UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

MODELO BASADO EN MINERÍA DE DATOS PARA PREDECIR LA CONDICIÓN DE SALUD DE LOS RECIÉN NACIDOS EN LA RED DE SALUD CHUCUITO – JULI EN EL PERIODO 2016 - 2018

TESIS

PRESENTADA POR:

RENE ARTURO CUTIPA CHAMBI

RUSSELL EULOGIO CARBAJAL VILCA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

PUNO – PERÚ

2019

UNIVERSIDAD NACIONAL EL ALTIPLANO - PUNO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

MODELO BASADO EN MINERÍA DE DATOS PARA PREDECIR LA CONDICIÓN DE SALUD DE LOS RECIÉN NACIDOS EN LA RED DE SALUD CHUCUITO - JULI EN EL PERIODO 2016 - 2018

TESIS PRESENTADA POR:

RENE ARTURO CUTIPA CHAMBI

RUSSELL EULOGIO CARBAJAL VILCA

PARA OPTAR EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS

APROBADO POR EL JURADO REVISOR CONFORMADO POR:

PRESIDENTE : .....
Mg. CARLOS BORIS SOSA MAYDANA

PRIMER MIEMBRO : .....
M.Sc. PABLO CESAR TAPIA CATACORA

SEGUNDO MIEMBRO : .....
M.Sc. MAGALI GIANINA GONZALES PACO

DIRECTOR / ASESOR : .....
M.Sc. WILLIAM EUSEBIO ARCAYA COAQUIRA

TEMA: Minería de datos

ÁREA: Ingeniería de Software, Bases de Datos e Inteligencia de Negocios

FECHA DE SUSTENTACIÓN: 12 SETIEMBRE DEL 2019

DEDICATORIA

A los que más se cansaron, más nunca se agotaron; mis padres.

A los que estaban ahí, mirando; mis hermanos.

A quienes estaban cerca, y ahora están lejos.

A quienes estaban lejos, y ahora están cerca.

René Cutipa

*A **Dios** por ser mi creador, por ser mi guía durante toda mi vida y por darme fortaleza y sabiduría.*

*Con todo mi amor y cariño a mis lindas hijas **Andreita y Camilita**, por ser mi motivación para poder superarme cada día más.*

*A mi amada esposa **Fanny Luz** por su comprensión, por ayudarme a construir mis sueños y por ocupar un lugar importante en mi corazón.*

*A mis queridos padres **Julia y Máximo**, por su apoyo incondicional en los peores momentos de mi vida y por enseñarme a ser un hombre de bien; y a mi querida hermana **Iris** por sus palabras de aliento que nunca le faltaron.*

Russell Carbajal

AGRADECIMIENTOS

Nuestro agradecimiento a la Escuela Profesional de Ingeniería de Sistemas de la Universidad Nacional del Altiplano, que, en sus aulas, junto a compañeros, amigos y docentes han sido parte de nuestra vida universitaria y formación profesional.

A los catedráticos que han intervenido en nuestra formación, por compartir sus conocimientos, absolver dudas; a los cuales debemos nuestra admiración y respeto.

Al Ms. Sc. William Eusebio Arcaya Coaquira, que, con sus aportes y dirección, ha impulsado el desarrollo de esta investigación.

A los miembros del Jurado Mg. Carlos Boris Sosa Maydana, Ms.Sc. Pablo Cesar Tapia Catacora, Ms.Sc. Magali Gianina Gonzales Paco; por ser partícipes de esta etapa.

Los autores

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

ÍNDICE DE ACRÓNIMOS

FUA: Formato Único de Atención.

FUR: Fecha de Última Regla.

SIS: Seguro Integral de Salud.

PS: Puesto de Salud.

CS: Centro de Salud.

RN: Recién Nacido.

HB: Hemoglobina.

SGRN: Semanas de Gestación de Recién Nacido.

KDD: Knowledge Discovery in Databases (Descubrimiento de Conocimiento en bases de datos)

CRISP-DM: Cross Industry Standard Process for Data Mining (Modelo de Proceso Estándar para Minería de Datos)

WEKA: Waikato Environment Knowledge Analysis (Entorno de Análisis De Conocimiento de la Universidad de Waikato)

ANN: Red Neuronal Artificial.

IPRESS: Instituciones Prestadoras de Servicios de salud.

ENDES: Encuesta Demográfica y de Salud Familiar.

OMS: Organización Mundial de la Salud.

ALC: América Latina y el Caribe.

DIRESA: Dirección Regional de Salud.

MINSA: Ministerio de Salud.

RESUMEN

La presente investigación titulada MODELO BASADO EN MINERÍA DE DATOS PARA PREDECIR LA CONDICIÓN DE SALUD DE LOS RECIÉN NACIDOS EN LA RED DE SALUD CHUCUITO – JULI EN EL PERIODO 2016 - 2018, tuvo como objetivo: aplicar un modelo basado en minería de datos para predecir la condición de salud de los recién nacidos en la Red de Salud Chucuito-Juli. Esta investigación es descriptiva y de tipo histórico pues requerimos conocer, analizar y evaluar nuestras variables en sus distintas etapas basados a un periodo de evaluación, asimismo, el diseño de la investigación es descriptiva simple, debido a que son observaciones al grupo experimental, al cual se han realizado las observaciones y obtenido los resultados, se ha tenido como población a 484 gestantes en la Red de Salud Chucuito-Juli, la misma que también es usada como muestra. Esta investigación ha desarrollado minería de datos usando la metodología CRISP-DM, teniendo a los árboles de decisión como técnica de minería de datos, con el uso de los algoritmos J48 y Logistic Model Tree. La investigación realizada establece que; en apoyo al diagnóstico y prevención los atributos procesados y discretizados posibilitan determinar la condición de salud de los recién nacidos; la misma que tiene un error absoluto de ± 0.1639, esto confirma la hipótesis que el modelo basado en minería de datos permite predecir la condición de salud del recién nacido en la Red de Salud Chucuito-Juli.

Palabras Clave: Condición de salud del recién nacido, Gestante, Minería de datos, Modelo predictivo.

ABSTRACT

This research entitled DATA-BASED MINING MODEL TO PREACH THE HEALTH CONDITION OF NEWBORDS IN THE RED DE SALUD CHUCUITO-JULI IN THE PERIOD 2016 - 2018, aimed to: apply a model based on data mining to predict the health condition of the newborns in Red de Salud Chucuito-Juli. This research is descriptive and of a historical type because we need to know, analyze and evaluate our variables in their different stages based on an evaluation period, also, the design of the research is simple descriptive, because they are observations to the experimental group, to which The observations have been made and the results obtained, 484 pregnant women in the Red de Salud Chucuito-Juli have been taken as a population, which is also used as a sample. This research has developed data mining using the CRISP-DM methodology, taking decision trees as a data mining technique, using the algorithms J48 and Logistic Model Tree. The research carried out establishes that; in support of diagnosis and prevention the attributes processed and discretized make it possible to determine the health condition of newborns; It has an absolute error of ± 0.1639, this confirms the hypothesis that the model based on data mining allows to predict the health condition of the newborn in Red de Salud Chucuito-Juli.

Keywords: Data mining, Newborn health condition, Predictive model, Pregnant.

CAPITULO I

INTRODUCCIÓN

El aseguramiento de las condiciones de salud, es una necesidad imperante a medida que surgen nuevos avances y al mismo tiempo nuevas amenazas a la misma. Esta investigación se centra en otorgar una herramienta de apoyo al diagnóstico para conocer antes de su nacimiento la probable condición de salud del recién nacido, esto mediante técnicas y herramientas de minería de datos.

En la actividad diaria de las atenciones que realiza la Red de Salud Chucuito – Juli, se presentan gestantes con diferentes características; estas características se abstraen y generan conocimiento; y a la vez esta pueda servir como apoyo a asegurar la condición de salud del recién nacido.

La presente investigación consta de: Capítulo I, se hace referencia al planteamiento del problema de investigación, en base a ello se formula el problema, la justificación de la investigación y los objetivos; en el Capítulo II, está referido a los antecedentes que tienen relación con el trabajo de investigación y el sustento teórico. El Capítulo III se menciona la metodología de la investigación, la población, los métodos, técnicas de medición de datos. En el Capítulo IV se muestran los resultados y discusión de la misma, el cual se llegó durante la investigación, así como la presentación de conclusiones y recomendaciones. Finalmente, se hace mención de las referencias bibliografía y anexos.

1.1. PLANTEAMIENTO DEL PROBLEMA

La mala condición de salud en un recién nacido es un problema que a través del tiempo permanece con indicadores altos, el cual se refleja en la morbilidad y mortalidad

de estos, a su vez la salud neonatal se halla estrechamente vinculada a la salud materna. La Organización Mundial de la Salud (OMS) define: “La salud es un estado de completo bienestar físico, mental y social, y no solamente la ausencia de afecciones o enfermedades” (INEI-ENDES, 2018). Dicho esto, durante el embarazo tanto la mujer materna como su futuro hijo afrontan distintos riesgos de salud, en ese sentido es importante que el seguimiento del embarazo lo realicen profesionales de salud calificados y con las herramientas necesarias y adecuadas.

La salud de los recién nacidos en el mundo entero tiene cifras alarmantes, en niños menores de cinco años que fallecen cada año, cerca del 40% son lactantes recién nacidos, es decir bebés de menos de 28 días o en periodo neonatal, estos fallecimientos se producen por lo general en países subdesarrollados que tienen escasos recursos para la atención de la salud, en gran parte estos recién nacidos se enferman y fallecen por no tener un adecuado control en el periodo gestacional.

Durante los últimos años la magnitud de la problemática neonatal en América Latina y el Caribe (ALC) ha llevado a cabo considerables avances en la disminución de la morbilidad y mortalidad post-neonatal; no obstante, el descenso de la morbilidad y mortalidad neonatal no ha seguido la velocidad de esta tendencia. Así mismo en nuestro país se exhibe una alta tasa de mortalidad y morbilidad neonatal, superiores a la de países desarrollados, y se muestra un mayor riesgo en los recién nacidos en zonas rurales y de bajo nivel de accesibilidad a las Instituciones Prestadoras de Servicios de salud (IPRESS).

El Instituto Nacional de Estadística e Informática - INEI, mediante la Encuesta Demográfica y de Salud Familiar (ENDES) 2012, 2013, 2014, 2015 y 2016 muestran la tasa de mortalidad neonatal de 5 años anteriores a la encuesta, obteniendo los resultados

de valor estimado de 10, 11, 10, 10 y 10 por cada mil nacidos vivos respectivamente, la información más reciente es el del 2017 donde se reportó 10 por cada mil nacidos vivos, esta última información proviene de un informe preliminar que se encuentra al 50% de la muestra; estos datos evidencian que la mortalidad neonatal en esos últimos 5 años se mantiene sin ninguna disminución o peor aún sin ninguna mejora.

Una de las regiones con más alto índice de mortalidad neonatal es la Región Puno, según los últimos datos de la ENDES 2017: Puno tiene la tasa de mortalidad neonatal más alta del país con 12 muertes de cada mil nacidos vivos, dichas muertes se registran más en el área rural que se estima aproximadamente en un 85%, además, de esta población el porcentaje de gestantes que recibieron 6 o más controles prenatales, es solo el 80,3%, siendo esta cifra también la más baja del país.

En la jurisdicción de la Red de Salud Chucuito – Juli, la situación del problema evidentemente no es distinta, en donde las muertes fetales son más que las muertes neonatales, ocurriendo estos hechos con mayor proporcionalidad en el área rural, ya que la provincia de Chucuito según el último resultado definitivo de los Censos Nacionales 2017, tiene el 27,1% de población urbana, y el 72,9% de población rural, concentrando el 12,0% de toda la población rural de la región Puno.

Las razones por las que se originó este problema con neonatos en la Red de Salud Chucuito – Juli, es la alta cantidad de madres gestantes con diversas enfermedades en la cual prevalece la anemia, que además se da más en el área rural, a esto se suma el exceso de madres gestantes atendidas en algunos centros de salud de la Red, no teniendo así un control y/o seguimiento adecuado en la mayoría de casos, tanto en el área rural como urbano.

De continuar este problema el pronóstico es negativo para las madres gestantes y peor aún para los recién nacidos, teniendo como efectos y consecuencias a recién nacidos con complicaciones de bajo peso, prematuros, con sepsis neonatal y/o con enfermedades neonatales diversas, es decir en malas condiciones de salud que pueden agravar su situación y quizás llevándolos hasta la muerte.

Por lo arriba mencionado la investigación propone un aporte a la identificación de patrones y monitoreo de este problema; es importante para el personal de salud conocer bien el estado de las madres gestantes, mediante un monitoreo adecuado y oportuno, el cual sea ajustado para cada control de gestación en las madres, es por eso este modelo para predecir las condiciones de salud de los recién nacidos, modelo que será basado con la ayuda del proceso de minería de datos, en la cual se evaluará señales que son factores de riesgo durante el periodo gestacional, y estos proporcionarán los indicadores que intervienen en la predicción de condición del recién nacido; por una parte este modelo predictivo nos posibilitará analizar el grupo de indicadores y arrojar una predicción que ayude a tener un mayor control y señales de alerta para procurar nacimientos en mejores condiciones, por otra parte será una herramienta que ayude al personal de salud a monitorear mejor el proceso de gestación y poder determinar en qué condiciones se tendrá a un recién nacido.

1.2. FORMULACIÓN DEL PROBLEMA

¿De qué manera el modelo basado en minería de datos permite predecir las condiciones de salud de los recién nacidos en la Red de Salud Chucuito – Juli?

1.3. JUSTIFICACIÓN DEL PROBLEMA

1.3.1. JUSTIFICACIÓN TEÓRICA

El presente estudio de investigación, al utilizar la tecnología minería de datos aporta aspectos teóricos en un área donde realmente había poco conocimiento en predecir observaciones futuras con precisión, como lo es en el área de la salud y específicamente en neonatología, en otras palabras, dando un nuevo conocimiento preliminar de predicción para la condición de salud de los recién nacidos.

1.3.2. JUSTIFICACIÓN METODOLÓGICA

Debido a que no se cuenta con suficientes estudios relacionados a esta investigación en particular, se ha propuesto y realizado este modelo predictivo con procedimientos supervisados, es decir donde llegamos a saber una respuesta (predicción). Modelo que fue obtenido con apoyo de la metodología CRISP-DM, su núcleo la minería de datos y algoritmos clave para descubrir patrones en los datos, todo esto hace a que la investigación sea singular para el campo de la neonatología.

1.3.3. JUSTIFICACIÓN PRÁCTICA

Existen muchos factores de riesgo durante un periodo gestacional que afectan a las condiciones en las que el nace un bebé. Este trabajo de investigación pretende evaluar todas esas señales e implementar un modelo predictivo que nos permita monitorizar y ver a futuro las condiciones en las que pudiera nacer el neonato. Entendemos que son diferentes factores los que intervienen en este proceso. Toda esta información nos permitirá analizar, evaluar y valorizar el grupo de indicadores y arrojar una correcta predicción que ayude a tener un mayor control y señales de alerta para procurar nacimientos en las mejores condiciones.

1.3.4. JUSTIFICACIÓN ECONÓMICA

Al tener un mejor seguimiento de madres gestantes bajo este modelo de predicción de condiciones de nacimiento, se van a cumplir indicadores planteados por el Seguro Integral de Salud (SIS), y eso implica una reducción de gastos en cada establecimiento de salud, ya que el porcentaje de morbilidad de madres e hijos reduciría, por otro lado también es evidente que las familias con madres gestantes y puérperas ya no tendrán que realizar gastos extras, ya que sabemos que la enfermedad de las mismas o los recién nacidos conllevaría gastos adicionales como enseres, transporte y lo más importante el tiempo, todo esto a pesar de contar con el seguro integral de salud.

1.3.5. JUSTIFICACIÓN SOCIAL

La investigación apoya y puede mostrarnos las señales de alerta del comportamiento durante el proceso gestacional, ya que, al tener un mejor seguimiento y monitoreo de madres gestantes, se va a reducir los índices de mortalidad y morbilidad tanto materna como neonatal, es decir esto significaría dar una mejor calidad de salud para las mujeres y los recién nacidos, lo cual es impacto positivo en la sociedad.

1.4. OBJETIVOS DE LA INVESTIGACIÓN

1.4.1. OBJETIVO GENERAL

Aplicar un modelo basado en minería de datos para predecir la condición de salud de los recién nacidos en la Red de Salud Chucuito - Juli.

1.4.2. OBJETIVOS ESPECÍFICOS

- Recopilar información histórica de madres en estado gestacional.
- Identificar patrones que intervienen en la predicción de condición del recién nacido.
- Clasificar la información bajo técnicas predictivas de minería de datos para obtener un modelo predictivo adecuado para definir la condición del recién nacido.

1.5. HIPÓTESIS

El uso del modelo basado en minería de datos permite realizar la predicción de la condición de salud de los recién nacidos en la Red de Salud Chucuito – Juli.

CAPITULO II

REVISIÓN DE LITERATURA

2.1. ANTECEDENTES DE LA INVESTIGACIÓN

2.1.1. A NIVEL NACIONAL

Ccopa, M. y Chavez, S. (2015) en la tesis MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA LA MEJORA EN LA TOMA DE DECISIONES DEL DEPARTAMENTO DE CIRUGÍA DEL HOSPITAL REGIONAL MANUEL NÚÑEZ BUTRÓN.

> Tuvieron por objetivo general desarrollar un Modelo Predictivo basado en Minería de Datos para la mejora en la toma de decisiones del Departamento Cirugía del Hospital Regional Manuel Núñez Butrón. Metodología de la investigación: de acuerdo a las características del problema, objetivo y la hipótesis es de tipo experimental, de diseño pre experimental, con pre-test y post-test, los diseños pre experimentales no presentan grupo control. Tuvo como población: Todos los datos de pacientes atendidos en las diferentes especialidades en el Hospital Manuel Núñez Butrón Puno, y muestra: Todos los datos de Pacientes atendidos en el periodo (2012-2013) en el Departamento de cirugía del Hospital Manuel Núñez Butrón Puno. Instrumentos: Encuestas: se aplicó encuestas a la Unidad de Estadística con el fin de obtener información requerida por el departamento de Cirugía. Fichas de observación: se elaboró una ficha de observación para registrar los datos del experimento con modelo y sin modelo. Resultados: Se desarrolló el Modelo Predictivo Basado en Minería de Datos para la Mejora en la Toma de Decisiones del Departamento de Cirugía del Hospital

Regional Manuel Nuñez Butrón, gracias al promedio del Coeficiente de Correlación de los indicadores hospitalarios de 81.34% que se obtuvo la evaluación del modelo es correcta mejorando así en la toma de decisiones en un porcentaje de 91.97%, y se logró comprender la situación actual del departamento de cirugía del Hospital Regional Manuel Núñez Butrón identificando los objetivos del departamento de cirugía, definiendo el problema de la minería de datos, identificando como caso de éxito el índice de precisión para la mejora en la toma de decisiones.

Saldaña, E. (2015) en la tesis MODELO PREDICTIVO DE MINERÍA DE DATOS DE APOYO A LA GESTIÓN HOSPITALARIA SOBRE LA MORBILIDAD DE PACIENTES HOSPITALIZADOS.

