Modelli di classificazione applicati al dataset dei funghi

Linda Malchiodi[1], Floriana Meluso[1], Sofia Monai[1]

Sommario

Le intossicazioni da funghi sono causate da una scorretta classificazione delle specie; le tossine possono causare diverse gravi sindromi in base alla quantità ingerita. Al fine di effettuare la giusta identificazione sono stati costruiti dei modelli di Machine Learning con l'obiettivo di prevedere, sulla base delle caratteristiche del fungo, se esso sia edibile. Inoltre, sono stati implementati metodi di classificazione per valutare la presenza consistente di funghi in base al territorio in cui crescono e alla loro commestibilità. Difatti, è noto che i diversi habitat nell'ecosistema favoriscono la crescita e la diffusione di funghi sia edibili sia velenosi. Supponendo reali i dati in esame, sono stati selezionati, dopo un'attenta valutazione e comparazione dei metodi, i modelli migliori in termini di performance e attendibilità per rispondere alle domande di ricerca. I risultati ottenuti consentiranno di fornire uno strumento in aiuto alla prevenzione di intossicazioni ed una guida che permetta di quantificare in base all'ambiente la presenza di funghi.

Keywords

Machine Learning – Classificazione – Funghi – Salute

[1]Università degli Studi di Milano – Bicocca, CdLM CLAMSES

Indice

Introduzione

Negli stati in cui i funghi sono molto consumati, un numero di intossicazioni rilevanti è censito ogni anno. Questo fenomeno è dovuto alla scorretta identificazione delle specie. Le tossine pericolose dei funghi sono in grado di causare diverse sindromi che possono essere fatali a seconda della quantità ingerita[1]. Pertanto una corretta identificazione del fungo, commestibile o velenoso, è importante per evitare incidenti. Il dataset impiegato per studiare questo fenomeno è composto da 8124 record e 23 variabili che descrivono le diverse peculiarità del fungo. Questo ipotetico campione corrisponde alle 23 specie di funghi lamellati nella famiglia Agaricus e Lepiota è tratto dalla Audubon Society Field Guide ai North American Mushrooms (1981). Ogni specie è identificata come commestibile o velenosa. La guida afferma chiaramente che non esiste una regola semplice, come ad esempio una singola caratteristica, per determinare la commestibilità di un fungo[2].

Dunque nella presente analisi, l'obiettivo primario è quello di prevedere, sulla base dei dati forniti relativamente alle caratteristiche del fungo, se esso sia edibile. Inoltre, l'obiettivo secondario è implementare un metodo di classificazione per identificare

l'abbondanza di funghi sulla base della velenosità e del loro habitat.

L'articolo è così costruito: inizialmente è stato introdotto il dataset ed un'analisi preliminare sui dati; successivamente sono stati presentati i modelli utilizzati e le misure di performance impiegate; infine sono state riportate le analisi ed i risultati ottenuti suddivisi secondo i due obiettivi introdotti precedentemente.

Presentazione del dataset

Descrizione delle variabili

Come anticipato il dataset è composto da 23 varabili categoriali nominali:

1. *class*: edibile (e) o velenoso (p);
2. *cap-shape*: forma del cappello;
3. *cap-surface*: superficie del cappello;
4. *cap-color*: colore del cappello;
5. *bruises*: ammaccature;
6. *odor*: odore;
7. *gill-attachment*: attaccamento delle lamelle;
8. *gill-spacing*: spaziatura delle lamelle;
9. *gill-size*: dimensione delle lamelle;
10. *gill-color*: colore delle lamelle;
11. *stalk-shape*: forma del gambo;
12. *stalk-root*: radice del gambo;
13. *stalk-surface-above-ring*: superficie del gambo sopra l'anello;
14. *stalk-surface-below-ring*: superficie del gambo sotto l'anello;
15. *stalk-color-above-ring*: colore del gambo sopra l'anello;
16. *stalk-color-below-ring*: colore del gambo sotto l'anello;
17. *veil-type*: tipo di velo;
18. *veil-color*: colore del velo;
19. *ring-number*: numero di anelli;
20. *ring-type*: tipo di anelli;
21. *spore-print-color*: impronta sporale;
22. *population*: numerosità;
23. *habitat*: terreno.

