Step-by-step

Acquisition de datos:

Los datos provienen del DANE y corresponden a la misión del "Empalme de las Series de Empleo, Pobreza y Desigualdad - MESE". Este conjunto de datos incluye cuatro bloques divididos en entrenamiento y prueba, tanto a nivel de hogar como individual. Es posible utilizar la variable *id* para vincular los hogares con los individuos. Los datos pueden obtenerse en el siguiente enlace [Kaggle - Uniandes BDML 2024 PS-2](https://www.kaggle.com/competitions/uniandes-bdml-2024-20-ps-2) o accediendo al repositorio en la sección *stores*.

Limpieza de datos:

Siga los scripts proporcionados para limpiar los datos extraídos. Ejecute el script de procesamiento test, o si prefiere trabajar con la base ya filtrada puede encontrarla en stores como en un archivo comprimido llamado “bases de dato finales”. En este paso también se realizan ciertas trasformaciones a los datos, se recomienda leer el script para entender los datos con los que se trabaja.

Modelos predictivos:

En scripts están todos los tipos de modelos utilizados en el trabajo, cada script se llama según el tipo de modelo implementado, si se busca replicar algunos de los resultados de algún modelo en especial es necesario entrar el script respectivo.

Para entrenar los modelos usamos los datos de train y separamos la muestra en 70:30 para calibrar la capacidad predictiva de los modelos. Una vez teníamos un modelo satisfactorio realizábamos una predicción sobre los datos Test. El objetico para calibrar los modelos a grandes rasgos fue aumentar el accuracy y reducir el RMSE.

Estadísticas Descriptivas:

En la sección de scripts el archivo de exploracion preliminar explica las primeras aproximaciones a los datos y como las variables se relacionaban entre si. Pueden seguirlo para replicar las gráficas descriptivas.