Censo poblacional y predicción de ingresos

...

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

Contenido

- Introducción y objetivo al Problema y objetivo
- Preprocesamiento de los datos
- Naive Bayes en el conjunto de datos
- C4.5 en el conjunto de datos
- Pruebas de los algoritmos
- Patrones encontrados

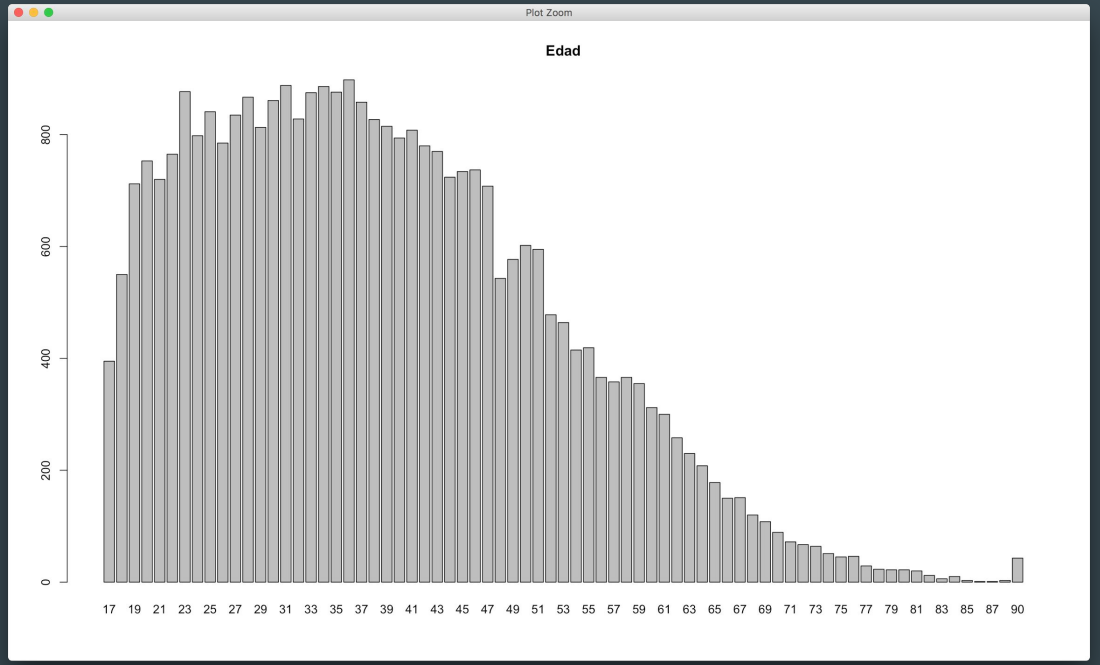
