Giovani ed alcool

Sara Nardi (876777), Andrea Cattaneo (8125585), Stefano Marzorati (830272), Federico Campanella (793652)

Sommario
Il consumo e l'abuso di alcol fra i giovani e gli adolescenti è un fenomeno preoccupante e in forte crescita a livello mondiale. La cultura del bere attualmente diffusa tra i giovani sfocia sempre più frequentemente verso il fenomeno del "binge-drinking", ossia il "bere per ubriacarsi". Tale problema sociale concerne l'abuso smisurato di alcool riservato a singole occasioni, in particolare nel fine settimana. Questo eccesso di alcool, oltre a comportare problemi di salute, può portare a comportamenti scorretti e addirittura mortali! Basti pensare ad esempio alla guida in stato di ebbrezza, la quale causa danni alla propria persona e agli altri. Pertanto, si è ritenuta importante un'analisi che mettesse in evidenza l'assunzione di alcool nel weekend, al fine di facilitare l'individuazione dei soggetti che abusano di alcolici e cercare di riportarli ad un consumo responsabile.

Il dataset impiegato per studiare questo fenomeno è *Student alcohol Consumption* disponibile sulla piattaforma Kaggle. Esso è composto da informazioni riguardanti gli alunni di una scuola secondaria iscritti alle lezioni di matematica, contenuti in un primo file csv da 395 record, e di portoghese, contenuti in un secondo file csv da 649 record ed è caratterizzato da 33 attributi. Dopo un attento studio del dataset sono stati selezionati gli attributi ritenuti maggiormente pertinenti e rilevanti per rispondere alla nostra domanda di ricerca:

- *Sex – sesso*
- *Age – età*
- *Address – ubicazione dello studente*
- *Pstatus – genitori conviventi o separati*
- *Studytime – ammontare delle ore di studio durante il fine settimana*
- *Failures – numero di bocciature*
- *Higher – volontà di proseguire gli studi*
- *Romantic – se ha una relazione sentimentale*
- *Famrel – qualità delle relazioni all'interno della famiglia*
- *Goout – frequenza di uscite con amici*
- *Health – attuale stato di salute*
- *Absences – numero di assenze dalle lezioni*
- *Walc – consumo di alcol durante il fine settimana*

La variabile sulla quale focalizziamo la nostra analisi (variabile target) è *Walc*, la quale descrive il consumo di alcool nel fine settimana dei giovani.

Lo studio è stato sviluppato come segue: dapprima è stata svolta un'attività di preprocessing, successivamente sono stati implementati alcuni modelli di classificazione e le relative misure di performance e infine si è provveduto a trarre delle conclusioni generali che fornissero una previsione utile per la sensibilizzazione di questo fenomeno all'interno degli istituti scolastici.

Indice

1. Preprocessing

Inizialmente è stata affrontata la fase di preprocessing. Tale processo ha come obiettivo la preparazione del dataset per la successiva classificazione. Nella presente analisi l'attività di preprocessing è stata sviluppata in tre sezioni:

Unione del dataset

Il dataset utilizzato è costituito da due differenti file csv: student-por.csv e student-mat.csv. In seguito all'importazione, effettuata mediante due differenti nodi CSV Reader, si è provveduto alla concatenazione, in modo da creare un unico dataset. Successivamente è seguita l' eliminazione dei duplicati, come suggerito dalla documentazione fornita, al fine di escludere 382 studenti che appartengono ad entrambi i dataset. In particolare, si sono cercate e rimosse le osservazioni che presentavano gli stessi valori per i seguenti attributi: *school, sex, age, address, famsize, Pstatus, Medu, Fedu, Mjob, Fjob, reason, nursery, internet*. In conclusione a questa operazione, si è ottenuto un dataset composto da 662 records.

Aggregazione di *Walc*

Il secondo passo è stato quello di aggregare le modalità della variabile *Walc*. In dettaglio, le 5 modalità dell'attributo sono state sintetizzate in una nuova variabile, *Walc-cons*, per mezzo del nodo *rule engine*, come segue: ai valori dell'attributo compresi tra 1 e 3 è stata assegnata la modalità "*alto*" e a quelli compresi tra 4 e 5 la modalità "*basso*".

