Tarea 2: — Métodos Estadísticos para el Manejo de Grandes Volúmenes de Datos

Profesor: Pedro Luiz Ramos — Pontificia Universidad Católica de Chile
Ayudante: Diego Andrés Bernal Soto — Facultad de Matemática

Ejercicio: La tragedia del naufragio del Titanic es uno de los naufragios más famosos de la historia.

El 15 de abril de 1912, durante su viaje inaugural, el RMS Titanic, considerado ampliamente ïnaundable", se hundió después de chocar con un iceberg. Desafortunadamente, no había suficientes botes salvavidas para todos a bordo, lo que resultó en la muerte de 1502 de los 2224 pasajeros y tripulantes.

Aunque había cierto elemento de suerte involucrado en la supervivencia, parece que ciertos grupos de personas tenían más probabilidades de sobrevivir que otros.

En este ejercicio, se solicita que construyas un modelo predictivo que responda a la pregunta: ¿Qué tipo de personas tenían más probabilidades de sobrevivir? Utilizando datos de los pasajeros (como edad, género, clase socioeconómica, etc.). El conjunto de datos está organizado de la siguiente manera:

1: Clase del Pasajero 2: Indicador de Supervivencia 3: Sexo 4: Edad 5: Número de hermanos y cónyuges a bordo 6: Número de padres e hijos a bordo 7: Tarifa pagada en libras (£) 8 : Embarcado - Puerto de Embarque

Tenga en cuenta que los datos han sido previamente limpiados para eliminar variables no deseadas y valores faltantes.

Para ello, utilice el archivo 'Tarea 2' disponible en Canvas.

a) Calcular las estadísticas descriptivas (media, desviación estándar) para el vector de variables.

b) Ajustar un modelo Naive Bayes utilizando una división que considere un conjunto de entrenamiento (80 %) y validación (20 %). Construir la matriz de confusión.

c) Repetir el procedimiento anterior utilizando un modelo de análisis discriminante lineal y cuadrático.

d) Repetir el procedimiento anterior utilizando un árbol de clasificación, random forest y XGBoosting. Indique cuáles fueron las variables más importantes para la variable respuesta.

e) Ajustar el mismo problema utilizando máquinas de soporte vectorial (SVM), considerando distintos tipos de kernel.

f) Ajustar el mismo modelo con una red neuronal, probando distintas estructuras de neuronas y capas, y comentar los resultados.

g) Discutir los resultados obtenidos y determinar cuál tuvo el mejor desempeño en términos de precisión (exactitud).

Data description:

Survival - Survival (0 = No; 1 = Yes).

Pclass - Passenger Class (1 = 1st; 2 = 2nd; 3 = 3rd)

Name - Name

Sex - Sex

Age - Age

Sibsp - Number of Siblings/Spouses Aboard

Parch - Number of Parents/Children Aboard

Ticket - Ticket Number

Fare - Passenger Fare

Cabin - Cabin

Embarked - Port of Embarkation (C = Cherbourg; Q = Queenstown; S = Southampton)