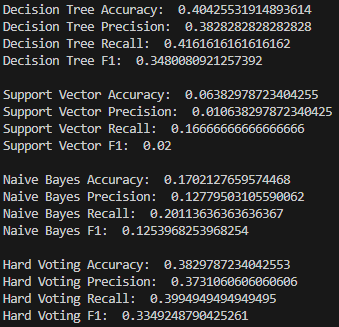
Bilanciamento

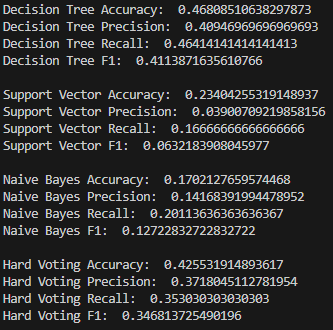
Poiché il dataset si è rivelato sbilanciato, occorre utilizzare delle tecniche di bilanciamento per ovviare al problema. Si nota infatti che le classi “*South America”* e *“Oceania”* sono sottorappresentate rispetto alle altre 4 – i loro dati risultano quindi più importanti per il corretto addestramento degli algoritmi e non conviene ridurli.

Le tecniche utilizzate sono 3: lo SMOTE *oversampling*, il random *oversampling* ed il random *undersampling*. Fra le tecniche di *oversampling* la prima risulta più precisa, e viene paragonata a quella casuale per verificare la sua capacità di creare una *decision boundary* più robusta. Poiché il numero di dati appartenenti alle classi minoritarie è ridotto si è preferito concentrarsi sulle tecniche di *oversampling*, quindi l’unica tecnica di *undersampling* valutata è il random *undersampling*.

**Ensemble**



Ensemble con SMOTE oversampling



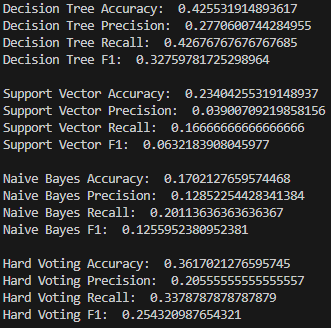
Ensemble con dati grezzi

*SMOTE*

Si può notare un impatto non trascurabile sulle prestazioni del modello, che subiscono una influenza negativa. In particolare, viene colpito il *Support Vector Classifier* che crolla dallo 0.234 di precisione fino allo 0.064. Questo, insieme ad una minore ma comunque significativa riduzione nella accuratezza del *Decision Tree* porta a risultati peggiori anche il classificatore multiplo.

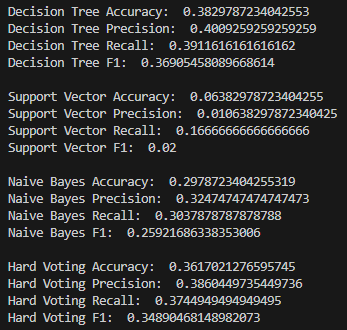
I dati ottenuti indicano quindi che la tecnica SMOTE non sia adatta al classificatore, e che il tentativo di generare più record su classi minoritarie porti ad una maggiore varianza. Questa peggiora le predizioni del modello, ed in realtà riesce solo a complicare il dataset

*Random oversampling*



Ensemble con Random Oversampling

Le affermazioni fatte per quanto riguarda la tecnica SMOTE vengono confermate dall’uso del metodo random, che mostra comunque un deterioramento delle capacità di classificazione maggiore rispetto a quella precedente, ma con una minore perdita di prestazioni per i classificatori base.



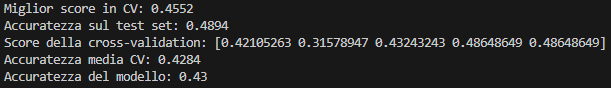
Ensemble con Random Undersampling

*Random undersampling*

Questa tecnica porta anch’essa ad un peggioramento delle prestazioni, riducendo significativamente l’accuratezza sia del *Decitsion Tree* che del *Support Vector Classifier*. In compenso il *Naive Bayes* guadagna molto da questo approccio, consentendo al metodo di ottenere risultati paragonabili a quelli del random *oversampling* – anche se con precisione e *recall* migliori.

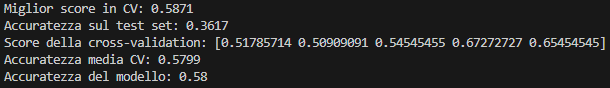
La causa di questo deterioramento è probabilmente la riduzione dei dati disponibili per l’addestramento dei classificatori, che si rivela troppo significativa su un dataset di dimensioni ridotte come quello analizzato.

