# **Analítica de Datos en Big Data**

# Primeros experimentos

Los experimentos realizados para comenzar han tenido los siguientes datasets:

* 1. Datos originales con 499449 elementos en train y 412636 en test, y 631 características, de donde:
  + Train: Clase 0 -> 349272 y clase 1 -> 150177
  + Test: Clase 0 -> 275091 y clase 1 -> 137545
* 2. Igual que el original, pero con los datos discretizados y con selección de 10 características.

Además, los modelos usados son:

* RF: Random Forest con 10 árboles, máxima profundidad de 4 y un max Bins de 32.
* NB: Naive Bayes con un lambda igual a 1.0
* DT: Decision Trees con máxima profundidad de 4 y un max Bins de 32.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datos + Modelo | Resultados | | | | | | | | Tiempo (s) |
| Test Error % | Precisión % | ROC % | TP | FP | FN | TN | TPR x TNR |
| 1 RF | 29,363 | 70,636 | 57,447 | 266876 | 8215 | 112951 | 24594 | 0,5267 | 19 |
| 1 NB | 29,374 | 70,625 | 60,726 | 248743 | 26348 | 94863 | 42682 | 0,4476 | 4 |
| 1 DT | 27,044 | 72,955 | 67,15 | 232632 | 42459 | 69137 | 68408 | 0,4757 | 15 |
| 2 RF | 30,027 | 69,972 | 57,079 | 263420 | 11671 | 112235 | 25310 | 0,4799 | 7 |
| 2 NB | 29,629 | 70,37 | 65,136 | 222376 | 52715 | 69548 | 67997 | 0,4291 | 2 |
| 2 DT | 29,104 | 70,895 | 61,767 | 245247 | 29844 | 90251 | 47294 | 0,4482 | 3 |

Donde podemos ver como los resultados, por lo general, se han mantenido, o han bajado un poco, con el conjunto de datos preprocesado, por lo que vamos a seguir preprocesando. Además, hemos guardado en nuestro directorio el dataset reducido con 10 características para volver a cargarlo y ver que se obtienen los mismos resultados, de forma que ha sido así.

Tenemos en los datos un imbalanceo, ya que tenemos 70-30 de ratio. Probaremos a hacer undersampling y oversampling para ver qué tal se comportan los modelos.

# Aplicación de RUS y ROS

En este apartado vamos a realizar balanceo de los datos con RUS, es decir, vamos a eliminar instancias de la clase mayoritaria para balancear el dataset. Además, vamos a aplicar ROS, es decir, vamos a crear instancias de la clase minoritaria para no tener desbalanceo de los datos. ROS lo he aplicado para tener el mismo número de instancias de las dos clases.

Por lo tanto, lo primero que se ha realizado ha sido el undersampling, quedando un dataset con los siguientes elementos en train:

* 149957 de la clase 0
* 150177 de la clase 1.

Por lo que hemos solucionado este problema de imbalanceo de los datos. Ahora, vamos a realizar las mismas pruebas que en el apartado anterior, de forma que vamos a tener los datasets:

* 3: Referente al dataset con undersampling y las 631 características.
* 4: Dataset con undersampling, pero haciendo selección de 10 características.

Los resultados obtenidos han sido:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datos + Modelo | Resultados | | | | | | | | Tiempo (s) |
| Test Error % | Precisión % | ROC % | TP | FP | FN | TN | TPR x TNR |
| 3 RF | 33,254 | 66,745 | 68,567 | 173584 | 101507 | 35715 | 101830 | 0,4153 | 19 |
| 3 NB | 35,569 | 64,43 | 64,486 | 176936 | 98155 | 48618 | 88927 | 0,3729 | 5 |
| 3 DT | 31,384 | 68,615 | 69,958 | 181366 | 93725 | 35780 | 101765 | 0,4348 | 13 |
| 4 RF | 34,796 | 65,203 | 67,738 | 165420 | 109671 | 33913 | 103632 | 0,4032 | 7 |
| 4 NB | 33,172 | 66,827 | 67,384 | 180772 | 94319 | 42562 | 94983 | 0,4061 | 2 |
| 4 DT | 31,603 | 68,396 | 67,079 | 195393 | 79698 | 50711 | 86834 | 0,4140 | 4 |

A la vista de estos resultados, vemos como la precisión ha empeorado, además la medida TPR x TNR ha bajado bastante, pero la medida ROC ha aumentado.

Por lo tanto, y teniendo en cuenta estos resultados, he decido no hacer más pruebas con el undersampling y profundizar más en la selección de características e intentar obtener un mejor resultado con el dataset reducido.

Ahora vamos a realizar oversampling, realizando los mismos experimentos. En train, una vez realizado ROS, tenemos:

* 349272 de la clase 0.
* 349742 de la clase 1.