Tuvo como objetivo Crear un modelo predictivo de minería de datos de apoyo a la gestión hospitalaria sobre la morbilidad de pacientes hospitalizados. Metodología: En esta investigación, se tomó como referencia la metodología CRISP-DM (Cross Industry Standard Process For Data Mining), que consiste en la comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. Instrumentos: se realizaron entrevistas de tipo no estructuradas, en forma espontaneas al personal de los diferentes servicios del hospital. Fuentes Documentales, recopilación de información, a través de fichas bibliográficas, estado del arte sobre el tema de tesis. Fuentes Datos, la recopilación de la información de los registros transaccionales de los últimos 8 años registrados en la base datos del hospital. Resultados: A lo largo de esta investigación se ha llevado a cabo una importante recopilación bibliográfica y revisión teórica sobre aspectos relacionados con el tema, que han permitido

conocer técnicas predictivas de series de tiempo estacionarias y no estacionarias, así como los métodos de pronósticos y suavizamiento de Box & Jenkins que incluye los Modelos AR (Auto-Regresivos), Modelos MA (Media Móvil), Modelos ARIMA (Auto regresivo Integrado con Media Móvil) y Modelos SARIMA (Auto regresivos Integrados con media móvil estacional); durante el proceso de preparación de los datos, se identificó los datos de origen, en dos bases datos transaccionales SQL Server (SYSFAR y GALENHOS), luego del proceso de Extracción Transformación y Carga (ETL), a través de consultas rápidas y técnicas de muestreo se detectó datos anómalos, eliminando o separando las tuplas, para posteriormente crear y/o cargar el Datamart con seis dimensiones y una tabla de hecho formando un modelo estrella, que sirve como repositorio para que finalmente durante un proceso de selección y transformación de variables, obtener los datos de entrada para el modelo.

Ticona M. (2018) en la tesis SISTEMA PARA LA PREDICCIÓN DE OBESIDAD EN LA ADOLESCENCIA UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS.

Tuvo como objetivo analizar, diseñar e implementar un software, con el uso de técnicas de minería de datos, que pueda predecir si un adolescente a determinada edad, va a padecer de obesidad en el futuro, es decir en la adultez y que diagnostique cual es el estado actual del paciente. Para ello intervienen diversas variables socio-culturales, familiares, genéticas y antropométricas. Metodología: Experimental, debido a que se requiere predecir posibilidades de enfermedad de obesidad en adolescentes y se pretende probar el uso de 3 técnicas de minería de datos y seleccionar la más adecuada. Población: Estudiantes, profesores de

educación física y padres, entre diversos colegios de la ciudad de Arequipa – Perú. Especialistas en el campo de la salud también deben ser incluidos. Muestra: estudiantes de 5 a 17 años de edad, entre diversos colegios de la ciudad de Arequipa – Perú. Resultados: Los algoritmos evaluados fueron J48, BayestNet, Multilayer Perceptron, ForestPA y NaiveBayes, obteniendo como mejor el algoritmo J48, con un porcentaje de precisión de 94.39%, y demostrando ser superior en otros indicadores. El algoritmo obtenido de las pruebas y comparaciones realizadas, fue implementado en una herramienta de software, con el objetivo de automatizar el proceso y evaluar a más personas para futuras investigaciones. Los resultados obtenidos de los algoritmos de clasificación de J48, BayesNet, MultilayerPerceptron, ForestPA y NaiveBayes para que posteriormente sean evaluados y saber cuál es el más adecuado para el presente caso. En cuanto a las cantidades de data analizadas y clasificadas, 660 registros de entrenamiento fueron usados; 10% fueron utilizados para la fase de entrenamiento y 90% para la fase de prueba, con la herramienta Weka, y el tipo de prueba Cross Validation. Es necesario saber que la cantidad de verdaderos positivos (TP) es equivalente a la cantidad de ejemplos que son verdaderos positivos y los falsos negativos (FP) son el número de ejemplos falsos negativos encontrados.

2.1.2. A NIVEL INTERNACIONAL

Abad, I. (2016) en la tesis MODELO PREDICTIVO DE PARTO PREMATURO BASADO EN FACTORES DE RIESGO.

Tuvo como objetivo: Diseñar un modelo matemático como instrumento que permita detectar el riesgo de parto prematuro, siendo extremadamente útil para

seleccionar a las pacientes con mayor probabilidad de padecerlo e intentar actuar en modo preventivo. Materiales y métodos: Se realizó un estudio observacional, retrospectivo y descriptivo de 481 nacimientos ocurridos entre el año 2015 y principios de 2016 en el Hospital Universitario Central de Asturias; de los cuales 257 fueron nacimientos pretérminos y 224 a término. Se llevó a cabo un análisis estadístico de los parámetros materno-fetales, recogidos a partir de las historias clínicas, y de los hábitos de sueño, conocidos gracias a la realización de una encuesta telefónica. A partir de las 223 pacientes más completas se realizó un modelo matemático aplicando el método de los árboles de decisión. Resultados: A partir del análisis de las variables se obtuvieron diferencias estadísticamente significativas en la dilatación, peso estimado, peso al nacer, tipo de concepción, domicilio, gestación múltiple (p=0.000), rotura prematura de membranas (p=0.001), interrupciones del sueño (p=0.021) y primiparidad (p=0.044) entre prematuros y control. Además, se obtuvo una herramienta que permite predecir el riesgo de parto prematuro en mujeres embarazadas. Conclusiones: Se ha diseñado el primer modelo matemático predictivo del riesgo de parto prematuro en gestantes, que permitirá actuar de manera preventiva para intentar reducir la incidencia de prematuridad en la sociedad actual.

Martínez, C. (2012) en la tesis APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA MEJORAR EL PROCESO DE CONTROL DE GESTIÓN EN ENTEL.

Tuvo como objetivo: Reducir el tiempo de cálculo de los indicadores de servicios privados de ENTEL, para lo cual se aplicó modelamiento multidimensional, técnicas de minería de datos y automatización de procesos, y de este modo poder entregar información más oportunamente. El presente estudio se enfoca en el

análisis de ingresos no percibidos en la empresa de telecomunicaciones ENTEL, dentro del proceso de provisión de servicios privados de telefonía, internet y comunicaciones a los clientes de mercados no residenciales. Dicho proceso es controlado mediante indicadores de gestión, obtenidos a partir de la transformación de datos de clientes y servicios. La generación de estos indicadores demanda tiempo y esfuerzo por parte de los analistas de la empresa, debido a que es un trabajo realizado en forma manual. La metodología de este trabajo se basa principalmente en las etapas del proceso conocido como Knowledge Discovery in Databases (KDD), implementadas de acuerdo a la metodología CRISP-DM, la cual es usada para el desarrollo de proyectos de minería de datos. Resultados: El trabajo realizado permitió una reducción del tiempo de obtención de los indicadores en un 78%, pasando de un total de 14 horas inicialmente a tan sólo 3 horas, logrando además estimar los ingresos perdidos mensualmente por servicios no facturados en un monto de MM $ 210, con un error de la estimación menor al 5%. Se espera que, con ayuda de este estudio, la empresa pueda tomar decisiones informadas y mejorar su capacidad de control del proceso de provisión de servicios privados, con el fin de regularizar su flujo de ingreso mensual.

Ortuño, M. (2018) en la tesis titulada MODELO DE ANÁLISIS PREDICTIVO PARA EL MEJORAMIENTO PRESUPUESTARIO DE LA PLANIFICACIÓN LOGÍSTICA DEL CONSEJO NACIONAL ELECTORAL DEL ECUADOR, BASADO EN EL USO DE TÉCNICAS DE MINERIA DE DATOS

Tuvo como objetivo: Construir un modelo de predicción mediante la obtención de patrones de comportamiento para la proyección de indicadores presupuestarios

del material electoral en el Consejo Nacional Electoral del Ecuador – CNE. Metodología: se utilizaron los tipos de investigación exploratoria y descriptiva para recopilar los datos del negocio; estas técnicas fueron aplicadas siguiendo los lineamientos de la metodología de desarrollo CRISP –DM específica para minería de datos. Los resultados muestran que se obtiene una precisión del 89,58 %, a diferencia del 25% obtenido antes de aplicar la solución propuesta. En consecuencia, podemos avizorar la efectividad del modelo, por lo que su aplicación permitirá un análisis eficiente en la toma de decisiones, y de esta forma mejorar la proyección presupuestaria y la optimización del recurso económico para las demás direcciones de la institución. Conclusiones: Se concluyó que las técnicas de patrones de comportamiento permiten mejorar las predicciones en los indicadores presupuestarios del material electoral que maneja la Dirección Nacional de Logística en el Consejo Nacional Electoral CNE. Se validó el modelo predictivo del presupuesto y se obtuvo un promedio de precisión del 89,50%. En base a estos datos se puede mencionar que la hipótesis planteada es verdadera, debido a que mejora notablemente la precisión que maneja el Consejo Nacional Electoral del Ecuador que es del 25% sin aplicar la solución propuesta.

Ayala, E. y Logacho, A. (2018) en la tesis titulada IDENTIFICAR UN MODELO DE DATA MINING PARA DESARROLLAR UN ANÁLISIS PREDICTIVO EN LA ADMINISTRACIÓN INTEGRAL DEL TRABAJO Y EMPLEO DE LAS EMPRESAS ECUATORIANAS

Tuvo como objetivo: Construir un modelo de minería de datos que permita predecir y caracterizar los datos de los sistemas transaccionales (Sistema de Administración Integral de Trabajo y Empleo (SAITE), Sistema Nacional de Control de Inspectores (SINACOI) y Sistema Integral Inspector 2.0 (SGI), para su posterior implementación y obtención de conocimiento que permita dar respuesta a indicadores y comportamiento de la información. La metodología Kimball nos guía en la construcción del data warehouse, los datos que se almacenan en el data warehouse nos sirve como insumo para determinar el modelo predictivo de minería de datos el cual es determinado con la aplicación de la metodología CRISP-DM en todas sus seis fases. Resultado: Investigación que identifica un modelo de data mining con la finalidad de desarrollar un análisis predictivo en la administración integral del trabajo y empleo de las empresas ecuatorianas, así como reconocer los patrones de comportamiento que se tienen las empresas en la contratación de talento humano. Conclusiones: La construcción de un Data Warehouse ayuda a mejorar el tratamiento de los datos para convertirlos en información, ya que los datos que se encuentran limpios e integrados, son utilizados para realizar explotación de datos con cualquier herramienta BI que sirven como apoyo a las autoridades del Ministerio de Trabajo en la toma de decisiones, adicional estos datos son utilizados para realizar el proyecto de minería de datos y de esta forma se evitan que los datos vayan al proyecto con ruido.

2.2. MARCO TEÓRICO

2.2.1. MINERÍA DE DATOS

Perez, C. (2007) indica:

> La minería de datos puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos. La disponibilidad de grandes volúmenes de información y el uso generalizado de herramientas informáticas ha transformado el análisis de datos orientándolo hacia determinadas técnicas especializadas englobadas bajo el nombre de minería de datos o Data Mining.

Asimismo, Ng, K. y Liu, H. (2000) precisan:

> El uso potencial del Data Mining en las empresas es identificar nuevas oportunidades de negocio, adaptar los productos ofrecidos o encontrar los clientes más valiosos con el fin de retenerlos, y de esta manera aumentar los ingresos y reducir las pérdidas o costos. Al determinar las características de los buenos clientes (profiling), las empresas pueden enfocarse en aquellos de características similares y diseñar productos o servicios acordes a sus necesidades.

2.2.2. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, “los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento**.”** (Weiss & Indurkhya, 1998)

"Los algoritmos supervisados o predictivos predicen el valor de un atributo (etiqueta) a partir de un conjunto de datos, conocidos otros atributos (atributos descriptivos)" (Tuya, Ramos, & Dolado, 2007).

"Los algoritmos no supervisados o de descubrimiento de conocimiento realizan tareas descriptivas como el descubrimiento de patrones y tendencias en los datos actuales (no utilizan datos históricos)" (Tuya, Ramos, & Dolado, 2007).

Tabla 2.1: Clasificación de las técnicas de minería de datos

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

Fuente: Traducido de Weiss, S. y Indurkhya, N. (1998)

2.2.2.1. TÉCNICAS DE MINERÍA SUPERVISADAS

2.2.2.1.1. ARBOLES DE DECISIÓN

"Es una técnica útil para problemas en los que se presentan decisiones secuenciales. Aunque esta técnica es de mayor utilidad para situaciones en que el riesgo está presente también es empleada en condiciones de certeza." (Yanet, 2012)

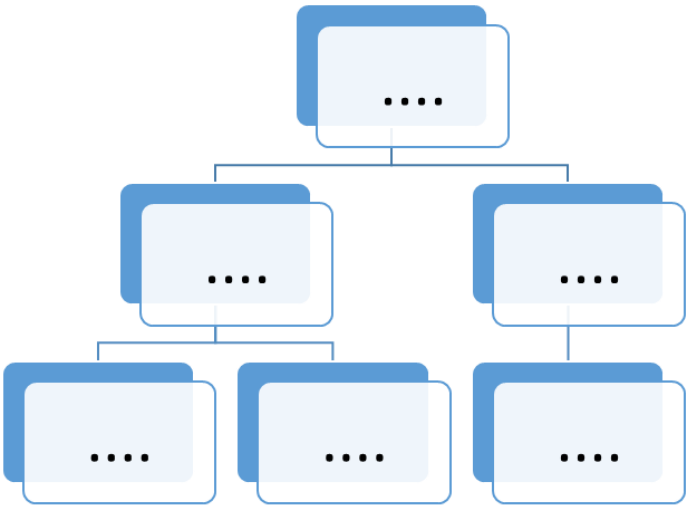
Asimismo, Edelstein H. (1999) indica que:

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

Son una manera de representar una serie de reglas que llevan hacia una clase o valor de los datos, y se utilizan para examinarlos y realizar predicciones. Los árboles de decisión poseen una estructura formada por:

- Nodos, que corresponden a los nombres o identificadores de los atributos que caracterizan al conjunto de datos. El nodo inicial o nodo raíz contiene la muestra total de atributos que definen a los datos.
- Ramas, representan a las variables de decisión o las condiciones que cumplen los objetos para separarse unos de otros.
- Hojas, que son finalmente los conjuntos o grupos de datos resultantes de la división que realiza el algoritmo.

Figura 2.1: Árbol de decisión



Elaborado por el equipo de trabajo

El algoritmo realiza una clasificación discreta de los objetos, determinando a qué clase pertenece, mediante la decisión de qué rama escoger. Para esto, se asume que los grupos o clases que se formarán serán disjuntas, es decir, una instancia u objeto no puede pertenecer a dos clases a la vez. Esta misma condición se cumplirá

para cada partición o sub-árbol que se forme, característica particular que tienen los árboles de decisión conocida como propiedad exhaustiva. Existen diversos algoritmos de aprendizaje que se pueden utilizar para obtener un árbol de decisión. El algoritmo utilizado puede determinar aspectos como la compatibilidad con el tipo de variables de entrada y salida, el procedimiento de evaluación de la distancia entre los grupos generados en cada división, y también la cantidad de ramas que se obtengan cada vez que un nodo se divide.

2.2.2.1.2. REDES NEURONALES

Kasabov, N. (1996) señala que:

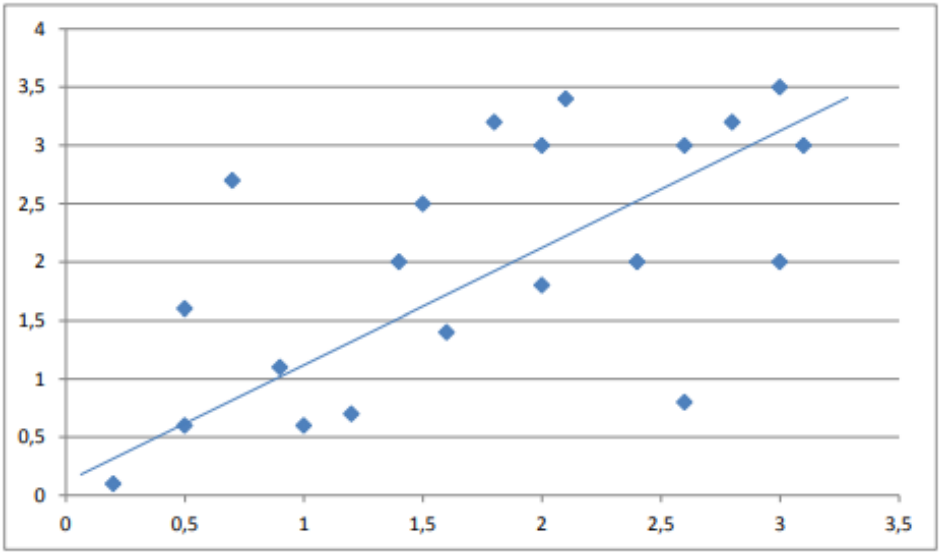
Este modelo de minería de datos es una de las estrategias más populares para aprendizaje supervisado y clasificación. Sin embargo, debido a la complejidad que posee, no se puede saber con exactitud el origen de sus resultados, lo que es una dificultad a la hora de explicar su funcionamiento. En un sentido directo, una red neuronal artificial (o denominada simplemente red neuronal, o ANN) consiste en procesar elementos (llamados neuronas) y las conexiones entre ellos con coeficientes(pesos) ligados a las conexiones, las cuales constituyen una estructura neuronal, y un entrenamiento y algoritmos recordatorios adjuntos a la estructura.

2.2.2.1.3. REGRESIÓN

La regresión, en la actualidad, según Gujarati, D. (2003):

El estudio de la dependencia de la variable dependiente, respecto a una o más variables (las variables explicativas), con el objetivo de estimar y/o predecir la media o valor promedio poblacional de la primera en términos de los valores conocidos o fijos (en muestras repetidas) de las últimas.

Figura 2.2: Regresión lineal



Elaborado por el equipo de trabajo

Devore, J. (2008) también indica que:

El modelo de pronóstico de regresión lineal permite hallar el valor esperado de una variable aleatoria a cuando b toma un valor específico. La aplicación de este método implica un supuesto de linealidad cuando la demanda presenta un comportamiento creciente o decreciente, por tal razón, se hace indispensable que previo a la selección de este método exista un análisis de regresión que determine la intensidad de las relaciones entre las variables que componen el modelo.

