa en la adaptación continua a circunstancias cambiantes. En ella no hay un ciclo de planificación-diseño-construcción del software, sino un ciclo especular colaborar y aprender.

#### 1.7.3.1 Ciclo de Vida

ASD utiliza un "cambio orientado hacia el ciclo de vida", que tiene tres componentes que son: especular colaborar y aprender.

**1) Especular**

Una primera fase de iniciación para establecer los principales objetivos y metas del proyecto en su conjunto y comprender las limitaciones (zonas de riesgo) con las que operará el proyecto.

En ASD se realizan estimaciones de tiempo sabiendo que pueden sufrir desviaciones. Sin embargo, estas son necesarias para la correcta atención de los trabajadores que se mueven dentro de plazos de forma que puedan priorizar sus tareas.

Se decide el número de iteraciones para consumir el proyecto, prestando atención a las características que pueden ser utilizadas por el cliente al final de la iteración. Son por tanto necesarios, marcar objetivos prioritarios dentro de las mismas iteraciones.

Estos pasos se puede volver a examinar varias veces antes de que el equipo y los clientes están satisfechos con el resultado.

**2) Colaborar**

Es la fase donde se centra la mayor parte del desarrollo manteniendo una componente cíclica. Un trabajo importante es la coordinación que asegure que lo aprendido por un equipo se transmite al resto y no tenga que volver a ser aprendido por los otros equipos.

**3) Aprender**

La última etapa termina con una serie de ciclos de colaboración, su trabajo consiste en capturar lo que se ha aprendido, tanto positivo como negativo. Es un elemento crítico para la eficacia de los equipos.

Jim Highsmith identifica cuatro tipos de aprendizaje en esta etapa:

- **Calidad del producto desde un punto de vista del cliente.** Es la única medida legítima de éxito, pero además, dentro de las metodologías ágiles, los clientes tienen un valor importante.
- **Calidad del producto desde un punto de vista de los desarrolladores.** Se trata de la evaluación de la calidad de los productos desde un punto de vista técnico. Ejemplos de esto incluyen la adhesión a las normas y objetivos conforme a la arquitectura.
- **La gestión del rendimiento.** Este es un proceso de evaluación para ver lo que se ha aprendido mediante el empleo de los procesos utilizados por el equipo.
- **Situación del proyecto.** Como paso previo a la planificación de la siguiente iteración del proyecto, es el punto de partida para la construcción de la siguiente serie de características.

(Jim Highsmith , 1990)

### *1.7.4 Métrica de Calidad ISO/IEC 9126*

La ISO/IEC9126 o ISO 9126 (estándar internacional para la evaluación de la calidad de producto de software) publicado en 1992, fue sustituido por el proyecto SQuaRE, ISO 25000:2005, el cual sigue los mismos conceptos con el modelo de calidad de producto propuesto por McCall.

**Funcionalidad -** Un conjunto de atributos que se relacionan con la existencia de un conjunto de funciones y sus propiedades específicas. Las funciones son aquellas que satisfacen las necesidades implícitas o explícitas.

### *1.7.5 Modelo Cocomo II*

El Modelo Constructivo de Costos es un modelo matemático de base empírica utilizado para la estimación de costos de software.

- Esfuerzo para completar una actividad
- Tiempo calendario se necesita para completar

El Modelo Constructivo de Costos (COCOMO, por su acrónimo del inglés COnstructive COst MOdel) es un modelo matemático de base empírica utilizado para estimación de costos1 de software. Incluye tres sub-modelos, cada uno ofrece un nivel de detalle y aproximación, cada vez mayor, a medida que avanza el proceso de desarrollo del software:

- Se presentan tres niveles: básico, intermedio y detallado.
- Pertenece a la categoría de modelos estimadores basados en estimaciones matemáticas.

## 1.8 Herramientas

Las herramientas a utilizar son las siguientes:

Windows 10, es el vigente sistema operativo desarrollado por Microsoft como parte de la familia de sistemas operativos Windows NT.6 Fue dado a conocer oficialmente en septiembre de 2014, seguido por una breve presentación de demostración en la conferencia Build 2014. (Wikipedia, 2014)

Lenguaje de Programación Java, es un lenguaje de programación de propósito general, concurrente, orientado a objetos. (Fernández, 2005)

NetBeans 8.2 es un entorno de desarrollo integrado libre, hecho principalmente para el lenguaje de programación Java. Existe además un número importante de módulos

para extenderlo. NetBeans es un producto libre y gratuito sin restricciones de uso. (Dorado, 2015)

Interfaz gráfica Weka VERSION 3.4.12, es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL. (Orallo, 2006)

Sublime Text es un editor de texto y editor de código fuente está escrito en C++ y Python para los plugins. Desarrollado originalmente como una extensión de Vim, con el tiempo fue creando una identidad propia, por esto aún conserva un modo de edición tipo vi llamado Vintage mode. (Wikipedia, Sublime Text 3, 2016)

## 1.9 Límites y Alcances

### *1.9.1 Limites*

- El Modelo de Minería de Datos solo podrá predecir el índice de crecimiento del cáncer de mama de mujeres de la ciudad de La Paz.
- El presente Modelo de Predictivo solo podrá predecir el índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos con un intervalo de 5 años.
- El modelo propuesto no puede resolver otro problema que no se encuentre dentro el contexto.

### *1.9.2 Alcances*

- Se diseñará un Modelo de Minería de Datos en base a los datos adquiridos del cáncer de mama de las mujeres de la ciudad de La Paz.
- Se desarrollara el modelo predictivo en base a Minería de Datos.
- La presente propuesta de investigación se utilizará técnicas de Minería de Datos el cual nos ayudará a predecir el índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos para un intervalo de 5 años.

## 1.10 Aportes

El Modelo Predictivo ayudará a predecir el índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, para la toma de decisiones para prevenir las muertes.

El Modelo Predictivo basado en Minería de Datos será diseñado con la finalidad de brindar información sobre el comportamiento del índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz.

La ejecución del Modelo Predictivo nos permitirá prevenir las futuras muertes por el cáncer de mama.

# Capitulo II: Marco Teórico

**Resumen**

En el presente capítulo se presentan conceptos sobre el problema del cáncer de mama, modelos y técnicas que se utilizaran así también se aplicaran metodologías y posteriormente se introduce la teoría de conjuntos aproximados y fundamentos de Minería de Datos.

## 2.1 La Problemática del Cáncer de mama

El cáncer de mama en Bolivia, se convirtió desde hace mucho tiempo en un problema serio. De acuerdo con el Ministerio de Salud se había manifestado que, cada dos días fallecía una mujer a causa de esta enfermedad, pues en la mayoría de los casos las portadoras de este acuden demasiado tarde a los servicios médicos.

Pero, debido a estas inquietudes desde hace ya varios años se buscaba pasar de un sistema de salud basado en la enfermedad a uno enmarcado en la promoción de la salud y la prevención de las enfermedades, ya que un sistema con enfermos satura la consulta oncológica con casos avanzados e incurables, por eso es necesario contar con un método de atención gineco-oncológica más ocupado en el tratamiento de enfermedades tempranas a través de un programa de promoción, prevención, detección, control, seguimiento y de rehabilitación. Actualmente en Bolivia, la información sobre el cáncer de mama es bastante escasa; los siguientes datos son el resultado de investigaciones aisladas de instituciones y personas sensibilizadas con la problemática de esta patología. Pocos servicios del Sistema Nacional de Salud disponen de la tecnología adecuada (mamografía) para el tamizaje y el diagnóstico del cáncer de mama, lo cual excluye a la

población femenina de bajos ingresos, por el alto costo de este examen (accesibilidad económica).

En la ciudad de La Paz, el cáncer de mama ocupa el segundo lugar con el 20,4%, mientras que en el departamento de Oruro representa el 11,6% del total; en Potosí, la incidencia del cáncer de mama alcanzó a un 12,5%.

En Bolivia, el 26,57% por cada 100 mil mujeres han desarrollado cáncer de mama, mientras que la mortalidad por esta patología alcanza a 8,71% de cada 100 mil mujeres.

## 2.2 Cáncer de Mama

### *2.2.1 ¿Qué es el Cáncer de mama?*

El cáncer de mama consiste en un crecimiento anormal y desordenado de las células del tejido mamario, según la Organización Mundial de la Salud (OMS), el cáncer comienza en una célula. La transformación de una célula normal en tumoral es un proceso multifásico (que tiene muchas fases) y suele consistir en la progresión de una lesión precancerosa a un tumor maligno, de ahí nace la importancia de ser detectada a tiempo. (OMS, 2010)

Este tumor maligno se origina en las células del seno. Un tumor maligno es un grupo de células cancerosas que pudiera crecer hacia (invadir) los tejidos circundantes o propagarse (hacer metástasis) a áreas distantes del cuerpo. Esta enfermedad ocurre casi por completo en las mujeres, pero los hombres también la pueden padecer. (OMS, 2010)

### *2.2.2 Sobre la detección temprana*

El diagnóstico temprano sigue siendo una importante estrategia de detección precoz, particularmente en los países de ingresos bajos y medios, donde la enfermedad se diagnostica en fases avanzadas y los recursos son muy limitados. Algunos datos sugieren que esta estrategia puede dar lugar a un "descenso del estadio TNM" (aumento de la 14 proporción de cánceres de mama detectados en una fase temprana) de la enfermedad, que la haría más vulnerable al tratamiento curativo. (OMS, 2010)

### *2.2.3 Factores de Riesgo*

El factor de riesgo es la condición que favorece o en la que predomina la ocurrencia de la enfermedad, dentro de los factores más importantes para el cáncer de mama son: el sexo femenino y la edad. Sin embargo, no todas las mujeres que lo padecen tienen factores identificables, existen algunos que son modificables y otros que no lo son.

A continuación se detalla los factores de riesgos:

**Tabla 2**

*Factores de riesgo del cáncer de mama*

| | **Factores de riesgo** |
|---|---|
| **Edad** | Más del 80 por ciento de los casos de cáncer de mama ocurren ente las mujeres de 50 años en adelante. |
| **Genero** | “Las mujeres tienen una mayor probabilidad que los hombres de padecer cáncer de mama. Esta enfermedad ocasiona el 31% del total de los casos de cáncer invasivo en las mujeres y menos del 1% en los hombres”. |
| **Genes** | Ciertas mutaciones genéticas incrementan el riesgo de cáncer de mama. |

| | |
|---|---|
| **Historia familiar** | Tener una madre, hermana, o hija que padezca cáncer de mama incrementa el riesgo. Historia personal: Una mujer con cáncer en una mama tiene mayor riesgo de desarrollar cáncer en la otra mama o en otro, lugar del mismo seno. |
| **Raza** | Las mujeres blancas, incluidas las latinoamericanas, tienen un riesgo mayor que las mujeres negras o asiáticas de desarrollar cáncer de mama. |
| **Sexo** | La incidencia del cáncer de mama aumenta con la edad. La mayoría de los casos se diagnostican en mujeres mayores de 40 años, el 80%, en mayores de 50. |
| **Menstruación / menopausia:** | Las mujeres que tuvieron su primera menstruación antes de los 13 años o a quienes les llegó la menopausia después de los 50 se encuentran entre las mujeres que presentan un mayor riesgo de padecer la enfermedad |
| **Factores de riesgo que pueden ser modificados embarazo / lactancia** | Las mujeres que tienen un bebé antes de los 30 años, aquellas que tienen bastantes hijos y las que amamantan están en menos riesgo de padecer cáncer de mama. |
| **Anticonceptivos orales** | El uso prolongado y a largo tiempo de las píldoras anticonceptivas puede incrementar ligeramente, en algunos casos, el riesgo de padecer cáncer de mama. |
| **Terapia de reemplazo hormona** | Utilizar una terapia de reemplazo hormonal combinada (estrógeno más progestero-na) por más de 5 años, incrementa el riesgo de padecer cáncer de mama. |
| **Peso** | La obesidad puede incrementar el riesgo de padecer cáncer de mama, especialmente después de la menopausia. Alcohol: Tomar regularmente más de dos unidades de alcohol por día incrementa el riesgo de tener cáncer de mama. |
| **Radiación** | Mujeres expuestas a altos niveles de radiación, particularmente en el área del pecho (por ejemplo como parte del tratamiento de radiación para tratar otros tipos de cáncer en la niñez) se encuentra en mayor riesgo. |
| **Ejercicio** | A pesar de que aún resulta un tema controversial, muchos estudios han sugerido que el ejercicio puede proteger contra esta patología. |

Fuente: (Elaboración Propia)

#### 2.2.3.1 Barreras de Información

Puesto que la falta de conocimientos y conciencia acerca del cáncer de mama se constituye en uno de los principales obstáculos por lo que las mujeres no se someten a exámenes de detección. Los programas deben llegar a aquellas mujeres de más alto riesgo con mensajes que las impulsen a buscar servicios de prevención de esta enfermedad de manera temprana.

### *2.2.4 Síntomas y Signos*

En la valoración de los signos se considera la edad, factores de riesgo, oscilaciones temporales, bilateralidad, exámenes previos, desencadenantes y otros.

a) Masa Palpable o Engrosamiento Unilateral

b) Secreción por el Pezón

c) Dolor

d) Síntomas Cutáneos

### *2.2.5 Autoexploración Mamaria*

No hay datos acerca del efecto del cribado mediante autoexploración mamaria. Sin embargo, se ha observado que esta práctica empodera a las mujeres, que se responsabilizan así de su propia salud. En consecuencia, se recomienda la autoexploración para fomentar la toma de conciencia entre las mujeres en situación de riesgo, más que como método de cribado. (OMS, 2010)

Independientemente del método de detección precoz utilizado, dos aspectos esenciales para el éxito de la detección precoz poblacional son una atenta planificación y un programa bien organizado y sostenible que se focalice en el grupo de población

adecuado y garantice la coordinación, continuidad y calidad de las intervenciones en todo el continuum asistencial. (OMS, 2010)

### *2.2.6 Cáncer de mama en Bolivia*

La cantidad de casos nuevos de cáncer en mujeres asciende a 7830 por año, siendo las 10 localizaciones más importantes las que presentamos a continuación por orden de frecuencia. (Departamento de ENT, 2013)

**Tabla 3**

*Tasas Estandarizadas de Cáncer en sexo femenino Bolivia Vs América Latina*

| LOCALIZACIÓN(SITE) | TASA ESTANDARIZADA BOLIVIA | TASA ESTANDARIZADA AMÉRICA LATINA |
|---|---|---|
| Cuello de útero | 80,6 | 23,2 |
| **MAMA** | **45,7** | **64,4** |
| Vesícula biliar | 16,6 | 4,1 |
| Estomago | 10,6 | 13,4 |
| Ovario | 9,5 | 9,7 |
| Glándula Tiroides | 9,4 | 11,7 |
| Bronquios y Pulmón | 8,3 | 9,6 |
| Cuerpo del Útero | 8,9 | 6,9 |
| Recto | 6,2 | 7,0 |
| Colon | 4,5 | 12,6 |
| **Todos los sitios** | 261,3 | 236,1 |

Fuente: (Registro Nacional de Cáncer – Programa Nacional ENT – Ministerio de Salud, 2011)

Los datos epidemiológicos más relevantes son los siguientes (Registro Nacional de Cáncer – Programa Nacional ENT – Ministerio de Salud, 2011):

- La tasa de incidencia de cáncer de mama (casos nuevos) en el país asciende a 45,7 x 100.000, dándonos un total de 1.371 casos nuevos al año.
- La prevalencia estimada para cáncer de mama (casos nuevos) en el país es de 3.222 casos en el 2011.

- La mortalidad por cáncer en mujeres está dada principalmente por el cáncer de cuello uterino con el 11,4 %, vesícula biliar con 10,1 %, mama representa el 6,9% y colon con 4,9 %, entre las más importantes.
- Las mujeres en edad comprendida entre los 40 a 59 años de edad son las que tiene el mayor porcentaje de ocurrencia del cáncer de mama, llegando al 54,6 % del total de la población femenina.

**Tabla 4**

*Porcentaje de casos de Cáncer de mama por edad Gestión 2011*

| RANGO DE EDAD (años) | PORCENTAJE DE CASOS |
|---|---|
| 20 a 29 | 3,1% |
| 30 a 39 | 6,7% |
| 40 a 49 | 27,6% |
| 50 a 59 | 27,0% |
| 60 a 69 | 16,6% |
| ≥70 | 19,0% |

Fuente: (Registro Nacional de Cáncer – Programa Nacional ENT – Ministerio de Salud, 2011)

En la **Tabla 4** se observa los casos de cáncer de mama en determinados rangos de edad, a partir de los 20 años

**Tabla 5**

*Tasas de incidencia de Cáncer en Mujeres Gestión 2012*

| LOCALIZACION (SITE) | GESTION 2012 | | |
|---|---|---|---|
| | **Tasa Cruda** | **Tasa Estandarizada** | **Tasa Truncada 35–64** |
| Cuello del Útero | 77,40 | 94,83 | 219,55 |
| **Mama** | **59,70** | **71,04** | **146,49** |
| Piel | 27,61 | 32,59 | 42,40 |
| Vesícula Biliar | 24,78 | 30,35 | 39,57 |
| Sitio Primario Desconocido | 28,64 | 21,77 | 29,21 |
| Estomago | 13,45 | 15,82 | 21,67 |
| Ovario | 11,80 | 15,37 | 27,33 |
| Glándula Tiroides | 10,62 | 11,37 | 17,90 |
| Colon | 9,67 | 11,61 | 17,90 |
| Sist. Hematopoyético y Retic. Endotelial | 9,44 | 10,88 | 12,95 |
| **Todos los sitios** | **323,74** | **388,79** | **696,33** |

Fuente: (Departamento de ENT, 2012)

En la **Tabla 5** se observan las tasas de incidencia de cáncer en la gestión 2012 de la ciudad de La Paz.

**Tabla 6**

*Tasa de Mortalidad por Cáncer en mujeres de la ciudad de La paz enero – junio 2017*

| Grupos de edad en años | Número de casos | Población | Tasa | Tasa con redondeo |
|---|---|---|---|---|
| De 0 – 4 | 1 | 39421 | 0,254 | 0 |
| De 5 – 14 | 4 | 80705 | 0,496 | 0 |
| De 15 – 44 | 27 | 191281 | 1,412 | 1 |
| De 45 – 64 | 51 | 68372 | 7,459 | 7 |
| De 65 o mas | 95 | 32889 | 5 | 29 |
| **Total** | **178** | **412668** | **4,313** | **4** |

Fuente: (CRISP-DM, 2000)

Entonces categorizamos la variable edad de 15-44 años la tasa de mortalidad por cáncer es 1, esto quiere decir que 1 /10.000 mujeres mueren por cáncer en ese rango de edad; de 45 a 64 años la tasa de mortalidad es 7 lo que es igual a 7 /10.000 de 45 a 64 años mueren a causa del cáncer y de 65 años a más la tasa de mortalidad fue 29 es decir 29/10.000.

## 2.3 Modelos Predictivos del Riesgo

El modelo más utilizado es el modelo de riesgo de GAIL. Expresa el riesgo de desarrollar cáncer de mama invasivo a los 5 años. Este es un modelo predictivo validado, que utiliza 5 factores: edad actual, edad de la menarquia, biopsias previas en la mama, edad del primer parto de hijo vivo e historia familiar de cáncer de mama en parientes de primer grado. Un riesgo a los 5 años de 1,66% o mayor se relaciona con un riesgo elevado de cáncer de mama y es de considerar para realizar estrategias de quimio-prevención.

## 2.4 Modelo

Un modelo es un esquema teórico de un sistema o realidad compleja que se elabora para facilitar su compresión y estudio, cada una de las modalidades, tipos o categorías que existen de algo, en ingeniería se le llaman modelos a representación gráfica o esquemática de una realidad, sirve para organizar y comunicar de forma clara los elementos que involucran un todo. (Mayor, Pablo Grach, 2001)

Un modelo es una abstracción es decir una simplificación utilizada para comprender mejor la realidad que representa, simplifica la realidad, suprimiendo detalles irrelevantes y reteniendo los aspectos esenciales, de modo que la esencia del sistema sea

mejor conocida y podamos hacer frente a su complejidad, en otras palabras, el modelo es independiente de su representación gráfica o textual. (Booch, 1999)

En resumen, un modelo es la representación concisa de una situación, ya sea de un objeto, un esquema teórico una abstracción matemática, puede ser una réplica exacta por eso representa un medio de comunicación más eficiente y efectivo.

### *2.4.1 Modelos de Datos*

Uno de los pasos cruciales en la construcción de una aplicación que maneje una base de datos, es sin duda, el diseño de la base de datos, en donde lo más importante son los modelos de datos.

Un modelo de datos es una herramienta intelectual (un conjunto de conceptos y reglas) que permiten estructurar los datos resultantes de la observación de la realidad, de forma que queden representadas todas sus propiedades, tanto estáticas como dinámicas. (Celma Casamayor,Mota, 2003)

La modelización consiste en representar el problema realizando múltiples abstracciones para asimilar toda la información de un problema, y de esta manera. (Iván López Montalbán, 2014)

En resumen, un modelo de datos es un conjunto de conceptos y reglas que permiten estructurar los datos resultantes.

### *2.4.2 Conocimiento*

El conocimiento es una mezcla de experiencia, valores, información y "saber hacer" que sirve como marco para la incorporación de nuevas experiencias e información, y es útil para la acción. Se origina y aplica en la mente de los conocedores. En las organizaciones con frecuencia no sólo se encuentra dentro de documentos o

almacenes de datos, sino que también está en rutinas organizativas, procesos, prácticas, y normas. Vamos a intentar realizar una primera definición de conocimiento que nos permita comunicar que queremos decir cuando hablamos de conocimiento dentro de las organizaciones. (Davenport y Prusak, 1999)

Hechos o medidas que describen características de objetos, eventos o personas, es la materia prima de la que se obtendrá la información.

### *2.4.3 Datos*

Datos: hechos o medidas que describen características de objetos, eventos o personas, es la materia prima de la que se obtendrá la información.

Un dato no dice nada sobre el porqué de las cosas, y por sí mismo tiene poca o ninguna relevancia o propósito. (Prusak, 1999)

### *2.4.4 Información*

Datos analizados y presentados en forma adecuada, de interés para un observador en un momento determinado. (Aguilar, 1999)

Como han hecho muchos investigadores que han estudiado el concepto de información, lo describiremos como un mensaje, normalmente bajo la forma de un documento o algún tipo de comunicación audible o visible. Como cualquier mensaje, tiene un emisor y un receptor. (Aguilar, 1999)

## 2.5 Minería de Datos

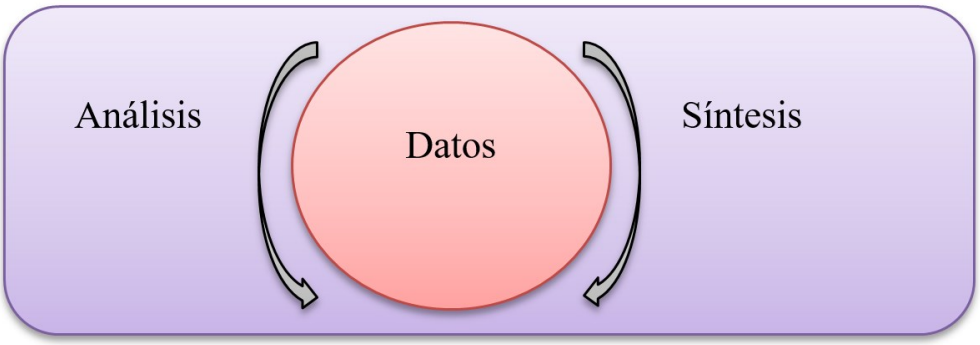
La minería de datos como proceso dual, según los autores [Fayyad, 1996] [Frawley, 1992], la Minería de datos (MDs) como análisis de información solo es un paso en todo el proceso de descubrimiento de conocimiento KDD.

La MDs como parte del proceso de descubrimiento de conocimiento y como a aplicación de algoritmos para obtener patrones más proclives traducida en conocimiento, elementalmente buscar predecir y describir. La predicción involucra el uso de algunas variables o atributos en el conjunto de datos para predecir otras variables de interés o carácter desconocidas. La descripción se enfoca en buscar patrones humanamente interpretables que divulguen a los datos. En ese sentido el uso de la minería de datos que puede entenderse como un proceso dual de síntesis (Predicción) y análisis (Descriptivo) sobre los datos.

Para obtener estos patrones y poder conseguir información relevante y de utilidad, la minería de datos dispone de varios métodos y algoritmos, que aplicados a grandes cantidades de datos son capaces de descubrir estos nuevos patrones tendencias ocultas. (Molina & Garcia, 2006)

**Figura 2**

*La Minería de Datos como proceso dual de análisis y síntesis sobre los datos*




Fuente: (Molina & García, 2006)

En Resumen, la Minería de Datos es un Conjunto de Técnicas y herramientas aplicada, con el objeto de predecir de forma automatizada en base a datos, el cual involucra varios pasos.

## 2.6 Clasificación de Datos

La clasificación se utiliza para clasificar un conjunto de datos basado en los valores de sus atributos. Por ejemplo, se podría clasificar a distintas personas para la otorgación de un préstamo en riesgo bajo, medio y alto, teniendo en cuenta información histórica de las mismas.

La clasificación encuentra las propiedades comunes entre un conjunto de objetos y los clasifica en diferentes clases, de acuerdo a un modelo de clasificación. Para construir este modelo, se utiliza un conjunto de entrenamiento, en el que cada instancia consiste en un conjunto de atributos y el valor de la clase a la cual pertenece. El objetivo de la clasificación es analizar los datos de entrenamiento y mediante un método supervisado, desarrollar una descripción o un modelo para cada clase utilizando las características disponibles en los datos. Esta descripción o modelo permite clasificar otras instancias, cuya clase es desconocida.

El método se conoce como supervisado debido a que, para el conjunto de entrenamiento, se conoce la clase de pertenencia y se le indica al modelo si la clasificación que realiza es correcta o no. La construcción del modelo se realimenta de estas indicaciones del supervisor. (Chen, 1996)

Los algoritmos mayormente utilizados para las tareas de clasificación son los algoritmos de inducción. En la actualidad existen numerosos enfoques de algoritmos de inducción y variedad en cada enfoque, el presente trabajo hará hincapié en aquellos orientados a generar árboles de decisión.

## 2.7 Aplicación de la Minería de Datos

**Tabla 7**

*Aplicación Minería de Datos*

| En internet | El mundo de los negocios: | En el mundo de las ciencias: |
|---|---|---|
| • **E-bussines:** Perfiles de clientes, publicidad dirigida, fraude.<br>• **Buscadores Inteligentes:** Generación de jerarquías, bases de conocimiento web.<br>• **Gestión del Tráfico de la Red:** Control de eficiencia y errores. | • **Banca:** Grupos de clientes, préstamos, oferta de productos.<br>• **Compañías de Seguros:** Detección de fraude, administración de recursos.<br>• **Marketing:** Publicidad dirigida, estudios de competencia. | • **Meteorología:** Tele conexiones (asociaciones espaciales), predicción.<br>• **Física:** Altas energías, datos de colisiones de partículas (búsqueda de patrones).<br>• **Bio-Informática:** Búsqueda de patrones en ADN, proyectos |

Fuente: (CRISP-DM, 2000)

### *2.7.1 Minería de Texto*

Examina una colección de documentos y descubre información no contenida en ningún documento individual de la colección; en otras palabras, trata de obtener información sin haber partido de algo. (Nasukawa & Nagano, 2001)

## 2.8 Análisis Olap

OLAP es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia de negocios (o Business Intelligence) cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

## 2.9 Técnicas de Minería de Datos Basadas en Aprendizaje Automático

Como ya se ha comentado, las técnicas de Minería de Datos (una etapa dentro el proceso completo de KDD intentan obtener patrones o modelos a partir de los datos recopilados. Decidir si los modelos obtenidos son útiles o no suele requerir una valoración subjetiva por parte del usuario. Las técnicas de Minería de

Datos se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o descriptivas.

Una técnica constituye el enfoque conceptual para extraer la información de los datos, y, en general es implementada por varios algoritmos. Cada algoritmo representa, en la práctica, la manera de desarrollar una determinada técnica paso a paso, de forma que es preciso un entendimiento de alto nivel de los algoritmos para saber cuál es la técnica más apropiada para cada problema. Asimismo es preciso entender los parámetros y las características de los algoritmos para preparar los datos a analizar.

Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones.

***Tabla 8***

*Técnicas de la Minería de Datos*

| | | | |
|---|---|---|---|
| **Técnicas** | **No supervisadas o Descriptivas** | **Clustering** | Numérico<br>Conceptual |
| | | **Asociación** | Probabilístico<br>A Priori |
| | **Supervisadas o Predictivas** | **Predictivas** | Regresión<br>Arboles de predicción<br>Estimador de núcleo |
| | | **Clasificación** | Tabla de decisión<br>Arboles de decisión<br>Introducción de reglas<br>Bayesianas<br>Basado en ejemplares<br>Redes neuronales<br>Lógica Borrosa<br>Técnica Genética |

Fuente: (CRISP-DM, 2000)

### *2.9.1 Aprendizaje Inductivo Supervisado*

En el aprendizaje inductivo supervisado existe un atributo especial, normalmente denominado clase, presente en todos los ejemplos que especifica si el ejemplo pertenece o no a un cierto concepto, que será el objetivo del aprendizaje. El atributo clase normalmente toma los valores + y -, que significan la pertenencia o no del ejemplo al concepto que se trata de aprender; es decir, que el ejemplo ejemplifica positivamente al concepto -pertenece al concepto o bien lo ejemplifica negativamente que no pertenece al concepto. Mediante una generalización del papel del atributo clase, cualquier atributo

puede desempeñar ese papel, convirtiéndose la clasificación de los ejemplos según los valores del atributo en cuestión, en el objeto del aprendizaje. (José M. Molina y Jesús García, 2012)

### *2.9.2 Aprendizaje Inductivo No Supervisado*

El aprendizaje inductivo no supervisado estudia el aprendizaje sin la ayuda del maestro; es decir, se aborda el aprendizaje sin supervisión, que trata de ordenar los ejemplos en una jerarquía según las regularidades en la distribución de los pares atributo-valor sin la guía del atributo especial clase. Éste es el proceder de los sistemas que realizan clustering conceptual y de los que se dice también que adquieren nuevos conceptos. Otra posibilidad contemplada para estos sistemas es la de sintetizar conocimiento cualitativo o cuantitativo, objetivo de los sistemas que llevan a cabo tareas de descubrimiento. (José M. Molina y Jesús García, 2012)

## 2.10 Modelo de Minería de Datos

### *2.10.1 Predicción*

Las predicciones se utilizan para prever el comportamiento futuro de algún tipo de entidad mientras que una descripción puede ayudar a su comprensión. De hecho, los modelos predictivos pueden ser descriptivos (hasta donde sean comprensibles por personas) y los modelos descriptivos pueden emplearse para realizar predicciones, de esta forma, hay algoritmos o técnicas que pueden servir para distintos propósitos, Por ejemplo, las redes de neuronas pueden servir para predicción, clasificación e incluso para aprendizaje no supervisado.

Los modelos predictivos siguen un aprendizaje supervisado, que consiste en aprender mediante el control de un supervisor o maestro que determina la respuesta que se desea generar del sistema. (Sivanandam, 2006)

El atributo a predecir se conoce como variable dependiente u objetivo, mientras que los atributos utilizados para realizar la predicción se llaman variables independientes o de exploración. (José M. Molina y Jesús García, 2012)

Este modelo se emplea para estimar valores futuros de variables de interés, el proceso se basa en la información histórica de los datos, mediante las cuales se predice un comportamiento de los datos, ya sea mediante clasificaciones, categorizaciones o regresiones. (Hernandez, 2006)

## 2.11 Técnicas de Minería de Datos

### *2.11.1 Regresión*

Esta técnica es muy parecida a la clasificación ya que a cada elemento se le asigna únicamente un valor de salida, con la diferencia de que este valor de salida es un valor numérico, es decir, puede ser un valor entero o real.

El análisis regresivo es una técnica utilizada para inter y extrapolar las observaciones, los cuales pueden clasificarse como regresión lineal o no lineal. Hablamos de modelo de regresión cuando la variable de respuesta y las variables explicativas son todas ellas cuantitativas. (Jose Hernandez Orallo, 2004)

### *2.11.2 Regresión Lineal*

La regresión lineal es una técnica estadística que intenta construir un modelo para los datos analizados, y a través de éste predecir los datos futuros. Este modelo cuantifica la relación entre dos variables continuas: “la variable dependiente o la variable que

intentamos predecir y la variable independiente o la variable predecible". (Jose Hernandez Orallo, 2004)

### *2.11.3 Regresión No Lineal*

En muchas ocasiones los datos no muestran una dependencia lineal. Esto es lo que sucede si, por ejemplo, la variable respuesta depende de las variables independientes según una función polinómica, dando lugar a una regresión polinómica que puede planearse agregando las condiciones polinómicas al modelo lineal básico. (Jose Hernandez Orallo, 2004)

### *2.11.4 Series Temporales*

Una serie temporal se define como una secuencia de $n$ observaciones o datos $xt$ ordenadas cronológicamente, sobre una característica (serie univariable) o sobre varias características (serie multivariable) de una unidad observable, tomadas en diferentes momentos. Las series temporales se caracterizan fundamentalmente por la gran numerosidad de los datos que la conforman, la alta dimensionalidad y la necesidad de su constante actualización. (César Soto Valero, 2016)

### *2.11.5 Forecast*

El forecasting o pronostico (como también es conocido) consiste en la previsión y estimación del producto o la demanda de un producto en el futuro, utilizando el histórico de ventas junto a las estimaciones y datos del departamento de marketing y todo y cualquier otro tipo de información útil para obtener una cifra más real posible.

### *2.11.6 Árboles de Decisión.*

Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos

diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema.

- Algoritmo ID3.
- Algoritmo C4.5

### *2.11.7 Modelos Estadísticos*

Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

### *2.11.8 Reglas de Asociación*

Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados. (Indurkhya, 1998)

- Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.
- Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos.

### *2.11.9 Redes Neuronales*

Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:

- El perceptrón.
- El perceptrón multicapa.
- Los mapas auto-organizados, también conocidos como redes de Kohonen.

## 2.12 Algoritmos de Minería de Datos

A continuación se describirá algunos algoritmos de Minería de Datos.

- **Algoritmo Cart**

Proviene de las siglas en inglés de "Classification and Regression Trees". Se caracteriza porque produce árboles binarios, es decir que cada nodo tiene sólo dos salidas posibles. La funcionalidad más importante de este algoritmo es la capacidad de generar árboles de regresión. (Ordoñez Briceño, 2013)

- **Algoritmo AQ15**

El AQ15 fue desarrollado por Michalski. Es un sistema de aprendizaje inductivo que genera reglas de decisión, donde el antecedente es una fórmula lógica. Una característica particular de este sistema es la inducción constructiva (constructive induction), es decir, el uso de conocimientos del dominio para generar nuevos atributos que no están presentes en los datos de entrada. (Magdalena, 2012)

- **Algoritmo CN2**

El sistema CN2, desarrollado por Clark y Niblett, es una adaptación del AQ15. La gran desventaja del AQ15 es que elimina los ruidos mediante pre y post procesamiento y no durante la ejecución del algoritmo. El objetivo del CN2 es, entonces, incorporar el manejo de datos ruidosos al algoritmo en sí. Combina entonces las técnicas de poda utilizadas en el ID3, con las técnicas de reglas condicionales utilizadas en el AQ15. (Magdalena, 2012)

- **RADIX/RX**

El sistema RX se utiliza para el descubrimiento de relaciones en bases de datos clínicas. La diferencia importante con otros sistemas es que incorpora la noción de tiempo: un dato es un conjunto de ejemplos que guardan información de un paciente en diferentes momentos, y los conocimientos generados son de naturaleza causal. El sistema divide su proceso de descubrimiento en dos etapas: primero genera hipótesis y, luego, utiliza técnicas avanzadas de estadística para validarlas. (Magdalena, 2012)

- **BACON**

El sistema BACON utiliza algoritmos de análisis de datos para descubrir relaciones matemáticas entre datos numéricos. Ha redescubierto leyes como la ley de Ohm para circuitos eléctricos y la ley de desplazamiento de Arquímides. Los datos de entrenamiento son numéricos y, normalmente, son generadas en algún experimento previo. (Magdalena, 2012)

- **Algoritmo SLIQ**

El algoritmo SLIQ (Supervised Learning In Quest) fue desarrollado por el equipo Quest de IBM. Este algoritmo utiliza los árboles de decisión para clasificar grandes cantidades de datos. El uso de técnicas de pre-ordenamiento en la etapa de crecimiento del árbol, evita los costos de ordenamiento en cada uno de los nodos. SLIQ mantiene una lista ordenada independiente de cada uno de los valores de los atributos continuos y una lista separada de cada una de las clases. (Magdalena, 2012)

- **Algoritmo A Priori**

La generación de reglas de asociación se logra basándose en un procedimiento de covering. Las reglas de asociación son parecidas, en su forma, a las reglas de

clasificación, si bien en su lado derecho puede aparecer cualquier par o pares atributo-valor. (Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo Prism**

PRISM es un algoritmo básico de aprendizaje de reglas que asume que no hay ruido en los datos. Sea t el número de ejemplos cubiertos por la regla y p el número de ejemplos positivos cubiertos por la regla. Lo que hace PRISM es añadir condiciones a reglas que maximicen la relación p/t (relación entre los ejemplos positivos cubiertos y ejemplos cubiertos en total). (Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo Part**

Uno de los sistemas más importantes de aprendizaje de reglas es el proporcionado por C4.5, explicado anteriormente. Este sistema, al igual que otros sistemas de inducción de reglas, realiza dos fases: primero, genera un conjunto de reglas de clasificación y después refina estas reglas para mejorarlas, realizando así una proceso de optimización global de dichas reglas. (Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo K-Means**

Es un algoritmo de clustering particional, uno de los métodos de clustering más conocidos y utilizados cuando todas las variables son de tipo cuantitativo. Funciona de forma iterativa, dividiendo óptimamente el conjunto inicial de datos en un número de clusters, el cual se indica como parámetro. (Álvarez, 2012)

- **Árboles de Decisión**

Corresponde a uno de los métodos inductivos de aprendizaje supervisado, el cual realiza divisiones sucesivas del conjunto de datos, utilizando algún criterio de selección, manteniendo organizada su estructura de forma jerárquica, con el fin de maximizar la

distancia entre los grupos de datos generados en cada iteración. (Álvarez, Algoritmos, 2013)

- **Algoritmo Two-Step Cluster**

Este algoritmo de clustering, a diferencia de muchos otros, fue diseñado para operar sobre conjuntos de datos muy grandes, y en comparación al algoritmo K-means tiene la ventaja de trabajar con variables continuas y categóricas, sin afectar la robustez de sus resultados. (Álvarez, Algoritmos, 2013)

- **Redes Neuronales Artificiales (Rna)**

Las redes neuronales son sistemas de procesamiento de datos, cuya estructura y diseño se basa en el proceso natural del funcionamiento del cerebro. Son muy interesantes de estudiar, ya que permiten modelar eficientemente problemas complejos, en los cuales se cuente con muchas variables predictoras. (Álvarez, Algoritmos, 2013)

- **Support Vector Machines (Svm)**

Es uno de los métodos de minería de datos más robustos y precisos, cuyo uso se ha vuelto muy popular para resolver problemas de clasificación y regresión. Su objetivo es encontrar la mejor función para clasificar un conjunto de datos, encontrando los hiperplanos que mejor dividan la muestra, maximizando el grado de separación entre las clases generadas. (Álvarez, Algoritmos, 2013)

- **Algoritmo 1R**

Es un clasificador muy sencillo, que únicamente utiliza un atributo para la clasificación. Sin embargo, aún hay otro clasificador más sencillo, el 0R, implementado en weka.classifers.ZeroR.java, que simplemente calcula la media en el caso de tener una

clase numérica o la moda, en caso de una clase simbólica. No tiene ningún tipo de opción de configuración. (Álvarez, Algoritmos, 2013)

- **Algoritmo Naive Bayesiano**

El algoritmo naive Bayesiano se encuentra implementado en la clase weka.classifiers.NaiveBayesSimple.java. No dispone de ninguna opción de configuración.

El algoritmo que implementa esta clase se corresponde completamente con el expuesto anteriormente. En este caso no se usa el estimador de Laplace, sino que la aplicación muestra un error si hay menos de dos ejemplos de entrenamiento para una terna atributo-valor-clase o si la desviación típica de un atributo numérico es igual a 0.

- **Algoritmo KNN**

En WEKA se implementa el clasificador KNN con el nombre IBk, concretamente en la clase weka.classifiers.IBk.java. Además, en la clase weka.classifiers.IB1.java hay una versión simplificada del mismo, concretamente un clasificador NN [Nearest Neighbor], sin ningún tipo de opción, en el que, como su propio nombre indica, tiene en cuenta únicamente el voto del vecino más cercano. (Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo Cobweb**

El algoritmo de COBWEB se encuentra implementado en la clase weka.clusterers.Cobweb.java. (Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo EM**

El algoritmo EM se encuentra implementado en la clase weka.clusterers.EM.java el algoritmo realiza un primer proceso consistente en obtener el número óptimo de clusters. (Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo ID3**

Este sistema ha sido el que más impacto ha tenido en la Minería de Datos. Desarrollado en los años ochenta por Quinlan, ID3 significa Induction Decision Trees, y es un sistema de aprendizaje supervisado que construye árboles de decisión. (Magdalena, 2012)

- **Algoritmo J48 o C4.5**

Es un algoritmo de inducción que genera una estructura de reglas o árbol a partir de subconjuntos (ventanas) de casos extraídos del conjunto total de datos de "entrenamiento".

El C4.5 es una extensión del ID3 que permite trabajar con valores continuos para los atributos, separando los posibles resultados en dos ramas: una para aquellos Ai<=N y otra para Ai>N. Este algoritmo fue propuesto por Quinlan en 1993. *(*Ordoñez Briceño, Karla Fernanda, 2013)

- **Algoritmo LMT (Logistic Model Tree)**

El LMT proporciona una descripción muy buena de los datos. Un LMT consiste básicamente en una estructura de un árbol de decisión con funciones de regresión logística en las hojas. (Garzon, Paula Andrea Vizcaino, 2008)

- **Algoritmo M5P (Árbol de regresión)**

Miguel Ángel Fuentes y Pablo Galarza citan “es un método de aprendizaje mediante árboles de decisión, utiliza el criterio estándar de poda M5”. (Garzon, Paula Andrea Vizcaino, 2008)

- **Algoritmo RandomTree**

Genera un modelo de árbol de decisión. La puntuación se predice basándose únicamente en este árbol de decisión. (Garzon, Paula Andrea Vizcaino, 2008)

- **Algoritmo NBTree (Naive Bayes Tree)**

Es un algoritmo hibrido, este genera un tipo de árbol de decisión, pero las hojas contienen un clasificador Naive Bayes construido. (Garzon, Paula Andrea Vizcaino, 2008)

- **Algoritmo REPTree (reducción de poda de errores)**

Genera múltiples árboles de decisión basados en diferentes iteraciones y elige el mejor árbol de decisión basado en el error cuadrático medio. El error cuadrático medio es una función de riesgo para calcular el árbol con la menor cantidad de errores. (Garzon, Paula Andrea Vizcaino, 2008)

- **Algoritmo RandomForest**

Según cita Francisco José Soltero y Diego José Bodas en su artículo Se basan en el desarrollo de muchos árboles de clasificación. Para clasificar un objeto desde un vector de entrada, se pone dicho vector bajo cada uno de los árboles del bosque. (Garzon, Paula Andrea Vizcaino, 2008)

- **Algoritmo SMOreg**

El algoritmo es seleccionado por el ajuste RegOptimizer. Es una implementación del algoritmo de optimización de secuencia mínima para formar un modelo de soporte de regresión de vectores, sus principales características son tratar con los valores perdidos y transformar atributos nominales en binarios. (Shevade, 2009)

## 2.13 Metodologías de Minería de Datos

### *2.13.1 Proceso de descubrimiento de Conocimiento en Base de Datos (Proceso KDD)*

El Proceso KDD (El Proceso de Descubrimiento de Conocimiento en Bases de Datos) es el proceso completo de extracción de información, que se encarga además de la preparación de los datos y de la interpretación de los resultados obtenidos. KDD se ha definido como "el proceso no trivial de identificación en los datos de patrones válidos, nuevos, potencialmente útiles, y finalmente comprensibles. Se trata de interpretar grandes cantidades de datos y encontrar relaciones o patrones.