Per una maggiore comprensione, in Figura 1 è rappresentata la composizione del fungo.

Figura 1 - Composizione del fungo

Dal risultato delle statistiche è emerso che l'unica variabile che presenta dei valori mancanti è *stalk-root*, che descrive 6 tipologie possibili di radici del gambo, di cui 2 non sono presenti nel dataset ed il 31% dei record è non registrato, quindi mancante.

Modelli e Misure di Performance

1. Modelli

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

- Modelli euristici: albero di decisione **J48**, implementato da Weka, consente di classificare anche i dati nominali; **Random Forest**, classificatore composto da molti alberi di decisione, è in grado di gestire anche i dati categoriali;
- Modelli di regressione: **Regressione logistica**, la variabile dipendente è di tipo dicotomico;
- Modelli di separazione: Sequential Minimal Optimization, risolve il problema che emerge durante l'addestramento del SVM, i kernel scelti sono: **polykernel** e **puk**; **Multilayer Perceptron**, reti neurali;
- Modelli probabilistici: **Naïve Bayes** basato sul teorema di Bayes può essere utilizzato anche combinando dati numerici e categoriali; **NBTree** genera un albero di decisione attraverso il classificatore Naïve Bayes.

2. Misure di performance

Nella nostra analisi sono stati utilizzati più criteri in grado di valutare la performance. In particolare, si è scelto di calcolare: Accuracy, Recall, Precision, F_1-measure e Area Under Curve (AUC) della curva Receiver Operating Characteristic (ROC). L'Accuracy indica la percentuale di osservazioni positive e negative previste correttamente e permette di selezionare l'istanza che garantisce la miglior performance sui record da prevedere. In particolare:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Dove 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.
In generale, i modelli con Accuracy più alta sono valutati come migliori. Tuttavia, la sola stima puntuale dell'Accuracy non basta per identificare un buon classificatore. Per tale motivo si considerano altri criteri

per ottenere una più completa valutazione della performance del classificatore.
L'indicatore Recall rappresenta la porzione di record positivi correttamente classificati dal modello. In particolare:

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

Un valore alto di Recall indica che pochi record positivi sono stati classificati in modo errato.
L'indicatore Precision descrive la frazione di record che sono effettivamente positivi tra tutti quelli predetti come tali. In particolare:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

Un valore alto di Precision determina un numero minore di falsi positivi. Queste ultime due misure possono essere in conflitto: è possibile che si costruiscano dei modelli che massimizzano unicamente una delle due misure, registrando, dunque, un contrasto nei risultati. La F_1-measure, media armonica tra Recall e Precision, consente di fornire un'interpretazione delle due metriche più ragionevole: un valore elevato garantisce che sia Recall sia Precision siano indicativamente alte.

$$F_1 - measure = \frac{2*Precision*Recall}{Precision+Recall} \quad (4)$$

Un ulteriore metodo utilizzato per valutare il modello di classificazione è la curva ROC che consente di rappresentare, sull'asse delle ordinate, la percentuale del numero totale di veri positivi (classe positiva effettivamente prevista come tale) e, sull'asse delle ascisse, la percentuale di falsi positivi (classe negativa erroneamente prevista come positiva). L'AUC rappresenta il valore dell'area sotto la curva ed è una buona misura di performance perché consente di definire la qualità del classificatore: più il valore è alto, migliore è il modello.

Analisi e risultati

In questa sezione sono presentate le analisi e i risultati ottenuti in merito alle due domande di ricerca.

1. Prima domanda di ricerca

L'attributo scelto come variabile risposta per la prima domanda di ricerca è *class*. La distribuzione è equa tra le modalità che assume (52% *edible* e 48% *poisonous*); dunque non siamo in presenza di un dataset sbilanciato. In questa sezione sono presentate le diverse tipologie di analisi effettuate: la classificazione con il metodo Holdout; la feature selection; la validazione con la rappresentazione degli intervalli di confidenza per l'Accuracy.

