

PRACTICA 1

Jaime Lorenzo Sánchez

21 de marzo de 2022

Índice general

1. Introducción	1
2. Ejercicio 1. Compare el rendimiento de al menos dos métodos no basados en boosting, como la mezcla por votación, bagging o random forests	2
2.1. Método bagging	2
2.2. Random Forest	4
2.3. Comparación de ambos clasificadores	5
3. Compare el rendimiento de al menos dos métodos basados en boosting, esto incluye las diferentes configuraciones de AdaBoost. Estude la evolución del error a medida que se añaden clasificadores para alguno de los problemas	7
3.1. Clasificador Naïve Bayes	9
3.2. Clasificador AdaBoost	9
3.3. Evolución del error	11
4. Ejercicio 3. Compare el rendimiento de los métodos de stacking con el mejor de los métodos de cada uno de los apartados anteriores.	13

Capítulo 1

Introducción

En esta práctica se estudiará el funcionamiento de diferentes métodos de ensembles. Se seleccionará un número suficiente de problemas, escogiendo aquellos con diferente número de clases, variables e instancias.

Los problemas elegidos son:

- Iris.data
- balance-scale.data
- breast cancer sklearn

El código de implementación utilizado es python3.

El método de entrenamiento será un 70 % de entrenamiento y un 30 % de test

Capítulo 2

Ejercicio 1. Compare el rendimiento de al menos dos métodos no basados en boosting, como la mezcla por votación, bagging o random forests

2.1. Método bagging

Para el método bagging, se ha empleado los siguientes datos:

- Número de estimadores: 100
- Características máximas: 10
- Muestras máximas: 4
- Estado aleatorio: 1
- Número de trabajos: 5

Aplicado este método, hemos obtenido los siguientes resultados con el archivo iris.data:

- Numero de elementos: 750

- Puntaje de la prueba del modelo: 0.956
- Puntaje de entrenamiento del modelo: 0.971

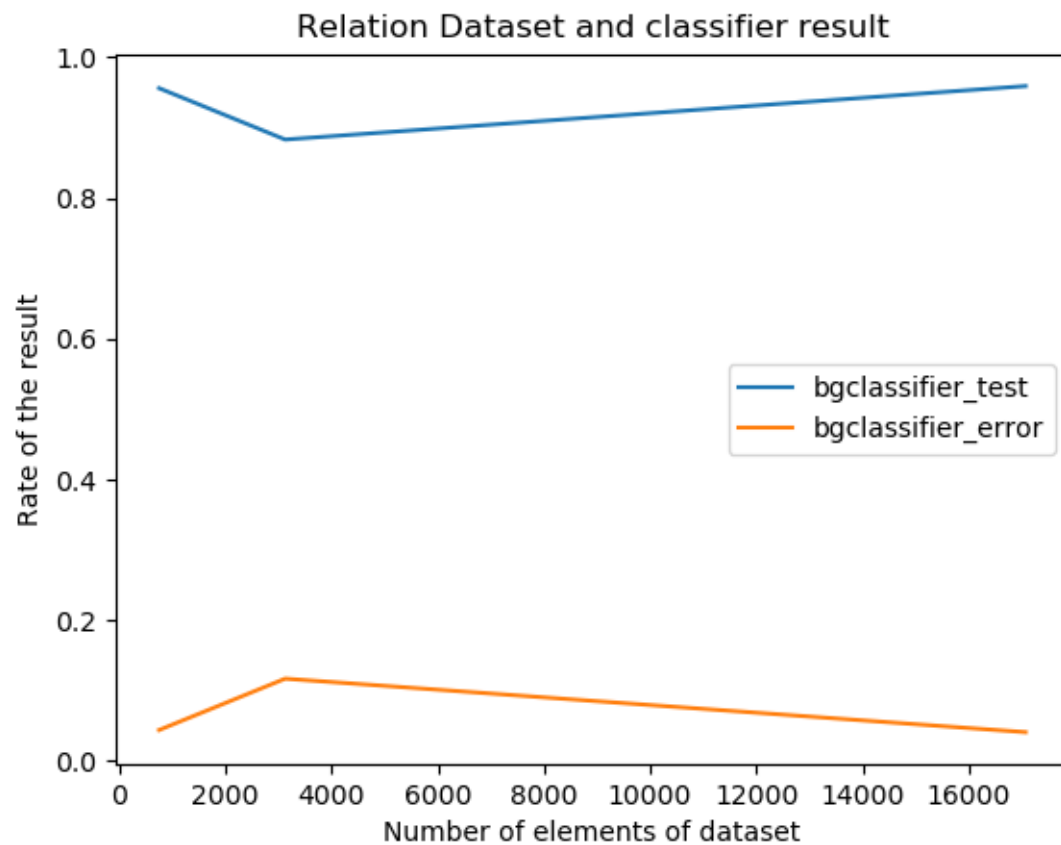
Aplicamos el método utilizado anteriormente con el archivo `balance-scale.data`, obteniendo los siguientes resultados:

- Numero de elementos: 3125
- Puntaje de la prueba del modelo: 0.883
- Puntaje de entrenamiento del modelo: 0.874

Aplicamos el método utilizado anteriormente con el `breast cancer sklearn dataset` y obtenemos los siguientes resultados (en este caso, el número máximo de muestras es de 4):

- Número de elementos: 17070
- Puntaje de la prueba del modelo: 0.959
- Puntaje de entrenamiento del modelo: 0.945

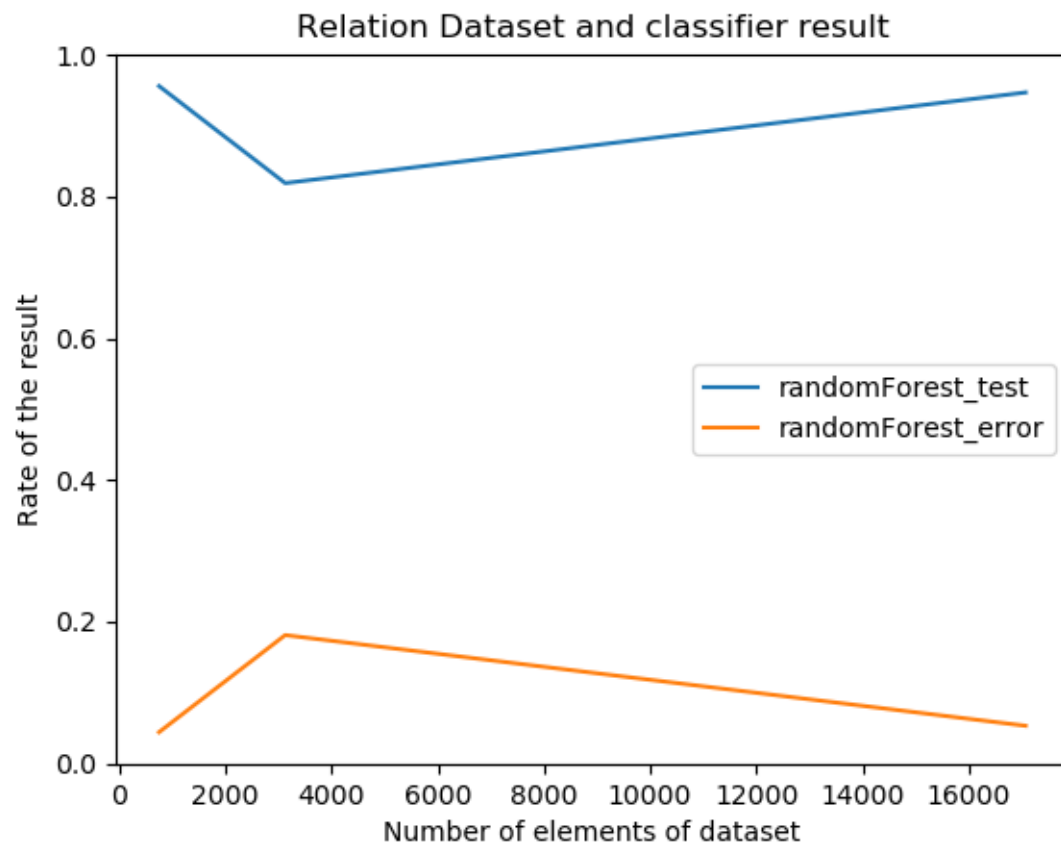
Para realizar un correcto estudio, se ha calculado el error cometido en cada dataset. A continuación, se muestra el funcionamiento del clasificador en función de los dataset utilizados:



Si observamos la gráfica, podemos observar que el clasificador utilizado ha obtenido unos resultados muy buenos, con un ratio de test elevado y con un ratio de error bajo.

2.2. Random Forest

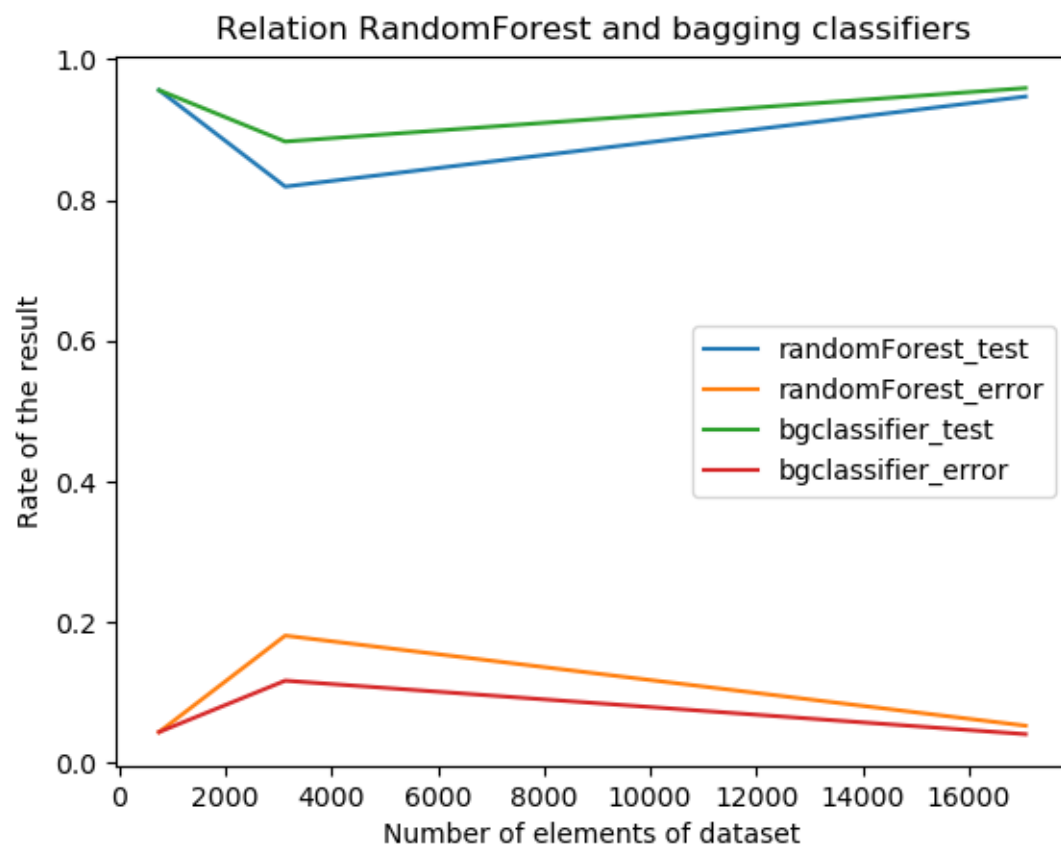
Se ha utilizado el algoritmo random forest con un número de 25 árboles a construir.



Si observamos la gráfica, podemos observar que el clasificador utilizado ha obtenido unos resultados muy buenos, con un ratio de test elevado y con un ratio de error bajo.

