ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

PROGRAMA DE MAESTRIA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TRABAJO DE TITULACIÓN, PREVIO A LA OBTENCIÓN DEL TITULO DE MAGISTER EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

TEMA: "ANÁLISIS PARA PREDICCIÓN DE COMPRA DE MEDICAMENTOS PARA EL ÁREA DE ONCOLOGÍA DEL HOSPITAL CARLOS ANDRADE MARIN APLICANDO TÉCNICAS DE MINERÍA DE DATOS"

AUTOR: SÁNCHEZ TAPIA EVELYN VALERIA

DIRECTOR: ING. MOLINA MARCO PHD.

SANGOLQUÍ - ECUADOR

2017

ii

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

PROGRAMA DE MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

CERTIFICACIÓN

Certifico que el trabajo de titulación, "ANÁLISIS PARA PREDICCIÓN DE COMPRA DE MEDICAMENTOS PARA EL ÁREA DE ONCOLOGÍA DEL HOSPITAL CARLOS ANDRADE MARIN APLICANDO TÉCNICAS DE MINERÍA DE DATOS" realizado por el señor/a EVELYN VALERIA SÁNCHEZ TAPIA, ha sido revisado en su totalidad y analizado por el software anti-plagio, el mismo cumple con los requisitos teóricos, científicos, técnicos, metodológicos y legales establecidos por la Universidad de Fuerzas Armadas ESPE, por lo tanto me permito acreditarlo y autorizar al señor ING. MARCO MOLINA PHD para que lo sustente públicamente.

Sangolquí, 20 de septiembre del 2017.

Ing. Marco Molina PhD.
DIRECTOR

UNIVERSIDAD DE LAS FUERZAS ARMADAS
INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

PROGRAMA DE MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORÍA DE RESPONSABILIDAD

Yo, EVELYN VALERIA SÁNCHEZ TAPIA, con cédula de identidad N° 172112607-4, declaro que este trabajo de titulación "ANÁLISIS PARA PREDICCIÓN DE COMPRA DE MEDICAMENTOS PARA EL ÁREA DE ONCOLOGÍA DEL HOSPITAL CARLOS ANDRADE MARIN APLICANDO TÉCNICAS DE MINERÍA DE DATOS" ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado los derechos intelectuales de terceros considerándose en las citas bibliográficas. Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance de la investigación mencionada.

Sangolquí, 20 de septiembre del 2017.

EVELYN VALERIA SÁNCHEZ TAPIA
C.C.: 172112607-4

ESPE

UNIVERSIDAD DE LAS FUERZAS ARMADAS

INNOVACIÓN PARA LA EXCELENCIA

VICERRECTORADO DE INVESTIGACIÓN INNOVACIÓN Y TRANSFERENCIA DE TECNOLOGÍA

CENTRO DE POSGRADOS

PROGRAMA DE MAESTRÍA EN GESTIÓN DE SISTEMAS DE INFORMACIÓN E INTELIGENCIA DE NEGOCIOS

AUTORIZACIÓN

Yo, EVELYN VALERIA SÁNCHEZ TAPIA,, autorizo a la Universidad de las Fuerzas Armadas ESPE publicar en la biblioteca Virtual de la institución el presente trabajo de titulación "ANÁLISIS PARA PREDICCIÓN DE COMPRA DE MEDICAMENTOS PARA EL ÁREA DE ONCOLOGÍA DEL HOSPITAL CARLOS ANDRADE MARÍN APLICANDO TÉCNICAS DE MINERÍA DE DATOS" cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, 20 de septiembre del 2017.

EVELYN VALERIA SÁNCHEZ TAPIA
C.C.: 172112607-4

DEDICATORIA

A Dios, por permitirme el haber llegado a este momento importante de mi formación profesional. Por los triunfos y momentos difíciles que me han enseñado a valorarlo cada día más. A mi esposo e hijo que ha sido un pilar fundamental para continuar mejorando como persona y profesional. A mis padres que con su tenacidad y lucha insaciable han hecho de ellos el gran ejemplo a seguir y destacar, no solo para mí, sino para mis hermanos y familia en general. A mí tutor de tesis, por su valiosa guía y asesoramiento en la realización de esta investigación.

Valeria

AGRADECIMIENTOS

Los resultados de este proyecto, están dedicados a todas aquellas personas que, de alguna forma, son parte de su culminación. Mi sincero agradecimiento está dirigido a mi familia por siempre brindarme su apoyo, tanto sentimental, como económico, pero, principalmente mi agradecimiento es para mi amado esposo quien ha apoyado y motivado mi formación académica creyendo en mí y nunca dudando de mi capacidad. A mis docentes a quienes les debo gran parte de mis conocimientos, gracias a su paciencia y enseñanza y finalmente un eterno agradecimiento a esta prestigiosa universidad que me ha preparado para un futuro competitivo.

Valeria

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El abastecimiento de medicamentos en forma oportuna y en las cantidades requeridas para atender las necesidades de salud de los pacientes del área de oncología, es uno de los factores críticos que afecta a la gestión del sistema de suministro del Hospital "Carlos Andrade Marín", para dar respuesta a esta problemática se busca el apoyo de la minería de datos que hoy en día está cobrando una relevancia creciente en las organizaciones para resolver problemas complejos, y, a través del procesamiento de volúmenes de datos descubre información valiosa para el negocio como por ejemplo el comportamiento de compra cuyo método de adquisición se enfoca en determinar la cantidad de producto a comprarse para un determinado periodo, satisfaciendo la demanda de la población. Este trabajo de investigación se centra en seguir una metodología de CRISP-DM para poder realizar una predicción de compra de medicamentos, aplicando técnicas de minería de datos como Modelo Lineal y Maquina de Vector de Soporte a la información proporcionada por el nosocomio, permitiendo obtener patrones de comportamiento en el movimiento de consumo de medicamentos oncológicos para garantizar su disponibilidad de acuerdo a las necesidades de los pacientes. Los resultados del modelo de predicción permiten realizar la planeación de la cantidad de medicamentos oncológicos para el Plan Anual de Contratación de la institución, permitiendo al personal de adquisiciones contar con el tiempo suficiente para seguir su proceso de compra normal, sin hacer reformas al presupuesto anual planificado.

Palabras Clave:

- **MINERÍA DE DATOS**
- **METODOLOGÍA DE CRISP-DM**
- **BASES DE DATOS**
- **MODELO LINEAL**
- **MÁQUINA DE VECTOR DE SOPORTE**

ABSTRACT

The provision of medicine at the appropriate time and at the required doses to tend to patients' medical needs in the oncology wing, is one of the critical factors distressing the supply system management in the "Carlos Andrade Marin" Hospital. To resolve this problem, we seek the support of data mining which is becoming increasingly relevant today in organizations to solve complex problems. Through the processing of large volumes of data, it discovers valuable information for a business for example the buying behavior whose method of acquisition focuses on determining the quantity of a product to be purchased for any given period of time, therefore satisfying public demand. This research work focuses on following the CRISP-DM methodology to be able to make a prediction of medicine purchase, applying techniques from data mining like Linear Model and Support Vector Machine to the information provided by the hospital therefore allowing to obtain patterns of behavior in the movement of consumption of oncological medicine to guarantee availability according to patient's needs. The results of the prediction model make it possible to plan the number of oncological drugs for the Annual Plan of Contracting for the institution. This will allow the purchasing staff to have enough time to follow their process of normal purchase, without making reforms to the planned annual budget.

Keywords

- **DATA MINING**
- **CRISP-DM METHODOLOGY**
- **DATABASES**
- **LINEAR MODEL**
- **SUPPORT VECTOR MACHINE**

CAPÍTULO I
ASPECTOS GENERALES

1.1 Introducción

Una de las necesidades actuales corresponde a la búsqueda de una solución para mitigar los problemas de las casas de salud administradas por el Instituto Ecuatoriano de Seguridad Social (IESS), dónde la demanda siempre está por encima de la oferta, ya que dichas entidades afrontan inconvenientes entre los que destacan la falta de camas, equipos y principalmente medicamentos en el servicio de farmacia, generando malestar en los afiliados, ya que al no existir un stock de los mismos, pueden retrasar y complicar el proceso de recuperación y tratamiento de los pacientes. De esta manera, el desabastecimiento de medicamentos es la razón por la cual existe la posibilidad de establecer una alternativa para mejorar la compra planificada de medicamentos cuyo método de adquisición se enfoca en determinar la cantidad de medicinas a comprarse para un determinado periodo, satisfaciendo la demanda de los pacientes.

Sobre la base de lo expuesto, una alternativa viable para lograr una solución definitiva al problema corresponde a implementar técnicas de predicción de minería de datos, las cuales ayudarán a generar recomendaciones apropiadas para la planeación de compra de medicamentos con escaso o nulo margen de error, con lo cual se garantiza su abastecimiento y el bienestar de los pacientes afiliados al IESS.

Las técnicas de minería de datos, aplicadas a la compra y almacenamiento de medicamentos, aportarán grandes ventajas para la administración de fármacos del hospital, puesto que se busca garantizar la disponibilidad de medicamentos, generando un mejor servicio en beneficio de la comunidad; adicionalmente, el proceso de compra planificada se realizará siguiendo el debido flujograma previamente establecido por las partes implicadas, que significa ahorro de recursos para la institución, tomando en cuenta que se busca que el presupuesto asignado vaya

de acuerdo a las necesidades, y no que se ajuste a un presupuesto, sin tomar en cuenta la demanda real de esta casa de salud.

Este documento se encuentra dividido en tres partes:

- La primera parte pone en contexto al lector y entrega una serie de conocimientos básicos acerca del descubrimiento de conocimiento en bases de datos y minería de datos, esto permite una conceptualización adecuada de los términos a usarse en la investigación.
- En la segunda parte se detalla paso a paso el proceso de elaboración de un modelo de acuerdo a la metodología seleccionada, para finalmente hacer una valoración de modelo obtenido verificando si satisface el objetivo de esta investigación.
- En la tercera parte se presentan las conclusiones y recomendaciones a las que se llega después de realizar la investigación.

1.2 Justificación e Importancia.

Las técnicas de minería de datos aplicadas a un determinado proceso de una institución, en este caso relacionado con la compra de medicamentos, aportan un gran beneficio al hospital, ya que pueden generar recomendaciones apropiadas para la planeación de compra de medicamentos con escaso o nulo margen de error, con lo cual se garantiza su abastecimiento y el bienestar de los pacientes afiliados al IESS que se realizan sus chequeos en el Hospital “Carlos Andrade Marín”.

De esta manera se aplica la minería de datos que se remite a un conjunto de herramientas y técnicas de análisis, que por medio de la identificación de patrones, extrae información interesante, novedosa y potencialmente útil de grandes bases de datos que puede ser utilizada como soporte para la toma de decisiones. Por lo mencionado, básicamente se requiere disponer de información histórica que se encuentre relacionada con la investigación como por ejemplo:

Registros históricos:

- Consumo actual de medicamentos.

- Saldos de medicamentos.
- Modalidad de la compra.
- Medicamentos Oncológicos que se encuentran en el listado del CNMB.
- Población Usuaria.
- Consumo histórico (cantidades despachadas por mes /semanal / anual; cantidades recetadas y no despachadas, etc.).
- Perfil epidemiológico (Registros de consultas, historias clínicas, egresos hospitalarios, certificados de defunción).

Además es importante conocer la forma cómo se relacionan las variables que integran los registros en la administración de medicamentos, ya que a partir de allí se pueden determinar situaciones particulares que no son apreciables a simple vista, y con ello detectar patrones que permitan sugerir medidas para optimizar recursos y evitar desabastecimiento de medicamentos.

1.3 Planteamiento del problema.

1.3.1 Formulación del problema.

Este proyecto de tesis tiene como propósito dar respuesta al siguiente cuestionamiento:

1. ¿Es posible desarrollar un mecanismo predictivo haciendo uso de alguna plataforma para la compra de medicamentos basado en análisis y patrones de registros históricos de medicamentos recetados?

1.3.2 Preguntas de Investigación.

Este proyecto de tesis tiene como propósito dar respuesta a los siguientes cuestionamientos:

1. ¿Es posible desarrollar un mecanismo predictivo haciendo uso de alguna plataforma para la compra de medicamentos, basado en el análisis de patrones descubiertos en registros históricos de medicamentos recetados?

2. ¿Es posible optimizar el stock de medicamentos, que evite un desabastecimiento y/o sobrante excesivo de medicamento haciendo uso de algún modelo de minería de datos?
3. ¿Hay la posibilidad de agilizar el proceso de compra de medicamentos planificados para un periodo futuro con alto nivel de confianza?

1.4 Limitaciones y Supuestos.

Al finalizar este trabajo de tesis, se espera contar con un sistema que guíe a la Coordinación General de Planificación y Estadística, áreas responsables del Plan Anual de Compras "PAC" y de realizar la planeación y solicitud de necesidades de los medicamentos oncológicos de la institución de acuerdo a sus requerimientos; no obstante también deben tener la capacidad de predecir los momentos en los que se deben realizar las reposiciones masivas de medicamentos, que permitirá al personal de adquisiciones contar con el tiempo suficiente para seguir su proceso de compra normal.

La investigación tiene como fin, utilizar la minería de datos haciendo uso de una plataforma para aplicar un modelo de predicción de compras de medicamentos elaborado sobre un conjunto de datos históricos seleccionados, con el fin de encontrar patrones que ayuden a determinar un stock con alto nivel de confianza.

- **Obtención de los datos:** La presente investigación se orienta fundamentalmente a identificar cuáles son las variables que influyen en la decisión de compra de medicamentos para el área de Oncología. De acuerdo con una investigación exploratoria en diferentes departamentos de este nosocomio, se definen una serie de variables que influyen en este proceso de compra y de las cuales se requiere información necesaria. Estas variables son:
    - Consumo actual de medicamentos.
    - Saldos de medicamentos.
    - Tiempo para el cual se efectúa la compra.
    - Modalidad de la compra.
    - Proveedores.

- Medicamentos Oncológicos que se encuentran en el listado del CNMB.

Por otro lado, los lineamientos que deben ser tomados en cuenta para determinar la cantidad de medicamentos que se requiere para cierto periodo son: la población usuaria, consumo histórico, perfil epidemiológico comparado con la definición de las necesidades del servicio, oferta del servicio y el presupuesto disponible.

- **Procesamiento de Datos:** En esta fase se enfoca en la limpieza y selección de datos, iniciando con la eliminación del mayor número posible de datos erróneos, inconsistentes e irrelevantes.

- **Construcción de Modelo:** La construcción del modelo requiere que los datos preparados puedan ser utilizados iterativamente, en otras palabras, poder aplicar algoritmos y técnicas sobre diferentes vistas "minables" y de esta manera descubrir patrones de comportamiento.

- **Evaluación del Modelo:** Este es el último paso donde se evalúa el modelo, no desde el punto de vista general, sino del cumplimiento de los objetivos del negocio. Se debe revisar el proceso teniendo en cuenta los resultados obtenidos, para repetir alguna fase en caso que se hayan cometido errores.

- **Interpretación de Resultados:** Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y aceptables para la solución del problema. Sin embargo, existen varias soluciones para un determinado problema, en este caso se selecciona la respuesta de mayor acercamiento a la solución del problema. Se debe realizar una valoración de los pasos previos a la construcción de los modelos en caso de que estos, no cumplan los objetivos esperados y crear nuevos modelos.

1.5 Objetivos.

1.5.1 Objetivo General.

El objetivo general de esta investigación es:

- Proponer un mecanismo, basado en BI, confiable para apoyar la toma de decisiones a la hora de comprar medicamentos en una institución hospitalaria.

1.5.2 Objetivos Específicos.

Los objetivos específicos de esta investigación son:

- Recopilar e integrar los datos históricos de la adquisición y administración de los medicamentos de pacientes oncológicos que reciben tratamiento en el Hospital "Carlos Andrade Marín.
- Diseñar un modelo de predicción que permita mejorar el proceso de compra de medicamentos para el área de Oncología en el Hospital "Carlos Andrade Marín.
- Validar la técnica, comprobando que ésta se ajusta a los requerimientos del problema planteado.
- Explicar los resultados obtenidos y obtener conclusiones aleccionadoras a partir de los mismos.

CAPÍTULO II

MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Marco teórico

2.1.1 Antecedentes históricos de la Minería de datos.

El concepto de Minería de Datos "*Data Mining*" no es reciente, ya que se lo viene utilizando desde hace algunos años atrás, donde profesionales estadísticos manejaban términos como *Data Fishing* "pesca de datos", o *Data Archaeology* "arqueología de datos", cuyo objetivo principal era encontrar correlaciones sin una hipótesis previa en una base de datos con ruidos (Pautsch, 2016).

Tampoco los modelos estadísticos empleados en la Minería de Datos son nuevos, ya que modelos como árboles de decisión fueron utilizados desde los 60's, mientras que las redes neuronales se conocen desde los 40's; sin embargo y debido a su complejidad ha tomado varios años de desarrollo para que su uso sea más sencillo.

Durante todo este tiempo las técnicas fueron evolucionando y empezaron a consolidarse y todo gracias a la ayuda de investigadores como: Gio Wiederhold, Piatetsky-Shapiro, Rakesh Agrawal y Robert Blum. (Bigado & Arruzazabala, 2003).

En 1989, Gregory Piatetsky-Shapiro acuñó el término Descubrimiento de Conocimiento en Bases de Datos KDD "Knowledge Discovery in Database", término que al poco tiempo se hizo famoso en la comunidad científica y académica. Años posteriores la Minería de Datos o "Data Mining" fueron tomando fuerza en el sector empresarial para constituirse en un apoyo esencial para la toma de decisiones (Consultores, 2016).

2.1.2 Antecedentes conceptuales y referenciales de la Minería de Datos.

- **Outlier:** Es un valor diferente al conjunto de datos de una muestra, es decir, es un dato cuyo patrón es inconsistente y se encuentra a una distancia anormal de otros valores con la relación que evidencia todo el conjunto de datos, sin embargo estos datos atípico "outliers" deben ser analizados y determinar cuál es un comportamiento normal de los datos para hacer una diferenciación de un comportamiento anormal. (Sematech, 2012)

- **Regresión lineal:** Es una técnica estadística que permite cuantificar la dependencia entre dos variables, es decir, es una función lineal que requiere de dos parámetros cuya tendencia es rectilínea. (Franco, 2001)

- **RMSE(Root Mean Squared Error):** La Raíz del Error Cuadrático Medio es un indicador que mide los resultados de predicción de la demanda que genera el modelo construido. (GEO, 2015)

- **MAE (Mean Absolute Error):** Error Absoluto Medio es un indicador que permite medir la cercanía de las predicciones a los resultados. (Julia, 2017)

2.2 Estado del Arte.

2.2.1 Descubrimiento de Conocimiento KDD.

2.2.1.1 Definición de KDD.

De acuerdo al avance tecnológico generado en la última década, existe una acumulación de grandes volúmenes de datos que promueven el desarrollo de nuevas teorías y herramientas computacionales para ayudar a los seres humanos a extraer información útil (conocimiento) de sus datos almacenados. Estas teorías y herramientas involucran el Descubrimiento de Conocimiento en Base de Datos, término que se resume con las siglas KDD (*Knowledge Discovery in Databases*).