1.1. Classificazione con metodo Holdout

In questa prima fase è stato impiegato il metodo dell'Holdout che si basa sulla partizione del dataset in due sottoinsiemi disgiunti attraverso un procedimento di stratified sampling in cui la variabile di stratificazione è *class*. Grazie a questo processo è stato ottenuto il traning set (67% dei record) ed il test set (33% dei record). I classificatori creati utilizzando gli 8 modelli precedentemente descritti sono stati addestrati con il training set e validati attraverso il test set.
Come anticipato in precedenza il dataset è composto da 23 variabili, pertanto è utile selezionare quelle più significative per il primo obiettivo: prevedere sulla base delle caratteristiche del fungo se esso sia edibile. Secondo la tradizione, la conoscenza etnomicologica dei funghi eduli è limitata ad un'analisi olfatto-visiva del corpo esterno. La loro identificazione si basa su caratteristiche come il colore e la forma del cappuccio, il colore e la forma del gambo, la dimensione delle lamelle, l'impronta sporale e l'odore[3]. Pertanto, le variabili utilizzate, in questa prima analisi, sono: *cap-shape*, *cap-surface*, *cap-color*, *odor*, *gill-size*, *gill-color*, *spore-print-color*.

Classificatore	Recall	Precision	F_1-measure	Accuracy	AUC
J48	1	0,994	0,997	0,997	1
Random Forest	1	0,994	0,997	0,997	0,999
Logistic	0,999	0,996	0,997	0,997	1
SMO poly	1	0,993	0,996	0,996	0,996
SMO puk	1	0,991	0,996	0,996	0,995
MLP	1	0,994	0,997	0,997	1
NaïveBayes	0,993	0,974	0,983	0,982	0,999
NBTree	1	0,994	0,997	0,997	1

Tabella 1 - Indicatori di performance sui diversi modelli (variabili input come da Letteratura)

I valori riportati in Tabella 1 non presentano differenze sostanziali tra i diversi classificatori. Inoltre, si nota che assumono tutti valori molto elevati, sintomo di una corretta classificazione.
I modelli quali *Regressione Logistica*, *SVM* e *MLP* non consentono, a livello teorico, l'utilizzo di variabili categoriali nominali come variabili input. Pertanto, è stata effettuata una seconda analisi, in cui nei modelli citati (*Logistic, SMO poly, SMO puk* e *MLP*) sono state binarizzate le variabili esplicative. I risultati ottenuti sono analoghi ai precedenti, quindi non sono riportati ulteriormente.

In seguito, è stata effettuata l'analisi scegliendo come input gli attributi: *gill-size*, *gill-spacing*, *gill-color*, *stalk-root*, *ring-type*, *spore-print-color* e *population*. I record che contengono i valori mancanti della variabile radici del gambo *(stalk root)* sono state eliminate e il dataset non è stato binarizzato.

Classificatore	Recall	Precision	F_1-measure	Accuracy	AUC
J48	0,979	0,997	0,988	0,986	0.999
Random Forest	0,982	0,997	0,989	0,987	1
Logistic	0,969	1	0,984	0,981	0.999
SMO poly	0,969	1	0,984	0,981	0.984
SMO puk	0,989	0,99	0,99	0,987	0.987
MLP	0,982	0,997	0,989	0,987	1
NaïveBayes	0,996	0,903	0,947	0,931	0.963
NBTree	0,989	0,994	0,991	0,989	1

Tabella 2 – Indicatori di performance sui diversi modelli (variabili selezionate)

In Tabella 2 sono riportate le misure di performance ottenute e anche in questo caso restano tutte molto alte. Il valore assunto dalla Recall nei modelli è diminuito, tranne per il *Naïve Bayes* in cui è aumentato. La Precision del *Naïve Bayes* è diminuita passando da 0,974 a 0,903. Le altre misure di performance (AUC, Accuracy e F_1-measure) hanno subìto variazioni leggerissime. In generale, il *Naïve Bayes* è stato l'unico modello che ha riscontrato un lieve peggioramento rispetto agli altri.

