Machine Learning
Brain Stroke

CODERHOUSE

Data Science

Alumnos:
Damian del Conte
Lucas Vissani

Descripción

Los accidentes cerebrovasculares (ACV o brain stroke en inglès) afectan en el año a más de 15 millones de personas, entre los afectados, 5 millones lamentablemente pierden la vida y otros 5 millones quedan con secuelas permanentes.

En algunas ocasiones se presenta sin antecedentes previos o no está arraigado a ninguna característica social o geopolítica, pero sabemos que algunos factores contribuyen a que estos sucedan.

Objetivo

Teniendo en cuenta diferentes características tanto socioculturales como mèdicas veremos el riesgo potencial de cada individuo en este trabajo y trataremos de evaluar los riesgos de un posible futuro ACV. Serán abordadas diferentes coyunturas de cada individuo, para luego trabajar en base a estas con procesos de ML y de esta manera poder determinar y estudiar el tipo de riesgo individual o grupal.

Exploración de datos

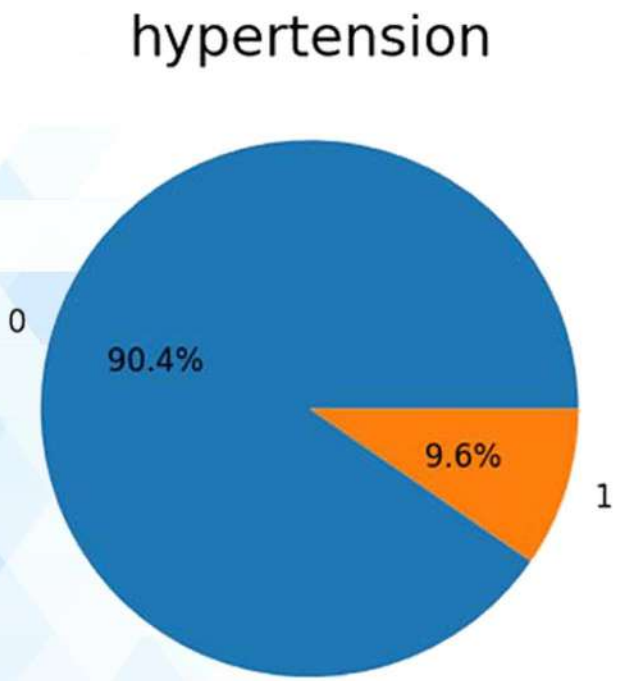
Tomamos un dataset de la página Kaggle, que contiene diferentes ítems relacionados con este estudio.
Es un dataset de uso público y gratuito

https://www.kaggle.com/code/tawfikbrahim/brain-stroke

Detallaremos algunos de los datos más importantes a continuación.

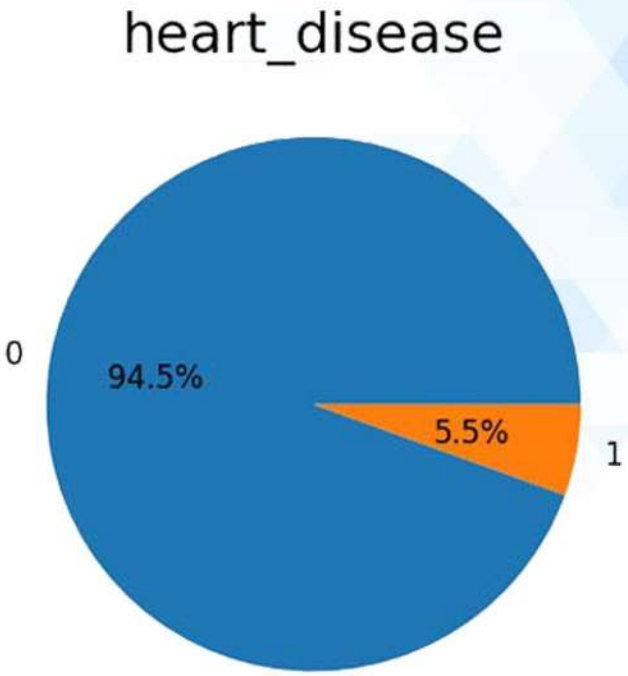
Hipertensión

Este término es el usado para describir la presión arterial alta, en algunos casos esta afección puede llevar a un accidente cerebrovascular, es por esto que es clave para el estudio que estamos realizando. Encontraremos el 0 como representación negativa y el 1 como representación positiva.



