PRÁCTICA

DATA MINING

Dolores Lorente Muñoz

Práctica DM

OBJETIVO

El objetivo de la práctica es abordar un problema de data mining *realista* siguiendo la metodología y buenas prácticas explicadas durante las clases teóricas.

Dicho lo cual, en estas instrucciones no se especifican los pasos exactos que el alumno tiene que llevar a cabo para realizar esta tarea con éxito. Es parte del trabajo aplicar las técnicas de procesamiento/transformación de variables que mejor se acondicionen al problema, identificar los modelos adecuados que proporcionen buenas prestaciones así como las variables potencialmente más relevantes, y elegir la métrica adecuada para contrastar los distintos modelos.

Las posibilidades son amplias, así que es recomendable abordar una aproximación incremental, esto es, comenzar por soluciones sencillas para progresivamente aumentar la complejidad de las técnicas utilizadas.

DATOS

Se trabajará con dos conjuntos de datos distintos, ambos datasets los ofrece SAS para las prácticas de data mining, Se pretende resolver dos problemas:

1. Modelización

Fuente: sashelp.bweight

Objetivo: **predecir el peso infantil al nacer un niño** (la target será la variable *weight*) a partir de las variables analíticas recogidas a tal fin.

Dataset:

Weight	Infant Birth Weight	Peso al nacer infantil (Target continua)
Black	Black Mother	Madre de color (1 si, 0 no)
Boy	Baby Boy	Bebé (1 niño, 0 niña)
CigsPerDay	Cigarettes Per Day	Número de Cigarrillos por día
Married	Married Mother	Madre casada (1 si, 0 no)
MomAge	Mother's Age	Edad de la madre (equivale el 0 a tener 25 años)
MomEdLevel	Mother's Education Level	Nivel de educación de la madre (de 0 a 3, siendo 0 nada)
MomSmoke	Smoking Mother	Madre fumadora (1 si, 0 no)
MomWtGain	Mother's Pregnancy Weight Gain	Embarazo de la madre pérdida/aumento de peso
Visit	Prenatal Visit	Visita prenatal (de 0 a 3, siendo el 0 ninguna visita)

2. Clasificación

Fuente: sashelp.birthwgt

Objetivo: **identificar a la población** que es más propensa a sufrir el efecto de la mediación del bajo peso al nacer sobre la mortalidad infantil (LowBirthWgt) a partir de las variables analíticas.

Dataset:

LowBirthWgt	**Low Birth Weigt**	**Bajo peso al nacer (Target dicotómica: yes, no)**

AgeGroup	Mother's Age Group	Grupo de edad (niveles:1, 2, 3)
Death	Death	Muerte (yes, no)
Drinking	Mother's Drinking	Madre alcoholica (yes, no)
Married	Married Mother	Madre casada (yes,no)
Race	Race	Raza (niveles: asian, black, hispanic, native, white)
Smoking	Smoking Mother	Madre Fumadora (yes, no)
SomeCollege	SomeCollege	Alguna educación superior (yes, no)

FORMATO DE ENTREGA

El alumno abordará los dos problemas en un mismo documento **Word y adjuntará el código desarrollado**, que será guardado en la carpeta drive correspondiente y cuyo enlace se incluirá en el fichero de revisión de prácticas.

Este archivo no debe de contener únicamente código, sino que **todos los pasos dados y las decisiones tomadas han de estar justificadas convenientemente**.

Se valorará positivamente por tanto, la comunicación de resultados (vía gráficas y explicaciones escritas), así como la interpretación de los resultados obtenidos.

FECHA DE ENTREGA

La fecha de entrega seguirá las condiciones fijadas en el Bootcamp.