Si consideri il seguente scenario di classificazione binaria in cui si vuole predire se una persona si iscriverà al corso di Inglese di livello B2

Età	Data di Nascita	Titolo di Studio	Diplomato?	Spesa Corsi	Parteciperà?
39	14/12/1784	Laurea Magistrale	Si	5000	Si
50	01/07/1973	Diploma	Si	2000	Si
15	02/09/2008	Licenza Media	No		No
23	19/04/2000	Diploma		1275	No
	30/03/1992	Diploma	Si	2000	No
42	15/11/1981	Laurea	Si		Si
17	28/10/2006	Licenza Media	No	750	Si

Esercizio 1

Indica se applicando il criterio di cancellazione delle righe si ottiene una distorsione nei dati. Motivare la risposta.

Soluzione:

Applicando la tecnica di cancellazione delle righe al dataset indicato si provocherà una distorsione dei dati rispetto lo scenario di riferimento. Infatti, verrebbero cancellate la terza, la quarta, la quinta e la sesta riga riducendo il dataset al seguente:

Età	Data di Nascita	Titolo di Studio	Diplomato?	Spesa Corsi	Parteciperà?
39	14/12/1784	Laurea Magistrale	Si	5000	Si
50	01/07/1973	Diploma	Si	2000	Si
17	28/10/2006	Licenza Media	No	750	Si

Tuttavia, il problema principale che si introduce in termini di distorsione non è tanto la riduzione delle osservazioni, quanto il fatto che non avremo più esempi di persone che hanno intenzione di partecipare al corso.

Esercizio 2

Applica tutti i possibili metodi deduttivi per ottenere alcuni dei valori mancanti. Riporta il dataset risultante.

Soluzione:

Nel dataset iniziale sono presenti tre tipi differenti di missing value, uno sull'attributo Età, uno sull'attributo diplomato e due sull'attributo Spesa corsi. Notiamo che si può applicare deduzione soltanto nei primi due casi.

Primo caso:

Quando per lo stesso record è presente il valore del titolo di studio è possibile *dedurre* l'età, applicando una semplice sottrazione tra la *current_date* e la *data_di_nascita*, e nel caso specifico si può dedurre che:

17/12/2023 – 30/03/1992 = 31

Per cui, il valore 31 può essere imputato nel quinto record sull'attributo Età.

Secondo caso:

Quando per lo stesso record è presente il valore della data di nascita è possibile *dedurre* il valore di diplomato, verificando semplicemente se il valore del titolo di studio è *licenza media* allora si può indicare

No, in Diplomato. Alternativamente, se il valore è *Diploma* o *Laurea Triennale* o *Laurea Magistrale* o *Dottorato di Ricerca media* allora si può indicare Si. Se anche per quell'attributo il valore è null, allora potremmo dedurre No per l'attributo diplomato solo se l'età è inferiore 17. Per cui nel caso specifico il valore Si può essere imputato nel quarto record sull'attributo Diplomato, visto che il valore, per lo stesso record, sull'attributo Titolo di Studio è proprio Diplomato.

Non possono essere fatte deduzioni esplicite per gli altri due missing value. Il dataset risultate dopo il processo di deduzione è il seguente.

Età	**Data di Nascita**	**Titolo di Studio**	**Diplomato?**	**Spesa Corsi**	**Parteciperà?**
39	14/12/1784	Laurea Magistrale	Si	5000	Si
50	01/07/1973	Diploma	Si	2000	Si
15	02/09/2008	Licenza Media	No		No
23	19/04/2000	Diploma	**Si**	1275	No
31	30/03/1992	Diploma	Si	2000	No
42	15/11/1981	Laurea	Si		Si
17	28/10/2006	Licenza Media	No	750	Si

Esercizio 3

Prova ad imputare i valori mancanti dell'attributo Spesa Corsi, sostituendo con i valori di media o moda. N.B.: devono essere riportate entrambe le tabelle risultati.

Soluzione:

Nel dataset sull'attributo Spesa corsi sono presenti due missing value, uno sulla terza e l'altro sulla sesta riga. Applichiamo i due metodo deterministici di imputazione indicati nella traccia:

Metodo 1: Media

Calcoliamo la media dei valori presenti in Spesa Corsi:

$$Media = \frac{5000 + 2000 + 1275 + 2000 + 750}{5} = 2205$$

A questo punto sostituiamo i missing value di Spesa Corsi con il valore della media. Otteniamo il seguente dataset.

Età	**Data di Nascita**	**Titolo di Studio**	**Diplomato?**	**Spesa Corsi**	**Parteciperà?**
39	14/12/1784	Laurea Magistrale	Si	5000	Si
50	01/07/1973	Diploma	Si	2000	Si
15	02/09/2008	Licenza Media	No	**2205**	No
23	19/04/2000	Diploma		1275	No
	30/03/1992	Diploma	Si	2000	No
42	15/11/1981	Laurea	Si	**2205**	Si
17	28/10/2006	Licenza Media	No	750	Si

Metodo 2: Moda

Verifichiamo qual è il valore maggiormente presente in Spesa Corsi. Nello specifico, l'unico valore che ha due occorrenze è il valore 2000. Esso rappresenta il valore della moda. A questo punto sostituiamo i missing value di Spesa Corsi con 2000. Otteniamo il seguente dataset.

Età	**Data di Nascita**	**Titolo di Studio**	**Diplomato?**	**Spesa Corsi**	**Parteciperà?**
39	14/12/1784	Laurea Magistrale	Si	5000	Si
50	01/07/1973	Diploma	Si	2000	Si
15	02/09/2008	Licenza Media	No	**2000**	No
23	19/04/2000	Diploma		1275	No
	30/03/1992	Diploma	Si	2000	No
42	15/11/1981	Laurea	Si	**2000**	Si
17	28/10/2006	Licenza Media	No	750	Si