Informe de Proyecto

POR:

Iván Daniel Salazar Alarcón

MATERIA

Fundamentos de Deep Learning

PROFESOR:

Raúl Ramos Pollan

UNIVERSIDAD DE ANTIOQUIA

Facultad de Ingeniería

Medellín 2023

Resumen:

El objetivo de este proyecto es plasmar mis conocimientos adquiridos durante el curso Fundamentos de Deep Learning (Ramos & Arias, 2020) aplicado a mi tema de investigación de readmisión hospitalaria. Dando un contexto del aplicativo en el que se quiere predecir la probabilidad que un paciente hospitalario sea readmitido, describiendo un poco los datos tomados de Diabetes 130-US hospitals for years 1999-2008 Data Set (Strack, y otros, 2014) y los procedimientos aplicados para ser apropiados en el uso de una arquitectura de Deep Learning con redes convoluciones para este problema en específico y finalmente se muestran los resultados obtenidos y las conclusiones de los mismos.

El reingreso hospitalario es el retorno no programado de un paciente dentro de un preespecificado período de tiempo después del alta hospitalaria, internacionalmente se asumen los siguientes 30 días como dicha ventana de tiempo (Wang, Shuwen and Zhu, & Xingquan, 2021).

Las causas detrás de los reingresos hospitalarios son diversas y muchas son evitables, por este motivo es importante abordar este tema de estudio, además que esta medida podría ser utilizada como indicador de calidad de atención al paciente y de otros factores considerables como su impacto económico en el sistema de salud.

Este dataset se encuentra disponible en UCI machine learning repository, el conjunto de datos representa 5 MB de información recolectada entre 1999 y 2008 de 130 hospitales de EE. UU. Incluye más de 100 mil registros de pacientes diagnosticados con diabetes contando con 50 características clínicas.

Estructura de los Notebooks

- **01 - exploración de datos.ipynb:**

  En este primer notebook es donde se hace la descarga de los datos desde la página en donde se encuentran disponibles, se llevó a cabo un análisis exploratorio de los datos de readmisión hospitalaria, evidenciando las características con variables tanto numéricas como categóricas en la composición de los datos, se muestran los primeros 5 registros:

