{desafío} latam_

Prueba 2: Analizado los crímenes en la Ciudad de Nueva York

En esta ocasión trabajaremos con datos públicos del departamento de policía de New York. El dataset es llamado `stop_and_frisk_data` y contiene información sobre interrogaciones y detenciones realizadas por le departamento de policia de NY en la vía pública. El diccionario de atributos se encuentra en el archivo `2009 SQF File Spec.xlsx`.

Para todo nuestro estudio utilizaremos los datos correspondientes al año 2009 como conjunto de entrenamiento y los datos del 2010 como conjunto de pruebas. Hay que hacer notar que los datos que estamos utilizando son un muestreo del 1 de la cantidad de registros reales que contiene el dataset, esta decisión fue tomada debido a los largos tiempos de entrenamiento y procesamiento que requiere el volumén de datos reales.

- Crea una carpeta de trabajo y guarda todos los archivos correspondientes (notebook, archivos auxiliares y csv).
- Una vez terminada la prueba, comprime la carpeta y sube el `.zip` a la sección correspondiente.

Objetivos

Para alcanzar el objetivo general, su trabajo se puede desagregar en los siguientes puntos:

1. Dado la gran cantidad de atributos, se le entrega un script `preproc_nyc_sqf.py` que normaliza la cantidad de atributos. Haga uso de la función `create_suitable_dataframe` para igualar el benchmark de los atributos.

2. Debe analizar de forma exploratoria los atributos. Reporte la cantidad de datos perdidos y presente su esquema de recodificación.

3. Generar un modelo predictivo que **condicional** a las características medidas del sospechoso, prediga si un determinado procedimiento concluirá en un arresto o no. Para ello, guíase por los siguentes lineamientos:
    - Entrene por lo menos 1 modelo que sea capaz de predecir si se producirá un arresto o no. Una vez que encuentre un modelo satisfactorio, reporte al menos dos métricas de desempeño.
    - Refine aquellos atributos relevantes. Encuentre por lo menos 30 atributos que explique la importancia relativa y ordénelos por orden de importancia.

- Finalmente, reporte la probabilidad que un individuo sea arrestado en uno de los cinco barrios, condicional al género y condicional a la raza.

4. Genere al menos cinco modelos predictivos que permitan determinar si el procedimiento policial concluirá en alguna acción violenta.
    - Para ello, debe generar un nuevo atributo como vector objetivo que indique cuándo hubo violencia o no. Éste debe ser creado a partir de atributos existentes que indiquen el tipo de violencia. El detalle de los atributos que se consideran violentos se detalla a continuación:

```
'pf_hands', 'pf_wall', 'pf_grnd', 'pf_drwep', 'pf_baton',
'pf_hcuff', 'pf_pepsp', 'pf_other'
```

5. Seleccione los 2 mejores modelos, serialicelos y envíelos a evaluación. Recuerde que el modelo serializado debe ser posterior al `fit`, para poder ejecutar `predict` en los nuevos datos.

6. La evaluación del modelo será realizada en función a un conjunto de datos reservados al cual no tienen acceso.

Evaluación

La siguiente rúbrica detalla los elementos de evaluación:

- Notebook (**20 puntos**): El notebook debe ser un reporte con la estrategia analítica, explicando los siguientes puntos:
    - La definición de los requerimientos, la definición del vector objetivo, la definición de las métricas a utilizar.(**3 puntos**)
    - Un análisis exploratorio (univariado y gráfico). Como mínimo, debe analizar el comportamiento del vector objetivo antes del preprocesamiento y posterior al procesamiento.(**5 puntos**)
    - La estrategia de preprocesamiento/feature engineering.(**2 puntos**)
    - La elección de los algoritmos a implementar, así como sus hiperparámetros. Un reporte sobre qué modelos enviarán a competencia. (**10 puntos**)

- Modelos serializados:
    - Los modelos deben estar serializados con la siguiente nomenclatura: `nombre_grupo-modelo-1` y `nombre_grupo-modelo-2`.

La evaluación de los modelos serializados se realizará en función al desempeño predictivo del modelo en un conjunto de datos externos.

La primera instancia es evaluar los dos modelos enviados a competencia por el grupo, preservando el mejor modelo para la competencia con los otros grupos.

El segundo paso es rankear según desempeño entre grupos.

> El alumno debe obtener un mínimo de 16 puntos para aprobar