Report
Gennaio, 2023

Studio sulla conoscenza del Machine Learning in Italia

Rosario Urso
Dipartimento di Scienze Politiche, Università degli studi di Napoli Federico II
E-mail: r.urso@studenti.unina.it

Summary: Il seguente elaborato è stilato per comprendere quale siano i fattori che impattano maggiormente sulla conoscenza del **Machine Learning** attraverso due differenti approcci:
Il primo riguarda l'uso dei modelli cumulativi, in particolare dei Proportional Odds Models (POM); nell'altro, invece, si ritrova l'applicazione dei modelli mistura, ovvero CUB e varianti.

Keywords: Machine Learning, Proportional Odds Model, CUB.

1. Introduzione

Il lavoro[1] oggetto d'esame concerne la conoscenza verso gli strumenti finanziari. In questo elaborato, è stato scelto di focalizzare l'attenzione sul Machine Learning. Per Machine Learning si intende *la scienza in grado di sviluppare algoritmi e modelli statistici utilizzati dai sistemi informatici per lo svolgimento di compiti senza istruzioni esplicite e basandosi, invece, su modelli e inferenza*.

[1] La corrispondente analisi è stata condotta dagli studenti dell'Università degli Studi di Napoli Federico II e dagli studenti dell'Università di Pavia nell'ambito del progetto VMG "Differenze di genere nella conoscenza e nell'uso dei prodotti Fintech: esiste un ruolo per la trasparenza?", parte del progetto europeo CA19130 Fintech and Artificial intelligence in Finance. I dati sono stati raccolti dall'Ottobre 2022 a Dicembre dello stesso anno. Il metodo di campionamento utilizzato è di tipo non probabilstico (campionamento a valanga).

Tale variabile dipendente è una risposta di tipo qualitativo su scala ordinale (presenta $m = 6$ categorie di risposta, dove con 0 il soggetto rispondente indica la non conoscenza in materia di Machine Learning e con 5 la perfetta conoscenza).

2. Struttura del dataset

Il dataset[2], che consta di 614 osservazioni, è composto da variabili di tipo qualitativo ordinale e di tipo quantitativo, sia discrete che continue.
Al fine di poter svolgere l'analisi qui stilata sono state considerate le seguenti variabili:
genere, età, componenti del nucleo familiare, una variabile rinominata *coinquilino* che indica le persone con le quali vive il rispondente (se vive da solo, con il partner, con i genitori, etc), la macroarea italiana di residenza (suddivisa in nord, centro, sud e isole), il titolo di studio (laureato o no), l'area disciplinare in cui il soggetto studia o ha studiato, il settore lavorativo (suddiviso in area umanistica e scientifica), l'ateneo frequentato (in Università degli Studi di Napoli Federico II ed altro), la condizione occupazionale (in studente, che include anche studente lavoratore, ed altro, nel quale rientrano lavoratori autonomi e dipendenti, casalinghe e disoccupati), il settore lavorativo distinto in settore STEM ed altro, benessere economico (che risulta essere una risposta di tipo ordinale da 1 a 10, dove con 1 e con 10 si indica rispettivamente la facilità/difficoltà dell'individuo nell'arrivare a fine mese) ed infine, le variabili di tipo ordinale Cloudcomputing e Machine Learning con risposte da 0 a 5, dove con 0 si indica la non conoscenza nella relativa materia e con 5 la perfetta conoscenza.

[2] Il dataset inizialmente conteneva 625 osservazioni, al quale sono stati rimossi 11 individui che non risiedevano in Italia.

3. Proportional Odds Models

3.1. Implementazione del modello

I Proportional Odds Models[3] appartengono alla classe dei Generalized Linear Models (GLM) e vengono applicati nel caso in cui la variabile dipendente assuma *m* modalità di risposta ordinate.
Nei POM così come nei *Continuation-Ratio Models* e negli *Adiacent Category Models*, che rappresentano altri due modelli cumulativi, si assume che ci sia una variabile latente soggiacente, che non è direttamente osservabile.
Essa racchiude l'insieme delle caratteristiche dell'individuo sottoposto a questionario, e per tale motivo, varia da rispondente a rispondente.
Si utilizza quindi un modello di regressione per tale variabile continua non osservata che si presume sia alla base della nostra risposta Y.
Sia $\boldsymbol{Y^*}$ la variabile latente e $\boldsymbol{x}$ le relative variabili esplicative. Supponiamo che Y^* vari attorno ad un parametro di posizione $\boldsymbol{\eta}$ (come la media) che dipende da $\boldsymbol{x}$ attraverso la relazione $\boldsymbol{\eta(x)} = \boldsymbol{\beta' x}$.
Da tale assunzione deriva che:

$$Y^* = \boldsymbol{\beta' x} + \epsilon \tag{1}$$

dove ϵ è la componente stocastica di errore.
Se si ipotizza che la funzione di ripartizione della variabile latente $\boldsymbol{Y^*}$ sia

$$P(Y^* \leq y^*|x) = G(y^* - \eta) = G(y^* - \beta' x) \tag{2}$$

il valore medio ha una distribuzione che segue quella della funzione logistica con $\mathbb{E}(\epsilon) = 0$
Essendo la variabile latente definita su un supporto continuo, esso viene diviso mediante dei *thresholds* o *cutpoints* in *m* classi

$$-\infty = \alpha_0 < \alpha_1 < ... < \alpha_m = +\infty \tag{3}$$

[3] I Proportional Odds Models sono stati proposti da McCullagh nel 1980

Condizione essenziale è che i *thresholds* seguano un ordinamento crescente per poter poi procedere alla stima del modello. Con l'imposizione di tale vincolo è stata discretizzato il supporto della latente.
L'imposizione dei *thresholds* consente di calcolare la probabilità che la dipendente $Y = j$ quando $\alpha_{j-1} < Y < \alpha_j$, ovvero si osserva che la risposta Y cade nella j-esima categoria quando la latente è compresa tra il $j - 1$ e il $j - esimo$ threshold.
Sotto la struttura della variabile latente:

$$P(Y \leq j|x) = P(Y^* \leq \alpha_j|x) = G(\alpha_j - \boldsymbol{\beta'x}) \tag{4}$$

dove G rappresenta la funzione di ripartizione della variabile casuale logistica, per cui applicando l'inversa della funzione logistica G^{-1} a (4) si ricava:

$$logit[P(Y \leq j|x)] = \alpha_j - \boldsymbol{\beta'x} \tag{5}$$

Per determinare la probabilità che la Y assuma il $j - esimo$ valore condizionatamente al valore dei regressori, si definisce:

$$\begin{aligned} \pi_j(x) = P(Y = j) &= P\left(\tau_j < \boldsymbol{x\beta} + \epsilon \leq \tau_{j+1}\right) \\ &= P(\tau_j - \boldsymbol{x\beta} < \epsilon \leq \tau_{j+1} - \boldsymbol{x\beta}) \\ &= P(\epsilon \leq \tau_{j+1} - \boldsymbol{x\beta}) - P(\epsilon \leq \tau_j - \boldsymbol{x\beta}) \end{aligned} \tag{6}$$

Tale probabilità equivale a calcolare la probabilità che il predittore sia compreso tra i due *thresholds*, che si completa nell'espressione in cui, esplicitando la funzione rispetto ai termini di errore, si ottiene la differenza tra le due funzioni di ripartizione calcolate rispetto ai *cutpoints* di riferimento.

3.2. Interpretazione del modello

Per quanto concerne l'interpretazione e per determinare l'impatto che hanno le covariate sulla conoscenza del Machine Learning, quando il link è di tipo *logistico*, ci si avvale dell'ausilio degli odds.

Si definisce $odds$ di un evento la probabilità che si verifichi l'evento successo diviso la probabilità dell'evento insuccesso.
Nei Proportional Odds Models, l'odds è dato da

$$odds_j\left(\boldsymbol{x}_i\right) = \frac{P_r\left(Y_i \leq j \mid \boldsymbol{x}_i\right)}{\mathrm{P}_r\left(Y_i > j \mid \boldsymbol{x}_i\right)} = \frac{P_r\left(Y_i \leq j \mid \boldsymbol{x}_i\right)}{1 - P_r\left(Y_i \leq j \mid \boldsymbol{x}_i\right)} \tag{7}$$

Ciò equivale a considerare la probabilità di avere una valutazione inferiore alla $j - esima$ categoria rispetto ad averne una superiore.
In presenza del link logistico, l'$odds_j$ sarà pari a

$$odds_j\left(\boldsymbol{x}_i\right) = exp(\alpha_j + \boldsymbol{\beta}'\mathbf{x}_i), \quad j = 1, \ldots, m \tag{8}$$

Il legame logistico, inoltre, ci consente di calcolare i log($odds$), che sono così definiti

$$logit[P(Y \leq j)] = \log \frac{P(Y \leq j)}{P(Y > j)} = \alpha_j - \boldsymbol{\beta}'\mathbf{x}_i, \quad j = 1, \ldots, m \tag{9}$$

Riguardo l'interpretazione:

- Se β_k > 0, è più probabile che Y cada delle categorie estreme della scala.
- se β_k < 0, si avrà una maggiore probabilità che la Y appartenga alle prime *j* categorie.

Rispetto agli *Adiacent Categori Models* e dei *Continuation Ratio Models* così come nei modelli per dati multinomiali, i POM risultano più parsimoniosi: tale modello ci restituisce le stime con i valori relativi ad α_j che rappresentano l'intercetta e i coefficienti di regressioni $\boldsymbol{\beta_k}$ associati alle relative variabili esplicative, che risultano costanti al variare di j, delle modalità di risposta.
Se il modello utilizzato fosse stato il Continuation Ratio Model, i parametri da stimare sarebbero stati k x m, ovvero tanti quante sono le modalità.

Nella quasi totalità dei casi vengono trattati modelli con più di una variabile esplicativa, per cui risulta interessante calcolare il cd. $log(oddsratio)$ (rapporto tra gli odds) per operare opportuni confronti.

$$\begin{aligned} &\operatorname{logit}[P(Y \leq j \mid \mathbf{x}_1)] - \operatorname{logit}[P(Y \leq j \mid \mathbf{x}_2)] \\ &= \log \frac{P(Y \leq j \mid \mathbf{x}_1)/P(Y > j \mid \mathbf{x}_1)}{P(Y \leq j \mid \mathbf{x}_2)/P(Y > j \mid \mathbf{x}_2)} = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2) \end{aligned} \tag{10}$$

da cui l'$oddsratio$

$$\frac{P(Y \leq j \mid \mathbf{x}_1)/P(Y > j \mid \mathbf{x}_1)}{P(Y \leq j \mid \mathbf{x}_2)/P(Y > j \mid \mathbf{x}_2)} = exp(\boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)) \tag{11}$$

Il log-odds ratio (e di conseguenza l'odds ratio) è proporzionale alla distanza tra x_1 e x_2. Inoltre, nel caso usuale in cui si consideri di cambiare un solo parametro alla volta di una unità, e^{β} rappresenta il rapporto di probabilità, come al solito nella regressione logistica. La differenza, tuttavia, è che $\boldsymbol{e^{\beta}}$ ora rappresenta un rapporto di quote cumulative, un odds ratio cumulativo.

