ESCUELA SUPERIOR POLITECNICA DEL LITORAL

ESTRUCTURAS DE DATOS

PROYECTO DE ESTRUCTURA

AVANCE

INTEGRANTES:

Eras Zamora Edwin Andrew

Holguin Wong Erick Weyling

Pazmiño Guerrero Gabriela Nicole

Vulgarin Punguil Jorge Adrian

PARALELO: 103

2020-2021

Contenido

Avance del proyecto

Selección de umbrales para los atributos no booleanos

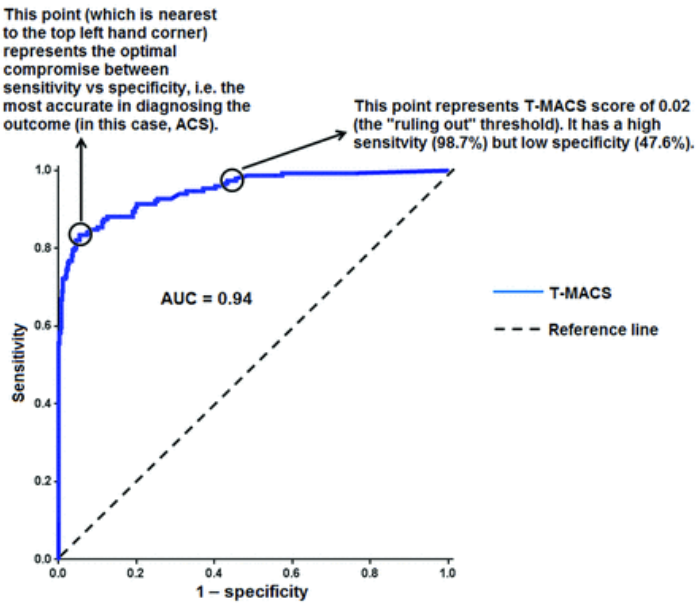
Curvas ROC

En la teoría de detección de señales, una curva ROC es una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador binario según se varía el umbral de discriminación. [1]

Para producir una curva ROC, la sensibilidad y especificidad de diferentes valores continuos son tabulados. El resultado se presenta en una lista de valores de prueba con su correspondiente sensibilidad y especificidad. Después, la curva ROC es producida al graficar la sensibilidad en el eje Y contra 1 - especificidad en el eje X. [2]

Como interpretar una curva Roc

Una curva ROC que sigue una curva diagonal indicada por la función y=x produce falsos resultados positivos al mismo ritmo que verdaderos resultados positivos. Por lo que se puede esperar que una prueba con una precisión razonable tenga una curva ROC en la parte superior izquierda del triángulo formado por la línea de la función y=x, como se muestra en la figura 1. [2]



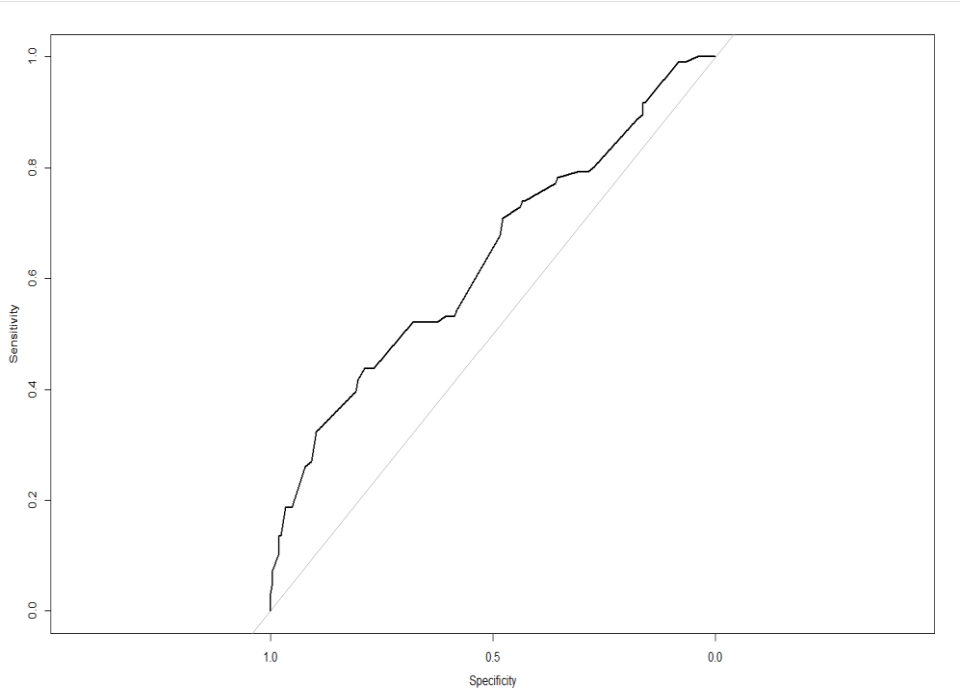
El área bajo la curva ROC, denominada AUC, es una medida global de la habilidad de una prueba para discriminar si una condición especifica está presente o no. Una AUC de 0.5 indica que la habilidad discriminatoria de una prueba es nula, lo que sería igual que obtener resultados al azar, mientras que un AUC de 1.0 simboliza una prueba con perfecta discriminación. [2]

