Trabajo Temas 7 y 8 MED (MUM)

Beatriz Coronado Sanz

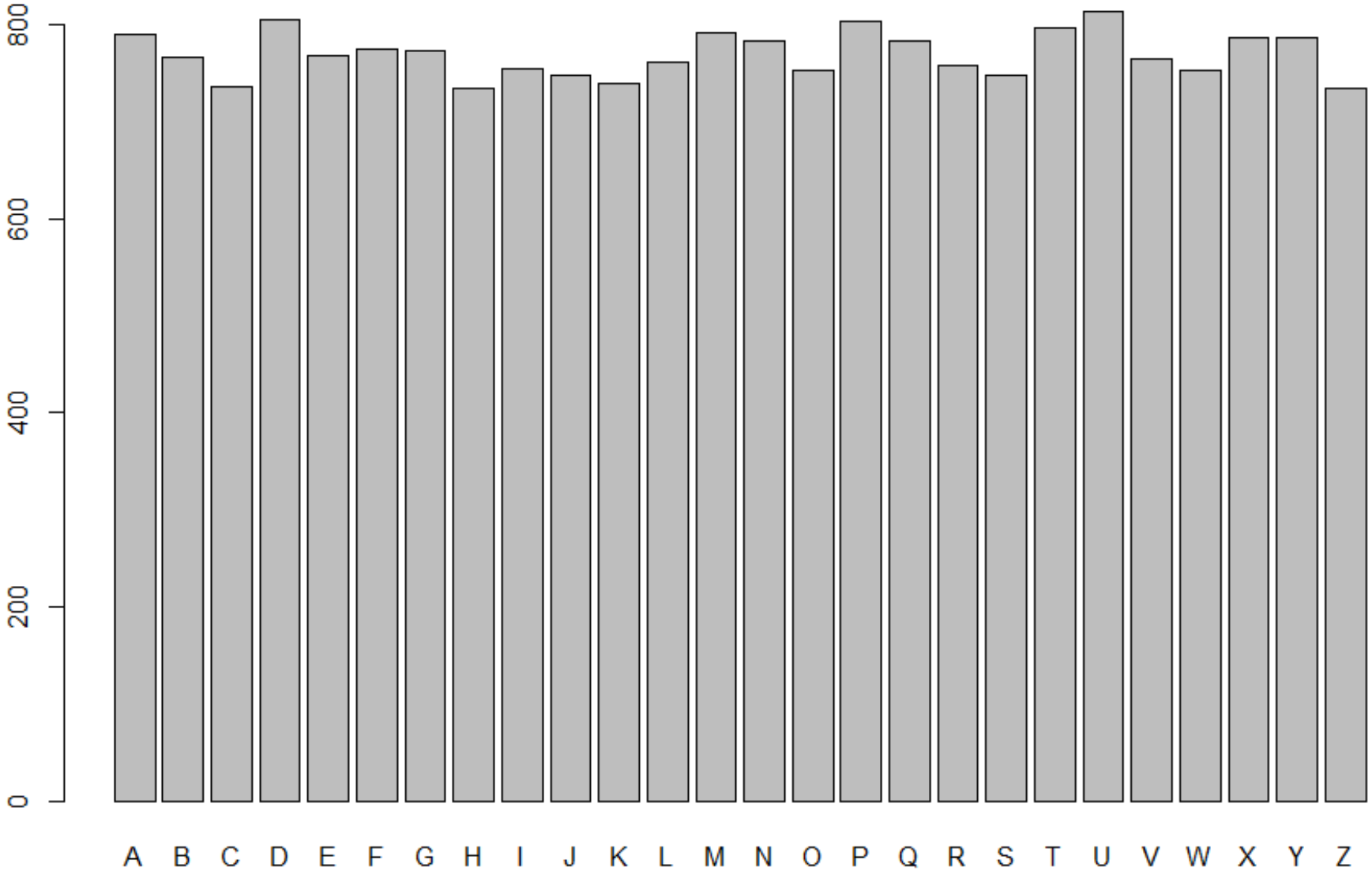
Curso 2018-2019

Índice

Data frame LetterRecognition

- Este data frame contiene imágenes de las distintas letras del alfabeto.
- Se usa para fabricar modelos que reconozcan unas letras de otras. Nosotros realizaremos modelos de clasificación binaria basados en la metodología CART y Random Forests.
- En nuestro caso vamos a trabajar con dos letras elegidas aleatoriamente (la I y la F).

Representación de LetterRecognition



Partición en conjunto de entrenamiento/test

- Trabajaremos con todos los casos de las dos letras escogidas.
- Partiremos este conjunto de datos en dos: un conjunto de entrenamiento (70 %) y un conjunto de test (30 %).
- Observamos que los datos del conjunto de entrenamiento son equilibrados: 539 para la letra F y 532 para la letra I
- Del mismo modo, los datos del conjunto test son equilibrados: 236 para la letra F y 223 para la letra I

