PROPUESTA DE MONOGRAFÍA

Título del proyecto	Predicción del bajo peso al nacer. Un estudio comparativo entre técnicas tradicionales de análisis y técnicas de Machine Learning.	
Estudiante 1		
Nombres completos	Diego Alejandro Salazar Blandon	e-mail: alejandro.salazar@udea.edu.co
		GitHub: https://github.com/dasb03/seminario_monografia

1. Descripción del problema

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 autónomo 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 con sigo enormes retos para el análisis estadístico tradicional basado en supuestos teóricos derivados de los fundamentos matemáticos, que subyacen en la mayoría de sus técnicas y trajo consigo un cambio de paradigma al privilegiar el poder predictivo de un modelo de análisis sobre el cumplimiento de supuestos teóricos (2). Esta considerable libertad ha permitido 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 que muchas de las 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 que 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 sulud. 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 mas usadas en salud como la regresión lineal o logística frente a las Maquinas de soporte vectorial, arboles decisión, procesos gaussianos y redes neuronales.

2. Descripción del dataset

Presente el origen de los datos, relacione la dirección de URL para acceder a esta (si lo tienes) y una breve descripción del dataset indicando la volumetría del mismo (número de archivos, filas, columnas) y las columnas o información que contiene. Si el número de características es muy grande, ponga solo las que se considere más representativas para dar una idea de cómo es el dataset.

Los datos están publicados en MeData (repositorio de datos oficiales de la alcaldía de Medellín) tienen licencia Creative Commons (CC), está compuesta por mas de 250.000 registros y 40 variables, algunas se describen en la tabla 1.

Variable	Tipo	Descripción	Obligatorio	Variable	Tipo	Descripción
ID	number	Número identificador	true	**IDFACTORRH**	string	Hemoclasificación del nacido vivo: Factor RH
AREANAC	string	Area del Nacimiento	true	**IDPERTET**	string	De acuerdo con la cultura, pueblo o rasgos físicos, el fallecido era o se reconocia como
COD_INSP	string	Centro poblado del nacimiento (inspección, corregimiento o caserío)	true	**NOM_PUEB**	string	¿A cuál pueblo indigena pertenece?
SIT_PARTO	string	Sitio de la Parto	true	**EDAD_MADRE**	number	Edad de la madre a la fecha del parto
OTRO_SIT	string	Otro sitio, ¿cúal?	true	**EST_CIVM**	string	Estado conyugal de la madre
COD_INST	string	Código de la institución de salud	true	**CODPTORE**	string	Departamento de residencia habitual de la madre
SEXO	string	Sexo del nacido vivo	true	**CODMUNRE**	string	Municipio de residencia habitual de la madre
PESO_NAC	number	Peso del nacido vivo, al nacer	true	**COD_BARRIRES**	string	Bario de residencia del fallecido
TALLA_NAC	number	Talla del nacido vivo, al nacer	true	**N_HIJOSV**	number	Número de hijos nacidos vivos que ha tenido la madre, incluido el presente
FECHA_NAC	date	Fecha del nacimiento	true	**FECHA_NACM**	date	Fecha de nacimiento del anterior hijo nacido vivo

Tabla 1: Descripción de variables en el dataset Nacimientos publicada en Medata.

Registro de los nacimientos de personas residentes 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.

3. Métricas de desempeño

Para el caso de la predicción del peso del bebe (modelos predictivos) 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 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.

4. 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.

5. Asesor: Se dio un primer encuentro con el profesor John Freddy Duitama, espero redondear mas la idea de la monografía en el curso de seminario para concretarlo con el profesor.

6. 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