Aprenentatge Automàtic

Treball de curs

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

FIB

Martí Ferret Vivó
Raúl García Fuentes
11 – 01 - 2019

Índex

Introducció

En aquesta pràctica hem posat a prova els nostres coneixements adquirits en l'assignatura de APA, primer hem triat un problema real a resoldre i hem treballat les dades fent pre-processat i diferents tipus de encoding adient per poder generar models de predicció. Hem usat 3 models lineals (KNN (k veïns propers), Naive Bayes i Regressió Logística) i 2 de no lineals (SVM (màquines de vector suport) i Random Forest), com a hipòtesi els mètodes no lineals haurien d'anar millor (tenir un percentatge d'encert més alt) que els mètodes lineals.

El problema que hem triat consisteix a decidir si una persona, que viu i treballa als Estats Units, ingressa més de 50.000$ a l'any, per resoldre'l disposem de dades com l'edat, el sexe o el país d'origen entre d'altres.

Aquest problema ja ha estat treballat per altres, per tant podrem comparar els nostres resultats amb els obtinguts prèviament. A continuació es mostra l'error de test generat amb diferents algorismes i tècniques de modelatge. (només es mostren els que hem utilitzat en aquesta i alguns que hem trobat interesants).

Mètode	**Error**
Nearest-neighbor (1 veí)	21.42
Nearest-neighbor (3 veïns)	20.35
Naive-Bayes	16.12
NBTree	14.10
Voted ID3 (0.6)	15.64
Voted ID3 (0.8)	16.47
T2	16.84
1R	19.54

Descripció de les dades

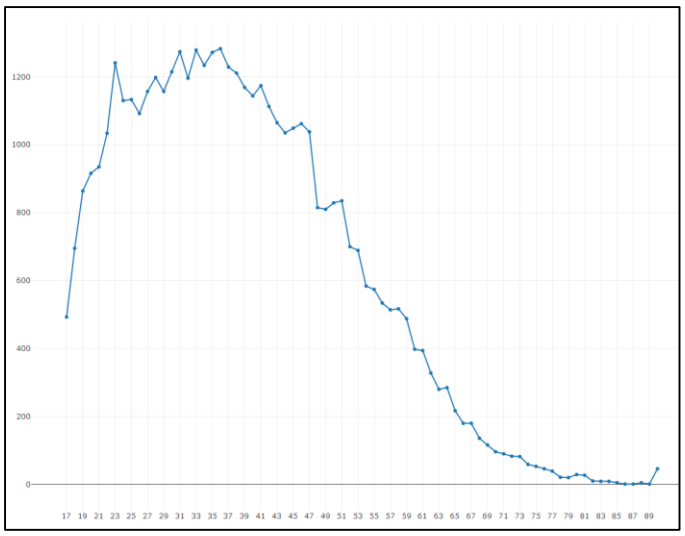
Per a realitzar la nostra pràctica hem fet servir un dataset que disposa de 14 variables, 6 d'elles contínues i les altres 8 categòriques.

A continuació es mostra una descripció estadística detallada de les dades:

Age:

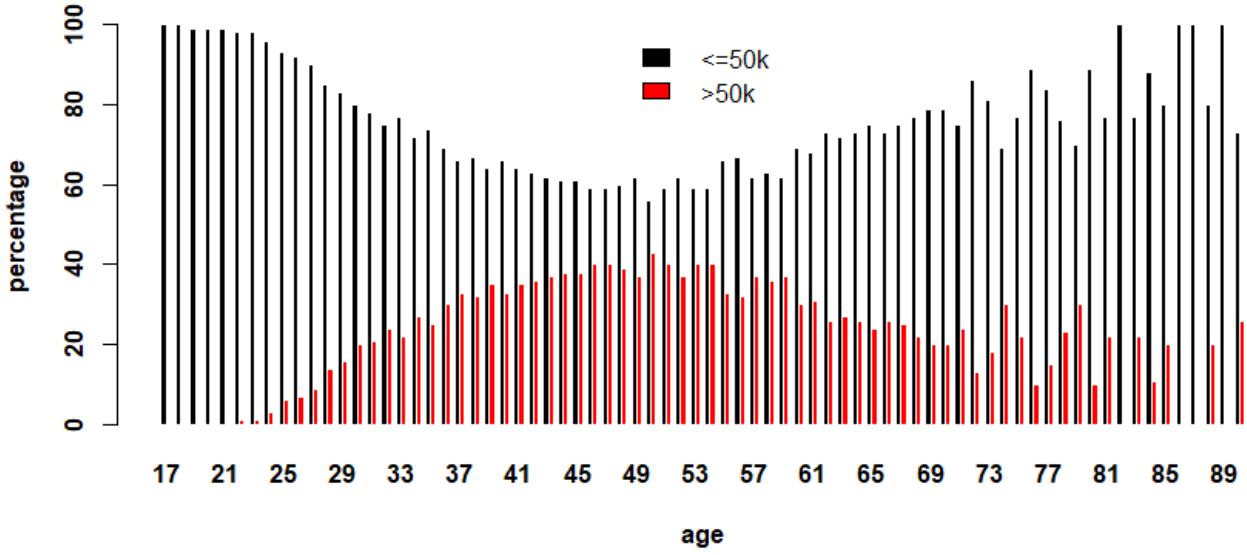
```
summary(TotalData$age)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
17.00   28.00   37.00   38.55   47.00   90.00
```

Variable continua, representa l'edat de la persona.



