Informe Proyecto final Curso fundamentos de Deep Learnin: Predicción de tiempo de estancia hospitalario a partir de notas medicas.

Daniel Sierra Botero

19 de noviembre del 2024

1. Introducción

Este informe detalla el desarrollo, implementación y evaluación de un modelo basado en redes neuronales recurrentes para la predicción de tiempos de estancia hospitalaria, abordando tanto una tarea de regresión como de clasificación binaria. El proyecto comprende varias iteraciones de diseño y optimización, evaluadas mediante métricas relevantes para cada tarea.

2. Estructura del repositorio

El proyecto está organizado en un conjunto de notebooks independientes, cada uno enfocado en una iteración específica del modelo. A continuación, se describe su estructura y contenido:

2.1 Estructura de los Notebooks

Cada notebook sigue un esquema uniforme que permite al usuario entender, ejecutar y analizar el comportamiento de la iteración correspondiente. La estructura general incluye:

- **Introducción:** Breve descripción de los objetivos y las modificaciones específicas de la iteración del modelo.

- **Preprocesado:** Código para cargar y preparar una muestra representativa de la base de datos, limitada a 1000 ejemplos para optimizar el tiempo de ejecución durante la evaluacion del proyecto.
- **Definición del Modelo:** Arquitectura específica para la iteración en cuestión.
- **Entrenamiento:** Entrenamiento del modelo con la muestra seleccionada.
- **Evaluación:** Resultados del modelo sobre el conjunto de datos de prueba derivado de la muestra.
- **Resultados sobre la Base de Datos Original:** Presentación de las métricas y gráficas obtenidas al evaluar el modelo entrenado sobre el conjunto completo de datos originales, para comparar con la evaluación representativa.

2.2 Organización Adicional en el Repositorio

En el repositorio se encuentra una carpeta denominada `otros`, que incluye archivos relevantes para reproducir y analizar los resultados obtenidos. El contenido es el siguiente:

- **Imágenes de resultados:** Gráficos generados al evaluar los modelos sobre la base de datos completa, organizados por iteración.
- **Modelos entrenados:** Archivos en formato `.h5` que contienen los pesos de las redes neuronales entrenadas, correspondientes a cada iteración.
- **Subset de datos procesados:**
  - `admission_subset.csv.gz`: Subset con datos de admisiones preprocesados.
  - `discharge_subset.csv.gz`: Subset con datos de egresos preprocesados.
  - `clean_subset.csv.gz`: Conjunto integrado y preprocesado, listo para el entrenamiento.

Las bases de datos completas utilizados en este proyecto se obtuvieron de la versión 2.2 de la base de datos MIMIC-IV. Específicamente:

- Los datos de admisiones, que contienen la información relacionada con el tiempo de estancia hospitalaria, provienen del recurso `https://physionet.org/content/mimiciv/2.2/`.
- Los datos de egresos, que incluyen el texto libre utilizado para generar embeddings, provienen de `https://physionet.org/content/mimic-iv-note/2.2/`.

Cabe mencionar que el acceso a la base de datos MIMIC-IV no es completamente libre. Sin embargo, los requisitos para obtenerlo son sencillos: crear una cuenta en la plataforma PhysioNet, verificarla y aprobar un breve test gratuito en línea sobre el manejo ético de datos médicos.

Esta organización permite ejecutar los notebooks fácilmente, analizar los resultados de las iteraciones y reproducir los experimentos utilizando tanto la muestra como el conjunto completo de datos.

3. Descripción de la Solución

3.1 Preprocesado de los Datos

Se llevaron a cabo las siguientes etapas:

- Tokenización y padding para transformar las notas médicas en secuencias de longitud fija.
- División de los datos en conjuntos de entrenamiento, validación y prueba, asegurando una distribución representativa de las variables objetivo.
- Creación de una matriz de embedding con vectores preentrenados (CBOW y Skip-Gram) utilizando la librería Gensim.

3.2 Arquitectura del Modelo

La base de todos los modelos consiste en:

- Una capa `Embedding` inicial, con pesos preentrenados y congelados.
- Dos capas LSTM (o bidireccionales en iteraciones específicas), con 128 y 64 unidades respectivamente.
- Capas `Dropout` del 30 % despues de cada capa LSTM para prevenir sobreajuste.

- Una capa densa final:
  - **Regresión**: activación lineal.
  - **Clasificación**: activación sigmoide.
- Configuración de compilación:
  - **Regresión**: funcion de pérdida de `mean squared error (MSE)`, métricas de `MAE` y `RMSE`.
  - **Clasificación**: funcion de pérdida de `binary crossentropy`, métricas de `accuracy` y `recall`.

