INSTRUCCIONES ENTREGA GRUPAL

El objetivo del trabajo es demostrar que se han adquirido el máximo de conocimientos de la asignatura por lo que hacer más no implica tener más nota: lo que hagáis, que tenga sentido, y sobre todo, que esté justificado, argumentado y explicado.

Cada grupo deberá realizar (al menos) dos análisis:

1. Predicción (variable objetivo continua):
   1. Con alguno de los datasets proporcionados en el mensaje envíado al campus, deberás implementar un análisis profundo y completo en el contexto de la regresión multivariante, con la mayor profundidad posible, con todo lo visto durante el curso.
   2. Será obligatorio realizar un análisis exploratorio detallado
   3. Será obligatorio realizar un análisis previo de correlación y colinealidad, tomando las decisiones que consideres.
   4. Tras lo anterior, será obligatorio realizar un primer ajuste saturado e interpretar lo que consideres.
   5. Deberás realizar una selección de modelos basada en AIC/BIC y comentar todo lo que consideres.
   6. Deberás realizar una regresión lineal ordinaria pero con una selección de variables basada en LASSO (eligiendo la penalización mediante validación cruzada k-folds).
   7. De los modelos d) e) f) deberás comentar las diferencias obtenidas.
   8. De los modelos d) e) f) deberás elegir 2 de ellos y realizar la diagnosis en ambos casos. ¿Cumplen las hipótesis? ¿Por qué es importante su cumplimiento? Deberás realizar también un ANOVA de ambos.
   9. Además deberás realizar dos ajustes, un ajuste ridge y LASSO (eligiendo la penalización mediante validación cruzada k-folds). En este último caso ya como modelo, no como selector de variables.
   10. Con los dos modelos de h) y los dos modelos de i) deberás comparar sus su calidad tanto en train como en test, así como comentar e interpretar sus coeficientes e interpretar la inferencia obtenida (como si se cumpliesen las hipótesis).
   11. En caso de que no se cumpliesen las hipótesis en los modelos h), no es obligatorio pero se valorará muy positivamente si se intenta transformar los datos para que se cumplan.
   12. Lo anterior es solo “lo mínimo” pero puedes hacer todo lo que consideres a partir de ahí.
2. Clasificación (variable objetivo binaria, pendiente de ver en clase):
   1. Se deberá realizar al menos una regresión logística con el dataset heart disease (con la función de enlace logit)
   2. El dataset contiene datos cardiovasculares del estado de Massachusetts de 4000 con una serie de variables, entre las que se encuentra la variable objetivo (CHD) que contiene un 1 si el paciente tiene un riesgo alto de tener un accidente cardiovascular en los próximos 10 años y 0 en caso contrario.
   3. Las predictoras son:
      1. Sex: sexo
      2. Age: edad (continua aunque se hayan truncado decimales)
      3. Current smoker: binaria (si el paciente es actualmente fumador)
      4. Cigs per Day: número de cigarros fumados diarios (en media).
      5. BP meds: binaria (si el paciente toma o no medicación para la tensión)
      6. Prevalent stroke: binaria (si el paciente ha tenido o no un infarto previo)
      7. Prevalent Hyp: binaria (si el paciente es o no hipertenso)
      8. Diabetes: binaria (si el paciente tiene o no diabetes)
      9. Tot chol: colesterol
      10. Sys/Dia BP: tensión sistólica y diastólica.
      11. BMI: índice de masa corporal
      12. Heart Rate: pulso (como si fuese continua)
      13. Glucose: niveles de glucosa (continua)
   4. El objetivo es no solo predecir un 1 y un 0 sino una probabilidad de accidente (como suele ser habitual, si supera el 50% será un 1, en caso contrario será un 0)
   5. Se deberá interpretar cada uno de los coeficientes obtenidos.
   6. Usando esa interpretación, si fueses médico, ¿qué 3 consejos daríais a un paciente para prevenir un accidente cardiovascular?
   7. Es obligatorio que proporciones, al menos, una métrica de error en train y otra en test.
   8. Si te da tiempo, en este caso la diagnosis e inferencia será opcional (aunque se valorará muy positivamente)
   9. Si será obligatorio que realicéis al menos una selección de variables a partir del modelo saturado. Opciones: BIC, AIC, LASSO.
   10. Lo anterior es solo “lo mínimo” pero puedes hacer todo lo que consideres a partir de ahí.