Introducción y Objetivo al problema

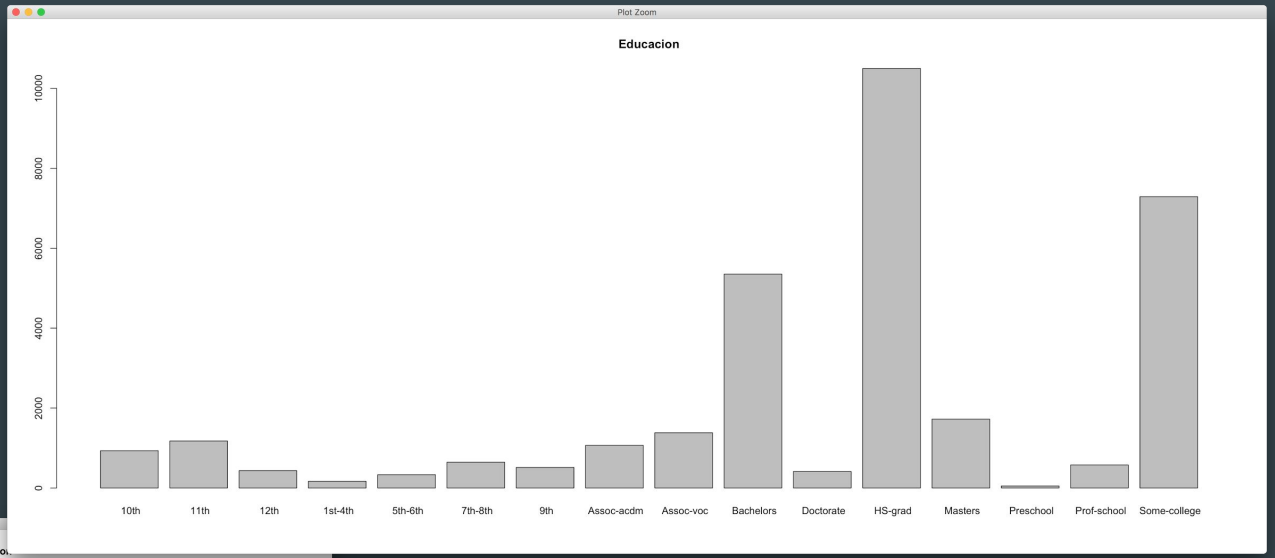
Se tiene una población de aproximadamente 32,000 individuos, tomada de un censo poblacional en el año de 1994. El objetivo es obtener un modelo de clasificación para predecir si una persona es capaz de generar más de $50K al año o no.

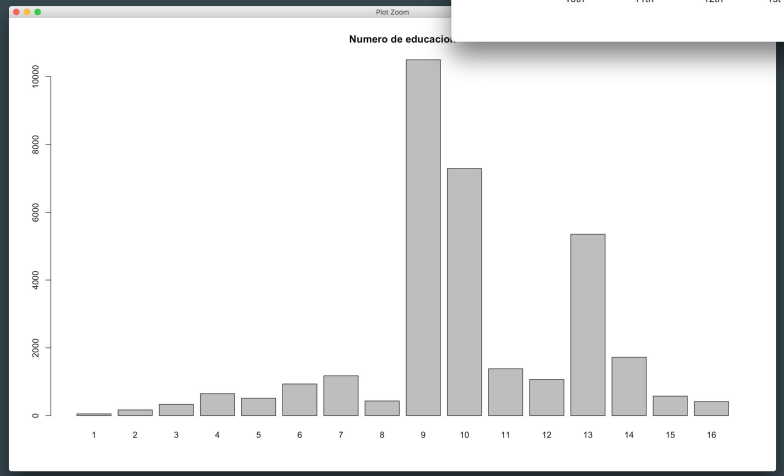
Se usaron los algoritmos:

- Naive Bayes
- C4.5

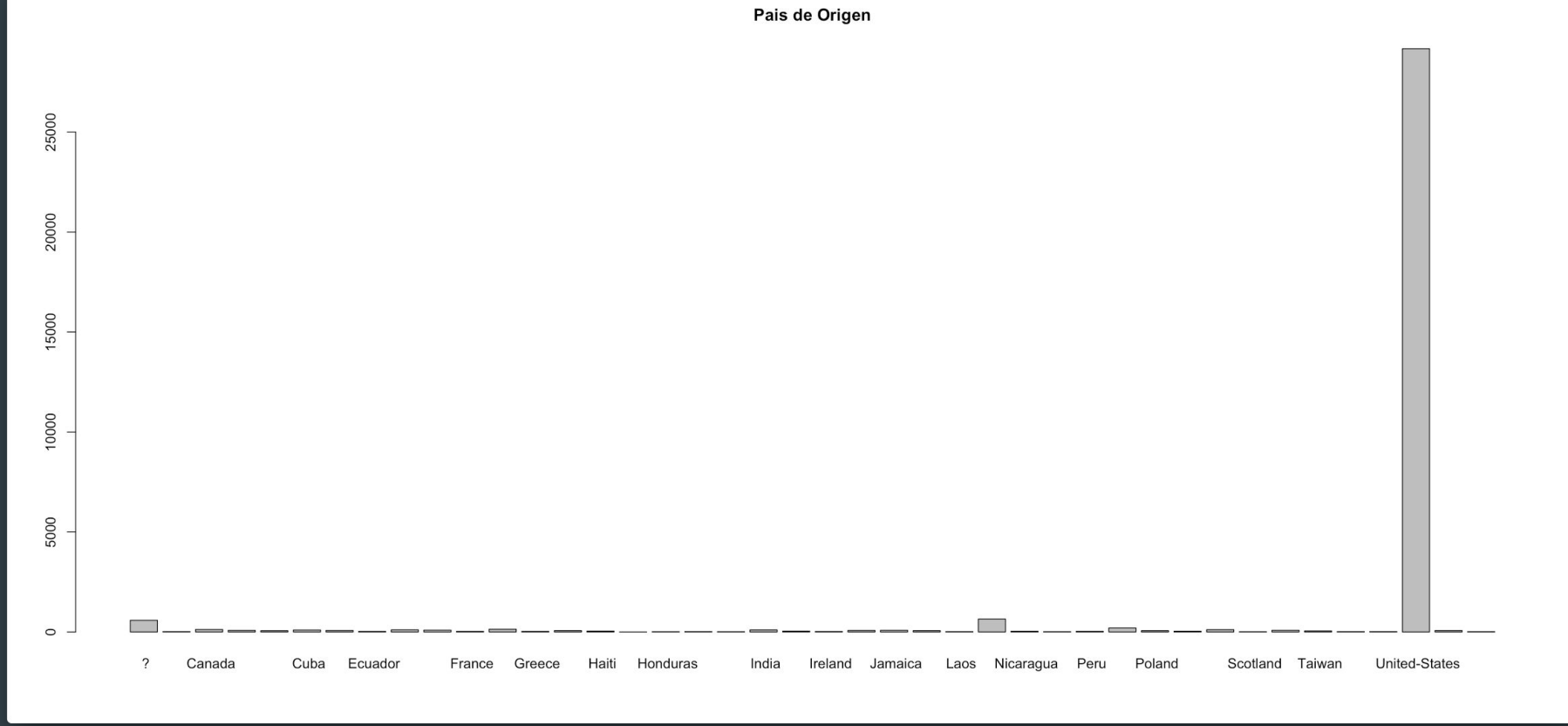
Visualización de los datos

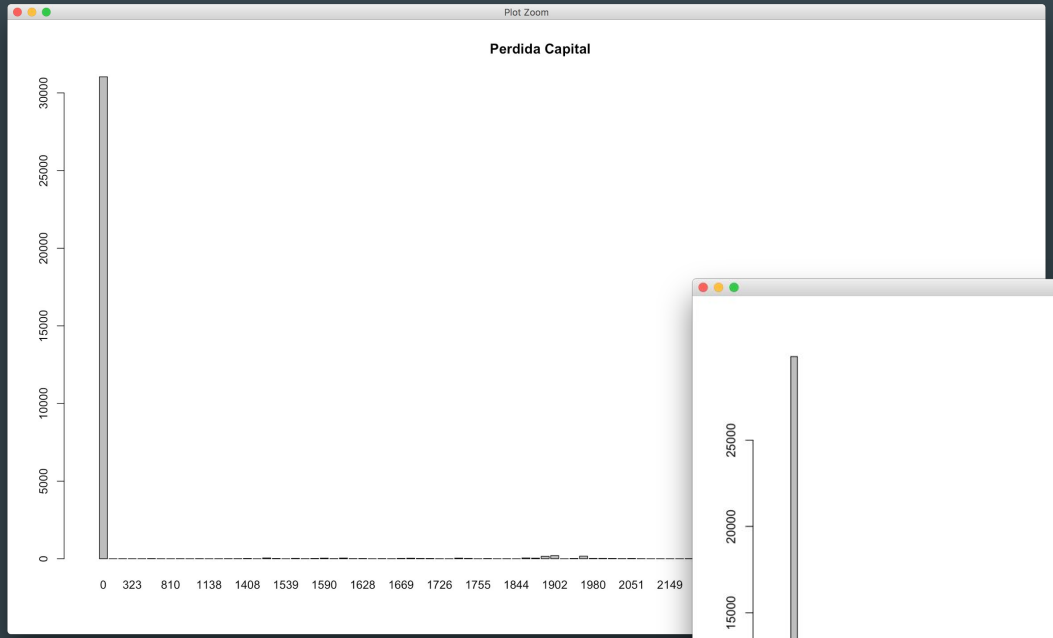


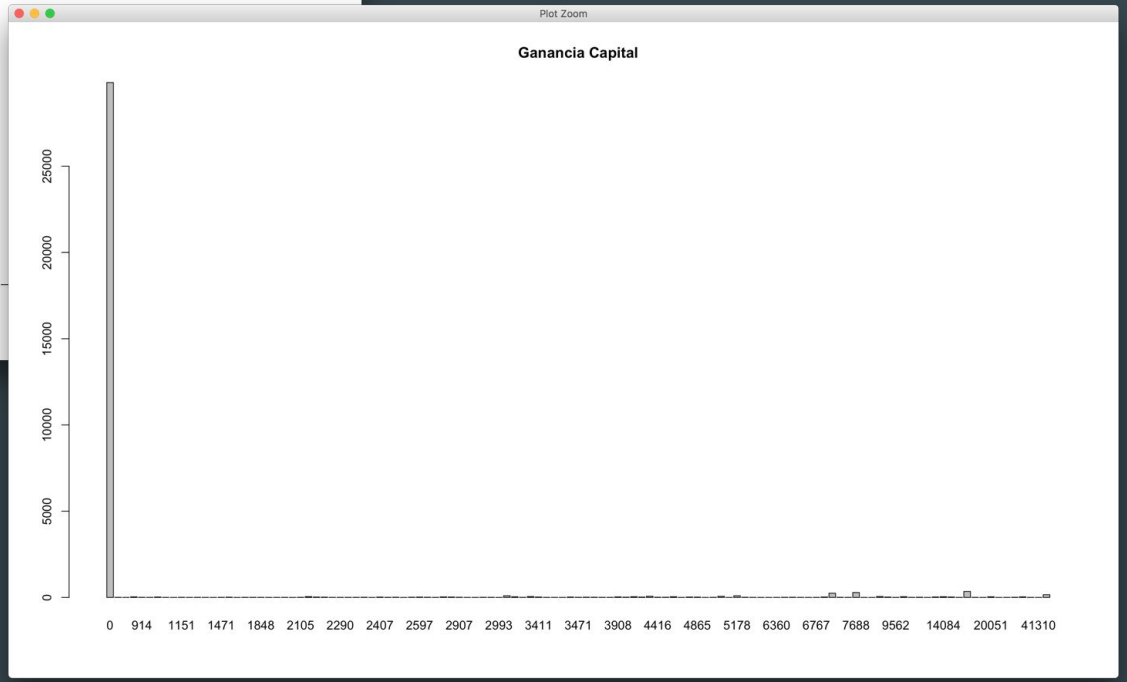

Plot Zoom
Educacion
0
2000
4000
6000
8000
10000
10th
11th
12th
1st-4th
5th-6th
7th-8th
9th
Assoc-acdm
Assoc-voc
Bachelors
Doctorate
HS-grad
Masters
Preschool
Prof-school
Some-college


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 el atributo Age por rangos de 10%
- Dividimos el atributo Hours Per Week por rangos de 5%
- 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.