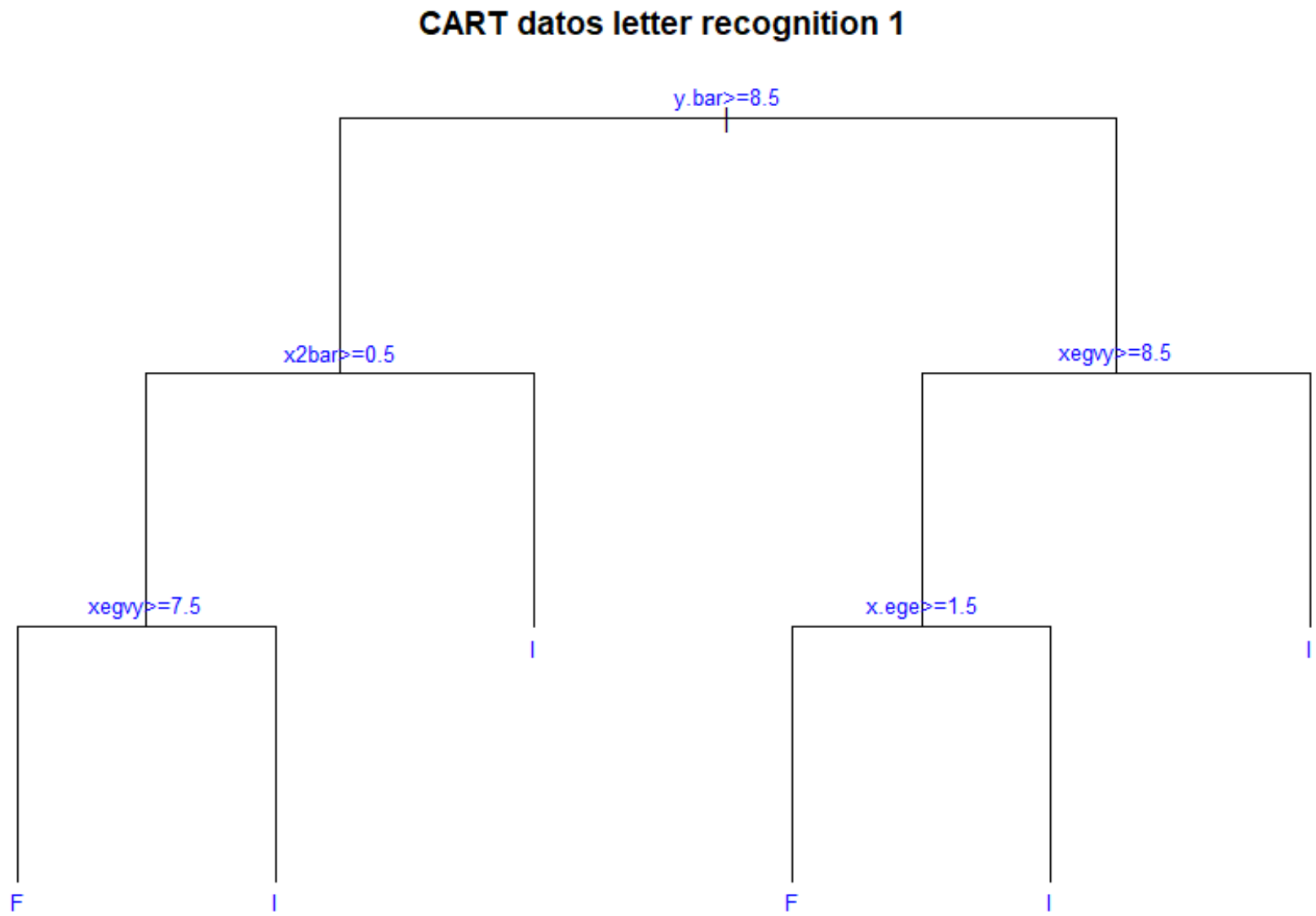
Primer modelo CART

- El primer modelo CART que estudiaremos será el que nos ofrezca la función *rpart* sobre el conjunto de entrenamiento sin ningún parámetro adicional.
- Los resultados de este modelo para el conjunto de test son:

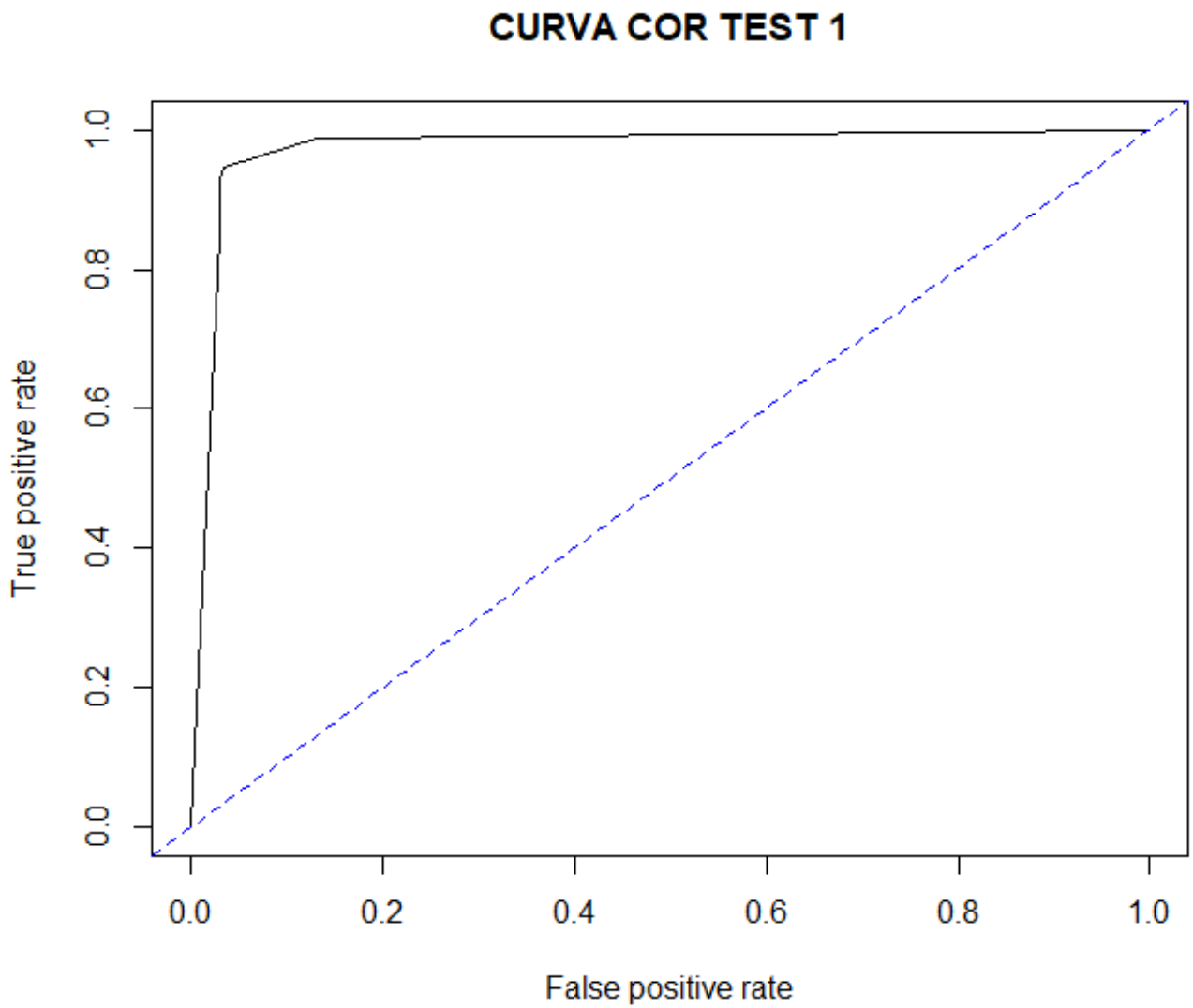
Real/Predicción	**F**	**I**
F	228	8
I	12	211

- El acierto para la letra F es del $96,61\,\%$ y del $94{,}62\,\%$ para la letra I. El acierto total es del $95{,}64\,\%$.