Para conseguirlo harán falta técnicas de aprendizaje automático, estadística, bases de datos, técnicas de representación del conocimiento, razonamiento basado en casos [CBR, Case Based Reasoning], razonamiento aproximado, adquisición de conocimiento, redes de neuronas y visualización de datos. Tareas comunes en KDD son la inducción de reglas, los problemas de clasificación y clustering, el reconocimiento de patrones, el modelado predictivo, la detección de dependencias, etc. KDD es un campo creciente: hay muchas metodologías del descubrimiento del conocimiento en uso y bajo desarrollo. Algunas de estas técnicas son genéricas, mientras otros son de dominio específico.

Los datos recogen un conjunto de hechos (una base de datos) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese

subconjunto). KDD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. (José M. Molina y Jesús García, 2012)

#### 2.13.1.1 Etapas del Proceso KDD

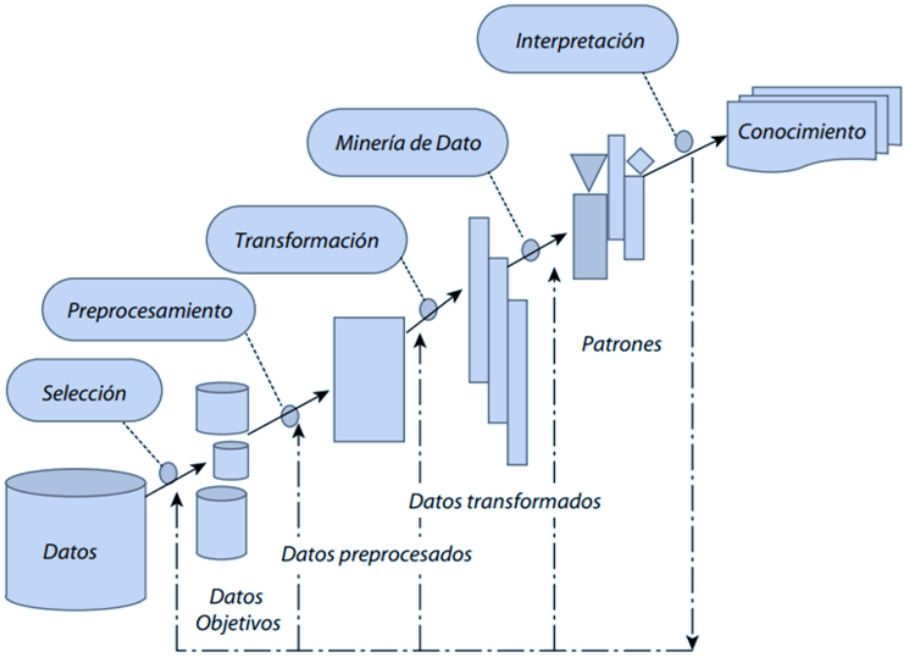
El proceso KDD es interactivo e iterativo, involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones.

Se resume en las siguientes etapas:

- Selección.
- Pre-procesamiento /limpieza.
- Transformación/reducción.
- Minería de datos (data mining).
- Interpretación/evaluación.

**Figura 3**

*Etapas del proceso KDD*




Fuente: (CRISP-DM, 2000)

Las Etapas del (Proceso KDD), tal como se muestra en la Figura 3 son:

**1) Etapa de selección**

En la etapa de selección, una vez identificado el conocimiento relevante y prioritario y definidas las metas del proceso KDD, desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos del negocio.

**2) Etapa de pre-procesamiento/limpieza**

En la etapa de pre-procesamiento/limpieza (data cleaning) se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos (missing y empty), datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. En esta etapa, es de suma importancia la interacción con el usuario o analista.

Los datos ruidosos (noisy data) son valores que están significativamente fuera del rango de valores esperados; se deben principalmente a errores humanos, a cambios en el sistema, a información no disponible a tiempo y a fuentes heterogéneas de datos. Los datos desconocidos empty son aquellos a los cuales no les corresponde un valor en el mundo real y los missing son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas gestores de bases de datos relacionales (sgbdr). En el proceso de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos.

### 3) Etapa de transformación/reducción

En la etapa de transformación/reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantesde los datos. (Fayyad et al.1996)

Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos horizontal o verticalmente. La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos (por ejemplo, edad por un rango de edades). La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (por ejemplo, edad y fecha de nacimiento). Se utilizan técnicas de reducción como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras. (Han y Kamber, 2001)

### 4) Etapa de minería de datos

El objetivo de la etapa minería de datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación (Quinlan, 1986) (Wang, Iyer y Scott, 1998), clustering (Ng y Han, 1994), (Zhang, Ramakrishnan, Livny, 1996), patrones secuenciales (Agrawal y Srikant, 1995) y asociaciones (Agrawal y Srikant, 1994), (Srikant y Agrawal, 1996), entre otras.

Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas, como por ejemplo predecir para nuevos clientes si son buenos o malos basados en su estado civil, edad, género y profesión, o determinar para nuevos estudiantes si desertan o no en función de su zona de procedencia, facultad, estrato, género, edad y promedio de notas. Entre las tareas predictivas están la clasificación y la regresión. Los modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos, como identificar grupos de personas con gustos similares o identificar patrones de compra de clientes en una determinada zona de la ciudad. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los clustering y las correlaciones.

**5) Etapa de interpretación/evaluación de datos**

En la etapa de interpretación/evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Esta etapa puede incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto. (Alvarado Pérez, J. C,, 2016)

### *2.13.2 Metodología Cross Industry Standard Process for Data Mining (CRISP-DM)*

CRISP-DM (Cross Industry Standard Process for Data Mining), es un modelo de proceso de minería de datos que describe una manera en la que los expertos en esta materia abordan el problema.

En 1993, líderes de la industria como Daimler Benz, SPSS de Inglaterra, OHRA de Holanda, NCR de Dinamarca, consorcio de empresas Europeas, y AG de Alemania construyeron el acrónimo CRISP-DM (Cross- Industry Standard Process for Data Mining), el cual tiene como finalidad proporcionar nuevas ideas a los que decidan trabajar con minería de datos. Esta metodología tiene la ventaja de que no ha sido construida de manera teórica y académica, sino que se basa en experiencias reales de cómo la gente hace proyectos (Moro, Laureano y Cortez, 2011) (Martínez y Podestá, 2014) (Raus, Vegega, Pytel y Pollo-Cattaneo, 2014). La metodología CRISP-DM provee una representación completa del ciclo de vida de un proyecto de DM, que se divide en seis fases, sus tareas y relaciones entre ellas.

#### 2.13.2.1 Ciclo de Vida de la Metodología CRISP-DM

CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software.

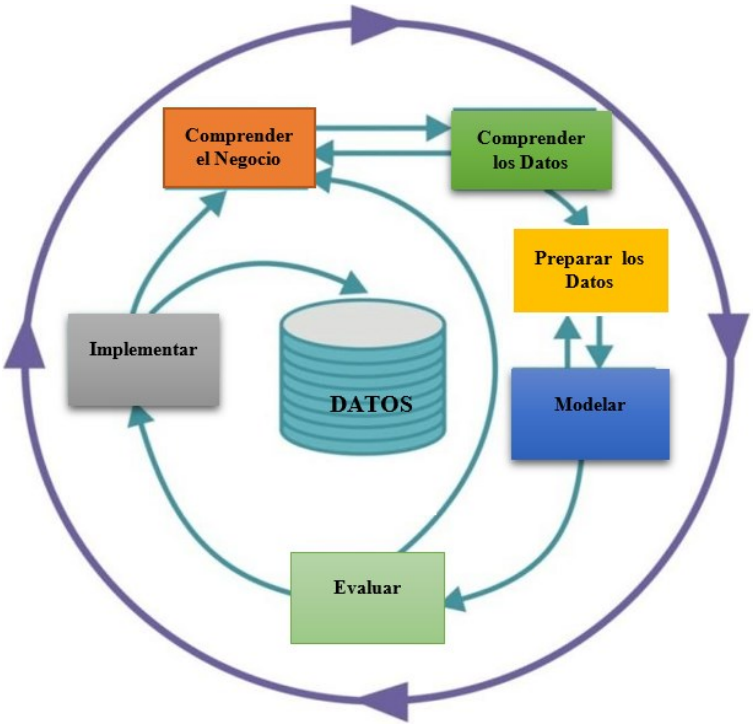
La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos.

**La secuencia de las fases no es rígida:** se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase, o qué tarea particular de una fase, hay que hacer después. Las flechas indican las dependencias más importantes y frecuentes.

El ciclo de vida del proyecto de minería de datos consiste en seis fases mostradas en la figura siguiente. *(Chapman, 2000)*

**Figura 4**

*Ciclo de Vida de la Metodología CRISP- DM*




Fuente: (CRISP-DM, 2012)

## a) Comprensión del negocio

Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.

**b) Comprensión de los datos**

La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

**c) Preparación de datos**

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final [los datos que serán provistos en las herramientas de modelado] de los datos en brutos iniciales.

**d) Modelado**

En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.

**e) Evaluación**

En esta etapa en el proyecto, usted ha construido un modelo (o modelos) que parece tener la alta calidad de una perspectiva de análisis de datos.

Antes del proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada.

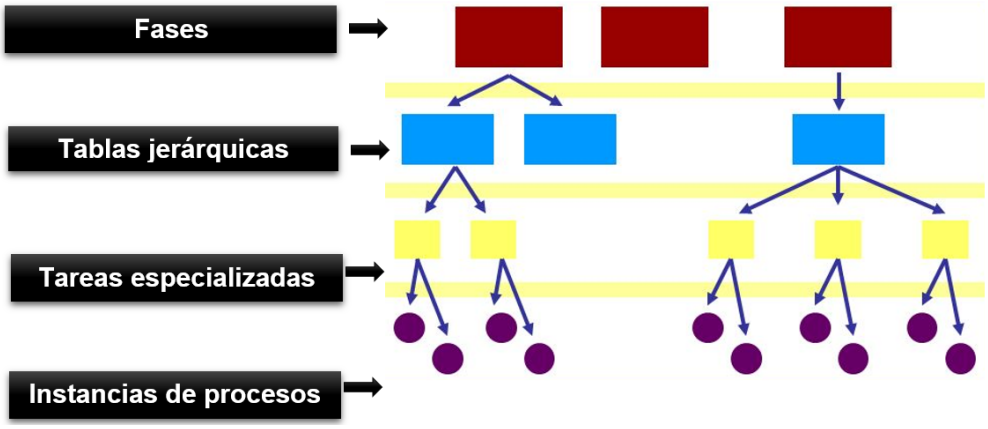
## f) Desarrollo

La creación del modelo no es generalmente el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento ganado tendrá que ser organizado y presentado en el modo en el que el cliente pueda usarlo. Ello a menudo implica la aplicación de modelos "vivos" dentro de un proceso de toma de decisiones de una organización, por ejemplo, en tiempo real la personalización de página Web o la repetida obtención de bases de datos de mercadeo.

Como ya sabemos la metodología CRISP consta de 4 niveles de abstracción organizados en forma jerárquica en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de minería de datos, en una serie de 6 fases:

**Figura 5**

*Esquema de los cuatro niveles de abstracción de la Metodología CRISP – DM*




Fuente: (CRISP-DM, 2000)

- **Fases:** Etapas del proceso
- **Tareas** genéricas: tareas generales, completas y estables
- **Tareas especializadas:** especificación de las tareas generales

- **Instancias de procesos**: acciones y decisiones concretas

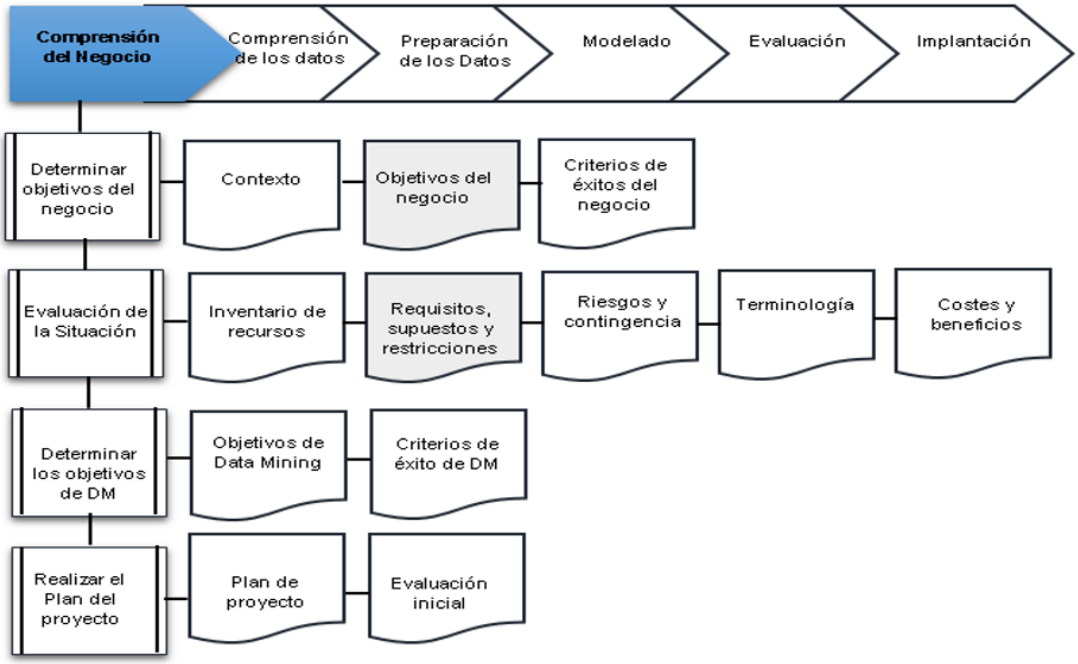
Las Metodología CRISP–DM (Cross Industry Standard Process for Data Mining) consta de 6 fases y cada una de ellas establecen tareas, las mismas que se describen a continuación:

### *2.13.3 Fases de la Metodología CRISP-DM*

**Fase 1.** Comprensión del negocio o problema.

**Figura 6**

*Fase de Comprensión del negocio*




Fuente: (CRISP-DM, 2000)

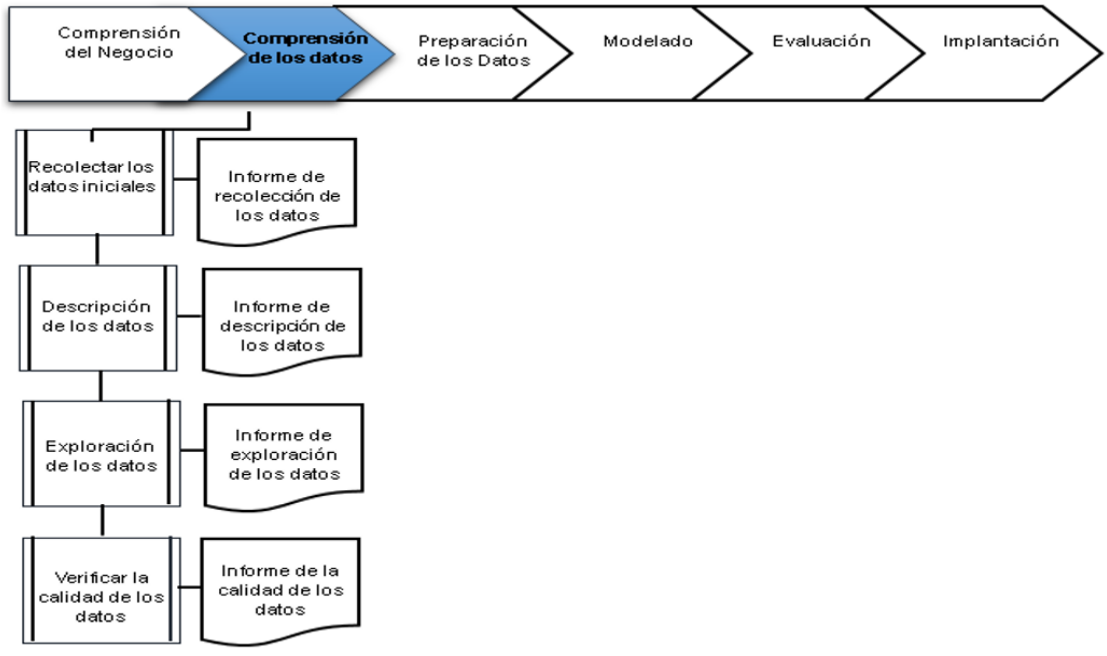
- **Determinar los objetivos.** Se determina cuál es el problema que se quiere resolver y por qué se usa minería de datos para dicho propósito; también se deben fijar los criterios de éxito. En cuanto a estos últimos, pueden ser de tipo cualitativo o de tipo cuantitativo.

- **Evaluar la situación actual.** En esta tarea se evalúan antecedentes y requisitos del problema, tanto en términos del negocio como en términos de la minería de datos. Algunos de los aspectos por tener en cuenta pueden ser el conocimiento previo acerca del tema, la cantidad de datos requeridos para resolver el problema.
- **Determinar los objetivos de la minería de datos.** Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de minería de datos.
- **Realizar el plan del proyecto.** Esta última tarea de la primera fase de CRISP-DM tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada uno de ellos.

**Fase 2.** Comprensión de los datos.

**Figura 7**

*Fase de Comprensión de los datos*



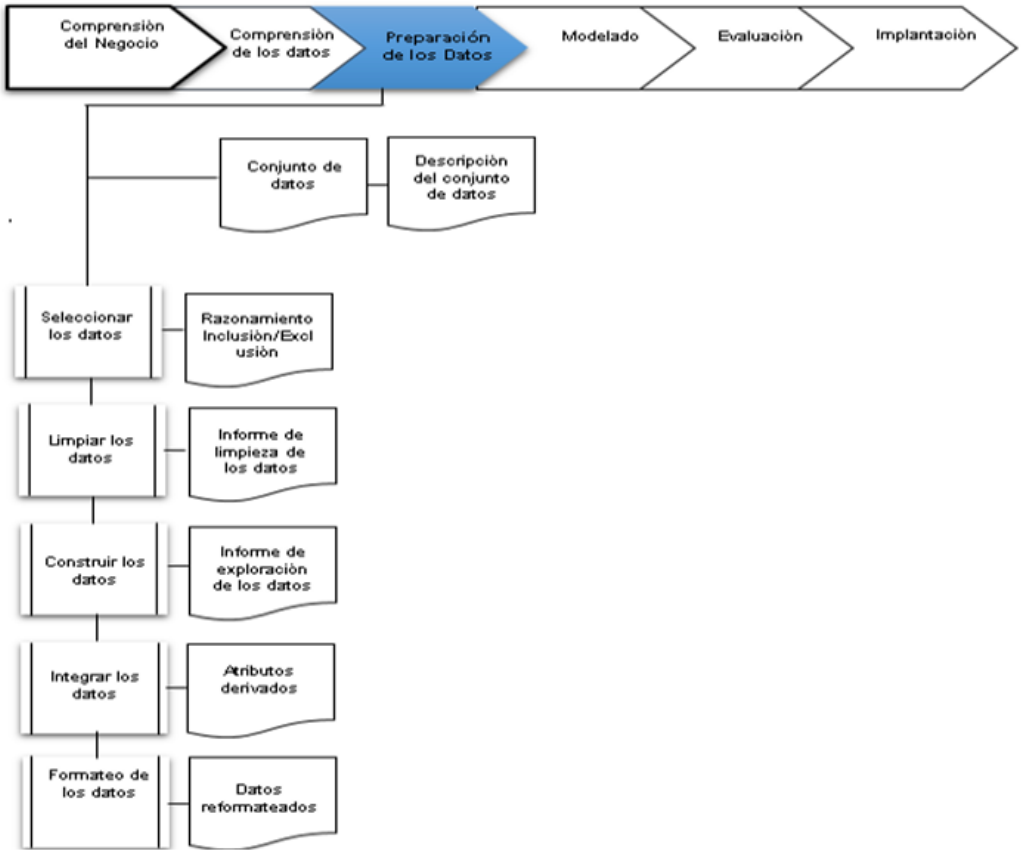

Fuente: (CRISP-DM, 2000)

- **Recolectar datos iniciales.** tiene como objetivo principal la recolección de datos iniciales y adecuación de datos para su posterior procesamiento. Se debe elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.
- **Describir los datos.** debemos describir los datos iniciales obtenidos, tales como número de registros y campos por registro, su identificación, el significado de cada campo y la descripción del formato inicial.
- **Explorar los datos.** Su finalidad es descubrir una estructura general para los datos. Involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos, se crean tablas de frecuencia y se construyen gráficos de distribución.
- **Verificar la calidad de los datos.** se realiza la verificación de los datos para determinar la consistencia de los valores de los campos, la cantidad y distribución de los valores nulos, encontrar valores fuera de rango que pueden ser ruido para el proceso. Se tiene como objetivo asegurar la completitud y corrección de los datos.

**Fase 3.** Preparación de los datos.

**Figura 8**

*Fase de Preparación de los Datos*




Fuente: (CRISP-DM, 2000)

- **Seleccionar los datos**. se selecciona un subconjunto de datos considerando la calidad de los datos, la limitación en el volumen o en los tipos de datos que están relacionadas con las técnicas de minería de datos seleccionadas.
- **Limpiar los datos.** existe una diversidad de técnicas aplicables a esta tarea con el fin de optimizar la calidad de los datos para prepararlos para la fase de modelación.
- **Estructurar los datos.** algunas de las operaciones a realizar en esta tarea puede ser la generación de nuevos atributos a partir de atributos ya existentes,

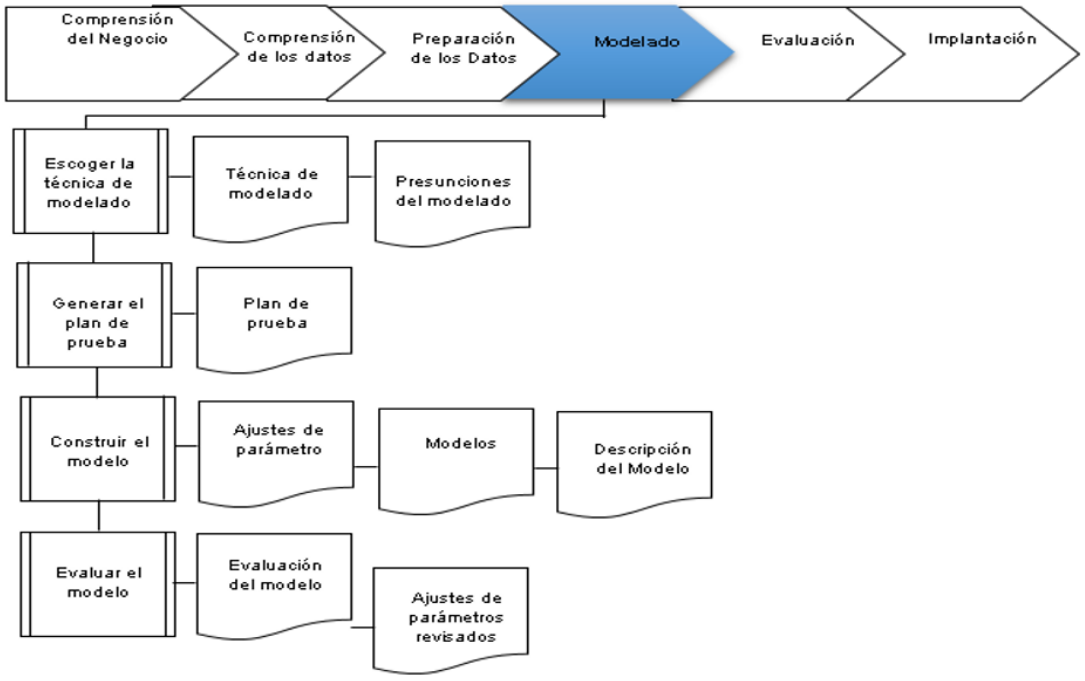
integración de nuevos registros o transformación de valores para atributos existentes.

- **Integrar los datos.** involucra la creación de nuevas estructuras, por ejemplo crear nuevos campos, nuevos registros, fusión de tablas o nuevas tablas.

**Fase 4:** Modelado.

***Figura 9***

*Fase de Modelado*




Fuente: (CRISP-DM, 2000)

- **Seleccionar la técnica del modelado.** se debe elegir una técnica de modelado más apropiado para el proyecto específico. Se pueden elegir de acuerdo a los siguientes criterios:

  1) Ser apropiada al problema

  2) Disponer de los datos adecuados

  3) Cumplir requisitos del problema

4) Tiempo adecuado para obtener un modelo

5) Conocimiento de la técnica

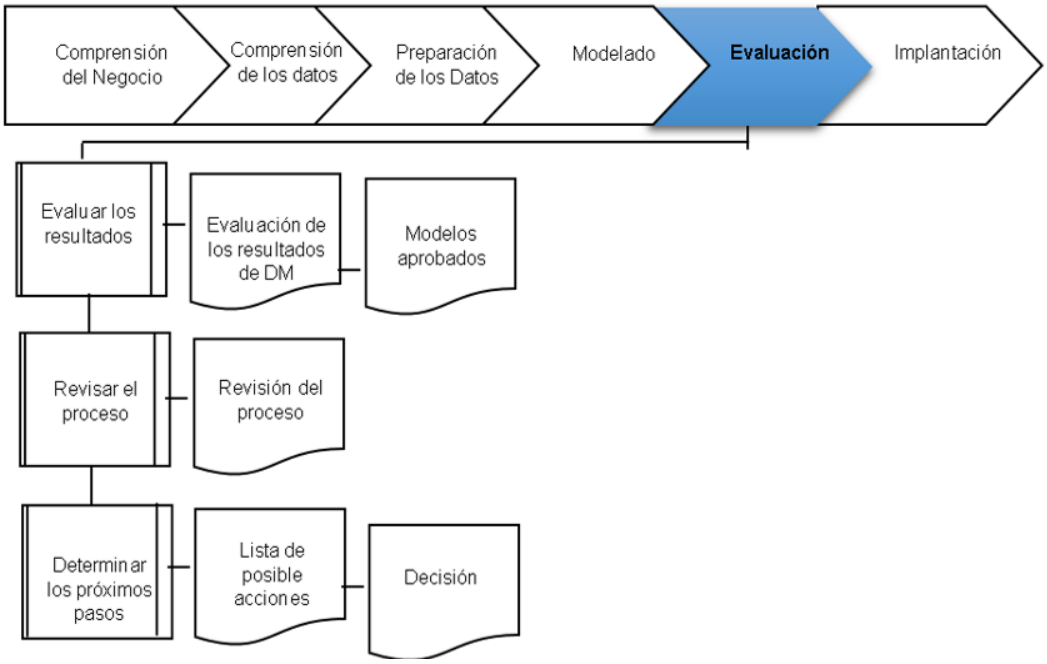
Por ejemplo si el problema es de clasificación podemos elegir de entre arboles de decisión, k-nearest neigbour o razonamiento basado en caos (CBR)

- **Generar plan de prueba.** se debe generar un plan para probar la calidad y validez del modelo construido. Por ejemplo, en una tarea como la clasificación es posible usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba.
- **Construir el modelo.** se ejecuta la técnica seleccionada sobre los datos preparados para generar uno o más modelos. Todas las técnicas del modelado tienen un conjunto de parámetros que determinan características del modelo a generar. La tarea de selección de los mejores parámetros es iterativa basado en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.
- **Evaluar el modelo.** se debe interpretar los modelos de acuerdo al conocimiento del dominio y los criterios de éxitos preestablecidos.
- **Evaluar el modelo.** En esta última tarea de esta fase de modelado los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en minería de datos aplican sus propios criterios (seguridad del conjunto de prueba, pérdida o ganancia de tablas, etc.).

**Fase 5:** Evaluación.

**Figura 10**

*Fase de Evaluación*



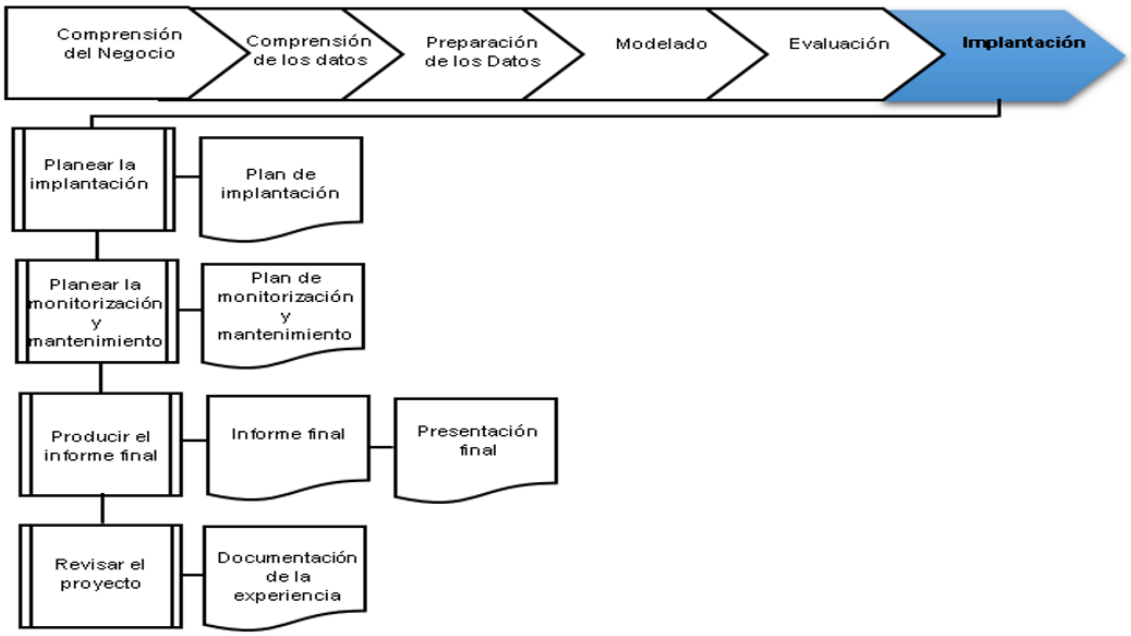

Fuente: (CRISP-DM, 2000)

- **Evaluar los resultados.** esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si es aconsejable probar el modelo o determinar si hay alguna razón de negocio para el cual, el modelo es deficiente.
- **Revisar el proceso.** consiste en la calificación del proceso entero de la minería de datos, con el objeto de identificar elementos que pudieran ser mejorados.
- **Determinar próximos pasos.** en caso de que no se han generado resultados satisfactorios, se podría decidirse por otra iteración desde la fase de preparación de datos o modelación con otros parámetros.

**Fase 6:** Implementación

**Figura 11**

*Fase de Implementación*




Fuente: (CRISP-DM, 2000)

- **Planear la implementación.** esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe estar documentado para su posterior implementación.
- **Monitorizar y mantener.** se debe preparar estrategias de monitorización y mantenimiento para ser aplicada sobre los modelos.
- **Informe final.** dependiendo del plan de implementación, esta puede ser un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto.

- **Revisar el proyecto.** se evalúa lo correcto y lo incorrecto.

### *2.13.4 Analogía entre las etapas del Proceso KDD y CRISP-DM*

Se puede decir que las etapas de esta metodología están relacionadas de alguna manera con las del proceso KDD, incluso se puede llegar a considerar CRISP-DM como una implementación del proceso KDD.

**Tabla 9**

*Analogía entre las etapas del proceso KDD y CRISP-DM*

| KDD | CRISP-DM |
|---|---|
| Identificación del problema en estudio | Comprensión del negocio |
| Selección e integración de los datos | Comprensión de los datos |
| Limpieza y pre-procesamiento de los datos | |
| Transformación de los datos | Preparación de los datos |
| Selección y aplicación de minería de datos | Modelamiento y Evaluación |
| Interpretación y Evaluación | |
| Post KDD | Despliegue del proyecto |

Fuente: (Basado en Paper KDD, 2012)

## 2.14 Métodologia Ágil ASD (Desarrollo de Software Adaptativo)

El método ágil ASD (Adaptive Software Development) traducido en español significa Desarrollo Adaptable de Software es un modelo de implementación de patrones ágiles para desarrollo de software. Al igual que otras metodologías ágiles, su funcionamiento es cíclico y reconoce que en cada iteración se producirán cambios e incluso errores.

El desarrollo de software adaptable (Adaptive Software Development - ASD) es una metodología de desarrollo que hace énfasis en aplicar las ideas que se originaron en el mundo de los sistemas complejos, adaptación continua del proceso al trabajo.

### *2.14.1 Características*

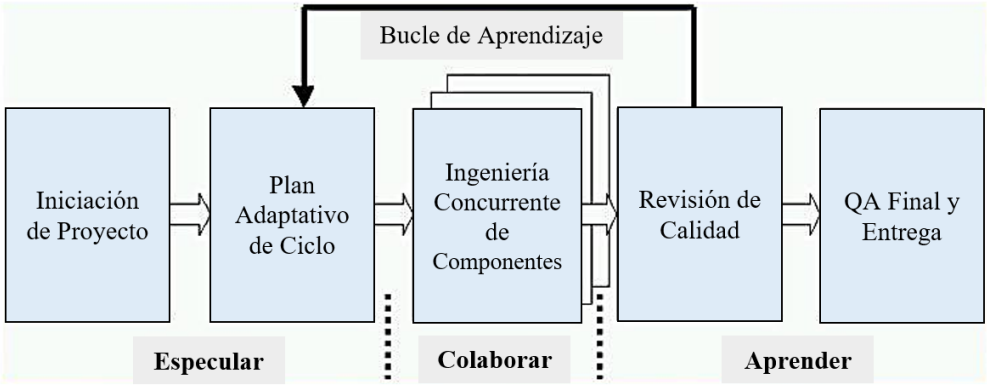
- Sus principales características del ASD son:
- Iterativo.
- Orientado a los componentes de software (la funcionalidad que el producto va a tener, características, etc.) más que a las tareas en las que se va a alcanzar dicho objetivo.
- Tolerante a los cambios.
- Guiado por los riesgos
- La revisión de los componentes sirve para aprender de los errores y volver a iniciar el ciclo de desarrollo

### *2.14.2 Ciclo de Vida*

ASD utiliza un "cambio orientado hacia el ciclo de vida", que tiene tres fases que son: especular colaborar y aprender.

**Figura 12**

*Fases de ASD (Desarrollo de Software Adaptativo)*




Fuente: (Highsmith, 1990)

- **Especular**

Una primera fase de iniciación para establecer los principales objetivos y metas del proyecto en su conjunto y comprender las limitaciones (zonas de riesgo) con las que operará el proyecto.

En ASD se realizan estimaciones de tiempo sabiendo que pueden sufrir desviaciones. Sin embargo, estas son necesarias para la correcta atención de los trabajadores que se mueven dentro de plazos de forma que puedan priorizar sus tareas.

Se decide el número de iteraciones para consumir el proyecto, prestando atención a las características que pueden ser utilizadas por el cliente al final de la iteración. Son por tanto necesarios, marcar objetivos prioritarios dentro de las mismas iteraciones.

Estos pasos se puede volver a examinar varias veces antes de que el equipo y los clientes están satisfechos con el resultado.

- **Colaborar**

Es la fase donde se centra la mayor parte del desarrollo manteniendo una componente cíclica. Un trabajo importante es la coordinación que asegure que lo aprendido por un equipo se transmite al resto y no tenga que volver a ser aprendido por los otros equipos.

- **Aprender**

La última etapa termina con una serie de ciclos de colaboración, su trabajo consiste en capturar lo que se ha aprendido, tanto positivo como negativo. Es un elemento crítico para la eficacia de los equipos.

Jim Highsmith identifica cuatro tipos de aprendizaje en esta etapa:

**Calidad del producto desde un punto de vista del cliente.** Es la única medida legítima de éxito, pero además, dentro de las metodologías ágiles, los clientes tienen un valor importante.

**Calidad del producto desde un punto de vista de los desarrolladores.** Se trata de la evaluación de la calidad de los productos desde un punto de vista técnico. Ejemplos de esto incluyen la adhesión a las normas y objetivos conforme a la arquitectura.

**La gestión del rendimiento.** Este es un proceso de evaluación para ver lo que se ha aprendido mediante el empleo de los procesos utilizados por el equipo.

**Situación del proyecto.** Como paso previo a la planificación de la siguiente iteración del proyecto, es el punto de partida para la construcción de la siguiente serie de características.

### *2.14.3 Ventajas y Desventajas*

**Ventajas**

- La tercera fase del ciclo de vida, revisión de los componentes, sirve para aprender de los errores y volver a iniciar el ciclo de desarrollo.
- Apunta hacia el Rapid Application Development (RAD), el cual enfatiza velocidad de desarrollo para crear un producto de alta calidad, bajo mantenimiento involucrando al usuario lo más posible.
- Utiliza información disponible acerca de cambios para mejorar el comportamiento del software.

**Desventajas**

- Aunque el ciclo entre el aprendizaje y la especulación es bueno permitiéndonos entregar productos con alta calidad, la prolongación de dicho ciclo por errores o cambios que no son detectados en reuniones anteriores afecta tanto a la calidad del producto como a su costo total.
- Dado a que es una metodología ágil implica no realizar procesos que son requeridos en las metodologías tradicionales o por lo menos no realizarlos en procesos diferentes, lo cual implica que empresas grandes las cuales necesitan llevar un mayor control a procesos y personas, tener tareas asignadas a un estado o proceso especifico, y en las cuales dicho incremento de procesos no afectan en gran medida al costo final del producto, para dichas empresas el elegir una metodología tradicional resulta mucho mas rentable tanto por el gran volumen de personal, de productos, y de costos que se manejan y para los cuales se tendrá un mayor control.

## 2.15 Métricas de Calidad

### *2.15.1 Norma ISO/IEC 9126*

Esta norma Internacional fue publicada en 1992, la cual es usada para la evaluación de la calidad de software, llamado “Information technology-Software product evaluation-Quality characteristics and guidelines for their use”; o también conocido como ISO 9126 (o ISO/IEC 9126). Este estándar describe 6 características generales: Funcionalidad, Confiabilidad, Usabilidad, Eficiencia, Mantenibilidad, y Portabilidad.

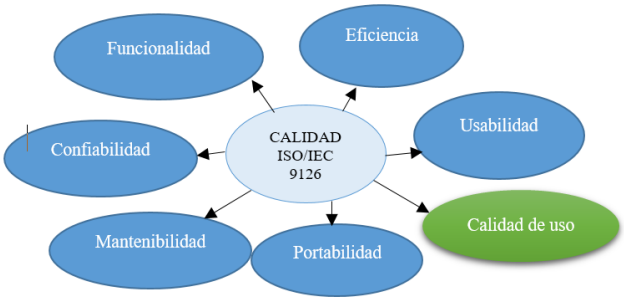
Los modelos de calidad para el software se describen así:

a) **Calidad interna y externa:** Especifica 6 características para calidad interna y externa, las cuales, están subdivididas.

b) **Calidad en uso:** Calidad en uso es el efecto combinado para el usuario final de las 6 características de la calidad interna y externa del software. Especifica 4 características para la calidad en uso.

**Figura 13**

*Norma de Evaluación ISO/IEC 9126*

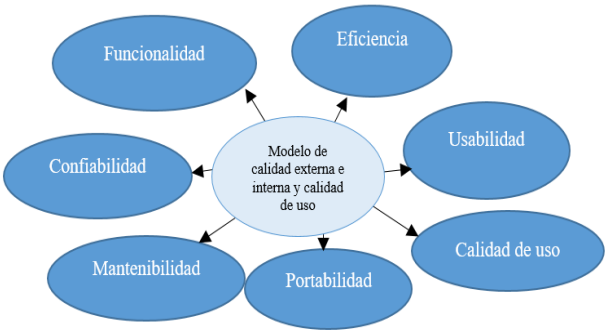



Fuente: (González, 2013)

Se establecen categorías para las cualidades de la calidad externa e interna y calidad en uso del software, teniendo en cuenta estos 7 indicadores (funcionalidad, confiabilidad, utilidad, eficiencia, capacidad de mantenimiento. *(González, 2013)*

**Figura 14**

*Evaluación Interna, externa y Calidad de Uso ISO/IEC 9126*




Fuente: (González, 2013)

Las definiciones se dan para cada característica y sub-característica de calidad del software que influye en la calidad.

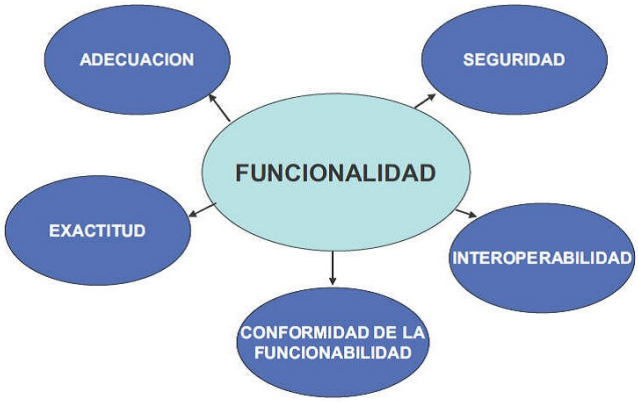
#### 2.15.1.1 Funcionalidad

Funcionalidad es la capacidad del software de cumplir y proveer las funciones para satisfacer las necesidades explícitas e implícitas cuando es utilizado en condiciones específicas.

A continuación se muestra la característica de Funcionalidad y las sub-características que cubre:

**Figura 15**

*Funcionalidad*




Fuente: (González, 2013)

La funcionalidad se divide en 5 criterios:

**Adecuación:** La capacidad del software para proveer un adecuado conjunto de funciones que cumplan las tareas y objetivos especificados por el usuario.

**Exactitud:** La capacidad del software para hacer procesos y entregar los resultados solicitados con precisión o de forma esperada.

**Interoperabilidad:** La capacidad del software de interactuar con uno o más sistemas específicos.

**Seguridad:** La capacidad del software para proteger la información y los datos de manera que los usuarios o los sistemas no autorizados no puedan acceder a ellos para realizar operaciones, y la capacidad de aceptar el acceso a los datos de los usuarios o sistemas autorizados

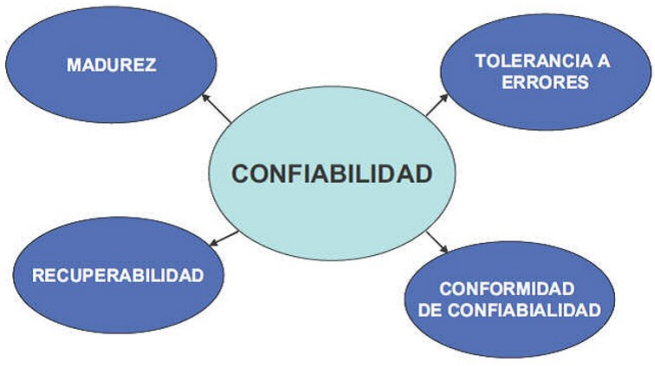
**Conformidad de la funcionalidad:** La capacidad del software de cumplir los estándares referentes a la funcionalidad.

#### 2.15.1.2 Confiabilidad

La confiabilidad es la capacidad del software para asegurar un nivel de funcionamiento adecuado cuando es utilizando en condiciones específicas. En este caso al confiabilidad se amplía sostener un nivel especificado de funcionamiento y no una función requerida.

**Figura 16**

*Confiabilidad*




Fuente: (González, 2013)

La confiabilidad se divide en 4 criterios:

**Madurez:** La capacidad que tiene el software para evitar fallas cuando encuentra errores. Ejemplo, la forma como el software advierte al usuario cuando realiza operaciones en la

unidad de diskett vacía, o cuando no encuentra espacio suficiente el disco duro donde esta almacenando los datos.

**Tolerancia a errores:** La capacidad que tiene el software para mantener un nivel de funcionamiento en caso de errores.

**Recuperabilidad:** La capacidad que tiene el software para restablecer su funcionamiento adecuado y recuperar los datos afectados en el caso de una falla.

