Entrega 1 proyecto Fundamentos de Deep Learning

Daniel Sierra Botero

Contexto de Aplicación

El tiempo de estancia hospitalaria (Length of Stay, LOS) se refiere al número de días que un paciente permanece en el hospital desde su admisión hasta el alta. Este indicador es crítico para la utilización de recursos y está altamente relacionado con el riesgo de eventos adversos. Estudios han demostrado que, cuando la estancia de un paciente supera los siete días (prolonged length of stay, pLoS), el riesgo de complicaciones intra-hospitalarias aumenta hasta diez veces. No solo esto incrementa la probabilidad de complicaciones, sino que también repercute en los costos para el paciente y el sistema de salud.
La capacidad de predecir con precisión el tiempo de estancia proporciona información valiosa que puede permitir implementar cambios que aumenten la eficiencia hospitalaria y la capacidad del sistema de salud, al mismo tiempo que se reducen los riesgos asociados. Sin embargo, mantener bases de datos estructuradas y actualizadas en los sistemas hospitalarios es un desafío significativo debido a la heterogeneidad de los datos, la falta de estandarización y la dificultad que presentan los profesionales de la salud para tomar datos de forma estructurada durante las consultas. En este contexto, extraer información directamente de las historias clínicas en formato de texto libre representa una oportunidad valiosa para facilitar el análisis y la predicción, aprovechando el contenido no estructurado que a menudo contiene insights cruciales sobre la estancia del paciente.

Objetivo

El objetivo principal de este proyecto es predecir el tiempo de estancia hospitalario de los pacientes a partir del texto libre de sus historias clínicas. En un inicio se plantea crear un modelo de regresión que pueda predecir de forma precisa el tiempo de estancia. Por otra parte, debido a la naturaleza de los datos y a la importancia que representan la detección de estancias largas también se desea explorar la construcción de un modelo de clasificación que identifique la estancia como corta o larga (mayor o menor a 7 días).

Dataset

El presente proyecto se centra en el análisis de datos obtenidos del dataset MIMIC IV versión 2.2, el cual está compuesto por textos libres extraídos de historias clínicas y diversas variables relacionadas con la estancia del paciente. Este conjunto de datos incluye aproximadamente 330,000 historias clínicas asignadas a estancias individuales, ocupando un tamaño aproximado de 3,400 MB. La riqueza de esta base de datos proporciona una oportunidad única para explorar las características de la estancia hospitalaria y su relación con los resultados de salud.
En cuanto a la distribución de las clases, es importante destacar que la variable objetivo puede ser categorizada en diferentes rangos, lo que influye en cómo se aborda el problema, ya sea como uno de regresión o clasificación. Se anticipa que la distribución de las clases será mucho más alta en las estancias cortas, disminuyendo significativamente a medida que se incrementa el tiempo de estancia, lo que presenta un desafío interesante para el modelado.

Métricas

Para evaluar el rendimiento del modelo, se utilizarán métricas específicas de machine learning. En el caso del modelo de regresión, se aplicará el error cuadrático medio (MSE), mientras que para los modelos de clasificación se considerarán métricas como la precisión, la recuperación y el F1-score, asegurando así la robustez del modelo. En este caso métricas más directamente relacionadas con el área como reducción de costos hospitalarios y en la mejora de la rotación de pacientes son inalcanzables en los márgenes de este proyecto ya que requerirían una aplicación del modelo en ambiente real.

Resultados Previos

En la literatura, se encuentran diversos modelos de predicción del tiempo de estancia enfocados en diferentes poblaciones. Por ejemplo, en cuidados intensivos, modelos basados en sistemas de puntuación de gravedad tradicionales como el "simplified acute physiology score II" (SAPS II) reportan AUCROC entre 0.667 [3] y 0.700 [4], mientras que el EuroSCORE alcanza un AUCROC de 0.729 [2]. Otros enfoques, como el uso de biomarcadores y exámenes individuales, muestran AUCROC que oscilan entre 0.67 y 0.73 [4]. Adicionalmente, modelos de regresión logística aplicados a poblaciones específicas, como pacientes de UCI tras cirugía de revascularización coronaria, presentan AUCROC entre 0.72 [2] y 0.78 [1].

References

[1] Christine Herman, Wojtek Karolak, Alexandra M. Yip, Karen J. Buth, Ansar Hassan, and Jean Francois Légarè. Predicting prolonged intensive care unit length of stay in patients undergoing coronary artery bypass surgery – development of an entirely preoperative scorecard. *Interactive CardioVascular and Thoracic Surgery*, 9:654–658, 10 2009.

[2] Katherine Meadows, Richard Gibbens, Dip Mathematical Statistics, Caroline Gerrard, and Alain Vuylsteke. Prediction of patient length of stay on the intensive care unit following cardiac surgery: A logistic regression analysis based on the cardiac operative mortality risk calculator, euroscore. *Journal of Cardiothoracic and Vascular Anesthesia*, 32:2676–2682, 2018.

[3] Jingyi Wu, Yu Lin, Pengfei Li, Yonghua Hu, Luxia Zhang, and Guilan Kong. Predicting prolonged length of icu stay through machine learning. *Diagnostics*, 11, 12 2021.

[4] Bernhard Zoller, Katharina Spanaus, Rahel Gerster, Mario Fasshauer, Paul A. Stehberger, Stephanie Klinzing, Athanasios Vergopoulos, Arnold von Eckardstein, and Markus Béchir. Icg-liver test versus new biomarkers as prognostic markers for prolonged length of stay in critically ill patients - a prospective study of accuracy for prediction of length of stay in the icu. *Annals of Intensive Care*, 4:1–5, 12 2014.