Práctica 3: Clasificación y evaluación de modelos

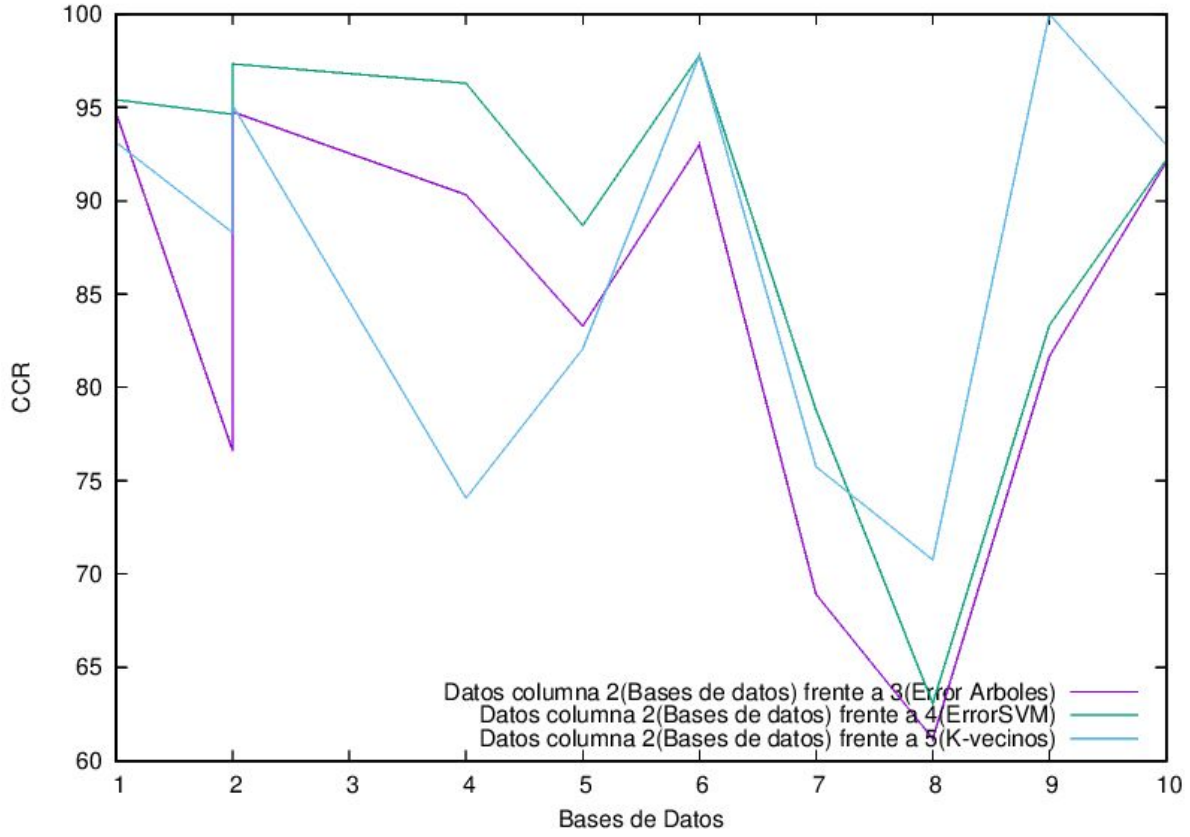
Asignatura: Introducción a la Minería de Datos, 4º Grado de Ingeniería Informática Escuela Politécnica Superior de Córdoba - Universidad de Córdoba 2020 - 2021

Trabajo realizado por:
-Antonio Gómez Giménez (32730338G)
i72gogia@uco.es

Índice:

Ejercicio 3

Los resultados obtenidos tras hacer un k-fold de 10 y usando 3 clasificadores (árboles, vecinos más cercanos y svm), sobre 10 bases de datos, son los siguientes:



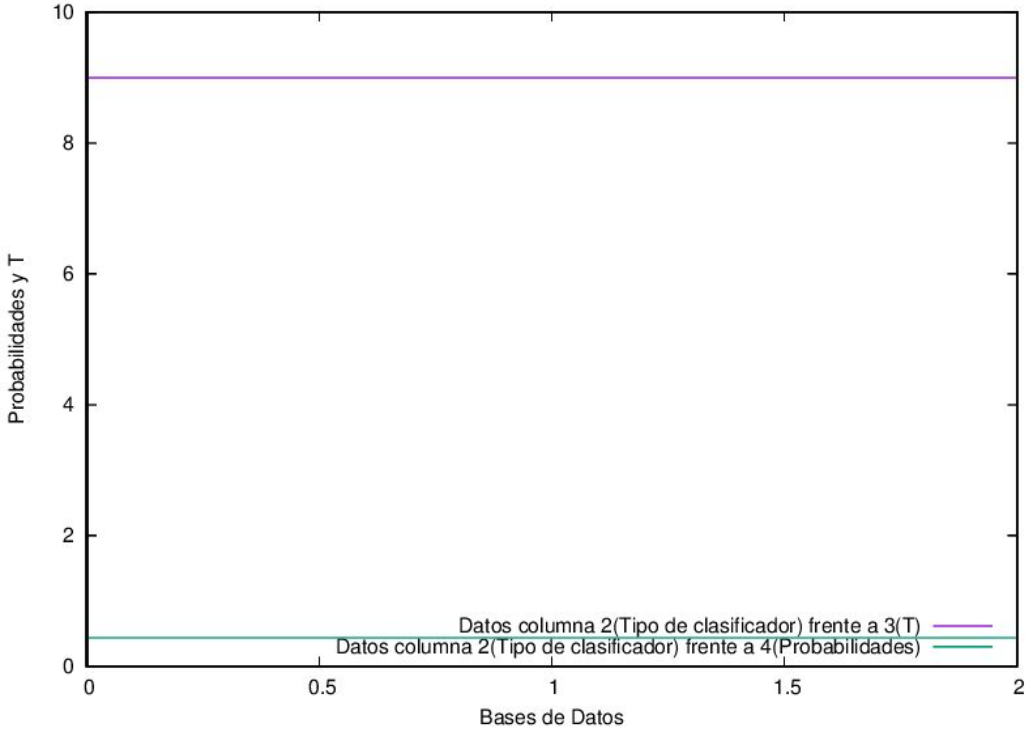
Las bases de datos son las siguientes:
1->vote_mod
2->soybean_mod
3->segment-challenge
4->wine
5->ionosphere
6->iris
7->diabetes
8->glass
9->labor_mod
10->segment-test_mod

El color morado representa el CCR resultante del clasificador de árboles para cada base de datos, el color verde es lo mismo pero para el clasificador SVM y por último, el color azules lo mismo pero para el clasificador de k-vecinos.

Ejercicio 4

Tras usar el test de Wilcoxon de comparación de dos algoritmos sobre todas nuestras bases de datos y aplicándolo a SVM y k-vecinos, el resultado es una T de 11 y un probabilidad de 0.10546875, esto lo que nos quiere decir es que si esta probabilidad es menor que la probabilidad que viene en las tablas de T de wilcoxon respecto a nuestra T, entonces descartamos la hipótesis nula y podemos decir que estos clasificadores entre sí, tienen diferencias significativas.

```
Statistics_wilcoxon= 11.0 ,  0.10546875
Statistics_friedman_vtree= 9.0 ,  0.43727418891386693
Statistics_friedman_vSVM= 9.0 ,  0.43727418891386693
Statistics_friedman_vKvecinos= 8.999999999999998 ,  0.43727418891386716
```



Tras realizar el rango de Friedman para cada clasificador y representarlo gráficamente ocurre algo similar a lo explicado anteriormente, si la probabilidad obtenida por friedman es menor o igual a la probabilidad crítica, entonces descartamos la hipótesis nula y podemos decir que las diferencias entre las diferentes configuraciones son significativas, como podemos observar, en nuestro caso son altas, y esto es lógico, ya que al no variar la configuración y solo modificarse respecto al k-fold, entonces no hay diferencias significativas. Esto es más lógico hacerlo a partir del ejercicio 6 comparando varía configuraciones donde por ejemplo en SVM podemos usar diferentes C.

Respecto a la gráfica se ha realizado sobre la base de datos Vote_mod, ya que para el resto de bases de datos sería repetirlo, como podemos observar, los resultados obtenidos, para todos los tipos de clasificadores son similares, esto es lógico ya que no se les modifican las variables y se basan únicamente en el k-fold.