Podem veure que l'edat de les persones de les nostres dades pren valors entre 17 i 90 anys, tot i que, com la informació ha estat extreta principalment de la població activa, l'edat mitjana es 38 anys, y el 75% dels entrevistats tenen 47 anys o menys.

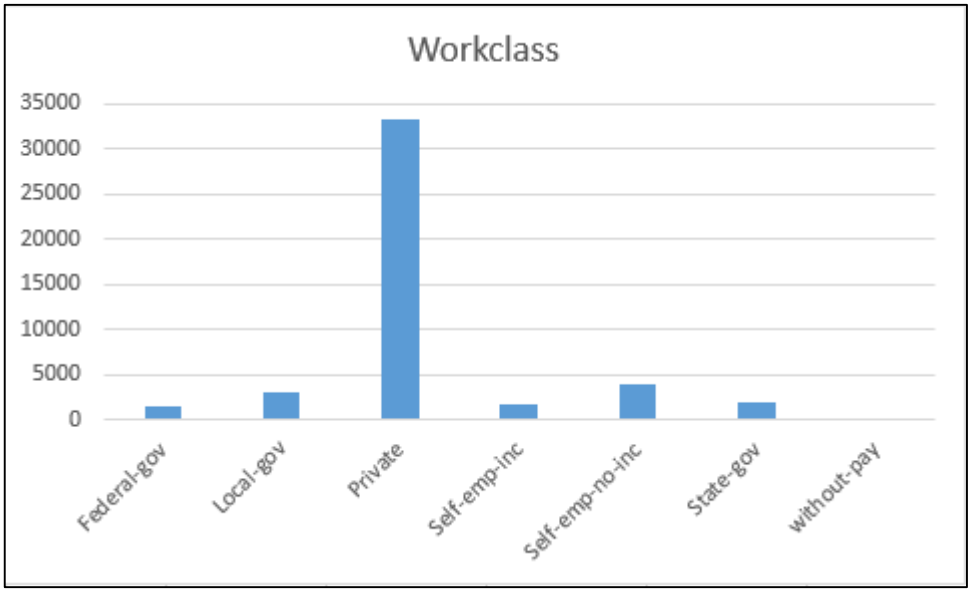
A la següent gràfica podem veure el percentatge d'entrevistats amb sou menor o major a 50.000$ anuals en funció de la seva edat.



Workclass:

Variable categòrica, representa la posició laboral de una persona (Treballador d'una empresa privada, treballador autònom, treballador de l'estat…)
Valors possibles:

Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.



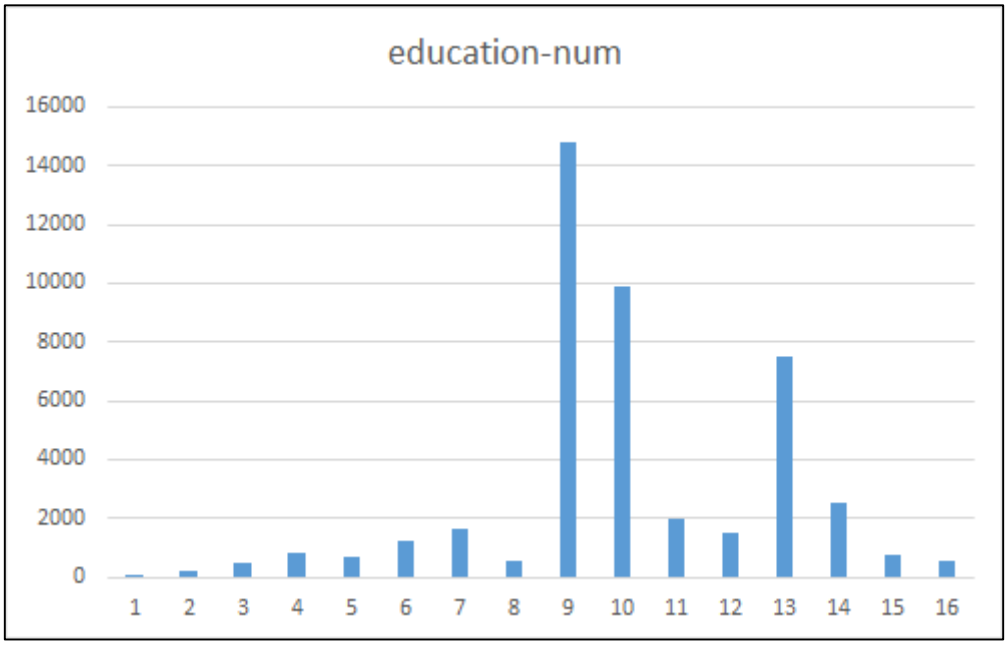
Hem decidit conservar aquesta variable tot i que no aporta gaire informació, ja que aproximadament el 75% dels entrevistats treballen a empreses privades.