**Decision Tree**



Decision Tree con dati grezzi

*SMOTE*

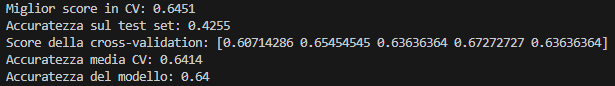


Decision Tree con SMOTE

Gli effetti della tecnica di *oversampling* sul classificatore risultano particolarmente interessanti, in quanto sembrano indicare un caso di *overfitting*. Nel corso delle *cross validation* il classificatore addestrato sulle versioni bilanciate mostra delle performance notevolmente migliori, fatto che influenza poi la sua accuratezza media – notevolmente superiore a quella del classificatore che lavora sui dati grezzi. Un’analisi dei risultati ottenuti sul test set mostra invece dei risultati opposti: un crollo di circa 12 punti percentuali nelle prestazioni su quest’ultimo.

Questi dati indicano che la tecnica di SMOTE introduce troppa variabilità nei dati sintetici, peggiorando sensibilmente le capacità del classificatore – si rivela dunque inadatta.

*Random Oversampling*

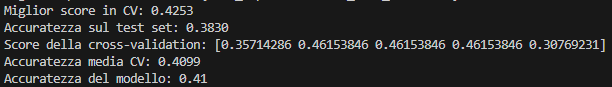


Decision Tree con Random Oversampling

I risultati ottenuti con il metodo random confermano il problema del precedente. I risultati ottenuti sui set di validazione sono molto migliori rispetto a quelli con SMOTE, e anche quelli sul test set si mostrano superiori. Questi ultimi rimangono tuttavia peggiori di quelli sui dati grezzi.

I risultati sembrano confermare che per gestire un dataset così sbilanciato le tecniche di *oversampling* siano inadatte.

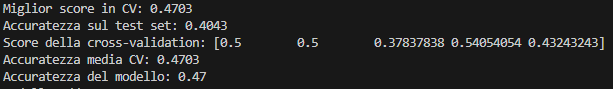
*Random Undersampling*



Decision Tree con Random Undersampling

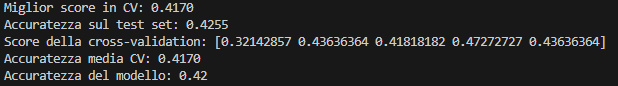
Il metodo di *undersampling* porta anch’esso ad una accuratezza notevolmente peggiore sul test set, ma in questo caso non migliora nemmeno quelle sul set di validazione. Risulta un metodo inadatto, in parte a causa della eccessiva riduzione dei valori per le classi maggioritarie necessaria per bilanciarle.

**Naïve Bayes**



Naïve Bayes con dati grezzi

*SMOTE*

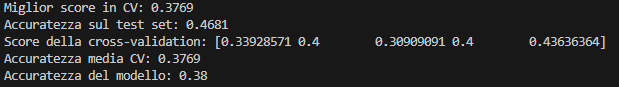


Naïve Bayes con SMOTE

Al contrario dell’albero decisionale con SMOTE si ottengono risultati peggiori nei 5 set di validazione generati dalla *cross-validation*, anche se il peggioramento è moderato rispetto ad altri analizzati in precedenza. Nel caso del test set si ha invece un leggero miglioramento, dovuto alla robustezza dei classificatori bayesiani rispetto al rumore generato dal metodo di generazione di dati sintetici.

Questo risultato porta a pensare che i metodi di bilanciamento possano avere degli effetti positivi su questo classificatore.

*Random Oversampling*

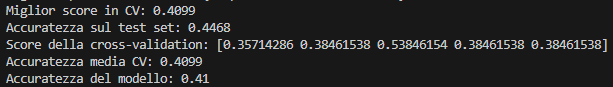


Naïve Bayes con Random Oversampling

I risultati ottenuti con il random *oversampling* confermano l’affermazione del paragrafo precedente. Gli effetti sono infatti gli stessi, ma il metodo porta a risultati più estremi – con un guadagno significativo di accuratezza nel test set ed una perdita altrettanto importante in quelli di validazione.

Si dimostra il metodo ottimale per questo classificatore, ma la variabilità dei risultati ottenibili dovuta alla sua natura casuale lo rende poco affidabile come metodo di miglioramento del modello.

