Utilizzo dell'algoritmo SVM sul dataset Adult

Corso di Modellizzazione Statistica, prof. M. Bilancia

Michele Di Nanni, mat. 729187

Il dataset *Adult*

Il dataset raccoglie differenti informazioni relative a determinati individui al fine di determinare se una persona guadagna una cifra superiore a 50000 dollari annuali o meno: il task di classificazione è quindi, quello di predire se un certo individuo, identificato tramite gli attributi, guadagnerà una cifra superiore o inferiore. Procediamo con effettuare l'analisi esplorativa del dataset(EDA), prima di lavorare con l'algoritmo SVM per la predizione.

1. Analisi esplorativa del dataset (EDA)

Descrizione del dataset

Il dataset è formato da 48842 osservazioni e da 14 attributi. Vediamo nel dettaglio questi ultimi:

1. *age*: variabile numerica indicante l'età di ogni individuo
2. *workclass*: variabile categorica indicante la categoria lavorativa del singolo individuo(ad esempio lavoratore autonomo, disoccupato, ecc. . .)
3. *fnlwgt*: variabile numerica indicante il peso di quanta parte della popolazione rappresenta quell'individuo
4. *education*: variabile categorica indicante il titolo di studio più alto ottenuto dall'individuo
5. *educational-num*: variabile numerica indicante il grado di istruzione
6. *marital-status*: variabile categorica indicante lo stato civile (celibe, divorziato, ecc. . .)
7. *occupation*: variabile categorica indicante la posizione attuale lavorativa dell'individuo
8. *relationship*: variabile categorica indicante la relazione dell'individuo nel nucleo familiare (moglie, marito, figlio/a, ecc..)
9. *race*: variabile categorica indicante l'etnia di ogni individuo
10. *sex*: variabile categorica indicante il sesso
11. *capital gain*: variabile numerica indicante il capital gain
12. *capital loss*: variabile numerica indicante il capital loss
13. *hours-per-week*: variabile numerica indicante le ore lavorative settimanali di ogni persona
14. *native-country*: variabile categorica indicante la provenienza originaria dell' individuo
15. *income*: variabile categorica indicante l'incasso (se $>$ 50K o se $<=$ 50K) [**v. di output**]

1.1. Analisi preliminare

Procediamo ad analizzare la presenza di valori mancanti ed al loro trattamento. I valori mancanti sono identificati dalla presenza del simbolo ? nel dataset.

```
##      workclass     occupation native_country
##              2              7             14
```

I dati mancanti si trovano nelle colonne *workclass*, *occupation*, *native_country*. Rimpiazziamo tali dati, in quanto si tratta di dati categorici, rispettivamente con la moda.

Procediamo con l'effettuare il mapping della variabile di output(*income*) che assumerà valore "No" nel caso in cui $\leq 50K$, "Yes" nel caso in cui $> 50K$:

```
     n    missing    distinct
 48842          0           2

Value         No   Yes
Frequency  37155 11687
Proportion 0.761 0.239
```

La proporzione di istanze di classe "Yes" è pari a circa il 24%, mentre la proporzione di istanze di classe "No" è pari a circa il 76%.

A causa della complessità computazionale molto alta nel caso delle *support vector machines*, abbiamo pensato di fare un *subsampling* del dataset, utilizzando una percentuale del 10% del dataset originale, mantenendo la stessa proporzione delle istanze appartenenti alle *labels*. Il subsampling è effettuato senza rimpiazzamento.

```
    n    missing    distinct
 4884          0           2

Value         No   Yes
Frequency   3705  1179
Proportion 0.759 0.241
```

1.2. Analisi esplorativa delle variabili numeriche

Visualizziamo, a questo punto, le statistiche di base delle variabili numeriche:

Nota: La variabile **education_num** indica il grado di istruzione di ogni singolo individuo. Poichè questo aspetto lo ritroviamo già all'interno della variabile *education*, decidiamo di eliminare tale variabile dalla nostra analisi. Inoltre, la variabile **Final weight(fnlwgt)** è poco esplicativa per la nostra analisi, in quanto ci stima il peso finale di quanta parte della popolazione rappresenta. Procediamo quindi a rimuoverla.

4 Variables 4884 Observations

adult.age

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4884	0	69	0.999	38.38	15.38	20	22	27	37	47	58	63

```
lowest : 17 18 19 20 21, highest: 81 82 84 88 90
```

adult.capital_gain

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4884	0	80	0.223	1049	2029	0	0	0	0	0	0	4650

```
lowest :     0   114   594   991  1055, highest: 20051 25236 27828 34095 99999
```

adult.capital_loss

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4884	0	59	0.122	79.28	152.5	0	0	0	0	0	0	0

```
lowest :    0  653  810  880 1138, highest: 2559 2603 2824 3004 3175
```

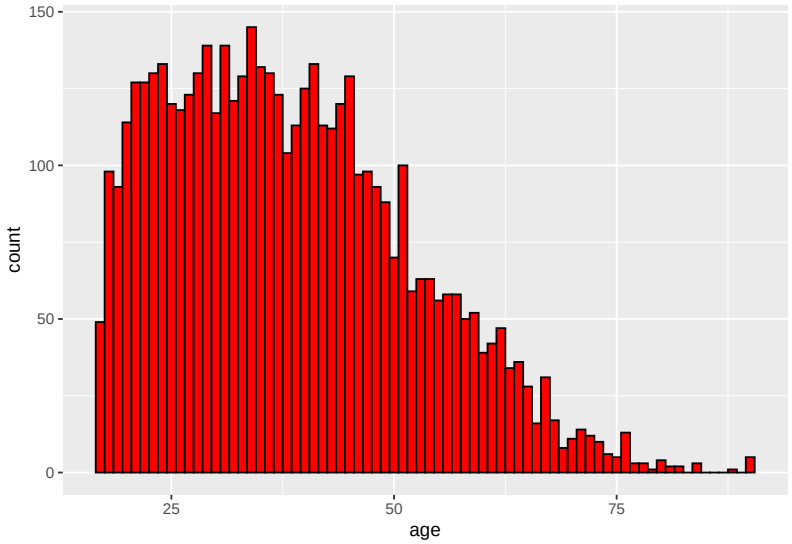
adult.hours_per_week

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4884	0	78	0.892	40.45	12.16	18	24	40	40	45	55	60

```
lowest :  1  2  3  4  5, highest: 84 90 96 97 99
```

