Enginyeria
UAB Universitat Autònoma de Barcelona

Universidad Autónoma de Barcelona

Escuela de Ingeniería

Grado en Ingeniería de Datos

Sistema Interactivo para el Seguimiento de Salud y Predicción de Hábitos

Alumno: Pau Montesinos Cáliz

Tutor: Oriol Cortés Comellas

Fecha: 25/05/2025

1. Desarrollo

El desarrollo de esta parte del proyecto ha sido en base a la creación y entreno de un modelo de aprendizaje profundo (Deep Learning). En un principio se había planteado la creación de dos modelos distintos (una MLP y una Sparse Attention-Based Neural Network). Debido a los malos resultados obtenidos con la MLP, en gran medida provocadas por un absurdo desbalanceo de las clases, se ha decidido trabajar con únicamente el primer modelo propuesto y ajustarlo de la mejor forma posible mediante diferentes técnicas:

- Random Oversampling: Copia aleatoria de filas de la clase minoritaria hasta llegar al mismo número de registros que la clase mayoritaria.
- SMOTE: Parecido al oversampling pero, en vez de la duplicación de filas, usa a los vecinos más cercanos para la generación de nuevos datos mediante la interpolación.
- Random Undersampling: Reducción aleatoria de registros de la clase mayoritaria, haciendo que el modelo esté balanceado con menos datos, pero con datos reales no sintéticos y sin duplicados.
- Tomek Links: Undersampling eliminando registros de la clase mayoritaria mediante el cálculo de la distancia euclidiana
- Establecimiento de pesos: Aumento del peso de los errores de la clase minoritaria. Siendo 0 la clase mayoritaria y 1 la clase minoritaria, un peso de 1:2 haría que penalizase el doble los errores en la clase 1.
- Establecimiento de un threshold: A través de la imposición de un threshold a las probabilidades calculadas por el MLP, se puede reducir el punto de corte entre clase 0 y clase 1.

Además de esta decisión de solo adaptar un modelo, también se ha eliminado uno de los datasets (Cancer). Este dataset ya se había puesto en cuestión anteriormente, más ha sido en este tramo del proyecto cuando se ha decidido eliminar completamente.

Esto ha resultado así debido a la naturaleza del dataset. Mientras que los datasets relacionados a los derrames cerebrales y las enfermedades del corazón representan tanto hábitos como datos de usuarios, el de cáncer nos muestra datos sobre un cáncer en concreto. Por otro lado, vemos que las columnas objetivo también son distintas. Mientras en los dos datasets escogidos el objetivo es el saber si ha tenido o va a tener cierta enfermedad, en el de cáncer el objetivo es ver si ese cáncer es benigno o maligno.

Dados estos hechos, y debido también a la naturaleza del proyecto (salud y hábitos), se ha considerado que eliminar este dataset de las variables a tener en cuenta era la mejor opción.

Dentro del modelo escogido también podemos ver diferentes estructuras desarrolladas en su creación. Aún esto, podemos ver una base simple: Una sucesión de capas con función de activación relu (convierte negativos a 0) acabando en una capa con función de activación sigmoid, convirtiendo el resultado en una probabilidad (rango de 0 a 1).

Entre las capas de las diferentes estructuras encontramos diferentes técnicas orientadas sobretodo a la generalización, overfitting y optimización del entrenamiento:

- Dropout: Apaga neuronas aleatoriamente. Evitando dependencias a neuronas específicas ayudamos al modelo a generalizar más.
- Batch Normalization: Normaliza la salida de las capas antes de pasar a la siguiente capa.
- Regularización: Penaliza los pesos grandes. Ayuda a que el modelo generalice más, no memorizando sino aprendiendo lo esencial para poder hacer más tarde la clasificación.
- Early Stopping: Detiene el entrenamiento si no hay mejoras en cierto número de épocas. De esta manera, se ahorran recursos y tiempo.
- Label Smoothing: Pasa las etiquetas de 0 o 1 a un valor no tan extremo (0.1 o 0.9 por ejemplo) para que el modelo no se vuelve demasiado confiado.