Memoria Trabajo Spark

Las primeras pruebas van a ser con los datos tratados de la siguiente forma:

* 1. Datos originales
* 2. Discretizados y con selección de 10 características
  + Clase 0 train -> 349272
  + Clase 1 train -> 150177
  + Clase 0 test -> 275091
  + Clase 1 test -> 137545

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datos | Configuración del modelo | | | | Resultados | | | Tiempo (s o m) |
| Modelo | Num Trees | Max Depth | Max Bins | Test Error (%) | Preccision (%) | ROC (%) |
| 1 | RF | 10 | 4 | 32 | 29,3348 | 70,6651 | 57,60 | 19s |
| 1 | NB |  |  |  | 29,3748 | 70,6251 | 60,72 | 6s |
| 1 | DT |  | 4 | 32 | 27,0446 | 72,9553 | 67,15 | 14s |
| 2 | RF | 10 | 4 | 32 | 29,6811 | 70,3188 | 57,72 | 7s |
| 2 | NB |  |  |  | 29,6297 | 70,3702 | 65,13 | 2s |
| 2 | DT |  | 4 | 32 | 29,1043 | 70,8956 | 61,76 | 3s |

Donde podemos ver como los resultados, por lo general, se han mantenido los resultados con el conjunto de datos preprocesado, por lo que vamos a seguir preprocesando.

Vemos como tenemos 631 características, y en estos ejemplos hemos probado a quedarnos con 10 características en los “Datos 2”, por lo que vamos a quedarnos por ejemplo con 300 de estas características, de forma que todos los cálculos se harán de una forma más rápida. Además, tenemos un imbalanceo ya que tenemos 70-30 de ratio. Probaremos a hacer oversampling para ver qué tal se comportan los modelos.