2.2.2.1.4. REGRESIÓN LINEAL SIMPLE

"La regresión lineal simple se basa en estudiar los cambios en una variable, no aleatoria, afectan a una variable aleatoria, en el caso de existir una relación funcional entre ambas variables que puede ser establecida por una expresión lineal." (Moral Peláez, 2012)

2.2.2.1.5. REGRESIÓN LINEAL MÚLTIPLE

"La regresión lineal permite trabajar con una variable a nivel de intervalo o razón, así también se puede comprender la relación de dos o más variables y permitirá relacionar mediante ecuaciones, una variable en relación a otras variables llamándose Regresión múltiple." (Rojo Abuín, 2017)

"La regresión lineal múltiple es cuando dos o más variables independientes influyen sobre una variable dependiente ejemplo: $Y=f(x,w,z)$." (Rojo Abuín, 2017)

2.2.2.1.6. SERIES TEMPORALES

Gonzales, M. (2009) indica que:

> Una serie temporal es una secuencia ordenada de observaciones cada una de las cuales está asociada a un momento de tiempo. (…) La mayoría de los métodos estadísticos elementales suponen que las observaciones individuales que forman un conjunto de datos son realizaciones de variables aleatorias mutuamente independientes. En general, este supuesto de independencia mutua se justifica por la atención prestada a diversos aspectos del experimento, incluyendo la extracción aleatoria de la muestra de una población más grande, la asignación aleatoria del tratamiento a cada unidad experimental, etc. Además, en este tipo de datos (tomamos una muestra aleatoria simple de una población más grande) el orden de las observaciones no tiene mayor importancia.

2.2.2.2. TÉCNICAS DE MINERÍA NO SUPERVISADAS

Molina, J. y García, J. (2011) indican que:

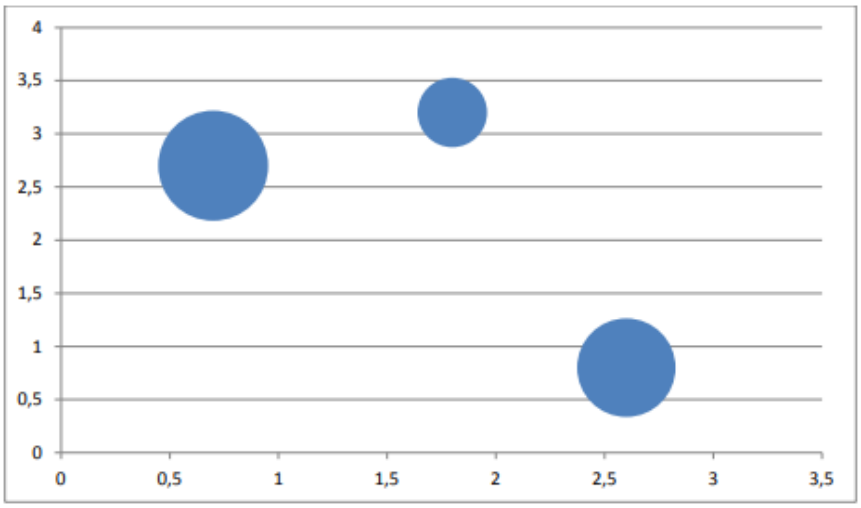
> Los métodos descriptivos no precisan registros de datos o sucesos, no dependen de los patrones obtenidos para detectar reglas, correlaciones y asociaciones.

Podemos obtener información prácticamente al momento de la información que tenemos.

Podemos obtener información prácticamente al momento de la información que tenemos.

El *Clustering* o agrupamiento, es un método mediante el cual descubrimos grupos y estructuras en los datos y que en cierta medida son parecidos o cumplen características similares sin utilizar estructuras conocidas en los datos.

Figura 2.3: Clustering o Agrupamiento



Fuente: Adaptado de Molina, J. y García, J. (2011)

La clasificación ABC, nos ayuda a clasificar los ítems en diferentes grupos basándose en valores y criterios cuantitativos. Por ejemplo, para clasificar a los comerciales de la empresa según el número de ventas realizadas, o del importe total de sus ventas.

Análisis asociativo, o comúnmente conocido como "análisis de la cesta de la compra" tiene como objetivo encontrar patrones, particularmente en procesos de

negocio, y formular reglas aplicables, como, por ejemplo, si un cliente compra hamburguesas, dicho cliente compra también pan de hamburguesa.

El análisis aproximativo incluye tres técnicas diferentes. Encontramos:

- Tablas ponderadas.
- Regresión lineal.
- Regresión no-lineal.

Aunque las técnicas trabajan de forma diferente, el objetivo final de todas ellas es el de aproximar un valor para un atributo específico.

2.2.3. EL PROCESO KDD

Maimon, O. y Rokach, L. (2010), indican que:

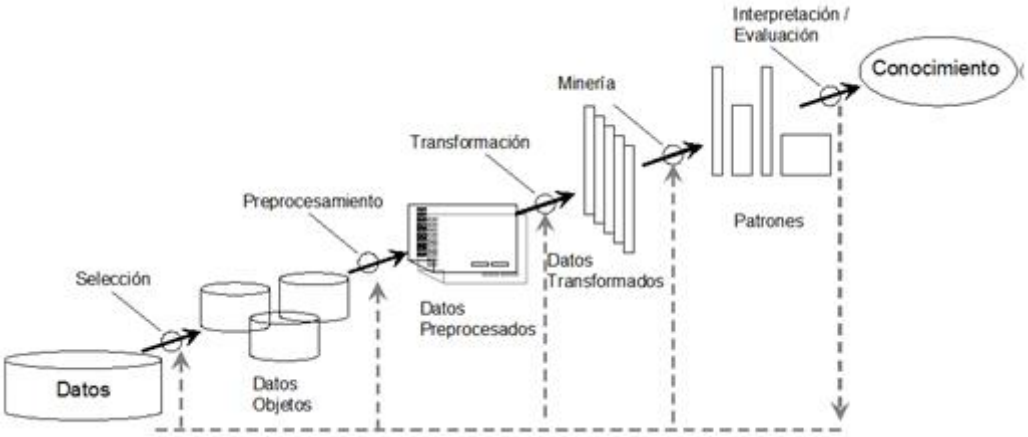
La Minería de Datos en realidad es el núcleo de todo un proceso llamado Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases – KDD) el cual es un proceso metodológico para encontrar un "modelo" válido, útil y entendible que describa patrones de acuerdo a la información, y como modelo entendemos que es la representación que intenta explicar ese patrón en los datos. Es importante mencionar que hablar de "modelo" como fórmula mágica no significa que existe una maestra para cualquier problemática, sino todo lo contrario, pues existen muchos métodos o algoritmos que podrían satisfacer las necesidades dependiendo de los objetivos del estudio y de los datos que se quieran analizar.

Estos pasos se dividen en 9 que son:

1. Abstracción del escenario.

2. Selección de datos.
3. Limpieza y preprocesamiento.
4. Transformación de los datos.
5. Elección de tareas de minería de datos.
6. Elección del algoritmo.
7. Aplicación del algoritmo.
8. Evaluación e interpretación.
9. Entendimiento del conocimiento.

Figura 2.4: El proceso KDD



Fuente: Maimon, O., & Rokach, L. (2010). The Process of Knowledge Discovery in Databases. [Imagen]

Antes de esto definamos un conjunto de datos, el cual es una colección de información, ya sea cuantitativa o cualitativa y que está compuesto por variables o atributos (columnas) que representan las propiedades de un fenómeno o suceso,

y casos (filas) que significan los diferentes sucesos que se presentaron en el escenario. Esto constituye la materia prima del KDD.

2.2.3.1. ABSTRACCIÓN DEL ESCENARIO

Maimon, O. y Rokach, L. (2010), indican:

No todo es matemática y estadística, sino entender la problemática a la que nos vamos a enfrentar y tener contexto para proponer soluciones viables y reales, ya que me ha tocado ver propuestas absurdas. Es importante conocer las propiedades, limitaciones y reglas del escenario en estudio, para posteriormente definir las metas a alcanzar.

2.2.3.2. SELECCIÓN DE LOS DATOS

Maimon, O. y Rokach, L. (2010), indican:

Del conjunto de datos recolectados y ya definidos los objetivos por alcanzar, se deben elegir datos disponibles para realizar el estudio e integrarlos en uno solo que puedan favorecer a llegar a alcanzar a los objetivos del análisis. Muchas veces esta información puede encontrarse en una misma fuente (centralizado) o pueden estar distribuidos.

2.2.3.3. LIMPIEZA Y PREPROCESAMIENTO

Maimon, O. y Rokach, L. (2010), indican:

En esta etapa se determina la confiabilidad de la información, es decir, realizar tareas que garanticen la utilidad de los datos. Para esto se hace la limpieza de datos (tratamiento de datos perdidos o remover valores atípicos). Esto implica eliminar variables o atributos con datos faltantes o eliminar información no útil para este

tipo de tareas como el texto (aunque puede utilizarse para hacer Minería de Texto, que es otro asunto).

2.2.3.4. TRANSFORMACIÓN DE LOS DATOS

"En esta etapa se mejora la calidad de los datos con transformaciones que involucran ya sea reducción de dimensionalidad (disminuir la cantidad de variables del conjunto de datos) o bien transformaciones como por ejemplo convertir los valores que son números a categóricos (discretización)." (Maimon & Rokach, 2010)

2.2.3.5. ELECCIÓN DE TAREA DE MINERÍA DE DATOS

Maimon, O. y Rokach, L. (2010), indican:

Fase en la que se refiere a elegir el paradigma apropiado de Minería de Datos, ya sea la clasificación, regresión o agrupación, según los objetivos que se haya planteado para la investigación (predicción o descripción), la primera ocupada para encontrar un modelo que sea utilizada para casos futuros y desconocidos; mientras que la segunda solo para observar su comportamiento.

2.2.3.6. ELECCIÓN DEL ALGORITMO DE MINERÍA DE DATOS

Maimon, O. y Rokach, L. (2010), indican:

Posteriormente se procede a seleccionar la técnica o algoritmo, o incluso más de uno para la búsqueda del patrón y obtener conocimiento. El meta-aprendizaje se enfoca en explicar la razón por la que un algoritmo funciona mejor en determinadas problemáticas, y para cada técnica existen diferentes posibilidades de cómo seleccionarlas. Cada algoritmo tiene su propia esencia, su propia manera de trabajar y obtener los resultados, por lo que es recomendable conocer las propiedades de aquellos candidatos a utilizar y ver cual se ajusta mejor a los datos.

2.2.3.7. APLICACIÓN DEL ALGORITMO

Maimon, O. y Rokach, L. (2010), indican:

Por fin, una vez seleccionado las técnicas el paso siguiente es aplicarlo a los datos ya seleccionados, limpiados y procesados. Es posible que la ejecución de los algoritmos sean varias intentando ajustar los parámetros que optimicen los resultados. Estos parámetros varían de acuerdo al método seleccionado.

2.2.3.8. EVALUACIÓN

Maimon, O. y Rokach, L. (2010), indican:

Una vez aplicado los algoritmos al conjunto de datos, procedemos a evaluar los patrones que se generaron y el rendimiento que se obtuvo para verificar que cumpla con las metas planteadas en las primeras fases. Para realizar esta evaluación existe una técnica que se llama Validación Cruzada, el cual realiza una partición de los datos dividiéndose en entrenamiento (que servirán para crear el modelo) y prueba (que serán utilizados para ver que en verdad funciona el algoritmo y realiza su trabajo bien).

2.2.3.9. APLICACIÓN

Maimon, O. y Rokach, L. (2010), indican:

Si todos los pasos se siguen correctamente y los resultados de la evaluación se satisfacen, la última etapa es simplemente aplicar el conocimiento encontrado al contexto y comenzar a resolver sus problemáticas. Si de lo contrario, los resultados no son satisfactorios entonces es necesario regresar a las anteriores etapas a realizar algún ajuste, analizando desde la selección de los datos hasta la etapa de evaluación.

2.2.4. MODELO PREDICTIVO

Hernández, J. Ramírez, M. y Ferri, C. (2004) indican:

El modelo predictivo se emplea para estimar valores futuros de variables de interés. El proceso se basa en la información histórica de los datos, mediante las cuales se predice el comportamiento de los datos, ya sea mediante clasificaciones, categorizaciones o regresiones. El atributo a predecir se le conoce como variable dependiente u objetivo, mientras que los atributos utilizados para realizar la predicción se llaman variables independientes o de exploración.

Sumathi, S. y Sivanandam, S. (2006) indican:

El modelado predictivo es a menudo un objetivo de alto nivel de la minería de datos en la práctica. Después de describir el problema del modelado predictivo, nos enfocamos en dos clases de algoritmos: métodos de árbol de decisión y máquinas de vectores de soporte. La entrada a los algoritmos de modelado predictivo es un conjunto de datos de registros de entrenamiento. El objetivo es construir un modelo que prediga un valor de atributo designado a partir de los valores de los otros atributos.

2.2.5. MODELO DESCRIPTIVO

Sumathi, S. y Sivanandam, S. (2006) indican:

En el modelo descriptivo se identifican patrones que describen los datos mediante tareas. Destacan que mediante este modelo se identifican patrones que explican o resumen el conjunto de datos, siendo estos útiles para explorar las propiedades de los datos examinados. Los modelos descriptivos siguen un tipo de aprendizaje no

supervisado, que consiste en adquirir conocimiento desde los datos disponibles, sin requerir influencia externa que indique un comportamiento deseado al sistema.

2.2.6. METODOLOGÍAS DE MINERÍA DE DATOS

Existen diferentes metodologías de aplicación de minería de datos, sin embargo, las metodologías más usadas son CRISP-DM y SEMMA.

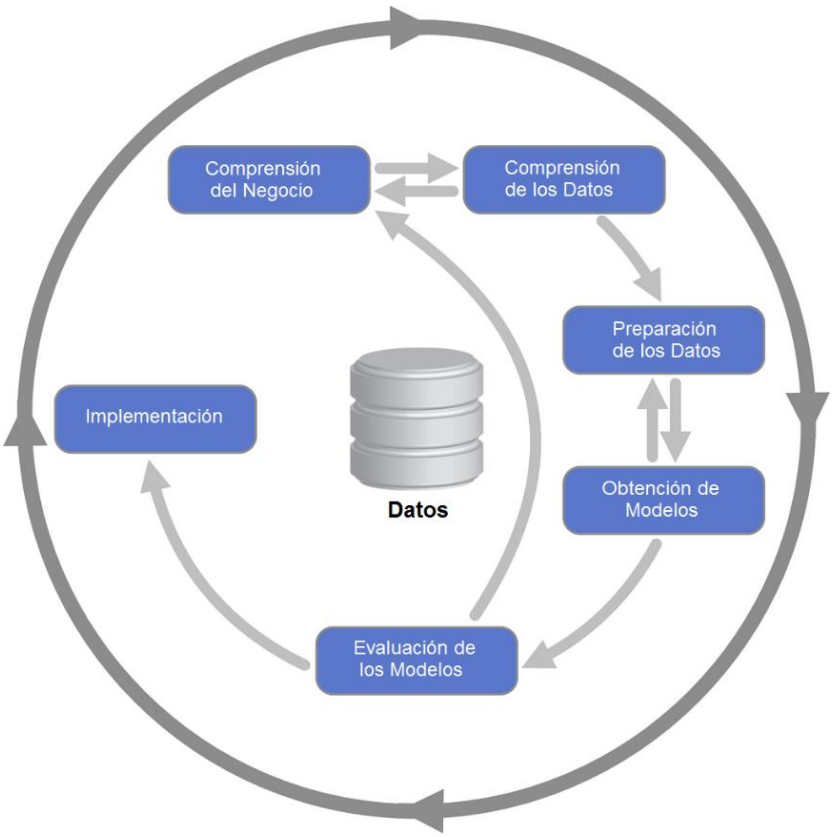
2.2.6.1. METODOLOGÍA CRISP-DM

IBM (2012), afirma:

> CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción no es posible identificar todas las relaciones; las relaciones podrían existir entre cualquier tarea según los objetivos, el contexto, y el interés del usuario sobre los datos.
>
> La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de vida del proyecto de minería de datos consiste en seis fases mostradas en la figura siguiente.

Figura 2.5: Ciclo de vida CRISP-Dm



Fuente: BIZMETRIKS. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen]. Recuperado de http://www.bizmetriks.com/metodologia.html

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

El círculo externo en la figura simboliza la naturaleza cíclica de los proyectos de análisis de datos. El proyecto no se termina una vez que la solución se despliega.

La información descubierta durante el proceso y la solución desplegada pueden producir nuevas iteraciones del modelo. Los procesos de análisis subsecuentes se beneficiarán de las experiencias previas.

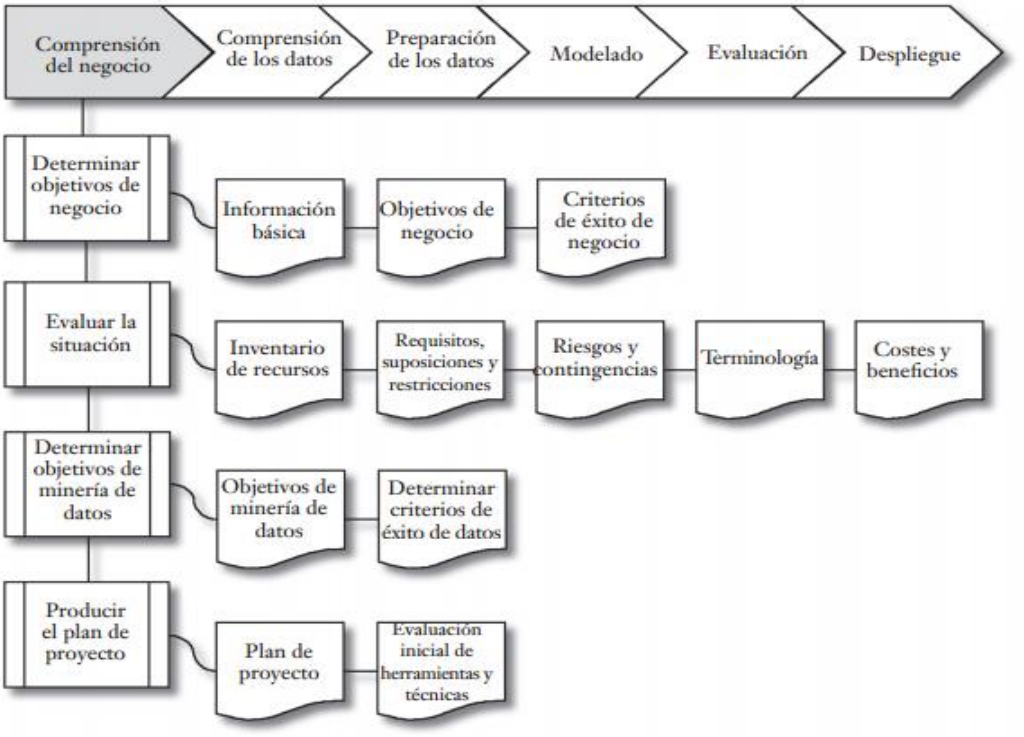
A continuación, vamos a describir brevemente cada una de las fases.

