UNIVERSIDAD NACIONAL DE MOQUEGUA

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

“MODELO DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE CASOS DE ANEMIA EN GESTANTES DE LA PROVINCIA DE ILO”

TESIS PRESENTADA POR EL BACHILLER:
SADAN EUSEBIO CONDORI BELLIDO

PARA OPTAR EL TÍTULO PROFESIONAL DE:
INGENIERO DE SISTEMAS E INFORMÁTICA

MOQUEGUA – PERÚ

2019

09

07

UNIVERSIDAD NACIONAL DE MOQUEGUA
CARRERA PROFESIONAL DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ACTA DE SUSTENTACIÓN

El Jurado suscribiente ha calificado el trabajo de Tesis:

TITULADO Modelo de Minería de Datos para la predicción de Casos de Anemia en gestantes de la provincia de Ilo

PRESIDENTE Mgr. Carlos Alberto Silva Delgado

MIEMBROS Dr. Anibal Fernando Flores Garcia
Mgr. José Antonio Guzmán Valdivia

ASESOR Msc. Hugo Euler Tito Churo

PRESENTADO POR EL BACHILLER Sadan Eusebio Condori Bellido

DE LA PROMOCIÓN: 2009

CUYOS RESULTADOS HAN SIDO LOS SIGUIENTES:

CALIFICATIVO DEL TRABAJO 13 (Trece)

CALIFICATIVO DE LA SUSTENTACIÓN 13 (Trece)

CALIFICATIVO FINAL 13 (Trece)

POR LO EXPUESTO, EL BACHILLER Sadan Eusebio Condori Bellido

HA SIDO DECLARADO EXPEDITO PARA QUE SE LE CONFIERA EL TÍTULO PROFESIONAL DE:

INGENIERO DE SISTEMAS E INFORMÁTICA

EN FE DE ELLO QUEDA ASENTADA LA PRESENTE ACTA

ILO, A LOS 24 DÍAS DE Octubre DE 2019

Presidente

Miembro

Miembro

DEDICATORIA

A Dios quien supo guiarme por el buen camino, dándome fuerzas para ser perseverante y no desvanecer en los problemas suscitados.

A mis padres por su apoyo incondicional, compresión, consejos y ayuda en los momentos difíciles, quienes me abastecieron de los recursos necesarios para poder concluir mi carrera universitaria.

AGRADECIMIENTOS

A nuestros miembros del jurado por sus aportes para la mejoría de nuestro trabajo de investigación. A todos los profesionales que formaron parte del juicio que nos brindaron su tiempo, dedicación y apoyo incondicional, ya que con su aporte científico y humano han colaborado en la realización de nuestro trabajo. Y un agradecimiento especial a todo el personal que trabaja en la Red Salud Ilo, por la confianza al brindarnos información y permitir el desarrollo de nuestro trabajo de investigación.

A mis compañeros y docentes de la universidad, quienes que me enseñaron a seguir adelante y cumplir mis objetivos.

RESUMEN

En la actualidad la anemia es una enfermedad que afecta al 24.8% de la población mundial, siendo los más afectados los niños en edad preescolar y las madres gestantes, esta realidad se da en la mayoría de los países del mundo, dada su relevancia, existen muchas investigaciones abordadas desde diferentes perspectivas, entre ellas, desde el enfoque de la ciencia de la computación a través de su línea de investigación denominada minería de datos que consiste en investigaciones de predicción y clasificación utilizando los diferentes algoritmos.

El presente trabajo de investigación pretende desarrollar un modelo de minería de datos predictivo aplicando las técnicas de Machine Learning, las que han aportado positivamente en el estudio de diferentes campos como la medicina. Al analizar el contexto del negocio, se planteó utilizar tres de la referidas técnicas para predecir futuros casos de madres gestantes con anemia, para lo cual se implementaron los algoritmos de Perceptrón Multicapa, Naive Bayer y Árbol de decisión J48, estos fueron entrenados sobre una base de datos histórica de 422 registros de madres gestantes con anemia de la Provincia de Ilo, el algoritmo que alcanzó mayor precisión fue el de Naive Bayes con un 89 %, seguido por el de árbol de decisión J48 con 79% y finalmente el perceptrón multicapa con un 62%.

El desarrollo del proyecto se basó en la metodología CRISP-DM para desarrollar cada una de las etapas que condujeron al resultado final.

ABSTRACT

At present, anemia is a disease that affects 24.8% of the world population, being the most affected children of preschool age and pregnant mothers, this reality occurs in most countries of the world, given its relevance, there are many researches approached from different perspectives, among them, from the approach of computer science through its line of research called data mining that consists of prediction and classification investigations using the different algorithms.

This research work aims to develop a predictive data mining model applying Machine Learning techniques, which have contributed positively in the study of different fields such as medicine. When analyzing the business context, it was proposed to use three of the aforementioned techniques to predict future cases of pregnant mothers with anemia, for which the algorithms of Multilayer Perceptron, Naive Bayer and Decision Tree J48 were implemented, these were trained on a base of historical data of 422 records of pregnant mothers with anemia in the Province of Ilo, the algorithm that reached the highest precision was that of Naive Bayes with 89%, followed by that of decision tree J48 with 79% and finally the multilayer perceptron with 62%.

The development of the project was based on the CRISP-DM methodology to develop each of the stages that led to the final result.

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El procesamiento de información a lo largo de los años sufrió un cambio radical entre los diversos modelos de clasificación, desde el clustering hasta las reglas de asociación, Machine learning o conocido en su momento como Data Fishing, este último tenía como objetivo que la máquina logrará un procesamiento semejante al de la mente humana.

Con el paso del tiempo las investigaciones toman nuevos rumbos que abren campos en el aprendizaje automático, Deep learning busca complementar la parte de volumen de información frente a los diversos problemas que la sociedad adquiere.

Es necesario determinar el problema que se combatirá para buscar las herramientas necesarias, en ese sentido, en la actualidad existen diversos frameworks que facilitan la manipulación de información. En redes neuronales existen plataformas como anaconda que integra python, c++, R y otros lenguajes de programación, esta se adapta a cualquier problema, además que existen librerías que ayudan a buscar la mejor solución posible, así como a representarlas gráficamente.

Existen diversas investigaciones en el campo de ML y DL aplicadas especialmente al campo médico debido a que los investigadores ponen como prioridad el alto índice de riesgo vital, desde la búsqueda de tratamientos idóneos hasta el diagnóstico médico.

Esta investigación pretende desarrollar un modelo predictivo en determinar porcentualmente los casos de anemia en las madres gestantes.

I. PLANTEAMIENTO DEL PROBLEMA

1.1. DESCRIPCIÓN DE LA REALIDAD PROBLEMÁTICA

A nivel mundial la anemia es considerada como una enfermedad potencial, según la Organización Mundial de la Salud (2013) "La anemia afecta a 1620 millones de personas, lo que corresponde a 24.8% de la población mundial", tal como se puede apreciar en la siguiente tabla:

Tabla 1: Prevalencia mundial de la anemia y número de personas afectadas

Grupo de población	**Prevalencia de la anemia**		**Población afectada**	
	El por ciento	95% CI	Número (en millones)	95%
Niños en edad pre escolar	47.4	45.7-49.1	293	283-303
Niños de edad escolar	25.4	19.9-30.9	305	238-371
Embarazadas	41.8	39.9-43.8	56	54-59
Mujeres no embarazadas	30.2	28.7-31.6	468	446-491
Varones	12.7	8.6-16.9	260	175-345
Ancianos	23.9	18.3-29.4	164	126-202
Población total	**24.8%**	**22.9-26.7**	**1620**	**1500-1740**

Fuente: de Benoist B et al., eds. Worldwide prevalence of anaemia 1993-2005. Base de datos mundial sobre la anemia de la OMS, Ginebra, Organización Mundial de la Salud, 2008.

En Perú, hasta el año 2017, según el Instituto Nacional de Estadística e Informática, se registró más del 40% de niños con anemia, esto estaría asociado al escaso consumo de hierro en la alimentación de las gestantes. Como se aprecia en la figura que sigue:

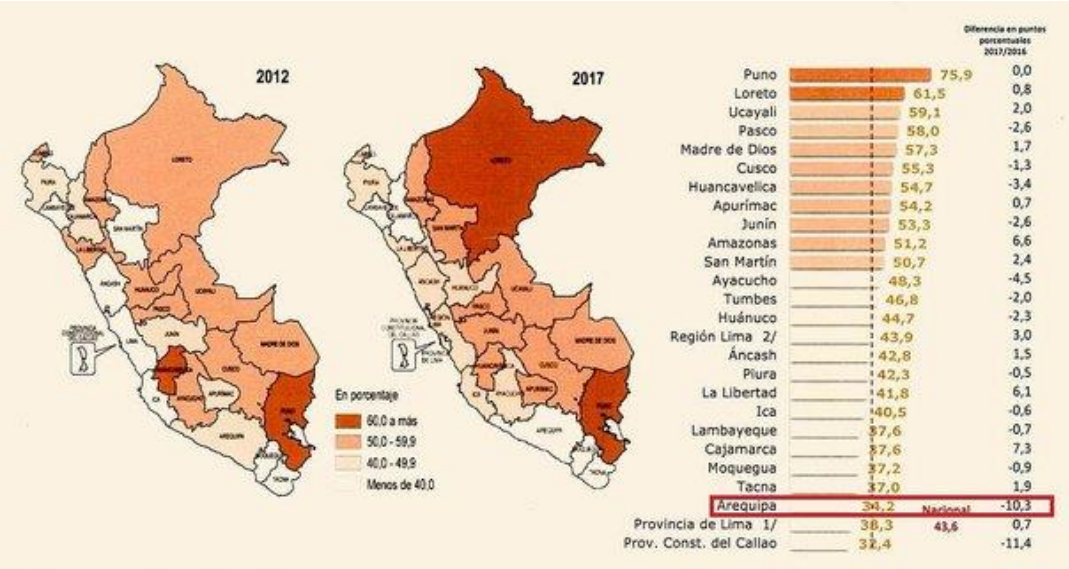


Figura 1: Proporción de niñas y niños de 6 a 36 meses de edad con anemia, según región, 2012-2017

La anemia, en la gestación se convierte en un problema que afecta tanto a la salud de la gestante como la del feto, en caso de no tomar las precauciones necesarias, las consecuencias a largo plazo pueden ser muy graves. Según los informes de la OMS, un quinto de la mortalidad perinatal y un décimo de la mortalidad materna en los países en desarrollo son atribuibles a la deficiencia de hierro.

Por otra parte, en ese estado de gravidez emerge una aparente anemia producto del aumento del volumen plasmático, el cual provoca una hemo dilatación que hace descender la cantidad de glóbulos rojos; así mismo aumentan los requerimientos de hierro, los que, si no se satisfacen, origina la anemia.

En la actualidad existen registros de gestantes que llegan al proceso de parto con un elevado riesgo de anemia que también padecería el feto.

Es necesario señalar que en el Perú existen reportes que aseguran que la carencia de hierro no solo afecta el peso del bebé al nacer y el estado inmunológico materno, sino que aumenta el riesgo de muerte durante el embarazo y parto.

Es por ello que el MINSA en su rol constante de mejorar la calidad de vida de las gestantes y los recién nacidos, requiere reforzar las estrategias de control y prevención de anemia en el periodo gestacional, atendiendo esa necesidad, el presente proyecto pretende aportar un modelo de minería de datos con cuyos resultados predictivos se puedan tomar decisiones coherentes y estratégicas.

De esta manera se contribuirá a la disminución de los casos de gestantes con anemia, mediante la detección anticipada a partir del contraste de historias clínicas de gestantes con anemia y gestantes sin anemia, la predicción se calculará de acuerdo a sus características del perfil, enfocándose en aquellas con alta probabilidad de padecer esta enfermedad. Teniendo esta información, el MINSA podría implementar estrategias o campañas preventivas que coadyuven a minimizar o erradicar el porcentaje de gestantes con esta enfermedad.

Por todo lo descrito en los párrafos precedentes es importante llevar a cabo el proyecto de investigación titulado: "MODELO DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE CASOS DE ANEMIA EN GESTANTES DE LA PROVINCIA DE ILO".

1.2. FORMULACIÓN DEL PROBLEMA

En la Provincia de Ilo, el tema de la salud es una preocupación constante dentro de los diferentes factores que influyen en el bienestar de las personas. Una de las principales preocupaciones del Ministerio de Salud son las gestantes.

1.2.1. Interrogante General

¿Se podrá implementar un modelo de minería de datos para la predicción de casos de anemia en madres gestantes para la provincia de Ilo?

1.2.2. Interrogantes Secundarias

¿Se podrá implementar un modelo de minería de datos aplicando la metodología CRISP-DM?

¿Se podrá validar el modelo de minería de datos a través de la validación cruzada?

1.3. JUSTIFICACIÓN

Dentro de uno de los objetivos del MINSA está el brindar un mejor servicio de atención a sus pacientes y en especial a las gestantes. Es bien sabido que la gestión de recursos para la atención de pacientes se basa en el promedio de atenciones médicas que posiblemente pueden darse y la información de apoyo a toma de decisiones. Las predicciones del presente trabajo permitirán tomar iniciativas para mejorar la atención a las gestantes que concurren a los establecimientos de salud.

El trabajo es factible de realizar por cuanto se ha considerado la información de historias clínicas y control de las madres gestantes almacenadas en las bases de datos del sistema WawaRed, por lo tanto, el presente estudio se justifica y resulta importante su implementación.

1.4. FORMULACIÓN DE OBJETIVOS

1.4.1. Objetivo General

Implementar un modelo de minería de datos para la predicción de casos de anemia en gestantes de la provincia de Ilo.

1.4.2. Objetivos Específicos

Proponer un modelo de minería de datos para la predicción de casos de anemia en gestantes de la provincia de Ilo aplicando CRISP-DM.

Validar el modelo de minería de datos aplicando validación cruzada.

1.5. FORMULACIÓN DE HIPÓTESIS

1.5.1. Hipótesis General

A través de la adecuada recopilación de diversas características de madres gestantes que han padecido anemia se podrá implementar un modelo de minería de datos para predecir casos de anemia.

1.5.2. Hipótesis Específicas

Aplicando la metodología CRISP-DM se podrá implementar un modelo de minería de datos para la predicción de casos de anemia en gestantes de la provincia de Ilo.

A través de la aplicación de la validación cruzada se podrá validar el modelo de minería de datos.

II. MARCO TEÓRICO

2.1. ANTECEDENTES

Existen diversos tipos de algoritmos de predicción en el actual estado de arte, de esta línea de investigación, algunos de los cuales se presentan a continuación: En el trabajo de (Abdullah & Al-Asmari, 2016) concluyen: “con la mejor precisión de predicción de tipos de anemia, con 93.27% de precisión con algoritmos SMO y J48 arboles de decisión, basada en un conjunto de datos de 41 pacientes”.

Es importante determinar qué factores son más influyentes en la anemia, en relación a eso, en la investigación de (Soto, 2016) concluye: “en determinación de los siguientes factores relevantes: edad gestacional, la edad de las gestantes, control prenatal, multíparas, primíparas, índice masa corporal, preeclampsia, eclampsia e intergenésico, cuáles son las principales variables como mayor índice de las madres con anemia en el Hospital Ginecobstetricia de San José de Callao - Lima”. Algunas de estas variables son consideradas en este trabajo de investigación.

