Consigna obligatorio
Analítica de negocios y Big Data
Marzo-2020

Se presenta una Base de Datos respecto al relevamiento de 6 variables en 150 hogares de Montevideo, el detalle de las variables se presenta a continuación:

- ID: Identificador único de las observaciones.
- Ant.Lab: Antigüedad Laboral promedio de los hogares.
- Hab: Cantidad de habitantes promedio en el hogar.
- Edu.Ter: Cantidad de años de Educación Terciaria en el hogar.
- Ingreso: Ingreso promedio de los hogares (valores cada 10.000 pesos).
- Nivel.SE: Nivel Socioeconómico, en tres categorías Bajo, Medio y Alto.

En base a la Base de Datos anterior, se pide:

1) En base a la información proporcionada, es posible elaborar un modelo que sirva para predecir el Ingreso? En caso de ser posible, detallar el mismo respetando las fases de un proyecto analítico (sugerencia, elaborar un modelo de regresión lineal).

   Deberá incluir los siguientes aspectos y sus **interpretaciones**:
   - Análisis descriptivo.
   - Análisis de correlación.
   - Al menos dos gráficos de dispersión de la variable dependiente contra otras variables y sus respectivas rectas de regresión simple. Al menos un gráfico de caja y un histograma.
   - Al menos una tabla de frecuencias absolutas (casos) y otra relativa (porcentajes).
   - Regresión con todos las variables.
   - Regresión con las variables finales. La regresión final deberá ser encontrada aplicando un método automático de selección de atributos.
   - Análisis de bondad de ajuste del modelo en testing.
   - Conclusión e interpretación del modelo.

2) En base a la información proporcionada, es posible elaborar un modelo que sirva para predecir cuando un hogar es de nivel socioeconómico Alto? En caso de ser posible, detallar el mismo respetando las fases de un proyecto analítico (sugerencia, elaborar un modelo de regresión logística y un árbol de clasificación).

   Deberá incluir los siguientes aspectos y sus **interpretaciones**:
   - Al menos dos gráficos de caja y dos histogramas relacionados con la variable a predecir.

- Al menos una tabla de frecuencias absolutas (casos) y otra relativa (porcentajes) que relacionen cualquier variable con la variable dependiente.
- Regresión con todos las variables.
- Regresión con las variables finales. La regresión final deberá ser encontrada aplicando un método automático de selección de atributos.
- Gráfico de la complejidad para el podado del árbol.
- Gráfico del árbol de clasificación.
- Determinación de las reglas inducidas por el árbol.
- Análisis de bondad de ajuste de ambos modelos en testing: área bajo la curva ROC y gráfico de curva ROC.
- Matriz de confusión en test para ambos modelos utilizando como punto de corte la probabilidad 0.5.
- Conclusión sobre los resultados y selección del mejor modelo.

3) El gerente del área nos informa que la variable nivel socioeconómico es muy complicada de medir en la práctica, por lo que no solicita si podemos hacer un modelo que nos clasifique el nivel socioeconómico sin considerar la misma en el armado del modelo (sugerencia, elaborar un modelo de cluster kmeans).

   Deberá incluir los siguientes aspectos y sus interpretaciones:
   - Justificación de la estrategia de estandarización de los datos si es que se realiza.
   - Justificación de la cantidad de clusters elegidos.
   - Diagramas de caja que relacionen todas las variables utilizadas con los clusters construidos. Todo en un panel.
   - Interpretación de los clusters.

Se espera que cada grupo entregue un documento (pdf) autocontenido donde se presenten los resultados obtenidos con los respectivos análisis y un código de R mediante el cual sea posible replicar los resultados del documento.