2.3. Comparación de ambos clasificadores

Una vez obtenido los resultados de cada clasificador, procedemos a realizar una comparación de ambos clasificadores.

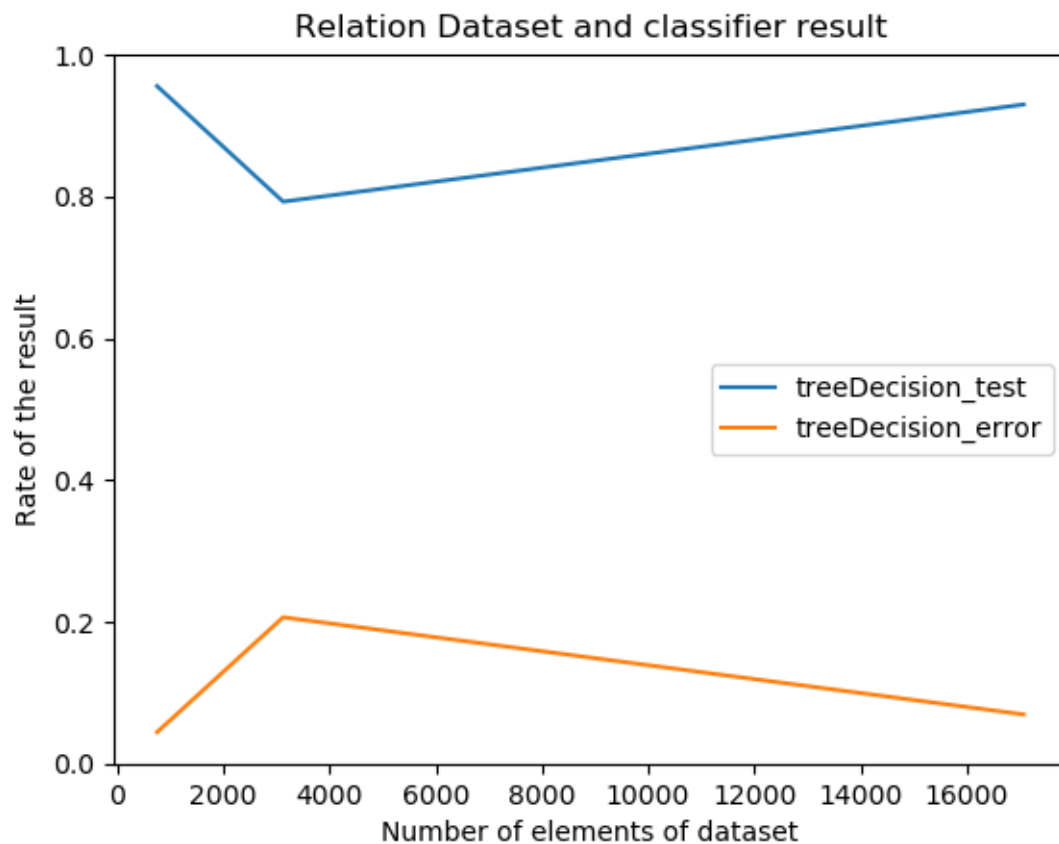


Si observamos la gráfica, podemos comprobar que el clasificador bagging aporta mejores resultados tanto a nivel de test como a nivel de error cometido.