I valori di performance alti per entrambe le classificazioni ci portano a concludere che, indipendentemente dal modello utilizzato scegliendo un numero di variabili significativo (maggiore di quattro) le misure di performance ottenute sono buone, al contrario, con solo una o due variabili, non performano in maniera adatta. In generale, nonostante i valori degli indicatori siano alti, è fondamentale che non ci siano falsi positivi cioè funghi velenosi classificati come edibili. Dunque, considerando la Precision, i modelli ottimali risultano essere *Logistic* e *SMO Poly*.

1.2. Feature selection

Con l'obiettivo di selezionare le variabili più performanti e ridurre il numero di attributi in input, migliorando l'interpretabilità dei dati, è stata applicata la feature selection con metodo CfsSubsetEval che effettua una valutazione dei singoli attributi prima di sottoporli al classificatore. In questo modo, tramite filtro multivariato, vengono selezionate le variabili che maggiormente influenzano la variabile risposta senza trascurare la correlazione tra le stesse.

1.2.1. Feature selection con CfsSubsetEval

Come effettuato per la classificazione con metodo Holdout, anche in questo caso, è stato partizionato il dataset iniziale in training set (67% dei record) ed in test set (33% dei record), grazie al procedimento di stratified sampling, selezionando *class* come variabile etichetta. Successivamente è stato impiegato il nodo Weka ***AttributeSelectedClassifier***, in cui sono stati sempre utilizzati i metodi CfsSubseEval e BestFirst, per ognuno degli 8 modelli precedentemente descritti. Le variabili individuate come ottimali sono: *odor*, *gill spacing*, *stalk surface above ring* e *veil color*.

In Tabella 3 sono riportati i risultati delle performance:

Classificatore	Recall	Precision	F_1-measure	Accuracy	AUC
J48	1	0,977	0,988	0,988	0,992
Random Forest	1	0, 977	0,988	0,988	0,992
Logistic	1	0,977	0,988	0,988	0,992
SMO poly	1	0,977	0,988	0,988	0,987
SMO puk	1	0,977	0,988	0,988	0,987
MLP	1	0,977	0,988	0,988	0,992
NaïveBayes	1	0,966	0,983	0,982	0,991
NBTree	1	0,977	0,988	0,988	0,992

Tabella 3 - Indicatori di performance sui diversi modelli (Feature selection su dataset iniziale con CfsSubsetEval)

In tutti i modelli il valore della Recall è sempre 1, quindi non ci sono falsi negativi cioè funghi edibili classificati come velenosi. Ad esclusione del *Naïve Bayes* tutti gli altri classificatori hanno una Precision di 0,977 ed una F_1-measure ed Accuracy di 0,988 mentre il *Naïve Bayes* ha dei valori leggermente più bassi: 0,966 di Precision, 0,983 di F_1-measure e 0,982 di Accuracy. I valori di AUC differenziano ulteriormente i classificatori: *SMO Poly* e *SMO Puk* registrano il valore più basso 0,987, il NaïveBayes 0,991 e tutti gli altri 0,992. Questo ci permette di concludere che utilizzando la feature selection le misure di performance sono ancora più precise rispetto alle analisi precedenti e non c'è una distinzione netta nel definire il classificatore migliore.

Il dataset iniziale presenta dei missing values e nonostante tra le variabili ottimali, selezionate con la feature selection, non sia presente *stalk root,* è stato deciso di sostituire i missing values di questa variabile categoriale utilizzando la moda condizionata alla variabile risposta. Questa decisione è stata presa al fine di ridurre la distorsione che una rimozione delle intere istanze del dataset avrebbe potuto generare in quanto le analisi sono rivolte ad un dataset di contenuta numerosità campionaria. Dunque, a partire dal nuovo dataset è stata implementata la feature selection. Le variabili risultanti come ottimali sono: *odor*, *gill spacing*, *stalk surface above ring* e *veil color*.