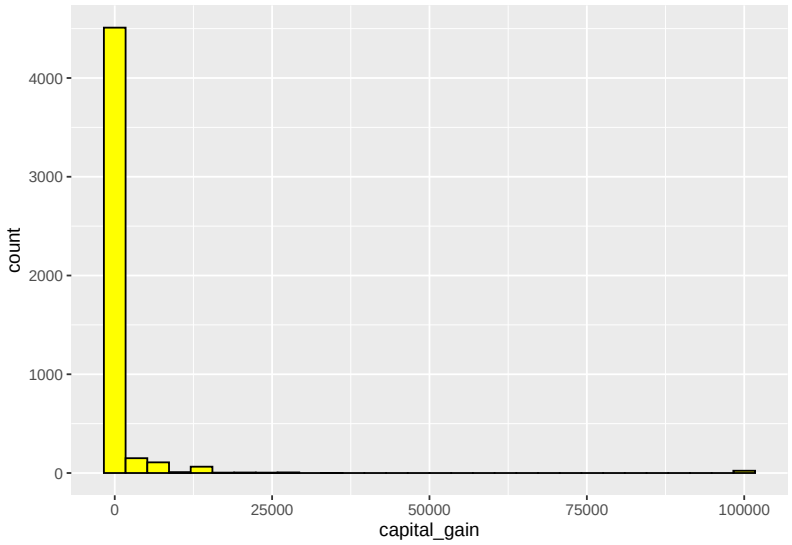
Osservazioni sulle variabili numeriche:

1. **Età**: Possiamo notare di come l'età media sia di *38* anni e la deviazione standard pari a *15.48* indica di quanto il valore si discosta da quello medio. L'età più piccola è 17, mentre quella più grande è 90. Dall'analisi dei quartili possiamo evincere che il 25% (primo quartile) delle osservazioni è un' età al di sotto di 28, mentre il 75% (terzo quartile) è un'età al di sotto di 48. Visualizziamo l'istogramma nella figura seguente:



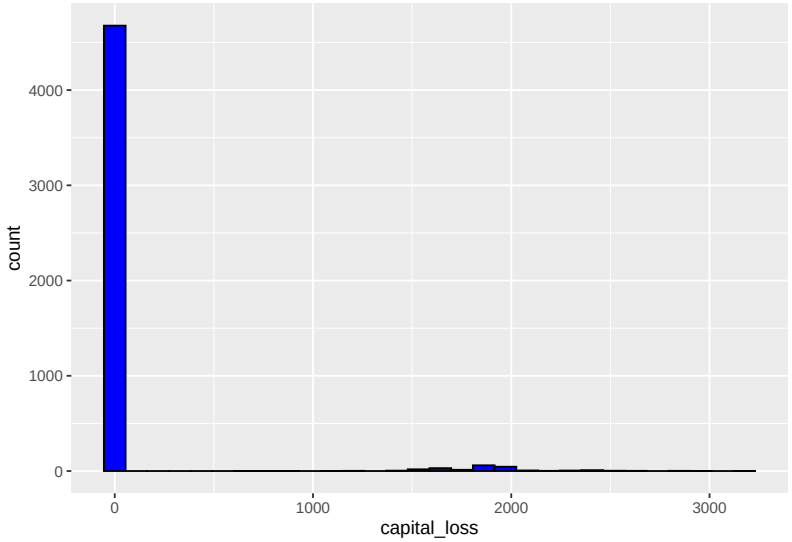
Quello che possiamo notare dalla distribuzione è sicuramente la presenza di asimmetria, con la conseguente presenza di coda a destra della distribuzione(*skewness*)

2. **capital-gain**: in questo caso la variabile assume valore medio di circa *1079*. Il secondo quartile (la mediana) è nullo, il che indica che la distribuzione è fortemente asimmetrica a destra. Sempre dai quartili possiamo desumere di come il *capital gain* si concentri attorno al valore 0 oppure attorno ad un valore molto alto: quindi un individuo può o non avere alcun guadagno oppure averne uno molto alto. Visualizziamo l'istogramma della distribuzione:



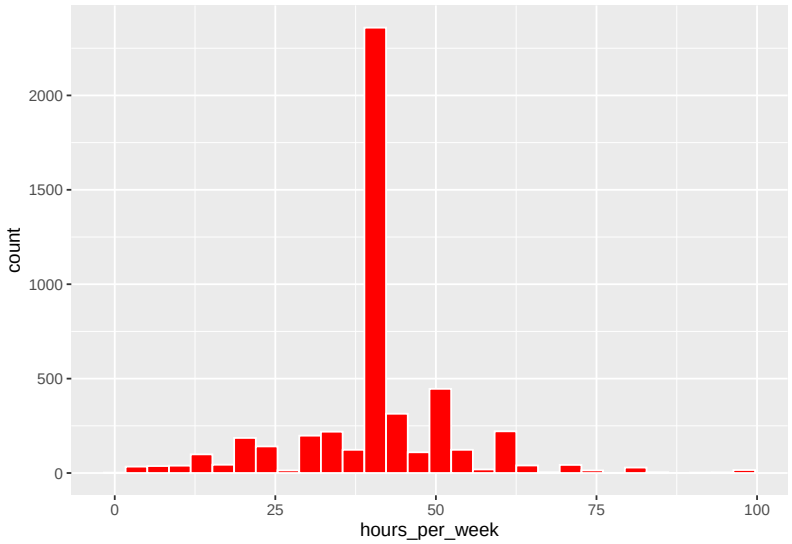
Possiamo notare di come l'istogramma mostri quanto affermato poc'anzi. Molti valori sono 0, mentre solo alcuni assumono valori di *capital-gain* alto