En la literatura actual existen algunas definiciones de KDD, sin embargo una de las definiciones más acertada menciona que la exploración de los datos almacenados utilizando técnicas de minería de datos como: estadística, detección de patrones y máquinas de aprendizaje permite obtener conocimiento tan valioso para las empresas. (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

De esta manera, el KDD tiene relación con el desarrollo de técnicas y métodos para dar sentido a los datos, en síntesis es un es proceso compuesto de varias fases que permiten adquirir conocimiento.

2.2.1.2 Proceso de KDD.

El término proceso implica que el KDD está compuesto por un conjunto de pasos que involucran la preparación de datos, búsqueda de patrones, evaluación del conocimiento y refinamiento, ya que los patrones descubiertos deben ser válidos con cierto grado de certeza que conduzca a algún beneficio al usuario. El proceso de KDD está centrado en el usuario y es altamente interactivo, y está guiado por la toma de decisiones del usuario y un agente inteligente, donde el desafío real es dar inteligencia al sistema para obtener conocimiento (Nigro, Xodo, Corti, & Terren, 2016).

Generalmente este proceso considera las siguientes etapas:

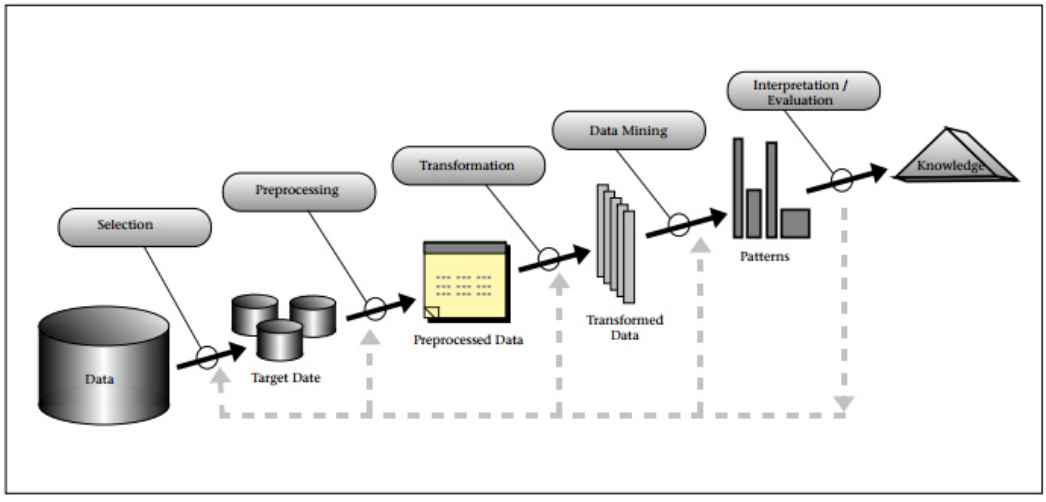

Figura 1. Proceso de Descubrimiento de Conocimiento
Fuente: (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

1. **Selección de datos:** Esta etapa consiste en entender cuál es el problema a resolver y cuáles son los objetivos que se desean alcanzar, esto determinará al investigador las fuentes de datos y que información debe ser extraída, buscando atributos apropiados de entrada que genere información relevante para llegar a la meta. (Diaz, Iribarrra, & Gutierrez, 2016)

2. **Preprocesamiento de datos:** En esta fase busca preparar y limpiar los datos extraídos desde las diferentes fuentes, este proceso utiliza una serie de estrategias para limpiar los datos sucios esto incluye datos incompletos, inconsistencias, datos en blanco que pueden afectar a las fases posteriores y contribuir a un análisis inexacto dando lugar a resultados incorrectos. (Diaz, Iribarrra, & Gutierrez, 2016)

3. **Transformación de datos:** Consisten en un tratamiento que se realiza a los datos principalmente en modificaciones sintácticas que permiten generar nuevas variables tomando como base las ya existentes. (Han & Kamber, 2001)

   Una de las ventajas de la transformación de los datos es que puede reducir el tiempo de búsqueda, ya que son de mejor compresión las reglas generadas por el algoritmo, sin embargo, también se puede tener perdida de información al realizar transformación de datos provocando una reducción de exactitud del conocimiento descubierto. (Diaz, Iribarrra, & Gutierrez, 2016)

4. **Minería de Datos:** Esta fase consiste en aplicar métodos inteligentes para encontrar una serie de patrones que expresen un comportamiento o tendencia en los datos. (Han & Kamber, 2001)

5. **Evaluación de los patrones:** En esta fase se debe validar el conocimiento adquirido probando los modelos creados, donde se identifican patrones

interesantes usando diferentes técnicas. (Nigro, Xodo, Corti, & Terren, 2016)

6. **Interpretación de resultados:** Consiste en comprender los resultados generados por los modelos evaluados en ciertas circunstancias es necesario regresar a pasos anteriores ya que los resultados no son los esperados. (Diaz, Iribarrra, & Gutierrez, 2016).

2.2.2 Minería de Datos.

Desde el punto de vista académico, el término "Minería de datos o *Data mining*" es una etapa que está integrada dentro de un proceso más grande llamado Descubrimiento de Conocimiento en Base de Datos; sin embargo, la minería de datos reúne las ventajas de algunas áreas como estadística, inteligencia artificial, computación gráfica y procesamiento masivo cuya fuente son las bases de datos.

Debido al crecimiento de los negocios donde el almacenamiento de sus datos es cada vez más importante y se ha convertido en un elemento estratégico para la toma de decisiones. En este sentido, las empresas han venido evolucionado y buscan agregar valor a la cantidad de datos que tienen almacenados en sus bases, razón por la cual han enfocado sus esfuerzos en automatizar los procesos y descubrir conocimientos que generen un valor agregado a sus productos y servicios; adicionalmente a este propósito se ha sumado la evolución tecnológica de los últimos años que a través del uso de software y hardware permiten el almacenamiento de grandes cantidades de datos y el análisis de los mismos.

2.2.2.1 Definición de Minería de Datos.

La bibliografía con respecto a minería de datos es bastante amplia, sin embargo se resalta las siguientes:

Los patrones interesantes pueden usarse para determinar algo nuevo o hacer predicciones. El proceso de descubrimiento del conocimiento se compone de varios

pasos, incluyendo la selección de datos para ser analizados y prepararlos, aplicar algoritmos de minería de datos para luego interpretar y evaluar los resultados. El término minería de datos se refiere al proceso de encontrar y utilizar los patrones interesantes en los datos (Benoît, 2002).

Data mining es un paso dentro del proceso KDD, que consiste en aplicar algoritmos de análisis y descubrimiento, que generan una enumeración de patrones particulares a partir de datos preprocesados (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

2.2.2.2 Algoritmos de Minería de Datos.

Existen dos grandes grupos de algoritmos de minería de datos: supervisados y no supervisados (Weiss & Indurkhya, 1998). Los algoritmos supervisados o predictivos son aquellos que predicen un valor de un atributo (etiqueta) a partir de un conjunto determinado de datos, estos algoritmos se usan en base a un entrenamiento y prueba, y permiten realizar predicciones de datos cuya etiqueta es desconocida. Por otro lado, los algoritmos no supervisados o de descubrimiento de conocimiento se dedican a encontrar patrones o tendencias sobre datos actuales que generan información valiosa, lo que permite tomar decisiones para obtener un beneficio del negocio o científico.

2.2.2.3 Tareas de Minería de Datos.

Las tareas de minería de datos se encuentran detalladas a continuación:

Tareas Descriptivas: Se encaminadas a representar un acumulado de datos.

- **Clustering:** Son conocidos como agrupamiento y permite la identificación de grupos donde sus elementos tienen una gran similitud y varias diferencias con los otros grupos.

  Clustering se caracteriza por usar una técnica que mide la similitud que está basada en los atributos que describen a cada objeto y se definen

habitualmente por la proximidad en un espacio multidimensional. (Zapata, 2011)

- **Reglas de Asociación:** Esta técnica se utiliza para determinar las posibles relaciones o correlaciones entre distintas acciones que son aparentemente independiente. Un ejemplo del uso de esta técnica es cuando se realizan análisis exploratorios, donde se buscan relaciones entre un conjunto de datos determinados que permitan predecir comportamientos y descubrir correlaciones y co-ocurrencias de eventos (Belinchón, s.f.).

Tareas Predictivas: están orientadas a predecir valores futuros desconocidos.

- **Predicción:** La técnica de predicción intenta predecir a partir de una muestra de datos los valores de una o más variables (Zapata, 2011).
- **Clasificación:** El proceso de la clasificación se encarga de dividir un conjunto de datos en grupos mutuamente excluyentes (Belinchón, s.f.).

2.2.2.4 Técnicas de Minería de Datos.

Se define a una técnica como un procedimiento para ejecutar una determinada tarea, en este caso particular se busca obtener conocimiento a partir de un conjunto de datos por medio de algoritmos. (Zapata, 2011)

Para poder resolver el problema de minería de datos es necesario conocer el algoritmo a ser aplicado, de esta manera preparar los datos que serán analizados y saber la técnica a ser utilizada, para construir un modelo que genere resultados óptimos.

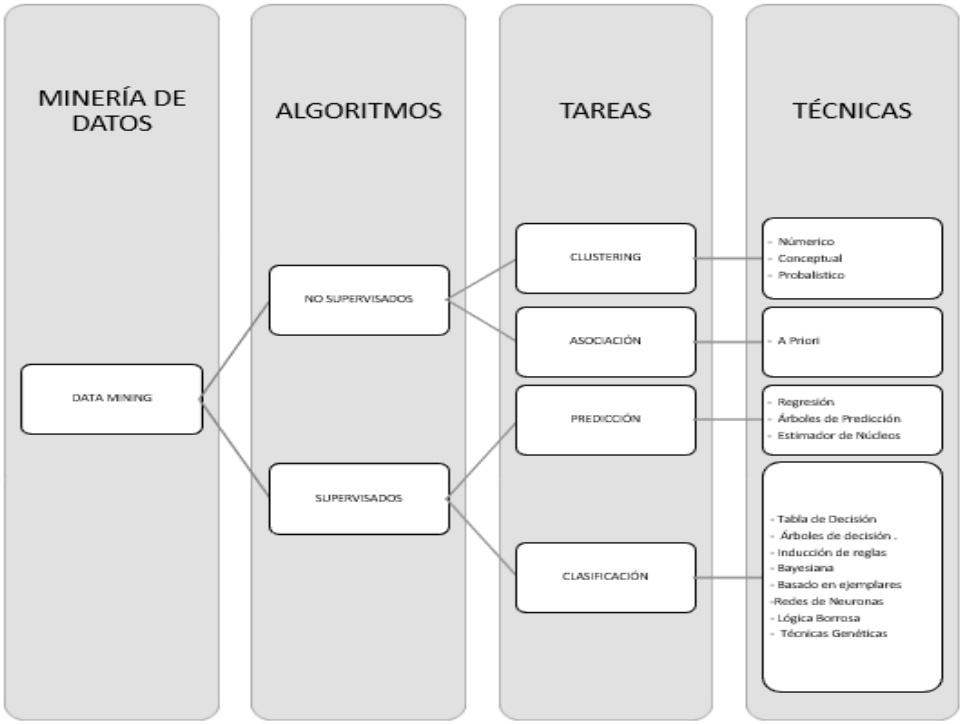

Figura 2. Técnicas de Minería de datos
Fuente: *(Zapata, 2011)*

A continuación se presentan de manera detallada las principales técnicas de minería de datos:

1. **Modelización estadística paramétrica.**

Explica el comportamiento de una variable que está extrayendo conocimiento a partir del uso de fórmulas algebraicas (Hernández, Ramírez, & Ferri, 2004). La regresión y la clasificación son dos de las tareas más utilizadas para este tipo de técnica, mientras que los algoritmos conocidos para dicha técnica son regresión lineal, logarítmica y logística.

La modelización lineal paramétrica en un contexto de minería de datos, utiliza métodos matemáticos ya existentes como regresión lineal múltiple (MLR), modelo lineal de ejemplos mixtos (LME) y mínimos cuadrados parciales (PLS). La ventaja de esta técnica es que puede ajustarse con una pequeña cantidad de datos, mientras que su desventaja corresponde a su rigurosidad que le dificulta adaptarse a un gran conjunto de datos, debido a que necesita gran nivel de procesamiento (Hernández, Ramírez, & Ferri, 2004).

2. Modelización estadística no paramétrica.

Funcionan al contrario de la técnica de modelización estadística paramétrica, ya que son capaces de modelar fenómenos complejos, ajustados a una gran cantidad de datos (Hernández, Ramírez, & Ferri, 2004). Al igual que la técnica anterior se utiliza en tareas de regresión y clasificación con métodos como k-ésimo vecino más cercano (k-NN) y vecinos más similar (MSN).

3. Reglas de asociación y dependencia.

Esta técnica consiste en que mediante reglas se determina el comportamiento entre los datos de las clases del dominio en función de la aparición conjunta de los valores de dos o más atributos (Hernández, Ramírez, & Ferri, 2004). Esta técnica expresa las combinaciones de valores de los atributos con mayor frecuencia, es decir, utiliza una tarea de asociación.

Una de sus principales características es que trabaja con grandes volúmenes de datos, y sus reglas generalmente trabajan con atributos nominales como por ejemplo el atributo edad (joven, adulto). Las reglas de asociación hacen uso del algoritmo A-priori que extrae los patrones de comportamiento.

4. Métodos Bayesianos.

Estos métodos usan una metodología para la inferencia y predicción, y en cierta instancia, para tomar decisiones que involucran cantidades inciertas. Adicionalmente una de las características de esta técnica es que usa distribuciones de probabilidad para cuantificar la incertidumbre de los datos que desea modelar (Hernández, Ramírez, & Ferri, 2004). Los métodos bayesianos utilizan las tareas de clasificación para extraer los patrones de comportamiento, mientras que los algoritmos que utiliza esta técnica son el clasificador bayesiano Naive, Bayes Net y el algoritmo EM.

5. Árboles de decisión.

Un árbol de decisión es un conjunto de nodos que mantienen una estructura jerárquica, donde cada uno establece una condición o regla que devuelve un

resultado de verdadero o falso según los atributos analizados. La decisión final se obtiene siguiendo las condiciones desde nodo raíz hasta alguno de los nodos hoja (Hernández, Ramírez, & Ferri, 2004). Las tareas que utiliza esta técnica son: clasificación, regresión y agrupamiento, esta técnica se basa en dos algoritmos: "divide y vencerás" como ID3/C4.5 o el CART, y "separa y vencerás" como CN2.

6. Redes neuronales artificiales.

Esta técnica trabaja en base a un modelo de entrenamiento de los valores que conectan un conjunto de neuronas de la red, y recibe como entrada un conjunto de datos. La neurona se activará si el resultado es superior a un determinado límite con el objetivo de comunicarse con otras neuronas (Hernández, Ramírez, & Ferri, 2004).

Esta técnica posee dos tipos de aprendizaje, ya que en primera instancia establece un aprendizaje supervisado que proporciona un conjunto de datos de entrada y la respuesta sirve para una regresión y clasificación. Por otro lado, en el aprendizaje no supervisado a la red, se provee de un conjunto de datos que debe auto-enseñarse para generar una respuesta, que sea útil en una tarea de agrupamiento.

7. Basadas en núcleos y máquinas de soporte vectorial.

Esta técnica intenta maximizar el margen entre los grupos o las clases formales, para lo cual se basan en transformaciones que pueden aumentar la dimensionalidad, llamados núcleos o kernels (Galán, 2015).

2.3 La metodología CRISP-DM

La metodología CRISP-DM surgió en 1999 en la Unión Europea conformada por un conjunto de empresas de diferentes sectores industriales que aplicaron sus experiencias prácticas para establecer una metodología standard para los proyectos de minería de datos (Niño, 2016).

Esta metodología es considerada una de las más completas de acuerdo a su flexibilidad para adaptarse a los proyectos de minería de datos, ya que en sus fases establece como tarea entender el negocio, recolectar los datos, explorarlos y

analizarlos para posteriormente obtener los modelos, evaluarlos y llegar a los resultados.

Para tener un mayor entendimiento, a continuación se muestra la imagen que describe las fases de esta metodología.

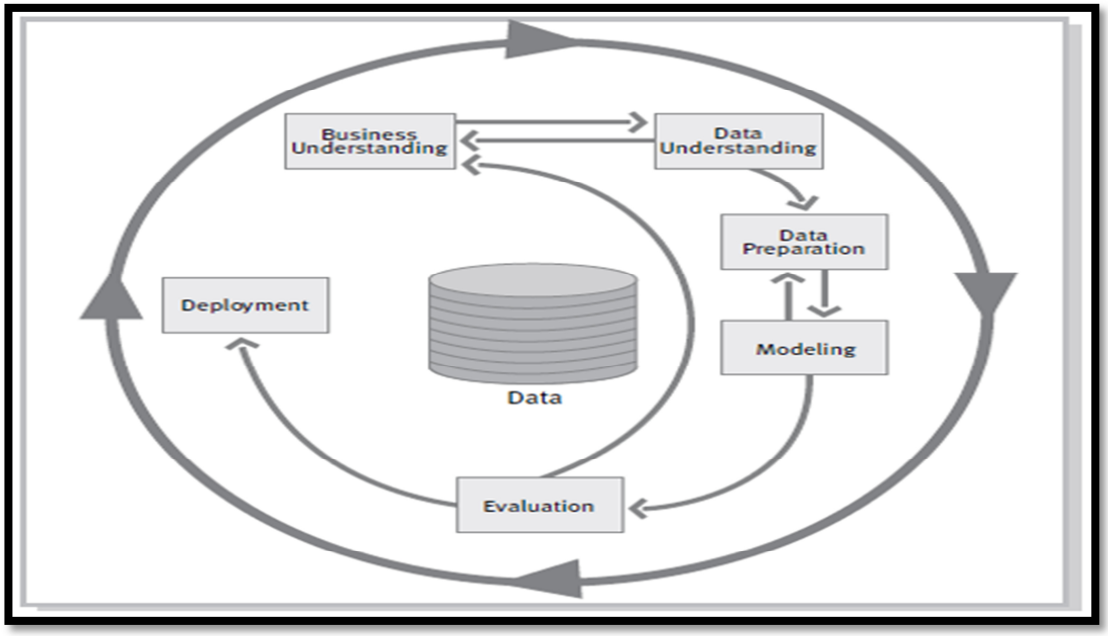

Figura 3. Fases de la metodología CRISP-DM
Fuente: (Chapman, y otros, 2016)

A continuación, se analizan en profundidad las tareas que se deben realizar en cada fase.

