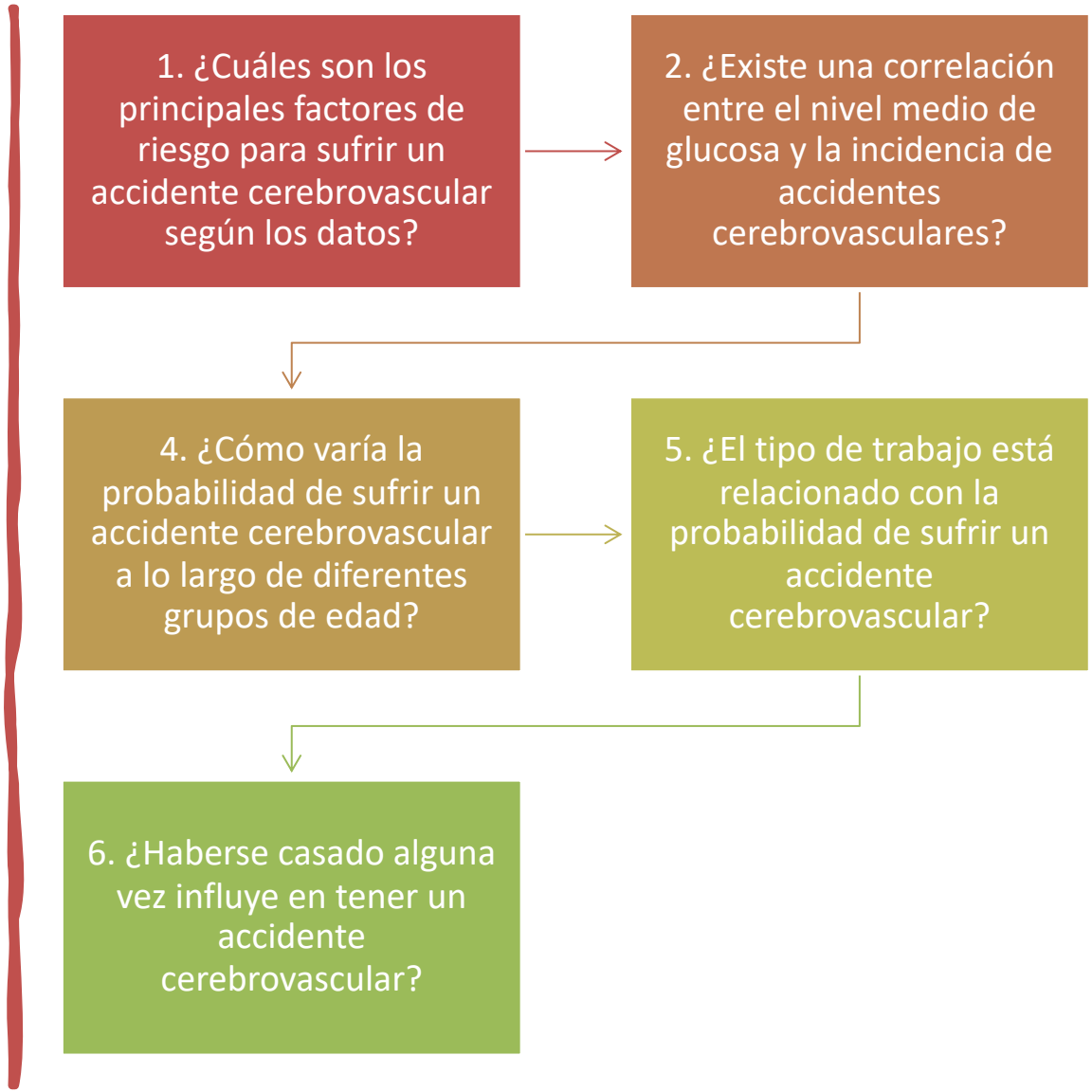
Análisis de Factores Contribuyentes a Accidentes Cerebrovasculares

Presentado por Carlos Hernandez Santes

Objetivo del Proyecto

- Explorar datos relacionados con accidentes cerebrovasculares para identificar factores de riesgo significativos.

Preguntas de Investigación



Análisis Exploratorio de Datos (EDA)

Distribución de Variables Categóricas: Se realizó un análisis de la distribución de las variables categóricas en el conjunto de datos. Esto incluye la visualización y el cálculo de porcentajes de diferentes categorías en variables como el género, el estado civil y el tipo de trabajo.

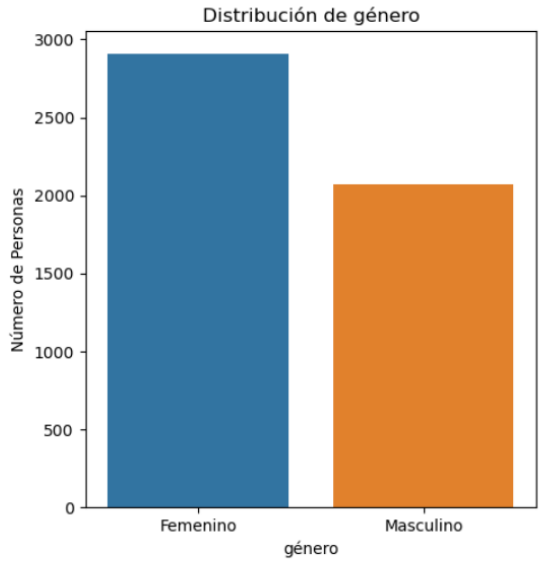
Observaciones de Variables Numéricas: Se llevó a cabo un análisis detallado de las variables numéricas, posiblemente incluyendo estadísticas descriptivas y visualizaciones como histogramas o gráficos de caja para entender la distribución de estas variables.

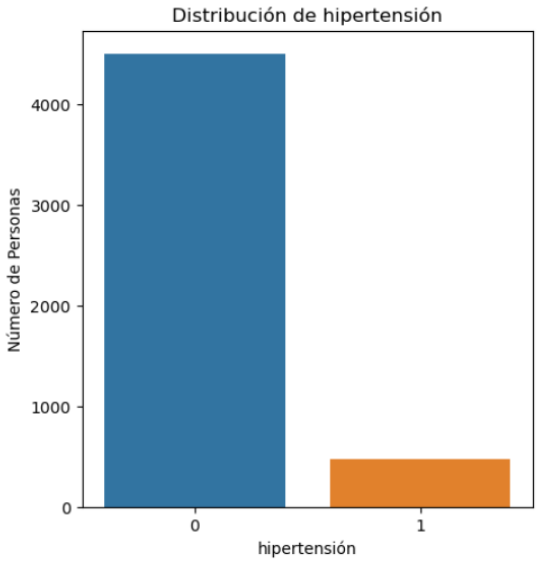
Relación entre Variables y Accidentes Cerebrovasculares: Se exploró cómo diferentes variables pueden estar relacionadas con la incidencia de accidentes cerebrovasculares. Esto podría haber incluido análisis bivariados o correlaciones.

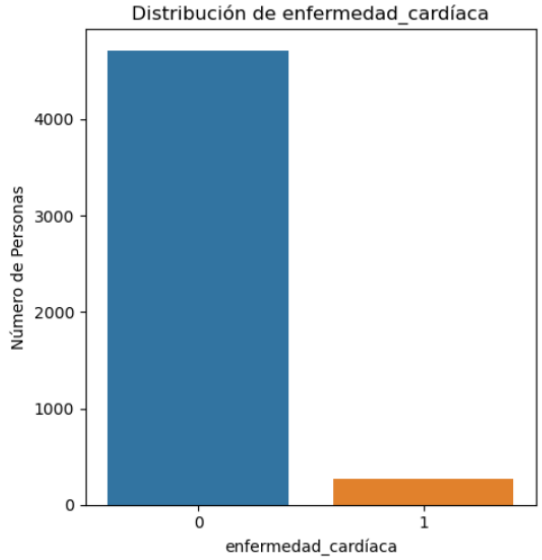
Gráficos de Densidad: Se utilizaron gráficos de densidad para observar la distribución de variables específicas, posiblemente comparando grupos con y sin antecedentes de accidente cerebrovascular.

