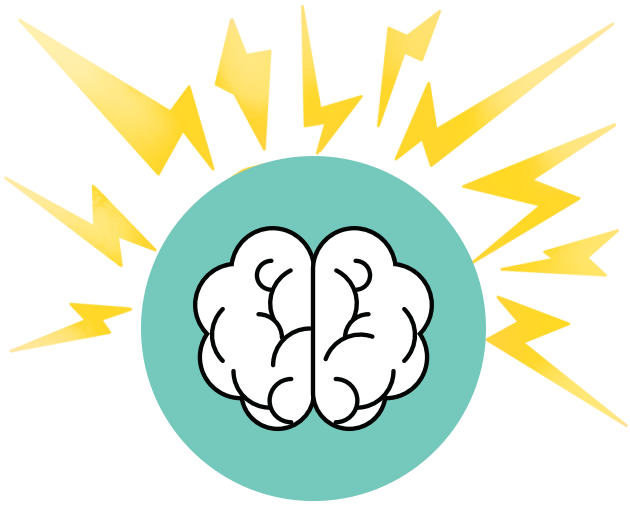
STROKE PREDICTION

Realizado por: Ángeles Torres, Alicia Ortega y Yamileth Cabrera

ÍNDICE

INTRODUCCIÓN

Cuando una parte del cerebro no recibe el oxígeno suficiente puede provocar discapacidades e incluso la muerte



Un Accidente cerebral, ACV, Apoplejía, Ataque cerebral, Derrame cerebral o Infarto cerebral sucede cuando alguno de los vasos sanguíneos que lleva sangre al cerebro se bloquea provocando la muerte de las células cerebrales por la falta de oxígeno.

El cerebro coordina todo lo que hacemos y cuando una parte de él no recibe el oxígeno suficiente, la parte del cuerpo controlada por esa área podría verse afectada provocando discapacidades e incluso la muerte. Entre más prolongado sea el tiempo durante el cuál el cerebro esta privado de oxígeno la discapacidad puede ser más severa y se aumentan las probabilidades de que sea permanente.

Actualmente, diversos organismos médicos a nivel mundial han buscado sensibilizar a la población sobre este padecimiento para que los pacientes puedan recibir el tratamiento adecuado a tiempo. También se enfatiza en la prevención considerando que los factores que pueden detonarlo frecuentemente se deben a hábitos modificables de los pacientes.

DESCRIPCIÓN DEL CASO

En este trabajo se analizó la información de más de 5,000 pacientes para determinar la relación que tiene su estilo de vida y estado de salud con la probabilidad de que tengan un infarto cerebral.

A partir de esta información se diseñó un modelo que permite conocer la probabilidad de que un paciente tenga un infarto cerebral considerando sus características actuales a partir del análisis de variables relacionadas con su historial médico y características personales.

OBJETIVOS

- Analizar la relación del estilo de vida de los pacientes con posibilidad de que tengan un infarto cerebral
- Diseñar un modelo que permita conocer la probabilidad de que los pacientes tengan un infarto cerebral en función de su estilo de vida y estado de salud

DATA ACQUISITION

La información utilizada para este proyecto es información extraída de Kaggle para fines didácticos y no se cuenta con la fuente de origen de la información. Los análisis y resultados del proyecto de acuerdo con el desarrollo del trabajo deben considerarse a discreción.

DESCRIPCIÓN DE LAS VARIABLES

Para cumplir con los objetivos de este trabajo contaremos con la información de 11 variables de relacionadas con características de los pacientes, cómo edad o género, estilo de vida e historial médico. A continuación, se describen cada una de ellas:

Información del paciente

Grupo de edad al que percenece el paciente y Género del paciente

Estado de Salud

Padecimiento de hipertensión, Antecedentes de enfermedades del corazón, Nivel de glucosa e Índice de Masa Corporal

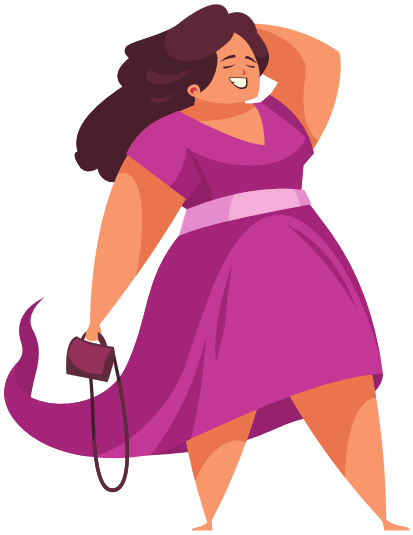
Estilo de vida

Estado civil, Tipo de trabajo, Tipo de residencia y Hábito de Fumar

DATA WRANGLING

En esta sección se analizó y adapto la información para poder utilizar posteriormente como fuente de datos para el modelo. Para ello se reviso la estructura y tamaño del dataset, se manejaron los datos nulos y se modelo la información de acuerdo a las características de cada variable

