Machine Learning

Se entrega el siguiente set de datos de una empresa del sector de salud, **HealthAnalytics.csv**, el cual es un dataset que cuenta con la información de una cadena de hospitales que tiene como objetivo crear la próxima generación de atención médica para sus pacientes, para esto ha contratado a usted para ayudarlo a lograr su visión. La compañía reúne a los mejores médicos y les permite brindar atención médica proactiva a sus pacientes. En este caso, su cliente quiere estudiar alrededor de una de las enfermedades críticas "**Stroke**". El **accidente cerebrovascular** es una enfermedad que afecta las arterias que conducen hacia y dentro del cerebro. Un accidente cerebrovascular ocurre cuando un vaso sanguíneo que transporta oxígeno y nutrientes al cerebro está bloqueado por un coágulo o estalla (o se rompe). Cuando eso sucede, parte del cerebro no puede obtener la sangre (y el oxígeno) que necesita, por lo que mueren las células cerebrales. En los últimos años, el **Cliente** ha capturado varios detalles de salud, demográficos y de estilo de vida sobre sus pacientes. Esto incluye detalles como la edad y el sexo, junto con varios parámetros de salud (por ejemplo, hipertensión, índice de masa corporal) y variables relacionadas con el estilo de vida (por ejemplo, el tabaquismo, el tipo de ocupación).

Debido a que usted está llevando una especialización en Machine Learning se requiere que los ayude a **IDENTIFICAR A LAS PERSONAS PROPENSAS A SUFRIR UN ACCIDENTE CARDIOVASCULAR.**

- Las variables que se disponibilizan son :

Variable	Definition
id	Patient ID
gender	Género del paciente
age	Edad del paciente
hypertension	Tenencia de hipertensión ?
heart_disease	Tenencia de enfermedad del corazón ?
ever_married	Está casado?
work_type	Tipo de ocupación del paciente
Residence_type	Tipo de área de residencia
avg_glucose_level	Nivel promedio de glucosa
bmi	Indice de masa corporal
smoking_status	El cliente fuma?
stroke	**0 - No stroke, 1 - Sufrió Stroke**

Entregables del TA:

1.- Identificar el problema de la naturaleza o del negocio que tiene la empresa, asociado a la necesidad o la razón de su situación actual.

El accidente cerebrovascular (stroke) representa una de las principales amenazas a la salud pública debido a su alta incidencia y consecuencias severas. Esta enfermedad ocurre cuando un vaso sanguíneo encargado de llevar oxígeno y nutrientes al cerebro se obstruye o se rompe, lo que impide el flujo adecuado de sangre a determinadas zonas del cerebro, provocando la muerte de las células cerebrales. Esta condición genera complicaciones neurológicas graves e incluso puede causar la muerte. En este contexto, surge la necesidad urgente de implementar mecanismos predictivos y preventivos que permitan identificar a personas con alto riesgo de padecer un accidente cerebrovascular, con el objetivo de reducir su incidencia y mitigar sus efectos en la población.

2.- Definir los objetivos de negocio que van a dar solución a la situación actual o van a ser un paliativo a su problema de negocio.

- Objetivo General
  - Desarrollar un modelo predictivo para identificar a las personas propensas a sufrir un accidente cardiovascular.
- Objetivos Específicos
  - Identificar variables clave que influyen en el riesgo de sufrir un accidente cerebrovascular.
  - Desarrollar un modelo de clasificación binaria que prediga si un paciente tiene alto riesgo de sufrir un accidente cerebrovascular
  - Implementar alertas o reportes automáticos para que el personal médico pueda intervenir de forma preventiva en los casos de alto riesgo.

3.- Los objetivos antes identificados los puede solucionar mediante machine learning basado en aprendizaje supervisado o aprendizaje no supervisado. Justifique su respuesta.

Para este caso se utilizará aprendizaje supervisado, específicamente un modelo de clasificación binaria, ya que la variable objetivo (stroke) es binariaa (0 o 1), indicando si un paciente ha sufrido o no un accidente cerebrovascular. Este tipo de problema se puede abordar con modelos como regresión logística, árboles de decisión, Random Forest o XGBoost, que permiten predecir la probabilidad de que un paciente esté en riesgo. Se recomienda iniciar con modelos interpretables como la regresión logística o árboles de decisión, y luego escalar a modelos más robustos como Random Forest si se busca mayor precisión

