

PREDICCION DE ACCIDENTES CEREBROVASCULARES

DATA SCIENCE II

Ignacio Festino

01

ABSTRACTO

El análisis de un conjunto de datos sobre pacientes con y sin accidente cerebrovascular (ACV) ofrece una oportunidad para identificar factores clave que contribuyen a este evento crítico. Este estudio se enfoca en variables específicas: edad, hipertensión, enfermedades cardíacas preexistentes, tipo de trabajo, lugar de residencia (rural o urbano), índice de masa corporal (IMC) y hábito de fumar. Este análisis se encuentra orientado al ámbito de la salud y especialistas del sistema nervioso.

02 METADATA

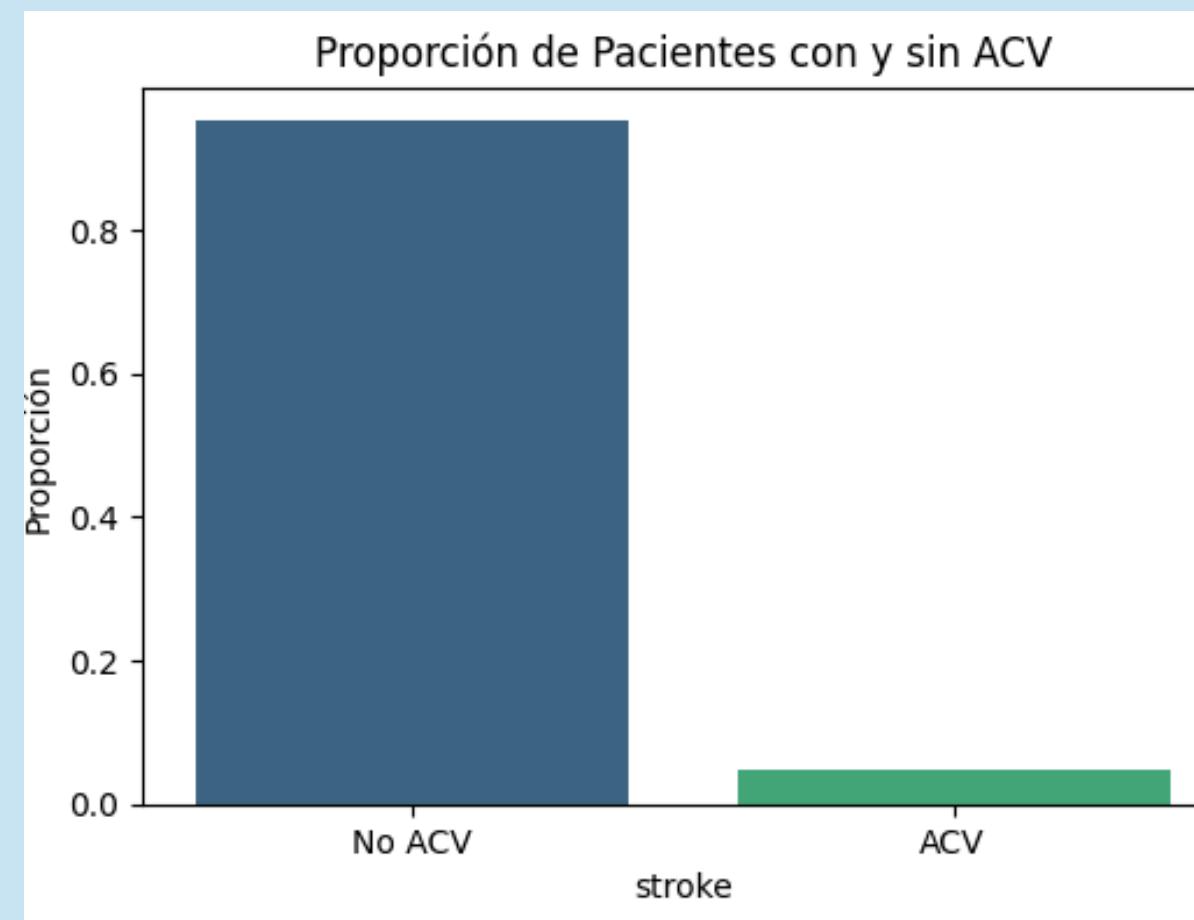
Columna	Tipo	Descripción
ID	Entero	Identificación única del paciente
Gender	Categórico	Género del paciente
Age	Numérico	Edad del paciente
Hypertension	Booleano	Indicador de hipertensión
Heart Disease	Booleano	Indicador de enfermedades cardíacas preexistentes
Ever Married	Booleano	Estado civil del paciente (casado alguna vez o no)
Work Type	Categórico	Tipo de empleo del paciente
Residence Type	Categórico	Tipo de área de residencia del paciente

