PREVENCIÓN DE ENFERMEDAD ACV

Contenidos

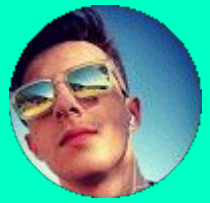
01

Empresa

Equipo

Martín Marino Uviedo

https://www.linkedin.com/in/mmarinouviedo/



Leandro Bruzzo

www.linkedin.com/in/leandro-bruzzo

Como parte de la consultora Uviedo-Bruzzo fuimos contactados por el Gobierno de la Ciudad de Buenos Aires con el objetivo de realizar un análisis que permita trabajar en la prevención de los ataques cerebrovasculares.

02

Temática y Objetivos generales

Temática

11%

De acuerdo a la Organización Mundial de la Salud (OMS), el ACV es la segunda causa de muerte a nivel global, responsable de aproximadamente **11%** de las muertes totales.

77%

El **77%** de los ACV corresponden a un primer evento, lo que pone de manifiesto la importancia de la prevención primaria.

Temática

8.6%

En Argentina, un estudio realizado en Junín indicó una prevalencia de 868 casos por cada 100 mil habitantes.

5%

Menos del 5% de los pacientes en el país consultan en el tiempo adecuado para recibir el tratamiento indicado.

Objetivo

Nuestro objetivo como organización es trabajar en identificar los factores de riesgo que nos permitan una detección temprana de posibles casos de accidentes cerebrovasculares.

A través de distintos indicadores con los que contamos en nuestra muestra, vamos a implementar un modelo de machine learning capaz de colaborar en el proceso preventivo de detección temprana.

03

Análisis del Dataset

Dataset

El dataset fue obtenido de la plataforma Kaggle desde el siguiente link : https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Información sobre el data set:
- Creador: fedesoriano - https://www.kaggle.com/fedesoriano

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1

Atributos

El dataset contiene una serie de atributos que nos brindarán información relevante sobre el paciente; desde su estado marital, nivel de glucosa en sangre, índice de masa corporal hasta si sufre de hipertensión o si sufrió un ACV

- **id**: identificador único
- **gender**: "Male", "Female" o "Other" (Masculino, Femenino u Otro)
- **age**: Edad del paciente
- **hypertension**: 0 si el paciente no tiene hipertensión , 1 si el paciente tiene hipertensión
- **heart_disease**: 0 si el paciente no tiene enfermedades del corazón, 1 si el paciente tiene enfermedades del corazón
- **ever_married**: "No" o "Yes" (No o Sí)

Atributos

- **work_type**: "children", "Govt_jov", "Never_worked", "Private" or "Self-employed" (Niño, trabajo en el sector público, nunca trabajó, privado o monotributista)
- **Residence_type**: "Rural" or "Urban"
- **avg_glucose_level**: nivel promedio de glucosa en la sangre
- **bmi**: Índice de masa corporal
- **smoking_status**: "formerly smoked", "never smoked", "smokes" or "Unknown"* (Ex fumador, nunca fumó, fuma, desconocido)
- **stroke: 1 si el paciente tuvo un ACV, 0 si no lo tuvo (Variable Target)**

Data Wrangling & EDA

Data Wrangling

Antes de continuar con el EDA y aplicaciones de modelos de Machine Learning, es necesario comprobar la utilidad del set de datos; si hay datos nulos, datos fuera de escala, con errores de escritura y eliminar variables innecesarias en caso de ser necesario.

Data Wrangling & EDA

Para preparar el set de datos de cara a los modelos de ML, hemos realizado unas tareas de limpieza y normalización de las variables para una mejor aplicación.

