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

El conjunto de datos DatosEleccionesEspaña.xlsx contiene información demográfica sobre los distintos municipios de España junto con los resultados que se obtuvieron en las últimas elecciones. Existen 7 posibles variables objetivo:

AbstentionPtge: Porcentaje de abstención

Izda Pct: Porcentaje de votos a partidos de izquierda (PSOE y Podemos)

Dcha Pct: Porcentaje de votos a partidos de derecha (PP y Ciudadanos)

Otros Pct: Porcentaje de votos a partidos distintos de PP, Ciudadanos, PSOE y Podemos

AbstencionAlta: Variable dicotómica que toma el valor 1 si el porcentaje de abstención es superior al 30 % y, 0, en otro caso.

Izquierda: Variable dicotómica que toma el valor 1 si la suma de los votos de izquierdas es superior a la de derechas y otros y, 0, en otro caso.

Derecha: Variable dicotómica que toma el valor 1 si la suma de los votos de derecha es superior a la de izquierda y otros y, 0, en otro caso.

El objetivo de esta práctica es obtener dos modelos de regresión (lineal y logística) seleccionando, de entre las 7 variables 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.