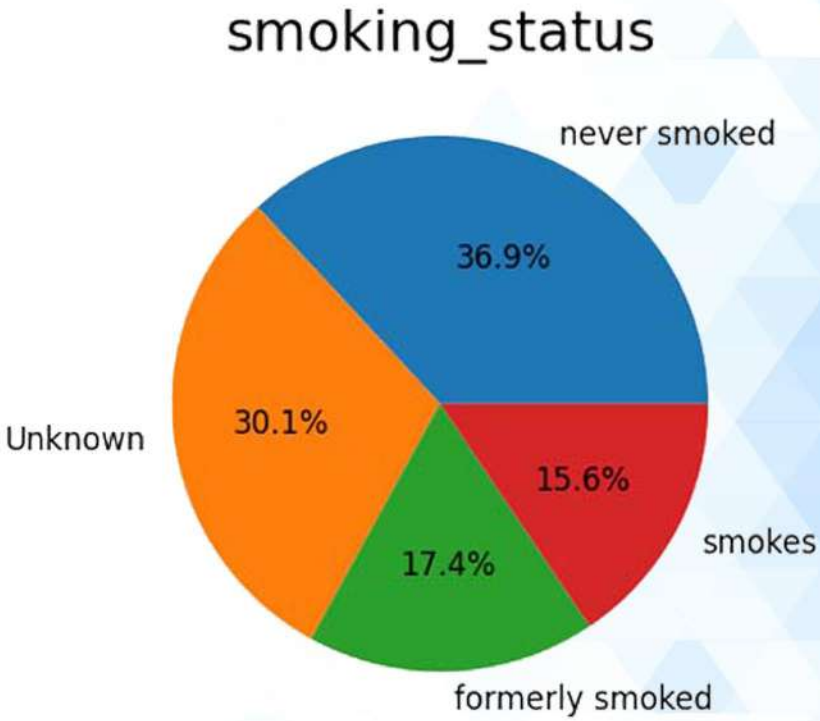
Afecciones cardiovasculares

Muchas afecciones cardiovasculares puede llevar a contraer hipertensiòn, tal como lo vimos antes, estas agravan el riesgo de sufrir un ACV, por lo tanto estudiamos cualquier patologìa que pueda dañar o alterar el funcionamiento cardìaco. Nuevamente encontraremos el 0 como respresentaciòn negativa y el 1 como representación positiva.



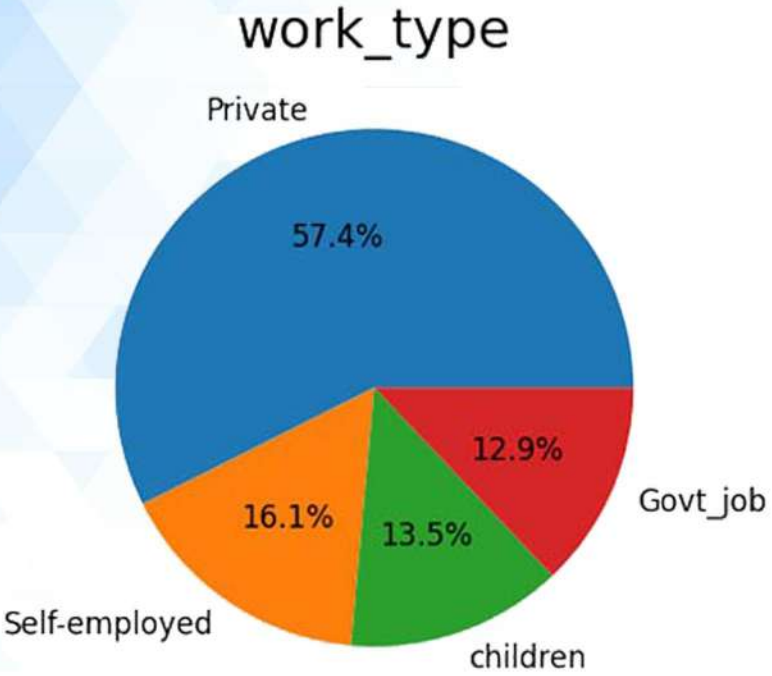
Fumador

El cigarrillo, como bien sabemos, puede producir varias patologías y afecciones. Entre ellas varias de características cardiovasculares que pueden derivar en un ACV.