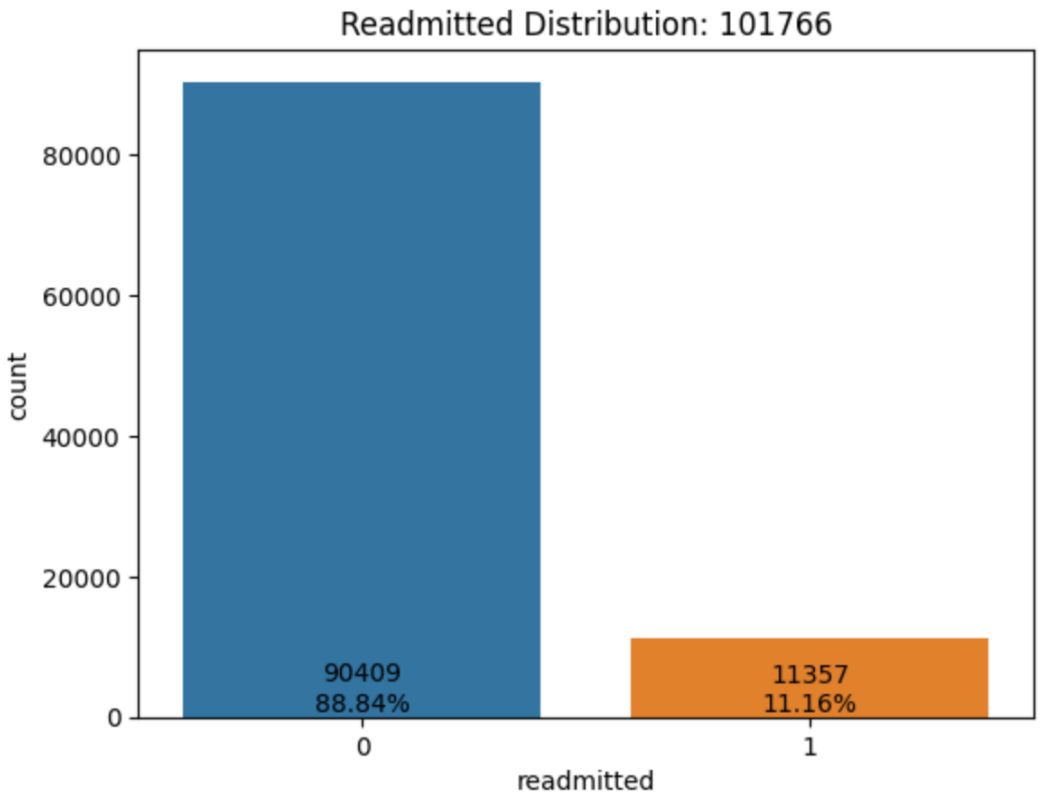
index	0	1	2	3	4

encounter_id	2278392	149190	64410	500364	16680
patient_nbr	8222157	55629189	86047875	82442376	42519267
race	Caucasian	Caucasian	AfricanAmerican	Caucasian	Caucasian
gender	Female	Female	Female	Male	Male
age	[0-10)	[10-20)	[20-30)	[30-40)	[40-50)
weight	?	?	?	?	?
admission_type_id	6	1	1	1	1
discharge_disposition_id	25	1	1	1	1
admission_source_id	1	7	7	7	7
time_in_hospital	1	3	2	2	1
payer_code	?	?	?	?	?
medical_specialty	Pediatrics-Endocrinology	?	?	?	?
num_lab_procedures	41	59	11	44	51
num_procedures	0	0	5	1	0
num_medications	1	18	13	16	8
number_outpatient	0	0	2	0	0
number_emergency	0	0	0	0	0
number_inpatient	0	0	1	0	0
diag_1	250.83	276	648	8	197
diag_2	?	250.01	250	250.43	157
diag_3	?	255	V27	403	250
number_diagnoses	1	9	6	7	5
max_glu_serum	None	None	None	None	None
A1Cresult	None	None	None	None	None
metformin	No	No	No	No	No
repaglinide	No	No	No	No	No
nateglinide	No	No	No	No	No
chlorpropamide	No	No	No	No	No
glimepiride	No	No	No	No	No
acetohexamide	No	No	No	No	No
glipizide	No	No	Steady	No	Steady
glyburide	No	No	No	No	No
tolbutamide	No	No	No	No	No
pioglitazone	No	No	No	No	No
rosiglitazone	No	No	No	No	No
acarbose	No	No	No	No	No
miglitol	No	No	No	No	No

troglitazone	No	No	No	No	No
tolazamide	No	No	No	No	No
examide	No	No	No	No	No
citoglipton	No	No	No	No	No
insulin	No	Up	No	Up	Steady
glyburide-metformin	No	No	No	No	No
glipizide-metformin	No	No	No	No	No
glimepiride-pioglitazone	No	No	No	No	No
metformin-rosiglitazone	No	No	No	No	No
metformin-pioglitazone	No	No	No	No	No
change	No	Ch	No	Ch	Ch
diabetesMed	No	Yes	Yes	Yes	Yes
readmitted	NO	>30	NO	NO	NO

De esta manera y según los artículos leídos y referenciados en la bibliografía, se identifican las características más importantes en la readmisión hospitalaria (Strack, y otros, 2014), se ve que algunos valores de las variables categóricas faltantes estaban como '?' y posteriormente se trató la variable de respuesta (readmitted) para obtener una salida binaria siendo '0' no readmito y '1' readmitido. De forma adiconal, con ayuda de gráficos descriptivos se pudo evidenciar el desbalanceo de esta clase dependiente:

ILUSTRACIÓN 1



Se guarda el nuevo dataset en Google drive, con los cambios de la distribución binaria de clases en la variable de respuesta y el filtrado de las características más relevantes en el estudio.

- **02 - preprocesado.ipynb:** En este notebook se realizó el preprocesamiento de los datos para su posterior uso en el modelo de predicción. Se llevaron a cabo tareas como: la descarga del nuevo dataset desde el drive, la limpieza de datos como el manejo de valores faltantes usando la técnica del valor que más se repite, la codificación de variables categóricas usando 3 técnicas según sea la naturaleza de los datos a codificar (Cavin, 2022):
  - One-Hot / Dummy Encoding.
  - Binary Encoding.
  - Label / Ordinal Encoding.