3. **capital-loss**: questo attributo è simile al precedente analizzato. La mediana è zero e sicuramente avremo la presenza di *skewness* nell'istogramma della distribuzione. Visualizziamo l'istogramma:



Possiamo quindi notare quanto affermato in precedenza: la presenza di valori che sono spesso nulli e l'asimmetria.

4. **hours-per-week**: il significato di questo attributo è quello di indicare le ore lavorative settimanali di ogni persona. Il valor medio si aggira attorno alle 40 ore di lavoro settimanali. Il minimo corrisponde a 1, il massimo a 99. Circa il 75% delle persone lavora all'incirca 45(o meno) ore alla settimana.



L'istogramma mostra la presenza di una concentrazione molto vasta di valori fra $30 - 40$ ore. La maggior parte delle persone lavora all'incirca 30/40 ore settimanali.

1.3. Analisi esplorativa delle variabili categoriche

Passiamo adesso all'analisi delle variabili categoriche:

8 Variables 4884 Observations

adult.workclass

n	missing	distinct
4884	0	8

```
lowest :  Federal-gov       Local-gov         Never-worked      Private           Self-emp-inc
highest:  Private           Self-emp-inc      Self-emp-not-inc  State-gov         Without-pay
```

Value	Federal-gov	Local-gov	Never-worked	Private
Frequency	148	305	1	3662
Proportion	0.030	0.062	0.000	0.750

Value	Self-emp-inc	Self-emp-not-inc	State-gov	Without-pay
Frequency	169	390	207	2
Proportion	0.035	0.080	0.042	0.000

adult.education

n	missing	distinct
4884	0	16

```
lowest :  10th          11th          12th          1st-4th       5th-6th
highest:  HS-grad       Masters       Preschool     Prof-school   Some-college
```

10th (145, 0.030), 11th (187, 0.038), 12th (61, 0.012), 1st-4th (24, 0.005), 5th-6th (67, 0.014), 7th-8th (95, 0.019), 9th (81, 0.017), Assoc-acdm (164, 0.034), Assoc-voc (217, 0.044), Bachelors (832, 0.170), Doctorate (51, 0.010), HS-grad (1537, 0.315), Masters (273, 0.056), Preschool (10, 0.002), Prof-school (74, 0.015), Some-college (1066, 0.218)

adult.marital_status

n	missing	distinct
4884	0	7

```
lowest :  Divorced              Married-AF-spouse      Married-civ-spouse     Married-spouse-absent  Never-married
highest:  Married-civ-spouse    Married-spouse-absent  Never-married          Separated              Widowed
```

Divorced (639, 0.131), Married-AF-spouse (2, 0.000), Married-civ-spouse (2233, 0.457), Married-spouse-absent (66, 0.014), Never-married (1626, 0.333), Separated (156, 0.032), Widowed (162, 0.033)

adult.occupation

```
     n    missing    distinct
  4884          0          14

lowest :  Adm-clerical      Armed-Forces      Craft-repair      Exec-managerial   Farming-fishing
highest:  Prof-specialty    Protective-serv   Sales             Tech-support      Transport-moving

Adm-clerical (562, 0.115), Armed-Forces (3, 0.001), Craft-repair (603, 0.123), Exec-managerial
(616, 0.126), Farming-fishing (154, 0.032), Handlers-cleaners (205, 0.042), Machine-op-inspct
(328, 0.067), Other-service (491, 0.101), Priv-house-serv (19, 0.004), Prof-specialty (909,
0.186), Protective-serv (96, 0.020), Sales (522, 0.107), Tech-support (152, 0.031),
Transport-moving (224, 0.046)
```

adult.relationship

```
     n    missing    distinct
  4884          0          6

lowest :  Husband         Not-in-family   Other-relative  Own-child       Unmarried
highest:  Not-in-family   Other-relative  Own-child       Unmarried       Wife

Value             Husband  Not-in-family Other-relative      Own-child      Unmarried
Frequency            1960           1298            150            765            474
Proportion          0.401          0.266          0.031          0.157          0.097

Value                Wife
Frequency             237
Proportion          0.049
```

adult.race

```
     n    missing    distinct
  4884          0          5

lowest :  Amer-Indian-Eskimo  Asian-Pac-Islander  Black              Other              White
highest:  Amer-Indian-Eskimo  Asian-Pac-Islander  Black              Other              White

Value      Amer-Indian-Eskimo Asian-Pac-Islander              Black              Other
Frequency                  46                160                497                 41
Proportion              0.009              0.033              0.102              0.008

Value                  White
Frequency               4140
Proportion             0.848
```

