**Ricerca dell’ottimo**

*Classificatore multiplo*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Selezione Attributi | Bilanciamento | Standardizzazione | Rimozione Outlier | Accuratezza |
| No | No | Non standardizzato | No | 0,43 |
| Chi-Squared | SMOTE | Standardizzato | Sì | 0,19 |
| Chi-Squared | SMOTE | Standardizzato | No | 0,21 |
| Chi-Squared | Random Over Sampling | Standardizzato | Sì | 0,19 |
| Chi-Squared | Random Over Sampling | Standardizzato | No | 0,21 |
| Chi-Squared | Random Under Sampling | Standardizzato | Sì | 0,19 |
| Chi-Squared | Random Under Sampling | Standardizzato | No | 0,12 |
| Mutual Information | SMOTE | Standardizzato | Sì | 0,47 |
| Mutual Information | SMOTE | Standardizzato | No | 0,43 |
| Mutual Information | Random Over Sampling | Standardizzato | Sì | 0,38 |
| Mutual Information | Random Over Sampling | Standardizzato | No | 0,4 |
| Mutual Information | Random Under Sampling | Standardizzato | Sì | 0,28 |
| Mutual Information | Random Under Sampling | Standardizzato | No | 0,32 |
| F - Classifier | SMOTE | Standardizzato | Sì | 0,45 |
| F - Classifier | SMOTE | Standardizzato | No | 0,43 |
| F - Classifier | Random Over Sampling | Standardizzato | Sì | 0,36 |
| F - Classifier | Random Over Sampling | Standardizzato | No | 0,4 |
| F - Classifier | Random Under Sampling | Standardizzato | Sì | 0,29 |
| F - Classifier | Random Under Sampling | Standardizzato | No | 0,28 |

Dalla tabella, prodotto della ricerca per la combinazione ottimale dei metodi di preprocessing per il classificatore multiplo, si possono notare delle tendenze interessanti.

La prima, e più evidente, è il peggioramento notevole delle prestazioni del classificatore con la selezione degli attributi effettuata con il metodo *Chi-Squared*. Poiché questo metodo sceglie gli attributi che risultano più legati alla classe, e nel dataset nessuno degli attributi mostra forti correlazioni con essa, è ragionevole ipotizzare che abbia effettuato una selezione eccessiva del numero di attributi, portando ad una riduzione dell’accuratezza del classificatore.

La seconda è che, gli altri due metodi portano a risultati simili, largamente determinati dal metodo di bilanciamento. Si può notare che *SMOTE* risulti la tecnica più adatta per il classificatore multiplo, poiché la variabilità che introduce nel dataset generando dati sintetici per classi fortemente sottorappresentate viene mediata grazie ai 3 classificatori base, rendendolo più robusto. Il random *oversampling* mostra performance simili, anche se leggermente ridotte – confermando l’ipotesi che il sovracampionamento sia il metodo da preferire per questo dataset. Il sottocampionamento random, come evidenziato nella sezione dedicata al bilanciamento, risulta inadatto in quanto rimuove troppi elementi per poter addestrare correttamente il modello.

Infine, si può notare come generalmente la rimozione dei valori estremi porti a una maggiore capacità di classificazione – con 2 eccezioni. La prima è legata al selettore *Chi-Squared*. In quel caso avere delle informazioni in più, anche se di bassa qualità per la classificazione, risulta positivo e porta a guadagni marginali. L’altra è dovuta al random *undersampling*: anche in questo caso la penuria di dati nel set di addestramento è fa si che l’aggiunta dei valori estremi sia un beneficio.

Non risulta sorprendente, quindi, che il risultato ottimale sia ottenuto usando il selettore con *Mutual Information* – capace di identificare dipendenze non lineari fra gli attributi per lasciare solo quelli che aggiungono la maggiore quantità di informazione – combinato con il metodo *SMOTE* e l’eliminazione dei valori estremi. Questo porta ad un moderato guadagno di prestazioni rispetto al dataset grezzo, da 0.43 a 0.47, giustificando lo sforzo per migliorare il dataset.

*Albero decisionale*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Selezione Attributi | Bilanciamento | Standardizzazione | Rimozione Outlier | Accuratezza |
| No | No | Non standardizzato | No | 0,46 |
| Chi-Squared | SMOTE | Standardizzato | Sì | 0,15 |
| Chi-Squared | SMOTE | Standardizzato | No | 0,13 |
| Chi-Squared | Random Over Sampling | Standardizzato | Sì | 0,26 |
| Chi-Squared | Random Over Sampling | Standardizzato | No | 0,26 |
| Chi-Squared | Random Under Sampling | Standardizzato | Sì | 0,26 |
| Chi-Squared | Random Under Sampling | Standardizzato | No | 0,26 |
| Mutual Information | SMOTE | Standardizzato | Sì | 0,4 |
| Mutual Information | SMOTE | Standardizzato | No | 0,26 |
| Mutual Information | Random Over Sampling | Standardizzato | Sì | 0,36 |
| Mutual Information | Random Over Sampling | Standardizzato | No | 0,4 |
| Mutual Information | Random Under Sampling | Standardizzato | Sì | 0,26 |
| Mutual Information | Random Under Sampling | Standardizzato | No | 0,19 |
| F - Classifier | SMOTE | Standardizzato | Sì | 0,43 |
| F - Classifier | SMOTE | Standardizzato | No | 0,38 |
| F - Classifier | Random Over Sampling | Standardizzato | Sì | 0,36 |
| F - Classifier | Random Over Sampling | Standardizzato | No | 0,43 |
| F - Classifier | Random Under Sampling | Standardizzato | Sì | 0,17 |
| F - Classifier | Random Under Sampling | Standardizzato | No | 0,28 |

Nel caso della ricerca per l’albero decisionale gli andamenti si dimostrano simili a quelli visti in precedenza. Il selettore *Chi-Squared* rimuove troppi attributi ed impatta negativamente la performance, mentre gli altri ottengono risultati simili. Il sottocampionamento ha anch’esso un’influenza negativa sia sul selettore *Mutual Information* che sul *F-Classifier*.

Si separa dal precedente poiché non si trovano combinazioni che portino ad un risultato migliore del dataset grezzo – indicando che le tecniche non siano adatte al classificatore. Da notare il fatto che per quanto riguarda l’albero decisionale l’effetto della rimozione dei valori estremi è molto meno prevedibile, migliorando o peggiorando i risultati in maniera significativa. Le eccezioni in questo caso sono quelle legate al primo selettore di attributi: in quei casi l’effetto è nullo o particolarmente ridotto.

Per questo classificatore i risultati migliori si ottengono a partire dai dati grezzi, poiché la rimozione di informazioni dovuta alla selezione degli attributi o al sottocampionamento e l’aggiunta di variabilità dovuta all’*oversampling* – non mediate tramite un meccanismo di voto – portano ad una complicazione dei dati per l’albero decisionale. In questo caso, lo sforzo si rivela inutile.