4. Descripción de las Iteraciones

Cada iteración se centró en experimentar con diferentes configuraciones, como se describe a continuación:

- **CBOW (`sg=0`) con 10 épocas**: Modelo base para establecer una línea de referencia en ambas tareas.
- **Skip-Gram (`sg=1`) con 20 épocas**: Evaluación de un embedding alternativo y un entrenamiento más extenso.
- **Capas Bidireccionales (`sg=0`) con 20 épocas**: Uso de LSTM bidireccionales para mejorar la captura de dependencias contextuales.

5. Descripción de Resultados

5.1 Resultados para Regresión

- **CBOW (`sg=0`, 10 épocas)**:
  - **MSE:** 23.70
  - **MAE:** 2.20
  - **RMSE:** 4.87
  - **R^2:** 0.57

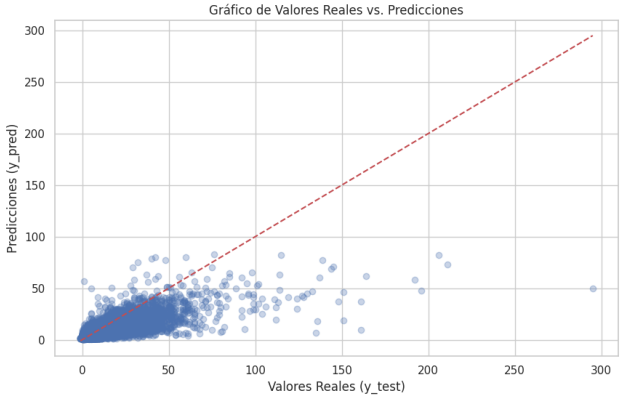

Figura 1: Gráfico de dispersión para el modelo CBOW (`sg=0`, 10 épocas).

- **Skip-Gram (`sg=1`, 20 épocas)**:
  - **MSE:** 18.01
  - **MAE:** 1.93
  - **RMSE:** 4.24
  - **R^2:** 0.67

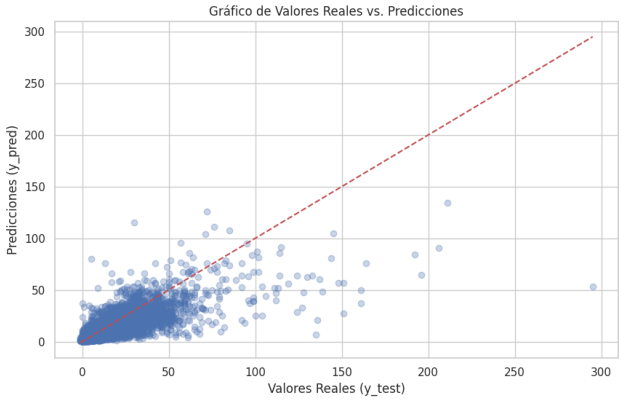

Figura 2: Gráfico de dispersión para el modelo Skip-Gram (`sg=1`, 20 épocas).

- **Bidireccional (`sg=0`, 20 épocas)**:
  - **MSE:** 19.53
  - **MAE:** 2.05

- **RMSE:** 4.41
- **R^2:** 0.64

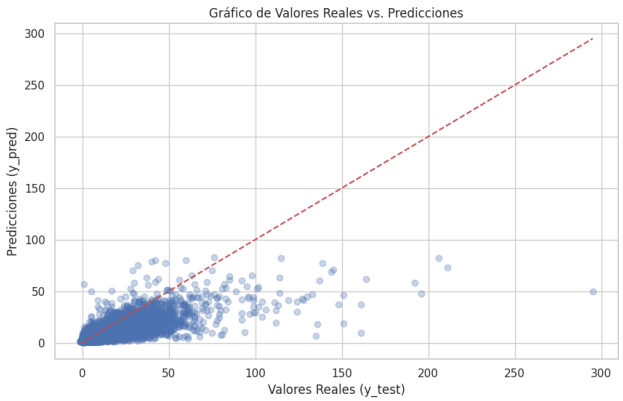

Figura 3: Gráfico de dispersión para el modelo Bidireccional (`sg=0`, 20 épocas).