Classificatore	Recall	Precision	F-measure	Accuracy	AUC
J48	1	0,98	0,99	0,989	0,993
Random Forest	1	0,98	0,99	0,989	0,994
Logistic	1	0,98	0,99	0,989	0,993
SMO poly	1	0,98	0,99	0,989	0,989
SMO puk	1	0,98	0,99	0,989	0,989
MLP	1	0,98	0,99	0,989	0,994
NaïveBayes	1	0,973	0,986	0,985	0,991
NBTree	1	0,98	0,99	0,989	0,994

Tabella 4 - Indicatori di performance sui diversi modelli (Feature Selection sul dataset senza missing con CfsSubsetEval)

Le misure di performance hanno subìto delle leggerissime variazioni in positivo sul dataset in cui sono stati sostituiti i missing values. A conclusione di queste analisi affermiamo che, anche in questo caso, tutti i modelli utilizzati sono performanti se si vuole classificare la commestibilità di un fungo.

1.3. Validazione e intervalli di confidenza

Date le misure di performance molto elevate si è pensato di modificare la partizione di dati. Dunque, in questa fase il dataset iniziale è stato diviso in 90% (partizione A) e 10% (partizione B), come in precedenza è stato impiegato il procedimento di stratified sampling in cui la variabile di stratificazione è *class.* Successivamente la partizione A (90% del dataset iniziale) è stata suddivisa ulteriormente in 67% dei record (A_train) e 33% dei record (A_test). I classificatori sono stati addestrati sulla partizione A_train e testati sia sulla partizione A_test sia sulla partizione B. Una volta implementato il nodo ***scorer,*** che calcola le misure di performance, è stato effettuato un confronto tra le diverse misure di Accuracy ottenute sui due dataset. Con l'obiettivo di comparare i diversi classificatori, è stato implementato un ***line plot*** che consente di raffigurare sul medesimo grafico i due livelli di Accuracy. In Figura 2 nella prima colonna è rappresentato il valore dell'Accuracy associato alla partizione A_test e nella seconda colonna quello riferito alla partizione B.

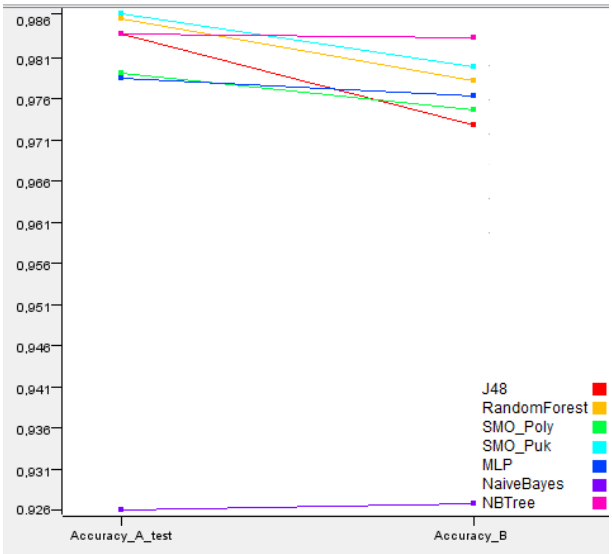

Figura 2 - Line Plot Accuracy sulle due partizioni

Si nota facilmente l'assenza di un modello migliore in assoluto: *SMO Puk* ha l'Accuracy più alta nella partizione A_test mentre per la partizione B il valore maggiore è registrato dal classificatore *NBTree*. Inoltre, ad esclusione del *Naïve Bayes* e dell'*NBTree* tutti gli altri classificatori registrano dei valori di Accuracy differenti a seconda della partizione. In Tabella 5 oltre ai valori di Accuracy è presente la variabile D costruita come differenza tra le due partizioni.

Classificatore	Accuracy_A_test	Accuracy_B	D
J48	0,984	0,973	0,011
Random Forest	0,986	0,979	0,007
Logistic	0,98	0,975	0,005
SMO poly	0,98	0,975	0,005
SMO puk	0,987	0,981	0,006
MLP	0,979	0,977	0,002
NaïveBayes	0,927	0,927	0
NBTree	0,984	0,984	0