3.3. Inferenza e verifica del modello

Successivamente alla fase di stima e interpretazione dei risultati del modello, bisogna controllare l'adeguatezza del modello attraverso indice di bontà di adattamento globale, test locali (per verificare sia rispettata l'assunzione di proporzionalità) e confronto tra diversi modelli. Per quanto riguarda il test di adattamento globale (in presenza di tabelle di contingenza), si specifica l'ipotesi nulla H_0 di buon adattamento ai dati contro l'ipotesi alternativa H_1 che denota un cattivo adattamento.
La statistica Test per la bontà di adattamento globale è

$$\mathrm{X}^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \tag{12}$$

con gradi di libertà pari al numero di logit cumulativi meno il numero di parametri del modello stimato $(c-1)(r-1)$.
Quando i dati raggiungono una numerosità accettabile e quando il modello contiene almeno una variabile esplicativa continua, si utilizzano test per la bontà di adattamento alternativi come quello di *Lipsitz et. Al* nel 1996. Tale test, che non è sempre calcolabile per piccoli campioni, è basato sull'approccio di *Hosmer-Lemeshow* per dati binari.
Nel 2004 *Pulkestenis* e *Robinson* nel 2004 hanno ovviato al problema della presenza nel modello di covariate continue andando a modificare la statistica χ^2

$$\chi^2 = \sum_{l=1}^{2} \sum_{k=1}^{K} \sum_{j=1}^{c} \frac{(O_{lkj} - E_{lkj})^2}{E_{lkj}} \tag{13}$$

dove l identifica i due sottogruppi in base alle categorie delle ordinali, K è il numero di esplicative osservate a causa delle covariate di natura categorica e c è il numero delle categorie della risposta. La distribuzione di riferimento per entrambe le statistiche è χ^2 con $(2K-1)(c-1)-p_{cat}-1$ gradi di libertà, dove p_{cat} è il numero di covariate categoriche.
Analiticamente per poter verificare tale assunzione, si possono seguire approcci diversi: con il test di *Brant* (1990) o con la verifica dei residui.
Il test di Brant compara le stime $\hat{\beta}$ per ogni predittore utilizzando il Test di *Wald*. Tale test verifica l'ipotesi di proporzionaltà dell'intero modello e delle singole variabili esplicative, per cui si rifiuterà l'ipotesi nulla di assunzione di proporzionalità con una statistica test χ^2 con $(k-2)p$ gradi di libertà. Rifiuteremo H_0 (*assumption proportional odds*) se il *p-value* è < 0.05, per cui si procederà poi a considerare *una semi-proportional assumption*.
Tale test presenta due tipi di problemi: nel caso sia presente un elevato k o p, i gradi di libertà tenderanno ad aumentare notevolmente con conseguente diminuzione del livello di potenza del test; in secondo luogo,

comprendere quali siano le cause della mancanza di adattamento non è agevole, in quanto, con la differenza delle stime, non viene manifestata la causa del mancato adattamento.
I motivi che portano al rifiuto del test sulla bontà di adattamento sono diversi. Si annovera: l'assenza di un importante covariata, di un effetto interazione o la presenza di problemi legati alla dimensione campionaria. Risulta, quindi, sensato comparare due modelli che differiscono, ad esempio, per una o più covariate o per un effetto interazione attraverso il test sul rapporto della verosimiglianza (**Likelihood Ratio Test**), che è un test per modelli annidati. Le ipotesi alla base del test sono

$$\begin{aligned} H_0 &: \beta_{k+1} = \cdots = \beta_p = 0 \\ H_1 &: \beta_{k+1} = \cdots = \beta_p \neq 0 \end{aligned} \tag{14}$$

La statistica test sarà rappresentata da

$$D(\hat{\theta}) = 2\left\{\ell^S - \ell(\hat{\theta})\right\} \tag{15}$$

dove ℓ^S rappresenta la *log-verosimiglianza* del modello saturo che presenta p parametri, mentre $\ell(\hat{\theta})$ rappresenta la *log-verosimiglianza* del modello annidato che presenta q parametri (con $p > q$).
La statistica Test si confronta con il percentile di una variabile casuale Chi-quadrato con $(q-p)$ gradi di libertà. Il test conduce al rifiuto dell'ipotesi nulla se il *p-value* < 0.05.
L'ipotesi H_0 impone che tutte le stime siano congiuntamente uguale a 0, e nel caso di rifiuto di tale ipotesi, verrà preferito il modello saturo: le stime $\beta_{k+1}, \beta_{k+2} \cdots \beta_p$ erano statisticamente significative.
Nel caso in cui si volesse effettuare un confronto per valutare quale tra i modelli con *link* differenti sia il migliore, si calcolano indici quali AIC (*Akaike Information Criterion*) e BIC (*Bayesian Information Criterion*) per modelli non annidati e si sceglie il modello che restituisce valori di **AIC/BIC** più bassi.

La costruzione della funzione di verosimiglianza ci consente di ricavare

le stime relative ai *threshold* e ai coefficienti

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n}\prod_{j=1}^{M} \Pr\left(Y_i = j \mid X_i\right)^{I[y_i=j]} \\ &= \prod_{i=1}^{n}\prod_{j=1}^{M} \left[F\left(\alpha_j - \boldsymbol{X_i\beta}\right) - F\left(\alpha_{j-1} - \boldsymbol{X_i\beta}\right)\right]^{I[y_i=j]} \end{aligned} \tag{16}$$

dove

$$\left[F(\alpha_j - \boldsymbol{X_i\beta}) - F(\alpha_{j-1} - \boldsymbol{X_i\beta})\right]^{I[y_i=j]} \tag{17}$$

rappresenta la differenza tra le funzioni di ripartizione e $\boldsymbol{I[y_i = j]}$ costituisce la variabile indicatrice, che assume valore 1 se è osservata la *j-esima* categoria, 0 altrimenti.
Come la regressione logistica per dati ordinari, tali equazioni non hanno una soluzione in forma chiusa e quindi si ricorre a metodi iterativi, quali algoritmo **E-M** (*Expectation-Maximization*)

4. CUB Models

4.1. Implementazione del modello

I modelli mistura, più specificamente i modelli CUB, **Combination of Discrete Uniform and Shifted Binomial** (Piccolo 2003), rappresentano un'alternativa ai modelli cumulativi, che, come specificato precedentemente, non sono modelli parsimoniosi in quanto richiedono la stima dei *threshold* e dei singoli parametri per ogni covariata.
Inoltre, i cumulative model non offrono una descrizione di ciò che viene definito *processo generatore dei dati*, ovvero l'insieme delle caratteristiche del rispondente che genera una determinata scelta. Per contro, l'idea che è alla base del CUB è quella di poter descrivere la risposta come una combinazione, sia degli aspetti percettivi che aspetti decisionali del rispondente.
Si ipotizza, quindi, che dietro percezione e decisione, ci siano 2 componenti:

- **Feeling**, che, a seconda degli argomenti, può avere sia un'accezione negativa che positiva. Può essere interpretato come accordo, felicità, percezione, preoccupazione, repulsione ed ha a che fare con la motivazione dell'individuo, con il suo livello di conoscenza e con il suo passato

- **Uncertanty**, che non deve essere intesa come incertezza campionaria, ma come incertezza del rispondente addizionata agli effetti di contesto, che possono riguardare il modo in cui viene raccolto e come viene erogato il questionario (*face-to-face*, telefonicamente, via mail), dalla lunghezza della scala, ma anche dalla propensione che ha il soggetto nel compilare il questionario, dalla stanchezza e dalla noia.

Il **CUB** supporta due variabili casuali discrete per le due componenti: la vc. Binomiale Traslata per la componente di feeling, mentre l'incertezza è descritta dalla vc. Uniforme.
Assumendo che R_i sia la risposta dell'i-esimo soggetto, la probabilità che il soggetto scelga la r-esima categoria dato le sue caratteristiche **xi**, **wi** attraverso la componente di feeling e uncertanty pesate è

$$\Pr\left(R_i = r \mid \boldsymbol{x}_i, \boldsymbol{w}_i\right) = \pi_i \left[b_r\left(\xi_i\right)\right] + \left(1 - \pi_i\right)\left[\frac{1}{m}\right] \qquad (18)$$
$$\forall r = 1, 2, \ldots, m, \forall i = 1, 2, \ldots, n$$

che rappresenta la **componente sistematica**, dove

$$b_r\left(\xi_i\right) = \left(\begin{array}{c} m-1 \\ r-1 \end{array}\right) \xi_i^{m-r} \left(1 - \xi_i\right)^{r-1}, \quad r = 1, 2, \ldots, m \qquad (19)$$

Il modello risulta ben definito (se il numero di categorie di risposta *m* > 3) perché $\pi \in (0,1)$, $\xi \in (0,1)$.
Nel caso in cui il valore associato a π risultasse 0, la probabilità di risposta sarà guidata esclusivamente dall'incertezza; nel caso opposto, invece, la componente di uncertanty non sarà presente e si potrebbe supporre che il soggetto abbia risposta con la massima consapevolezza.
$\boldsymbol{\pi}$ quindi riesce a stabilire quanto feeling e uncertanty è presente nella

scelta di ogni singolo rispondente.
Per cui, $(\mathbf{1}-\boldsymbol{\pi})$ misura il peso dell'incertezza delle risposte perché aumenta con l'importanza dell'Uniforme discreta nella mistura, mentre $\boldsymbol{\pi}$ misura il peso del feeling, ovvero esso sarà tanto più pesante quanto più bassa sarà l'indecisione ad essa associato.

