Revista de investigación de sistemas e informática
14(2): 89-98 (2021)
DOI: https://doi.org/10.15381/risi.v14i2.23150

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Aplicación Web Basado en Minería de Datos usando la Técnica de Naive Bayes para la Predicción de la obesidad en edad infantil en los Hospitales Públicos de Lima

Web Application Based on Data Mining using the Naive Bayes Technique for the Prediction of childhood obesity in Public Hospitals of Lima

Nicole Emily Becerra Romero
nicole.becerra@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos.
Lima, Perú

Ana María Huayna Dueñas
https://orcid.org/0000-0001-7726-8206
ahuaynad@unmsm.edu.pe
Universidad Nacional Mayor de San Marcos.
Lima, Perú

Recibido: 05/01/2022 - Aceptado: 15/02/2022 - Publicado: 28/02/2022

RESUMEN

La obesidad infantil es una enfermedad que actualmente causa mucha preocupación a nivel mundial y provoca distintas comorbilidades en los niños como lo son la diabetes y los trastornos respiratorios, además de ser un factor de riesgo para el COVID-19. Por estas razones es que existen varias investigaciones que buscan predecir esta enfermedad utilizando distintas técnicas de Minería de Datos como Árboles de Decisión, Regresión Logística, Sistemas Neuro – Difusos entre otros. El presente trabajo de investigación realiza la predicción de que un menor de 5 años padezca de obesidad en un futuro usando la técnica Naive Bayes; el conjunto de datos para implementar el modelo contó con 770 registros y 27 variables extraídas del aplicativo e-Qhali. Las pruebas fueron efectuadas sobre 317 registros obteniendo un modelo con 72% de precisión y 93% de sensibilidad, y al comparar la técnica Naive Bayes con otras técnicas de clasificación como Regresión Logística, Random Forest y SVM, esta alcanzó el mayor porcentaje de sensibilidad.

Palabras clave: Modelos Predictivos; Naive Bayes; Obesidad Infantil; Inteligencia Artificial; KDD.

ABSTRACT

Childhood obesity is a disease that currently causes much concern worldwide and causes different comorbidities in children such as diabetes and respiratory disorders, in addition to being a risk factor for COVID19. For these reasons, there are several investigations that seek to predict this disease using different Data Mining techniques such as Decision Trees, Logistic Regression, Neuro-Fuzzy Systems, among others. The present research work makes the prediction that a child under 5 years old will suffer from obesity in the future using the Naive Bayes technique; The data set to implement the model had 770 records and 27 variables extracted from the e-Qhali application. The tests were performed on 317 records obtaining a model with 72% precision and 93% sensitivity, and when comparing the Naive Bayes technique with other classification techniques such as Logistic Regression, Random Forest and SVM, it reached the highest percentage of sensitivity.

Keywords: Predictive Models; Naive Bayes; Childhood Obesity; Artificial Intelligence; KDD.

I. INTRODUCCIÓN

Desde los años ochenta, los niños han tenido la inclinación por consumir alimentos distintos a los acostumbrados en la familia; desean consumir comida “chatarra” en lugar de la alimentación tradicional que recibían antes; como consecuencia surgieron los distintos desequilibrios nutricionales como son el sobrepeso y la obesidad, considerados “los problemas más graves en el siglo XXI” así los definió la OMS (2016).La obesidad es responsable de 4.7 millones de muertes prematuras cada año a nivel mundial, siendo 4 veces más el número de muertos en accidentes de tránsito y 5 veces más el número de fallecidos por VIH.