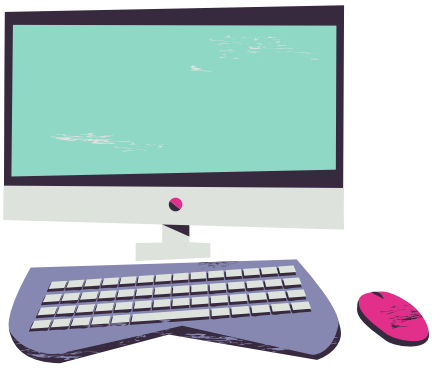
SOBRE LOS PACIENTES



- Contamos con información de 5,100 pacientes
- El promedio de edad de los pacientes es de 43 años. Desde niños menores de un año hasta adultos mayores de 82 años
- El promedio de masa corporal de los pacientes es de 29 lo cual corresponde a sobrepeso. Sin embargo, la desviación estándar es alta.
- El promedio de nivel de glucosa de los pacientes es de 109 lo que se considera normal. Sin embargo, la desviación estándar es muy grande.

SOBRE EL DATASET

- Contamos con información de 11 variables que describen el estilo de vida y estado de salud de los pacientes
- La única variable con datos faltantes era el índice de masa corporal
- La edad es una variable de tipo flotante y para pacientes menores a dos años tenia decimales
- No hay información duplicada en el dataset



ANÁLISIS EXPLORATORIO DE DATOS (EDA)

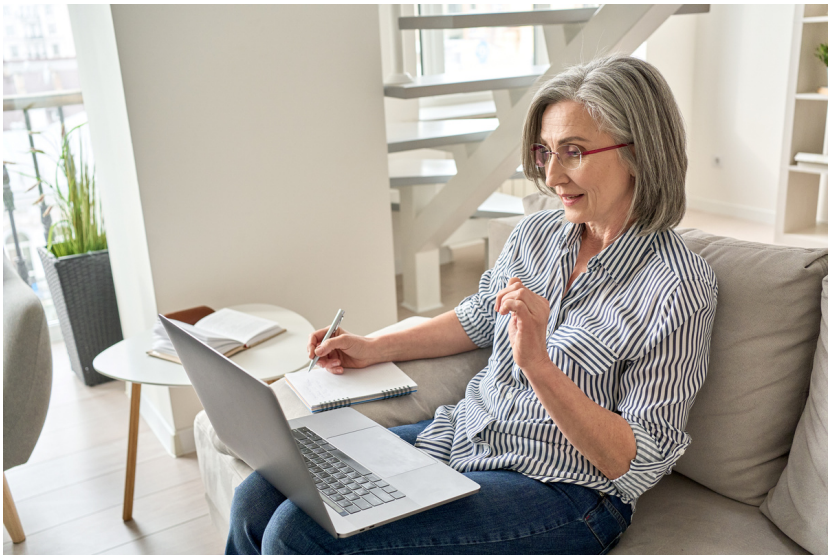
El Análisis Exploratorio de Datos nos permite examinar la estructura de los datos, así como la relación que existe entre las variables permitiendo detectar datos atípicos, manejar datos ausentes y dar forma al dataset de acuerdo con el objetivo del trabajo. Para realizar este análisis se examinaron los siguientes puntos:

- Outliers
- Datos faltantes
- Correlaciones
- Distribución de las variables

Para ello realizamos 3 tipos de análisis de las variables: indivualmente, en comparación con la variable objetivo y entre ellas que corresponden al análisis Univariado, Bivariado y Multivariado respectivamente.

ANÁLISIS UNIVARIADO

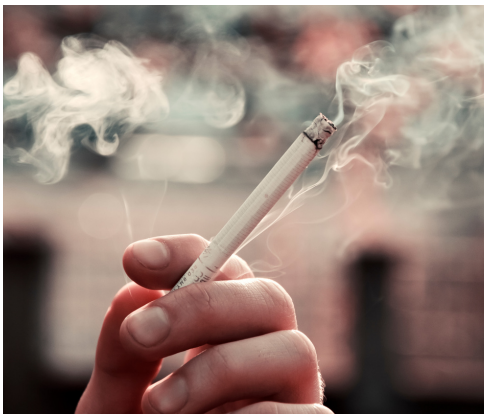
La muestra está compuesta mayormente por mujeres mayores de 50 años. La mayoría están casadas y trabajan en el sector privado. Sobre su estado de salud la mayoría de los pacientes no han sufrido algún ACV en el pasado, tienen sobrepeso, sus niveles de glucosa son normales, no tienen enfermedades del corazón ni hipertensión.