Formateamos los valores de las columnas residence_type y ever_married a tipo booleano (Es decir, valores binarios 0 y 1) para una mejor performance del modelo y ahorro de memoria.

ever_married	work_type	Residence_type
Yes	Private	Urban
Yes	Self-employed	Rural
Yes	Private	Rural
Yes	Private	Urban
Yes	Self-employed	Rural

ever_married	work_type	residence_type
1	Private	0
1	Self-employed	1
1	Private	1
1	Private	0
1	Self-employed	1

Data Wrangling & EDA

En la variable smoking_status presentamos valores ‘Unknown’, dado que no podemos asumir si esa persona es fumadora o no, junto a su frecuencia; se ha decidido eliminar los datos sacrificando cantidad por calidad.

```
#Vemos los valores unicos para la columna de smoking status y posteriormente excluimos los desconocidos
df['smoking_status'].unique()
```

✓ 0.5s

```
array(['formerly smoked', 'never smoked', 'smokes', 'Unknown'],
      dtype=object)
```

```
df.smoking_status.value_counts()
```

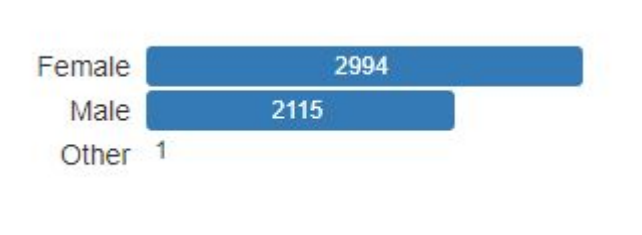
✓ 0.4s

```
never smoked       1892
Unknown            1544
formerly smoked     885
smokes              789
Name: smoking_status, dtype: int64
```

Como se puede observar, la categoría ‘unknown’ representa una gran cantidad de nuestro set de datos

Data Wrangling & EDA

En la variable gender, presentamos datos de géneros male, female y other; éste último sólo presenta un registro, por lo tanto será eliminado para conservar una sola columna a la hora de convertir los datos a binarios (valores más eficaces para los algoritmos). En el caso de que en el futuro se ingresen datos pertenecientes a esta categoría de forma significativa, los mismos serán incluidos en el algoritmo.



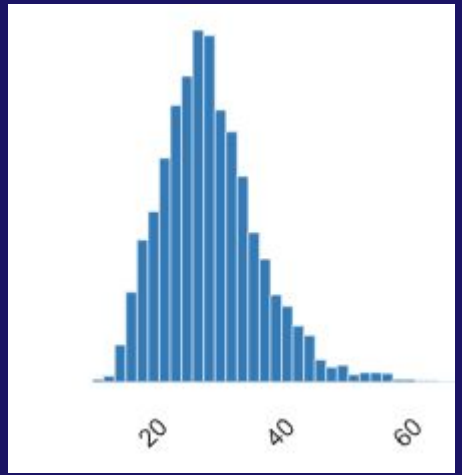
Asimismo, se puede observar la buena distribución de datos en los géneros female y male. Esto es crucial ya que si no hubiera dicha distribución, el algoritmo predeciría mejor un género frente a otro.