4.2. Interpretazione del modello

Mediante il link di tipo logistico, è possibile specificare l'effetto di covariate ($\boldsymbol{x}_i$, $\boldsymbol{w}_i$) sul feeling e sull'incertezza

$$\begin{cases} \text{logit}(\pi_i) = \log(\frac{\pi_i}{1-\pi_i}) = x_i\beta; \\ \text{logit}(\xi_i) = \log(\frac{\xi_i}{1-\xi_i}) = w_i\gamma; \end{cases} \iff \begin{cases} \pi_i = \frac{1}{1+e^{-x_i\beta}} \\ \xi_i = \frac{1}{1+e^{-w_i\gamma}} \end{cases} \tag{20}$$

Tale funzione legame (che rappresenta un link robusto per dati di tipo ordinale) rende inoltre possibile ottenere un'interpretazione semplice ed immediata tra le covariate dei rispondenti e i parametri $(\mathbf{1}-\boldsymbol{\pi})$ e $(\mathbf{1}-\boldsymbol{\xi})$ che rappresentano, rispettivamente, il peso della componente incertezza e la misura di feeling

$$\begin{cases} \text{logit}\,(1-\pi_i) = -\beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_p x_{ip} \\ \text{logit}\,(1-\xi_i) = -\gamma_0 - \gamma_1 w_{i1} - \gamma_2 w_{i2} - \ldots - \gamma_q w_{iq} \end{cases} \tag{21}$$

É possibile esplicitare il modello CUB senza covariate quando si vuole descrivere completamente la distribuzione con un livello di feeling ed incertezza

$$\Pr(R=r) = \pi\left[\binom{m-1}{r-1}(1-\xi)^{r-1}\xi^{m-r}\right] + (1-\pi)\left[\frac{1}{m}\right] \tag{22}$$

$$\forall r = 1,2,\ldots,m, \forall i = 1,2,\ldots,n$$

dove $\boldsymbol{\pi}$ e $\boldsymbol{xi}$ sono le medie dei parametri dei soggetti.
Si noti che il valore atteso dipende sia da $\boldsymbol{\pi}$ che da $\boldsymbol{xi}$:

$$\mathbb{E}(R) = \frac{m+1}{2} + (m-1)\pi\left(\frac{1}{2} - \xi\right) \tag{23}$$

4.3. Inferenza e verifica del modello

L'approccio inferenziale per i cub si basa sul *metodo della massima verosimiglianza*, che ci assicura stime e test dei parametri asintoticamente efficienti.

$$\log L(\pi, \xi) = \sum_{r=1}^{m} n_r \log\left(p_r(\pi, \xi)\right) \tag{24}$$

essendo

$$p_r(\pi, \xi) = \Pr(R = r \mid \pi, \xi)$$

Tuttavia, la tipologia di modello adottato (essendo una mistura) permette di sfruttare l'algoritmo E-M per poter ottenere una convergenza sicura.
Per operare confronti tra il modello CUB senza covariate e modello che presenta le covariate significative (quindi modelli annidati), si inizia stimando il modello senza covariate **(22)**.
Successivamente all'aggiunta di tutte le variabili, attraverso un approccio *backward*, si procederà alla rimozione, prima su una componente poi sull'altra, delle variabili non risultate significative.
Per tali confronti ci si avvale del Test sul rapporto di verosimiglianza, dove come riferimento teorico viene confrontato il modello nullo con il modello saturo.

$$\Delta Dev = 2\ell_S - \ell_{00} \tag{25}$$

dove ℓ_S rappresenta la *log-verosimiglianza* del modello saturo (modello completo, ovvero utilizzando il maggior numero di parametri) ed ℓ_{00} indica la *log-verosimiglianza* del modello nullo.
Di maggiore interesse, è confrontare due modelli che differiscono tra di

loro per delle covariate o per un effetto interazione.
Si vuole confrontare un modello M_1 con un modello M_2 (M_2 annidato in M_1).
Così come illustrato nei Proportional Odds Models (**14**), si vuole sottoporre a test l'ipotesi che i parametri relativi all'uncertanty e al feeling siano pari a 0, contro l'alternativa che ce ne sia almeno uno statisticamento diverso.
Il confronto in termini di bontà di adattamento avviene attraverso le devianze

$$\Delta Dev = 2Dev_{M_1} - Dev_{M_2} \tag{26}$$

con Dev_{M_1} e Dev_{M_2} rispettivamente la devianza del modello M_1 e M_2, da cui

$$\Delta Dev = (\ell_{M_1} - \ell_{00}) - (\ell_{M_2} - \ell_{00}) \tag{27}$$

Tale valore si confronta con una statistica Test Chi-uadrato $\chi^2_{M_1-M_2}$ (gradi di libertà pari alla differenza tra il modello $M_!$ e quello annidato)

$$\Delta Dev \sim \chi^2_{M_1-M_2} \tag{28}$$

Rifiuto il modello M_2 se $\Delta Dev > \chi^2_{M_1-M_2}$; non rifiuterò invece il modello annidato se ΔDev non risulterà significativo.
Oltre al criterio basato sulla verosimiglianza, è possibile utilizzare un diverso indice globale di fitting

$$\text{Diss } = \frac{1}{2}\sum_{r=1}^{m}\left|f_r - p_r(\hat{\boldsymbol{\theta}})\right| \tag{29}$$

Tale indice è chiamato **Indice di dissimilarità**, [Diss $\in$ (0,1)] e rappresenta un confronto tra frequenze osservate e frequenza teoriche.
Diss < 0.08, ci suggerisce un buon adattamento tra frequenze osservate e teoriche; per contro, un indice più alto di quella soglia denota un cattivo adattamento.

4.4. Modello CUB con effetto shelter

Una diversa formulazione dei modelli CUB adottati è rappresentato dal cd. modello CUB con *effetto shelter*.
Tale modello viene adottato nel caso in cui sia presente una forte asimmetria nella distribuzione a causa di un eccesso di risposte in una determinata categoria c, categoria in cui i rispondenti si rifugiano (da qui la parola shelter).
Ciò può accadere per diversi motivi: dalla noia, dal desiderio di privacy, dalla formulazione errata della domanda, dalla numerazione della scala e così via.
Questo modello rappresenta una generalizzazione del modello CUB, poiché presenta gli stessi parametri $\boldsymbol{\pi}$ e $\boldsymbol{\xi}$, più un parametro $\boldsymbol{\delta}$ che misura l'effetto dello shelter

$$\Pr(R = r \mid \boldsymbol{\theta}) = \delta \left[D_r^{(s)}\right] + (1-\delta)\left[\pi^* b_r(\xi) + (1-\pi^*)\frac{1}{m}\right] \qquad (30)$$
$$\forall r = 1, 2, \ldots, m$$

dove

$$D_r^{(c)} = \begin{cases} 1, \text{ se } r = c \\ 0, \text{ altrimenti} \end{cases} \qquad r = 1, 2, \ldots, m \qquad (31)$$

$\boldsymbol{D_r^{(c)}}$ è una variabile degenere che assume valore 1 se la massa è osservata nella categoria in cui si conserva l'eccesso di frequenza, 0 se si osserva nelle restanti categorie.
I modelli CUB sono innestati nei modelli CUB in cui si osserva la presenza shelter. Infatti, quando δ_i = 0 si ritorna alla formulazione standard del CUB **(18)**
Se il modello implica la presenza di possibili covariate si avrà:

$$\pi_i(\boldsymbol{\beta}) = \frac{1}{1+e^{-\boldsymbol{y}_i\boldsymbol{\beta}}}; \qquad \xi_i(\boldsymbol{\gamma}) = \frac{1}{1+e^{-\boldsymbol{w}_i\gamma}}; \qquad \delta_i(\boldsymbol{\omega}) = \frac{1}{1+e^{-\boldsymbol{x}_i\boldsymbol{\omega}}} \qquad (32)$$

5. Analisi Empirica

La variabile di risposta Machine Learning che assume 6 categorie di risposta (dove 0 esprime la non conoscenza e 5 la perfetta conoscenza) è così rappresentata:

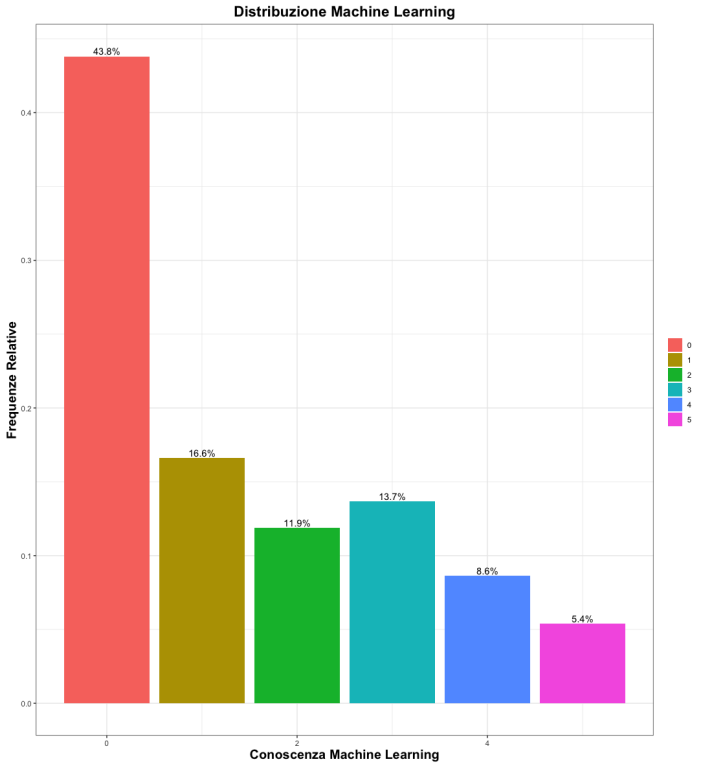

Figure 1. Distribuzione Machine Learning

Dal grafico in Figura (1) emerge che meno della metà dei rispondenti (43,8%) non ha conoscenza sul Machine Learing. Tale percentuale aumenta notevolmente, se si considera anche una conoscenza precaria del fenomeno oggetto di studio.
Infatti, se si osservano anche la prima e la seconda categoria di risposta, la percentuale supera la soglia del **70%**, ovvero 7 rispondenti su 10 detengono una competenza in ambito Machine Learning piuttosto bassa. Solo

poco più del **5%** dispone di una perfetta conoscenza.
Tuttavia, si è ritenuto opportuno osservare il comportamento della variabile dipendente rispetto al genere del rispondente ed emerge quanto segue:

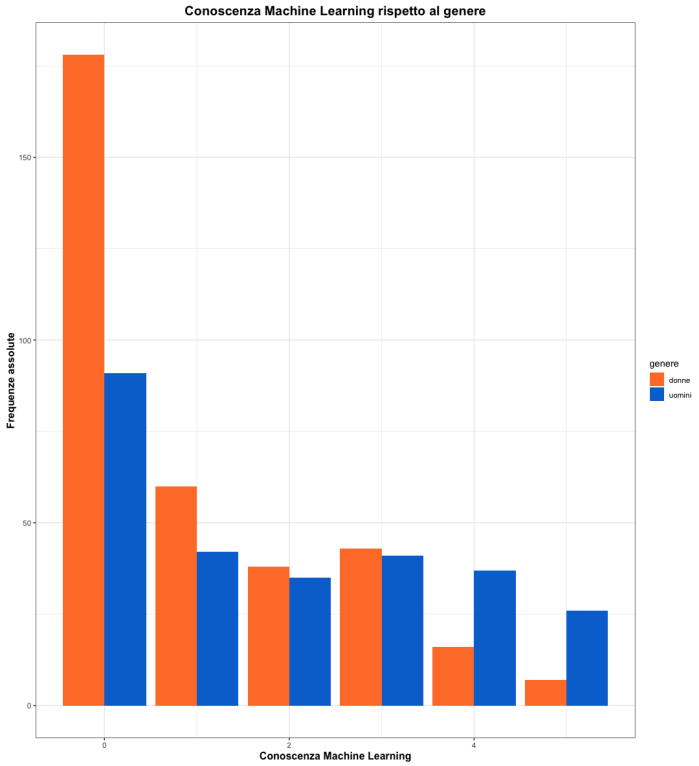

Figure 2. Distribuzione Machine Learning per genere

I rispondenti di genere femminile tendono ad avere una conoscenza più superficiale rispetto agli uomini, e ciò si può evincere dal grafico in prossimità della categoria 0, in cui circa il **60%** ritene di non conoscere affatto il Machine Learning rispetto a circa il 35% degli uomini, così come risulta dalla Figura (2).

