PROPUESTA DE MONOGRAFÍA

Título del proyecto	**Predicción de ocurrencia de accidentes cerebrovasculares**	
Estudiante 1		
Nombre completo	Andrés Julián Espinal Benjumea	e-mail: ajulian.espinal@udea.edu.co
		GitHub: https://github.com/AndresEspinal
Nombre asesor	Yony Fernando Ceballos	e-mail: yony.ceballos@udea.edu.co

1. Descripción del problema

El ministerio de salud colombiano define a los accidentes cerebrovasculares como "*fenómenos agudos que se deben sobre todo a obstrucciones que impiden que la sangre fluya hacia el cerebro. […] El 60% de los pacientes con accidentes cerebrovasculares mueren o quedan con discapacidad"*. (Ministerio de Salud de Colombia, 2020). Además, la OMS asegura que esta enfermedad es la segunda causa de muerte a nivel mundial[1].

Debido a los datos anteriormente mencionados y a la importancia de conocer las causas y conexiones que llevan a sufrir este accidente, se decide realizar su estudio partiendo de los datos que proporciona kaggle. Estos datos cuentan con variable de salida por lo cual se hará una clasificación de los datos.

2. Descripción del dataset

La base de datos se puede encontrar en kaggle en el siguiente vínculo:

https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

- **id**: Se refiere a un código único que tiene cada paciente.
- **gender**: Aquí se encuentran 3 géneros los cuales son "Male", "Female" y "Other"
- **age**: Hace referencia a la edad del paciente.
- **hypertension**: La base de datos clasifica con 0 si el paciente no tiene hipertensión y con 1 si el paciente sufre hipertensión.
- **heart_disease**: En esta variable se clasifica con 0 si el paciente no tiene ninguna enfermedad cardíaca y con 1 si el paciente padece una enfermedad cardíaca.
- **ever_married**: Esta variable explica si el paciente está casado con "Yes" y con "No" si no lo está.
- **work_type**: Se divide en si trabajó con niños como "children", si obtuvo un trabajo en el gobierno "Govt_jov", si nunca trabajó "Never_worked", si trabajó en el sector privado "Private" o por el contrario trabajó como independiente "Self-employed".
- **Residence_type**: Se divide en si la zona de residencia es rural "Rural" o urbana "Urban".
- **avg_glucose_level:** Dato numérico que mide el nivel promedio de glucosa en sangre.

[1] Datos obtenidos en https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

- **bmi**: Dato numérico que muestra el índice de masa corporal.
- **smoking_status**: Columna que especifica si el paciente ya había fumado anteriormente "formerly smoked", si nunca fumó "never smoked", si en la actualidad fumaba "smokes" o si la información no está disponible para el paciente como "Unknown".
- **stroke**: Es la variable respuesta y nos indica con 1 si el paciente tuvo un accidente cerebrovascular o con 0 si no lo tuvo.

  La base de datos cuenta con 12 columnas y 5.110 filas.

3. Métricas de machine learning:

Debido a que se desea resolver un problema de clasificación, las métricas de regresión que se plantean utilizar son el recall y la precisión, esto es porque estamos hablando de un caso médico y la base de datos cuenta con un desbalance entre 4.87% para registros positivos y 95.13% para registros negativos. Es importante predecir bien los casos positivos que es la condición primordial del ejercicio y por sus pocos registros.
Además, este ejercicio vendrá acompañada con otras medidas como la curva ROC, la exactitud, f1-score para otorgar una mejor clasificación.

4. Métricas de negocio:

Se espera que el modelo arroje una predicción cercana al 85-90% para recall y precisión, dado que se desea clasificar si se va a sufrir de un accidente cerebro vascular, un diagnóstico o duda temprana de un posible caso de accidente cerebrovascular permite ganar tiempo valioso para diagnosticar un posible caso. Con ello se puede llegar a determinar si es necesario aplicar correctivos que permitan salvar la vida del paciente.

5. Referencias

Ministerio de Salud de Colombia. (30 de 10 de 2020). *https://www.minsalud.gov.co/*. Obtenido de https://www.minsalud.gov.co/: https://www.minsalud.gov.co/Paginas/Conozca-como-prevenir-los-accidentes-cerebrovasculares.aspx#:~:text=Los%20accidentes%20cerebrovasculares%20pueden%20da%C3%B1ar,los%20AVC%20prematuros%20son%20prevenibles.

OMS. (09 de 12 de 2020). https://www.who.int. Obtenido de https://www.who.int/: https://www.who.int/es/news-room/fact-sheets/detail/the-top-10-causes-of-death

Ramírez, J. (19 de 07 de 2018). medium.com. Obtenido de medium.com: https://medium.com/bluekiri/curvas-pr-y-roc-1489fbd9a527

sitiobigdata. (27 de 05 de 2019). sitiobigdata.com. Obtenido de sitiobigdata.com: https://sitiobigdata.com/2019/05/27/aprendizaje-automatico-seleccionando-metricas-regresion/#