Cuando queremos seleccionar un umbral se debe considerar lo que se necesita de una prueba, tomando en cuenta el valor de los falsos positivos y verdaderos positivos. El enfoque más para seleccionar un umbral es buscar un punto de corte en la gráfica (1-especificidad) vs. Sensibilidad que de igual peso a la importancia de ambos ejes. [2]

Implementación en R

Para poder obtener los umbrales de los atributos estudiados, se usarán funciones de la librería pRoc en rStudio.

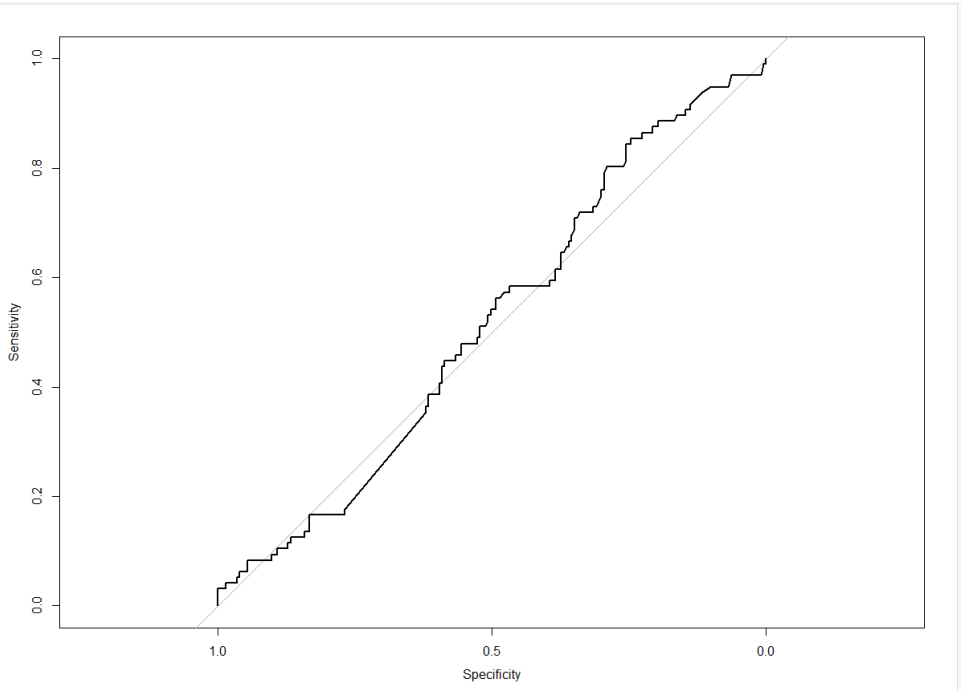
- **Umbral de la edad en base al evento de muerte**



Por medio de las funciones usadas en R, se obtiene el siguiente umbral:

```
> #Umbral de edad
> threshold2(tabla$age, tabla$DEATH_EVENT)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] 67.5
>
```