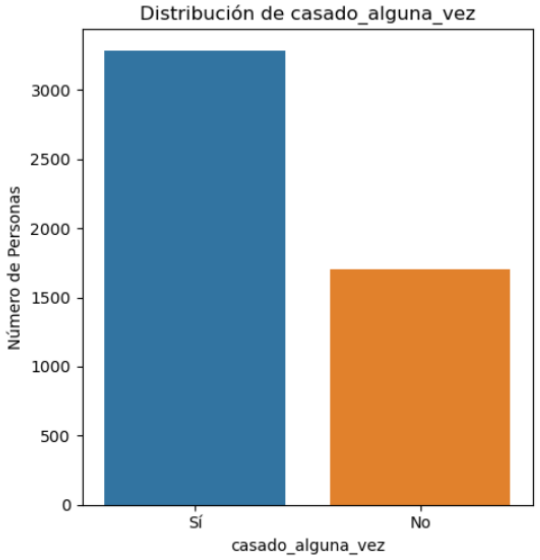
Estadísticas categoricas resumidas

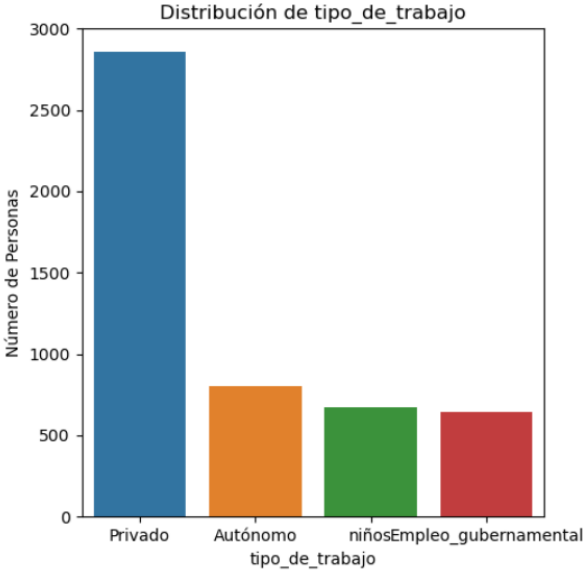
- Género: 2 valores únicos (Hombre Y Mujer)
- Casado alguna vez: 2 valores únicos (Sí, No)
- Tipo de Trabajo: 4 tipos únicos (Privado, trabajador del gobierno, Autoempleado, Ama de casa cuidando a los hijos)
- Tipo de Residencia: 2 tipos únicos (Urbano, Rural)
- Estado de Fumar: 4 estados únicos (Deconocido, Fumador, Exfumador, nunca fumo)

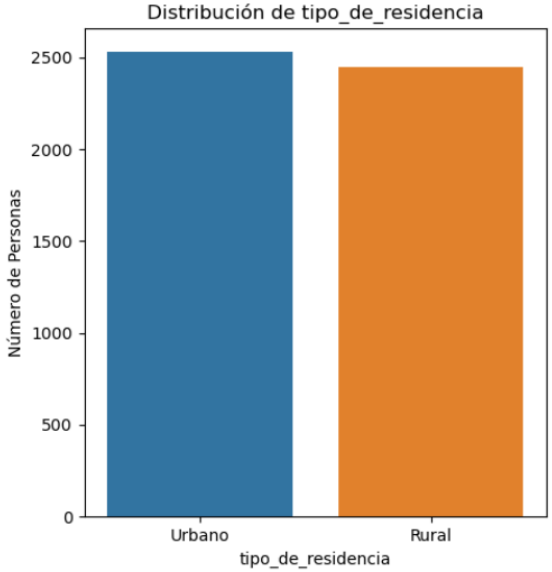

Distribución de género
Número de Personas
3000
2500
2000
1500
1000
500
0
Femenino
Masculino
género

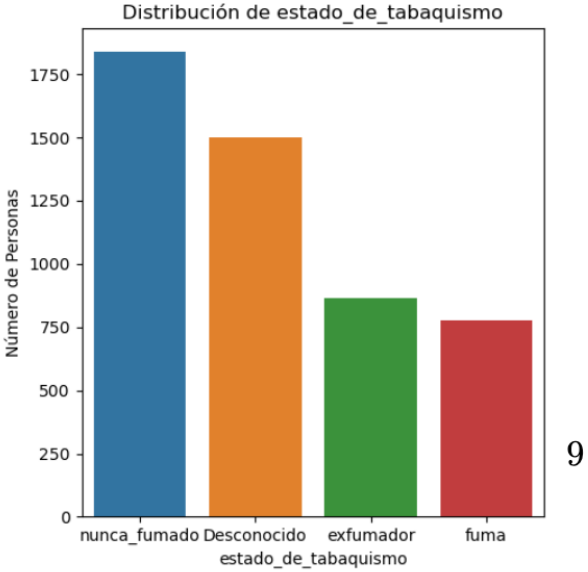

Distribución de hipertensión
Número de Personas
4000
3000
2000
1000
0
0
1
hipertensión

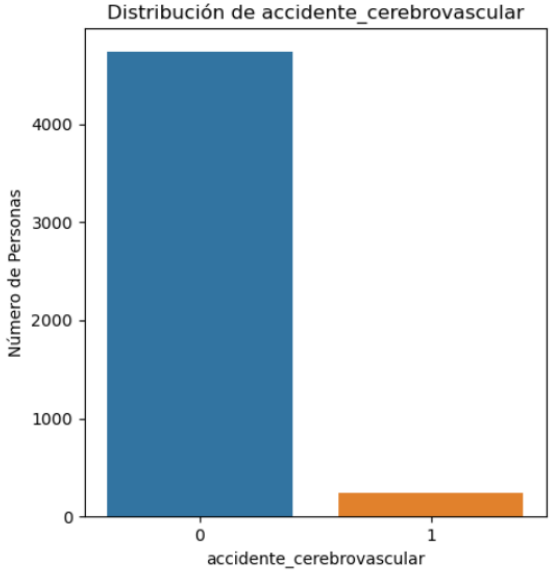

Distribución de enfermedad_cardíaca
Número de Personas
4000
3000
2000
1000
0
0
1
enfermedad_cardíaca


Distribución de casado_alguna_vez
Número de Personas
3000
2500
2000
1500
1000
500
0
Sí
No
casado_alguna_vez


Distribución de tipo_de_trabajo
Número de Personas
3000
2500
2000
1500
1000
500
0
Privado
Autónomo
niños
Empleo_gubernamental
tipo_de_trabajo


Distribución de tipo_de_residencia
Número de Personas
2500
2000
1500
1000
500
0
Urbano
Rural
tipo_de_residencia


Distribución de estado_de_tabaquismo
Número de Personas
1750
1500
1250
1000
750
500
250
0
nunca_fumado
Desconocido
exfumador
fuma
estado_de_tabaquismo