En el estudio de (Gallego, Fernanda Navarro, & Castillo, 2015) se aplicó técnicas de minería de datos para “análisis de riesgo de mujeres gestantes de la población Manizaleña, específicamente K-means, Clúster, reglas de asociación y método de correlación, el trabajo se desarrolló con la metodología CRISP-DM “Por lo que el desarrollo de la presente investigación también de basa en la metodología CRISP-DM”.

En la investigación de (Ponce & Margain, 2016) concluyen: “el uso de minería de datos puede ser una herramienta de apoyo para identificar patrones de comportamiento en las personas, para predecir la predisposición de una enfermedad determinada”.

Considerando a (Retamar et al., n.d.) “se desarrolló la investigación con el objetivo de seleccionar y reducir la dimensionalidad de data de los nacimientos ocurridos entre 2009 y 2017, se experimentaron con tres algoritmos: Class Balancer, SMOTE y Spread Sub Sampled, para balanceo de clases, para la clasificación se aplicó arboles de decisión, J48, REP Tree y Random Tree”

Al decir de (Ponce & Margain, 2016) “mediante la aplicación de técnicas de minería de datos se identificó las características de los casos de muerte fetal y se determinó los factores de riesgo que inciden en el hecho”.

Del mismo modo, según (Sanap, Nagori, & Kshirsagar, 2011) “ Se aplicó algoritmo árbol de decisión de C4.5 para clasificación con 99.42 de precisión de los 514 instancias y 3 instancias incorrectas. Mientras que el Support Vector Machine alcanzo el 88.13%. Las técnicas de clasificación de minería de datos pueden proporcionar asistencia para hacer el diagnóstico o la clasificación de la anemia en función del recuento sanguíneo completo”.

En la investigación. (Sanap et al., 2011) presenta un proceso mejorado de selección secuencial de la función ADD-Left Remove-Right (ALRR) y el algoritmo de clasificación Gausnominal para predecir las clases de enfermedad de anemia (leve, no anémica y severa o moderada) según las técnicas de minería de datos.

“El mejor modelo obtenido J48 (94.39%) fue entrenado a partir del análisis de la comparación con otros modelos de minería de datos. El modelo Multilayer Perceptrón obtuvo un puntaje de 93.18% de precisión, Redes Neuronales, presenta un buen nivel de precisión y predicción, quizás con el análisis de mayores registros de información,

habría obtenido mejores resultados; Se esperaba que la técnica BayesNet y NaiveBayes obtenga mejores resultados, de acuerdo a la documentación revisada este se acomoda mejor en cuanto a temas de salud" (Sanap et al., 2011).

2.2. BASES TEÓRICAS

2.2.1. Minería de Datos

El siglo XXI es considerado por algunos autores de prestigio como la era de la información, puesto que el quehacer diario de los individuos genera información abrumadora que es almacenada en bases de datos de diferentes sistemas de información, esta masa de información valiosa podría ayudar a instituciones donde se toman decisiones sobre algún tema en específico, como el presente estudio, que pretende predecir casos de anemia de madres gestantes de un centro de salud.

En este contexto la tecnología de minería de datos tiene un roll muy importante, según (Leskovec, Rajaraman, & Ullman, 2014) "la definición más comúnmente aceptada de "minería de datos" es el descubrimiento de "modelos" para datos". Sin embargo, para (Han, Kamber, & Pei, 2012) minería de datos es "descubrimiento de conocimiento a partir de datos", además plantea una serie de pasos de minería de datos, que continuación se presenta:

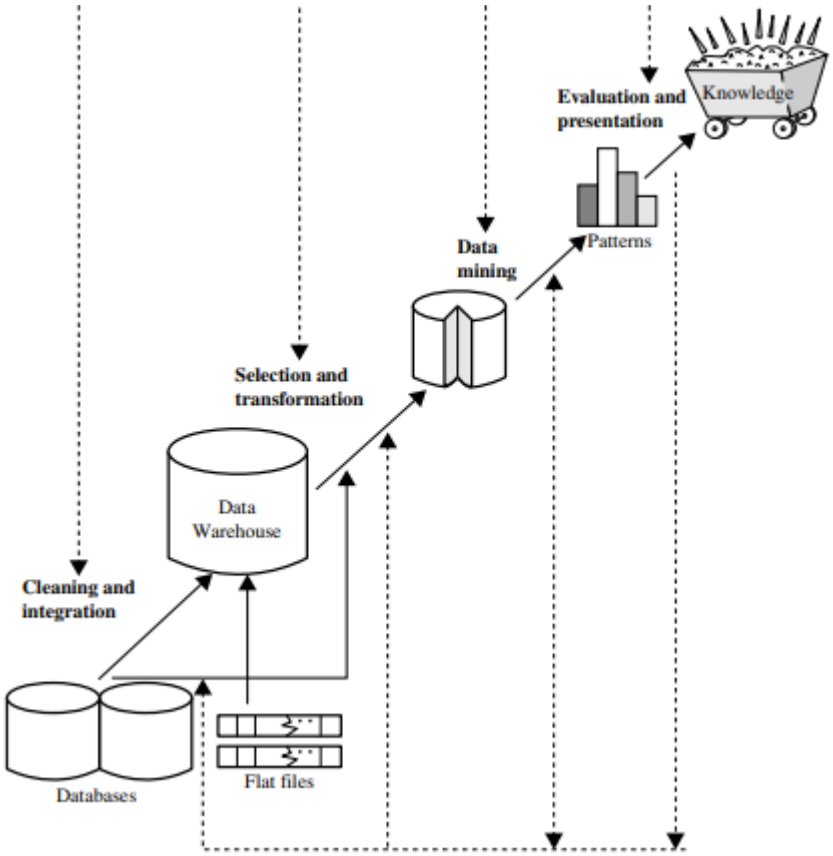


Figura 2: Pasos de minería de datos para descubrir el conocimiento.
Fuente: Data Mining Concepts and Techniques (Han, Kamber, & Pei, 2012)

1. Limpieza de datos. – una fase que consiste en remover datos inconsistentes y ruidosos.
2. Integración de data. – Donde se puede combinar diferentes datos.
3. Selección de data. - Una fase donde los datos relevantes son recuperadas de los bases de datos para el análisis.
4. Transformación de data. - Una fase donde la data es transformada y consolidada en un formato apropiado para la operación con técnicas de minería de datos.

5. Minería de datos. - La fase esencial de aplicación de técnicas para extracción de patrones interesante.
6. Modelo de evaluación. – La fase donde se identifica a patrones de representación de conocimientos basada en características.
7. Presentación de conocimiento. - Es la etapa de presentación de conocimientos y visualización.

Descubrir patrones de comportamiento en grandes volúmenes de datos acumulados en las empresas, es posible a la aplicación de minería de datos, estos datos pueden ser aprovechado para su uso adicional, por ejemplo, para la toma de buenas decisiones. Entonces se puede definir como un proceso de descubrimiento de nuevas y significativas relaciones de patrones de un conjunto de grandes cantidades de datos.

Sin embargo, la presente investigación es desarrollado basado en la metodología de CRISP-DM, una metodología validada y aplicada por muchos investigadores.

2.2.2. CRISP-DM (Proceso Estándar de la Industria Cruzada para la Minería de Datos).

Según (Piatetsky, s.f.), en la encuesta realizada por KDnuggets, la metodología más usada para desarrollo de proyectos de minería de datos y ciencia de datos es CRISP-DM, como se puede apreciar en la Figura 3.

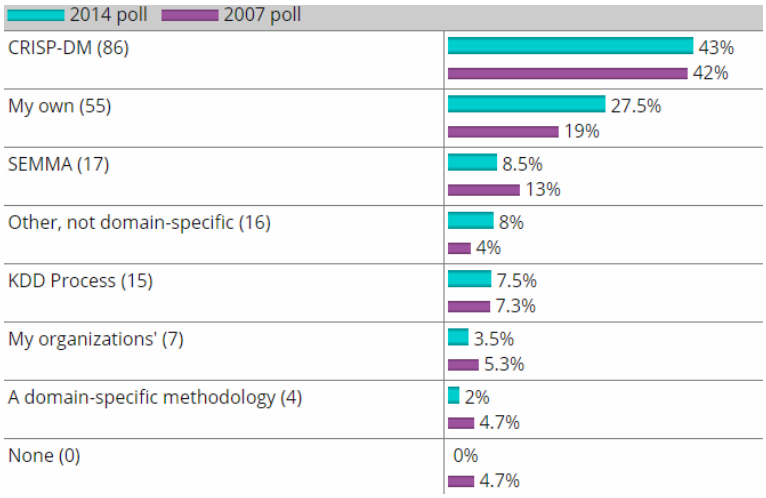


Figura 3: CRISP-DM, sigue siendo la mejor metodología para proyectos de análisis, minería de datos o ciencia de datos.

Fuente: KDnuggets.

La metodología más popular para desarrollo de proyectos de minería de datos consta de 6 fases, cada una de estas conduce a desarrollar de manera adecuada el proyecto de este contexto, como se puede apreciar en la Figura 4.

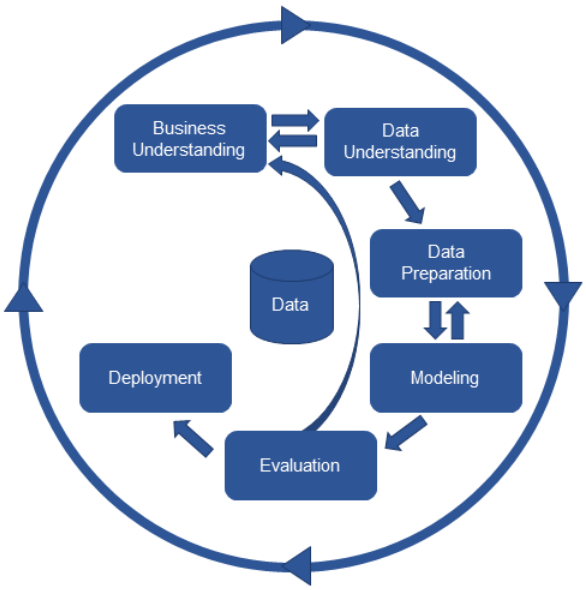


Figura 4: Metodología CRSP-DM.

Fuente: KDnuggets.

1. **Comprensión de negocio.**

   Para desarrollar un proyecto de minería de datos es importante entender bien el modelo de negocio del sector en el que se va a desarrollar el proyecto de minería de textos.

   Principalmente se “entienden los objetivos del negocio y requerimiento de perspectiva del negocio” (Wirth & Hipp), también se planifica los objetivos del proyecto según el contexto.

2. **Comprensión de data.**

   Es la fase donde se inicia con la recolección de los datos, análisis e identificación de los datos de calidad, también se identifican los posibles problemas.

3. **Preparación de data.**

   Es el proceso de estructuración de la data set, pero las principales actividades consisten en realizar la “selección de data, limpieza de los datos, pre estructuración de la data, integración de la data” (Brown, s.f.) y normalización de la data para que se puedan aplicar algoritmos de minería de datos.

4. **Modelado.**

   Es la fase donde se aplican las técnicas de minería de datos para descubrir el comportamiento de patrones, también es la etapa en la que se diseñan

pruebas, pero principalmente de construyen modelos y finalmente se evalúan los modelos seleccionados para el proyecto.

5. **Evaluación.**

   En esta etapa se evalúan los resultados obtenidos a partir de la aplicación de los modelos de minería de datos, también se hace una revisión al proceso y finalmente se terminan los siguientes nuevos pasos.

6. **Desarrollo.**

   Proceso de integración del proyecto al negocio, haciendo el uso de las metodologías de integración, pero antes se debe informar los resultados a los involucrados en el negocio y revisión del mismo.

2.2.3. Técnicas de Minería de Datos

En esta línea de ciencias de la computación existen diversos algoritmos basados en modelos matemáticos, estadísticos y probabilísticos, los cuales se clasifican como sigue:

Técnicas descriptivas

Las técnicas descriptivas tienen el objetivo de descubrir y explicar la relación de variables, “estas mostrarán nuevas relaciones entre las variables o excepciones de acuerdo a la empresa en que se utilice este proceso” (Virseda Benito & Román Carrillo).

- **Descripción de clases.**

  Existen tres características de este tipo, caracterización de la data, que consiste en hacer resúmenes de características generales, discriminación de data, lo cual consiste en comparación de características generales de los objetos de una clase con respecto a otro.

- **Análisis de asociación.**

  Es el descubrimiento de reglas de asociación de datos, este proceso que busca correlaciones relevantes de un conjunto de datos.

- **Análisis de clusters.**

  En síntesis, análisis de asociación, consiste en la agrupación de objetos maximizando la similitud dentro de una clase y minimizando la similitud entre clases.

Técnicas predictivas

Las técnicas predictivas según (Espino Timón, 2017) “consiste en la extracción de información existente en los datos y su utilización para predecir tendencias y patrones de comportamiento, pudiendo aplicarse sobre cualquier evento desconocido, ya sea en el pasado, presente o futuro”, a continuación, se presentan algunas técnicas predictivas más destacadas.

- **Clasificación y predicción.**

  Son técnicas de clasificación de clases basada en datos de variables, también se puede hacer predicción de un nuevo ítem en una de las clases del modelo, estos trabajan con datos cuantitativos y categóricos.

- **Arboles de decisión.**

  Existen diversas definiciones sobre Árboles de decisión, según (Unidad de Informacion y Analisis Financiero, 2014) "Un árbol de decisión es un modelo de clasificación que divide un conjunto de análisis, buscando el mayor grado de pureza entre los grupos resultantes".

- **Redes neuronales.**

  Según (Espino Timón, 2017) "Las redes neuronales se utilizan cuando no se conoce la naturaleza exacta de la relación entre los valores de entrada y de salida", sin embargo en los últimos tiempos "el término red neuronal ha evolucionado para abarcar una gran clase de modelos y métodos de aprendizaje" (Friedman, Tibshirani, & Hastie, 2017).

2.2.4. Redes Neuronales

Hay variadas definiciones, en el aspecto bilógico es "una simple unidad procesadora que recibe y combina señales desde y hacia otras neuronas" (Basogain Olabe).

"En esencia, se aplica un conjunto de entradas a la neurona, cada una de las cuales representa una salida a otra neurona. Cada entrada se multiplica por su

"peso" o ponderación correspondiente análoga al grado de conexión de la sinapsis. Todas las entradas ponderadas se suman y se determina el nivel de excitación o activación de la neurona. Una representación vectorial del funcionamiento básico de una neurona artificial se indica según la siguiente expresión de la ecuación". (Basogain Olabe, pág. 17).

Una red neuronal es un sistema conformado por las neuronas, capas y redes, como se puede apreciar en la figura.

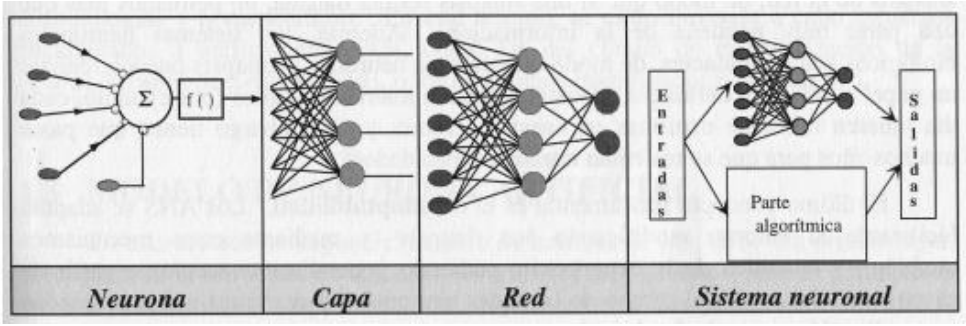


Figura 5: Partes del sistema de una red neuronal.