Columna	Tipo	Descripción
Avg Glucose Level	Numérico	Nivel promedio de glucosa en sangre del paciente
BMI	Numérico	Índice de masa corporal del paciente
Smoking Status	Categórico	Estado de fumador del paciente
Stroke	Booleano	Indicador de si el paciente ha tenido un accidente cerebrovascular (ACV)

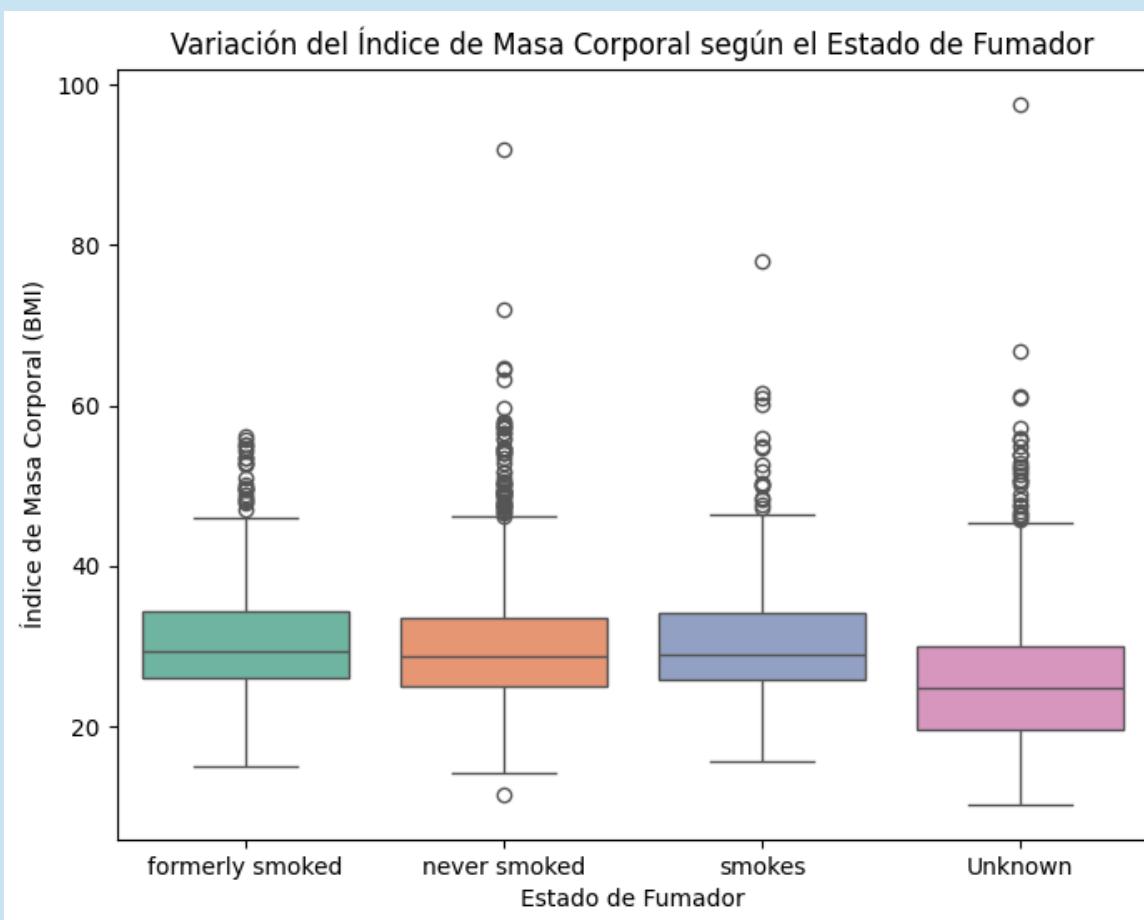
03 PREGUNTAS E HIPOTESIS

PREGUNTAS	HIPOTESIS
¿CUÁLES SON LOS FACTORES MÁS SIGNIFICATIVOS ASOCIADOS CON LA OCURRENCIA DE UN ACV ENTRE LAS VARIABLES CONSIDERADAS?	LA HIPERTENSIÓN Y LAS ENFERMEDADES CARDÍACAS PREEXISTENTES ESTÁN FUERTEMENTE ASOCIADAS CON UN MAYOR RIESGO DE ACV
¿CÓMO INFLUYEN LAS CARACTERÍSTICAS INDIVIDUALES, COMO LA HIPERTENSIÓN, ENFERMEDADES CARDÍACAS, Y EL IMC, EN EL RIESGO DE ACV?	UN IMC ELEVADO Y EL HÁBITO DE FUMAR AUMENTAN SIGNIFICATIVAMENTE LA PROBABILIDAD DE SUFRIR UN ACV
