Trabajo Práctico N°3: Clasificador Naive-Bayes

Introducción al Aprendizaje Automatizado

Alumno: Navall, Nicolás Uriel. N-1159/2.

2)

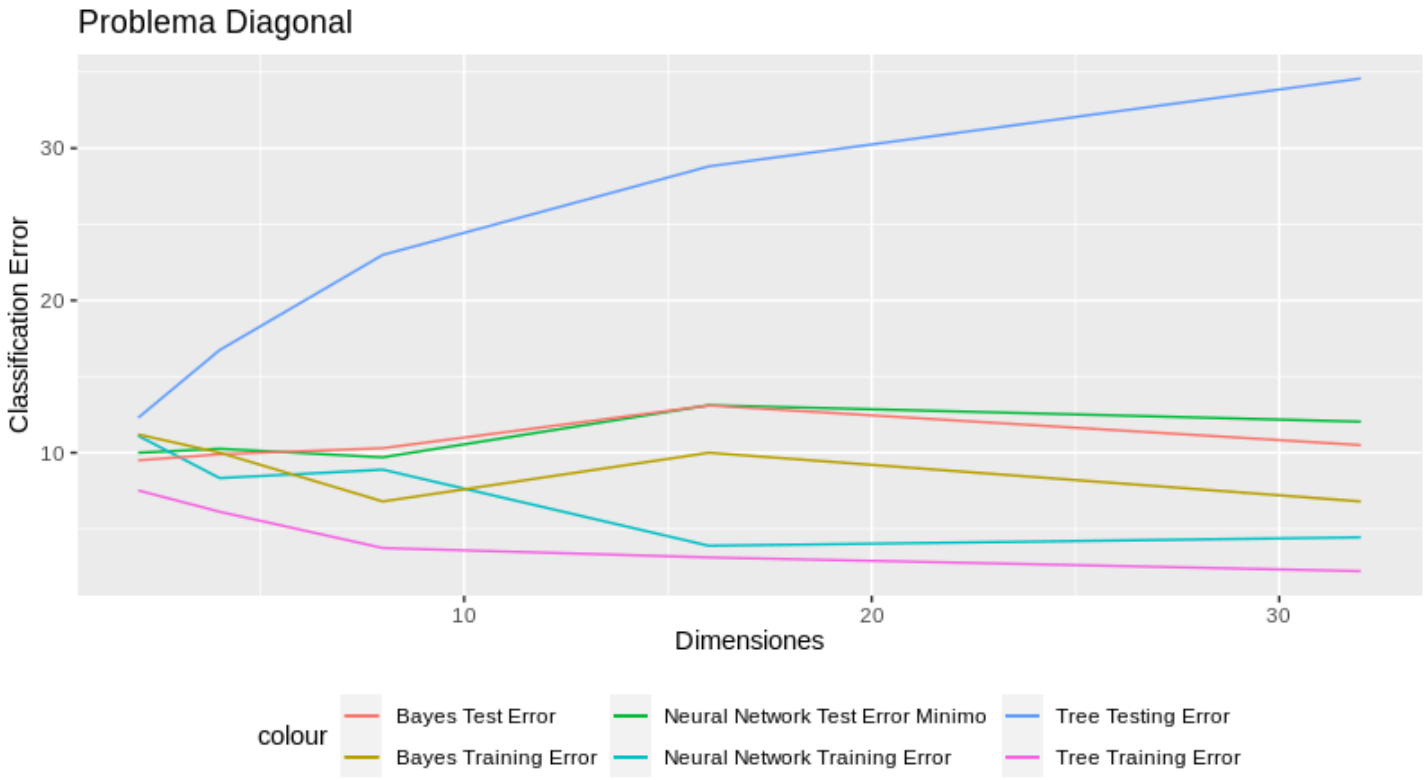

Figura 2.1

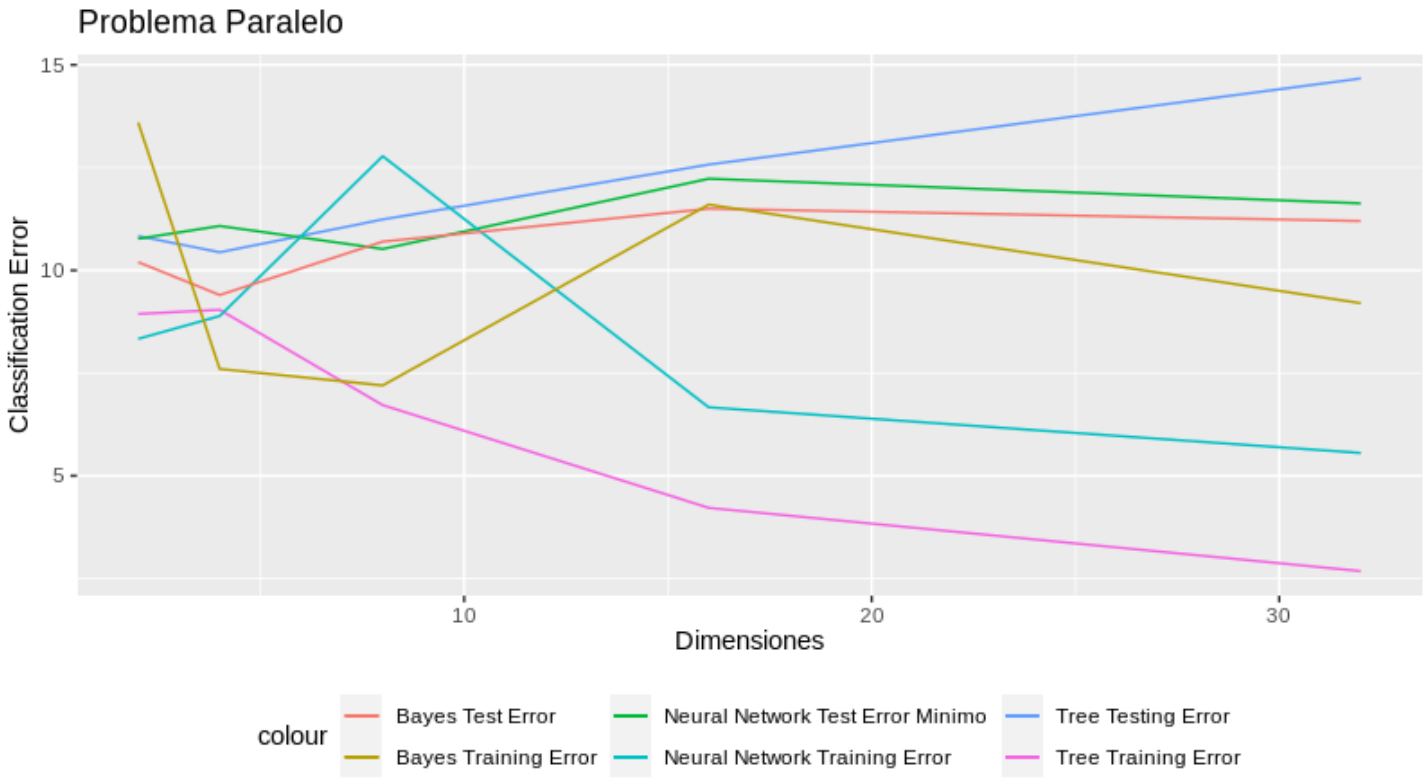

Figura 2.2

Tiene sentido que el error de test de bayes nos de muy buenos resultados en ambos problemas, ya que el método de naive bayes, a diferencia de las redes neuronales o los arboles de clasificacion que son métodos discriminativos, es generativo, por lo cual en lugar de generar “reglas” busca modelar la distribución de las clases. Y como los datos son generados siguiendo el modelo de una normal y para modelar la distribución de la muestra se utiliza una distribución gaussiana, entonces la clasificación resulta ser muy precisa.

Además, el hecho de que los datos son generados por una gaussiana, implica que no hay una regla precisa que pueda utilizarse para clasificar los datos, y es por eso que utilizar un método probabilístico como Naive Bayes es tan ventajoso.