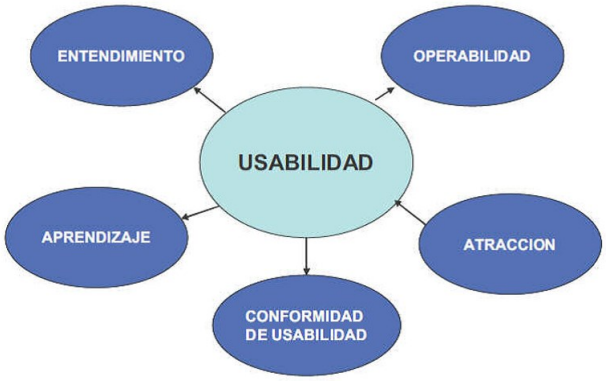
**Conformidad de la fiabilidad:** La capacidad del software de cumplir a los estándares o normas relacionadas a la fiabilidad.

#### 2.15.1.3 Usabilidad

La usabilidad es la capacidad del software de ser entendido, aprendido, y usado en forma fácil y atractiva. Algunos criterios de funcionalidad, fiabilidad y eficiencia afectan la usabilidad, pero para los propósitos de la ISO/IEC 9126 ellos no clasifican como usabilidad. La usabilidad está determinada por los usuarios finales y los usuarios indirectos del software, dirigidos a todos los ambientes, a la preparación del uso y el resultado obtenido.

**Figura 17**

*Usabilidad*




Fuente: (González, 2013)

La usabilidad se divide en 5 criterios:

**Entendimiento:** La capacidad que tiene el software para permitir al usuario entender si es adecuado, y de una manera fácil como ser utilizado para las tareas y las condiciones particulares de la aplicación. En este criterio se debe tener en cuenta la documentación y de las ayudas que el software entrega.

**Aprendizaje:** La forma como el software permite al usuario aprender su uso. También es importante considerar la documentación.

**Operabilidad:** La manera como el software permite al usuario operarlo y controlarlo.

**Atracción:** La presentación del software debe ser atractiva al usuario. Esto se refiere a las cualidades del software para hacer más agradable al usuario, ejemplo, el diseño gráfico.

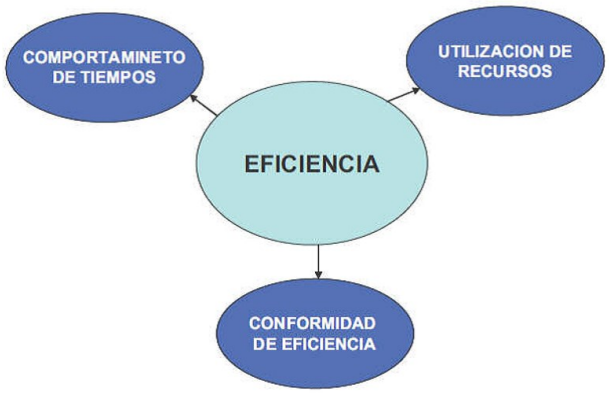
**Conformidad de uso:** La capacidad del software de cumplir los estándares o normas relacionadas a su usabilidad.

#### 2.15.1.4 Eficiencia

La eficiencia del software es la forma del desempeño adecuado, de acuerdo a al número recursos utilizados según las condiciones planteadas. Se debe tener en cuenta otros aspectos como la configuración de hardware, el sistema operativo, entre otros.

**Figura 18**

*Eficiencia*




Fuente: (González, 2013)

La eficiencia se divide en 3 criterios:

**Comportamiento de tiempos:** Los tiempos adecuados de respuesta y procesamiento, el rendimiento cuando realiza su función en condiciones específicas. Ejemplo, ejecutar el procedimiento más complejo del software y esperar su tiempo de respuesta, realizar la misma función pero con más cantidad de registros.

**Utilización de recursos:** La capacidad del software para utilizar cantidades y tipos adecuados de recursos cuando este funciona bajo requerimientos o condiciones establecidas. Ejemplo, los recursos humanos, el hardware, dispositivos externos.

**Conformidad de eficiencia:** La capacidad que tiene el software para cumplir con los estándares o convenciones relacionados a la eficiencia.

#### 2.15.1.5 Capacidad de Mantenimiento

La capacidad de mantenimiento es la cualidad que tiene el software para ser modificado. Incluyendo correcciones o mejoras del software, a cambios en el entorno, y especificaciones de requerimientos funcionales.

**Figura 19**

*Capacidad de Mantenimiento*


Fuente: (González, 2013)

El mantenimiento se divide en 5 criterios:

**Capacidad de ser analizado:** La forma como el software permite diagnósticos de deficiencias o causas de fallas, o la identificación de partes modificadas.

**Cambiabilidad:** La capacidad del software para que la implementación de una modificación se pueda realizar, incluye también codificación, diseño y documentación de cambios.

**Estabilidad:** La forma como el software evita efectos inesperados para modificaciones del mismo.

**Facilidad de prueba:** La forma como el software permite realizar pruebas a las modificaciones sin poner el riesgo los datos.

**Conformidad de facilidad de mantenimiento:** La capacidad que tiene el software para cumplir con los estándares de facilidad de mantenimiento.

### 2.15.1.6 Portabilidad

La capacidad que tiene el software para ser trasladado de un entorno a otro.

**Figura 20**

*Portabilidad*


Fuente: (González, 2013)

La usabilidad se divide en 5 criterios:

**Adaptabilidad:** Es como el software se adapta a diferentes entornos especificados (hardware o sistemas operativos) sin que implique reacciones negativas ante el cambio. Incluye la escalabilidad de capacidad interna (Ejemplo: Campos en pantalla, tablas, volúmenes de transacciones, formatos de reporte, etc.).

**Facilidad de instalación:** La facilidad del software para ser instalado en un entorno específico o por el usuario final.

**Coexistencia:** La capacidad que tiene el software para coexistir con otro o varios software, la forma de compartir recursos comunes con otro software o dispositivo.

**Reemplazabilidad:** La capacidad que tiene el software para ser remplazado por otro software del mismo tipo, y para el mismo objetivo. Ejemplo, la remplazabilidad de una nueva versión es importante para el usuario, la propiedad de poder migrar los datos a otro software de diferente proveedor.

**Conformidad de portabilidad:** La capacidad que tiene el software para cumplir con los estándares relacionados a la portabilidad.

#### 2.15.1.7 Calidad en Uso

Calidad en uso es la calidad del software que el usuario final refleja, la forma como el usuario final logra realizar los procesos con satisfacción, eficiencia y exactitud. La calidad en uso debe asegurar la prueba o revisión de todas las opciones que el usuario trabaja diariamente y los procesos que realiza esporádicamente relacionados con el mismo software.

**Figura 21**

*Calidad en uso*


Fuente: (González, 2013)

La calidad de uso se divide en 4 criterios:

**Eficacia:** La capacidad del software para permitir a los usuarios finales realizar los procesos con exactitud e integridad.

**Productividad:** La forma como el software permite a los usuarios emplear cantidades apropiadas de recursos, en relación a la eficacia lograda en un contexto específico de uso. Para una empresa es muy importante que el software no afecte al productividad del empleado

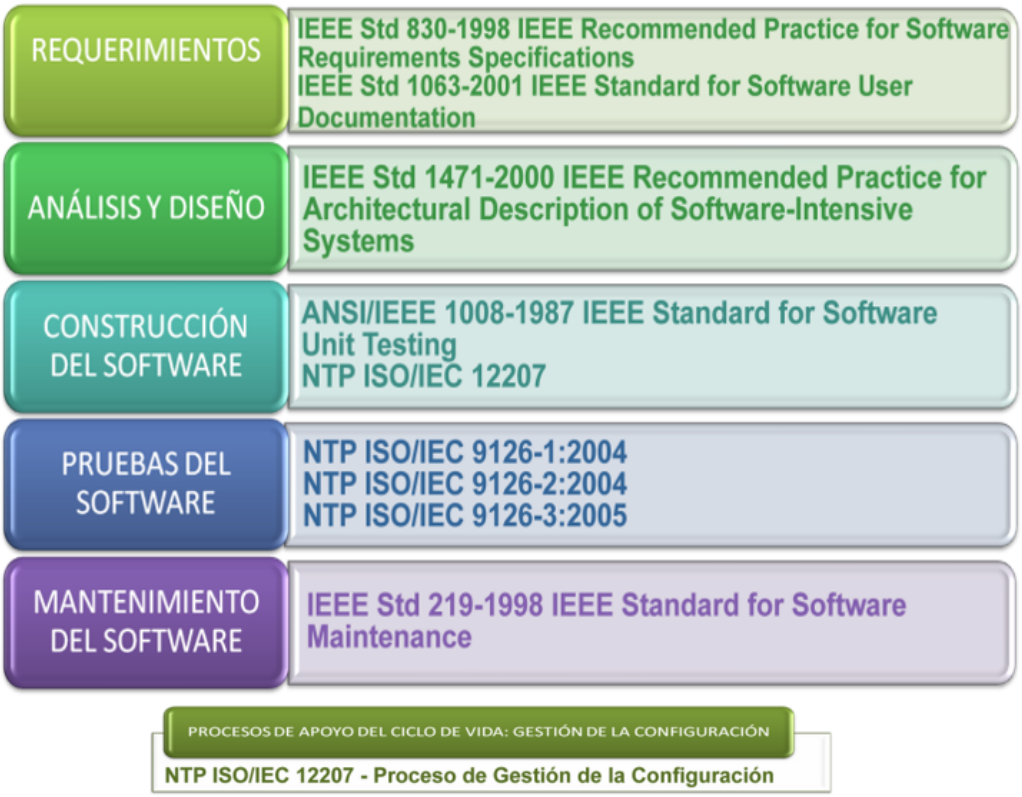
**Seguridad:** Se refiere al que el Software no tenga niveles de riesgo para causar daño a las personas, instituciones, software, propiedad intelectual o entorno. Los riesgos son normalmente el resultado de deficiencias en la funcionalidad (Incluyendo seguridad), fiabilidad, usabilidad o facilidad de mantenimiento.

**Satisfacción:** La satisfacción es la respuesta del usuario a la interacción con el software, e incluye las actitudes hacia el uso del mismo. A continuación se describe un cuadro donde podemos resumir las características y cada uno de sus atributos, este cuadro le ayudara a visualizar el proceso de evaluación.

### *2.15.2 Estándares para el Ciclo de Vida del Software*

**Figura 22**

*Estándares para el Ciclo de vida del software*




Fuente: (Jadisha Yarif, 2010)

## 2.16 Modelo Cocomo II

Los tres modelos de COCOMO II se adaptan tanto a las necesidades de los diferentes sectores, como al tipo y cantidad de información disponible en cada etapa del ciclo de vida de desarrollo, lo que se conoce por granularidad de la información. Estos tres modelos son:

**Figura 23**

*Modelo Cocomo II*


Fuente: (INGESIS II)

- **Modelo de composición de aplicación.** Utilizado durante las primeras etapas de la Ingeniería del software, donde el prototipado de las interfaces de usuario, la interacción del sistema y del software, la evaluación del rendimiento, y la evaluación de la madurez de la tecnología son de suma importancia.
- **Modelo de fase de diseño previo:** Utilizado una vez que se han estabilizado los requisitos y que se ha establecido la arquitectura básica del software.
- **Modelo de fase posterior a la arquitectura:** Utilizado durante la construcción del software.

### *2.16.1 Objetivos para la Construcción de Cocomo II*

- Desarrollar un modelo de estimación de costo y cronograma de proyectos de software que se adaptara tanto a las prácticas de desarrollo.
- Construir una base de datos de proyectos de software que permitiera la calibración continua del modelo, y así incrementar la precisión en la estimación.
- Implementar una herramienta de software que soportara el modelo.
- Proveer un marco analítico cuantitativo y un conjunto de herramientas y técnicas que evaluaran el impacto de las mejoras tecnológicas de software sobre los costos y tiempos en las diferentes etapas del ciclo de vida de desarrollo.

Los tres modelos de COCOMO II se adaptan tanto a las necesidades de los diferentes sectores, como al tipo y cantidad de información disponible en cada etapa del ciclo de vida de desarrollo, lo que se conoce por granularidad de la información. Estos tres modelos son:

### *2.16.2 Modelos de Estimación*

- Este Modelo se crea como alternativa a la estimación del tamaño de un producto software mediante LDC (Líneas de Código Fuente).
- El método de estimación de puntos de función se utiliza para determinar el tamaño del software.
- Están orientadas a la función es decir se centra en la funcionalidad o utilidad del programa.

Las ecuaciones que se utilizan en los tres modelos son:

**Tabla 10**

*Modelo de Estimación*

| | |
|---|---|
| E | a $(Kl)^b$ * (X), en personas - mes |
| Tdev | c $(E)^d$, en meses |
| P | E/Tdev, en personas |

Fuente (CRISP-DM, 2000)

**Dónde:**

**Tabla 11**

*Variables del Modelo de Estimación*

| | |
|---|---|
| E | Es el esfuerzo requerido por el proyecto en persona – mes |
| Tdev | Es el tiempo requerido por el proyecto, en meses |
| P | Es el número de personas requerido por el proyecto |
| a, b, c, y d | Son constantes con valores definidos en una tabla según cada sub-modelo. |
| Kl | Es la cantidad de líneas de código, en miles |
| M(X) | Es un multiplicador que depende de 15 atributos |

Fuente (CRISP-DM, 2000)

### *2.16.3 Características Generales*

Pertenece a la categoría de modelos estimadores basados en estimaciones matemáticas. Está orientado a la magnitud del producto final, midiendo el "tamaño" del proyecto, en función de la cantidad de líneas de código, principalmente.

## 2.17 Herramientas

### *2.17.1 Sistema Operativo Windows 10*

Windows 10 es un sistema operativo desarrollado por la empresa Microsoft para teléfonos inteligentes, tabletas y computadoras personales. Pertenece a la familia de

sistemas operativos Windows. La versión anterior es Windows 8.1. (Wikipedia, Windows 10, 2014)

### *2.17.2 Netbeans 8.2*

NetBeans 8.2 es un entorno de desarrollo integrado libre, hecho principalmente para el lenguaje de programación Java. Existe además un número importante de módulos para extenderlo. NetBeans IDE1 8.2 es un producto libre y gratuito sin restricciones de uso.

NetBeans es un proyecto de código abierto de gran éxito con una gran base de usuarios, una comunidad en constante crecimiento. Sun MicroSystems fundó el proyecto de código abierto NetBeans en junio de 2000 y continúa siendo el patrocinador principal de los proyectos (Actualmente Sun Microsystems es administrado por Oracle Corporation). (Wikipedia, Netbeans, 2016)

### *2.17.3 Java*

Java es un lenguaje de programación de alto nivel, orientado a objetos multi-hebra, usado para escribir tantos programas auto-contenidos.

[Jorge L. Ortega; Notas de Introducción al Lenguaje de Programación Java; 2004].

Java es un lenguaje de programación creado por Sun Microsystems, (empresa que posteriormente fue comprada por Oracle) para poder funcionar en distintos tipos de procesadores. Su sintaxis es muy parecida a la de C o C++, e incorpora como propias algunas características que en otros lenguajes son extensiones: gestión de hilos, ejecución remota, etc. (Oscar Belmonte, 2004)

### *2.17.4 Sublime Text*

Es una de las herramientas más populares en la actualidad tanto para desarrolladores web como para maquetadores. Es gratuito para uso esporádico (y barato si quieres usarlo profesionalmente), ligero, multiplataforma, y cuenta con un abundante catálogo de plugins. (Wikipedia, Sublime Text 3, 2016)

### *2.17.5 Weka*

WEKA (Waikato Environment for Knowledge Analysis) [Waikato, 2011] es un software libre distribuido bajo licencia GNU-GPL, escrito en Java y desarrollado por la Universidad de Waikato, Nueva Zelanda.

WEKA contiene una colección de herramientas de visualización y algoritmos para el análisis de datos y modelado predictivo, junto con una interfaz gráfica para poder acceder fácilmente a sus funcionalidades. (Hernández, 2014)

# Capitulo III: Marco Aplicativo

***Resumen***

En el siguiente capítulo se llevara en práctica, los conceptos teorizados del anterior capitulo para la construcción del Modelo de Minería de Datos, para la Predicción del índice de crecimiento del cáncer de mama en las mujeres de entre 20 a 40 años de la ciudad de La Paz, donde se explicó en el anterior capitulo la metodología, algoritmos y herramientas a utilizar para llegar al objetivo planteado.

## 3.1 Aplicación de Técnicas de Minería de Datos en la Construcción y Validación del Modelo Predictivo

En el siguiente trabajo se aplica cada una de las fases que comprende el proceso de la metodología CRISP–DM (Cross Industry Standard Process for Data Mining) y KDD (Knowledge Discovery in Databases).

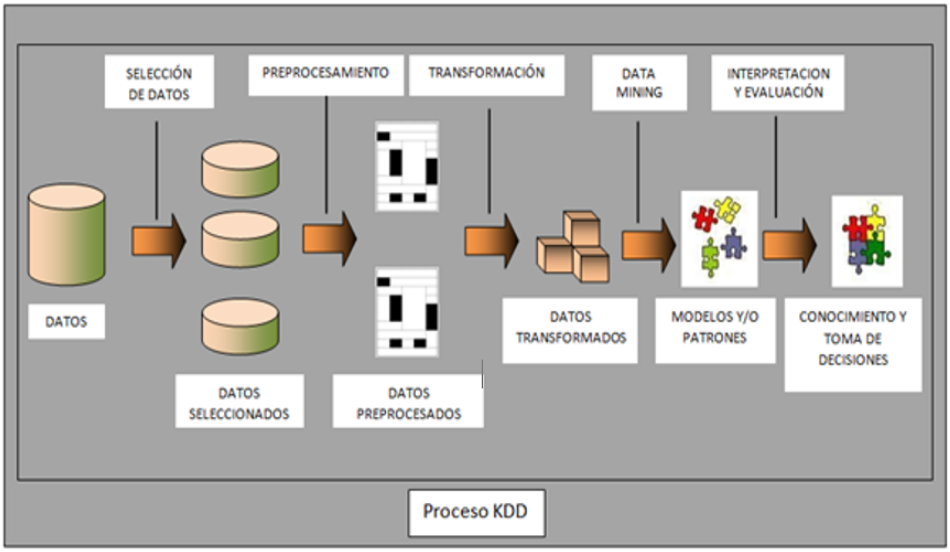
En la construcción del Modelo Predictivo del índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos, se aplicara las siguientes metodologías CRISP-DM y KDD, en base a datos recolectados de diferentes gestiones y para el desarrollo de prototipo del Modelo Predictivo se utiliza la metodología ASD (DESARROLLO DE SOFTWARE ADAPTABLE).

### *3.1.1 Comprensión del Negocio (Fase I)*

A continuación se realiza el esquema de solución para el Modelo de Predicción del índice de crecimiento del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos.

**Figura 24**

*Esquema de la solución del modelo de Minería de Datos*




Fuente: (CRISP-DM, 2000)

### *3.1.2 Comprensión de los Datos (Fase II)*

En esta fase se realizó la recolección inicial de los datos que están relacionados con el problema del índice de crecimiento del cáncer de mama de las mujeres de la ciudad de La Paz, para lo cual se hizo las solicitudes de información a diferentes hospitales e instituciones de salud.

### 3.1.2.1 Datos recolectados del cáncer de mama de la ciudad de La Paz

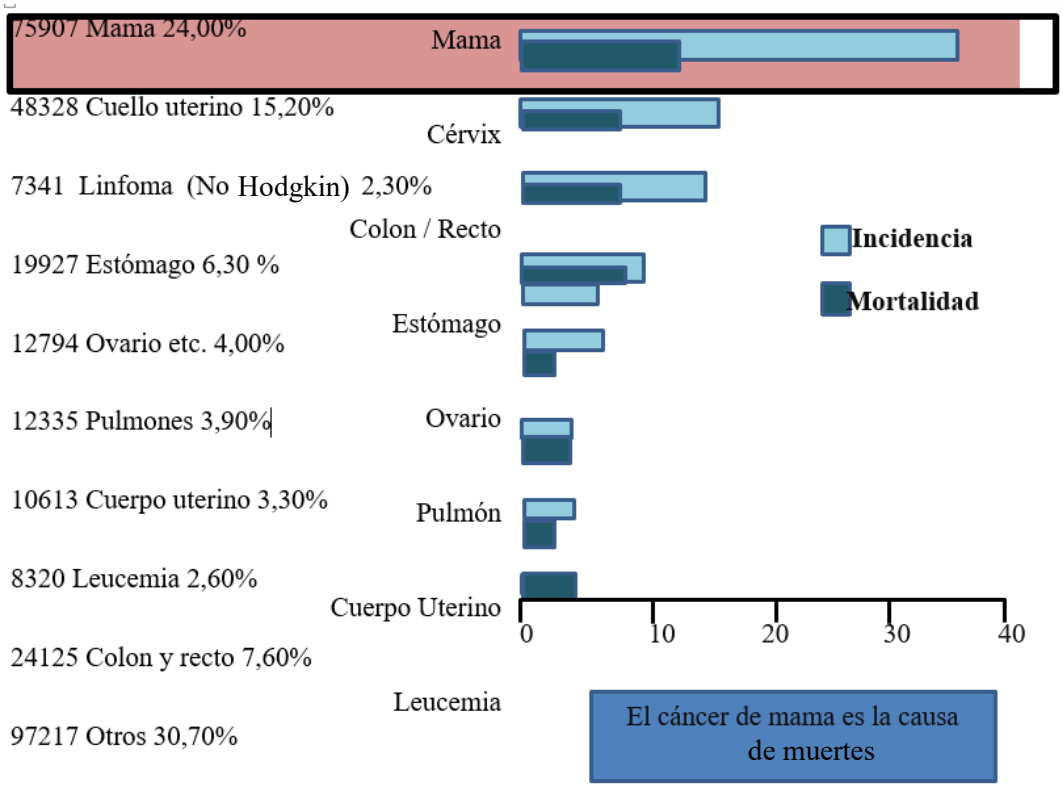
En Latinoamérica, el cáncer de mama ocupa el primer lugar en incidencia con un 24% del total de los cánceres en la región (le sigue el cáncer de cuello uterino, con un 15,2%), lo que constituye un problema importante de salud pública.

A continuación se muestran los datos obtenidos sobre el índice de crecimiento del cáncer de mama de las diferentes instituciones de salud de la ciudad de La Paz.

*(GLOBOCAN 2002, IARC)*

**Figura 25**

*Incidencia y mortalidad por tipo de cáncer*




Fuente: (GLOBOCAN 2002, IARC)

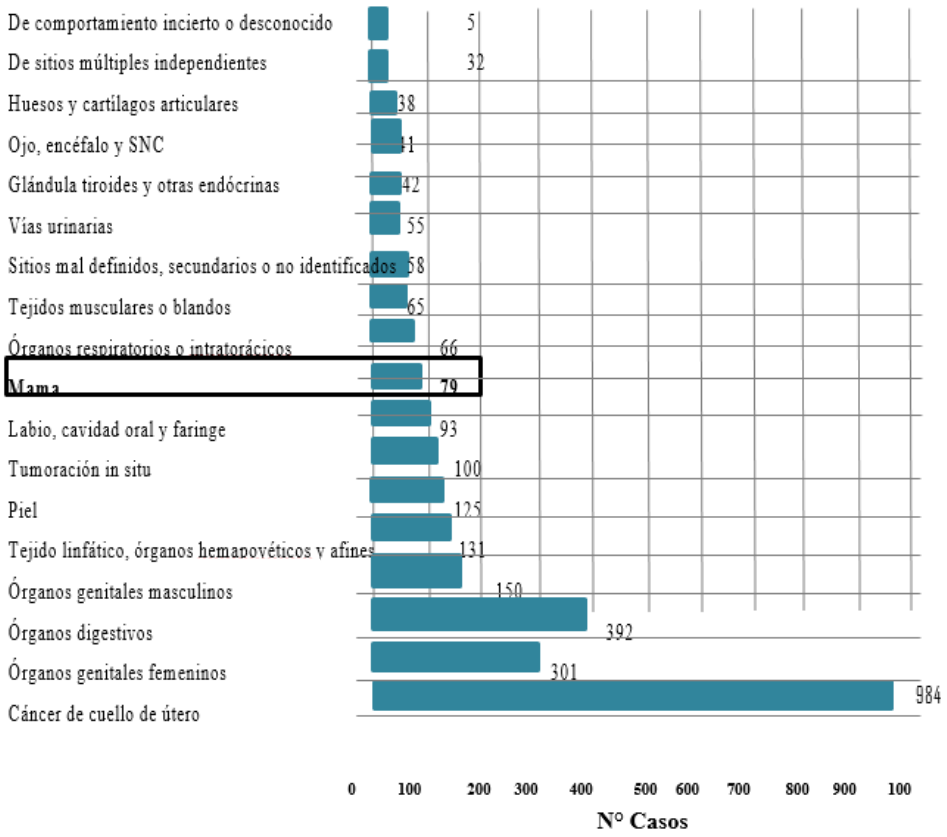
Como se observa en la **Figura 25** el cáncer de mama se encuentra en primer lugar en incidencia y mortalidad.

La gran mayoría de la población femenina en riesgo desconoce las características y evolución del cáncer de mama; la técnica del auto-examen de mama es el paso inicial dentro de la ruta de diagnóstico de este tipo de cáncer.

En Bolivia, la información sobre el trabajo realizado con relación al cáncer de mama es bastante escasa; los datos que a continuación se detallan son el fruto de investigaciones aisladas de instituciones y personas sensibilizadas con la problemática de esta patología

**Figura 26**

*Incidencia y mortalidad por cáncer de mama en La Paz (Accesibilidad Económica)*




Fuente: (Narda Navarro, 1995-2001)

En la **Figura 26** se muestra el número casos que inciden en cuanto al cáncer de mama.

Pocos servicios del Sistema Nacional de Salud disponen de la tecnología adecuada (mamografía) para el diagnóstico del cáncer de mama, lo cual excluye a la población femenina de bajos ingresos, por el alto costo de este examen (accesibilidad económica).

Sobre la base de los datos proyectados por el INE y si la edad de examen de mama se haría a partir de los 30 años debido a recurrencia de casos en poblaciones cada vez más jóvenes, y considerando que es deseable llegar a la totalidad de la población femenina en riesgo (30-69 años), tenemos un panorama cuanto más alarmante:

**Tabla 12**

*Bolivia: Población femenina de 30 a 69 años de edad que debería realizarse examen de mama por día.*

| **La Paz** | **1.701** |
|---|---|
| Oruro | 267 |
| Potosí | 422 |
| Cochabamba | 1.229 |
| Chuquisaca | 331 |
| Tarija | 278 |
| Pando | 31 |
| Beni | 1.006 |
| SCZ | 1.383 |

Fuente: (Elaboración propia sobre la base de datos de proyecciones de población del INE. LP. , 2010)

Como se observa por los datos de la **Tabla 12**, es impensable que se pueda llegar a toda la población en edad de realizarse examen de mama, ni siquiera a la mitad de la población. Es por ello que la prevención debería ser el principal objetivo de cualquier Plan. Por tanto, de cara a las nuevas responsabilidades que las Gobernaciones y Municipios tienen en la protección a la salud de acuerdo a las atribuciones que la Ley

Marco de autonomías y descentralización les confiere, es menester compartir responsabilidades con el compromiso conjunto de todos los espacios de gestión pública en salud, para hacer frente a esta enfermedad como la primera causa de mortalidad de mujeres porque si el tema no se lo asume con voluntad política traducida en presupuestos, la deuda social continuará teniendo un rostro de mujer y sobre todo de mujer pobre e indígena.

**Tabla 13**

*Población incidencia y mortalidad con cáncer de mama*

| TEMA | PLAN 2004-2008 | PLAN 2009-2015 |
|---|---|---|
| Mortalidad por cáncer de mama<br>Por cada 100 mil mujeres | 153 | Cáncer de mama |
| N° muertes | 661 (22,2/100.000) | Según OMS 1.665 por año (4,5 mujeres mueren por día). |
| Edad de mayor incidencia | 35 a 64 años | Se han presentado casos de cáncer de mama en edad de 30 años |
| Lugar de incidencia y tasas | La Paz con 60,9 | Se mantiene el mismo dato |

Fuente: (Elaboración propia INE. LP. , 2010)

Como observamos en la **Tabla 13**, al parecer la incidencia ha ido en incremento y no deja de llamar la atención la falta de información que el Plan actual no incorpora.

Respecto al cáncer de mama, lamentablemente el tema no ha sido considerado con anterioridad al Plan actual y ello no permite un conocimiento más acercado del tema en nuestro país. Las entrevistas a profesionales oncólogos, contemplado en la intervención defensorial, arroja información como la que sigue: “El un tema recién considerado por el Ministerio de Salud, todavía no tenemos datos de screening (poner en pantalla),

educación, cobertura, edades, zonas, de acuerdo a nuestra realidad, además es bueno recalcar, que en los diferentes Seguros de Salud de Bolivia,

La revisión de registros del Hospital del seguro social Militar con sede en La Paz, brindó registro de ambas patologías pero las mismas no están desagregadas por tipo de patología y solo hacen referencia a La Paz y no a nivel nacional. Tampoco se tiene, y esto a nivel general, de un sistema estadístico que no solo registre a las fallecidas. De acuerdo a la información recabada de profesionales en este nosocomio, "las pacientes que acuden para su internación, generalmente son aquellas que recibirán terapia con quimioterapia, una cirugía oncológica o las que se encuentran con progresión de enfermedad y están con terapia paliativa. La mayor parte de estas pacientes siempre vuelven a reinternarse tres veces por mes o de acuerdo a protocolo de tratamiento.". Esto mismo ratifica la aseveración que desde distintos espacios se hace, acerca de no abordar la temática únicamente desde un punto de vista biologista.

**Tabla 14**

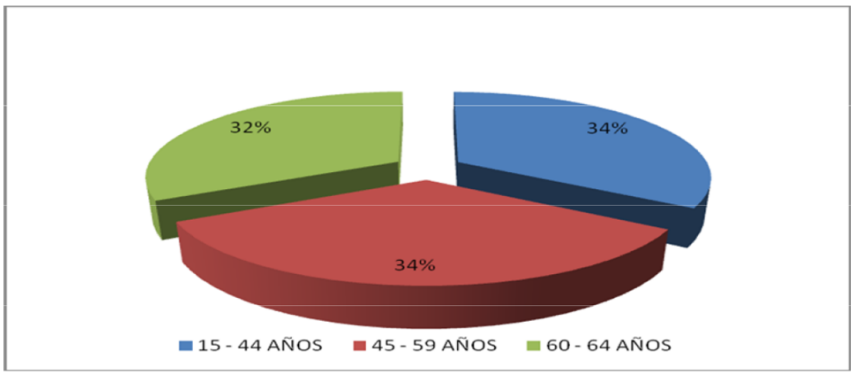
*COSSMIL: Defunciones en patologías de cáncer de mama según grupo de edad. 2005-2008*

| GRUPO DE EDADES | DEFUNCIONES |
|---|---|
| 15 - 44 AÑOS | **2** |
| 45 - 59 AÑOS | **4** |
| 60 - 64 AÑOS | **0** |
| 65 O MAS | **3** |
| TOTAL | **9** |

Fuente: (COSSMIL. La Paz, 2008)

**Figura 27**

*Cossmil: prevalencia de cáncer por grupos de edad. 2005-2008*




Fuente: (Hospital Militar. La Paz. 2008)

La información disponible no permite disgregar el tipo de cáncer aunque la fuente se refiere a mujeres.

- Plan de prevención control y seguimiento del cáncer de mama (2009-2015)

**Indicadores:**

**Tabla 15**

*Indicadores*

| | |
|---|---|
| **1** | Cobertura de detección: Proporción de mujeres con examen clínico de mama positivo<br>Total mujeres en edad de riesgo (35 – 64 años)<br><br>No de estudios mamográficos positivos<br>Total de mamografías realizadas |
| **2** | Porcentaje de mujeres con biopsias positivas, confirmadas por histopatología<br><br>Nº de mujeres con biopsias positivas confirmadas por histopatología<br>Total de biopsias realizadas<br>Porcentaje de mujeres con cáncer de mama tratadas |

| | |
|---|---|
| Nº de mujeres con diagnóstico de cáncer tratadas | |
| N° total de mujeres con diagnóstico de cáncer | |
| Porcentaje de mujeres seguidas después del tratamiento de cáncer | |
| Nº mujeres con cáncer tratadas y con seguimiento mínimo de 12 meses | |
| N° total de mujeres con tratamiento de cáncer | |

Fuente: (CRISP-DM, 2000)

**Proyecciones:**

La población objetivo del Plan Nacional de Prevención y Control de cáncer de mama 2009-2015 es la población femenina entre 20 y 40 años de edad.

**Tabla 16**

*Población femenina proyectada*

| Gestión | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Beni | 45.331 | 46.894 | 48.457 | 50.021 | 51.584 | 53.147 | 54.710 |
| Chuquisaca | 74.885 | 76.576 | 78.267 | 79.958 | 81.648 | 83.339 | 85.030 |
| Cochabamba | 226.497 | 233.265 | 240.032 | 246.800 | 253.567 | 260.335 | 267.103 |
| **La Paz** | **386.252** | **396.349** | **406.445** | **416.542** | **426.638** | **436.735** | **446.832** |
| Oruro | 60.839 | 61.787 | 62.736 | 63.684 | 64.632 | 65.580 | 66.529 |
| Pando | 6.271 | 6.588 | 6.906 | 7.223 | 7.540 | 7.857 | 8.175 |
| Potosí | 96.022 | 96.636 | 97.250 | 97.864 | 98.478 | 99.092 | 99.706 |
| Santa Cruz | 313.232 | 327.660 | 342.087 | 356.515 | 370.942 | 385.369 | 399.797 |
| Tarija | 62.907 | 65.204 | 67.501 | 69.798 | 72.095 | 74.393 | 76.690 |
| **BOLIVIA** | **1.272.237** | **1.310.959** | **1.349.681** | **1.388.404** | **1.427.126** | **1.465.847** | **1.504.599** |

Fuente: (Plan Nacional contra el Cáncer, 2016)

A partir de estos datos demográficos, se pueden realizar proyecciones de cobertura anual que se quiere lograr de la **Tabla 16**.

Si se limita el número de mamografías por mujer durante los cinco años de duración del Plan Nacional, la cobertura total debería llegar a más de 50%, es decir que más de una mujer en edad de riesgo sobre dos se habría hecho realizar una mamografía en estos cinco años.

**Tabla 17**

*Número esperado de mujeres con mamografía*

| **Cobertura Anual** | **5%** | **10%** | **15%** | **20%** | **25%** | **30%** | **35%** |
|---|---|---|---|---|---|---|---|
| **Gestión** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** |
| Beni | 2.265 | 4.689 | 7.268 | 10.004 | 12.896 | 15.941 | 19.148 |
| Chuquisaca | 3.744 | 7.657 | 11.740 | 15.991 | 20.412 | 24.494 | 29.760 |
| Cochabamba | 11.324 | 23.326 | 36.004 | 49.360 | 63.391 | 78.100 | 93.486 |
| **La Paz** | **19.302** | **39.604** | **60.996** | **83.308** | **106.659** | **131.020** | **156.391** |
| Oruro | 3.041 | 6.178 | 9.410 | 12.736 | 16.158 | 19.674 | 23.285 |
| Pando | 313 | 658 | 1.035 | 1.444 | 1.885 | 2.357 | 2.861 |
| Potosí | 4.801 | 9.663 | 14.587 | 19.597 | 24.619 | 29.727 | 34.897 |
| Santa Cruz | 15.661 | 32.776 | 51.313 | 71.303 | 92.735 | 115.610 | 139.928 |
| Tarija | 3.145 | 6.520 | 10.125 | 13.959 | 18.023 | 22.317 | 26.841 |
| **BOLIVIA** | **63.596** | **131.071** | **202.478** | **277.702** | **356.778** | **439.240** | **526.597** |

Fuente: (Plan Nacional contra el Cáncer, 2016)

**Tabla 18**

*Número de cáncer de mama detectados y esperados con proyecciones*

| **Gestión** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** |
|---|---|---|---|---|---|---|---|
| Beni | 1 | 2 | 2 | 3 | 4 | 5 | 6 |
| Chuquisaca | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| Cochabamba | 3 | 6 | 10 | 13 | 17 | 20 | 24 |
| **La Paz** | **5** | **11** | **16** | **22** | **28** | **35** | **41** |
| Oruro | 1 | 2 | 3 | 4 | 5 | 5 | 6 |
| Pando | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| Potosí | 2 | 2 | 4 | 5 | 7 | 8 | 9 |
| Santa Cruz | 4 | 9 | 14 | 19 | 25 | 31 | 37 |
| Tarija | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| **BOLIVIA** | **19** | **37** | **56** | **73** | **99** | **117** | **138** |

Fuente: (Plan Nacional contra el cáncer, 2016)

La incidencia de cáncer de mama por 100.000 mujeres en Bolivia es de 26,57 y la mortalidad por esta patología es de 8,71 por 100.000 (GLOBOCAN 2000 IARC CANCER BASE) 26,57/ 100.000 = 0,0002657 x Población proyectada.

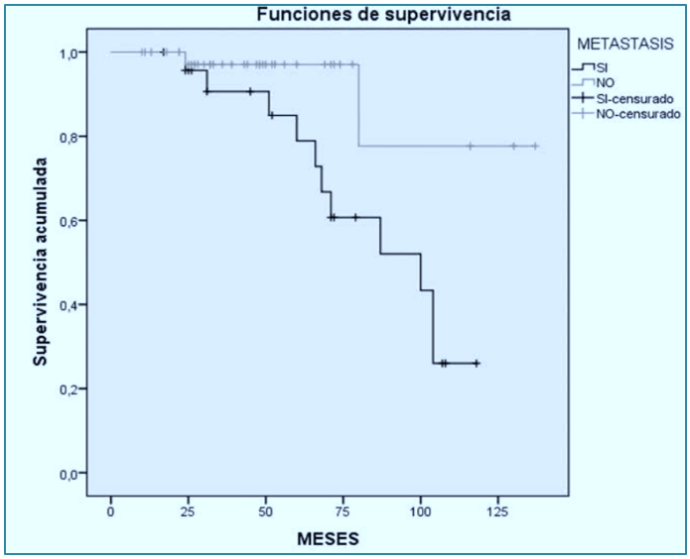
**Análisis de supervivencia de pacientes con cáncer de mama: en el Hospital Seguro Social Universitario, gestión 2000 A 2016**

Entre la Población asegurada al HSSU, con cáncer de mama de la gestión 2000 a 2016, la edad más frecuente es mayor a 51 años, mientras que el Tipo de cáncer más usual es el Hormono (+) con un 64%, el diagnostico. Histopatológico con más frecuencia es el Estadio IV-B con un 33% y el 62 % de las pacientes No presentan metástasis.

El Promedio de supervivencia de los pacientes que no desarrollaron metástasis fue mayor en relación al grupo de los que tuvieron metástasis. La diferencia encontrada es significativa Log Rank (Mantel-Cox) = 5,4 p=0.020, mayor sobrevida atribuida a la no presencia de metástasis.

**Figura 28**

*Comparación de Supervivencia según Metástasis, Pacientes con cáncer de mama, HSSU, Gestión 2000 a 2016*




Fuente: (Hospital Seguro Social Universitario, 2017)

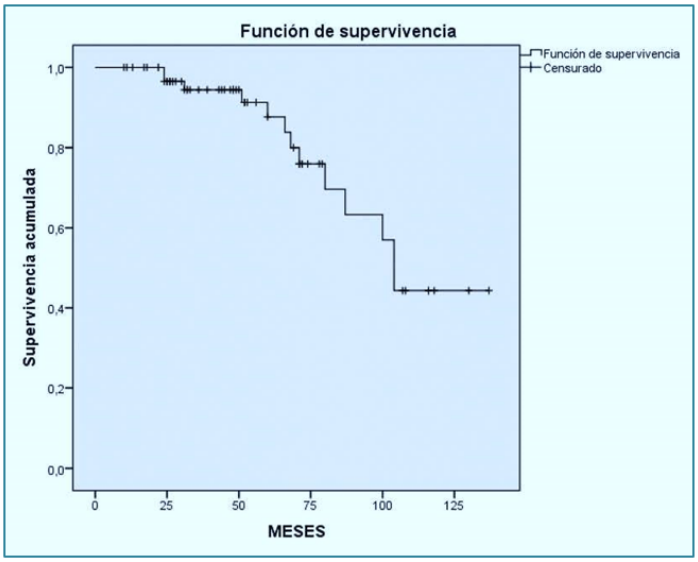
En la **Figura 28** se puede observar, que la diferencia encontrada es mayor la supervivencia en los pacientes que tienen metástasis en comparación a aquellas que no tienen.

En promedio las pacientes tuvieron una sobrevida de 104 meses (8,7 años), coincidiendo con la media de supervivencia, por lo que el 50% de las pacientes tuvieron una sobrevida de 104 meses o menos, y la mitad restante un tiempo de vida mayor, disminuyendo este tiempo a en el 75% delos casos a 80 meses (6,7 años) (percentil 75).

La figura Nº 28 confirma y muestra con mayor claridad el tiempo medio de supervivencia de pacientes del HSSU con cáncer de mama es 104 meses (8,7 años).

**Figura 29**

*Supervivencia de Pacientes con cáncer de mama, HSSU, Gestión 2000 a 2016*




Fuente: (Hospital Seguro Social Universitario, 2017)

En la **Figura 29** el tipo de cáncer de mama, más frecuente en el Hospital Seguro Social Universitario es el hormono (+) 64%, seguido en un 30% por el tipo Herb New.

**Tabla 19**

*Cáncer de mama de Pacientes del HSSU Gestión 2000 a 2016*

| TIPO DE CANCER | ESTADO | | | | TOTAL | |
|---|---|---|---|---|---|---|
| | MUERTO | | VIVO | | | |
| | F | % | F | % | F | % |
| HERB 2 NEW | 2 | 15,4 | 17 | 33,3 | 19 | 29,7 |
| HORMONO (+) | 11 | 84,6 | 30 | 58 | 41 | 64,1 |
| TRIPLE (-) | 0 | 0 | 4 | 7,8 | 4 | 6,3 |
| **TOTAL** | **13** | **100** | **51** | **100** | **64** | **100** |

Fuente: (Hospital Seguro Social Universitario, 2017)

El promedio de sobrevida de los pacientes que recibieron Exemestano fue de 99,5 meses (8.3 años) con un IC. 83,2 – 115.78.

**Tabla 20**

*Tiempo de Supervivencia de Pacientes con cáncer de mama Hormono (+) que recibieron Tratamiento con Exemestano, en el Hospital Seguro Social Universitario, de la Gestión 2000 A 2016*

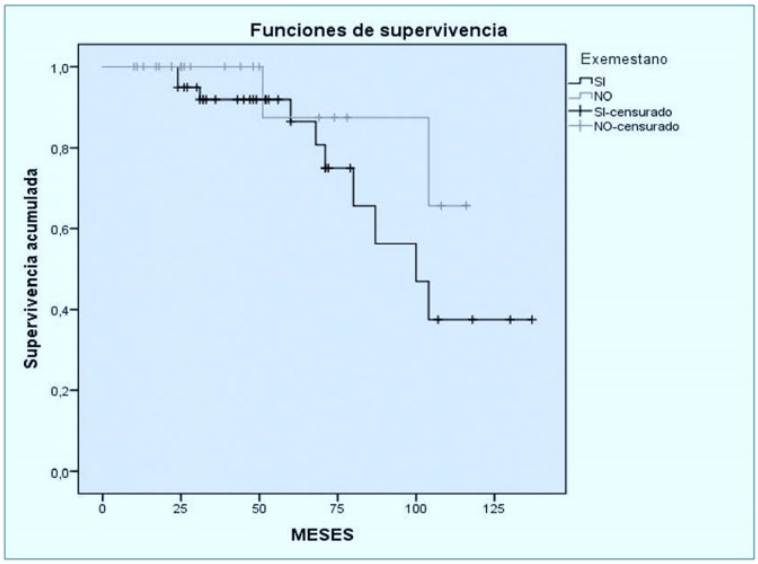
| Exemestano | Media | Error Standart | Confianza de 95% | |
|---|---|---|---|---|
| | | | Límite Inferior | Límite superior |
| SI | 99,528 | 8,295 | 83,269 | 115,787 |
| NO | 105,25 | 7,598 | 90,359 | 120,141 |
| **GLOBAL** | **105,554** | **6,882** | **92,065** | **119,042** |

Fuente: (Hospital Seguro Social Universitario, 2017)

Al comparar con el grupo que no recibió el tratamiento no se evidencio diferencias significativas (Long Rank 1,602 p=0,206), como se puede observar en el **Tabla Nº 20**.