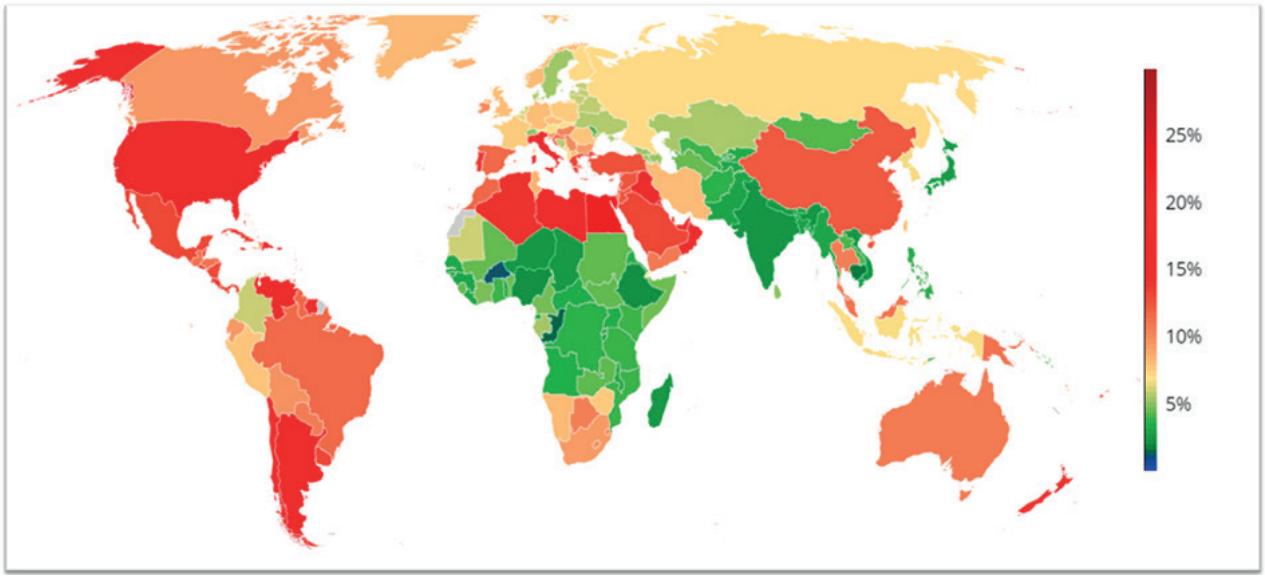
Actualmente, la obesidad es considerada uno de los factores de riesgo de mortalidad más resaltante en los pacientes con COVID-19, ya que esta altera el sistema inmunitario, disminuye la función pulmonar y dificulta expandir los pulmones. Según el Ministerio de Salud (MINSA-PERU,2020) , el 85.5% de pacientes fallecidos con diabetes e hipertensión por COVID-19 hasta agosto del 2020 padecían obesidad; estas cifras no solo se presentaron en el Perú; en New York personas con COVID-19 cuyo IMC era mayor a 35 kg/m2 tuvieron 3.6 veces mayor posibilidad de ser admitidos en UCI que pacientes con un IMC menor a 30 kg/m2 y en Francia se observó que el riesgo de ingresar a UCI y requerir ventilación mecánica fue 7 veces mayor en aquellos pacientes con COVID-19 con un IMC alto que en los pacientes con un IMC dentro del rango saludable (Guija y Guija, 2020).

La OMS (2016) indicó que, en todo el mundo, el número de lactantes y niños pequeños (de 0 a 5 años) que padecieron sobrepeso u obesidad aumentó de 32 millones en 1990 a 42 millones en 2013. Solamente en la Región de África, el número de niños con sobrepeso u obesidad aumentó de 4 a 9 millones en el mismo período. La prevalencia de esta patología cambia de un país a otro, entre las principales causas de obesidad identificadas por la entidad de salud en el continente africano resaltan la falta de acceso a alimentos saludables y ausencia de nutricionistas que puedan indicar las pautas necesarias para una correcta alimentación infantil.

En la Figura 1 podemos visualizar el estado de la obesidad infantil en niñas de 5 a 17 años en las distintas regiones del mundo en el año 2016, observando así que países que comprenden el continente americano, Oceanía y algunos países de África se encuentra con los valores más altos a nivel mundial, es por ello la gran importancia de enfocar estrategias inmediatas para poder solucionarlo, y mejor aún, prevenirlo antes de tiempo. Senthilingam (2017) explica que los altos porcentajes de prevalencia de obesidad observados en el mapa se deben principalmente a la “desigualdad de actividad” que se calcula obteniendo la diferencia entre los que más y menos caminan en una población, mientras mayor sea la diferencia mayor será su tasa de obesidad, la aplicación móvil “Azumio Argus” permitió medir los pasos diarios de las personas y el top 5 de países por “desigualdad de actividad” fue Arabia Saudita, Australia, Canadá, Egipto y Estados Unidos.

Figura 1

Mapa de prevalencia del sobrepeso en niñas y adolescentes a nivel mundial [4]



En América, el país con mayor porcentaje de obesidad infantil es Argentina con una 20% de prevalencia en niños y 13% en niñas; mientras que Perú se encuentra en el decimos lugar. Orgaz (2019) indicó que los altos porcentajes de prevalencia de sobrepeso y obesidad en nuestro continente se dan por el desarrollo económico en las áreas rurales lo cual conlleva un menor gasto energético en la población. Son principalmente los padres los responsables de esta situación al no ser conscientes del exceso de peso en sus hijos y no acudir a un especialista en busca de orientación.

La situación del Perú no es distinta a la de otros países latinoamericanos, en los últimos 36 años la prevalencia de obesidad se ha cuadruplicado (2% a 8%) y el sobrepeso ha variado de 11% a 27%. Según el Centro de Estudios de Alimentación y Nutrición (CENAN, 2019) solo en 4 años, del 2015 al 2019, el porcentaje de obesidad en menores de 5 años varió de 1.5% a 1.9%, y en los departamentos como Lambayeque, La libertad y Ayacucho tuvieron una gran variación en la suma de sus porcentajes de sobrepeso y obesidad como se puede observar en la Figura 2.

