

Tarea ML

Realice las siguientes actividades

1. Con los datos de COVID hacer una selección de variables:
 - Utilizar el AIC con forward, backward y stepwise.
 - Utilizar los criterios de BIC y Mallow Cp.
 - Si tuviese que eliminar dos variables cuales serían y porqué. Use SelectKBest.
2. Ajuste una Máquina de Soporte de Vectorial a los datos del ejercicio anterior.
 - Aplique OneHotEncoder a las variables categóricas.
 - Ajuste un kernel lineal
 - Compare su resultado de ese mismo subconjunto contra un modelo de regresión logística.
 - De los modelos anteriores obtenga la matriz de confusión y la curva ROC.
 - Haga una optimización de hiperpametros priorizando la precisión.
 - Haga una gráfica de Precision-Recall vs Threshold
3. El archivo “sonar.csv” contiene 111 patrones obtenidos haciendo rebotar señales de sonar en un cilindro de metal en varios ángulos y bajo diversas condiciones. El archivo también contiene 97 patrones obtenidos de rocas en condiciones similares. Cada patrón es un conjunto de 60 números en el rango de 0.0 a 1.0. Cada número representa la energía dentro de una banda de frecuencia particular, integrada durante un cierto período de tiempo. La etiqueta asociada con cada registro contiene la letra “R” si el objeto es una roca y “M” si es una mina (cilindro de metal). Los números en las etiquetas están en orden creciente de ángulo.
 - Ajuste un modelo de árbol de clasificación para predecir la variable de respuesta, si es una roca o una mina.
 - Haga optimización de hiperparametros, variando los la profundidad del árbol, el máximo número de nodos hojas y el criterio de impureza.
 - Haga un análisis costo complejidad y gráfique *cpp_alpha* vs impureza, numero de nodos hoja, y finalmente la gráfica de alpha vs accuracy en la muestra de entrenamiento y de prueba. ¿Qué concluye acerca del valor de *cpp_alpha*?
 - Ajuste un modelo de Random Forest, Bagging y Boosting a los datos de sonar.
 - Obtenga las matrices de confusión y curva ROC.
4. Haga un análisis de clústeres con kmeans y dbscan con los datos de “coris”. El análisis deberá incluir el método del codo y las métricas de validación interna.