2.2.6.1.1. COMPRENSIÓN DEL NEGOCIO

"Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto. Después se convierte este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos." (BIZMETRIKS, 2013)

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

Figura 2.6: Fase CRISP-DM – Comprensión del negocio



Fuente: Bizmetriks. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen]. Recuperado de http://www.bizmetriks.com/metodologia.html

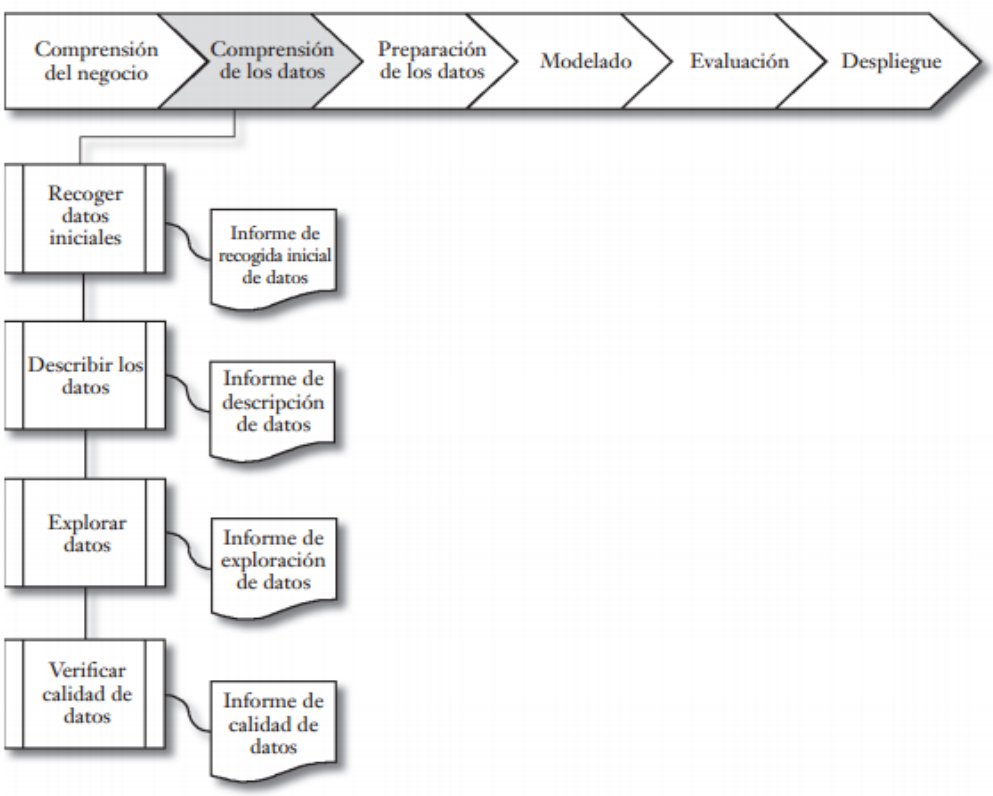
2.2.6.1.2. COMPRENSIÓN DE LOS DATOS

Según BIZMETRIKS (2013), indica:

La fase de entendimiento de datos comienza con la colección de datos inicial y continúa con las actividades que permiten familiarizarse con los datos, identificar los problemas de calidad, descubrir conocimiento preliminar sobre los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

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

Figura 2.7: Fase CRISP-DM – Comprensión de los datos



Fuente: Bizmetriks. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen].

Recuperado de http://www.bizmetriks.com/metodologia.html

2.2.6.1.3. PREPARACIÓN DE LOS DATOS

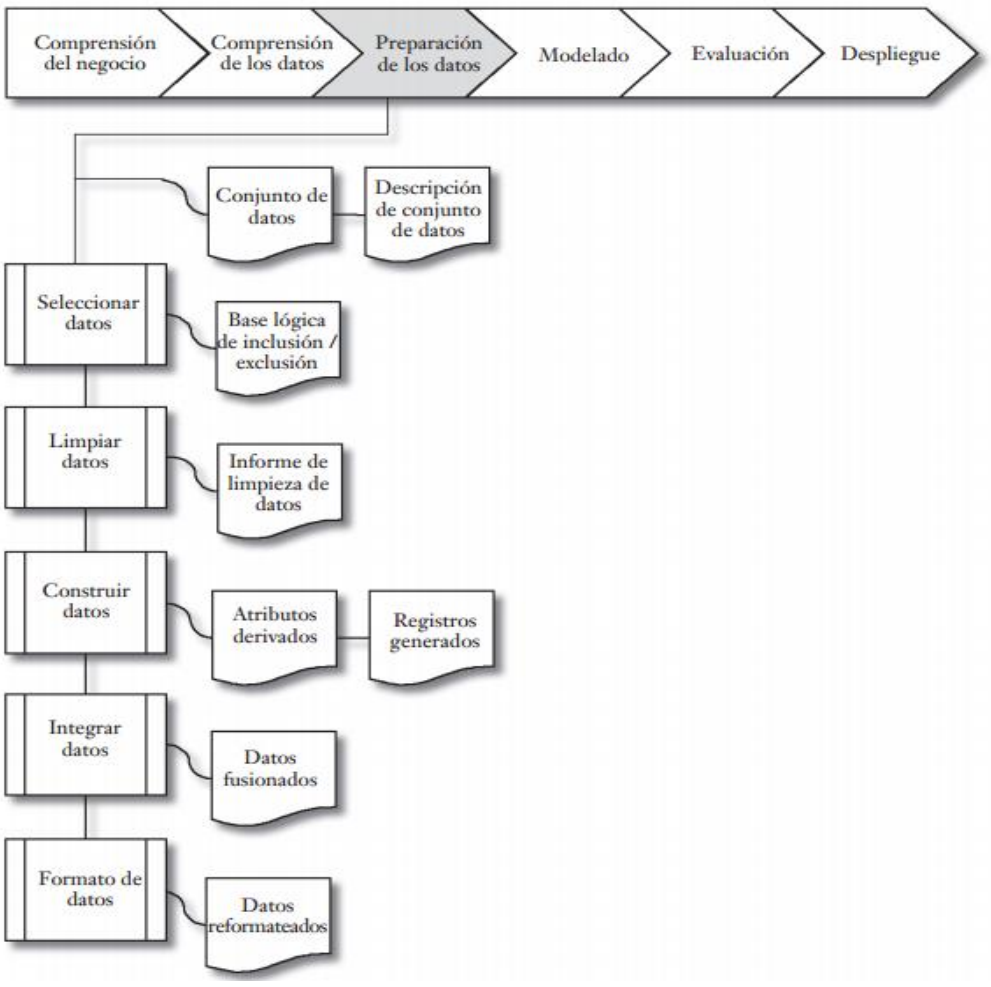
Según BIZMETRIKS (2013), indica:

La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto final de datos (los datos que se utilizarán en las herramientas de modelado) a partir de los datos en bruto iniciales. Las tareas incluyen la

selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.

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

Figura 2.8: Fase CRISP-DM – Preparación de los datos



Fuente: Bizmetriks. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen].

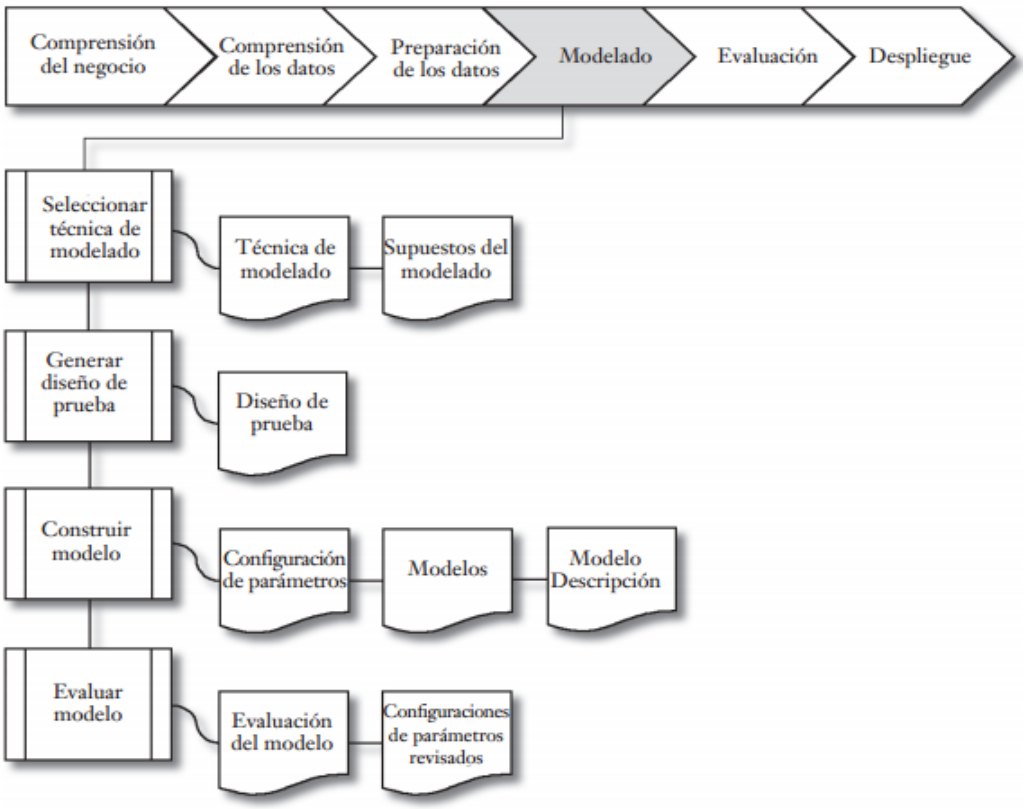
Recuperado de http://www.bizmetriks.com/metodologia.html

2.2.6.1.4. MODELADO

Según BIZMETRIKS (2013), indica:

En esta fase, se seleccionan y aplican las técnicas de modelado que sean pertinentes al problema (cuantas más mejor), y se calibran sus parámetros a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Por lo tanto, casi siempre en cualquier proyecto se acaba volviendo a la fase de preparación de datos.

Figura 2.9: Fase CRISP-DM – Modelado



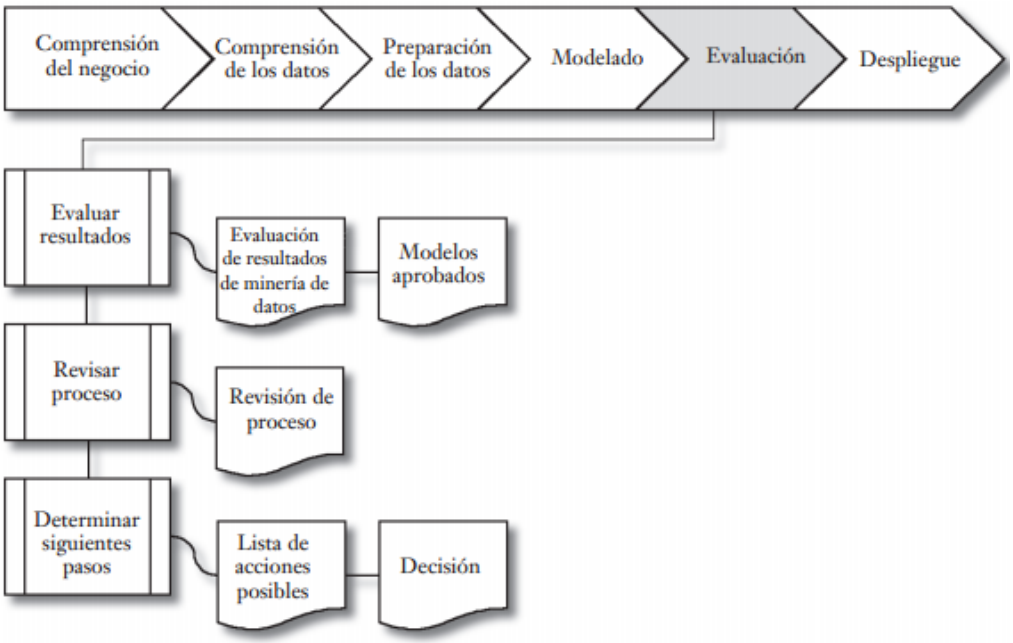
Fuente: Bizmetriks. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen]. Recuperado de http://www.bizmetriks.com/metodologia.html

2.2.6.1.5. EVALUACIÓN

En esta etapa en el proyecto, se han construido uno o varios modelos que parecen alcanzar calidad suficiente desde la perspectiva de análisis de datos, BIZMETRIKS (2013), indica:

> Antes de proceder al despliegue final del modelo, es importante evaluarlo a fondo y revisar los pasos ejecutados para crearlo, comparar el modelo obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no haya sido considerada suficientemente. Al final de esta fase, se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

Figura 2.10: Fase CRISP-DM – Evaluación



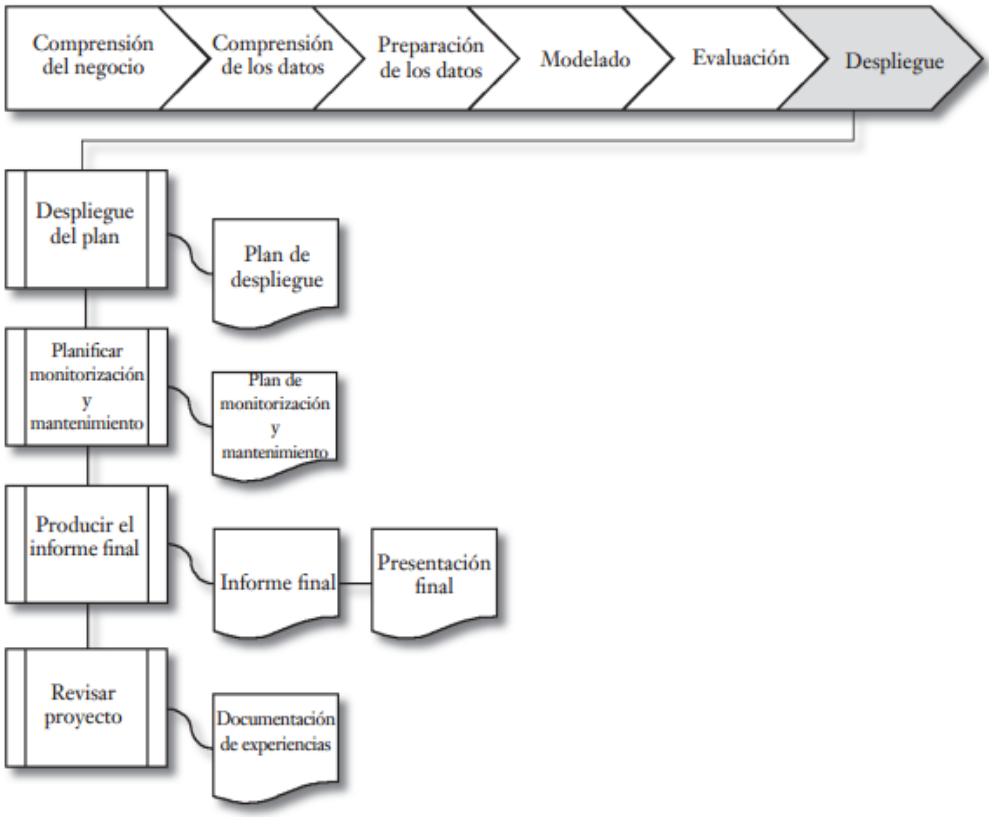
Fuente: Bizmetriks. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen]. Recuperado de http://www.bizmetriks.com/metodologia.html

2.2.6.1.6. DESPLIEGUE

Según BIZMETRIKS (2013), indica:

Generalmente, la creación del modelo no es el final del proyecto. Incluso si el objetivo del modelo es de aumentar el conocimiento de los datos, el conocimiento obtenido tendrá que organizarse y presentarse para que el cliente pueda usarlo. Dependiendo de los requisitos, la fase de desarrollo puede ser tan simple como la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

Figura 2.11: Fase CRISP-DM – Implementación



Fuente: Bizmetriks. (2013). Bizmetriks - Descubriendo conocimiento. [Imagen]. Recuperado de http://www.bizmetriks.com/metodologia.html

2.2.6.2. METODOLOGÍA SEMMA

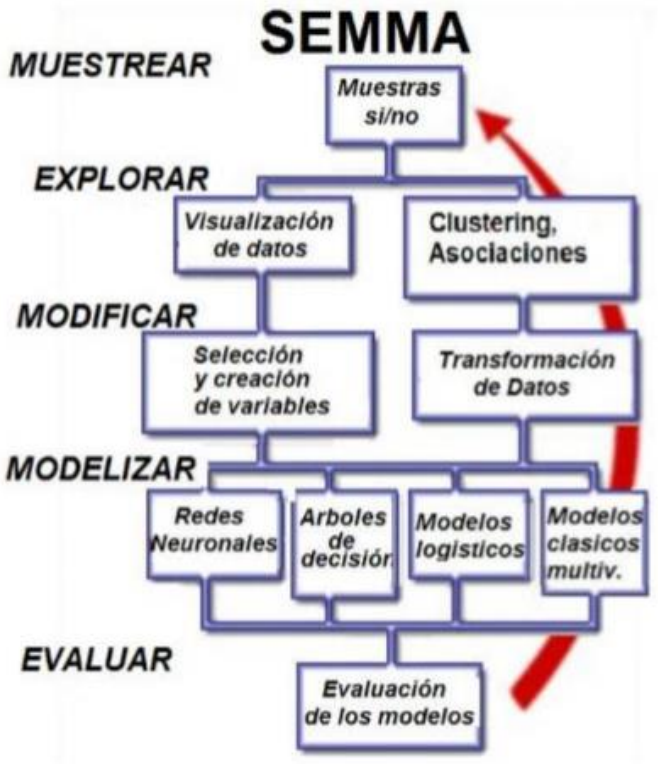
SAS Institute Inc. (2017), define:

La minería de datos como el proceso de muestreo, exploración, modificación, modelado y evaluación (SEMMA) de grandes cantidades de datos para descubrir patrones previamente desconocidos que pueden utilizarse como una ventaja comercial. El proceso de minería de datos es aplicable en una variedad de industrias y proporciona metodologías para problemas comerciales tan diversos como detección de fraude, hogar, retención y desgaste de clientes, marketing de bases de datos, segmentación de mercado, análisis de riesgos, análisis de afinidad, satisfacción del cliente, predicción de bancarrota y cartera análisis.

- El software Enterprise Miner es un producto integrado que proporciona una solución comercial integral para la minería de datos.
- Una interfaz gráfica de usuario (GUI) proporciona una interfaz fácil de usar para el proceso de minería de datos SEMMA:
- Muestree los datos creando una o más tablas de datos. Las muestras deben ser lo suficientemente grandes como para contener la información significativa, pero lo suficientemente pequeñas como para procesarlas.
- Explore los datos buscando relaciones anticipadas, tendencias no anticipadas y anomalías para obtener comprensión e ideas.
- Modifique los datos creando, seleccionando y transformando las variables para enfocar el proceso de selección del modelo.

- Modele los datos utilizando las herramientas analíticas para buscar una combinación de datos que prediga de manera confiable el resultado deseado.
- Evalúe los datos evaluando la utilidad y confiabilidad de los hallazgos del proceso de minería de datos.

Figura 2.12: Esquema SEMMA



Fuente: Universidad Complutense Madrid (2016). *Proceso de Metodología SEMMA*. [Imagen].

Puede incluir o no todos los pasos de SEMMA en su análisis, y puede ser necesario repetir uno o más de los pasos varias veces antes de estar satisfecho con los resultados. Después de completar la fase de evaluación del proceso SEMMA, aplica la fórmula de puntuación de uno o más modelos de campeón a nuevos datos que pueden contener o no el objetivo. Obtener nuevos datos que no están

disponibles en el momento de la capacitación del modelo es el resultado final de la mayoría de los problemas de minería de datos.