El sobrepeso y la obesidad en el Perú son predominantes en Lima Metropolitana y la costa peruana, lo que se podría explicar en la mayor urbanización y desarrollo económico que conllevan cambios en los estilos de vida y provocan modificaciones en los patrones de alimentación y actividad física.

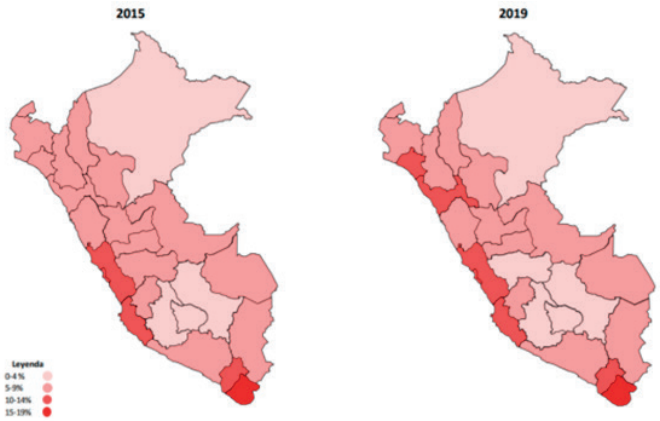
Debido al gran problema mundial que significa la obesidad infantil actualmente, son muchas las iniciativas que buscan prevenir que un niño con obesidad se convierta en un adulto con mayor riesgo de sufrir de diabetes o hipertensión. En las investigaciones actuales sobre la predicción de obesidad infantil, Alva et al. (2020) desarrollaron un modelo de inferencia difusa compuesta por 81 reglas que permitió detectar tempranamente el sobrepeso y obesidad en niños y adolescentes, este modelo obtuvo una precisión promedio de 95.5%. En otro trabajo realizado por Hammond et al. (2019) se utilizó distintos modelos de regresión lineal y se obtuvo un AUC promedio de 78.9% basándose en la historia clínica del niño y la madre. Por último, Ticona (2018) implementó un software utilizando la técnica J48 y obtuvo un 94.39% de precisión, esta técnica fue superior a técnicas como BayestNet, Multilayer Perceptron, y ForestPA en el proceso de validación.

En el presente trabajo se desarrolla una aplicación web usando la técnica de Naive Bayes para proporcionar una mejor precisión en la literatura, esta técnica en comparación con otras es más simple, computacionalmente eficiente, requiere pocos datos para el entrenamiento, no tiene muchos parámetros y es naturalmente robusta a los datos faltantes y al ruido, por lo tanto, se considera como uno de los clasificadores más útiles para respaldar las decisiones de los médicos (Al-Aidaroos, Bakar, & Othman, 2018).

Nuestra propuesta fue puesta a prueba realizando, utilizando un total de 129 casos reales. Los resultados le otorgan al trabajo desarrollado una tasa de éxito del 96% así como un valor de 0.9864 en términos del indicador AUC, además, sugieren que las Redes Bayesianas alcanzan un alto rendimiento, así como también ofrecen transparencia durante el proceso de inferencia, algo que no sucede con muchas otras técnicas, y que son ideales para afrontar problemas de predicción.

Figura 2

Porcentaje de peso (sobrepeso + obesidad) en niños menores de 5 años. Del 2015 al 2019 (CENAN, 2019)



II. ESTADO DEL ARTE

Umoh e Isong (2015) propusieron el diseño para la metodología de un sistema de diagnóstico y control de obesidad como instrumento para contrarrestar el gran aumento de calorías consumidas por las personas. El sistema llamado FESDMO se basó en lógica difusa debido a la incertidumbre de los datos que se manejaban, utilizaron reglas difusas de tipo Mamdani lo que permitió diagnosticar el grado nutricional exacto del paciente (saludable, sobrepeso, obesidad) y fue validado en Matlab.

Suca, Córdova, Condori y Cayra (2016) presentaron su artículo en el cual se desarrolló un modelo para clasificar a hombres y mujeres de 6- 17 años que sufren de este mal. Los resultados obtenidos demostraron que el modelo obtuvo una exactitud 83.65% para hombres y 76.13% en mujeres para la predicción de casos de obesidad.

Suca, Córdova, Condori, Cayra y Sulla (2016) en su artículo evaluaron distintas técnicas de Machine Learning como los árboles de decisión C4.5, J48, Máquinas de Vectores de Soporte (SVM), Naive Bayes y Back Propagation. Como resultado se observó que los árboles de decisión obtuvieron un mayor porcentaje de instancias correctas que la técnica SVM y Redes Neuronales.