Representación gráfica del árbol obtenido en el primer modelo CART



Curva COR para el primer modelo CART

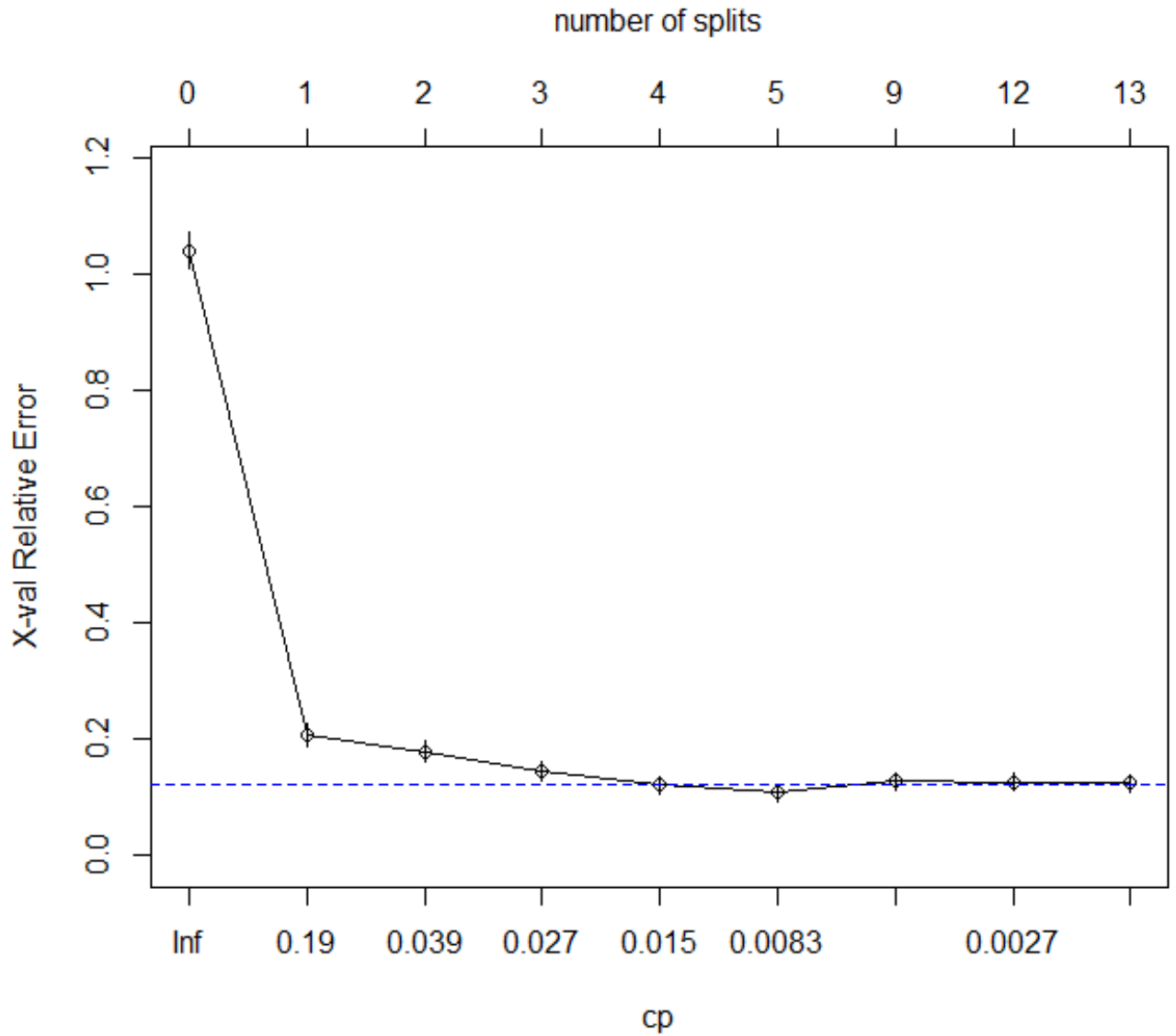


Siendo su AUC igual a 0,97.

Conclusiones del primer modelo CART

- Aunque este primer modelo ya es bastante bueno, vamos a intentar obtener un nuevo modelo en donde la impureza sea menor.
- Para ello añadiremos a la función *rpart* el parámetro $cp = 0{,}01$ y así no se parará hasta que las divisiones tengan un mínimo de impureza igual a esa cantidad.