Fuente: Redes Neuronales (Larranaga, Inza, & Moujahid).

Redes neuronales artificiales monocapa.

Es el tipo de red neuronal conformada de una sola capa, según (Jorge Matich, 2001) "en las redes monocapa, se establecen conexiones entre las neuronas que pertenecen a la única capa que constituye la red. Las redes mono capas se utilizan generalmente en tareas relacionadas con lo que se conoce como auto asociación

(regenerar información de entrada que se presenta a la red de forma incompleta o distorsionada)".

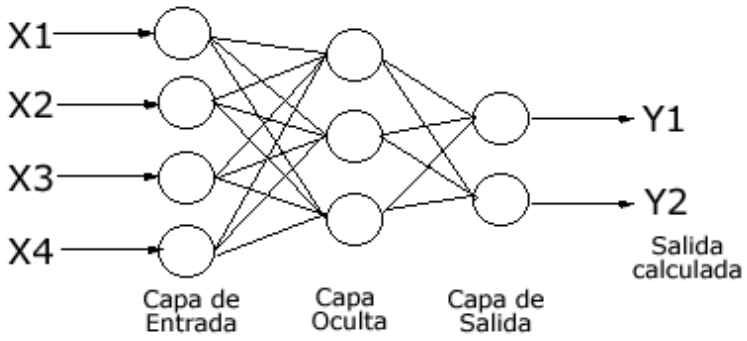


Figura 6: Estructura de la red neuronal monocapa.

Redes neuronales artificiales multicapa.

Las redes neuronales multicapa son formadas por más de una capa e interconectadas entre ellas, según (Basogain Olabe) "la salida de una capa es la entrada de la siguiente capa. Se ha demostrado que las redes multicapa

presentan cualidades y aspectos por encima de las redes de una capa simple."

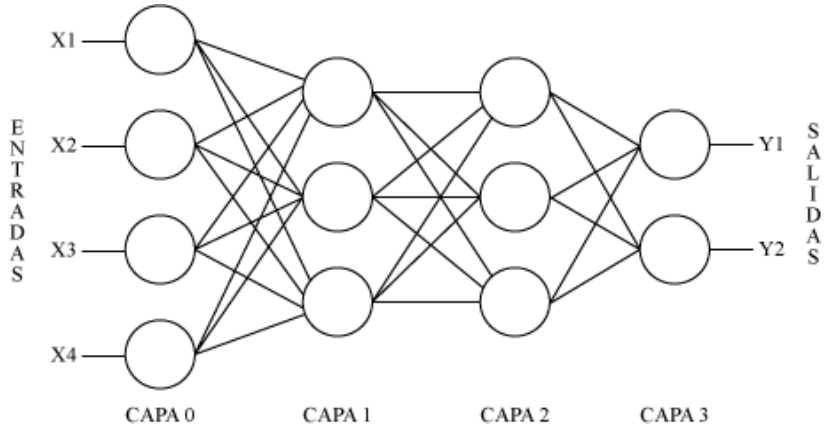


Figura 7: Estructura de la red neuronal multicapa.

La red neuronal se hace compleja según la cantidad de capas ocultas y neuronas por capa, esta complejidad hace que el algoritmo pueda ser robusto, sin embargo, su aplicación puede variar aplicando la hiper-parametrización como el pre-procesamiento de datos.

Existe un problema general aplicando redes neuronales, conocido como malos resultados, radica en ajuste de aprendizaje que generalmente tienen problemas de overfiting y underfitting, uno de los factores es la poca data en la muestra del entrenamiento y son sobre ajustados con los algoritmos de optimización, también se aplica el error cuadrático medio para tener un valor alto y pueda minimizarse en cada época con la retro propagación.

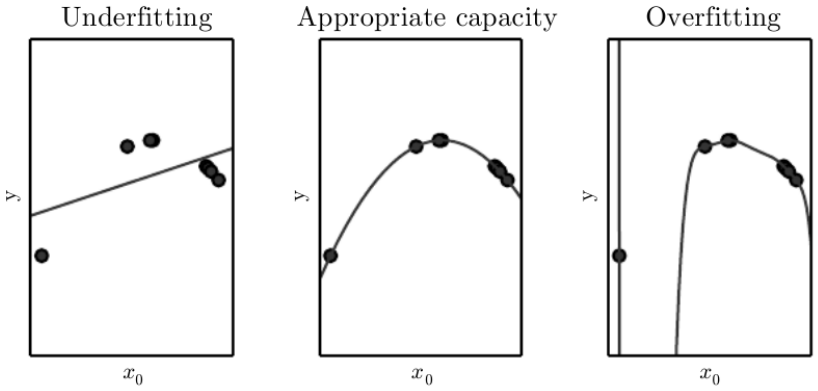


Figura 8: Sobre ajuste

2.2.5. Función de Activación

La función de activación es quien da forma a la salida en el hiperplano, es la definición de tipo no lineal y existen diversas funciones que son empleadas para una regresión o clasificación.

2.2.6.1. Función lineal

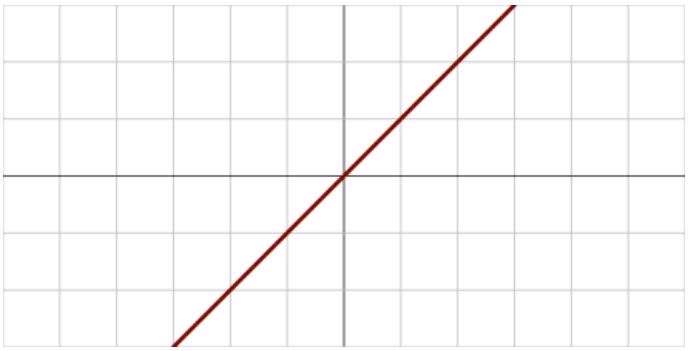


Figura 9: Función lineal

Fuente: Elaboración Propia

2.2.6.2. Función Binar Step

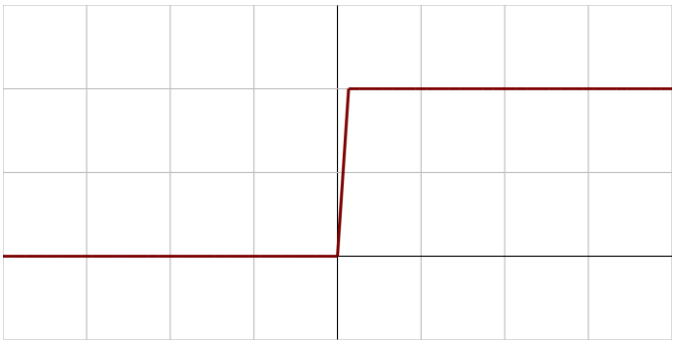


Figura 10: Función Binar Step

Fuente: Elaboración Propia

Fórmula:

$$f(x) = \begin{cases} 0 \; for \; x < 0 \\ 1 \; for \; x > 0 \end{cases} \quad (1)$$

2.2.6.3. Función Logística

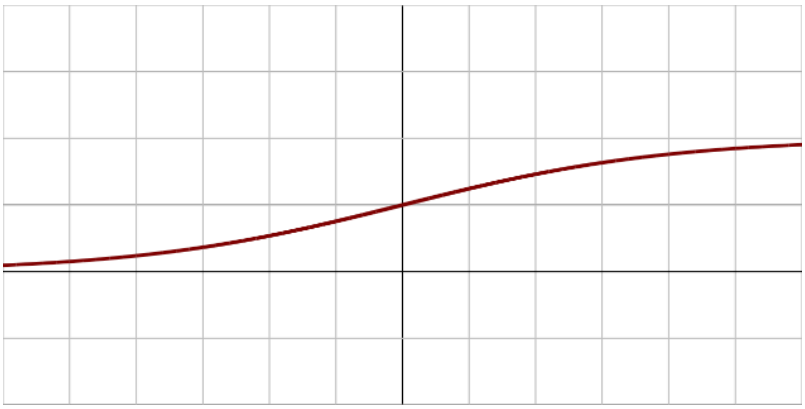


Figura 11: Función Logística

Fuente: Elaboración Propia

Fórmula:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (2)$$

2.2.6.4. Función Tangente Hiperbólica

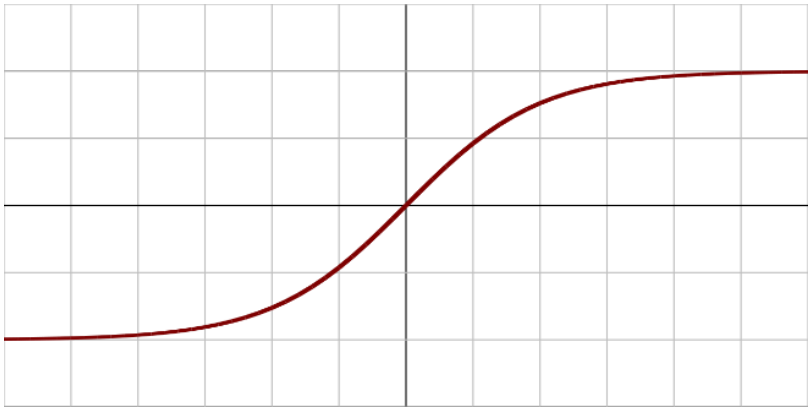


Figura 12: Función Tangente Hiperbólica

Fuente: Elaboración Propia

Fórmula:

$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad (3)$$

2.2.6.5. Función Arco Tangente

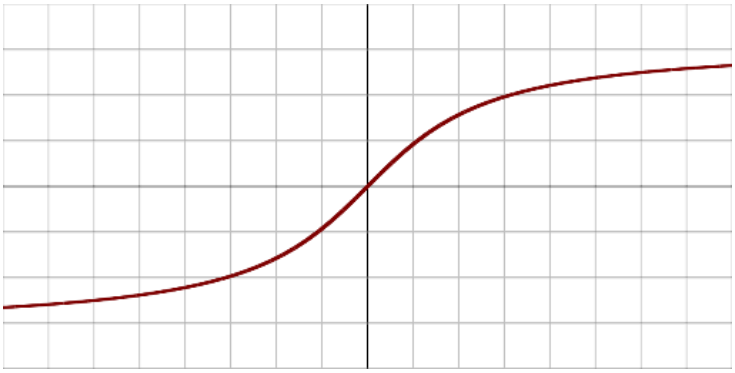


Figura 13: Función Arco Tangente

Fuente: Elaboración Propia

Fórmula:

$$f(x) = Tan^{-1}(x) \qquad (4)$$

2.2.6.6. Función Gaussian

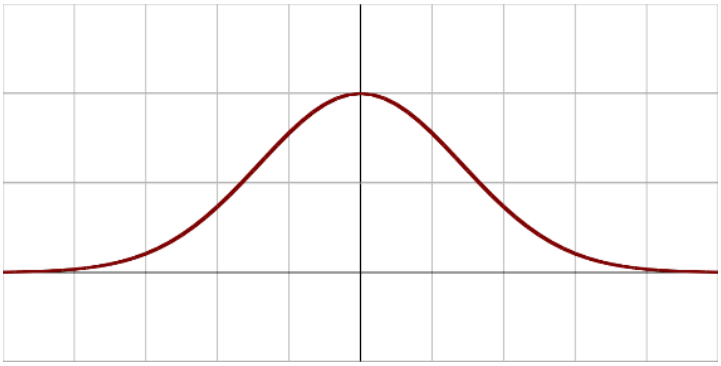


Figura 14: Función Gaussian

Fuente: Elaboración Propia

2.2.6. Árbol de decisión J48.

Es un algoritmo de árbol de decisión que permite analizar decisiones basadas en el uso de resultados y probabilidades asociadas, “es un conjunto de condiciones o reglas organizadas en una estructura jerárquica, de tal manera que la decisión final se puede determinar siguiendo las condiciones que se cumplen desde la raíz hasta alguna de sus hojas”. (Vizcaino Garzon, 2008).

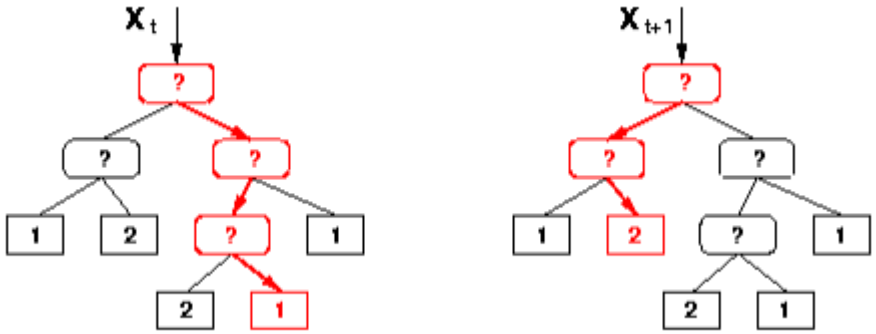


Figura 15: Representación gráfica de árbol de decisión

Fuente: (Vizcaino Garzon, 2008)

Los árboles de decisión son compuestos por un nodo de decisión, que indica que una decisión necesita tomarse en ese punto del proceso, está representado por un cuadrado. Nodo de probabilidad, que indica que en ese punto del proceso ocurre un evento aleatorio, está representado por un círculo. Rama, que muestra los distintos caminos que puede emprender cuando se toma una decisión o cuando ocurre algún evento aleatorio.

Otro algoritmo de árbol de decisión J48 utiliza el método Gini para crear puntos divididos. Para hallar el valor de Gini se aplica de la siguiente fórmula:

$$Gini(D) = 1 - \sum_{i=1}^{m} Pi^2$$

Donde, pi es la probabilidad de que una tupla en D pertenezca a la clase Ci.

El índice de Gini considera una división binaria para cada atributo. Puede calcular una suma ponderada de la impureza de cada partición. Si una división binaria en el atributo A divide los datos D en D1 y D2, el índice Gini de D es:

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D)_1 + \frac{|D_2|}{|D|} Gini(D)_2 + \frac{|D_n|}{|D|} Gini(D)_n$$

En el caso de un atributo de valor discreto, el subconjunto que proporciona el índice de Gini mínimo para el elegido se selecciona como un atributo de división. En el caso de los atributos de valor continuo, la estrategia es seleccionar cada par de valores adyacentes como un posible punto de división y punto con un índice de Gini más pequeño elegido como el punto de división.

2.2.7. Naive Bayes.

Naive Bayes es un algoritmo de clasificación, basada en probabilidad de Teorema de Bayes, principalmente es usada para clasificación de textos con multidimensional, pero también es efectiva en construir modelos predictivos. "La Clasificación Bayesiana representa un método de aprendizaje supervisado, así como un método estadístico para la clasificación. Asume un modelo probabilístico subyacente y nos permite capturar la incertidumbre sobre el modelo de una manera basada en principios al determinar las probabilidades de

los resultados. Puede resolver problemas de diagnóstico y predicción" (Collins, pág. 3)

Teorema de Bayes.

También conocido como leyes bayesianas, es un método de cálculo de probabilidad condicional, la probabilidad de un evento se basa en conocimientos previos del evento, la ecuación matemática se puede ver a continuación:

$$P\ (A|B)\ = \frac{P(B|A)P(A)}{P(B)}$$

Donde:

P (A | B): Probabilidad (probabilidad condicional) de ocurrencia del evento A dado que el evento B es verdadero.

P (A) y P (B): Probabilidades de ocurrencia del evento A y B respectivamente.