Morlán et al. (2017), en su investigación identificaron de manera anticipada el riesgo de que un niño padezca de obesidad infantil, se realizó un estudio de 242 niños y basándose en 14 parámetros antropométricos se construyó el modelo de regresión logística, como resultado se obtuvo un modelo para niñas con 96.65% de sensibilidad y el modelo de niños obtuvo 96.3%.

Sulla et al. (2018) buscaron clasificar la obesidad en niños y adolescentes varones entre los 6 a 17 años utilizando redes neuronales y lógica difusa. Se eligió el modelo neuro-difuso ANFIS que se encuentra en el toolbox de Matlab el cual incluye un completo conjunto de características tanto para la fusificación, defusificación, entrenamiento, pruebas y validación. Las pruebas experimentales realizadas mostraron un 96.96% de exactitud en la clasificación y un 3.04% de error.

Ticona (2018) utilizó el método de Árboles de Decisión J48 implementado con la herramienta Weka para la predicción de obesidad infantil, este obtuvo una precisión de 94.39% y fue superior a otros métodos como BayesNet, ForestPA y Multilayer Perceptron, Hammond et al. (2019) utilizaron registros médicos electrónicos de los dos primeros años de vida e información del historial médico de las madres, entrenó varios algoritmos de aprendizaje automático para realizar una clasificación binaria y regresión entre los cuales se varió el set de variables para cada modelo. Al finalizar la investigación, se concluyó que las variables de peso, talla, IMC entre 19 y 24 meses y la última medida de IMC registrada antes de los dos años fueron las características más importantes para la predicción, los modelos con mejor desempeño obtuvieron un AUC del 81.7% para las niñas y del 76.1% para los niños.

Por último, Rossman et al. (2021) en su artículo diseñó un modelo de predicción dirigido a identificar a los niños con alto riesgo de obesidad antes de esta ventana de tiempo, prediciendo la obesidad a los 5-6 años de edad según los datos de los primeros 2 años de vida de 136,196 niños, el modelo se implementó con el método Gradient Boosting obteniendo un AUC de 80.4%.

III. METODOLOGÍA

La metodología aplicada se basó en KDD, considerando 4 principales fases, la extracción de datos que es primordial, el tratamiento de inconsistencias realizado para limpiar la información, la obtención y creación de nuevas variables y finalmente la aplicación de la técnica Naive Bayes.

3.1 Extracción de datos

El conjunto de datos utilizado se recogió de la web E-qhali del MINSA, se extrajo un total de 244 mil historias clínicas de menores atendidos entre el 2011 y 2021, de los cuales quedaron 771 registros luego de realizar los filtros de completitud de variables, filtro por edad ya que es necesario que el paciente sea mayor a 5 años y además se filtró solo a los pacientes que tengan como mínimo dos consultas de seguimiento de peso.

3.2 Tratamientos de Inconsistencias

Realizar un buen tratamiento a las variables es lo más importante al momento de construir un modelo predictivo, para ello fue necesario conocer la definición de cada variable. En nuestro caso contamos con 27 variables iniciales que se clasifican en 3 etapas de la vida del paciente: Los antecedentes e información de la madre, Información y características del parto, consultas de seguimiento de peso y talla como se muestra en la Figura 3.

Un modelo de Naive Bayes debe tener todas las variables homogeneizadas; es decir, todas deben

Figura 3

Lista de Variables con definición

VARIABLE	DESCRIPCION	TIPO
	ANTECEDENTES	
Fecha de nacimiento	Fecha de nacimiento del niño	Fecha
# de embarazo	Numero de embarazo de la madre en la que nació el niño	Continuo
# de atenciones prenatales	Numero de atenciones prenatales que tuvo la madre durante el embarazo	Continuo
Lugar de atenciones prenatales	Centro médico donde se atendió	Discreto
	PARTO	
Condición del parto	El parto pudo ser Cesárea, Espontaneo, Instrumentado u Otro	Discreto
Sexo	El sexo del paciente puede ser “H” o “M”	Discreto
Lugar del parto	Lugar de atención del parto, puede ser Domicilio, Centro Médico, etc.	Discreto
Parto atendido por	Indica quien fue quien atendió el parto, si un familiar, técnico o profesional de la salud	Discreto
Edad gestacional del nacimiento	Semanas de embarazo antes de dar a luz	Continuo
Peso al nacer (gr)	Peso del niño al nacer	Continuo
Talla al nacer (cm)	Talla del niño al nacer	Continuo
Perímetro Cefálico al nacer	Perímetro cefálico al nacer	Continuo
Peso para la edad gestacional	Puede ser Pequeño, Adecuado o Grande	Discreto
Perímetro torácico al nacer	Perímetro torácico del niño al nacer	Continuo
Índice de apgar_1	Índice de APGAR en el 1’	Continuo
Índice de apgar_5	índice de APGAR en el 5’	Continuo
Enfermedad congénita	Indica si nació o no con una enfermedad congénita	Booleano
Índice CPP	Indica si existió contacto piel a piel	Booleano
Índice AC	Indica si existió Alojamiento Conjunto al nacer	Booleano
Requirió hospitalización	Indica si nació o no con una enfermedad congénita	Booleano
	POSTPARTO	
Fecha_Atencion_1	Fecha del primer control de crecimiento	Date
Peso_consulta_1	Peso del primer control de crecimiento	Continuo
Talla_consulta_1	Talla del primero control del crecimiento	Continuo
Resultado hemoglobina	Resultado de hemoglobina en alguna consulta	Continuo
Fecha_Atencion_2	Fecha del segundo control de crecimiento	Date
Peso_consulta_2	Peso del segundo control de crecimiento	Continuo
Talla_consulta_1	Talla del segundo control del crecimiento	Continuo