Reducción de la impureza del segundo modelo CART gráficamente



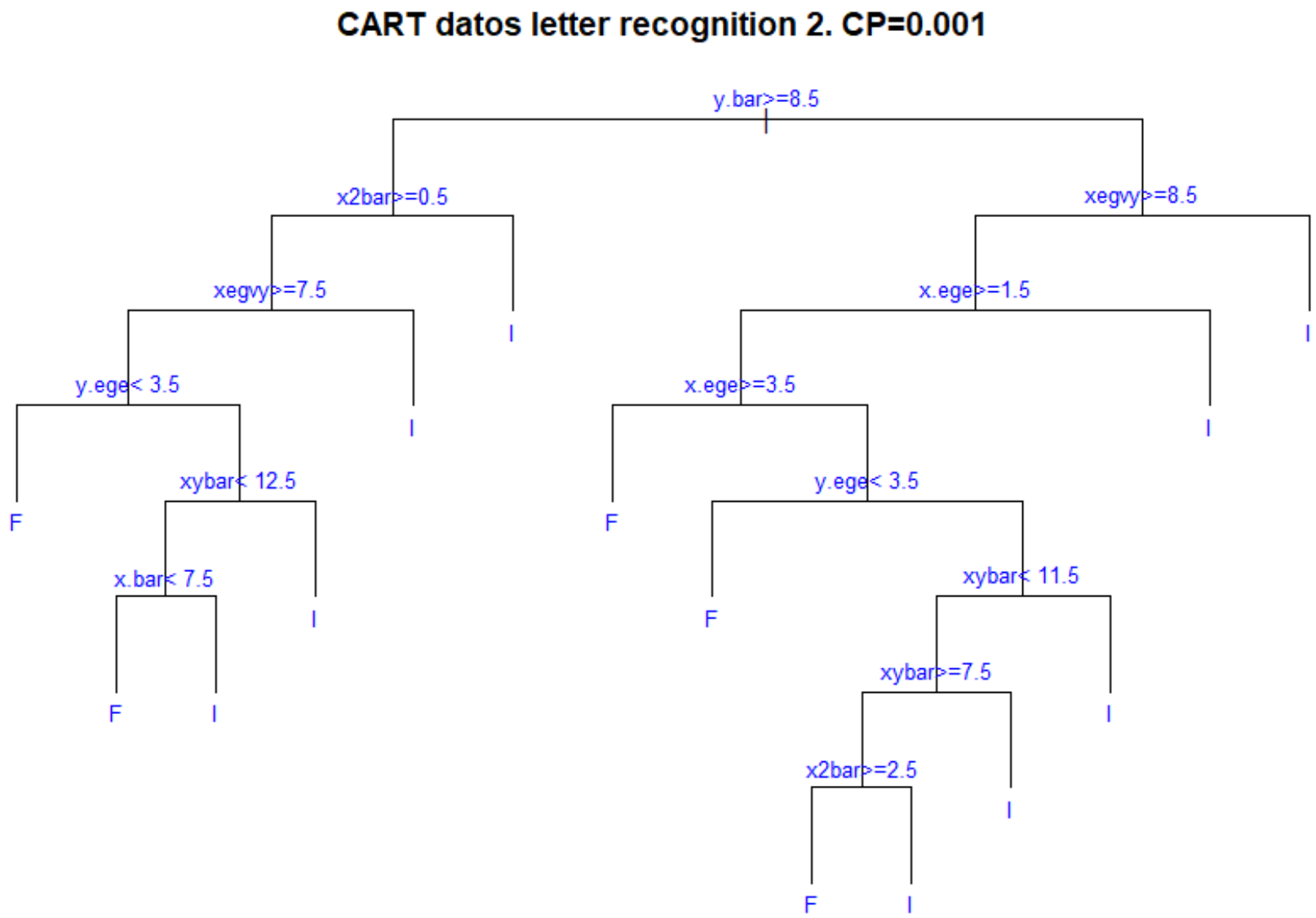
Resultados del segundo modelo CART

- Los resultados de este modelo para el conjunto de test son:

Real/Predicción	**F**	**I**
F	218	18
I	3	220

- El acierto para la letra F es del $92,27\,\%$ y del $98{,}65\,\%$ para la letra I. El acierto total es del $95{,}42\,\%$.
- Observamos que el acierto para la letra I mejora respecto al primer modelo pero empeora para la letra F. El acierto total permanece prácticamente constante.

Representación gráfica del árbol obtenido en el segundo modelo CART



Conclusiones del segundo modelo CART

- Hemos observamos que el árbol obtenido es muy grande, por lo que podría sobreajustar nuestro problema.
- Para solucionar esto utilizaremos la regla 1-ES para saber por donde tenemos que cortar este árbol para que el error VC sea mínimo.
- De esta forma obtenemos nuestro tercer y último modelo para la metodología CART.

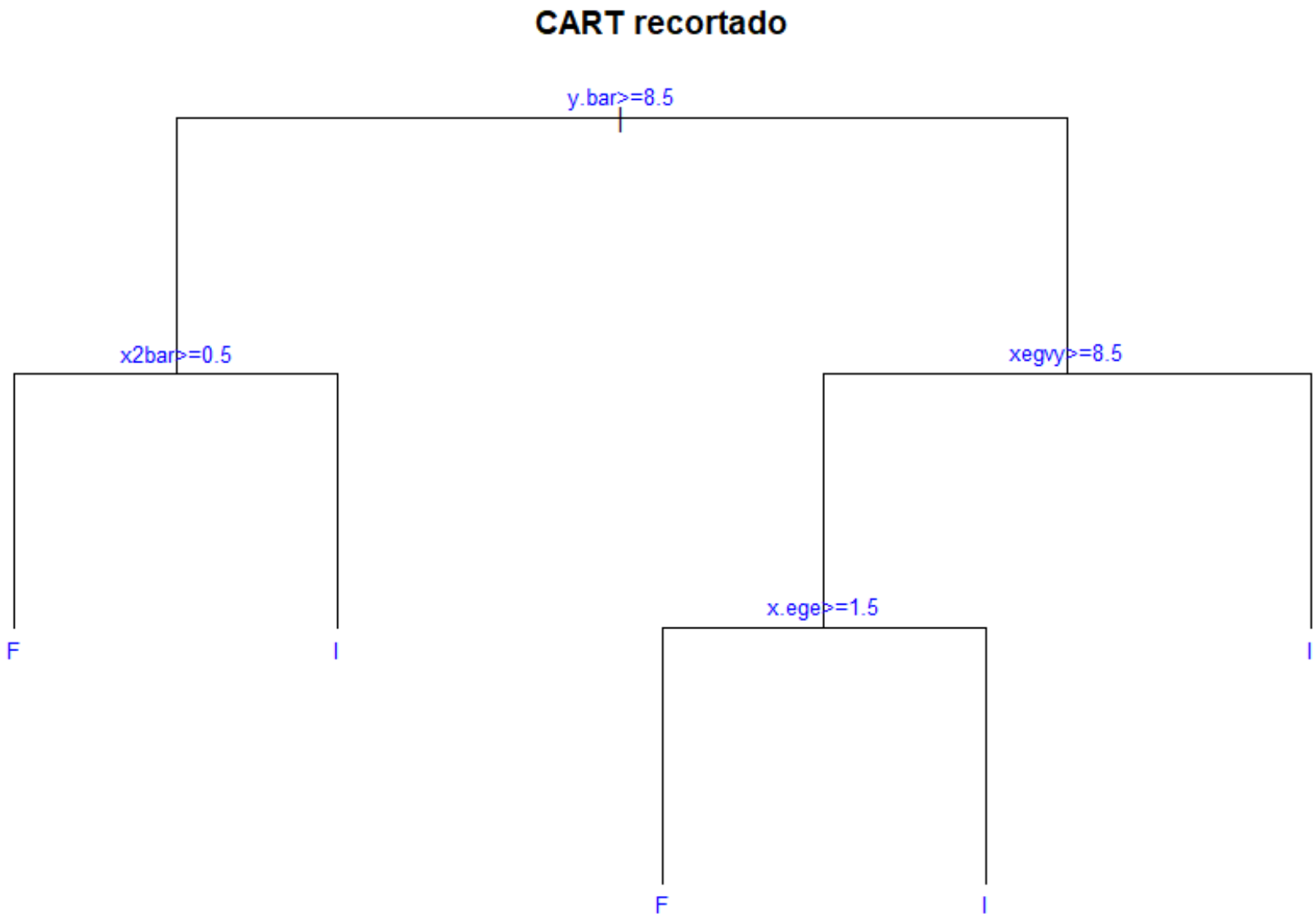
Tercer modelo CART

- Se construye con la función *rpart* a partir del corte óptimo del árbol del segundo modelo.
- Los resultados de este modelo para el conjunto de test son:

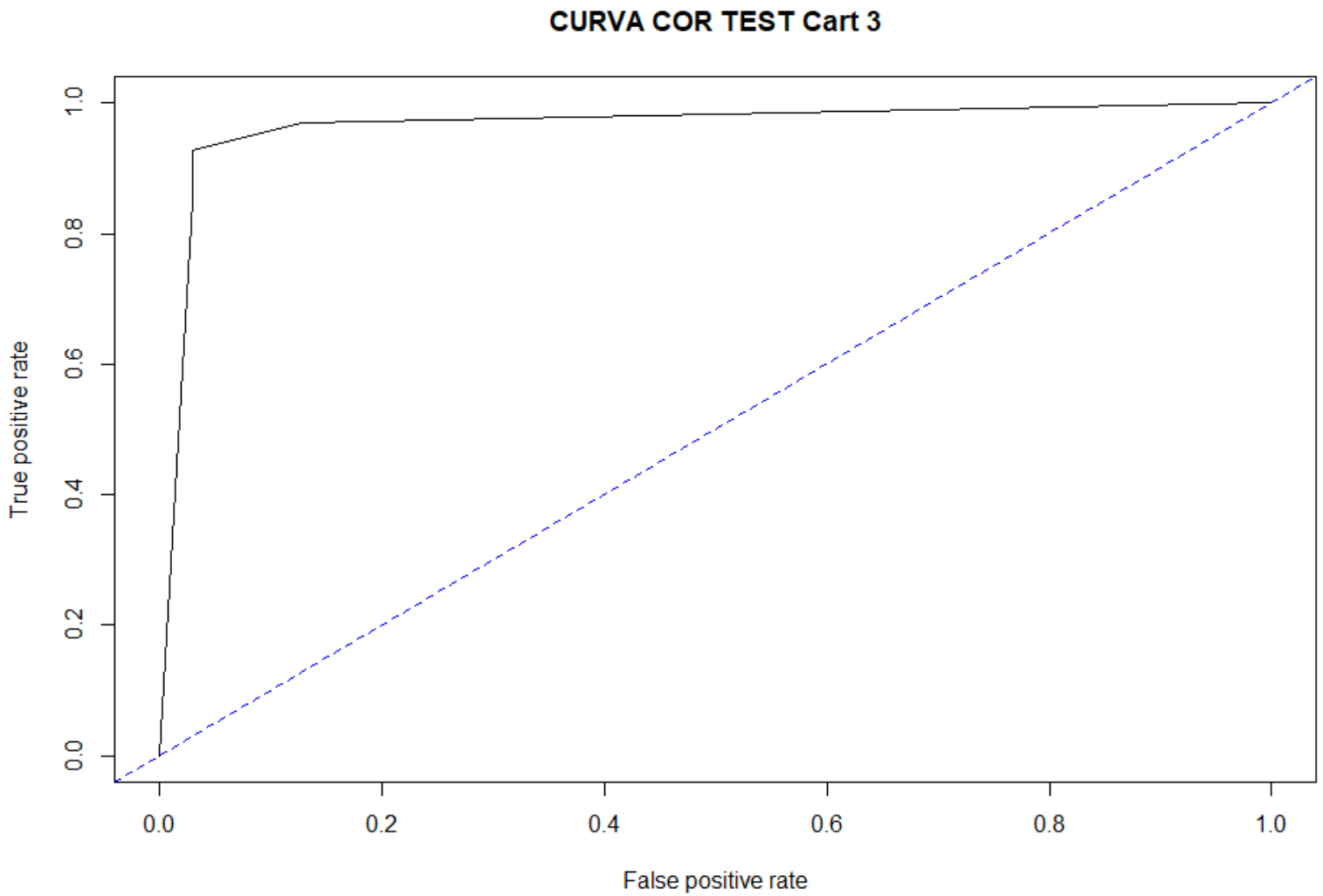
Real/Predicción	**F**	**I**
F	229	7
I	16	207

- El acierto para la letra F es del $97,03\,\%$ y del $92{,}83\,\%$ para la letra I. El acierto total es del $94{,}99\,\%$.
- Observamos que el acierto total es menor que en los otros dos modelos y que se acierta más en la letra F que en la letra I.

Representación gráfica del árbol obtenido en el tercer modelo CART



Curva COR para el tercer modelo CART



Siendo su AUC igual a 0,96.

Conclusiones finales de la metodología CART

- Hemos tenido la suerte, o la desgracia, de que las dos letras escogidas son bastante parecidas entre si. Por lo que todos los modelos sesgan un poco y reconocen más una de las dos letras.
- Aun así, observamos que todos los modelos aciertan por encima del 94 %.
- Además, seguramente nuestros modelos 1 y 3 sean muy robustos a la entrada de nuevos datos.

Curiosidades de los modelos CART obtenidos

- Como curiosidad,se ha añadido en el código una lista de reglas de clasificación para cada uno de los modelos obtenidos.
- También se ha añadido una lista de importancia de las variables para cada modelo. Se observa que las variables *y.bar*, *x.bar* y *x2ybr* son las más importantes para los tres modelos pero sus valores de importancia fluctúan un poco entre modelos. A partir de aquí los resultados se vuelven más heterogéneos.

Primer modelo RF

- Creamos el primer modelo RF con la función *randomForest*
- Los resultados en el conjunto de entrenamiento para este modelo son:

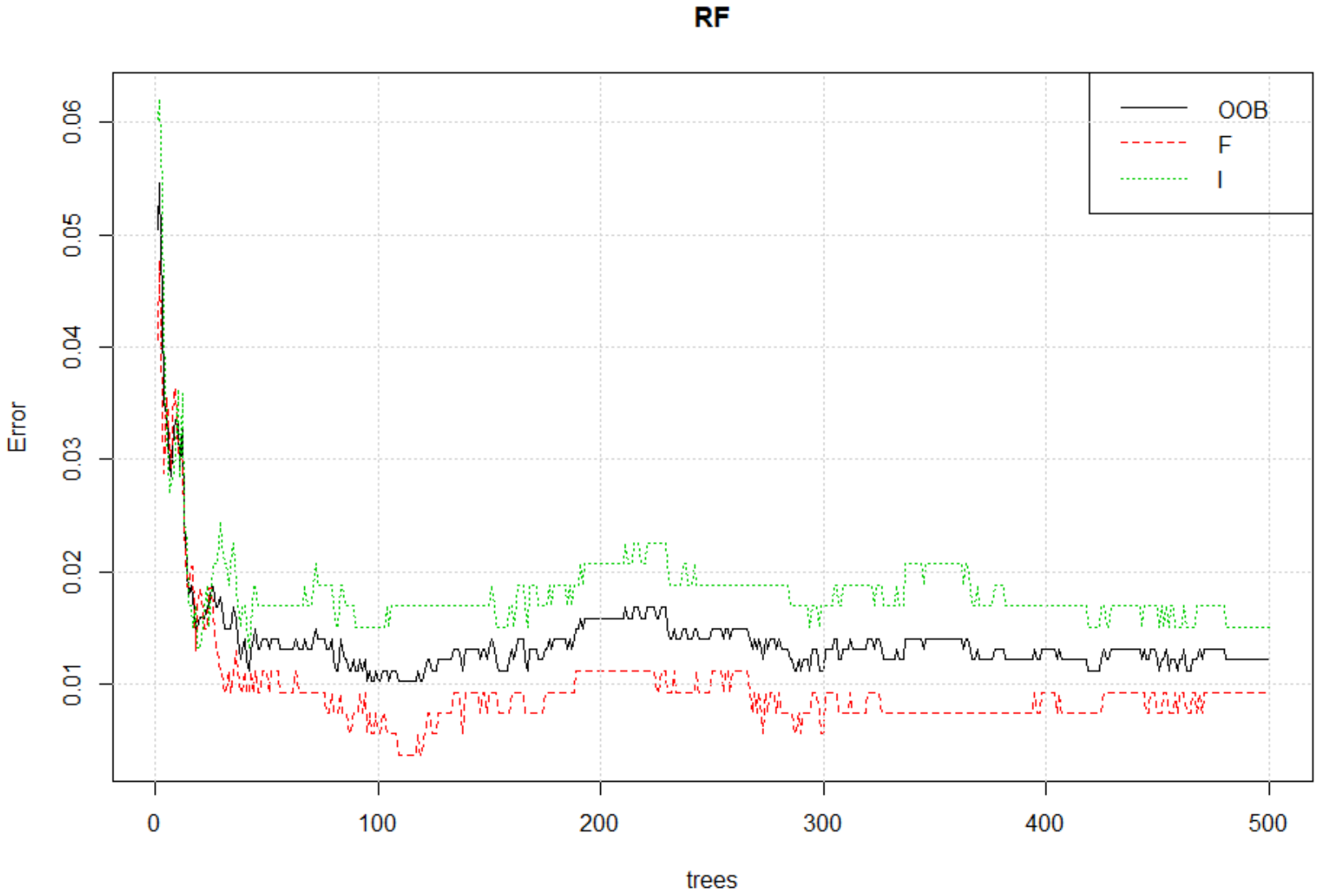
Real/Predicción	**F**	**I**	**Error por clase**
F	533	6	0.011
I	9	523	0.017

- Los resultados en el conjunto de test para este modelo son:

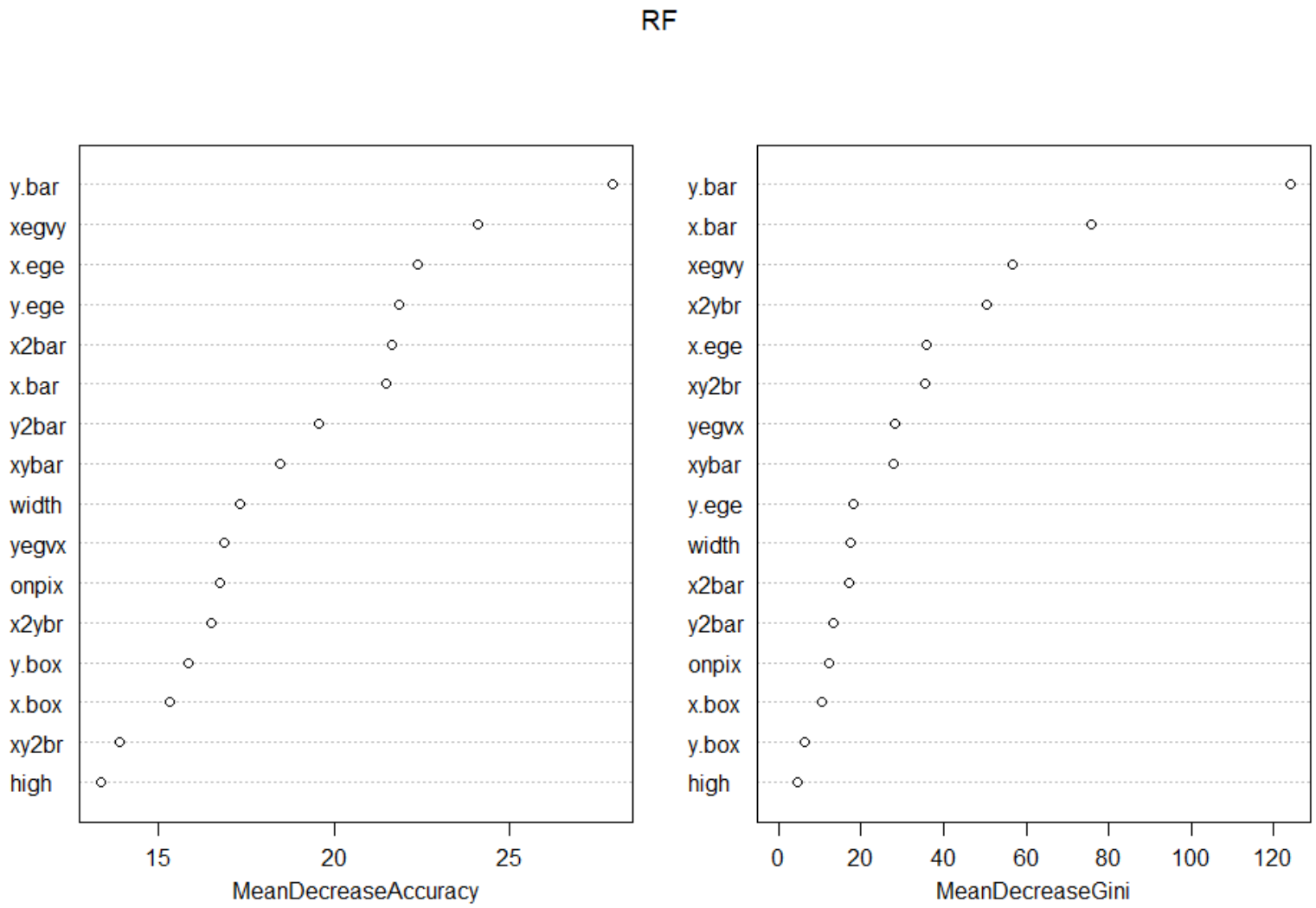
Real/Predicción	**F**	**I**
F	233	3
I	3	220

- El acierto para la letra F es del 98, 73 % y del 98,65 % para la letra I. El acierto total es del 98,69 %.
- En la siguiente gráfica observamos que el error de la letra I siempre es un poco mayor que el de la letra F.