2.3.1 Comprensión del negocio o problema

De acuerdo con la metodología CRISP-DM, en esta fase se debe comprender el negocio para entender los procesos y tener una idea clara antes de definir los objetivos del negocio (Chapman, y otros, 2016).

Para tener una idea general de las actividades que se deben realizar en esta fase, se muestra la Figura 4.

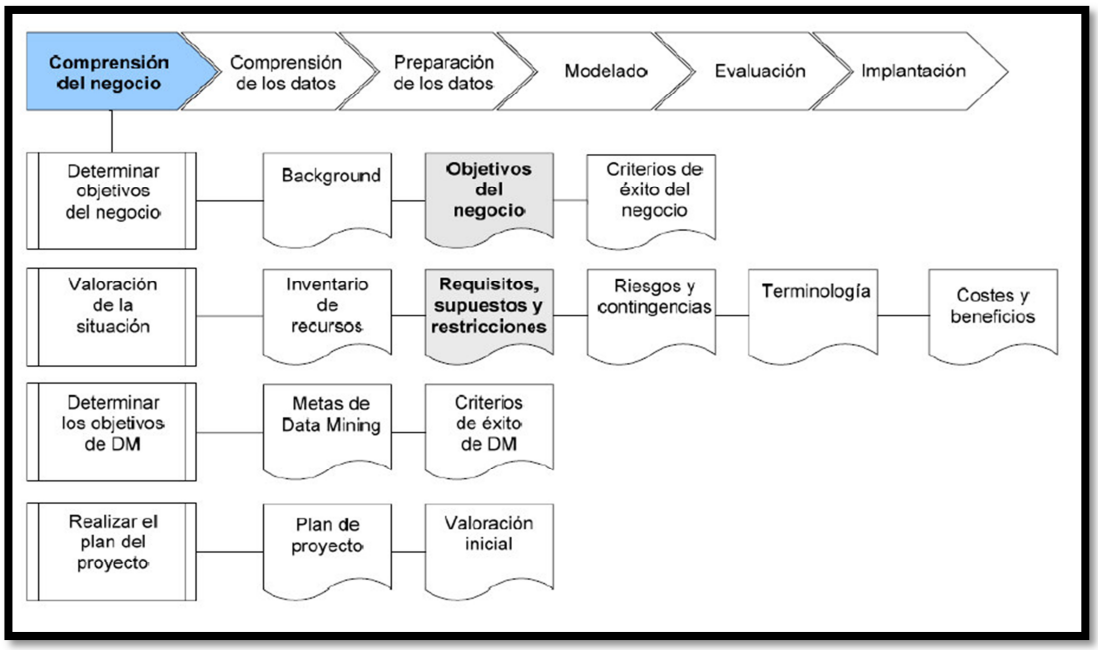

Figura 4. Fase de comprensión del negocio
Fuente: *(Yoshibauco, 2011)*

A continuación se hace una breve descripción de las tareas que se realizan en la fase de Comprensión del Negocio.

2.3.1.1 Determinar los objetivos del negocio.

En esta tarea se debe definir los objetivos que persigue la institución para resolver su problemática. Lo sustancial de esta fase es determinar al inicio del proyecto, los factores que pueden influir en el resultado, permitiendo conocer la situación de la institución, que no solo ayuda a identificar cuáles son los objetivos del negocio, sino que además contribuye a determinar los recursos de la organización que pueden usarse durante el desarrollo del proyecto (Chapman, y otros, 2016).

2.3.1.2 Valoración de la situación.

En esta tarea, los antecedentes y requisitos del problema deben ser evaluados tanto en términos de negocio como de minería de datos.

2.3.1.3 Determinar los objetivos de DM.

La metodología CRISP-DM establece que se debe definir los objetivos de minería de datos los cuales son muy segmentados, medibles y deben ayudar a resolver los objetivos definidos en el menor tiempo posible (Chapman, y otros, 2016).

2.3.1.4 Realizar el plan del proyecto.

En esta última tarea de la fase de comprensión del negocio, se debe definir un plan con el tiempo que tomará para ejecutarse cada una de las siguientes fases (Chapman, y otros, 2016).

2.3.2 Comprensión de los datos.

La metodología CRISP-DM establece que se deben entender los datos relacionados a los objetivos de negocio y de minería de datos para solicitar solo aquella información que sea útil para cumplir con los objetivos de la minería, y por ende llegar a la conclusión de cada uno de ellos.

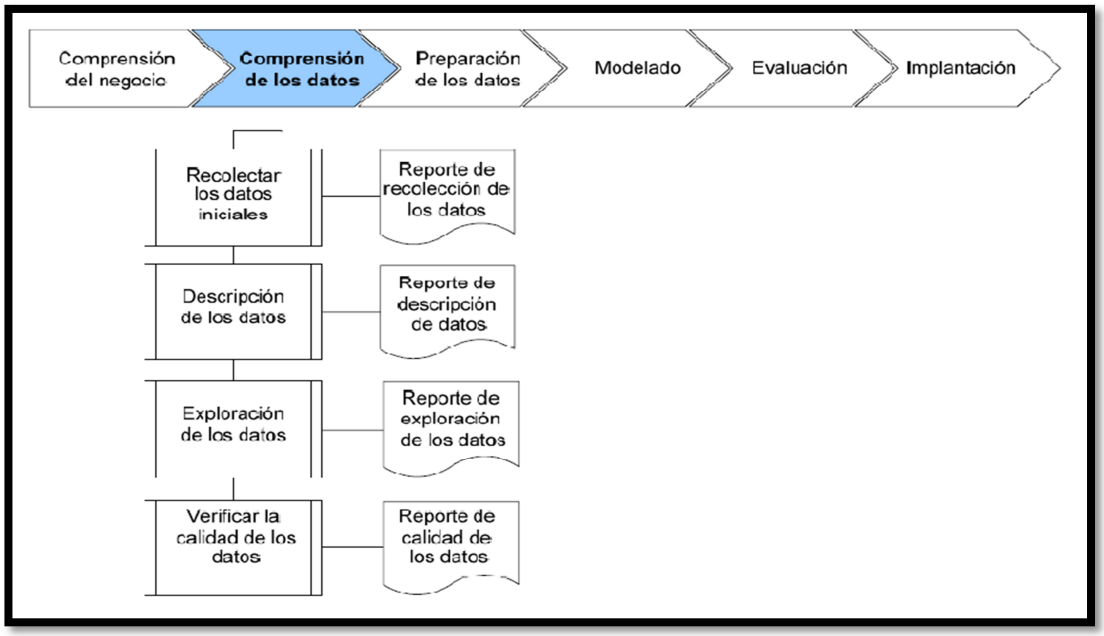

Figura 5. Fase de comprensión de los datos
Fuente: *(Yoshibauco, 2011)*

A continuación, se definen las tareas descritas por la metodología CRISP-DM, para llevar a cabo la comprensión de los datos (Chapman, y otros, 2016).

2.3.2.1 Recolectar datos iniciales.

Esta actividad tiene como objetivo que el investigador recolecte la información necesaria relacionada al problema y describa cómo realizó el proceso de recolección (Chapman, y otros, 2016).

2.3.2.2 Describir los datos.

Se deben describir los datos recolectados por parte del investigador como la funcionalidad de los campos, el tipo de datos y la identificación (Chapman, y otros, 2016).

2.3.2.3 Inspeccionar los datos.

Esta tarea tiene como finalidad que el investigador explore la información de la institución, utilizando técnicas estadísticas como: tablas de frecuencia, diagramas de caja, histogramas, entre otras, que ayuden a organizar los datos para que revelen datos relevantes para la investigación.

2.3.2.4 Verificar la calidad de los datos.

En esta última tarea, se debe analizar los datos con la finalidad de verificar la consistencia de la información. Esto significa que si encontramos valores nulos, datos anormales o que no pertenezcan al formato establecido, el investigador deberá tomar una decisión para que afecte en lo mínimo en los resultados obtenidos. (Chapman, y otros, 2016).

2.3.3 Preparación de los datos.

La preparación de los datos consiste en tener el conjunto de registros que serán utilizados para ser investigados. Esta fase es de extrema importancia debido a que

estos datos serán usados para realizar regresiones y/o entrenar modelos, por tal motivo, es importante que en esta fase se tenga extrema precaución. (Chapman, y otros, 2016).

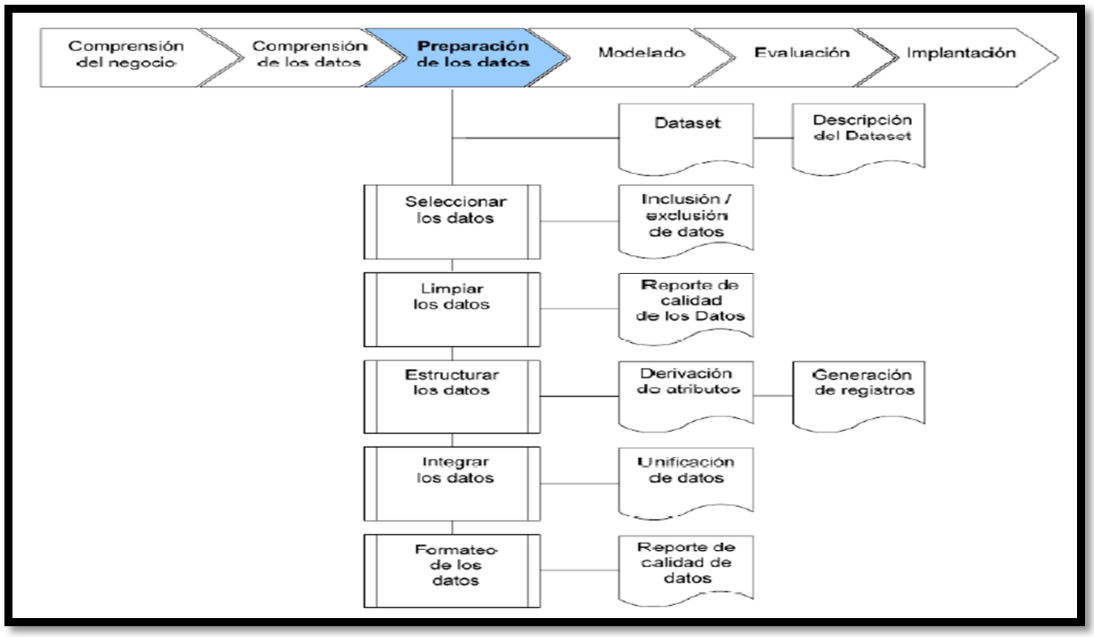

Figura 6. Fase de preparación de los datos
Fuente: *(Yoshibauco, 2011)*

A continuación, se definen los conceptos en cada una de las tareas dentro de esta fase de preparación de los datos.

2.3.3.1 Seleccionar los datos.

Se toma un conjunto de datos de los registros obtenidos para ser analizados.

2.3.3.2 Limpiar los datos.

En esta tarea se deben limpiar los datos, aquellos que ya fueron identificados, y que deben ignorarse para que no representen problemas al momento que los algoritmos comiencen el análisis de la información. Si se mantienen los datos nulos, posiblemente los resultados se vean afectados.

2.3.3.3 Estructurar los datos.

De acuerdo al a metodología CRISP-DM, se deben construir datos que sean de extrema importancia para la investigación, siempre que sea posibles, es decir, se deben formar datos a partir de los ya existentes.

2.3.3.4 Integrar los datos.

Una vez que se han descartado errores y se han construido los datos, estos se pasan a una tabla céntrica o resumen para la investigación. Puede elaborarse una o más tablas resúmenes dependiendo del caso, ya que el objetivo es facilitar el análisis de la información y los tiempos de procesamiento (Chapman, y otros, 2016).

2.3.3.5 Formatear los datos.

En caso de que los campos a ser analizados tengan problemas de formato, estos deben llevarse a un solo formato para evitar problemas de conversión, incluso en ciertos casos puede ocurrir que existan tabulaciones innecesarias o saltos de línea en los datos. En resumen, todos estos aspectos deben ser chequeados antes de proceder a aplicar los algoritmos (Chapman, y otros, 2016).

2.3.4 Modelado.

La metodología CRISP-DM sugiere al investigador, para construir un modelo que se ajuste al proyecto de minería de datos debe ser selectivo con la técnica a usarse, además que los requisitos que debe cumplir corresponden a los siguientes:

- Ser coherentes con el problema.
- Disponer de datos adecuados.
- Cumplir con los requisitos del problema.
- Tiempo adecuado para obtener el modelo.
- Conocimiento de la técnica.

En la siguiente gráfica de CRISP-DM se indican las actividades a realizar.

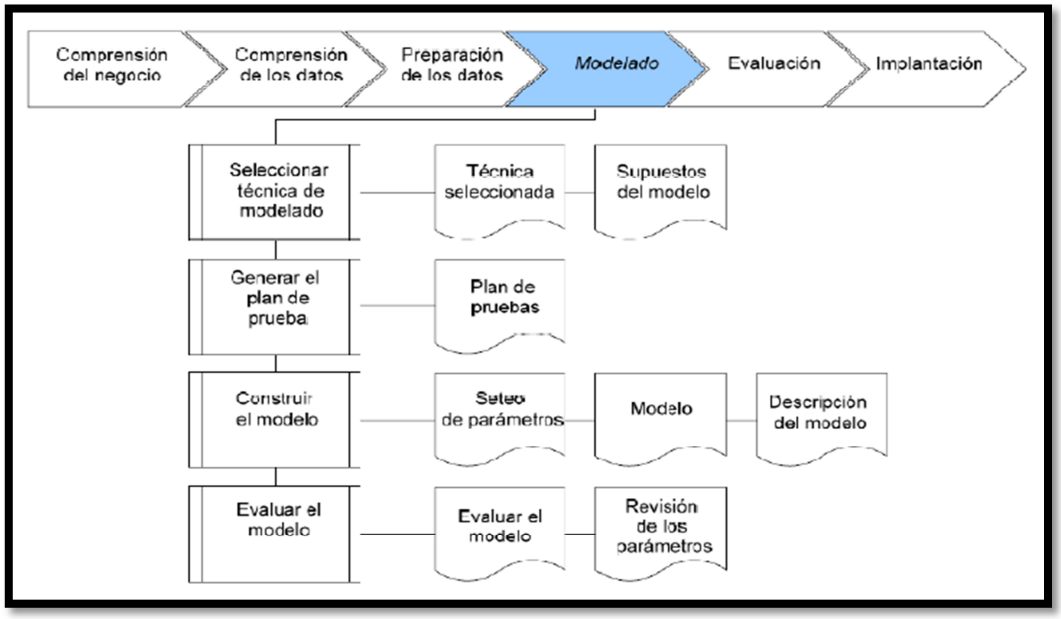

Figura 7. Fase de modelado
Fuente: *(Yoshibauco, 2011)*

A continuación se describen las actividades a realizarse en esta fase.

2.3.4.1 Seleccionar técnica del modelado.

Tal como lo sugiere la metodología CRISP-DM, al inicio de esta fase se debe seleccionar una técnica de modelado apropiada para el proyecto específico, incluso en algunos casos los investigadores seleccionan más de una técnica con el objetivo de compararlas y seleccionar la que mejor se adecue al problema.

2.3.4.2 Generar plan de prueba.

El plan de pruebas sirve para determinar la calidad y validez del modelo. Es aquí donde el investigador debe seleccionar las técnicas que debe aplicar para medir los errores generados por los modelos (Chapman, y otros, 2016).

2.3.4.3 Construir el modelo.

En esta actividad se construyen los modelos en base a las técnicas seleccionadas. Su selección se basa en la justificación de cada uno de ellos, es decir, el que tenga mejores resultados (Chapman, y otros, 2016).

2.3.4.4 Evaluar el modelo.

En esta etapa se evalúan los modelos obtenidos desde el punto de vista de los objetivos de la minería de datos establecidos para el proyecto (Chapman, y otros, 2016).

2.3.5 Evaluación.

En esta fase, la metodología CRISP-DM determina que deben evaluarse los objetivos de negocio en relación con los de minería de datos, verificando que se hayan cumplido y sean beneficiosos para la institución. Además, se debe describir de forma objetiva si los modelos obtenidos son útiles para su aplicación en la institución, y por ende, pasar a su explotación. En caso de que el investigador después de su análisis determine que los objetivos propuestos son imposibles de realizar, se debe redefinir los objetivos institucionales, y en consecuencia replantear los correspondientes a la minería de datos (Chapman, y otros, 2016).

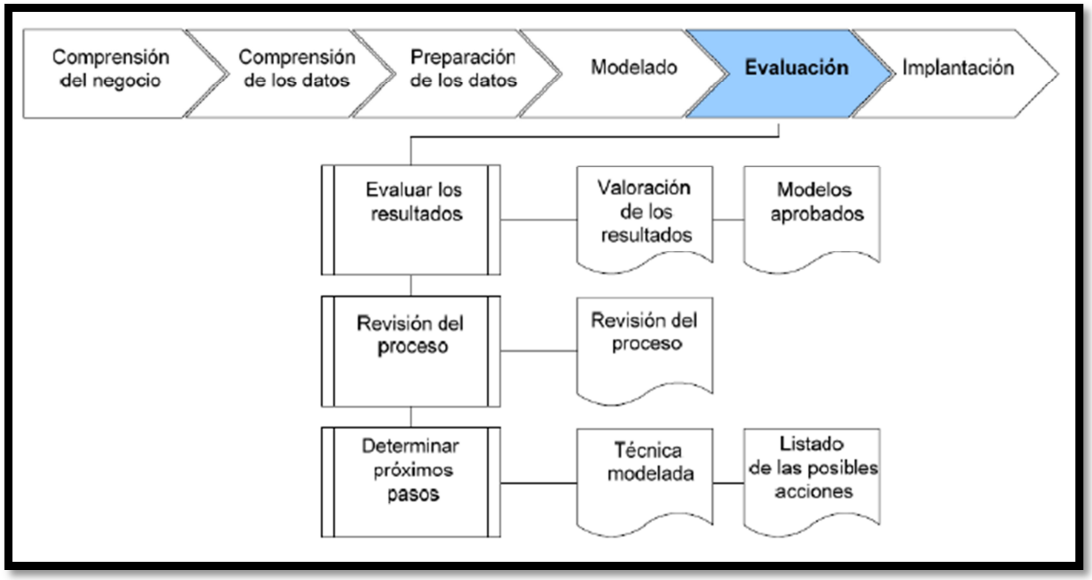

Figura 8. Fase de Evaluación
Fuente: (Chapman, y otros, 2016)

En esta fase, las actividades a realizarse son las siguientes:

2.3.5.1 Evaluar los resultados.

Se verifica que los modelos obtenidos hayan cumplidos con los criterios de éxito, conformados tanto por la minería de datos y los objetivos que persigue la institución (Chapman, y otros, 2016).