Riduzione dimensionalità dataset

La riduzione della dimensionalità apporta notevoli vantaggi all'analisi, in quanto consente di migliorare il funzionamento degli algoritmi utilizzati per l'apprendimento, l' interpretabilità dei risultati e consente di ridurre il tempo e la memoria di computazione.
A tal proposito, si è creato un nuovo attributo, *edu-lev*, dall'aggregazione di *Fedu* e *Medu*. L'attributo ottenuto, riguardante il livello di educazione familiare, assume le seguenti modalità: *alta* e *bassa*. In particolare, alla classe alta appartengono i record per i quali la somma di *Fedu* e *Medu* ha valore tra 6 e 10, mentre i casi in cui tale somma è inferiore a 6 sono assegnati alla classe *bassa*.
Concordemente a quanto esposto si sono rimossi gli attributi *Fedu* e *Medu*. Inoltre, sono stati rimossi gli attributi ritenuti ridondanti: *G1* e *G2*, in quanto aggregati in *G3* da colui che ha costruito il dataset; *Walc*, sostituito da *Walc-cons*; *Dalc*, poichè si è ritenuto più rilevante il consumo di alcolici nel weekend, in virtù della giovane età dei soggetti considerati.

2. Dataset sbilanciato

Il dataset utilizzato presenta una distribuzione sbilanciata della variabile target, in particolare la classe positiva (la classe *alto*) è rappresentata solo dal 21% delle osservazioni.
All'interno del workflow sono state utilizzate alcune tecniche per cercare di sopperire allo sbilanciamento. Per ridurre la distorsione è stata applicata la tecnica dell'*equal size sampling*, una tipologia di *random undersampling*. Tale metodologia risulta essere particolarmente utile in caso di sbilanciamento del dataset in quanto rimuove in modo casuale record appartenenti alla classe negativa (classe con presenza maggiore nel dataset), al fine di ottenere un campione che contenga tutte le osservazioni appartenenti alla classe positiva ed un ugual numero di osservazioni appartenenti alla classe negativa. Tuttavia, tale tecnica comporta la perdita di molte informazioni in quanto una parte del dataset viene trascurata.
Un altro metodo valutativo appropriato in presenza di dataset sbilanciato è l'analisi dei costi. Tale indagine è stata implementata utilizzando la tecnica del *cost sensitive learning*. Questo metodo associa alla matrice di confusione, che descrive le performance del classificatore, una matrice dei costi, che assegna ad ogni istanza della matrice di confusione uno specifico peso.
Inoltre, in presenza di un problema di sbilanciamento è doveroso prestare attenzione a quale misura di performance utilizzare per testare le prestazioni dei classificatori. Infatti, la misura di accuratezza non è significativa per la valutazione, in quanto tratta ogni classe con la stessa importanza, trascurando il fatto che si ritiene maggiormente interessante la classe rara in casi di questo tipo.

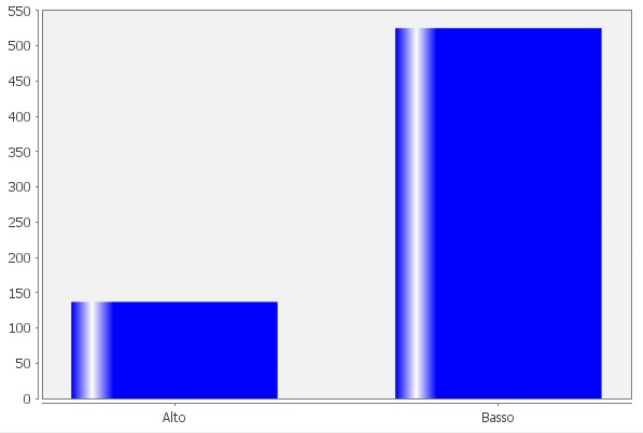

Figura 1. Distribuzione del dataset

3. Classificatori

In questo studio sono state implementate svariate tecniche di classificazione con lo scopo di individuare la più adatta, sulla base dei dati disponibili:

- Modelli euristici: L' algoritmo euristico è un particolare tipo di algoritmo progettato per trovare una soluzione approssimata, in quanto il risultato viene ottenuto cercando di equilibrare gli obiettivi di maggiori accuratezza, completezza, ottimizzazione, e velocità di computazione. Tra questa categoria di modelli è stata

posta attenzione sui classificatori *Random Forest* e *J48*, entrambi implementati dall'ambiente *Weka*.

- Modelli probabilistici: Tali modelli si basano sulla diretta manipolazione delle probabilità. In particolare si poggiano sul ragionamento bayesiano, il quale si fonda sul presupposto che le quantità di interesse sono disciplinate da distribuzioni di probabilità e che le decisioni ottimali possono essere assunte in seguito all'analisi congiunta di queste probabilità e dei dati osservati. Nel workflow Knime sono stati sviluppati il *Naive Bayes* (*Weka*) e *NBTree* (*Weka*).