*Random Undersampling*

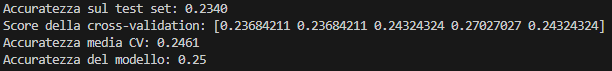


Naïve Bayes con Random Undersampling

Il random *undersampling* porta, con questo classificatore, ad ottenere dei risultati intermedi rispetto a quelli delle due tecniche viste in precedenza. Si ha infatti un miglioramento delle prestazioni nel set di test, accompagnato da un peggioramento di quelle nei set di validazione. In entrambi i casi, per quanto la differenza con i risultati ottenuti sul dataset grezzo non sia trascurabile, non raggiunge nemmeno gli estremi del metodo precedente.

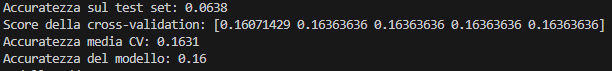
I risultati ottenuti indicano che le tecniche di bilanciamento sono adatte al classificatore bayesiano, e che nonostante un apparente peggioramento nella fase di validazione, ci si possa aspettare un moderato incremento dell’accuratezza nella classificazione.

**Support Vector Classifier**



SVC con dati grezzi

*SMOTE*

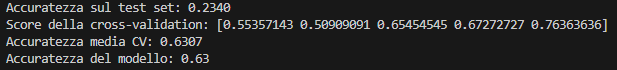


SVC con SMOTE

La tecnica SMOTE si rivela del tutto inadatta per l’uso con lo SVC. Si ottiene infatti un crollo dell’accuratezza sul test set di oltre 15 punti percentuali, e le sue prestazioni sui set di validazione – per quanto migliori – risultano comunque notevolmente peggiorate rispetto a quelle sui dati grezzi.

Il classificatore si dimostra particolarmente vulnerabile alla variabilità introdotta nel dataset dai dati sintetici, il che indica che tecniche che ne fanno uso siano inadatte.

*Random Oversampling*

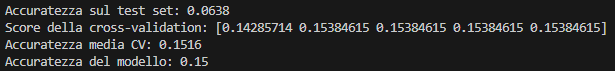


SVC con Random Oversampling

I risultati ottenuti applicando questa tecnica risultano interessanti, in quanto non si ha un vero guadagno in termini di capacità di classificazione sui dati del set di prova. Tuttavia, è possibile notare un marcato miglioramento su quelli dei set di validazione.

In questo caso, i risultati ottenuti non sembrano indicare un problema di *overfitting*, quanto una incapacità del metodo di bilanciamento di influenzare in maniera positiva i risultati – rivelandosi così inefficacie

*Random Undersampling*



SVC con Random Undersampling

Poiché il classificatore si basa sulla massimizzazione del margine tra le classi, ridurre troppo i dati che utilizza per calcolare il confine tra di esse gli risulta fatale. Si ottengono infatti risultati paragonabili a quelli ottenuti con il metodo SMOTE. Si dimostra quindi anch’esso inadatto al classificatore preso in esame.

**KNN**



KNN con dati grezzi

*SMOTE*



KNN con SMOTE

L’uso della tecnica SMOTE non sembra sortire effetti di alcun tipo sulla precisione del classificatore KNN, che mantiene prestazioni simili a quelle ottenute con i dati grezzi sia per quanto riguarda i set di validazione, sia per quelle relative al set di prova. È probabile che il metodo per il bilanciamento, in quanto basato sulla creazione di valore sintetici e partendo da classi fortemente sottorappresentate, crei questi nuovi record in punti che non si differenziano abbastanza da quelli precedenti. Questo fa si che i record ai quali risultano più vicini i valori di test non siano diversi – spiegando i risultati ottenuti e rivelandosi inefficace nel migliorare la precisione del classificatore.

*Random Oversampling*



KNN con Random Oversampling

L’affermazione del paragrafo precedente viene dimostrata dai risultati ottenuti con il random *oversampling*: la duplicazione dei record delle classi sottorappresentate porta a dei risultati migliori nel set di validazione, in quanto si trovano in posizioni identiche. Per quanto riguarda il set di prova, il risultato invariato mostra che il classificatore risulta insensibile alle tecniche di bilanciamento che fanno uso di *oversampling* quando si devono gestire classi con differenze di popolazione importante.

*Random Undersampling*



KNN con Random Undersampling

La tecnica di random *undersampling* si mostra invece l’unica con degli effetti sulle effettive capacità del classificatore. Porta infatti un moderato aumento della accuratezza nel test set, in cambio di una forte riduzione di capacità in quelli di validazione.

Questo risultato è dovuto alla forte diminuzione dei dati disponibili in ciascuna classe: gli esempi necessari per un modello di *rote learning* come il KNN sono ulteriormente ridotti quando si fa uso dei set di validazione, portando a performance minori. Nel caso invece del test di prova la rimozione di elementi che potevano portare a risultati errati si dimostra essere un beneficio ed il modello si rivela più prestante.