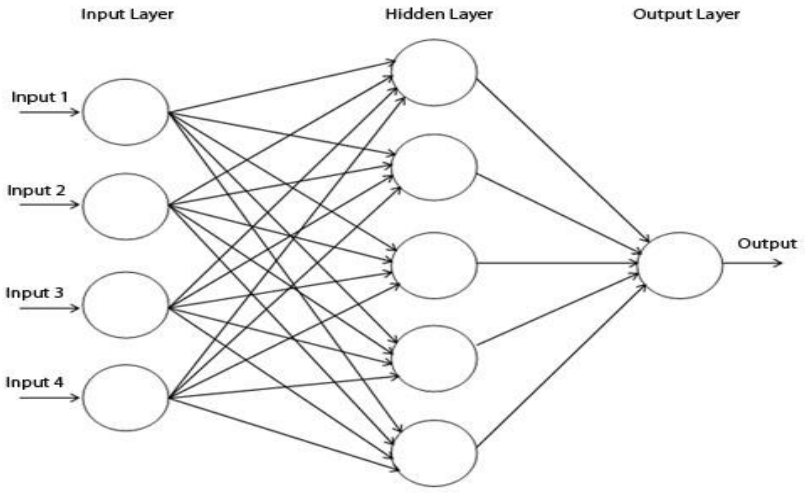
P (B | A): Probabilidad de la ocurrencia del evento B dado que el evento A es verdadero.

2.2.8. Perceptrón Multicapa

Es una red neuronal de múltiples capas para resolver problemas no separables linealmente, "es un tipo que asocia las variables de entrada con las salidas, el método más conocido es propagación del error hacia atrás, lo cual consiste en organizar una representación del conocimiento" (Palmer, Jiménez, & Montaño, 2001).

La arquitectura tiene tres principales capas, la capa de entrada, las capas ocultas y capa de salida, es de conexiones entre neuronas de son siempre adelante, es decir no existen conexiones laterales, ni hacia atrás, es decir siempre será hacia la salida, como se puede apreciar en la Figura N° 16.

Figura 16: Arquitectura de perceptrón multicapa.



2.2.9. Algoritmos de Optimización

2.2.7.1. Momento

La gradiente descendiente tiene problemas con la búsqueda de los mínimos locales, en estos escenarios la gradiente oscila por las pendientes, un método para acelerar la dirección de oscilación es la optimización por momento.

$$v_j = n\, v_{t-1} + \alpha \nabla_w J(\theta)$$

$$\theta = \theta - v_t$$

La ecuación 5, representa la gradiente y el impulso del momento representado como $'n'$ se multiplica por el coeficiente denominado momento del mínimo local v_t.

2.2.7.2. Adagrad

Es un algoritmo de optimización basada en la gradiente, con la finalidad de adaptarse a la velocidad de aprendizaje de los parámetros realizando actualizaciones en cada paso. Estos parámetros trabajan con los mínimos locales para reducir el error cuadrático medio en el entrenamiento de la red.

$$\Delta\theta_j = -\frac{R\ M\ S\ [\Delta\theta]_{t-1}}{R\ M\ S\ [g]_t} g_t$$

$$\theta_{t+1} = \theta_t - \Delta\theta_t$$

La tasa de aprendizaje se establece con la regla de actualización, por medio del error cuadrático medio RMS.

2.2.7.3. RMSprop

Es un método de tasa de aprendizaje adaptativo de manera independiente, que divide la velocidad de aprendizaje por el promedio decreciente exponencial de los gradientes cuadrados, similar al Adagrad.

$$E[g^2]_t = 0.9\ E[g^2]_{t-1} + 0.1 g_t^2$$

$$\theta_{t+1} = \theta_t - \frac{n}{\sqrt{E[g^2]_t} + \in} g_t$$

2.2.7.4. Adam

Método estocástico que calcula la tasa de aprendizaje por parámetro y almacena el resultado del decaimiento exponencial del gradiente por los medios cuadrados, similar al RMSprop, también mantiene el decreciente exponencial de los gradientes pasados por medio del impulso y su comportamiento radica en el sesgo para la actualización de los parámetros según la regla de actualización del primer y segundo momento.

$$m'_t = \frac{m_t}{1 - \beta_1^t}$$

$$v'_t = \frac{v_t}{1 - \beta_2^t}$$

Donde m_t y v_t, son los mínimos locales del primer y segundo momento.

2.2.10. Anemia

Anemia es una enfermedad en la sangre en estado líquido que afecta a la salud y la calidad de vida de los seres humanos, existen diversos tipos de anemia como: Anemia por deficiencia de hierro, anemia perniciosa, anemia aplasia y la anemia hemolítica. Esta enfermedad puede afectar a diversas edades.

Causas de anemia

"Los glóbulos rojos contienen hemoglobina, una proteína que transporta oxigeno por el cuerpo del ser humano, cuando mejor producción de estos glóbulos rojos, o se destruye demasiados, puede generar malestar al organismo del ser humano y por ende causar cansancio otros síntomas" (US Department of Health and Human Services, 2011).

A continuación, se presenta algunas posibles signos y síntomas de la anemia:

- Cansancio o debilidad
- Piel pálida o amarilla
- Desaliento o mareo
- Sed en exceso
- Sudor
- Puso débil y rápido
- Respiración rápida
- Problemas en el corazón

En el Perú según la Norma Técnica – Manejo Terapéutico y Preventivo de la Anemia en Niños, Adolescentes, Mujeres Gestantes y Puérperas, reporta lo siguiente:

Población	Con Anemia Según niveles de Hemoglobina (g/dL)			Sin anemia según niveles de Hemoglobina
Niños				
Niños Prematuros				
1ª semana de vida	≤ 13.0			>13.0
2ª a 4ta semana de vida	≤ 10.0			>10.0
5ª a 8va semana de vida	≤ 8.0			>8.0
Niños Nacidos a Término				
Menor de 2 meses	< 13.5			13.5-18.5
Niños de 2 a 6 meses cumplidos	< 9.5			9.5-13.5
	Severa	Moderada	Leve	
Niños de 6 meses a 5 años cumplidos	< 7.0	7.0 - 9.9	10.0 - 10.9	≥ 11.0
Niños de 5 a 11 años de edad	< 8.0	8.0 - 10.9	11.0 - 11.4	≥ 11.5
Adolescentes				
Adolescentes Varones y Mujeres de 12 - 14 años de edad	< 8.0	8.0 - 10.9	11.0 - 11.9	≥ 12.0
Varones de 15 años a más	< 8.0	8.0 - 10.9	11.0 - 12.9	≥ 13.0
Mujeres NO Gestantes de 15 años a más	< 8.0	8.0 - 10.9	11.0 - 11.9	≥ 12.0
Mujeres Gestantes y Puérperas				
Mujer Gestante de 15 años a más (*)	< 7.0	7.0 – 9.9	10.0 - 10.9	≥ 11.0
Mujer Puérpera	< 8.0	8.0 – 10.9	11.0 - 11.9	≥ 12.0

Figura 17: Valores normales de concentración de hemoglobina y niveles de anemia en Niños, Adolescentes, Mujeres Gestantes y Puérperas (hasta 1,000 msnm)

Fuente: Ministerio de Salud (MINSA)

2.3. DEFINICIÓN DE TÉRMINOS

ANN = Red neuronal artificial

CNN = Red convolucional

RNN = Red recurrente

MLP = Multilayer Perceptrón

BP = Back Propagación

DL = Deep Learning

ML = Machine Learning

DM = Data Mining

III. MARCO METODOLÓGICO

3.1. TIPO Y DISEÑO

El proyecto está basado en una investigación pre-experimental, en este tipo de investigación se analiza una sola variable y prácticamente no existe ningún tipo de control. No existe manipulación de la variable independiente ni se utiliza grupo control (Ávila Baray, 2006). En una investigación pre-experimental el tipo de diseño consiste en administrar un tratamiento a un determinado grupo y después aplicar una medición en una o más variables orientado a identificar algunos patrones que permitan pronosticar un aproximado de nacimientos para el MINSA de la provincia de Ilo. Para el diseño de contrastación de la hipótesis, se usará lo siguiente. Grupo (Rango 50 meses comprendidos desde 2018-2019), Grupo experimental con X.

G X O

Donde:

O: Medición de los sujetos del grupo (Entrenamiento, validación)

G: Grupo de sujetos (Grupo1, Grupo2)

X: Aplicación del Estímulo (Modelo de minería de datos)

Aplicando las técnicas de machine learning se pretende realizar un diseño que entrene y valide datos del algoritmo propuesto.

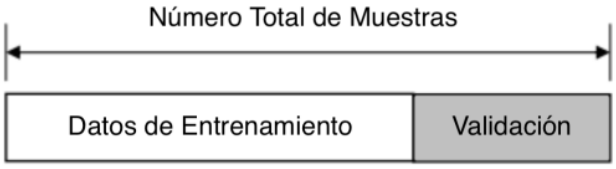


Figura 18: Modelo Experimental

Fuente: Elaboración Propia

3.2. NIVEL DE INVESTIGACIÓN

Se pretende realizar una investigación tecnológica aplicada, de tipo experimental, con diseño cuasi pre experimental. El nivel de investigación será descriptivo, explicativo, predictivo y relacional.

3.3. OPERACIONALIZACIÓN DE VARIABLES

3.3.1. Variable Dependiente

Minería de datos.

3.3.2. Variable Independiente

Predicción de caso de anemia en madres gestantes.

Tabla 2: Operacionalización de variables

Variable	Tipo	Indicador	Escala de medición
Minería de datos	Independiente	Probabilístico	Probabilístico
Predicción de casos de anemia en madres gestantes	Dependiente	Lugar Nacimiento	SI/NO
		Educación	
		Ocupación	
		Estado Civil	
		Tipo Seguro	
		Número Gestaciones	
		Hijos Vivos	
		Menarquia	
		Duración Menstruación	
		IMC	
		Número Abortos	
		Edad Gestacional	
		Tipo Socioeconómico	
		Class	

Fuente: Modelo CRISP-DM

3.4. POBLACIÓN Y MUESTRA

3.4.1. Población

La información abarca los casos de anemia en niños, adultos y gestantes, en los periodos 2018 y 2019 de la región Moquegua, sin embargo, se observó en el

grupo de gestantes un crecimiento y se hizo énfasis en los casos de anemia de los mismos.

3.4.2. Muestra

Se toma como muestra los casos de anemia en gestantes de la provincia de Ilo, recolectando información de todos los centros de salud con casos de anemia en gestantes.

3.5. DISEÑO EXPERIMENTAL

Variables	
Variable independiente:	Variable dependiente:
Modelo de minería de datos basada en CRISP-DM	**Predicción de casos de anemia en madres gestantes**
Comprensión de negocio	Revisión de PEI Revisión de ROF
Comprensión de data	Determinación de variables de data.
Preparación de data	Limpieza de data y transformación de datos.
Modelado	Implementación del modelo basado en CRISP-DM.
Evaluación	Evaluación del modelo implementado.
Desarrollo	

3.6. TÉCNICAS E INSTRUMENTOS PARA LA RECOLECCIÓN DE DATOS

Se recolecto los datos acumulados desde 2018 -2019 en Wawared, bajo previo coordinación y autorización de dato para la presente investigación.

3.7. MÉTODOS Y TÉCNICAS PARA LA PRESENTACIÓN Y ANÁLISIS DE DATOS

Técnicas de minería de datos.

3.8. VALIDACIÓN Y CONFIABILIDAD DE LOS INSTRUMENTOS

Son datos recopilados de casos reales de madres gestantes con anemia y/o sin anemia por Wawared de la provincia de Ilo.

IV. PRESENTACIÓN DE RESULTADOS

4.1. COMPRENSION DE NEGOCIO

4.1.1. Comprensión de MOF de departamento de ginecología y obstetricia

Según el organigrama del departamento de ginecología y obstetricia, la máxima autoridad es el director de programa sectorial, seguido por el técnico administrativos y entre otras divisiones según corresponda, los objetivos del departamento es prestar cuatro tipos de servicio, servicio de ginecología, servicio de obstetricia, servicio producción humano y servicio de obstétrica, la dirección estratégica está conformada por el director sectorial de ginecología y obstetricia, los principales actividades o servicios son soportadas por técnico administrativo, obstetras y ginecólogos según el organigrama del departamento, por lo tanto, el técnico administrativo tiene un rol de soporte muy importante para la toma de decisiones estratégicas del sector, puesto que registra todo tipo de información

en el sistema Wawared, en el presente proyecto pretende usar los datos de madres gestantes con anemia y sin anemia para la predicción de casos de anemia en gestantes de la provincia de Ilo.

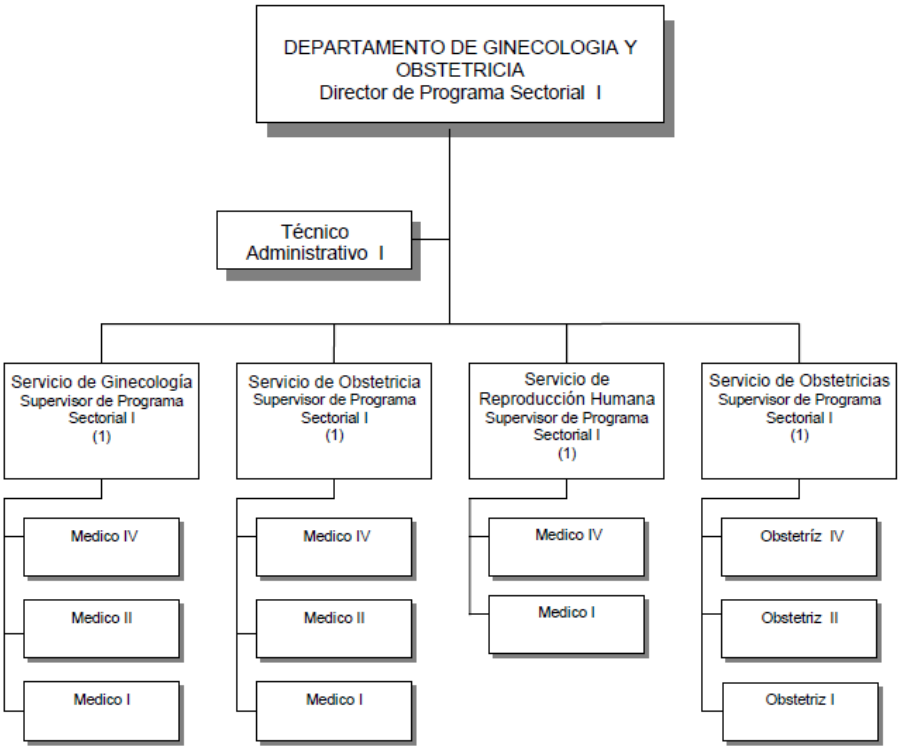


Figura 19: Organigrama del Departamento de Ginecología y Obstetricia

Fuente: MOF de MINSA

Según manual de organización y funciones del departamento génico obstetricia en sus funciones específicas 4.12 y 4.13 determina que el técnico administrativo se encarga de: Elaborar diariamente el HIS (Registro Diario de Atenciones y otras actividades) de Partos, Cesáreas, AQV y Recién Nacidos atendidos en el Servicio. Remisión mensual a la Unidad de Estadística e Informática de la Institución el reporte HIS de partos, cesáreas, Recién Nacidos, Defunción Peri natal (Óbito Fetal), AQV y Certificados de Nacimiento.

Estas funciones específicas y otras tienen como objetivo: Recepcionar, registrar, clasificar y distribuir la documentación recibida y los que genera el departamento en forma oportuna para la toma de decisiones por parte de la Jefatura. Ejecutar y supervisar las actividades de gran complejidad de apoyo secretarial, en cumplimiento con los objetivos funcionales de la Institución.