Education-num:

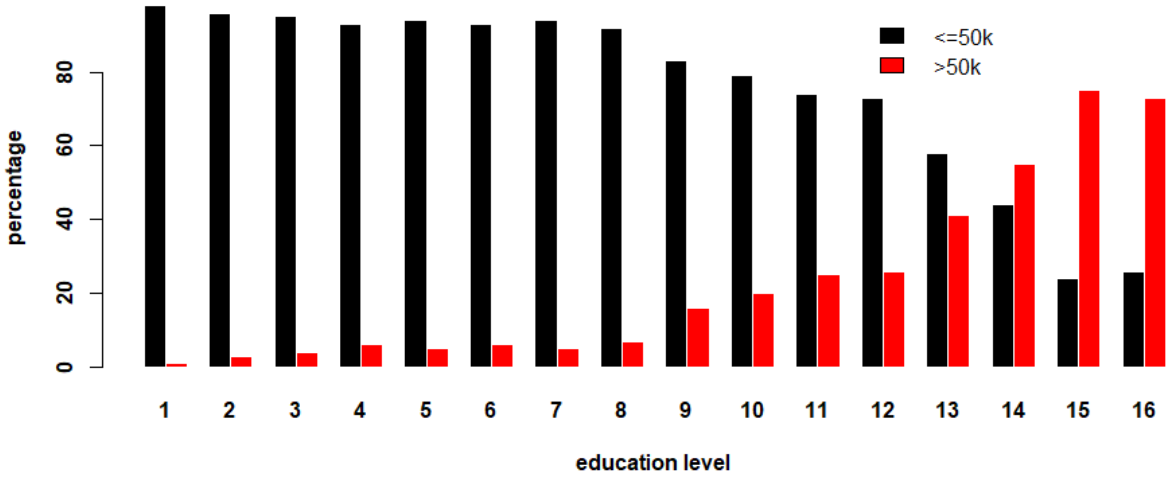
Variable continua que representa el nivell màxim d'escolarització assolit per l'entrevistat. Aquesta variable pren valors entre 1, que representa el nivell d'estudis més baix y 16, el més alt. Com a observació podem destacar que un 50% de la gent ha completat estudis superiors al nivell 10 (estudis superiors al nivell mitjà).

```
summary(edu)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00    9.00   10.00   10.12   13.00   16.00
```

Al següent gràfic podem veure la distribució dels entrevistats segons aquesta variable:



En el següent barplot veiem el percentatge de la gent que guanya més de 50.000$ l'any en funció del nivell màxim d'estudis assolit. Com era d'esperar, el percentatge de gent amb bons salaris és proporcional al nivell d'estudis.

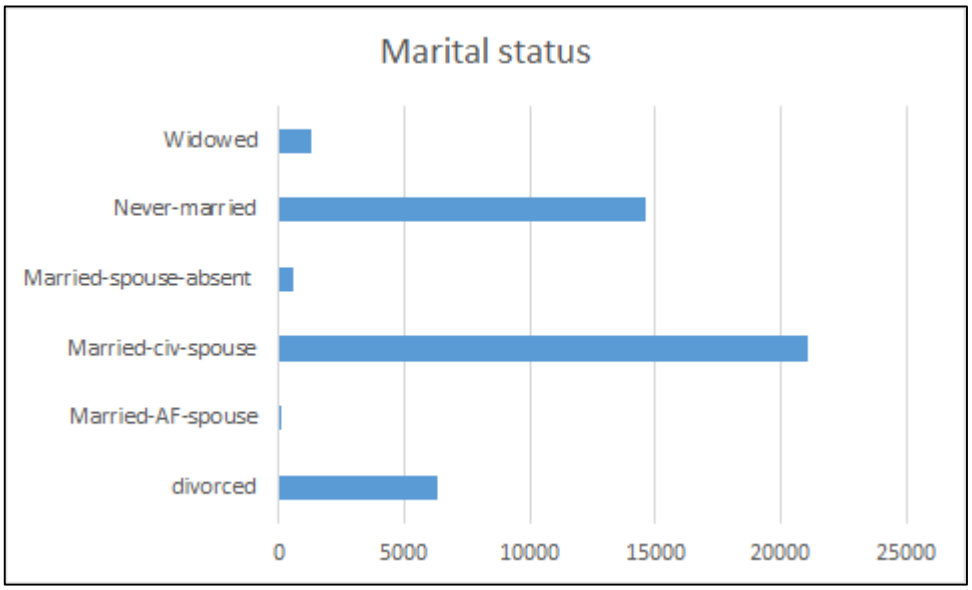


Marital-status:

Variable categòrica que representa la situació sentimental de l'entrevistat.

Valors possibles:

Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse

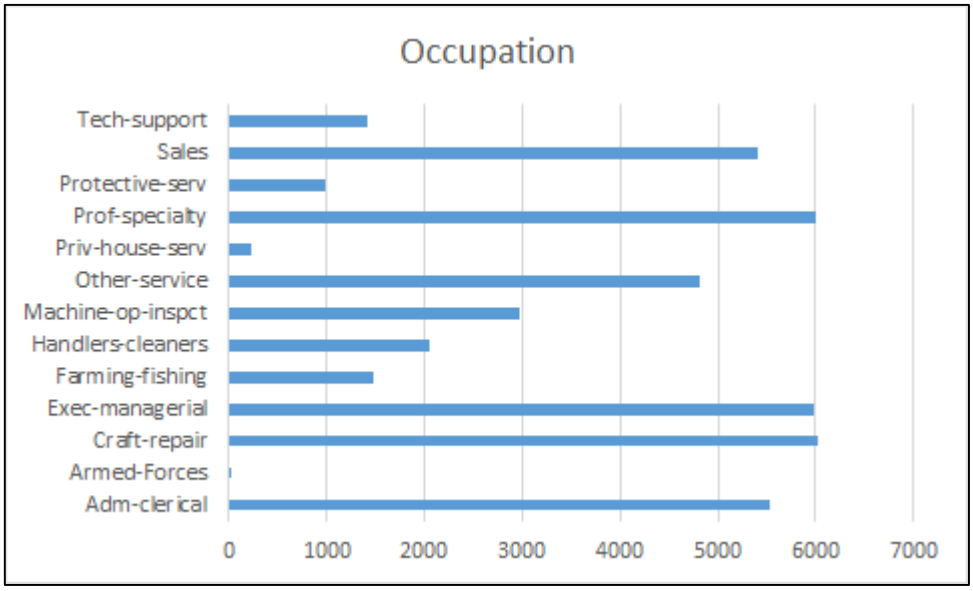


Occupation:

Variable categòrica que representa l'ofici de l'individu.