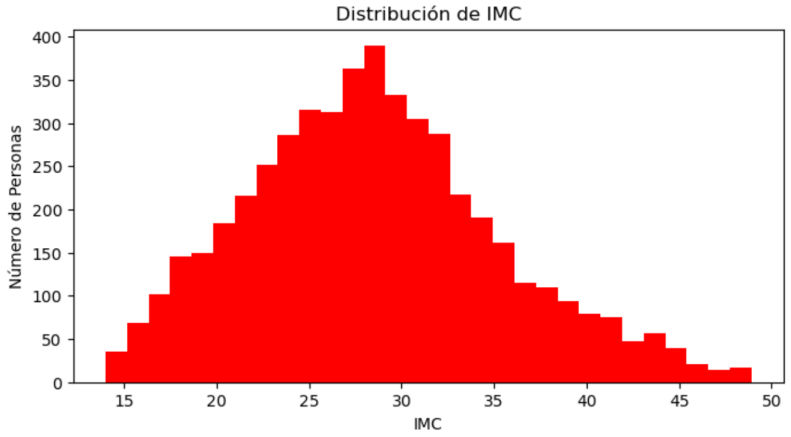

Distribución de accidente_cerebrovascular
Número de Personas
4000
3000
2000
1000
0
0
1
accidente_cerebrovascular

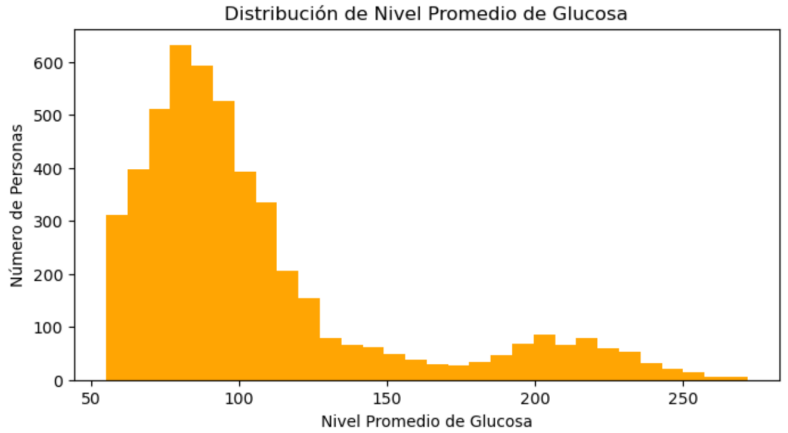
Consideraciones para Modelos de Aprendizaje Automático

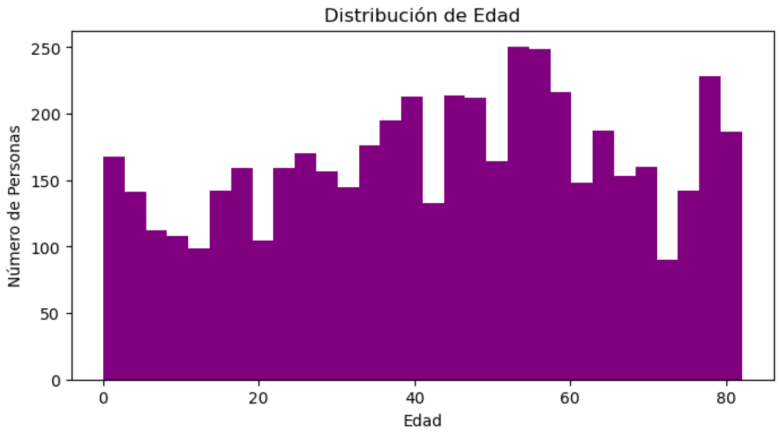
Uno de los principales desafíos con este conjunto de datos es el desequilibrio en la variable objetivo (stroke). Un número significativamente menor de individuos han experimentado un derrame cerebral en comparación con los que no lo han hecho. Este desequilibrio podría llevar a un modelo que está sesgado hacia la clase mayoritaria, resultando en un rendimiento deficiente en la identificación de casos de derrame cerebral.

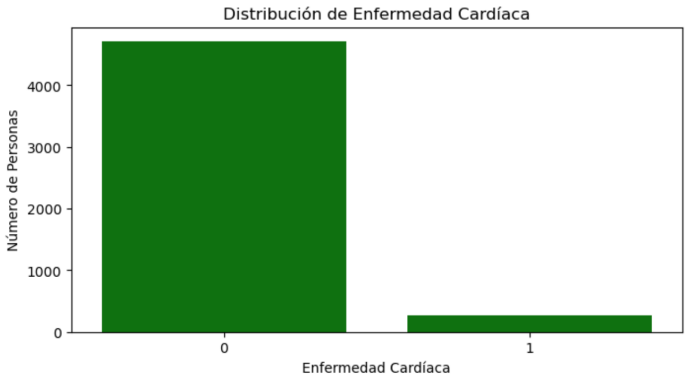
Estadisticas numericas resumidas

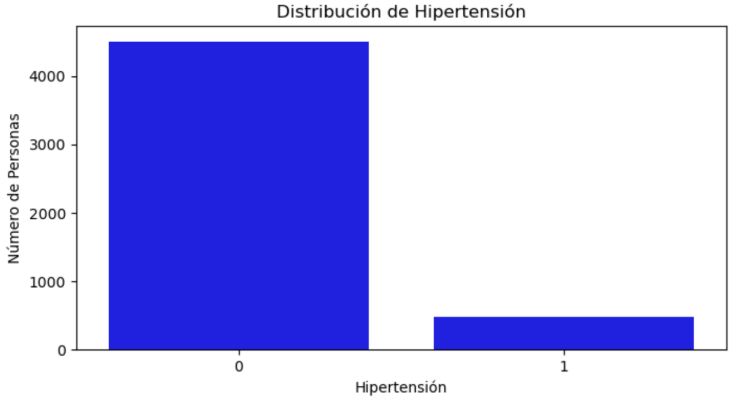
- Edad: Varía desde 0.08 hasta 82 años, con un promedio de aproximadamente 43.42 años.
- Hipertensión: Alrededor del 9.6% de los individuos tienen hipertensión.
- Enfermedad Cardíaca: Alrededor del 5.5% de los individuos tienen enfermedad cardíaca.
- Nivel Promedio de Glucosa: Varía desde 55.12 hasta 271.74, con un promedio de aproximadamente 105.94.
- IMC: Varía desde 14 hasta 48.9, con un promedio de aproximadamente 28.5.


Distribución de IMC
Número de Personas
IMC
0
50
100
150
200
250
300
350
400
15
20
25
30
35
40
45
50


Distribución de Nivel Promedio de Glucosa
Número de Personas
Nivel Promedio de Glucosa
0
100
200
300
400
500
600
50
100
150
200
250


Distribución de Edad
Número de Personas
Edad
0
50
100
150
200
250
0
20
40
60
80


Distribución de Enfermedad Cardíaca
Número de Personas
0
1000
2000
3000
4000
0
1
Enfermedad Cardíaca


Distribución de Hipertensión
Número de Personas
0
1000
2000
3000
4000
0
1
Hipertensión

Observaciones Detalladas de Variables Numéricas

Edad: Aunque la mayoría de los individuos están en el rango de 40 a 60 años, hay una representación significativa de todas las edades. Esto podría tener implicaciones en cómo la edad se relaciona con el riesgo de derrame cerebral.

Hipertensión: La minoría de los individuos en el conjunto de datos tiene hipertensión. Esto plantea preguntas sobre cómo esta variable podría interactuar con otras para influir en el riesgo de derrame cerebral.

Enfermedad Cardíaca: Similar a la hipertensión, la enfermedad cardíaca es poco común pero podría ser un factor crítico. Sería interesante examinar cómo esta variable se relaciona con la edad y otros factores de riesgo.

