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

- Para realizar esta prueba debes haber estudiado previamente todo el material disponibilizado correspondiente al módulo.
- Una vez terminada la prueba, comprime la carpeta que contiene el desarrollo de los requerimientos solicitados y sube el .zip en el LMS.
- Puntaje total: 10 puntos.
- Desarrollo prueba:
    - La prueba se debe desarrollar de manera Individual.
    - Para la realización de la prueba necesitarás apoyarte del archivo Apoyo Prueba2.zip.

Contexto

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 el departamento de policía de NY en la vía pública. El diccionario de atributos se encuentra en el archivo `2009 SQF File Spec.xls`.

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 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 volumen de datos reales.

Objetivos

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

1. Debe analizar de forma exploratoria los atributos. Reporte la cantidad de datos perdidos y presente su esquema de recodificación. Tenga presente que lo que observe en el análisis exploratorio debe guiar su proceso de ingeniería de atributos, por lo que se le recomienda que piense en aspectos de las variables involucradas que puedan afectar el proceso mencionado.

2. Reporte la probabilidad de que un individuo sea arrestado en uno de los cinco barrios, condicional al género y a la raza. Concluya, ¿qué implicancias éticas tienen algunas conclusiones de lo que observa?.

3. Entregue un modelo predictivo que prediga efectivamente si un determinado procedimiento concluirá en un arresto o no. Para ello, guíate por los siguientes lineamientos:

    - Entrene por lo menos 3 modelos que sean capaces 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 con alguna estrategia que crea conveniente y reporte por lo menos 5 atributos relevantes para realizar la predicción.

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.

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

Evaluación

La siguiente rúbrica detalla los elementos que se evaluarán en el entregable final (Hito 3) de la prueba, es decir, una vez que su trabajo esté completo:

- Notebook **(9 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.**(1 Punto)**
    - Un análisis exploratorio (univariado y gráfico). Como mínimo, debe analizar el comportamiento del vector objetivo antes del preprocesamiento y posterior al procesamiento.**(3 Puntos)**
    - La estrategia de preprocesamiento/feature engineering.**(3 Puntos)**
    - La elección de los algoritmos a implementar, así como sus hiper parámetros. Un reporte sobre qué modelos se seleccionarán. **(2 Puntos)**
- Modelos serializados **(1 Punto)**:
    - Los modelos deben estar serializados con la siguiente nomenclatura: `nombre_grupo-modelo-1` y `nombre_grupo-modelo-2`.

El alumno debe obtener un mínimo de 7 puntos para aprobar.