Valors que pren la variable:

Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

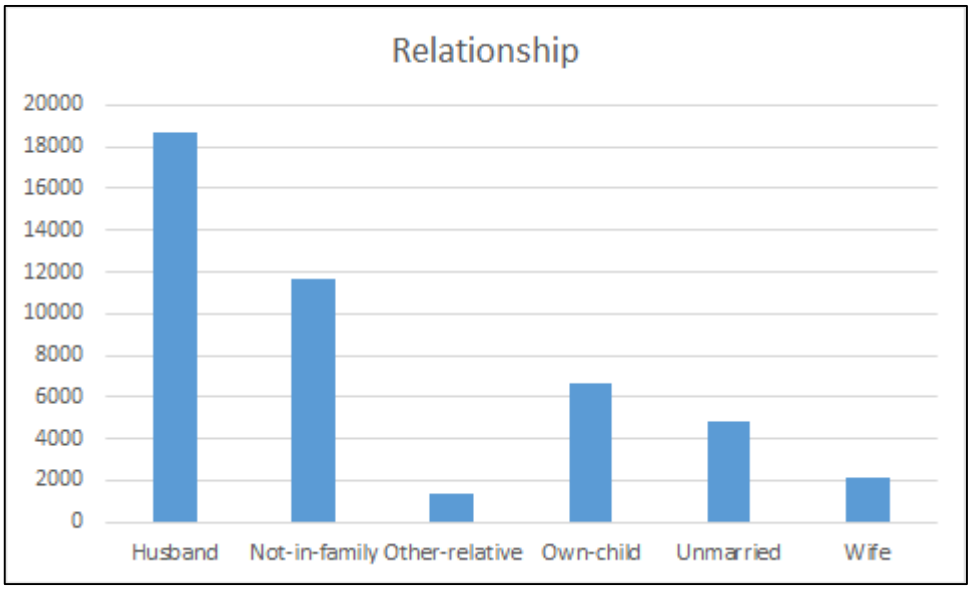


Relationship:

Variable categòrica que representa el rol familiar de l'individu.

Valors possibles:

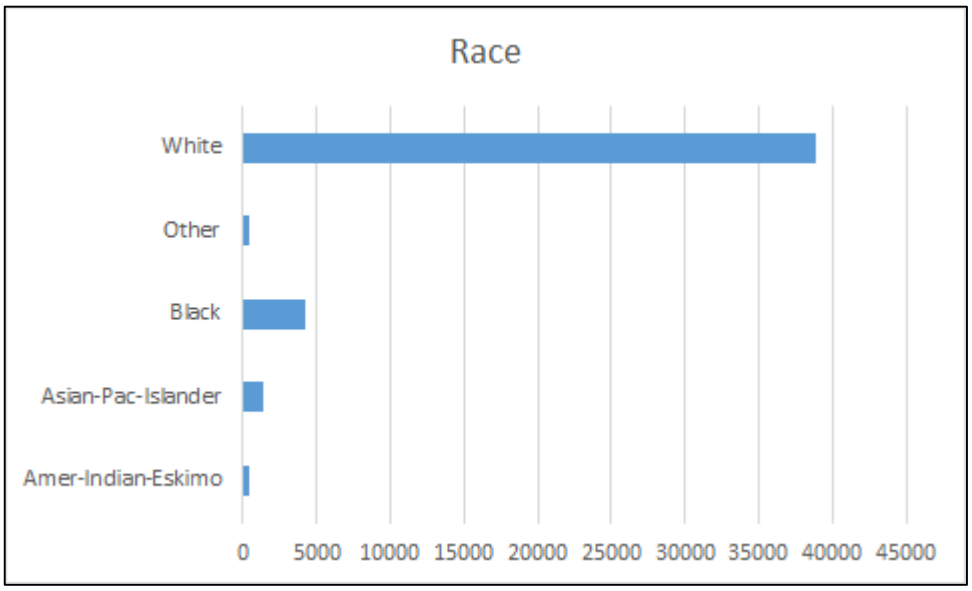
Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.