**Figura 30**

*Tiempo de Supervivencia de Pacientes con cáncer de mama Hormono (+) que recibieron Tratamiento con Exemestano, en el Hospital Seguro Social Universitario, Gestión 2000 a 2016*




Fuente: (Hospital Seguro Social Universitario, 2017)

El promedio del tiempo de supervivencia de los pacientes Trastuzumab fue de (8,6 años) 103,28 meses (IC 86,54 – 120,024), al comparar el tiempo de sobrevida con otros tratamientos se evidencia que existe diferencia (Long Rank 0,048 p = 0,048), situación que se corrobora en la **Figura 30**.

**Tabla 21**

*Tiempo de Supervivencia de Pacientes con cáncer de mama Herb New (-) que recibieron Tratamiento con Trastuzumab, en el Hospital Seguro Social Universitario, Gestión 2000 A 2016*

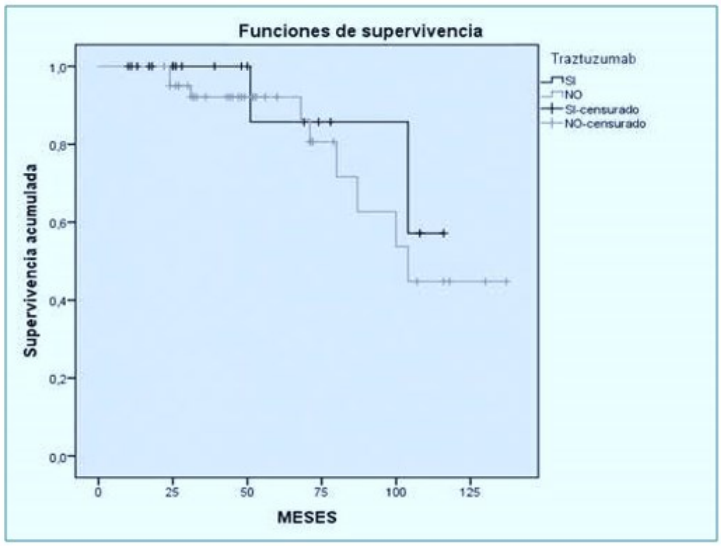
| Trastuzumab | | Error estándar | Intervalo de confianza de 95% | |
|---|---|---|---|---|
| | | | Límite inferior | Límite Superior |
| SI | 103,286 | 8,54 | 86,548 | 120,024 |
| NO | 94,673 | 8,059 | 88,877 | 120,469 |
| **Global** | **107,457** | **6,867** | **93,998** | **120,917** |

Fuente: (Hospital Seguro Social Universitario, 2017)

En la **Tabla 21** se puede observar el mayor tiempo de sobrevida de pacientes que recibieron Trastuzumab, en comparación de aquellos que no recibieron el tratamiento.

**Figura 31**

*Tiempo de Supervivencia de Pacientes con cáncer de mama Herb New (-) que recibieron Tratamiento con Trastuzumab, en el Hospital Seguro Social Universitario, Gestión 2000 a 2016*




Fuente: (Hospital Seguro Social Universitario, 2017)

Según el método actuarial al realizar la Tabla de Mortalidad de pacientes con cáncer de mama, se confirma que la mediana de tiempo de supervivencia de las pacientes con CA de mama es de 104,13 meses (8,7 años). **Figura 31**.

**Tabla 22**

*Tabla de Mortalidad de Pacientes con cáncer de mama, HSSU, de la Gestión 2000 a 2016*

| Intervalo | Vivos | Censurados | Expuestos | Muertos | Proporción muertos | Proporción Supervivencia | Supervivencia Acumulada |
|---|---|---|---|---|---|---|---|
| 0-12 | 64 | 2 | 63,000 | 0 | 0,00 | 1,00 | 1,00 |
| 12- 24 | 62 | 5 | 59,500 | 0 | 0,00 | 1,00 | 1,00 |
| 24 - 36 | 57 | 13 | 50,500 | 3 | 0,06 | 0,94 | 0,94 |
| 36- 48 | 41 | 6 | 38,000 | 0 | 0,00 | 1,00 | 0,94 |
| 48- 60 | 35 | 9 | 30,500 | 1 | 0,03 | 0,97 | 0,91 |
| 60- 72 | 25 | 4 | 23,000 | 4 | 0,17 | 0,83 | 0,75 |
| 72- 84 | 17 | 5 | 14,500 | 1 | 0,07 | 0,93 | 0,70 |
| 84- 96 | 11 | 0 | 11,000 | 1 | 0,09 | 0,91 | 0,64 |
| 108- 120 | 6 | 4 | 4,000 | 0 | 0,00 | 1,00 | 0,44 |
| 120- 132 | 2 | 1 | 1,500 | 0 | 0,00 | 1,00 | 0,44 |
| 132 | 1 | 1 | 0,500 | 0 | 0,00 | 1,00 | 0,44 |

Fuente: (Hospital Seguro Social Universitario, 2017)

La supervivencia de pacientes con cáncer de mama, tratados en el Hospital Seguro Social Universitario, de la Gestión 2000 a 2016, tuvieron un promedio de tiempo de sobrevida de 8,6 años. Los pacientes con cáncer de mama Hormono (+) que recibieron tratamiento con exemestano, después de Adriamicina y Ciclofosfamida el promedio del tiempo de supervivencia fue de 99,5 meses (8,3 años); y en pacientes con cáncer de mama Herb New (-) tratados con Trastuzumab, después de recibir Adriamicina y Ciclofosfamida fue de 103,28 meses (8,6 años). Se confirma según el método actuarial y tablas de vida, que la mediana de tiempo de supervivencia, de los pacientes con cáncer de mama, tratados en el Hospital Seguro Social Universitario fue de 104,13 meses (8,7 años).

- **Precisión del problema**

**Investigación Exploratoria**

Los resultados fueron analizados de manera cualitativa y cuantitativa para identificar factores de riesgo de afrontamiento y vivencia de la investigación exploratoria.

**Resultados Cuantitativos**

**Tabla 23**

*Características socio-demográficas y económicas de las pacientes*

| ***VARIABLE N=20*** | **FRECUENCIA** | **%** |
|---|---|---|
| ***EDAD*** | | |
| *25 a34* | **1** | **5%** |
| *35 a 44* | **3** | **15%** |
| *45 a 54* | **4** | **20%** |
| *55 a 64* | **6** | **30%** |
| *65 a más* | **6** | **30%** |
| ***ESTADO CIVIL*** | | |
| *Soltera* | **4** | **20%** |
| *Casada* | **11** | **55%** |
| *Unión estable* | **1** | **5%** |
| *Divorciada* | **4** | **10%** |
| *Viuda* | **2** | **10%** |
| ***NUMERO DE HIJOS*** | | |
| *Ninguno* | **2** | **10%** |
| *Uno* | **4** | **20%** |
| *Dos* | **6** | **30%** |
| *Tres* | **3** | **15%** |
| *Cuatro* | **3** | **15%** |
| *Cinco o más* | **2** | **10%** |
| ***OCUPACION*** | | |
| *Estudiante* | | |
| *Profesora* | **4** | **20%** |
| *Profesional* | **3** | **15%** |
| *Comerciante* | **3** | **15%** |
| *Ama de casa* | **5** | **25%** |
| *Rentista/magisterio* | **4** | **20%** |
| *Otros* | **1** | **5%** |
| ***TOTAL*** | **20** | **100%** |

Fuente: (Elaboración propia, 2008)

Interpretación Resultados **Tabla 23** Características socio demográficas y económicas de las pacientes

- Del total de mujeres el 60% están afectadas con cáncer de mama, comprendidas en el grupo etareo a partir de los 55 años y más lo que demuestra que el grupo de estudio es a mayor edad mayor la frecuencia de cáncer de mama.
- El 55% de estas mujeres son casadas, el 20% eran solteras y sólo el 5% son de unión estable. Demostrando que el porcentaje mayor se presenta en mujeres casadas que posiblemente tendría más apoyo del cónyuge.
- Un 30% de las mujeres tienen 2 hijos, el 20% 1 hijo y el 15% es madre de 3 hijos. La condición de madre en esta situación ayuda el afrontamiento y vivencias de la paciente con cáncer de mama.
- Del total de mujeres con cáncer de mama el 40% son profesoras y rentistas de magisterio, el 25% son amas de casa, un 15% son comerciantes y profesionales de distintas ramas.

**Tabla 24**

*Apoyo brindado por la pareja, familia, equipo de salud y sociedad*

| *VARIABLE N=20* | **FRECUENCIA** | **%** |
|---|---|---|
| ***RELACIÓN CON LA PAREJA*** | | |
| Buena | 8 | 40% |
| Distante | 3 | 15% |
| Unida | 5 | 25% |
| No contesta | 4 | 20% |
| ***RELACIÓN CON LOS HIJOS*** | | |
| Buena | 15 | 75% |
| Regular | 1 | 5% |
| Unida | 4 | 20% |
| ***APOYO DE LA FAMILIA*** | | |
| SI | 12 | 60% |
| NO | 4 | 20% |
| Desconoce | 4 | 20% |
| ***EQUIPO DE SALUD*** | | |
| Atenta | 4 | 20% |
| Desatenta | 7 | 35% |
| Normal | 9 | 45% |
| **TOTAL** | **20** | **100%** |

Fuente: (Elaboración propia)

Interpretación Resultados **Tabla 24** Apoyo brindado por la pareja, familia, equipo de salud y sociedad.

- La investigación determinó que el 40% la relación con la pareja es buena, un 25% es unida, el 20% no contesta y el 15% es distante, sin embargo, estos indicadores son negativos para la mujer que tiene que afrontar y vivir con esta enfermedad.
- El 75% muestra que la relación es buena con los hijos, el 20% es unida, sin embargo el 5% es regular.

- El 60% la familia brinda apoyo a la mujer con cáncer de mama y 20% desconoce o no le da apoyo respectivamente, este indicador también es negativo.
- Las pacientes con cáncer de mama califican la atención normal del equipo de salud con 45%, el 35% la atención es desatenta el mismo que es un factor negativo, sólo el 25% la atención del equipo de salud es atenta donde se reconoce.
- El afrontamiento de la enfermedad depende mucho del apoyo que reciban por parte de su pareja, sin embargo para que surja aquello es necesario que exista comunicación entre ellos, aparentemente el apoyo brindado según resultados cuantitativos es positivo de parte de la pareja, familia y el mismo equipo de salud que ayuda en el afrontamiento.

**Tabla 25**

*Aceptación, Autoestima y Afrontamiento de la mujer con cáncer de mama*

| ***VARIABLE N=20*** | **FRECUENCIA** | **%** |
|---|---|---|
| ***ACEPTACIÓN*** | | |
| SÍ | 15 | 75% |
| NO | 5 | 25% |
| ***AUTOESTIMA*** | | |
| Plenamente | 6 | 30% |
| Parcialmente | 7 | 35% |
| Aparentemente | 7 | 35% |
| ***AFRONTAMIENTO*** | | |
| Ansiedad | 1 | 5% |
| Miedo | 5 | 25% |
| Depresión | 3 | 15% |
| Dependencia | 1 | 5% |
| Perturbación | 4 | 20% |
| Rabia | 2 | 10% |
| Otros | 4 | 20% |
| **TOTAL** | **20** | **100%** |

Fuente: (Elaboración propia)

Interpretación Resultados **Tabla 25** Aceptación, Autoestima y Afrontamiento de la mujer con cáncer de mama.

- La aceptación es un aspecto que se relaciona con el afrontamiento de la enfermedad, este proceso de aceptación para todo paciente es difícil, ya que ello implica una aceptación positiva o negativa. En las pacientes de cáncer de mama encuestadas, se determinó que el 75% aceptan su enfermedad. Asimismo el 25% no acepta.
- El 35% afirman que su autoestima cambio aparentemente y parcialmente, aspecto que interviene en el proceso de afrontamiento y aceptación del cáncer de mama. Mientras que el 30% indican que su cambio fue plenamente aspecto que influye en el modo de vida de las pacientes.
- La investigación detalla que el sentimiento de mayor presencia es el miedo con un 25%, la depresión un 15%, perturbación el 20% y Rabia el 10% por insatisfacción de los resultados de laboratorio aspectos que van relacionados con el afrontamiento y vivencias de la paciente frente al cáncer.

- **Encuestas previas de mujeres con cáncer de mama**

Se determinó como casos–tipo a 20 mujeres diagnosticadas con cáncer de mama que cumplen los criterios de inclusión, a quienes se les aplicó una encuesta semi-estructurada, con respuestas de selección múltiple.

Con el propósito de ampliar la información se aplicó a 6 personas dispuestas a colaborar en la investigación con la entrevista en profundidad.

**Tabla 26**

*Características de las 6 informantes*

| Nº | INFORMANTES | EDAD | ESTADO CIVIL | Nº DE HIJOS | OCUPACIÓN | SESIONES DE ENTREVISTA |
|---|---|---|---|---|---|---|
| **1** | Rosa | 43 | Casada | 3 | Profesora de técnicas | 3 |
| **2** | Azucena | 40 | Divorciada (15 años) | 2 | Costurera | 2 |
| **3** | Hortensia | 42 | Casada | 2 | Profesora | 2 |
| **4** | Margarita | 68 | Casada | 6 | Agricultora | 3 |
| **5** | Clavel | 38 | Casada | 1 | Costurera | 2 |
| **6** | Gladiolo | 31 | Casada | 1 | Profesora rural | 2 |

Fuente: (Elaboración propia)

**Las informantes fueron elegidas bajo los siguientes criterios**

- Pacientes diagnosticadas con cáncer de mama
- Mastectomizadas
- Que reciben quimioterapia de combinación
- De asistencia regular a la consulta
- De participación voluntaria en la investigación

**La situación de la mortalidad por cáncer en la mujer de los municipios de La Paz y El Alto en el primer semestre del año 2017.**

El fallecimiento de mujeres por cáncer en la ciudad de La Paz, determino que la Proporción de mortalidad fue (235/2760) * 100; con una tasa de mortalidad por cáncer 8,5 es decir 8.5% del total de muertes entre mujeres para el 2017 se deben algún tipo de cáncer.

**Tabla 27**

*Tasa de Mortalidad por cáncer en mujeres de la ciudad de La Paz Enero – junio 2017*

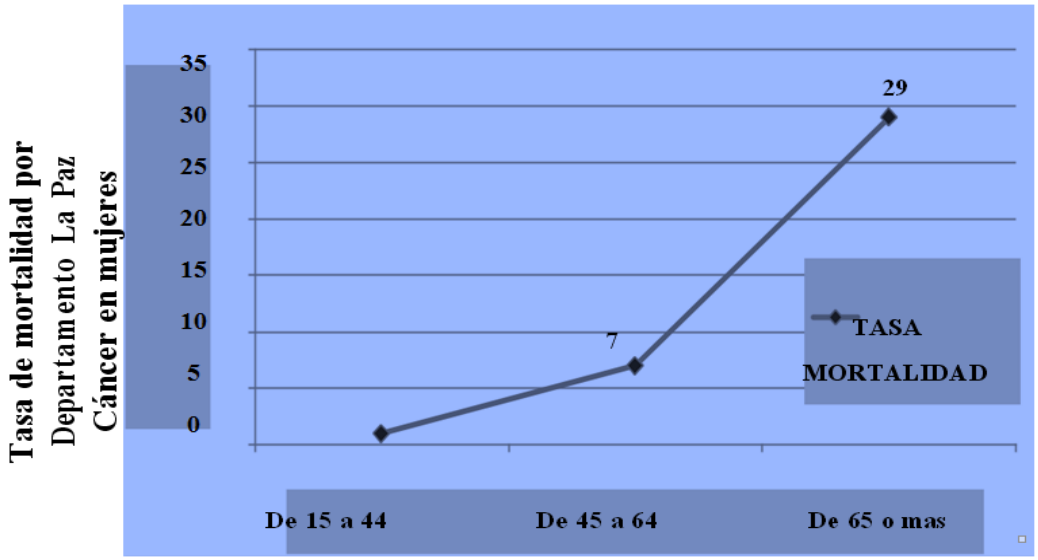
| GRUPOS DE EDAD EN AÑOS | NUMERO CASOS | POBLACIÓN | TASA | TASA CON REDONDEO |
|---|---|---|---|---|
| De 0 a 4 | **1** | **39421** | **0,254** | **0** |
| De 5 a 14 | **4** | **80705** | **0,496** | **0** |
| De 15 a 44 | **27** | **191281** | **1,412** | **1** |
| De 45 a 64 | **51** | **68372** | **7,459** | **7** |
| De 65 o mas | **95** | **32889** | **28,885** | **29** |
| **Total** | **178** | **412668** | **4,313** | **4** |

Fuente: (Elaboración propia con datos del estudio)

Durante el año 2017 la tasa de mortalidad por cáncer en la mujer en la ciudad de La Paz es de 4,31, es decir que de cada 10.000 mujeres que residen en la ciudad de la Paz 4 murieron a causa de algún tipo de cáncer.

**Figura 32**

*Tasas de mortalidad por cáncer de mama en mujeres del Municipio de La Paz enero - junio 2017*




Fuente: (Elaboración propia con datos del estudio)

**Edad en años de muertes por cáncer mujeres de la paz**

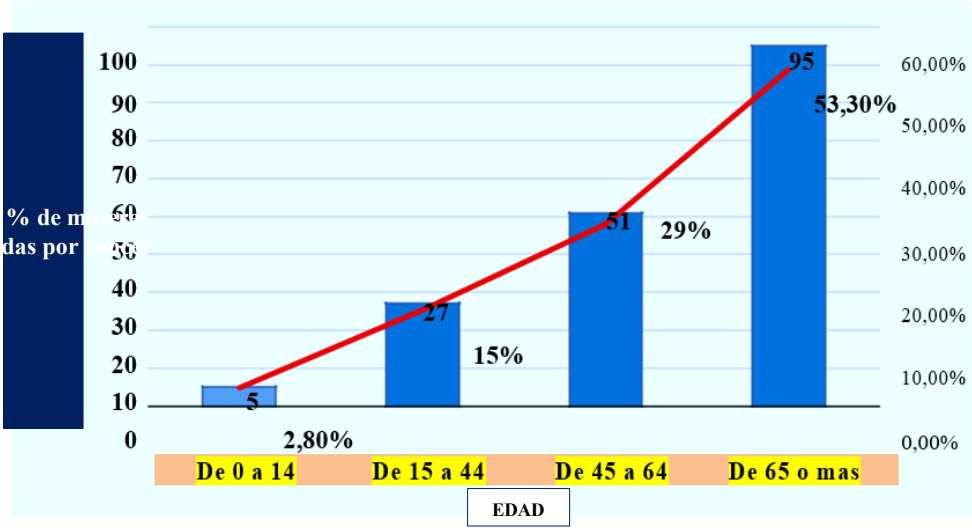
Entonces categorizamos la variable edad de 15-44 años la tasa de mortalidad por cáncer es 1, esto quiere decir que 1 /10.000 mujeres mueren por cáncer en ese rango de edad; de 45 a 64 años la tasa de mortalidad es 7 lo que es igual a 7 /10.000 de 45 a 64 años mueren a causa del cáncer y de 65 años a más la tasa de mortalidad fue 29 es decir 29/10.000.

**Edad de fallecimiento por cáncer en mujeres de municipios de La Paz y El Alto durante el primer semestre de la gestión 2017.**

Estratificamos la edad de las fallecidas y encontramos algunos datos de importancia, para fines prácticos encontramos que 0 a 14, de 15 a 44, 45 a 65 y de 65 años en adelante.

**Figura 33**

*Relación de mujeres fallecidas de cáncer de mama por grupos de edad en el municipio de La Paz enero-junio2017*




Fuente: (Elaboración propia con datos del estudio)

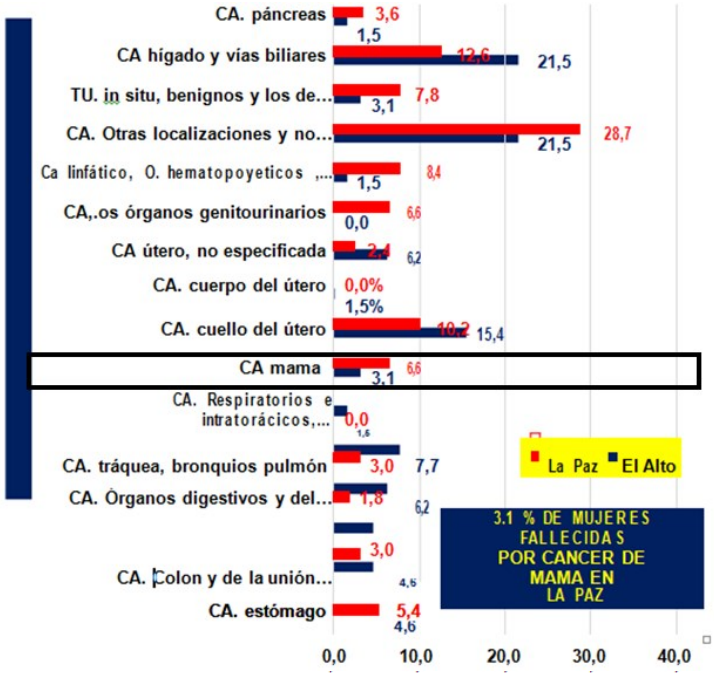
Encontramos que aproximadamente un (53,3 %) de las fallecidas tenían 65 años o más seguidas casi con un poco más de la mitad la incidencia mayor (29%) de las fallecidas

correspondían a las edades de 45 a 64 años, un dato llamativo y creo que es realmente el fruto de esta investigación a considerar que existe un (15%) es decir 15 de cada 100 fallecidas dentro la edad de 15 a 44 años fallece por algún tipo de cáncer, edad socio demográficamente muy importante con muchos aspecto debatibles como ejemplo que se trata de mujeres consideradas económicamente activas que fallecieron a causa de algún tipo de cáncer, en las cuales evidentemente existe una gran pérdida: en años , en economía y en la estructura conformacional de una sociedad de un departamento tan importante para el País.

- **La localización anatómica frecuente de cáncer en las mujeres de los municipios de La Paz y El Alto en el primer semestre de 2017.**

**Figura 34**

*Incidencia de localización anatómica de cáncer en las fallecidas de los municipios de La Paz y El Alto en el primer semestre de 2017*



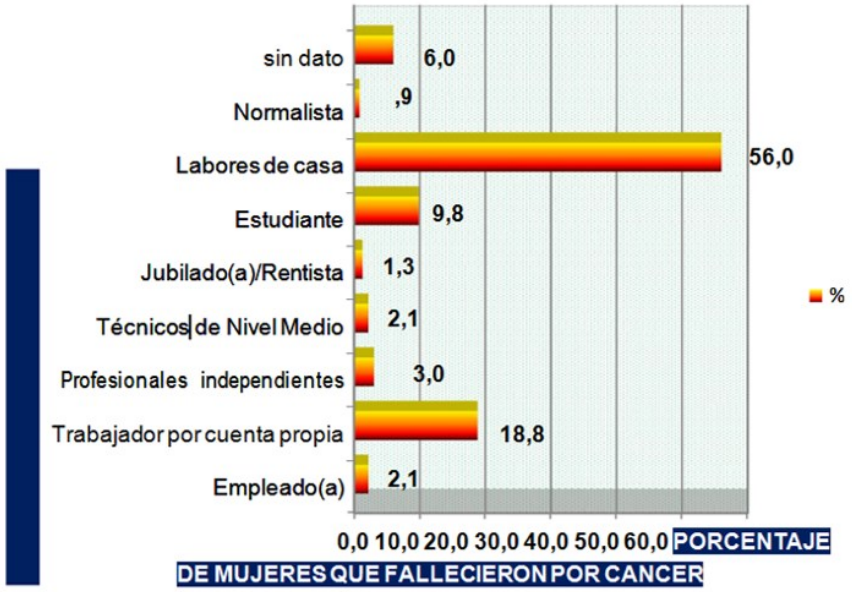

Fuente: (Elaboración propia con datos del estudio)

En quinto lugar, para La Paz, lo comparten cáncer de mama y cáncer de órganos genitourinarios (6,6%) de, para la ciudad del El Alto cáncer de cuerpo de Útero (6,2%) y cáncer de órganos digestivos y peritoneo (6,2%).

**El grado de instrucción y ocupación con el cáncer en las mujeres fallecidas en los municipios de La Paz durante primer semestre de 2017.**

**Figura 35**

*Distribución porcentual de ocupación y grado de instrucción de las mujeres que fallecieron por cáncer en la ciudad de La Paz enero -junio 2017*



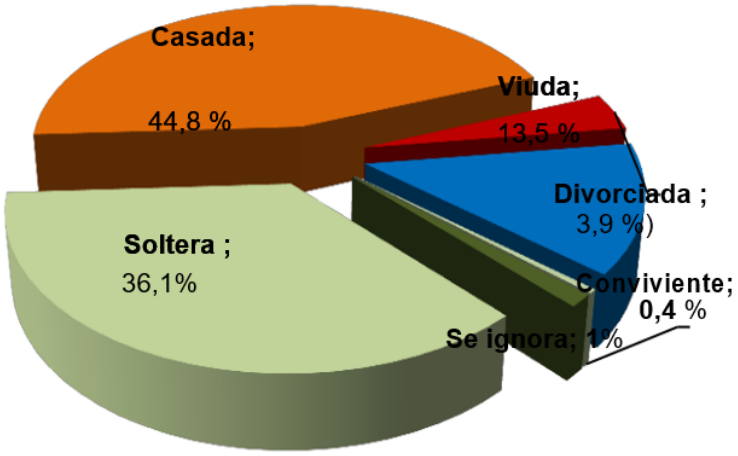

Fuente: (Elaboración propia con datos del estudio)

De las fallecidas por algún tipo de cáncer tenían como ocupación principal labores de casa (56%), un (18,8 %) de las fallecidas trabajaban por cuenta propia, el (9,8 %) eran estudiantes; (2,1%) técnico medio, (3%) siendo profesionales independientes; (2,1%) denominadas empleados y (1,3 %) dentro de la población jubilada o rentista.

**Estado Civil de las fallecidas por cáncer en los municipios de La Paz en el primer semestre 2017.**

**Figura 36**

*Distribución porcentual por estado civil de las mujeres fallecidas por algún tipo de Cáncer en las ciudades de La Paz Enero – Junio 2017*




Fuente: (Elaboración propia con datos del estudio)

El (44,8 %) de las fallecidas por algún tipo de cáncer estaban casadas, un (36,1%) eran solteras, un (13,5 %) eran viudas, divorciadas (3,9 %), se ignora (1%) y convivientes (0,4%).

En la **Figura 36** muestra el estado de mujeres fallecidas por cáncer de mama.

**Se determinó los APVP (años potenciales de vida pérdida) e IAPVP (Índice de Años Potenciales de Vida Perdidos) a causa de cáncer de mama en las mujeres de los municipios de La Paz en el primer semestre de 2017**

**Tabla 28**

*Cálculo de los APVP y de IAPVP debido al cáncer de mama por grupos de edad en las mujeres para el municipio de La Paz, enero –junio 2017*

| Edad en quinnquenios (1) | Punto o medio del intervalo(PMI) (2) | 76-PMI (3) | Nro muertes (4) | APVP (5) (3x4) | Numero de de Habitantes (6) | Indice APVP (7) (5/6)x1000 |
|---|---|---|---|---|---|---|
| **< 1 año** | 0,5 | 75,5 | 0 | 0 | **7899** | **0,0** |
| **1 - 4** | 2,5 | 73,5 | 1 | 73,5 | **31522** | **2,3** |
| **5 - 9** | 7,5 | 68,5 | 2 | 137 | **39877** | **3,4** |
| **10 - 14** | 12,5 | 63,5 | 2 | 127 | **40828** | **3,1** |
| **15 -19** | 17,5 | 58,5 | 1 | 58,5 | **38892** | **1,5** |
| **20 - 24** | 22,5 | 53,5 | 0 | 0 | **35598** | **0,0** |
| **25 - 29** | 27,5 | 48,5 | 1 | 48,5 | **32320** | **1,5** |
| **30 - 34** | 32,5 | 43,5 | 5 | 217,5 | **30903** | **7,0** |
| **35 - 39** | 37,5 | 38,5 | 7 | 269,5 | **28562** | **9,4** |
| **40 - 44** | 42,5 | 33,5 | 13 | 435,5 | **25006** | **17,4** |
| **45 - 49** | 47,5 | 28,5 | 6 | 171 | **21699** | **7,9** |
| **50 - 54** | 52,5 | 23,5 | 13 | 305,5 | **18362** | **16,6** |
| **55 - 59** | 57,5 | 18,5 | 13 | 240,5 | **15451** | **15,6** |
| **60 - 64** | 62,5 | 13,5 | 19 | 256,5 | **12860** | **19,9** |
| **65 - 68** | 66,5 | 9,5 | 19 | 180,5 | **8543** | **21,1** |
| **69 - 72** | 70,5 | 5,5 | 25 | 137,5 | **7238** | **19,0** |
| **73 - 75** | 74 | 2 | 11 | 22 | **4419** | **5,0** |
| **76 a mas** | 76 | 0 | 40 | 0 | **12688** | **0,0** |
| **Total** | | | **178** | **2680,5** | **412668** | **150,9** |

Fuente: (Elaboración propia con datos del estudio)

Los Años Potenciales de Vida Perdidos (APVP) constituyen un indicador que ilustra sobre la pérdida que sufre la población de La Paz como consecuencia de la muerte de personas jóvenes o de fallecimientos prematuros a causa del cáncer. Sin duda se trata de una medida del impacto relativo que ejercen diversas enfermedades como el Cáncer sobre la sociedad.

La cifra de los años potenciales de vida pérdida a consecuencia de una causa determinada es la suma, en todas las personas que fallecen por esta causa, de los años que habrían vivido si se hubieran cumplido las esperanzas de vida normales que poseían. Como se vio el cáncer afecta a las mujeres de 45 años para arriba, el cuadro antecedente llama la atención un índice de 17,4 IAPVP en mujeres paceñas que oscilan 40-44 años, ahí ponemos a juicio la definición de (PEA) población económicamente activa, que implica años doblemente perdidos.

Generalmente se opta por APVP-65, lo cual define la muerte prematura como la que ocurre antes de los 65años. Sin embargo, también debe promoverse el uso de las edades 75 y quizás 85 años. A medida que aumenta la longevidad, sino se cuentan las edades después de los 65 años, se está ignorando la carga de enfermedad de las enfermedades crónicas sobre la población, y probablemente también se reduce la variación observada entre poblaciones.

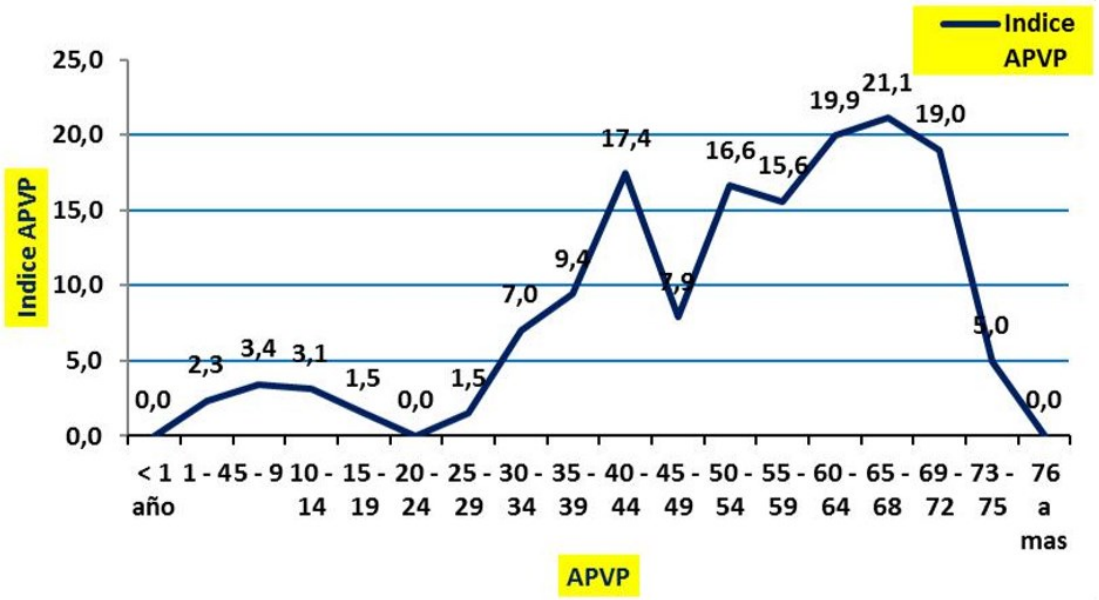
Siendo la carga de enfermedad crónica un claro indicador del estado de salud de la población, este enfoque puede no ser deseable. Otra medida empleada con frecuencia es APVP–Esperanza de Vida, o APVP (EV).De esta manera, todos los grupos de edad contribuyen a la APVP, dado que existe una esperanza de vida específica a cada grupo de edad. Este enfoque da origen a una medida más compleja de la APVP, en la cual la

esperanza de vida cambia con la edad y a lo largo del tiempo. Además, existen controversias sobre qué es más apropiado, si utilizar la esperanza de vida al nacer o utilizar la esperanza de vida al momento de la muerte. Se han usado ambos enfoques. Otros más han usado variantes de APVP que solamente cuentan los años de vida perdidos entre las edades de 15 y 65  Así, cincuenta años es el máximo que puede perder un individuo. Esta medida incorpora una declaración acerca del valor de la vida a una edad determinada. Puede ser útil desde el punto de vista económico, pero no lo es para diferenciar entre las poblaciones más y menos saludables.

Que después de explicar este acápite esta investigación reportó que en promedio para el municipio de La Paz IAPVP es de 15,9 años por cada mujer fallecida a causa de algún tipo de cáncer.

**Figura 37**

*Distribución del índice de Años Potenciales de Vida Perdidos, de las mujeres del municipio de La Paz, a causa cáncer para enero –junio 2017.*




Fuente: (Elaboración propia con datos del estudio)

Las mujeres fallecieron a causa de algún tipo de cáncer y la **Figura 37** refleja que hay cuatro picos, dos en ascenso y uno en descenso, los ascensos coinciden con 17,4 años a edades 40-44 años; uno en descenso 7,9 años a la edad de 45-46 años; otro ascenso 16,6 años a la edad de 50 años; y el último ascenso 21,2 años a la edad de los 65 años este último coincide con la esperanza de vida al nacer.

**Estadística inferencial: pruebas estadísticas de correlación.**

**Tabla 29**

*Relación de localización del cáncer de mama en la ciudad de La Paz con el Grado de Instrucción.*

| LOCALIZACIÓN DEL CANCER | NIVEL DE ESCOLARIDAD | | | |
|---|---|---|---|---|
| | PRIMARIA | SECUNDARIA | OTRO | TOTAL |
| **Ca. de mama en la mujer** | **6** | **1** | **3** | **10** |
| Ca. cuello/cuerpo del útero | 9 | 11 | 6 | 26 |
| Ca. del útero, parte no especificada | 1 | 3 | 3 | 7 |
| Ca. Órganos genitourinarios | 2 | 1 | 2 | 5 |
| **TOTAL** | **18** | **16** | **14** | **48** |

Fuente: (Elaboración propia con datos del estudio)

Para evaluar el comportamiento de una correlación de la variable localización cáncer y grado de escolaridad encontramos según la fuerza de correlación Pearson con una distribución normal, un valor de 0,049 lo que muestra relación significativa veamos:

**Tabla 30**

*Correlación de la variable localización de cáncer y grado de escolaridad en las mujeres fallecidas del municipio de La Paz durante el primer semestre 2017*

| ESTADÍSTICO | VALOR | GL | SIG. ASINTÓTICA (BILATERAL) |
|---|---|---|---|
| **Chi-Cuadrado De Pearson** | 16,125[A] | 5 | 0,049 |
| **Razón De Verosimilitudes** | 16,782 | 5 | 0,044 |
| **Asociación Lineal Por Lineal** | 10,971 | 1 | 0,034 |
| **N De Casos Válidos** | **48** | | |

Fuente: (Elaboración propia con datos del estudio)

El valor de probabilidad calculado Chi-cuadrado = 0,049 es relativamente menor al valor de significancia del 5%, concluyéndose que existe una relación significativa entre el tipo de Cáncer y el nivel de escolaridad.

Por tanto, el nivel de escolaridad es influyente sobre las causas de muerte por cáncer de mama.

Como vemos en la **Tabla 30**, vimos las localizaciones de cáncer asociados a nivel de escolaridad, añadimos a las variables cáncer genitourinario tratando de asociar a la ocupación de las fallecidas encontramos lo siguiente:

**Tabla 31**

*Relación de la variable localización de cáncer y ocupación en las mujeres fallecidas del municipio de La Paz durante el primer semestre 2017*

| | OCUPACIÓN | | | |
|---|---|---|---|---|
| **LOCALIZACIÓN DE CÁNCER** | **TRABAJADOR POR CUENTA PROPIA** | **LABORES DE CASA** | **OTRO** | **TOTAL** |
| **Ca. de mama en la mujer** | **2** | **7** | **4** | **13** |
| Ca. cuello/cuerpo del útero | 8 | 13 | 4 | 25 |
| Ca. del útero, parte no especificada | 3 | 3 | 1 | 7 |
| Ca. otros órganos genitourinarios | 1 | 3 | 2 | 6 |
| **TOTAL** | **14** | **26** | **11** | **51** |

Fuente: (Elaboración propia con datos del estudio)

La fuerza de asociación de estas dos variables según la prueba de Chi-cuadrado.

**Tabla 32**

*Correlación de la variable localización de cáncer y ocupación en las mujeres fallecidas del municipio de La Paz y El Alto durante el primer semestre 2017*

| **ESTADÍSTICO** | **VALOR** | **GL** | **SIG. ASINTÓTICA (BILATERAL)** |
|---|---|---|---|
| **Chi-Cuadrado De Pearson** | 13,282[A] | 6 | 0,03772711 |
| **Razón De Verosimilitudes** | 13,304962 | 6 | 0,03769711 |
| **Asociación Lineal Por Lineal** | 10 | 1 | 0,04836066 |
| **N De Casos Válidos** | **51** | | |

Fuente: (Elaboración propia con datos del estudio)

El valor de probabilidad calculado Chi-cuadrado = 0,038 es menor al valor de significancia del 5%, concluyéndose que existe una relación significativa entre el tipo de Cáncer genitourinario y la ocupación.

Por tanto, la ocupación es influyente sobre las causas de muerte por cáncer en órganos genitourinarios de la mujer. La variable tipo y localización de cáncer en órganos genitourinarios y la ocupación se correlacionan o se asocian.

Por otro lado, correlacionamos la variable localización de cáncer y el estado conyugal encontramos lo siguiente:

**Tabla 33**

*Relación entre tipo y localización del cáncer y estado conyugal ciudad de La Paz, enero a junio 2017.*

| **LOCALIZACIÓN DEL CANCER** | **ESTADO CONYUGAL** | | |
|---|---|---|---|
| | **Sin pareja** | **Con pareja** | **TOTAL** |
| **Ca. de mama en la mujer** | **6** | **7** | **13** |
| Ca. cuello/cuerpo del útero | 16 | 11 | 27 |
| Ca. del útero, parte no especificada | 5 | 2 | 7 |
| Ca. Órganos genitourinarios | 4 | 2 | 6 |
| **TOTAL** | **31** | **22** | **53** |

Fuente: (Elaboración propia con datos del estudio)

La fuerza de asociación de estas dos variables localización de cáncer y estado conyugal según la prueba de Chi-cuadrado**.**

**Tabla 34**

*Correlación de la variable localización de cáncer y ocupación en las mujeres fallecidas del municipio de La Paz durante el primer semestre 2017*

| ESTADÍSTICO | VALOR | GL | SIG. ASINTÓTICA (BILATERAL) |
|---|---|---|---|
| **Chi-Cuadrado de Pearson** | 14,469[A] | 3 | 0,0689375 |
| **Razón de Verosimilitudes** | 14,480441 | 3 | 0,0686792 |
| **Asociación Lineal por Lineal** | 19,33259 | 1 | 0,0334017 |
| **N De Casos Válidos** | **53** | | |

Fuente: (Elaboración propia con datos del estudio)

El valor de probabilidad calculado Chi-cuadrado = 0,69 es mayor al valor de significancia del 5%, concluyéndose que No existe una relación significativa entre estado conyugal y el tipo y localización de Cáncer en órganos genitourinarios.

Por tanto, el estado conyugal no es influyente sobre la causa de muerte por Cáncer.

**Situación de mortalidad por cáncer en la mujer de La Paz primer semestre 2017.**

De acuerdo a los resultados la mortalidad con respecto a 2009 en relación a grupo etéreo se aprecia un panorama similar.

Este estudio analiza la mortalidad por cáncer en mujeres que en 2009 presentó una tasa de 4,6 /10.000 comparado 2017 Pudimos observar una tasa de 4.3 / 10.000 para la ciudad de La Paz, que en este período de 8 años hubo un descenso discreto de la tasa global de mortalidad por cáncer en la mujer existiendo una razón de tasa 1 es decir la tasa actual es 1% menos que el año 2009. Esta tendencia es contraria a lo estimado.

**Edad promedio de las fallecidas por cáncer en la ciudad de La Paz.**

Categorizados los grupos etéreo para el estudio de referencia a edad promedio de las fallecidas por cáncer fue 64,1 años en nuestro estudio detectamos el promedio de las fallecidas superan los 65 años seguidas del rango de 45 a 64 años.

Se refleja en este estudio datos sobre cáncer de mama primera causa de muerte en el mundo, contrapone los datos obtenidos en los municipios de La Paz la tasa de mortalidad corresponde a un (6,6%).

**Situación de mortalidad por cáncer en la mujer en los municipios de La Paz.**

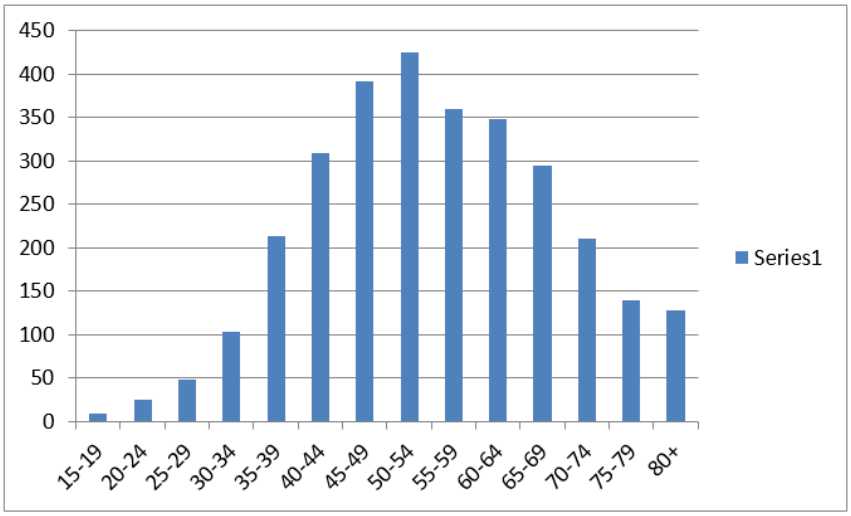
La tasa de mortalidad por cáncer en La Paz para el 2017 fue de 4,31, es decir que de 4 de 10.000 mujeres mueren por de cáncer mama.

**Situación epidemiológica del cáncer a nivel nacional (2016-2018)**

**Figura 38**

*Distribución de cáncer de mama por grupos de edad*

*2016-2018*




Fuente: (Ministerio de Salud , 2018)

En las ultimas 3 gestiones 2016- 2018, en el número total de defunciones por cáncer de mama fue de 587 lo que representa una media de 199 defunciones por año, el mayor porcentaje de fallecimientos ocurren en estados muy avanzados.