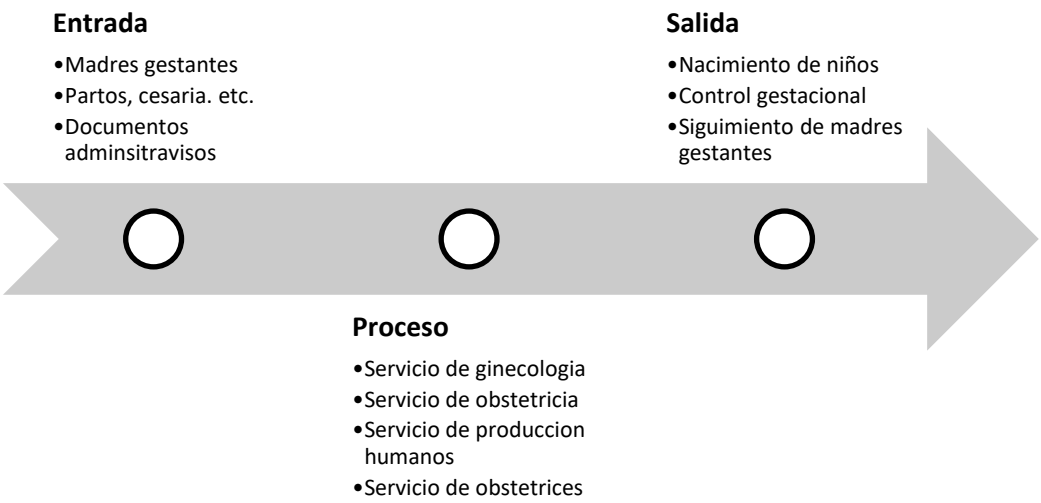


Figura 20: Modelo de negocio del departamento de Ginecología y Obstetricia

Fuente: MOF de MINSA

4.2. COMPRENSION DE DATA

4.2.1. Recolección de data y descripción

La primera etapa de metodología CRISP-DM describe la alineación de los objetivos del proyecto con la investigación, por ende, el levantamiento de la información fue recolectada mediante el sistema Wawared que comprende del periodo 2018 al 2019, dicha información se alinea con los objetivos del proyecto y la información adquirida servirá para analizarla y poder interpretar el modelo que se desea diseñar.

La información recolectada comprende los siguientes indicadores:

Tabla 3: Diccionario de Datos

Variable	**Descripción**	**Categoría**
Edad	Edad del paciente	No tiene
Lugar Nacimiento	Departamento de nacimiento	Amazonas Arequipa Ica Lima Madre De Dios Moquegua Puno Tacna Ucayali
Educación	Tipo de Educación de la gestante	Primaria Secundaria Superior no universitaria
Ocupación	Tipo de trabajo de la gestante	Ama de casa Comerciante Estudiante Obrero Otros Profesional Sin ocupación
Estado Civil	Estado civil de la gestante	Casada Conviviente Soltera
Tipo Seguro	Tipo de seguro de la gestante	SIS Otros

Número Gestaciones	Cantidad de embarazos	No tiene
Hijos Vivos	Cantidad de hijos con vida	No tiene
Menarquia	Día de menstruación	No tiene
Duración Menstruación	Días de duración de la menstruación	No tiene
IMC	Refleja el peso y talla	No tiene
Número Abortos	Cantidad de Abortos de la gestante	No tiene
Edad Gestacional		No tiene
Tipo Socioeconómico	Tipo socioeconómico de la gestante	No Pobre Pobre
Class	Clasificador de anemia	Anemia No Anemia

Fuente: Elaboración Propia

La fase de Análisis de Datos comprende, la distribución estadística, gráfico de cajas y otras instancias.

La recolección de Datos descrita en la tabla 3, es utilizada para determinar los rangos cuantitativos, de esta manera limita un rango que posteriormente es analizada haciendo una inferencia estadística, los datos recolectados como se observa en la ilustración (14) especifica el tipo de valor que tienen las etiquetas de las cuales necesitan ser convertidas en un valor cuantitativo para poder realizar analizar la data y sirva como base al modelo predictivo.

Tabla 4: Datos no procesados

	Lugar_nacim	Educacion	Ocupacion	Estado_civil	Tipo_seguro	Embarazos_p	Embarazos_p	Menarquia_e	d
29	Ucayali	Secundaria	Ama de casa	Casada	SIS	3	3	12	
19	Tacna	Secundaria	Estudiante	Conviviente	SIS	2	2	13	
28	Lima	Superior no	Comerciante	Soltera	SIS	1	1	14	
21	Moquegua	Superior no	Estudiante	Conviviente	SIS	1	1	10	
38	Moquegua	Superior	Ama de casa	Casada	SIS	1	1		
27	Madre De Di	Superior no	Ama de casa	Conviviente	SIS	1	1	13	
25	Arequipa	Superior no	Otros	Soltera	SANIDAD	0	0		
21	San Martin	Secundaria	Ama de casa	Soltera	SIS	2	1		
34	Moquegua	Superior no	Ama de casa	Casada	SIS	4	3	12	
21	Callao	Superior	Estudiante	Conviviente	SIS	0	0		
40	Moquegua	Secundaria	Ama de casa	Conviviente	SIS	2	2		
27	Cusco	Secundaria	Empleada	Conviviente	OTROS	1	1	14	
30	Puno	Secundaria	Ama de casa	Conviviente	SIS	3	2	14	
30	Puno	Primaria	Comerciante	Conviviente	SIS	2	2	13	
30	Loreto	Secundaria	Ama de casa	Conviviente	SIS	3	3		
27	Ancash	Secundaria	Ama de casa	Casada	SIS	1	1	11	
36	Moquegua	Superior no	Ama de casa	Conviviente	SIS	2	2	11	
19	Moquegua	Superior no	Estudiante	Soltera	SIS	0	0	11	
20	Moquegua	Secundaria	Ama de casa	Soltera	SIS	1	1	9	
32	Lambayeque	Secundaria	Ama de casa	Conviviente	SIS	3	3	15	
31	Ica	Superior	Obrero	Conviviente	SIS	2	2	12	
30	Piura	Superior no	Ama de casa	Conviviente	SIS	3	3	15	

4.2.2. Preparación de Datos

Para preparar la data se lleva una descripción estadística de todas las etiquetas para determinar la media y desviación estándar, también se muestran los gráficos de la data para verificar la igualdad de partición en la clase a predecir, un factor clave en esta fase es el sobre muestreo en la clase minoritaria y se identifica en la observación de la data.

Tabla 5. Descripción Estadística

	count	Mean	std	min	25 %	50 %	70 %	max
EDAD	294	27.9	5.9	15.0	23.0	27.0	32.0	46.0
LUGAR NACIMIENTO	294	12.8	4.8	0.0	13.0	14.0	14.0	23.0
EDUCACION	294	1.7	0.8	0.0	1.0	2.0	3.0	3.0
OCUPACION	294	1.3	2.0	0.0	0.0	0.0	3.0	7.0
ESTADO CIVIL	294	1.1	0.6	0.0	1.0	1.0	2.0	2.0
TIPO SEGURO	294	3.4	1.1	0.0	4.0	4.0	4.0	4.0
NUMERO GESTACIONES	294	1.5	1.4	0.0	0.0	1.0	2.0	8.0
HIJOS VIVOS	294	1.1	1.1	0.0	0.0	1.0	2.0	7.0
MENARQUIA	294	12.7	1.5	8.0	12.0	13.0	13.0	18.0
DURACION MENSTRUACION	294	6.1	6.9	0.0	3.2	4.0	5.0	30.0
IMC	294	1.9	0.9	0.0	1.0	2.0	2.0	5.0
NUMERO ABORTOS	294	0.3	0.6	0.0	0.0	0.0	0.7	3.0
EDAD GESTACIONAL TRIMENSTRAL	294	1.7	0.6	1.0	1.0	2.0	2.0	3.0
TIPO SOCIOECONOMICO	294	0.7	0.4	0.0	1.0	1.0	1.0	1.0
CLASS	204	0.5	0.5	0.0	0.0	0.5	1.0	1.0

Fuente: Elaboración Propia

Parte de la preparación es ver la distribución de la etiqueta predictiva, una forma de realizar es graficar si la clase está balanceada debido que un desbalanceo no permite realizar una predicción deseada, sin embargo, existen métodos como el oversampling y undersampling que dan soporte a las investigaciones con problemas de balanceo, la estandarización de la clase a predecir se refleja en la siguiente imagen.

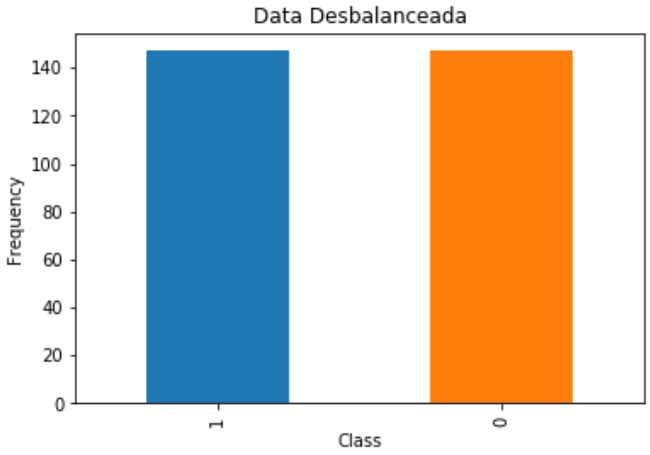


Figura 21: Data Balanceada

Fuente: Elaboración Propia

Se realiza al verificar el ruido en una data, es esencial, esto se realiza mediante un diagrama de caja que verifica la distribución y normalización de la data.

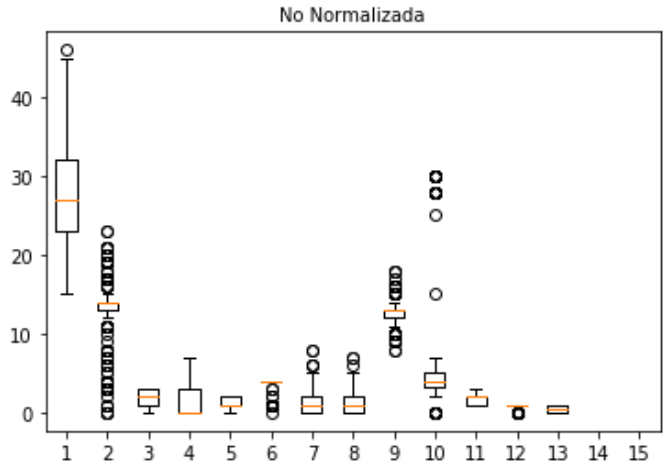


Figura 22: Diagrama de Caja - No Normalizado

Fuente: Elaboración Propia

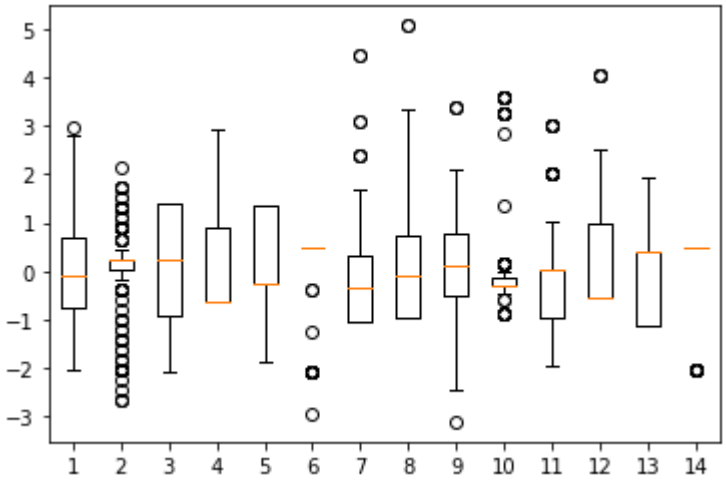


Figura 23: Diagrama de Caja – Normalizada

Fuente: Elaboración Propia

Por último, se realiza un diagrama de las variables en el hiperplano para graficar la clasificación de la misma.

4.3. PREPARACION DE DATA

4.3.1. Estructuración de data

En mención de en estructuración de data se realiza un conteo de los rangos de cada etiqueta, con la finalidad de saber los máximos y mínimos en el diagrama de caja y saber los cuartiles que contienen ruido. En Scikit Learn existe una función, labelEncoder es un método que transforma la etiqueta a un valor numérico con lo cual es parte del procesamiento de clasificar la información para posteriormente introducirla al modelo predictivo y este a su vez refleje el valor deseado de la clase tendenciosa.

Tabla 6: Procesamiento de Data

Variable	**Rango**
Edad	0 al 100
Lugar Nacimiento	0 al 8
Educación	0 al 3
Ocupación	0 al 6
Estado Civil	0 al 2
Tipo Seguro	0 – 1
Número Gestaciones	0 al 7
Hijos Vivos	0 al 6
Menarquia	0 al 16
Duración Menstruación	0 al 30
IMC	0 al 5
Número Abortos	0 al 5
Edad Gestacional	0 al 3
Tipo Socioeconómico	0 – 1
Class	0 – 1

Fuente: Elaboración Propia

Realizada la categorización, como resultado obtendremos una data procesada con valores numéricos que refleja el valor de cada registro, este tipo de procesamiento es parte de una normalización de datos.