Tabella 5 – Confronto valori di Accuracy

Pertanto, con l'obiettivo di effettuare un'analisi più approfondita su queste differenze (visibili in Tabella 5), sono stati calcolati gli intervalli di confidenza sull'Accuracy ottenuta nella prima partizione. Gli intervalli di confidenza sono stati calcolati secondo Wilson a livello di confidenza del 95%, di seguito è riportata la formula[4]:

$$\left(\frac{acc+\frac{z^2_{1-\frac{\alpha}{2}}}{2N}-z_{1-\frac{\alpha}{2}}\sqrt{\frac{acc}{N}-\frac{acc^2}{N}+\frac{z^2_{1-\frac{\alpha}{2}}}{4N}}}{1+\frac{z^2_{1-\alpha/2}}{N}}, \frac{acc+\frac{z^2_{1-\frac{\alpha}{2}}}{2N}+z_{1-\frac{\alpha}{2}}\sqrt{\frac{acc}{N}-\frac{acc^2}{N}+\frac{z^2_{1-\frac{\alpha}{2}}}{4N}}}{1+\frac{z^2_{1-\alpha/2}}{N}} \right)$$

Equation 1 - Formula dell'intervallo di confidenza di Wilson

Gli intervalli di confidenza dell'Accuracy sulla partizione A_test e il valore dell'Accuracy sulla partizione B sono stati rappresentati utilizzando i **boxplot** (Figura 3).

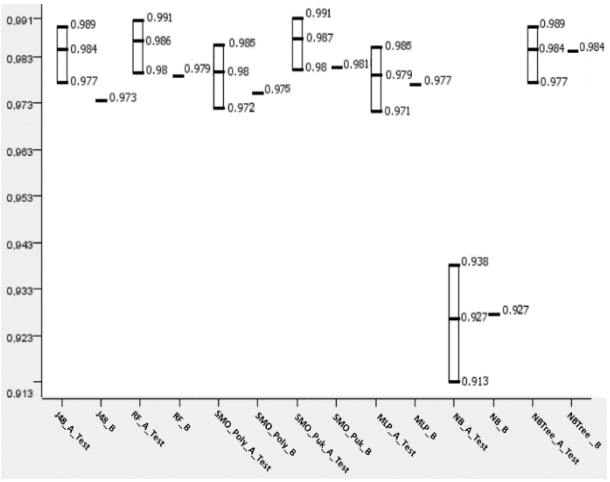

Figura 3 – Boxplot IC Accuracy A_test e Accuracy B

Come si osserva dalla Figura 3, per i classificatori *J48* e *Random Forest*, il valore di Accuracy testato sulla partizione B non ricade all'interno dell'intervallo di confidenza costruito per l'Accuracy testata sulla partizione A_test. Difatti, per questi due modelli, il **line plot** in Figura 2 restituisce le linee con pendenza più evidente indice del maggior divario tra le due accuratezze.

Inoltre, tra il modello *Logistic* e *SMO Poly*, che hanno gli stessi valori sia sulla partizione A_test sia su B, si è scelto di raffigurare solo *SMO Poly*. Concludendo, come ci aspettavamo dal ***line plot*** (Figura 2) il valore dell'Accuracy nella partizione B è compreso nell'intervallo di confidenza valutato sulla partizione A_test per i classificatori: *SMO Poly*, *SMO Puk*, *MLP*, *Naïve Bayes* e *NBTree*; al contrario del *J48* e del *Radom Forest* per i quali la variazione dell'Accuracy tra la partizione A_test e B è maggiore rispetto agli altri casi (Tabella 5). Quindi, tenendo in considerazione l'Accuracy i modelli migliori sono *NBTree* e *SMO Puk*.

2. Seconda domanda di ricerca

È noto che i funghi crescano in differenti habitat ma la maggior parte di essi è concentrata negli ambienti boschivi[5]. Il dataset in esame conferma questa teoria, infatti tra i possibili valori che *habitat* può assumere, *wood* (bosco) è la categoria che registra più osservazioni. In generale i diversi habitat nell'ecosistema favoriscono la presenza e l'abbondanza delle diverse tipologie di funghi[6]. Pertanto, nella seconda domanda di ricerca è stata classificata la numerosità dei funghi in base al territorio in cui crescono e alla loro tipologia, sfruttando le variabili *population* e *habita*t come input.

