Tarea Sistemas Inteligentes

La siguiente Tarea se puede hacer de manera individual o en equipos de máximo 3 integrantes.

1. Realiza una investigación comparativa, de máximo 1 cuartilla, sobre los pros y contras del uso de Python o R en el área de ciencia de datos.

2. El siguiente ejercicio utiliza la base de datos de Titanic con información de los pasajeros del famoso barco que se hundió el 15 de abril de 1912. Originalmente Kaggle lanzó el reto para predecir quién sobrevive, de acuerdo a la información de cada pasajero. Los datos fueron divididos en los archivos Train y Test. La diferencia es que el archivo Test.csv no tiene la variable Survived ya que esta información la utiliza Kaggle para seleccionar al ganador. Así, para este ejercicio por el momento solamente utilizaremos el archivo train.csv de la siguiente liga:

   https://www.kaggle.com/c/titanic/data

   Con base a la información del archivo train.csv, contesta los siguientes incisos:

   a) Indica el total de variables del archivo y el tipo de cada variable (numérica o categórica).
   b) Como se desea usar la información de las variables para construir un modelo de aprendizaje automático que nos ayude a determinar si un pasajero sobrevive o no, determina si alguna o algunas de dichas variables debieran ser eliminadas por considerar que no proporcionan información relevante. Justifica la decisión de las variables que elimines. En los incisos siguientes usa el conjunto de datos simplificado, es decir, sin las variables que eliminaste.
   c) Haz un análisis para determinar si existen datos perdidos. De ser así, indica qué variables y cuántos datos perdidos tiene cada una.
   d) Haz un análisis numérico y gráfico para determinar qué decisión tomar sobre dichos datos perdidos: eliminarlos, sustituirlos para algún valor representativo, etc. Indica y justifica la decisión que tomes.
   e) En particular, como habrás observado, la variable Age tiene casi un 20% de datos perdidos. Independientemente de la decisión que tomaste en el inciso anterior sobre qué hacer con estos datos perdidos, para esta Tarea en lo sucesivo deberás realizar dos modelos:

      Primer modelo, A: usando la variable Age y sustituyendo los datos perdidos por el valor representativo que mejor consideres;

      Segundo modelo, B: eliminando la variable Age.
   f) Iniciando con el primer modelo, A, realiza una partición del conjunto de datos en un 80% entrenamiento y un 20% de prueba. Verifica que las proporciones de cada partición tienen valores aproximados con respecto al total de sobrevivientes del archivo inicial.

g) Usando el conjunto de entrenamiento del inciso anterior, obtener un Árbol de Decisión para el primer modelo, A, de este ejercicio. Realiza una breve descripción que resuma dicho diagrama.
h) Obtener la matriz de confusión de este primer modelo A. Con base a dicha matriz:
    I. ¿Qué porcentaje de predicciones fueron correctas y qué porcentaje incorrectas?
    II. ¿Qué porcentaje de Verdaderos Positivos (VP) y Verdaderos Negativos (VN) se obtuvieron? ¿Cuál es el significado de cada uno de estos términos?
    III. ¿Qué porcentaje de Falsos Negativos (FN) y Falsos Positivos (FP) se obtuvieron? ¿Qué significado tiene cada uno de estos? ¿Cuál de estos dos errores consideras que pudiera ser el de mayor importancia y tratar de disminuirlo?
i) Repite los incisos “g”, “h”, “i” cinco veces, para obtener 5 modelos con la misma partición.
j) Obtener la matriz de confusión promedio resultante de los 5 modelos anteriores: promediando cada una de las entradas correspondientes de las 5 matrices, es decir, VN, VP, FN y FP.
k) Repite los incisos “f” a “j” usando ahora una partición de 70% en el conjunto de entrenamiento y 30% del conjunto prueba con los datos del primero modelo, A.
l) Repite los incisos “f” a “j” usando ahora una partición de 60% en el conjunto de entrenamiento y 40% del conjunto prueba con los datos del primero modelo, A.
m) Repite los incisos “f” a “l” usando ahora los datos del segundo modelo, B.
n) Realiza un resumen con base a los resultados obtenidos. En particular, indica por ejemplo cuál de todos los modelos generados consideras es el más adecuado para hacer la predicción con datos nuevos.