Algoritmos de Aprendizaje Automático para Predecir Anemia en Niños y Jovenes en México

El uso de modelos de aprendizaje automático permite a través de factores sociodemográficos y clínicos obtener una clasificación más exacta de niños y jóvenes con factores asociados a la anemia, donde estos factores de riesgo son utilizados como características de entrenamiento para el algoritmo. **Objetivo:** Desarrollar un modelo de aprendizaje automático el cual permita clasificar a los pacientes de acuerdo a su condición. **Materiales y Métodos:** Los datos fueron extraídos de la base de datos de la ENSANAUT(Encuesta Nacional de Salud y Nutrición) la cual contiene una encuesta realizada en 2012. En este estudio se muestra un total de 3921 de personas que padecen de anemia. De este estudio se creo una muestra utilizando métodos estadísticos, permitiendo tener una base de datos balanceada. La metodología propuesta comprende cuatro etapas, la primera es el preprocesamiento la cual se enfoca en limpiar los datos y crear un conjunto de datos balanceado. Posteriormente, se pasa a la etapa de extracción de características, la cual obtiene las variables más representativas para describir el fenómeno. Posteriormente se hace uso de algoritmos de aprendizaje automático con el fin de entrenar un modelo que permita describir el problema. Finalmente, se realizó una evaluación de los algoritmos en términos de precisión y AUC. **Resultados:** Se encontró que el algoritmo SVM, logró una AUC de 0.95 con una precisión promedio de 0.98, mientras que RF obtuvo una AUC de 0.92 con un precisión promedio de 0.92. Y la RL obtuvo una AUC de 0.93 con una precisión promedio de 0.98. **Conclusiones:** Concluimos que los métodos de ML se pueden considerar como una herramienta la cual permite la predicción de anemia en niños y jóvenes. Estas herramientas son sencillas de utilizar y permitirían a partir de un sistema de información facilitar las labores en telemedicina y asistencia remota.

Palabras clave: *anemia, niños, jovenes, ENSANAUT, aprendizaje automático, toma de decisiones*

1. Introducción

1.1. Trabajos Relacionados

2. Materiales y Métodos

Para este trabajo, se emplea la siguiente metodología resumida en los siguientes pasos. Donde, cada fase puede ser visualizada en la figura 1.

Figura 01: Metodología propuesta para la resolución del problema presentado.

2.1. Descripción de los Datos

El conjunto de datos fue recolectado del repositorio de la Encuesta Nacional de Salud y Nutrición (2012). El cual contiene tres bases de datos, la primera consiste en la población menor a 5 años, la segunda muestra contiene los datos recolectados de niños entre 5 y 11 años de edad. Finalmente, el último conjunto de datos contiene la muestra recolectada de la población adolescente de niños entre 12 y 19 años de edad.

Rango de Edades	Anémicos	Conteo
n > 5	1	1729
	0	5841
5 < n <= 11	1	1462
	0	12404
12 < n <= 19	1	730
	0	10908
Total	1	3921
	0	29153
	---	33074

Tabla 01: Casos encontrados con presencia de anemia de acuerdo a la distribución de edades presentadas por el conjunto de datos.

2.1.1. Descripción de las Características

Este conjunto de datos está conformado por 33 características las cuales contienen información sociodemográfica dividida de la siguiente manera.

2.2. Preprocesamiento de Datos

En esta fase, tras haber realizado la recolección de los datos se procede a la preparación para adaptarlos a técnicas de aprendizaje automatizado y búsqueda de relaciones entre los factores que componen los datos. De esta manera se podrá realizar una mejor aplicación de

algoritmos y técnicas de visualización. Para esto, se realizan cuatro actividades principales las cuales consisten en:

1. **Estructuración de los Datos:** Esta actividad consiste en dos partes. La primera es eliminar las características no necesarias tales como características repetidas o de poca relevancia tales como identificadores. La segunda parte consiste en generar nuevos campos a partir de otros existentes o bien fusionar conjuntos donde se tienen valores en común a través de los registros.
2. **Selección de Datos:** En esta parte la tarea principal es seleccionar un subconjunto de los datos basados en criterios tales como la completitud de los datos, corrección de los datos, limitación en el volumen de datos y establecimiento de los tipos de datos a manejar.
3. **Limpieza de Datos:** Para esta actividad se espera limpiar los datos ya que se pueden encontrar discrepancias con la calidad de los datos, normalización de los datos, discretización de campos numéricos y manejo de valores ausentes, así como complementar el manejo del volumen de datos al reducir el volumen.
4. **Integración de los Datos:** Finalmente la integración de los datos consiste en realizar transformaciones sintácticas de los datos sin modificar su significado con el fin de manejar de forma más ágil los datos obtenidos.

2.3. Descripción de Características*

2.4. Análisis de Clasificación

Con el apoyo de la estadística tradicional, el aprendizaje automático permite obtener y procesar más datos de una forma en la cual se pasan por una fase de entrenamiento. Esta permite a través de la implementación de algoritmos reconocimiento de patrones estadísticos y el establecimiento de relaciones nuevas entre posibles patrones sin la necesidad de la creación de hipótesis. Usando el aprendizaje automático como base para minar datos o mejor dicho la actividad de obtener nuevos patrones a través de un conjunto de datos, esto permite mejorar y medir a través de diversas métricas de empleadas para su medición y sustento.

2.4.1. Regresión Logística