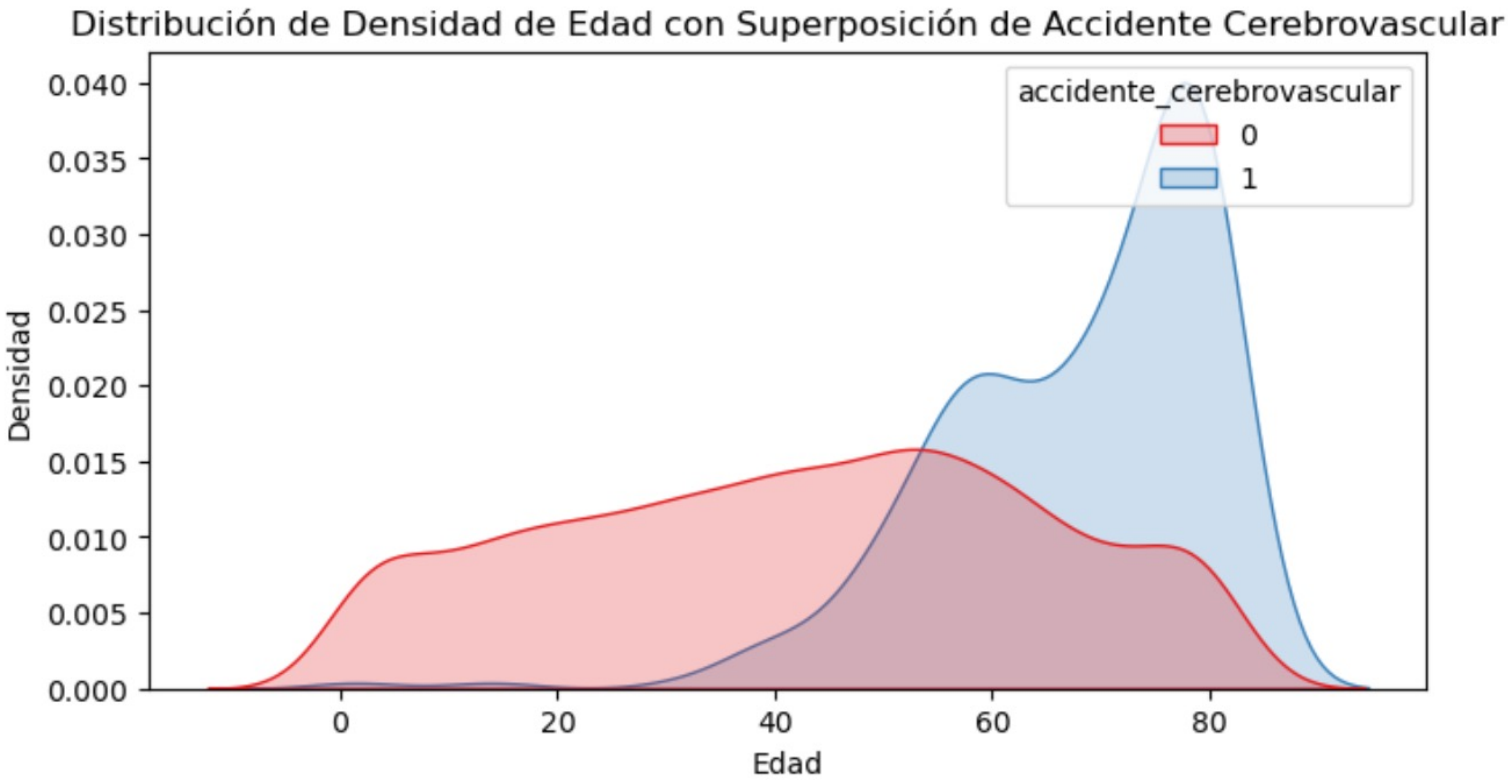
Nivel Promedio de Glucosa: La mayoría de los individuos tienen niveles de glucose en un rango que se consideraría normal o pre-diabético. Sin embargo, hay una cantidad significativa de valores atípicos en el extremo superior que podrían ser de interés clínico.

IMC: El IMC tiende a agruparse en el rango que se consideraría como sobrepeso, lo que podría tener su propia serie de implicaciones para la salud.

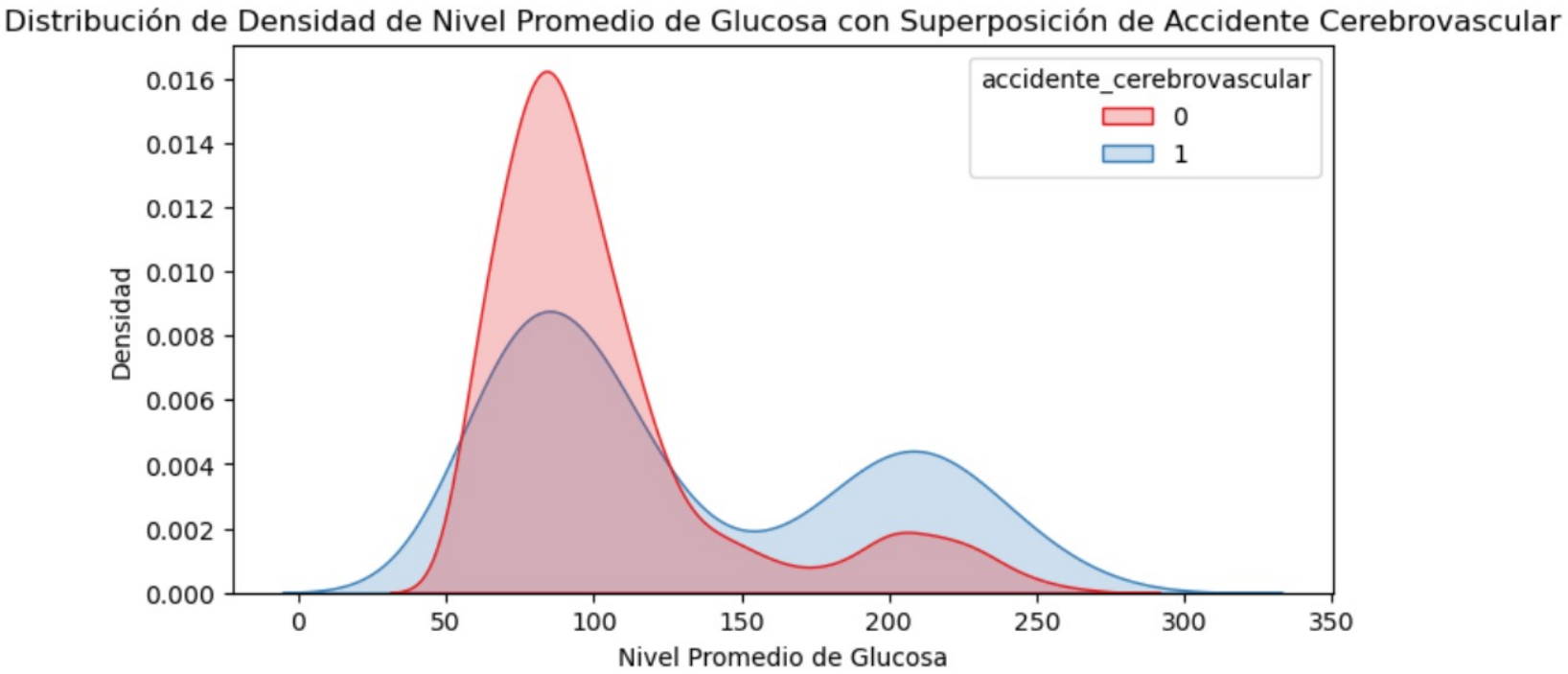
Exploración de la Relación entre Variables y Accidentes Cerebrovasculares

En esta sección, me centramos en entender cómo diferentes variables se relacionan con la probabilidad de sufrir un accidente cerebrovascular. Utilizando tanto gráficos de dispersión para las variables numéricas como gráficos de barras y de violín para las variables categóricas y binarias.

