

# Prueba y Comparación de Naive Bayes y C4.5

Juan Carlos López López

Adolfo Marín Arriaga

Luis Rodrigo Rojo Morales

3 de diciembre de 2016

## 1. Introducción

Para hacer las pruebas de estos algoritmos usamos el conjunto de datos que esta destinado a probar los clasificadores, este se puede obtener en: <https://archive.ics.uci.edu/ml/datasets/Adult> y en nuestro repositorio se encuentra en: /Entregable3/NaiveBayes/AdultDataSetTest.csv. Dicho conjunto tiene 16,281 registros, los cuales cada uno tiene los mismos atributos que el conjunto de datos original.

## 2. Naive Bayes

El script para probar este clasificador se encuentra en /Entregable3/NaiveBayes/TestAdultDataSet.java el objetivo de este script es cargar el conjunto de datos prueba, aplicar el clasificador y dar los datos de cuantos clasifico bien y cuantos mal, para así poder hacer la matriz de confusión, la cual queda de la siguiente manera:

	> 50K	≤ 50K
> 50K	3023	823
≤ 50K	2201	10234

Con esta matriz podemos observar que el clasificador predijo 13257 bien y 3024 mal, a 823 personas le dijimos que iban a ganar ≤ 50K pero en realidad ganaron > 50K, mientras que a 2201 personas les dijimos que iban a ganar > 50K pero ganaron ≤ 50K.

El clasificador en general tiene una exactitud del 81.37 %, mientras que individualmente predice a los que ganan más de 50K con una exactitud de 75.92 % y a los que ganan menos o 50K con una exactitud de 83.05 %

## 3. C4.5

El script para probar este clasificador se encuentra en /Entregable3/C4.5/TestAdultDataSet.java el objetivo de este script es cargar el conjunto de datos prueba, aplicar el clasificador y dar los datos de cuantos clasifico bien y cuantos mal, para así poder hacer la matriz de confusión, la cual queda de la siguiente manera:

	> 50K	≤ 50K
> 50K	2168	1648
≤ 50K	689	11746

Con esta matriz podemos observar que el clasificador predijo 13914 bien y 2337 mal, a 1648 personas le dijimos que iban a ganar ≤ 50K pero en realidad ganaron > 50K, mientras que a 689 personas les dijimos que iban a ganar > 50K pero ganaron ≤ 50K.

El clasificador en general tiene una exactitud del 85.46 %, mientras que individualmente predice a los que ganan más de 50K con una exactitud de 56.81 % y a los que ganan menos o 50K con una exactitud de 94.45 %

## 4. Comparación

Comparando ambas técnicas de clasificación podemos concluir que naive bayes predice mejor a los  $> 50K$  mientras que c4.5 predice mejor a  $\leq 50K$ , pero en general la técnica de clasificación árboles C4.5 tiene un poco de más precisión.

	Naive Bayes	C4.5
Clasificó Bien	13,257	13,914
Clasificó Mal	3,024	2,337
Precisión en general	81.7%	85.46%
Precisión $> 50K$	57.86%	75.92%
Precisión $\leq 50K$	74.43%	94.45%