La medesima analisi è stata svolta in Figura (3) rispetto all'età dei singoli rispondenti, attraverso l'ausilio del *Raincloud Plot*, che rappresenta

uno strumento di visualizzazione dei dati intuitivo e robusto.
Da tale rappresentazione grafica, si può desumere come la popolazione più anziana rispetto alla media di riferimento (circa 24 anni) possegga una conoscenza più bassa, se non assente, sul Machine Learning (così come evidenziabile dalle code presenti lungo la categoria 0 ed 1).
Tuttavia, come emerge dalla nube dei punti (che rappresentano le osservazioni), i rispondenti con età maggiore di 30 rappresentano meno del 10% del campione osservato, per cui risulta difficile generalizzare quanto affermato pocanzi.

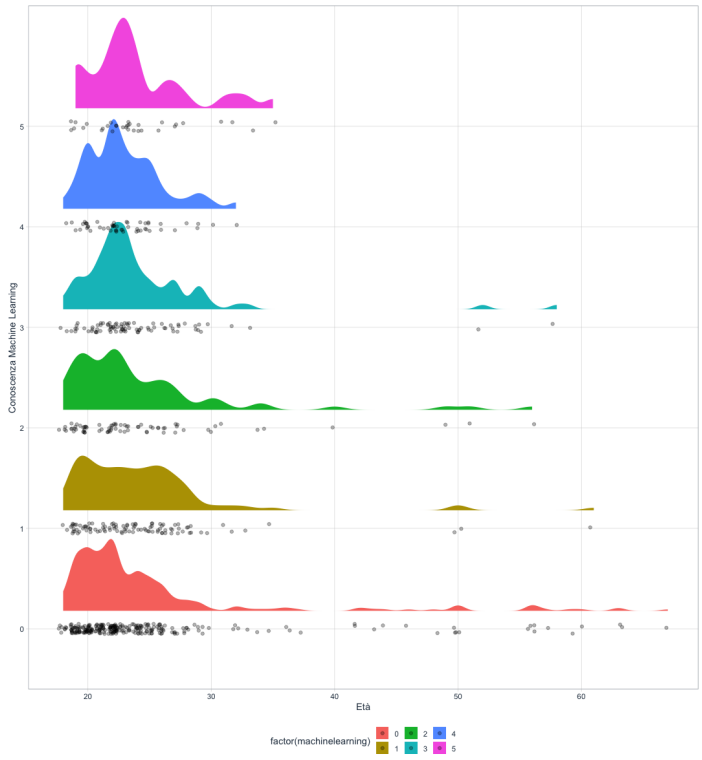

Figure 3. Raincloud plot di Machine Learning ed età

Infine, è stata condotta un'analisi esplorativa per comprendere la relazione tra la variabile dipendente e la conoscenza del Cloudcomputing, ovvero

la variabile di tipo ordinale considerata, attraverso un plot a mosaico (particolarmente utile per il confronto tra gruppi).
Nel plot a mosaico in Figura (4), la larghezza delle colonne è proporzionale al numero di osservazioni per ciascun livello della variabile inserita sull'asse delle ascisse, mentre l'altezza delle barre in verticale è proporzionale al numero di osservazioni della seconda variabile per ciascun livello della prima variabile.

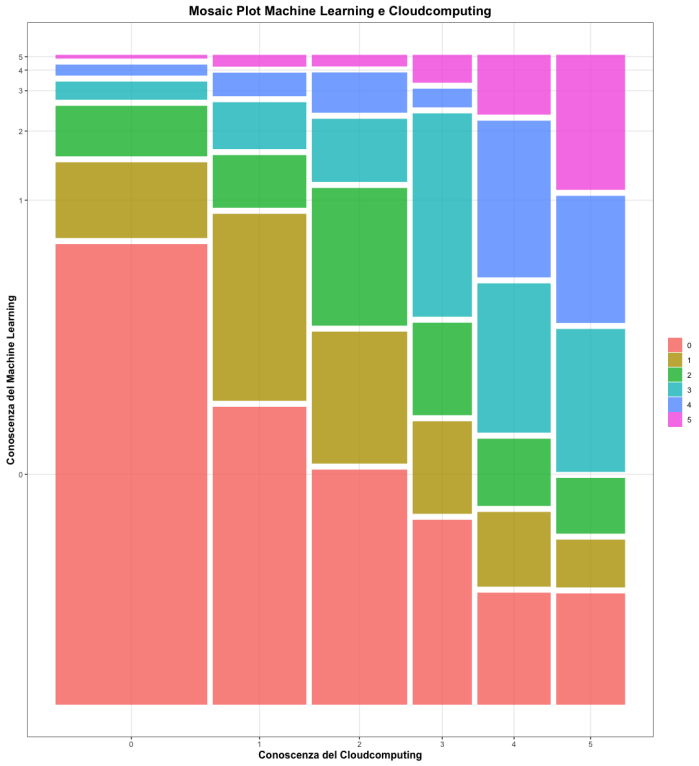

Figure 4. Plot a mosaico tra Machine Learning e Cloudcomputing

Si noti come la non conoscenza del Machine Learning sia spesso associata alla non conoscenza del Cloudcomputing. Ciò lo si può osservare in corrispondenza del valore 0 dove la barra risulta più alta rispetto alle altre. Anche in prossima della conoscenza perfetta, emerge come le due

variabili siano tra loro particolarmente collegate.

Prima di procedere alla stima del modello POM, è stato opportuno individuare un criterio di selezione della variabili attraverso la correlazione per ranghi di Spearman (Figura 5) utilizzata per valutare la relazione per variabili di tipo qualitativo ordinale.
Per questa analisi, ci si è avvalsi dell'uso di 2 batterie di domande del questionario sempre con riferimento alla conoscenza/non conoscenza au determinate materie

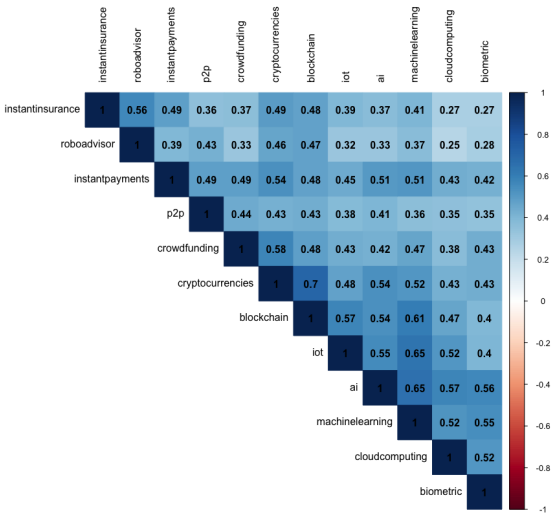

Figure 5. Matrice di correlazione

Tale indice, così come l'indice di correlazione di Pearson, varia – 1 e + 1

- Valori vicini a +1 indicano una perfetta relazione positiva tra i ranghi

- Valori prossimi allo 0 indicano l'assenza di relazione tra i ranghi
- Valori vicini -1 identificano una perfetta relazione inversa tra i ranghi

In questo caso studio, la correlazione tra Machine Learning e le altre variabili di tipo ordinale è molto alta, e ciò era prevedibile anche dalla Figura (4) vista precedentemente.
Ai fini della successiva stima del modello, è stato ritenuto opportuno considerare la variabile Cloudcomputing che risultava particolarmente correlata, sia in termini statistici (indice pari a **0.52**) che in quelli prettamente applicati.
I *p-value* che ci restiuiscono i ranghi sono tutti molto bassi (*p-value* < 2.2e-16), il quale portano al rifiuto dell'ipotesi H_0 che il coefficiente ρ sia pari a 0. Ciò a conferma dell'alta correlazione tra le variabili.
Per misurare, invece, la forza del legame tra il Machine Learning e le variabili di tipo qualitativo non ordinali, ci si serve dello studio sull'associazione attraverso l'indice normalizzato di Cramer che varia tra 0 e 1.

- Un indice prossimo a 0 denota una scarsa associazione tra le variabili
- Un indice vicino a 1 mostra una perfetta associazione tra le variabili

Si osserva un indice V di Cramer moderatamente apprezzabile tra genere e Machine Learning (≈ 0.27) e tra ML e area in cui il rispondente studia o ha studiato (≈ 0.32) .
Tale associazione è confermata anche dal Test del Chi-quadrato, che mette a confronto le seguenti due ipotesi:

- Ipotesi nulla H_0 di indipendenza fra le due variabili
- Ipotesi alternativa H_1 di associazione fra le variabili

Considerato un livello di significatività α pari a 0.05, si rifiuterà l'ipotesi nulla H_0 di indipendenza se il valore osservato sarà maggiore del percentile di una vc Chi-quadrato con $(n-1)(k-1)$ gradi di liberta, dove n e k sono il numero di classi delle due variabili.
Osservato che il valore del χ^2 tra Machine Learning e genere è pari 43.328

e il percentile della vc chi quadrato con 5 gradi di libertà è 11.070, si rifiuta l'ipotesi nulla di indipendenza con un *p-value* pari a $3.171e-08$. Medesima applicazione è stata effettuata tra Machine Learning e area studio, ed anche in questo caso, vi è stato il rifiuto di H_0 con una statistica test pari a 60.585, percentile della vc χ^2 con 5 gradi di libertà 11.070 e *p-value* di $9.201e-12$.
Infine, per misurare il legame con l'unica variabile quantitativa continua, l'età, è stato utilizzato il coefficiente di correlazione di Pearson che ha evidenziato un legame esiguo tra Machine Learning ed età. Ciò in riferimento al problema legato alla numerosità del campione per età maggiori di 30 già discusso nella Figura (3)