¿EXISTEN DIFERENCIAS EN EL RIESGO DE ACV SEGÚN EL TIPO DE TRABAJO, EL LUGAR DE RESIDENCIA, Y EL HÁBITO DE FUMAR?	EL TIPO DE TRABAJO Y EL LUGAR DE RESIDENCIA (RURAL VS. URBANO) INFLUYEN EN EL RIESGO DE ACV, CON POSIBLES VARIACIONES SEGÚN EL ENTORNO DE TRABAJO Y LAS CONDICIONES DE VIDA.

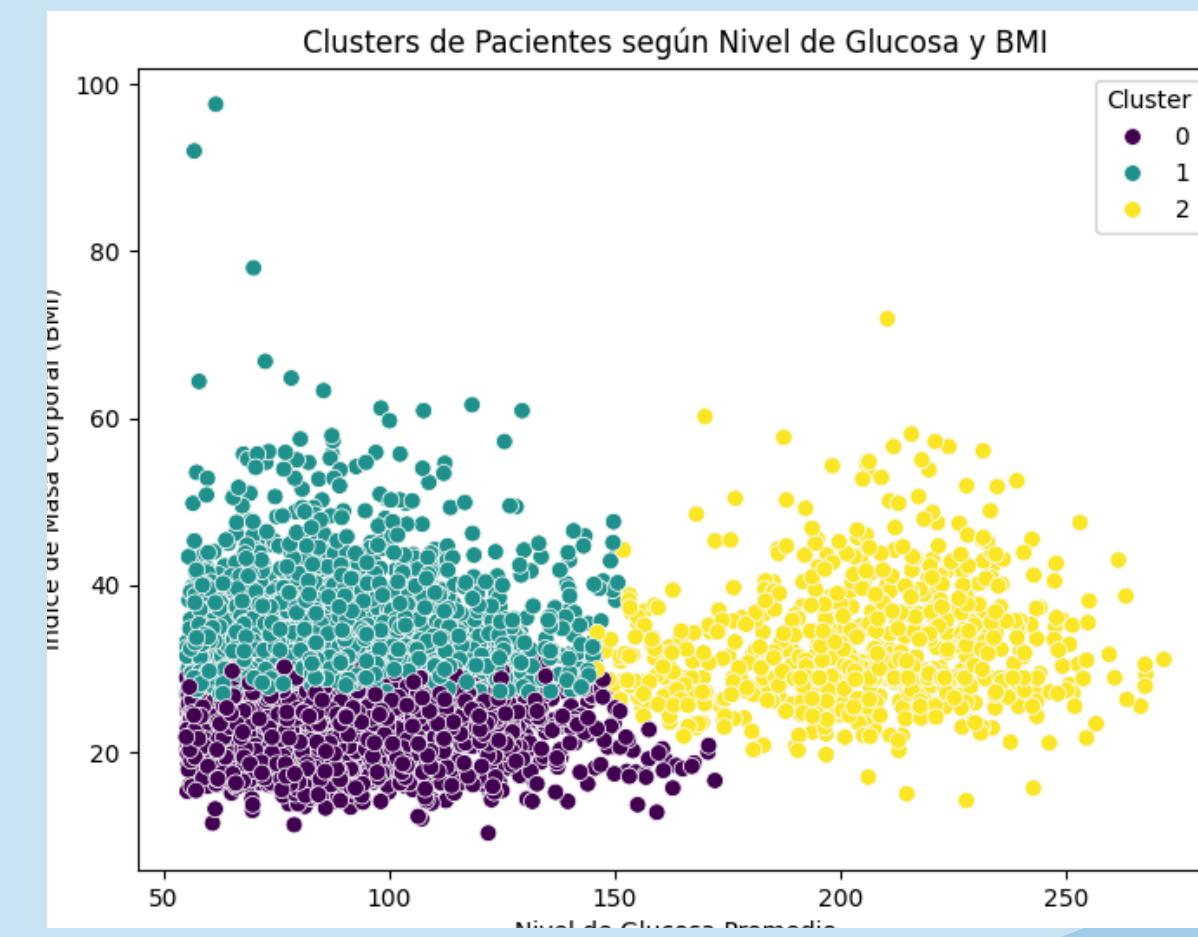
04 VISUALIZACIONES



ANALISIS UNIVARIADO
Casos de ACV vs NO
ACV



ANALISIS BIVARIADO
Variacion IMC vs Estado
de Fumador



**ANALISIS
MULTIVARIADO**
Pacientes según nivel de
glucosa y IMC

05 INSIGHTS

1

Se encuentra una gran mayoria de casos donde los pacientes no sufrieron ACV por lo que dificulta la predicciones

