**DIMENSIONALITY REDUCTION**

Dai risultati e dalle considerazioni fatte in precedenza risulta certa la presenza nel dataset di più attributi troppo correlati tra loro (le informazioni che contengono sono troppo simili, es: la popolazione del 2022 con quella del 2021) e soprattutto di attributi poco utili al fine della predizione (non permettono di suddividere abbastanza efficacemente le classi). Perciò ora, attraverso le opportune tecniche di riduzione dimensionalità, osserveremo come cambiano le prestazioni a seguito della riduzione.

Le tecniche che abbiamo scelto di utilizzare sono tutte basate su **SelectKBest** il quale esegue uno scoring di tutti le features e permette di selezionare le prime k migliori. Tale scoring è stato eseguito su 3 test statistici:

1. con **mutual\_info**: una volta fornitogli il set di record da analizzare, ‘mutual\_info\_classif’ permette di misurare la quantità di informazione condivisa tra ogni feature e la variabile target. Si scelgono infine le prime k features che hanno la quantità di informazione condivisa maggiore.
2. Con **test chi2**: una volta fornitogli il set di record da analizzare, il test chi-quadro (chi2) misura l'indipendenza tra ogni feature e la variabile target. Calcola la differenza tra le frequenze osservate e quelle attese per ciascuna feature. Le prime k features con i valori di chi-quadro più alti, che indicano una maggiore dipendenza dalla variabile target, vengono selezionate.
3. Con test **ANOVA F-value:** una volta fornitogli il set di record da analizzare, il test (f\_classif) calcola il rapporto tra la varianza tra le classi e la varianza all'interno delle classi per ogni feature. Un valore F più alto indica una maggiore capacità della feature di separare le classi in modo significativo. Infine, vengono selezionate le prime k feature con i valori F più elevati, poiché risultano più efficaci nella discriminazione tra le classi.

Queste tre tecniche di riduzione sono state selezionate perché:

* La riduzione tramite mutual info è utile perché in grado di catturare relazioni non lineari tra le caratteristiche e la variabile target. Inoltre, è abbastanza robusto e discreto su una diversa varietà di distribuzioni dei dati e può essere applicato sia a caratteristiche continue che discrete. Misurando quanto una caratteristica riduce l'incertezza sulla variabile target, seleziona le caratteristiche più informative, che è ottimo per ottenere migliori performance.
* La riduzione tramite test chi2 risulta molto utile invece per individuare tutti quegli attributi che hanno una forte relazione statistica con la variabile target, e che quindi permettono sicuramente in maniera più efficace di distinguere le classi durante la predizione.
* La riduzione tramite test ANOVA F-value è utile perché, essendo un test basato sull’analisi della varianza, permette di selezionare le caratteristiche che, mostrando una maggiore differenza di varianza tra le classi, aiutano a distinguerle più efficacemente.

**CLASSIFICATORE MULTIPLO**

Senza preprocessing:

Immagine che contiene testo, schermata, Carattere, menu

Il contenuto generato dall'IA potrebbe non essere corretto.

Preprocessing con test chi2:

Immagine che contiene testo, schermata, Carattere, menu

Il contenuto generato dall'IA potrebbe non essere corretto.

A seguito della riduzione di dimensionalità mediante test chi2 è evidente che sia il Nayve Bayes che l’SVM in pratica non percepiscono alcun miglioramento nelle performance, da ciò si può intuire che questa tecnica non sembra particolarmente efficace con questi classificatori. La stessa cosa non può essere detta per l’albero decisionale che vede calare la sua accuratezza e le altre statistiche di più del 10% (quindi un deciso calo di performance).

Più o meno la stessa perdita di performance la osserviamo poi nella predizione finale del classificatore ensemble e ciò sottolinea che la riduzione tramite test chi2, a causa dell’inefficacia sull’albero decisionale, non sia ottimale con il nostro classificatore composto.