5.1. Proportional Odds Model

Il primo modello implementato è il Proportional Odds Model (POM). Dalla formula specificata in **(4)** si stima così il modello con tutte le covariate, che rappresenta il cd. modello saturo. A partire da esso, si procede nuovamente alla stima escludendo le variabili che inizialmente sono risultate non significative.
Il modello che così si ottiene è il seguente:

$$\begin{aligned} Pr\left(Y_i \leq j\right) = G\left(\hat{\alpha}_j - 0.8258Ge_i - 0.5937Stu_i \right. \\ \left. +0.7663Area_i - 0.7227Sett_i - 0.6534Cloud_i\right) \end{aligned} \tag{33}$$

dove Ge_j rappresenta la variabile dicotomica che identifica il genere, Stu_j il il titolo di studio, $Area_i$ l'area disciplinare, $Sett_i$ il settore lavorativo e $Cloud_i$ il livello di conoscenza del Cloudcomputing.
I *thresholds* $\hat{\alpha}_j$ stimati sono

$$\alpha_1 = 1.2377 \quad \alpha_2 = 2.1986 \quad \alpha_3 = 2.9883 \quad \alpha_4 = 4.2004 \quad \alpha_5 = 5.5179$$

che risultano statisticamente significativi.
Per verificare che le covariate inserite nel modello saturo siano effettivamente non significative (e che quindi il modello migliore sia quello con meno parametri) è stato operato un Test sul rapporto di verosimiglianza

(così come illustrato **(14)**), in cui è stato confrontato il modello saturo con il modello specificato in **(33)**.
Indicando con ℓ_{M_2} la *log-verosimiglianza* del modello annidato e con ℓ_{M_1} la *log-verosimiglianza* del modello che annida, si ricava

$$\Delta Dev = 2(\ell_{M_1} - \ell_{M_2})$$

Tale valore va confrontato con il percentile della vc χ^2 con gradi di libertà pari alla differenza tra i parametri del modello saturo e quello annidato. Il valore della statistica χ^2 è 7.3475, con un *p-value* pari a 0.3936.
Dato che $\Delta Dev < \chi^2_{(0.05,7)}$ non rifiutiamo l'ipotesi nulla e quindi il modello da utilizzare sarà M_2.

Per verificare l'assunzione di proporzionalità è possibile stimare diversi test per la bontà di adattamento. In questo elaborato è stato stimato:

- Il test di Brant, in cui nell'ipotesi H_0 viene specificata la corretta assunzione di proporzionalità; per converso, l'ipotesi H_1 riflette la situazione in cui la proportional assumption non viene rispettata.

Variabili	X^2	Gradi di libertà	p-value
Omnibus	25.19	20	0.19
Genere	4.66	4	0.32
Studio	1.78	4	0.78
Area Studio	5.23	4	0.26
Settore Lavoro	6.95	4	0.14
Cloudcomputing	5.25	4	0.26

Il test, che viene confrontato con una vc Chi-quadrato con gradi di libertà pari a $(k-2)p$, conduce per tutte le covariate, compresa quella *Omnibus* che contempla l'assunzione di proporzionalità del modello complessivo, al non rifiuto dell'ipotesi nulla. Inoltre, il *p-value* risulta sensibilmente maggiore di 0.05 per tutte le variabili esplicative.

- Un secondo test stimato è quello di *Lipsitz*. Tale test ci suggerisce di non rifiutare l'ipotesi nulla H_0 di adattamento perfetto (*p-value* è superiore alla soglia del 5%, ovvero $0.5829 > 0.050$).
  Ciò significa che, come confermato anche dal test di Brant, il modello adottato soddisfa l'ipotesi di assunzione di proorzionalità.

- Infine, anche Il test di *Hosmer* e *Lemeshow* ci suggerisce di non rifiutare l'ipotesi nulla H_0 di corretta specificazione del modello.
  Sempre considerando la soglia del *p-value* pari allo 0.05, in questo test rileviamo un *p-value* pari a 0.0864, che ci conduce appunto a non rifiutare l'ipotesi nulla.

Dal modello sopra ottenuto, è possibile interpretare le stime per la categoria j-esima con l'ausilio di *odds*, *odds ratio* ed effetti marginali medi.
Grazie agli *odds*, è possibile ottenere una prima ed immediata interpretazione grazie ai segni di parametri stimati.
Ne deriva che il logit della probabilità di conoscenza verso il Machine Learning:

- Aumenta per gli uomini

- Aumenta per coloro che possiedono una laurea

- Aumenta se il settore lavorativo è quello appartenente alle categorie STEM

- Aumenta per ogni incremento unitario sulla scala relativa alla conoscenza del Cloudcomputing

- Diminuisce se l'area disciplinare in cui l'individuo studia o ha studiato è quella umanistica.

Le interpretazioni effettuate sono ottenute fermo restando l'impatto delle altre covariate.
Per calcolare l'impatto, dal punto di vista quantitativo, delle covariate sulla dipendente, si ricorre agli *odds ratio* $e^{(\beta_j)}$

	OR
Genere	-0.56
Studio	-0.45
Area Studio	1.15
Settore Lavoro	-0.51
Cloudcomputing	-0.48

- L'essere uomo piuttosto che donna riduce gli odds in favore delle prime categorie del 56%
- L'essere laureato piuttosto che diplomato riduce gli odds in favore delle prime categorie del 45%
- L'aver studiato in un'area umanistica piuttosto che in un'area scientifica aumenta gli odds del 115% di ritrovarsi nelle prime categorie
- Il lavorare in un settore STEM piuttosto che in un altro settore diminuisce gli odds del 51% di collocarsi nelle prime categorie
- Un punto in più sulla scala della conoscenza del Cloudcomputing diminuisce gli odds del 48% in favore delle prime categorie

Come per gli odds, le interpretazioni effettuate sono ottenute fermo restando l'impatto delle altre covariate.
Per quanto riguardo gli effetti marginale medi (AME) in riferimento alla prima categoria di risposta

	AME
Genere	-0.146
Studio	-0.104
Area Studio	0.137
Settore Lavoro	-0.124
Cloudcomputing	-0.112

ne deriva quanto segue:
La probabilità media di selezionare la prima categoria di risposta diminuisce

del 14.6% per gli uomini rispetto alle donne, diminuisce del 10.4% per i laureati rispetto ai diplomati, aumenta del 13.7% per coloro che provengono da area umanistica piuttosto che scientifica, diminuisce del 12.4% per i rispondenti che lavorano in discipline STEM piuttosto che non STEM e diminuisce del 11.2% per ogni incremento unitario della variabile Cloud-computing.

Per misurare la bontà di adattamento del modello ai dati, è stato utilizzato lo Pseudo R2 di McFadden, che risulta pari a 0.15, ed è un valore soddisfacente dato che si è vicini alla situazione di addattamento ottimale (*0.20-0.40*). Ciò è confermato anche dall'indice R2 di Mckelvey e Zavoina, pari a *0.40*.
In aggiunta, è stata calcolato l'indice di corretta capacità previsiva attraverso la matrice di confusione per valutare le prestazioni del modello. Tale indice confronta i valori predetti dal modello e quelli osservati

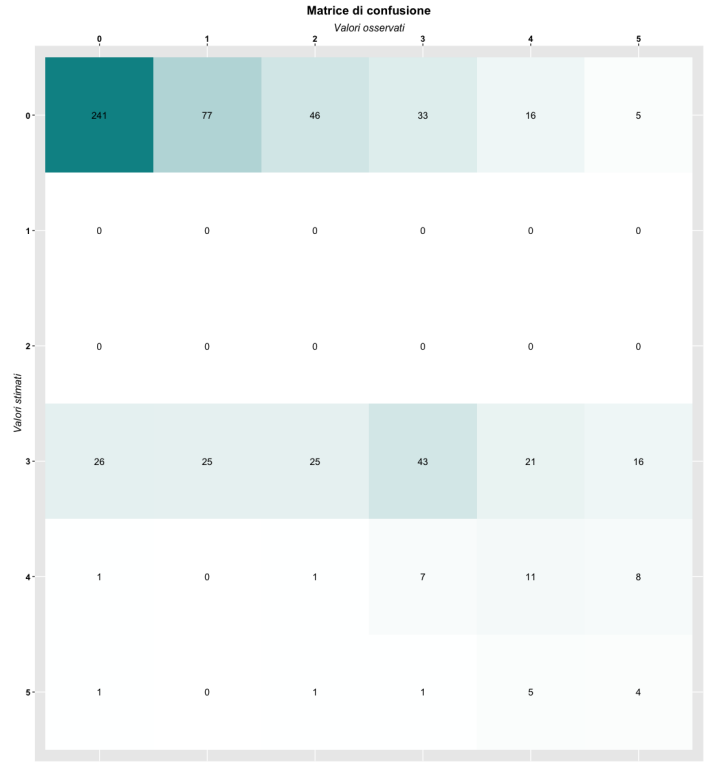

Figure 6. Matrice di confusione

L'indice ci restituisce un valore di 0.49, ovvero il modello riesce a stimare correttamente circa la metà delle osservazioni. Come si evince dalla Figura (6), il modello non riesce a cogliere le risposte sulla categoria di risposta 1 e 2; in compenso, riesce a predire correttamente circa il **60%** delle risposte sulla categoria 0 che denota la scarsa conoscenza.

Di seguito sono stati analizzati i seguenti profili.
A parità di tutte le altre covariate, sono stati osservati 4 profili facendo variare il genere ed il settore lavoro. Si considera quindi:

- L'individuo di sesso maschile con settore lavorativo STEM
- L'individuo di sesso maschile con settore lavorativo in discipline

non STEM

- L'individuo di sesso femminile con settore lavorativo STEM
- L'individuo di sesso femminile con settore lavorativo in discipline non STEM

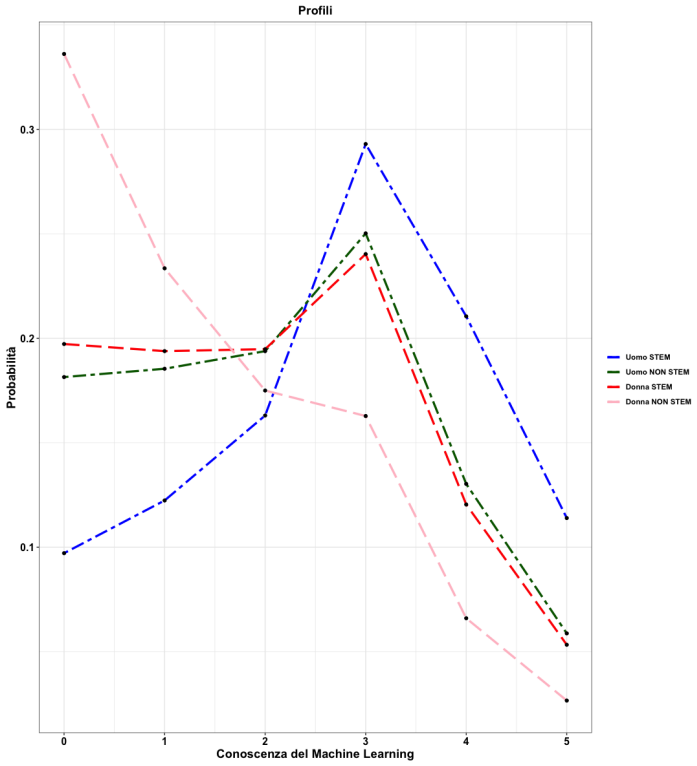

Figure 7. Profili POM

Dalla Figura (7) emerge che, a prescindere da quale sia il settore lavorativo relativo al rispondente, per gli uomini si rileva una conoscenza maggiore riguardo il Machine Learning rispetto alle donne e ciò lo si può osservare dalle prime e dalle ultime categorie (nella categoria di risposta 5 emerge una sottile differenza tra l'uomo con settore lavorativo non in discipline STEM e la donna che lavora in un settore STEM).