El proceso de minería de datos SEMMA está impulsado por un diagrama de flujo de proceso, que puede modificar y guardar. La GUI está diseñada de tal manera que el analista de negocios que tiene poca experiencia estadística puede navegar a través de la metodología de minería de datos, mientras que el experto cuantitativo puede ir "detrás de escena" para ajustar y ajustar el proceso analítico.

Enterprise Miner contiene una colección de herramientas de análisis sofisticadas que tienen una interfaz común fácil de usar que puede usar para crear y comparar múltiples modelos. Las herramientas estadísticas incluyen clustering, mapas autoorganizados (mapas de Kohonen), selección de variables, árboles, regresión lineal y logística, y redes neuronales. Las herramientas de preparación de datos incluyen detección de valores atípicos, transformaciones variables, imputación de datos, muestreo aleatorio y partición de conjuntos de datos (en conjuntos de datos de entrenamiento, prueba y validación). Las herramientas de visualización avanzadas le permiten examinar rápida y fácilmente grandes cantidades de datos en histogramas multidimensionales y comparar gráficamente los resultados de modelado.

2.2.7. HERRAMIENTAS DE MINERÍA DE DATOS

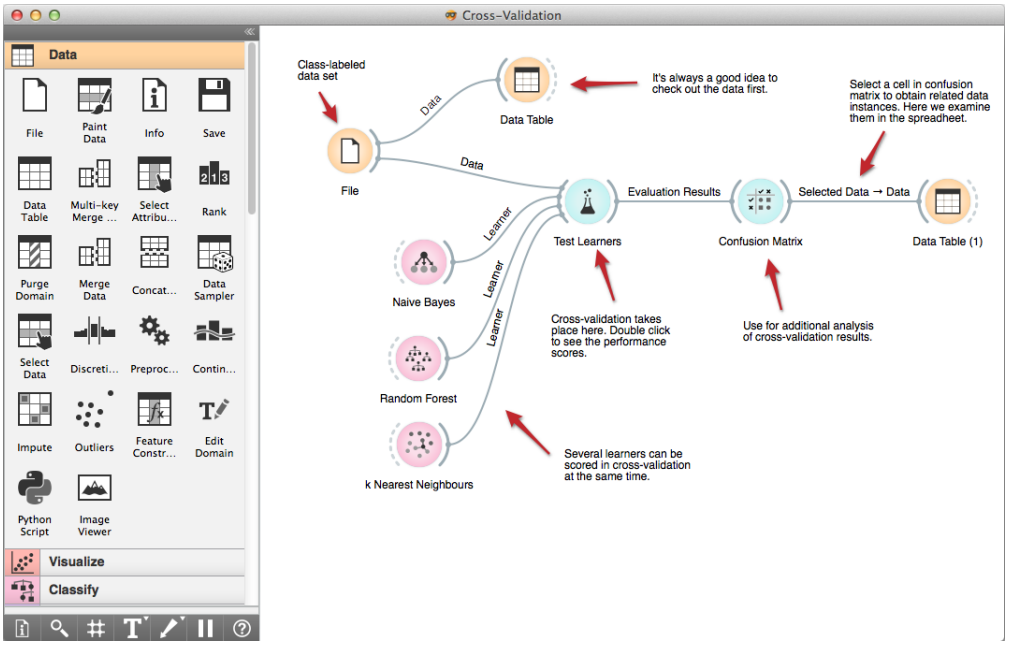
2.2.7.1. ORANGE DATA MINING

De acuerdo a Predictive Analitics Today (2019):

Orange es una herramienta de análisis y visualización de datos de código abierto. Orange se desarrolla en el Laboratorio de Bioinformática de la Facultad de

Informática y Ciencias de la Información de la Universidad de Ljubljana, Eslovenia, junto con la comunidad de código abierto. La minería de datos se realiza mediante programación visual o secuencias de comandos de Python.

Figura 2.13: Interfaz de Orange Data Mining



Fuente: PAT RESEARCH (n.d.). Orange Data Mining Interface. [Imagen]. Recuperado de https://www.predictiveanalyticstoday.com

La herramienta tiene componentes para aprendizaje automático, complementos para bioinformática y minería de texto y está repleta de características para análisis de datos. Orange es una biblioteca de Python. Los scripts de Python pueden ejecutarse en una ventana de terminal, entornos integrados como PyCharm y PythonWin, o shells como iPython. Orange consiste en una interfaz de lienzo en la que el usuario coloca widgets y crea un flujo de trabajo de análisis de datos. Los widgets ofrecen funcionalidades básicas como leer los datos, mostrar una

tabla de datos, seleccionar características, predecir el entrenamiento, comparar algoritmos de aprendizaje.

2.2.7.1.1. CARACTERÍSTICAS

- Código abierto.
- Visualización interactiva de datos.
- Programación visual.
- Admite capacitación práctica e ilustraciones visuales.
- Complementos extiende la funcionalidad.

2.2.7.1.2. BENEFICIOS

- Para todos: principiantes y profesionales.
- Ejecute análisis de datos simples y complejos.
- Cree gráficos hermosos e interesantes.
- Utilícelos en una clase de análisis de datos.
- Acceda a funciones externas para análisis avanzados.

2.2.7.2. R SOFTWARE ENVIROMENT

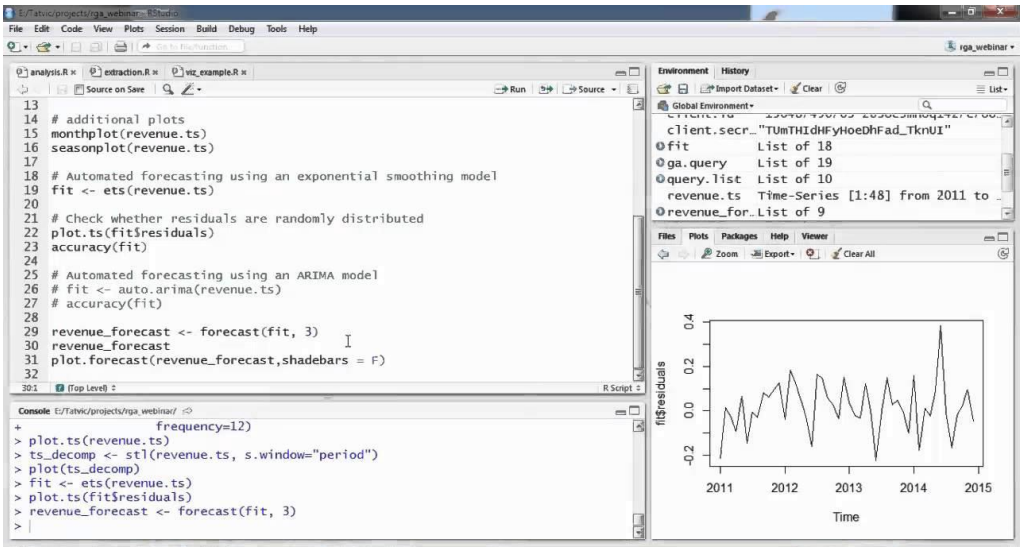
De acuerdo a Predictive Analitics Today (2019):

R es un entorno de software libre para computación estadística y gráficos. Compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS. R es un conjunto integrado de instalaciones de software para la manipulación de datos, el cálculo y la visualización gráfica. Algunas de las funcionalidades incluyen una instalación eficaz de manejo y almacenamiento de datos, un conjunto de operadores para cálculos en matrices, en particular matrices, una colección grande, coherente e integrada de herramientas intermedias para el

análisis de datos, facilidades gráficas para el análisis de datos y visualización directamente en la computadora o en papel, y un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario e instalaciones de entrada y salida.

R es en gran medida un vehículo para desarrollar nuevos métodos de análisis de datos interactivos. Se ha desarrollado rápidamente y se ha ampliado con una gran colección de paquetes. El lenguaje R es ampliamente utilizado entre los estadísticos y mineros de datos para desarrollar software estadístico y análisis de datos.

Figura 2.14: Interfaz R Software Enviroment



Fuente: PAT RESEARCH (n.d.). R Software Enviroment Interface. [Imagen].

Recuperado de https://www.predictiveanalyticstoday.com

2.2.7.2.1. CARACTERÍSTICAS

- Código abierto - Software gratuito.
- Proporciona una amplia variedad de estadísticas (modelado lineal y no lineal, pruebas estadísticas clásicas, análisis de series de tiempo, clasificación, agrupamiento) y técnicas gráficas.
- Facilidad de manejo y almacenamiento de datos efectivos.
- Conjunto de operadores para cálculos en matrices, en particular matrices.
- Colección grande, coherente e integrada de herramientas intermedias para el análisis de datos.
- Facilidades gráficas para el análisis y visualización de datos en pantalla o en papel.
- Lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas e instalaciones de entrada y salida.

2.2.7.2.2. BENEFICIOS

- Trae análisis a sus datos.
- Se ejecuta en una amplia variedad de plataformas: UNIX, Windows, MacOS.
- Software estadístico ampliamente utilizado.
- Fácil de aprender.
- Natural y expresivo.
- Capacidades reconocidas para visualizar datos.

2.2.7.3. WEKA

Es la herramienta que en base a las referencias consultadas es propicia para la elaboración de modelos de esta investigación, de acuerdo a Predictive Analitics Today (2019):

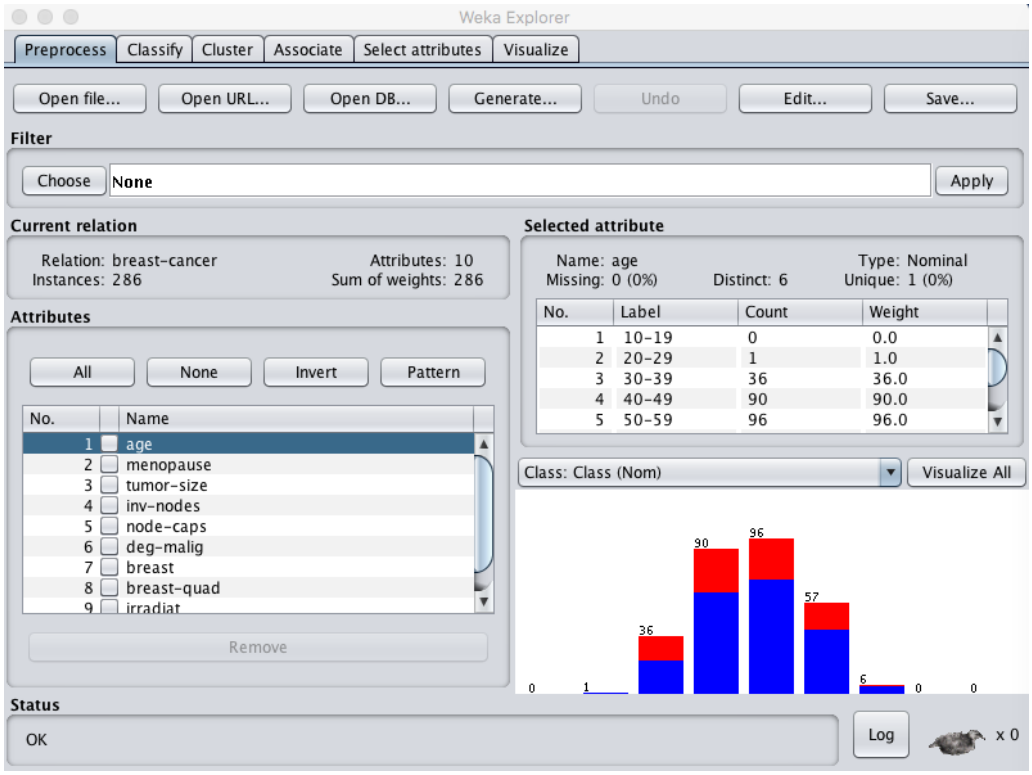
> Weka es una colección de algoritmos de aprendizaje automático para tareas de minería de datos. Los algoritmos se pueden aplicar directamente a un conjunto de datos o invocar desde su propio código Java. Las características de Weka incluyen aprendizaje automático, minería de datos, preprocesamiento, clasificación, regresión, agrupación, reglas de asociación, selección de atributos, experimentos, flujo de trabajo y visualización. Weka está escrito en Java, desarrollado en la Universidad de Waikato, Nueva Zelanda. Todas las técnicas de Weka se basan en el supuesto de que los datos están disponibles como un solo archivo plano o relación, donde cada punto de datos se describe mediante un número flujo de atributos. Weka proporciona acceso a bases de datos SQL utilizando Java Database Connectivity y puede procesar el resultado devuelto por una consulta de base de datos. No es capaz de minería de datos multirrelacionales.
>
> La interfaz de usuario principal de Weka es el Explorer, a la misma funcionalidad también se puede acceder a través de la interfaz de Knowledge Flow basada en componentes y desde la línea de comandos. También está el Experimentador, que permite la comparación sistemática del rendimiento predictivo de los algoritmos de aprendizaje automático de Weka en una colección de conjuntos de datos. La interfaz del Explorador presenta varios paneles que proporcionan acceso a los componentes principales del banco de trabajo, como el panel de preproceso que facilita la importación de datos, el panel de clasificación permite al usuario aplicar

algoritmos de clasificación y regresión, el panel asociado proporciona acceso a los estudiantes de reglas de asociación, el panel de clúster da acceso para las técnicas de agrupamiento, el panel seleccionar atributos proporciona algoritmos para identificar los atributos más predictivos en un conjunto de datos.

Weka proporciona un conjunto completo de herramientas de procesamiento previo de datos, algoritmos de aprendizaje y métodos de evaluación, interfaces gráficas de usuario y un entorno para comparar algoritmos de aprendizaje. Los datos se pueden importar desde un archivo en varios formatos, como ARFF, CSV, C4.5, binario. Los datos también se pueden leer desde una URL o desde una base de datos SQL (usando JDBC). Las herramientas de pre-procesamiento en WEKA se denominan "filtros" y hay filtros disponibles para la discreción, normalización, re-muestreo, selección de atributos, transformación y combinación de atributos.

Los esquemas de aprendizaje implementados son árboles y listas de decisión, clasificadores basados en instancias, máquinas de vectores de soporte, perceptrones de múltiples capas, regresión logística, redes de Bayes. Los meta clasificadores incluidos son códigos de salida de embolsado, refuerzo, apilamiento, corrección de errores, aprendizaje ponderado localmente. Los esquemas implementados son k-Means, EM, Cobweb, X-means, FarthestFirst. Los clústeres se pueden visualizar y comparar con los clústeres "verdaderos". Apriori puede calcular todas las reglas que tienen un soporte mínimo dado y exceden una confianza dada. En Weka, las fuentes de datos, clasificadores, etc. son beans y pueden conectarse gráficamente.

Figura 2.15: Interfaz de WEKA



Fuente: PAT RESEARCH (n.d.). WEKA Interface. [Imagen]. Recuperado de https://www.predictiveanalyticstoday.com

2.2.7.3.1. CARACTERÍSTICAS

- Preprocesamiento de datos.
- Clasificación de datos.
- Regresión de datos.
- Agrupación de datos.
- Reglas de asociación de datos.
- Visualización de datos.

2.2.7.3.2. BENEFICIOS

- Portátil.
- De uso gratuito.
- Fácil de usar.
- Adaptado para crear nuevas formas de diseños de aprendizaje automático.
- Contiene herramientas con múltiples usos.
- Cursos en línea gratuitos disponibles.
- Profesores altamente educados, capacitados y comprometidos.
- Libros y publicaciones extremadamente ingeniosos disponibles.
- Últimas tendencias en inteligencia artificial.

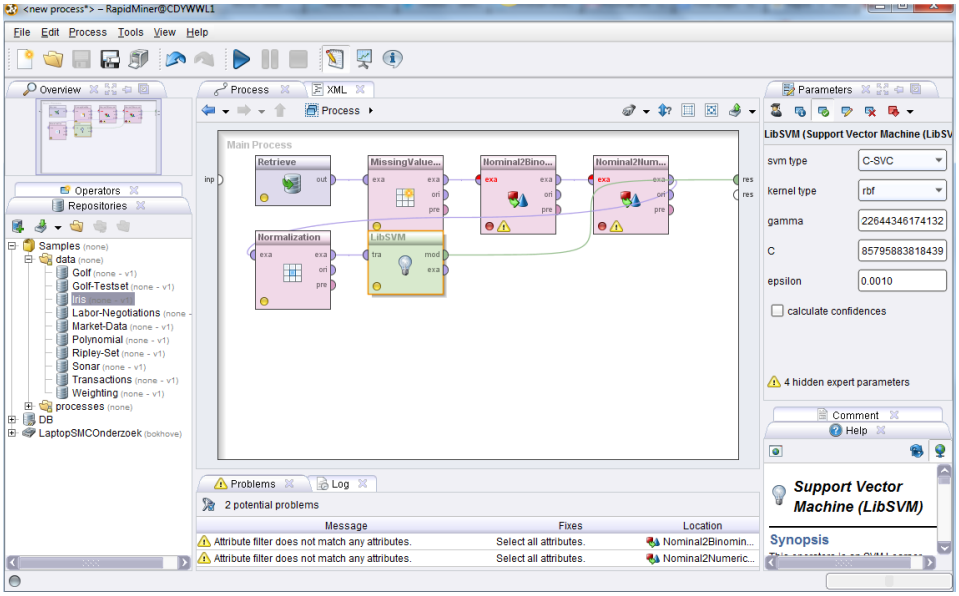
2.2.7.4. RAPIDMINER

Predictive Analitics Today (2019), indica:

RapidMiner Studio ofrece una gran cantidad de funcionalidades para acelerar y optimizar las tareas de exploración, combinación y limpieza de datos, reduciendo el tiempo dedicado a importar y disputar sus datos. RapidMiner proporciona un entorno integrado para la preparación de datos, aprendizaje automático, aprendizaje profundo, minería de texto y análisis predictivo. Se utiliza para aplicaciones comerciales y de negocios, así como para investigación, educación, capacitación, creación rápida de prototipos y desarrollo de aplicaciones, y admite todos los pasos del proceso de aprendizaje automático, incluida la preparación de datos, la visualización de resultados, la validación y optimización de modelos. Cientos de aprendizaje automático, análisis de texto, algoritmos de modelado predictivo, automatización y funciones de control de procesos lo ayudan a construir mejores modelos más rápido que nunca. RapidMiner Studio (las de

datos: 10,000), RapidMiner Server (2 GB de RAM) y RapidMiner Radoop (limitado a un solo usuario) están disponibles en la edición de inicio con limitaciones.

Figura 2.16: Interfaz RapidMiner



Fuente: PAT RESEARCH (n.d.). Rapid Miner Interface. [Imagen]. Recuperado de https://www.predictiveanalyticstoday.com

2.2.7.4.1. CARACTERÍSTICAS

- La multitud de algoritmos de clasificación y regresión facilitan el aprendizaje supervisado.
- La amplia gama de algoritmos de agrupación, similitud y segmentación admiten el aprendizaje no supervisado.
- La integración perfecta de los scripts R y Python en los flujos de trabajo proporciona una mayor extensibilidad.
- Capacidades de modelado y algoritmos de aprendizaje automático.

2.2.7.4.2. BENEFICIOS

- Conéctese a cualquier fuente de datos, cualquier formato, a cualquier escala.
- Descubra rápidamente patrones o problemas de calidad de datos.
- Cree el conjunto de datos óptimo para el análisis predictivo.
- Limpie de manera experta los datos para algoritmos avanzados.