La variabile *population*, caratterizzata da sei diverse modalità (*abundant*, *clustered*, *numerous*, *scattered*, *several*, *solitary*) è stata ridotta nella sua dimensione a due categorie: *pochi* e *tanti*. Il raggruppamento è stato effettuato attraverso il nodo Knime **Rule Engine** con il seguente criterio logico: *abundant*, *numerous*, *clustered*, *several*, codificati come t (*tanti*) e *solitary* e *scattered* come p (*pochi*). Quando si considera la classificazione binaria, si è soliti identificare i casi rari come "Classe Positiva" mentre la classe maggioritaria come "Classe Negativa". Tuttavia, in questo studio, sulla base della domanda di ricerca, la classe positiva è la modalità *tanti* che costituisce la classe più numerosa. Si osserva che la nuova variabile risposta, *popclass*, non è significativamente sbilanciata nelle sue modalità: il 64% dei record (*popoclass*=tanti) contro il 36% dei record (*popoclass*=pochi).

2.1. Holdout e validation

Una volta costituita la nuova variabile risposta, *popclass*, è stato impiegato il metodo dell'Holdout: training set (67% dei record) e test set (33% dei record). I classificatori utilizzati, corrispondenti ai 5 modelli precedentemente descritti, sono: *J48*, *Logistic*, *SMO Poly*, *SMO Puk*, *Naïve Bayes*, che sono stati addestrati sul training set e validati attraverso il test set.

In Tabella 6 sono riportati i valori delle misure di performance.

Classificatore	Recall	Precision	F_1-measure	Accuracy	AUC
J48	0,876	0,685	0,769	0,67	0,7074
Logistic	0,868	0,678	0,762	0,65	0,7197
SMO poly	0,675	0,755	0,713	0,65	0,6466
SMO puk	0,876	0,685	0,769	0,67	0,5870
NaïveBayes	0,846	0,686	0,758	0,66	0,7157

Tabella 6 – Indicatori di performance sui diversi modelli

L'indicatore F_1-measure è un buon indicatore da considerare poiché rappresenta la media armonica tra Recall e Precision. Seguendo tale criterio, i modelli migliori risultano essere *J48* e *SMO Puk*, come si evince osservando gli intervalli di confidenza al 95% costruiti per l'indicatore F_1-measure (Figura 4). Tale risultato è in accordo con l'analisi dell'Accuracy.

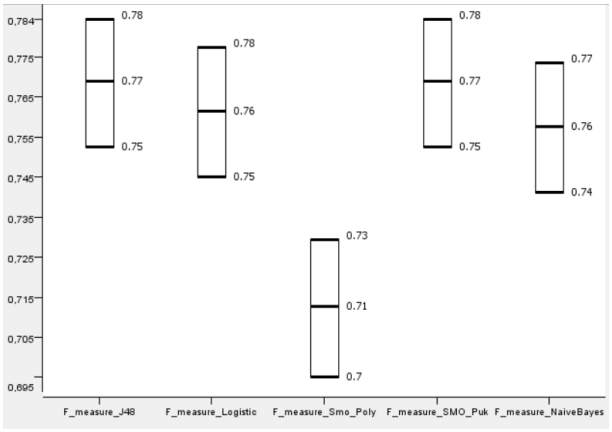

Figura 4 - Boxplot F_1-measure

Infine, per un'ulteriore valutazione della performance dei classificatori sono state analizzate e confrontate le corrispondenti curve **ROC**. Una caratteristica interessante di questa tecnica grafica è quella di non tener conto della distribuzione della variabile risposta.