ser o categóricas o continuas, en nuestro caso vamos a convertir todas las variables a continuas con el fin de aprovechar el valor que pueden aportar las variables numéricas como el peso, la talla, los resultados de hemoglobina al modelo.

Luego de tener nuestro conjunto de datos limpio y transformados, es necesario revisar que variables tienen datos vacíos; las variables con un bajo porcentaje de información son eliminadas y las que cuentan con solo algunos datos faltantes son llenados con el promedio de los datos existentes agrupados por la variable “Peso para la edad gestacional”. Por último, una vez que se tiene el conjunto de datos listo, con apoyo del experto se identifica los “outliers” para eliminar aquellos registros que posiblemente fueron llenados con información errónea:

3.3 Obtención de Variables

Luego de definir las variables con apoyo del experto se creó la variable TARGET utilizando la edad, el sexo, el peso y la talla del niño en la última consulta con la finalidad de clasificarlo en OBESO o NO OBESO. El proceso para la creación de esta variable se inicia con el cálculo del IMC como se observa en la Figura 4, luego se procede a verificar la situación nutricional apoyándonos de las tablas de percentiles que nos brinda la OMS dependiendo del peso y la edad del paciente.

3.4 Aplicación de Naive Bayes

Luego de haber culminado con la creación del dataset, se procedió a realizar un benchmarking con distintas técnicas de minería de datos basándonos

en 8 criterios de evaluación (Figura 5). Cada criterio de evaluación contó con valores y puntajes para evaluar cada técnica como se muestra en la Figura 6, Naive Bayes obtuvo 18 puntos, siendo superior sobre las otras técnicas por su rápida interpretabilidad, su manejo al ruido en la información y su bajo nivel de complejidad.

Para implementar el modelo utilizamos la librería sklearn de Python, esta librería nos brinda módulos para convertir variables, separar las muestras, entrenar el modelo con los datos y también nos facilita métricas como la matriz de confusión y la precisión del modelo.