Race:

Variable categòrica que indica la raça de l'individu. Pren els valors de:

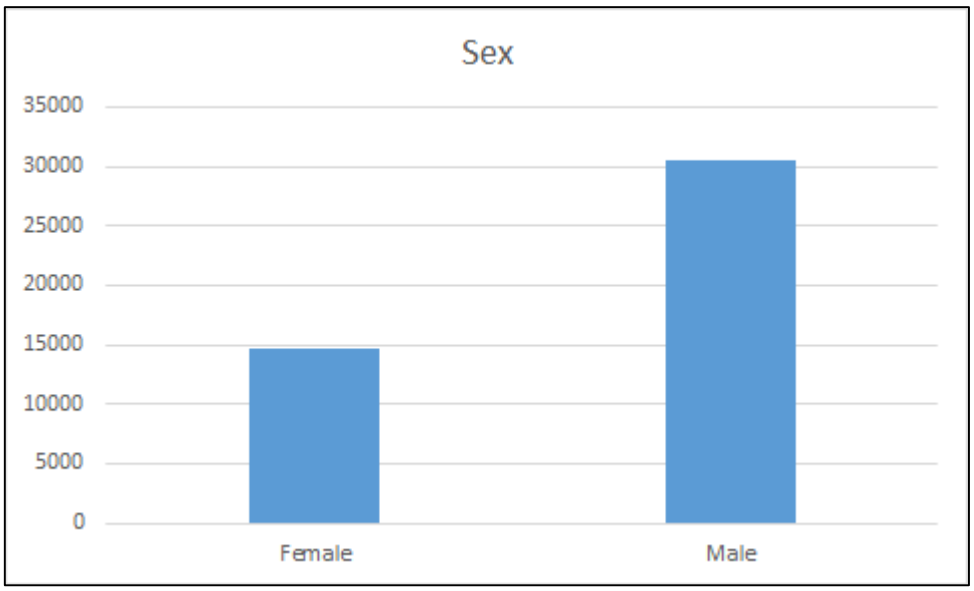
White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.



Sex:

Variable categòrica que representa el sexe de l'individu (Home / Dona):

Female, Male.

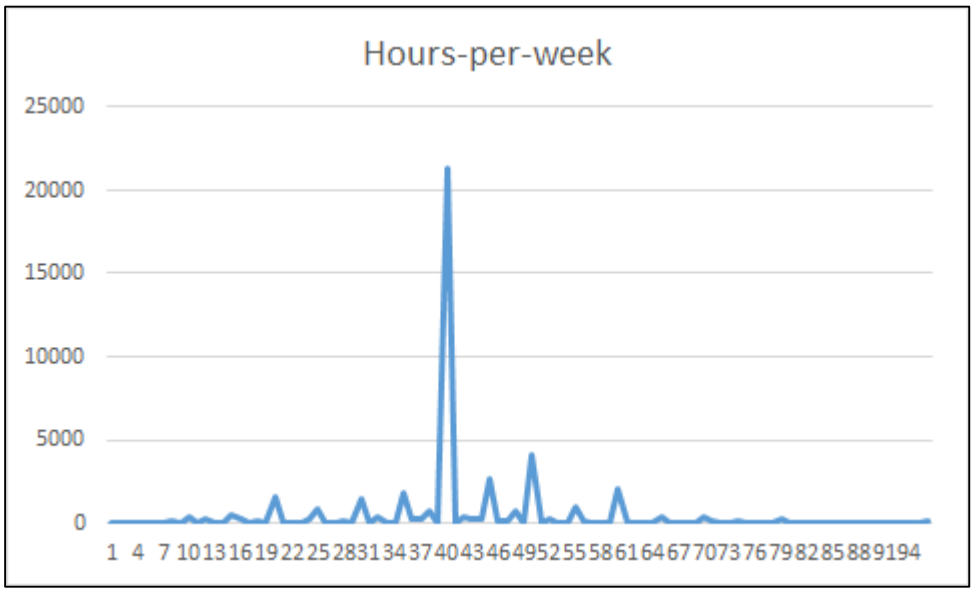


Hours-per-week:

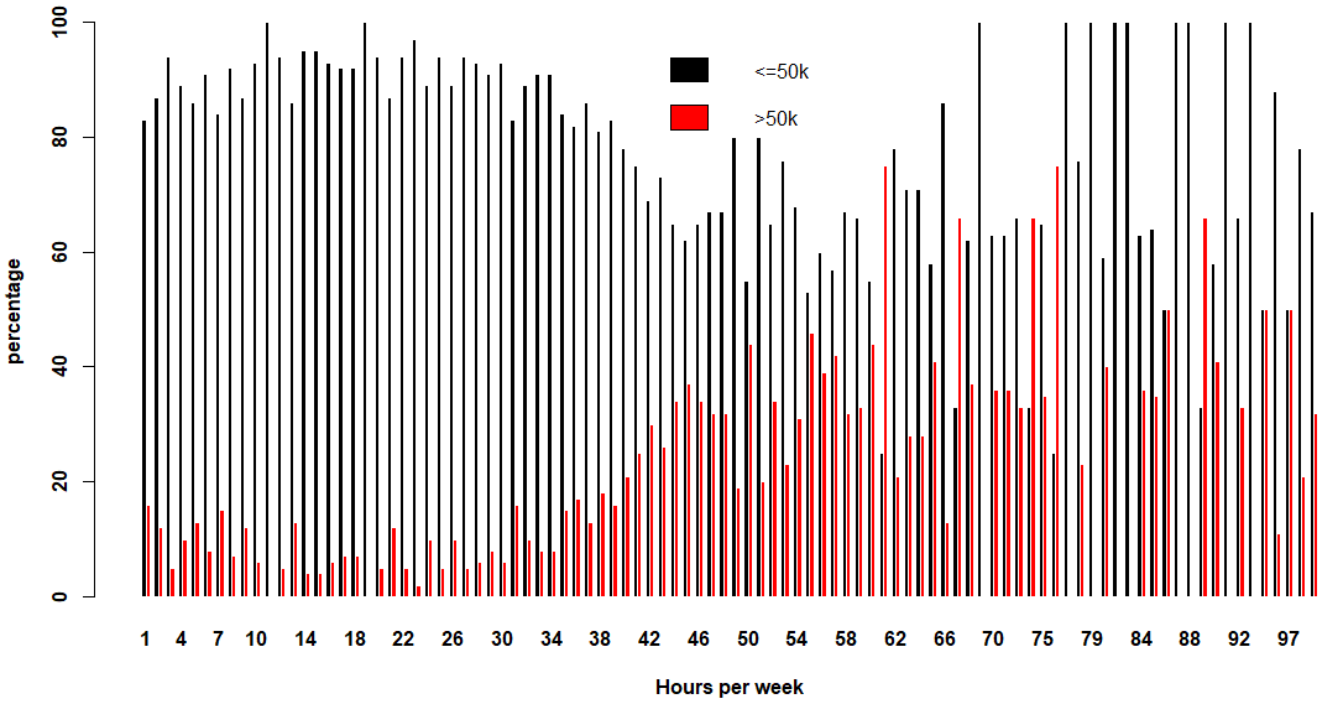
Variable continua que representa les hores que l'individu treballa a la setmana.

```
summary(TotalData$hours.per.week)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   40.00   40.00   40.94   45.00   99.00
```