Preprocessing con mutual info:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Mediante riduzione tramite mutual info notiamo che le predizioni dell’albero decisionale questa volta vengono peggiorate in maniera inferiore rispetto al caso precedente e al contrario il Nayve Bayes sembra aver tratto un leggero beneficio da questa tecnica di riduzione. I risultati del classificatore ensemble in generale peggiorano, soprattutto nella precisione, di conseguenza anche questa tecnica non è adatta il nostro classificatore.

Preprocessing con f\_classif:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Attraverso quest’ultima tecnica di riduzione i risultati sembrano essere molto simili a quelli ottenuti precedentemente, infatti, l’albero decisionale cala leggermente nelle performance mentre il Nayve bayes trae un leggero vantaggio da questa riduzione dimensionale. Tutto ciò porta a un’accuratezza finale leggermente maggiore al prezzo però di un decremento nella precisione del classificatore.

**KNN CUSTOM**

Senza preprocessing:



Preprocessing con test chi2:



La tecnica di riduzione tramite test chi2 non sembra aver dato dei buoni risultati, al contrario sembra che il KNN reagisca male a questa tecnica visto questo decremento delle performance sia nel training set con cross validation sia poi nel test set.

Come avviene per altri tipi di classificatori, la riduzione della dimensionalità dovrebbe teoricamente apportare benefici anche a questo tipo di modelli. Quindi, riteniamo che il calo nelle prestazioni possa essere attribuito alla mancanza di informazioni utili a fini predittivi, derivante dalla selezione degli attributi. In altre parole, la riduzione potrebbe aver eliminato variabili che, pur essendo considerate meno significative, risultavano comunque utili per una corretta separazione delle classi.

Preprocessing con mutual info:



Anche la riduzione tramite mutual info sembra riconfermare che questo classificatore non stia ottenendo alcun beneficio da queste riduzioni di dimensionalità. Il calo prestazionale è deciso e anche molto simile al caso precedente, soprattutto sul test set. Per quanto riguarda il validation set, si ha anche in questo caso un peggioramento di performance e ciò ci conferma che probabilmente durante la riduzione si sono perse anche informazioni utili per la predizione.

Preprocessing con f\_classif:



La tecnica con test ANOVA dimostra di ottenere risultati ancora peggiori delle tecniche precedenti. Questa quindi fra le tre tecniche utilizzate è quella meno adatta al nostro classificatore e conferma infine che nessuna delle tecniche di riduzione permette al classificatore di ottenere una performance migliore rispetto al dataset grezzo. In quest’ultimo caso poi si ha un netto peggioramento nelle performance sul validation set che sottolinea che il modello non è riuscito a adattarsi bene agli attributi selezionati e, come conseguenza, questo sicuramente ha peggiorato ulteriormente le predizioni sul test set.

**NAIVE BAYES**

Senza preprocessing:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Preprocessing con test chi2:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

A seguito della riduzione effettuata sulla base del test chi2 si nota un grosso peggioramento sia nelle predizioni sul test set sia su quelle del training set ottenute durante cross-validation, osservando in entrambi i casi un calo prestazionale di circa 15%. Tra i precedenti classificatori, insieme all’albero decisionale dell’ensemble, sembra che sia tra i modelli di classificazione meno adatti per questa tecnica di riduzione.

Per questo classificatore la riduzione dovrebbe risultare molto utile, anche perché elimina potenziali features condizionalelmente dipendenti tra loro (e che quindi violano la sua assunzione di dipendenza condizionale tra gli attributi). Il forte calo però sembra sottolineare invece che gli attributi eliminati erano effettivamente utili al classificatore e che quindi una loro eliminazione porti a perdere informazioni significative.

Preprocessing con mutual info:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Rispetto al caso precedente, la riduzione mediante mutual info sembra migliorare le predizioni sul test set di un 2% ma comunque portare a un sostanzioso peggioramento per quelle effettuate sul validation set.