2.3.5.2 Revisión del proceso.

La metodología sugiere que debería revisarse si los procesos que se han empleado son los más adecuados, caso contrario sugiere mejorarlos (Chapman, y otros, 2016).

2.3.5.3 Determinar próximos pasos.

En dicha fase, el investigador debe optar por seguir o realizar una iteración sobre la preparación de los datos, siempre y cuando los resultados no sean satisfactorios para la investigación que se está llevando a cabo.

2.3.6 Implantación.

Una vez que el modelo ha sido probado y validado, los datos se transforman en conocimiento para la empresa. Posteriormente el investigador debe documentar y presentar los resultados obtenidos para que los administradores del negocio hagan el mejor uso de ellos. Adicionalmente se debe establecer las pautas de mantenimiento y monitoreo en caso de llegarse a implementar dicho modelo (Chapman, y otros, 2016).

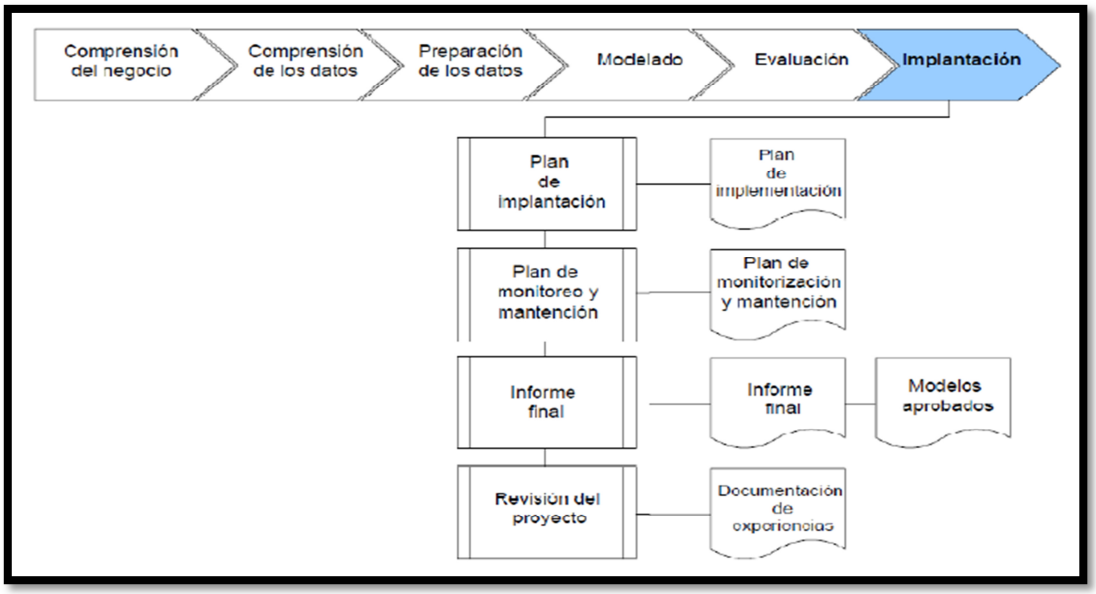

Figura 9. Fase de Implantación
Fuente: (Chapman, y otros, 2016)

Las tareas que se llevan a cabo en esta fase son las siguientes:

2.3.6.1 Plan de implantación.

Para llevar acabo la implantación del modelo de minería de datos en la empresa, el investigador en colaboración con la empresa debe planear una estrategia y obtener un documento donde se definan los pasos a seguir.

2.3.6.2 Plan de monitoreo y mantención.

En esta tarea, el investigador debe presentar un plan con las pautas de monitoreo del modelo y mantenimiento del mismo con el objetivo de prepararse y saber qué hacer en caso de que sea necesaria una intervención (Chapman, y otros, 2016)

2.3.6.3 Informe final.

El investigador debe presentar un resumen de los puntos más importantes de la investigación y la experiencia obtenida, y presentar un informe con los resultados obtenidos (Chapman, y otros, 2016).

CAPÍTULO III
METODOLOGÍA APLICADA

3.1 Uso de la Metodología CRISP-DM.

En este capítulo se aplica cada una de las fases de la metodología CRISP-DM en relación al problema planteado, para lo cual inicialmente se definen los objetivos desde el punto de minería de datos, los mismos que deben ser comprendidos, ya que al analizarlos permite obtener modelos y pasar a su evaluación. A continuación se describen las actividades a realizarse en cada punto:

3.2 Comprensión del negocio o problema.

Esta tarea es una de las más relevantes de la minería de datos, ya que es aquí donde se define los objetivos desde el punto de vista técnico (minería de datos), es decir, se busca obtener la mayor información de los objetivos comerciales, tarea que no resulta del todo sencilla, pero que permite reducir los riesgos, especificando problemas, objetivos y recursos. De esta manera, los objetivos que se determinan en esta sección deben ser medibles y resolverse lo más rápido posible.

3.2.1 Determinar los objetivos del negocio.

A continuación se presentan los objetivos comerciales de minería de datos:

- Desarrollar un modelo que determine el comportamiento de las tendencias de la adquisición de medicamentos oncológicos por parte del Hospital "Carlos Andrade Marín" a sus pacientes durante el año 2016.
- Desarrollar un modelo para predecir la tendencia del número de pacientes del Hospital "Carlos Andrade Marín" con perfil epidemiológico, que probablemente consumirán medicamentos oncológicos durante el año 2016.
- Predecir la cantidad óptima de adquisición de medicamento oncológico por parte del Hospital "Carlos Andrade Marín" durante el año 2016.

3.2.1 Valoración de la situación.

El Hospital "Carlos Andrade Marín" (HCAM) es una institución médica de tercer nivel administrada por el Instituto Ecuatoriano de Seguridad Social (IESS) al igual que otras casas de salud, donde la demanda es alta en comparación con la oferta que genera una serie de inconvenientes como la falta de camas, equipos, y medicamentos en el servicio de farmacia, provocando con ello malestar en los pacientes, ya que al no existir un stock de estos productos, los procesos de recuperación se retrasan afectando a su salud.

En este sentido y enfocándose en el problema de desabastecimiento de medicamentos es importante tomar en cuenta que en los dos últimos años se ha generado una reducción del 20% del presupuesto destinado a la compra de estos productos, debido en parte a que las compras son validadas a nivel jerárquico de la institución, donde no se prioriza este aspecto, y por tanto la adquisición de fármacos se ajusta al presupuesto asignado. Es por este contexto descrito, que esta investigación surge y por tanto a continuación se definen los objetivos de la minería de datos que serán enfocados al problema, pero de una forma integral para que puedan ser alcanzables y medibles.

3.2.2 Determinar los objetivos de DM.

Ahora que los objetivos comerciales han sido determinados, se definen los objetivos relativos a la minería de datos:

- Desarrollar un modelo que determine el comportamiento de las tendencias de la adquisición de medicamentos oncológicos, por ejemplo Filgrastim, por parte del Hospital "Carlos Andrade Marín" durante el año 2016.
- Desarrollar un modelo para predecir la tendencia del número de pacientes del Hospital "Carlos Andrade Marín" con perfil epidemiológico, que consumirán el medicamento oncológico Filgrastim durante el año 2016.
- Predecir la cantidad óptima de adquisición de medicamento oncológico Filgrastim por parte del Hospital "Carlos Andrade Marín" durante el año 2016.

3.2.3 Realizar el plan de proyecto.

A continuación se planea la realización de las siguientes actividades en los tiempos definidos:

- Comprensión de datos: 1 semana.
- Recolección de los datos: 4 semanas.
- Descripción de los datos: 1 semana.
- Exploración de los datos, verificarlos, limpiarlos y formatearlos: 6 semanas.
- Modelado: 8 semanas.
- Evaluación del modelo: 2 semanas.

3.2.4 Evaluación de la técnica y selección de la herramienta.

Para el análisis de los datos se tiene una serie de herramientas tales como:

- RapidMiner.
- Weka.
- R.
- Excel.
- SPSS.

Para el almacenamiento de la información se cuenta con un listado de las siguientes bases:

- Mysql.
- Postgresql.
- SqlServer.

Al analizar todas las herramientas y considerando que la institución donde se está aplicando dicha investigación pertenece al sector público, se utilizará la herramienta *opensource* R para el análisis de datos, además de Postgresql como gestor de base de datos. Es importante anotar que R es una poderosa herramienta estadística *open source* y además cuenta con una inmensa comunidad científica que la actualiza

constantemente, razón por la cual es la herramienta más adecuada para la realización de minería de datos.

Finalmente y una vez que se han seleccionado las herramientas necesarias para el trabajo, se debe escoger las técnicas con las cuales se harán las predicciones de los datos, además que debido a que R incorpora un gran cantidad de técnicas para análisis y predicción de datos, se ha procedido a utilizar la técnica de Regresión Lineal (LM) que permite formular un comportamiento a partir de conjunto de datos, ya que la técnica de Máquinas de Vector de Soporte (SVM) cumple el mismo objetivo que la técnica LM, no obstante la técnica SVM contribuye a generar un menor rango de error en sus predicciones.

3.3 Comprensión de datos.

En esta etapa se recolecta la información relacionada con los objetivos planteados de minería de datos. Aquí se recolectará, describirá, explorará y verificará la calidad de los datos antes de pasar a la fase de modelado.

3.3.1 Recolectar los datos iniciales.

La información que fue entregada por parte del Hospital "Carlos Andrade Marín" corresponde a archivos de Excel que contiene datos que se observan en las siguientes tablas:

Tabla 1.
Descripción de las tablas

Tabla	Descripción
MEDICAMENTO	Contiene los nombres de los medicamentos oncológicos.
DIAGNOSTICO_PACIENTE	Contiene el diagnóstico de los pacientes que se han hecho chequear en el Hospital "Carlos Andrade Marín".
MOVIMIENTO_ONCOLOGICO	Contiene los movimientos de los fármacos que se entregaron a los pacientes.

La información detallada está relacionada con los objetivos de la minería de datos que se desea resolver. En este sentido se va a describir cada una de ellas a continuación.

Diagrama relacional

En el siguiente diagrama se observa las relaciones de las entidades.

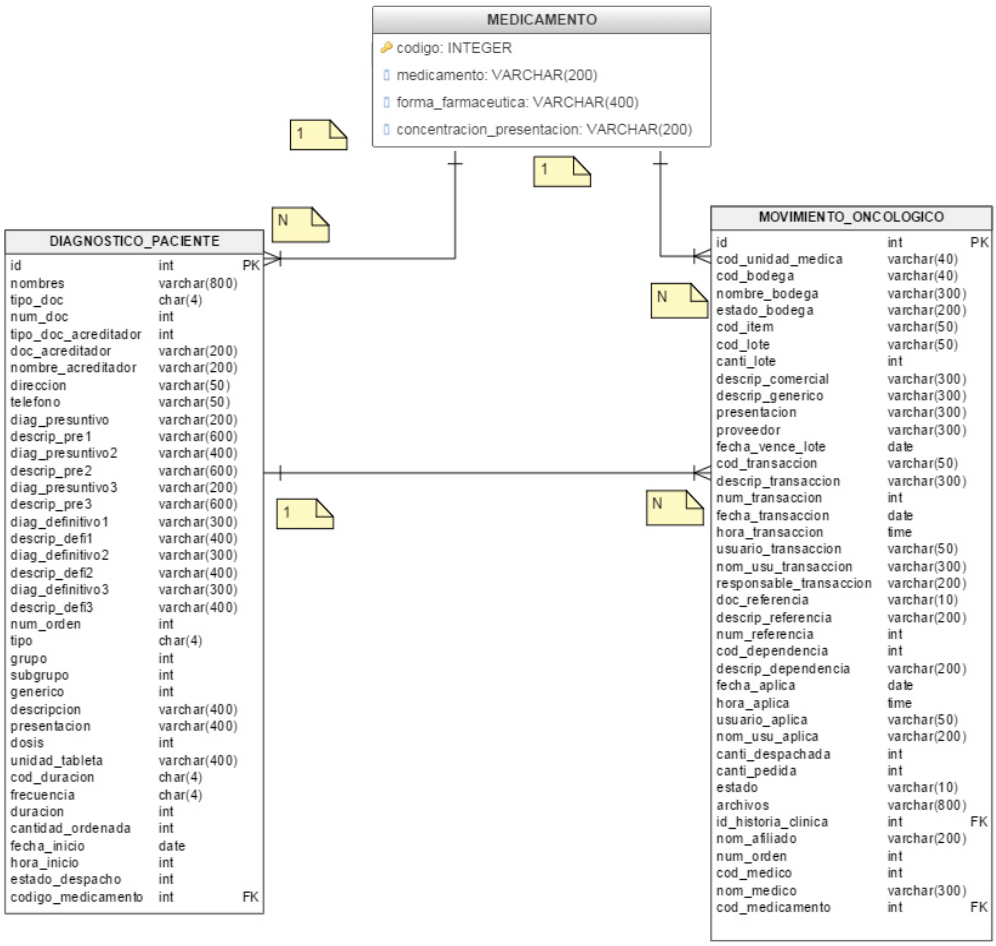

Figura 10. Diagrama de entidad relación

A continuación se expone las relaciones entre las distintas tablas:

- La relación entre “MEDICAMENTO y DIAGNOSTICO_PACIENTE” es de uno a muchos.
- La relación entre “MEDICAMENTO y MOVIMIENTO_ONCOLOGICO” es de uno a muchos.

- La relación entre "DIAGNOSTICO_PACIENTE y MOVIMIENTO_ONCOLOGICO" es de uno a muchos.

3.3.2 Describir los datos.

En esta etapa se describen los campos de las tablas involucradas en el proceso de investigación para tener una mayor comprensión de la información almacenada.

Tabla 2.
Tabla de Medicamento

MEDICAMENTO		
CAMPO	**TIPO DE DATO**	**DESCRIPCIÓN**
Código	Entero	Campo secuencial identificador de la tabla.
Medicamento	Cadena	Campo que almacena el nombre del medicamento.
forma_farmaceutica	Cadena	Campo que almacena la forma farmacéutica del medicamento.
concentracion_presentacion	Cadena	Campo que almacena la presentación del medicamento.

Tabla 3.
Tabla de Diagnóstico_Paciente

DIAGNÓSTICO_PACIENTE		
CAMPO	**TIPO DE DATO**	**DESCRIPCIÓN**
Id	Entero	Campo secuencial identificador de la tabla.
Nombre	Cadena	Campo que almacena el nombre completo del paciente.
tipo_doc	Entero	Campo que almacena la identificación del documento del

CONTINÚA

		paciente como cédula, pasaporte, etc.
num_doc	Cadena	Campo que almacena el número de documento.
tipo_doc_acreditador	Cadena	Campo que almacena el tipo de documento acreditado.
doc_acreditador	Cadena	Campo que almacena el nombre del documento acreditado.
nombre_acerditador	Cadena	Campo que almacena el nombre del acreditador.
Dirección	Cadena	Campo que almacena la dirección del paciente
Teléfono	Cadena	Campo que almacena el teléfono del paciente.
diag_presuntivo	Cadena	Campo que almacena el diagnóstico presuntivo del paciente.
descrip_pre1	Cadena	Campo que almacena la descripción del diagnóstico.
diag_presuntivo2	Cadena	Campo que almacena el diagnóstico presuntivo del paciente.
descrip_pre2	Cadena	Campo que almacena la descripción del diagnóstico.
diag_presuntivo3	Cadena	Campo que almacena el diagnóstico presuntivo del paciente.
descrip_pre3	Cadena	Campo que almacena la descripción del diagnóstico.
diag_definitivo1	Cadena	Campo que almacena el diagnóstico definitivo.
descrip_defi1	Cadena	Campo que almacena la

CONTINÚA

		descripción del diagnóstico definitivo.
diag_definitivo2	Cadena	Campo que almacena el diagnóstico definitivo.
descrip_defi2	Cadena	Campo que almacena la descripción del diagnóstico definitivo.
diag_definitivo3	Cadena	Campo que almacena el diagnóstico definitivo.
descrip_defi3	Cadena	Campo que almacena la descripción del diagnóstico definitivo.
num_orden	Entero	Campo que almacena el número de orden.
Tipo	Entero	Campo que almacena el tipo de paciente.
Grupo	Entero	Campo que almacena el grupo que pertenece el paciente.
Subgrupo	Entero	Campo que almacena el subgrupo que pertenece el paciente.
Generico	Entero	Campo que describe si el medicamento es genérico.
Descripcion	Cadena	Campo que almacena la descripción de la receta.
Presentacion	Cadena	Campo que almacena la presentación del medicamento.
Dosis	Entero	Campo que almacena la dosis del medicamento.
unidad_tableta	Cadena	Campo que almacena las unidades de tableta que se entrega al paciente.
cod_duracion	Cadena	Campo que almacena código de

CONTINÚA

		duración.
Frecuencia	Cadena	Campo que almacena la frecuencia.
Duración	Entero	Campo que almacena el tiempo en que deben terminarse las tabletas el paciente.
cantidad_ordenada	Entero	Campo que almacena la cantidad ordenada de medicamento.
fecha_inicio	Fecha	Campo que almacena la fecha de registro.
hora_inicio	Entero	Campo que almacena la hora de registro.
estado_despacho	Entero	Campo que almacena el estado del historial clínico del paciente.
codigo_medicamento	Entero	Campo que almacena el código del medicamento otorgado al paciente. Este se relaciona con la tabla medicamento.

Tabla 4.
Tabla de Movimiento_Oncologico

MOVIMIENTO_ONCOLOGICO		
CAMPO	**TIPO DE DATO**	**DESCRIPCIÓN**
Id	Entero	Campo secuencial identificador de la tabla.
cod_unidad_medica	Cadena	Campo que almacena el nombre del medicamento.
cod_bodega	Cadena	Campo que almacena el código de la bodega.
nombre_bodega	Cadena	Campo que contiene el nombre de la bodega.