adult.sex

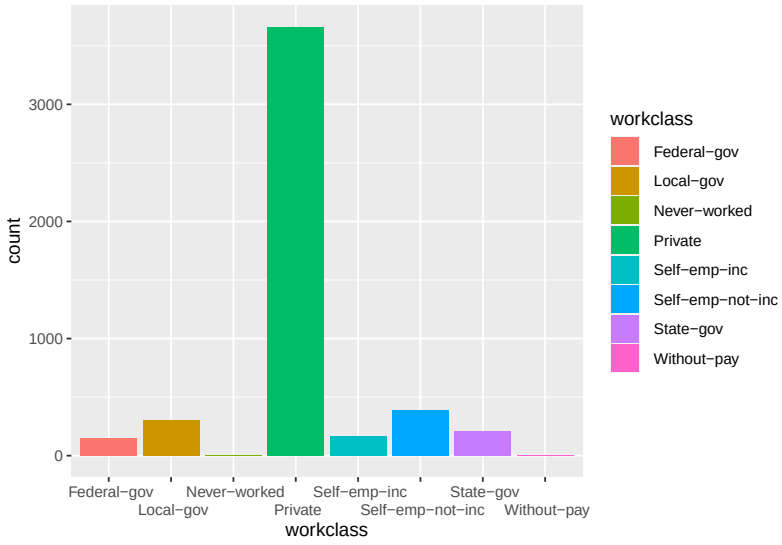
```
     n    missing    distinct
  4884          0          2

Value      Female   Male
Frequency    1632   3252
Proportion  0.334  0.666
```

adult.native_country

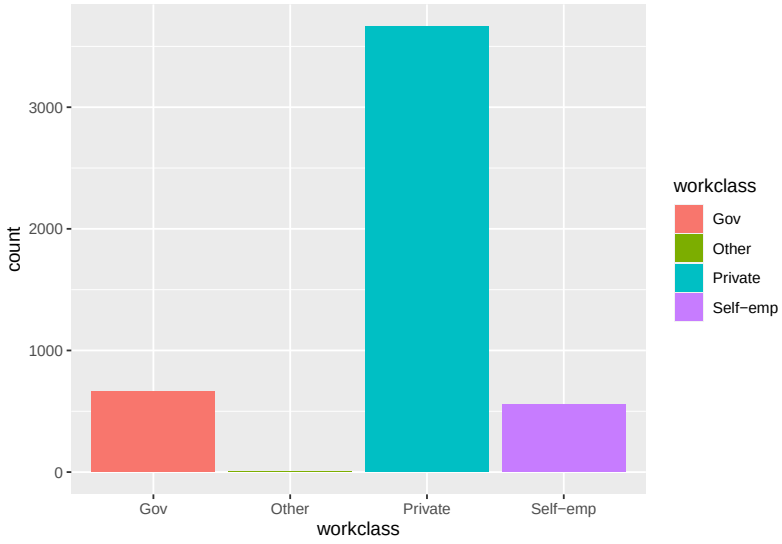
```
     n    missing    distinct
  4884          0         40

lowest :  Cambodia         Canada           China            Columbia         Cuba
highest:  Thailand         Trinadad&Tobago  United-States    Vietnam          Yugoslavia
```

Considerazioni sulle variabili categoriche:

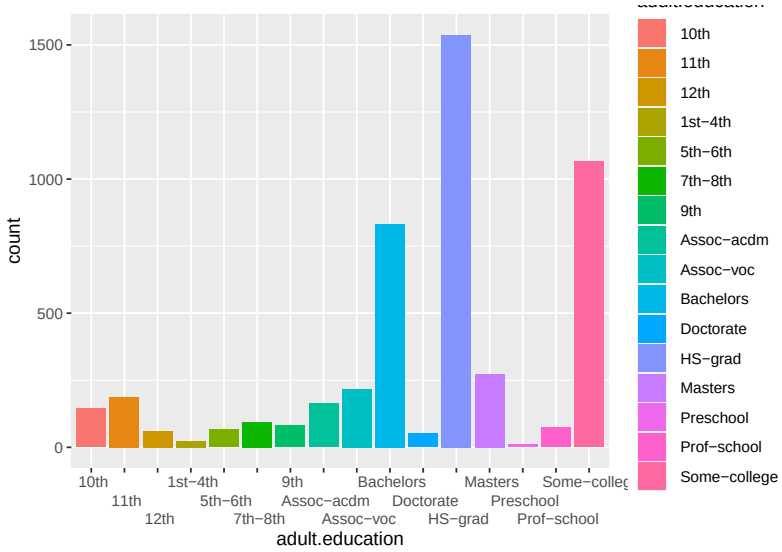
1. **workclass**: questa variabile indica la categoria di lavoro di ogni persona. Possiamo notare da una prima lettura della tabella sovrastante, di come il valore modale più alto sia *Private*. Visualizziamo nel *barplot* seguente le frequenze di ogni categoria lavorativa:



Ci sono 8 categorie di lavoro, la cui categoria predominante è *"private"*. Poichè le persone che non hanno mai lavorato sono davvero poche(ca. 10) e poichè alcuni valori sono molto simili fra loro(es. "Federal-gov" e "Local-gov"), possiamo riassumere tutte queste variabili in 4 differenti livelli: *lavoro "statale"*, *lavoro "autonomo"*, *lavoro "privato"* e *"altro"*.