Por lo que ya hemos solucionado el imbalanceo de los datos. Si ahora tenemos los datasets:

* 5: Referente al dataset con oversampling y las 631 características.
* 6: Dataset con oversampling, pero haciendo selección de 10 características.

Los resultados obtenidos han sido:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datos + Modelo | Resultados | | | | | | | | Tiempo (s) |
| Test Error % | Precisión % | ROC % | TP | FP | FN | TN | TPR x TNR |
| 5 RF | 33,065 | 66,934 | 68,673 | 174565 | 100526 | 35914 | 101631 | 0,4170 | 25 |
| 5 NB | 35,519 | 64,48 | 64,515 | 177183 | 97908 | 48660 | 88885 | 0,3733 | 5 |
| 5 DT | 31,56 | 68,439 | 69,935 | 180044 | 95047 | 35182 | 102363 | 0,4338 | 19 |
| 6 RF | 32,48 | 67,519 | 68,289 | 181501 | 93590 | 40437 | 97108 | 0,4164 | 7 |
| 6 NB | 32,413 | 67,586 | 67,931 | 184029 | 91062 | 42686 | 94859 | 0,4141 | 2 |
| 6 DT | 31,729 | 68,27 | 68,007 | 189250 | 85841 | 45087 | 92458 | 0,4188 | 3 |

Donde podemos ver como los resultados, con respecto a RUS, son un poco mejores, pero siguen sin ser como los del dataset completo. Vamos a seguir, por lo tanto, haciendo pruebas preprocesando el dataset.

# Selección de características 200

El siguiente experimento es sobre selección de características, 200 concretamente, por lo que tenemos el siguiente conjunto de datos:

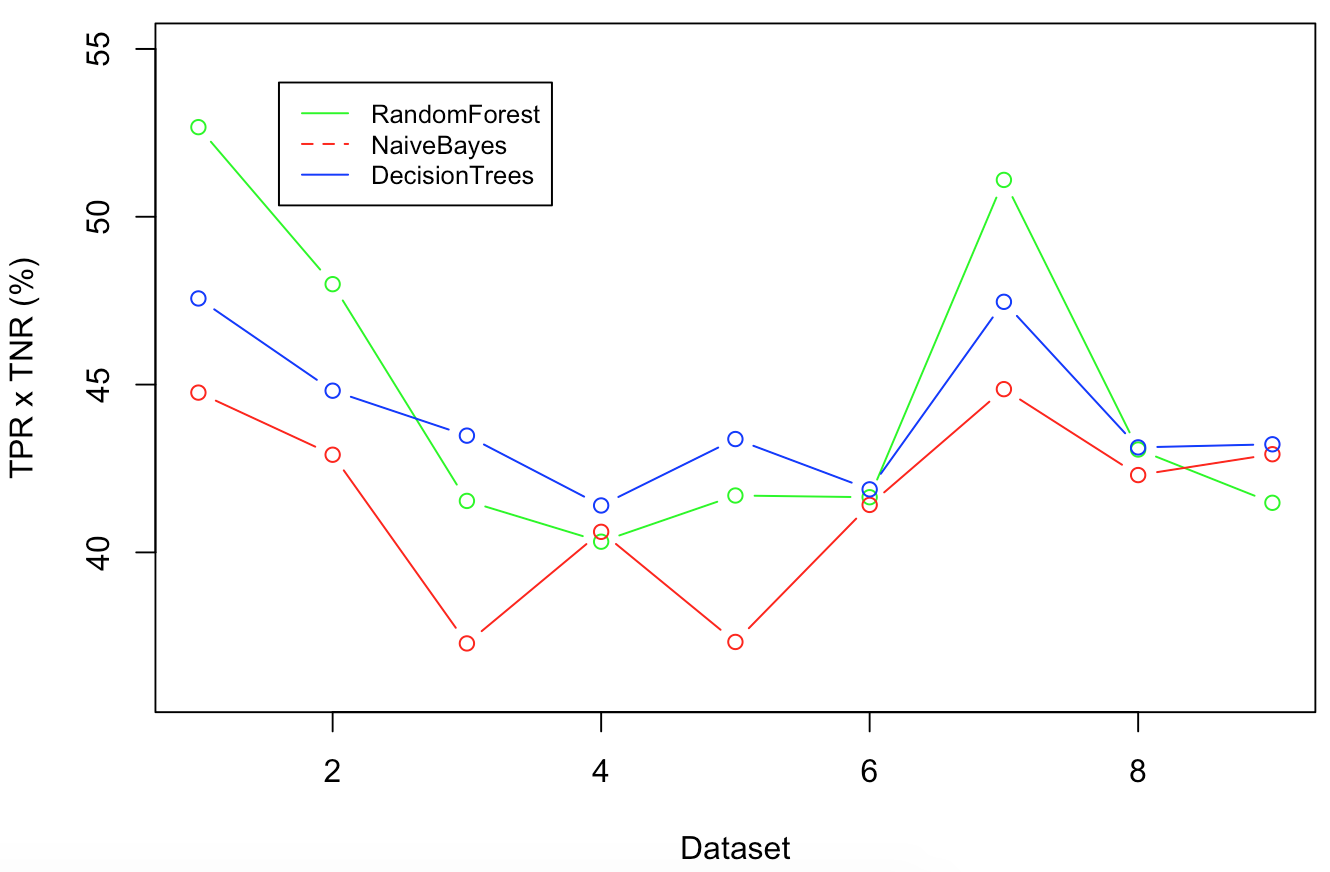
* 7: Referente al dataset original con una selección de 200 características.
* 8: Dataset con RUS y una selección de 200 características.
* 9: Dataset con ROS y una selección de 200 características.

