Esame di Analisi dei dati 8 febbraio 2024

Avete 1h e 40m!

Risposte errate nelle domande a risposta multipla e vero/falso pesano negativamente sulla valutazione.

Nome e cognome	Matricola	1

1 Si dica quali affermazioni sono vere e quali false.

[V] [F] L'algoritmo delle k-medie è un metodo di raggruppamento mediante procedimento iterativo

[V] [F] Il campionamento a stadi si utilizza quando le unità da estrarre ad ogni stadio sono simili tra loro per una o piu carattersitiche

[V] [F] Se in un modello di regressione la variabile di risposta è fortemente correlata con le variabili indipendenti si parla di multicollinearità

[V] [F] Nel caso di dati quantitativi, un possibile criterio di imputazione dei dati mancanti è l'uso della media

2 Indicare 4 fasi consecutive del protocollo Cross-Industry Standard Process for Data Mining (CRISP-DM)

☐ 1) ..... ☐ 2) .....

☐ 3) ..... ☐ 4) .....

3 Si dica quali affermazioni sono vere e quali false o completare le risposte.

[V] [F] In un modello di regressione la variabile di risposta non può essere dicotomica

[V] [F] La silhouette è un indice di bontà di una partizione dei dati

[V] [F] Per effettuare un'analisi in componenti principali le variabili dovrebbero essere correlate tra loro

[V] [F] La matrice di correlazione è una matrice quadrata, ossia ha ugual numero di righe e colonne

4 Descrivere brevemente i principali vantaggi e svantaggi del repertimento dati dalle varie fonti

5 Il reparto marketing di una grande azienda assicurativa operante negli Stati Uniti vuole valutare l'efficacia di una campagna pubblicitaria relativa ad un nuova polizza assicurativa in termini di persone che la sottoscrivono (variabile Y).
Per fare cio viene realizzata un'indagine campionaria e vengono raccolte informazioni in merito all'ammontare (in dollari) speso per la pubblicità su vari media e il numero di sottoscrittori che ne è conseguito. In particolare le variabili osservate sono le seguenti 5:

- TV - migliaia di dollari spesi per annunci pubblicitari in TV
- Internet - migliaia di dollari spesi per annunci pubblicitari su siti web e social networks
- Mailing - migliaia di dollari spesi per invio di annunci pubbllicitari via mail
- Members - numero di persone (in migliaia) che hanno sottoscritto il contratto assicurativo a seguito della campagna pubblicitaria
- Region - regione geografica degli Stati Uniti in cui la cammpagna pubblcitaria è stata implementata

I risultati del modello sono i seguenti:

```
#> Call:
#> lm(formula = Members ~ TV + Internet + Mailing + Region, data = marketing)
#>
#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -17.1643  -1.4354   0.5728   2.3103   6.0815
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  5.415167   0.723235   7.487 2.44e-12 ***
#> TV           0.045750   0.001396  32.767  < 2e-16 ***
#> Internet     0.188204   0.008652  21.753  < 2e-16 ***
#> Mailing     -0.001185   0.005907  -0.201    0.841
#> RegionEast   1.017463   0.674725   1.508    0.133
#> RegionSouth  0.829555   0.646325   1.283    0.201
#> RegionWest   0.242106   0.664486   0.364    0.716
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 3.37 on 193 degrees of freedom
#> Multiple R-squared:  0.8988, Adjusted R-squared:  0.8957
#> F-statistic: 285.8 on 6 and 193 DF,  p-value: < 2.2e-16
```

a. Sappiamo che la reference category della variabile Region è Central. Cosa vuol dire? In quale coefficiente del modello di regressione si osserva l'effetto di tale modalità?

b. Commentare il valore dell'indice R^2

c. Commentare il p-value del test congiunto sui coefficienti del modello

d. Commentare le stime dei coefficienti del modello in relazione ai dati del problema. A quali conclusioni dovrebbe pervenire il reparto marketing dell'azienda?

6 I seguenti dati riguardano la "solvibilità" o "affidabilità creditizia" (ovvero la capacità di rimborsare i prestiti erogati) dei clienti di una banca tedesca.
Utilizzando un opportuno modello di regressione la banca cerca di prevedere la probabilità che un prestito venga rimborsato sulla base di alcune caratteristiche economiche o personali del cliente.
Per stimare tale probabilità viene usato un dataset in possesso della banca di $n = 1000$ prestiti emessi. Ad ogni prestito è associata una variabile di risposta definita come:
$y = 0$ il cliente ha rimborsato il prestito $y = 1$ il cliente non ha rimborsato il prestito.
Altre caratteristiche osservate sui clienti sono le seguenti:

- acc = stato della situazione delle giacenze in banca (no= nessuna giacenza attiva; bad=giacenza sotto soglia; good=giacenza sopra soglia)
- duration = durata del credito in mesi
- amount = importo del credito in migliaia di euro
- moral = comportamento di rimborso precedenti prestiti (0 = cattivo pagatore ossia mancati rimborsi pregressi; 1 = buon pagatore ossia nessun mancato rimborso pregresso)
- intuse = uso previsto del prestito (1 = privato; 0 = affari)

Di seguito i risultati di un primo modello stimato sui dati:

```r
mod1=glm(y~ acc+duration+amount+moral+intuse,family=binomial(link=logit))
summary(mod1)

##
## Call:
## glm(formula = y ~ acc + duration + amount + moral + intuse, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8876  -0.8440  -0.4628   0.9629   2.3620
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.284402   0.302579  -0.940 0.347255
## accgood     -1.337748   0.201127  -6.651 2.91e-11 ***
## accno        0.617659   0.174728   3.535 0.000408 ***
## duration     0.033233   0.007746   4.290 1.78e-05 ***
## amount       0.045875   0.064092   0.716 0.474134
## moral       -0.986066   0.250891  -3.930 8.49e-05 ***
## intuse      -0.425536   0.158272  -2.689 0.007174 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

a. Di che modello statistico si tratta?

b. Quali sono le variabili che hanno impatto positivo, negativo o nullo sulla probabilità di solvibilità dei clienti?

c. Interpretare i valori dei coefficienti del modello dopo averne valutato la significatività