CONTINÚA

estado_bodega	Cadena	Campo que contiene el estado de la bodega donde se encuentra el medicamento.
cod_item	Cadena	Campo que mantiene el código de barra del medicamento.
cod_lote	Cadena	Campo que almacena el código del lote de la bodega.
canti_lote	Entero	Campo que almacena la cantidad de elementos en el stock de la bodega.
descrip_comercial	Cadena	Campo que almacena la descripción comercial del medicamento.
descrip_generico	Cadena	Campo que almacena la descripción del medicamento genérico.
Presentación	Cadena	Campo que almacena la presentación del medicamento.
Proveedor	Cadena	Campo que almacena el nombre del proveedor del medicamento.
fecha_vence_lote	Fecha	Campo que almacena la fecha de vencimiento del lote.
cod_transaccion	Entero	Campo que almacena el código de la transacción.
descrip_transaccion	Cadena	Campo que almacena la descripción de la transacción.
num_transaccion	Entero	Campo que almacena el número de la transacción.
fecha_transaccion	Fecha	Campo que almacena la fecha de la transacción.
hora_transaccion	Entero	Campo que almacena la hora

CONTINÚA

		de transacción.
usuario_transaccion	Cadena	Campo que almacena la transacción del usuario.
nom_usu_transaccion	Cadena	Campo que almacena el nombre del usuario que realizó la transacción.
responsable_transaccion	Cadena	Campo que almacena el responsable de la transacción.
doc_referencia	Cadena	Campo que almacena el documento de referencia.
descrip_referencia	Cadena	Campo que almacena la descripción de la referencia.
num_referencia	Entero	Campo que almacena el número de referencia.
cod_dependencia	Entero	Campo que almacena el código de la dependencia.
descrip_dependencia	Cadena	Campo que almacena la descripción de la dependencia.
fecha_aplica	Fecha	Campo que almacena la fecha que aplica el movimiento.
hora_aplica	Entero	Campo que almacena la hora que aplica el movimiento.
usuario_aplica	Cadena	Campo que almacena el usuario que realiza el movimiento.
nom_usu_aplica	Cadena	Campo que almacena el nombre del usuario que realiza el movimiento.
canti_despachada	Entero	Campo que almacena la cantidad despachada.
canti_pedida	Entero	Campo que almacena la cantidad solicitada.

CONTINÚA

Estado	Cadena	Campo que almacena el estado del pedido.
Archivos	Cadena	Campo que almacena si el pedido se encuentra dentro del periodo activo.
id_historia_clinica	Entero	Campo que almacena el código de la historia clínica del paciente.
nom_afiliado	Cadena	Campo que almacena el nombre del afiliado, es decir el paciente.
num_orden	Entero	Campo que almacena el número de la orden.
cod_medico	Entero	Campo que almacena el código del médico.
nom_medico	Cadena	Campo que almacena el nombre del médico.
cod_medicamento	Entero	Campo que almacena el código del medicamento que el doctor ha recetado al paciente.

Una vez descrito los campos de las tablas proporcionadas por parte del hospital, se procede a seleccionar las tablas y los campos, en base a los objetivos de la investigación. En este caso, las tablas seleccionadas son:

- DIAGNOSTICO_PACIENTE
- MOVIMIENTO_ONCOLOGICO

A continuación se expone las tablas que van a servir para la investigación y se detalla cada uno de los campos escogidos para el análisis.

Tabla 5.
Tabla de Diagnostico_Paciente

DIAGNOSTICO_PACIENTE		
CAMPO	**TIPO DE DATO**	**DESCRIPCIÓN**
Id	Entero	Campo auto numérico de la tabla
id_historia_clinica	Entero	Campo que almacena el id de la historia clínica.
nombre_paciente	Cadena	Campo que almacena el nombre del paciente.
identificacion_paciente	Cadena	Campo que almacena la identificación del paciente.
descripcion_medicamento	Cadena	Campo que almacena la descripción del medicamento.
presentacion_medicamento	Cadena	Campo que almacena la descripción del medicamento.
cantidad_ordenada_medicamento	Entero	Campo que almacena la cantidad ordenada de medicamento por parte del paciente.
fecha_inicio	Fecha	Campo que almacena la fecha del registro.

Tabla 6.
Tabla de Movimiento_Oncologico

MOVIMIENTO_ONCOLÓGICO		
CAMPO	**TIPO DE DATO**	**DESCRIPCIÓN**
Id	Entero	Campo auto numérico de la tabla.
cod_unidad	Cadena	Campo que almacena el código de unidad del medicamento.
codigo_item	Cadena	Campo que almacena el código de barra del medicamento.

CONTINÚA

descripcion_medicamento	Cadena	Campo que contiene el nombre del medicamento.
presentacion_medicamento	Cadena	Campo que contiene la presentación del medicamento.
Proveedor	Cadena	Campo que almacena el nombre del proveedor del medicamento.
nombre_paciente	Cadena	Campo que almacena el nombre del afiliado o paciente a quién se otorgó el medicamento.
cantidad_ordenada_medicamento	Entero	Campo que almacena la cantidad del medicamento que se entregó al paciente.
fecha_movimiento	Fecha	Campo que almacena la fecha de movimiento del medicamento.

Las tablas 6 y 7 con sus respectivos campos fueron diseñadas a partir de la información proporcionada por el hospital, y son las que se utilizarán para el análisis de datos. Por tanto se necesita exportar los datos de Excel a la base de datos PostgreSQL a través de la herramienta de exportación con los campos deseados.

3.3.3 Inspeccionar los datos.

En esta sección se exploran los datos, situación para lo cual se aplican pruebas estadísticas básicas que revelan sus propiedades, ya que al crear gráficos de frecuencia y distribución, se desea tener la noción de que sucede con esta información. Para llevar a cabo esta actividad se realizan gráficas de la información de las tablas que interesan y que corresponden a: **MOVIMIENTO_ONCOLOGICO** y **DIAGNOSTICO_PACIENTE.**

En las figuras de la 11 a la 17 se resume la información básica referente al movimiento del medicamento Filgrastim.

Pacientes atendidos mensualmente que requirieron Filgrastim.

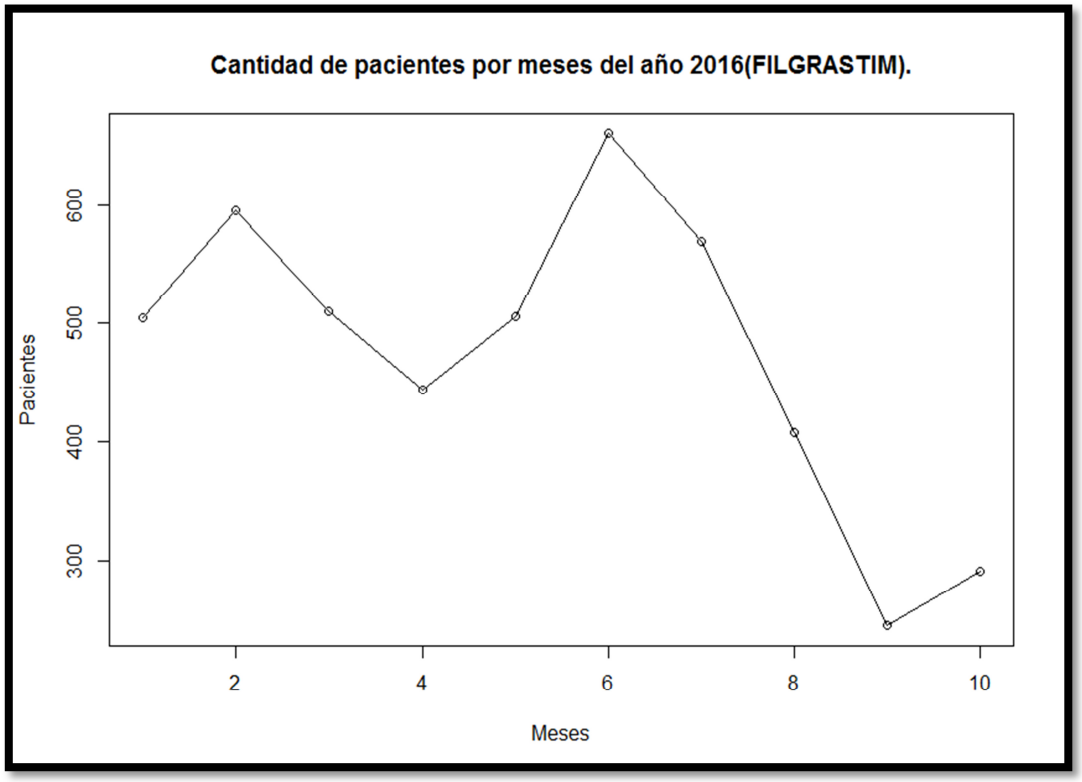

Figura 11. Pacientes atendidos por mes del año 2016 (Filgrastim)

En la Figura 11 se observa que el año 2016 inicia con una fuerte demanda de 500 personas solicitando este medicamento. Luego se incrementa a 600 personas para el mes de febrero; sin embargo hay un descenso en marzo, mientras que en julio existe una aceleración de la demanda del medicamento, que para agosto en adelante desacelera (Ver anexo 1, literal A.)

En base a esta observación se deduce que los meses de mayor demanda de este medicamento son: enero, febrero, junio y julio. Y asimismo, los meses con menor demanda son: marzo, abril, mayo, agosto, septiembre, octubre, noviembre y diciembre.

Total de medicamentos solicitados por mes del año 2016

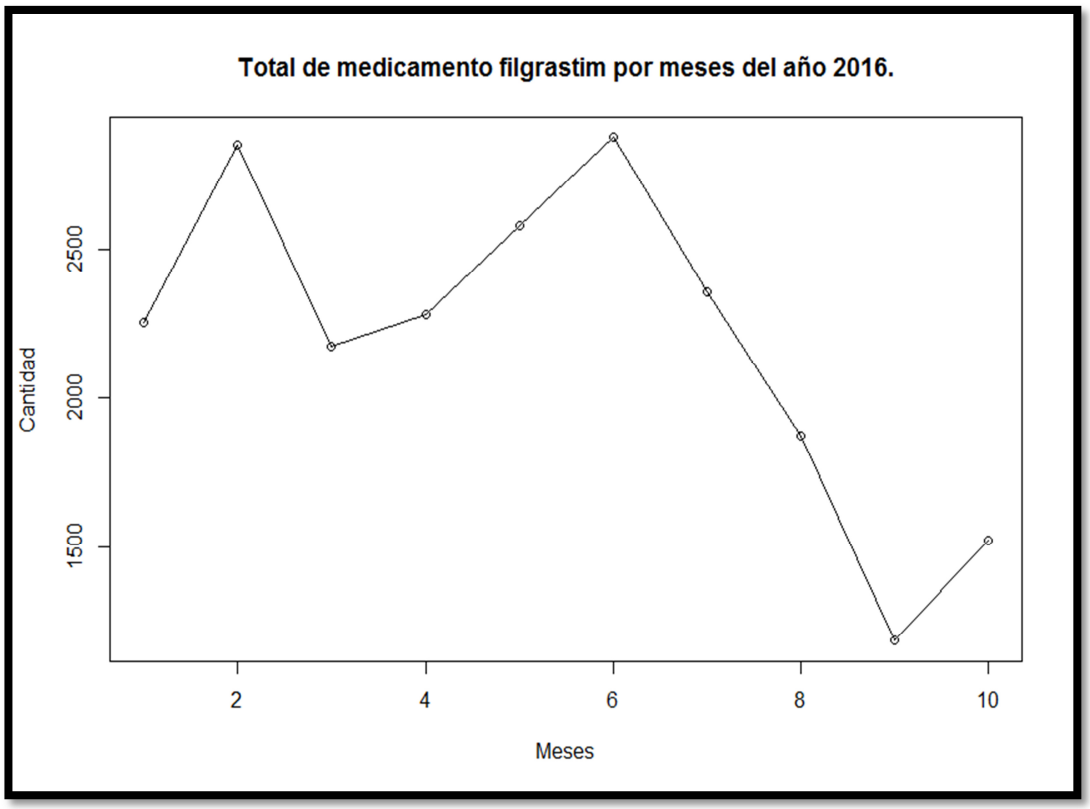

Figura 12. Total de medicamento de Filgrastim por meses del año 2016

La Figura 12 muestra que febrero es un mes con mayor cantidad de demanda del medicamento, tal como se observaba en la Figura 11. Luego en junio y julio se vuelve a observar una aceleración en la demanda que supera las 2500 unidades de medicamento de Filgrastim, pero a finales de año vuelve a caer la demanda. De esta manera se observa que el comportamiento es brusco al inicio y a mediados de año, mientras que a finales de año, el proceso de demanda baja bruscamente frenando los pedidos. (Ver anexo 1, literal B.)

Promedio Mensual de Filgrastim

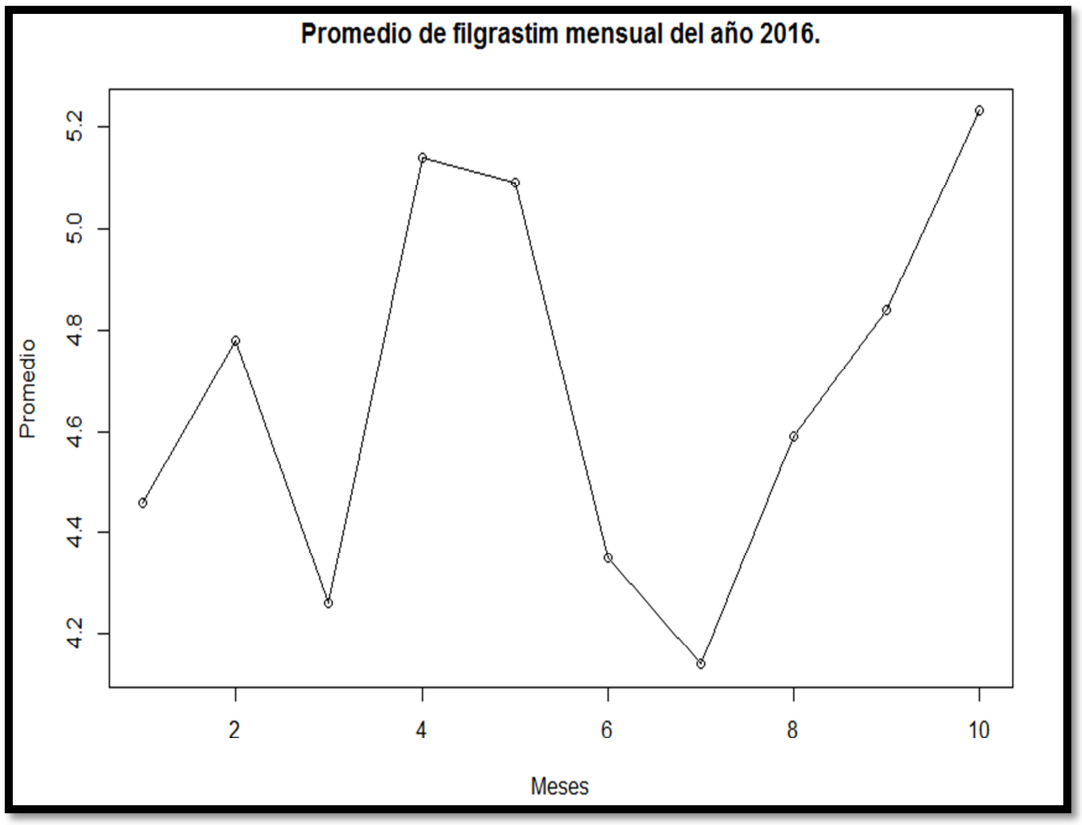

Figura 13. Promedio Mensual de Medicamento Filgrastim del año 2016.

Si se observa la Figura 13, se nota que los meses en que las personas adquieren más de cinco medicamentos se encuentran entre: abril, mayo, septiembre, octubre, noviembre y diciembre. Mientras que en enero, febrero, marzo, junio y julio, las personas acceden a menos de cinco medicamentos. (Ver anexo 1, literal C.)

Histograma de cantidades solicitadas de Filgrastim

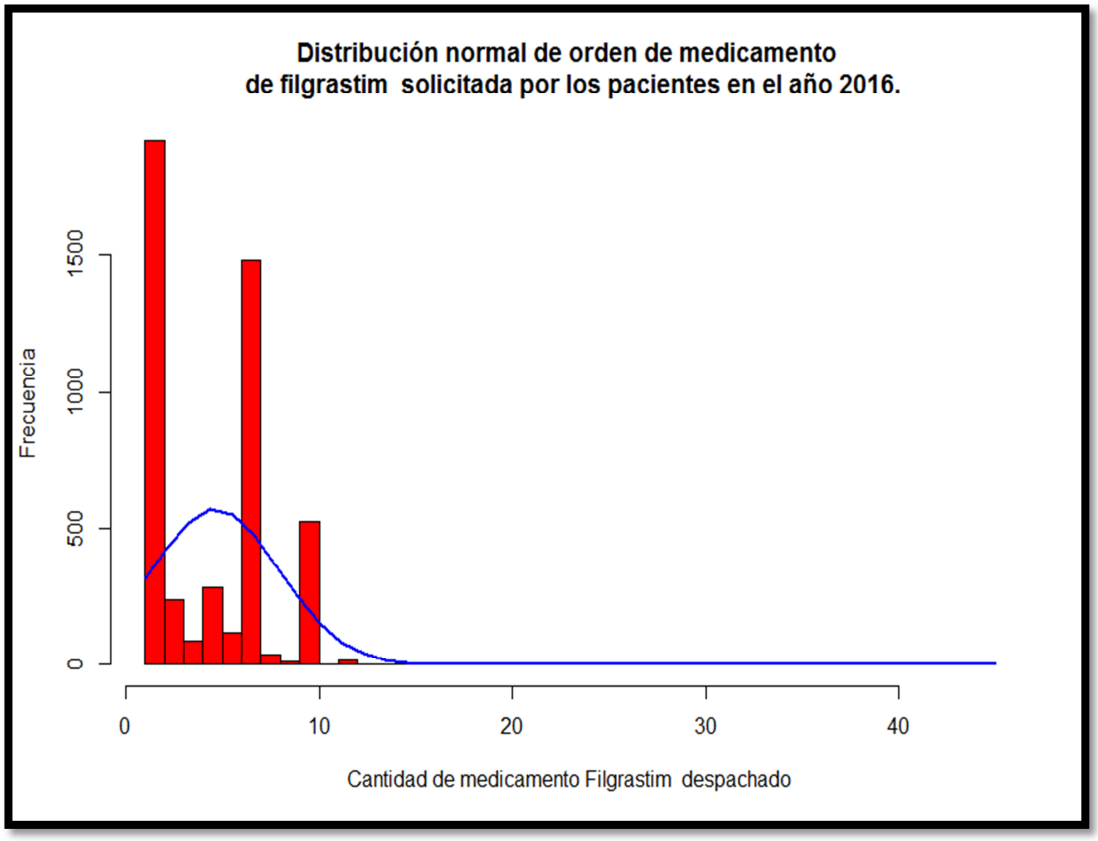

Figura 14. Histograma con distribución normal de orden de medicamento de Filgrastim solicitada por los pacientes en el año 2016.

En la Figura 14, se observa que existe una columna que supera a todas y que indica que constan más de 1500 pedidos de Filgrastim de una unidad por paciente. Mientras que en segundo lugar se observa que entre 1000 y 1500 solicitudes han llevado como pedido entre 7 y 7,5 unidades de Filgrastim. Finalmente la cantidad que ha llevado 10 unidades de Filgrastim se sitúa en menos de 500 pedidos. (Ver anexo 1, literal D.)

Si se observa detenidamente el histograma existe una línea azul que representa una distribución normal de los medicamentos, y que indica que en promedio los pacientes realizan frecuentemente pedidos de cinco unidades a nivel nacional.

Cantidad de pedido de Filgrastim durante el 2016.