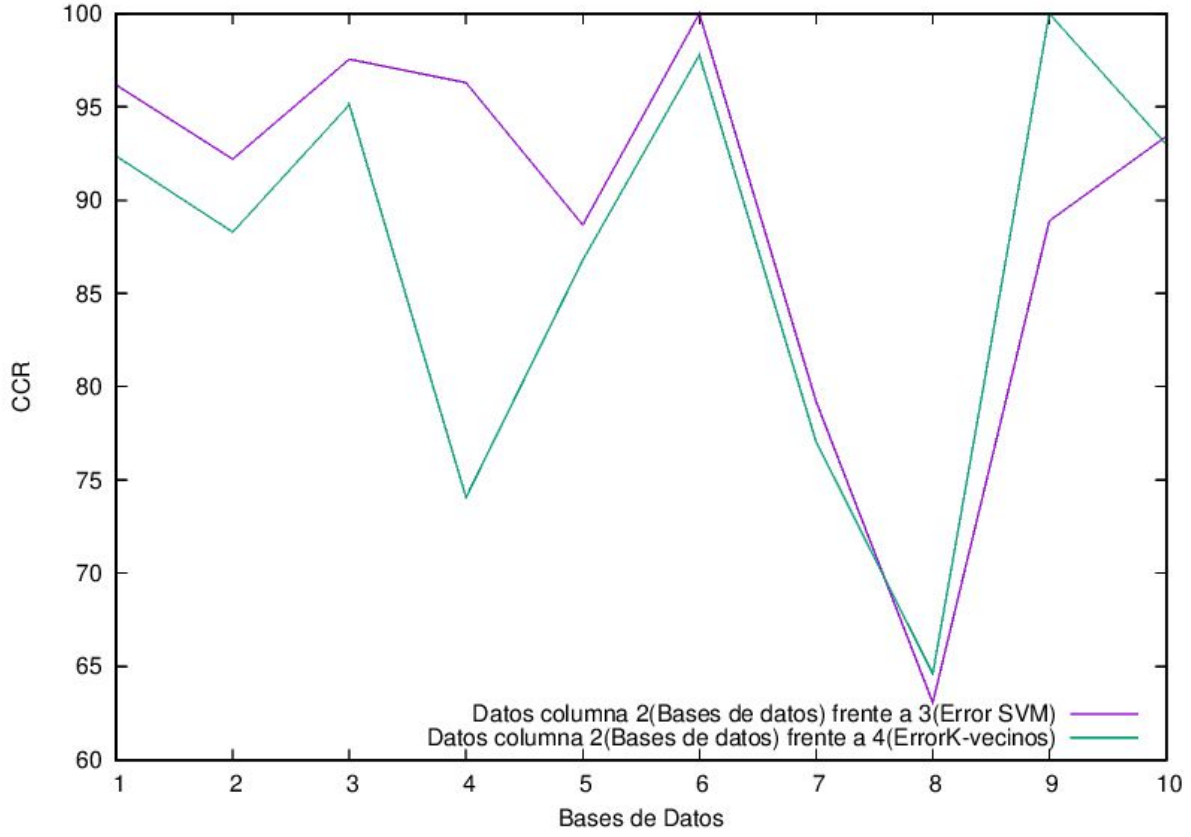
Por último, aplicamos el test de Iman-Davenport sobre los tres clasificadores obteniendo:

```
Statistics_Iman-Davenport= 12.800000000000011 ,  0.0016615572731739255
```

por consiguiente, y como hemos visto antes, encontramos una diferencia significativa entre los tres clasificadores.

Ejercicio 6

Para este ejercicio se ha realizado igual que el ejercicio 3 pero solo con los clasificadores SVM y k-vecinos, usando grid Search para optimizar ciertos parámetros.



Como se puede apreciar, han mejorado un poco los resultados y esto es lógico ya que con esta función se busca optimizar ciertos parámetros permitiendo mejorar el ccr. Pero el tiempo de cómputo se dispara tardando mucho más en obtener los resultados.
Sobre todo esta mejora se nota en el clasificador SVM.