3)

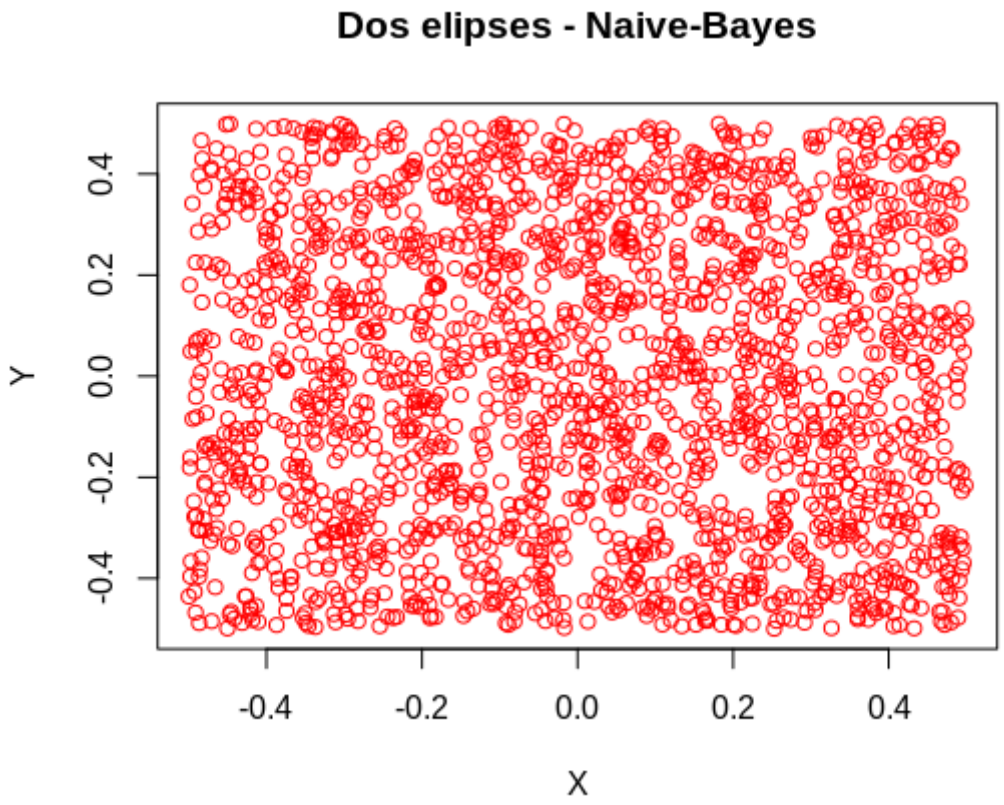

Figura 3.1

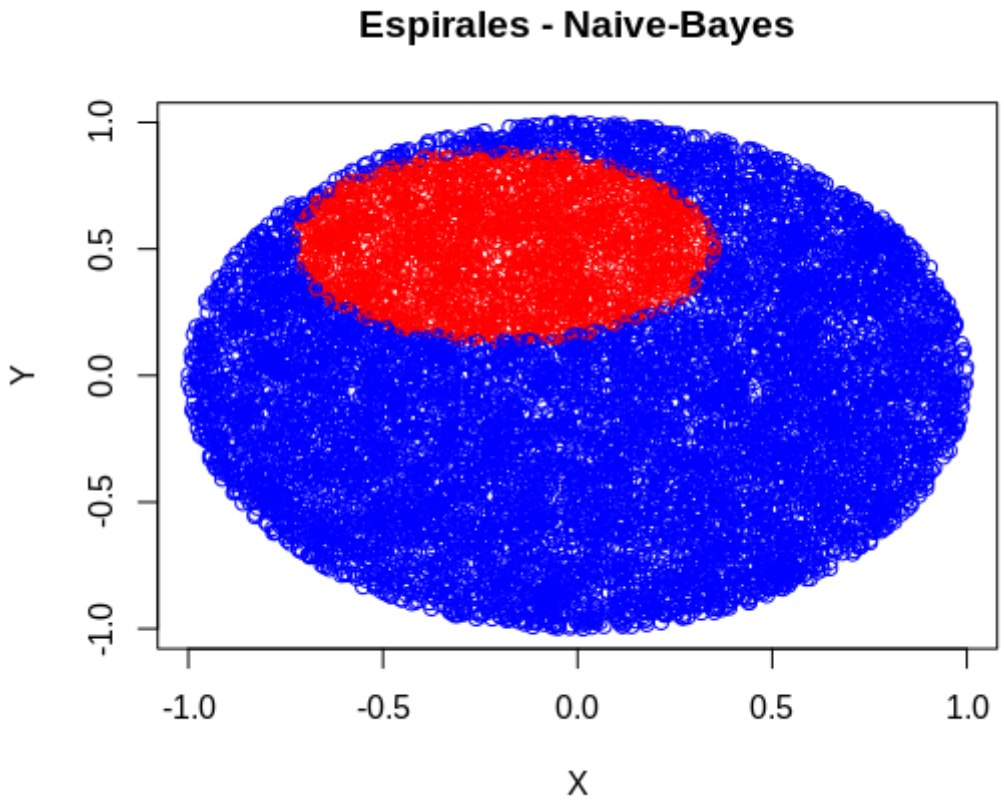

Figura 3.2

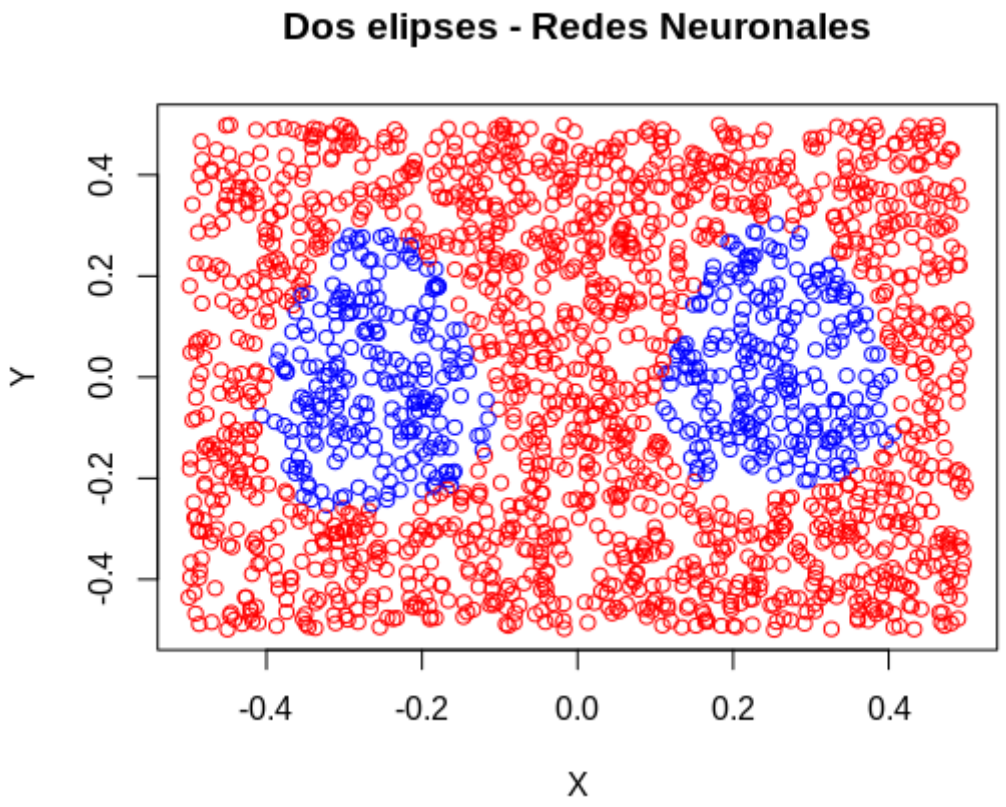

Figura 3.3

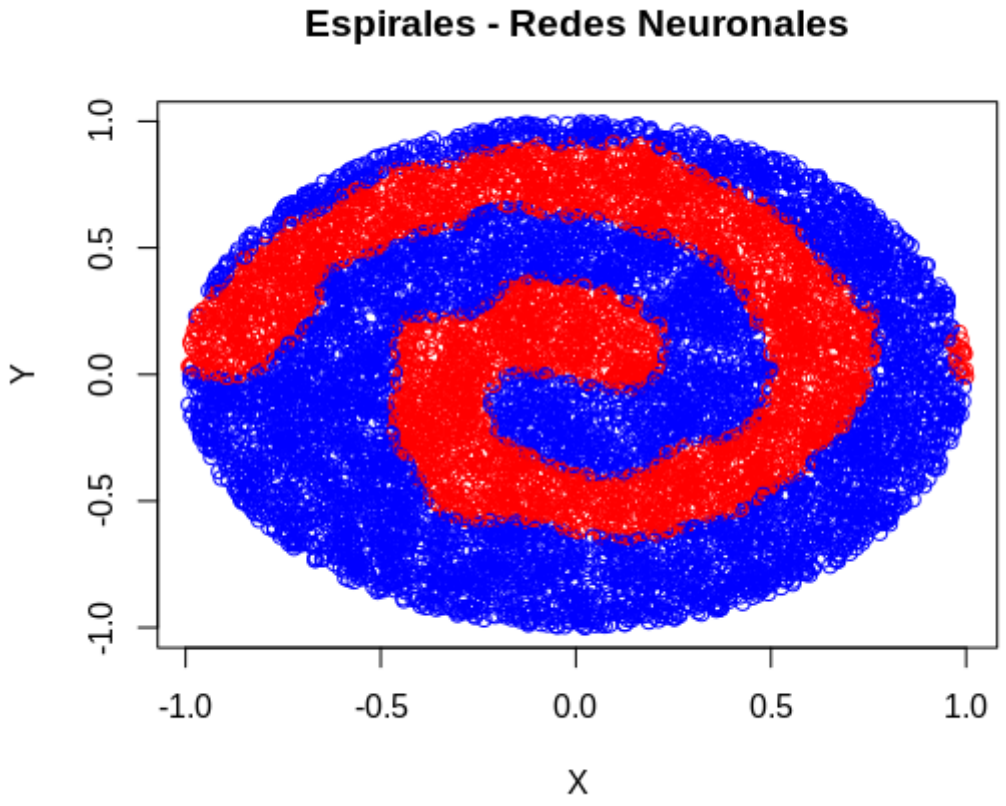

Figura 3.4

Mencionamos anteriormente que el método está modelando el clasificador a partir de una gaussiana, pero esto no resulta beneficioso para problemas cuyos datos no fueron generados siguiendo una distribución normal. En su lugar, los datos de los dos problemas presentados fueron generados por “reglas”, razón por la

cual la implementación de naive-bayes con gaussianas palidece en comparación a las redes neuronales en estos casos.

4)

