Introducción a la ciencia de datos: Informe final

Jesús Enrique Cartas Rascón

Classification: Hayes-Roth Dataset

Estudio del dataset

Antes de empezar si quiera a explorar, introduciremos brevemente el contexto del dataset, así como algunas de las propiedades más fundamentales del mismo, con el objetivo de encauzar nuestra exploración y análisis en la mejor dirección posible.

Contexto

El dataset Hayes-Roth es un conjunto de datos sintético hecho para examinar el rendimiento de algoritmos de clasificación. Según la documentación del archivo, se compone de 5 clasificadores enteros, y una variable de salida, `Class`, que, muy evidentemente, es nuestra variable objetivo. La idea es clasificar la `Class` de una muestra en función de las otras variables: `Hobby, Age, EducationalLevel, MaritalStatus`.

Al ser sintético, hubo absoluto control sobre su creación. No posee datos perdidos, y la distribución `RNI` era originalmente `(0/4/0)`, perfecto para entrenar algoritmos de clasificación. Al parecer, según el autor del dataset que vamos a utilizar:

> I've replaced the actual values of the attributes (i.e., hobby has values chess, sports and stamps) with numeric values. I think this is how the authors' did this when testing the categorization models described in the paper. I find this unfair. While the subjects were able to bring background knowledge to bear on the attribute values and their relationships, the algorithms were provided with no such knowledge.

nos comenta que ha cambiado la naturaleza de las variables, sustituyendo los valores nominales por números. Así que nuestra nueva distribución es `RNI (0/0/4)`.

No existe mucha más información al respecto, aunque no necesitamos mucho más. En conclusión, tenemos que ser capaces de averiguar el atributo Class en función de los otros 4.

Hipótesis

Al ser un dataset sintético y carecer de valores nominales, no podemos establecer una hipótesis acerca de qué variable puede afectar a la salida más significativamente. Deberemos proceder con un análisis exploratorio directamente para ver qué información podemos sacar.

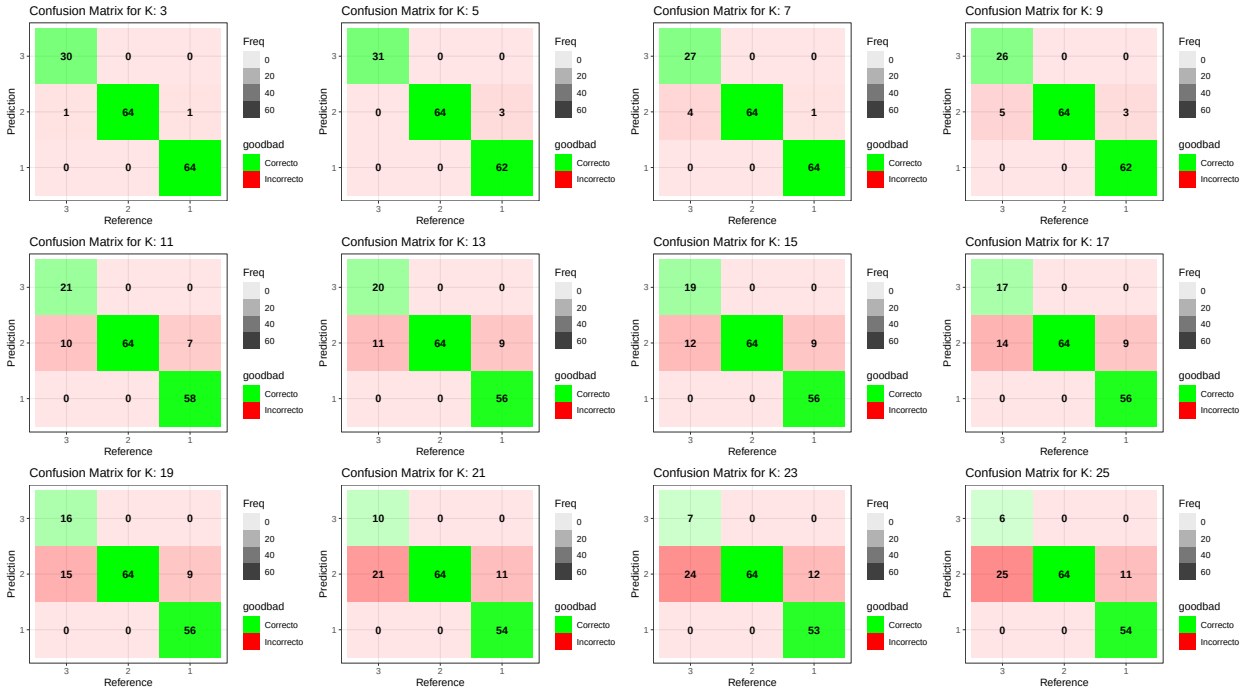

Figure 1: Confusion matrices for K