Figura 4

Proceso para Cálculo de la Variable Target

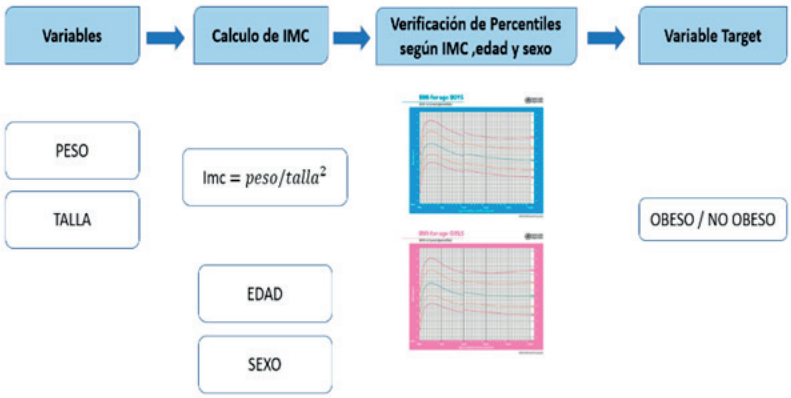


Figura 5

Criterios de Evaluación

ID	CRITERIO	VALOR	DESCRIPCIÓN	PUNTAJE
A.	Naturaleza de Datos	Continuo o Discreto	El método sólo admite que el dato de entrada sea discreto o que sea variable.	1
		Ambos	El método es adaptable a ambos tipos de datos.	2
B.	Cantidad de Datos de Entrenamiento	Alto	Se requiere de un alto número de datos de entrenamiento para alcanzar un alto rendimiento.	1
		Bajo	No se requiere de muchos datos de entrenamiento para alcanzar un alto rendimiento.	2
C.	Interpretabilidad	No interpretable	El método no cuenta con un razonamiento simbólico y representación semántica.	1
		Compleja	El método es descriptivo y requiere cierta interpretación.	2
		Fácil	Se presentan los resultados de manera visual o al menos de manera que su interpretación sea muy clara.	3
D.	Velocidad	Baja	El método requiere de un alto costo computacional, incluso si la cantidad de datos a manipular no es alta, por lo que los tiempos de respuesta son lentos.	1
		Alta	El método no siempre requiere de un alto costo computacional, dependiendo de la cantidad de datos a manipular, los tiempos de respuesta pueden ser incluso inmediatos.	2
E.	Precisión	Baja	El método tiene una precisión ≤ 70%.	1
		Media	La precisión está entre 70% y 85%.	2
		Alta	El método tiene un indicador de precisión mayor a 85%	3
F.	Manejo de Ruido	Bajo	El método tiende a cometer imprecisiones cuando recibe datos con alta presencia de ruido.	1
		Medio	Significa que la técnica es capaz de reconocer algunos datos ruidosos, sin embargo, no es constante y su tasa de precisión puede verse afectada notablemente.	2
		Alto	El método maneja eficientemente cualquier dato recibido, almacenado o editado sin que esto afecte la precisión del resultado.	3
G.	Área de Dominio	Alta	Se requiere conocimiento completo del dominio que abarca el sistema.	1
		Media	Se requiere una delimitación en el dominio de trabajo para la construcción.	2
		Baja	La construcción no requiere ningún dominio de conocimiento.	3
H.	Nivel de Complejidad	Alta	El método utiliza algoritmos complejos que requieren de una alta curva de aprendizaje y son difíciles de implementar si no se cuenta con experiencia previa.	1
		Media	La implementación de los algoritmos tiene una curva media de aprendizaje, no son complicados de implementar, pero si requieren experiencia previa para su buen entendimiento.	2
		Baja	Los algoritmos utilizados son sencillos e intuitivos, con una curva de aprendizaje alta, y no requieren experiencia previa para ser implementados.	3

Figura 6

Benchmarking de técnicas de Minería de Datos

CRITERIO/TECNICA	SISTEMAS NEURO– DIFUSOS	REGRESION LOGISTICA	NAIVE BAYES	RANDOM FOREST	GRADIENT BOOSTING	ID3	C4.5
Naturaleza de los Datos	2	1	2	2	2	2	2
Cantidad de datos de entrenamiento	1	2	2	1	1	1	1
Interpretabilidad	3	3	3	2	1	2	2
Velocidad	2	2	2	1	1	1	1
Precisión	2	2	2	3	3	2	3
Manejo de Ruido	2	1	2	3	3	2	2
Área de Dominio	1	2	2	3	3	3	3
Nivel de Complejidad	2	3	3	2	1	1	2
TOTAL	15	15	18	17	16	14	15

Sklearn nos permite implementar 3 tipos distintos de algoritmo Naive Bayes según el tipo de información y objetivo que tenemos, para nuestro modelo utilizamos el Naive Bayes Gaussiano que recibe información continua y asume que las variables tienen una distribución normal. Una vez seleccionado el modelo se procedió a definir los inputs, de las 27 variables del paciente y la probabilidad de cada clase como entradas, y estas serán utilizadas por el modelo Naive Bayes, el cual nos devolverá la clase predicha para el niño o niña (Obeso o No Obeso) con mayor probabilidad dependiendo de las variables que se ingresen.

Luego de la definición del modelo, con el 59% de los registros se realiza el entrenamiento y con el 41% restante se realiza la validación, luego de haber entrenado el modelo este se guardará en un archivo llamado “finalized_model.sav” que nos permitirá luego cargarlo para poder conectarlo con la aplicación web que nos servirá de interfaz para que el experto pueda realizar las predicciones a sus nuevos pacientes.

IV. RESULTADOS

La evaluación de la técnica Naive Bayes se realizó sobre el 41% de los datos recolectados (317 registros), y la selección de las métricas que debemos calcular para dicha evaluación se basó en el tipo de variable que tenemos que predecir (Target) y ya que esta es un conjunto finito y discreto que clasifica cada registro como niño “Obeso” o “No obeso” las métricas más adecuadas fueron la matriz de confusión, la precisión, la sensibilidad y la exactitud.

La matriz de confusión (Figura 7) nos muestra 4 valores importantes:

- 214 casos Verdaderos Negativos
- 89 casos Falsos Positivos
- 1 caso de Falso Negativo
- 13 casos de Verdaderos Positivos

Con los valores resultantes de la matriz de confusión se calcula la exactitud del modelo (71.6%), la sensibilidad (92.85%) y la precisión (99.53%)

Las técnicas de Regresión Logística, Random Forest, KNN y XGB fueron utilizadas para evaluar y comparar con los resultados del modelo propuesto. Estas técnicas se implementaron en Python utilizando la librería sklearn, además nos apoyamos en la técnica de validación Cross-Validation que nos permitió realizar varias iteraciones con distintos subconjuntos de datos debido a que nuestro Dataset no es muy grande.

