

**Tarea 2. Introducción a Analítica. Escuela de Estadística. Facultad de Ciencias. Universidad Nacional de Colombia, Sede Medellín**

**Modo y fecha de entrega:** La entrega de las tareas es por medio electrónico, vía email, en formato pdf y se deben enviar a jcsalaza@unal.edu.co el lunes 7 de septiembre de 2020 hasta las 6:00pm; se usará el reloj de mi PC para llevar el registro de entrega. Tareas que se entreguen entre las 6:01 y 6:30pm se calificarán sobre 4:00, y tareas que se entreguen desde las 6:31pm se calificarán como reprobadas. Se pueden conformar grupos de máximo 4-5 personas. VER REVERSO DE LA PÁGINA.

1. (10 pts. Teórico) Considere el estadístico leverage:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^n (x_{i'} - \bar{x})^2}$$

Demuestre que

$$\frac{1}{n} \leq h_{ii} \leq 1$$

2. (50 pts. Práctico) Considere el conjunto de datos anexo (bank.csv) el cual tiene 17 variables. Asuma que el supervisor es la variable loan.
- a) Cree un conjunto de datos de entrenamiento del 75 % y el restante 25 % trátelo como datos de test o de prueba
  - b) Con los datos de entrenamiento, implemente Naive Bayes usando loan como el supervisor y las demás como predictores.
  - c) Con los datos de entrenamiento, implemente Knn usando loan como el supervisor y las demás como predictores. Ensaye con varios valores de K y reporte solo uno de acuerdo a su preferencia. Observe que algunas variables son categóricas y se deben crear variables dummies.
  - d) Con los datos de entrenamiento, implemente Regresión logística usando loan como el supervisor y las demás como predictores.
  - e) Con los datos de entrenamiento, implemente LDA usando loan como el supervisor y las demás como predictores.
  - f) Con los datos de entrenamiento, para cada uno de los métodos anteriores, calcule el training-MSE, la matriz de confusión y grafique la curva ROC.
  - g) Use los respectivos ajustes de cada uno de los modelos anteriores y con el conjunto de prueba, calcule el test-MSE, la matriz de confusión y grafique la curva ROC.
  - h) ¿Con cuál modelo observó un mejor desempeño y porqué?

3. (40 pts. Práctico) Considere el conjunto de datos anexo (costumer loan details.csv) el cual tiene 12 variables incluyendo el ID. Asuma que el supervisor es la variable income.
- a) Cree un conjunto de datos de entrenamiento del 75 % y el restante 25 % trátelo como datos de test o de prueba
  - b) Con los datos de entrenamiento, implemente Knn (con al menos tres valores para K) usando income como el supervisor y debts como predictor. Grafique e interprete.
  - c) Con los datos de entrenamiento, implemente regresión lineal simple usando income como el supervisor y debts como predictor. Grafique e interprete.
  - d) Use los respectivos ajustes de cada uno de los modelos anteriores y con el conjunto de prueba, calcule el **test-MSE**. ¿Qué observa?
  - e) Usando todos los datos y regresión lineal múltiple seleccione un modelo usando forward, backward y stepwise.
  - f) Seleccione uno de los modelos del paso anterior y responda con argumentación la pregunta: ¿ajusta bien dicho modelo?