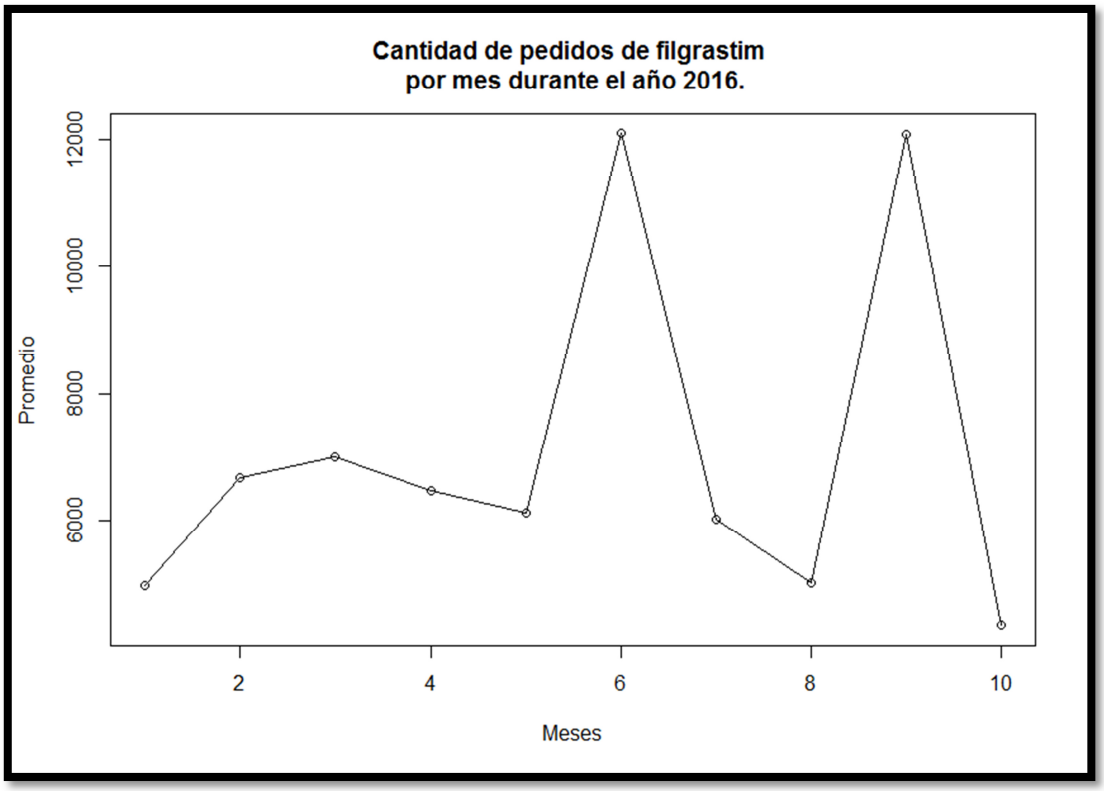

Figura 15. Cantidad de pedidos de Filgrastim por mes durante el año 2016.

En la Figura 15 se analiza la tendencia anual de la cantidad promedio de pedidos de Filgrastim por mes. Al observar la gráfica, se aprecia una tendencia baja que va desde enero hasta mayo, y que en su tope más alto alcanza aproximadamente los 7.000 pedidos del medicamento. Luego la demanda de este medicamento sube para el mes de junio, luego baja en los dos meses siguientes, y vuelve a subir la demanda del producto para el mes de noviembre, mientras que para diciembre vuelve a disminuir. (Ver anexo 1, literal E.)

Histograma de Filgrastim del año 2016

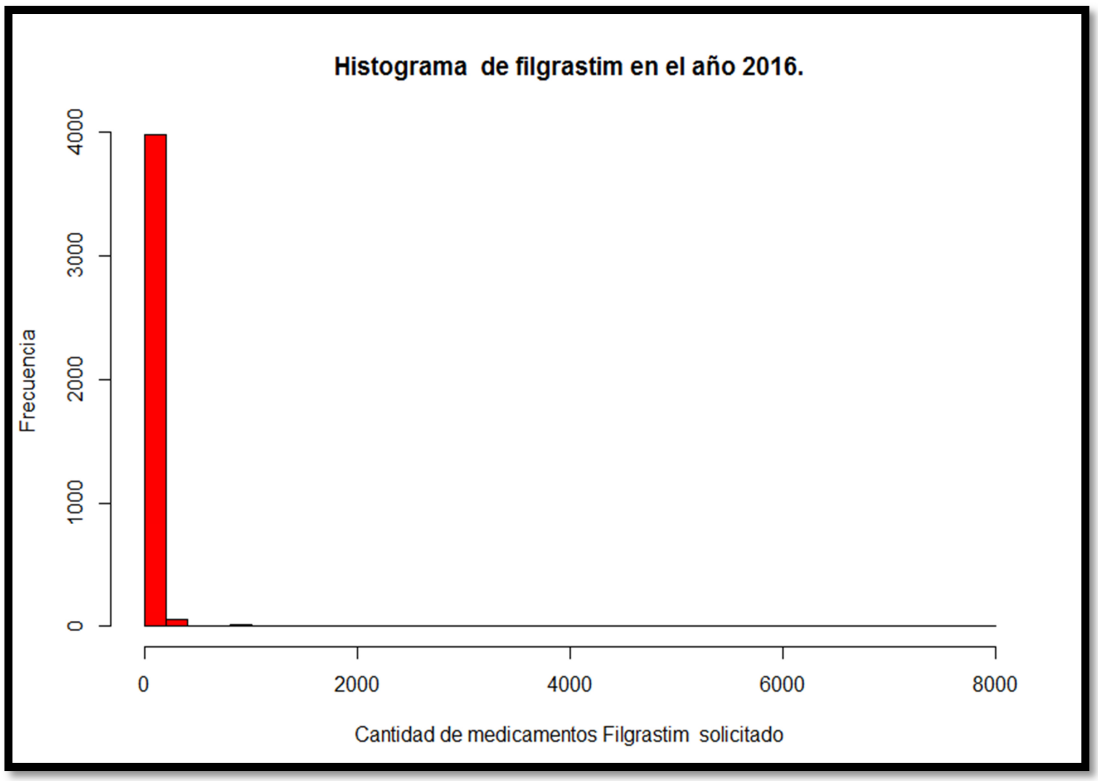

Figura 16. Histograma de Filgrastim en el año 2016.

En este histograma de la Figura 16, se observa que hay una columna inicial que predomina sobre el resto que son prácticamente nulas. Al analizar los datos, se establece que aproximadamente 4000 veces se pidieron entre 1 y 5 unidades del medicamento Filgrastim. Con esta información se deduce que la mayoría de las personas solicitan entre 1 y 5 medicamentos durante los meses de enero, junio y octubre tal como lo muestra la Figura 15. (Ver anexo 1, literal F.)

3.3.4 Verificar la calidad de los datos.

En esta sección se verifica la calidad de los datos para tener una noción de la calidad de la data obtenida, es decir, se desea detectar los "outliers" que influyen en los resultados de los modelos, y precisamente por esta razón se identifica los datos

influyentes, cuya representación permitirá tomar una decisión para ignorarlos o tomarlos en cuenta para el desarrollo de los modelos.

A continuación se exponen los diagramas de caja:

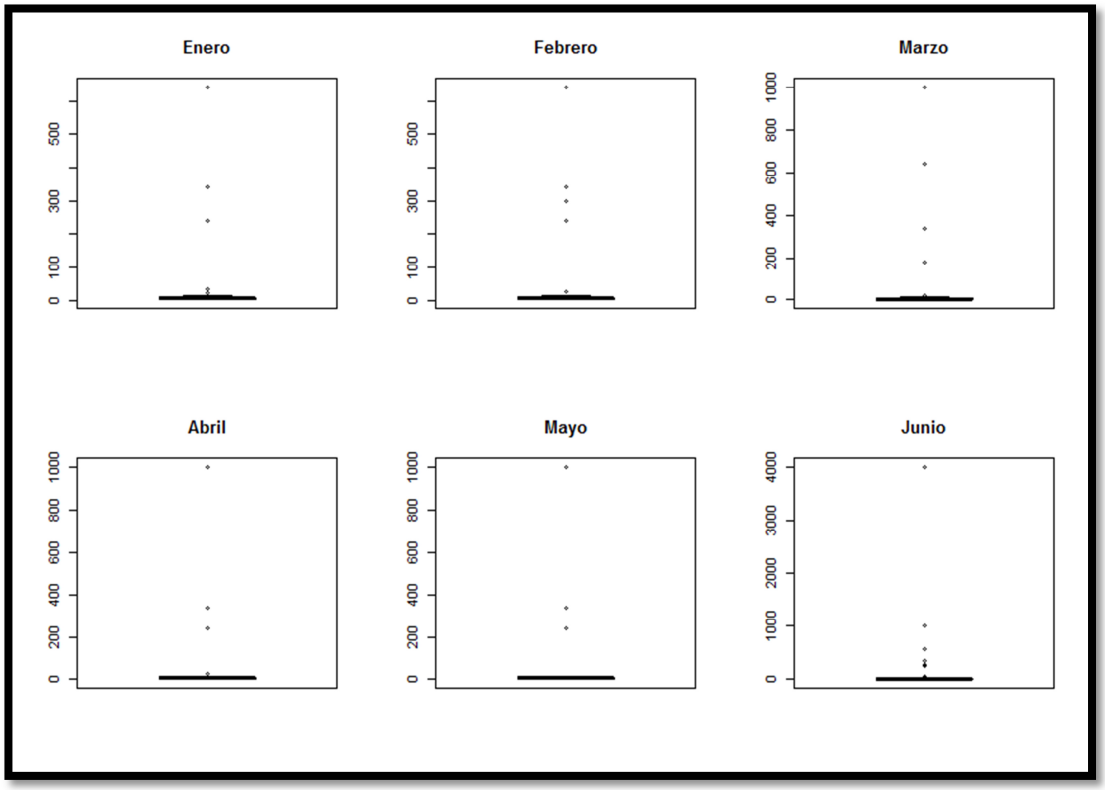

Figura 17. Diagrama de cajas de Movimientos Oncológicos de ene-jun del 2016

En la Figura 17 se nota que en enero y en el resto de meses existen pedidos que van desde 200 en adelante, y se realizan de forma esporádica. Precisamente estos registros son los que afectan a la distribución de los datos, y que están presentes en todos los meses afectando a los resultados. Esta información significa que si en enero se pidió un promedio de 5 a 6 medicamentos de Filgrastim, a causa de otro pedido de 200 en adelante, estos datos afectan los resultados, razón por la cual se debería tomar en cuenta para el análisis de los modelos, o simplemente ignorarlos para que no altere los resultados finales. (Ver anexo 2, literal A.)

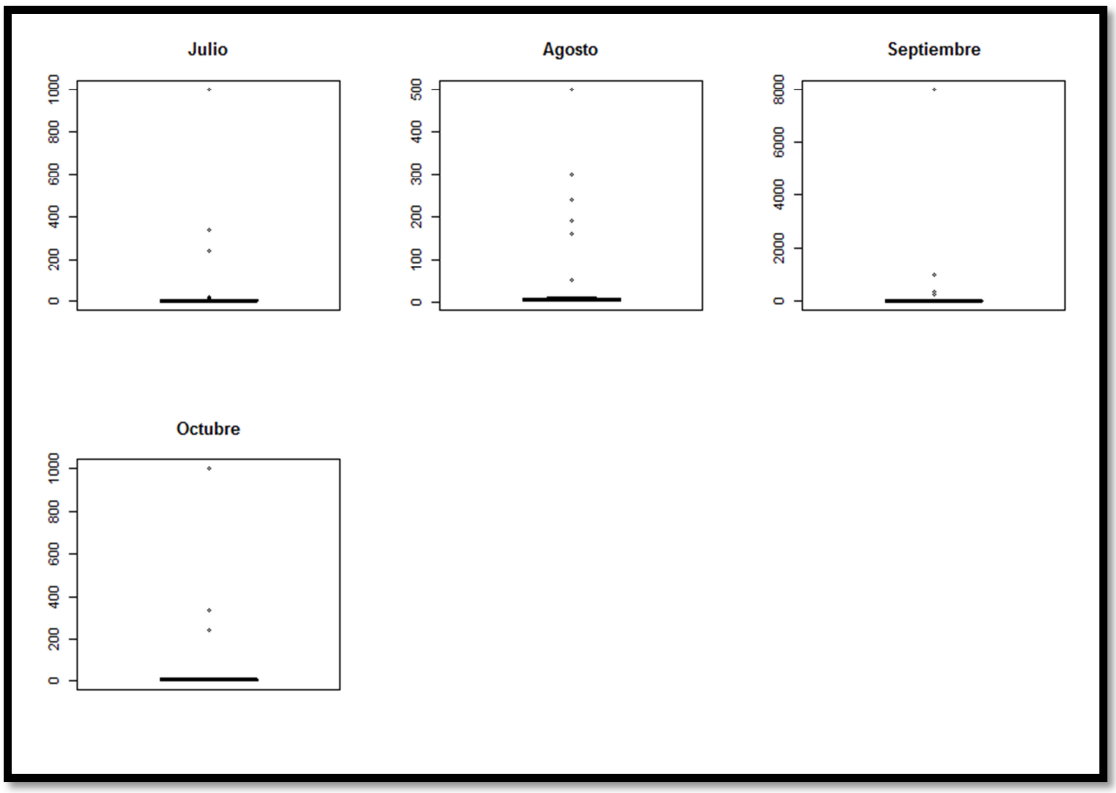

Figura 18. Diagrama de cajas de Movimientos Oncológicos de jul-oct del 2016

En la figura 18 se observa la misma tendencia de pedidos, mientras que los picos que se observaban en las figuras 15 y 16 se reflejan en los diagramas de caja, es decir, se trata del mismo comportamiento de la Figura 17. (Ver anexo 2, literal A.)

Al constatar que en la Figura 17 y 18 existen datos atípicos, se realiza una representación gráfica de los datos que se están analizando.

Para la obtención de porcentajes de datos outliers, se aplica algoritmos de R para identificar cuál es la cantidad de datos que están fuera del rango, y después se toma en cuenta el total de registros, transformándolos en porcentajes para su representación.

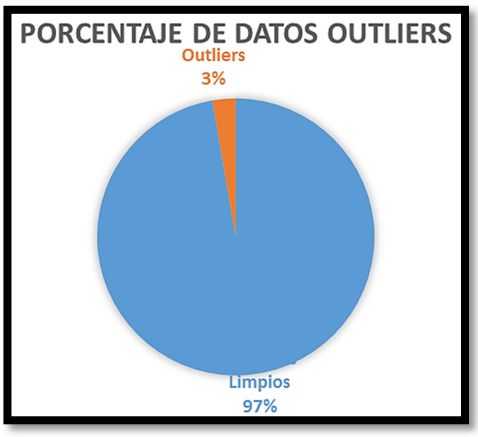

Figura 19. Representación de datos outliers sobre el universo de datos

En la Figura 19, se observa que los datos atipicos representan el 3% del conjunto de datos, y por lo tanto se debe tomar una decisión si son o no considerados (Ver anexo 2 Literal B). En el caso de que los datos atípicos sean considerados implicaría alteraciones en la predicción, es decir, el modelo que se va a construir se vería directamente afectado. Por esta razón se toma la decisión de excluir los datos atípicos por ser un porcentaje mínimo. De esta manera y una vez que se encontró y separaron los datos atipicos del universo se procedió a regenerar los diagramas de caja para establecer los resultados finales.

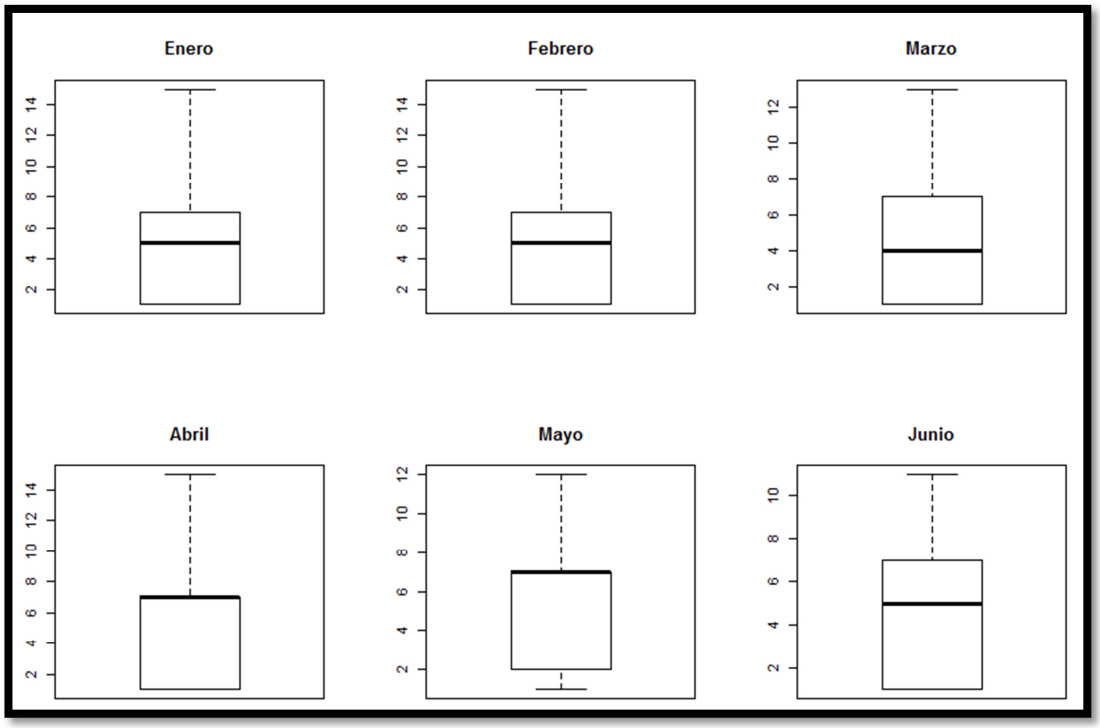

Figura 20. Corrección de datos outlier – Enero a Junio

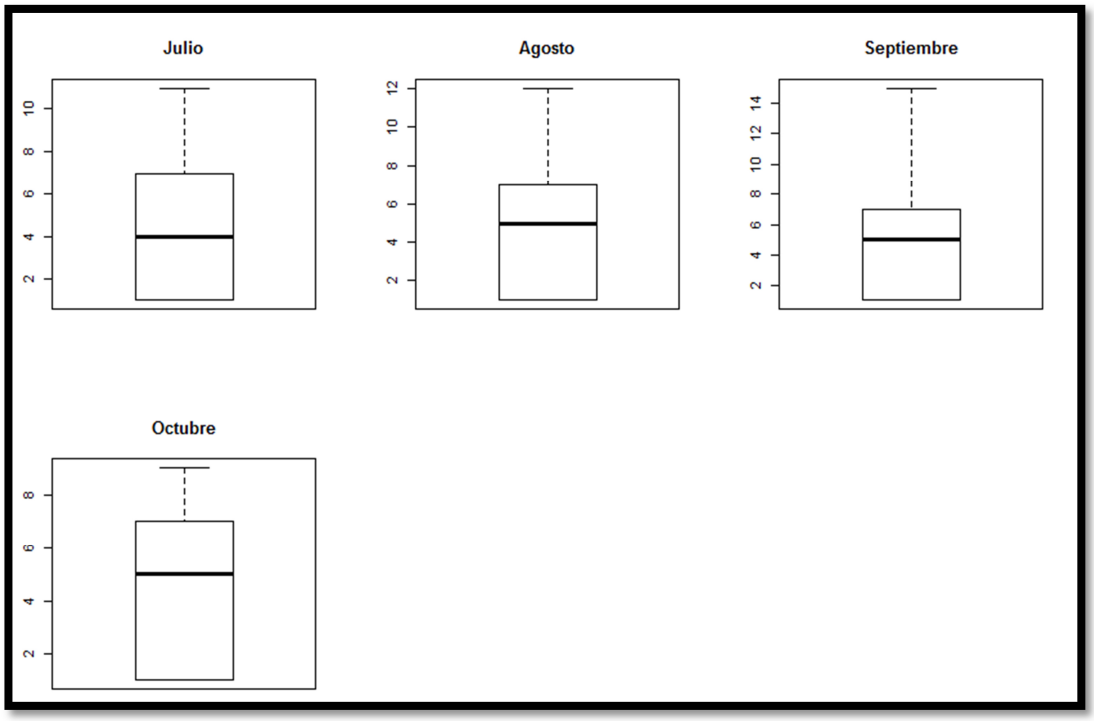

Figura 21. Corrección de datos Outlier – Julio a Octubre

Al observar la Figura 20 y 21, se evidencia que los datos atípicos desaparecen y queda normalizado los datos de los diagramas de caja. (Ver anexo 2, literal A.)