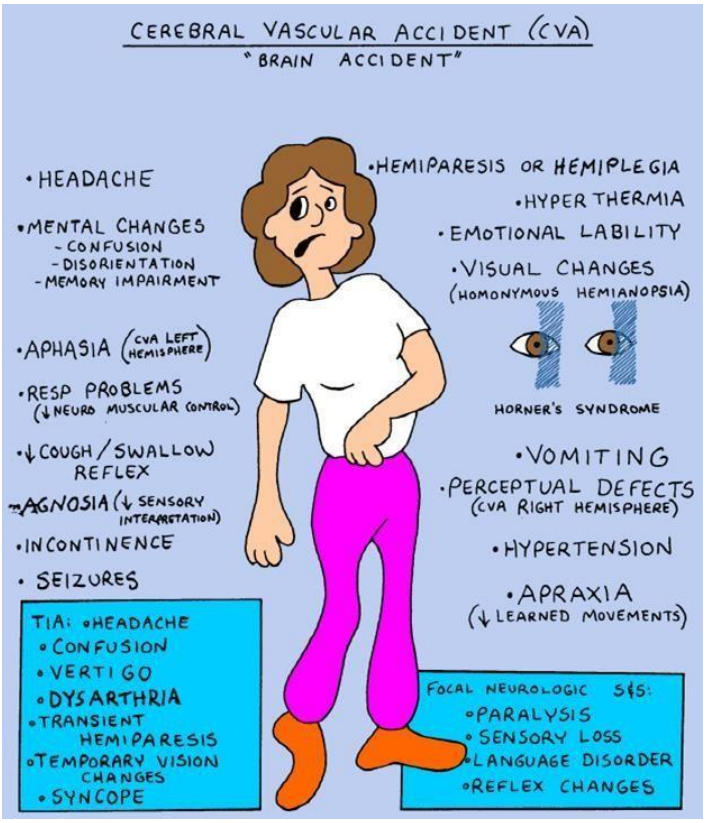
4.- ¿Qué tipo de variables se utilizan en el problema de negocio? Es decir que dominios identifica. Ejemplo: Dominio sociodemográfico, dominio de facturación, dominio de reclamos.

- Sociodemográfico: Gender, age, ever_married, Residence_type
- Antecendentes medicos: hypertension, heart_disease, avg_glucose_level, Bmi

- Estilo de vida: Work_type, smoking_status

5.- De las variables entregadas en la BBDD, identifique:

- La variable **objetivo, respuesta o target**.
  - Stroke
- Las variables **independientes o drivers**.
  - Gender, age, ever_married, Residence_type, hypertension, heart_disease, avg_glucose_level, Bmi, Work_type, smoking_status

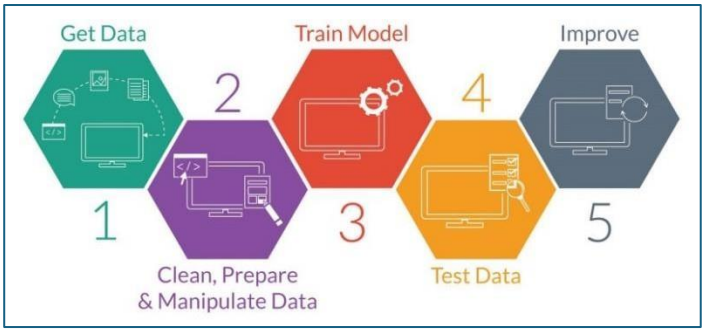


6.- Construya mediante las distintas técnicas de ingeniería de variables, nuevas características que les permitan tener mejores relaciones o mayor explicabilidad con la variable objetivo.

Se crearon nuevas variables paramejorar la predictibilidad del modelo:

- **Binarización de variables:** Se binarizó ever_married, gender, work_type y Residence_type para que los modelos pudieran procesarlas de manera más eficiente.
- **Agrupación por riesgo clínico:** Se crearon categorías para age (age_group), bmi (bmi_category) y avg_glucose_level (glucose_category) basándose en umbrales médicos.
- **Variables de interacción:** Se construyeron variables que capturan la combinación de riesgos, como multi_condition_risk (para hipertensión y enfermedad cardíaca) y metabolic_risk (para glucosa alta y sobrepeso).

- **Variable ordinal de riesgo:** Se creó health_risk, una variable ordinal que resume el riesgo de un paciente en categorías de 'High', 'Moderate' y 'Low' según su hipertensión y enfermedad cardíaca.



7.- De los drivers o features seleccionados cuáles de éstos son los más importantes para determinar la probabilidad de cada persona de sufrir un accidente cerebro vascular.