- Modelli di separazione: I modelli di separazione utilizzati nella presente analisi sono il *Support Vector Machine* e l' *Artificial neural network*. Il primo implementato dal nodo *SMO* (*Weka*), il quale utilizza una funzione kernel *Puk*, normalizza gli attributi e sfrutta un algoritmo di ottimizzazione minima sequenziale per addestrare un classificatore del tipo support vector. Il secondo, effettuato mediante il nodo *MLP*(*Weka*), il quale implementa un percettrone multistrato che classifica le istanze attraverso una procedura di backward propagation error.

4. Misure di performance

Uno degli aspetti più importanti dell'apprendimento supervisionato è la valutazione delle misure di performance del modello di classificazione sviluppato. Tale peculiarità è stata considerata anche nella presente analisi, in quanto sono state calcolate le principali misure di performance.
L'*Accuracy* indica la percentuale di osservazioni positive e negative previste correttamente e permette di selezionare l'istanza che garantisce la miglior previsione sui record ignoti[1]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

Una buona misura di accuratezza raggiunta da un classificatore è rappresentata dalla percentuale dei dei record di test che sono stati classificati correttamente.
Tuttavia, in presenza di un dataset sbilanciato, come in questo caso, la sola stima puntuale dell'*Accuracy* risulta essere insufficiente e poco significativa nei riguardi della valutazione delle prestazioni del modello.

Pertanto, al fine di ottenere una migliore valutazione della bontà dei classificatori sviluppati nel presente modello sono state calcolate anche *Recall*, *Precision*, *F-Measure* e *Area Under Curve (AUC)* della curva *Receiver Operating Characteristic (ROC)*:

- *Recall*: tale indicatore misura la frazione di record positivi correttamente predetti dal modello di classificazione e, pertanto, viene anche comunemente denominata "*True Positive Rate*" o sensibilità del classificatore ed è calcolata come segue:

  $$Recall = \frac{TP}{TP+FN}$$

  Un elevato valore[2] di tale misura indica una bassa percentuale di record positivi classificati in modo errato.

- *Precision*: la *Precision* è quella quantità che determina la porzione di record che risulta effettivamente positiva nel gruppo che il modello di classificazione ha dichiarato come positiva:

  $$Precision = \frac{TP}{TP+FP}$$

  Dalla formulazione di tale misura si evince che più essa è elevata[3], minore è il numero dei record classificati come falsi positivi.
  Un modello di classificazione ottimale ha valori di *Precision* e *Recall* pari a 1 o al 100%. Tuttavia, precisione e sensibilità sono spesso correlate inversamente: quando migliora la precisione, peggiora la sensibilità del modello, e viceversa.
  Il più delle volte è consigliata la ricerca di un punto intermedio di equilibrio, il quale si presta molto bene al confronto tra classificatori: F-Measure.

- *F-Measure*: La media armonica delle misure di *Precision* e *Recall* è l'indicatore *F-Measure*:

  $$F-Measure = \frac{2(Recall * Precision)}{Recall + Precision}$$

  Essa varia nell'intervallo [0,1]: al tendere a 1 di tale misura, vi è garanzia di *Recall* sia *Precision* elevate.

- *Area Under Curve*: una *curva ROC* è una curva di probabilità, un grafico che mostra le prestazioni di un modello di classificazione a tutte le soglie di classificazione. In particolare, l'abbassamento della soglia di classificazione valuta più elementi come positivi di quelli che effettivamente lo sono, aumentando così sia i

[1] Le performance dei modelli di classificazione sono basate sul conteggio dei record correttamente e scorrettamente predetti dal classificatore e tali conteggi sono rappresentati tipicamente in una matrice di confusione. In particolare, TP e TN indicano il numero di istanze classificate correttamente come appartenenti rispettivamente alla classe positiva e negativa; FP e FN indicano il numero di istanze positive e negative classificate erroneamente.

[2] Il massimo valore di *Recall* è 1

[3] Il massimo valore di *Precision* è 1

falsi positivi che i veri positivi.
L'*AUC* misura l'intera area bidimensionale sotto l'intera *curva ROC* e rappresenta il grado o misura di separabilità, intesa come capacità del modello di distinguere tra classi. Dunque, maggiore è l'*AUC*, migliore è il modello nel prevedere 0 come 0 e 1 come 1.