Rara vez se observa esta patología antes de los 25 años de edad por lo que se tiene un porcentaje mínimo en ese grupo de edad.

Los grupos de edad donde se registraron mayores casos de cáncer de mama son en el grupo de 40 hasta 64 años, donde se registran 1833 casos.

El grupo de edad donde se registraron mayores casos de fallecimientos por cáncer de mama es en el grupo de 40 hasta 64 años registrándose 283 muertes en segundo lugar los mayores de 65 años y en menor número de grupos de mujeres menores de 39.

La tasa de letalidad por cáncer de mama es de 19.5 %.

En el departamento de La Paz en las últimas 3 gestiones se registraron 1025 casos de cáncer de mama, para la gestión 2016 se registraron 384 casos, 2017 se registraron 369 y la gestión 2018 se registraron 272 casos nuevos de cáncer de mama.

En las ultimas 3 gestiones 2016 -2018 se registraron 185 defunciones por causa del cáncer de mama.

### *3.1.3 Preparación de los Datos (Fase III)*

En esta fase se describe las tareas que se realizan para la construcción de la tabla general de los datos la Incidencia de mujeres con cáncer de mama en la ciudad de La Paz.

1) **Selección de datos:** Una vez obtenido los datos históricos de las diferentes instituciones de salud, se realizó la selección de datos en común de las gestiones 2009 – 2018.

2) **Limpieza de datos:** Una vez obtenido los datos en común, se utilizó la técnica de normalización de datos, eliminando los campos con valores faltantes y se redujo el volumen de la cantidad de datos obteniendo la media de campos en común por cada gestión.

3) **Construcción de datos:** Se convirtieron los datos históricos del cáncer de mama en total de factor de riesgo por edades de 20 a 40 años de las gestiones 2009 al 2018.

4) **Integración de datos:** Una vez hecha la selección, limpieza y construcción de los datos históricos, se procedió a la integración de los datos de las gestiones 2009 – 2018 en una tabla general.

5) **Formateo de datos:** Una vez obtenida la tabla general de los datos históricos, se observó las inconsistencias en algunos campos donde se realiza el formateo de datos sin modificar el valor de los datos.

#### 3.1.3.1 Datos Estandarizados del cáncer de mama

Una vez realizado el proceso de preparación de datos se obtienen los datos ya estandarizados sobre el índice de crecimiento del cáncer de mama de mujeres de edades entre 20 a 40 años de la ciudad de La Paz, de las gestiones 2009 - 2018, además se realizó la clasificación de los parámetros en dos los cuales son:

- **N° Casos con cáncer de mama según factor de riesgo**

Aplicado las 3 primeras fases de la metodología CRISP-DM y las 3 primeras fases del Proceso KDD.

A continuación se muestra los datos estandarizados, donde se describe los parámetros según su clasificación.

**Tabla 35**

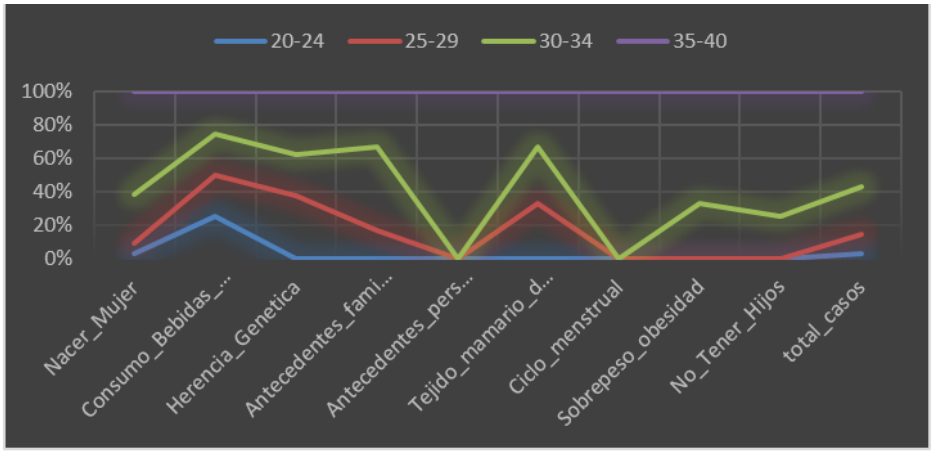
*Datos estandarizados del cáncer de mama de mujeres dividido por factor de riesgo y edades de 20 a 40 años de las gestiones 2009-2018 de la ciudad de La Paz*

| ID | Gestión | edad | Nacer Mujer | Consumo Bebidas Alcohólicas | Herencia genética | Antecedentes familiar | Antecedentes personales | Tejido mamario denso | Ciclo menstrual | Sobrepeso obesidad | No tener hijos | Total casos | Sexo | Tipo de riesgo |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 2009 | 20-24 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | F | bajo |
| 2 | 2009 | 25-29 | 2 | 2 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 8 | F | medio |
| 3 | 2009 | 30-34 | 10 | 1 | 2 | 3 | 0 | 1 | 0 | 1 | 2 | 20 | F | alto |
| 4 | 2009 | 35-40 | 21 | 1 | 3 | 2 | 1 | 1 | 3 | 2 | 6 | 40 | F | alto |
| 5 | 2010 | 20-24 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | F | bajo |
| 6 | 2010 | 25-29 | 2 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 6 | F | medio |
| 7 | 2010 | 30-34 | 5 | 2 | 3 | 1 | 0 | 2 | 1 | 2 | 3 | 19 | F | alto |
| 8 | 2010 | 35-40 | 19 | 1 | 2 | 3 | 2 | 1 | 1 | 3 | 2 | 34 | F | alto |
| 9 | 2011 | 20-24 | 2 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | F | bajo |
| 10 | 2011 | 25-29 | 3 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 6 | F | medio |
| 11 | 2011 | 30-34 | 9 | 2 | 1 | 1 | 1 | 0 | 2 | 3 | 2 | 21 | F | alto |
| 12 | 2011 | 35-40 | 17 | 2 | 1 | 3 | 1 | 1 | 4 | 3 | 4 | 36 | F | alto |
| 13 | 2012 | 20-24 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | F | bajo |
| 14 | 2012 | 25-29 | 2 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 7 | F | medio |
| 15 | 2012 | 30-34 | 6 | 3 | 1 | 1 | 0 | 2 | 1 | 1 | 4 | 19 | F | alto |
| 16 | 2012 | 35-40 | 12 | 3 | 2 | 0 | 1 | 3 | 1 | 2 | 6 | 30 | F | alto |
| 17 | 2013 | 20-24 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | F | bajo |
| 18 | 2013 | 25-29 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 4 | F | medio |
| 19 | 2013 | 30-34 | 5 | 2 | 1 | 1 | 0 | 2 | 2 | 2 | 3 | 18 | F | alto |
| 20 | 2013 | 35-40 | 16 | 2 | 4 | 1 | 1 | 1 | 2 | 3 | 4 | 34 | F | alto |
| 21 | 2014 | 20-24 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | F | bajo |
| 22 | 2014 | 25-29 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | F | bajo |
| 23 | 2014 | 30-34 | 6 | 1 | 2 | 1 | 1 | 2 | 1 | 0 | 6 | 16 | F | medio |
| 24 | 2014 | 35-40 | 14 | 2 | 1 | 3 | 1 | 2 | 3 | 1 | 5 | 32 | F | alto |
| 25 | 2015 | 20-24 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | F | alto |
| 26 | 2015 | 25-29 | 3 | 1 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 9 | F | bajo |
| 27 | 2015 | 30-34 | 7 | 2 | 1 | 1 | 1 | 0 | 2 | 1 | 2 | 17 | F | medio |
| 28 | 2015 | 35-40 | 10 | 3 | 1 | 2 | 1 | 3 | 1 | 1 | 5 | 27 | F | bajo |
| 29 | 2016 | 20-24 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | F | medio |
| 30 | 2016 | 25-29 | 2 | 2 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 7 | F | alto |
| 31 | 2016 | 30-34 | 6 | 1 | 3 | 1 | 0 | 0 | 2 | 1 | 1 | 15 | F | alto |
| 32 | 2016 | 35-40 | 9 | 3 | 1 | 4 | 1 | 3 | 0 | 1 | 5 | 27 | F | bajo |
| 33 | 2017 | 20-24 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | F | medio |
| 34 | 2017 | 25-29 | 3 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 5 | F | alto |
| 35 | 2017 | 30-34 | 4 | 0 | 2 | 1 | 0 | 1 | 0 | 1 | 2 | 11 | F | alto |
| 36 | 2017 | 35-40 | 9 | 2 | 1 | 4 | 1 | 0 | 0 | 2 | 5 | 24 | F | bajo |
| 37 | 2018 | 20-24 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | F | medio |
| 38 | 2018 | 25-29 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 4 | F | alto |
| 39 | 2018 | 30-34 | 5 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 9 | F | alto |
| 40 | 2018 | 35-40 | 6 | 1 | 3 | 4 | 2 | 3 | 1 | 0 | 2 | 22 | F | bajo |

Fuente: (Elaboración propia en base a datos obtenidos)

**Figura 39**

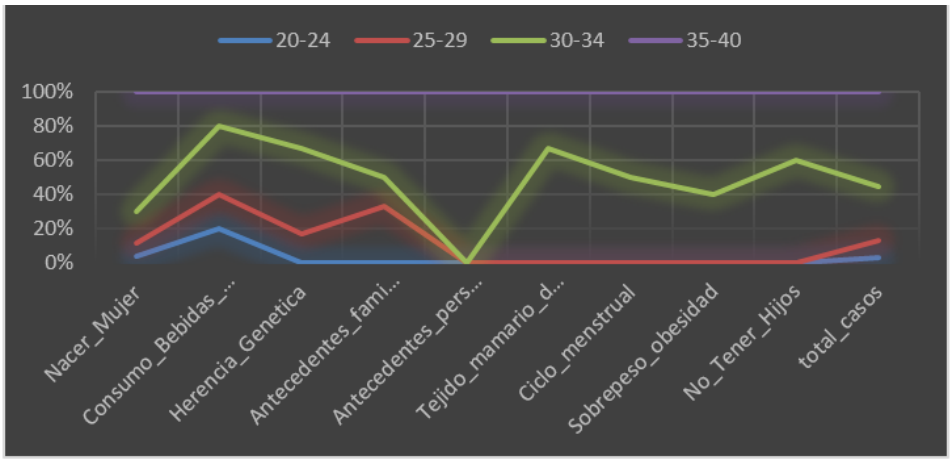
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2009*




Fuente: (Elaboracion Propia)

**Figura 40**

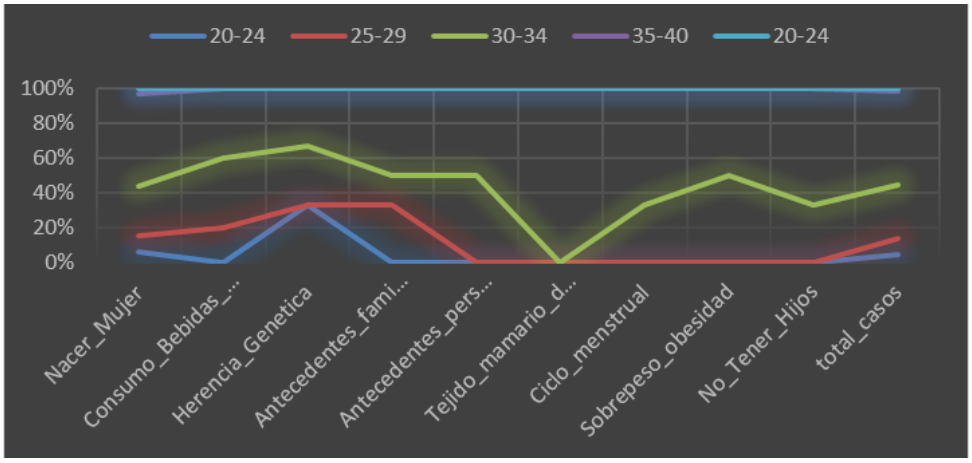
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2010*




Fuente: (Elaboracion Propia)

**Figura 41**

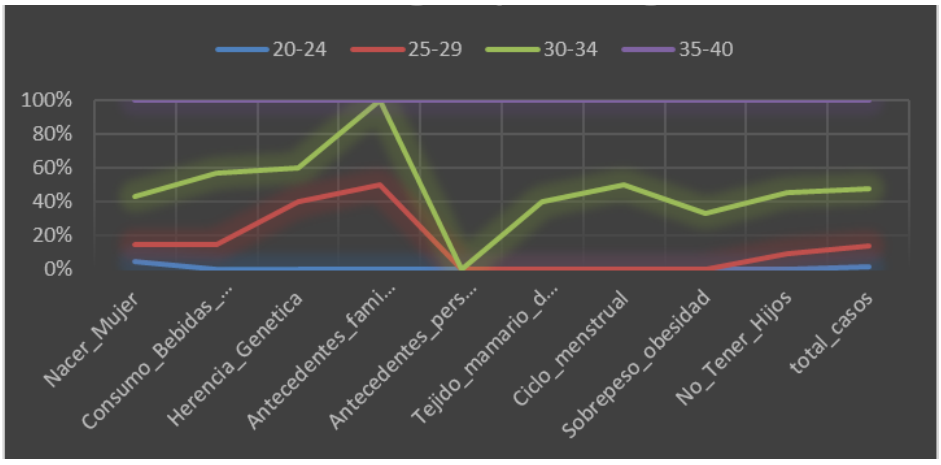
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2011*




Fuente: (Elaboracion Propia)

**Figura 42**

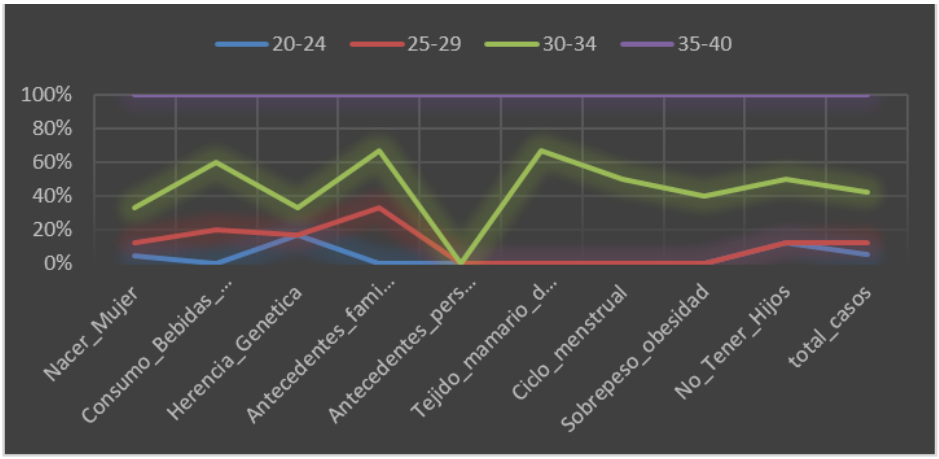
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2012*




Fuente: (Elaboracion Propia)

**Figura 43**

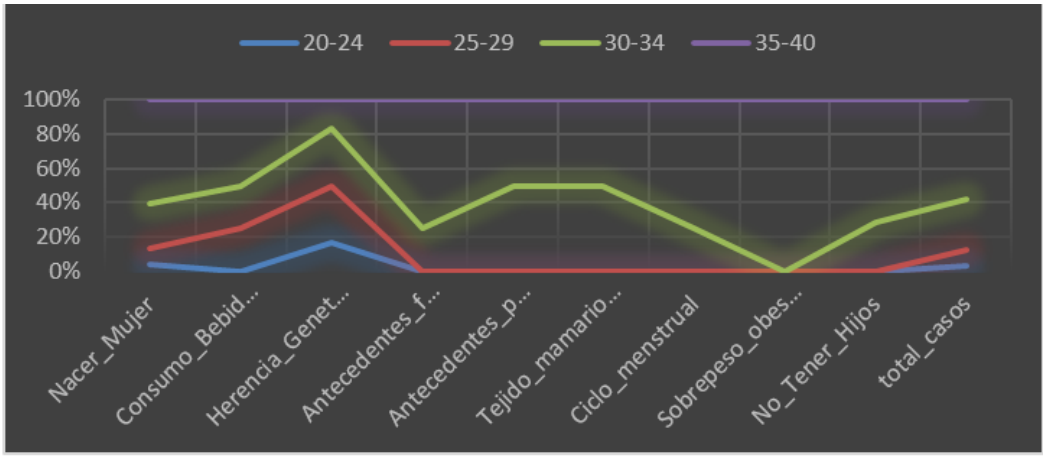
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2013*




Fuente: (Elaboracion Propia)

**Figura 44**

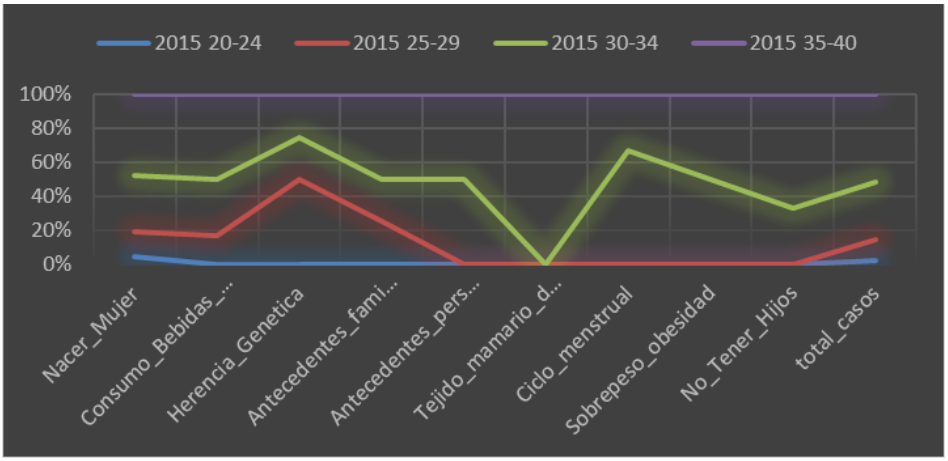
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2014*




Fuente: (Elaboracion Propia)

**Figura 45**

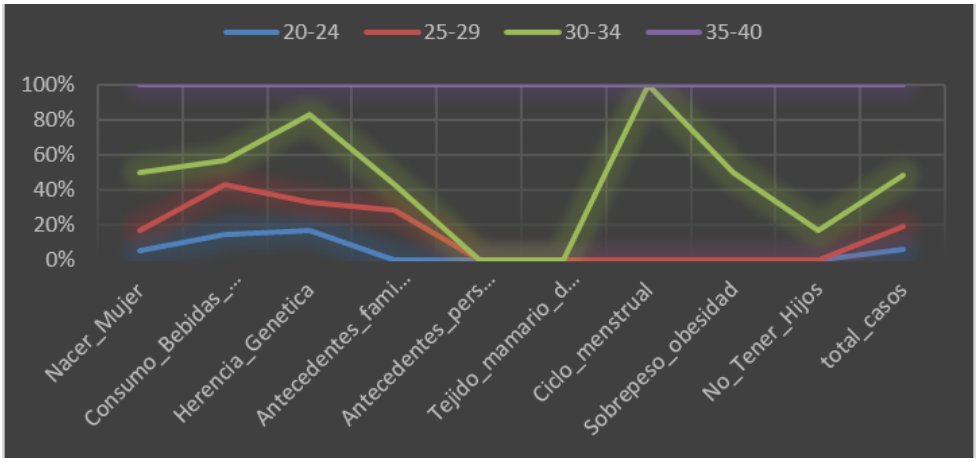
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2015*




Fuente: (Elaboracion Propia)

**Figura 46**

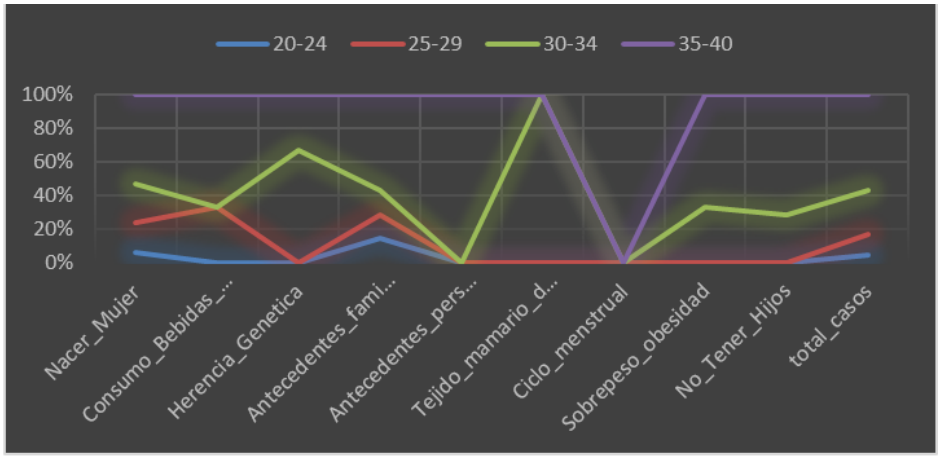
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2016*




Fuente: (Elaboracion Propia)

**Figura 47**

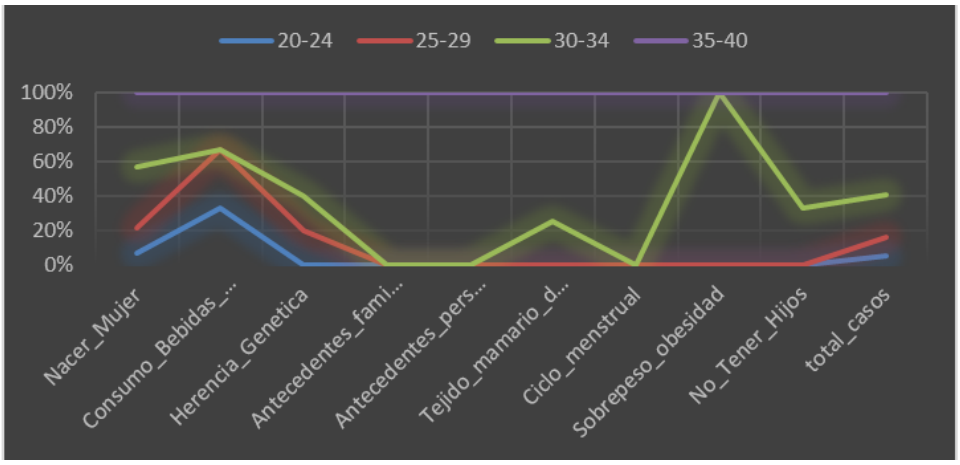
*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2017*




Fuente: (Elaboracion Propia)

**Figura 48**

*Distribución del cáncer de mama según el factor de riesgo de edades entre 20 a 40 años de la gestión 2018*




Fuente: (Elaboracion Propia)

**Tabla 36**

*Datos estandarizados del cáncer de mama dividido en total de edades entre 20 a 40 años y total de casos de 20 a 69 años de las gestiones 2009-2018 de la ciudad de La Paz*

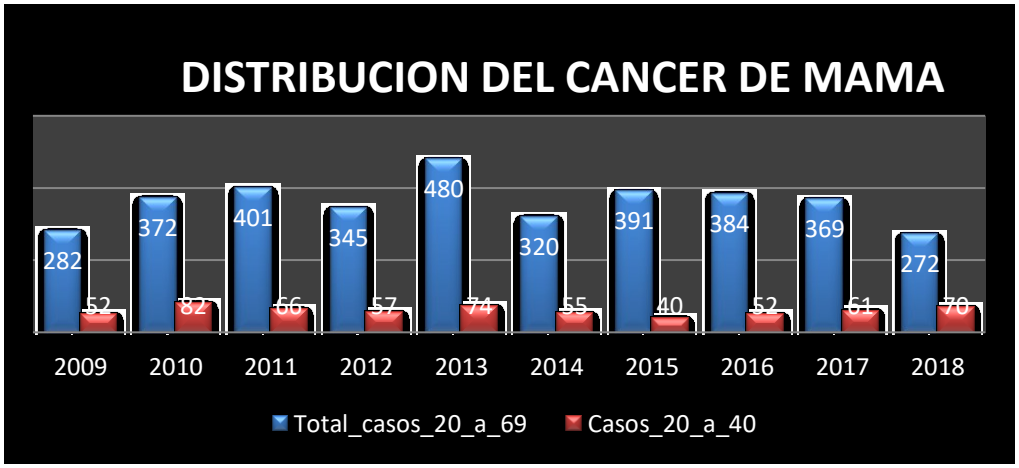
| ID | Año | Total_Casos_20_a_69 | Casos_20_a_40 |
|---|---|---|---|
| 1 | 2009 | 282 | 52 |
| 2 | 2010 | 372 | 82 |
| 3 | 2011 | 401 | 66 |
| 4 | 2012 | 345 | 57 |
| 5 | 2013 | 480 | 74 |
| 6 | 2014 | 320 | 55 |
| 7 | 2015 | 391 | 40 |
| 8 | 2016 | 384 | 52 |
| 9 | 2017 | 369 | 61 |
| 10 | 2018 | 272 | 70 |

Fuente: (Elaboración propia en base a datos obtenidos)

En la **Tabla 36** se describe los parámetros de las gestiones 2009, 2010, 2011, 2013, 2014, 2015, 2016, 2017 y 2018, datos que fueron recolectados de las distintas instituciones de salud y organizaciones mencionadas anteriormente.

**Figura 49**

*Distribución del cáncer de mama según gestiones (2009 - 2018) del total de casos de edades entre 20 a 69 años y edades entre 20 a 40 años*




Fuente: (Elaboracion Propia)

### *3.1.4 Modelado (Fase IV)*

En esta fase se procedió a creación el modelo, seleccionando las técnicas y algoritmos de Minería de Datos más apropiados al problema.

#### 3.1.4.1 Aplicación del Proceso (KDD)

Para generar el Modelo de Minería de Datos se aplica la fase de Minería de Datos del Proceso de Descubrimiento de Conocimiento en Base de Datos (KDD) seleccionando las técnicas y algoritmos adecuados para el Modelo Predictivo del índice de crecimiento del cáncer de mama para así descubrir patrones de información.

#### 3.1.4.2 Técnicas y Algoritmos de Minería de Datos

A continuación se describe las técnicas y algoritmos de Minería de Datos que se utilizaran para el modelado, que tiene integrado la herramienta Weka.

**Tabla 37**

*Selección de Técnicas y Algoritmos de Minería de Datos para el Modelado*

| TECNICAS | ALGORITMOS |
|---|---|
| • **Arboles de Decisión**<br>• **Series de Tiempo** | REPTree<br>RandomTree<br>J48<br>MultilayerPerceptron<br>M5P |

Fuente: (Elaboracion Propia)

#### 3.1.4.3 Estructura del Archivo Arff

A continuación se muestra con el Editor de Texto Sublime Text, la estructura de los Archivos con extensión .arff, los casos de cáncer de mujeres de edades entre 20 a 40 años.

- **Datos recolectados del cáncer de mama de edades entre 20 a 40 años según factor de riesgo y tipos de riesgo ( formato .arff)**

```
@Relation Casos_cancer_mama
@attribute ID numeric
@attribute Anyo Date yyyy
@attribute edad numeric
@attribute Nacer_Mujer numeric
@attribute Consumo_Bebidas_Alcoholicas numeric
@attribute Herencia_Genetica numeric
@attribute Antecedentes_familiar numeric
@attribute Antecedentes_personal numeric
@attribute Tejido_mamario_denso numeric
@attribute Ciclo_menstrual numeric
@attribute Sobrepeso_obesidad numeric
@attribute No_Tener_Hijos numeric
@attribute total_casos numeric
@attribute Sexo {F,M}
@attribute Tipo_Riesgo {bajo,medio,alto}
@data
1 ,2009,20,1,0,0,0,0,0,0,0,0,1,F,bajo
2 ,2009,22,0,1,0,0,0,0,0,0,0,1,F,bajo
3 ,2009,23,0,0,1,0,0,0,0,0,0,1,F,bajo
4 ,2009,26,1,1,0,0,0,0,0,0,0,2,F,bajo
5 ,2009,27,1,0,0,0,0,0,0,0,0,1,F,medio
6 ,2009,28,1,0,1,0,0,0,0,0,0,2,F,medio
7,2009,28,1,1,0,0,0,0,0,0,0,2,F,medio
8 ,2009,29,1,0,1,0,0,0,0,0,0,2,F,medio
9 ,2009,30,1,0,0,0,0,0,0,0,0,1,F,medio
10 ,2009,31,1,0,0,1,0,0,0,0,1,2,F,medio
11 ,2009,31,1,1,0,0,0,0,0,0,0,2,F,medio
12 ,2009,32,1,0,0,0,0,1,0,0,0,2,F,medio
13 ,2009,33,1,0,1,0,0,0,0,0,1,3,F,medio
14 ,2009,33,1,1,0,0,0,1,0,0,0,3,F,medio
15 ,2009,34,1,0,0,0,0,0,0,0,0,1,2,F,alto
16 ,2009,35,1,0,0,0,0,0,0,0,0,1,F,alto
17 ,2009,35,1,0,0,0,0,0,0,0,0,1,F,alto
18 ,2009,36,1,0,0,0,0,0,0,0,0,1,F,alto
19 ,2009,36,1,0,0,1,0,0,0,0,1,3,F,alto
20 ,2009,36,1,1,0,0,0,0,0,0,0,2,F,alto
21 ,2009,37,1,0,0,0,1,1,0,0,0,3,F,alto
22 ,2009,37,1,0,1,0,0,0,0,0,1,3,F,alto
23 ,2009,37,1,0,0,0,0,0,0,0,0,1,F,alto
24 ,2009,37,1,0,0,0,0,0,0,0,1,2,F,alto
25 ,2009,38,1,1,1,0,0,1,0,0,0,4,F,alto
26 ,2009,38,1,0,0,0,0,0,0,0,0,1,F,alto
27 ,2009,38,1,0,0,0,0,0,0,1,1,3,F,alto
28 ,2009,39,1,0,0,1,0,0,0,0,1,3,F,alto
29 ,2009,39,1,0,0,0,0,1,0,0,0,2,F,alto
30 ,2009,39,1,0,1,0,0,0,0,0,0,2,F,alto
31 ,2009,39,1,1,0,0,0,0,0,0,0,2,F,alto
32 ,2009,40,1,0,0,0,0,0,0,0,0,1,F,alto
33 ,2009,40,1,0,0,1,0,0,0,0,1,3,F,alto
34 ,2010,23,1,0,1,0,0,0,0,0,0,2,F,bajo
35 ,2010,24,1,0,0,0,0,0,0,0,0,1,F,bajo
```

```
36 ,2010,25,1,0,0,1,0,0,0,0,0,2,F,bajo
37 ,2010,25,1,1,0,1,0,0,0,0,0,3,F,bajo
38 ,2010,26,1,0,0,0,0,0,0,0,0,1,F,bajo
39 ,2010,30,1,0,0,0,0,0,0,1,1,3,F,medio
40 ,2010,31,1,1,0,0,1,0,0,0,0,3,F,medio
41 ,2010,32,1,0,1,0,0,0,1,0,1,4,F,medio
42 ,2010,33,1,0,0,0,0,0,0,0,0,1,F,medio
43 ,2010,33,1,1,0,1,0,0,1,1,1,6,F,medio
44 ,2010,33,1,0,0,0,0,0,0,0,0,1,F,medio
45 ,2010,34,1,0,0,0,0,0,0,1,1,3,F,alto
46 ,2010,35,1,0,0,0,0,0,0,1,0,0,2,F,alto
47 ,2010,35,1,0,0,0,1,1,0,0,1,4,F,alto
48 ,2010,36,1,1,1,0,0,1,0,1,1,6,F,alto
49 ,2010,36,1,0,0,0,0,0,0,0,0,1,F,alto
50 ,2010,37,1,0,0,0,0,0,1,1,1,4,F,alto
51 ,2010,38,1,0,0,1,0,0,0,0,0,2,F,alto
52 ,2010,38,1,1,0,0,0,1,0,0,0,3,F,alto
53 ,2010,38,1,0,0,0,0,0,0,1,1,0,3,F,alto
54 ,2010,39,1,0,0,0,0,0,0,1,1,1,4,F,alto
55 ,2010,40,1,0,0,0,0,0,0,0,1,0,2,F,alto
56 ,2011,23,0,1,0,0,0,0,0,0,0,1,F,bajo
57 ,2011,26,1,1,1,0,0,0,0,0,0,3,F,bajo
58 ,2011,29,1,0,1,1,0,0,0,0,0,3,F,medio
59 ,2011,31,1,1,0,0,0,0,1,0,1,4,F,medio
60 ,2011,32,1,0,0,1,0,1,0,1,0,4,F,medio
61 ,2011,32,1,1,1,0,0,0,0,0,1,4,F,medio
62 ,2011,34,1,0,1,0,0,1,0,0,1,4,F,alto
63 ,2011,34,1,1,0,0,0,0,0,0,1,3,F,alto
64 ,2011,35,1,0,0,0,0,0,0,0,0,1,F,alto
65 ,2011,35,1,1,0,0,0,0,1,0,0,3,F,alto
66 ,2011,36,1,0,0,1,0,0,1,0,0,3,F,alto
67 ,2011,37,1,1,0,0,0,0,0,0,1,3,F,alto
68 ,2011,37,1,0,0,0,0,0,1,1,0,0,3,F,alto
69 ,2011,37,1,1,1,0,0,0,1,0,1,5,F,alto
70 ,2011,37,1,0,0,0,0,0,0,1,1,3,F,alto
71 ,2011,38,1,1,1,0,0,1,0,0,0,4,F,alto
72 ,2011,39,1,0,0,0,1,1,0,1,1,5,F,alto
73 ,2011,39,1,0,0,1,0,0,0,0,1,3,F,alto
74 ,2011,40,1,0,0,1,0,0,0,0,1,3,F,alto
75 ,2012,21,1,1,0,0,0,0,0,0,0,2,F,bajo
76 ,2012,25,1,0,0,1,0,0,0,0,0,2,F,bajo
77 ,2012,28,1,1,1,0,0,0,0,0,0,3,F,medio
78 ,2012,31,1,0,0,1,0,0,1,0,0,3,F,medio
79 ,2012,32,1,1,0,0,0,0,0,0,0,2,F,medio
80 ,2012,33,1,0,1,0,1,0,0,1,1,5,F,alto
81 ,2012,34,1,1,0,0,0,0,0,1,1,4,F,alto
82 ,2012,34,1,0,0,0,0,0,1,0,1,3,F,alto
83 ,2012,35,1,1,0,0,0,0,0,0,0,2,F,alto
84 ,2012,35,1,0,0,0,0,0,1,0,0,0,2,F,alto
85 ,2012,36,1,0,1,0,0,0,0,1,0,3,F,alto
86 ,2012,37,1,1,0,0,1,0,1,0,1,5,F,alto
87 ,2012,37,1,0,1,1,0,0,0,1,1,5,F,alto
88 ,2012,37,1,0,1,0,0,0,1,1,1,5,F,alto
89 ,2012,38,1,0,0,1,0,0,1,1,1,5,F,alto
90 ,2012,39,1,0,0,0,0,0,0,0,1,2,F,alto
91 ,2012,39,1,0,1,0,0,0,0,0,1,3,F,alto
92 ,2013,20,1,0,1,0,0,0,0,0,0,2,F,bajo
```

```
93 ,2013,25,1,1,1,0,0,0,0,0,0,3,F,bajo
94 ,2013,27,1,0,1,0,0,0,0,0,0,2,F,medio
95 ,2013,30,1,1,0,0,0,1,1,0,1,5,F,medio
96 ,2013,32,1,0,1,0,0,0,0,0,1,3,F,medio
97 ,2013,33,1,0,1,0,0,1,0,0,1,4,F,medio
98 ,2013,33,1,0,0,1,1,0,0,0,1,4,F,medio
99 ,2013,34,0,0,0,0,0,0,0,0,1,1,F,alto
100 ,2013,34,0,0,0,0,0,0,0,0,1,1,F,alto
101 ,2013,36,1,0,1,0,0,0,1,0,1,4,F,alto
102 ,2013,36,1,0,0,1,0,0,0,1,0,3,F,alto
103 ,2013,37,1,0,1,0,0,0,0,1,0,3,F,alto
104 ,2013,38,1,0,0,1,0,0,0,0,0,2,F,alto
105 ,2013,38,1,1,0,1,0,0,0,0,1,3,F,alto
106 ,2013,39,1,1,0,0,0,0,0,0,0,1,F,alto
107 ,2013,39,1,1,0,0,1,1,0,0,1,5,F,alto
108 ,2013,39,1,0,0,0,0,1,0,0,1,3,F,alto
109 ,2013,40,1,1,0,0,0,0,1,0,0,3,F,alto
110 ,2013,40,1,0,0,0,0,0,1,0,1,3,F,alto
111 ,2014,23,1,1,0,0,0,0,0,0,0,2,F,bajo
112 ,2014,24,0,1,1,0,0,0,0,0,0,2,F,bajo
113 ,2014,26,1,1,1,1,0,0,0,0,0,3,F,bajo
114 ,2014,26,1,0,0,0,0,0,0,0,0,1,F,bajo
115 ,2014,28,1,0,1,0,0,0,0,0,0,2,F,medio
116 ,2014,30,1,0,0,0,0,0,0,0,0,1,F,medio
117 ,2014,32,1,1,1,0,0,0,1,0,1,5,F,medio
118 ,2014,32,1,1,0,0,1,0,0,0,1,3,F,medio
119 ,2014,34,1,0,0,1,0,0,1,1,1,5,F,alto
120 ,2014,34,0,0,0,0,0,0,0,0,1,1,F,alto
121 ,2014,35,1,0,1,0,0,1,0,0,1,3,F,alto
122 ,2014,35,1,1,0,1,0,1,0,0,0,4,F,alto
123 ,2014,36,1,0,0,1,0,0,1,0,1,4,F,alto
124 ,2014,38,1,0,0,0,0,0,0,1,1,3,F,alto
125 ,2014,39,1,1,0,0,0,1,0,1,0,4,F,alto
126 ,2014,39,1,1,0,0,0,1,0,1,0,4,F,alto
127 ,2014,39,1,0,0,0,0,1,0,0,1,3,F,alto
128 ,2014,40,1,0,0,0,1,0,0,0,1,3,F,alto
129 ,2015,23,1,1,0,0,0,0,0,0,0,2,F,bajo
130 ,2015,24,0,0,0,1,0,0,0,0,0,1,F,bajo
131 ,2015,26,1,1,0,0,0,0,0,0,0,2,F,bajo
132 ,2015,28,1,0,1,0,0,0,0,1,0,3,F,medio
133 ,2015,29,0,1,0,0,0,0,0,1,0,2,F,medio
134 ,2015,31,1,0,1,0,0,0,0,0,1,3,F,medio
135 ,2015,31,1,1,1,1,0,0,0,1,0,1,5,F,medio
136 ,2015,33,1,0,0,0,0,0,1,0,1,3,F,medio
137 ,2015,34,0,0,1,1,0,0,0,1,1,4,F,alto
138 ,2015,34,0,0,0,0,0,0,0,0,1,1,F,alto
139 ,2015,35,1,0,1,1,1,0,0,0,0,4,F,alto
140 ,2015,35,1,0,0,1,0,0,0,0,1,3,F,alto
141 ,2015,37,1,1,0,1,0,1,0,0,1,5,F,alto
142 ,2015,37,1,0,0,1,0,1,0,0,0,3,F,alto
143 ,2015,39,1,1,0,1,0,1,0,1,1,6,F,alto
144 ,2015,39,1,0,0,1,0,0,0,1,1,4,F,alto
145 ,2015,40,1,0,0,1,0,1,0,1,0,4,F,alto
146 ,2016,22,1,1,0,0,0,0,0,0,0,2,F,bajo
147 ,2016,24,0,0,1,0,0,0,0,0,0,1,F,bajo
148 ,2016,27,1,0,1,1,0,0,0,0,0,3,F,medio
149 ,2016,28,1,1,0,1,0,0,0,0,0,3,F,medio
```

```
150 ,2016,29,0,1,0,0,0,0,0,0,0,1,F,medio
151 ,2016,30,1,0,1,0,0,0,1,1,1,5,F,medio
152 ,2016,30,1,0,0,1,0,0,0,0,1,3,F,medio
153 ,2016,31,1,0,1,0,0,0,0,0,1,3,F,medio
154 ,2016,34,0,1,1,0,0,0,0,0,1,3,F,alto
155 ,2016,35,1,0,1,1,1,1,0,0,1,6,F,alto
156 ,2016,35,1,0,0,1,0,1,0,1,0,4,F,alto
157 ,2016,36,1,1,0,1,1,1,0,1,0,6,F,alto
158 ,2016,37,1,0,0,1,1,0,0,1,1,5,F,alto
159 ,2016,39,1,1,0,0,0,1,0,0,1,4,F,alto
160 ,2016,40,0,0,0,1,0,1,0,0,1,3,F,alto
161 ,2017,20,1,0,0,0,0,0,0,0,0,1,F,bajo
162 ,2017,23,0,1,0,0,0,0,0,0,0,1,F,bajo
163 ,2017,24,0,0,1,0,0,0,0,0,0,1,F,bajo
164 ,2017,25,1,0,0,0,0,0,0,0,0,1,F,bajo
165 ,2017,26,1,0,0,0,0,0,0,0,0,1,F,bajo
166 ,2017,29,1,1,0,0,0,0,0,0,0,2,F,medio
167 ,2017,31,1,0,0,1,0,0,0,0,1,3,F,medio
168 ,2017,33,1,0,1,0,0,0,0,1,0,3,F,medio
169 ,2017,33,1,0,1,0,0,1,0,0,0,3,F,medio
170 ,2017,34,1,0,0,0,0,0,0,0,1,2,F,alto
171 ,2017,35,1,1,0,0,0,0,0,1,0,3,F,alto
172 ,2017,36,1,1,0,1,1,0,0,0,0,4,F,alto
173 ,2017,36,1,0,0,1,1,0,0,0,1,4,F,alto
174 ,2017,37,1,0,0,0,0,0,0,0,1,2,F,alto
175 ,2017,38,0,1,0,0,0,0,0,0,1,2,F,alto
176 ,2017,38,1,0,1,1,0,0,0,0,1,4,F,alto
177 ,2017,39,1,0,0,1,0,0,0,1,1,4,F,alto
178 ,2017,40,0,0,0,0,0,0,0,0,1,1,F,alto
179 ,2018,20,1,0,0,0,0,0,0,0,0,1,F,bajo
180 ,2018,21,0,1,0,0,0,0,0,0,0,1,F,bajo
181 ,2018,25,1,0,1,0,0,0,0,0,0,2,F,bajo
182 ,2018,26,0,1,0,0,0,0,0,0,0,1,F,bajo
183 ,2018,28,1,0,0,0,0,0,0,0,0,1,F,medio
184 ,2018,30,1,0,0,0,0,0,0,0,1,2,F,medio
185 ,2018,31,1,0,0,0,0,1,0,0,0,2,F,medio
186 ,2018,32,1,0,1,0,0,0,0,0,0,2,F,medio
187 ,2018,32,1,0,0,0,0,0,0,0,1,2,F,medio
188 ,2018,34,1,0,0,1,0,0,0,0,0,2,F,alto
189 ,2018,35,1,0,0,0,0,1,0,0,1,3,F,alto
190 ,2018,36,1,0,1,1,0,1,0,0,0,4,F,alto
191 ,2018,37,1,0,0,1,0,1,0,0,0,3,F,alto
192 ,2018,38,1,1,1,1,0,0,1,0,0,5,F,alto
193 ,2018,39,1,0,1,0,1,0,0,0,0,3,F,alto
194 ,2018,40,1,0,0,0,0,0,0,0,0,1,F,alto
194 ,2018,40,0,0,0,0,1,0,0,0,0,1,F,alto
194 ,2018,40,0,0,0,0,0,0,0,0,1,1,F,alto
```

Fuente: (Elaboracion Propia)

- **Total casos del cáncer de mama de edades entre 20 a 40 años y el total de cáncer de mama de edades entre 20 a 69 años de las gestiones 2009-2018 (formato .arff)**

```
@Relation Casos_Totales_Casos_Edades
@attribute ID numeric
@attribute Total_Casos numeric
@attribute Casos_20_a_40 numeric
@attribute Anyo Date yyyy
@data
1,282,52,2009
2,372,82,2010
3,401,66,2011
4,345,57,2012
5,480,74,2013
6,320,55,2014
7,391,40,2015
8,384,52,2016
9,369,61,2017
10,272,70,2018
```

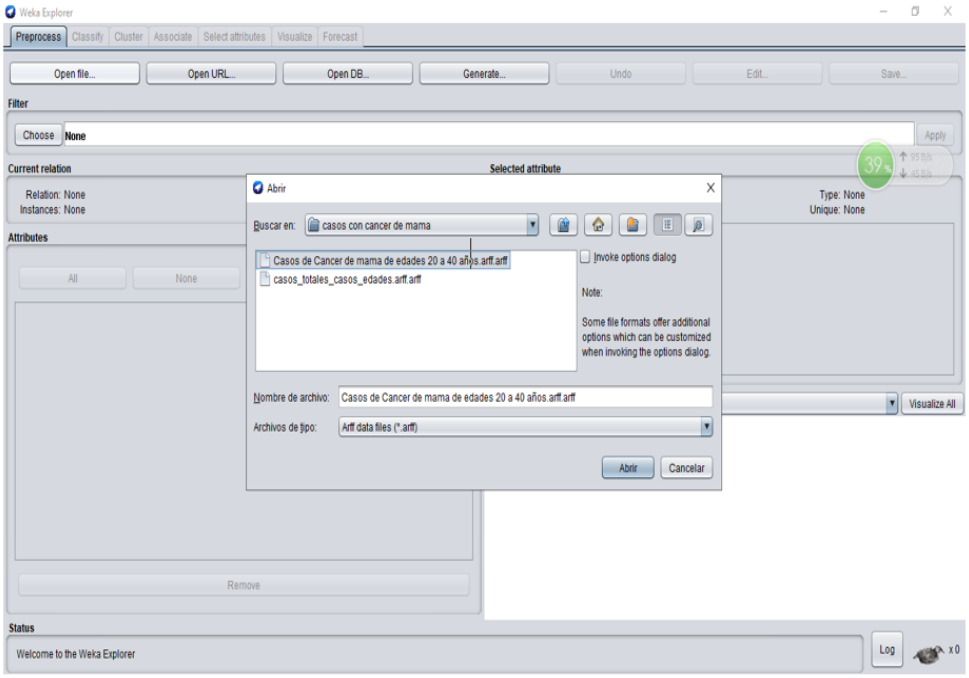
Fuente: (Elaboracion Propia)

#### 3.1.4.4 Entrenando en la Herramienta Weka

Para el entrenamiento del modelo se aplicara los algoritmos en la herramienta Weka.