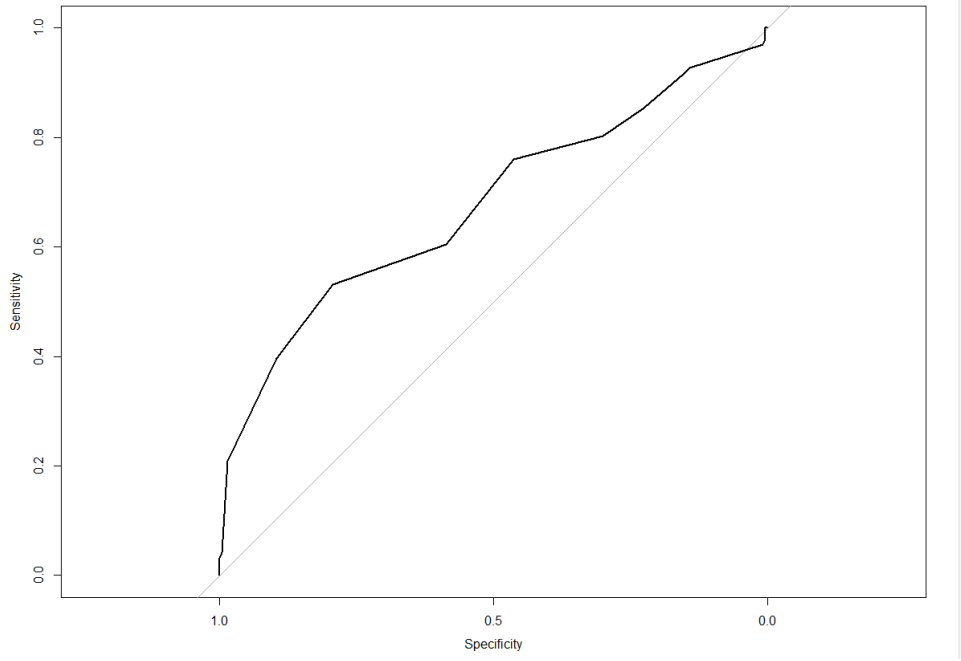
- **Umbral de CPK_enzyme en base al evento de muerte**



Por medio de las funciones usadas en R, se obtiene el siguiente umbral:

```
> #Umbral de CPK_enzyme
> threshold2(tabla$CPK_enzyme, tabla$DEATH_EVENT)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] 103.5
>
```

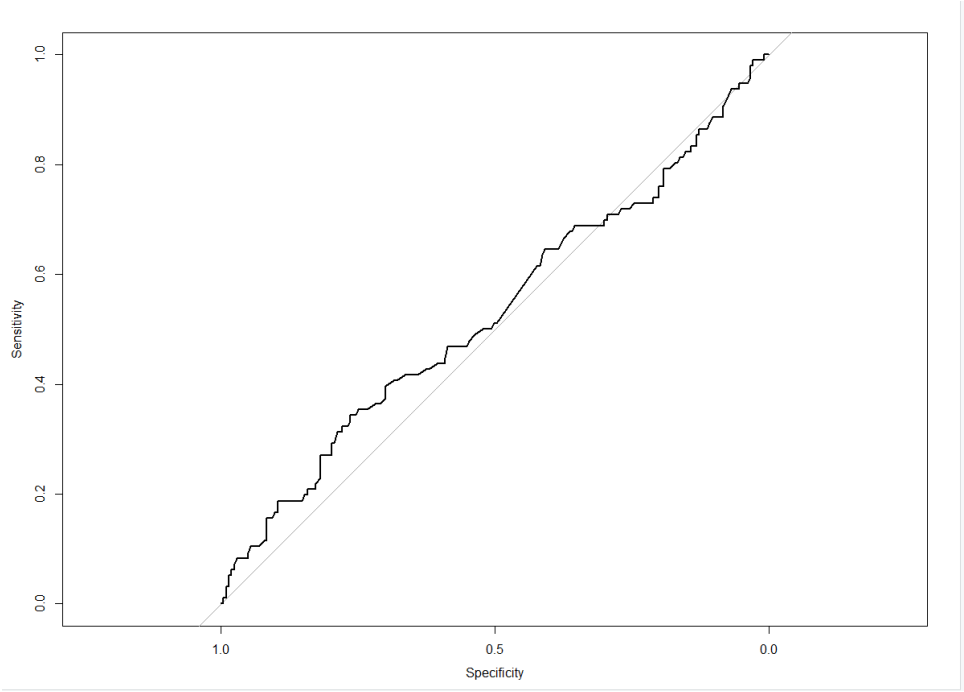
- **Umbral de Ejection_fraction en base al evento de muerte**



Por medio de las funciones usadas en R, se obtiene el siguiente umbral:

```
> #Umbral de ejection_fraction
> threshold2(tabla$ejection_fraction, tabla$DEATH_EVENT)
Setting levels: control = 0, case = 1
Setting direction: controls > cases
[1] 32.5
>
```