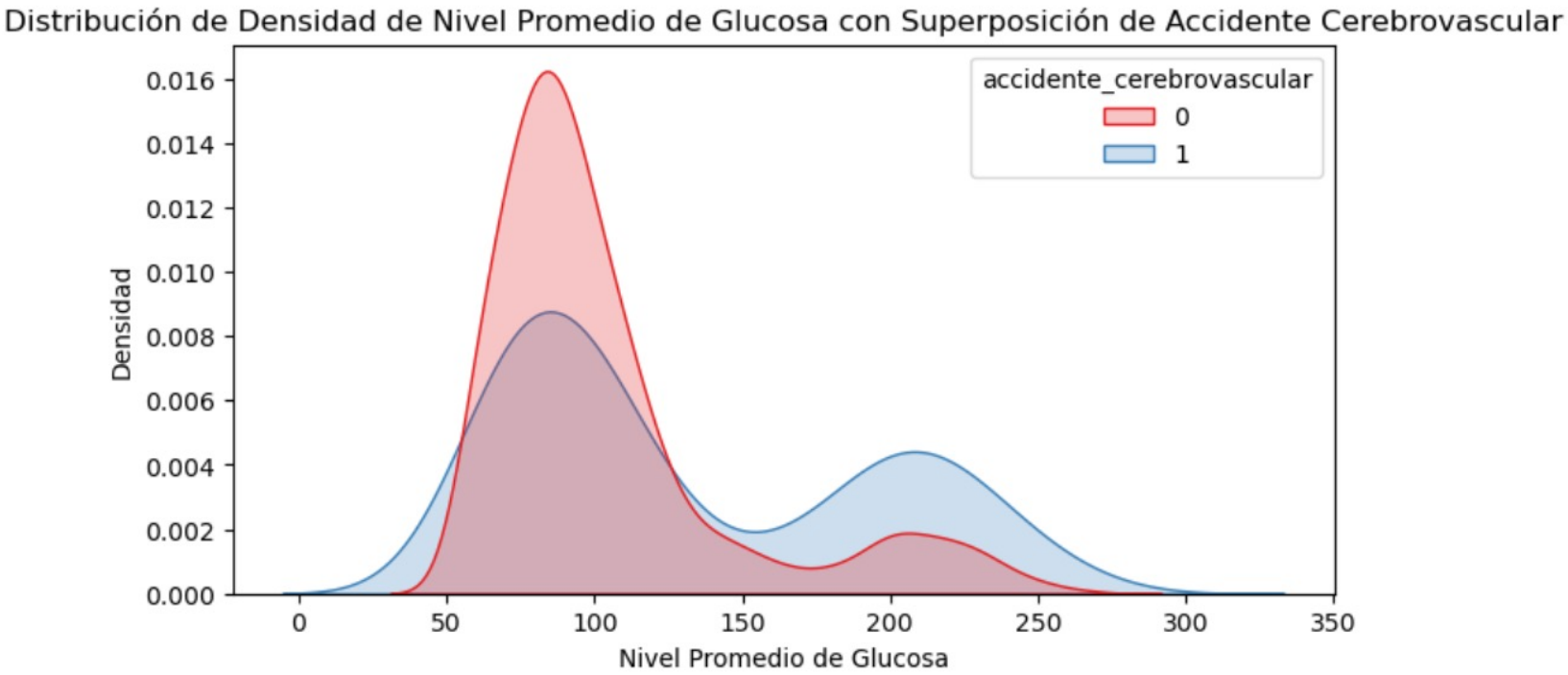
Edad vs Accidente Cerebrovascular



Nivel Promedio de Glucosa vs Accidente Cerebrovascular



IMC vs Accidente Cerebrovascular



Observaciones sobre los Gráficos de Densidad con Superposición de Accidente Cerebrovascular

Edad (age): El gráfico muestra claramente que la incidencia de accidentes cerebrovasculares es mucho más alta en personas mayores. La densidad de personas sin accidentes cerebrovasculares es bastante alta en las edades más jóvenes y disminuye a medida que aumenta la edad.

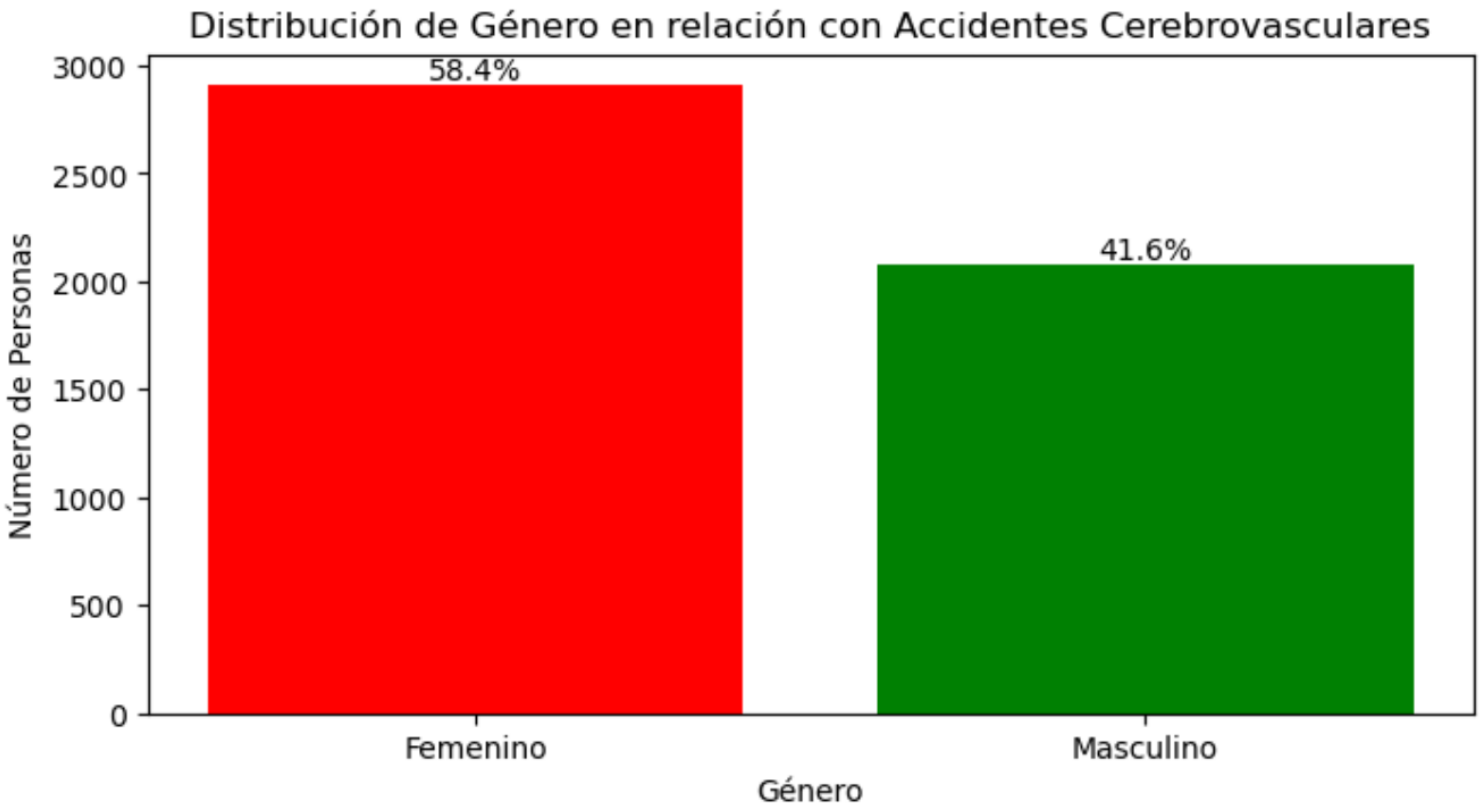
Nivel Promedio de Glucosa (avg_glucose_level) Hay dos picos notables para las personas con accidentes cerebrovasculares, uno en niveles bajos de glucosa y otro en niveles más altos. Esto podría sugerir que tanto niveles bajos como altos de glucosa podrían estar relacionados con un mayor riesgo de accidente cerebrovascular. La mayoría de las personas sin accidentes cerebrovasculares tienen un nivel de glucosa en un rango más "normal".