**Figura 50**

*Cargado de los archivos .arff en Weka*




Fuente: (Elaboracion Propia)

En la **Figura 50**, se muestra como se procedió a cargar los archivos en la herramienta Weka.

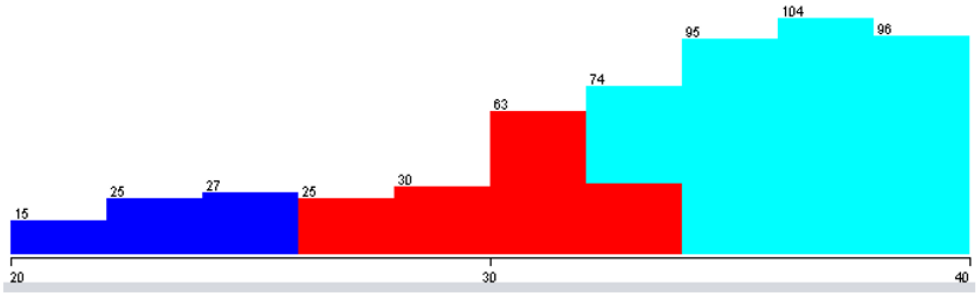
### 3.1.4.5 Aplicación de Técnicas y Algoritmos en Weka

Se aplican los algoritmos seleccionados en la herramienta Weka, para el Modelo de Predicción de Minería de Datos.

- **Tipo de riesgo del cáncer de mama de mujeres de 20 a 40 años aplicado en WEKA**

**Figura 51**

*Tipo de riesgo por edad del cáncer de mama de mujeres entre 20 a 40 años de las gestiones 2009 - 2018 aplicado en WEKA*




Fuente: (Elaboracion Propia)

Se obtuvo en la **Figura 51** el número total del cáncer de mama de mujeres de 20 a 40 años y por tipo de riesgos que padecen.

**Tabla 38**

*Total de casos por tipos de riesgos de 20 a 40 años de las gestiones 2009-2018 de la ciudad de La Paz*

| Edad | Tipo de Riesgo | Total de casos |
|---|---|---|
| 20-26 | Bajo | 67 |
| 27-34 | Medio | 118 |
| 32-34 | Medio | 31 |
| 32-34 | Alto | 43 |
| 35-40 | Alto | 295 |

Fuente: (Elaboracion Propia)

En la **Tabla 38** se puede observar los totales de casos según los tipos de riesgo (Bajo, Medio y Alto) dividido por edades.

#### 3.1.4.6 Aplicando Algoritmos de Minería de Datos

Para el entrenamiento del Modelo se aplica los Algoritmos REPTree, RandomTree y J48 con los datos preparados del cáncer de mama de las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, de las gestiones 2009 -2018.

**1) Resultado obtenido del entrenamiento del Modelo aplicado con el algoritmo REPTree para la toma de decisiones.**

```
=== Run information ===

Scheme:       weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L
-1 -I 0.0

Relation:     Casos_cancer_mama-
weka.filters.unsupervised.attribute.Remove-R1-2,14

Instances:    554
Attributes:   12
              edad
              Nacer_Mujer
              Consumo_Bebidas_Alcoholicas
              Herencia_Genetica
              Antecedentes_familiar
              Antecedentes_personal
              Tejido_mamario_denso
              Ciclo_menstrual
              Sobrepeso_obesidad
              No_Tener_Hijos
              total_casos
              Tipo_Riesgo
Test mode:    10-fold cross-validation
=== Classifier model (full training set) ===


REPTree
============

edad < 33.5
|   edad < 26.5 : bajo (44/0) [23/0]
|   edad >= 26.5 : medio (104/5) [50/0]
edad >= 33.5 : alto (221/0) [112/0]

Size of the tree : 5

Time taken to build model: 0.01 seconds
```

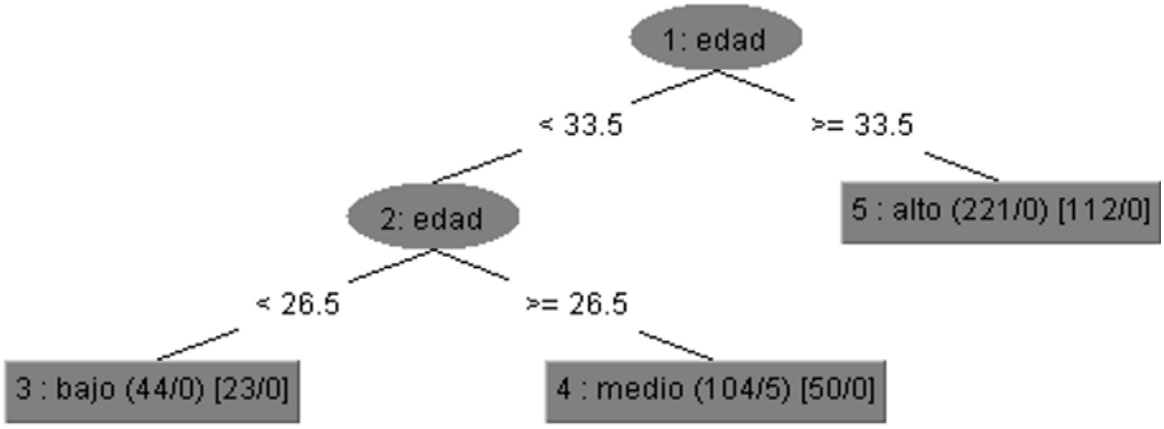
```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         549               99.0975 %
Incorrectly Classified Instances         5                0.9025 %
Kappa statistic                          0.9834
Mean absolute error                      0.0117
Root mean squared error                  0.0768
Relative absolute error                  3.233  %
Root relative squared error             18.0886 %
Total Number of Instances              554

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     bajo
               1,000    0,012    0,968      1,000    0,983      0,978    0,990     0,943     medio
               0,985    0,000    1,000      0,985    0,993      0,981    0,992     0,996     alto
Weighted Avg.  0,991    0,003    0,991      0,991    0,991      0,983    0,992     0,982

=== Confusion Matrix ===

   a   b   c   <-- classified as
  67   0   0 |   a = bajo
   0 149   0 |   b = medio
   0   5 333 |   c = alto
```

Fuente: (Elaboracion Propia)

En los resultados obtenidos con el algoritmo REPTre, de los casos de cáncer de mama de edades entre 20 a 40 años de la ciudad de La Paz, se muestra en una matriz de confusión los tres tipos de riesgo y la cantidad de mujeres por edad, donde 67 se encuentra en riesgo bajo, 149 se encuentra en riesgo medio, y 333 en riesgo alto y 5 podrían estar en riesgo medio.

**Figura 52**

*Árbol de decisión por edad y tipo de riesgo con algoritmo REPTree*




Fuente: (Elaboracion Propia)

## 2) Resultado obtenido del entrenamiento del Modelo aplicado con el Algoritmo RandomTree para la toma de decisiones.

```
=== Run information ===

Scheme:       weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1
Relation:     Casos_cancer_mama-weka.filters.unsupervised.attribute.Remove-R1-
2,14
Instances:    554
Attributes:   12
              edad
              Nacer_Mujer
              Consumo_Bebidas_Alcoholicas
              Herencia_Genetica
              Antecedentes_familiar
              Antecedentes_personal
              Tejido_mamario_denso
              Ciclo_menstrual
              Sobrepeso_obesidad
              No_Tener_Hijos
              total_casos
              Tipo_Riesgo
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===


RandomTree
==========

Consumo_Bebidas_Alcoholicas < 0.5
|   edad < 33.5
|   |   Ciclo_menstrual < 0.5
|   |   |   Nacer_Mujer < 0.5
|   |   |   |   Herencia_Genetica < 0.5
|   |   |   |   |   Antecedentes_familiar < 0.5
```

```
|   |   |   |   |   |   Tejido_mamario_denso < 0.5
|   |   |   |   |   |   |   edad < 26.5 : bajo (2/0)
|   |   |   |   |   |   |   edad >= 26.5
|   |   |   |   |   |   |   |   Antecedentes_personal < 0.5
|   |   |   |   |   |   |   |   |   edad < 32.5 : medio (21/0)
|   |   |   |   |   |   |   |   |   edad >= 32.5
|   |   |   |   |   |   |   |   |   |   Sobrepeso_obesidad < 0.5 : medio (6/1)
|   |   |   |   |   |   |   |   |   |   Sobrepeso_obesidad >= 0.5 : medio (3/1)
|   |   |   |   |   |   |   |   Antecedentes_personal >= 0.5
|   |   |   |   |   |   |   |   |   edad < 32.5 : medio (2/0)
|   |   |   |   |   |   |   |   |   edad >= 32.5 : medio (2/1)
|   |   |   |   |   |   |   Tejido_mamario_denso >= 0.5 : medio (7/0)
|   |   |   |   |   |   Antecedentes_familiar >= 0.5
|   |   |   |   |   |   |   edad < 26 : bajo (4/0)
|   |   |   |   |   |   |   edad >= 26 : medio (10/0)
|   |   |   |   |   Herencia_Genetica >= 0.5
|   |   |   |   |   |   edad < 26 : bajo (10/0)
|   |   |   |   |   |   edad >= 26
|   |   |   |   |   |   |   edad < 32.5 : medio (17/0)
|   |   |   |   |   |   |   edad >= 32.5 : medio (5/1)
|   |   |   |   Nacer_Mujer >= 0.5
|   |   |   |   |   edad < 26.5 : bajo (28/0)
|   |   |   |   |   edad >= 26.5
|   |   |   |   |   |   edad < 32.5 : medio (42/0)
|   |   |   |   |   |   edad >= 32.5 : medio (11/1)
|   |   |   Ciclo_menstrual >= 0.5 : medio (9/0)
|   |   edad >= 33.5 : alto (304/0)
Consumo_Bebidas_Alcoholicas >= 0.5
|   No_Tener_Hijos < 0.5
|   |   edad < 33.5
|   |   |   edad < 26.5 : bajo (23/0)
|   |   |   edad >= 26.5 : medio (19/0)
|   |   edad >= 33.5 : alto (28/0)
|   No_Tener_Hijos >= 0.5 : alto (1/0)

Size of the tree : 41

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         547               98.7365 %
Incorrectly Classified Instances         7                1.2635 %
Kappa statistic                          0.9767
Mean absolute error                      0.0138
Root mean squared error                  0.0994
Relative absolute error                  3.8262 %
Root relative squared error             23.4092 %
Total Number of Instances              554


=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,985    0,000    1,000      0,985    0,992      0,991    0,993     0,987     bajo
               0,993    0,015    0,961      0,993    0,977      0,968    0,988     0,951     medio
               0,985    0,005    0,997      0,985    0,991      0,977    0,990     0,991     alto
Weighted Avg.  0,987    0,007    0,988      0,987    0,987      0,977    0,990     0,980

=== Confusion Matrix ===

   a   b   c   <-- classified as
```

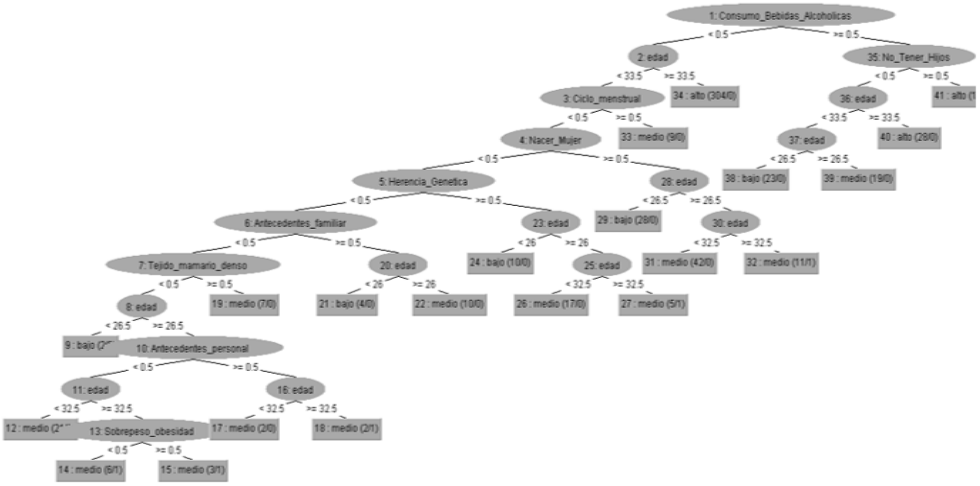
```
66   1   0 |   a = bajo
 0 148   1 |   b = medio
 0   5 333 |   c = alto
```

Fuente: (Elaboracion Propia)

En los resultados obtenidos con el algoritmo RandomTree, de los casos de cáncer de mama de edades entre 20 a 40 años de la ciudad de La Paz se muestra en una matriz de confusión los 3 tipos de riesgo y la cantidad de mujeres por edad 66 se encuentra en riesgo bajo, 148 se encuentra en riesgo medio ,1 se encuentra en riesgo alto y 333 en riesgo alto y 5 podrían estar en riesgo medio.

**Figura 53**

*Árbol de decisión por edad y tipo de riesgo con algoritmo RandomTree*




Fuente: (Elaboracion Propia)

Se visualiza en la **Figura 53 y Anexo A3** se muestra los resultados mediante un Árbol de decisión entrenado con el algoritmo RandomTree, donde se tiene como principal factor de riesgo (no tener hijos) de edades >=33 años, además el tipo de riesgo en padecer el cáncer de mama es Alto, por lo que se recomienda hacerse exámenes médicos y exploratorios del cáncer de mama en la mujeres de edades entre 33 a 40 años para la detección temprana.

**3) Resultado obtenido del entrenamiento del Modelo aplicado con el Algoritmo J48 para la toma de decisiones.**

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     Casos_cancer_mama-
weka.filters.unsupervised.attribute.Remove-R1-
weka.filters.unsupervised.attribute.Remove-R1
Instances:    554
Attributes:   13
              edad
              Nacer_Mujer
              Consumo_Bebidas_Alcoholicas
              Herencia_Genetica
              Antecedentes_familiar
              Antecedentes_personal
              Tejido_mamario_denso
              Ciclo_menstrual
              Sobrepeso_obesidad
              No_Tener_Hijos
              total_casos
              Sexo
              Tipo_Riesgo
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

edad <= 33
|   edad <= 26: bajo (67.0)
|   edad > 26: medio (154.0/5.0)
edad > 33: alto (333.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         549               99.0975 %
Incorrectly Classified Instances         5                0.9025 %
Kappa statistic                          0.9834
Mean absolute error                      0.0117
Root mean squared error                  0.0768
Relative absolute error                  3.233  %
```

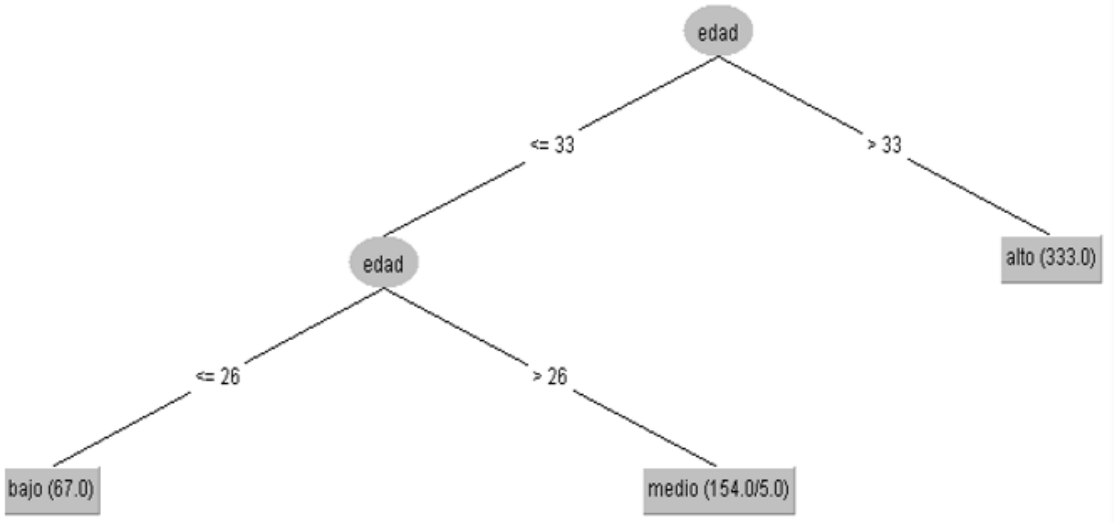
```
Root relative squared error                18.0886 %
Total Number of Instances                554
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     bajo
               1,000    0,012    0,968      1,000    0,983      0,978    0,990     0,943     medio
               0,985    0,000    1,000      0,985    0,993      0,981    0,992     0,996     alto
Weighted Avg.  0,991    0,003    0,991      0,991    0,991      0,983    0,992     0,982

=== Confusion Matrix ===

   a   b   c   <-- classified as
  67   0   0 |   a = bajo
   0 149   0 |   b = medio
   0   5 333 |   c = alto
```

Fuente: (Elaboracion Propia)

En los resultados obtenidos con el algoritmo J48 de los casos de cáncer de mama de edades entre 20 a 40 años de la ciudad de La Paz se muestra en una matriz de confusión los tres tipos de riesgo y la cantidad de mujeres por edad 67 se encuentra en riesgo bajo, 148 se encuentra en riesgo medio y 1 se encuentra en riesgo alto, y 333 en riesgo alto y 5 podrían estar en riesgo medio.

**Figura 54**

*Árbol de decisión por edad y tipo de riesgo con algoritmo J48*




Fuente: (Elaboracion Propia)

**4) Resultado obtenido con el algoritmo M5P, para la predicción a futuro de las gestiones (2009-2018) por un intervalo de 5 años.**

```
=== Run information ===

Scheme:
      M5P -M 4.0

Lagged and derived variable options:
      -F Total_Casos,Casos_20_a_40 -L 1 -M 5 -G Anyo

Relation:     Casos_Totales_Casos_Edades-
weka.filters.unsupervised.attribute.Remove-R1
Instances:    10
Attributes:   3
              Total_Casos
              Casos_20_a_40
              Anyo

Transformed training data:

             Total_Casos
             Casos_20_a_40
             Anyo-remapped
             Lag_Total_Casos-1
             Lag_Total_Casos-2
             Lag_Total_Casos-3
             Lag_Total_Casos-4
             Lag_Total_Casos-5
             Lag_Casos_20_a_40-1
             Lag_Casos_20_a_40-2
             Lag_Casos_20_a_40-3
             Lag_Casos_20_a_40-4
             Lag_Casos_20_a_40-5
             Anyo-remapped^2
             Anyo-remapped^3
             Anyo-remapped*Lag_Total_Casos-1
             Anyo-remapped*Lag_Total_Casos-2
             Anyo-remapped*Lag_Total_Casos-3
             Anyo-remapped*Lag_Total_Casos-4
             Anyo-remapped*Lag_Total_Casos-5
             Anyo-remapped*Lag_Casos_20_a_40-1
             Anyo-remapped*Lag_Casos_20_a_40-2
             Anyo-remapped*Lag_Casos_20_a_40-3
             Anyo-remapped*Lag_Casos_20_a_40-4
             Anyo-remapped*Lag_Casos_20_a_40-5

Total_Casos:
M5 pruned model tree:
(using smoothed linear models)

Lag_Total_Casos-2 <= 369.5 : LM1 (3/10.966%)
Lag_Total_Casos-2 >  369.5 :
|   Lag_Total_Casos-2 <= 387.5 : LM2 (4/13.383%)
|   Lag_Total_Casos-2 >  387.5 : LM3 (3/20.378%)
```

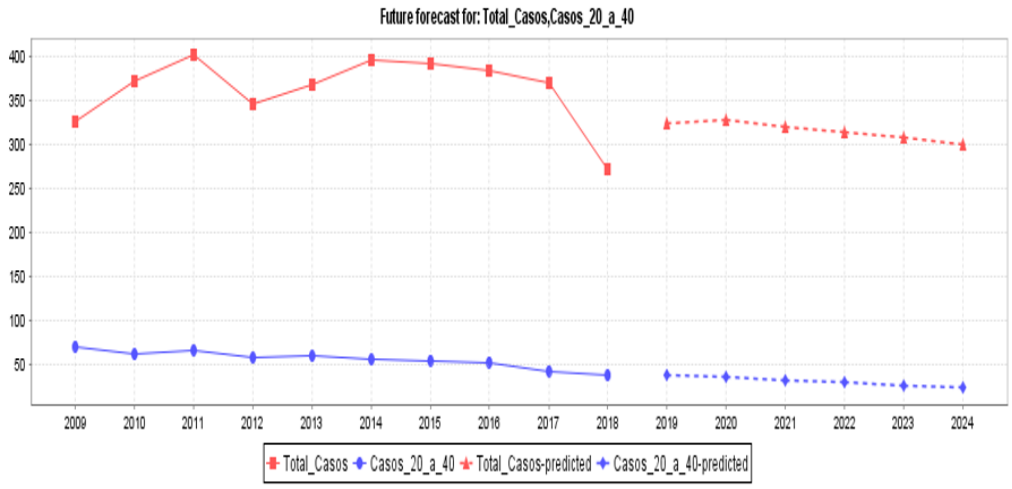
```
LM num: 1
Total_Casos =
      2.1601 * Lag_Casos_20_a_40-1
      + 244.0716

LM num: 2
Total_Casos =
      0.6386 * Lag_Total_Casos-2
      + 3.0005 * Lag_Casos_20_a_40-1
      - 60.8492

LM num: 3
Total_Casos =
      0.674 * Lag_Total_Casos-2
      + 2.8302 * Lag_Casos_20_a_40-1
      - 61.7247

Number of Rules : 3

Casos_20_a_40:
M5 pruned model tree:
(using smoothed linear models)
LM1 (10/34.334%)

LM num: 1
Casos_20_a_40 =
      -3.1071 * Anyo-remapped
      + 69.2888

Number of Rules : 1

=== Future predictions from end of training data ===
Time    Total_Casos_20_a_69 Casos_20_a_40
2009            282            52
2010            372            82
2011            401            66
2012            345            57
2013            480            74
2014            320            55
2015            391            40
2016            384            52
2017            369            61
2018            272            70
2019*           323            38
2020*           326            35
2021*           319            31
2022*           313            28
2023*           306            25
```

Fuente: (Elaboracion Propia)

Se demuestra los resultados obtenidos con el algoritmo M5P de la incidencia del cáncer de mama de las mujeres de la ciudad de La Paz, de los casos totales por gestiones y edades entre 20 a 40 años, la predicción realizada es de un intervalo de 5 gestiones (2019 – 2023).

**Figura 55**

*Resultados obtenidos de la predicción del total de Casos de las Mujeres de edades entre 20 a 40 años, aplicado con el algoritmo M5P*




Fuente: (Elaboracion Propia)

**5) Resultado obtenido con el algoritmo RandomTree, para la predicción a futuro de las gestiones (2009-2018) para un intervalo de 5 años.**

```
=== Run information ===

Scheme:
                            RandomTree -K 0 -M 1.0 -V 0.001 -S 1

Lagged and derived variable options:
                            -F Total_Casos,Casos_20_a_40 -L 1 -M 5 -G
Anyo

Relation:     Casos_Totales_Casos_Edades
Instances:    10
Attributes:   4
              ID
              Total_Casos
              Casos_20_a_40
```

```
              Anyo

Transformed training data:

              Total_Casos
              Casos_20_a_40
              Anyo-remapped
              Lag_Total_Casos-1
              Lag_Total_Casos-2
              Lag_Total_Casos-3
              Lag_Total_Casos-4
              Lag_Total_Casos-5
              Lag_Casos_20_a_40-1
              Lag_Casos_20_a_40-2
              Lag_Casos_20_a_40-3
              Lag_Casos_20_a_40-4
              Lag_Casos_20_a_40-5
              Anyo-remapped^2
              Anyo-remapped^3
              Anyo-remapped*Lag_Total_Casos-1
              Anyo-remapped*Lag_Total_Casos-2
              Anyo-remapped*Lag_Total_Casos-3
              Anyo-remapped*Lag_Total_Casos-4
              Anyo-remapped*Lag_Total_Casos-5
              Anyo-remapped*Lag_Casos_20_a_40-1
              Anyo-remapped*Lag_Casos_20_a_40-2
              Anyo-remapped*Lag_Casos_20_a_40-3
              Anyo-remapped*Lag_Casos_20_a_40-4
              Anyo-remapped*Lag_Casos_20_a_40-5

Total_Casos:

RandomTree
==========

Lag_Casos_20_a_40-3 < 78
|   Anyo-remapped*Lag_Total_Casos-1 < 3198.56
|   |   Lag_Total_Casos-1 < 396
|   |   |   Anyo-remapped^3 < 0.5 : 282 (0.52/-0)
|   |   |   Anyo-remapped^3 >= 0.5
|   |   |   |   Lag_Casos_20_a_40-1 < 53.5
|   |   |   |   |   Lag_Total_Casos-3 < 400 : 369.9 (1.43/1.89)
|   |   |   |   |   Lag_Total_Casos-3 >= 400 : 380.4 (1.43/30.24)
|   |   |   |   Lag_Casos_20_a_40-1 >= 53.5 : 395.62 (1.86/24.85)
|   |   Lag_Total_Casos-1 >= 396
|   |   |   Anyo-remapped < 1.5 : 282 (0.22/0)
|   |   |   Anyo-remapped >= 1.5
|   |   |   |   Lag_Casos_20_a_40-3 < 59 : 345 (1/0)
|   |   |   |   Lag_Casos_20_a_40-3 >= 59 : 320 (1/0)
|   Anyo-remapped*Lag_Total_Casos-1 >= 3198.56 : 273 (1.11/9)
Lag_Casos_20_a_40-3 >= 78 : 441.5 (1.43/4228.65)

Size of the tree : 17

Casos_20_a_40:

RandomTree
```

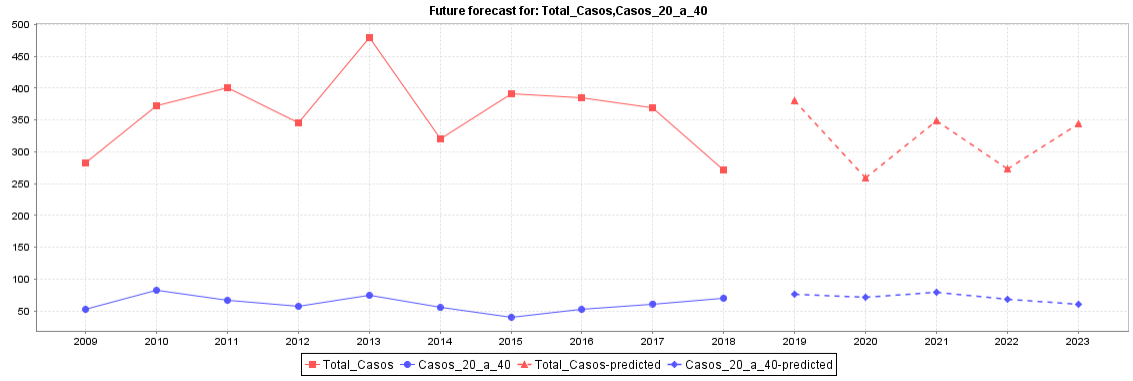
```
==========

Lag_Total_Casos-4 < 332.5
|   Lag_Casos_20_a_40-1 < 63.5
|   |   Anyo-remapped^2 < 0.5 : 52 (0.26/0)
|   |   Anyo-remapped^2 >= 0.5
|   |   |   Lag_Total_Casos-1 < 313.5 : 82 (0.33/-0)
|   |   |   Lag_Total_Casos-1 >= 313.5 : 72 (2/4)
|   Lag_Casos_20_a_40-1 >= 63.5 : 60.55 (0.74/26.35)
Lag_Total_Casos-4 >= 332.5
|   Anyo-remapped < 2.5
|   |   Lag_Total_Casos-1 < 327 : 72 (1/200)
|   |   Lag_Total_Casos-1 >= 327 : 61.33 (1/43.56)
|   Anyo-remapped >= 2.5
|   |   Lag_Total_Casos-2 < 435.5
|   |   |   Anyo-remapped < 7.5
|   |   |   |   Lag_Casos_20_a_40-1 < 53 : 52 (1/0)
|   |   |   |   Lag_Casos_20_a_40-1 >= 53 : 55.8 (1.67/0.96)
|   |   |   Anyo-remapped >= 7.5 : 61 (1/0)
|   |   Lag_Total_Casos-2 >= 435.5 : 40 (1/0)

Size of the tree : 19

=== Future predictions from end of training data ===
Time   Total_Casos_20_a_69 Casos_20_a_40
2009            282            52
2010            372            82
2011            401            66
2012            345            57
2013            480            74
2014            320            55
2015            391            40
2016            384            52
2017            369            61
2018            272            70
2019*           380            76
2020*           258            71
2021*           348            79
2022*           272            67
2023*           343            60
```

Fuente: (Elaboracion Propia)

Se muestra los resultados obtenidos con el algoritmo RandomTree de la incidencia del cáncer de mama de las mujeres de la ciudad de La Paz de los casos totales por gestiones y edades entre 20 a 40 años, la predicción realizada es de un intervalo de 5 gestiones (2019 - 2023).

**Figura 56**

*Resultados obtenidos del total de casos de mujeres de edades entre 20 a 40 años, aplicado con el algoritmo RandomTree*




Fuente: (Elaboracion Propia)

**6) Resultado obtenido con el algoritmo MultilayerPerceptron, para la predicción a futuro de las gestiones (2009-2018) para un intervalo de 5 años.**

```
=== Future predictions from end of training data ===
Time    Total_Casos_20_a_69 Casos_20_a_40
2009            282             52
2010            372             82
2011            401             66
2012            345             57
2013            480             74
2014            320             55
2015            391             40
2016            384             52
2017            369             61
2018            272             70
2019*           410             76
2020*           269             67
2021*           384             84
2022*           280             74
2023*           323             81
```

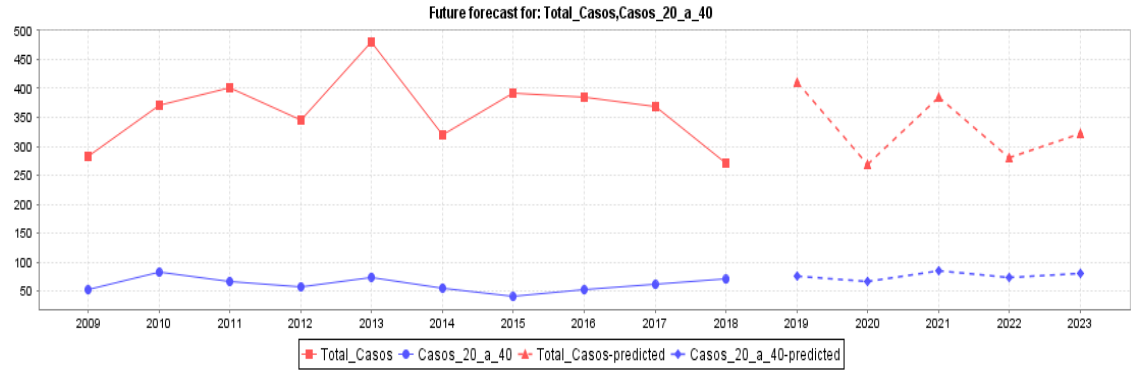
Fuente: (Elaboracion Propia)

Se muestra los resultados obtenidos con el algoritmo MultilayerPerceptron de la incidencia del cáncer de mama de las mujeres de la ciudad de La Paz, de los casos totales

por gestiones y edades entre 20 a 40 años, la predicción realizada es de un intervalo de 5 gestiones (2019 - 2023).

**Figura 57**

*Resultados obtenidos de la predicción del total de Casos de las Mujeres de edades entre 20 a 40 años, aplicado con el algoritmo MultilayerPerceptron*




Fuente: (Elaboracion Propia)

### *3.1.5 Evaluación (Fase V)*

En esta fase se mostrara la evaluación de cada algoritmo entrenado en el Modelo.

- **Evaluación de los resultados obtenidos con el Algoritmo REEPTre**

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         549               99.0975 %
Incorrectly Classified Instances         5                0.9025 %
Kappa statistic                          0.9834
Mean absolute error                      0.0117
Root mean squared error                  0.0768
Relative absolute error                  3.233  %
Root relative squared error             18.0886 %
Total Number of Instances              554

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     bajo
                 1,000    0,012    0,968      1,000    0,983      0,978    0,990     0,943     medio
                 0,985    0,000    1,000      0,985    0,993      0,981    0,992     0,996     alto
Weighted Avg.    0,991    0,003    0,991      0,991    0,991      0,983    0,992     0,982
```

```
=== Confusion Matrix ===
   a   b   c   <-- classified as
  67   0   0 |   a = bajo
   0 149   0 |   b = medio
   0   5 333 |   c = alto
```

Fuente: (Elaboracion Propia)

- **Evaluación de los resultados obtenidos con el Algoritmo RandomTree**

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         547               98.7365 %
Incorrectly Classified Instances         7                1.2635 %
Kappa statistic                          0.9767
Mean absolute error                      0.0138
Root mean squared error                  0.0994
Relative absolute error                  3.8262 %
Root relative squared error             23.4092 %
Total Number of Instances              554
=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,985    0,000    1,000      0,985    0,992      0,991    0,993     0,987     bajo
               0,993    0,015    0,961      0,993    0,977      0,968    0,988     0,951     medio
               0,985    0,005    0,997      0,985    0,991      0,977    0,990     0,991     alto
Weighted Avg.  0,987    0,007    0,988      0,987    0,987      0,977    0,990     0,980

=== Confusion Matrix ===
   a   b   c   <-- classified as
  66   1   0 |   a = bajo
   0 148   1 |   b = medio
   0   5 333 |   c = alto
```

Fuente: (Elaboracion Propia)

- **Evaluación de los resultados obtenidos con el Algoritmo J48**

```
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances         549               99.0975 %
Incorrectly Classified Instances         5                0.9025 %
Kappa statistic                          0.9834
Mean absolute error                      0.0117
Root mean squared error                  0.0768
```

```
Relative absolute error                  3.233  %
Root relative squared error             18.0886 %
Total Number of Instances               554

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     bajo
               1,000    0,012    0,968      1,000    0,983      0,978    0,990     0,943     medio
               0,985    0,000    1,000      0,985    0,993      0,981    0,992     0,996     alto
Weighted Avg.  0,991    0,003    0,991      0,991    0,991      0,983    0,992     0,982

=== Confusion Matrix ===
   a   b   c   <-- classified as
  67   0   0 |   a = bajo
   0 149   0 |   b = medio
   0   5 333 |   c = alto
```

Fuente: (Elaboracion Propia)

- **Comparativa de Resultados Evaluados**

En la tabla 39, podemos observar un cuadro comparativo de los algoritmos de Minería de Datos ya entrenados en el modelo y procedemos a evaluar cada algoritmo entrenado mediante los resultados obtenidos.

**Tabla 39**

*Cuadro comparativo de los resultados obtenidos con los Algoritmos entrenados para el Modelo Predictivo*

| Ecuación | Instancias clasificadas correctamente | Error de la Media Absoluta | Error Absoluto Relativo | Error cuadrático medio de raíz |
|---|---|---|---|---|
| **REPTree** | 99.0975% | 0.0117 | 3.233% | 0.0768 |
| **RandomTree** | 98.7365% | 0.0138 | 3.8262% | 0.0994 |
| **J48** | 99.0975% | 0.0117 | 3.233% | 0.0768 |

Fuente: (Elaboracion Propia)

En la **Tabla 39** se puede observas los resultados obtenidos del entrenamiento de los algoritmos REPTree, RandomTree y J48, se enmarco los algoritmos REPTree y J48 los cuales se consideran factibles para el Modelado por las instancias clasificadas correctamente que tienen un valor alto y el Error Absoluto relativo que tienen un valor mínimo a comparación del Algoritmo RandomTree.

Una vez realizados estos pasos, se guardan los modelos creados con el fin de trabajar posteriormente en la etapa de implementación desarrollado que nos permita alcanzar los objetivos planteados.

- **Evaluación de los resultados obtenidos del índice de crecimiento del cáncer de mama de las mujeres, de las gestiones 2019-2023, la predicción realizada por un intervalo de 5 años con el Algoritmo M5P**.