2.2.8. CONTROL DE GESTANTE

Para conocer la evolución y características del proceso de gestación, de acuerdo al Gobierno del Perú – Ministerio de Salud (2018):

> El control prenatal en las gestantes es muy importante para llevar un embarazo sano. Y uno de los procedimientos que contribuyen a tener un buen control prenatal es la ecografía obstétrica, lo cual permitirá observar el desarrollo fetal y las características de la placenta durante la gestación.
>
> Según datos de la Dirección de Salud Sexual y Reproductiva del Ministerio de Salud (Minsa), una gestante debe realizarse al menos tres ecografías en el transcurso de su gravidez, a cargo de un médico especializado en el diagnóstico por imágenes.
>
> La primera ecografía debe ser lo más cercano al diagnóstico del embarazo, utilizando el transductor vaginal (con el cual se puede tener un diagnóstico más preciso de la edad gestacional) o en su defecto la ecografía abdominal, si ha transcurrido el I trimestre o las primeras 12 semanas de embarazo.
>
> La segunda ecografía puede realizarse entre las 21 y 25 semanas de gestación, para conocer el desarrollo del bebé, su posición dentro del vientre y el estado de la placenta.

La tercera ecografía va hacia el final del embarazo (35 a 37 semanas de edad gestacional) y tiene por objeto definir la maduración de la placenta y la posición probable para el parto, así como las características del líquido amniótico.

Otros datos que muestra este procedimiento son: el desarrollo del feto, si probablemente tiene una malformación y cuan vital está, por lo tanto, las gestantes pueden realizarse este procedimiento con toda seguridad, según lo indique el médico tratante.

El especialista advirtió que este tipo de exámenes deben hacerse en establecimientos de salud del sector público o privado que cuenten con un médico especializado en diagnóstico por ecografía general u obstétrica.

- Los especialistas recomiendan que la gestante debe acudir a los controles prenatales mensualmente hasta los 7 primeros meses.
- Cada 15 días entre el 7° y 8° mes.
- Cada semana en el 9° mes de gestación.

2.2.9. MEDIDAS GENERALES DE PREVENCIÓN DE ANEMIA

"La anemia es un problema multifactorial cuyos efectos permanecen en todo el ciclo de la vida. Las medidas de prevención y de tratamiento contempladas en esta norma ponen énfasis en un abordaje integral e intersectorial." (MINISTERIO DE SALUD DEL PERÚ, 2017)

Las medidas de prevención que recomienda el Ministerio de salud son las siguientes (MINISTERIO DE SALUD DEL PERÚ, 2017):

- El equipo de salud debe realizar la atención integral en el control de crecimiento y desarrollo, atención prenatal y puerperio, incluyendo el

despistaje de anemia, a todos los niños, adolescentes, mujeres gestantes y puérperas que reciben suplementos de hierro, en forma preventiva o terapéutica.

- Se debe brindar una adecuada consejería a la madre, familiar o cuidador del niño, adolescente, y a las mujeres gestantes y puérperas, sobre las implicancias y consecuencias irreversibles de la anemia; la importancia de una alimentación variada y con alimentos ricos en hierro de origen animal; y la importancia de la prevención o tratamiento de la anemia.
- Se pondrá énfasis en informar a los padres de niños y adolescentes, a mujeres gestantes y puérperas sobre los efectos negativos de la anemia en el desarrollo cognitivo, motor y el crecimiento, con consecuencias en la capacidad intelectual y de aprendizaje (bajo rendimiento en la escuela o estudios, entre otros) y motora (rendimiento físico disminuido) y con repercusiones incluso en la vida adulta (riesgo de padecer enfermedades crónicas).

2.2.9.1. EN LA GESTACIÓN

"Educación alimentaria que promueva la importancia de una alimentación variada incorporando diariamente alimentos de origen animal como: sangrecita, hígado, bazo y otras vísceras de color oscuro, carnes rojas, pescado." (MINISTERIO DE SALUD DEL PERÚ, 2017)

"Suplementación de la gestante y puérpera con Hierro y Ácido Fólico a partir de la semana 14 de gestación hasta 30 días post-parto." (MINISTERIO DE SALUD DEL PERÚ, 2017)

2.2.9.2. EN EL PARTO

"Pinzamiento y corte tardío del cordón umbilical, a los 2 – 3 minutos después del nacimiento en el recién nacido a término y sin complicaciones. Inicio de la lactancia materna dentro de la primera hora de nacimiento, de manera exclusiva hasta los 6 meses y prolongada hasta los 2 años de edad." (MINISTERIO DE SALUD DEL PERÚ, 2017)

2.2.9.3. PRIMERA INFANCIA, NIÑEZ Y ADOLESCENCIA

MINISTERIO DE SALUD DEL PERÚ (2017) indica que:

Alimentación Complementaria desde los 6 meses de edad durante la niñez y adolescencia que incluya diariamente alimentos de origen animal como sangrecita, bazo, hígado, carnes rojas, pescado, ya que son las mejores fuentes de hierro hemínico.

Suplementación preventiva con Hierro a niños prematuros a partir de los 30 días de nacido y a niños nacidos a término desde el 4to mes hasta los 35 meses.

En localidades con prevalencia de anemia infantil, mayor al 20%, se suplementará a las adolescentes mujeres escolares, en dosis semanal para prevenir la anemia por un periodo de 3 meses por año.

2.2.10. MEDICIÓN DE CONCENTRACIÓN DE HEMOGLOBINA

MINISTERIO DE SALUD DEL PERÚ (2017) indica que:

La medición de la concentración de hemoglobina es la prueba para identificar anemia.

Todo Establecimiento de Salud, de acuerdo al nivel de atención, debe contar con uno de los métodos anteriormente descritos y sus respectivos insumos para la

determinación de hemoglobina o hematocrito. Se deberá realizar el control de calidad de los datos obtenidos por cualquiera de estos métodos. En el caso de hemoglobina, se contará con una solución patrón de concentración de hemoglobina conocida.

Tabla 2.2: Valores normales de concentración de Hb y niveles de anemia

POBLACIÓN	CON ANEMIA HB (G/DL)			SIN ANEMIA
	SEVERA	MODERADA	LEVE	HB (G/DL)
Niño menor de 2 meses		<13.5		13.5-18.5
Gestante 15 años a más.	< 7.0	7.0 – 9.9	10.0 – 10.9	≥ 11.0
Puérpera	< 8.0	8.0 – 10.9	11.0 – 11.9	≥ 12.0

Fuente: Ministerio de Salud (2017). Valores normales de concentración de hemoglobina y niveles de anemia en Niños, Adolescentes, Mujeres Gestantes y Puérperas (hasta 1,000 msnm)

CAPITULO III

MATERIALES Y MÉTODOS

3.1. LUGAR DE ESTUDIO

Esta investigación se realizó en la jurisdicción de atención de la Red de Salud Chucuito - Juli con sus diferentes establecimientos de salud.

- **Hospital Rafael Ortiz Ravines**
- **Microred Pomata** (Batalla, Pomata, Ampatiri, Lampa Grande, Huapaca San Miguel, Tuquina, Collini, Tambillo)
- **Microred Desaguadero** (Desaguadero, Huacullani, Pisacoma, Kelluyo, Carancas, Bajo Llallahua, Callaza, Chacocollo, Alto Llallahua, Totoroma)
- **Microred Zepita** (Zepita, Ancoputo, Izani, Parco Patacollo, Tasapa Patacollo, Sicuyani, Molino Kapia, Villa Chimu, Alto Ayrihuas)
- **Microred Molino** (Pueblo Libre, Pasiri, Caspa Central, Santiago, Molino, Yacango, Ccajje, Casimuyo, Keruma, Rosario De Sorapa, Challapampa, San Juan De Yarihuani, Callacami)

3.2. POBLACIÓN

La población de estudio ha sido definida por las gestantes con domicilio en la jurisdicción de atención de la Red de Salud Chucuito – Juli que cuentan con SIS con posterior parto y producto recién nacido, y está conformada por 484 registros almacenados durante el periodo 2016-2018.

3.3. MUESTRA

Las gestantes dentro la jurisdicción no cumplen con asistir a sus controles y/o exámenes y en muchos casos no se les ubica en las direcciones que registran, por lo que para la investigación se ha tomado en cuenta los siguientes factores:

- Paciente con FUA 015 Diagnostico de embarazo positivo.
- Gestante con FUA 011 Examen de laboratorio de la gestante.
- Gestante con FUA 009 Atención prenatal.
- Paciente con FUA 054 Atención de parto vaginal o FUA 055 Cesárea.

La muestra utilizada es de tipo poblacional, y en función de los criterios a tomar son de 484 registros en el periodo 2016 al 2018, seleccionados por conveniencia para el estudio.

3.4. MÉTODO DE INVESTIGACIÓN

Esta investigación de acuerdo a la caracterización del problema es de tipo descriptivo, ya que en base a datos reales de seguimiento a gestantes se desarrolló el modelo predictivo que nos permite predecir la condición de salud de los recién nacidos.

También se tiene el método es histórico pues requerimos conocer, analizar y evaluar nuestras variables en sus distintas etapas basados a un periodo de evaluación el mismo que estará controlado por diferentes factores de temporada.

El diseño de la investigación es descriptiva simple, pues se han tomado en cuenta la recolección de datos producto de los criterios de muestra, por lo que son observaciones al grupo experimental y en base a su comportamiento se han realizado las observaciones y obtenido los resultados.

Esquema: O → G

Donde:

- G: Grupo experimental (Muestra)
- O: Observación El grupo experimental estará conformado por datos de control de gestantes que cumplan con los criterios de muestra.

3.5. METODOLOGÍA DE MINERÍA DE DATOS

La metodología a utilizar para el desarrollo de minería de datos es CRISP-DM, que tiene como propósito construir variables que sirva como fuente de información para crear el modelo, siguiendo un conjunto de pasos que guíen el proceso que se debe seguir, comprende las siguientes fases:

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación

3.6. TÉCNICA DE MINERÍA DE DATOS

Para el desarrollo de esta investigación se ha seleccionado la técnica de árboles de decisión. “Corresponde a uno de los métodos inductivos de aprendizaje supervisado, el cual realiza divisiones sucesivas del conjunto de datos, utilizando algún criterio de selección, manteniendo organizada su estructura de forma jerárquica, con el fin de maximizar la distancia entre los grupos de datos generados en cada iteración.” (Edelstein, 1999)

Los árboles de decisión a diferencia de otras técnicas (Jiawei, Kamber, & Pein, 2012)

- Facilitan la interpretación de los datos.
- Proporcionan un alto grado de comprensión del conocimiento utilizado en la toma de decisiones.
- Explican el comportamiento respecto a
- una determinada tarea de decisión.
- Reducen el número de variables independientes.
- Permiten establecer la selección del algoritmo de minería de datos.

Todos los algoritmos de clasificación tienen dos etapas: entrenamiento y test. La primera de ellas ajusta el algoritmo de clasificación con una parte del conjunto de datos (conjunto de entrenamiento) y posteriormente se evalúa dicho algoritmo en la etapa de test con el conjunto de datos de test. La división del conjunto de datos suele ser 70% para el entrenamiento y 30% para la evaluación. (Jimenez Vázquez & Gómez Bertoli, 2009)

Los algoritmos utilizados para esta técnica de minería de datos son los siguientes:

3.6.1. ALGORITMO J48

Quinlan R. (1993) indica:

J48 es una implementación open source en lenguaje de programación Java del algoritmo C4.5 en la herramienta Weka de minería de datos.

C4.5 construye árboles de decisión desde un grupo de datos de entrenamiento de la misma forma en que lo hace ID3, usando el concepto de entropía de información. Los datos de entrenamiento son un grupo S = s1,s2,… de ejemplos ya clasificados. Cada ejemplo si = x1,x2,… es un vector donde x1,x2,… representan los atributos o características del ejemplo. Los datos de entrenamiento

son aumentados con un vector C = c1,c2,… donde c1,c2,… representan la clase a la que pertenece cada muestra.

En cada nodo del árbol, C4.5 elige un atributo de los datos que más eficazmente dividen el conjunto de muestras en subconjuntos enriquecidos en una clase u otra. Su criterio es el normalizado para ganancia de información (diferencia de entropía) que resulta en la elección de un atributo para dividir los datos. El atributo con la mayor ganancia de información normalizada se elige como parámetro de decisión. El algoritmo C4.5 divide recursivamente en sub listas más pequeñas.

3.6.2. ALGORITMO LMT

Witten, I. Frank, E. (2005) indica:

Un modelo logístico de árbol consiste básicamente en una estructura de árbol de decisión estándar con funciones de regresión logística en las hojas. Como un modelo de árbol, es un árbol de regresión con funciones de regresión en las hojas. Como en un árbol de decisión ordinario, una prueba en uno de los atributos está asociada con cada nodo interno. Para un atributo nominal con k valores, el nodo tiene k nodos secundarios y los casos son clasificados hacia abajo en una de las k ramas dependiendo del valor del atributo. Para atributos numéricos, el nodo tiene dos nodos secundarios y la prueba consiste en comparar el valor del atributo con un umbral: un caso es ordenado hacia abajo en la rama izquierda si su valor para el atributo es más pequeño que el umbral, y en la rama derecha en el caso contrario.

"De una manera formal un modelo logístico de árbol consta de una estructura de árbol compuesto por un conjunto de nodos internos o no terminales N y un conjunto de hojas o nodos terminales T." (Landwehr, Mark, & Frank, 2006)

CAPITULO IV

RESULTADOS Y DISCUSIÓN

4.1. RESULTADOS

4.1.1. RECOPILAR INFORMACIÓN HISTÓRICA DE MADRES EN ESTADO GESTACIONAL.

4.1.1.1. COMPRENSIÓN DEL NEGOCIO

La Red de Salud Chucuito - Juli, en sus diferentes puestos y centros de salud realiza la atención a la población y para este estudio se toman las atenciones a gestantes y posterior resultado de recién nacido.

4.1.1.1.1. OBJETIVOS DEL NEGOCIO

Como ya se ha mencionado son la predicción de datos para los nuevos recién nacidos en base a la información de la gestante. La información histórica de las gestantes de la Red de Salud Chucuito - Juli ayuda a identificar patrones de comportamiento y poder realizar una predicción más fiable de la condición de salud del recién nacido. Para este caso se han definido los siguientes objetivos:

- Hacer predicción del valor de concentración de hemoglobina del recién nacido.
- Determinar si es un caso de riesgo para el recién nacido.

Estos informes son de importancia ya que ayudan a poder focalizar a gestante y asegurar las mejores condiciones de salud para el recién nacido. De esta forma podemos ayudar al profesional encargado de la atención a tomar decisiones acerca de la necesidad de suplementos o medicamentos.

4.1.1.1.2. EVALUACIÓN DE LA SITUACIÓN

Se cuenta con una base de datos en el gestor SQL Server con información histórica y detallada de las gestantes del año 2016 al 2018; se cuenta con datos de control de gestación en todos los códigos prestacionales. Las tablas dentro de la base de datos que tenemos son:

Tabla 4.1: Descripción de las tablas de base de datos

Tabla	Descripción
I_atediagnosticos	Tabla donde se encuentran los diagnósticos por atención (máximo 5 por atención)
I_ateinsumos	Tabla en la que se registra los insumos materiales no medicamentos que se utilicen en la atención al paciente.
I_atemedicamentos	Tabla en la que se registra los medicamentos recetados y/o aplicados al paciente.
I_atencion	Tabla en la que se registra la atención bajo un código de FUA y un código prestacional
I_atencionser	Tabla en la que se registra el servicio en el cual se realiza la atención.
I_ateprocedimientos	Tabla en la que se registra los procedimientos, exámenes y sus respectivos resultados.
I_atern	Tabla temporal de relación de recién nacido con puérpera.
I_atesmi	Tabla en la que se registra datos de tamizaje del paciente.

Elaborado por el equipo de trabajo

4.1.1.1.3. REALIZAR EL PLAN DE TRABAJO

El proyecto de minería de datos se ha programado y organizado de la siguiente manera:

- Etapa 1. Análisis de la estructura de datos y la información de las bases de datos.
- Etapa 2. Ejecución de consultas para tener muestras representativas de los datos.
- Etapa 3. Preparación de los datos.
- Etapa 4. Selección de las técnicas de modelado y ejecución con los datos.
- Etapa 5. Analizar resultados de aplicación de técnicas de modelado.
- Etapa 6. Interpretación de los resultados obtenidos de la aplicación de técnicas de modelado.
- Etapa 7. Presentación de resultados.

4.1.1.2. COMPRENSIÓN DE LOS DATOS

En esta segunda fase de la metodología CRISP-DM se realiza la recolección inicial de los datos para poder establecer un primer contacto con el problema, familiarizarse con los datos y averiguar su calidad, así como identificar las relaciones más evidentes para formular las primeras hipótesis.

4.1.1.2.1. RECOLECTAR LOS DATOS INICIALES

Los datos utilizados son referentes a pacientes; y los datos registrados en sus diferentes atenciones; para el caso de esta investigación se han escogido las atenciones con relación al tratamiento de gestantes, parto y del recién nacido.

- 009 - Atención prenatal.
- 011 - Exámenes de laboratorio completo de la gestante.
- 015 - Diagnóstico del embarazo.
- 029 - Tamizaje Neonatal.
- 050 - Atención inmediata del recién nacido normal.
- 054 - Atención de parto vaginal.
- 055 – Cesárea.

Para cada código prestacional, se tiene registrado datos de acuerdo al guía de correcto llenado de FUA que cada profesional maneja. Para el presente trabajo de investigación se tiene dos grupos de trabajo.

GESTANTE

- 009 – Atención prenatal.
- 011 – Exámenes de laboratorio completo de la gestante.
- 015 – Diagnóstico del embarazo.
- 054 – Atención de parto vaginal
- 055 – Cesárea.

RECIÉN NACIDO

- 029 – Tamizaje Neonatal.
- 050 – Atención inmediata del recién nacido normal.

Esta distinción nos ayuda a definir la información vamos a utilizar para poder realizar la predicción de condición de salud del recién nacido.

4.1.2. IDENTIFICAR PATRONES QUE INTERVIENEN EN LA PREDICCIÓN DE CONDICIÓN DEL RECIÉN NACIDO

4.1.2.1. PREPARACIÓN DE LOS DATOS

4.1.2.1.1. EXTRACCIÓN DE LOS DATOS

El primer paso consistió en recolectar los datos de acuerdo al código prestacional de cada FUA.

CÓDIGO PRESTACIONAL 009

Se debe de precisar fecha de probable parto a partir de la Fecha de Ultima Regla (FUR) de la gestante.

La norma técnica indica que para el caso de atención a la gestante la cantidad de controles necesarios se define por:

Tabla 4.2: Periodo de controles.

Regla de atención	Cantidad
01 atención/mensual hasta 32ss	07 controles
01 atención/quincenal entre las 33ss y 36ss	02 controles
01 atención/semanal desde la 37ss hasta el parto	04 controles

Fuente: Ministerio de Salud del Perú (2017)

También es requerido datos de tamizaje como son peso, talla e índice de masa corporal IMC.