Índice de Masa Corporal (bmi) Para el índice de masa corporal, no se observa una diferencia significativa en la densidad entre las personas con y sin accidentes cerebrovasculares. Sin embargo, hay una ligera acumulación en el extremo más bajo del índice de masa corporal para las personas con accidentes cerebrovasculares.

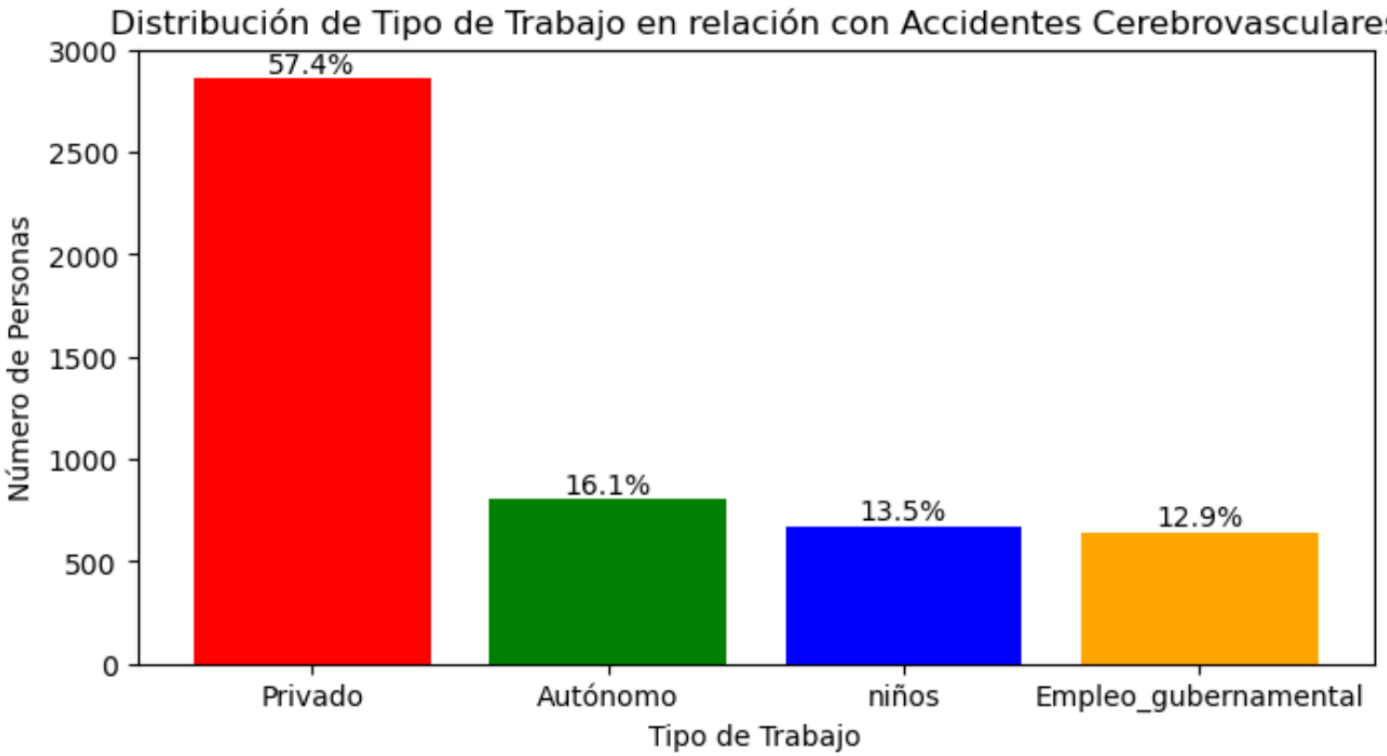
Variables Categóricas y Binarias

- Esta exploración visual nos ayuda a formar hipótesis iniciales sobre qué variables podrían ser más importantes para predecir los accidentes cerebrovasculares. Sin embargo, es crucial tener en cuenta que la visualización es sólo el primer paso. Se requerirá un análisis más profundo, posiblemente utilizando técnicas de aprendizaje automático, para validar estas hipótesis.