```
Number of Rules: 1

=== Future predictions from end of training data ===
Time    Total_Casos_20_a_69 Casos_20_a_40
2018              272          70
2019*             323          38
2020*             326          35
2021*             319          31
2022*             313          28
2023*             306          25
```

Fuente: (Elaboracion Propia)

Se puede observar los resultados de la predicción obtenidos con el algoritmo M5P del total de casos de edades entre 20 a 69 años de diferentes gestiones, en la gestión 2019 se observa que va en aumento los casos a 323, 2020 también va aumentando de apoco a 326 casos, 2021 los casos van reduciendo 319, 2022 los casos reducen a 313 y 2023 va reduciendo los casos a 306.

En cuanto a las edades de 20 a 40 años de las mujeres de la ciudad de La Paz, en 2019 va aumentando de apoco a 38 casos, 2020 va reduciendo a 35, 2021 los casos bajan a 31 casos, 2022 los casos van bajando a 28 y para el 2023 el caso bajara razonablemente hasta 25 casos.

- **Evaluación de los resultados obtenidos del índice de crecimiento del cáncer de mama de las mujeres de las gestiones 2019-2023, la predicción realizada por un intervalo de 5 años con el Algoritmo RandonTree.**

```
=== Future predictions from end of training data ===
Time    Total_Casos_20_a_69 Casos_20_a_40
2018            272             70
2019*           380             76
2020*           258             71
2021*           348             79
2022*           272             67
2023*           343             60
```

Fuente: (Elaboracion Propia)

Se puede observar la predicción del total de casos de edades ente 20 a 69 años de diferentes gestiones, en la gestión 2019 se observa que va en aumento los casos a 380, 2020 el número de caso es de 258, 2021 el número de caso es de 348, 2022 el número de caso es de a 272 y 2023 va en aumento a 343.

En cuanto a las edades de 20 a 40 años en 2019 va aumentando a 76 casos, 2020 va reduciendo a 71, 2021 los casos aumentan a 79, 2022 los casos van bajando a 67 y para el 2023 el caso bajara razonablemente a 60 casos.

- **Evaluación del índice de crecimiento del cáncer de mama de las gestiones 2009-2018 la predicción realizada por un intervalo de 5 años con el Algoritmo MultilayerPerceptron.**