Por lo que los resultados obtenidos han sido:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datos + Modelo | Resultados | | | | | | | | Tiempo (s) |
| Test Error % | Precisión % | ROC % | TP | FP | FN | TN | TPR x TNR |
| 7 RF | 27,688 | 72,311 | 61,548 | 258134 | 16957 | 97297 | 40248 | 0,5110 | 9 |
| 7 NB | 28,77 | 71,229 | 68,681 | 209962 | 65129 | 53589 | 83956 | 0,4486 | 3 |
| 7 DT | 27,207 | 72,792 | 66,351 | 235677 | 39414 | 72856 | 64689 | 0,4747 | 5 |
| 8 RF | 31,496 | 68,503 | 69,495 | 182995 | 92096 | 37868 | 99677 | 0,4306 | 9 |
| 8 NB | 31,368 | 68,631 | 68,292 | 190670 | 84421 | 45015 | 92530 | 0,4230 | 4 |
| 8 DT | 31,913 | 68,086 | 69,823 | 177744 | 97347 | 34339 | 103206 | 0,4313 | 5 |
| 9 RF | 32,908 | 67,091 | 68,333 | 177729 | 97362 | 38430 | 99115 | 0,4148 | 11 |
| 9 NB | 30,819 | 69,18 | 68,666 | 193134 | 81957 | 45217 | 92328 | 0,4293 | 4 |
| 9 DT | 31,788 | 68,211 | 69,87 | 178517 | 96574 | 34597 | 102948 | 0,4322 | 7 |

Si nos enfocamos en la medida TPR x TNR, vemos como con el conjunto original, se ha obtenido un resultado muy parecido al mejor hasta el momento obtenido con todas las características y el dataset completo. Con respecto a ROS y RUS, de momento se han obtenido solo peores resultados que con el dataset original.

Veamos ahora una gráfica sobre cómo van los experimentos en función de la medida TPR x TNR:



De forma que vemos como el modelo que mejores resultados ha dado hasta ahora es el Random Forest. Los mejores resultados se obtienen en los Datasets 1 y 7, que corresponden a los datasets originales sin aplicar RUS ni ROS. Podemos decir, además, que la selección de las 200 características ha dado un resultado muy parecido al completo. Viendo esto, voy a realizar más selección de características para intentar conseguir mejores resultados o parecidos.