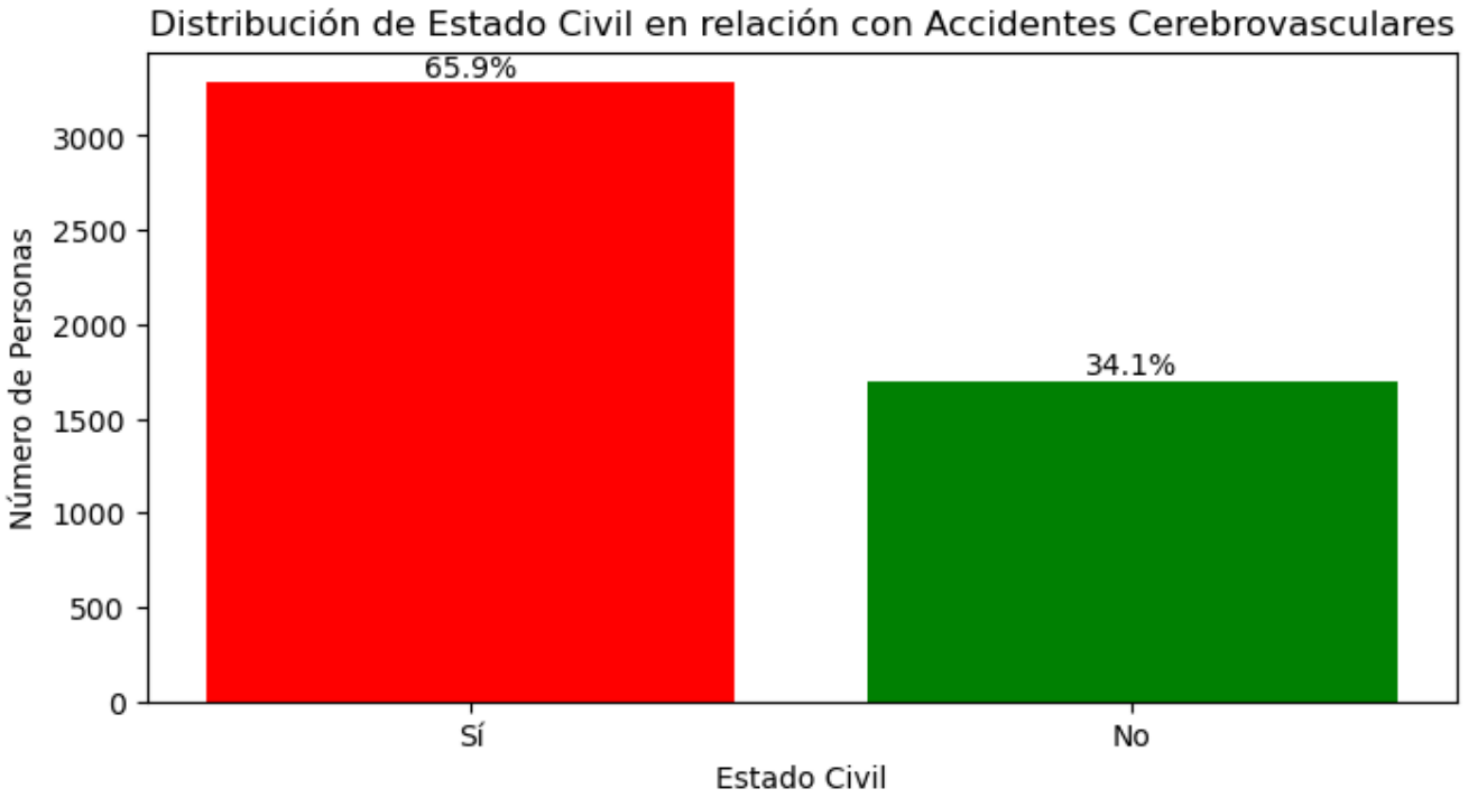
Género


Distribución de Género en relación con Accidentes Cerebrovasculares
Número de Personas
3000
2500
2000
1500
1000
500
0
58.4%
41.6%
Femenino
Masculino
Género