Luego de haber evaluado los 5 modelos, se realizó la comparación de las métricas por cada técnica, y la que obtuvo mayor exactitud fue la Regresión Logística con 96.21%, esto significa que fue la técnica con mayor cantidad de aciertos al momento de predecir, pero esto no nos demuestra que tan buena es diferenciando la clase “Obeso” de la clase “No obeso”.

También se comparó la precisión entre técnicas, Naive Bayes fue la que obtuvo 99.53%, el mayor valor para la clase “No obeso”, las otras técnicas obtuvieron un resultado bastante cercano como se muestra en la Figura 8.

Finalmente, al comparar la sensibilidad entre técnicas, Naive Bayes obtuvo un 92% (Figura 9), lo que superó por mucho a las otras. Al tener nuestro

Figura 7

Matriz de Confusión de Naive Bayes

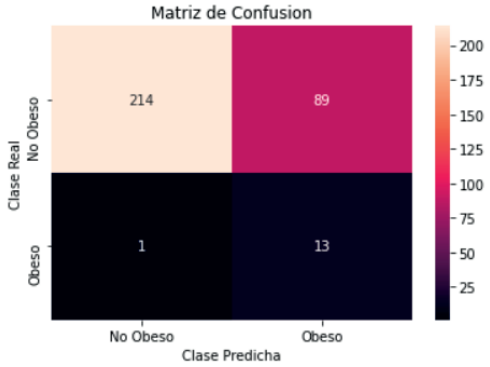


Figura 8

Comparación de Precisión entre técnicas

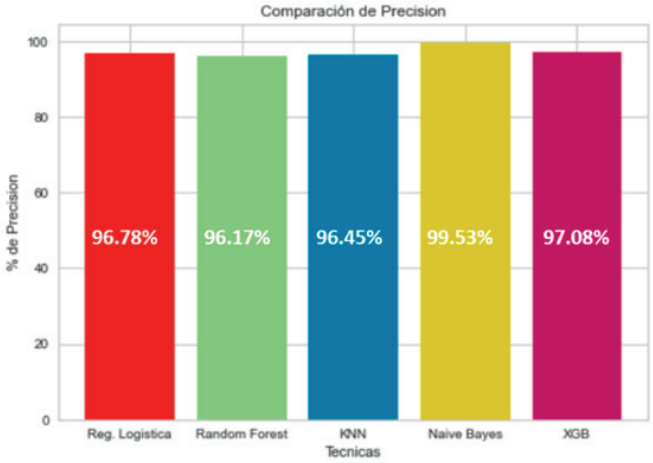
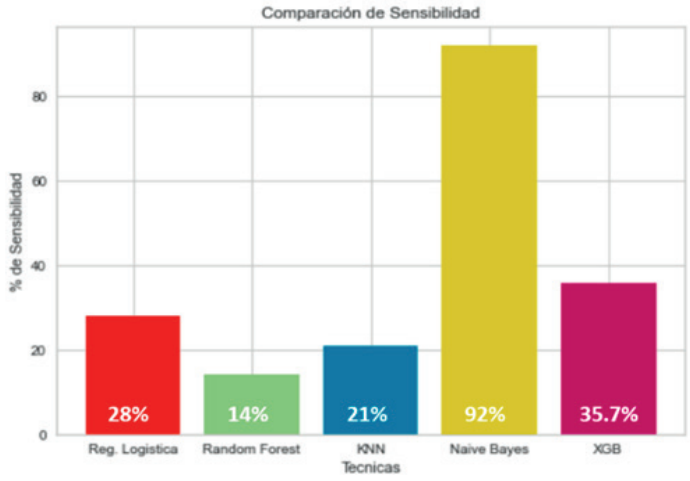


Figura 9

Comparación de Sensibilidad entre técnicas



conjunto de datos clases desbalanceadas debido a la naturaleza del problema donde es mucho más común encontrar niños no obesos, la métrica de sensibilidad es la más indicada para señalar si nuestro modelo está prediciendo bien estos pocos casos de niños que sufrirán de obesidad.

V. CONCLUSIONES

Se recolectó la información de historias clínicas del aplicativo e-Qhali del MINSA para realizar los filtros necesarios y armar el Dataset que cumpla con los requerimientos de que el paciente sea mayor a 5 años y cuente con la información clínica suficiente, el resultado fueron 771 registros con 27 variables, las variables con mayor importancia fueron la condición del parto, el peso en la consulta de seguimiento, la edad de nacimiento gestacional y la variable que indica si el menor necesitó hospitalización al nacer.