Además, se debe registrar el número de control prenatal CPN, edad gestacional en semanas.

En la parte de diagnósticos se debe de precisar el código CIE10, que de acuerdo sea el caso será Z340 si es el primer embarazo ó Z348 si fuera otro número de embarazo.

Respecto a los medicamentos, hasta la semana 13 se le hace entrega del ácido fólico; y desde la semana 14 se le hace entrega de ácido fólico y sulfato ferroso.

Si fuera el caso se registra también los procedimientos 85018 Hemoglobina, 86592 Prueba de Sífilis cualitativa y 86701 HIV – 1 Anticuerpos.

CÓDIGO PRESTACIONAL 011

La norma técnica indica que para este código prestacional se debe de precisar fecha de probable parto a partir de la Fecha de Ultima Regla (FUR) de la gestante. También es requerido datos de tamizaje como son peso, talla y edad gestacional en semanas.

En la parte de diagnósticos se debe de precisar el código CIE10, será Z017 Examen de laboratorio.

Para este caso los exámenes de laboratorio son obligatorios, pues este FUA es atendido por un profesional Biólogo de Servicio de Laboratorio.

Los exámenes que se deben de registrar son 85018 Hemoglobina, 86592 Prueba de Sífilis cualitativa y 86701 HIV – 1 Anticuerpos.

CÓDIGO PRESTACIONAL 054

La norma técnica indica que se debe de registrar la fecha de parto, lo cual puede diferir de la fecha calculada bajo la FUR.

Se debe de registrar el peso y talla de la gestante, asimismo la edad gestacional lo cual definirá si fuese un parto prematuro y/o reflejará errores en la precisión de la FUR, ya que es un dato que la gestante proporciona.

Respecto al diagnóstico se debe de usar el código:

- O80.0 Parto único espontáneo, presentación cefálica de vértice.
- O80.1 Parto único espontáneo, presentación de nalgas o podálica.
- O80.8 Parto único espontáneo, otras presentaciones.
- O80.9 Parto único espontáneo, sin otra especificación.

Respecto a los procedimientos se debe de registrar el valor de hemoglobina.

4.1.2.2. LIMPIEZA DE DATOS

La limpieza de datos inicia con la eliminación de los registros que no cuentan con parto en la Red de Salud Chucuito - Juli.

CÓDIGO PRESTACIONAL 009

Para el caso del código prestacional 009 el valor de hemoglobina que se registra es bajo la aplicación de reactivo a una muestra de sangre; esto se realiza en los establecimientos de salud periféricos, asimismo debemos de precisar que estos datos son muy volátiles y prima según norma los datos de laboratorio, por lo que no serán considerados en este código prestacional.

Asimismo, se debe de tener en cuenta los FUAS que tienen registrado los valores de peso, talla y valor de hemoglobina del recién nacido.

Tabla 4.3: Atributos seleccionados para CP 009

ATRIBUTO	DESCRIPCIÓN
EDAD	Edad de la gestante
EDAD_GEST	Semanas de gestación
PESO	Peso obtenido de acuerdo a la semana de gestación
TALLA	Talla obtenida de acuerdo a la semana de gestación
AF	Entrega de Ácido Fólico
AF_SF	Entrega de Ácido Fólico + Sulfato Ferroso
SGRN	Semanas de gestación del RN
HB_RN	Valor de hemoglobina del recién nacido

Elaborado por el equipo de trabajo

CÓDIGO PRESTACIONAL 011

Asimismo, relacionamos los registros que tenemos del código prestacional 011 con el FUA del parto de código prestacional 054 y su respectivo FUA de código prestacional 050 del recién nacido; al igual que para el caso del código prestacional 009; muchas gestantes no tienen el parto en la Red de Salud Chucuito - Juli; por lo mismo no podemos acceder a los datos de parto ni del recién nacido.

Para la limpieza de datos se ha tomado en cuenta los partos sea prematuro o a terminó de los cuales se tiene datos que podamos usar.

Asimismo, no podemos considerar los FUAS en los cuales no se tiene los valores de hemoglobina, y por último solo podemos considerar los registros que cuenten con datos de recién nacido que tengan valores de hemoglobina.

Los valores que tenemos son los siguientes:

Tabla 4.4: Atributos seleccionados para CP 011

ATRIBUTO	DESCRIPCIÓN
EDAD	Edad de la gestante
EDAD_GEST	Semanas de gestación
PESO	Peso obtenido de acuerdo a la semana de gestación
TALLA	Talla obtenida de acuerdo a la semana de gestación
HB	Valor de hemoglobina según laboratorio
SGRN	Semanas de gestación del RN
PESO_RN	Peso del recién nacido
TALLA_RN	Talla del recién nacido
HB_RN	Valor de hemoglobina del recién nacido

Elaborado por el equipo de trabajo

4.1.2.3. SELECCIÓN DE DATOS PARA CONSTRUCCIÓN DE MODELO

A continuación, se va a realizar la correlación de los datos obtenidos con las consultas finales, en este caso se realiza la correlación hacia las variables objetivo que son los datos del recién nacido.

CÓDIGO PRESTACIONAL 009

Para esto se aplica la correlación de variables de forma individual usando la siguiente ecuación:

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

Donde, $\bar{x}$ y $\bar{y}$ son el promedio de x e y respectivamente.

Tabla 4.5: Correlación de atributos por CP 009

	SGRN	PESO_RN	TALLA_RN	HB_RN
EDAD	0.075	0.082	0.184	0.041
EDAD_GEST	0.042	0.049	0.033	0.003
PESO	0.130	0.283	0.229	0.012
TALLA	0.145	0.080	0.080	0.027
AF_SF	-0.088	0.003	-0.007	0.093

Elaborado por el equipo de trabajo

Podemos observar que el peso de la gestante tiene correlación baja con el estado de gestación del recién nacido, asimismo el peso de la madre tiene correlación positiva con el peso del recién nacido. Un poco más desplazado observamos que la edad de la gestante influye ligeramente en la talla del recién nacido, y peso de la gestante influye un poco más en el peso del recién nacido.

CÓDIGO PRESTACIONAL 011

De igual manera lo calculado con el código prestacional 009, calculamos la correlación en pares de variables cruzando los atributos de gestante y de recién nacido.

Tabla 4.6: Correlación de atributos por CP 011

	SGRN	PESO_RN	TALLA_RN	HB_RN
EDAD	0.1170	0.1427	0.1507	0.0367
EDAD_GEST	-0.0094	0.0790	-0.0181	-0.0100
HB	0.0523	0.0005	0.0093	0.0834

Elaborado por el equipo de trabajo

4.1.3. CLASIFICAR LA INFORMACIÓN BAJO TÉCNICAS PREDICTIVAS DE MINERÍA DE DATOS PARA OBTENER UN MODELO PREDICTIVO ADECUADO PARA DEFINIR LA CONDICIÓN DEL RECIÉN NACIDO

4.1.3.1. MODELADO

La base de obtención de un buen modelo, se basa en la calidad de datos escogidos y su selección adecuada.

Agrupamos los datos para aplicar los algoritmos de clasificación y hacer las comparaciones pertinentes.

Para esta etapa se usará los algoritmos en el 70% de 484 registros que son 339 y posteriormente usar el 30% restante en la evaluación del mismo.

Para esta investigación se ha discretizado la condición de salud de acuerdo a los parámetros siguientes:

Tabla 4.7: Discretización de condición de salud del recién nacido

CAT.	PESO RN (G)	TALLA RN (CM)	SGRN	HB_RN
MUY BUENO	>3500	>50	RNPostT, RNT	>16
BUENO	3000-3500	45-50	RNPosT, RNT	14-16
REGULAR	2500-3000	40-45	RNT, RNTP	12-14
MALO	2000-2500	35-40	RNT, RNTP	10-12
MUY MALO	<2000	<35	RNTP	<10

Elaborado por el equipo de trabajo

Agrupamos los datos para aplicar los algoritmos de clasificación y hacer las comparaciones pertinentes.

Para esta etapa se usará los algoritmos en el 70% de 484 registros que son 339 y posteriormente usar el 30% restante en la evaluación del mismo.

4.1.3.1.1. ALGORITMO J48

De acuerdo a la clasificación que nos muestra este algoritmo, tenemos los siguientes resultados:

```
Correctly Classified Instances    240          70.7965 %
Incorrectly Classified Instances  99           29.2035 %
Kappa statistic                                0.5501
Mean absolute error                            0.1639
Root mean squared error                        0.2863
Relative absolute error                        60.3508 %
Root relative squared error                    77.7864 %
Total Number of Instances                      339
```

4.1.3.1.2. ALGORITMO LMT

De acuerdo a la clasificación que nos muestra este algoritmo, tenemos los siguientes resultados:

```
Correctly Classified Instances         237               69.9115 %
Incorrectly Classified Instances       102               30.0885 %
Kappa statistic                          0.5375
Mean absolute error                      0.1832
Root mean squared error                  0.2939
Relative absolute error                 67.4294 %
Root relative squared error             79.841  %
Total Number of Instances               339
```

4.1.3.1.3. COMPARACIÓN ENTRE LOS RESULTADOS DE LOS ALGORITMOS J48 Y LMT

Luego de aplicar los algoritmos J48 y LMT, al conjunto de datos de entrenamiento, se obtuvieron los resultados de la Tabla 7, en la cual se presenta el resultado de instancias correctamente clasificadas y el error absoluto de ambos métodos.

Tabla 4.8: Comparativa de algoritmos de clasificación

ALGORITMO	INSTANCIAS CORRECTAS	ERROR ABSOLUTO
J48	70.79%	0.1639
LMT	69.91%	0.1832

Elaborado por el equipo de trabajo

4.1.3.1.4. COMPARACIÓN DE MATRICES DE CONFUSIÓN

Para evaluar los resultados de los clasificadores, nos vamos a basar en el porcentaje de instancias clasificadas incorrectamente y en la matriz de confusión. Ésta matriz es de vital importancia en problemas de clasificación debido a que nos ofrece información del tipo de error de clasificación de los algoritmos. (Jimenez & Gómez, 2009). A continuación, se presentan las matrices de confusión de ambos algoritmos.

Figura 4.1: Matriz de confusión weka.classifiers.trees.J48

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
  98  27   0   2   0 |   a = BUENO
  27 106   0   2   1 |   b = REGULAR
   2   5   2   2   0 |   c = MUY MALO
  15   7   0  21   0 |   d = MUY BUENO
   6   2   0   1  13 |   e = MALO
```

Elaborado por el equipo de trabajo

Figura 4.2: Matriz de confusión weka.classifiers.trees.LMT