	Edad	Lugar_nac	Educacion	Ocupacion	Estado_civ	Tipo_segu	NUMERO	HIJOS VIV	Menarquia	duracion_	Talla	Peso_habi	IMC	abortos	El
1	29	21	1	0	0	4	3	3	12	5	149	83	4	0	
2	19	19	1	3	1	4	2	2	13	7	158	61.5	1	0	
3	28	11	3	1	2	4	1	1	14	5	150	65	2	0	
4	21	14	3	3	1	4	1	1	10	7	156	64.4	2	0	
5	38	14	2	0	0	4	1	1	13	4	153	63	2	0	
6	27	13	3	0	1	4	1	1	13	4	156	89.5	4	0	
7	25	3	3	5	2	3	0	0	13	4	165.2	77.7	2	0	
8	21	18	1	0	2	4	2	1	13	4	155.2	69	2	1	
9	34	14	3	0	0	4	4	3	12	4	157.8	95.4	4	1	
10	21	23	2	3	1	4	0	0	13	4	147.2	54	1	0	
11	40	14	1	0	1	4	2	2	13	4	155.4	64.8	2	0	
12	27	4	1	2	1	1	1	1	14	30	158	64.3	2	0	
13	30	17	1	0	1	4	3	2	14	30	152.4	83	4	1	
14	30	17	0	1	1	4	2	2	13	7	154.9	77.5	3	0	
15	30	12	1	0	1	4	3	3	13	4	144.3	55.7	2	0	
16	27	1	1	0	0	4	1	1	11	7	146.2	53.3	1	0	
17	36	14	3	0	1	4	2	2	11	30	155.4	60.8	2	0	
18	19	14	3	3	2	4	0	0	11	4	157	66	2	0	
19	20	14	1	0	2	4	1	1	9	3	153.3	86	4	0	
20	32	10	1	0	1	4	3	3	15	28	156.7	61.5	2	0	
21	31	7	2	4	1	4	2	2	12	28	159	58.8	1	0	
22	30	16	3	0	1	4	3	3	15	30	163.9	83	3	0	

Figura 24: Data Procesada

4.4. MODELADO

4.4.1. Construcción del Modelo Perceptrón Multicapa.

La construcción del Modelo de una MLP, se grafica con 3 capas:

- La primera capa corresponde a los valores de entradas que son las 14 variables analizadas y procesadas en las fases de la metodología CRISP DM.
- La segunda capa corresponde a la capa oculta, que es parte del proceso en las redes neuronales para realizar una retropropagación y el modelo sea óptimo, esta capa consta de 7 neuronas
- Por último, la capa deseada a predecir es de 1 neurona cuyo valor es 0 o 1

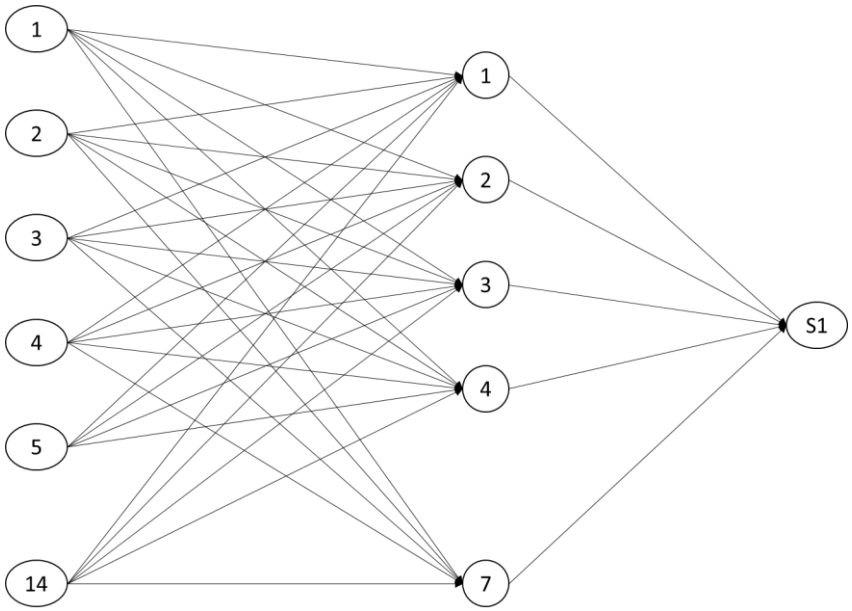


Figura 25: Red MLP

Fuente. Elaboración Propia

En Python, existen frameworks como keras que está desarrollada en base a redes neuronales, dentro del cual tiene un modelo clasificador, el código fuente de la red MLP en keras es el siguiente:

```
def Model():
    classifier = Sequential()
    classifier.add(Dense(units = 14, kernel_initializer='uniform', activation = 'relu', input_dim = 14))
    classifier.add(Dense(units = 7, kernel_initializer='uniform', activation = 'relu'))
    classifier.add(Dense(units = 1, kernel_initializer='uniform', activation = 'sigmoid'))
    classifier.compile(optimizer = 'adam' , loss = 'mean_squared_error',  metrics = ['acc','mse'])
    return classifier
```

Figura 26: Red Neuronal en Keras

Fuente: Elaboración Propia

La imagen describe, un modelo secuencial de un clasificador por capa, la función Dense describe las variables de 1 capa, en la primera capa describe el valor 'units' como las neuronas de la primera capa, la capa de entrada consta de 14 etiquetas, la inicialización de pesos mediante la derivada 'kernel' es de rango de

-0.5 a +0.5 debido que presenta mayor valor de error cuadrático medio y es óptimo para realizar una convergencia. La función de activación relu y por último tiene una dimensión de 14, similar a las neuronas de la capa principal y en este caso cada Dense identifica 1 capa de la red MLP

Por último, la compilación el clasificador es mediante un optimizador “Adam” que permite realizar los ajustes con el error cuadrático medio y la función ‘mean_squared_error’ carga dicho ajuste, también se evalúa la red MLP con las métricas que se describen en la etapa de evaluación del modelo.

Como ajuste se realiza una correlación de las etiquetas para determinar las más influyentes y sea el patrón de las futuras madres con anemia, esta correlación determina las variables con más del 70% que son relevantes para el modelo predictivo, como resultado se obtiene:

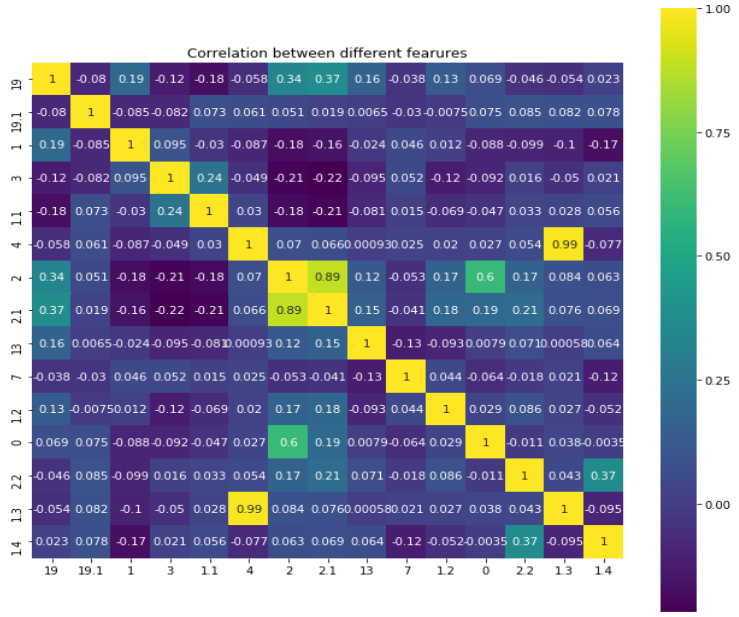


Figura 27: Correlación de Etiquetas

Fuente: Elaboración Propia

El gráfico muestra que las variables Edad, Educación, Ocupación, Número_Gestaciones,Hijos_Vivos,Tipo_Socioeconomico,Edad_gestacional_trimetral son las más influyentes en el modelo de red MLP, sin embargo este patrón de variables será variado cuando se incremente la data.

4.4.2. Evaluación del Modelo Perceptrón Multicapa

La red neuronal trabaja con una validación cruzada y entre los resultados destacan los siguientes gráficos.

El modelo de precisión refleja que entre la data de entrenamiento y testeo existen una precisión del 80%, está dentro de los parámetros de modelos de machine learning.

Ilustración 1: Model Accuracy

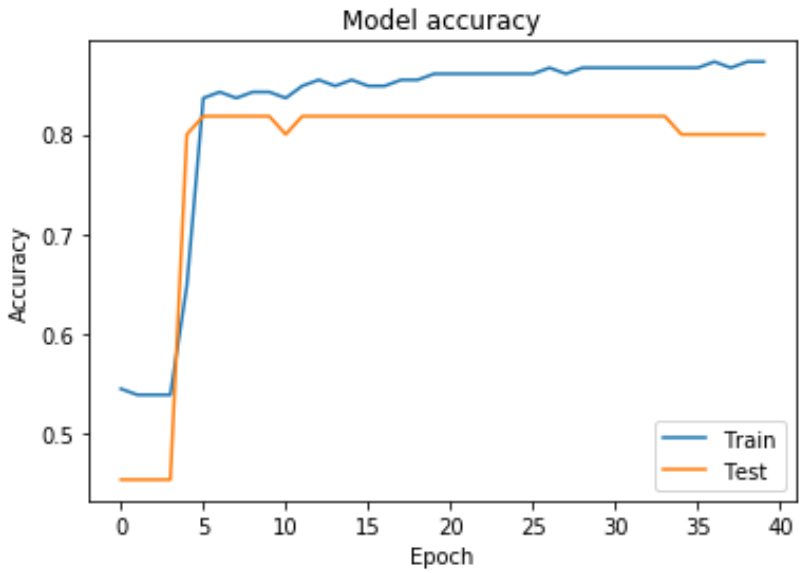


Figura 28:Model Accuracy
Fuente: Elaboración Propia

La red neuronal grafica la pérdida del error por medio del error cuadrático medio y aplicando el algoritmo de optimización de Adam, muestra que el siguiente grafico realiza la convergencia buscando los mínimos locales en una duración de 40 épocas.

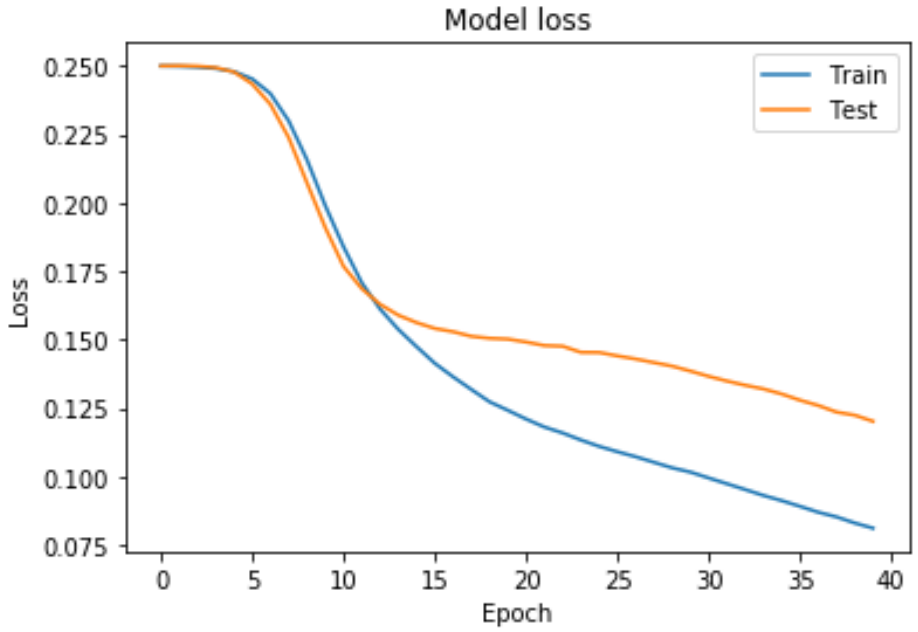


Figura 29: Exactitud – Época

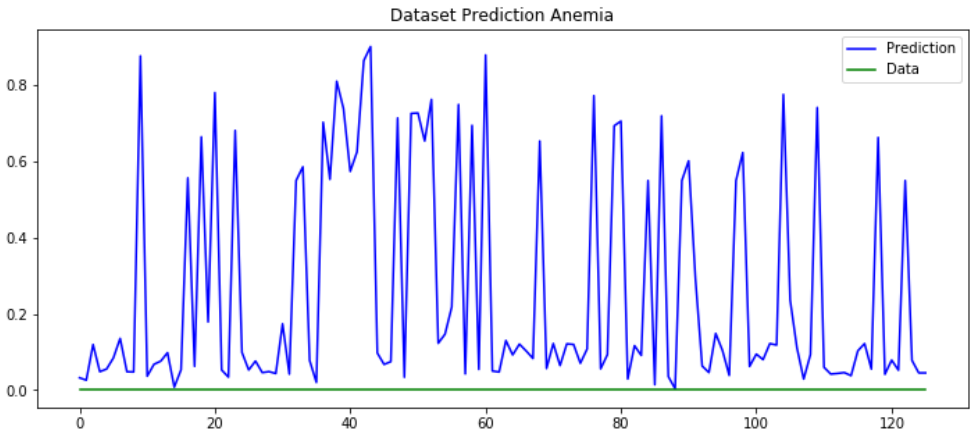
Fuente: Elaboración Propia

4.4.2.1 Explotación

El modelo neuronal, demostró resultados satisfactorios de los cuales se agregó data que no fue procesada por la red para determinar los resultados latentes de las madres gestantes con anemia.

Los resultados fueron los siguientes:

Figura 30: Validación de la Red Neuronal



Fuente: Elaboración Propia

Se hace el observa que la red neuronal frente a la nueva data tiene una precisión del 80%.

4.4.2.2 Validación cruzada.

La validación cruzada se aplica en diversos modelos de regresión y clasificación, se realiza una partición de la data por bloques con la finalidad que cada iteración de la red, la data entrenada y procesada varíe y los resultados siempre sean distintos, esta función determina que el modelo pase diversas pruebas de testeo y que tenga una precisión por encima del 70%.

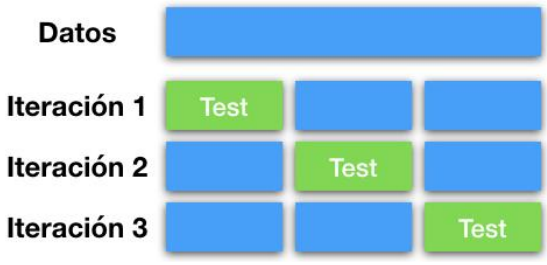


Figura 31: Modelo de Partición

Fuente: Elaboración Propia

4.4.2.3 Matriz de clasificación de MLP.

Una forma de validar modelos de precisión es por medio de la matriz de clasificación y estás se determinar por las siguientes fórmulas:

La precisión.

$$Precisión_i = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ki}}$$

Recall.

$$Recall_i = \frac{N_{ii}}{\sum_{k=1}^{n} N_{ik}}$$

F1 Score.

$$F - Score_i = \frac{2\ x\ Precision\ _i * \ Recall_i}{Precision_i + Recall_i}$$

El modelo obtuvo el siguiente reporte de métricas.

Reporte de Metricas:

	precision	recall	f1-score	support
0	0.67	0.56	0.61	39
1	0.59	0.69	0.63	35
accuracy			0.62	74
macro avg	0.63	0.62	0.62	74
weighted avg	0.63	0.62	0.62	74

Figura 32: Reporte de Métricas

Fuente: Elaboración Propia

4.4.3. Construcción del Modelo Naive Bayes.

Para hacer la predicción se necesita calcular la probabilidad que hay en una instancia de datos que pertenezca cada variable o clase. El proceso sigue los siguientes 4 pasos: Calcular la función de la densidad de probabilidad gaussiana, las probabilidades de las variables, predicción y precisión de la estimación.

$$P\ (A|B)\ = \frac{P(B|A)P(A)}{P(B)}$$

Donde:

P (A / B): la probabilidad de ocurrencia del evento A, dado el evento B, ya ha ocurrido.

P (A) - Probabilidad de ocurrencia del evento A.

P (B) - Probabilidad de ocurrencia del evento B.

P (B / A) - Probabilidad de ocurrencia del evento B, dado que el evento A ya ha ocurrido.

