UNIVERSIDAD DEL PACÍFICO
Departamento Académico de Economía
Machine Learning para Economistas
Segundo semestre de 2022
Profesor: F. Rosales, JP: R. Arauco

PRÁCTICA CALIFICADA 4

Indicaciones Generales:

Esta práctica calificada es sobre modelos basados en árboles. Es para la casa y se debe desarrollar en grupos de 2 ó 3 personas. Para su desarrollo usted deberá utilizar la data en `cm_dataset.csv`. Se le recomienda usar sólo las variables listadas en el archivo `cm_dictionary.xlsx`.

Contexto:

El 21 porciento de las mujeres en el mundo se casa antes de cumplir los 18 años. A pesar de cierto progreso para lidiar con este problema, el matrimonio infantil (MI) sigue siendo una práctica común, en particular en el sur de Asia. Los datos en `cm_dataset.csv` contienen indicadores a nivel regional (regional), información obtenida a través de entrevistas (household), e información obtenida mediante fuentes terceras (georeferenced) que podrían estar vinculadas a la presencia de matrimonio infantil a nivel familiar. La descripción de las variables y su agrupamiento por tipo de variable está en el archivo `cm_dictionary.xlsx`. La variable objetivo es la variable binaria `married_`, donde "1" es un marcador que indica que al menos una persona en el hogar contrajo matrimonio antes de cumplil 18 años, y "0" indica lo contrario. Considere que el objetivo de este trabajo es ayudar a los tomadores de decisiones a construir un modelo que prediga la presencia de matrimonio infantil para diseñar políticas que ayuden a reducirlo.

Preguntas:

1. Inspección de datos
   (a) ¿Cuál es la proporción de MI por país? ¿Y en todos los países? (1 pts.)
   (b) ¿Qué hipótesis le sugieren los datos? ¿Alguna variable debería estar relacionada con MI, alguna interacción entre las variables? (3 pts.)
   (c) ¿Hay alguna particularidad adicional en la data que valga la pena resaltar? (3 pts.)
2. Modelo basado en árboles
   (a) La base de datos está desbalanceada ¿Cómo lidiará con este problema? (2 pts.)
   (b) ¿Bagging o Boosting? ¿Por qué? (2 pts.)
   (c) ¿Qué métrica de desempeño seleccionó? ¿Por qué? (2 pts.)
   (d) Reporte la matriz de confusión en test, para un train/test split de 70/30. Reporte también otras métricas que considere relevantes. (2 pts.)
3. Interpretación de resultados
   (a) ¿Cuáles son las variables más importantes para predecir la incidencia de MI? (2 pts.)

(b) ¿Qué acciones de política se podrían tomar para mitigar el MI? (3 pts.)

Indicaciones Específicas:

- Prepare su código en python como jupyter notebook. Asegúrense de que sea ordenado y fácil de entender. Utilicen comentarios, tablas y gráficos cuando lo consideren necesario.
- Envíe un único archivo comprimido con sus resultados a `francisco@brein.pe`. En el tema del mensaje coloque `MLE-PC4`. En el cuerpo del mensaje indique los integrantes del grupo.
- La fecha límite de entrega es el día domingo 20 de Noviembre a las 11:59pm. Cualquier entrega posterior será penalizada.