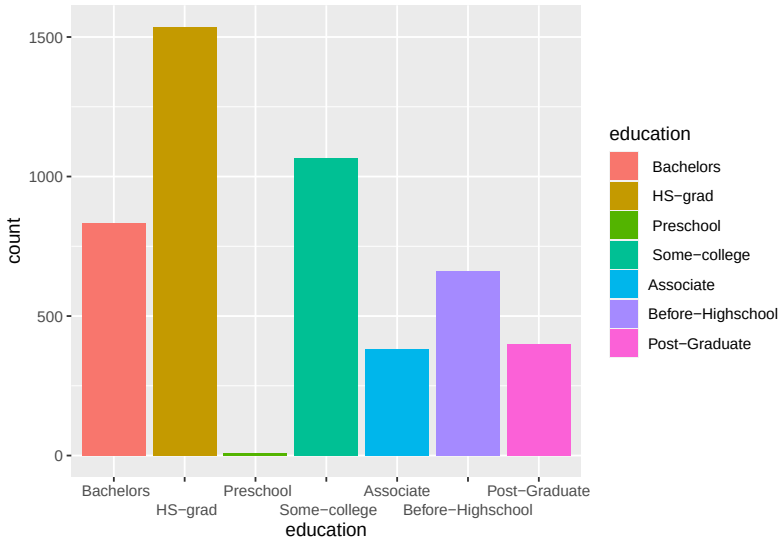


2. **education**: questa variabile può assumere 16 differenti valori. Visualizziamo nel diagramma a barre seguente le frequenze

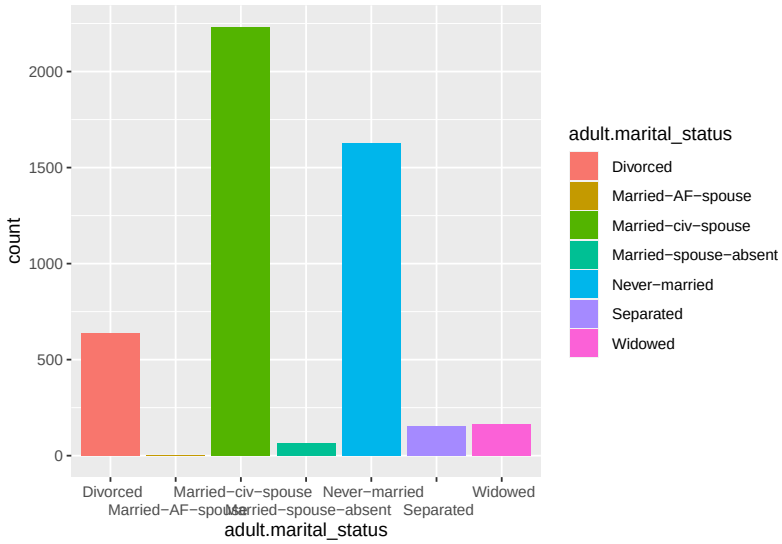


La frequenza più alta è relativa ad *HS-grad*. Anche in questo caso possiamo sintetizzare alcuni valori della variabile:

- Dal primo al dodicesimo grado riassumiamo ad etichettiamo con "Before-Highschool"
- I college biennali sono riassunti con *Associate*(titolo di studio che richiede due anni dopo la high-school)
- I *master*, *dottorati* e le *Prof-school* sono riassunti nell'attributo *Post-Graduate*

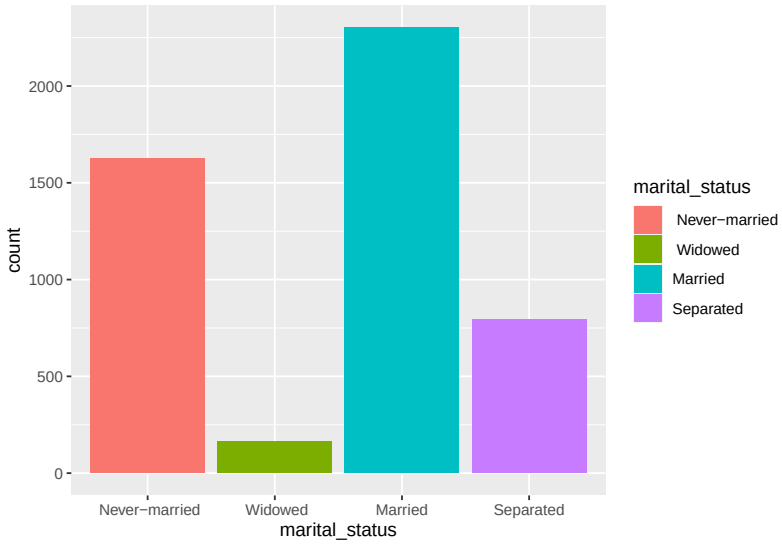


3. **marital-status**: possiamo notare di come abbiamo 7 diversi tipi di stato civile. Visualizziamo il diagramma a barre



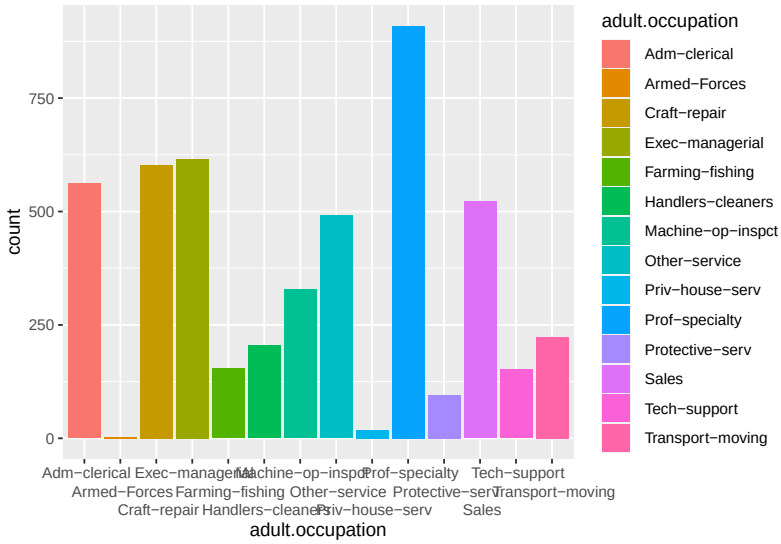
Possiamo pensare di riassumere:

- *Divorced* e *Separated* con un unico valore *Separated*
- Tutte le variabili che hanno come prefisso "*Married*" con *Married*

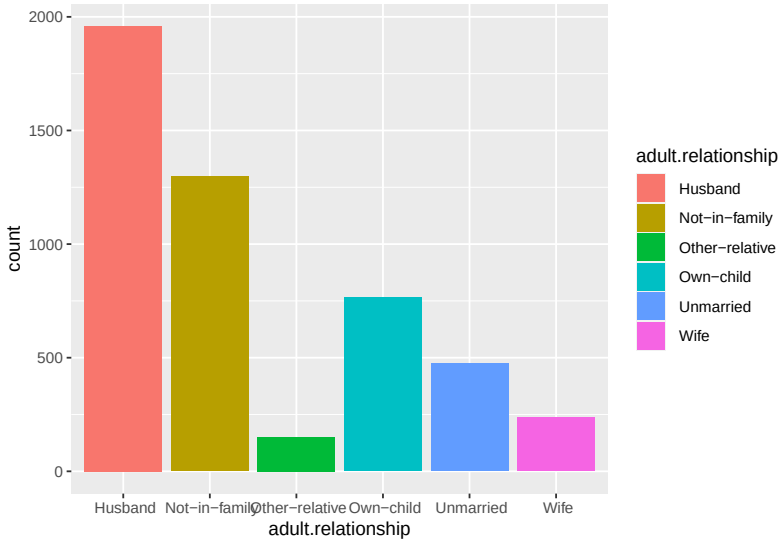


La maggior parte delle persone è sposata.

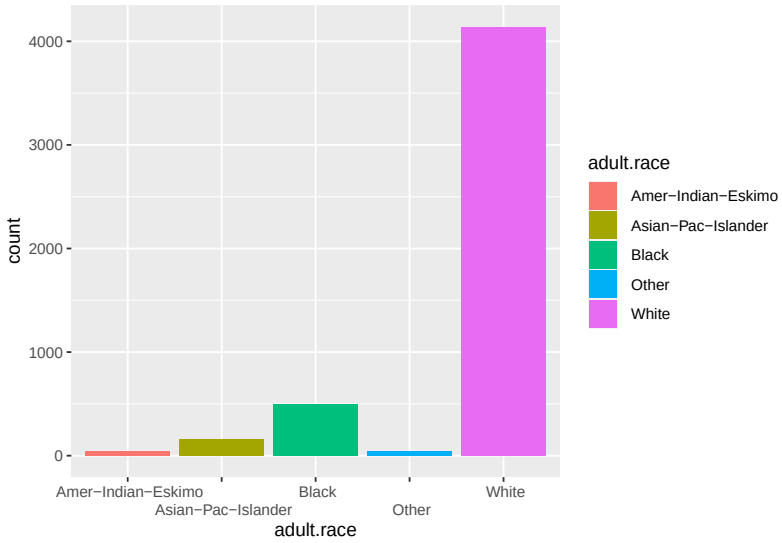
4. **occupation**: questa variabile indica l'occupazione lavorativa di ogni singolo individuo. Il valore modale più alto è denotato dal valore *Prof-specialty*. Visualizziamo il diagramma a barre:



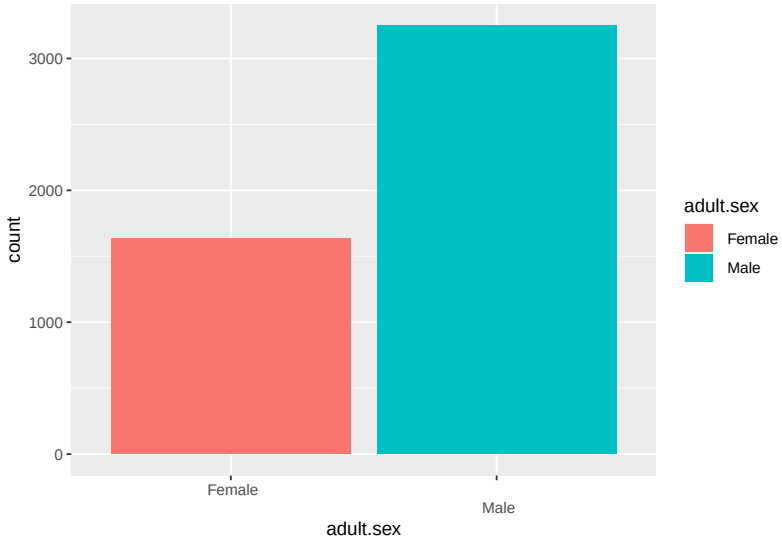
5. **relationship**: l'attributo indica la relazione dell'individuo nel nucleo familiare. Ci sono 6 valori unici nel dataset. Visualizziamo il diagramma a barre



6. **race**: attributo che indica la razza dell'individuo. Possiamo notare la presenza di 5 valori unici, la cui maggior parte è *white*, a cui segue *black*. Di seguito il diagramma a barre:

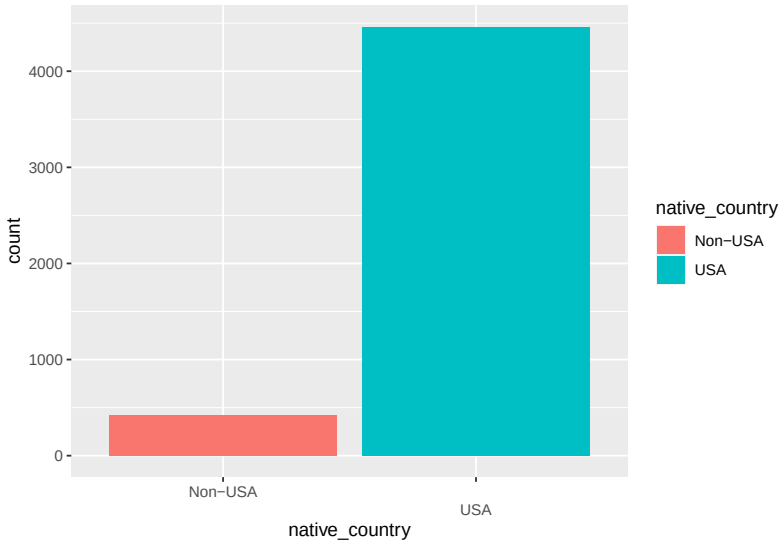


7. **sex**: attributo indicante il sesso di un singolo individuo. Il diagramma a barre è il seguente:



Possiamo notare di come il sesso maschile prevalga.