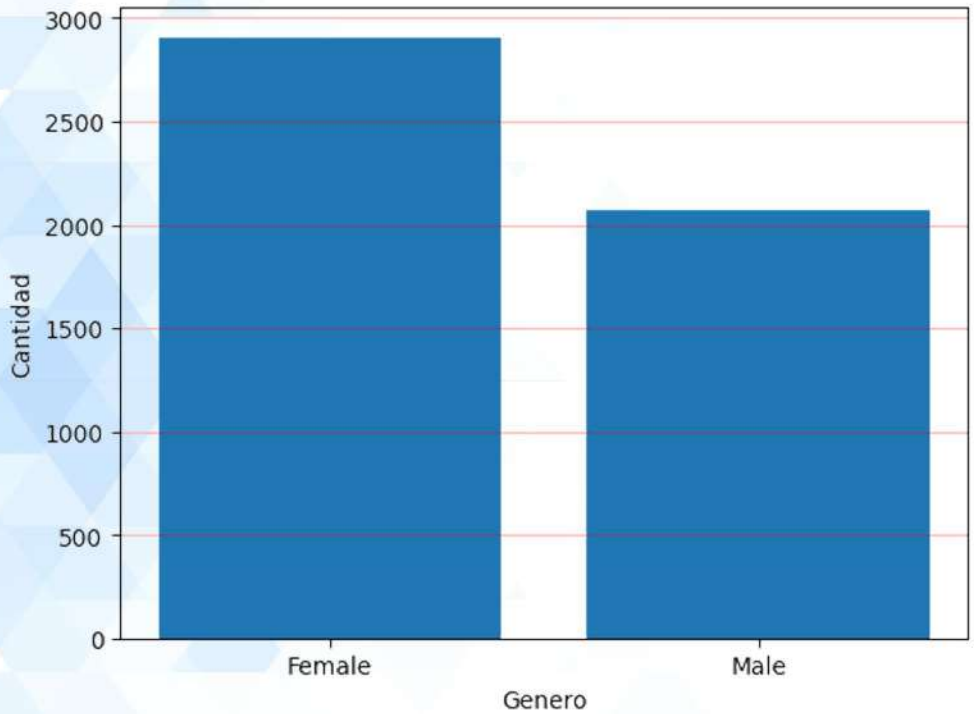
Tipo de trabajo

Este apartado, por más redundante que parezca, tiene que ver con el estrés laboral. El estrés es una de las causas más frecuentes de ACV en el mundo. Los trabajos con altos niveles de estrés suelen provocar o contribuir en el desarrollo de estas afecciones.



Género

El gráfico nos va a mostrar la cantidad de hombres y mujeres estudiados en este proyecto. En cuánto a su redundancia en el mismo, un dato a tener en cuenta es que los accidentes cerebrovasculares son bastante más frecuentes en el género masculino que en el género femenino.



Si bien encontramos que estos son los datos más relevantes, el dataset nos aporta también otros datos que no dejan de ser útiles a la hora de explorar más profundamente las causales de los accidentes cerebrovasculares.
A continuación veremos la tabla con el resto de ítems.

Gender	Hyper tension	Heart disease	Ever married	Work type	Residence type	Smoking status
Male	0 (No)	0 (No)	Yes	Private	Urban	Never smoked
Female	1 (Yes)	1 (Yes)	No	Self employed	Rural	Never smoked
				Children		Smokes
				Govt job		Unknown

Elección del algoritmo

Si bien sabíamos de antemano que la regresión lineal no era adecuada para este estudio, decidimos demostrar esto en el campo de acción. No es la mejor opción para predecir variables binarias como "stroke"(0 y 1), ya que la regresión lineal está más orientada a predecir valores continuos en lugar de clasificaciones binarias.

Los modelos que seleccionamos para este caso fueron Random Forest y regresión Logìstica.

Podemos divisar que ambos modelos arrojaron resultados similares, variando solo en decimales al final. Notaremos que la regresiòn logìstica dió mejores resultados(94% para Random Forest y 95% para regresiòn logìstica en cuánto a l a accuracy).

Podemos notar que utilizando el Label Encoding para transformar el dato de si el paciente fumaba o no se puede observar una mínima mejora en el accuracy del modelo. Agregar este dato al modelo de machine learning no agrega una carga significativa al modelo lo cual nos lleva a pensar que se puede agregar tranquilamente.

Conclusión

-Podemos ver que la hipertensión afecta de manera totalmente directa y exponencial a la formación de accidentes cerebrovasculares.

-El tabaquismo también obtiene injerencia de manera negativa. Visualizamos una suba en el promedio al colocar esta característica.

-Gracias al estudio realizado sobre este dataset en términos tanto analíticos como de Machine Learning podremos utilizarlo tanto para prevención, como para el estudio arraigado a esta patología cada vez más frecuente en nuestra sociedad.