IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Predicción del bajo peso al nacer. Un estudio comparativo entre técnicas tradicionales de análisis y técnicas de Machine Learning. (junio 2023)

Salazar, Diego A[1]

[1]Facultad de Enfermería, Universidad de Antioquia Medellín-Colombia

Autor de correspondencia: Salazar, Diego A. (e-mail: alejandro.salazar@udea.edu.co).

RESUMEN Los avances tecnológicos de las últimas décadas han traído consigo importantes retos en ámbito investigativo, se dispone de grandes de volúmenes de datos y de capacidades computacionales que eran impensables hace algunas décadas. En este punto los algoritmos computacionales y las técnicas estadísticas tradicionales han evolucionado de manera acelerada como disciplina en lo que hoy se conoce como Machine Learning o Aprendizaje de Maquinas. Esa evolución rápida a supuesto cambios en la forma de investigar; se dispone de nuevas y mejoradas herramientas para el procesamiento y análisis de datos. Particularmente en la investigación en el área de la salud esto ha supuesto un cambio de paradigma donde las técnicas estadísticas tradicionales basadas en supuestos teóricos, que permiten la demostración matemática de la efectividad del método y su interpretabilidad en el área, son retadas por mejoras o técnicas algorítmicas que carecen la mayoría de las veces de estos atributos, pero han demostrado ser efectivas en sus resultados. En esta investigación se pretende comparar dos técnicas tradicionales y muy empleadas en área de salud como lo son la Regresión Lineal y la Regresión Logística con algunas técnicas de Machine Learning, en el caso particular de la predicción del bajo peso al nacer, esto con el fin hacer más cercanas las técnicas algorítmicas al personal sanitario mostrando potencialidades y debilidades en un caso en particular de interés en salud y así puedan a futuro tomar decisiones razonadas sobre uso de estas técnicas en sus investigaciones particulares.

PALABRAS CLAVE Bajo peso al nacer, Aprendizaje de máquinas, Regresión, Predicción

INTRODUCTION

Los avances tecnológicos de las últimas décadas han enfocado grandes esfuerzos en lograr simular la inteligencia humana por medio de máquinas (Inteligencia Artificial); el Machine Learning o aprendizaje de maquina es una rama especifica que busca emular por medio de algoritmos computacionales la capacidad que tenemos los humanos de aprender con la experiencia que se puede materializar en este contexto en grandes volúmenes de datos (1).

Esta perspectiva de aprendizaje trajo consigo enormes retos para el análisis estadístico tradicional basado en supuestos teóricos derivados de los fundamentos matemáticos, que subyacen en la mayoría de sus técnicas lo cual ha implicado un cambio de paradigma al privilegiar el poder predictivo de un modelo de análisis sobre el cumplimiento de supuestos teóricos (2). Esta considerable libertad permitió importantes avances en el Machine Learning, el cual se caracteriza por agrupar un conjunto de técnicas estadísticas tradicionales y algoritmos computacionales en busca de optimizar el análisis de datos para la toma de decisiones y cuyo alcance desbordó su intencionalidad inicial y abrió la puerta para que sea aplicado en diferentes y diversas áreas del conocimiento como la medicina (3) o el derecho (4) por mencionar algunas. La implementación del Machine Learning en cualquier disciplina particular implica cuestionamientos teóricos de dichas técnicas que avalen su uso, donde el poder predictivo por sí solo ya no es suficiente, con el agravante de que muchas de estas estas técnicas computacionales han evolucionado a tal punto que se convierten en una caja negra que siempre ofrece resultados, pero donde no se entiende muy bien lo que

pasa al interior como sí ocurre con la mayoría de las técnicas estadísticas convencionales.
A pesar de lo anterior en el área de la salud en general, el uso de técnicas Machine Learning se abren camino poco a poco para análisis de datos, con un alto potencial para transformar ciencias epidemiológicas (2). No obstante, y paralelo a esto, es relevante avanzar en el estudio del alcance y las limitaciones de dichas técnicas que brinden mayor seguridad para su uso en este contexto particular y permitan por ejemplo distinguir cual algoritmo es más efectivo en la predicción de eventos en salud o características en los datos mejoran el desempeño de uno u otro modelo y así poder aportar a la formalización y estandarización de su uso en el área de la salud. El presente estudio busca aportar en esta dirección tomando un evento en salud importante como es el bajo peso al nacer en un contexto local y comparar las potencialidades y desventajas de las técnicas estadísticas tradicionales más usadas en salud como la regresión lineal o logística frente a las Maquinas de soporte vectorial, arboles decisión, procesos gaussianos y redes neuronales. Particularmente en esta entrega solo compara la Regresión Lineal basada en estimación de sus parámetros en mínimos cuadrados con una mejora computacional propuesta en el ámbito de Machine Learning, donde se usan técnicas algorítmicas para estimación de los parámetros basados en entropía y también se aplica en la comparación inicial con un Random Forest Regression.