Capítulo 3

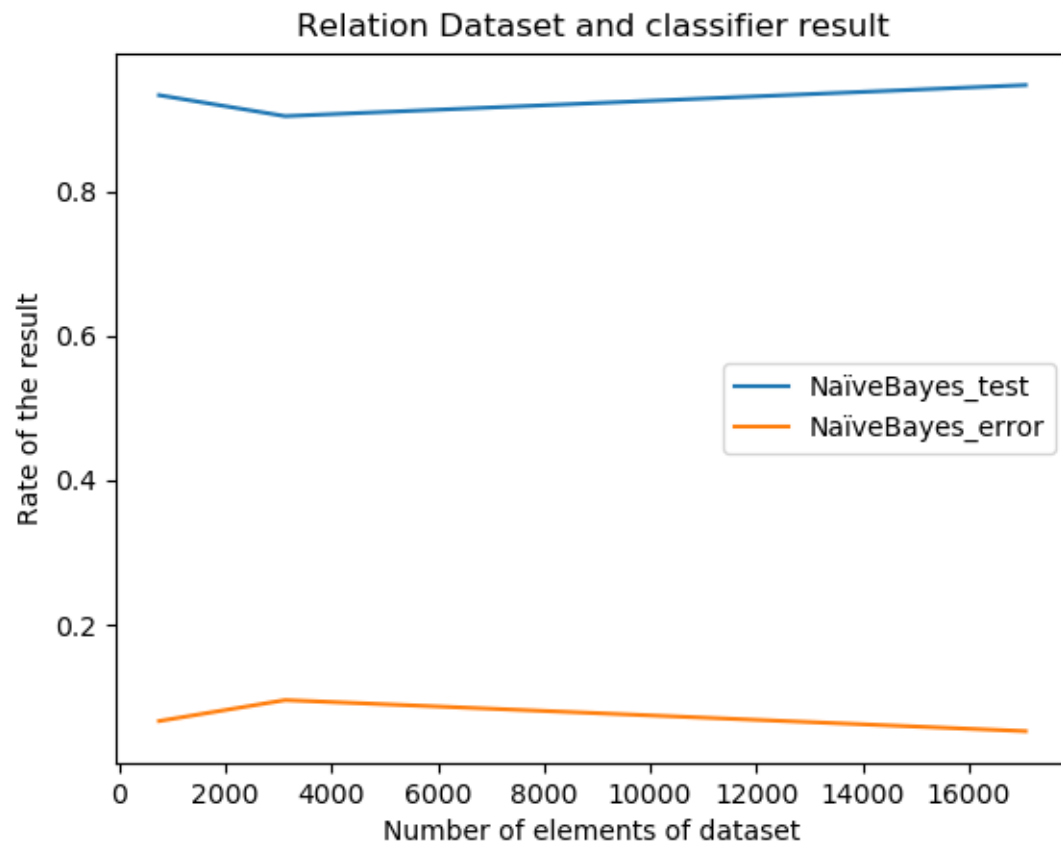
Compare el rendimiento de al menos dos métodos basados en boosting, esto incluye las diferentes configuraciones de AdaBoost.

Estudie la evolución del error a medida que se añaden clasificadores para alguno de los problemas



Si observamos la gráfica, observamos que el ratio de test para los dataset utilizados es elevado (se encuentra en el intervalo $[0.8,1]$); y el ratio de error cometido es bajo (se encuentra en el intervalo $[0,0.2]$).

