Censo poblacional y predicción de ingresos

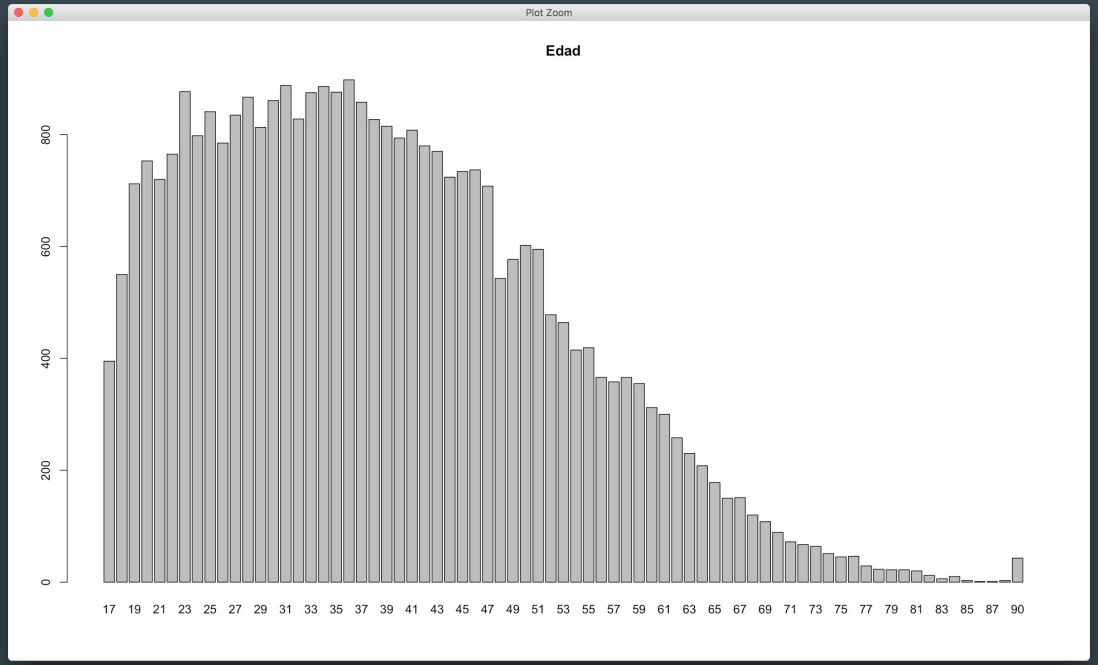
...

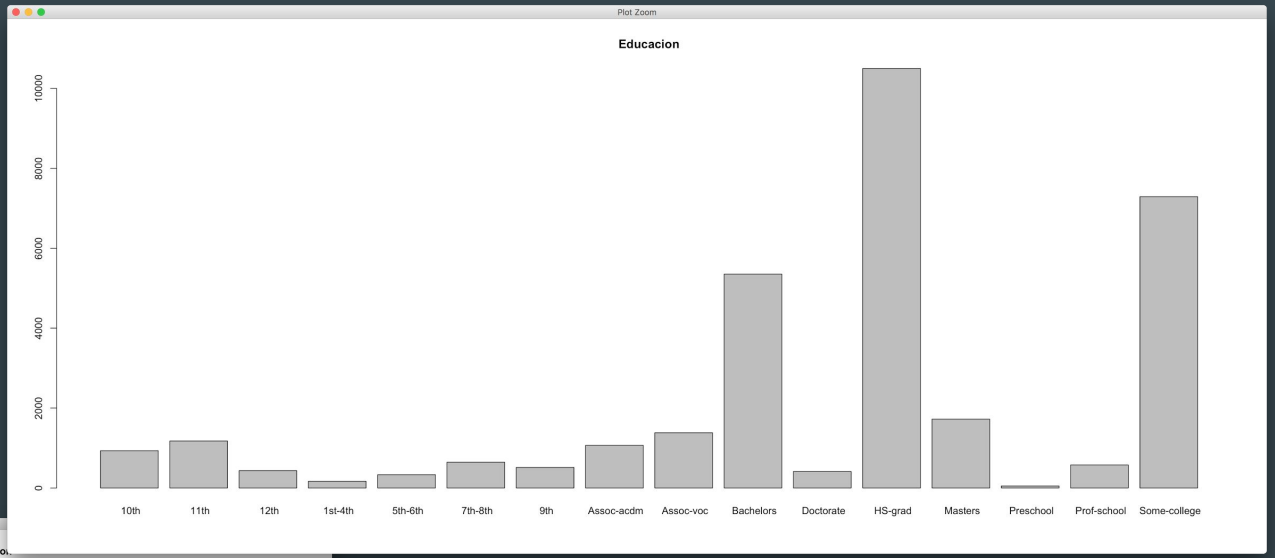
- Adolfo Marín Arriaga
- Juan carlos López López
- Luis Rodrigo Rojo Morales

