**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.

*KNN*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Selezione Attributi | Bilanciamento | Standardizzazione | Rimozione Outlier | Accuratezza |
| No | No | Non standardizzato | No | 0,32 |
| Chi-Squared | SMOTE | Standardizzato | Sì | 0,21 |
| Chi-Squared | SMOTE | Standardizzato | No | 0,21 |
| Chi-Squared | Random Over Sampling | Standardizzato | Sì | 0,23 |
| Chi-Squared | Random Over Sampling | Standardizzato | No | 0,23 |
| Chi-Squared | Random Under Sampling | Standardizzato | Sì | 0,29 |
| Chi-Squared | Random Under Sampling | Standardizzato | No | 0,19 |
| Mutual Information | SMOTE | Standardizzato | Sì | 0,28 |
| Mutual Information | SMOTE | Standardizzato | No | 0,28 |
| Mutual Information | Random Over Sampling | Standardizzato | Sì | 0,32 |
| Mutual Information | Random Over Sampling | Standardizzato | No | 0,36 |
| Mutual Information | Random Under Sampling | Standardizzato | Sì | 0,26 |
| Mutual Information | Random Under Sampling | Standardizzato | No | 0,26 |
| F - Classifier | SMOTE | Standardizzato | Sì | 0,3 |
| F - Classifier | SMOTE | Standardizzato | No | 0,27 |
| F - Classifier | Random Over Sampling | Standardizzato | Sì | 0,34 |
| F - Classifier | Random Over Sampling | Standardizzato | No | 0,36 |
| F - Classifier | Random Under Sampling | Standardizzato | Sì | 0,3 |
| F - Classifier | Random Under Sampling | Standardizzato | No | 0,36 |

Per quanto riguarda il KNN si può notare che il metodo di selezione degli attributi che porta ai migliori risultati è lo *F-Classifier*. Questo sistema, che valuta la varianza dell’attributo all’interno di una classe e tra le classi, consente di mantenere nel dataset quelli che sono meglio in grado di separare le classi – e quindi la cui distanza ha il maggiore effetto sulla classificazione. Gli altri metodi portano a risultati moderatamente peggiori, ma paragonabili.

Tra i metodi di bilanciamento quello che risulta più efficace è il sovracampionamento random, che solo considerando il selettore *Chi-Squared* non porta ai risultati migliori. In quel caso, considerando la penuria di attributi sui quale calcolare le distanze, il modello risulta meno in grado di gestire la variabilità aggiunta dai metodi di *oversampling*, e si ottengono risultati migliori mantenendo il maggior numero possibile di record veri senza generarne di nuovi.

L’eliminazione degli outlier ha degli effetti moderati nel caso del KNN: poiché questi valori sono così lontani dagli altri, hanno una limitata capacità di influenzare le classi che vengono assegnate ai nuovi.

Per questo modello, i risultati migliori si ottengono con lo *F-Classifier* il sovracampionamento random e senza togliere i valori estremi, il che porta ad un moderato guadagno di 0.04 in termini di capacità di classificazione per il modello

*Naïve Bayes*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Selezione Attributi | Bilanciamento | Standardizzazione | Rimozione Outlier | Accuratezza |
| No | No | Non standardizzato | No | 0,4 |
| Chi-Squared | SMOTE | Normalizzato | Sì | 0,38 |
| Chi-Squared | SMOTE | Normalizzato | No | 0,4 |
| Chi-Squared | Random Over Sampling | Normalizzato | Sì | 0,4 |
| Chi-Squared | Random Over Sampling | Normalizzato | No | 0,42 |
| Chi-Squared | Random Under Sampling | Normalizzato | Sì | 0,4 |
| Chi-Squared | Random Under Sampling | Normalizzato | No | 0,36 |
| Mutual Information | SMOTE | Normalizzato | Sì | 0,19 |
| Mutual Information | SMOTE | Normalizzato | No | 0,32 |
| Mutual Information | Random Over Sampling | Normalizzato | Sì | 0,19 |
| Mutual Information | Random Over Sampling | Normalizzato | No | 0,17 |
| Mutual Information | Random Under Sampling | Normalizzato | Sì | 0,15 |
| Mutual Information | Random Under Sampling | Normalizzato | No | 0,15 |
| F - Classifier | SMOTE | Normalizzato | Sì | 0,19 |
| F - Classifier | SMOTE | Normalizzato | No | 0,17 |
| F - Classifier | Random Over Sampling | Normalizzato | Sì | 0,19 |
| F - Classifier | Random Over Sampling | Normalizzato | No | 0,17 |
| F - Classifier | Random Under Sampling | Normalizzato | Sì | 0,17 |
| F - Classifier | Random Under Sampling | Normalizzato | No | 0,17 |

Questo classificatore richiede l’uso della normalizzazione al posto della standardizzazione, in quanto l’algoritmo richiede valori esclusivamente positivi per funzionare, il che influenza le prestazioni dei metodi rispetto a quanto visto nel caso degli altri classificatori.

Per il *Naïve Bayes* si nota una inversione di tendenza rispetto a quelli visti in precedenza. In questo caso la forte riduzione di attributi che si ottiene utilizzando il selettore *Chi-Squared* risulta ottimale, mentre gli altri metodi portano ad una forte riduzione di prestazioni, quasi sempre dimezzando il risultato ottenuto sul dataset grezzo.

Applicato il selettore, i metodi di bilanciamento hanno degli effetti minori sull’accuratezza del classificatore bayesiano, senza poter distinguere un metodo che risulti particolarmente più adatto. I dati migliori si ottengono con il sovracampionamento random, ma il guadagno è sufficientemente ridotto da poter essere attribuibile ad una buona, seppur casuale, scelta dei valori da duplicare.

L’eliminazione dei valori estremi ha anch’essa una influenza minore: poiché il classificatore è basato sulla probabilità gli *outlier* hanno una influenza ridotta, e difficilmente prevedibile.

Si conclude quindi che la combinazione ottimale di tecniche per questo classificatore è una selezione *Chi-Squared* con random *oversampling* e senza la rimozione dei valori estremi.