Data Wrangling & EDA

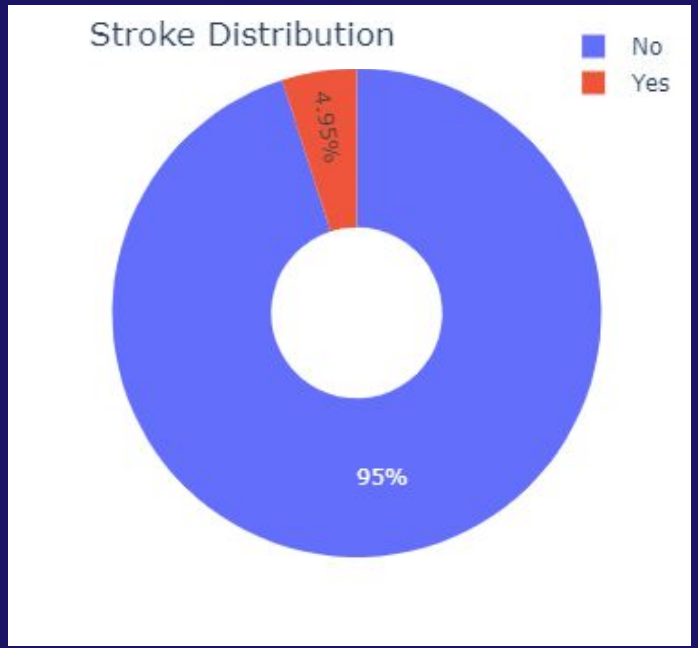
La columna BMI (índice de masa corporal) presenta una cantidad de 140 datos nulos, dado que la cantidad de datos se vio reducida considerablemente al eliminar el estado de fumador ‘unknown’, se procedió a completar dichos valores NaN con el promedio de los mismos. Manteniendo así la mayor cantidad posible con la mejor calidad posible.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3566 entries, 0 to 5108
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 3566 non-null   int64  
 1   gender             3566 non-null   object 
 2   age                3566 non-null   float64
 3   hypertension       3566 non-null   int64  
 4   heart_disease      3566 non-null   int64  
 5   ever_married       3566 non-null   int64  
 6   work_type          3566 non-null   object 
 7   residence_type     3566 non-null   int64  
 8   avg_glucose_level  3566 non-null   float64
 9   bmi                3566 non-null   float64
 10  smoking_status     3566 non-null   object 
 11  stroke             3566 non-null   int64  
dtypes: float64(3), int64(6), object(3)
memory usage: 362.2+ KB
```

Luego de los filtros y formateos aplicados, el dataset presenta 3566 datos válidos, sin nulos.
Estos están en la etapa final de limpieza y en este punto podemos comenzar con el EDA para obtener respuestas y observar posibles patrones en los datos.



Exploratory Data Analysis

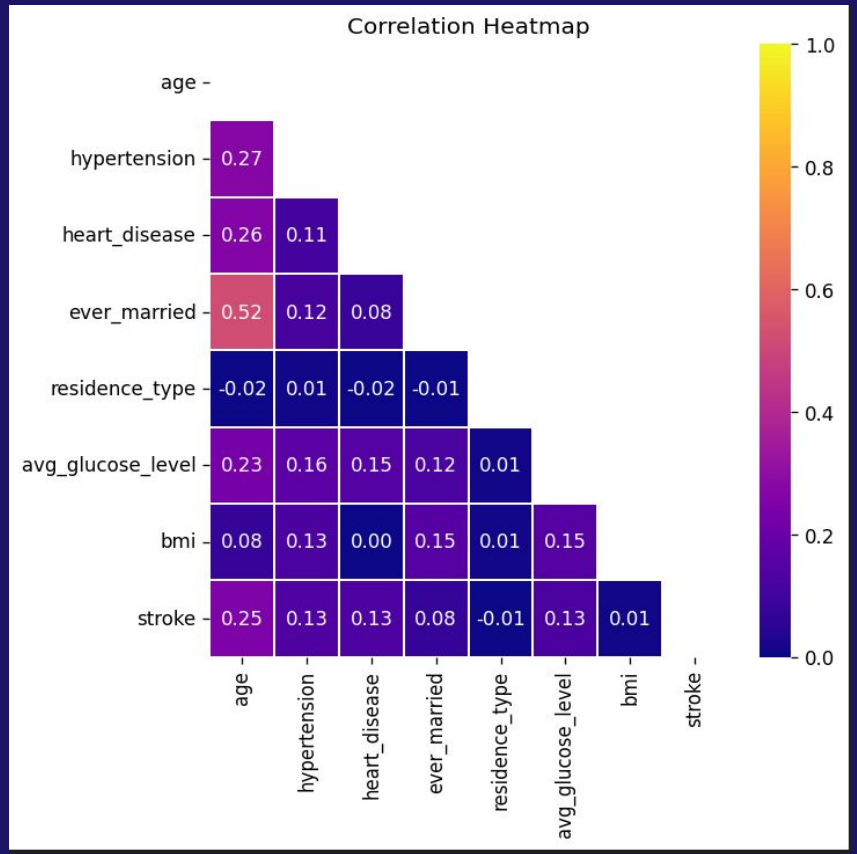


Para comenzar con el EDA, hemos realizado un gráfico de torta para observar la proporción del dataset entre personas que han sufrido un ACV y quienes no.

Como se observa en el gráfico, existe una desproporción sumamente importante.