5.2 Resultados para Clasificación Binaria

- **CBOW (`sg=0`, 10 épocas)**:
  - **Accuracy:** 89.9 %
  - **Recall:** 69.8 %
  - **AUC-ROC:** 0.94

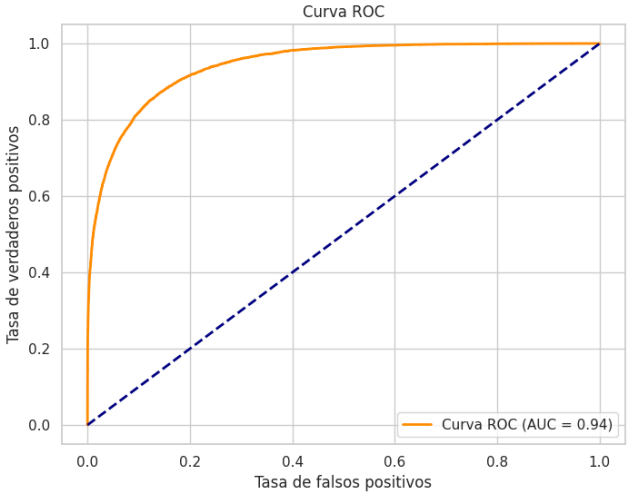

Figura 4: Curva ROC para el modelo CBOW (`sg=0`, 10 épocas).

- **Skip-Gram (`sg=1`, 20 épocas)**:
  - **Accuracy:** 89.5 %
  - **Recall:** 74.9 %
  - **AUC-ROC:** 0.94

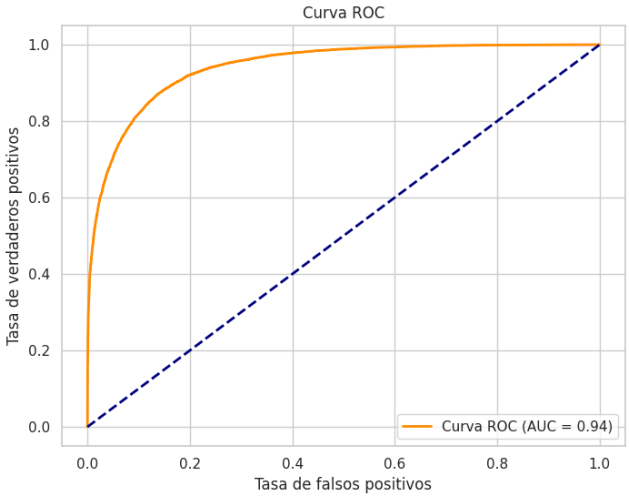

Figura 5: Curva ROC para el modelo Skip-Gram (`sg=1`, 20 épocas).

- **Bidireccional (`sg=0`, 20 épocas)**:
  - **Accuracy:** 90.0 %
  - **Recall:** 71.8 %
  - **AUC-ROC:** 0.95

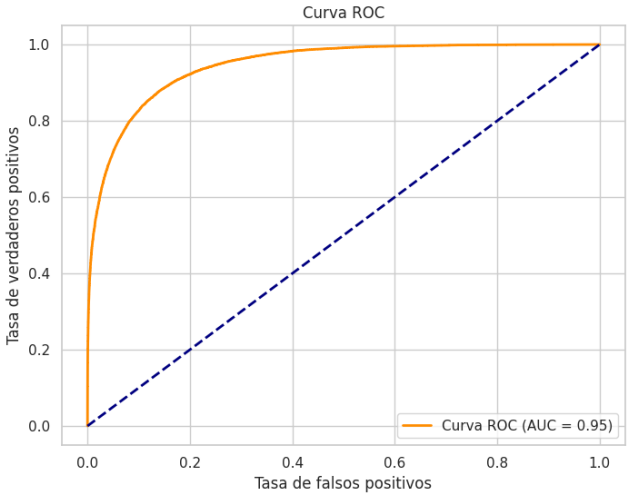

Figura 6: Curva ROC para el modelo Bidireccional (`sg=0`, 20 épocas).

Se puede observar que la implementación de redes bidireccionales no mostró una mejora significativa en el desempeño con respecto a las iteraciones previas, observándose resultados muy similares entre las últimas dos configuraciones evaluadas tanto para la regresión como para la clasificación binaria. Como línea de trabajo futura, sería interesante explorar un modelo multimodal que incorpore no solo el texto libre procesado, sino también las numerosas variables categóricas disponibles en MIMIC IV, relacionadas tanto con las características del paciente como con el contexto hospitalario, lo que podría mejorar la capacidad predictiva del modelo.