Inoltre, per quanto riguarda la prima categoria di risposta, ovvero la non conoscenza del fenomeno di riferimento, emerge che più del **30%** del donne appartenenti al settore lavorativo non STEM non conosca il Machine Learning. Tale interpretazion è stata effettuata a parità delle altre covariate presenti nel modello, ovvero avendo imposto il conseguimento di una laurea, lo studio o l'aver studiato in un'area scientifica ed una risposta sulla conoscenza del Cloudcomputing pari al valor medio.

5.2. CUB

Come alternativa ai modelli cumalativi, è stato utilizzato un approccio mediante i modelli CUB. Inizialmente, è stato stimato un modello senza covariate, sia sulla categoria del *feeling* che su quella dell'*uncertanty*.
Il modello restituisce i valori complessivi di π e ξ, che sono

$$\pi = 0.42318 \qquad\qquad \xi = 0.95735$$

Dalla Figura (8), si osserva la presenza di una misura di interesse, soddisfazione $(\mathbf{1 - \xi})$ piuttosto bassa, nonostante il livello di incertezza $(\mathbf{1 - \pi})$ non sia particolarmente elevato. Inoltre, si rileva un indice di dissimilarità soddisfacente pari a 0.058.
Dopodichè, è stata ottenuta una nuova stima del modello inserendo solo le covariate prima sulla componente di Feeling poi su quella di uncertanty ed eliminando, ad ogni step, quelle che non risultavano significative.

Tuttavia, prima di procedere alla specificazione dei profili, si può verificare se, a seguito dell'inclusione/esclusione di covariate nel modello, il test sul rapporto della verosimiglianza ci consenta di rifiutare l'ipotesi in cui i parametri siano tutti statisticamente pari a 0.
Si confronta inizialmente il modello stimato con le covariate solo sul feeling. Si considera con ℓ_{feel} la *log-verosimiglianza* con tutte le covariate sul feeling e con ℓ_{feel_1} la *log-verosimiglianza* con le sole covariate significative e si osserva che

$$\ell_{feel} = -823.1025 \qquad\qquad \ell_{feel_1} = -824.312$$

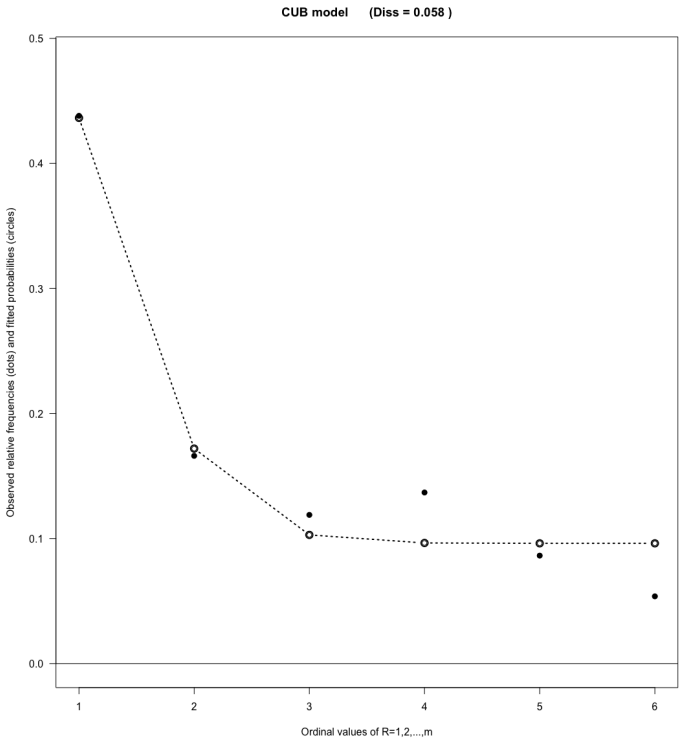

Figure 8. Modello CUB

La statistica χ^2 risulta pari a $2(\ell_{feel} - \ell_{feel_1})$
Essa si distribuisce come una vc Chi-quadrato con 14 - 10 gradi di libertà $\chi^2_{(0.05,4)}$, ed il valore sarà pari a 9.488.
Non si rifiuta l'ipotesi nulla H_0; risulterà quindi preferibile il modello più parsimonioso (le 4 covariate erano statisticamente uguali a 0).
Medesima verifica è stata effettuata con le covariate presenti solo sull'incertezza, dove si considera ℓ_{unc} il modello con tutte le covariate sulla componente di incertezza ed ℓ_{unc_1} il modello con solo le covariate significative.

$$\ell_{unc} = -853.4125 \qquad \ell_{unc_1} = -858.3396$$

Anche in questo caso non si rifiuta l'ipotesi H_0 e si preferisce il modello con meno variabili, $\Delta Dev < \chi^2_{(0.05,7)}$ $(10 < 14.067)$.

È stato così stimato il modello completo, che prevede l'inserimento sia di variabili sull'*uncertanty* che sul *feeling*. Procedendo alla stima del modello, il genere, il titolo di studio e il settore lavorativo sono risultate non significative. E' stata così effettuata un'altra stima del modello senza queste covariate.
Il confronto con il LRT fa propendere verso l'utilizzo del modello più parsimonioso, dal momento che $\Delta Dev < \chi^2_{(0.05,3)}$, ossia $(0.4 < 7.815)$. Tale modello, quindi, rappresenta il modello finale.
Considerando il modello finale in cui sono presenti solo le covariate significative su entrambe le componenti, si nota che:

- Il livello di indecisione nelle risposte diminuisce se il soggetto rispondente studia o ha studiato in area scientifica e aumenta per ogni livello (sulla scala ordinale) di conoscenza in più in cloudcomputing.

- Il livello di soddisfazione aumenta se si è uomini, se si ha residenza nel nord e nel centro italia, se si è laureati, se si studia in area scientifica, se si è iscritti presso l'Università degli studi di Napoli Federico II, se il settore in cui si lavora è quello STEM e aumenta per ogni livello (sulla scala ordinale) di conoscenza in più in Cloudcomputing.

In riferimento ai profili considerati nel Proportional Odds Models, sono stati costruiti anche in questo caso i medesimi profili per poter effettuare un confronto.
In Figura (9), le considerazioni sono le stesse effettuate per i POM, dove emerge sostanzialmente una netta differenza tra uomini e donne.
Le valutazioni fatte sono state compiute per un individuo laureato, residente nel Sud Italia, che studia o ha studiato nell'area scientifica, frequentante l'Università degli Studi di Napoli Federico II, che lavora in un'area STEM e che possiede una conoscenza media riguardo il Cloudcomputing.

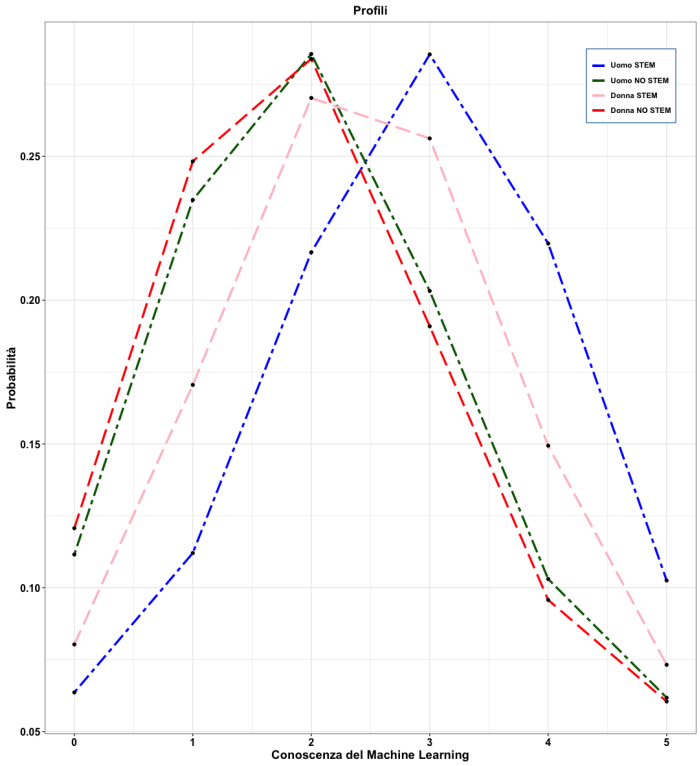

Figure 9. Profili CUB

5.3. CUB con shelter

Visto l'accentramento delle risposte sulla categoria 0, è stato stimato anche il modello cub con l'effetto *shelter*:

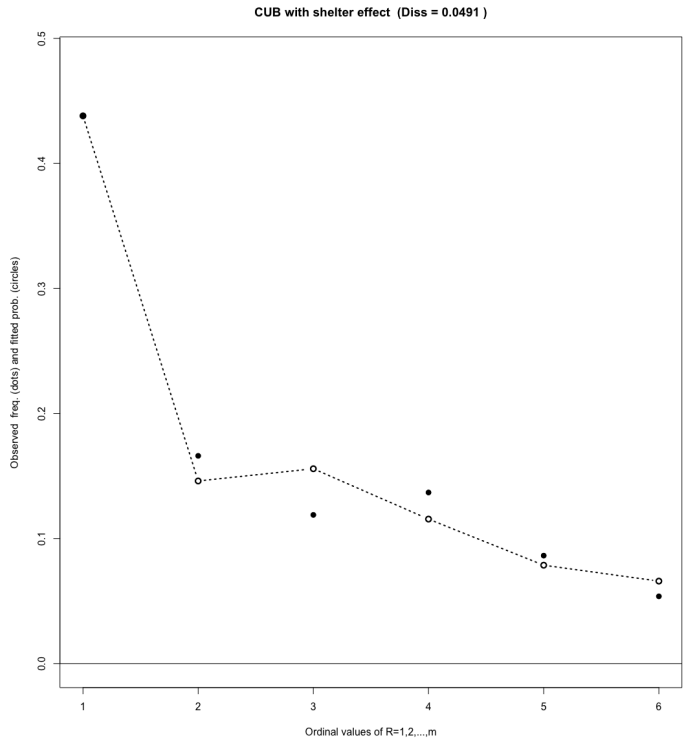

Figure 10. Modello CUB con shelter

Rispetto al CUB, è presente un migliore adattamento tra frequenze osservate e quelle teoriche come risulta dalla Figura (10).
L'indice di dissimilarità infatti risulta più piccolo (0.0491 rispetto a 0.058 del CUB).
Il livello dell'incertezza è inoltre più contenuto ed aumenta sostanzialmente la misura di soddisfazione, con la differenza che in questo modello è presente un parametro aggiuntivo δ che rileva l'accentramento sulla prima categoria

$$\pi = 0.4112866 \qquad \xi = 0.6411016 \qquad \delta = 0.3445616$$

Per poter comprendere l'impatto che ha lo shelter sul modello CUB, è stata utilizzata l'intera batteria di variabili dipendenti che riguarda la

conoscenza nei relativi ambiti attraverso:

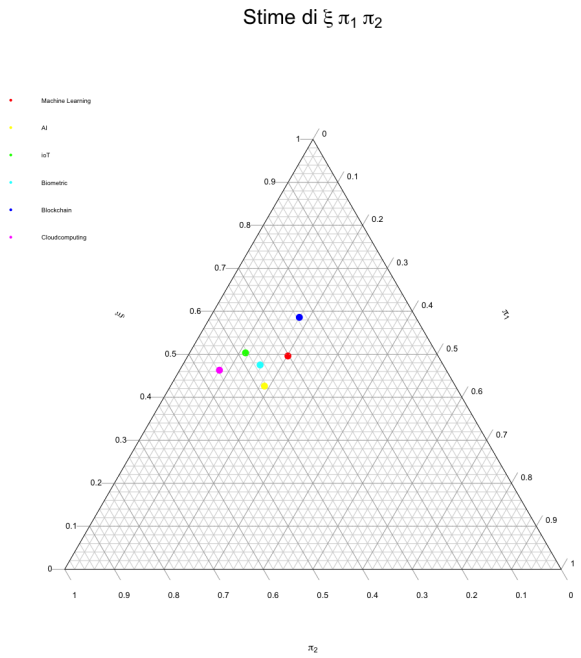

Figure 11. Plot effetto shelter

Con la seguente Figura (11), che presenta le stime di $\boldsymbol{\pi_1}$, $\boldsymbol{\pi_2}$ e $\boldsymbol{\delta}$, emerge che tutte le variabili presentano un parametro δ molto alto: ciò riflette la probabilità che anche nelle altre domande i rispondenti si rifugino nella categoria 0. Inoltre, si nota come tutte le domande presenti all'interno della batteria presentino anche un livello di incertezza piuttosto elevato. Graficamente, l'impatto dello shelter è dimostrato con il grafico che segue

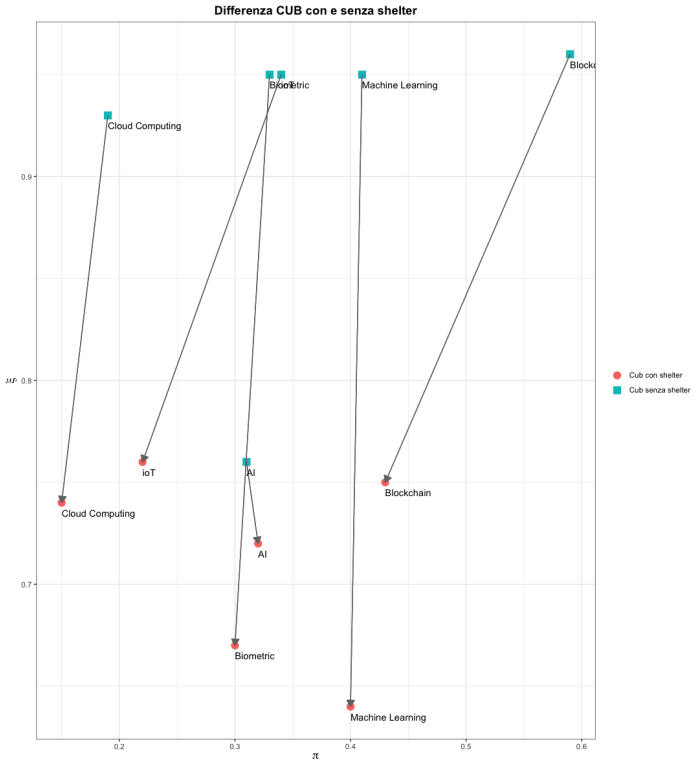

Figure 12. Plot differenza cub e cub con shelter

Dalla Figura (12), si evince come la presenza dello shelter riduca $\boldsymbol{\xi}$, e di conseguenza faccia aumentare il livello di soddisfazione $\mathbf{1} - \boldsymbol{\xi}$. Inoltre, si osserva anche una discreta riduzione del parametro $\boldsymbol{\pi}$ per quasi tutte le variabili considerate.

Per operare un confronto, dato che il CUB con shelter differisce con il CUB solo per la variabile studio, è stato utilizzato il LRT, dove ℓ_{cub} è la *log-verosimiglianza* del modello CUB e ℓ_{she} è *log-verosimiglianza* del modello CUB con shelter.
La statistica test ΔDev sarà pari a 13.41 ed è confrontata con il valore della $\chi^2_{(0.05,2)}$ (pari a 5.991); si rifiuterà quindi l'ipotesi H_0 e il modello

migliore sarà quello meno parsimonioso, ovvero il modelo CUB con lo shelter effect.
Infine, sono stati individuati gli stessi profili costruiti in precedenza, dove sostanzialmente si ricavano le stesse considerazioni, ma l'accentramento sulla categoria 0 è diminuito rispetto al CUB con la presenza della covariata studio, così come si osserva in Figura (13)

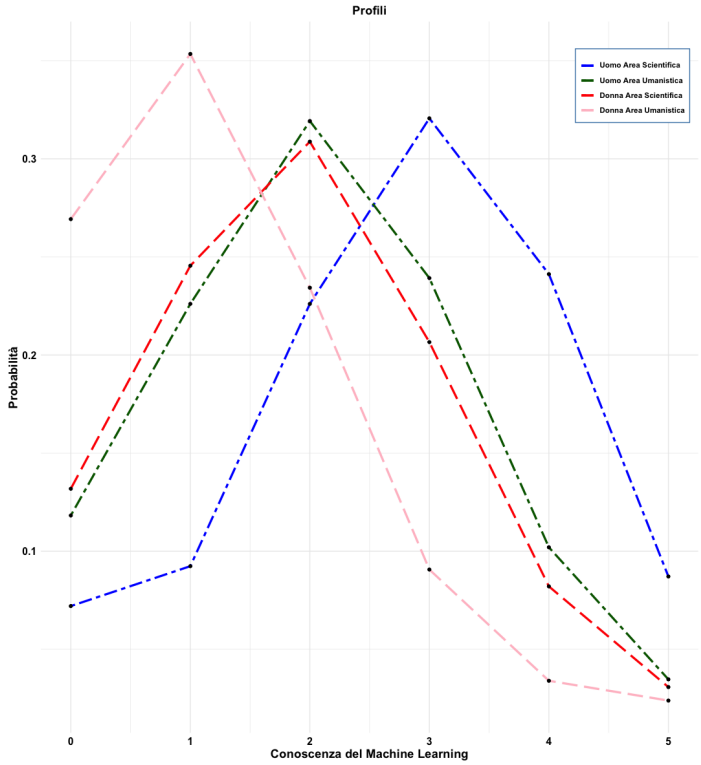

Figure 13. Profili CUB con shelter

6. Conclusioni

Alla fine dell'analisi, il modello più idoneo all'obiettivo che si vuole perseguire risulta il Proportional Odds Model, che presenta un indice **BIC** più piccolo rispetto a quello del CUB con shelter (*1678.49 < 1705.84*).
Tuttavia, il modello CUB è riuscito a cogliere quelle che sono le motivazioni che si celano dietro al processo decisionale, a tutto quel meccanismo che porta un soggetto alla scelta di una domanda piuttosto che un'altra.
La lunghezza, la complessità e la tipologia del questionario stesso probabilmente hanno accentuato l'accentramento sulla categoria 0, concentrazione che è stata giustificata dal titolo di studio, dove è emerso che i rispondenti meno istruiti tendono a *rifugiarsi* con una probabilità più alta rispetto a coloro che possiedono (almeno) una laurea triennale.
In conclusione, in questo modello, così come negli altri modelli stimati, le considerazioni a cui si giunge sono pressochè le medesime: che si stimi un Proportional Odds Model, un CUB o un CUB con effetto shelter, sulla conoscenza del Machine Learning, ed in generale in materie attinenti al campo informatico/scientifico, emerge un considerevole *gender gap* e ciò a prescindere dal background conoscitivo e lavorativo del soggetto rispondente.

7. Riferimenti

Salvan Alessandra, Nicola Sartori, e Luigi Pace. Springer, Milano, 20 *"Modelli Lineari Generalizzati"*, Modelli Lineari Generalizzati. Springer, Milano, 2020.

Alan Agresti, *Categorical Data Analysis*, Vol792, John Wiley Sons, 2012.

Piccolo Domenico e Rosaria Simone, *The class of CUB models: statistical foundations, inferential issues and empirical evidence.*, Statistical Methods Applications 28.3 (2019): 389-435.

Iannario Maria, *Modelling shelter choices in a class of mixture models for ordinal responses*, Statistical Methods and Applications, 21(1) (2012), 1–22.

