Unidad 6 - Ajustes y modelos

Fundamentos de ciencia de datos

Ejercicio N° 1

El dataset `student_data.csv` contiene información sobre el promedio de horas semanales de estudio que un conjunto de estudiantes dedicó para la preparación de un examen y la calificación final recibida luego de rendirlo.

1. Realice un gráfico que le permita visualizar la relación existente entre las calificaciones finales y las horas de estudio del grupo de estudiantes. ¿Cómo describiría dicha relación a partir de lo observado?
2. A partir de los datos, ajuste un modelo de regresión lineal para la calificación final del examen en función de las horas semanales de estudio.
3. Escriba dicho modelo en forma desarrollada e interprete cada uno de los coeficientes que lo componen.

Ejercicio N° 2

Utilizando el dataset `dataset_rendimiento.csv`, que contiene información sobre 235 estudiantes universitarios de segundo año de una institución universitaria en relación a las horas semanales promedio de estudio dedicadas, el tipo de desayuno que consumen antes de los exámenes y el rendimiento medio en los exámenes parciales:

1. Ajuste un modelo de regresión lineal que permita predecir el rendimiento promedio de estudiantes de segundo año a partir del número de horas semanales promedio de estudio y el tipo de desayuno que consumen.

2. ¿Cuál/es de las variables incluidas contribuye/n significativamente a explicar el rendimiento de los estudiantes? Justifique.
3. Utilizando el modelo ajustado, prediga el rendimiento promedio que presentará un estudiante que dedica, en promedio, 5.5 horas semanales de estudio y posee la costumbre de consumir un desayuno saludable previo a las instancias de examen.

Ejercicio N° 3

El dataset `USA_Housing.xlsx` contiene la siguiente información sobre un conjunto de 5000 viviendas en Estados Unidos:

- **avg_area_income:** ingreso promedio de los residentes de la ciudad en la que se encuentra la vivienda.
- **avg_area_house_age:** antigüedad promedio de las casas que se encuentran en la misma ciudad en la que se encuentra la vivienda.
- **avg_area_number_of_rooms:** número promedio de ambientes en las casas que se encuentran en la misma ciudad en la que se encuentra la vivienda.
- **avg_area_number_of_bedrooms:** número promedio de habitaciones en las casas que se encuentran en la misma ciudad de la vivienda.
- **area_population:** población de la ciudad de la vivienda.
- **price:** precio de venta de la vivienda.
- **address:** domicilio de la vivienda.

PARTE I - Análisis exploratorio

1. Importe el dataset al entorno de trabajo y realice cualquier tarea de limpieza y adecuación del mismo que considere necesaria para su posterior análisis.
2. Realice un gráfico que le permita visualizar la distribución del precio de venta de las residencias que componen el dataset. En base al gráfico realizado, ¿cómo describiría la distribución de la variable en cuanto a sus características de simetría?
3. Construya un gráfico que le permita evaluar el grado de asociación lineal que existe entre las distintas variables cuantitativas que componen el dataset. En base al mismo, identifique la/s variable/s que se encuentran más fuertemente

correlacionadas e informe e interprete la medida de asociación lineal correspondiente.

4. Elija uno de los pares de variables que identificó en el ítem anterior y realice un gráfico que le permita visualizar la relación general que existe entre las mismas.

PARTE II - Ajuste de modelo

El objetivo principal de esta segunda parte es analizar si es factible ajustar un modelo de regresión lineal que permita predecir el precio de venta de una vivienda en Estados Unidos a partir de la cantidad de personas que viven en la ciudad en la que ésta se ubica y su ingreso promedio, así como las características generales de las viviendas en dicho lugar (antigüedad promedio/número promedio de ambientes/número de habitaciones promedio).

1. Comience ajustando el modelo completo, incluyendo la totalidad de las variables explicativas o independientes que resultan de interés. ¿Cuál/es de las variables incluidas contribuye/n significativamente a explicar la calidad de los vinos? Justifique.
2. Ajuste un nuevo modelo tomando en cuenta su respuesta a la pregunta anterior.

- En base al ajuste realizado, escriba dicho modelo en forma desarrollada e interprete cada uno de los coeficientes.
- ¿Qué medida utilizaría para comparar los dos modelos en cuanto a su bondad de ajuste? Interprete dicha medida para el modelo final.
- Realice un gráfico de valores predichos vs. valores reales y comente lo observado.