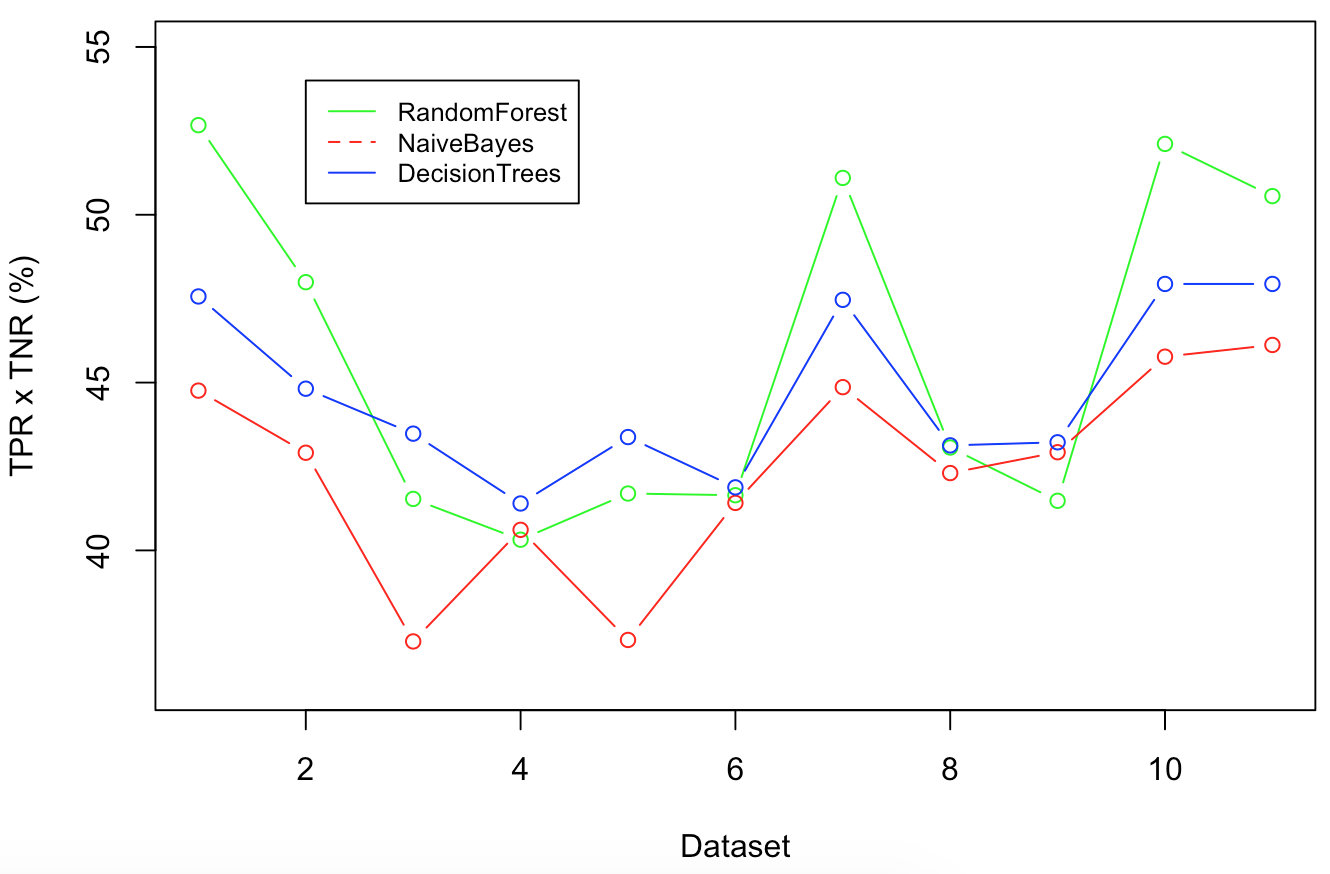
# Selección de características 150 y 100

Si comparamos la selección de 150 características sobre el dataset original con la selección de 100 características tenemos que los dataset utilizados son:

* 10: Dataset con selección de 150 características
* 11: Dataset con selección de 100 características

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Datos + Modelo | Resultados | | | | | | | | Tiempo (s) |
| Test Error % | Precisión % | ROC % | TP | FP | FN | TN | TPR x TNR |
| 10 RF | 29,692 | 70,307 | 56,884 | 267257 | 7834 | 114689 | 22856 | 0,5211 | 8 |
| 10 NB | 28,096 | 71,903 | 68,924 | 214185 | 60906 | 55032 | 82513 | 0,4577 | 3 |
| 10 DT | 27,79 | 72,209 | 63,218 | 248108 | 26983 | 87691 | 49854 | 0,4794 | 5 |
| 11 RF | 29,707 | 70,292 | 57,158 | 265635 | 9456 | 113126 | 24419 | 0,5056 | 8 |
| 11 NB | 27,776 | 72,223 | 68,241 | 220590 | 54501 | 60114 | 77431 | 0,4612 | 3 |
| 11 DT | 27,79 | 72,209 | 63,218 | 248108 | 26983 | 87691 | 49854 | 0,4794 | 4 |

Veamos la gráfica para comparar los resultados de una forma gráfica:



Por lo tanto, vemos como el Dataset con 150 características ha funcionado casi igual que el original, con una pequeña diferencia.

El siguiente paso a probar será el uso de ROS con diferentes tasas de balanceo. Anteriormente la tasa que se había usado era 100, de forma que se tenían el mismo número de valores para la clase minoritaria que la mayoritaria. Si usamos 90, tendremos unos pocos valores en la minoritaria menos que en la mayoritaria. Además, usaré 110, 120 y 130.

# Aplicación de ROS

Como he comentado, voy a usar ROS con valores de 90, 110, 120 y 130. Además, todos ellos serán sobre el dataset con 150 características para acelerar los cálculos. Estos dataset son:

* 12: ROS 90, quedando de la clase 0 y de la clase 1.
* 13: ROS 110, quedando de la clase 0 y de la clase 1.
* 14: ROS 120, quedando de la clase 0 y de la clase 1.
* 15: ROS 130, quedando de la clase 0 y de la clase 1.