5. Analisi e risultati

5.1 Holdout

In questa prima analisi del dataset si è applicato il metodo del *Holdout*, che è una delle tecniche comunemente utilizzate per stimare le misure di performance dei classificatori. Suddetta metodologia prevede la ripartizione arbitraria del dataset in *Training set* (67% delle osservazioni) e *Test set* (il restante 33%). Allo scopo di impedire che la modalità più frequente dell'attributo target domini nel campionamento, si è scelto di attuare un campionamento stratificato.
I risultati ottenuti dall'analisi sono i seguenti:

	Recall	Precision	F-Measure	Accuracy	AUC
J48	0.556	0.568	0.562	0.822	0.696
NBTree	0.467	0.656	0.545	0.840	0.716
NaiveBayes	0.333	0.469	0.390	0.785	0.750
SMOPuk	0.244	0.647	0.355	0.817	0.605
RandomForest	0.444	0.541	0.488	0.808	0.756
MLP	0.556	0.510	0.532	0.799	0.806
Logistic	0.378	0.739	0.500	0.845	0.857

Tabella 1. Risultati Holdout

Soffermando l'attenzione sui singoli classificatori, risulta essere maggiormente performante in termini di *Recall* il classificatore *Multilayer Perceptron*, mentre è l'*NBTree*, con riguardo alla misura di *Precision*, il classificatore preferibile tra quelli sviluppati. In aggiunta, la tabella evidenzia come, in termini di *Recall*, i valori di *SMOPuk*, *NaiveBayes* e *Logistic* sono bassi e poco significativi.
Inoltre, è possibile osservare che tutti i classificatori, ad eccezione del *Multilayer Perceptron*, presentano valori inferiori della *Recall* rispetto alla *Precision*, questo causato dal fatto che tendono a classificare negativamente i record in maniera eccessiva. In relazione alla *F1-measure* si osserva che i valori maggiormente elevati sono quelli dei modelli *J48*, *NBTree*, *MLP*, al contrario un valore poco significativo è quello relativo al *RandomForest*.
In termine di *Area Under Curve*, può contraddistinguersi per l'elevato valore il classificatore *Logistic* (0.857).

5.2 Cross Validation

Al fine di provare ad ottenere una misura più attendibile delle capacità dei classificatori, si è provveduto ad utilizzare il metodo della *k-folds cross validation*, il quale considerando tutti i valori del dataset è da ritenersi maggiormente preciso rispetto all' *Holdout*. Questo metodo si sostanzia nella suddivisione del dataset in *k* partizioni, di dimensioni pressoché uguali, ognuna delle quali con possibilità di essere utilizzata una volta come set di controllo e *k-1* volte per addestrare il modello.
Conseguentemente, le misure di performance di un classificatore nella *cross validation* corrispondono alla media delle misure di performance ottenute in ogni ciclo.

	Recall	Precision	F-Measure	Accuracy	AUC
J48	0.453	0.626	0.525	0.831	0.719
NBTree	0.445	0.555	0.494	0.811	0.756
NaiveBayes	0.445	0.496	0.469	0.792	0.785
SMOPuk	0.307	0.689	0.424	0.828	0.635
RandomForest	0.489	0.583	0.532	0.822	0.767
MLP	0.394	0.505	0.443	0.795	0.745
Logistic	0.416	0.620	0.498	0.826	0.813

Tabella 2. Risultati Cross Validation

In termini di *Accuracy* i modelli migliori sarebbero *SMOPuk* e *J48*, come mostrato dalla tabella 2. Tuttavia, come già preventivamente sottolineato, la caratteristica di sbilanciamento del dataset utilizzato rende l'*Accuracy* una misura inaffidabile e pertanto si è fatto ricorso alle misure di *Recall*, *Precision*, *F-measure* e *AUC* della *curva ROC*. Soffermando l'attenzione sulla *Recall*, il modello migliore risulta essere il *Random Forest*, che assume il valore più elevato (0,489), seguito dai modelli *J48* (0,453) e, con pari valore (0,445) *NBTree* e *NaiveBayes*. Inoltre, è interessante notare che la *Recall* è sempre inferiore a 0,5, ciò significa che meno del 50% dei valori positivi sono predetti correttamente. Dei quattro modelli con *Recall* migliore il *NaiveBayes* è quello con la peggior performance in termini di *Precision* (0,445), mentre il migliore è il *J48* (0,626), seguito da *Random Forest* (0,583) e *NBTree* (0,555). Inoltre, dalla tabella in figura è possibile mettere in evidenza che la sensibilità del modello di classificazione, in riferimento all'applicazione della tecnica di *Cross Validation*, è maggiore rispetto alla capacità di *Recall* dello stesso, per tutti gli inducer sviluppati.
Valutando i modelli in termini di efficienza complessiva attraverso la *F-measure*, il miglior classificatore sviluppato è il *Random Forest* (0,532), seguito da *J48* (0,525), *NBTree* (0,494) e *NaiveBayes* (0,469). In conclusione, è possibile affermare che i quattro classificatori più performanti assumono valori di *Recall*, *Precision* e *F-Measure* abbastanza vicini tra loro.
In riferimento a tutti i modelli selezionati, la misura *AUC* risulta superiore a 0,5 e questo permette di concludere che è maggiormente performante l'utilizzo di classificatori, rispetto