3.4 Preparación de los datos.

Esta sección se enfoca en preparar los datos, es decir, seleccionar aquellos que serán analizados para luego filtrarlos por un proceso de limpieza; no obstante si aparecieran anomalías de integridad se procederá a construir tales datos y definirlos en un mismo formato.

3.4.1 Selección de los datos.

En esta sección, la metodología sugiere seleccionar el conjunto de datos a ser analizados para entrenar los modelos. En este caso, se trabajará con toda la información de la muestra obtenida, es decir el 100% de los datos obtenidos.

A continuación se muestra los atributos de la tabla MOVIMIENTO_ONCOLÓGICO:

Tabla 7.

Atributos de la tabla Movimiento_Oncológico

MOVIMIENTO_ONCOLOGICO
id
cod_unidad
código_item
descripción_medicamento
presentación_medicamento
proveedor
fecha_movimiento
cantidad
nombre_afiliado

3.4.2 Limpiar los datos.

Debido a que los datos que fueron transferidos no presentan anomalías como tipos de datos diferentes a los que deberían estar en la columna de las tablas, se considera que no es necesario realizar una revisión o limpieza de los mismos.

3.4.3 Estructurar los datos.

Atributos derivados: Al no existir ningún tipo de transformación de los campos, no existe la necesidad de crear un campo a partir de los existentes, porque la información proporcionada es necesaria para llevar a cabo el cumplimiento del objetivo propuesto, razón por la cual no es necesario crear un campo derivativo.

Registros generados: En este apartado y tomando en cuenta que los datos que serán analizados son los necesarios, no existe motivo para crear o completar registros con nuevos atributos.

3.4.4 Integrar datos.

Con el objetivo de tener datos centralizados y describir su tipo y atributo para que al momento de analizarlos no existan errores, a continuación se detallan los datos de la tabla creada “MOVIMIENTO_ONCOLÓGICO”.

Tabla 8.
Tabla de Movimiento_Oncológico

MOVIMIENTO_ONCOLÓGICO	
CAMPOS	**TIPO DE DATO**
Id	Numérico
cod_unidad	Texto
codigo_item	Texto
descripción_medicamento	Texto
presentación_medicamento	Texto
proveedor	Texto
fecha_movimiento	Fecha
cantidad	Numérico
nombre_afiliado	Texto

3.4.5 Formatear datos.

De acuerdo a la metodología CRISP-DM, cuando los datos que se proporcionan para la investigación contemplan valores fuera de su contexto, es necesario proveer de un formato a los mismos, es decir, si se tiene un campo numérico donde se encuentran caracteres que lo representen, entonces se considera que hay que formatear la columna. En este caso particular, los datos de los campos de los archivos son apropiados porque los datos que se tienen son los adecuados para cada una de las columnas, motivo por el cual no se tuvo que formatear ninguno de los campos.

3.5 Modelado.

Esta sección de modelado es la fase más relevante de la metodología CRISP-DM porque consiste en seleccionar la técnica de modelado, seleccionar la prueba, ejecutarla y verificar sus resultados, comparando cada modelo entre sí.

3.5.1 Seleccionar técnica de modelado.

Este ítem se enfoca en seleccionar los modelos que serán aplicados para cumplir con los objetivos definidos; cabe recalcar que la metodología establece mínimo dos modelos para compararlos entre sí, y seleccionar el que genere menor cantidad de errores.

Para el propósito de esta investigación se ha seleccionado los siguientes modelos:

- LM: Modelo de regresión lineal.
- SVM: Máquina de vector de soporte.

3.5.2 Generar el plan de prueba.

Medir los errores de las pruebas generadas siempre ha sido un trabajo difícil pero a la vez beneficioso para saber el modelo por el cual se debe optar, y una vez que se lo haya seleccionado éste cuente con la menor cantidad de errores posibles. Para la medición de los errores en cada modelo se utilizará la raíz del error cuadrático medio RMSE (GEO, 2015) y el error absoluto medio MAE (Julia, 2017).

- Error cuadrático medio.

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(y_j - \hat{y}_j\right)^2}$$

Figura 22. Fórmula del error cuadrático medio

Esta fórmula se basa en obtener la diferencia entre el valor real y el obtenido por el modelo para luego elevarlo al cuadrado y dividirlo para el número de muestras, y del resultado final se extrae la raíz cuadrada.

- Error absoluto medio

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

Figura 23. Fórmula del error absoluto medio

El error absoluto medio consiste en medir las diferencias del valor absoluto entre el valor real y el obtenido por el modelo para luego dividirlo por el número de muestras.

3.5.3 Construir el modelo.

El objetivo de esta fase es resolver cada uno de los objetivos de minería de datos planteados, es decir, se probará los modelos seleccionados para posteriormente evaluarlos.

- **Objetivo 1.** Desarrollar un modelo que determine el comportamiento de las tendencias de la adquisición de medicamentos oncológicos, por ejemplo Filgrastim, por parte del Hospital "Carlos Andrade Marín", durante el año 2016.

Modelo de regresión lineal

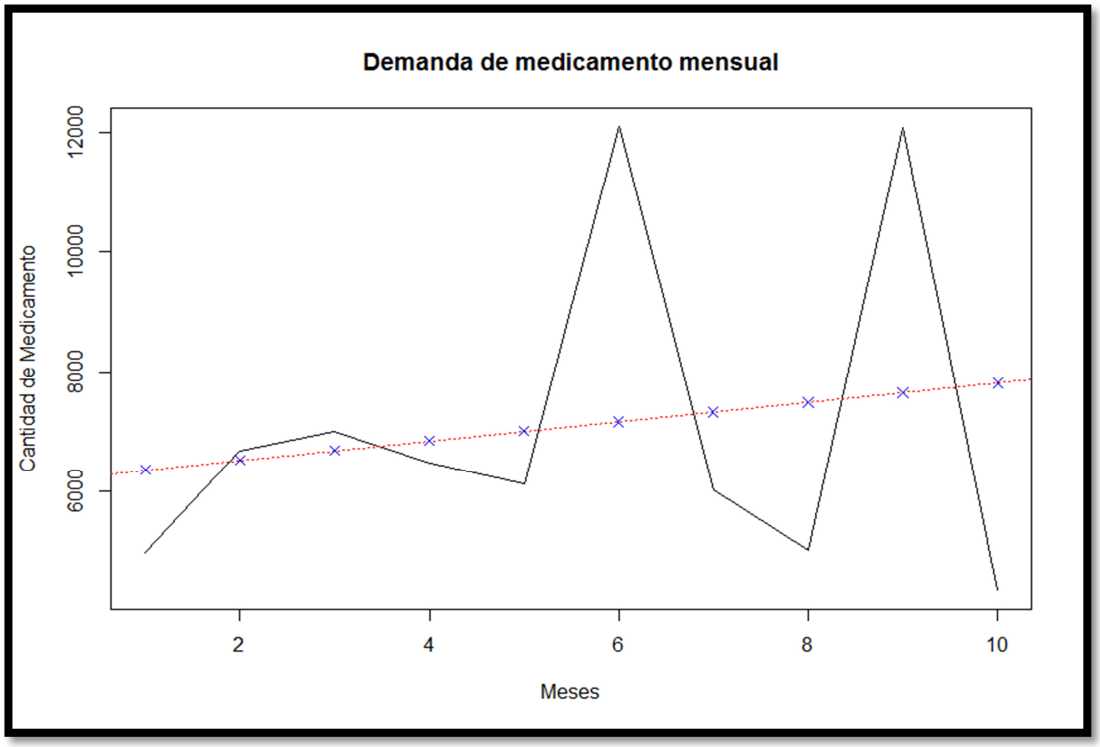

Figura 24. Modelo de regresión lineal de la demanda mensual de medicamentos del año 2016

El modelo lineal establece una ecuación que se describe a continuación (Ver anexo 3, literal A.):

$$Y = 6194.3 + 162.1X$$

Dónde:

Y: Total de productos entregados mensualmente

X: Meses

Tabla 9. Tabla de resultados del modelo de regresión lineal

RESULTADOS DEL MODELO DE REGRESIÓN LINEAL	
Intercepción:	6194.3
Variable:	162.1
RMSE:	2578.721
MAE:	1968.067

Modelo de regresión SVM

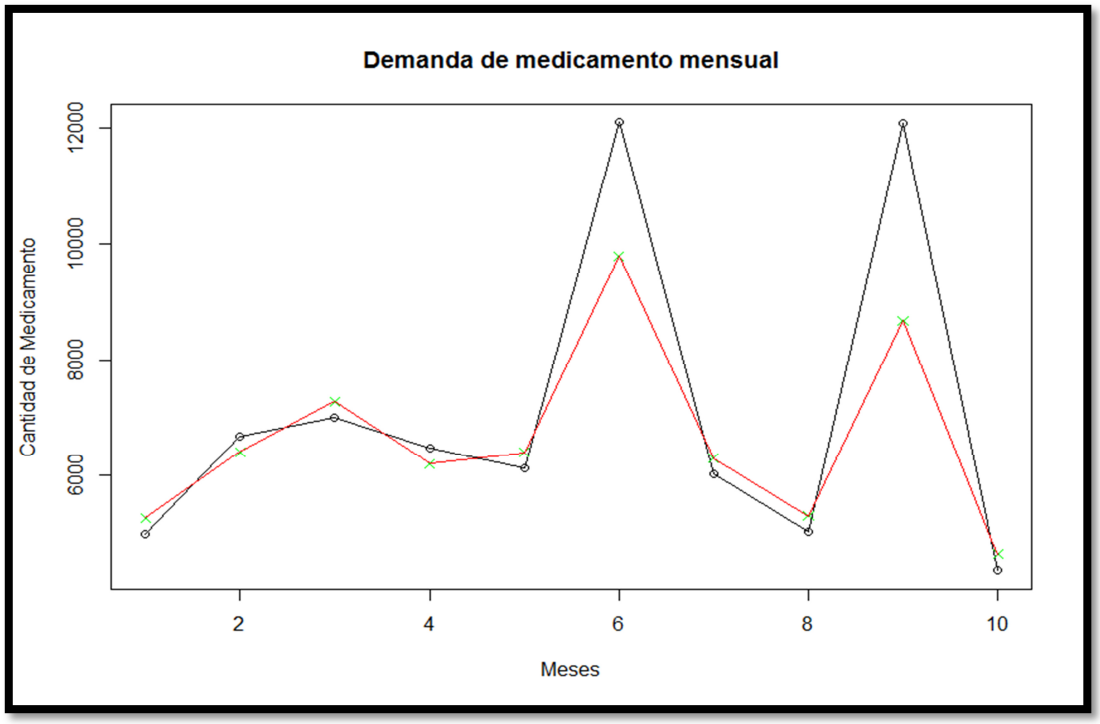

Figura 25. Modelo de regresión SVM de la demanda mensual de medicamentos del año 2016

Resumen del modelo de regresión SVM (Ver anexo 3, literal B.):

Tabla 10. Tabla de resultado del modelo de regresión SVM

RESULTADO DEL MODELO DE REGRESIÓN SVM	
RMSE:	1319.851
MAE:	790.5101

Nota: SVM se destaca por no tener ecuaciones debido a que es un modelo de minimización, el cual se ajusta a los valores reales.

- **Objetivo 2**. Desarrollar un modelo para predecir la tendencia del número de pacientes del Hospital "Carlos Andrade Marín" con perfil epidemiológico, que consumirán el medicamento oncológico Filgrastim, durante el año 2016.

Para llevar a cabo este modelo se realizará una regresión de la cantidad de pacientes que fueron atendidos mensualmente durante el año 2016 que hayan solicitado el medicamento "Filgrastim".

Modelo de regresión lineal

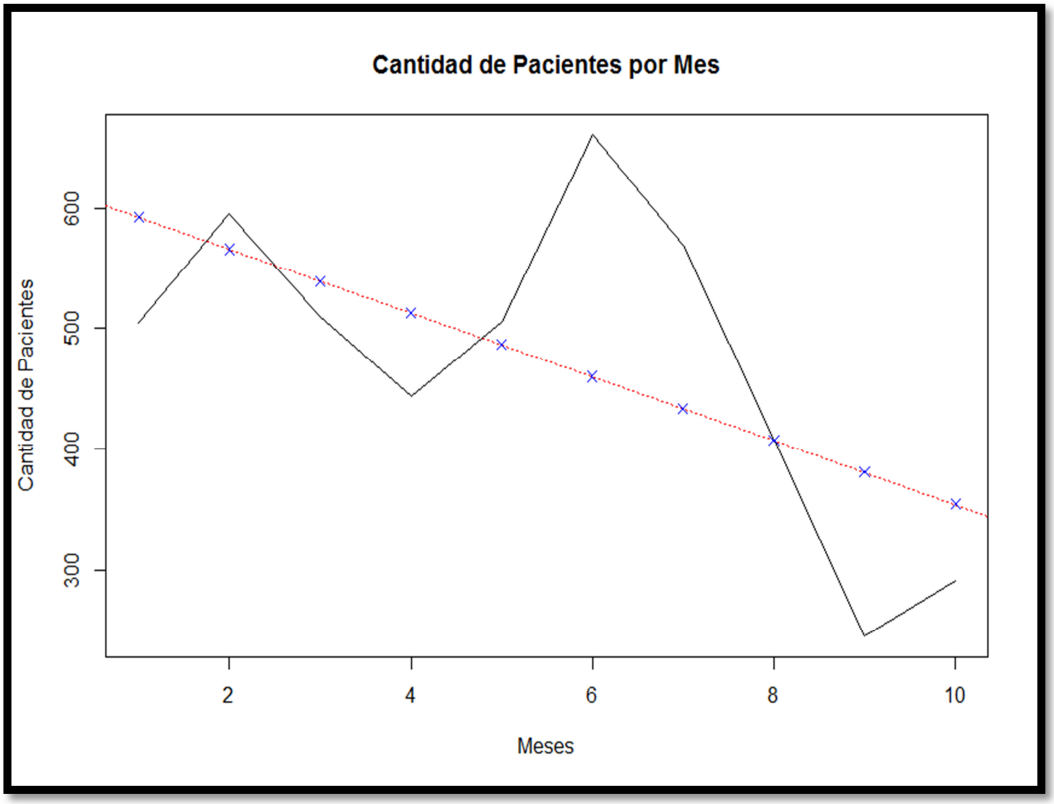

Figura 26. Modelo de regresión lineal de la cantidad de pacientes por mes

El modelo lineal presenta una ecuación que se describe a continuación (Ver script del anexo 3, literal C.):

$$Y = 618.53 - 26.41X$$

Dónde:

Y: total de pacientes atendidos

X: meses

Tabla 11.
Tabla de resultados del modelo de regresión lineal

RESULTADOS DEL MODELO DE REGRESIÓN LINEAL	
Intercepción:	618.53
Variable:	26.41
RMSE:	97.57
MAE:	76.94

Modelo de regresión SVM

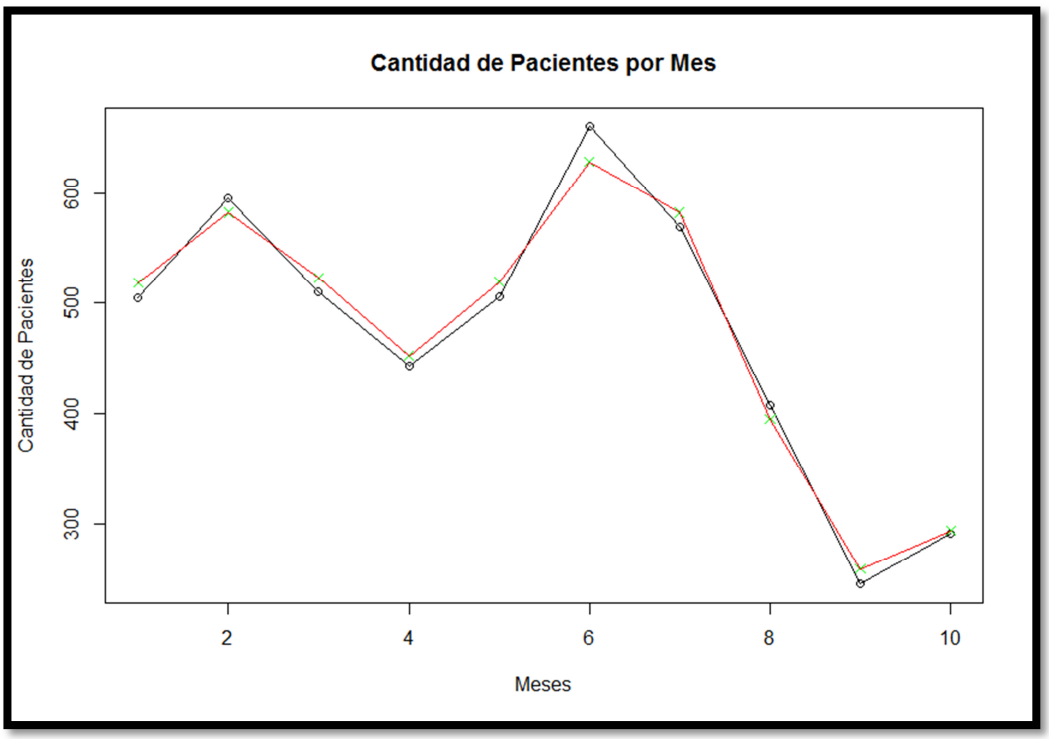

Figura 27. Modelo de regresión SVM de la cantidad de pacientes por mes

Resumen del modelo de regresión SVM (Ver script del anexo 3, literal D.):

Tabla 12.
Tabla de resultado del modelo de regresión SVM

RESULTADOS DEL MODELO DE REGRESIÓN SVM	
RMSE:	15.22
MAE:	13.50

- **Objetivo 3**. Predecir la cantidad óptima de Adquisición de Medicamento Oncológico Filgrastim por parte del Hospital "Carlos Andrade Marín" durante el año 2016.

