Máster en Big Data y Business Analytics
Universidad Complutense de Madrid
Minería de datos y Modelización Predictiva

El conjunto de datos DatosImpuestos_Tarea.xlsx contiene información socio-económica de una selección aleatoria simple de ayuntamientos. Estos ayuntamientos son identificados con su CIF, pero se ha omitido la información de esta variable. Este conjunto de datos, además, contiene información sobre las viviendas y las personas que habitan en el municipio asociado al correspondiente ayuntamiento. Para estos ayuntamientos se dispone de los impagos asociados a los distintos impuestos municipales en el último año, lo que da lugar a las posibles variables objetivo:

ActEconom_Cuali: Variable dicotómica que toma el valor 1 si el impago del impuesto de actividad económica fue superior el año en curso que el anterior y, 0, en otro caso.

ActEconom: Porcentaje de impagos impuesto de actividad económica.

Basura: Porcentaje de impagos impuesto de basura.

Vehiculo: Porcentaje de impagos impuesto de vehículos.

Vivienda: Porcentaje de impagos impuesto de vivienda.

Basura_Cuali: Variable dicotómica que toma el valor 1 si el impago del impuesto de basura fue superior el año en curso que el anterior y, 0, en otro caso.

Vehiculo_Cuali: Variable dicotómica que toma el valor 1 si el impago del impuesto de vehículos fue superior el año en curso que el anterior y, 0, en otro caso.

El objetivo de esta práctica es obtener dos modelos de regresión (lineal y logística) seleccionando, de entre las variables objetivo anteriores, una variable objetivo continua y otra binaria, respectivamente (no olvides rechazar las variables que no escojas como objetivo en cada modelo). Antes de desarrollar los modelos de predicción, es necesario llevar a cabo un proceso de depuración de los datos. Los pasos a seguir para la realización de la práctica son:

1. Introducción al objetivo del problema y las variables implicadas.
2. Importación del conjunto de datos y asignación correcta de los tipos de variables.
3. Análisis descriptivo del conjunto de datos. Número de observaciones, número y naturaleza de variables, datos erróneos etc.

4. Corrección de los errores detectados.
5. Análisis de valores atípicos. Decisiones.
6. Análisis de valores perdidos. Imputaciones.
7. Transformaciones de variables y relaciones con las variables objetivo.
8. Detección de las relaciones entre las variables input y objetivo.
9. Construcción del modelo de regresión lineal.
    - Selección de variables clásica
    - Selección de variables aleatoria
    - Selección del modelo ganador
    - Interpretación de los coeficientes de dos variables incluidas en el modelo, una binaria y otra continua
    - Justificar porqué es el mejor modelo y medir la calidad del mismo

10. Construcción del modelo de regresión logística.
    - Selección de variables clásica
    - Selección de variables aleatoria
    - Selección del modelo ganador
    - Determinar el punto de corte óptimo
    - Interpretación de los coeficientes de dos variables incluidas en el modelo, una binaria y otra continua
    - Justificar porqué es el mejor modelo y medir la calidad del mismo

Se entregará un informe en PDF (máximo 20 páginas, la portada y el índice no están incluidas, cualquier página adicional no se tendrá en cuenta) en el que se explicarán detalladamente los pasos seguidos incluyendo los códigos y salidas más relevantes. Imprescindible mostrar los modelos finales (summary). Es muy importante comentar y justificar razonadamente las decisiones que se toman.

La puntuación de la tarea está dividida en tres partes:

- Depuración de datos (3,3 puntos). Todos los apartados de esta parte tienen la misma puntuación.
- Regresión Lineal (3,3 puntos). Todos los apartados de esta parte tienen la misma puntuación.
- Regresión Logística (3,4 puntos). Todos los apartados de esta parte tienen la misma puntuación.