ad un "no model". Il miglior caso, quello che prevede un valore di *AUC* maggiormente prossima ad 1, è rappresentato dal classificatore *Logistic* regression (0,813), mentre il peggiore è quello che riguarda la performance di *SMOpuk* (0.635). Confrontando i risultati ottenuti dalle tecniche di *Holdout* e *Cross Validation*, è possibile notare, in riferimento alla *Recall* della *Cross Validation*, una diminuzione di tale misura soltanto per *J48*, *NBTree*, *MLP*, mentre aumenta in tutti gli altri casi. Tale evidenza dimostra una tendenza dell' *Holdout* a sovrastimare la *Recall* dei primi tre classificatori e a sottostimarla negli altri. Per quanto concerne la misura di *Precision* risulta che l'*Holdout* tende a sottostimarla con *NBTree*, *MLP* e *Logistic*; al contrario, esibisce una sovrastima negli altri casi, ad eccezione di quella riferita al *NaiveBayes* rimasta invariata.
Inoltre, la metodologia Holdout tende a sovrastimare *F-measure* per *J48*, *NBTree* e *MLP*, mentre propende a una sottostima negli altri classificatori. Infine, in termini di *AUC*, è rinvenibile una sovrastima della tecnica *Holdout* con riferimento ai classificatori *MLP* e *Logistic*.

5.3 Feature Selection

In seguito, si è provveduto all'applicazione del processo di *feature selection*. I metodi di selezione delle variabili hanno lo scopo di ridurre il numero di variabili input a quelle ritenute più utili per un modello, al fine di prevedere la variabile target. In altre parole, la selezione delle features si concentra principalmente sulla rimozione di attributi, non informativi o ridondanti, dal modello.
L'applicazione di tale procedura consente di ottenere innumerevoli vantaggi, quali la riduzione dei costi della raccolta dei dati, la riduzione del tempo di inferenza, ossia del tempo richiesto dal classificatore per predire il valore dell'attributo di classe, aumentare l'interpretabilità ed incrementare l'*Accuracy*.
Tuttavia, come già sottolinato preventivamente, disponendo di un dataset sbilanciato la misura dell'
Accuracy è poco significativa. Per ovviare parzialmente a questo problema, si è applicato il nodo *EqualSizeSampling* sul train set, il quale consente di bilanciare il dataset da utilizzare per addestrare i classificatori.
Per quanto concerne la suddetta analisi, si è ritenuto importante mettere in evidenza se gli attributi soggettivamente ritenuti rilevanti ai fini della predizione, fossero effettivamente quelli maggiormente legati alla variabile target da presagire, secondo la funzione obiettivo scelta. Per gli scopi illustrati precedentemente è stata appunto condotta l'analisi della *feature selection*, in particolare attraverso l'approccio del filtro multivariato. Dunque sono stati utilizzati i metodi *CfsSubsetEval* e *BestFirst*, per ognuno dei 7 modelli precedentemente descritti, in modo da effettuare una valutazione dei singoli attributi prima di sottoporli al classificatore. Il tutto si è svolto eseguendo il nodo *AttributeSelectedClassifier*, proprio dell'ambiente *Weka*. L'output di tale nodo mostra gli attributi che maggiormente influenzano la variabile risposta, senza trascurare la correlazione tra gli stessi: *sex, studytime, famsup, goout, grade*. Si ottiene quindi il nuovo dataset composto dai 5 attributi selezionati dall'analisi di feature e la variabile da presagire, al quale si applicano le tecniche di *Holdout* e *Cross Validation*, in modo da stimare le misure di prestazione dei classificatori scelti nel modello e quindi evidenziare la presenza o meno di *overfitting* e *undefitting* , e compararle successivamente con quelle ottenute dai classificatori senza l'utilizzo della metodologia di *feature selection*.

L'output ottenuto applicando la tecnica dell' *Holdout* è il seguente:

	Recall	**Precision**	**F-Measure**	**Accuracy**	**AUC**
J48	0.667	0.508	0.577	0.799	0.766
NBTree	0.733	0.398	0.516	0.717	0.796
NaiveBayes	0.711	0.432	0.538	0.749	0.812
SMOPuk	0.711	0.471	0.566	0.776	0.752
RandomForest	0.689	0.392	0.500	0.717	0.754
MLP	0.578	0.491	0.531	0.790	0.793
Logistic	0.822	0.457	0.587	0.763	0.833

