Seminario de Proyectos

4ª. Resumen. Generalidades de una Base de Datos de Investigación

Luis Fernando Izquierdo Berdugo

4 de mayo de 2025

Tipo de algoritmo seleccionado

Considerando las necesidades del proyecto, se seleccionan las redes neuronales artificiales, particularmente los modelos de Deep Learning y Redes Neuronales Recurrentes (RNN). Estas últimas son especialmente útiles cuando se trabaja con secuencias temporales de datos clínicos, como es el caso de los registros electrónicos de salud (EHR).

Las redes neuronales permiten identificar patrones complejos entre múltiples variables clínicas, incluso cuando estos no siguen una relación lineal. Además, los modelos de Deep Learning han demostrado ser eficaces para tareas de clasificación en contextos médicos, como la detección temprana de enfermedades cardiovasculares.

Características necesarias para el dataset

Dado que se usarán redes neuronales profundas y recurrentes, las características que debe cumplir el conjunto de datos son:

- **Multivariado**: que contenga varias variables por paciente (clínicas, demográficas, etc.).
- **Estructurado y secuencial**: idealmente en formato longitudinal con registros temporales.
- **Limpio y completo**: con valores faltantes imputados y errores corregidos.
- **Anotado**: cada ejemplo debe indicar si el paciente desarrolló o no enfermedad cardiovascular.
- **Balanceado**: para evitar sesgos hacia una sola clase.
- **Normalizado**: para asegurar una escala uniforme entre las variables numéricas.

Aspectos generales para la construcción de la base de datos

Fuente(s) de información

- Base de datos privada generada por el Hospital Juárez de México.
- Alternativamente, se puede utilizar MIMIC-III, una base de datos pública con información de salud de más de 40,000 pacientes.

Manera o algoritmo para obtener los datos

- Extracción mediante scripts en Python, usando librerías como pandas, SQLAlchemy y psycopg2.
- Preprocesamiento que incluye limpieza, codificación categórica, normalización y manejo de valores faltantes.

Sistema de gestión de base de datos (SGBD)

- Se utilizará MySQL para almacenamiento estructurado.
- Para prototipado local, se podrá usar SQLite o Pandas DataFrames.

Variables a incluir en la base de datos

- Datos demográficos: edad, sexo, nivel educativo, ocupación.
- Indicadores clínicos: presión arterial, frecuencia cardiaca, glucosa, colesterol total, HDL, LDL, triglicéridos.
- Estilo de vida: tabaquismo, alcoholismo, sedentarismo.
- Antecedentes personales y familiares.
- Diagnóstico final.
- Timestamps para análisis temporal (en caso de usar RNN).

Reglamentación de uso de datos

- Firma de acuerdo de confidencialidad con el Hospital Juárez de México.
- Anonimización de datos personales conforme a la Ley General de Protección de Datos Personales en Posesión de Sujetos Obligados (LGPDPP) en México.

Referencias

Baesens, B. (2014). *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. John Wiley & Sons. pp. 13 a 34.

Ascolano Ruiz, F., Cazorla Quevedo, M. A., Alfonso, M. I., Colomina Pardo, O. y Lozano Ortega, M. A. (2003). *Inteligencia artificial: modelos, técnicas y áreas de aplicación*. Editorial Paraninfo. pp. III-VIII