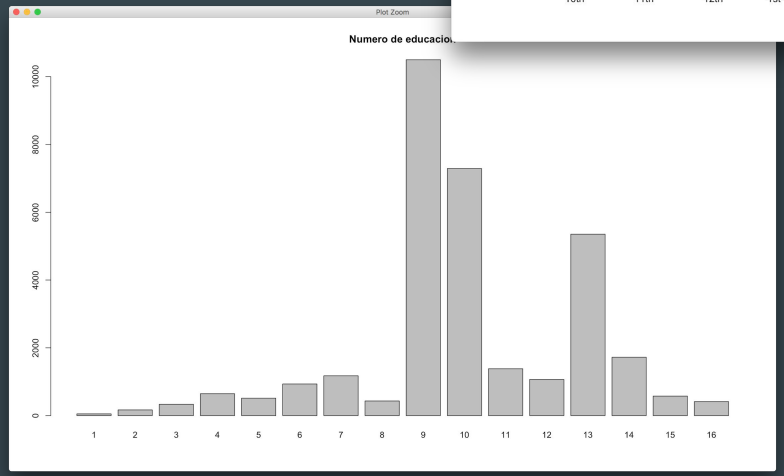
Contenido

- Preprocesamiento de los datos
- Naive Bayes en el conjunto de datos
- C4.5 en el conjunto de datos
- Pruebas de los algoritmos
- Patrones encontrados