Tabella 3. Feature Selection Holdout

Il suddetto risultato mette in evidenza quanto la misura di *Recall* sia in generale maggiore della misura di *Precision* e questo significa che i classificatori utilizzati nel modello tendono maggiormente ad evitare la presenza di falsi negativi, rispetto a quella di falsi positivi. Inoltre, è possibile notare che i valori di *AUC* sono tutti maggiori di 0,7 e questo indica che il modello ha una buona capacità di separabilità tra le classi. Con riguardo ai singoli classificatori, quello preferibile in termini di *Recall* è *Logistic* con il valore 0.882. Quest'ultimo indica che più dell'80% dei valori positivi vengono predetti correttamente da tale classificatore. Invece, in relazione alla misura di *Precision*, colui che assume un valore più elevato è *J48* (0.508) e pertanto tale classificatore classifica correttamente il 50% dei valori positivi predetti tra tutto l'insieme di valori positivi.

	Recall	**Precision**	**F-Measure**	**Accuracy**	**AUC**
J48	0.401	0.705	0.512	0.841	0.719
NBTree	0.453	0.713	0.554	0.849	0.806
NaiveBayes	0.394	0.607	0.478	0.822	0.810
SMOPuk	0.372	0.689	0.483	0.835	0.664
RandomForest	0.445	0.488	0.466	0.789	0.742
MLP	0.445	0.735	0.555	0.852	0.813
Logistic	0.401	0.679	0.505	0.837	0.824

Tabella 4. Feature Selection Cross validation

In relazione alla tecnica di *Cross validation*, dalla tabella 4, emerge che la misura di *Precision* è maggiore rispetto a quella di *Recall*, quindi si tende a commettere più errori come falsi

negativi.
Inoltre, comparando i singoli classificatori in merito alla misura di *Precision*, emerge che il *MultilayerPerceptron* è quello che assume valore più elevato, mentre la maggiore sensibilità è quella del classificatore *NBTree*.
In riguardo alle restanti misure di performance, *F-Measure* e *AUC*, è possibile osservare rispettivamente che *NBTree* e *Logistic* sono gli inducer da preferire.
Comparando la misura *F-Measure* della Cross Validation con quella dell'*Holdout* è possibile constatare risultati pressoché analoghi, in quanto anche nel caso della *Cross Validation* tale misura assume valori abbastanza lontani da 1, riflettendo uno sbilanciamento tra i valori di *Recall* e *Precision*.

5.4 Validazione e intervalli di confidenza

Al fine di svolgere una valutazione maggiormente scrupolosa dei classificatori utilizzati si è suddiviso il dataset, tramite il procedimento di *stratified sampling*, nella *partizione_A*, contenente l'80% dei record, e nella *partizione_B* costituita dal restante 20%, considerando come variabile di stratificazione *Walc-cons*.
La *partizione_A* precedentemente ottenuta è stata ulteriormente scissa in *training set* e *test set*. In particolare, al primo è stato assegnato il 67% delle osservazioni della *partizione_A* e al secondo la restante porzione. Analogamente a quanto utilizzato precedentemente nella *Feature Selection*, è stato applicato l'*equal size undersampling* al *training set* della *partizione_A* e su di esso sono stati allenati i modelli di classificazione. Il dataset è stato poi testato sul test set della *partizione_A* e sulla *partizione_B*. Suddetta classificazione è stata effettuata applicando il metodo *Holdout*. In seguito, si è provveduto alla comparazione della *F-Mesure* ottenuta sulle due diverse partizioni. Tale confronto è stato attuato mediante un grafico *line plot*, il quale risulta essere particolarmente esplicativo, in quanto permette la visualizzazione sia dei valori della prima partizione sia quelli della seconda partizione e pertanto consente di notare il cambiamento del valore della *F-measure* sulle due suddivisioni.