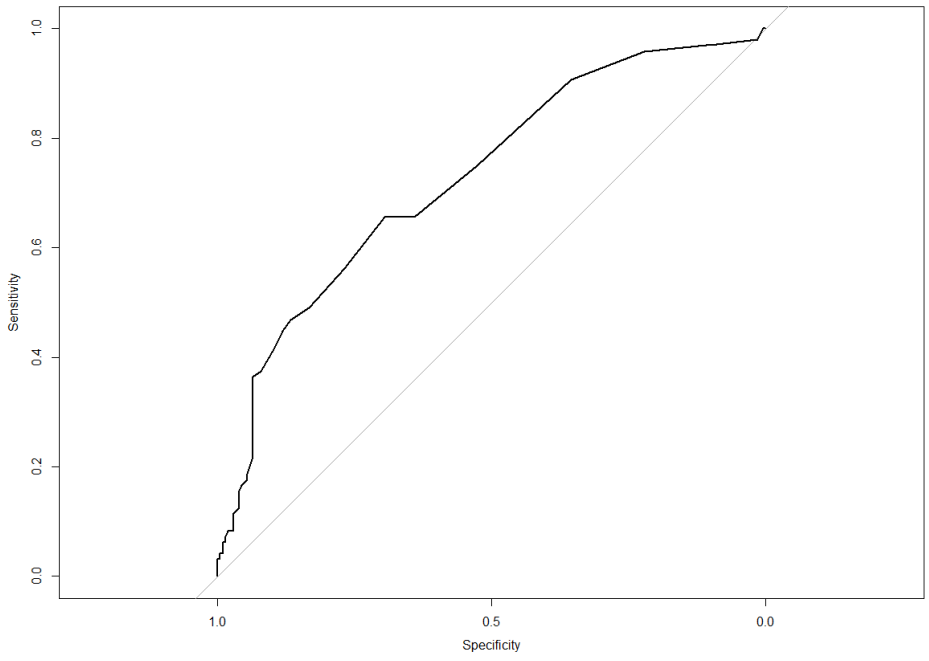
- **Umbral de Platelets en base al evento de muerte**



Por medio de las funciones usadas en R, se obtiene el siguiente umbral:

```
> #Umbral de platelets
> threshold2(tabla$platelets, tabla$DEATH_EVENT)
Setting levels: control = 0, case = 1
Setting direction: controls > cases
[1] 217500
>
```

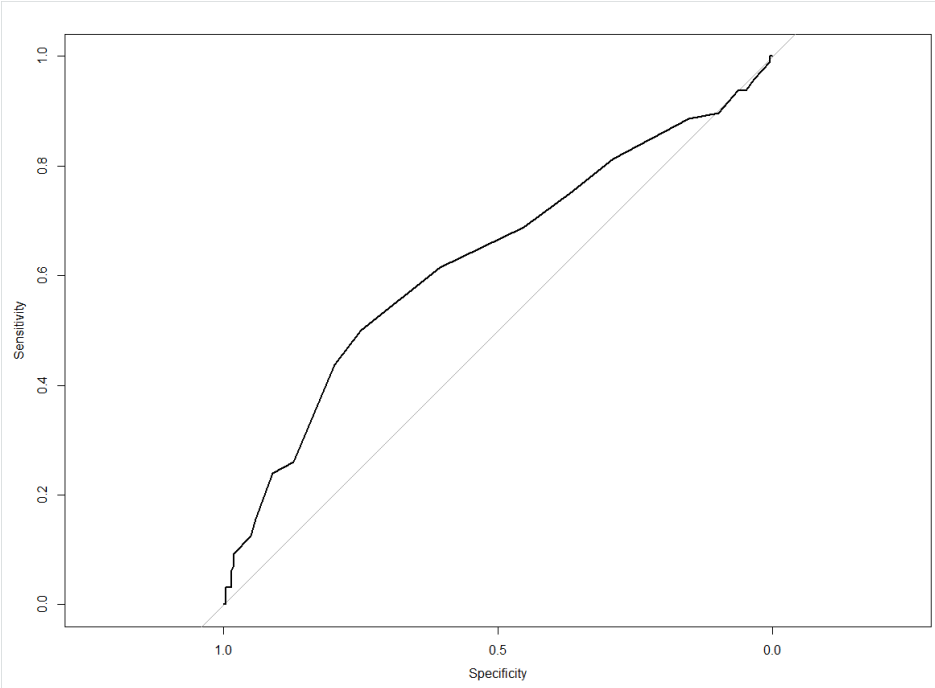
- **Umbral de Serum_creatinine en base al evento de muerte**



