ntic master
revolucionamos la comunicación

UNIVERSIDAD COMPLUTENSE DE MADRID

Tarea Minería de Datos

‘Práctica de Evaluación’

[redacted]

La entrega se realizará en un único NoteBook de Jupyter identificado con nombre y apellidos donde se Consignarán los títulos para los 4 ejercicios y se desarrollarán los códigos necesarios, añadiendo comentarios explicativos.

Ejercicio 1

Con el conjunto de datos "***FEV_data.csv***", información en fev.doc (live.com), que se encuentra en la carpeta de datos de la documentación, y sabiendo que la variable objetivo es '*fev*', se pide:

1- Reporte descriptivo de los datos. Dimensiones del dataset, número de variables continuas y categóricas. Distribuciones. Comentarios generales.

2- Decide si se descarta de inicio alguna de las variables de cara al modelado.

3- Ajusta el mejor modelo de regresión lineal con variables originales y sin interacciones entre las variables.

4- Ajusta el mejor modelo de regresión lineal con variables originales y con interacciones entre las variables que resulten relevantes.

5- Comparación de ambos modelos por validación cruzada repetida. ¿Cuál de ellos tiene mejor comportamiento en generalización?

Ejercicio 2

Con los datos de la serie de "***defunciones.xlsx***" que se encuentra en la carpeta de datos de la documentación, se pretende ajustar el mejor modelo de series temporales a cualquiera de las series allí contenidas (elige 1).

1- Lectura y representación de la serie. Descomposición. Conclusiones.

2- Partición training y test (los últimos 2 años de datos).

3- Mejor modelo de suavizado exponencial. ¿Pasa el test residual de Ljung.Box?

4- Mejor modelo ARIMA. ¿Pasa el test residual de Ljung.Box?

5- Comparación. En relación al MAPE en el conjunto de tets, ¿Qué modelo resulta más preciso en sus predicciones?

Ejercicio 3

Con los datos de ***wisconsin.xlsx*** que se encuentran en la carpeta Datos de la documentación.

1- Leer el archivo

2- Obtener la matriz de correlaciones entre las variables numéricas. Conclusiones.

3- Sabiendo que se trata de predecir la variable binaria diagnosis ('M,'B'). ¿Cuál sería el modelo adecuado de predicción? ¿Qué problemas auguras a la luz de la información sobre las correlaciones del archivo?

4- Como posible estrategia se plantea la realización de un ACP para la reducción de dimensiones y un modelo de predicción adecuado utilizando como predictores las componentes principales resultantes de tal forma que se retenga al menos el 70% de la variabilidad del archivo.

   a. Valora la adecuación muestral a priori

   b. Realiza el ACP sobre las numéricas. ¿Cuántas componentes se deberían considerar para cumplir el criterio mencionado? Interpreta la componente 1.

   c. Crea el input con estas componentes (matriz de scores o puntuaciones) y ajusta un modelo de predicción adecuado para la variable objetivo *diagnosis*.

   d. Conclusiones del modelo por validación cruzada repetida. Interpreta el parámetro de la componente 1.

Ejercicio 4

El conjunto de datos ***DatosEleccionesEspaña.xlsx*** contiene información demográfica sobre los distintos municipios de España junto con los resultados que se obtuvieron en un proceso electoral. En esta práctica, el objetivo fundamental es crear grupos de comunidades autónomas en función de ciertas características previamente seleccionadas.

1- Escoge 10 variables numéricas presentes en el conjunto de datos y realiza un ligera depuración sobre este subconjunto de variables (Valores raros, Outliers, NAs..)

2- Agrega los valores por CCAA, distinguiendo entre valores relativos (agregación por la media) y valores absolutos (agregación por suma)

3- Valora la necesidad de escalar los datos y decide el tipo de distancia a aplicar

4- Explora los métodos de clustering jerárquico para estos datos y decide el tipo de Linkage más adecuado.

5- Toma una decisión sobre el número de clusters a considerar y realiza un análisis cluster mediante el método k-means.

6- Muestra el biplot (puede ser proyección sobre dos de las variables que resulten relevantes o proyección sobre las dos primeras componentes principales de la matriz de entrada) y comenta los grupos formados. Interpreta los centroides.