La variable pren valors entre 1 i 99 hores de treball a la setmana, tot i que el 50% dels individus treballa entre 40 i 45 hores a la setmana, que correspon a una jornada laboral "estàndard".



En el següent plot podem veure que hi ha una relació directa entre l'augment del número de hores treballades setmanalment y el percentatge d'individus que cobra més de 50.000$ l'any treballant aquestes hores.

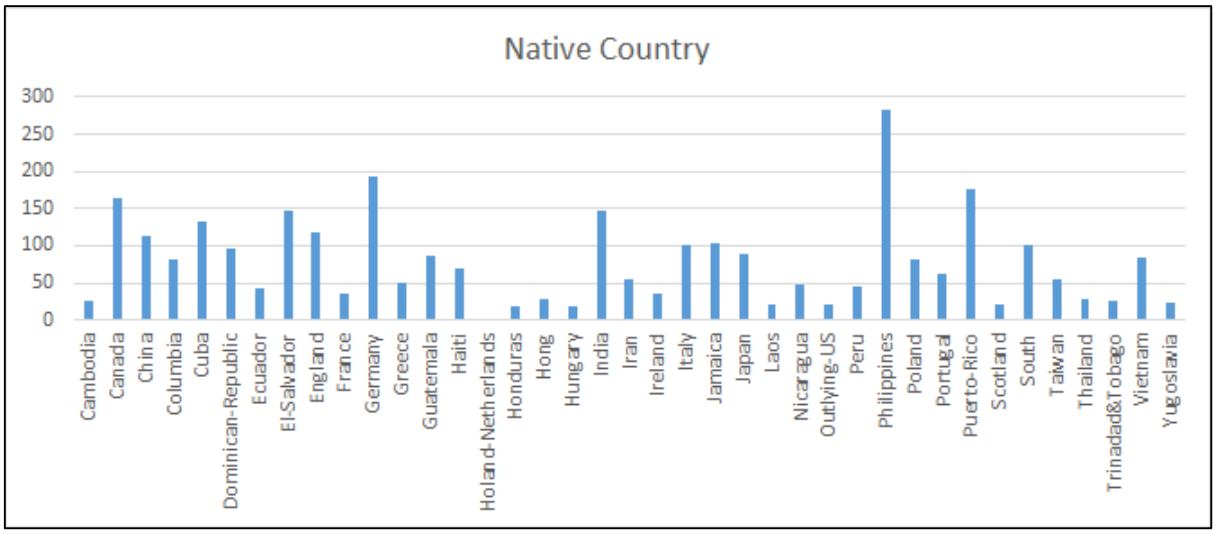


Native-country:

Variable categòrica que indica el país d'origen de l'individu. Pren els valors:

United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

El següent plot mostra la distribució dels individus segons el seu país d'origen.



**El plot anterior no conté els països "United-States" i "Mexico" ja que contenen valors molt elevats i dificultarien la visualització de les dades de la resta de països. (Mexico = 903, United-States = 41292)*

D'aquestes variables n'hem descartat 4:

- **fnlwgt** (La estimació del número de persones amb les mateixes característiques) ja que era una dada irrellevant.
- **capital-gain** i **capital-loss:** ja que en la majoria d'observacions els seus valors eren iguals a 0.
- **education**: ja que era una variable redundant.

Posteriorment hem fet un pre-processat de les dades per eliminar totes aquelles observacions que tenien valors nuls en alguna de les seves variables.