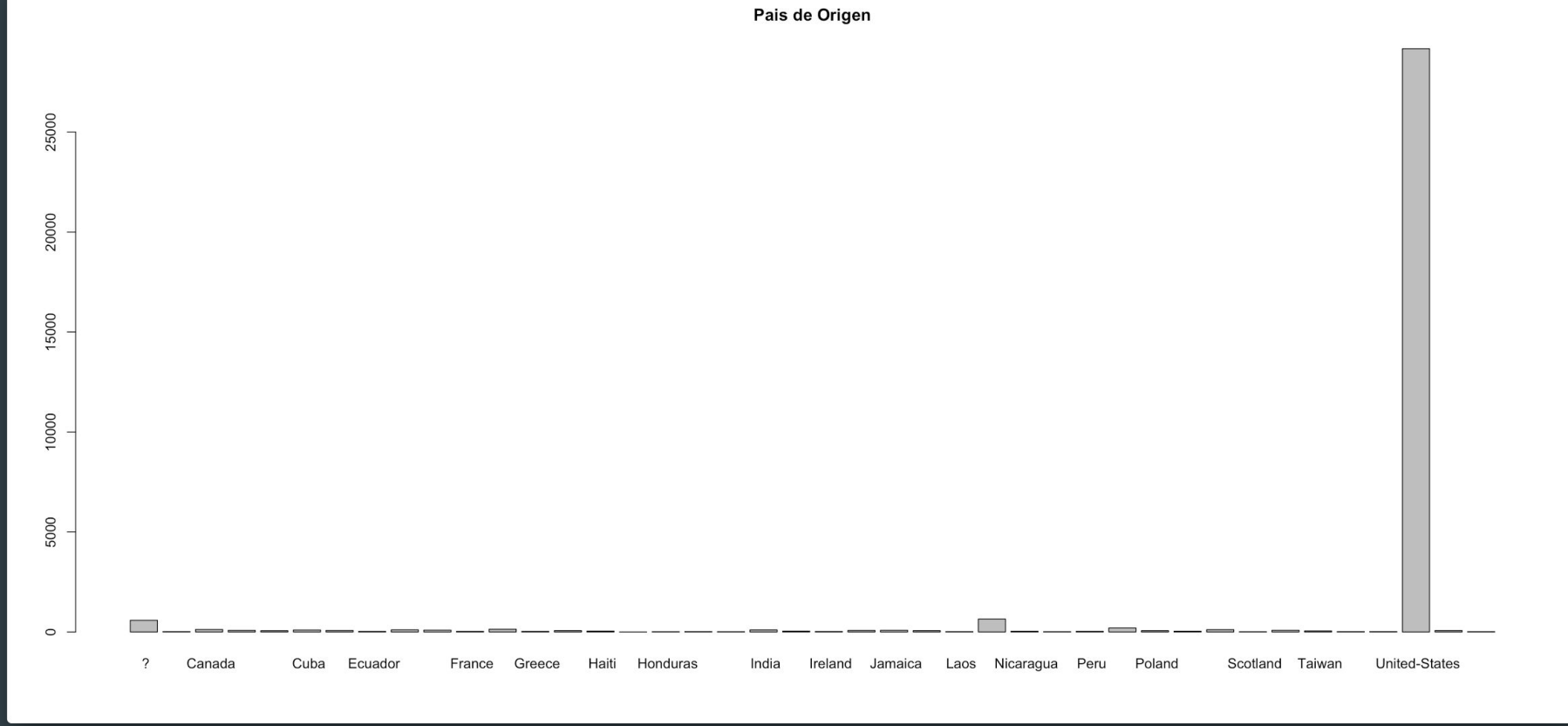
Preprocesamiento de los datos

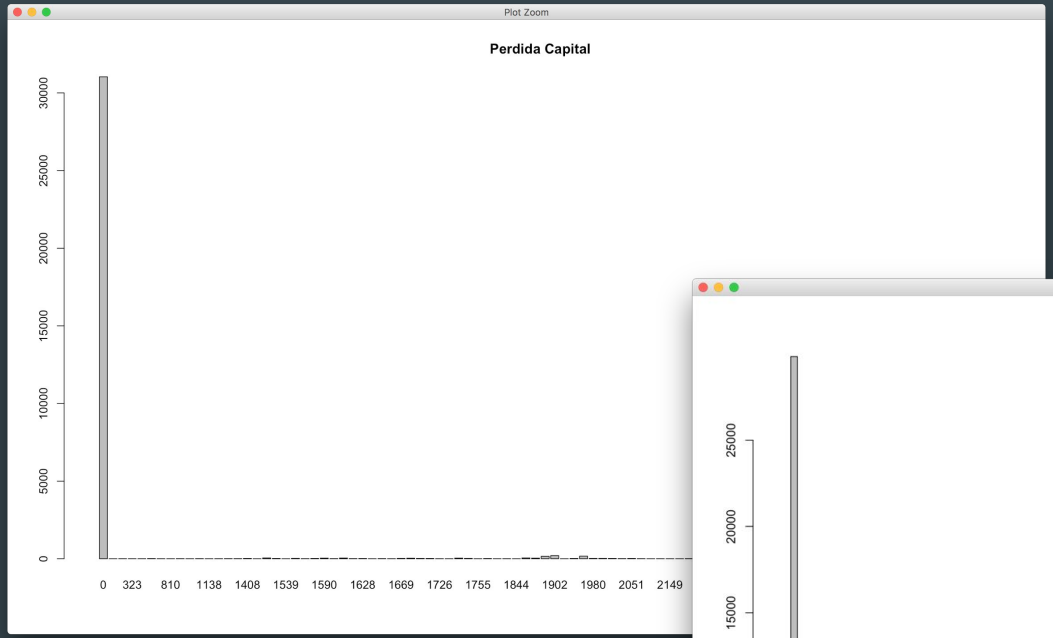


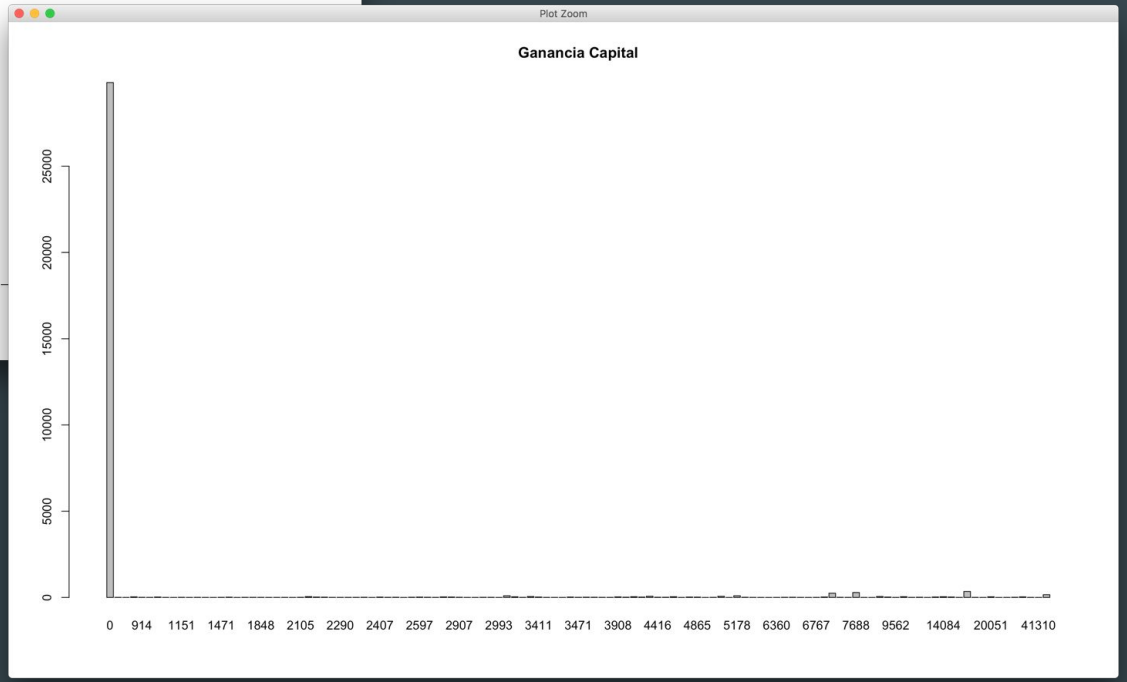

Plot Zoom
Educacion
0
2000
4000
6000
8000
10000
10th
11th
12th
1st-4th
5th-6th
7th-8th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Preschool
Prof-school
Some-college