```
=== Confusion Matrix ===

   a   b   c   d   e   <-- classified as
 102  22   1   2   0 |   a = BUENO
  36  98   0   2   0 |   b = REGULAR
   3   2   5   1   0 |   c = MUY MALO
  15   9   0  19   0 |   d = MUY BUENO
   8   1   0   0  13 |   e = MALO
```

Elaborado por el equipo de trabajo

Para el algoritmo J48, según se muestra en la Figura 17, se tiene que, de 127 instancias con condición de salud “Bueno”, 98 fueron clasificadas correctamente y 29 presentaron errores.

Para el algoritmo LMT, según se muestra en la Figura 18, se tiene que, de 127 instancias con condición de salud “Bueno”, 102 fueron clasificadas correctamente y 25 presentaron errores.

4.1.3.2. EVALUACIÓN DEL MODELO

Previamente se presentaron los resultados de aplicación del algoritmo J48 y LMT en la herramienta WEKA, como se mencionó los datos de entrenamiento representan un 70% del conjunto de datos original, y el 30% restante son datos de prueba.

4.1.3.2.1. COMPARACIÓN DE PRECISIÓN ENTRE ALGORITMOS J48 Y LMT

En la siguiente tabla se realiza la comparación de los resultados basados en el entrenamiento del 70% del conjunto de datos original con el 30% destinado para ser los datos de conjunto de pruebas.

Tabla 4.9: Comparación Entrenamiento – Prueba (J48 y LMT)

ALG.	ENTRENAMIENTO (339)		PRUEBAS (145)	
	CLASIFICADOS	ERROR ABS.	CLASIFICADOS	ERROR ABS.
J48	70.79%	0.1639	69.58%	0.1767
LMT	69.91%	0.1832	68.89%	0.2066

Elaborado por el equipo de trabajo

Para ambos algoritmos, tanto las instancias correctamente clasificadas y el error absoluto del entrenamiento y pruebas, son muy próximos; de lo mismo se puede deducir que si el número de instancias correctas se incrementa, el error absoluto disminuye.

Ambos algoritmos nos proporcionan similares márgenes de instancias correctamente clasificadas y error absoluto. Se puede deducir que 7 de cada 10 instancias cumplen con un patrón de control y son correctamente clasificadas en el modelo.

El modelo escogido y probado en esta investigación es el proporcionado por el algoritmo J48.

4.2. DISCUSIÓN

A partir de los resultados que hemos tenido, acepamos nuestra hipótesis alternativa general que establece que el uso del modelo basado en minería de datos permite realizar la predicción de la condición de salud de los recién nacidos en la Red de Salud Chucuito-Juli.

Estos resultados guardan relación con lo que afirman Ccopa M.y Chavez S. (Ccopa & Chavez, 2015) aplicado en la toma de decisiones del departamento de cirugía del hospital regional Manuel Núñez Butrón, Saldaña E. (Saldaña, 2015) aplicado en morbilidad de pacientes hospitalizados, Ticona M. (Ticona, 2018) en obesidad en la adolescencia y Abad I. (Abad, 2016) aplicado en partos prematuros, quienes señalan que el conjunto de técnicas y tecnologías de la minería de datos permiten explorar grandes bases de datos de manera automática, ayudando a predecir de manera eficiente en el área de la salud. Estos autores indican que bajo un modelo basado en minería de datos se logra predecir factores de riesgo que colaboren con el profesional de la salud. Estos estudios son acordes con lo que en esta investigación se halla.

Por otro lado, en lo que respecta al uso y aplicación de la metodología Crisp-DM, los autores Ccopa M. y Chavez S. (Ccopa & Chavez, 2015) y Saldaña E. (Saldaña, 2015) determinan que es la que mejor opción que se ajusta para este tipo de estudios en la línea de investigación de la salud, con lo que concuerda con lo que en este estudio se señala. Además, Ticona M. (Ticona, 2018) en su estudio: Sistema para la predicción de obesidad en la adolescencia utilizando técnicas de minería de datos, al igual que en la presente investigación, indica que el uso del algoritmo J48 con apoyo y utilización de la herramienta Weka son óptimos para la predicción en la salud, colaborando a tener mejores resultados que te pueden permitir actuar de manera preventiva.

CONCLUSIONES

PRIMERO: La investigación realizada establece que, en apoyo al diagnóstico y prevención los atributos procesados y discretizados posibilitan determinar la condición de salud de los recién nacidos; la misma que tiene un error absoluto de ± 0.1639, esto confirma la hipótesis que el modelo basado en minería de datos permite predecir la condición de salud del recién nacido en la Red de Salud Chucuito-Juli.

SEGUNDO: Se evidencia que, según el modelo, 7 de cada 10 gestantes cumplen con los patrones de comportamiento según las indicaciones y medicación que se le otorga.

TERCERO: Los datos recolectados con los códigos prestacionales 009 Atención prenatal y 011 Exámenes de laboratorio completo de las gestantes, permiten representar el comportamiento de los atributos de las gestantes hacia el posterior parto con producto de recién nacido vivo.

CUARTO: Los modelos predictivos obtenidos en esta investigación permiten observar en cuanto intervienen los atributos recolectados de la gestante en la condición de salud y sus atributos del recién nacido.

QUINTO: El análisis de los datos usando técnicas predictivas de minería de datos permite la obtención de modelos de minería de datos que evalúan la condición de salud del recién nacido.

RECOMENDACIONES

Para la institución, se debe de valorar el proceso de digitación de FUAs, pues en la limpieza de datos se evidencia algunos errores de digitación que bajo las normas técnicas del Ministerio de Salud se han tenido que obviar, pues generan valores que escapan de la correlación de datos y pueden influir en una equivocada interpretación.

Los factores atributo que se recopilan de la gestante influyen en el recién nacido, muchos atributos del recién nacido dependen de otros factores. Los mismos que no pueden ser identificados pues en los FUAs correspondientes no se tiene un esquema de recolección por patrones.

La investigación futura se debe de centrar en el monitoreo de comportamiento del desarrollo del niño, pues se tiene códigos prestacionales que recolectan más atributos, y de acuerdo a la norma técnica se podría observar la evolución del niño.

REFERENCIAS BIBLIOGRÁFICAS

Abad, I. (2016). *MODELO PREDICTIVO DE PARTO PREMATURO BASADO EN FACTORES DE RIESGO.* Oviedo - España: UNIVERSIDAD DE OVIEDO.

Ayala Rosero, E. J., & Logacho Fernández, A. I. (2018). *Identificar un Modelo de Data Mining para Desarrollar un Análisis Predictivo en la Administración Integral del Trabajo y Empleo de las Empresas Ecuatorianas.* Sangolquí, Ecuador: Tesis de Maestría .

BIZMETRIKS. (2013). *bizmetriks-Descubriendo Conocimiento*. Recuperado el 2019, de http://www.bizmetriks.com/metodologia.html

Caisés Amalguer, Y., & Navarro Rodriguez, R. (2010). Transformación de Características para la Minería de Datos. *Ciencias Holguín*.

Ccopa, M., & Chavez, S. (2015). *MODELO PREDICTIVO BASADO EN MINERÍA DE DATOS PARA LA MEJORA EN LA TOMA DE DECISIONES DEL DEPARTAMENTO DE CIRUGÍA DEL HOSPITAL REGIONAL MANUEL NÚÑEZ BUTRÓN.* Puno: Universidad Nacional del Altiplano.

Devore, J. L. (2008). *Probabilidad y Estadística para Ingenierías y Ciencias.* México: Cengage Learning Editores.

Edelstein, H. (1999). *Introduction to Data Mining and Knowledge Discovery.* Estados Unidos: Two Crows Corp.

Gobierno del Perú - Ministerio de Salud. (25 de Abril de 2018). *Plataforma digital única del Estado Peruano*. Obtenido de

https://www.gob.pe/institucion/minsa/noticias/5841-minsa-gestantes-deben-realizarse-al-menos-tres-ecografias-durante-todo-el-embarazo

Gonzales Casimiro, M. P. (2009). *Análisis de series temporales: Modelos ARIMA.* Vasco: Universidad del País Vasco.

Gujarati, D. N. (2003). *Econometria.* United States: Mc Graw Hill.

Hernández Orallo, J., Ramiréz Quintana, M. J., & Ferri Ramírez, C. (2004). *Introducción a la minería de datos.* Madrid: Prentice Hall.

IBM. (2012). *IBM SPSS Modeler CRISP-DM Guide.* USA: IBM.

INEI-ENDES. (2018). *Indicadores de Resultados de los Programas Presupuestales, 2013-2018 – Primer Semestre.* Lima.

Jiawei, H., Kamber, M., & Pein, J. (2012). *DATA MINING Concepts and Techniques.* USA: Elsevier.

Jimenez Vázquez, E., & Gómez Bertoli, D. (2009). *Sistema de localización en redes Wi-Fi con Weka.* Madrid, España: Universidad Carlos III de Madrid.

Jimenez, E., & Gómez, D. (2009). *Sistema de localización en redes Wi-Fi con Weka.* Madrid, España: Universidad Carlos III de Madrid.

Kasabov, N. K. (1996). *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering.* Cambridge: Massachusetts Institute of Technology.

Landwehr, N., Mark, H., & Frank, E. (2006). Logistic Model Trees. En *Proceedings of the* (págs. 14-21). Holanda: Kluwer Academic Publishers.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledge Discovery Handbook.* Estados Unidos: Springer.

Mártinez Álvarez, C. A. (2012). *APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA MEJORAR EL PROCESO DE CONTROL DE GESTIÓN EN ENTEL.* Santiago de Chile: Universidad de Chile.

MINISTERIO DE SALUD DEL PERÚ. (2017). *NORMA TÉCNICA - MANEJO TERAPÉUTICO Y PREVENTIVO DE LA ANEMIA EN NIÑOS, ADOLESCENTES, MUJERES GESTANTES Y PUÉRPERAS.* Lima: MINSA.

Molina, J., & García, J. (2011). *Técnicas de análisis de datos.*

Moral Peláez. (2012). *Modelos de regresión:Llineal Simple y Regresión Logogística.* España.

Ng, K., & Liu, H. (2000). Customer Retention via Data Mining. *Artificial Intelligence Review.*, 570.

Ortuño Ulco, M. T. (2018). *MODELO DE ANÁLISIS PREDICTIVO PARA EL MEJORAMIENTO PRESUPUESTARIO DE LA PLANIFICACIÓN LOGÍSTICA DEL CONSEJO NACIONAL ELECTORAL DEL ECUADOR, BASADO EN EL USO DE TÉCNICAS DE MINERIA DE DATOS.* Ecuador: Universidad de las Fuerzas Armadas ESPE.

Pérez López, C. (2007). *Minería de datos: técnicas y herramientas.* Madrid: Paraninfo.

Predictive Analitics Today. (12 de 03 de 2019). *PAT Research.* Obtenido de https://www.predictiveanalyticstoday.com/compare/orange-data-mining-vs-r-software-environment-vs-weka-data-mining-vs-rapidminer-starter-edition/

Quinlan, R. (1993). *C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.* CA: San Mateo.

Rojo Abuín, J. M. (2017). *Regresión lineal múltiple. Madrid.* Madrid.

Saldaña, E. (2015). *MODELO PREDICTIVO DE MINERIA DE DATOS DE APOYO A LA GESTION HOSPITALARIA SOBRE LA MORBILIDAD DE PACIENTES HOSPITALIZADOS.* Trujillo: UNIVERSIDAD PRIVADA ANTENOR ORREGO.

SAS Institute Inc. (2017). *SAS® Enterprise Miner 14.3: Reference Help.* USA: SAS.

Sumathi, S., & Sivanandam, S. N. (2006). *Introduction to Data Mining and its Applications.* Heidelberg, Berlin: Springer.

Ticona, M. (2018). *SISTEMA PARA LA PREDICCIÓN DE OBESIDAD EN LA ADOLESCENCIA UTILIZANDO TÉCNICAS DE MINERÍA DE DATOS.* Arequipa: Universidad Católica de Santa María.

Tuya, J., Ramos, I., & Dolado, J. (2007). *Técnicas Cuantitativas para la Gestión en la Ingeniería del Software.* España: Netbiblo.

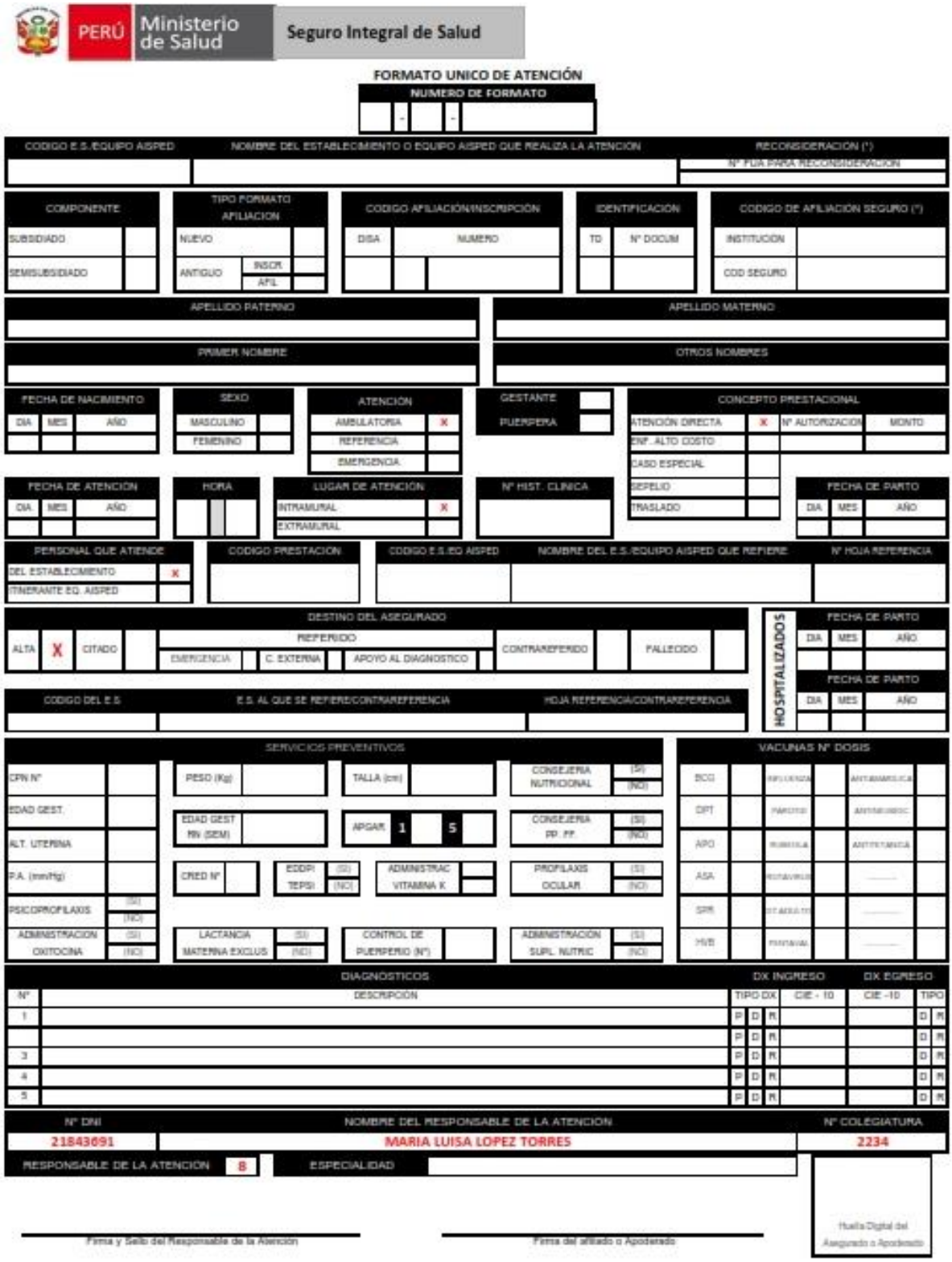
Weiss, S., & Indurkhya, N. (1998). *Predictive Data Mining A Practical Guide.* Estados Unidos: Morgan Kaufmann Publishers, Inc.

Witten, I., & Frank, E. (2005). *DATA MINING Practical Machine Learning Tools and Techniques.* San Franciso, CA: Elsevier Inc.

Y. L. (2012). *Árboles de Decisiones.* Sevilla.

ANEXOS

ANEXO 1. FORMATO ÚNICO DE ATENCIÓN (FUA)



PERÚ Ministerio de Salud — Seguro Integral de Salud

FORMATO UNICO DE ATENCIÓN

NUMERO DE FORMATO

CODIGO E.S./EQUIPO AISPED | NOMBRE DEL ESTABLECIMIENTO O EQUIPO AISPED QUE REALIZA LA ATENCION | RECONSIDERACION (*) — N° FUA PARA RECONSIDERACION

COMPONENTE: SUBSIDIADO, SEMISUBSIDIADO | TIPO FORMATO AFILIACION: NUEVO, ANTIGUO (INSOR, AFIL.) | CODIGO AFILIACION/INSCRIPCION: DISA, NUMERO | IDENTIFICACION: TD, N° DOCUM | CODIGO DE AFILIACION SEGURO (*): INSTITUCION, COD SEGURO

APELLIDO PATERNO | APELLIDO MATERNO

PRIMER NOMBRE | OTROS NOMBRES

FECHA DE NACIMIENTO: DIA, MES, AÑO | SEXO: MASCULINO, FEMENINO | ATENCION: AMBULATORIA X, REFERENCIA, EMERGENCIA | GESTANTE, PUERPERA | CONCEPTO PRESTACIONAL: ATENCIÓN DIRECTA X, ENF. ALTO COSTO, CASO ESPECIAL, REPELIO, TRASLADO — N° AUTORIZACION, MONTO

FECHA DE ATENCION: DIA, MES, AÑO | HORA | LUGAR DE ATENCION: INTRAMURAL X, EXTRAMURAL | N° HIST. CLINICA | FECHA DE PARTO: DIA, MES, AÑO

PERSONAL QUE ATIENDE: DEL ESTABLECIMIENTO X, ITINERANTE EQ. AISPED | CODIGO PRESTACION | CODIGO E.S./EQ AISPED | NOMBRE DEL E.S./EQUIPO AISPED QUE REFIERE | N° HOJA REFERENCIA

DESTINO DEL ASEGURADO: ALTA X, CITADO | REFERIDO: EMERGENCIA, C. EXTERNA, APOYO AL DIAGNOSTICO | CONTRAREFERIDO | FALLECIDO

HOSPITALIZADOS: FECHA DE PARTO: DIA, MES, AÑO | FECHA DE PARTO: DIA, MES, AÑO

CODIGO DEL E.S | E.S. AL QUE SE REFIERE/CONTRAREFERENCIA | HOJA REFERENCIA/CONTRAREFERENCIA

SERVICIOS PREVENTIVOS: CPN N°, EDAD GEST., ALT. UTERINA, P.A. (mm/Hg), PSICOPROFILAXIS (SI/NO), ADMINISTRACION OXITOCINA (SI/NO) | PESO (Kg) | TALLA (cm) | CONSEJERIA NUTRICIONAL (SI/NO) | EDAD GEST RN (SEM) | APGAR 1 5 | CONSEJERIA PP. FF (SI/NO) | CRED N° | EDDP/TEPSI (SI/NO) | ADMINISTRAC VITAMINA K | PROFILAXIS OCULAR (SI/NO) | LACTANCIA MATERNA EXCLUS (SI/NO) | CONTROL DE PUERPERIO (N°) | ADMINISTRACIÓN SUPL. NUTRIC (SI/NO)

VACUNAS N° DOSIS: BCG, DPT, APO, ASA, SPR, HVB

DIAGNOSTICOS

N°	DESCRIPCIÓN	DX INGRESO: TIPO DX (P D R), CIE - 10	DX EGRESO: CIE -10, TIPO (D R)
1		P D R	D R
		P D R	D R
3		P D R	D R
4		P D R	D R
5		P D R	D R

N° DNI	NOMBRE DEL RESPONSABLE DE LA ATENCION	N° COLEGIATURA
21843691	MARIA LUISA LOPEZ TORRES	2234

RESPONSABLE DE LA ATENCIÓN: B | ESPECIALIDAD

Firma y Sello del Responsable de la Atención

Firma del afiliado o Apoderado

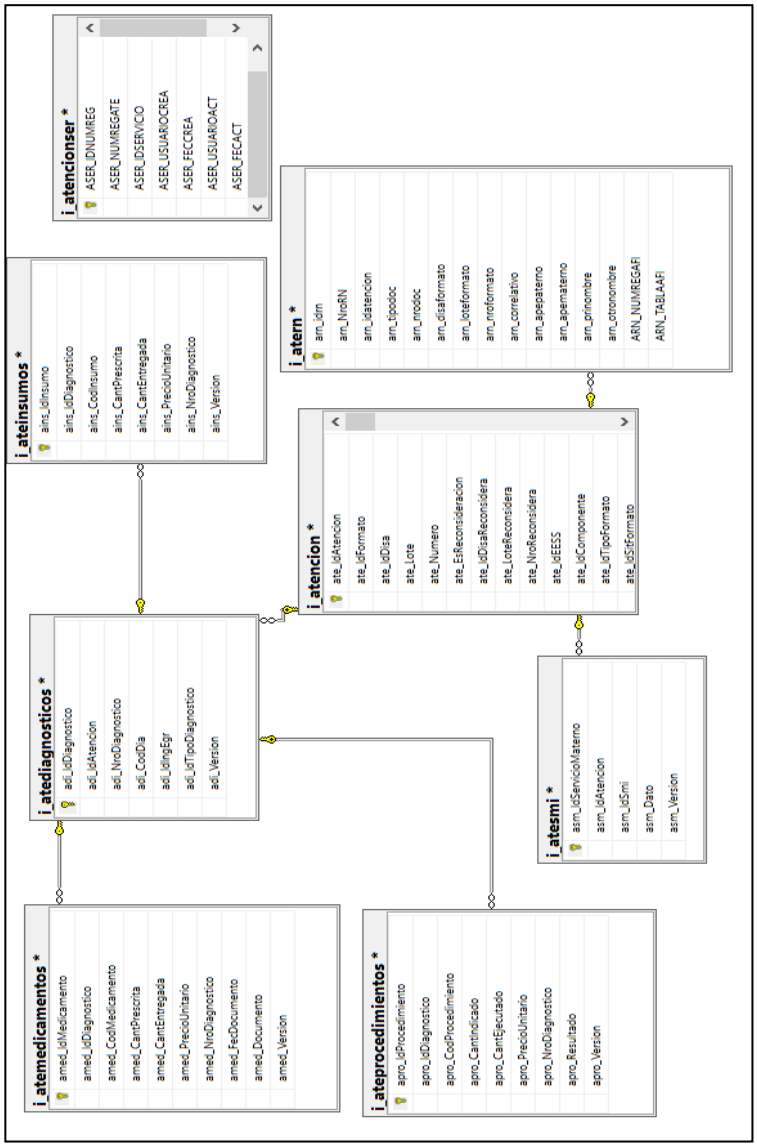
Huella Digital del Asegurado o Apoderado

ANEXO 2. CÓDIGOS PRESTACIONALES.

001 - Control de crecimiento y desarrollo en menores entre 0 - 4 años
002 - Control del recién nacido con menos de 2,500 gr, prematuro, con secuelas al nacer
005 - Consejería nutricional para niñas o niños en riesgo nutricional y desnutrición
007 - Suplemento de micronutrientes
008 - Profilaxis antiparasitaria
009 - Atención prenatal
010 - Atención del puerperio normal
011 - Exámenes de laboratorio completo de la gestante
013 - Exámenes de ecografía obstétrica
015 - Diagnóstico del embarazo
016 - Atención temprana para menores de 36 meses
017 - Atención Integral del adolescente
018 - Salud reproductiva (planificación familiar)
019 - Detección trastorno agudeza visual y ceguera
020 - Salud Bucal
021 - Prevencion de caries
022 - Detección de problemas en Salud Mental
023 - Deteccion precoz de cancer de prostata (PSA)
024 - Detección precoz de cáncer cérvico-uterino
025 - Detección precoz de cancer de mama (Mamografía)
026 - Tratamiento profiláctico para gestante positiva a prueba rápida/ELISA VIH
027 - Tratamiento profilactico a niños expuestos al VIH
029 - Tamizaje Neonatal
050 - Atención inmediata del recién nacido normal
051 - Internamiento del RN con patología no quirurgica
052 - Internamiento con intervención quirúrgica del RN
053 - Tratamiento de VIH-SIDA (0-19a)
054 - Atención de parto vaginal
055 - Cesárea
056 - Consulta externa
057 - Obturación y curación dental simple
058 - Obturación y curación dental compuesta
059 - Extracción dental (exodoncia)
060 - Atención extramural urbana y periurbana (Visita domiciliaria)
061 - Atención en tópico
062 - Atención por emergencia
063 - Atención por emergencia con observación
064 - Intervención medico-quirúrgica ambulatoria
065 - Internamiento en EESS sin intervención quirúrgica
066 - Internamiento con intervención quirúrgica menor
067 - Internamiento con intervención quirúrgica mayor

068 - Internamiento con Estancia en la Unidad de Cuidados Intensivos (UCI)
069 - Transfusión sanguínea o hemoderivados
070 - Atención odontológica especializada
071 - Apoyo al diagnóstico
074 - Tratamiento de ITS en adolescentes, adultos y adultos mayores
075 - Atención extramural rural (Visita domiciliaria)
111 - Asignación por Alimentación
112 - Sepelio para Óbito fetal (Muerte Intraútero)
113 - Sepelio para Niñas/os
114 - Sepelio para Adolescentes y Adultos
116 - Sepelio para Recién Nacidos
117 - Traslado de Emergencia
118 - Control de crecimiento y desarrollo en menores entre 5 - 9 años
119 - Control de crecimiento y desarrollo en entre de 10 - 11 años
200 - Atención de rehabilitación (post fractura y/o post esguince)
900 - Prótesis dental removible
901 - Apoyo al Tratamiento
902 - Atención Preconcepcional
903 - Atenciòn Integral de Salud del Adulto Mayor
904 - Atención Integral de Salud del Joven y Adulto
906 - Consulta externa por profesionales no médicos ni odontólogos
907 - Atención por Telesalud
S01 - (****)
S02 - Salud Escolar (***)

ANEXO 3. DIAGRAMA DE BASE DE DATOS.



ANEXO 4. MATRIZ DE CONSISTENCIA Y OPERACIONALIZACIÓN DE VARIABLES

PROBLEMA	OBJETIVO	HIPÓTESIS	VARIABLES	DIMENSIONES	INDICADORES
¿De qué manera el modelo basado en minería de datos permite predecir las condiciones de salud de los recién nacidos en la red de salud Chucuito – Juli?	Aplicar un modelo basado en minería de datos para predecir la condición de salud de los recién nacidos en la RED SALUD CHUCUITO - JULI.	El uso del modelo basado en minería de datos permite realizar la predicción de la condición de salud de los recién nacidos en la Red de salud Chucuito – Juli.	**Variable independiente:** Modelo predictivo.	Comprensión del negocio.	Objetivos del negocio. Evaluación de la situación.
				Compresión de los datos.	Entender el problema existente en la información transaccional, analizándola y seleccionando los campos pertinentes de las tablas seleccionadas.
				Procesamiento de los datos.	ETL extracción, transformación y carga de los datos de la muestra seleccionada. Limpiar los datos de la muestra seleccionada. Selección de Atributos para el análisis del algoritmo.
				Modelado.	Identificación del posible modelo. Estimación del modelo. Diagnóstico del modelo. Pronostico del modelo.
			Variable dependente: Condición de salud del recién nacido.	Análisis	Datos Históricos. Predicción.