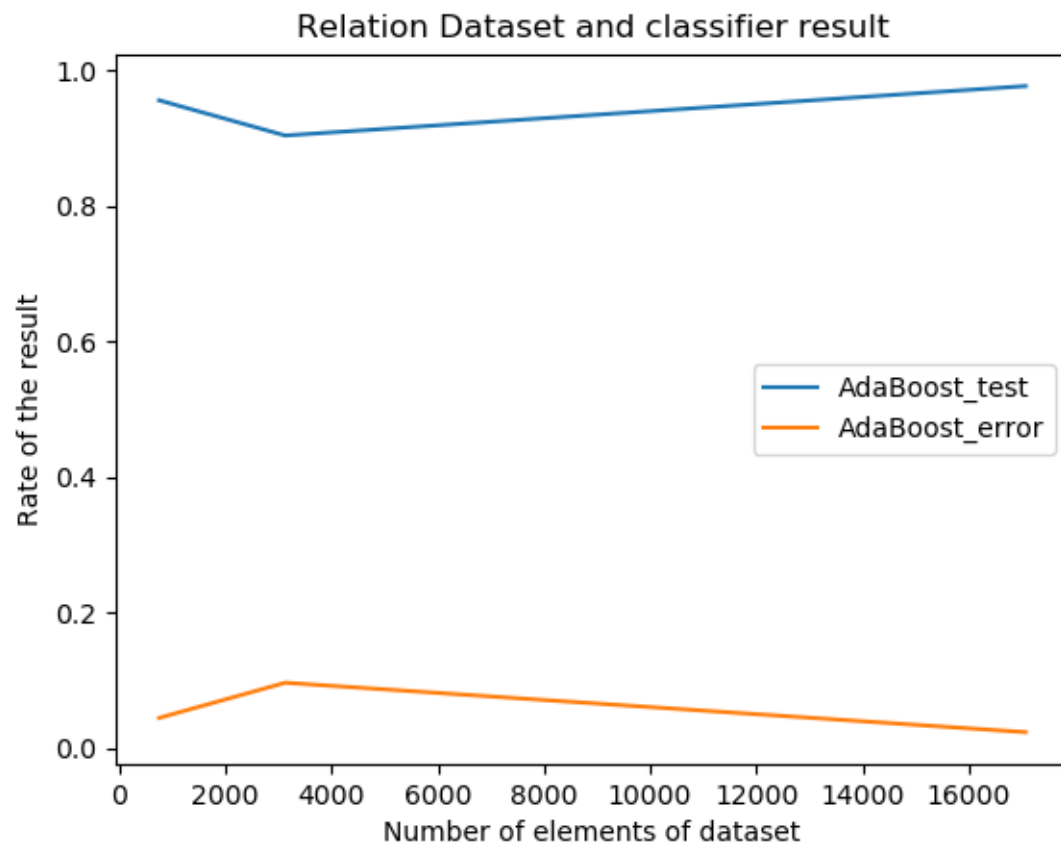
3.1. Clasificador Naïve Bayes



Si observamos la gráfica, observamos que el ratio de test para los dataset utilizados es muy elevado (entre 0.9 y 1); mientras que el ratio de error cometido es muy bajo (entre 0 y 0.1)