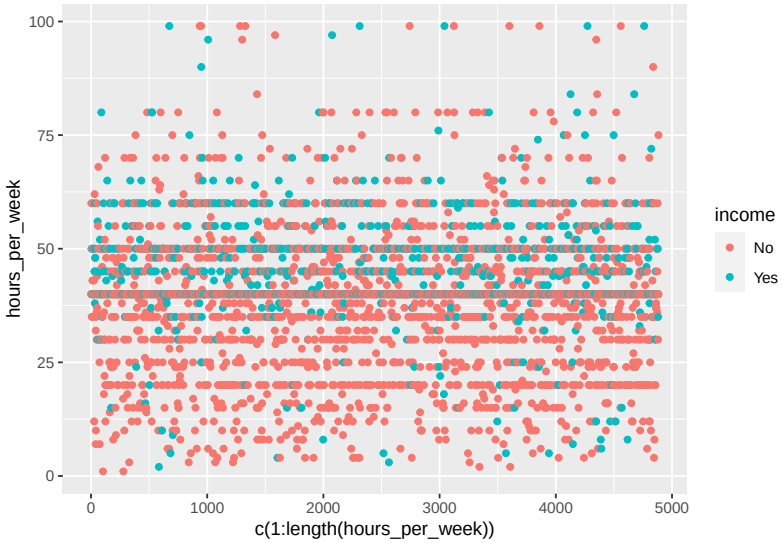
8. **native_country**: attributo indicante la provenienza di origine del singolo individuo. Poichè abbiamo differenti valori, procediamo prima nel sintetizzare con "cittadino di provenienza statunitense e non". Visualizziamo il diagramma a barre con i dati "trasformati":



Come si può evincere dal diagramma a barre, la maggior parte delle persone ha origini statunitensi.

1.4. Outlier detection

Procediamo a verificare la presenza di valori *outlier*(anomali) tra le variabili numeriche. Visualizziamo la distribuzione della variabile **hours per week** rispetto alle classi:



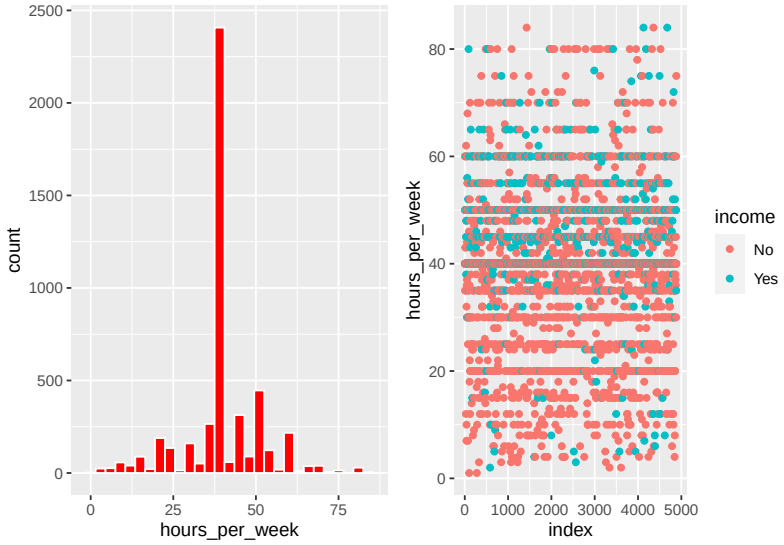
Possiamo notare con evidenza la presenza di outlier: l' anomalia principale risiede nel fatto che è praticamente impossibile che un individuo lavori così tante ore a settimana.

Effettuando un test χ^2 possiamo notare di come il valore 99 sia un *outlier*.

```
##
##  chi-squared test for outlier
##
## data:  adult$hours_per_week
## X-squared = 22.655, p-value = 1.939e-06
## alternative hypothesis: highest value 99 is an outlier
```

Inoltre, supponendo che una persona lavori 12 ore al giorno **tutti** i giorni(aspetto praticamente improbabile),

avremmo un monte ore pari ad $84(12 \times 7)$ ore settimanali. Tuttavia, nel dataset sono presenti valori maggiori di 84, pertanto, tollerando ad esempio qualche caso in cui si lavori effettivamente più di 84 ore, considereremo *outliers* i valori che si trovano nel range $[90 - 99]$. Procediamo a sostituirli rispettivamente con la media. Visualizziamo, dunque, il diagramma a barre.



Notiamo di come i valori anomali sono stati rimossi.

2. Applicazione dell'algoritmo SVM sul dataset

Iniziamo con il standardizzare le variabili numeriche all'interno del dataset, affinchè abbiano media nulla e varianza unitaria.

Procediamo con il partizionare il dataset in training e test set: la percentuale scelta è 80% per il dataset di training e 20% per quello di test.