```
=== Future predictions from end of training data ===
Time   Total_Casos_20_a_69 Casos_20_a_40
2018            272             70
2019*           410             76
2020*           269             67
2021*           384             84
2022*           280             74
2023*           323             81
```

Fuente: (Elaboracion Propia)

Se puede observar los resultados obtenido la predicción del total de casos de edades entre 20 a 69 años de diferentes gestiones, en la gestión 2019 se observa que va aumentando el caso a 410, 2020 también va reduciendo a 269 casos, 2021 los casos va en aumento a 384 casos, 2022 los casos reducen a 280 y 2023 va en aumento hasta 323 casos.

En cuanto a las edades de 20 a 40 años en 2019 va aumentando a 76 casos, 2020 va reduciendo a 67 casos, 2021 los casos aumentan a 84 casos, 2022 los casos son de 74 casos y para el 2023 el caso aumenta a 81 casos.

### *3.1.6 Implantación (Fase VI)*

#### 3.1.6.1 Aplicación de la Metodología ASD (Desarrollo de Software Adaptativo)

Para el desarrollo del Modelo Predictivo del índice de crecimiento del cáncer de mama en mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en minería de datos se aplicara las fases de la metodología ASD (Desarrollo de Software Adaptativo).

**1) Fase de Especulación**

Esta es la fase inicial del ciclo de vida ASD, que comienza con la Especulación esta consiste en el inicio del proyecto y la planificación de las características del mismo, para

este fin sin duda el primer paso es la captura de requisitos, para posteriormente realizar una planificación adecuada del proyecto.

**Iniciación del proyecto**

La finalidad del presente trabajo es desarrollar un Modelo Predictivo donde se puede obtener el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años.

Para la elaboración del Modelo predictivo se presentan los siguientes requerimientos de software.

**Tabla 40**

*Requerimientos mínimos de software para el Modelo Predictivo*

| Requerimiento | Software | Versión |
|---|---|---|
| Sistema Operativo | Windows | 10 |
| Lenguaje de Programación | Java | 8u111 |
| Entorno de Desarrollo Integrado | Netbeans IDE | 8.2 |
| Interfaz grafica | Weka | 3.8 |
| Editor de Texto | SublimeText | 3 |
| Herramienta UML | StarUML | 3.2.2 |

Fuente: (Elaboracion Propia)

**Planeación de los ciclos**

Se describe los requerimientos funcionales del Modelo de Predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años en la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años.

**Tabla 41**

*Requerimientos funcionales del Modelo Predictivo*

| N° | FUNCIÓN |
|---|---|
| R1 | El Modelo Predictivo realizara la predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos. |
| R2 | Realiza el entrenamiento del Modelo Predictivo del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años, basado en Minería de Datos |
| R3 | El Modelo Predictivo realiza el entrenamiento de un algoritmo seleccionado |
| R4 | Se obtendrá el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años con un intervalo de 5 años |
| R5 | El Modelo Predictivo generara los resultados del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años, basado en Minería de Datos por un intervalo de 5 años |
| R6 | El Modelo predictivo obtiene resultados mediante grafica (Árbol de Decisión) |

Fuente: (Elaboracion Propia)

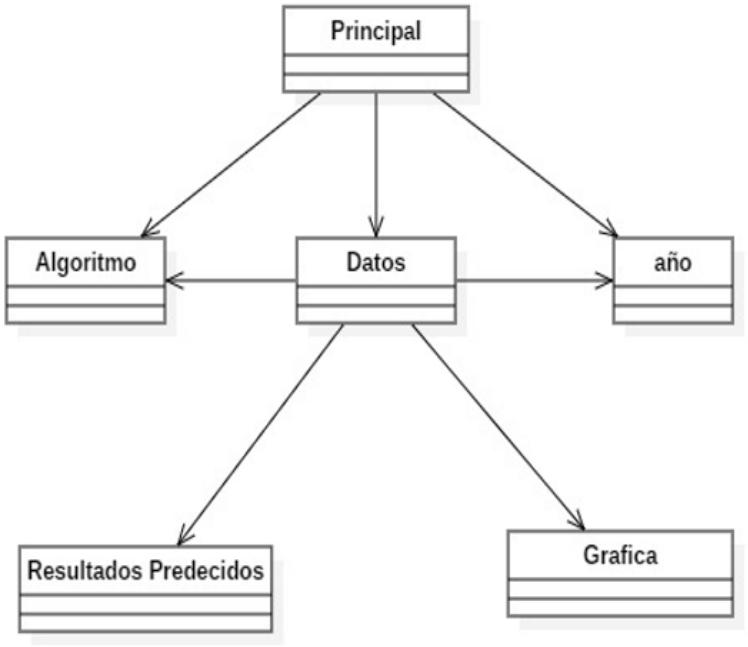
## 2) Colaborar

Se define las características del software y el entorno de trabajo para generar las historias de usuario y los ciclos repetitivos.

**R1.** El Modelo Predictivo realizara la predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos.

**Figura 58**

*Diagrama de Clases del Modelo Predictivo*

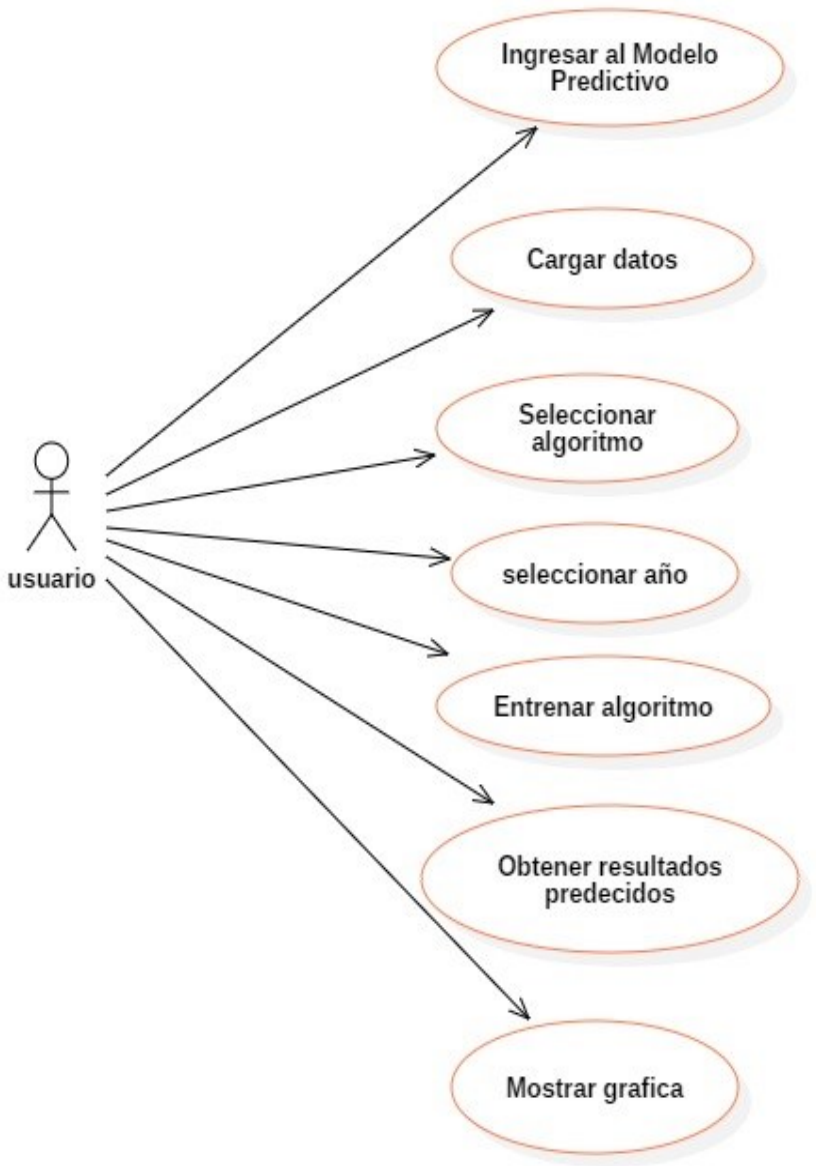



Fuente: (Elaboracion Propia)

**R2.** Realiza el entrenamiento del Modelo Predictivo del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años, basado en Minería de Datos.

**Figura 59**

*Diagrama de casos de uso general para el entrenamiento del Modelo Predictivo, basado en Minería de Datos*

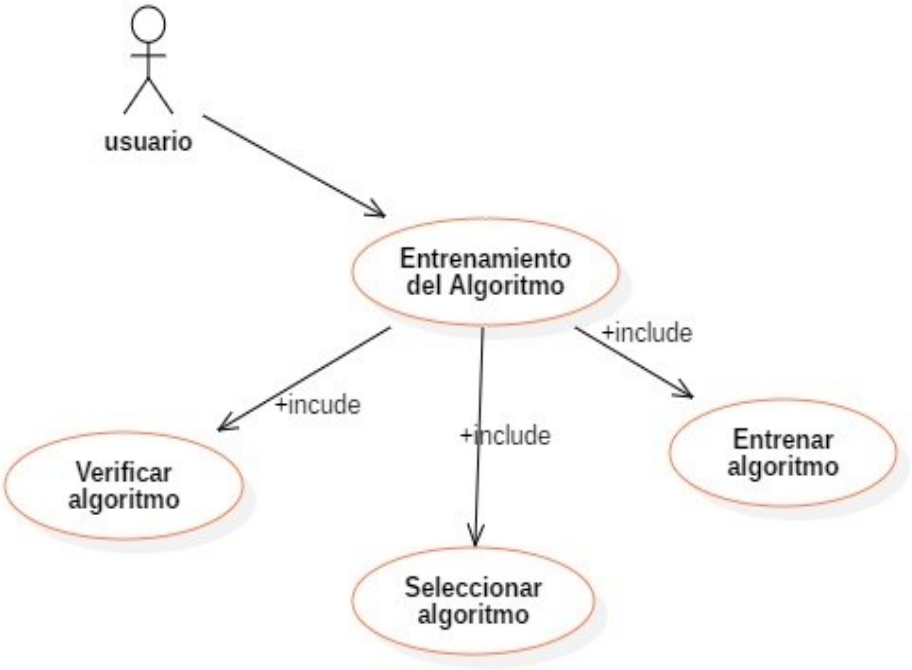



Fuente: (Elaboracion Propia)

**R3.** El Modelo Predictivo realiza el entrenamiento de un algoritmo seleccionado

**Figura 60**

*Diagrama de caso de uso para el entrenamiento del algoritmo de Minería de Datos*

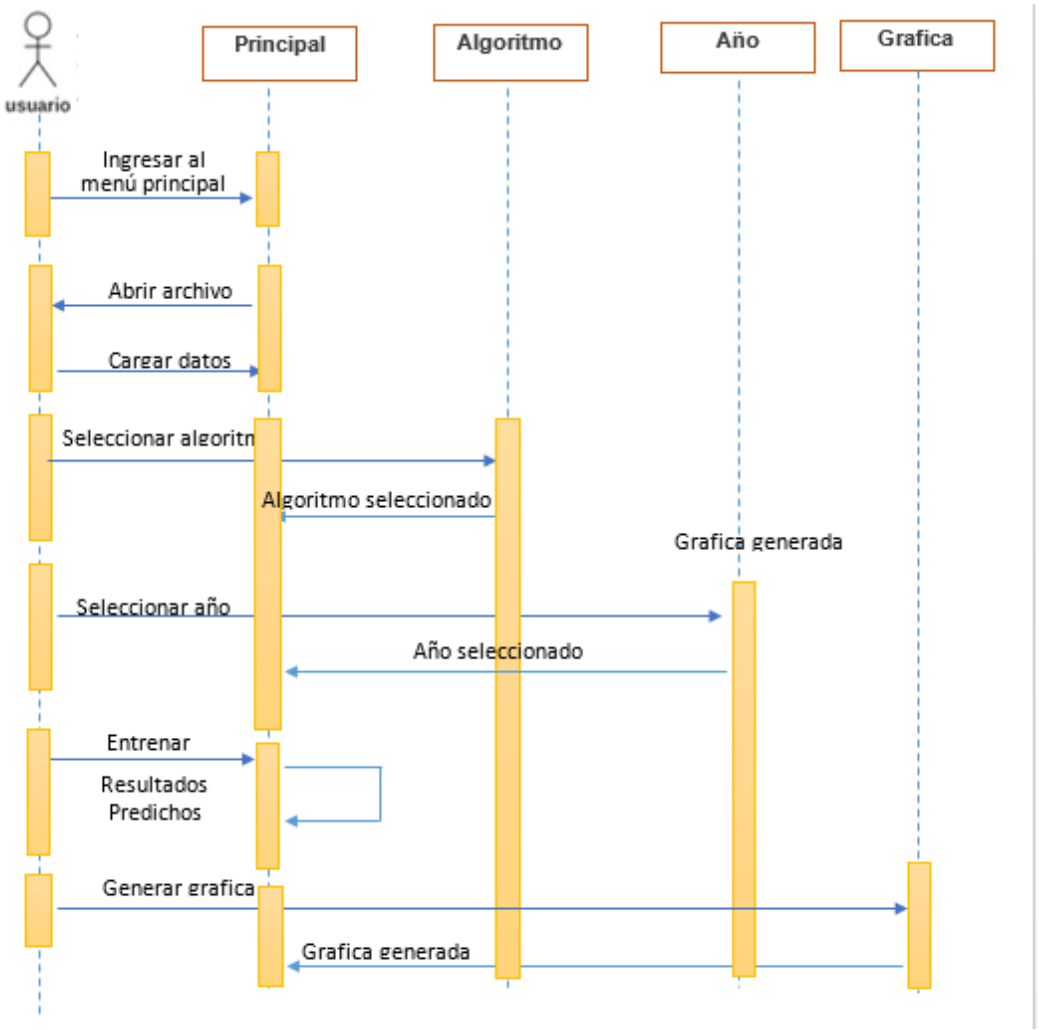



Fuente: (Elaboracion Propia)

**R4.** Se obtendrá el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años con un intervalo de 5 años.

**Figura 61**

*Diagrama de secuencia del Modelo de Predictivo*



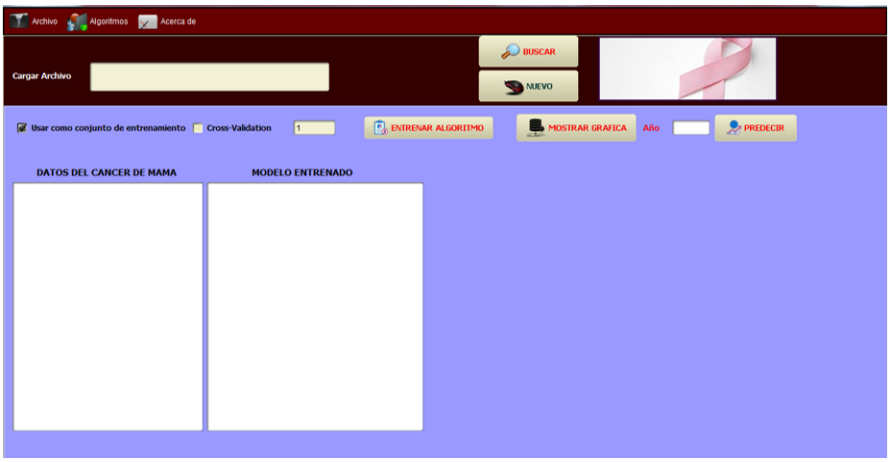

*Fuente: (Elaboracion Propia)*

### 3) Aprender

En esta fase se verifica la calidad del resultado del Modelo Predictivo que realizo con las iteraciones R1, R2, R3, R4 se procede a realizar el Menú Principal del Modelo Predictivo.

**Figura 62**

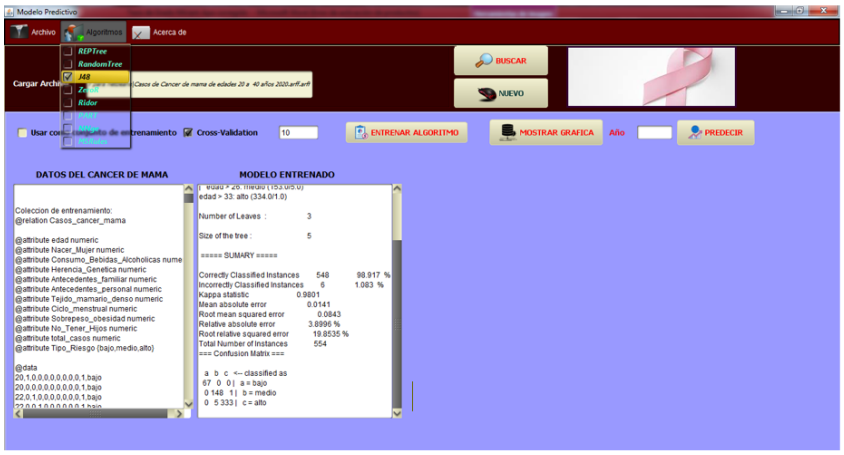
*Pantalla Principal del Modelo Predictivo del cáncer de mama*




Fuente: (Elaboracion Propia)

**Figura 63**

*Resultados obtenidos con el Algoritmos J48*

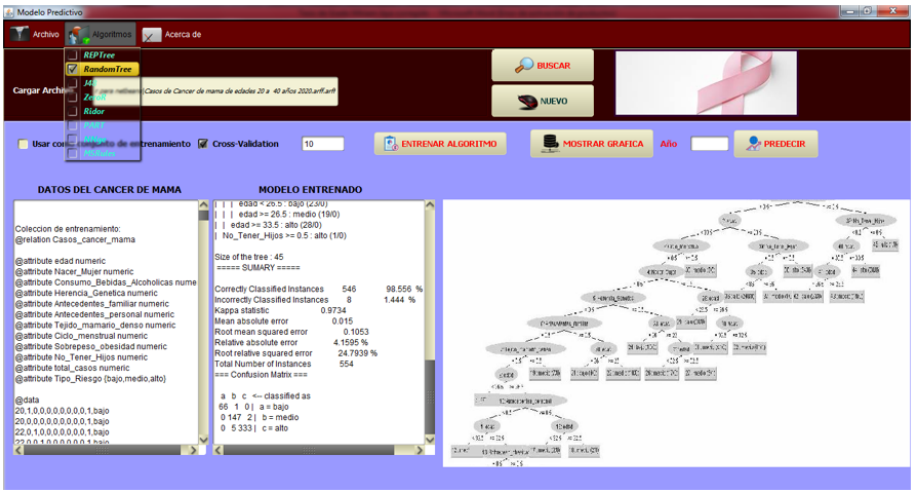



Fuente: (Elaboracion Propia)

**R5.** El Modelo Predictivo generara los resultados del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años, basado en Minería de Datos por un intervalo de 5 años.

**Figura 64**

*Resultados obtenidos con el Algoritmos RandomTree*

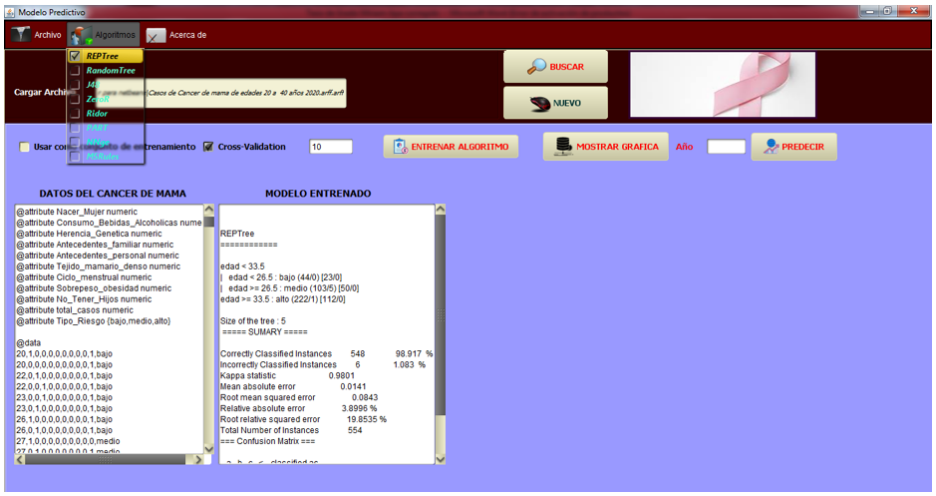



Fuente: (Elaboracion Propia)

**R6.** El Modelo predictivo obtiene resultados mediante grafica (Árbol de Decisión)

**Figura 65**

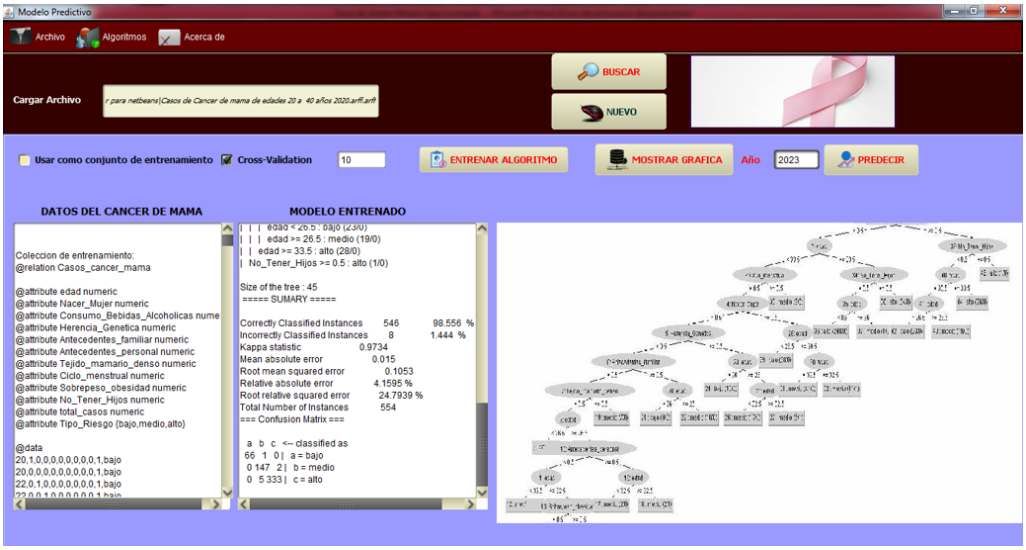
*Resultados obtenidos con el Algoritmos REPTree*




Fuente: (Elaboracion Propia)

**Figura 66**

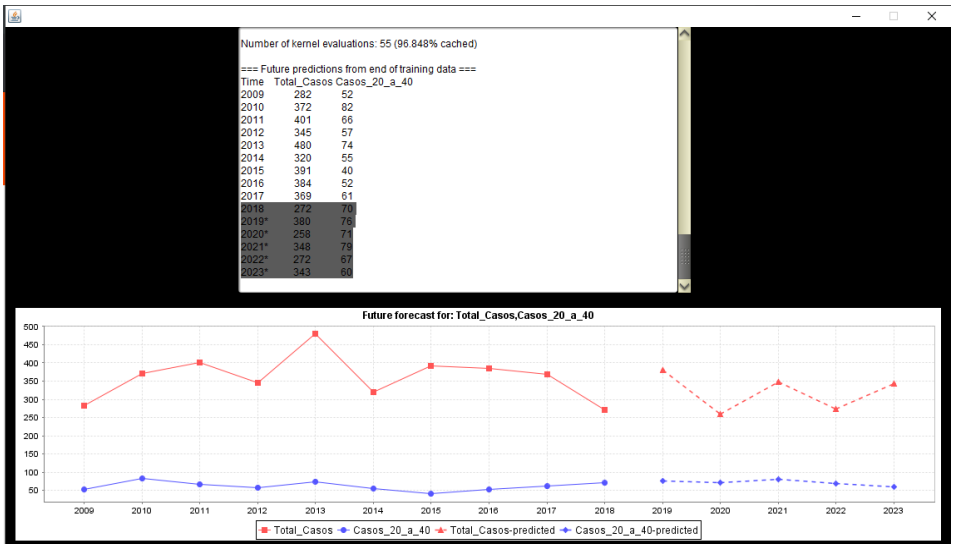
*Entrenamiento del Modelo Predictivo mediante el factor de riesgo del cáncer de mama para la toma de decisiones*




Fuente: (Elaboracion Propia)

***Figura 67***

Resultados obtenidos con el algoritmo RandomTree del índice de crecimiento del cáncer de mama de mujeres, basado en datos históricos por un intervalo de 5 años




Fuente: (Elaboracion Propia)

- **Prueba de caja negra**

**Tabla 42**

*Aplicando la prueba de caja negra*

| NRO | PREGUNTAS | 1<br>25% | 2<br>50% | 3<br>75% | 4<br>100% |
|---|---|---|---|---|---|
| 1 | El entrenamiento del Modelo Predictivo es inmediato | | | | X |
| 2 | Se carga los datos a entrenar de manera inmediata | | | | X |
| 3 | Se visualiza los resultados obtenidos de manera efectiva | | | | X |
| 4 | Se puede entrenar el modelo con otro algoritmo | | | X | |
| 5 | Se puede visualizar mediante un gráfico los resultados del entrenamiento del modelo | | | | X |
| | **TOTAL** | | | **150** | **300** |

Fuente: (Elaboracion Propia)

Se tiene:

R= 100+100+100+75+100/5= 475/5=**95%**

Es la aceptación del usuario, entonces cumple las expectativas del usuario.

## 3.2 Métrica de Calidad del Software ISO/IEC 9126

El modelo de calidad establecido en la primera parte del estándar, ISO 9126, clasifica la calidad del software en un conjunto estructurado de características y sub características de la siguiente manera:

**Tabla 43**

*Aplicación de la Métrica de Calidad Externa e interna ISO/IEC 9126-1*

| Características | Pregunta central | Sub-Características | Medida | Total % |
|---|---|---|---|---|
| **Funcionabilidad** | ¿Se satisfacen las funciones implícitas y explicitas requeridas en el software? | Idoneidad | 0,95 | 91% |
| | | Exactitud | 0,90 | |
| | | Interoperabilidad | 0,90 | |
| | | Seguridad | 0,90 | |
| | | Cumplimiento de normas. | 0,90 | |
| **Fiabilidad** | ¿Puede mantener el nivel de rendimiento bajo ciertas condiciones y por cierto tiempo? | Madurez | 0,95 | 95% |
| | | Recuperabilidad | 0,95 | |
| | | Tolerancia a fallos | 0,95 | |
| **Usabilidad** | ¿El software es fácil de usar y aprender? | Aprendizaje | 0,98 | 95% |
| | | Comprensión | 0,95 | |
| | | Operatividad | 0,95 | |
| | | Atractividad | 0,90 | |
| **Eficiencia** | ¿Es rápido y minimalista en cuanto al uso de servicio? | Comportamiento en el tiempo | 0,80 | 85% |
| | | Comportamiento de recursos | 0,90 | |
| **Mantenibilidad** | ¿Es fácil de modificar y verificar? | Estabilidad | 0,95 | 91% |
| | | Facilidad de análisis | 0,90 | |
| | | Facilidad de cambio | 0,90 | |
| | | Facilidad de pruebas | 0,90 | |
| **Portabilidad** | ¿Es fácil de transferir de un ambiente a otro? | Capacidad de instalación | 0,95 | 91% |
| | | Capacidad de reemplazamiento | 0,95 | |
| | | Adaptabilidad | 0,80 | |
| | | Co-Existencia | 0,95 | |
| | | **Evaluación Global - Promedio** | **0,92** | **91%** |

Fuente: (Elaboracion Propia)

Se tiene la calidad global del Modelo de Minería de Datos, que es del 91%, lo que quiere decir que de cada 100 personas que usen el Modelo Predictivo de Minería de Datos, 91 quedarán conformes y satisfechas con él.

Dado que el porcentaje de Calidad Global se encuentra entre 60 y 100%, se puede decir que el Modelo Predictivo de Minería de Datos tiene un nivel de aceptación satisfactorio.

## 3.3 Evaluación de Costos y Beneficios

### *3.3.1 Método Cocomo II*

A continuación se describe la evaluación de los costos y beneficios para el modelo, donde se aplica el método COCOMO II.

Aplicando las ecuaciones

$E = a\ (Kl)^b * (X)$, en personas - mes

$Tdev = c\ (E)^d$, en meses

$P = E/Tdev$, en personas

E = es el esfuerzo requerido por el proyecto en persona – mes

Tdev = es el tiempo requerido por el proyecto, en meses

P = es el número de personas requeridas por el proyecto

a, b, c, y d = son constantes con valores definidos en una tabla según cada sub-modelo

Kl = es la cantidad de líneas de código, en mil

M(X) = es un multiplicador que depende de atributos

**Se tiene un total de líneas de código Klcd = 1300**

Así:

Kl = Klcd/1000

Kl = 1000/1000 = 1500 Kl

En este caso es el tipo orgánico será más apropiado ya que el número de líneas de código no supera los 50 Kl

Se utiliza la siguiente tabla para poder obtener una primera aproximación rápida del esfuerzo, y hace uso de la siguiente tabla de constantes para calcular distintos aspectos de costes.

**Tabla 44**

*Constante para calcular aspectos de costes*

| MODO | a | b | c | d |
|---|---|---|---|---|
| **Orgánico** | **3.2** | **1.05** | **2.50** | **0.38** |
| Semilibre | 3.00 | 1.12 | 2.50 | 0.35 |
| Rígido | 3.60 | 1.20 | 2.50 | 0.32 |

Fuente: (Elaboración en base a Cocomo II)

**Calculando el esfuerzo del desarrollo se aplica el modo orgánico**

$E = a\ (Kl)^b * (X)$, en personas - mes

E = 3.2 * (1300)^1.05 * 0.549

E = 3.2 * 1,86 * 0.549

E = 3.26 = 3 personas

**Calculando el tiempo de desarrollo**

$Tdev = c(E)^d$, en meses

T = 2.5 * 3.26 ^ 0.38

$T = 2.5 * 1.56$

$T = 3.91 = 3 \text{ a } 4 \text{ meses}$

**Estimando los costos**

Se realiza un supuesto en cuanto a la estimación del costo del Modelo Predictivo

La hora por trabajar = 4,37$

Día =4,37$* 4 horas = 17,48$ /día

Semana = 17,48$ *5 días= 87,4$

Al mes 87,4$ * 4 semanas = 349.6$* 3 personas = 1040$/mes pagadas a 3 personas

Calculamos 349.6$ * 3 meses *3 personas = 3146$, el costo del producto elaborado en 3 meses.

Calculamos el costo del producto en Bs. (3146$* 6,5 Bs.)/1$ = 20449 Bs.

# Capitulo IV: Prueba de Hipótesis

**Resumen**

Para el presente capítulo se realiza el cálculo de la prueba de Hipótesis planteada para el Modelo Predicción del índice de crecimiento del cáncer de mama en la mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años, donde se utiliza la hipótesis descriptiva, con el propósito de hallar el valor de la eficiencia del Modelo Predictivo.

## 4.1 Formulación de la Hipótesis

El Modelo Predictivo con la ayuda de Minería de Datos demuestra que existe un alto índice de crecimiento del cáncer de mama en mujeres de edades entre 20 a 40 años en la ciudad de La Paz con una eficiencia de 95% por un intervalo de 5 años.

## 4.2 Estado de la Hipótesis

Para la demostración de la hipótesis planteada se utilizó la Estadística Descriptiva.

Formular la hipótesis nula: $H_o$

Formular la Hipótesis alternativa: $H_1$, dependiendo el contexto o contenido del problema de estudio.

## 4.3 Calculo de la Hipótesis

Para demostrar la hipótesis planteada, se realiza las pruebas en base a los datos recolectados sobre el índice de crecimiento del cáncer mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, tomadas mediante muestras.

### *4.3.1 Hipótesis:*

Hi: El Modelo de Predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años permite predecir con un nivel de confianza del 95%

### *4.3.2 Hipótesis Nula:*

Ho: El Modelo de Predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años no permite predecir con un nivel de confianza del 95%

### *4.3.3 Hipótesis Alternativa:*

$H_1$: El Modelo de Predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años permite predecir el nivel de confianza del 95%.

Se realizó 200 pruebas para la primera muestra del Modelo Predictivo, 180 pruebas fueron correctas y 20 pruebas fueron incorrectas.

N=200
X=190
P = ?

$$P = \frac{X}{N}$$

$$P = \frac{190}{200} \rightarrow P = 0.95 * 100 = 95\%$$

$$(1 - P) \approx (1 - 0.95) = 0.05 * 100\% = 5\%$$

$H_0 = P \neq 95\%$

$H_i = P = 95\%$

Se realizó 150 pruebas para la primera muestra del Modelo Predictivo, 180 pruebas fueron correctas y 100 pruebas fueron incorrectas.

N = 100

X = 90

**p** =?

$$\mathbf{p} = \frac{X}{N}$$

$$\mathbf{p} = \frac{90}{100} \quad \rightarrow \quad \mathbf{p} = 0.9 * 100 = 90\%$$

$$(1 - p) = (1 - 0.9) = 0.1 \approx 0, 1 * 100 = 10\%$$

**Aplicación de la Distribución Estándar Normal**

Se formula de la siguiente manera:

$$Z = \frac{(\mathbf{p} - P)}{\sqrt{\frac{P\,(1-P)}{n}}}$$

$$Z = \frac{(0.90 - 0.95)}{\sqrt{\frac{0.95\,(1-0.95)}{100}}}$$

$$Z = \frac{(-\,0.05)}{\sqrt{\frac{0.95\,(0.05)}{100}}}$$

$$Z = -\,\frac{0.05}{\sqrt{(0.0004)}}$$

$$Z = -\,\frac{0.05}{0.021}$$

$$Z = -\,2.38$$

Hallando el Intervalo del Nivel de Confianza

$$\alpha = 5\% \approx 95\%$$

$$z_{\left(1 - \frac{\alpha}{2}\right)}$$

$$z_{\left(1 - \frac{0.05}{2}\right)}$$

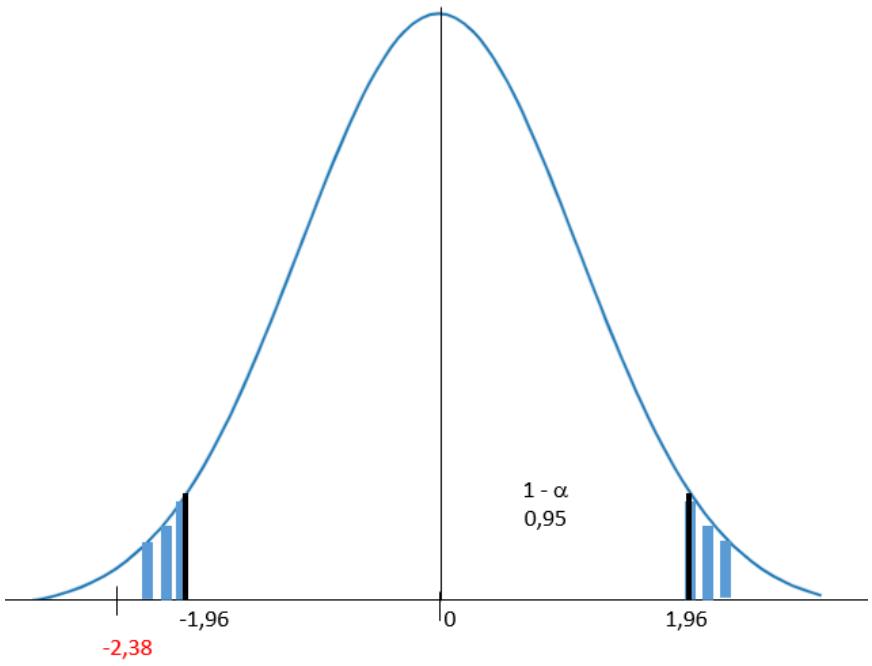
$$z_{(1 - 0.0025)} \Longrightarrow z_{0.975} \Longrightarrow z = 1.96$$

El intervalo de probabilidad comprendido entre -z y z, para 1.96 el intervalo es del 95% (0.95)

*Zona critica izquierda* *-1.96*
*Zona critica derecho* *1,96*

**Figura 68**

*Resultado de la Distribución normal*




Fuente: (Elaboracion Propia)

$$P\left[\mathbf{p} - z_{\alpha/2}\frac{\sqrt{P(1-P)}}{n}\frac{\sqrt{N-n}}{N-1} < P < \mathbf{p} + z_{\alpha/2}\frac{\sqrt{P(1-P)}}{n}\frac{\sqrt{N-n}}{N-1}\right] = 1 - \alpha$$

$$P\left[0.90 - (1.96)\,(0.021) < P < 0.90 + (1.96)\,(0.021)\right] = 95\%$$

$$P\left[0.90 - 0.04 < P < 0.90 + 0.04\right] = 95\%$$

$$P\left[0.86 < P < 0,\ 94\right] = 95\%$$

Se obtiene como resultado el cálculo del nivel de confianza de Z = - 2.38 por lo que cae en la zona de rechazo, por lo que se rechaza la Hipótesis Nula $H_0$, y se acepta la Hipótesis Alternativa $H_1$, con un intervalo de confianza de 95 %.

# Capítulo V: Conclusiones y Recomendaciones

**Resumen**

Para el presente capitulo se describen las conclusiones a las que se llegó con el presente trabajo de investigación, alcanzando los objetivos planteados, además se describen las recomendaciones necesarias para una futura investigación sobre el trabajo.

## 5.1 Conclusiones

En el presente Trabajo de Investigación: "Modelo de Predicción del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos, el cual se llegó a las siguientes conclusiones mediante los objetivos planteados en el capítulo I.

- Se investigó y analizo los datos que se recopilaron de las diferentes instituciones de salud sobre el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, con el propósito de prevenir futuras muertes.
- Se diseñó un Modelo de Predicción aplicando Algoritmos de Minería de Datos.
- Para el desarrollo del Modelo Predictivo se aplicaron técnicas de series de Tiempo y Arboles de Decisión para visualizar el comportamiento del factor de riesgo del y el índice de crecimiento del cáncer mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz.

- Para el desarrollo del Modelo Predictivo se aplicó la metodología ágil ASD(Software de Desarrollo Adaptable) y además se utilizaron herramientas libres como ser: WEKA, Java, Netbeans 8.2
    - Mediante la aplicación de los Algoritmos de Minería de Datos REPTree, RandonTree y J48, se obtuvo como principal factor de riesgo (no tener hijos) de edades >=33 años, además el tipo de riesgo en padecer el cáncer de mama es Alto.
- Para el entrenamiento del Modelo predictivo se aplicaron las metodologías CRISP-DM y la metodología de Proceso KDD para el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, basado en Minería de Datos por un intervalo de 5 años.

## 5.2 Recomendaciones

A partir del estudio realizado sobre el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz, se plantean las siguientes recomendaciones para futuras investigaciones relacionadas al presente trabajo de investigación.

- Para futuras investigaciones sobre el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La paz, se recomienda contar con una base de datos completa de las diferentes instituciones de salud que cuenta la ciudad de La Paz, para obtener una mejor precisión en cuanto a la predicción.
- Se recomienda utilizar otros métodos de Predicción, para la comparación de resultados del índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz.
- Se recomienda estudiar otras alternativas de técnicas de Modelado, con el propósito de tener otras opciones para la toma de decisiones sobre el índice de crecimiento del cáncer de mama en las mujeres de edades entre 20 a 40 años de la ciudad de La Paz.
- Se recomienda estudiar a profundidad los algoritmos de Minería de Datos que no se utilizaron en el presente trabajo de investigación.

## Bibliografía


- Pardo Igor. (2015). Plan Nacional de Prevención, Control y Seguimiento del cáncer de mama. La Paz - Bolivia: Ministerio de Salud y Deportes.
- Velasco Mallea, Alison Fabiola; Clavijo Cárdenas, Edgar Palmiro; Huanca Quisbert, Carmen Rosa. (2016). Control e Información para la Detección temprana del Cáncer de Mama mediante Tecnología Móvil. La Paz -Bolivia
- Ríos Medrano, Vania Marcela; Bejarano Carvajal Sergio. (2016). Desgaste emocional como causa de la aceleración del proceso de la enfermedad en pacientes mujeres con cáncer de mama del Hospital de Clínicas de La Paz. La Paz –Bolivia.
- Propiedad de IBM. (2012). Manual KDD de IBM SPSS Modeler. En IBM (56). Estados Unidos: IBM SPSS Modeler
- Mamani Chávez, Sol María; Tancara Cuentas, Wilfredo Dacio. (2018). Situación de la mortalidad por cáncer en la mujer en los municipios La Paz-El Alto en el primer semestre del 2017. La Paz –Bolivia.
- Camacho Centellas Sandro Saúl. (2016). Método heurístico para el diagnóstico de cáncer de mama basado en minería de datos. La Paz -Bolivia
- Matta Catacora Patricia Evelyn (2011) “Sistema web para el control de ventas y facturación usando agentes inteligentes”, importadora de fármacos. La Paz - Bolivia
- Luna Gustavo. (2015). Plan Nacional de prevención, control y seguimiento del cáncer de mama. La Paz-Bolivia.

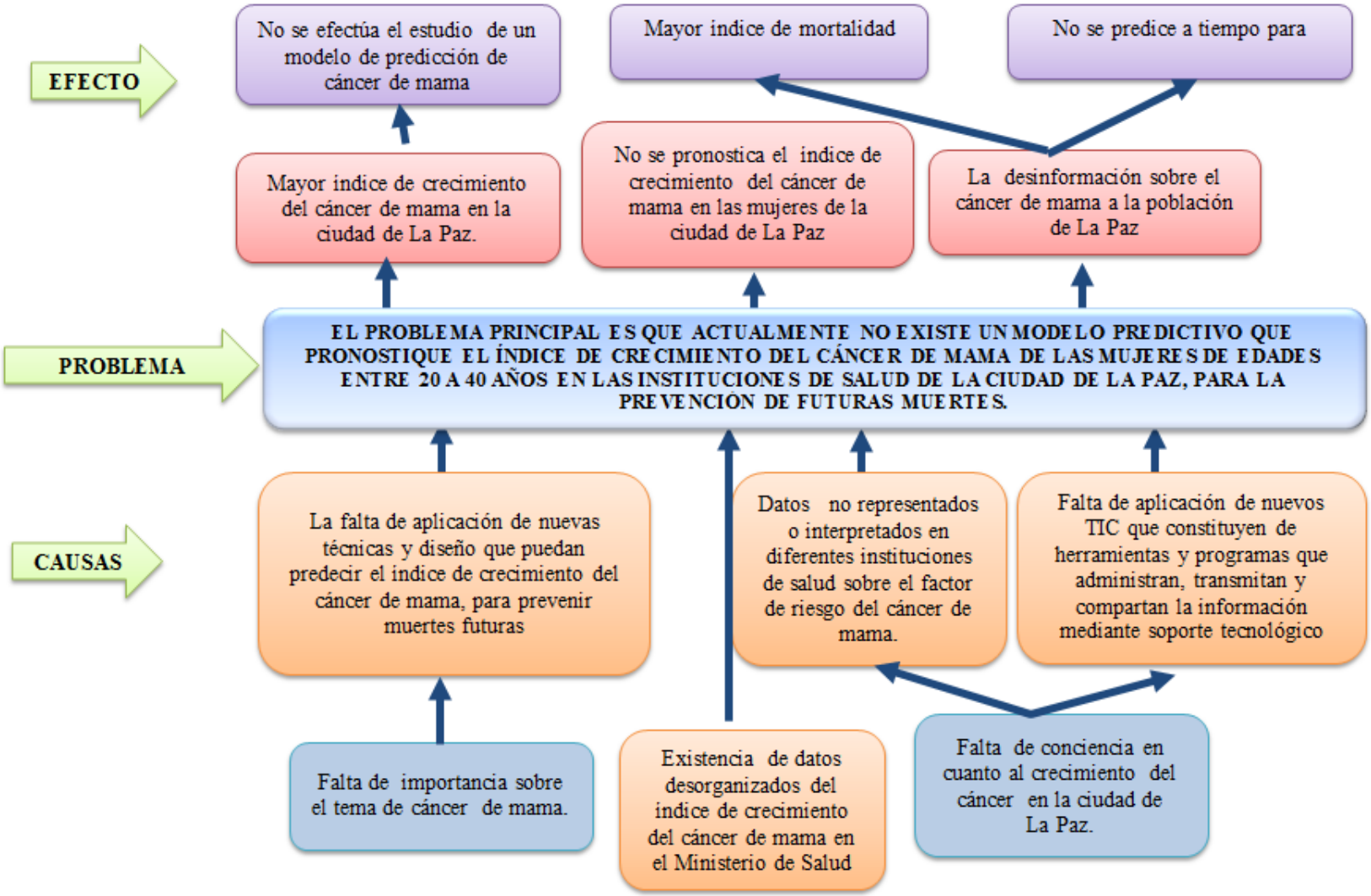
- Valle Calderón Freddy. (2016). Estudio de la Mortalidad materna. La Paz - Bolivia: Memorias del C.A.I.
- Ministerio de Salud del Estado Plurinacional de Bolivia. (2019). Programa Nacional de lucha contra el cáncer. En 5p. La Paz, Bolivia.
- Zamora Ezequiel. (2014). Desarrollo de Software Adaptable. Barinas Estado Barinas.
- Molina José M. y García Jesús. (2012). Esquema del proceso de KDD.
- Chapman Peter. (2000). CRISP-DM.
- García Morate Diego. (2000). Manual de WEKA. Nueva Zelanda.
- Gómez Jiménez Enrique y Moreno Núñez Jhonatan. (2019). Fundamentos de Programación java con netbeans 8.2. 520p. 2019: Ed. Alfa Omega.
- Molina José M. y García Jesús. (2012). Técnicas de Análisis de Datos.
- Hernández Orallo José. (2004). Introducción a la Minería de Datos.
- Martínez Álvarez Clemente Antonio. (2012). Aplicación de Técnicas de Minería de Datos para mejorar el Proceso de Control de Gestión en Entel.
- De la Gálvez M. Alberto, Tamayo C. Carlos, Calani L.Franz Perfil de Mortalidad en la ciudad de La Paz 2009.1ª ed. La Paz –Bolivia.
- Ministerio de Salud y Deportes, Instituto Nacional de Estadística. Encuesta PostCensal de Mortalidad Materna. La Paz: MSD/INE; abril 2004.

**Webgrafía:**

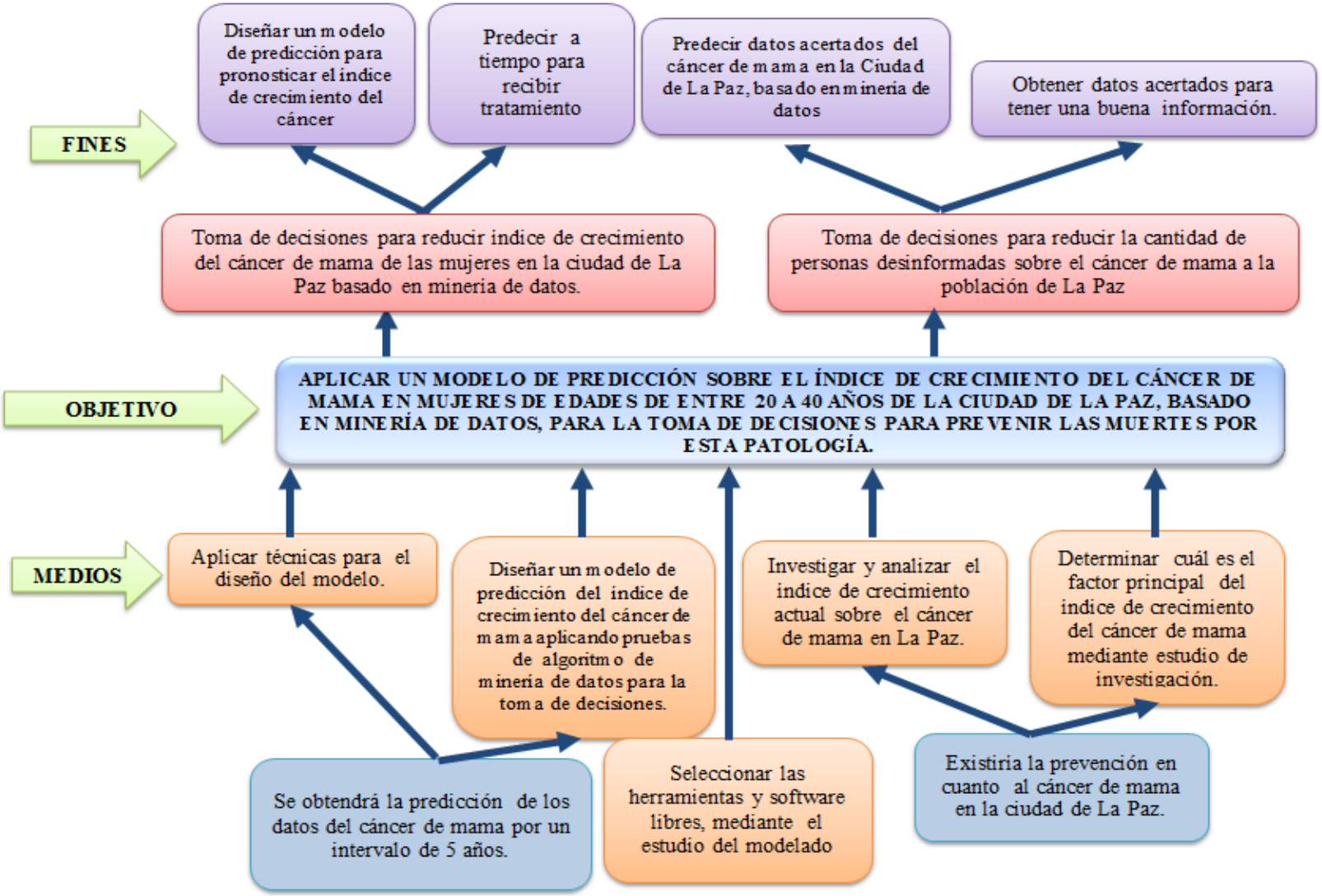

- El Cáncer de Mama en la ciudad de La Paz, https://www.cies.org.bo/aprende/cancer/cancer-de-mama.
- El crecimiento del cáncer es un problema esta enfermedad es un enemigo silencioso, https://www.fubolcancer.com/, 03 de junio de 2019
- El cáncer de mama es uno de los tipos de cáncer más comunes, https://www.cies.org.bo/aprende/cancer/cancer-de-mama, 03 de junio de 2019.
- Java (Lenguaje de Programación), https://es.wikipedia.org/wiki/Java_(lenguaje_de_programaci%C3%B3n), 03 de junio de 2019.
- Cross Industry Standard Process for Data Mining, https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining , 21 de octubre del 2015.
- Wikipedia. (2018). ISO/IEC 9126. 27 de agosto, de Wikipedia Sitio web: https://es.wikipedia.org/wiki/ISO/IEC_9126
- Método Científico, https://es.wikipedia.org/wiki/M%C3%A9todo_cient%C3%ADfico, 30 de octubre del 2015.
- Wikipedia. (2018). ISO/IEC 9126. 27 de agosto, de Wikipedia Sitio web: https://es.wikipedia.org/wiki/ISO/IEC_9126

# Anexos

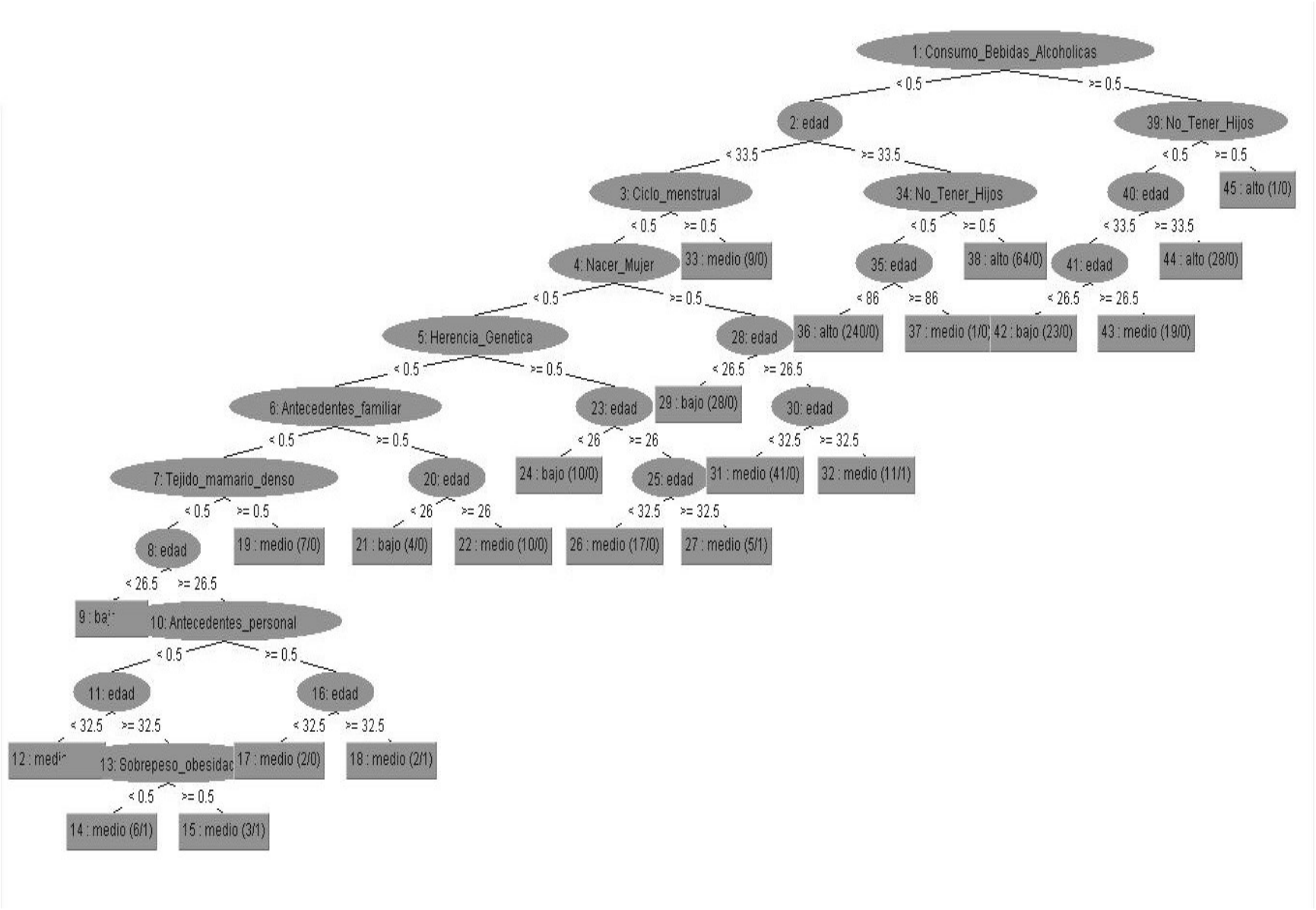
**A1. Árbol de Problemas.**



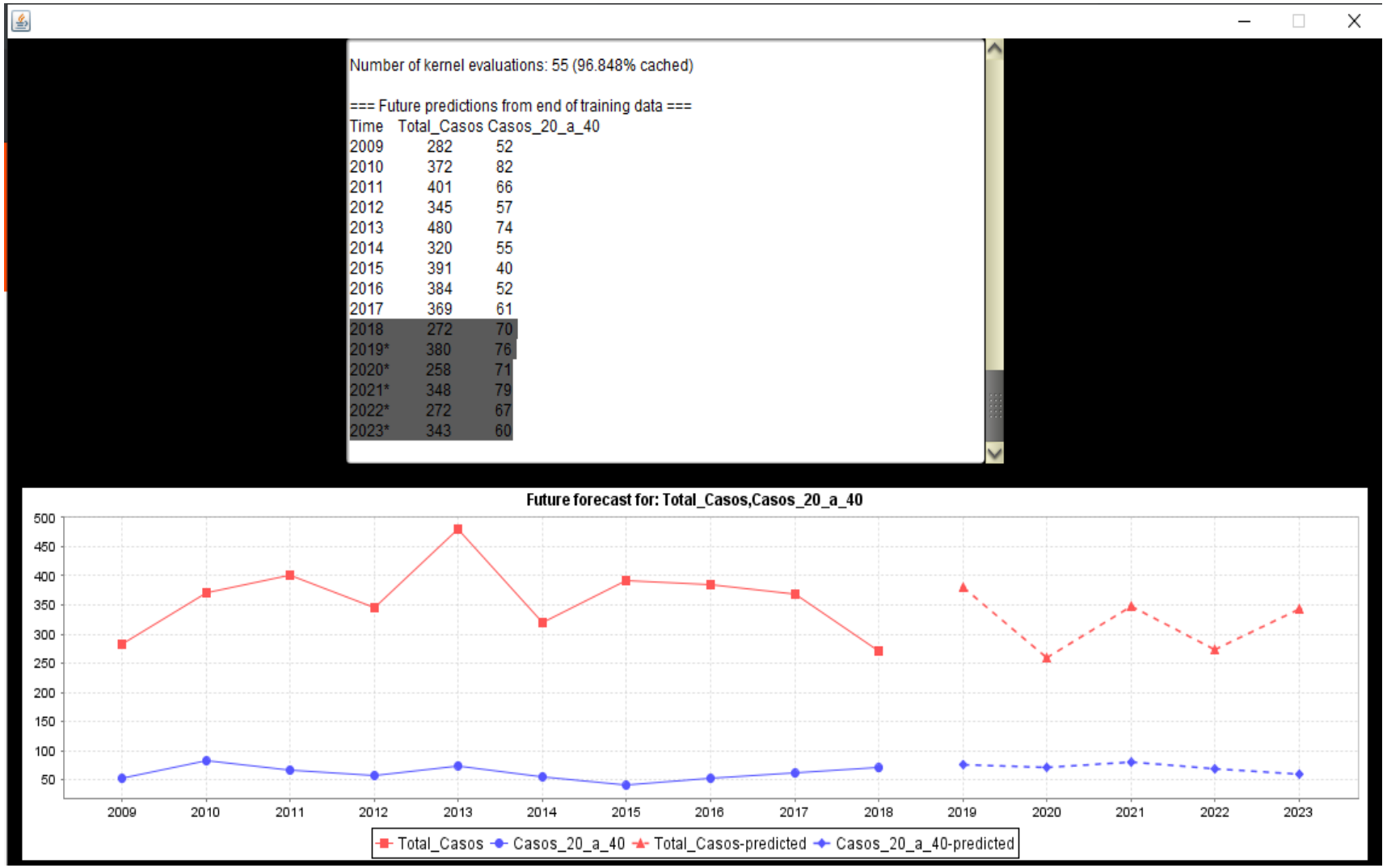
**A2. Árbol de Objetivos.**



**A3. Árbol de Decisión.**



**A4. Predicción del índice de crecimiento del cáncer de mama de mujeres de entre 20 a 40 años aplicado con el algoritmo RandomTree.**



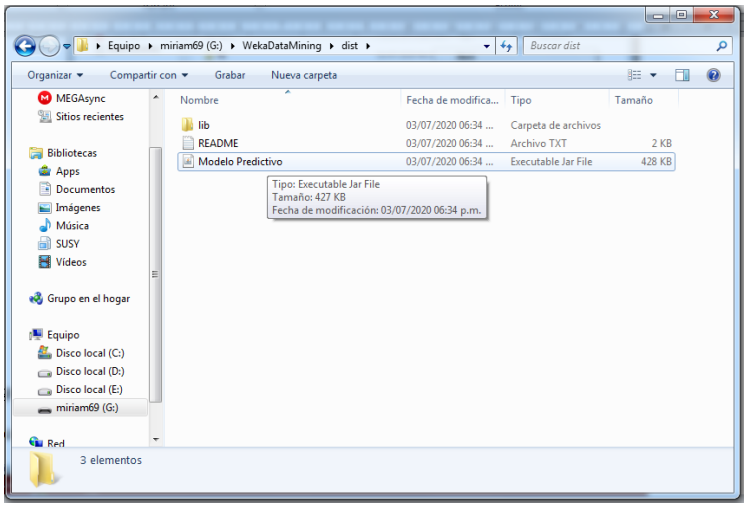
Number of kernel evaluations: 55 (96.848% cached)

=== Future predictions from end of training data ===

| Time | Total_Casos | Casos_20_a_40 |
|---|---|---|
| 2009 | 282 | 52 |
| 2010 | 372 | 82 |
| 2011 | 401 | 66 |
| 2012 | 345 | 57 |
| 2013 | 480 | 74 |
| 2014 | 320 | 55 |
| 2015 | 391 | 40 |
| 2016 | 384 | 52 |
| 2017 | 369 | 61 |
| 2018 | 272 | 70 |
| 2019* | 380 | 76 |
| 2020* | 258 | 71 |
| 2021* | 348 | 79 |
| 2022* | 272 | 67 |
| 2023* | 343 | 60 |

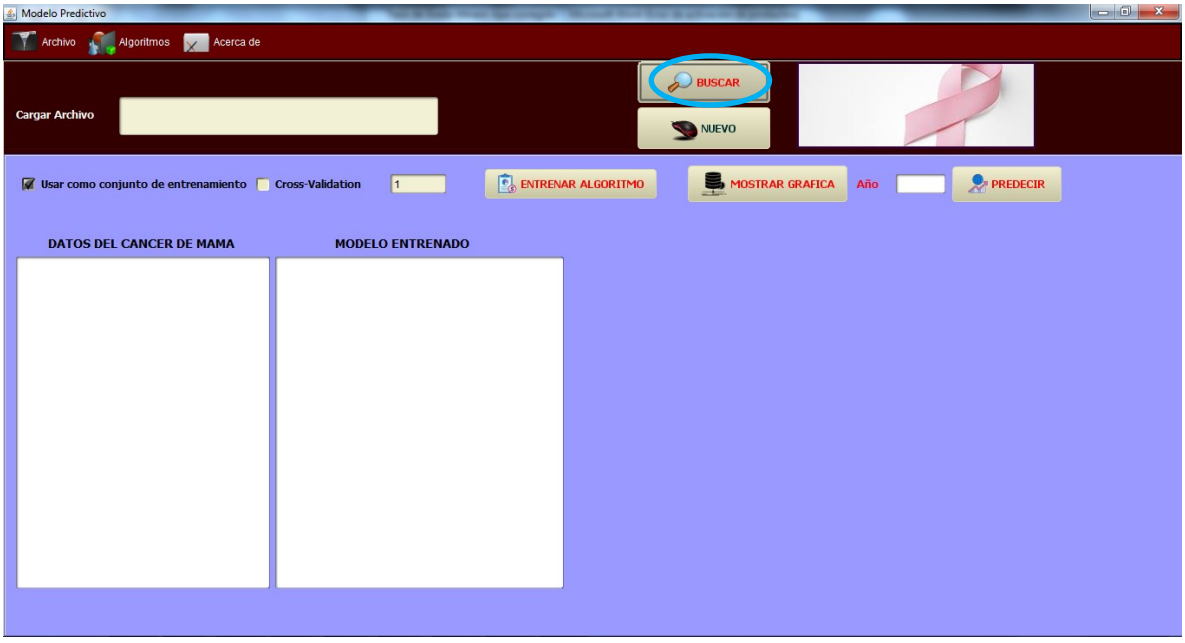
## A5. Manual de Usuario.

Primer paso abrir el software ejecutable (Modelo Predictivo)

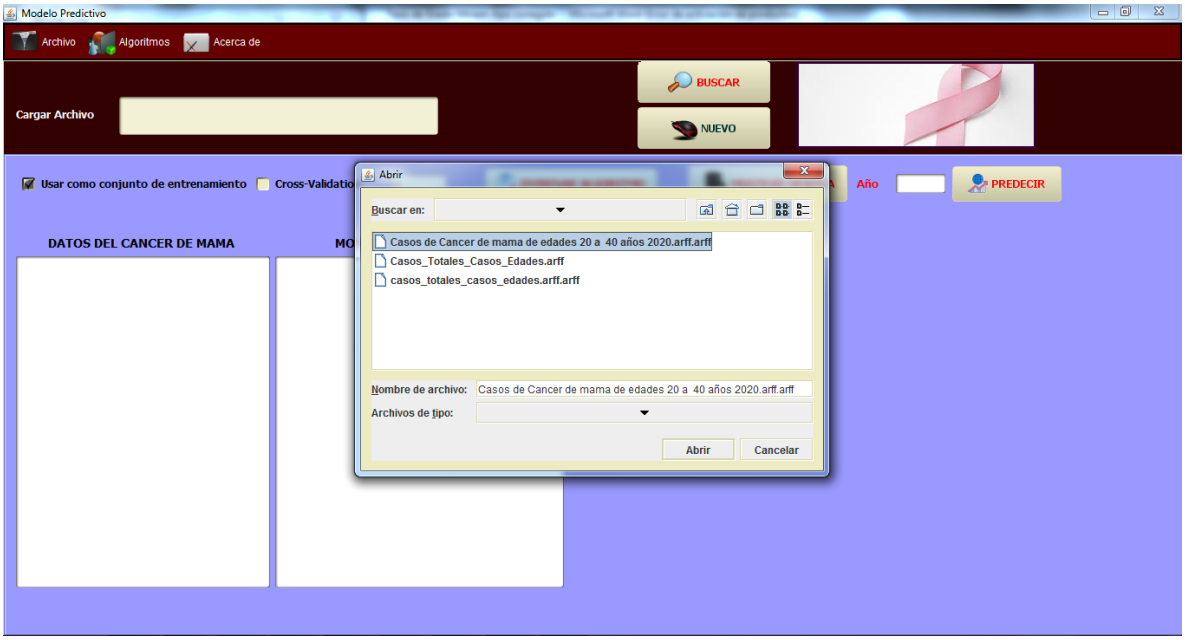



Paso 2: Para cargar el archivo (.arff)

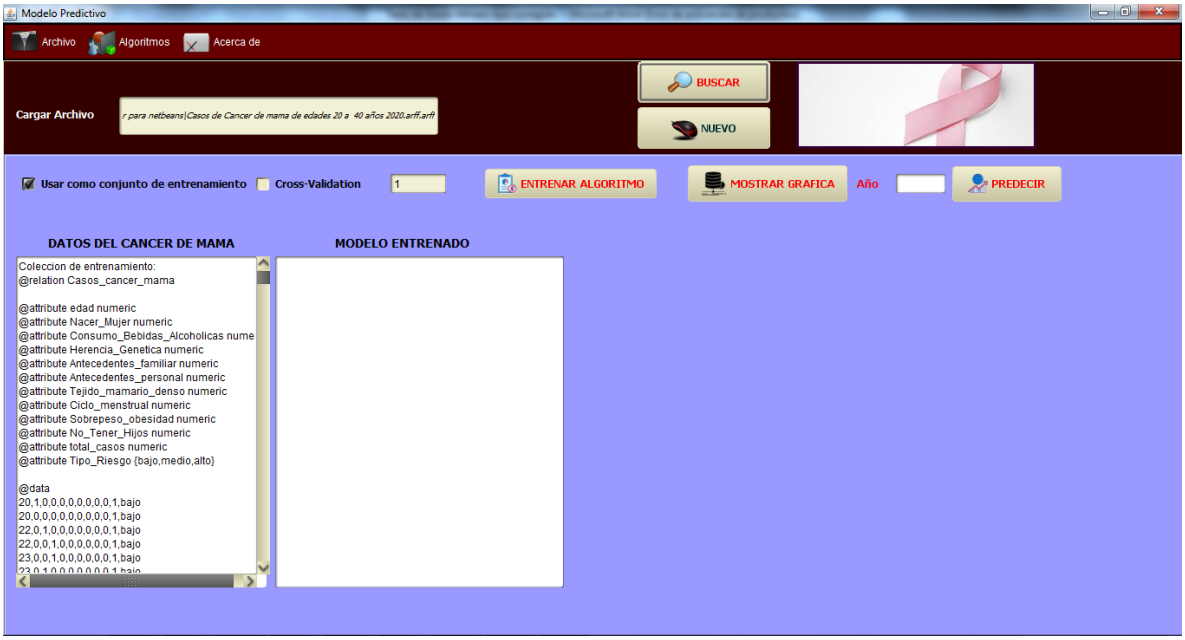
Hacer clic izquierdo en el botón Buscar



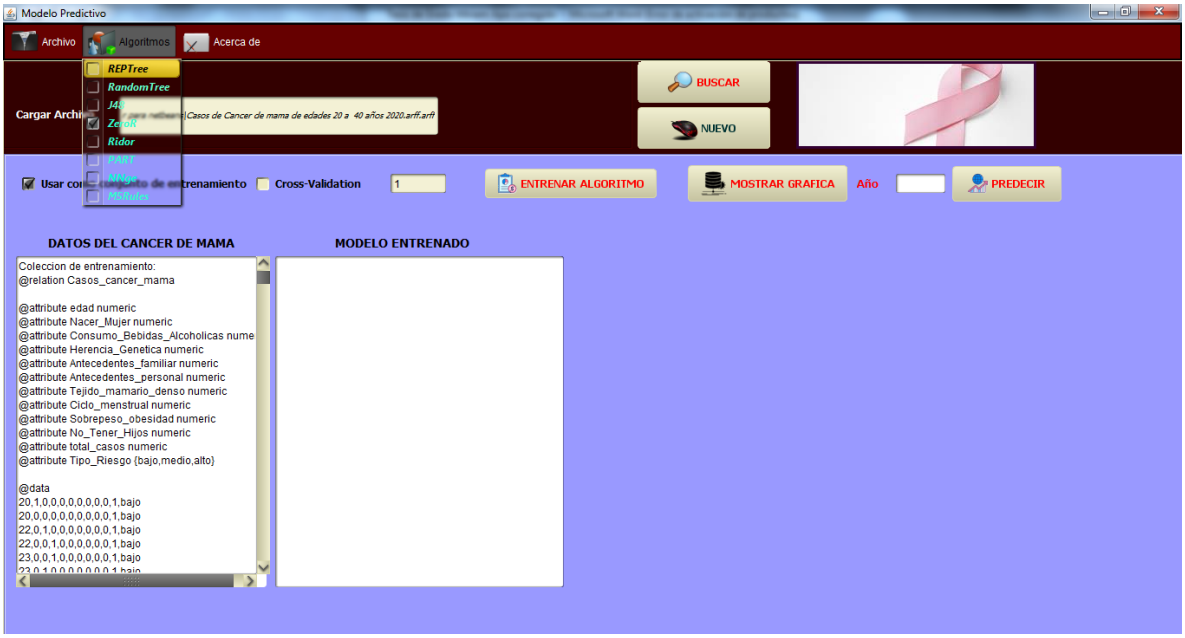

Buscar el archivo .arff y luego seleccionar y abrir



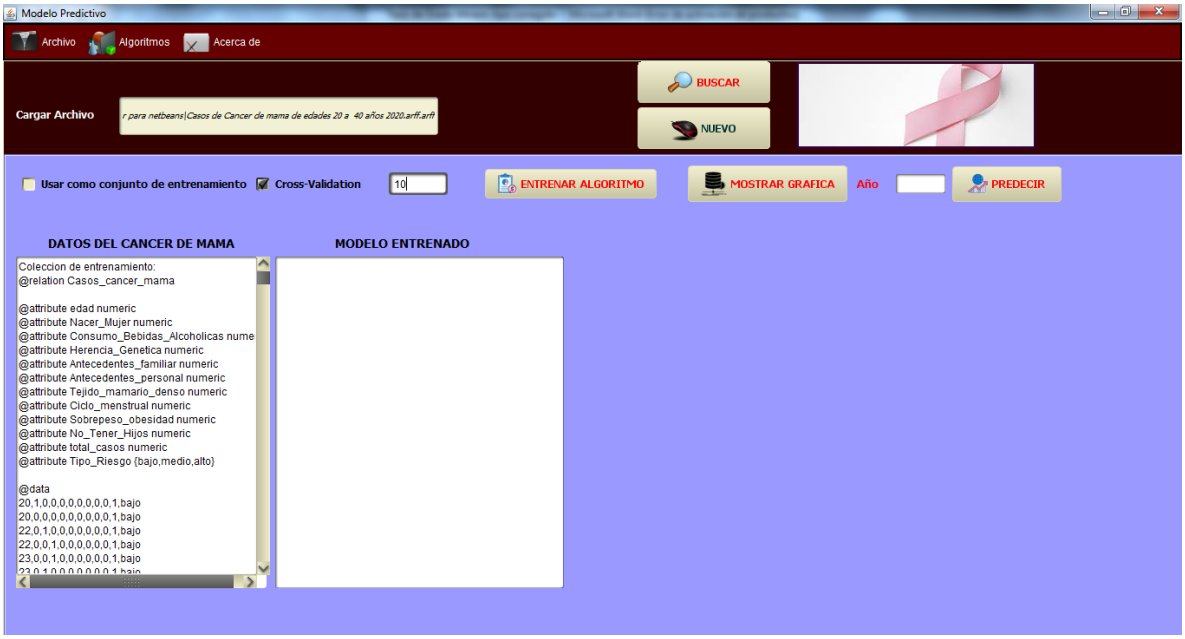

A continuación muestra el archivo .arff (datos del índice de crecimiento del cáncer de mama de mujeres de edades entre 20 a 40 años)



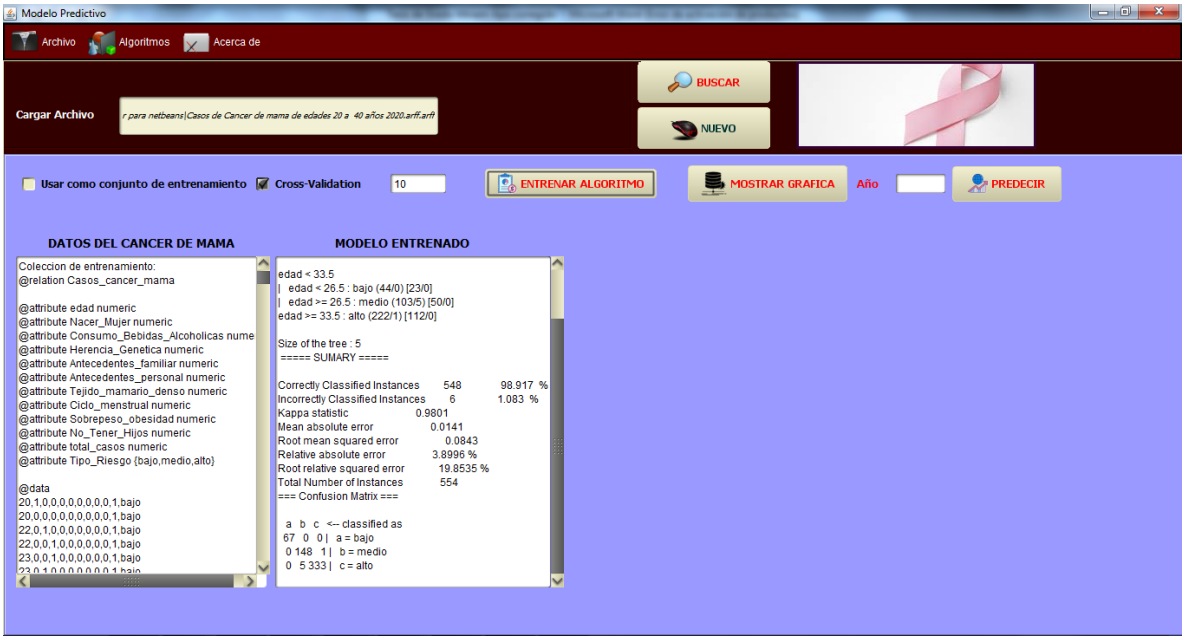

Una vez ya cargado los datos se procede a seleccionar un algoritmo para el entrenamiento del Modelo




Una vez ya seleccionado el algoritmo se selecciona Cross-Validation (Validación Cruzada) y luego se pone el número de la validación cruzada

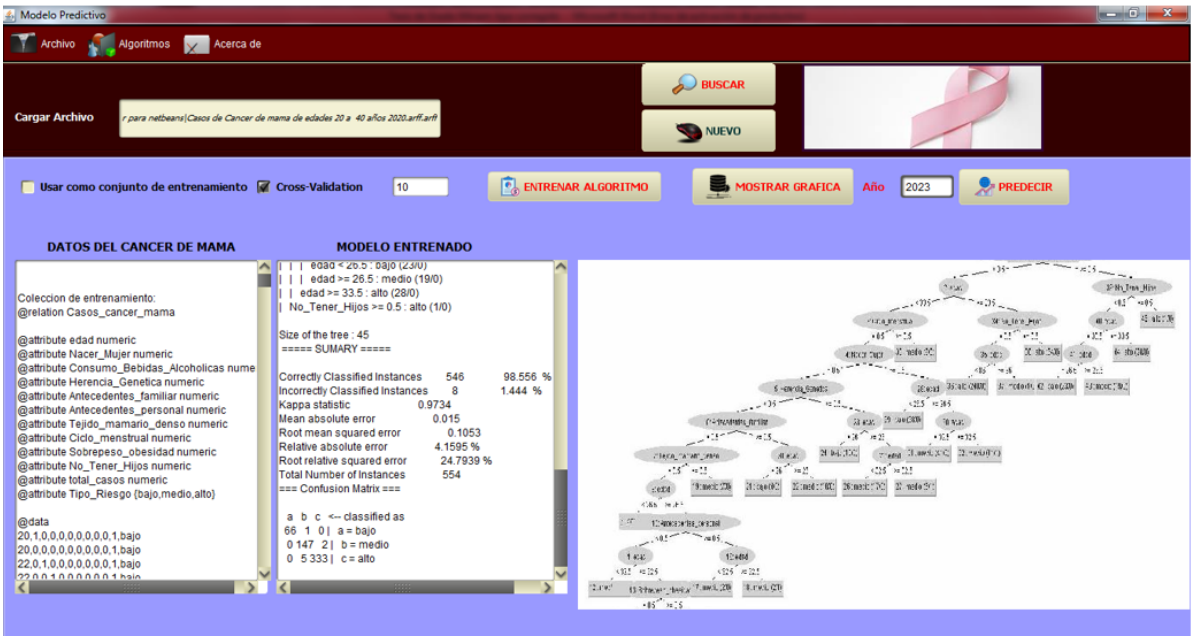



Luego se procede a hacer clic izquierdo al botón ENTRENAR ALGORITMO y muestra resultados del entrenamiento




Y luego hacemos clic en mostrar gráfica y nos muestra el árbol de decisión.

Finalmente ponemos el año y luego el botón Predecir  el resultado nos muestra observar en el Anexo A4.

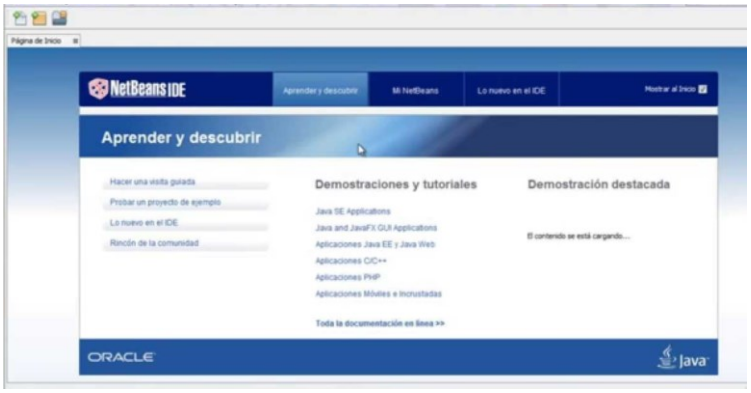


**A6. Manual Técnico.**

A continuación se muestra las siguientes herramientas libres que se instalaron para el desarrollo del Modelo Predictivo.

**Primero**: Instalar JDK Java Development Kit


**Segundo:** instalar NetBeans IDE 8.2




**Tercero** importar librerías de WEKA

GOBIERNO AUTÓNOMO DEPARTAMENTAL DE LA PAZ
SERVICIO DEPARTAMENTAL DE SALUD
RECIBIDO
30 ENE 2020 12:07
DIRECCIÓN
Nº REG. ..... Nº FOJAS. 1 FS
HOSPITAL DE LA MUJER
LA PAZ - BOLIVIA

*El Alto, 27 de enero de 2020*
*CITES/C.I.S./No 001/2019*

*Señores:*
*HOSPITAL DE LA MUJER*
*Presente.-*

***REF.: SOLICITUD DE INFORMACIÓN***

*De mi mayor consideración:*

*Al interior del plan de estudios de la carrera de Ingeniería de Sistemas, la universitaria* ***Miriam Apaza Ajnota*** *con* ***C.I.: 7063340 LP.*** *y* ***R.U.: 8002798*** *, se encuentra cursando la materia "TALLER DE LICENCIATURA II", y uno de los requisitos centrales consiste en contar con su autorización respectiva para realizar el trabajo de Tesis de Grado*

*Apreciaría mucho tenga la gentileza de proporcionar toda la información referida al tema del proyecto de grado para obtener los requerimientos de sus necesidades e incluirlas en el producto software a construir denominado "MODELO DE PREDICCIÓN DEL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA DE LAS MUJERES DE LA CIUDAD DE LA PAZ Y EL ALTO"*

*Seguro de contar con su colaboración, saludo a usted con el mayor respeto y consideración.*

*Atentamente,*

Lic. David Carlos Mamani Quispe
DIRECTOR
CARRERA INGENIERÍA DE SISTEMAS
U.P.E.A.

DIRECCION CARRERA INGENIERIA DE SISTEMAS
UNIVERSIDAD PUBLICA DE EL ALTO
El Alto - Bolivia

*DCMQ/ic*
*c.c.:Archivo C.I.S.*

***¡¡¡Con ideas claras construiremos una nueva Universidad!!!***

# Universidad Pública de El Alto

Creada por Ley 2115 del 5 de Septiembre de 2000 y Autónoma por Ley 2556 del 12 de Noviembre de 2003

## Ingeniería de Sistemas

***El Alto, 27 de enero de 2020***
*CITES/C.I.S./No 006/2019*

RECIBIDO
30 ENE 2020
INSTITUTO NACIONAL DE ESTADÍSTICA
N° 41

*Señores:*
*INSTITUTO NACIONAL DE ESTADÍSTICAS -INE*
*Presente.-*

***REF.: SOLICITUD DE INFORMACIÓN***

*De mi mayor consideración:*

*Al interior del plan de estudios de la carrera de Ingeniería de Sistemas, la universitaria* ***Miriam Apaza Ajnota*** *con* ***C.I.: 7063340 LP.*** *y* ***R.U.: 8002798*** *, se encuentra cursando la materia "TALLER DE LICENCIATURA II", y uno de los requisitos centrales consiste en contar con su autorización respectiva para realizar el trabajo de Tesis de Grado*

*Apreciaría mucho tenga la gentileza de proporcionar toda la información referida al tema del proyecto de grado para obtener los requerimientos de sus necesidades e incluirlas en el producto software a construir denominado "MODELO DE PREDICCIÓN DEL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA DE LAS MUJERES DE LA CIUDAD DE LA PAZ Y EL ALTO"*

*Seguro de contar con su colaboración, saludo a usted con el mayor respeto y consideración.*

*Atentamente,*

Ing. David Carlos Mamani Quispe
DIRECTOR
CARRERA INGENIERÍA DE SISTEMAS
U.P.E.A.

DIRECCIÓN CARRERA INGENIERIA DE SISTEMAS
UNIVERSIDAD PÚBLICA DE EL ALTO
El Alto - Bolivia

*DCMQ/ic*
*c.c.:Archivo C.I.S.*

***¡¡¡Con ideas claras construiremos una nueva Universidad!!!***

*El Alto, 27 de enero de 2020*
*CITES/C.I.S./No 002/2019*

MINISTERIO DE SALUD
RECIBIDO
13 FEB 2020
Nº:
Hora: 12:28
DIRECCIÓN GENERAL DE SERVICIOS

*Señores:*
*MINISTERIO DE SALUD*
*Presente.-*

***REF.: SOLICITUD DE INFORMACIÓN***

*De mi mayor consideración:*

*Al interior del plan de estudios de la carrera de Ingeniería de Sistemas, la universitaria* ***Miriam Apaza Ajnota*** *con* ***C.I.: 7063340 LP.*** *y* ***R.U.: 8002798*** *, se encuentra cursando la materia "TALLER DE LICENCIATURA II", y uno de los requisitos centrales consiste en contar con su autorización respectiva para realizar el trabajo de Tesis de Grado*

*Apreciaría mucho tenga la gentileza de proporcionar toda la información referida al tema del proyecto de grado para obtener los requerimientos de sus necesidades e incluirlas en el producto software a construir denominado "MODELO DE PREDICCIÓN DEL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA DE LAS MUJERES DE LA CIUDAD DE LA PAZ Y EL ALTO"*

*Seguro de contar con su colaboración, saludo a usted con el mayor respeto y consideración.*

*Atentamente,*

Ing. David Carlos Mamani Quispe
DIRECTOR
CARRERA INGENIERÍA DE SISTEMAS
U.P.E.A.

DIRECCIÓN CARRERA INGENIERÍA DE SISTEMAS
UNIVERSIDAD PÚBLICA DE EL ALTO
El Alto - Bolivia

*DCMQ/ic*
*c.c.:Archivo C.I.S.*

***¡¡¡Con ideas claras construiremos una nueva Universidad!!!***

La Paz, Febrero de 2020

Señor:
Lic Yuri Miranda Gonzales
**DIRECTOR GENERAL EJECUTIVO**
**INSTITUTO NACIONAL DE ESTADISTICAS (INE)**
Presente:

**Ref.- SOLICITUD DE INFORMACIÓN DEL CANCER DE MAMA DE LAS MUJERES DE LA CIUDAD DE LA PAZ Y EL ALTO**

Mediante la presente hacerle llegar mis cordiales saludos al mismo tiempo desearle éxitos en las funciones que desempeña.

El motivo de la presente es para solicitar **INFORMACION Y DATOS ESTADISTICOS DEL INDICE DE CRECIMIENTO DEL CÁNCER DE MAMA EN MUJERES DE LAS CIUDADES DE LA PAZ Y EL ALTO , DE LAS GESTIONES 2005 AL 2019 DE LOS DIFERENTES CENTROS HOSPITALARIOS**, en detalle:

- Por edades.
- Causas por la cual se origina esta enfermedad
- Consecuencias que sufren al padecer este tipo de cáncer
- Por tipos de cáncer de mama.
- Índice de crecimiento de mortalidad del cáncer de mama de las gestiones 2005 al 2019.
- Y otros datos que existieran.

Seguro de contar con su colaboración me despido con las consideraciones más distinguidas.

Atentamente,

**Miriam Apaza Ajnota**
**C.I. 7063340 L.P.**
**R.U. 8002798**

La Paz, Marzo de 2020

Señora:
Lic. Martha Trigo Mercado
**DIRECTORA EJECUTIVA**
**FUNDACION BOLIVIANA CONTRA EL CANCER**
Presente:

**Ref.- SOLICITUD DE INFORMACIÓN EN DETALLE DEL CANCER DE MAMA DE LAS MUJERES DE LA CIUDAD DE LA PAZ**

Mediante la presente hacerle llegar mis cordiales saludos al mismo tiempo desearle éxitos en las funciones que desempeña.

El motivo de la presente es para solicitar **INFORMACION Y DATOS ESTADISTICOS DE LA TASA DE CRECIMIENTO DEL CÁNCER DE MAMA EN MUJERES DE LAS CIUDADES DE LA PAZ , DE LAS GESTIONES 2005-2006-2007-2008-2009-2010-2011-2012-2013-2014-2015-2016-2017-2018-2019** en detalle:

- Todas las causas por la cual se origina el cáncer de mama
- Consecuencias que sufren al padecer este tipo de cáncer
- Datos por edades y por sexo
- Índice de crecimiento de mortalidad del cáncer de mama de las gestiones 2005 al 2019.
- Y otros datos que existieran.

**Nota**: Esta información solicito con el fin de realizar mi Tesis de Grado con el tema "**MODELO DE PREDICCION DEL INDICE DE CRECIMIENTO DEL CANCER DE MAMA DE LA CIUDAD DE LA PAZ, BASADO EN MINERIA DE DATOS",** en la Carrera Ingeniería de Sistemas de la Universidad Publica El Alto.

Seguro de contar con su colaboración me despido con las consideraciones más distinguidas.

Atentamente,

**Miriam Apaza Ajnota**
**C.I. 7063340 L.P.**
**Celular: 67112352**

FUNDACION BOLIVIANA CONTRA EL CANCER
12 MAR. 2020
RECIBIDO
La Paz - Bolivia

# AVAL DE CONFORMIDAD

El Alto, 10 Julio de 2020

Señor:

**HONORABLE CONCEJO DE CARRERA**
**INGENIERIA DE SISTEMAS**
Presente

**Ref.: AVAL DE CONFORMIDAD**

De mi mayor consideración Honorable concejo:

Mediante la presente tengo bien de comunicarle mi conformidad de la Tesis de Grado "MODELO DE PREDICCIÓN SOBRE EL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA EN LAS MUJERES DE EDADES ENTRE 20 A 40 AÑOS DE LA CIUDAD DE LA PAZ, BASADO EN MINERÍA DE DATOS" que propone la postulante Univ. Miriam Apaza Ajnota con C.I. 7063340 LP, para su defensa pública, evaluación correspondiente a la materia Taller de Licenciatura II, de acuerdo al reglamento vigente de la carrera Ingeniería de Sistemas de la Universidad Pública de El Alto.

Sin otro particular, reciba saludos cordiales.

Atentamente,

---

Ing. Marisol Arguedas Balladares
TUTOR METODOLÓGICO

**AVAL DE CONFORMIDAD**

El Alto, Junio de 2020

Sin otro particular, reciba cordiales saludos

Atentamente.

Señor:

Ing. David Carlos Mamani Quispe
**DIRECTOR**
**CARRERA INGENIERIA DE SISTEMAS**

Presente.-

Ing. Enrique Flores Baltazar

**Ref.- Aval de Conformidad**

Distinguido Director:

Mediante la presente tengo a bien comunicarle mi conformidad del Perfil de Tesis de Grado, **TUTOR INTELIGENTE PARA EL "MODELO DE PREDICCIÓN SOBRE EL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA EN LAS MUJERES DE EDADES ENTRE 20 A 40 AÑOS DE LA CIUDAD DE LA PAZ, BASADO EN MINERÍA DE DATOS",** que propone la postulante Miriam Apaza Ajnota, con Cedula de Identidad 7063340 L.P., para su defensa publica y su evaluación correspondiente de la materia Taller I, de acuerdo al Reglamento Vigente de la Carrera Ingeniería de Sistemas de la Universidad Pública de El Alto.

Sin otro particular, reciba saludos cordiales.

Atentamente,

Ing. Enrique Flores Baltazar
TUTOR ESPECIALISTA

# AVAL DE CONFORMIDAD

El Alto, 07 Julio de 2020

Señora:
Ing. Marisol Arguedas Balladares
**TUTOR METODOLÓGICO TALLER DE LICENCIATURA II**
Presente

**Ref.: AVAL DE CONFORMIDAD**

Distinguida Ingeniera:

Mediante la presente tengo bien de comunicarle mi conformidad de la Tesis de Grado "MODELO DE PREDICCIÓN SOBRE EL ÍNDICE DE CRECIMIENTO DEL CÁNCER DE MAMA EN LAS MUJERES DE EDADES ENTRE 20 A 40 AÑOS DE LA CIUDAD DE LA PAZ, BASADO EN MINERÍA DE DATOS" que propone la postulante Miriam Apaza Ajnota con C.I. 7063340 LP, para su defensa publica , evaluación correspondiente a la materia Taller de Licenciatura II, de acuerdo al reglamento vigente de la carrera Ingeniería de Sistemas de la Universidad Pública de El Alto.

Sin otro particular, reciba saludos cordiales.

Atentamente,

---

Ing. Julio Mamani Luque
TUTOR REVISOR