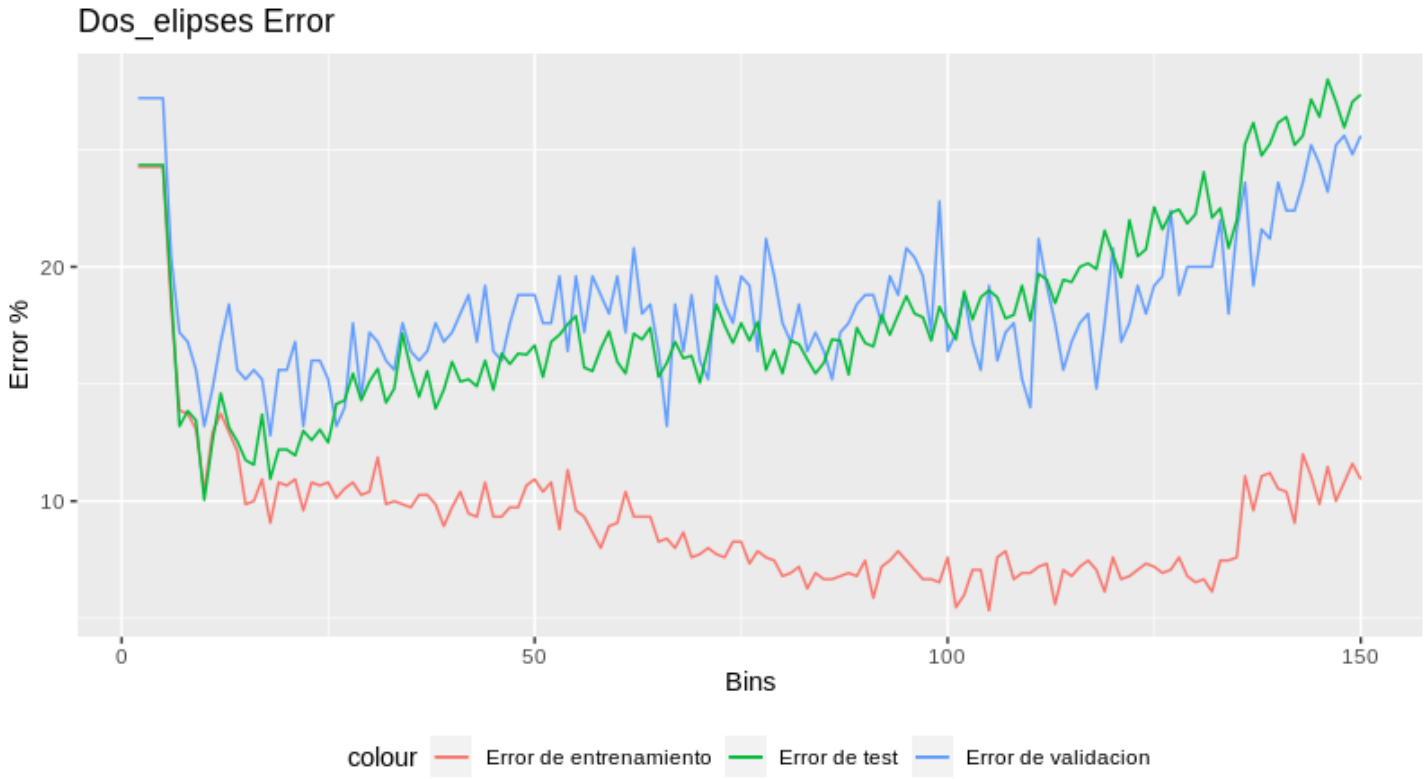

Figura 4.1

Figura 4.2

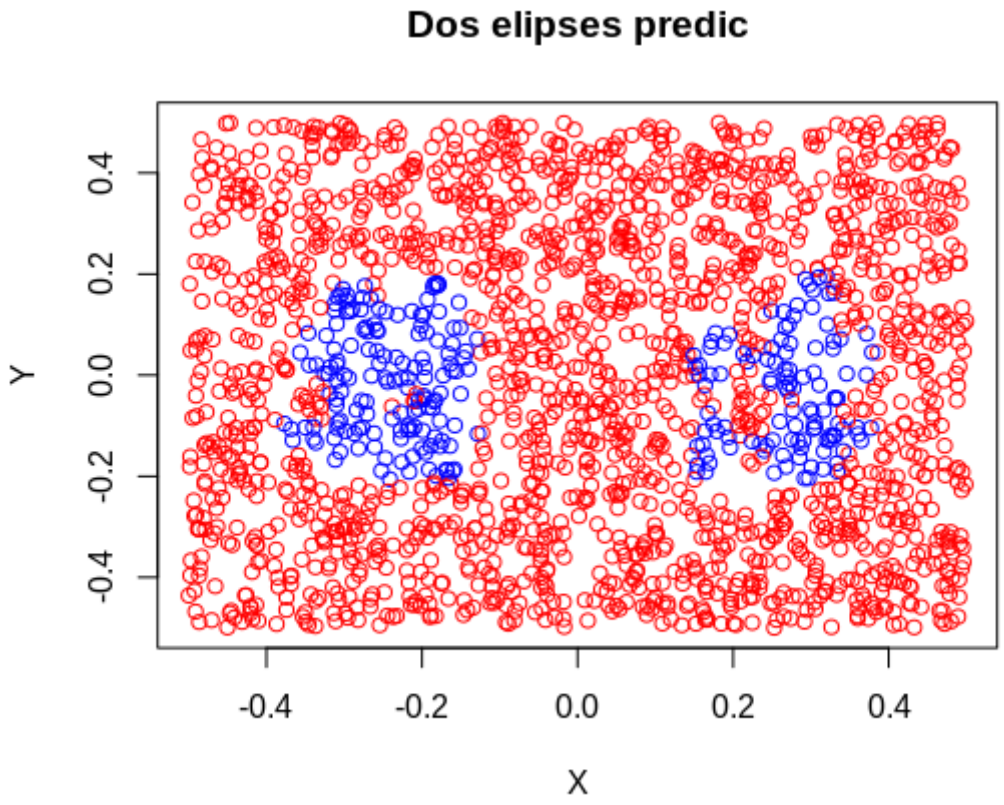

Figura 4.3

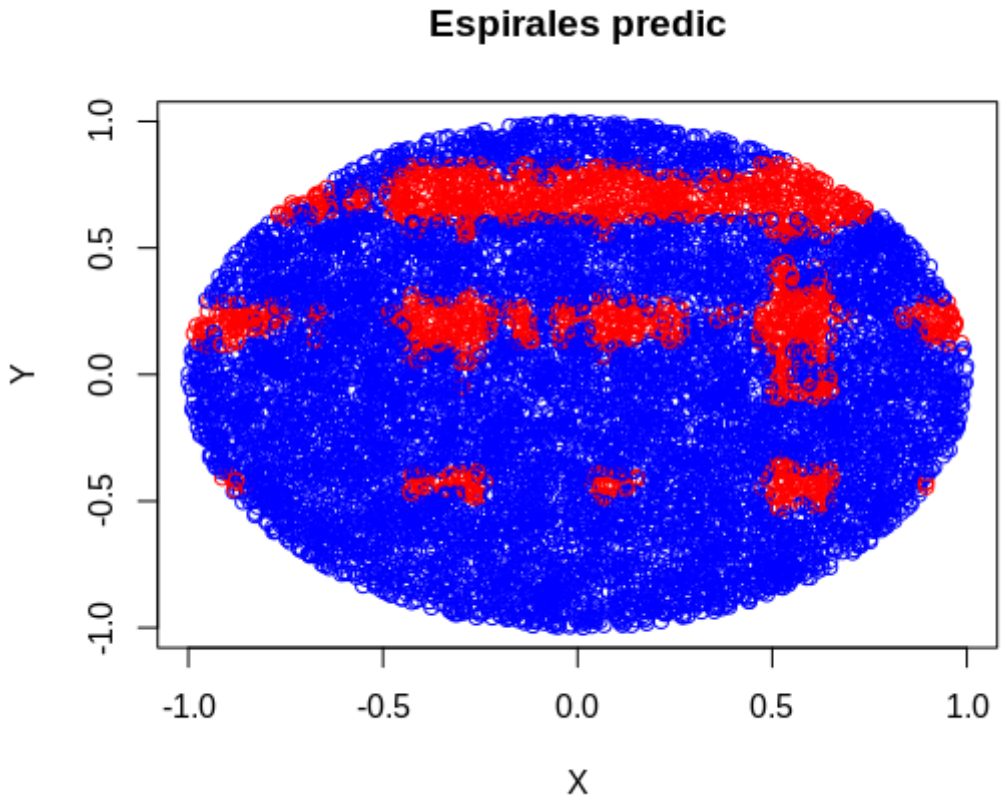

Figura 4.4

El resultado visto en la figura 4.4, el problema del espiral, está tan alejado de la clasificación real por que el clasificador está construido de forma tal que toma las probabilidades de cada componente de los valores de forma independiente, cuando en este problema las probabilidades de las componentes ‘x’ e ‘y’ están intrínsecamente relacionadas.

El overfitting que se puede observar en las figuras 4.1 y 4.2 se produce ya que al usar muchos bins estamos separando las componentes de los datos de entrenamiento en muchos “grupos”, por lo que puede que en un bin de una clase nos falten muestras y la probabilidad de la clase en este punto sea menor comparadas con la de la otra clase, afectando al cálculo final de probabilidad de esa clase.