Reporte sobre el modelo de Clasificación mediante algoritmo de Regresión Logística

Teniendo en cuenta los conjuntos de datos “train” y “test” queremos predecir la clase a la que corresponde un municipio cuando se tienen unas variables predictoras tales como “VELOCIDAD_SUBIDA” y “VELOCIDAD_BAJADA” se quiere predecir la etiqueta (en este caso la clase) sobre la columna “MATRICULA”.

A continuación se describe de manera general el proceso implementado en el modelo predictivo, el cual utiliza un modelo de regresión logística para clasificar datos a futuro.
A continuación se detalla cada uno de los pasos realizados en el código:

1. Importación de librerías:
Se importan las bibliotecas necesarias para la manipulación de datos (`pandas`), la división de conjuntos de datos (`train_test_split`), la regresión logística (`LogisticRegression`), el escalado de características (`StandardScaler`), y la evaluación del modelo (`accuracy_score`, `classification_report`, `confusion_matrix`).

2. Carga de datasets:
Se cargan los conjuntos de datos desde archivos CSV y Excel utilizando `pd.read_csv` y `pd.read_excel`, respectivamente. El conjunto de entrenamiento se almacena en la variable `train` y el de prueba en `test`.

3. Limpieza de datos:
Se limpia la columna `'MATRICULA'` del conjunto de entrenamiento para eliminar espacios adicionales utilizando el método `str.strip()`.
Se define un diccionario de reemplazo (`reemplazos`) para convertir las categorías de la columna `'MATRICULA'` en valores numéricos que faciliten la clasificación.

4. Reemplazo de valores en la columna 'MATRICULA':
Se aplican los reemplazos definidos en el diccionario a la columna `'MATRICULA'`, convirtiendo las categorías en números y asignando `None` a los valores que son "No aplica".

5. Eliminación de filas con valores NaN:
Se eliminan las filas del conjunto de entrenamiento que tienen valores NaN en la columna `'MATRICULA'` para asegurar que el modelo no se vea afectado por datos incompletos.

6. Verificación de cambios en la columna 'MATRICULA':
Se imprime un conteo de los valores en la columna `'MATRICULA'` para verificar que la limpieza se realizó correctamente.

7. Selección de características y variable objetivo:
Se definen las características (variables independientes) en `X`, que incluyen `'ZONA_URBANA'`, `'ZONA_RURAL'`, `'VELOCIDAD_SUBIDA'`, y `'VELOCIDAD_BAJADA'`, mientras que la variable objetivo (dependiente) se almacena en `y`.

8. División de datos:
Los datos se dividen en conjuntos de entrenamiento y prueba utilizando `train_test_split`, con un 20% de los datos reservados para la prueba y un `random_state` establecido para reproducibilidad.

9. Escalado de datos: Se aplica `StandardScaler` para escalar las características de velocidad (`''VELOCIDAD_SUBIDA''` y `''VELOCIDAD_BAJADA''`), tanto en el conjunto de entrenamiento como en el de prueba, para normalizar los datos y mejorar el rendimiento del modelo.

10. Verificación de dimensiones de los conjuntos: Se imprimen las dimensiones de los conjuntos de entrenamiento y prueba para confirmar que la división se realizó correctamente.

11. Creación del modelo de Regresión Logística: Se inicializa un modelo de regresión logística utilizando `LogisticRegression()`.

12. Ajuste del modelo: Se entrena el modelo utilizando el conjunto de entrenamiento a través del método `fit`, donde se ajustan los parámetros del modelo a los datos.

13. Predicción: Se realizan predicciones en el conjunto de prueba utilizando el método `predict`.

14. Evaluación del modelo: Se evalúa el rendimiento del modelo calculando la precisión (`accuracy_score`), la matriz de confusión (`confusion_matrix`), y el informe de clasificación (`classification_report`), que proporciona métricas detalladas sobre el rendimiento del modelo.

15. Resultados: Se crea un DataFrame para almacenar las predicciones junto con los valores reales del conjunto de prueba. Se exportan los resultados de las predicciones a un archivo CSV llamado `''resultados_predicciones.csv''`.
El informe de clasificación se convierte en un DataFrame y se exporta a un archivo Excel llamado `''informe_clasificacion.xlsx''`.

Conclusiones:

El modelo de regresión logística ha mostrado un excelente desempeño en la clasificación de la columna `''MATRICULA''`, logrando una precisión del 100% en el conjunto de prueba. Después de la limpieza de datos, se encontraron 776 instancias de la clase `0.0`, 183 de la clase `1.0` y 45 de la clase `2.0`, reflejando un desbalance en la distribución de clases. La matriz de confusión indica que no hubo falsos positivos ni falsos negativos, y el informe de clasificación muestra que todas las métricas (precisión, recall y f1-score) son de 1.0 para las clases relevantes.