

EVALUACIÓN N°5 (ABPRO)

Plan Formativo	Nivel de Dificultad:
Ciencia de Datos	Medio / Alto
Nombre del proyecto: Evaluación Módulo 5	Tema: Aprendizaje de máquina supervisado
Objetivo del proyecto: (Competencias del Módulo):	Elaborar un modelo predictivo a partir de un set de datos utilizando técnicas de aprendizaje de máquina supervisado implementados en lenguaje Python para resolver un problema.
Ejecución: Grupal	
Descripción del ejercicio	
<p>Utilizaremos un set de datos de las policias de New York del año 2009 y 2010. (2009_1perc.csv y 2010_1perc.csv) los cuales nos darán información de los procedimientos policiales realizados. Además se le entregará el diccionario de variables para que pueda consultar que significa cada categoría dentro de las variables.</p> <ul style="list-style-type: none"> La variable respuesta 'arstmade' informa si los procedimientos policiales han terminado en arresto o no, y el objetivo será realizar un modelo de Machine Learning para predecir si un futuro procedimiento terminará en arresto. <p>1.- Enliste todas las librerías que utilizará Nota: Se recomienda ir actualizando la lista conforme las necesidades vaya teniendo durante el desarrollo de la prueba</p> <p>2.- Importación y revisión de los datos Importe ambos sets. Dado que la fuente de datos proviene de la misma base, tienen las mismas columnas. Consolide ambos sets y reporte una exploración <i>básica</i> de los datos (numero de filas/columnas, tipos de datos, estadísticas básicas, casos perdidos)</p> <p>3.- Preprocesamiento de datos Habrá notado que los datos parecen tener ciertas inconsistencias. Siga los siguientes pasos para limpiar este set:</p> <ul style="list-style-type: none"> 3.1 Obtenga una lista con todas las variables categoricas que tengan entre 2 y 99 categorías (inclusive). (hint: son las variables tipo categoricas) 3.2 Reemplace las siguientes clases faltantes: <p>-Si alguna categoría de las columnas officrid, offshld o offverb es igual a "" cambielo a 'N' y en caso contrario dejelo como 'Y'</p>	

-Si alguna categoría de las columnas sector, trhsloc o beat es igual a "" (o NA, dependiendo de como haya catgado la base de datos), cambielo a 'U' y en caso contrario mantenga su valor

Nota, los valores significan {N: No, Y: Yes, U: Unknown}

- 3.3 Transforme las columnas ht_feet junto con ht_inch en una única columna (de la forma "ht_feet.ht_inch") llamado 'meters' (hint: transforme con el siguiente cálculo: metros = (pies+pulgadas)*0.3048)
- 3.4 Note que la fecha viene en un formato MMDDAAAA en la columna datestop. Genere 2 nuevas columnas llamadas month y year que solo tenga el mes y el año respectivamente.

Hint: Explore la libreria lubridate <https://www.rstudio.com/resources/cheatsheets/>

- 3.5 Filtre su DataFrame y solo deje las columnas seleccionadas en el punto 3.1, el mes, el año, los metros y la edad. Luego solo deje los registros cuyas edades esten entre 18 y 100 años, ambos inclusive.

4.- Análisis exploratorio

- 4.1 Estudie la variable respuesta por si sola (arstmade), puede ayudarse de un gráfico. Comente
- 4.2.- Estudie la relación de la variable respuesta en comportamiento con la raza (race), comente.
- 4.3 Estudie la relación de la variable respuesta en comportamiento con la sexo (sex), comente.
- 4.3 Estudie la relación de la variable respuesta en comportamiento con la sexo y la edad en su conjunto, comente.
- 4.4 Recodifique la variable respuesta a 1 y 0. Donde 0 es N y 1 es Y
- 4.5 Muestre en un gráfico la probabilidad que un individuo sea arrestado,condicional al género y a la raza. ¿qué implicancias éticas tienen algunas conclusiones de lo que observa?.

5.- Determinar si el procedimiento policial concluirá en alguna acción violenta.

Los atributos que tienen el prefijo pf (['pf_hands'], ['pf_wall'], ['pf_grnd'], ['pf_drwep'], ['pf_ptwep'], ['pf_baton'], ['pf_hcuff'], ['pf_pestsp'] y ['pf_other']) indican si hubo fuerza fisica utilizada por el oficial al momento del procedimiento, con la marca 'Y'.

Genere una nueva variable llamada 'violencia' la cual sea 1 si en cualquiera de las 9 variables pf hubo alguna 'Y', y 0 en otro caso. Luego indique el porcentaje de casos que terminaron con violencia.

6.- Modelación

- 6.1 Genere las variables dummies correspondientes (Tenga cuidado de no utilizar variables que expliquen lo mismo, ¡recuerde que acaba de crear una variable a partir de otras!, además recuerde que creó una variable numérica que es una categoría :)). Luego genere los sets de train-test utilizando el año 2009 para entrenar, y el año 2010 para testear.

- 6.2 Entrene 2 modelos de clasificación y reporte el mejor modelo bajo algún criterio.

Bonus (20 pts)

¿Que puede hacer para mejorar la predicción de los modelos?

Requerimientos de los participantes		
Conocimientos previos <ul style="list-style-type: none">• Inferencia Estadística	Actitudes para el trabajo <ul style="list-style-type: none">• Cumplimiento de plazos• Buenas prácticas de codificación• Diseño y Estructura• Trabajo en equipo• Optimización del tiempo	Valores Tiempo de resolución. Enfoque al requerimiento. Estructura de Solución.
Duración del proyecto	1 semana	
Tips o listado de Preguntas Guía		
- Modelos Supervisados		
Productos para obtener durante la realización del proyecto		
Documento de conclusión con análisis de la información.		
Especificaciones de desempeño		



Deberá realizar la actividad según requerimientos técnicos y en un plazo máximo de 1 clase; el resultado deberá ser un documento de conclusión con análisis.