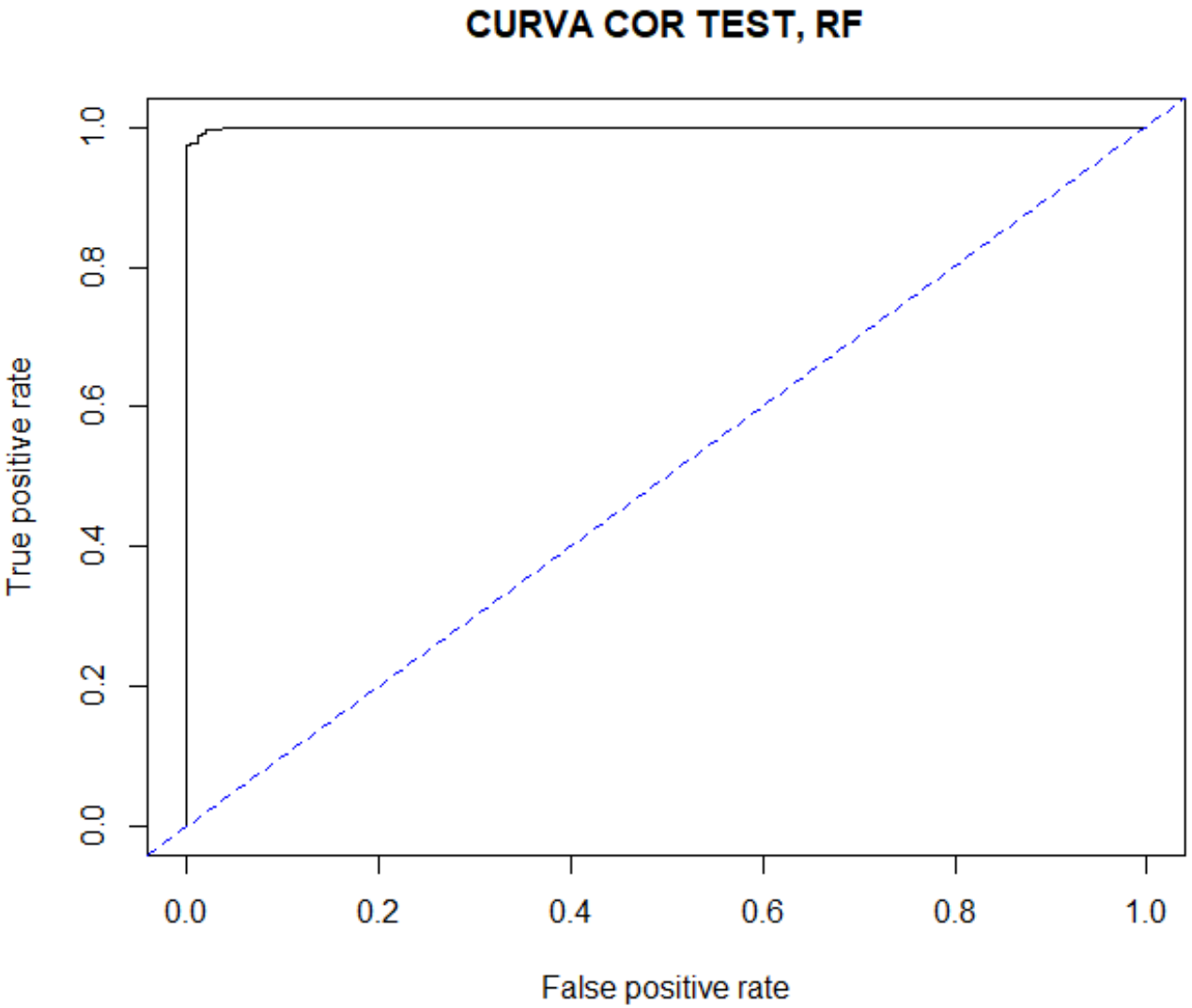
Representación gráfica de los errores en el primer modelo RF



