Evaluación domiciliaria

1. Consignas para la entrega de la evaluación

- El trabajo debe estar redactado a modo de informe en un editor de texto a su elección.
- No debe contener el código utilizado.
- Toda "salida" de software y/o figura que se incluya en el informe deberá contribuir al análisis del problema, deberá ser comentada adecuadamente y referenciada en el texto.
- Las conclusiones y comentarios deben apoyarse en tablas y/o gráficos y salidas del software.
- Entregar el examen con su nombre y apellido en el encabezado.
- Entregar dos archivos. Un pdf con el informe y un archivo R con los comandos usados.
- Ambos archivos deberán tener el siguiente nombre *MAA2025-Nombre_Apellido*. Por ejemplo: *MAA2025-Juan_Perez*.
- Fecha y horario de entrega: Viernes 25 de julio hasta las 23.55 hs.

2. Evaluación

1. En el archivo *Restaurantes.csv* se ha recopilado información sobre el precio promedio de una cena y distintas variables de interés que pueden influir en dicho precio, con el objetivo de elaborar un modelo de regresión que permita predecir el precio de la cena. Las variables relevadas son:

   - *precio*: precio promedio (en dólares) de una cena con bebida y propina incluida.
   - *cal1*: calificación otorgada por los clientes a la comida (escala de 0 a 30).
   - *decor*: calificación otorgada a la ambientación y decoración del restaurante (escala de 0 a 30).
   - *servicio*: calificación del servicio recibido (escala de 0 a 30).
   - *lugar*: variable categórica que indica si el restaurante está al este o al oeste de una determinada avenida.
   - *cal2*: segunda medición independiente sobre la calidad de la comida, provista por un evaluador externo en una escala comparable.

   a) Proponer un modelo de regresión lineal que relacione el *precio* con la variable *cal1*. ¿Detecta algún punto influyente? En caso afirmativo analice un modelo sin este punto y reporte los cambios que observa.

   b) Dar la ecuación del modelo considerando lo estudiado en el ítem anterior. ¿Son significativos los parámetros del modelo?

   c) Realice un scatter plot donde se visualicen los puntos, la recta encontrada y bandas de confianza del 95 % para la media.

2. Considere los cambios realizados en el problema anterior.

   a) Ajuste un modelo de regresión lineal que explique el precio en función de todas las variables continuas presentes en el archivo ¿Detecta algún problema de multicolinealidad? ¿Cómo se manifiesta? En caso afirmativo ¿qué decisión toma respecto de mantener o no las variables correlacionadas en el modelo?

b) En base a la decisión tomada, elija el mejor modelo y analice la validez de los supuestos del mismo.

3. Ajuste ahora un modelo considerando las variables *precio*, *cal1* y *lugar* con el objetivo de estudiar el impacto de esa calificación de acuerdo al lugar donde se ubica el restaurante.

   a) ¿Es significativa la interacción? ¿Hay multicolinealidad debida a la interacción? Si hubiera multicolinealidad, realice los cambios necesarios para resolverla.

   b) Escriba los modelos resultantes en cada grupo, interprete los parámetros del modelo y realice un scatter plot con la recta ajustada en cada grupo.

4. El archivo *DatosPSA_Prostata.csv* contiene datos de 97 hombres con cáncer de próstata que se examinaron antes de una operación. Cada observación corresponde a a los datos de un paciente y contiene las siguientes variables:

   - *lcavol*: log del volumen del cáncer
   - *lpeso*: log del peso de la próstata
   - *edad*: edad del paciente
   - *lhiper*: log de hiperplasia benigna
   - *iv*: invasión vesical (binaria)
   - *lpc*: log de penetración capsular
   - *gleason*: puntaje de Gleason
   - *pgg45*: porcentaje de células con patrón 4 o 5
   - *lpsa*: log del nivel del antígeno prostático específico (PSA) – **variable respuesta**

   a) Realice un ajuste del modelo completo. ¿Todas las variables predictoras resultan significativas considerando un nivel de significación del 5 %? ¿Detecta algún punto influyente y multicolinealidad? ¿Es significativo el modelo? ¿Cuánto es el porcentaje de variabilidad explicada por el modelo?

   b) Defina los conjuntos de entrenamiento (0.80 %) y test (0.20 %) considerando la semilla definida por *set.seed*(11). Ajuste tres modelos diferentes de regresión lineal múltiple utilizando R^2, BIC y AIC como criterios para la selección de variables. ¿Qué variables resultan elegidas de acuerdo a cada criterio?

   c) Implementar una validación cruzada de $k = 5$ folds y estime las diferentes métricas en cada modelo. Presente la información en una tabla que permita visualizar las diferencias entre las diferentes metodologías.

   d) Ajustar el modelo final con las variables seleccionadas para cada uno de los criterios, usando el conjunto completo de entrenamiento. Evaluar su desempeño en el conjunto de prueba sin volver a entrenar, reportando las siguientes métricas:
      - RMSE (Root Mean Squared Error)
      - MAE (Mean Absolute Error)
      - R^2 (Coeficiente de determinación)

   e) ¿el modelo ajustado se desempeña bien sobre datos nuevos? ¿Hay sobreajuste?

   f) ¿Qué modelo elige? Justifique su respuesta.