UNIVERSIDAD ICESI
Analítica de Datos
Aníbal Sosa y Andrés Aristizábal

Proyecto de clase

Descripción

El dataset que se va a analizar contiene datos de 299 pacientes con insuficiencia cardíaca en el año 2015. Se dispone del diccionario de datos siguiente:

- age: age of the patient (years)
- anaemia: decrease of red blood cells or hemoglobin (boolean)
- high blood pressure: if the patient has hypertension (boolean)
- diabetes: if the patient has diabetes (boolean)
- ejection fraction: percentage of blood leaving the heart at each contraction (percentage)
- platelets: platelets in the blood (kiloplatelets/mL)
- sex: woman or man (binary)
- serum creatinine: level of serum creatinine in the blood (mg/dL)
- serum sodium: level of serum sodium in the blood (mEq/L)
- smoking: if the patient smokes or not (boolean)
- time: follow-up period (days)
- [target] death event: if the patient deceased during the follow-up period (boolean)

La idea es poder predecir si el paciente fallece a partir de un modelo basado en los demás campos.

Puntos a desarrollar

1. **Limpieza y EDA**: Verifiquen si hay problemas de calidad de datos.
   Se espera una primera sección de evaluación de la calidad de los datos y de entendimiento de la relación entre las variables predictivas y la variable objetivo (**OJO!** Solo poner gráficos y análisis de las relaciones importantes, **menos es más!**)
2. **Modelos predictivos**: Entrenen modelos predictivos (al menos 3 familias de modelos) que permitan estimar si el paciente muere a partir de los valores de las demás variables. Escoja el mejor modelo, buscando sus parámetros óptimos.
   Se espera una sección donde se establezca el protocolo de evaluación y los procesos de entrenamiento y evaluación de los modelos.
3. **Cambio de representación del dataset:** Considerando todas las variables (menos death event), realice un análisis de componentes principales (PCA), escogiendo el número de componentes necesarios para conservar el 95% de la representación original.
4. **Caracterización de los jugadores:** Con los datos en su nueva representación de PCs, realice una segmentación, estableciendo el mejor número de clusters entre 3 y 5. Caracterice los clusters con respecto a las variables originales (incluyendo death event).

Rúbrica de puntuación

UNIVERSIDAD ICESI
Analítica de Datos
Aníbal Sosa y Andrés Aristizábal

Calidad de datos	Visualización de datos	Extracción de intuiciones de los datos	Entendimiento de los datos y limpieza	Protocolo de entrenamiento y evaluación de modelos	Entrenamiento de los 3 modelos	Transformación de los datos por PCA	Caracterización de perfiles de pacientes	TOTAL PROYECTO
0.3	0.7	0.7	0.3	0.3	1.0	0.7	1.0	5.0