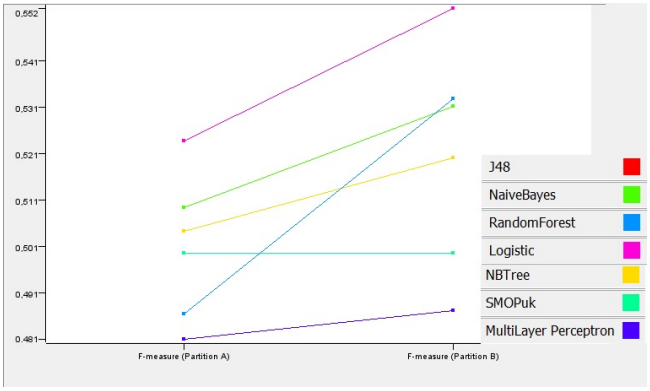

Figura 2. F-measure Partizioni

Dalla figura 2 è possibile notare come quasi tutti i classificatori abbiano valori maggiori di *F-measure* nella *partizione_B*, ciò è maggiormente evidente per il classificatore *Random Forest*, che presenta una linea maggiormente inclinata, mentre gli unici classificatori che mantengono costante il valore di *F-Measure* sono *J48* e *SMO puk*.
Si può notare anche che in termini di *F-measure* c'è un classificatore migliore degli altri, questo è il *Logistic* che sia nella *partizione_A* che nella *partizione_B*, ha performance migliori rispetto agli altri. Continuando nell'approfondimento dell'analisi, si sono calcolati gli intervalli di confidenza, con un livello di confidenza del 95%, sul valore della *F-measure* ottenuta nella prima partizione, la *partizione_A*. Il computo degli intervalli di confidenza è stato svolto secondo *Wilson* e nella figura 3 ne è stata fornita una visualizzazione.

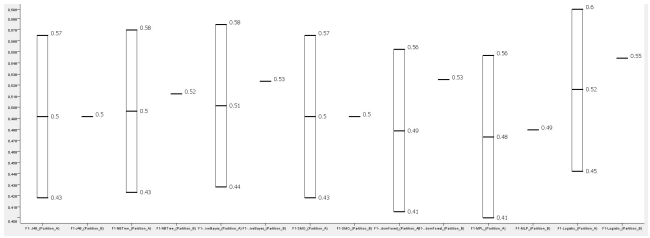

Figura 3. Box-plot intervalli di confidenza

Dalla figura si evince la presenza di tutti i valori della *F-measure* calcolati sulla *Partizione_B* situati all'interno dell'intervallo di confidenza. Due di questi valori, *J48* e *SMO puk*, sono addirittura coincidenti con quelli ottenuti sulla *Partizione_A*, concordemente a quanto emerso precedentemente nel line plot, in quanto erano quelli che presentavano una linea orizzontale. Inoltre, uniformemente a quanto rappresentato nella figura 2, tutti i valori della *partizione_B* sono collocati nella parte alta dell'intervallo, a dimostrazione che in tale partizione la misura di *F-measure* è maggiore nella *Partizione_A*.

5.5 Cost sensitive learner

Come detto precedentemente, per sopperire alle distorsioni create da un dataset sbilanciato si è utilizzato un *cost sensitive classifier*. In particolare, è stato utilizzato un nodo *Weka*, il *CostSensitiveClassifier*. L'analisi dei costi è stata svolta applicando il metodo dell'*Holdout*, attraverso la scelta arbitraria di assegnazione del 67% dei record al *training set*, mentre il restante al *test set*.
Al fine di scegliere la matrice di costo da utilizzare, si è adottata una tecnica Brute Force, modificando i valori dei costi di FP e FN, così da ottimizzare la *Recall*. Questo procedimento è stato attuato mediante l'uso dei nodi *Parameter Optimization Loop Start* e *Parameter Optimization Loop End*, i quali hanno permesso di provare tutte le combinazioni di costi con

un range da 0 a 10.
Tuttavia, questo metodo di selezione della matrice di costo è da ritenersi non ottimale per l'elevata forza computazionale che richiede e non solo. Infatti, per avere una maggiore sicurezza che i valori dei costi scelti siano corretti, sarebbe opportuno consultare un esperto di dominio.

Coerentemente alla domanda di ricerca posta nella presente analisi, è stata scelta l'ottimizzazione della misura di sensibilità dei classificatori rispetto alle altre misure. Infatti, si ritiene maggiormente dannosa la mancata identificazione di un ragazzo che consuma troppo alcol (FN), piuttosto che l'identificazione di un ragazzo che consuma pochi alcolici come se fosse uno che ne consuma troppi (FP).

	Recall	**Precision**	**F-Measure**	**Accuracy**	**AUC**
J48	0.711	0.372	0.489	0.694	0.700
NBTree	0.844	0.215	0.342	0.333	0.523
NaiveBayes	0.778	0.299	0.432	0.580	0.653
SMOPuk	0.244	0.647	0.355	0.817	0.605
RandomForest	0.911	0.261	0.406	0.452	0.622
MLP	0.733	0.398	0.516	0.717	0.723
Logistic	0.911	0.313	0.466	0.571	0.697