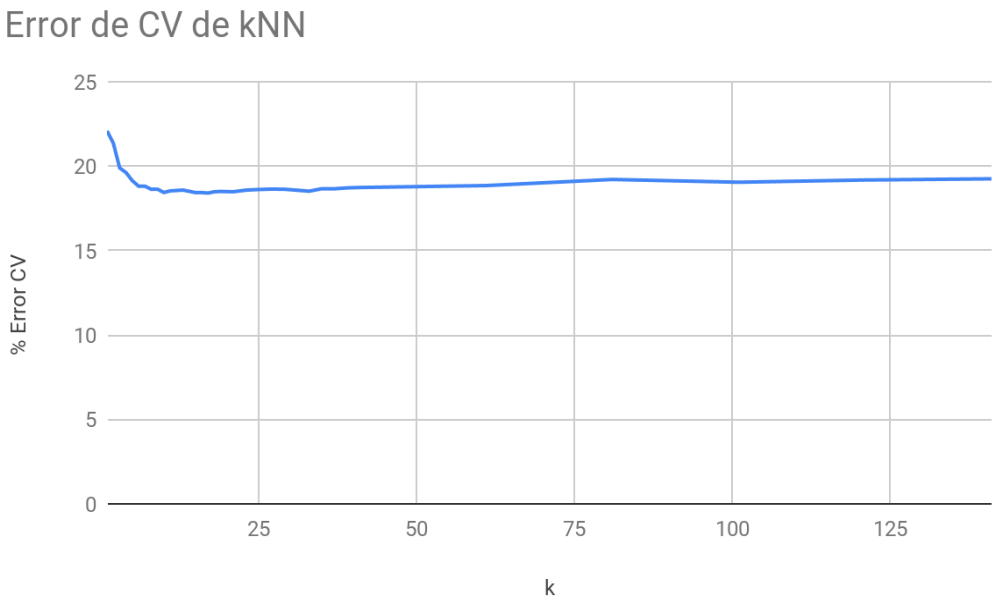
A partir d'aquí, com que tenim moltes variables que són categòriques hem aplicat **One-Hot-Encoding** per utilitzar els mètodes que no suporten aquest tipus de variables. En total surten sobre 103 variables, 3 contínues i les altres categòriques amb One-Hot Encoding. Tot i així, la en la majoria de mètodes utilitzem les dades com a categòriques (factors en R).
En el nostre cas les dades de training i les de test ja venien separades (66% training i 33% test), per tant hem utilitzat aquest mostreig. A les dades de test també hem aplicat el mateix pre-processat.

Mètodes Lineals

1. Nearest-neighbor

El primer mètode lineal que hem utilitzat és el de veïns més propers. Per fer el modelatge, hem utilitzat la rutina "knn" del paquet "class", aquest mètode, no suporta tenir variables categòriques, per tant, utilitzem les dades generades prèviament amb one-hot-encoding.

Per optimitzar el paràmetre **k** (nombre de veïns més propers a tenir en compte), hem fet diferents execucions amb validació creuada partint de k = 1 fins a k = sqrt(N), (on N es el nombre d'observacions de les dades de training). L'error de cross validation de les diferents execucions de k es mostra a la següent gràfica.



Per tant, experimentalment, tal com es pot apreciar a la gràfica anterior, veiem que el millor valor de k és 10, a partir de d'aquí l'error no varia notablement. Procedim a fer la crida definitiva amb el paràmetre de k corresponent i avaluem el predictor amb les dades de Test. La matriu de confusió obtinguda és:

		Real	
		<=50	**>50**
Predit	**<=50**	10.113	1.575
	>50	1.247	2.125

L'**error de test** obtingut pel mètode de k veïns més propers amb k = 10, és **18,74%**.

Comparat amb els resultats d'estudis anterior veiem que el nostre és millor, això segurament és degut a que en el nostre hem optimitzat el valor de k (agafem més veïns i no només 1).

2. Naive-Bayes

El segon mètode lineal utilitzat per resoldre el problema de *Adults* és el Naive-Bayes. Aquest sí que ens permet utilitzar entrades amb variables categòriques, per tant utilitzem les dades sense el one-hot-encoding.

Sabem que en aquest mètode no tenim paràmetres a optimitzar així que simplement hem creat un model amb les dades de training i l'hem avaluat amb les dades de test. Els resultats obtinguts es mostren a la matriu de confusió següent.

	Real		
Predit		**<=50**	**>50**
	<=50	9.434	946
	>50	1.926	2.754

L'**error de test** obtingut amb aquest mètode és **19.07%**, veiem però que hem obtingut resultats pitjors que en els estudis realitzats prèviament amb el mateix mètode. Això podria ser degut a les rutines utilitzades o a les variables seleccionades per nosaltres a l'hora de generar els models.

3. Logistic Regression

Finalment hem aplicat la típica regressió logística com a últim mètode lineal. Hem utilitzat les dades com a categòriques (ja que la rutina ho suporta), sabem que el mètode retorna una probabilitat de pertànyer a la classe 0 o 1 (sent 0 com cobra menys de 50k a l'any i 1 cobra més), per tant hem decidit a posar el llindar P a 0,5 (donant com a classe 1, >50k, tot aquell que tingui una P igual o superior a 0.5).

Avaluant el model amb les dades de test hem obtingut la següent matriu de confusió:

	Real		
Predit		**<=50**	**>50**
	<=50	10.602	758
	>50	1.820	1880

L'**error de test** obtingut amb el mètode de regressió logística és **17.12%**, veiem doncs que dels mètodes lineals utilitzat en la nostra pràctica aquest és el que ens retorna menys error.

Mètodes no Lineals

Un cop hem experimentat amb els mètodes lineals, procedim a resoldre el problema usant mètodes no lineals. N'hem triat 2: màquines de suport vectorial i random forest. A continuació es mostren els resultats obtinguts.

1. Support Vector Machines

El primer mètode no lineal treballat per resoldre el problema *Adults* és el de Support Vector Machines, hem usat la funció svm de la llibreria e1071 de R. Per tal de que sigui no lineal hem utilitzat el famós kernel RFB.