La implementación de la técnica se puede ver continuación:

```
def calculateProbability(x, mean, stdev):
    exponent = math.exp(-(math.pow(x-mean,2)/(2*math.pow(stdev,2))))
    return (1 / (math.sqrt(2*math.pi) * stdev)) * exponent

def calculateClassProbabilities(summaries, inputVector):
    probabilities = {}
    for classValue, classSummaries in summaries.items():
        probabilities[classValue] = 1
        for i in range(len(classSummaries)):
            mean, stdev = classSummaries[i]
            x = inputVector[i]
            probabilities[classValue] *= calculateProbability(x, mean, stdev)
    return probabilities
def predict(summaries, inputVector):
    probabilities = calculateClassProbabilities(summaries, inputVector)
    bestLabel, bestProb = None, -1
    for classValue, probability in probabilities.items():
        if bestLabel is None or probability > bestProb:
            bestProb = probability
            bestLabel = classValue
    return bestLabel

def getPredictions(summaries, testSet):
    predictions = []
    for i in range(len(testSet)):
        result = predict(summaries, testSet[i])
        predictions.append(result)
    plt.plot(predictions)
    #plt.plot(testSet)
    plt.show()
```

Figura 33: Código de programación de Naive Bayes

4.4.4. Evaluación del Modelo de Naive Bayes

La precisión de predicción de Naive Bayes alcanzo el 89%, siendo la más alta predicción de la presente investigación, de una total de 294 de registros balanceados con SMOTE.

```
def getAccuracy(testSet, predictions):
    correct = 0
    for i in range(len(testSet)):
        if testSet[i][-1] == predictions[i]:
            correct += 1
    return (correct/float(len(testSet))) * 100.0
def main():
    filename = 'balancesadancsv.csv'
    splitRatio = 0.67
    dataset = loadCsv(filename)
    trainingSet, testSet = splitDataset(dataset, splitRatio)
    print(('Split {0} rows into train={1} and test={2} rows').format(len(dataset),
    # prepare model
    summaries = summarizeByClass(trainingSet)
    # test model
    predictions = getPredictions(summaries, testSet)
    accuracy = getAccuracy(testSet, predictions)
    print(('Accuracy: {0}%').format(accuracy))
main()
```

```
Split 294 rows into train=196 and test=98 rows
Accuracy: 89.79591836734694%
```

Figura 34: Cálculo de precisión de Naive Bayes

En al siguiente Figura se refleja la predicción de variables con Naive B

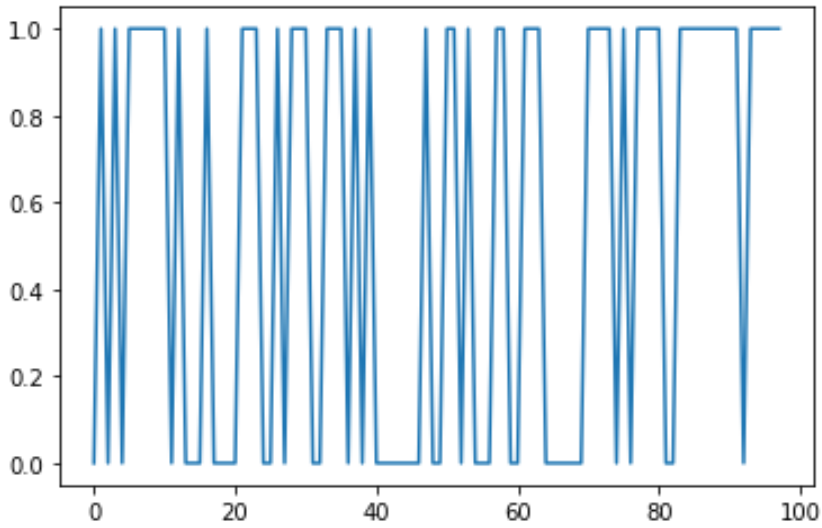


Figura 35. Representación gráfica del resultado de predicción con Naive Bayes

4.4.5. Construcción del Modelo Árbol de decisión.

Extracción de las variables más redundantes aplicando “Feature Importance”, que es el grado de importancia de las variables para construir un modelo

probabilístico, el cual es una clasificación artificial donde el número de variables (n_features) son la cantidad de los valores de entradas al árbol y el gráfico arroja como resultado la importancia de las características informativas de (n_informative).

```
import numpy as np
import matplotlib.pyplot as plt

from sklearn.datasets import make_classification
from sklearn.ensemble import ExtraTreesClassifier

# Build a classification task using 3 informative features
X, y = make_classification(n_samples=10000,
                           n_features=14,
                           n_informative=14,
                           n_redundant=0,
                           n_repeated=0,
                           n_classes=2,
                           random_state=0,
                           shuffle=False)

# Build a forest and compute the feature importances
forest = ExtraTreesClassifier(n_estimators=250,
                              random_state=0)

forest.fit(X, y)
importances = forest.feature_importances_
std = np.std([tree.feature_importances_ for tree in forest.estimators_],
             axis=0)
indices = np.argsort(importances)[::-1]
```

Figura 36: Código de programación de Árbol de decisión

En la figura N° 38 se puede apreciar una lista de las variables de entrada, la importancia de los cuales contrasta con la figura N° 39 y estos tienen la siguiente correspondencia:

'EDAD'.	feature 0
LUGAR_NACIMIENTO'	feature 1
EDUCACION'	feature 2
OCUPACION'	feature 3
ESTADO_CIVIL'	feature 4
TIPO_SEGURO'	feature 5
NUMERO_GESTACIONES'	feature 6
HIJOS_VIVOS'	feature 7
MENARQUIA'	feature 8
DURACION_MENSTRUACION'	feature 9
IMC'	feature 10
'NUMERO_ABORTOS',	feature 11
EDAD_GESTACIONAL_TRIMESTRAL'	feature 12
'TIPO_SOCIOECONOMICO'	feature 13

```
X = data.iloc[:,0:14].values
y = data.iloc[:,-1].values
columns = ['EDAD','LUGAR_NACIMIENTO','EDUCACION','OCUPACION','ESTADO_CIVIL','TIPO_SEGURO','NUMERO_GESTACIONES',
           'HIJOS_VIVOS','MENARQUIA','DURACION_MENSTRUACION','IMC','NUMERO_ABORTOS','EDAD_GESTACIONAL_TRIMESTRAL',
           'TIPO_SOCIOECONOMICO','Class']
X_train , X_test, y_train, y_test = train_test_split(X, y, test_size = 0.25, random_state = 0)
#4 8 12 7
```

Figura 37: Nombres de las variables de entrada

Como resultado se obtiene, el ranking de importancia de las variables, las cuales deberían ser prioridad analizarlas para dar soporte a la investigación cuyo objetivo es determinar las causas de anemia y el ranking, refleja el valor de cada variable.

```
Feature ranking:
1. feature 4 (0.090244)
2. feature 8 (0.087227)
3. feature 12 (0.082800)
4. feature 7 (0.077870)
5. feature 10 (0.076812)
6. feature 11 (0.073403)
7. feature 9 (0.072768)
8. feature 2 (0.071752)
9. feature 13 (0.070856)
10. feature 6 (0.063622)
11. feature 0 (0.063081)
12. feature 1 (0.059472)
13. feature 5 (0.056419)
14. feature 3 (0.053673)
```

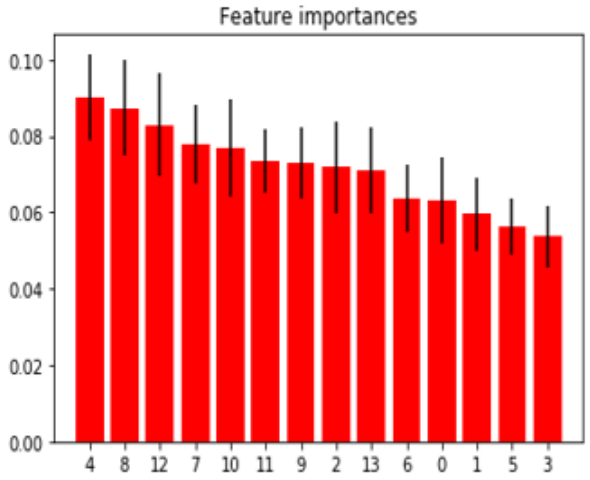


Figura 38: Representación gráfica de las variables más influyentes

Conocer los parámetros de un árbol de decisión es fundamental debido a que explica el resultado que estamos obteniendo, la optimización del árbol es Gini, que hace se selección de las variables diferentes por medio de la entropía.

$$Gini(D) = 1 - \sum_{i=1}^{m} Pi^2$$

Para determinar el valor probabilístico de cada nodo, se aplica la siguiente fórmula.

$$Gini_A(D) = \frac{|D_1|}{|D|} Gini(D)_1 + \frac{|D_2|}{|D|} Gini(D)_2 + \frac{|D_n|}{|D|} Gini(D)_n$$

La ganancia de información es la disminución de la entropía, esto quiere decir que a mayor entropía el árbol de decisión no será eficaz, para determinar el grado de importancia de las variables.

$$Info(D) = -\sum_{i=1}^{m} Pi\ log_2\ Pi$$

Conociendo el grado de importancia de las variables, se procede a construir el modelo de árbol de decisión.

A continuación, se puede apreciar la importancia de cada una de las variables de la data de experimentación en la presente investigación, también se puede expresar como variables más influyentes según el cálculo con la técnica de árbol de decisión J48, que a la vez alcanzó un 79% de precisión, estos resultados son representación de influencia de manera general, sin especificar las clases que podría tener un problema, Sin embargo se puede apreciar en la Ilustración N° 28 los aspectos más detallados por clases.

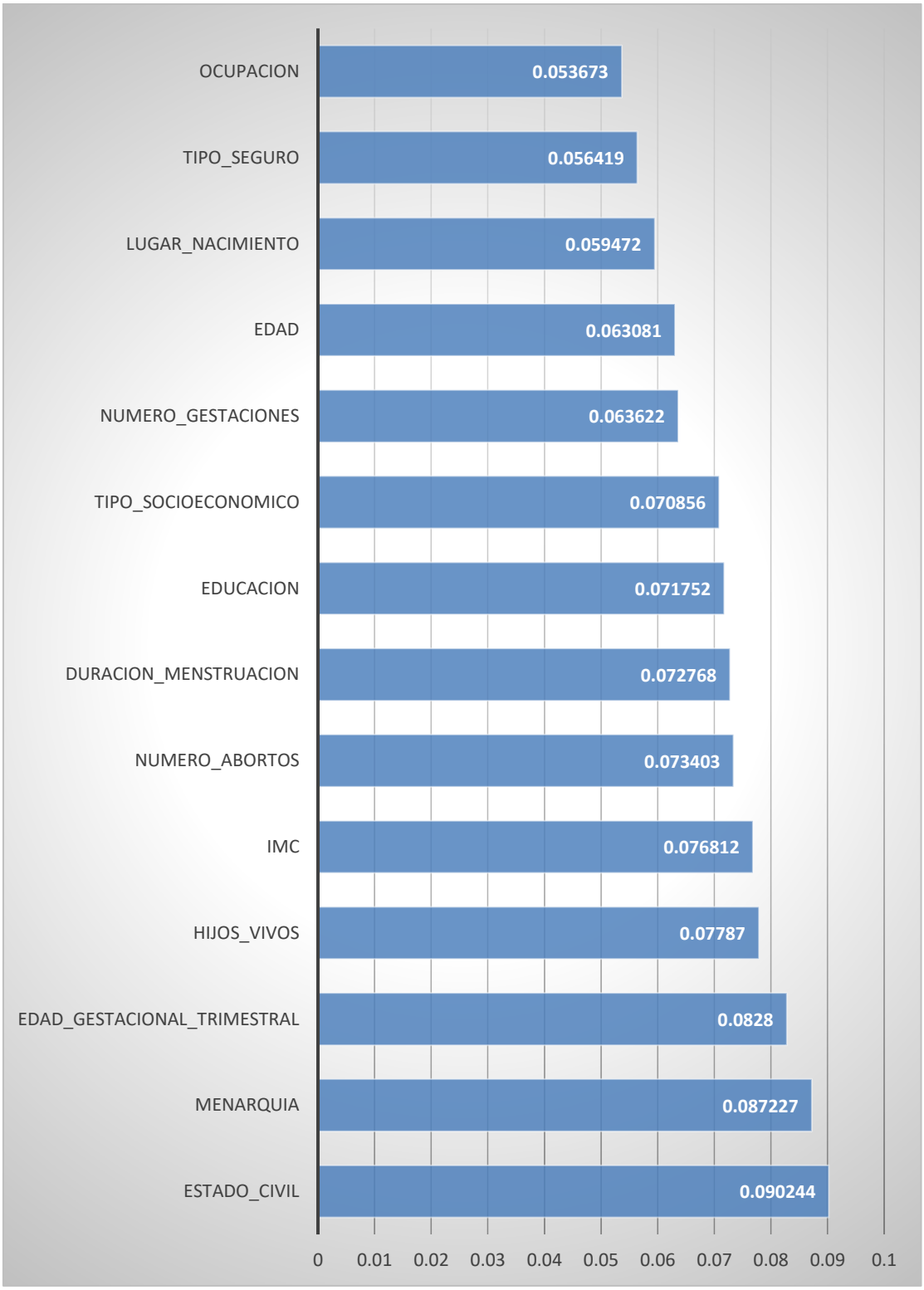


Figura 39: Variables más influyentes en la predicción de anemia en madres gestantes de Ilo

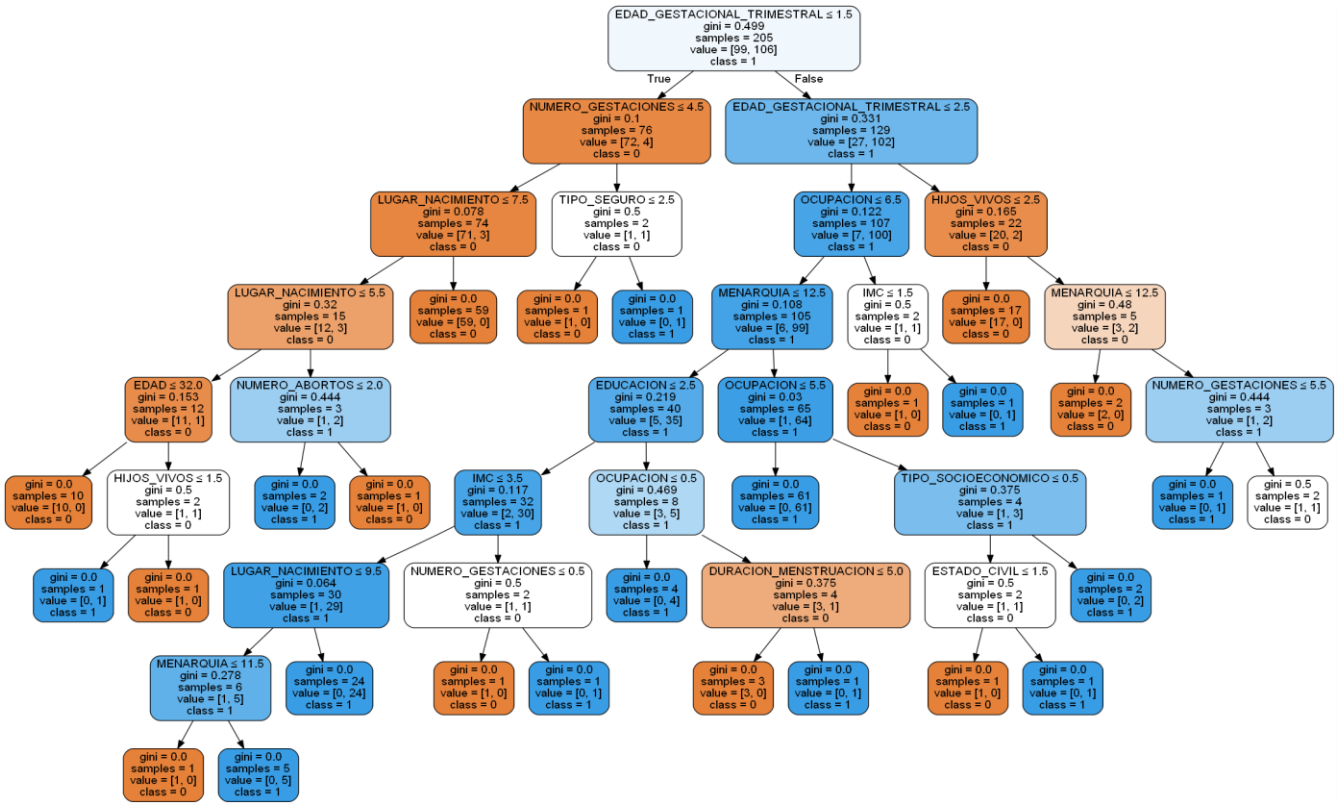


Figura 40: Árbol de decisión

4.4.6. Evaluación del Modelo de árbol de decisión.

Del modelo estimado, reflejan los siguientes resultados:

```
Matriz de Confusión:
 [[36 12]
 [ 7 34]]
```

```
Reporte de Metricas:
             precision    recall  f1-score   support

          0       0.84      0.75      0.79        48
          1       0.74      0.83      0.78        41

avg / total       0.79      0.79      0.79        89
```

Las imágenes mostradas reflejan el grado de precisión del árbol de decisión, para dar realce a la investigación se muestrea la siguiente predicción de una data que no ingreso al modelo.