Para encontrar la cantidad óptima de pedido de medicamento "Filgrastim" se realizará una sumatoria de cada uno de los modelos (Lineal y SVM), teniendo como insumo la tabla "movimiento_oncologico", donde se registran cada uno de los pedidos solicitados por los pacientes.

$$\boldsymbol{total} = \sum_{\boldsymbol{n=1}}^{\boldsymbol{meses=8}} (\mathbf{f(x)})$$

Figura 28. Ecuación para obtener la cantidad de medicamento

Descripción de los valores:

Tabla 13.
Tabla de descripción de los valores de la ecuación

DESCRIPCIÓN DE LOS VALORES	
Meses	Representa el número de meses a ser analizados.
f(x)	Representa a los valores predichos por los modelos tanto lineal como SVM.
N	Representa la iteración inicial de la sumatoria. En este caso comenzará a sumar desde el primer mes hasta el octavo.
Total	Almacena el resultado de la sumatoria.

A continuación se presentan los resultados de los modelos lineales y de máquina de vector de soporte (Ver anexo 3, literal E y F):

Tabla 14.
Tabla de resultados de la sumatoria de los modelos de regresión LM y SVM

RESULTADOS DE LA SUMATORIA DE LOS MODELOS DE REGRESIÓN LM Y SVM			
Modelo Regresión LM		**Modelo de regresión SVM**	
Total	70.857	Total	66.271,7

Al observar la Tabla 14, se establece que entre los modelos existe una diferencia de 4.585,3 (cantidad de medicamento Filgrastim), que significa que se debe recurrir al modelo que tenga menor cantidad de error tanto en MAE y RMSE, para seleccionar al modelo que sirva como predictor en las adquisiciones de medicamentos.

3.5.4 Evaluar el modelo.

En la siguiente tabla se realiza una comparativa de los modelos con respecto a la cantidad de errores generados en sus valores de predicción.

Tabla 15.
Tabla de resultados de errores medidos en los modelos propuestos

RESULTADOS DE ERRORES MEDIDOS EN LOS MODELOS PROPUESTOS				
	Objetivo 1		**Objetivo 2**	
	LM	**SVM**	**LM**	**SVM**
Error absoluto medio (MAE)	1968,067	790,51	76,94	13,50
Raíz del error cuadrático medio (RMSE)	2578,72	1319,85	97,57	15,22

Como se puede observar en la Tabla 15, se define la cantidad de error absoluto medio y el error cuadrático medio de cada modelo en cada uno de los objetivos. En esta parte no se menciona el objetivo 3, ya que este depende de la selección del mejor modelo del objetivo 1 para su respuesta.

Para seleccionar el modelo que satisfaga las necesidades de predicción se debe seleccionar aquel que tenga menor cantidad de errores en MAE y RMSE. Por ello, la selección de modelos se realizará en la sección de evaluación que viene a continuación.

3.6 Evaluación.

Esta sección se enfoca en seleccionar el mejor modelo para cada objetivo en base a los resultados obtenidos, que contribuirán a la selección de la opción más adecuada para la institución.

3.6.1 Evaluar los resultados.

- **Objetivo 1**

De acuerdo a la Tabla 16, el modelo lineal obtiene en MAE un valor de 1968.067 mientras que el SVM tiene 790.51, que significa que SVM tiene 2.49 veces menos cantidad de error en MAE. Si se analiza la parte de RMSE en LM se determina que tiene 2578.72 de error, mientras que SVM tiene 1319.85, datos que significan que SVM tiene 1.95 veces menos cantidad de error en RMSE que LM.

- **Objetivo 2**

En este objetivo se analizan los resultados de la Tabla 16 del objetivo 2, por lo cual se inicia analizando el error generado MAE en cada uno de los modelos. El MAE en LM es de 76.94, mientras que en SVM es de 13.50 lo cual significa que SVM tiene 5.70 veces menor cantidad de error que el modelo LM. Si se analiza el RMSE generado en LM se tiene 97.57, mientras que SVM tiene 15.22, que equivale a 6.41 veces menor cantidad de error que el modelo lineal.

- **Objetivo 3**

Este objetivo depende de la selección de cualquiera de los modelos que se escoja en el objetivo 1. No obstante y debido a los resultados que son sobresalientes en el modelo SVM, desde ya se tiene la respuesta para este objetivo.

Modelos aprobados.

Dado los argumentos en cada uno de los objetivos tanto desde la minería y del negocio se ha procedido a aprobar los modelos SVM en cada uno de los objetivos, ya

que sus valores son de mayor aproximación y presentan menor cantidad de errores tanto en MAE y RMSE.

Tabla 16.
Modelos seleccionados para los objetivos planteados

OBJETIVOS	MODELO SELECCIONADO
Objetivo 1	SVM
Objetivo 2	SVM

Una vez confirmado los modelos a ser utilizados en cada uno de los objetivos, se puede afirmar que para el cumplimiento del objetivo 3 se utilizará el modelo SVM para la predicción de cantidad de medicamento a solicitar.

3.6.2 Revisión del proceso.

En esta parte, la metodología CRISP-DM genera flexibilidad al investigador sobre todo si puede mejorar algún aspecto en particular. En este caso, se ha realizado mejoras en los scripts de análisis en R y se ha investigado funciones que han ayudado a minimizar la cantidad de código para los modelos.

3.6.3 Determinar los próximos pasos.

Los siguientes pasos corresponden a la implantación de los objetivos definidos y el establecimiento de los lineamientos para su implantación.

3.7 Implantación.

El objetivo de esta sección es explicar a la institución como poner en marcha el proyecto desarrollado y los resultados obtenidos durante la investigación. Además de los objetivos antes mencionados, se añade la implementación de un plan de estrategia que fusionado a un informe, resumen las mejoras del proyecto en un futuro, así como las dificultades encontradas durante la investigación.

3.7.1 Plan la implantación.

Para llevar a cabo este proyecto en la institución es necesario que se cuente con acceso directo a la base de datos de las tablas de "movimiento_oncológico", "historial_pacientes" y "medicamentos", con la finalidad de acceder a toda la información siempre que constituya una base de respaldo para no crear conflicto en la base transaccional. De allí en adelante se seguirá con la fase de análisis de datos hasta la evaluación de los resultados, que se ha venido desarrollando en el proyecto. Otro punto a tratar corresponde a tener un servidor solo dedicado al análisis de datos y que tenga instalado R para que ejecute los modelos desarrollados durante la investigación.

3.7.2 Plan de monitoreo y mantención.

Es importante monitorizar y mantener el aplicativo de acuerdo con la frecuencia que se cambian los datos e ingresan nuevos pedidos por parte de pacientes, por esta razón es necesario realizar un mantenimiento de los modelos de minería de datos cada seis meses para obtener resultados más acercados a la realidad, siempre que no se cambien los objetivos de la minería de datos establecidos en esta investigación.

Como plan de mantenimiento y monitorización se establecen los siguientes lineamientos:

- Cada once meses, se debe realizar un respaldo de la base de datos (movimiento_oncología, medicamentos, historial_pacientes) transaccional para actualizar la predicción de los modelos.
- Las tablas de respaldo deben ser volcadas en postgresql porque los modelos realizan el análisis de esa base de datos.
- Cada modelo entrenado debe ser guardado con su respetivo respaldo con la finalidad de tener un historial de los modelos actualizados.
- Realizar una comparación del nuevo modelo entrenado con el anterior, con la finalidad de que el nuevo modelo tenga mejor desempeño que el anterior. Y los resultados de la comparación deben ser almacenados para tener respaldo.

3.8 Informe Final.

La metodología CRISP-DM ha permitido seguir paso a paso los lineamientos para no desenfocarse de los objetivos establecidos en la minería de datos, que ha permitido contribuir con éxito el desarrollo de un modelo que determine el comportamiento de las tendencias de adquisición de medicamentos oncológicos "Filgrastim", asimismo se determinó la tendencia de pacientes que consumirán medicamento "Filgrastim". Finalmente se llegó a la predicción total de medicamentos a solicitar de tal forma que sea óptima para la institución.

De los tres objetivos definidos, el más laborioso fue el objetivo 3, debido a que depende del modelo seleccionado del objetivo 1. Pero para los próximos pasos es más sencillo porque ya se tiene el modelo seleccionado y se disminuirá el trabajo.

Ahora analizando los puntos para llegar al objetivo se debe referir que:

La primera etapa fue la interpretación de la información y su volcado en la base de datos postgresql, que luego fue analizada con el programa R, dónde se generó un cierto grado de dificultad, no obstante y conforme se fue avanzando, se fue facilitando el proceso.

Otro punto fue el de los datos atípicos que se encontró en la información, debido a que el bodeguero, a través de una sola orden, solicitaba gran cantidad de medicamentos, razón por la cual se decidió no considerar este aspecto, porque afectaría al modelo debido a la variación de pedidos de 5 a 7 medicamentos, y de pronto la aparición de cantidades de 600; no obstante este aspecto fue controlado gracias al análisis de los datos de diagrama de cajas que fueron muy útiles para encontrar estas variaciones.

Adicionalmente, se realizó el análisis de los objetivos, estableciendo la forma de medir la cantidad de errores generados por cada modelo, además de investigar los modelos a ser aplicados durante esta investigación. Para medir los errores se utilizó

el error absoluto medio y el error cuadrático medio, mientras que la codificación de los scripts se realizó en el programa R.

Finalmente, una vez que los modelos fueron codificados y se obtuvo la data para entrenar los modelos, se investigó e interpretó los resultados de los modelos a fin de tomar decisiones en beneficio de la institución, y así presentar los resultados obtenidos de la investigación.

3.9 Revisar el proyecto.

De acuerdo a la metodología CRISP-DM, esta etapa se enfocó en realizar una radiográfica de los procesos realizados, y revisar la posibilidad de establecer mejoras, tomando apuntes para las próximas ejecuciones que permitan mejorar el proceso de minería.

Otro parte que también se considera laboriosa fue el proceso de exploración de los datos, debido a que se debe analizar distintas aristas de la información para extraer conocimiento relacionado a los objetivos definidos, proceso que es necesario para encontrar puntos de soporte para los objetivos definidos, tal es el caso de los diagramas de caja que se realizaron para encontrar pedidos que ascendían a 500 por persona, y así como aquellos desarrollados en el resto del proceso de exploración.

CAPÍTULO IV
CONCLUSIONES Y RECOMENDACIONES

4.1 Conclusiones.

- El uso de la metodología CRISP-DM en este trabajo de investigación permitió crear un modelo de minería de datos que se adaptó al objetivo.
- En la fase de comprensión del negocio se realizó reuniones con el personal responsable del manejo de medicamentos del nosocomio, que permitió conocer la problemática de la institución en el proceso de gestión de medicamentos, acción que ayudó a reducir riesgos, clarificando los problemas, objetivos del negocio y recursos.
- Debido al valor informativo adquirido en los pasos previos, la recolección de los datos contribuyó a identificar las necesidades de la institución, clarificando cual debía ser la información histórica relacionada a la investigación como consumo de medicamentos, perfil epidemiológico, y medicamentos oncológicos que se encuentran en el listado del CNMB, datos que han permitido crear un modelo de predicción de compra planificada de medicamentos para futuros periodos.
- La exploración de los datos a través de tablas y gráficas permitió el análisis de la información que determinó el comportamiento de los pedidos de medicamento y la tendencia de pacientes que son atendidos mensualmente, además se descubrió un aumento de flujo de pacientes oncológicos en los meses de febrero, junio y julio, razón por la cual el despacho de medicamento aumentó en este periodo; otro aspecto relevante descubierto en esta exploración de datos es que la cantidad promedio de medicamento por paciente varía entre 4 y 5 unidades.
- Los diagramas de cajas y bigotes aplicados al análisis de datos permitieron descubrir la existencia de datos atípicos en el movimiento de medicamentos oncológicos, así como pedidos que variaban entre 200 a 500 para un solo responsable, que responde a tratamientos realizados en el hospital; además al realizar una depuración de los datos atípicos se

identificó que los pedidos de medicamentos oncológicos tienden a una asimetría positiva, cuya cantidad de pedidos es superior a 4.

- En cuanto a las técnicas aplicadas para la creación del modelo, se determinó que el resultado que entregó el modelo SVM utiliza un proceso de optimización que establece el valor de los parámetros que permite que el algoritmo tenga el valor más óptimo.
- Las técnicas para medir el error generado en cada modelo sirvió para seleccionar el mejor modelo para la investigación.
- Se encontró que los modelos SVM son mejores a los LM debido que se acercan más al comportamiento de los datos originales.
- De acuerdo al análisis, las máquinas de soporte presentan una mayor aproximación a los comportamientos de tendencia tanto de medicamento como de pacientes.
- Se evidenció el uso de los modelos SVM y LM, y se pudo resolver los dos primeros objetivos, no obstante el que presentó menor cantidad de errores fue el SVM.
- Finalmente el uso de modelo SVM predijo con mayor exactitud la cantidad de medicamento a solicitar para el siguiente periodo

4.2 Recomendaciones.

- Se recomienda realizar cada 11 meses un nuevo entrenamiento de los modelos con el objetivo de mejorar las predicciones del modelo.
- Cuando se realice el entrenamiento de los modelos, se debe tenerlos en un servidor aparte de donde se encuentra la base transaccional.
- Para el entrenamiento de los modelos se debe tener una maquina adecuada para hacer el procesamiento.
- El modelo entrenado no necesita volver a entrenarse porque ya tiene todas las condiciones necesarias para predecir por sí solo, sin recurrir a la data.
- Es recomendable no manipular el script de los modelos entrenados sin que previamente haya sido capacitado, para evitar una mala interpretación en los resultados.

- Cada vez que el modelo haya sido alterado, se debe realizar una copia de este.
- Cada modelo entrenado debe compararse con sus predecesores con la finalidad de observar si existe una mejora sustancial, y dependiendo de ello tomar el criterio más adecuado para ponerlo en marcha o no.
- Se recomienda utilizar los modelos SVM porque son los más precisos de acuerdo a los datos obtenidos en la investigación realizada.

BIBLIOGRAFÍA

- Barrios, M. (2010). Modelo del Negocio. *Americana*.
- Belinchón, Y. (s.f.). *Minería de datos.* Obtenido de http://www.it.uc3m.es: http://www.it.uc3m.es/~jvillena/irc/practicas/10-11/15mem.pdf
- Benoît, G. (2002). Data Mining. *Annual Review of Information Science and Technology*, 265-310.
- Bigado, V., & Arruzazabala, M. (Septiembre de 2003). *exa.unne.edu.ar*. Obtenido de /exa.unne.edu.ar: http://exa.unne.edu.ar/informatica/SO/MonografiaMD.PDF
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (30 de 12 de 2016). *The modelo agency*. Obtenido de CRISP-DM 1.0: https://www.the-modeling-agency.com/crisp-dm.pdf
- Consultores, W. (16 de 06 de 2016). *¿Qué es Data Mining?* Obtenido de Webmining Powering Welb Intelligence: http://www.webmining.cl/2011/04/que-es-data-mining/
- Diaz, B., Iribarrra, F., & Gutierrez, M. (4 de Noviembre de 2016). *Minería de Datos*. Obtenido de http://mineriadatos1.blogspot.com/2013/06/descubrimiento-del-conocimiento-kdd-el.html
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (8 de Enero de 1996). *From Data Mining to Knowledge Discovery in Databases.* Obtenido de http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf
- Franco, A. (Diciembre de 2001). *Regresión Lineal*. Obtenido de www.sc.ehu.es: http://www.sc.ehu.es/sbweb/fisica/cursoJava/numerico/regresion/regresion.htm
- Galán, V. (2015). *Aplicación de la metodología CRISP-DM en un proyecto de minería de datos en el entorno universitario.* Madrid.
- GEO, T. (15 de Junio de 2015). *Proyección de Demanda*. Obtenido de Gestión de Operaciones: https://www.gestiondeoperaciones.net/proyeccion-de-demanda/calculo-de-la-raiz-del-error-cuadratico-medio-o-rmse-root-mean-squared-error/

- Han, J., & Kamber, M. (2001). *KDD: Proceso de Extracción de conocimiento*. Obtenido de Webmining Powering Web Intelligence: http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/
- Hernández, J., Ramírez, M., & Ferri, C. (2004). *Introducción a la Minería de Datos.* España: Pearson.
- Julia. (25 de Mayo de 2017). *Error absoluto medio*. Obtenido de Kaggle: https://www.kaggle.com/wiki/MeanAbsoluteError
- Nigro, Ó., Xodo, D., Corti, G., & Terren, D. (5 de Septiembre de 2016). *KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario* . Obtenido de http://sedici.unlp.edu.ar/bitstream/handle/10915/21220/Documento_completo.pdf?sequence=1
- Niño, M. (30 de 12 de 2016). *CRISP-DM: Metodología para proyectos de Data Mining.* Obtenido de http://www.mikelnino.com/2015/09/crisp-dm-metodologia-proyectos-data-mining.html
- Pautsch, J. (13 de 06 de 2016). *UNNE*. Obtenido de http://exa.unne.edu.ar/informatica/SO/TFAGermanPAUTSCHFinal.pdf
- Sematech, N. (Abril de 2012). *Engineering Stadistics Handbook*. Obtenido de e-Handbook of Statistical Methods: http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm
- Weiss, S., & Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide.* Morgan Kaufmann.
- Yoshibauco. (27 de Abril de 2011). *Algoritmos TDIDT aplicado al Analisis de suelo.* Obtenido de yoshibauco.wordpress.com: https://yoshibauco.wordpress.com/2011/04/27/empezando-con-las-etapas-de-crisp-dm/
- Zapata, S. (25 de Marzo de 2011). *Técnicas de Minería de Datos basadas en Aprendizaje Automático.* Obtenido de santiagozapatakdd.wordpress.com: https://santiagozapatakdd.files.wordpress.com/2011/03/curso-kdd-full-cap-3.pdf