1. CONJUNTO DE DATOS

   Para la resolución del examen se proporciona una tabla que contiene características de personas que aplicaron para solicitar un crédito, adicional a esa información contamos con una variable "tgt" que representa a nuestra variable objetivo que toma el valor de 1 si se otorgó el crédito y 0 si se rechazó la solicitud. :

	ID_CLIENT	ID_SHOP	SEX	MARITAL_STATUS	AGE	QUANT_DEPENDANTS	EDUCATION	FLAG_RESIDENCIAL_PHONE	AREA_CODE_RESIDENCIAL_PHONE
0	2	15	F	S	18.0	0	NaN	Y	31
1	4	12	F	C	NaN	0	NaN	N	31
2	5	16	F	S	28.0	0	NaN	Y	31
3	6	24	M	S	26.0	0	NaN	N	31
4	7	55	F	S	22.0	0	NaN	Y	31
...	...	...	...	...	...	...	...	...	...
50995	29197	16	F	S	72.0	0	NaN	Y	31
50996	43802	23	F	S	31.0	0	NaN	Y	31
50997	1881	4	F	C	51.0	0	NaN	Y	31
50998	42857	25	F	C	71.0	0	NaN	Y	50

   El diccionario de datos se anexa en un excel donde podrá encontrar la descripción de cada variable.

   Tabla "Ocupaciones"

PROFESSION_CODE	PROFESSION
999	Healthcare Practitioners and Technical
950	Healthcare Practitioners and Technical
13	Educational Instruction and Library
205	Production
703	Educational Instruction and Library
...	...
152	Production
516	Life, Physical, and Social Science
50	Educational Instruction and Library
850	Life, Physical, and Social Science
726	Educational Instruction and Library

Tabla que contiene la profesión además del código de profession

2. Calidad de datos

   - Realice el etiquetado de las variables de acuerdo a su tipo
   - Revisar y eliminar duplicados , mantenga el primer elemento de los duplicados
   - Completitud
   - Revisión de valores fuera de la naturaleza de la variables (no validos) y conversion a NaN
   - Completitud
   - Realice la limpieza de variables y haga transformaciones a tipo de dato int o float en continuas (haga normalización de categorías si es necesario)
   - Eliminación de variables que posean un completitud inferior al 80%
   - Cruce con la tabla de ocupaciones

   Los resultados o hallazgos encontrados en cada punto de la sección de calidad de datos , debe estar comentada dentro del código, por ejemplo para el primer punto : "Contamos con 4 variables de tipo continuo que se muestran a continuación : ... "

3. Análisis Exploratorio de Datos

   Realice análisis interesantes sobre los datos proporcionados , generé gráficas representativas. Los análisis deben estar acompañados de una descripción corta dentro del código.

   TRATAMIENTO I
   Realice dos copias de la tabla limpia y haga lo siguiente con una de las tablas que debe llevar por nombre tratamiento_1:

4. Datos anómalos

   - Realice la identificación de datos anómalos y elimine aquellos que sean catalogados como outliers.
   - Se debe añadir los gráficos del histograma antes de la remoción y después de la remoción de outliers de todas las variables continuas.

5. Datos faltantes

   - Genere su conjunto de entrenamiento y prueba estratificado , donde el conjunto de prueba tenga el 30% de la información
   - Realice la imputación de valores ausentes , sobre las variables que lo requieran (para las continuas seleccione uno de los métodos posibles y realice KS para conocer el mejor valor a ocupar)
   - En el PDF agregue las variables imputadas y el valor que se le asignó a los valores faltantes, además de mencionar que método utilizó.

6. Ingeniería de datos

   - Genere nuevas variables a partir de la información que posee
   - En el código indique qué variables se crearon

7. Reducción de dimensiones

   - Elimine variables a través de los métodos vistos en clase
   - Indique en el código que variables se eliminaron
   - Tome solo las variables continuas y utilice PCA para generar una visualización en 3D de las variables continuas con la distinción de los valores de la target. Además ¿cuánta varianza explican esos tres componentes de las variables continuas?

Al final del primer tratamiento debe tener una tabla de entrenamiento y una de prueba con las variables "limpias" y las nuevas variables creadas.

TRATAMIENTO II – OPCIONAL (PUNTOS EXTRA)

Con la segunda copia que debe llevar por nombre tratamiento_2 realice lo siguiente:

1. Transformación Entropica:

   - Convertir continuas a categóricas
   - Valores faltantes a una nueva categoría denominada "Missings"
   - Generar conjunto entrenamiento y prueba, donde el conjunto de prueba contenga en 30% de la información de la tabla tratamiento_2
   - Obtener los WOE de cada variable calculados solo con el conjunto de entrenamiento y aplicado a los conjuntos de entrenamiento y prueba
   - Obtener el IV de todas las variables
   - Obtener el tamaño del efecto con V_Cramer

   Al terminar el flujo deberá tener dos tablas: entrenamiento y prueba con los valores woe de cada variable además de la target.

FEEDBACK
En classroom tendrán una tarea llamada feedback
Por favor aporte comentarios sobre el curso (se asignará una tarea en classroom donde podrá añadir sus comentarios), el ponente y las clases. El objetivo es poder mejorar los contenidos y el desarrollo del modulo.

ENTREGABLES

- Código en python (Notebook) , limpio, ordenado , comentado y bien estructurado, sin errores en el código.

Consideraciones

- El feedback es obligatorio para poder asentar calificación