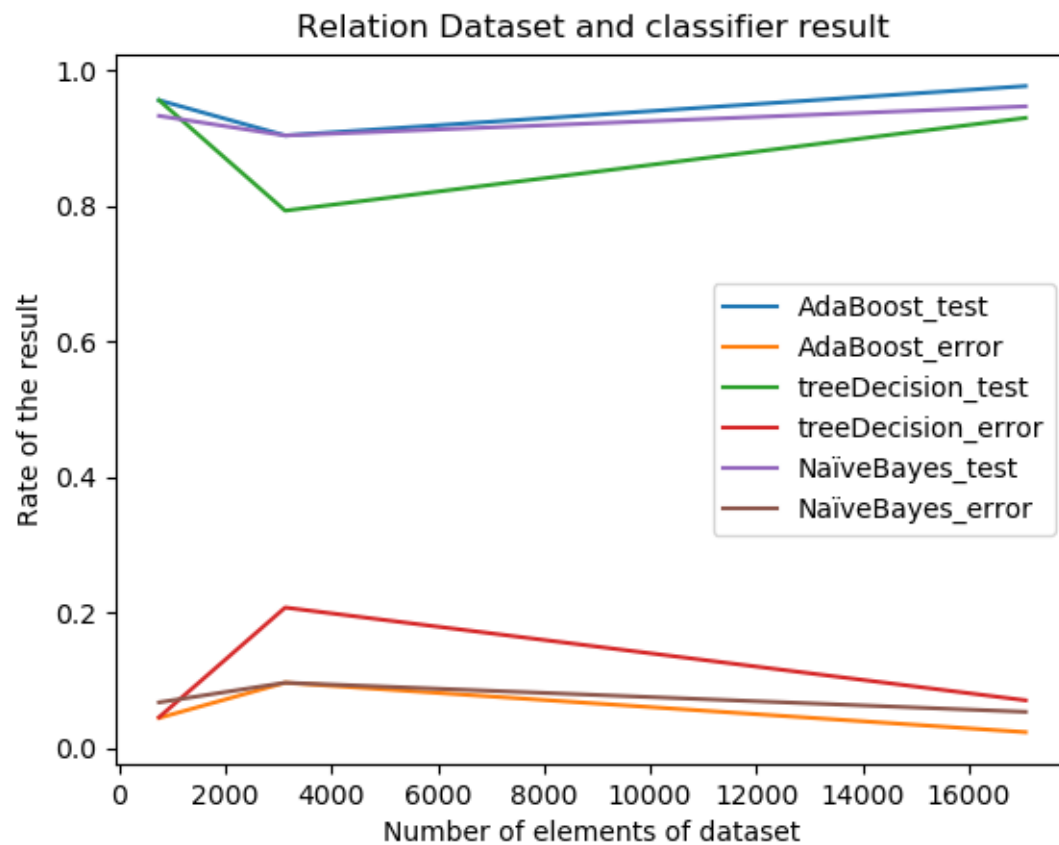
3.2. Clasificador AdaBoost

En la realización del clasificador, se han utilizado 200 estimaciones.



Si observamos la gráfica, vemos que con este clasificador hemos obtenido unos resultados muy buenos tanto para los test como el error cometido.

A continuación, realizamos un estudio de los datos obtenidos por cada clasificador:

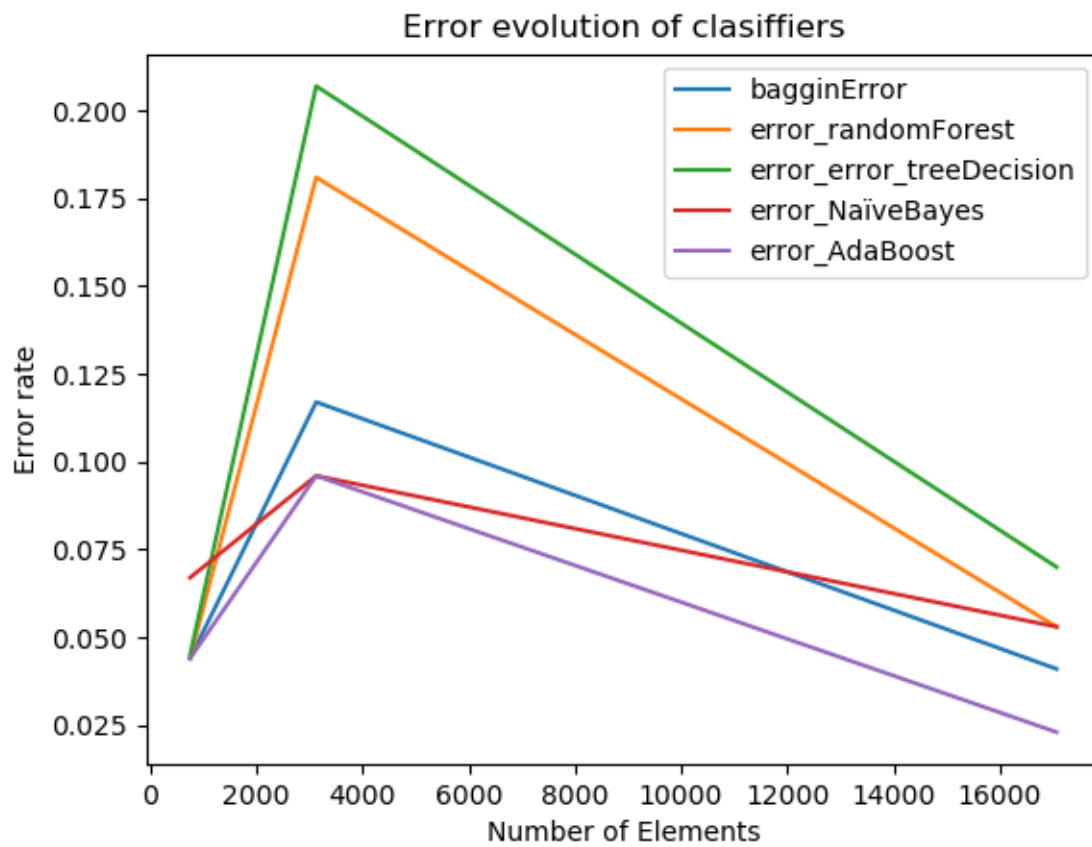


Si observamos la gráfica, podemos observar que para los problemas empleados el orden de mejor a peor clasificador es el siguiente:

1. AdaBoost
2. Naïve Bayes
3. TreeDecision

3.3. Evolución del error

A continuación, se muestran las gráficas de la evolución del error para los problemas empleados:



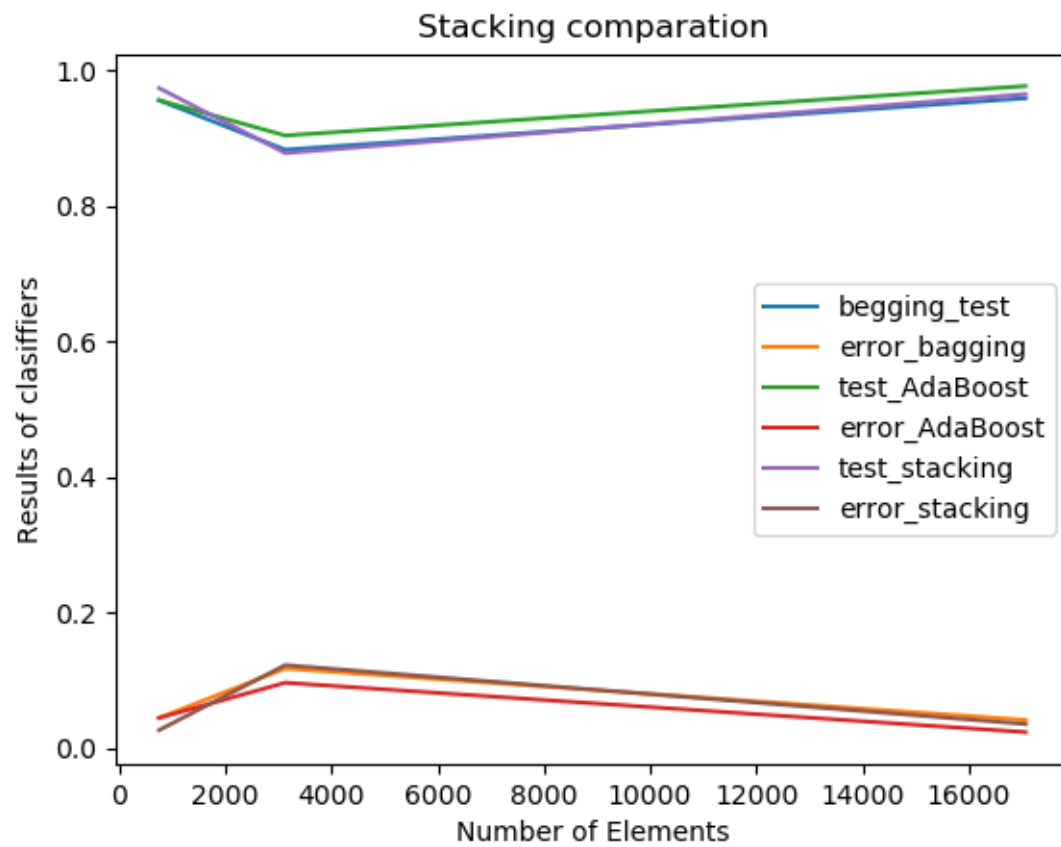
Si observamos la gráfica, vemos que a medida que añadimos un clasificador el error cometido aumenta.

Capítulo 4

Ejercicio 3. Compare el rendimiento de los métodos de stacking con el mejor de los métodos de cada uno de los apartados anteriores.

Para la comparación de los métodos, realizaremos la comparación del rendimiento de los métodos de stacking con los clasificadores siguientes:

- Clasificador Bagging
- Clasificador AdaBoost



Si observamos la gráfica, comprobamos que el método stacking tiene un mejor rendimiento que el clasificador Bagging, pero tiene un peor rendimiento que el clasificador AdaBoost.