Obteniendo el siguiente dataset mostrando las primeras muestras:

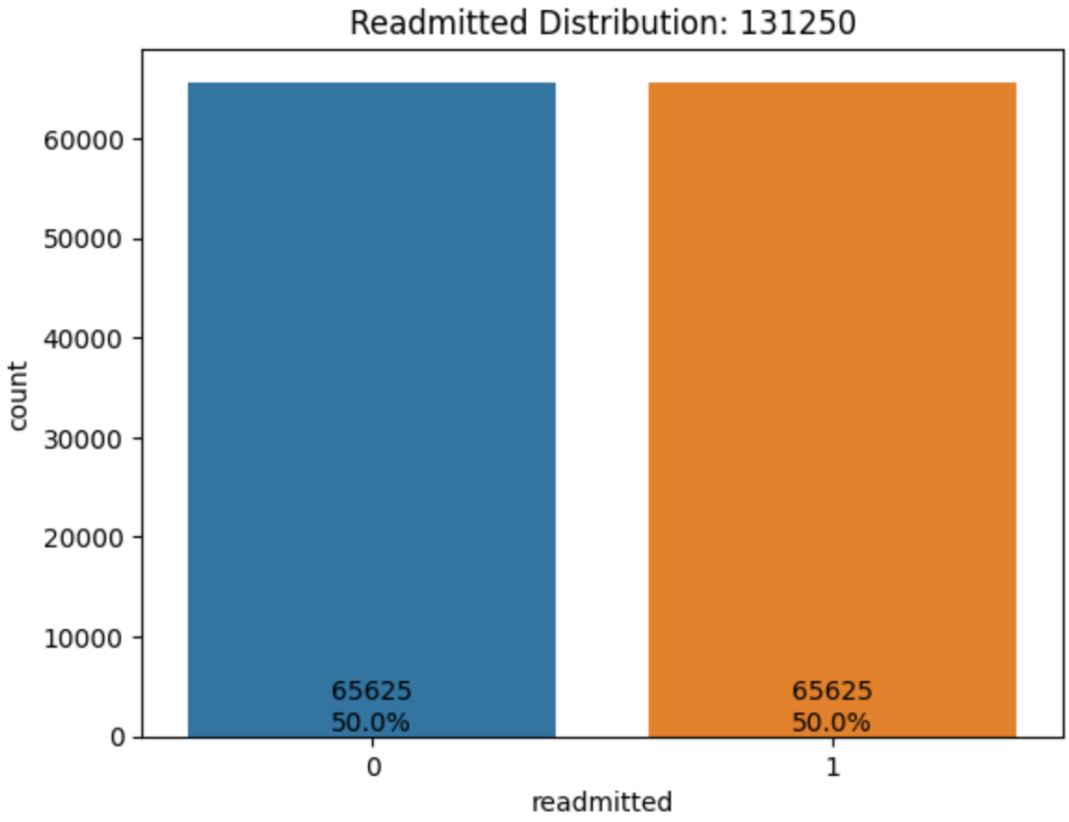
index	**0**	**1**	**2**	**3**	**4**
gender	0	0	0	1	1
age	0	1	2	3	4
admission_type_id	6	1	1	1	1

discharge_disposition_id	25	1	1	1	1
admission_source_id	1	7	7	7	7
time_in_hospital	1	3	2	2	1
number_diagnoses	1	9	6	7	5
num_lab_procedures	41	59	11	44	51
num_procedures	0	0	5	1	0
num_medications	1	18	13	16	8
max_glu_serum	0	0	0	0	0
A1Cresult	0	0	0	0	0
metformin	0	0	0	0	0
glimepiride	0	0	0	0	0
glipizide	0	0	1	0	1
glyburide	0	0	0	0	0
pioglitazone	0	0	0	0	0
rosiglitazone	0	0	0	0	0
insulin	0	3	0	3	1
change	1	0	1	0	0
diabetesMed	0	1	1	1	1
readmitted	0	0	0	0	0
race_AfricanAmerican	0	0	1	0	0
race_Asian	0	0	0	0	0
race_Caucasian	1	1	0	1	1
race_Hispanic	0	0	0	0	0
race_Other	0	0	0	0	0

Vemos que ya todas las variables son numéricas, ya resuelto este problema, se continúa con el tratado del desbalance de clases, acá se trató con 2 técnicas:

- Submuestreo de la clase más dominante.
- Sobremuestrear la clase desequilibrada.