8. Appendice

```
Call:
polr(formula = as.factor(dataset$machinelearning) ~ genere +
    eta + componenti + macroarea + studio + areastudio + unina +
    lavoro + settorelavoro + benessereeconomico + cloudcomputing,
    data = dataset, method = "logistic")

Coefficients:
                             Value Std. Error t value
genereuomini              0.850724    0.16436  5.1761
eta                      -0.003797    0.01491 -0.2547
componenti                0.118840    0.10658  1.1150
macroareanord            -0.072964    0.29625 -0.2463
macroareasudisole        -0.460889    0.33583 -1.3724
studiolaurea              0.633874    0.16407  3.8635
areastudioareaumanistica -0.726853    0.17205 -4.2245
uninaunina                0.211236    0.24032  0.8790
lavorostudenti            0.077729    0.22815  0.3407
settorelavoroSTEM         0.720279    0.25697  2.8029
benessereeconomico        0.020832    0.03513  0.5931
cloudcomputing            0.652543    0.05199 12.5524

Intercepts:
    Value   Std. Error t value
0|1  1.6231  0.7262     2.2350
1|2  2.5930  0.7316     3.5442
2|3  3.3866  0.7393     4.5810
3|4  4.6042  0.7531     6.1138
4|5  5.9276  0.7697     7.7013

Residual Deviance: 1606.943
AIC: 1640.943
```

```
Call:
polr(formula = as.factor(dataset$machinelearning) ~ genere +
    studio + areastudio + settorelavoro + cloudcomputing, data = dataset,
    method = "logistic")

Coefficients:
                           Value Std. Error t value
genereuomini              0.8258    0.16164   5.109
studiolaurea              0.5937    0.15928   3.728
areastudioareaumanistica -0.7663    0.16921  -4.529
settorelavoroSTEM         0.7227    0.25539   2.830
cloudcomputing            0.6534    0.05125  12.748

Intercepts:
    Value   Std. Error t value
0|1  1.2377  0.2010     6.1582
1|2  2.1986  0.2136    10.2927
2|3  2.9883  0.2302    12.9801
3|4  4.2004  0.2644    15.8883
4|5  5.5179  0.3152    17.5039

Residual Deviance: 1614.29
AIC: 1634.29
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 18  Maxiter= 500
=========================================================================
Uncertainty
    Estimates     StdErr     Wald
pai 0.7715766 0.03139267 24.57824
=========================================================================
Feeling
                             Estimates     StdErr        Wald
constant                    5.77152577 0.74381826   7.7593225
genereuomini               -0.91056833 0.16787644  -5.4240387
eta                        -0.00198313 0.01715517  -0.1155995
componenti                 -0.13447736 0.09939970  -1.3528950
macroarea1centro           -1.09345573 0.37991381  -2.8781679
macroarea1nord             -0.92032310 0.26405132  -3.4853948
studiolaurea               -0.62282050 0.17072639  -3.6480623
areastudio1areascientifica -1.02034163 0.20670438  -4.9362362
uninaunina                 -0.83407852 0.28038670  -2.9747436
lavorostudenti              0.12329386 0.23545849   0.5236331
settorelavoroSTEM          -0.82626477 0.21999750  -3.7557916
benessereeconomico         -0.01380873 0.03864712  -0.3573028
cloudcomputing             -0.72341876 0.05576110 -12.9735388
=========================================================================
Log-lik            = -823.1053
Mean Log-likelihood= -1.340562
-------------------------------------------------------------------------
AIC       = 1674.211
BIC       = 1736.09
ICOMP     = 1673.3
=========================================================================
Elapsed time= 1.041 seconds =====>>> Mon Jan 23 19:51:59 2023
=========================================================================
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 18  Maxiter= 500
=========================================================================
Uncertainty
    Estimates    StdErr     Wald
pai 0.7682975 0.0315658 24.33955
=========================================================================
Feeling
                            Estimates     StdErr       Wald
constant                    5.2361890 0.40100736  13.057588
genereuomini               -0.9275700 0.17013730  -5.451891
macroarea1centro           -1.0715847 0.37656752  -2.845664
macroarea1nord             -0.8824199 0.25739600  -3.428258
studiolaurea               -0.6146110 0.16401372  -3.747314
areastudio1areascientifica -1.0503039 0.21129766  -4.970731
uninaunina                 -0.7869776 0.27753991  -2.835547
settorelavoroSTEM          -0.8276534 0.22265377  -3.717221
cloudcomputing             -0.7202078 0.05278731 -13.643578
=========================================================================
Log-lik            = -824.312
Mean Log-likelihood= -1.342528
-------------------------------------------------------------------------
AIC       = 1668.624
BIC       = 1712.824
ICOMP     = 1660.825
=========================================================================
Elapsed time= 0.475 seconds =====>>> Mon Jan 23 19:56:40 2023
=========================================================================
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 51  Maxiter= 500
=========================================================================
Uncertainty
                              Estimates     StdErr       Wald
constant                     5.574979641 1.70804593  3.2639518
genereuomini                -1.308763318 0.41433622 -3.1586988
eta                         -0.006677337 0.03035622 -0.2199660
componenti                  -0.336500076 0.27191138 -1.2375358
macroarea1centro            -0.213344068 0.71873688 -0.2968319
macroarea1nord              -0.379422202 0.54473608 -0.6965248
studiolaurea                -1.530611063 0.42675907 -3.5865930
areastudio1areascientifica -0.915893594 0.38142509 -2.4012411
uninaunina                   0.200661923 0.52351094  0.3833003
lavorostudenti              -1.067300039 0.56842097 -1.8776577
settorelavoroSTEM           -3.546414740 1.69897178 -2.0873888
benessereeconomico           0.052759725 0.08488522  0.6215420
cloudcomputing              -1.134342185 0.18279798 -6.2054417
=========================================================================
Feeling
    Estimates      StdErr     Wald
csi 0.9583459 0.008683317 110.3663
=========================================================================
Log-lik            = -853.4216
Mean Log-likelihood= -1.389938
-------------------------------------------------------------------------
AIC       = 1734.843
BIC       = 1796.723
ICOMP     = 1740.226
=========================================================================
Elapsed time= 2.66 seconds =====>>> Mon Jan 23 20:00:58 2023
=========================================================================
```

```
==========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
==========================================================================
 m= 6  Sample size: n= 614  Iterations= 25  Maxiter= 500
==========================================================================
Uncertainty
                           Estimates    StdErr      Wald
constant                    3.524153 0.4794503  7.350404
genereuomini               -1.073788 0.3751829 -2.862038
studiolaurea               -1.265437 0.3751905 -3.372786
areastudio1areascientifica -1.149117 0.3640996 -3.156052
settorelavoroSTEM          -3.180881 1.5082656 -2.108966
cloudcomputing             -1.054314 0.1524221 -6.917067
==========================================================================
Feeling
    Estimates      StdErr     Wald
csi 0.9539793 0.008983318 106.1945
==========================================================================
Log-lik            = -858.3396
Mean Log-likelihood= -1.397947
--------------------------------------------------------------------------
AIC       = 1730.679
BIC       = 1761.619
ICOMP     = 1732.617
==========================================================================
Elapsed time= 0.431 seconds =====>>> Mon Jan 23 20:04:15 2023
==========================================================================
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 29  Maxiter= 500
=========================================================================
Uncertainty
                            Estimates    StdErr       Wald
constant                    1.5210164 0.3154288  4.8220600
genereuomini                0.1001958 0.3675790  0.2725830
studiolaurea               -0.1501753 0.3389078 -0.4431155
areastudio1areascientifica  0.9709807 0.4300563  2.2577991
settorelavoroSTEM          -0.2740979 0.7372719 -0.3717732
cloudcomputing             -0.3942174 0.1039014 -3.7941497
=========================================================================
Feeling
                            Estimates     StdErr       Wald
constant                    5.4190745 0.41875769  12.940836
genereuomini               -0.9130039 0.18401712  -4.961516
macroarea1centro           -1.0984778 0.38240241  -2.872570
macroarea1nord             -0.9922988 0.25951012  -3.823739
studiolaurea               -0.5627894 0.17850238  -3.152840
areastudio1areascientifica -1.2807491 0.29200240  -4.386091
uninaunina                 -0.9773352 0.28038099  -3.485740
settorelavoroSTEM          -0.7777593 0.22722984  -3.422787
cloudcomputing             -0.7017149 0.05546801 -12.650805
=========================================================================
Log-lik            = -814.4793
Mean Log-likelihood= -1.326513
-------------------------------------------------------------------------
AIC       = 1658.959
BIC       = 1725.258
ICOMP     = 1644.233
=========================================================================
Elapsed time= 1.434 seconds =====>>> Mon Jan 23 20:06:33 2023
=========================================================================
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 25  Maxiter= 500
=========================================================================
Uncertainty
                            Estimates    StdErr      Wald
constant                    1.5070420 0.2705213  5.570880
areastudio1areascientifica  0.9148394 0.3851230  2.375447
cloudcomputing             -0.4031985 0.1022130 -3.944689
=========================================================================
Feeling
                            Estimates     StdErr       Wald
constant                    5.3892251 0.41428131  13.008613
genereuomini               -0.8790982 0.17489540  -5.026423
macroarea1centro           -1.0916177 0.37701751  -2.895403
macroarea1nord             -0.9889346 0.26204484  -3.773914
studiolaurea               -0.5905921 0.17358485  -3.402325
areastudio1areascientifica -1.2580851 0.28193315  -4.462352
uninaunina                 -0.9612546 0.28138746  -3.416125
settorelavoroSTEM          -0.7987978 0.21703755  -3.680459
cloudcomputing             -0.7001536 0.05518967 -12.686315
=========================================================================
Log-lik            = -814.6829
Mean Log-likelihood= -1.326845
-------------------------------------------------------------------------
AIC       = 1653.366
BIC       = 1706.406
ICOMP     = 1640.683
=========================================================================
Elapsed time= 0.987 seconds =====>>> Mon Jan 23 20:09:30 2023
=========================================================================
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 91  Maxiter= 500
=========================================================================
=========================================================================
     Estimates     StdErr      Wald
pai1 0.2695730 0.04476941  6.021367
pai2 0.3858653 0.05240038  7.363789
csi  0.6411016 0.03702702 17.314428
=========================================================================
Alternative parameterization
        Estimates     StdErr      Wald
paistar 0.4112866 0.07067701  5.819242
csi     0.6411016 0.03702702 17.314428
delta   0.3445616 0.02597114 13.267097
=========================================================================
Log-lik            = -959.7831
Mean Log-likelihood= -1.563165
Log-lik(UNIFORM)   = -1100.14
Log-lik(saturated) = -953.9499
Deviance           = 11.66634
-------------------------------------------------------------------------
AIC       = 1925.566
BIC       = 1938.826
ICOMP     = 1921.373
=========================================================================
Elapsed time= 0.004 seconds =====>>> Mon Jan 23 20:12:36 2023
=========================================================================
```

```
=========================================================================
=====>>> CUB  model    <<<=====   ML-estimates via E-M algorithm
=========================================================================
 m= 6  Sample size: n= 614  Iterations= 68  Maxiter= 500
=========================================================================
Uncertainty
                            Estimates    StdErr      Wald
constant                    1.4626274 0.2936896  4.980180
areastudio1areascientifica  0.9466881 0.4564721  2.073923
cloudcomputing             -0.2887779 0.1308102 -2.207610
=========================================================================
Feeling
                            Estimates     StdErr       Wald
constant                    4.4354106 0.21283092  20.840067
genereuomini               -0.8254285 0.10757648  -7.672945
macroarea1centro           -0.9878833 0.22589179  -4.373259
macroarea1nord             -0.8014514 0.16265854  -4.927202
studiolaurea               -0.2938299 0.11447227  -2.566821
areastudio1areascientifica -0.8646336 0.12288979  -7.035846
uninaunina                 -0.7447468 0.16965082  -4.389880
settorelavoroSTEM          -0.6978172 0.15817838  -4.411584
cloudcomputing             -0.6862845 0.03435841 -19.974282
=========================================================================
Shelter effect
             Estimates    StdErr      Wald
constant     -1.444212 0.2995256 -4.821666
studiolaurea -1.748880 0.7927015 -2.206228
=========================================================================
Log-lik            = -807.9768
Mean Log-likelihood= -1.315923
-------------------------------------------------------------------------
AIC       = 1643.954
BIC       = 1705.834
ICOMP     = 1635.289
=========================================================================
```