2

La mayoria de casos de ACV se concentran en el rango etario entre 75-85 años

3

Habiendo hecho un cluster podemos evidenciar que grupos de acuerdo a sus caracteristicas y datos fisiologicos son mas propensos a sufrir un ACV

06

INGENIERIA DE ATRIBUTOS

- **Clasificación de Niveles de Glucosa:** Categorización en 'Bajo', 'Normal', 'Alto' .
- **Codificación de Género:** Conversión de valores de 'gender' a numéricos
- **Cálculo del Factor de Riesgo:** Creación de una variable 'risk_factor' multiplicando 'age' y 'avg_glucose_level'.
- **Rango Saludable de IMC**
- **Normalización y Estandarización:** Normalización de 'age' y 'avg_glucose_level'
- **Codificación del Estado de Fumador**
- **Manejo de Valores Faltantes:** Imputación de valores faltantes en 'bmi' con la mediana y creación de una indicación de valores faltantes.

07

ENTRENAMIENTO Y TESTEO

- **Regresión Logística:** Se entrenó un modelo de regresión logística con LogisticRegression.
- **Árbol de Decisión:** Se entrenó un modelo de Árbol de Decisión con DecisionTreeClassifier.
- **Predicciones:** Se realizaron predicciones en el conjunto de prueba con ambos modelos (y_pred_log_reg para regresión logística y y_pred_dt para el árbol de decisión).
- **Matriz de Confusión:** Evaluación de la precisión y errores de los modelos mediante la matriz de confusión.
- **Reporte de Clasificación:** Obtención de métricas de evaluación detalladas (precisión, recall, F1-score) con classification_report.
- **Precisión (Accuracy):** Cálculo de la precisión de ambos modelos en el conjunto de prueba con accuracy_score.
- Realizamos **validación cruzada** con cross_val_score para ambos modelos.

08 OPTIMIZACION

- **Regresión Logística:** Se utilizó RandomizedSearchCV para optimizar los hiperparámetros C y solver. Se encontró que los mejores hiperparámetros son log_reg_random.best_params_.
- **Árbol de Decisión:** Se utilizó RandomizedSearchCV para optimizar los hiperparámetros max_depth, min_samples_split y min_samples_leaf. Los mejores hiperparámetros encontrados son dt_random.best_params_.



09 SELECCION DE MODELO

REGRESION LOGISTICA

- MEJORES HIPERPARÁMETROS: {'C': 3.845401188473625, 'SOLVER': 'LIBLINEAR'}
- MATRIZ DE CONFUSIÓN: 960 0 62 0
- PRECISIÓN (ACCURACY): 0.94
- PRECISION: 0.94 (PARA LA CLASE 0), 0.00 (PARA LA CLASE 1)
- RECALL: 1.00 (PARA LA CLASE 0), 0.00 (PARA LA CLASE 1)
- F1-SCORE: 0.97 (PARA LA CLASE 0), 0.00 (PARA LA CLASE 1)
- MSE: 0.0607
- ROC-AUC: 0.85

ARBOL DE DECISION

- MEJORES HIPERPARÁMETROS: {'MAX_DEPTH': 4, 'MIN_SAMPLES_LEAF': 4, 'MIN_SAMPLES_SPLIT': 5}
- MATRIZ DE CONFUSIÓN: 923 37 54 8
- PRECISIÓN (ACCURACY): 0.91
- PRECISIÓN: 0.94 (PARA LA CLASE 0), 0.18 (PARA LA CLASE 1)
- RECALL: 0.96 (PARA LA CLASE 0), 0.13 (PARA LA CLASE 1)
- F1-SCORE: 0.95 (PARA LA CLASE 0), 0.15 (PARA LA CLASE 1)
- MSE: 0.0897
- ROC-AUC: 0.80

10

CONCLUSION

La Regresión Logística fue seleccionada como el mejor modelo basado en varias métricas de rendimiento clave. Este modelo mostró una precisión más alta y un menor error cuadrático medio (MSE), indicando una mayor capacidad para predecir correctamente sin errores significativos. Además, el área bajo la curva (ROC-AUC) de la Regresión Logística fue superior, lo que demuestra una mejor discriminación entre las clases.



MUCHAS
GRACIAS

Ignacio Festino