A. COMPRENSIÓN DEL PROBLEMA DE APRENDIZAJE AUTOMÁTICO

La Organización Mundial de la Salud (OMS) define el Bajo Peso al Nacer (BPN) como aquel recién nacido que tiene un peso inferior o igual a 2.500 gramos sin importar su edad gestacional o motivo del bajo peso. Esta definición se basa en observaciones epidemiológicas según las cuales los niños que pesan menos de 2.500g tienen aproximadamente 20 veces más probabilidades de morir que los bebés con un peso mayor (5).

El BPN es una de las principales causas de mortalidad neonatal y también se asocia con problemas posteriores a esta etapa a mediano y largo plazo. Los niños con bajo peso presentan mayor incidencia de déficit neurológico, alteraciones del crecimiento, deficiencias cognitivas y enfermedades crónicas no transmisibles. Entre las causas del BPN se ha encontrado evidencia de relación con factores sociodemográficos asociados a los padres, la cantidad de consultas prenatales entre otros. Es por esto por lo que se considera un indicador de salud pública en la atención de salud de la madre gestante y el recién nacido, debido a la relación que tiene con las condiciones de vida, la salud de la población y de los cuidados médicos de la mujer gestante. En este sentido contar modelo que permita anticiparse al BPN identificando los factores riesgo que se pueden intervenir permitiría reducir las tasas de mortalidad infantil y mejorar los indicadores de salud.

Diferentes estudios han abordado la predicción del BPN, desde diferentes perspectivas. Benjumea et al (6) usaron variables morfológicas y fisiológicas de madre, encontrando que los predictores más relevantes fueron la circunferencia del brazo y la pantorrilla entre el segundo y el tercer trimestre y el peso de la madre en el primero y segundo trimestre usando regresión lineal y arboles de clasificación y reportaron un R^2 de 0.245. Por su parte Agudelo et al (7) se basaron variables sociodemográficas y factores de la madre y del recién nacido. Encontraron como variables relevantes el nivel educativo de la madre y asistir a más de 4 controles prenatales. Usaron un modelo de Regresión Logística y no reportaron indicadores de bondad y ajuste del modelo. En otro estudio se empleó la información contenida en los certificados de nacido vivo y se aplicó un modelo regresión logística para determinar los predictores asociados. Entre sus hallazgos se destacan como variables predictoras el sexo del bebe, nacer por cesaría, el nivel educativo de la madre y la cantidad de controles prenatales, nuevamente no se reportan métricas de desempeño del modelo (8).

En este estudio se usan variables obtenidas de los certificados del nacido vivo y publicadas en el sitio de datos abiertos Colombia y corresponden a los nacimientos ocurridos en Medellín entre 2012 y 2020. La variable objetivo es el peso del bebe, en este sentido la misma base puede usarse para usar modelos de regresión y clasificación al dicotomizar la variable según el BPN. El objetivo de la investigación es comparar diferentes técnicas y encontrar el modelo que mejor permita predecir exponiendo en cada caso ventajas y desventajas de cada técnica.

B. ENTRENAMIENTO Y EVALUACIÓN DE MODELOS

Los datos están publicados en MeData (repositorio de datos oficiales de la alcaldía de Medellín) tienen licencia Creative Commons (CC), está compuesta por más de 250.000 registros y 40 variables de los nacimientos en el municipio de Medellín, comprendidos en los años 2012 al 2020 (http://medata.gov.co/dataset/nacimientos). El evento de interés de esta base es el peso del recién nacido.

Métricas de desempeño

Para el caso de la predicción del peso del bebe (modelos predictivos de regresión) se contempla usar el MAE (Error Cuadrático Medio) definido por la formula

$$MAE = \frac{1}{M}\sum_{m=1}^{M-1}|y_m - \widehat{y_m}|$$

donde M es el número de muestras de los datos de prueba y m es la m-ésima muestra de los datos de prueba. Y el coeficiente de determinación R^2 definido por

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})}{\sum_{i=1}^{n}(Y_i - \bar{Y})}$$

Y para el caso del bajo peso al nacer (clasificación) se contempla usar
Tasa de Clasificación Errónea o TCE que se define como:

$$TCE = \frac{NCE}{NOBS}$$

donde NCE corresponde al número de clasificaciones erradas por la técnica en el conjunto de validación y NOBS corresponde al número de observaciones en el conjunto de validación.

Criterio de desempeño

Se pretende tener como línea base el desempeño en las métricas de los métodos tradicionales regresión lineal y logística multivariada y se evalúa con respecto a las métricas arrojadas por lo métodos algoritmos.