Por medio de las funciones usadas en R, se obtiene el siguiente umbral:

```
> #Umbral de serum_creatinine
> threshold2(tabla$serum_creatinine, tabla$DEATH_EVENT)
Setting levels: control = 0, case = 1
Setting direction: controls < cases
[1] 1.19
>
```

- **Umbral de Serum_sodium en base al evento de muerte**



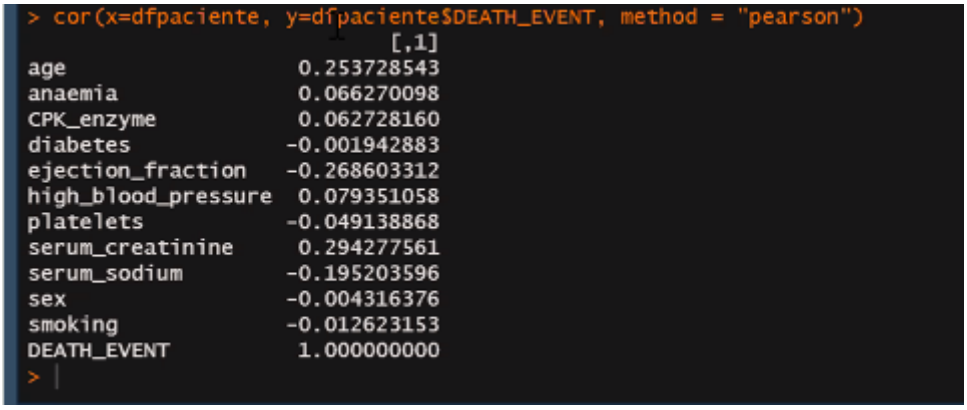
Por medio de las funciones usadas en R, se obtiene el siguiente umbral:

```
> #Umbral de serum_sodium
> threshold2(tabla$serum_sodium, tabla$DEATH_EVENT)
Setting levels: control = 0, case = 1
Setting direction: controls > cases
[1] 135.5
>
```

Primeros 3 atributos de mayor importancia para crear un árbol de decisión a partir del dataset dado

Para escoger los 3 atributos utilizamos la función cor() en R, que genera una matriz de correlación entre las variables especificadas. La matriz resultante mostrara las correlaciones entre cada atributo con respecto a DEATH_EVENT. El coeficiente de correlación obtenido nos indica que tan relacionada están los atributos con la variable DEATH_EVENT y aunque sean valores pequeños ciertas variables tienen valores más altos en relación con los demás.

En este caso las variables con mayor coeficiente de correlación, sea positivo o negativo, son las que se relacionan más con DEATH_EVENT y nos indican en nuestro análisis cuál de ellas tiene mayor importancia, por lo que escogemos los atributos serum_creatinine, ejection_fraction y age.



Formas de cargar/representar el dataset en un programa en Java

- **Opción 1: Objeto de java**

Una forma de representar el dataset podría ser por medio de la implementación de un objeto en java, en el cual los atributos del objeto serán listas que almacenen los datos de las columnas del dataset.

- **Opción 2: HashMap**

Otra opción es usar un HashMap para representar el dataset. Para esta implementación, el nombre de la columna del dataset correspondería la clave y el valor serían los datos de las columnas ingresadas como listas.

- **Opción 3: Listas**

La última opción expuesta para representar el dataset seria almacenar todos los datos de cada columna en una lista diferente y manipular los datos por separado.

Una vez analizado los datos la opción que elegimos para la representación de los datos es la primera que expusimos, representar el dataset como un objeto en java. De esta manera se utiliza correctamente el paradigma de la Programación Orientada a Objetos. A más de esto, esta opción nos permite la fácil manipulación de los datos, proceso que se volvería tedioso en el caso de que los datos se almacenen en una estructura como un HashMap o una Lista

Trabajos citados

[1] Z. H. Hoo, J. Candlish y D. Teare, «What is an ROC curve?,» Emergency Medicine Journal, 2017.

[2] «Signal detection theory and ROC analysis in psychology and diagnostics: collected papers,» Swets, 1996.