Esto es crucial a la hora de aplicar los modelos de ML, ya que estos requerirán de configuraciones específicas para tratar estos datos desbalanceados.

Exploratory Data Analysis



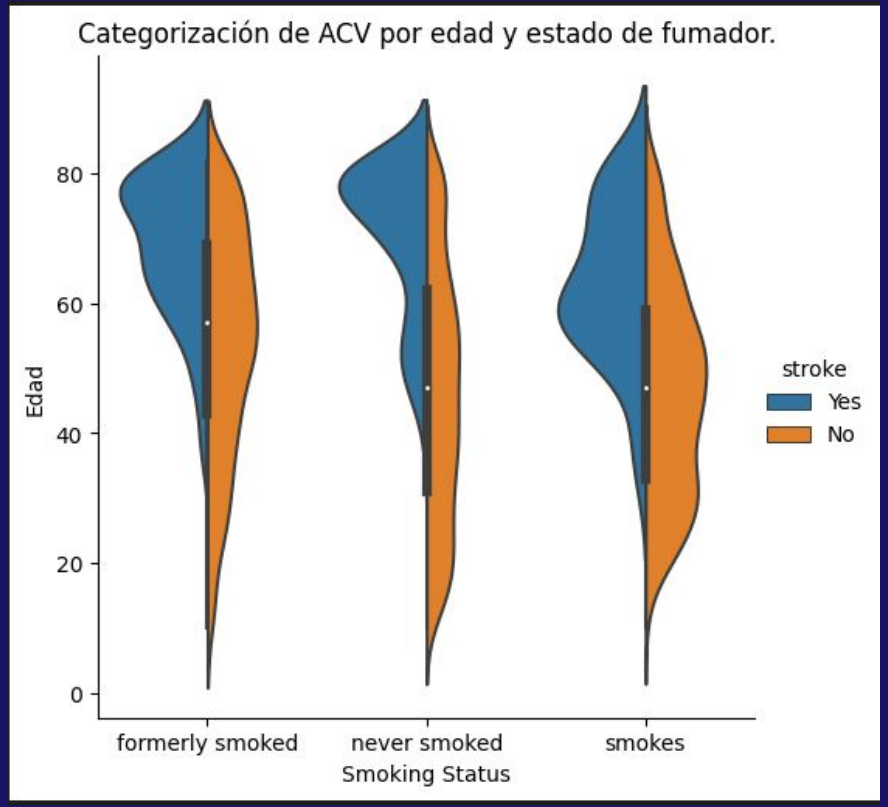
Realizamos un mapa de correlación para observar tendencias entre las variables.

Se observa una correlación negativa de la variable residence_type con respecto al resto de los datos, por lo tanto la misma será eliminada.

Asimismo, se puede observar como ever_married y age tienen un gran coeficiente, pero este no es de valor ya que sólo indica que a mayor edad, más gente presenta matrimonio.

Sin embargo, podemos observar correlación entre la edad y padecer hipertensión, problemas de corazón, un mayor nivel de glucosa en sangre y sufrir un ACV

Exploratory Data Analysis



Decidimos realizar un gráfico de violín para observar la distribución de stroke en base a la edad y el estado de fumador.

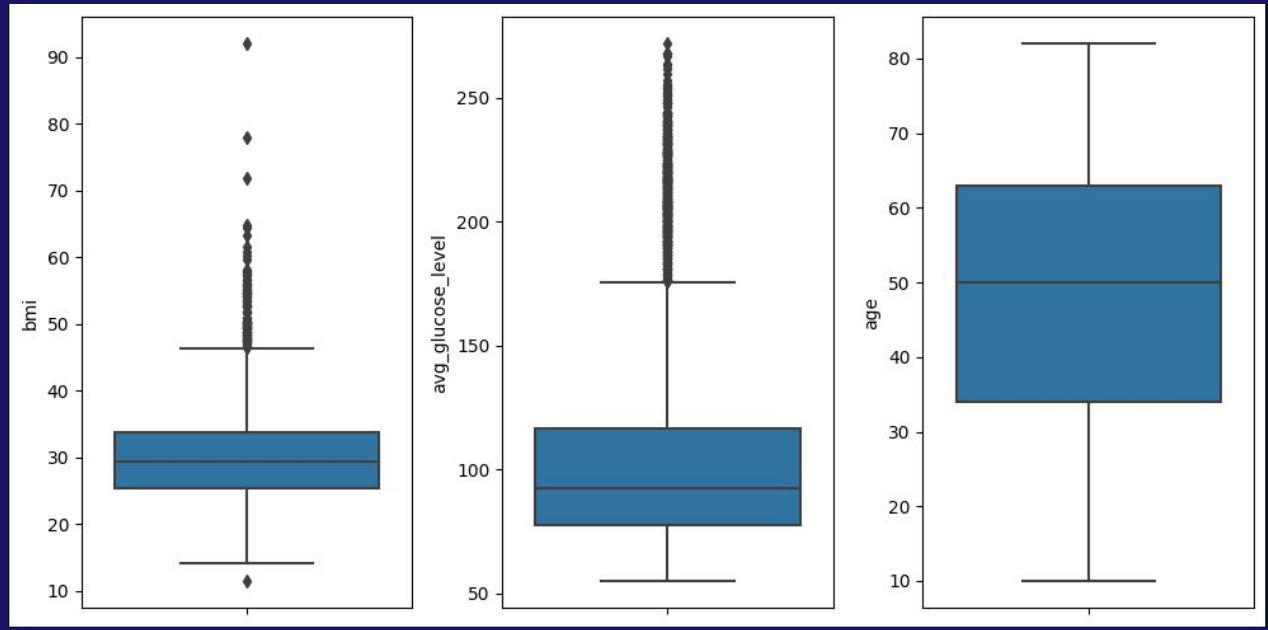
Se puede observar que a mayor edad si el estado de fumador es alto, se es más propenso a sufrir un ACV frente a quien no fuma, o lo hace a un nivel mucho más moderado.

Exploratory Data Analysis

Gráfico de boxplot para observar los outliers presentados en las variables bmi, avg_glucose_level y age.
Se pueden observar los valores atípicos como así también la mediana de los mismos.

Mediana age: 50
Mediana glucose: 90
Mediana bmi: 29

Si bien tanto bmi como avg_glucose_level tienen una gran cantidad de valores atípicos, estos se conservarán ya que son datos válidos.



04

Implementación de modelos ML

Selección de algoritmos

Teniendo en cuenta la característica de nuestros datos y el objetivo de nuestro análisis, nos inclinamos por un algoritmo de clasificación y uno de regresión que nos permita entender si una persona es más propensa a sufrir un ACV.