Con la información recolectada se desarrolló la aplicación web con el framework Flask de Python para que el experto pueda registrarse, ingresar los datos del paciente y obtener como resultado la predicción y su respectiva probabilidad.

Finalmente, una vez realizada la validación contra las técnicas de regresión logística, Random Forest, KNN y XGB, se pudo concluir que el modelo Naive Bayes resultó ser mejor en precisión y sensibilidad obteniendo 99.53% y 93% respectivamente.

VI. REFERENCIAS

[1] Alva, L., Laria, J., Ibarra, S., Castán, J., & Terán, J. (2020). A proposed diffuse model to determine overweight and obesity in children and adolescents. Revista Chilena de Nutrición, 545-551.

[2] Al-Aidaroos, K., Bakar, A., & Othman, Z. (2018). Medical data classification with Naive Bayes approach. Information Technology Journal, 1166-1174.

[3] CENAN. (2019). Estado Nutricional de Niños Peruanos menores de 5 años 2019. Obtenido de https://web.ins.gob.pe/sites/default/files/Archivos/cenan/van/sala_nutricional/sala_1/2020/sala_situacional_estado_nutricional_ninos_menores_de_5_anos_sienhis_2019.pdf

[4] Guija, E., & Guija, H. (2020). La obesidad como factores de riesgo para COVID-19. Obtenido de https://medicina.usmp.edu.pe/images/noticias_eventos/2020/Obsidad-covid19.pdf

[5] Hammond, R., Athanasiadou, R., Curado, S., Aphinyanaphongs, Y., Abrams, C., Messito, M., . . . Elbel, B. (2019). Predicting childhood obesity using electronic health records and publicly available data. PLoS One.

[6] MINSA-PERU. (2020). El 85% de pacientes fallecidos con comorbilidades por Covid-19 padecían obesidad. [Nota de Prensa]. Obtenido de https://www.gob.pe/institucion/minsa/noticias/286005-el-85-5-de-pacientes-fallecidos-con-comorbilidades-por-covid-19-padecian-obesidad

[7] Morlan, L., de Arriba, A., de Francisco, R., Martínez, I., de Francisco, M., Pascual, J., . . . Ferrández, Á. (2017). Modelo estadístico para la prevención precoz de desarrollo de sobrepeso/obesidad en población infantil. 73-80.

[8] NCDRISC. (2016). Data Visualisations. Obtenido de Recuperado de http://ncdrisc.org/data-visualisations.html

[9] OMS. (2016). Establecimiento de áreas de acción prioritarias para la prevención de la Obesidad Infantil [Archivo PDF]. Obtenido de https://apps.who.int/iris/bitstream/handle/10665/250750/9789243503271-spa.pdf;sequence=1

[10] Orgaz, C. (14 de 05 de 2019). Los países de América Latina donde más ha crecido la obesidad. Obtenido de https://www.bbc.com/mundo/noticias-america-latina-48258937

[11] Senthilingam, M. (14 de 07 de 2017). Estos son los paises mas obesos del Mundo. Obtenido de CNN: https://cnnespanol.cnn.com/2017/07/14/estos-son-los-paises-mas-obesos-del-mundo/

[12] Suca, C., Córdova, A., Condori, A., & Cayra, J. (2016). Modelo difuso para la predicción de casos de obesidad empleando el árbol GFID3 generalizado. Research in Computing Science, 9-22.

[13] Suca, C., Córdova, A., Condori, A., Cayra, J., & Sulla, J. (2016). Comparación de Algoritmos de Clasificación para la Predicción de Casos de Obesidad Infantil.

[14] Sulla, J., Soto, C., Cardenas, R., Huancco, L., & Alfara, L. (2018). Application of the ANFIS Neuro-Fuzzy model for the classification of obesity in children and adolescents. 16° LACCEI International Multi - Conference for Engineering Education and Technology. Arequipa. Obtenido de Recuperado de http://repositorio.unsa.edu.pe/handle/UNSA/6154

[15] Rossman, H., Shilo, S., Barbash-Hazan, S., Shalom, N., Hadar, E., D. Balicer, R., . . . Segal, E. (2021). Prediction of Childhood Obesity from Nationwide Health Records. The Journal of Pediatrics, 132-140.

[16] Ticona, M. (2018). Sistema para la predicción de obesidad en la adolescencia utilizando técnicas de minería de datos [Tesis de Bachiller]. Universidad Catolica de Santa María.

[17] Umoh, U., & Isong, E. (2015). Design Methodology of Fuzzy Expert System for the Diagnosis and Control of Obesity. Computer Engineering and Intelligent Systems.

Fuentes de financiamiento:

Propia.

Conflictos de interés:

Los autores declaran no tener conflictos de interés.