Tabella 5. Risultati Cost Sensitive Classifier

Valutando i risultati rispetto all'applicazione dei modelli senza la matrice di costo, in generale, è possibile constatare un aumento della *Recall*, che si assesta a livelli molto alti. L'unico modello che peggiora in termini di sensibilità è *SMO Puk*, il quale presenta un valore minore di quello ottenuto con la tecnica dell'*Holdout*. I modelli che più rispecchiano le esigenze della domanda di ricerca sono il *Random forest* e il *Logistic*, in quanto esibiscono una *Recall* di 0,911.
Da un'ulteriore analisi dei risultati, è possibile notare che ottimizzando la *Recall* è stata penalizzata la *Precision*, che è diminuita in tutti i classificatori. Tale evidenza è verificata, seppur in modo minore, anche per le altre misure di performance, come *F-measure* e *AUC*.
In conclusione è possibile affermare che il *cost sensitive* ha condotto a ottimi risultati in termini di *Recall*, ma ha condotto a un calo delle prestazioni generali dei classificatori.
Si rende necessario evidenziare che i costi ottenuti usando il metodo *Brute Force* per l'ottimizzazione della *Recall* hanno indotto ad avere all'interno della cost matrix un costo maggiore per i FP rispetto ai FN. Questo sembra essere l'unico modo per massimizzare la sensibilità dei classificatori, nonostante sia contrastante con il ragionamento logico formulato, secondo il quale dovrebbe essere assegnato un costo maggiore per i FN, in modo tale da "etichettare" quest'ultimo come errore più grave.

6. Conclusioni

La presente analisi è stata svolta con l'obiettivo di trovare il classificatore più adatto tra quelli selezionati per lo studio, con lo scopo di predire il consumo di alcolici tra i giovani durante il fine settimana.
I classificatori che paiono essere più adeguati a tale predizione sono *NBTree*, *Random Forest* e *MLP*. Infatti, tali inducer ottengono i valori di *Recall* e delle altre misure di prestazione tra i più elevati.
Le altre tecniche di classificazione sono state scartate per svariati motivi. Ad esempio, *SMO Puk* e *Logistic*, nonostante presentino valori adeguati di *AUC*, *F-Measure* e *Precision*, esibiscono valori di recall spesso inferiori a 0,4 e tendenzialmente tra i peggiori, confrontandoli con gli altri classificatori.
In seguito all'attuazione della procedura di *Feature Selection*, il modello *Naive Bayes* ha mostrato un miglioramento di performance. In generale, è possibile affermare che la *Feature Selection* ha apportato un miglioramento complessivo delle prestazioni delle tecniche di classificazione utilizzate.
In conclusione, si vuole sottolineare l'importanza della previsione del consumo di alcolici da parte degli studenti nel weekend, in quanto, sebbene il dataset suggerisca una maggior porzione di individui che assume alcool in moderate quantità, è comunque presente una frazione di essi che ne abusa. La formulazione di tale modello di classificazione auspica l'applicazione di esso per studi analoghi, in modo da poterli utilizzare congiuntamente per una radicata sensibilizzazione e conseguente estirpazione del dannoso fenomeno sociale trattato.

7. Appendice

Il dataset utilizzato per la domanda di ricerca posta nella presente analisi è composto da numerose variabili. Al fine di renderlo maggiormente chiaro e comprensibile viene riportata di seguito la lista esaustiva di tutti gli attributi in esso presenti.
Attributi numerici:

- *Age - età*
- *Medu - livello di educazione della madre*
- *Fedu - livello di educazione del padre*
- *Traveltime - tempo impiegato per arrivare a scuola*
- *Studytime - ore di studio nel weekend*
- *Failures - numero di bocciature*
- *Famrel - qualità della relazione familiare*
- *Freetime - tempo libero dopo la scuola*
- *Goout - tempo libero impiegato con gli amici*

- *Dalc - consumo di alcool durante la settimana*
- *Walc - consumo di alcool durante il weekend*
- *Health - stato di salute*
- *Absences - numero di assenze*
- *G1 - media dei voti nel primo semestre*
- *G2 - media dei voti nel secondo semestre*
- *G3 - voto finale*

Attributi binari:

- *School - quale scuola frequenta*
- *Sex - sesso*
- *Address - tipo di ubicazione dello studente*
- *Famsize - numero di componenti della famiglia*
- *Pstatus - stato di convivenza dei genitori*
- *Schoolsup - supporto educativo extra*
- *Famsup - sostegno educativo familiare*
- *Paid - lezione extra per le seguenti materie: Matematica e Portoghese*
- *Activities - attività extracurriculari*
- *Nursery - se ha frequentato scuola materna*
- *Higher - volontà di iscriversi a una scuola superiore*
- *Internet - disponibilità di internet a casa*
- *Romantic - se ha una fidanzata*

Attributi nominali:

- *Mjob - lavoro della madre*
- *Fjob - lavoro del padre*
- *Reason - motivo di scelta della scuola*
- *Guardian - chi è il tutore dello studente*

Riferimenti Bibliografici:

- *https://www.kaggle.com/uciml/student-alcohol-consumption*
- *https://www.medicalive.it/alcol-tra-i-giovani-impatto-sociale-e-sanitario*