Tablas de frecuencias

Edad

17-21y bigint	22-25y bigint	26-29y bigint	30-33y bigint	34-37y bigint	38-41y bigint	42-45y bigint	46-50y bigint	51-57y bigint	58-88y bigint	AGE-? bigint
3130	3281	3300	3452	3518	3245	3008	3167	3095	3322	43

Hours Per Week

1-18h bigint	19-18h bigint	25-30h bigint	30-24h bigint	35-37h bigint	38-39h bigint	40h bigint	41-44h bigint	45h bigint	46-49h bigint	50h bigint	51-59h bigint	60-64h bigint	65-89h bigint	HPW? bigint
1690	1579	1976	1487	1666	514	15217	618	1824	677	2819	1058	1519	927	139

Tablas de frecuencias (cont.)

Native Country

United-States bigint	Foreigner bigint	COUNTRY-? bigint
29170	2808	583

Capital Loss

CL-CERO bigint	CL-MAYOR-CERO bigint
31042	1519

Capital Gain

CG-CERO bigint	CG-MAYOR-CERO bigint
29849	2712

Tablas de frecuencias (cont.)

Race

White bigint	Asian-Pac-Islander bigint	Amer-Indian-Eskimo bigint	Other bigint	Black bigint
27816	1039	311	271	3124

Education

10th bigint	11th bigint	12th bigint	1st-4th bigint	5th-6th bigint	7th-8th bigint	9th bigint	Assoc-acdm bigint	Assoc-voc bigint	Bachelors bigint	Doctorate bigint	HS-grad bigint	Masters bigint	Preschool bigint	Prof-school bigint	Some-college bigint
933	1175	433	168	333	646	514	1067	1382	5355	413	10501	1723	51	576	7291

Occupation

Tech-support bigint	Craft-repair bigint	Other-service bigint	Sales bigint	Exec-managerial bigint	Prof-specialty bigint	Handlers-cleaners bigint	Machine-op-inspct bigint	Adm-clerical bigint	Farming-fishing bigint	Transport-moving bigint	Priv-house-serv bigint	Protective-serv bigint	Armed-Forces bigint	? bigint
928	4099	3295	3650	4066	4140	1370	2002	3770	994	1597	149	649	9	1843

Naive Bayes en el conjunto de datos

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

Naive Bayes en el conjunto de datos

```
Clasifico 3024 mal.
Clasifico 13257 bien.
Hay  3846 >50K.
Hay  12435 <=50K.
Hay  5224 predicciones >50K.
Hay  11057 predicciones <=50K.
Predijo 2201 como >50k pero son <=50k.
Predijo 823 como <=50k pero son >50k.
Por lo tanto hay: 3023 predicciones de >50K bien y
10234 predicciones de <=50K bien.
```

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

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

C4.5 en el conjunto de datos

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

Pruebas de los algoritmos (Naive Bayes)

Predicción

Reales

	$> 50K$	$\leq 50K$
$> 50K$	3023	823
$\leq 50K$	2201	10234

Matriz de Confusión

Pruebas de los algoritmos (Naive Bayes)

- 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 81.7 %, 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 %

	$> 50K$	$\leq 50K$
$> 50K$	3023	823
$\leq 50K$	2201	10234

Pruebas de los algoritmos (C4.5)

Predicción

Reales

	$> 50K$	$\leq 50K$
$> 50K$	2168	1648
$\leq 50K$	689	11746

Matriz de Confusión

Pruebas de los algoritmos (C4.5)

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 %

	$> 50K$	$\leq 50K$
$> 50K$	2168	1648
$\leq 50K$	689	11746

Tabla Comparativa

	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 ≤ 50K	74.43%	94.45%

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, Masters 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 ninguna de estas o muy pocas del segundo grupo y ninguna del primero es muy probable que caiga en la clase <=50k.

Repositorio

https://github.com/rodrigorojo/ProyectoFinalMineria