ÁNALISIS BIVARIADO

- El análisis de nuestra muestra corrobora lo que dice la información médica sobre el aumento en los casos con antecedentes de ACV de pacientes con problemas cardíacos, diabetes, colesterol alto y el tabaquismo .

- Algunos estudios sugieren que los hombres presentan mayores casos de infartos cerebrales, sin embargo en nuestra muestra el género no influye en el número de casos



- Uno de los factores por los que se incrementan los casos de infartos cerebrales es por el tabaquismo. En este caso vemos que efectivamente aquellos pacientes que han fumado en el pasado presentan mayor número de casos

ÁNALISIS MULTIVARIADO

- Uno de los factores de riesgo de tener un infarto cerebral es la edad del paciente, lo cual se confirma en nuestra muestra.

- Cuando los pacientes tienen más de una enfermedad al mismo tiempo los casos de infarto cerebral disminuyen. Esto podría confirmar que lo que influye más en la probabilidades de un infarto cerebral esta más relacionado con el estilo de vida del paciente que con sus enfermedades preexistentes

- Cuando consideramos más de un aspecto en el estilo de vida de los pacientes los casos de antecedentes de infarto cerebral aumentan. Específicamente se incrementan para aquellos pacientes que están casados, los que nunca han fumado y los que viven en zonas urbanas.

ELECCIÓN DEL MODELO

Para este proyecto debido a que nuestra variable target es de tipo categórica nos enfrentamos ante un problema de clasificación para poder predecir la clase más probable de un elemento en función del conjunto de variables que tenemos. En este caso específicamente buscamos predecir la probabilidad de que un paciente tenga un ACV.

Para poder encontrar el modelo que mejor predijera la probabilidad de que un paciente tuviera un ACV se entreno una muestra de 2 mil pacientes y se probaron 7 técnicas de modelaje: Árbol de decisión, Regresión Logística, Random Forest, Adaboost, Gradient Boosting, Light GBM y Xgboost

Las metricas de cada modelo fueron:

Modelo	Exactitud	Precisión	Sensibilidad	Score F1	AUC
Árbol de decisión	77.86	73.73	87.63	79.91	0.8216
Regresión Logística	89.34	91.18	87.22	89.15	0.9579
Random Forest	93.90	93.12	94.81	93.96	0.9806
Adaboost	85.92	83.49	89.54	86.41	0.9321
Gradient Boosting	86.17	85.35	87.32	86.33	0.9337
Light GBM	93.99	92.72	95.47	94.07	0.9796
Xgboost	93.95	92.85	95.22	94.02	0.9802

Modelo	Verdaderos Positivos	Falsos Negativos
Árbol de decisión	1,070	151
Regresión Logística	1,065	156
Random Forest	1,152	63
Adaboost	1,088	127
Gradient Boosting	1,061	154
Light GBM	1,060	55
Xgboost	1,157	56

CONCLUSIONES

En conclusión, los pacientes con problemas cardíacos, diabetes, colesterol alto y el tabaquismo son los que tienen más antecedentes de ACV. Las variables que más influencia tienen son la edad del paciente, antecedentes de enfermedades del corazón o hipertensión y su estado civil.

Por otro lado, es que disminuyen las probabilidades de que alguien pueda sufrir un ACV son cuándo son jóvenes y tienen niveles normales de glucosa

Después de revisar las métricas de certeza de cada uno de los modelos se eligió a Light GBM por obtener los mejores niveles. El modelo podrá:

Certeza en la predicción vs equivocaciones

La proporción de casos en los que el modelo acertó que la paciente tenía probabilidad de tener un ACV y realmente su historial así lo demostraba vs los pacientes en los que el modelo lo predijo, pero fue erróneo es de **95%**. Eso significa que el **95.47%** de los casos que fueron detectados por el modelo realmente eran casos con antecedentes de ACV

Porcentaje de predicciones correctas

En el **92.72%** de las predicciones que el modelo hace son correctas, incluyendo predicciones de pacientes que podrían tener un ACV como aquellos que no tienen probabilidad

Predicciones correctas de pacientes con probabilidad de tener un ACV

De los 2 mil pacientes en los que el modelo entreno, en el **47.73%** de los casos el modelo acertó aquellos el paciente que podría tener un infarto cerebral y efectivamente tenían el antecedente de ACV