Tipo de Trabajo



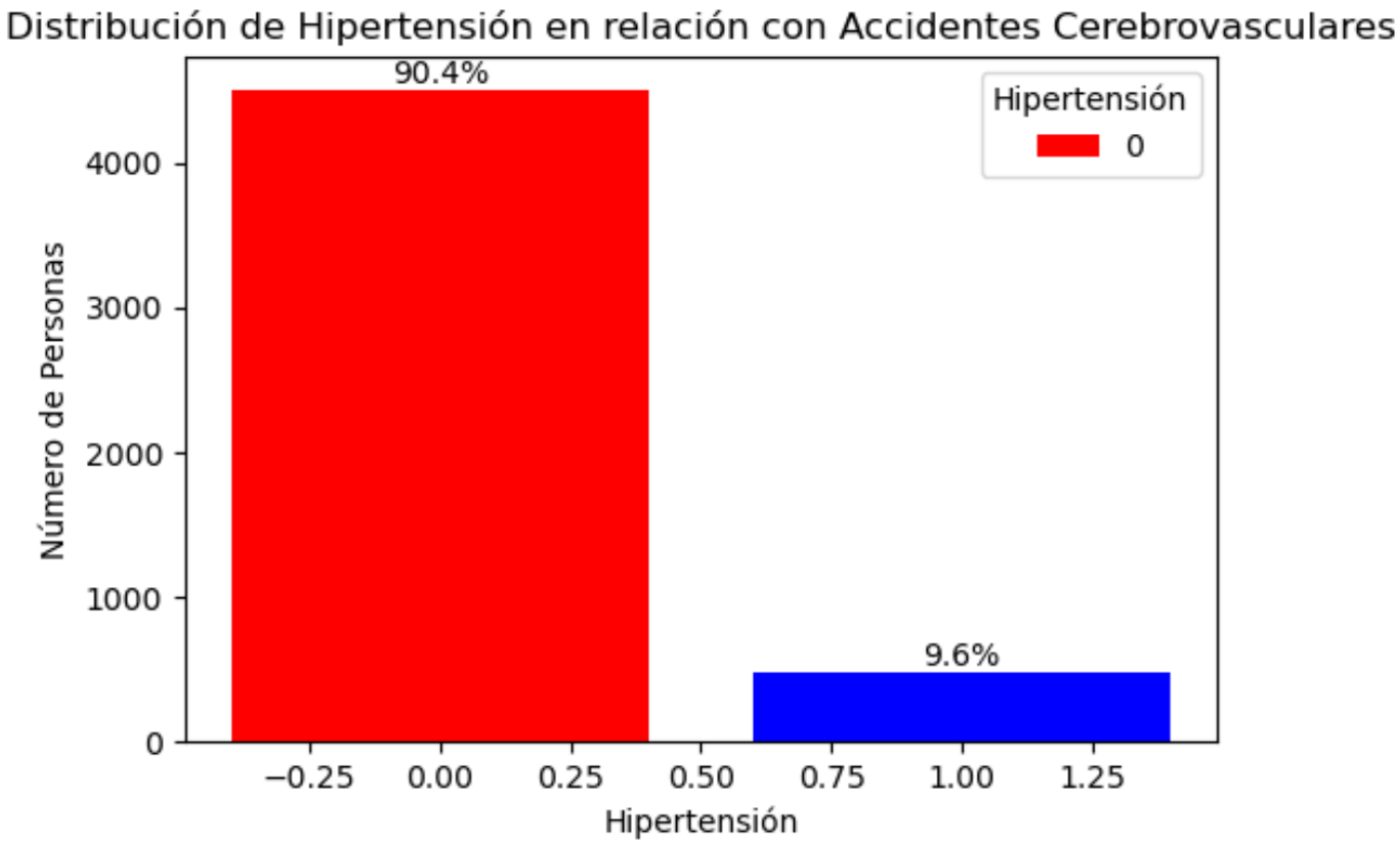
Estado Civil



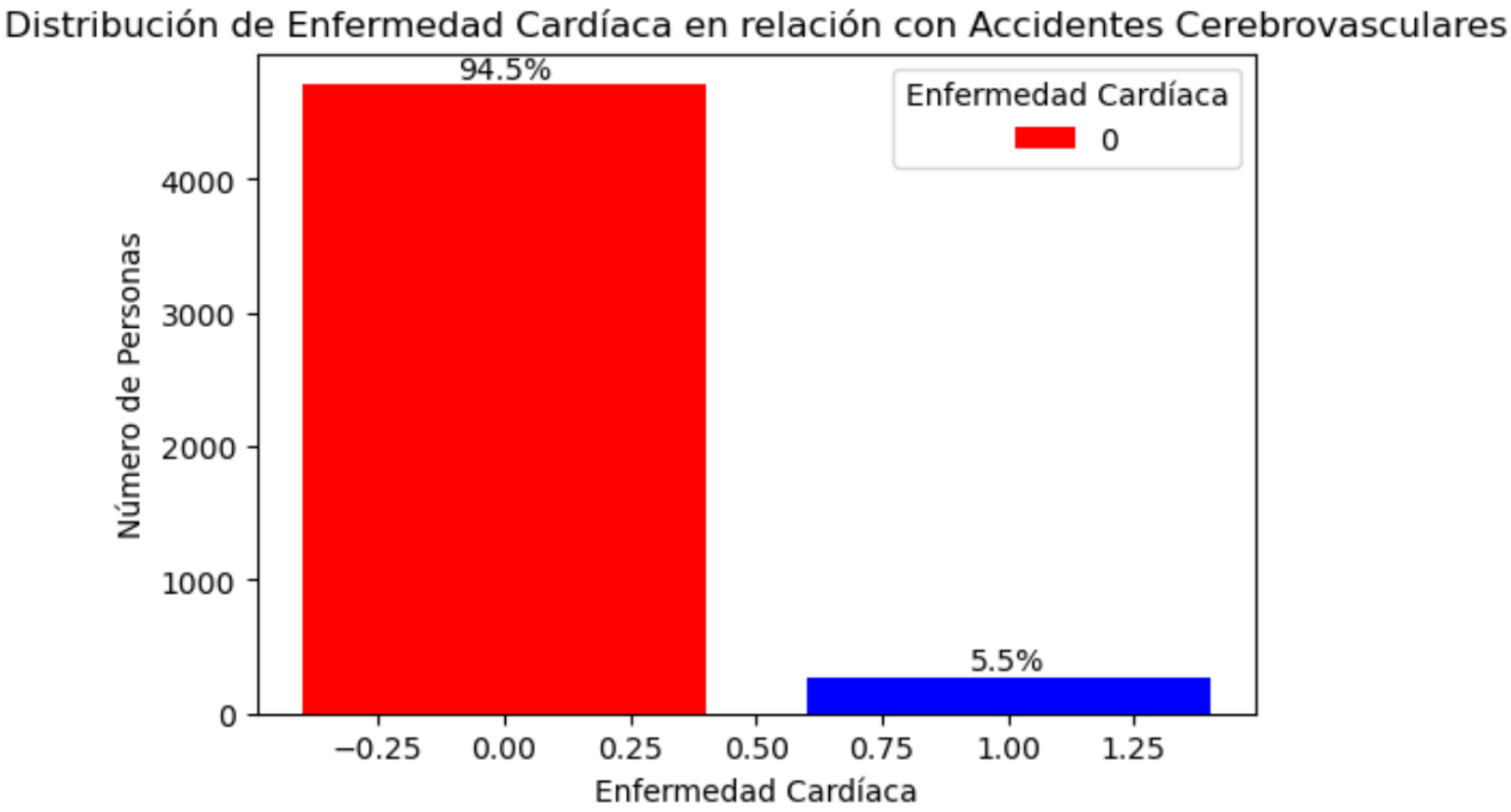
Insights Importantes

- Genero: La distribución de género muestra una presencia significativa de mujeres en comparación con los hombres en el dataset. Sería interesante explorar si hay una correlación significativa entre el género y la prevalencia de accidentes cerebrovasculares.
- Tipo de trabajo: La mayoría de las personas en el dataset trabajan en el sector privado. Podría ser relevante investigar si diferentes tipos de trabajos están asociados con diferentes niveles de riesgo para accidentes cerebrovasculares.
- Alguna vez Casado: Una gran proporción de las personas en el dataset ha estado casada en algún momento. Investigar la relación entre el estado civil y los accidentes cerebrovasculares podría ofrecer

Hipertensión



Enfermedad Cardíaca



Insights Adicionales para Variables Binarias

Hipertensión: Una mayoría significativa de las personas en el dataset no tiene hipertensión. Sin embargo, sería relevante explorar si la presencia de hipertensión está correlacionada con un mayor riesgo de accidentes cerebrovasculares.

Enfermedad Cardíaca: Al igual que con la hipertensión, una gran mayoría de las personas en el dataset no tiene enfermedades cardíacas. Explorar la relación entre enfermedades cardíacas y accidentes cerebrovasculares podría ofrecer insights adicionales.

Contestando a las preguntas iniciales:

- 1. ¿Cuáles son los principales factores de riesgo para sufrir un accidente cerebrovascular según los datos?

  R: La edad, entre mas grande eres mas propenso a sufrir uno al pasar los 50 años, tambies es mucho mas probable si tienes un trabajo en el sector privado y si Tambien alguna vez te has Casado, estos son los factores que mas pesan basandonos en lo datos.
  Contestando a las preguntas inciales:

2. ¿Existe una correlación entre el nivel medio de glucosa y la incidencia de accidentes cerebrovasculares?

R: Hay dos picos notables para las personas con accidentes cerebrovasculares, uno en niveles bajos de glucosa y otro en niveles más altos. Esto podría sugerir que tanto niveles bajos como altos de glucosa podrían estar relacionados con un mayor riesgo de accidente cerebrovascular. La mayoría de las personas sin accidentes cerebrovasculares tienen un nivel de glucosa en un rango más “normal”.

3. ¿Cómo varía la probabilidad de sufrir un accidente cerebrovascular a lo largo de differentes grupos de edad?

R: Los datos muestran claramente que la incidencia de accidentes cerebrovasculares es mucho más alta en personas mayores. La densidad de personas sin accidentes cerebrovasculares es bastante alta en las edades más jóvenes y disminuye a medida que aumenta la edad.

4. ¿El tipo de trabajo está relacionado con la probabilidad de sufrir un accidente cerebrovascular?

R: Si definitivamente se puede observar que el trabajo en un entorno coporativo rpivado aumenta la probablidad de tener un accidente cerebrovascular.

5. ¿Haberse casado alguna vez influye en tener un accidente cerebrovascular?

R: Si se puede observer que incrementa las probabilidades habserse Casado, aunque no se puede inferior mas proque no se sabe si ya no esta Casado ni los motivos de esta separacin, ni los años de Casado ni otros factores que podrian ser releavantes y entrariamos en un sesgo de informacion incompleta

Conclusiones del modelo ML

- Usamos OneHotEncoder para las variables categóricas, lo que crea nuevas columnas indicadoras binarias para cada categoría. Para las variables numéricas, aplicamos StandardScaler para es tandarizar las características eliminando la media y escalando a la varianza de la unidad

- Para seleccionar los mejores parámetros para nuestro modelo Random Forest, utilizamos GridSearchCV

Mejor Puntuación de Validación Cruzada (F1)

La puntuación F1 combina la precisión y la recuperación en una sola métrica, que es particularmente útil en situaciones de clases desequilibradas. La mejor puntuación F1 obtenida fue de **0.07005362827109404**. Esta es una puntuación bastante baja, lo que sugiere que el modelo no está realizando una buena distinción entre las clases, particularmente la clase minoritaria (los casos positivos de accidente cerebrovascular)

Aplicación de la técnica SMOTE

Después de obtener un puntaje F1 inicial con nuestro conjunto de datos desequilibrado, vamos a aplicar una técnica de oversampling conocida como SMOTE (Synthetic Minority Over-sampling Technique) para aumentar el número de instancias de la clase minoritaria en nuestro conjunto de entrenamiento. Esto debería ayudar a mejorar el rendimiento del modelo, especialmente en lo que respecta a la sensibilidad y el puntaje F1.

Usaremos el mejor estimador encontrado por GridSearchCV para asegurarnos de que estamos uti lizando los parámetros más óptimos identificados previamente. Después del reentrenamiento, hare mos predicciones sobre nuestro conjunto de validación y calcularemos el puntaje F1 para evaluar el impacto del reequilibrio de clases

Puntaje F1 después de aplicar SMOTE: 0.10344827586206896

Conclusión Final del Modelo de Machine Learning

A lo largo de este ejercicio, hemos construido y evaluado un modelo de Machine Learning uti lizando un conjunto de datos sobre casos de accidentes cerebrovasculares. Nuestro objetivo era predecir la probabilidad de que un individuo sufra un accidente cerebrovascular basándonos envarias características demográficas y de salud

Este ejercicio ha sido un ejemplo valioso de los desafíos reales que enfrentamos en el campo de la ciencia de datos, particularmente cuando se trabaja con datos desequilibrados en un context crítico como el de la salud. Aunque los resultados no fueron óptimos, proporcionan una base sólida sobre la cual se pueden construir futuros modelos y una experiencia de aprendizaje significativa en la aplicación práctica de técnicas de Machine Learning.