Parcial 2 parte 1 de analítica de datos (3.5)

Utilizando la base de datos de diabetes proporcionada para este examen realizar:

1. Realizar gráficas de dispersión entre variables para entender un poco la relación entre ellas.
2. Calcular la matriz de correlación del conjunto de datos original.
3. Estadísticos de cada columna como lo son la media, mediana, moda, kurtosis y asimetría. Mencionar en base de estos valores si dicha variable o columna tiene tendencia de ser una distribución normal o no.
4. Generar dos datasets, uno con valores atípicos y otro sin ellos. Los dos se utilizarán para entrenar modelos.
5. Calcular la matriz de correlación del conjunto sin atípicos.
6. Los modelos por entrenar tienen que ser validados por medio de una validación cruzada con K igual a 7, 9 y 11. SE DEBE GARANTIZAR LA HOMOGENEIDAD DE LOS DATOS AL MOMENTO DE REALIZAR LA VALIDACIÓN CRUZADA.
7. Se debe imprimir la matriz de confusión por cada validación del numeral anterior
8. Calcular el desempeño de cada modelo usando sensibilidad, precisión y especificidad.
9. Deben realizar al menos 4 modelos por algoritmo utilizado y decidir en base de la curva ROC cual es mejor.
10. Fusionar por esquema de votación la salida de algoritmos clasificadores, los cuales pueden ser: SVM, regresión logística, árboles de decisión, KNN y redes neuronales. Nota 1: usar al menos tres de ellos.

NOTA 1: CADA MODELO NO DEBE SUPERAR MÁS DE 7 variables.

NOTA 2: ÚNICAMENTE SE PUEDE ENTREGAR EL PARCIAL EN NOTEBOOK DE JUPYTER.

NOTA 3: REALIZAR UNA PRESENTACIÓN DEL CÓDIGO DE MANERA VIRTUAL O PRESENCIAL (1.5)

FECHA DE ENTREGA DEL NOTEBOOK: 20 DE MAYO

FECHA DE SUSTENTACIÓN:22 - 24 DE MAYO