Nos dio mejor resultado la segunda opción, corrigiendo este problema:



- **03 - arquitectura de línea de base.ipynb:** En este notebook se implementó una arquitectura de red convolucional (CNN) como línea de base para el modelo de predicción de readmisión hospitalaria.

  Inicialmente se carga la data procesada. Para asegurarse de que el modelo generalice y no se ajuste demasiado, los datos se dividieron en dos partes: 20% para pruebas y 80% para entrenamiento y se normalizaron para estén en una escala común y estandarizada.

  Posteriormente se desarrolló el modelo de redes neuronales convolucionales, dado que el clasificador distingue entre dos clases (0 como no readmitido y 1 como readmitido), se eligió la función de activación Sigmoid para la capa de salida. Se eligió ReLU para todas las demás para conducir a un cálculo eficiente y menos problemas de gradiente (Hammoudeha, Al-Naymat, Ghannamb, & Obieda, 2018), teniendo inicialmente 2 capas convolucionales con 16 filtros de tamaño 3 y 8 filtros de tamaño 5 respectivamente para tratar de extraer información de diferentes dimensiones, adicional entre las capas convoluciones se agrega una de MaxPooling para disminuir la dimensionalidad a la hora de hacer la transición a las capas fully

connected a las que se les aplicó una capa de dropout entre ellas para regularización y finalmente tendrán una neurona de salida con el porcentaje de predicción.

El modelo se compila con algoritmo de optimización Adam y se quiere reducir la función de perdida binary_crossentropy descrita por:

$$\mathcal{L}(\hat{y}, y) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)$$

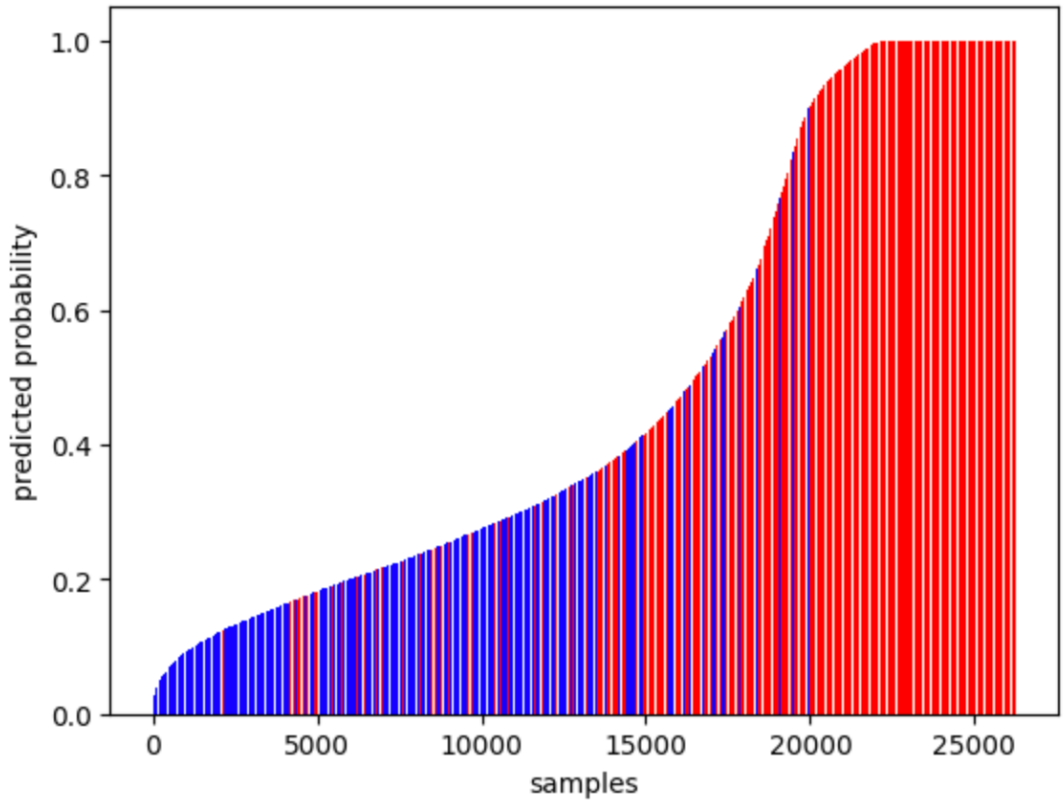
En el entrenamiento se utilizó la técnica de EarlyStopping para evitar el sobreajuste, cuando el error de validación aumenta para máximo 3 iteraciones, el entrenamiento se detiene.