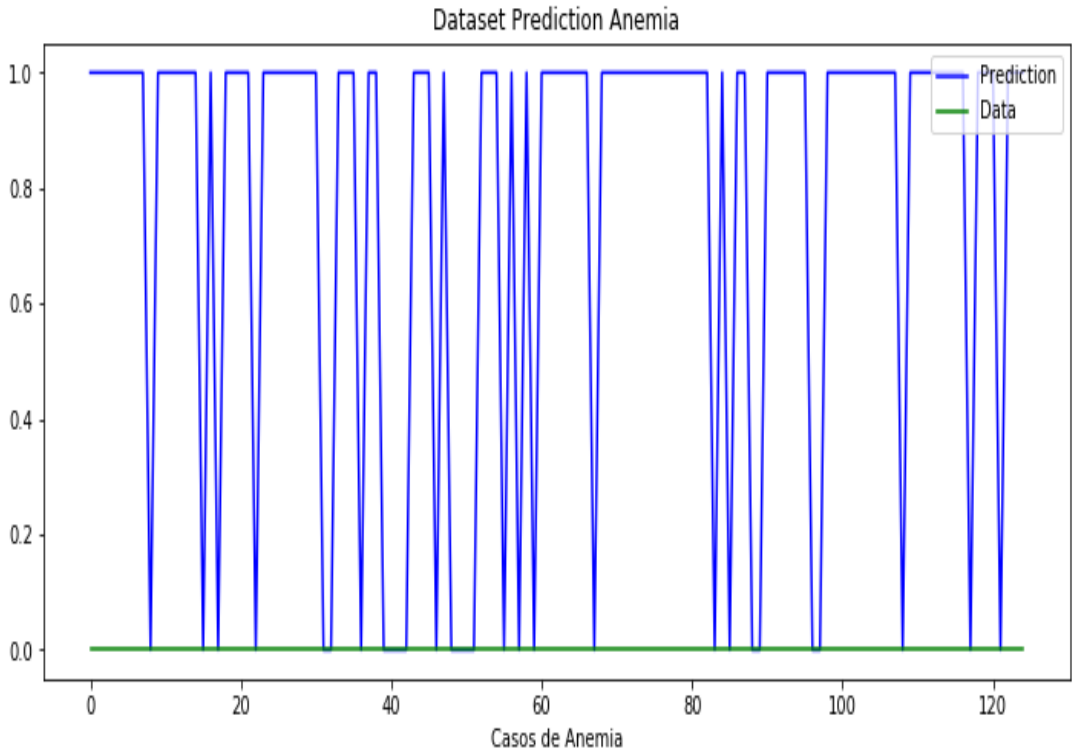


Figura 41: Representación de grafica de resultados de Árbol de decisión

4.5. EVALUACION

4.5.1. Evaluación de precisión de las técnicas.

En el presente proyecto de minería de datos se experimentó tres principales algoritmos, Redes neuronales perceptrón multicapa, Naive Bayes y árbol de decisión, entre los cuales la técnica Naive Bayes alcanzó a la mejor precisión con el 89%, seguido por árbol de decisión con 79% de precisión, como se puede aprecias en la tabla N 07.

Tabla 7: Resultados de las técnicas de minería de datos

Ítem	Técnica	precisión
1	Perceptrón Multicapa	62%
2	Naive Bayes	89%
3	Árbol de decisión	79%

4.6. IMPLEMENTACION

EL modelo queda como propuesta para Red de Salud de la provincia de Ilo, puesto que se ha utilizado los datos de la dicha institución, exactamente del departamento de ginecología y obstetricia, la implementación y aplicación queda a decisión de las autoridades pertinentes.

V. CONCLUSIONES

Se logro Implementar un modelo de minería de datos para la predicción de casos de anemia en gestantes de la provincia de Ilo, la cual permitió clasificar entre madres gestantes con anemia y sin anemia, así como determinar las características más relevantes que requiere reforzar las estrategias de control y prevención de anemia en el periodo gestacional.

Se aplico la metodología de CRISP-DM para:

- Clasificar madres gestantes con anemia y sin anemia aplicando los tres algoritmos de predicción; Perceptrón multicapa, Árbol de decisión, Naive Bayes.
- Se determinó el ranking de las variables de las madres gestantes con anemia y sin anemia con la técnica de árbol de decisión J48, siendo las tres variables más influyentes, el estado civil, menarquia y la edad gestacional, con dicha técnica se alcanzó un 79% de precisión.
- Después de aplicar la validación cruzada en el modelo se obtuvo el siguiente reporte de métricas: Perceptrón multicapa (62%), Árbol de decisión (79%), Naive Bayes (89%), siendo este último el de mejor precisión.

VI. TRABAJOS FUTUROS

- Realizar periódicamente el levantamiento de información de casos de anemia en gestantes con el objetivo de alimentar la base de datos y pueda realizar mejores pruebas en la red neuronal, de la misma forma modificar las derivadas de la función para su proceso.

- Unir los casos de anemia de niños y adultos a la base de datos con la finalidad de tener una capa de salida multivariable, generará mejor aporte a la investigación, también existirá el sobre muestreo en la clase a predecir, la complejidad de la red neuronal se modificará y la predicción será multiclase.

VII. RECOMENDACIONES

- Prestar atención a la aplicación de modelos de minería de datos en las instituciones del sector público, esto ayudaría a prevenir enfermedades y su tratamiento a tiempo.
- Aprovechar la información acumulada de los pacientes para la predicción de algunas enfermedades haciendo el uso de algoritmos clasificadores y predictores.
- Usar el Naive Bayes para predicción sobre los datos cuantitativos y categóricos de casos similares a la presente investigación.
- Utilizar como conocimiento el presente trabajo para futuras investigaciones relacionadas con esta materia.

VIII. REFERENCIAS

Brown, M. (s.f.). *Dummies*. Obtenido de Dummies: https://www.dummies.com/programming/big-data/phase-6-of-the-crisp-dm-process-model-deployment/

Palmer, A., Jiménez, R., & Montaño, J. J. (2001). Tutorial sobre Redes Neuronales Artificiales: El Perceptrón Multicapa. *Facultad de Psicología. Universitat de les Illes Balears.* .

Virseda Benito, F., & Román Carrillo, J. (s.f.). Minería de datos y aplicaciones .

Abdullah, M., & Al-Asmari, S. (2017). Anemia types prediction based on data mining classification algorithms. *Communication, Management and Information Technology – Sampaio de Alencar*.

Basogain Olabe, X. (s.f.). REDES NEURONALES ARTIFICIALES Y SUS APLICACIONES. *Escuela Superior de Ingeniería de Bilbao, EHU*.

Collins, M. (s.f.). *The Naive Bayes Model, Maximum-Likelihood Estimation, and the EM Algorithm.*

Espino Timón, C. (2017). *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso.*

Friedman, J., Tibshirani, R., & Hastie, T. (2017). *The Elements of Statistical Learning Data Mining, Inference, and Prediction.*

Gallego, J. A., Navarro, L. F., & Castillo, L. F. (2015). APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN ATENCIÓN PRIMARIA EN SALUD (APS) PARA EL

Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques.* Waltham, MA 02451, USA: Morgan Kaufmann ELSEVIER.

Jorge Matich, D. (2001). *Redes Neuronales: Conceptos Básicos y Aplicaciones.* Universidad Tecnológica Nacional – Facultad Regional Rosario Departamento de Ingeniería QuímicaGrupo de Investigación Aplicada a la Ingeniería Química (GIAIQ).

Larranaga, P., Inza, I., & Moujahid, A. (s.f.). Redes Neuronales. *Departamento de Ciencias de la Computacion e Inteligencia Artificial Universidad del Paıs Vasco–Euskal Herriko Unibertsitatea*.

Leskovec, J., Rajaraman, A., & Ullman, J. (2014). *Mining of Massive Datasets.* Cambridge University Press.

Organiazación Mundial de la Salud. (2008). *Sistema de Información Nutricional sobre Vitaminas y Minerales (VMNIS)*. Obtenido de https://www.who.int/vmnis/database/anaemia/anaemia_data_status_t2/es/

Parneet, K., Manpreet, S., & Gurpreet Singh, J. (2015). Classification and prediction based data mining algorithms to predict slow learners in education sector. *ScienceDirect*.

Piatetsky, G. (s.f.). *KDnuggets*. Obtenido de KDnuggets: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Ramirez, J. S. (2018). Factores asociados a anemia en gestantes hospitalizadas en el servicio de ginecoobstetricia del Hospital “San José” Callao - Lima. *Repositorio de la Universidad Ricardo Palma*.

Retamar, S., De Babtista, A., Ramos, L., Nuñes, J., Savoy, F., & De Garcia, L. (2018). Minería de datos para la detección de factores de influencia en el test de Apgar.

Toscano de la Torre, B., Ponce, J., Margain, L., & Lopez-Espinoza, R. (2016). Aplicación de Minería de Datos para la Identificación de Factores de Riesgo Asociados a la Muerte Fetal. *Conference: VIII Congreso Internaconal en Ciencias Computacionales - CICOMP 2016*.

Unidad de Informacion y Analisis Financiero. (2014). *TÉCNICAS DE MINERÍA DE DATOS PARA LA DETECCIÓN Y PREVENCIÓN DEL LAVADO DE ACTIVOS Y LA FINANCIACIÓN DEL TERRORISMO (LA/FT).*

US Department of Health and Human Services. (2011). *Guia breve sobre anemia.*

Vizcaino Garzon, P. A. (2008). *Aplicacion de tecnicas de induccion de arboles de decision a problemas de clasificacion mendiante uso de Weka.* Bogota: Fundacion Universitaria Konrad Lorenz, Facultad de Ingenieria de Sistemas .

Wirth, R., & Hipp, J. (s.f.). CRISP-DM: Towards a Standard Process Model for Data Mining.

ANEXOS

"AÑO DE LA LUCHA CONTRA LA CORRUPCIÓN Y LA IMPUNIDAD"

SOLICITO: AUTORIZACION DE EJECUCION DE PROYECTO DE TESIS

ILO 05 Julio del 2019

Señor (a) : M.C. PERCY HUANCAPAZA CHAMBI

Director Ejecutivo de la Red de Salud Ilo

YO, Sadan Eusebio Condori Bellido Identificada con DNI Nº 46925817 con domicilio en Calle Loreto H - 4 Ante Ud. con el debido respeto me presento y digo:

Es grato dirigirme a Ud. A fin de saludarlo cordialmente y solicito información para mi tesis que se encuentra desarrollando el proyecto de investigación "MODELO DE MINERÍA DE DATOS PARA LA PREDICCIÓN DE CASOS DE ANEMIA EN GESTANTES DE LA PROVINCIA DE ILO" con la finalidad de obtener mi tesis de Ingeniería de Sistemas e Informática de la Universidad Nacional de Moquegua.

Solicita nos brinde su AUTORIZACION a cerca de la aceptación y facilidad de conceder realizar la ejecución del proyecto de tesis en mención.

POR LO EXPUESTO:

Ruego a Ud. M.C. PERCY HUANCAPAZA CHAMBI

SADAN EUSEBIO CONDORI BELLIDO

DNI: 46925817

GOBIERNO REGIONAL MOQUEGUA
Gerencia Regional de Salud Moquegua
Red de Salud Ilo

"AÑO DE LUCHA CONTRA LA CORRUPCION E IMPUNIDAD"

Ilo, 21 de octubre del 2019

CARTA Nº 001 -2019-GRM/GRSM/RED SALUD ILO/OSIC

Señor.
BACH. SADAN EUSEBIO CONDORI BELLIDO
PRESENTE.

ASUNTO: ACEPTACION DE EJECUCION DE PROYECTO DE TESIS

Es grato dirigirme a Ud., para saludarlo cordialmente y a la vez comunicarle que hemos recibido su solicitud s/n, donde nos solicita autorización para desarrollar su proyecto de investigación de tesis denominado "MODELO DE MINERIA DE DATOS PARA LA PREDICCION DE CASOS DE ANEMIA EN GESTANTES DE LA PROVINCIA DE ILO", AÑO 2019, motivo por el cual nuestra institución se complace en aceptar y brindarle todas las facilidades que el caso amerite a fin de que pueda utilizar la información requerida y de esta manera poder sustentar su tesis para que pueda optar el título de Ingeniero en Sistemas e Informática .

Sin otro particular. Es propicia la ocasión para hacerle llegar mis deferencias personales.

Atentamente,

REGION MOQUEGUA RED DE SALUD ILO DIRECCION DIRES MOQUEGUA

GOBIERNO REGIONAL MOQUEGUA
GERENCIA REGIONAL DE SALUD MOQUEGUA
RED SALUD ILO
M.C. [illegible]
DIRECTOR EJECUTIVO RED DE SALUD ILO

PHCH/DESI
EAMT/OSIC
DCHS/UPS
c.c. Archivo

III. MATRIZ DE CONSISTENCIA:

PROBLEMA GENERAL:	OBJETIVO GENERAL:	HIPÓTESIS GENERAL:
¿Es posible implementar un modelo de predicción del diagnóstico de la anemia en pacientes gestantes utilizando técnicas de Minería de Datos aplicando la metodología CRISP-DM según registros del MINSA de la Provincia de ILO?	Desarrollar el modelo de predicción del diagnóstico de la anemia en pacientes gestantes utilizando técnicas de Minería de Datos aplicando la metodología CRISP-DM según registros del MINSA de la Provincia de ILO.	Se podrá realizar un modelo de predicción del diagnóstico de la anemia en pacientes gestantes utilizando técnicas de Minería de Datos aplicando la metodología CRISP-DM según registros del MINSA de la Provincia de ILO.
PROBLEMAS ESPECÍFICOS:	**OBJETIVOS ESPECÍFICOS:**	**HIPÓTESIS ESPECÍFICAS:**
¿Cuáles son los algoritmos adecuados de Minería de Datos para la predicción del diagnóstico de la anemia en pacientes gestantes según registros del MINSA de la Provincia de ILO? ¿Qué técnicas de validación se emplearán para validar el modelo de predicción del diagnóstico de la anemia en pacientes gestantes según registros del MINSA de la Provincia de ILO?	Encontrar los algoritmos adecuados de Minería de Datos para la predicción del diagnóstico de la anemia en pacientes gestantes según registros del MINSA de la Provincia de ILO. Validar con técnicas de Minería de Datos el modelo de predicción del diagnóstico de la anemia en pacientes gestantes según registros del MINSA de la Provincia de ILO.	Es posible aplicar algoritmos adecuados de Minería de Datos para la predicción del diagnóstico de la anemia en pacientes gestantes según registros del MINSA de la Provincia de ILO. Es posible validar con técnicas de Minería de Datos el modelo de predicción del diagnóstico de la anemia en pacientes gestantes según registros del MINSA de la Provincia de ILO