C. Resultados y discusión

En este primer experimento se realizo una limpieza de los datos eliminando las variables que no son de interés o no presentan ninguna utilidad en el modelado. Se identifican y eliminan los valores nulos o perdidos y registros duplicados. En este punto la base pase de tener 250000 registros a 146814 y de 40 variables a tan solo 21. Luego se procede a etiquetar correctamente las variables con la ayuda del formato de registro del recién nacido, ya que en la base original las variables categóricas no estaban codificadas por números.

Luego de este apartado se genera una visualización de los datos separando las variables categóricas de las cuantitativas y de la variable de salida. Posterior a esto se hace un tratamiento a los datos atípicos se crean las variables dummies se corre un el modelo tradicional de regresión obteniendo un valor base de referencia de R^2=0.683.

Al correr un regresión lineal usando técnicas de entropía en la estimación de los parámetros y usando validación cruzada el R^2=0.683 permanece estable.

Al usar un Random Forest (Regresión) inicial sin ajustar los parámetros el coeficiente de correlación disminuye un poco R^2=0.594. Luego de afinar el modelo se logra tener con esta técnica usando validación cruzada un R^2=0.6363 un valor ligeramente inferior al obtenido con modelo inicial de regresión. Es importante resaltar que el tiempo de cómputo del modelo de usando Arboles de Decisión es considerablemente mayor. Por lo que en este primer experimento la técnica convencional demostró mayor fortaleza en tiempo de ejecución y métrica de despeño.

Resta evaluar otras técnicas de regresión como Maquinas de Soporte Vectorial que no se pudieron evaluar en este primer experimento por los tiempos de ejecución y también comparar las técnicas de clasificación basadas en algoritmos de Machine Learning con el método convencional de la regresión logística.

REFERENCIAS

[1] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, Justin Lessler, What is Machine Learning? A Primer for the Epidemiologist, American Journal of Epidemiology, Volume 188, Issue 12, December 2019, Pages 2222–2239, https://doi.org/10.1093/aje/kwz189

[2] Wiemken TL, Kelley RR. Machine Learning in Epidemiology and Health Outcomes Research. Annu Rev Public Health. 2 de abril de 2020;41(1):21-36.

[3] Núñez Reiz A, Armengol de la Hoz MA, Sánchez García M. Big Data Analysis y Machine Learning en medicina intensiva. Medicina Intensiva. octubre de 2019;43(7):416-26.

[4] Salinas A. Sobre la implementación de algoritmos de Machine Learning en las ciencias penales y sus implicaciones jurídicas. Revista Mexicana De Ciencias Penales. 5 de septiembre de 2020;12:191-204

[5] World Health Organization & United Nations Children's Fund (UNICEF). (2004). Low birthweight : country, regional and global estimates. World Health Organization. https://apps.who.int/iris/handle/10665/43184

[6] Benjumea Rincón, María Victoria, et al. "La prediccion del bajo peso y del peso insuficiente al nacer mediante la antropometria materna." Hacia la Promocion de la Salud, Jan. 2009, pp. 35+.

[7] Agudelo Pérez Sergio, Maldonado Calderón María, Plazas Vargas Merideydy, Gutiérrez Soto Isabel, Gómez Ángela, Díaz Quijano Diana. Relación entre factores sociodemográficos y el bajo peso al nacer en una clínica universitaria en Cundinamarca (Colombia). Salud, Barranquilla [Internet]. 2017 Aug [cited 2023 June 18] ; 33(2): 86-97. Available from: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-55522017000200086&lng=en

[8] Márquez-Beltrán, Marlon & Vargas Hernández, Jhonny & Quiroga-Villalobos, Edwin & Pinzón-Villate, Gloria. (2013). Análisis del bajo peso al nacer en Colombia 2005-2009. Revista de Salud Pública. 15. 626-637.

Salazar DA. Matemático, con maestría en Ciencias Estadística. Enfocado en aplicaciones a las ciencias de la salud, en docencia e investigación. Docente asociado a la Facultad de Enfermería de la Universidad Antioquia. Integrante activo de los grupos de investigación de Políticas Sociales y Servicios de Salud de la Universidad Antioquia, Investigación en Estadística Universidad Nacional de Colombia, Sede Medellín y del grupo de Neurociencias-Universidad Nacional de Colombia. Investigador asociado según clasificación Minciencias https://scienti.minciencias.gov.co/cvlac/visualizador/generarCurriculoCv.do?cod_rh=0001392302.
https://scholar.google.es/citations?hl=es&pli=1&user=XHxxCLMAAAAJ.
https://orcid.org/0000-0002-8724-7705.