Plot Zoom
Numero de educacio
0
2000
4000
6000
8000
10000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Plot Zoom


Pais de Origen
0
5000
10000
15000
20000
25000
?
Canada
Cuba
Ecuador
France
Greece
Haiti
Honduras
India
Ireland
Jamaica
Laos
Nicaragua
Peru
Poland
Scotland
Taiwan
United-States


Plot Zoom
Perdida Capital
30000
25000
20000
15000
10000
5000
0
0 323 810 1138 1408 1539 1590 1628 1669 1726 1755 1844 1902 1980 2051 2149


Plot Zoom
Ganancia Capital
25000
20000
15000
10000
5000
0
0 914 1151 1471 1848 2105 2290 2407 2597 2907 2993 3411 3471 3908 4416 4865 5178 6360 6767 7688 9562 14084 20051 41310

Preprocesamiento de los datos

- Eliminamos los atributos education_num y fnlwgt
- Dividimos los atributos numéricos por rangos.
- Creamos las tablas de frecuencias.
- En native_country minimizamos a dos valores “United-States” y “Foreigner”
- En Capital Gain y Loss lo mismo pero con 0 y mayor a 0
- Se creó una tabla de todos los atributos con sus frecuencias.

Naive Bayes en el conjunto de datos

- NaiveBayes.java
- NBAdultDataSet.java
- rNaiveBayes.r

Naive Bayes en el conjunto de datos

```
Clasifico 3033 mal
Clasifico 13248 bien
Hay  3846 >50K
Hay  12435 <=50K
Hay  5027 predicciones >50K
Hay  11254 predicciones <=50K
```

```
> table(ok)
ok
FALSE  TRUE
 2851 13430
> table(datest$V15)

<=50K  >50K
12435  3846
> table(pr)
pr
<=50K  >50K
13552  2729
```

Naive Bayes en el conjunto de datos

Lo que nos dicen estos resultados es que hay 12435 personas de la clase <=50k y hay 3846 de la clase >50k. El primero clasificó 11254 personas en la clase <=50k y 5027 en la clase >50k.

El segundo clasificó 13552 personas en la clase <=50k y 2729 en la clase >50k.

El primero clasificó bien a 13248 personas y mal a 3033 personas. El segundo clasificó bien a 13430 personas y mal a 2851 personas. El primer clasificador acertó en el 81.37 % de los casos.

El segundo clasificador acertó en el 82.49 % de los casos.

C4.5 en el conjunto de datos

- AlgoritmoC45.java
- AdultC45.java
- c45.r

Pruebas de los algoritmos

Para Probar Naive Bayes:

- TestAdultDataSet.java

Para Probar C4.5:

-

Pruebas de los algoritmos (Naive Bayes)

Predicción

Reales

	$> 50K$	$\leq 50K$
$> 50K$	2920	926
$\leq 50K$	2107	10328

Matriz de Confusión

Pruebas de los algoritmos (Naive Bayes)

Con esta matriz podemos observar que el clasificador predijo 13248 bien y 3033 mal, a 926 personas le dijimos que iban a ganar ≤ 50K pero en realidad ganaron > 50K, mientras que a 2107 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 %

	$> 50K$	$\leq 50K$
$> 50K$	2920	926
$\leq 50K$	2107	10328

Pruebas de los algoritmos (C4.5)

Predicción

Reales

Matriz de Confusión

Patrones Encontrados

Los atributos que hacen que una persona tenga más probabilidad de caer en la clase >50k son si tiene un capital gain mayor que cero, un grado de educación Doctorate, Maters o Prof-School, una clase de trabajo de self-emp-inc, otras que también aumentan la probabilidad de que caiga en >50k pero menos que las anteriores son: tener entre 41 y 60 años, tener un grado de educación Bachelors, trabajar entre 41 y 80 horas a la semana, tener un estado civil Married-civ-spouse o Married-AF-spouse, tener ocupación de Exec-managerial, o Prof-speciality, tener en Relationship Wife o Husband, tener una clase de trabajo Federal-gov. Si no tiene ningúna de estas o muy pocas del segundo grupo y ningu´na del primero es muy probable que caiga en la clase <=50k.

Repositorio

https://github.com/rodrigorojo/ProyectoFinalMineria