Con esta lógica, probamos cuatro algoritmos distintos:

- Logistic Binary Regression
- Random Forest
- KNN
- Decision tree

Si bien buscaremos aplicar todos los modelos, seleccionaremos el más apto para nuestro objetivo.

Selección de algoritmos

Para lograr un rendimiento óptimo de los mismos, se utilizó GridSearchCV para la mejor opción de hiper parámetros; pudiendo así obtener el mejor rendimiento de los mismos.
Se ha aplicado el parámetro class_weight= “balanced” en los algoritmos compatibles para poder trabajar con un set de datos desbalanceado. Esto normalmente hace que los resultados tengan un puntaje más bajo, pero se trabaja de una mejor forma.
El fin es poder predecir ambos casos, tanto como si va a sufrir un ACV, o si no.

Métrica de evaluación

La métrica que elegimos para evaluar el modelo es:

Recall → `tp / (tp + fn)`

Teniendo en cuenta que nuestro objetivo es determinar si una persona tiene riesgo de sufrir un ACV, decidimos optar por esta métrica que nos ayuda a clasificar todas las muestras positivas.

Resultados


Random forest
Weighted avg
0.75
Decision tree
Weighted avg
0.63
Logistic Regression
Weighted avg
0.74

Conclusiones

Como resultado de la investigación realizada a partir de los datos proporcionados, recomendamos implementar las siguientes medidas:

1. Trabajar en campañas de prevención, especialmente apuntadas a personas con factores de riesgo tales como: bmi alto, glucosa alta, hipertensión.
2. Concientizar sobre las consecuencias negativas de ser un fumador activo y el aumento de posibilidades de sufrir un ACV que esto conlleva.
3. Implementar el modelo de ML en hospitales públicos con el fin de generar una detección temprana de este tipo de patologías.

¡Gracias!