Durante el desarrollo del proyecto, se llevaron a cabo varias iteraciones para mejorar la arquitectura del modelo y optimizar sus hiperparámetros. Se realizaron pruebas con diferentes configuraciones de capas y número de filtros en las capas convolucionales. También se experimentó con diferentes tasas de aprendizaje y algoritmos de optimización, como RMSprop y Adam.

Además, se aplicaron técnicas de regularización, como la capa de dropout, para evitar el sobreajuste del modelo. Se realizaron pruebas con diferentes valores de dropout para encontrar el equilibrio entre la capacidad de generalización y el rendimiento del modelo.

Resultados Obtenidos

Los resultados obtenidos fueron evaluados utilizando una matriz de confusión para analizar la precisión de las predicciones del modelo, obteniendo una precisión aceptable para ambas clases de alrededor del 80%. Además, se generó una gráfica de distribución de probabilidades ordenadas de menor a mayor, lo que permitió analizar la separación entre clases y definir el umbral de predicción para evitar que una clase tenga mejor probabilidad de predicción que otra.



Viendo que hay una marcada separación de clases, teniendo a la izquierda con las barras azules los pacientes con menos probabilidad de ser readmitidos, mientras que a la derecha con las barras rojas son los pacientes con más probabilidad a ser readmitidos.

Cómo acciones adicionas se podría analizar de mejor manera la selección de características del conjunto de datos haciendo análisis estadísticos propios validando que efectivamente estas características tengas influencia sobre la variable de respuesta, también hacer un análisis de las barras rojas que se encuentras en la parte azul y viceversa para validar si son valores atípicos y obtener mejor precisión.

BIBLIOGRAFÍA

Wang, Shuwen and Zhu, & Xingquan. (2021). Predictive modeling of hospital readmission: challenges and solutions. (IEEE/ACM Transactions on Computational Biology and, Ed.) *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 19*, 2975--2995.

Strack, Beata and DeShazo, Jonathan P and Gennings, Chris and Olmo, Juan L and Ventura, Sebastian and Cios, . . . John N. (2014). Impact of HbA1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*.

Van Walraven, Carl and Dhalla, Irfan A and Bell, Chaim and Etchells, Edward and Stiell, Ian G and Zarnke, . . . Alan J. (2010). Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Can Med Assoc, 182*, 551-557.

Rajkomar, Alvin and Oren, Eyal and Chen, Kai and Dai, Andrew M and Hajaj, Nissan and Hardt, . . . Mimi and others. (2018). Scalable and accurate deep learning with electronic health records. *Nature Publishing Group UK London*, 18.

Huang, Y., Talwar, A., Chatterjee, S., & Aparasu, R. (2021). Application of machine learning in predicting hospital readmissions: a scoping review of the literature. BMC medical research methodology 21.1.

Ramos, R., & Arias, J. (2020). *Fundamentos de Deep Learning*. (Universidad de Antioquia) Recuperado el 2023, de https://rramosp.github.io/2021.deeplearning/intro.html

Strack, B., DeShazo, J., Gennings, C., Olmo, J., Ventura, S., Cios, K., & Clore, J. (2014). Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. *BioMed Research International*, 11.

Cavin, A. (2022). *medium*. Obtenido de 6 Ways to Encode Features for Machine Learning Algorithms: https://towardsdatascience.com/6-ways-to-encode-features-for-machine-learning-algorithms-21593f6238b0

Hammoudeha, A., Al-Naymat, G., Ghannamb, I., & Obieda, N. (2018). Predicting Hospital Readmission among Diabetics using Deep Learning. *ScienceDirect*.