Importancia de las variables en el primer modelo RF



Curva COR para el primer modelo RF

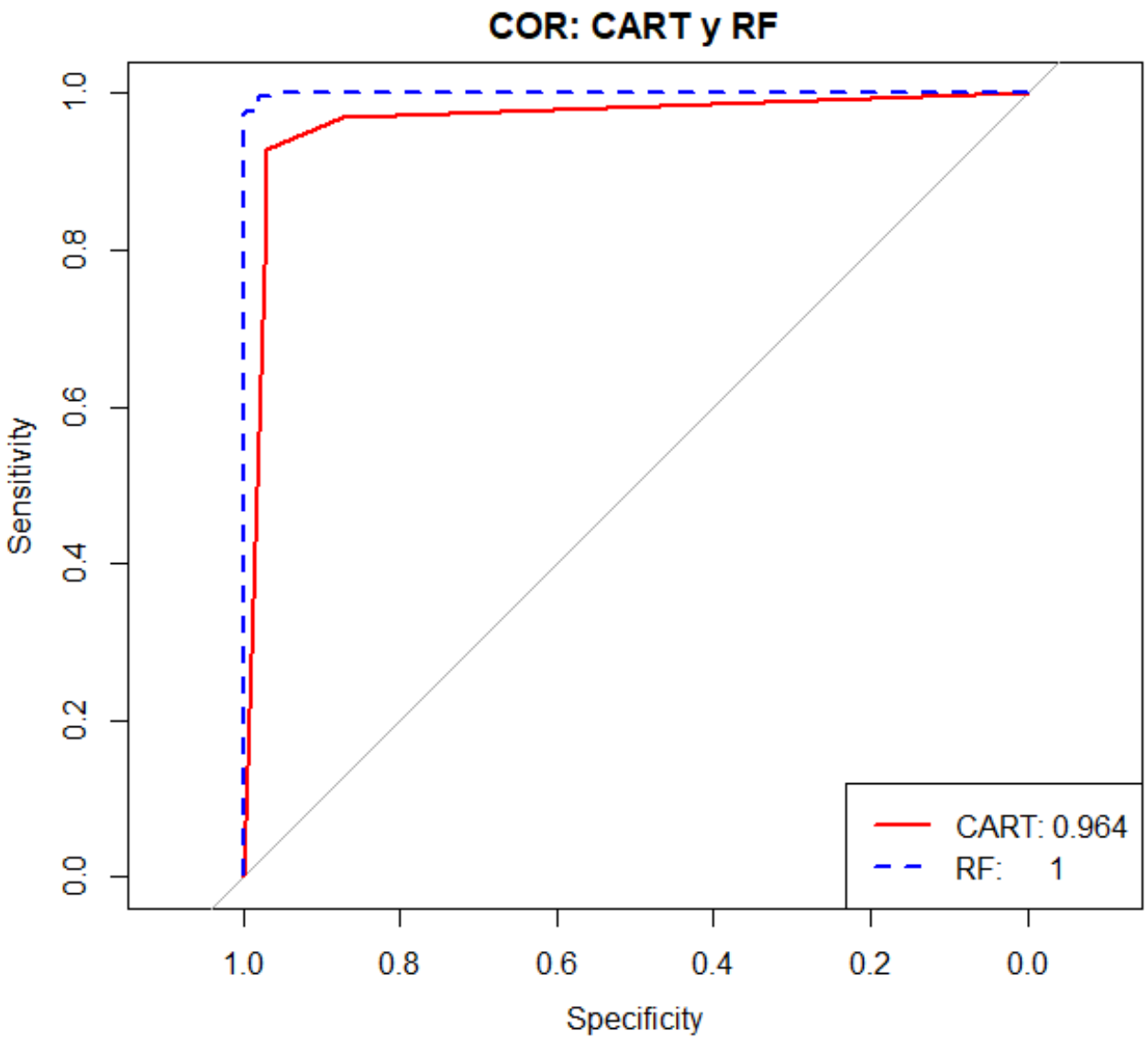


Siendo su AUC prácticamente 1.

Conclusiones del primer modelo RF

- Observamos que la estrategia de construir 500 árboles y elegir para un nuevo dato la clase mayoritaria da bastante mejores resultados que la metología CART.
- En este caso no se produce un sesgo hacia ninguna de las clases y los resultados globales están por encima del 98 %.
- Vemos en la siguiente diapositiva una comparativa de las curvas COR para este modelo de RF y el tercer modelo de la metodología CART.

Comparación entre RF y el modelo 3 de CART



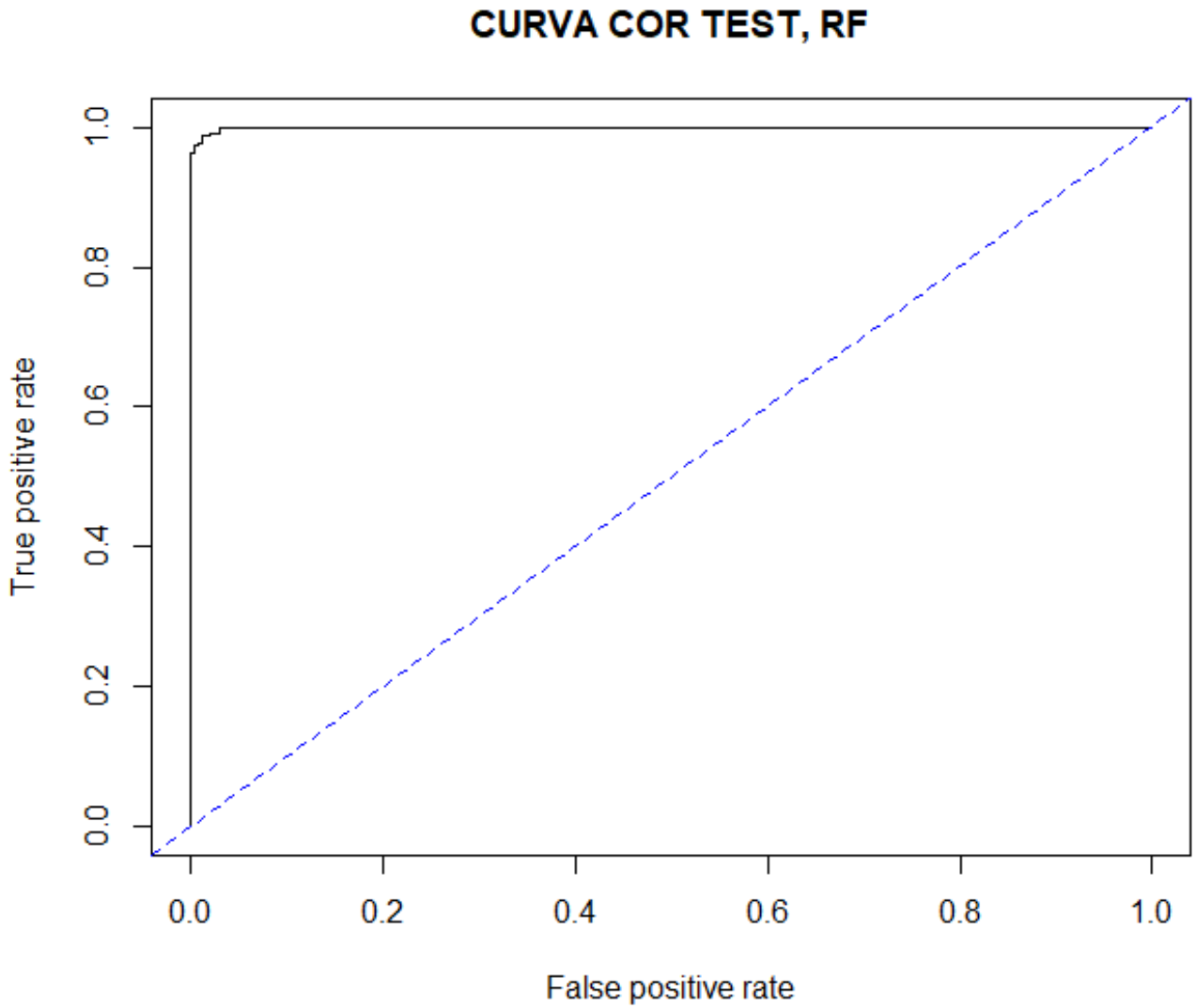
Segundo modelo RF

- Por último, vamos a construir un modelo RF usando procesamiento paralelo.
- Los resultados en el conjunto de test para este modelo son:

Real/Predicción	**F**	**I**
F	232	4
I	3	220

- El acierto para la letra F es del $98,31\,\%$ y del $98{,}65\,\%$ para la letra I. El acierto total es del $98{,}47\,\%$.
- Observamos que los resultados son parecidos a los obtenidos con el primer modelo.

Curva COR para el segundo modelo RF



Siendo su AUC prácticamente 1.

Conclusiones finales

- Hemos comprobado que ambas metodologías son satisfactorias a la hora de reconocer dos letras del alfabeto, incluso cuando estas letras se parecen mucho.
- Tendríamos que comprobar si estos resultados se mantienen aumentando el número de letras a reconocer.