Se identificó, a través de un análisis de selección de características como Boruta, que los "drivers" más importantes para determinar la probabilidad de sufrir un accidente cerebrovascular son:

- **age**: La edad es un factor de riesgo fundamental. El riesgo de accidente cerebrovascular aumenta significativamente a medida que una persona envejece.
- **avg_glucose_level**: El nivel promedio de glucosa en sangre es un indicador crítico de riesgo, ya que los niveles altos están directamente relacionados con la diabetes, una de las principales causas de accidente cerebrovascular.
- **bmi**: El Índice de Masa Corporal (IMC) es un factor importante, ya que el sobrepeso y la obesidad son condiciones que aumentan la probabilidad de sufrir un accidente cerebrovascular.
- **smoking_status**: El tabaquismo es un factor de riesgo conocido para enfermedades cardiovasculares, incluyendo el accidente cerebrovascular. Esta variable capta el impacto del estilo de vida del paciente.
- **work_type**: El tipo de trabajo puede influir en el riesgo de accidente cerebrovascular. Factores como el estrés laboral, la actividad física y el sedentarismo asociados a la ocupación son relevantes para la salud cardiovascular.

8.- ¿Qué conclusiones obtienes desarrollando un modelo con la totalidad de las variables y uno con las variables más relevantes? Comente sus resultados.

Se demostró que usar al solo las variables más relevantes para entrenar los modelos, en lugar de todas las variables disponibles, es una estrategia más efectiva. Esto se debe a que la precisión global (accuracy) es una métrica engañosa en un problema de clasificación con clases desbalanceadas. En el caso de predecir accidentes cerebrovasculares (stroke), donde la mayoría de los pacientes no los sufren, un modelo puede alcanzar una alta precisión simplemente prediciendo siempre la clase mayoritaria.

Sin embargo, al usar solo las variables más importantes (age, avg_glucose_level, bmi, smoking_status, work_type), el modelo puede aprender a identificar mejor los patrones específicos que conducen a un stroke. Esto se refleja en métricas más críticas como la sensibilidad (recall) y la precisión (precision), que miden la capacidad del modelo para identificar correctamente a los pacientes de riesgo y la proporción de predicciones positivas que fueron correctas, respectivamente. Un modelo con un buen recall es fundamental en este contexto clínico, ya que ayuda a evitar falsos negativos (pacientes de riesgo que no son detectados)

9.- Entrene y valide con la información proporcionada un algoritmo de Machine Learning para solucionar la problemática planteada.

La regresión Logística es la mejor opción entre los tres modelos analizados. A pesar de su accuracy, su capacidad para detectar los casos de stroke (evitando falsos negativos) es muy superior. En un contexto médico, es mucho más importante identificar a un paciente en riesgo que tener una alta precisión en todas las predicciones. El recall es la métrica más importante en este caso, y la regresión logística es el modelo que la satisface.

Modelo	Accuracy	Recall
Logistic Regression	74.25%	78.13%
Decision Tree Classifier	69.71%	84.82%
Random Forest	73.75%	79.46%
AdaBoost	87.74%	30.36%

10.- Cuáles son sus principales recomendaciones o hallazgos en la aplicación del laboratorio

El objetivo principal fue desarrollar un modelo predictivo para identificar a pacientes con riesgo de sufrir un accidente cerebrovascular (stroke). El desafío más significativo fue el desbalanceo de la variable objetivo, donde la inmensa mayoría de los pacientes no sufre un stroke. Esto llevó a que métricas como la accuracy fueran engañosas.

- Ingeniería de Características

  Se utilizaron técnicas de ingeniería de variables para crear nuevas características más predictivas. Se agruparon variables continuas como la edad y el IMC en categorías de riesgo clínico, y se crearon variables de interacción para capturar el riesgo combinado de condiciones como la hipertensión y la enfermedad cardíaca. Estas variables mejoraron la capacidad del modelo para encontrar patrones significativos.

- Métricas Clave

  La accuracy no fue una métrica confiable. El recall se identificó como la métrica más crítica, ya que mide la capacidad del modelo para identificar a los pacientes en riesgo, evitando falsos negativos.

- Selección del Modelo

  Se probaron varios modelos de clasificación. Aunque modelos como Random Forest y Adaboost mostraron una accuracy muy alta, su recall fue de igual o menos que 0.05, lo que los hizo inútiles para la detección de stroke. El modelo de Regresión Logística, a pesar de su menor accuracy, demostró ser el más efectivo, logrando un recall superior a 0.75. Esto indica que es el único modelo que tiene la capacidad de identificar a una parte significativa de los pacientes en riesgo.