Su quest’ultimo si nota però che con questa tecnica i risultati sono comunque meno variabili (rispetto alle predizioni ottenute con le altre tecniche di riduzione) e si attestano più o meno tutte sul 40% di accuratezza. Questo, insieme all’accuratezza superiore, sottolinea che scegliere gli attributi in base all’informazione mutua sia sicuramente il metodo più efficace per il nostro classificatore (tra quelli osservati).

Preprocessing con f\_classif:

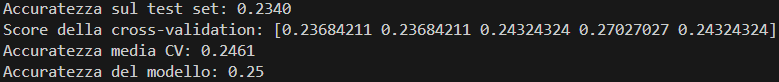
Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Con questa tecnica si può osservare una performance risultante molto simile a quella che si ottiene su dataset grezzo, con un leggerissimo decremento delle prestazioni durante la cross-validation e un decremento del 4% sul test set. Questa tecnica, quindi, non è adatta al nostro classificatore.

**SUPPORT VECTOR MACHINE**

Senza preprocessing:



Preprocessing con test chi2:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

La riduzione della dimensionalità tramite test chi2 non ha comportato neanche la minima differenza di performance su questo classificatore. La riduzione di dimensionalità dovrebbe portare all’SVM una maggiore stabilita e la possibilità di concentrarsi sulle caratteristiche più utili ai fini della predizione.

Si deduce quindi che gli attributi eliminati non erano per niente significativi per questo modello e che già l’SVM li ignorava durante le predizioni.

Preprocessing con mutual info:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Anche dopo questa tecnica di riduzione il risultato è il medesimo ottenuto precedentemente. Neanche scegliere gli attributi che forniscono le maggiori informazioni sulle classi è utile per incrementare la sua accuratezza.

Preprocessing con f\_classif:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Infine, con la riduzione mediante f\_classif, osservando che i risultati sono anche in questo caso identici a quelli ottenuti precedentemente, possiamo quindi asserire che nessuna delle 3 tecniche di riduzione di dimensionalità sia efficace per questo classificatore.

**DECISION TREE:**

Senza preprocessing:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Preprocessing con test chi2:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Con questa tecnica l’accuratezza sul test set crolla quasi del 50% rispetto a quella che si ottiene con il dataset grezzo mentre, durante la cross-validation sul training set, sembra ci sia solo un ‘leggero’ calo di performance. Questi risultati non ci sorprendono dato che abbiamo già osservato, nel classificatore ensemble, che l’albero decisionale non reagisce bene alla riduzione con test chi2. Possiamo dedurre quindi che si sia adattato troppo ai record di training diminuendo così la sua capacità di generalizzare (nel training set il calo di performance potrebbe essere invece dato dalla perdita di informazioni durante la riduzione).

Preprocessing con mutual info:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

A seguito della riduzione tramite mutual info l’accuratezza sul test set è leggermente diminuita, così come quella risultante dalla media delle accuratezze ottenute durante la cross-validation. La variabilità nei risultati sul validation set mostra una minore stabilità del classificatore (potrebbe essere dovuta agli attributi selezionati che non permettono un adeguata divisione tra le classi e, come nella tecnica precedente, alla perdita di informazioni utili per l’induzione dell’albero).

Preprocessing con f\_classif:

Immagine che contiene testo, schermata, Carattere

Il contenuto generato dall'IA potrebbe non essere corretto.

Anche con la riduzione tramite f\_classif i risultati sia sul test set (anche se in maniera maggiore rispetto al caso precedente) sia durante la cross-validation sembrano leggermente peggiori. Quindi anche questa tecnica peggiora le performance di questo classificatore e di conseguenza possiamo constatare che quest’ultimo non beneficia da nessuna delle tre tecniche di riduzione (dovuto come già detto ad attributi non ottimi per l’induzione dell’albero e la perdita di informazioni durante la riduzione).