Avendo a che fare con variabili categoriche all'interno del dataset, possiamo convertire tali variabili in variabili "dummy", che assumeranno il valore 1 se una particolare caratteristica è vera, e 0 altrimenti.

Procediamo con l'utilizzo di *caret*, utilizzando kernel differenti.

2.1 kernlab e kernel lineare

Procediamo ad utilizzare un kernel lineare e il calcolo in parallelo con 3 *cores*. Utilizziamo inoltre una $5 - fold$ cross-validation e, attraverso il parametro *tuneGrid*, otteniamo il valore del parametro C ottimale.

```
##   C
## 4 2
```

```
##    user  system elapsed
##    6.81    0.15  378.66
```

Il valore migliore di C ottenuto è pari ad 2. Possiamo visualizzare anche tempo di calcolo.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  684  82
##        Yes  57 153
##
```

```
##                Accuracy : 0.8576
##                  95% CI : (0.8341, 0.8789)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : 1.976e-14
##
##                   Kappa : 0.5958
##
##  Mcnemar's Test P-Value : 0.04179
##
##             Sensitivity : 0.9231
##             Specificity : 0.6511
##          Pos Pred Value : 0.8930
##          Neg Pred Value : 0.7286
##              Prevalence : 0.7592
##          Detection Rate : 0.7008
##    Detection Prevalence : 0.7848
##       Balanced Accuracy : 0.7871
##
##        'Positive' Class : No
##
```

Otteniamo, con un kernel lineare, una accuratezza di circa l' 86%.

2.2 e1071 e kernel lineare

Procediamo ad utilizzare un kernel lineare utilizzando la libreria *e1071* e il calcolo in parallelo con 3 *cores*. Utilizziamo inoltre una $5 - fold$ cross-validation, standardizziamo i dati numerici, e attraverso il parametro *tuneGrid* otteniamo il valore del parametro C ottimale.

```
##   cost
## 1 0.25
```

```
##    user  system elapsed
##    7.28    0.17  755.69
```

Notiamo di come i tempi di calcolo risultino essere maggiori con l'utilizzo di *e1071*.

Passiamo alla fase predittiva:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  684  82
##        Yes  57 153
##
##                Accuracy : 0.8576
##                  95% CI : (0.8341, 0.8789)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : 1.976e-14
##
##                   Kappa : 0.5958
##
##  Mcnemar's Test P-Value : 0.04179
##
##             Sensitivity : 0.9231
##             Specificity : 0.6511
```

```
##          Pos Pred Value : 0.8930
##          Neg Pred Value : 0.7286
##              Prevalence : 0.7592
##          Detection Rate : 0.7008
##    Detection Prevalence : 0.7848
##       Balanced Accuracy : 0.7871
##
##        'Positive' Class : No
##
```

L'accuratezza predittiva risulta pari all' 86%, consistentemente con i risultati ottenuti rispetto alla libreria precedente.

2.3 kernlab e kernel gaussiano

In questa sezione procediamo ad addestrare il modello attraverso una *support vector machine* basata su kernel gaussiano. Innanzitutto, poichè la complessità nel calcolo attraverso il metodo *train()* è molto alta, calcoliamo una stima dell'iperparametro attraverso il metodo **sigest()** di *kernlab*.

```
##         sigma   C
## 10 0.01241182 128
```

```
##    user  system elapsed
##    3.71    0.25   70.83
```

Con un kernel gaussiano otteniamo che il parametro C migliore ottenuto è 128 ed i tempi sono ridotti poichè abbiamo impostato il valore di *sigma* in partenza, senza far lavorare il *train()* alla ricerca dell'iperparametro ottimale, in quanto una sua stima l'abbiamo ottenuta dalla funzione *sigest()*, fornita all'interno della libreria *kernlab*.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  693  81
##        Yes  48 154
##
##                Accuracy : 0.8678
##                  95% CI : (0.845, 0.8885)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6203
##
##  Mcnemar's Test P-Value : 0.004841
##
##             Sensitivity : 0.9352
##             Specificity : 0.6553
##          Pos Pred Value : 0.8953
##          Neg Pred Value : 0.7624
##              Prevalence : 0.7592
##          Detection Rate : 0.7100
##    Detection Prevalence : 0.7930
##       Balanced Accuracy : 0.7953
##
##        'Positive' Class : No
##
```

Con un kernel gaussiano otteniamo una accuratezza di circa 87%.

2.4 kernlab e kernel polinomiale

Procediamo a questo punto ad utilizzare un kernel polinomiale, utilizzando *kernlab*.

```
##    degree      scale   C
## 22      3 0.01241182 256
```

```
##    user  system elapsed
##    5.56    0.13  120.71
```

Possiamo notare di come i tempi di calcolo non siano stati molto alti e soprattutto il considerare come grado ottimale del polinomio separatore, il grado 3. Il costo migliore ottenuto è $2^8 = 256$ per quanto riguarda il parametro C. Procediamo con la predizione.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  No Yes
##        No  694  86
##        Yes  47 149
##
##                Accuracy : 0.8637
##                  95% CI : (0.8406, 0.8846)
##     No Information Rate : 0.7592
##     P-Value [Acc > NIR] : 3.382e-16
##
##                   Kappa : 0.6049
##
##  Mcnemar's Test P-Value : 0.0009842
##
##             Sensitivity : 0.9366
##             Specificity : 0.6340
##          Pos Pred Value : 0.8897
##          Neg Pred Value : 0.7602
##              Prevalence : 0.7592
##          Detection Rate : 0.7111
##    Detection Prevalence : 0.7992
##       Balanced Accuracy : 0.7853
##
##        'Positive' Class : No
##
```

Anche in questo caso, l'accuratezza predittiva è di circa l'86%, in generale potremo dire una buona accuratezza.