Com a paràmetre, *epsilon* l'hem fixat a 0.01. Els resultats d'error obtingut es mostren a continuació:

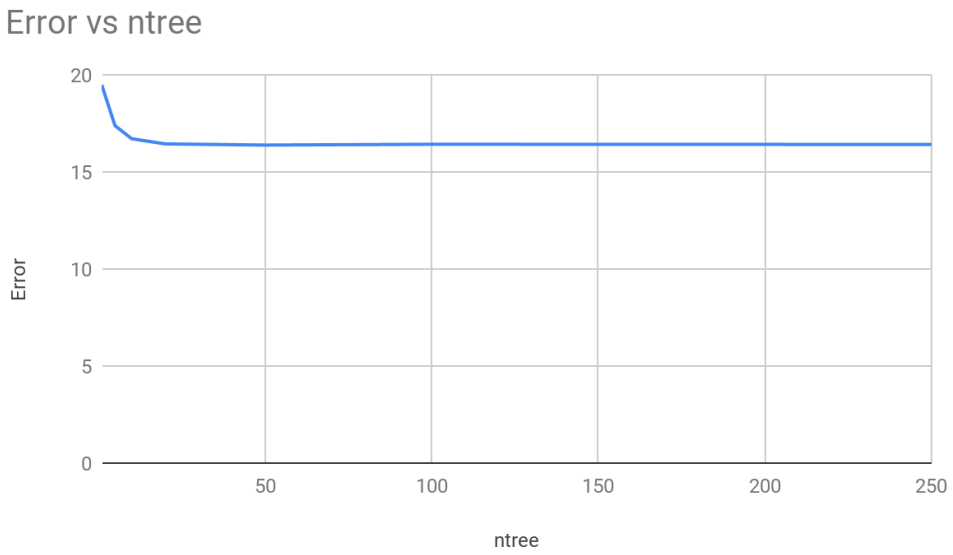
		Real	
		<=50	**>50**
Predit	**<=50**	10572	1760
	>50	788	1940

L'**error de test** obtingut amb aquest mètode és **16.9%**, no el podem comparar amb altres estudis utilitzant el mateix mètode però comparat amb els mètodes lineals usats anteriorment podem apreciar una millora notable.

2. Random Forest

El segon mètode no lineal que hem utilitzat és el Random Forest. La funció triada per generar el model és la que ve al paquet randomForest de R, en aquesta també li podem entrar les dades com a categòriques i així ho hem fet.

Per optimitzar el nombre d'arbres (paràmetre *ntree* de la funció) hem fet diferents execucions amb diferents valors per veure com es comportava l'error de predicció. Hem obtingut la següent gràfica.



Doncs experimentalment hem vist com no aconseguim reduir l'error a partir de 20 arbres (ntree = 20), per tant hem considerat que aquest valor és l'òptim en relació error i temps de computació. Per a que l'algorisme no tingui en compte la proximitat hem posat el paràmetre *proximity* a *FALSE*. Els resultats finals es mostren a continuació.

	Real		
Predit		**<=50**	**>50**
	<=50	10445	915
	>50	1599	2101

L'**error de test** obtingut pel mètode Random Forest és **16.6%**. No podem comparar aquests resultats amb el mateix mètode d'estudis anteriors ja que no disposem d'aquest, comparat amb els altres mètodes en general sembla que el random forest és dels millors.

Model final i conclusions

Segons els resultats obtinguts prèviament el millor model (el que ha tingut menys error en les dades de test), és el de **Random Forest**, amb un error del **16.6%**, com era d'esperar els mètodes no lineals han tret millors resultats que els mètodes lineals.

Per altra banda, si comparem el nostre millor model amb els resultats obtinguts per altres estudis veiem que el nostre no és dels millors.

Amb aquesta pràctica hem pogut realitzar un estudi estadístic a partir d'unes dades que ens van semblar interessants. Vam fer servir un dataset que classificava una sèrie d'individus segons si els seus ingressos anuals superaven o no els 50.000 dòlars. D'aquests individus disposàvem d'informació com el sexe, l'edat, l'educació obtinguda, el país on van néixer...

Vam decidir utilitzar aquest dataset perquè entre altres coses, tenia gairebé 50.000 instàncies amb aproximadament un 7% de valors nuls i a més, al ser un dataset conegut, ja havia sigut estudiat per altres estadístics i disposàvem del seus resultats. Vam fer servir aquests resultats de manera orientativa, ens servien per saber si ens apropàvem o no als valors correctes segons el mètode que fèiem servir.

Un cop aplicats tots els mètodes es pot apreciar que l'error ve causat principalment pels casos de classificació de ">50k", ja que al dataset representen menys d'un 30% i per tant, per als nostres models és més probable el cas "<=50k" per a uns paràmetres similars.

Una de les principals limitacions durant la realització del projecte va ser el llenguatge de programació, ja que no hem fet servir gaire *RStudio* a la universitat. Això ens va suposar una gran pèrdua de temps a l'hora de manipular, modificar, o representar les dades degut a que aquest llenguatge té algunes particularitats de tipus de dades que ens suposaven problemes al executar les rutines.

Referències

El dataset emprat en la pràctica va ser obtingut de http://archive.ics.uci.edu/ml/

A més, per la codificació en R, ens va ser útil la documentació trobada a les webs:

https://machinelearningmastery.com
https://www.datamentor.io
https://stackoverflow.com

I sobretot haver assistit als laboratoris de APA i disposar dels arxius treballats a classe ens ha sigut de gran ajuda.