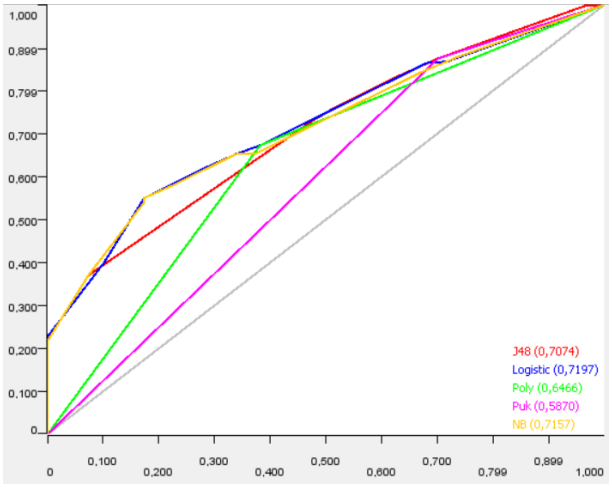

Figura 5 - ROC Curve

In Figura 5 sono rappresentate le curve **ROC** relative ai 5 classificatori presi in esame, si ritiene che tutti i modelli utilizzati sono buoni in quanto le aree sottese alle curve sono maggiori dell'area sottesa alla retta rappresentante il modello "Zero Rule" cioè il modello casuale che non aggiunge nessuna informazione. Nel nostro caso in corrispondenza del 40% di True Positive i modelli *Logistic*, *J48* e *Naïve Bayes* restituiscono all'incirca il 10% di False Positive mentre, per esempio, il modello *SMOPuk* restituisce più del 30% di istanze classificate erroneamente come positive risultando quindi meno performante (Figura 5).
L'area sotto ciascuna curva è definita AUC. In relazione a questa misura, le curve migliori risultano essere quelle dei modelli *Logistic* e *Naïve Bayes*. Tuttavia, non si osserva la presenza di un classificatore migliore in assoluto, cioè la cui curva sia sempre al di sopra delle altre.
In conclusione, tenendo in considerazione l'F_1-measure i modelli preferibili sono *J48* e *SMO Puk*; tra i due, analizzando i valori dell'AUC, è preferibile il modello J48. Difatti, il classificatore *SMO Puk* presenta un valore di AUC inferiore rispetto a tutti gli altri classificatori (Tabella 6).

Conclusioni e sviluppi futuri

In relazione al primo obiettivo, nella classificazione con metodo Holdout sono state scelte differenti variabili di input. Nel primo caso secondo la letteratura[3] sono state selezionate le variabili: *cap-shape*, *cap-surface*, *cap-color*, *odor*, *gill-size*, *gill-color*, *spore-print-color*; nel secondo caso invece: *gill-size*, *gill-spacing*, *gill-color*, *stalk-root*, *ring-type*, *spore-print-color* e *population*. Le misure di performance dei modelli di classificazione studiati presentano valori molto elevati in entrambi i casi. È fondamentale evitare i modelli che classificano i funghi velenosi come commestibili, pertanto basandoci sull'indicatore Precision, selezioniamo come modello migliore il Logistic per entrambi i casi e lo SMO Poly solo per il secondo caso. Successivamente è stata effettuata una feature selection sia sul dataset con i missing values sia sul dataset in cui sono stati sostituiti. Le variabili risultanti come ottimali sono: *odor, gill spacing, stalk surface above ring* e *veil color*. In questo caso è stato riscontrato un adattamento quasi perfetto dei modelli ai dati. Ciò ci farebbe concludere che l'utilizzo di un modello piuttosto che un altro sia del tutto analogo. Con l'obiettivo di eseguire un'analisi più approfondita, è stato nuovamente partizionato il dataset in A_train, A_test e B. I modelli migliori considerando l'Accuracy sono *NBTree* e *SMO Puk*.
Infine, relativamente al secondo obiettivo, ovvero quello di fornire una previsione sulla quantità di funghi che popolano un territorio con determinate caratteristiche, la scelta è ricaduta sul modello J48.

Il tema affrontato può suggerire ulteriori sviluppi futuri: per identificare la possibilità di intossicazioni, devono essere condotti studi più approfonditi, analisi cliniche e sperimentali. A tal proposito, tecniche di Machine Learning come la *Cluster Analysis* possono rivelarsi utili ad individuare raggruppamenti di funghi con caratteristiche simili. Inoltre, disponendo dell'informazione relativa ai vari livelli di tossicità del fungo, può essere utile effettuare un'analisi dei costi. Difatti, la scorretta classificazione comporterebbe un rischio troppo elevato e da evitare.

Referenze

1. Saviuc Ph., Garon D., Danel V., Richard J. M., *Intoxications par les cortinaires. Analyse des cas de la littérature*, Néphrologie Vol. 22 n° 4, 2001, pp. 167-173

2. https://www.kaggle.com/uciml/mushroom-classification

3. Ukwuru MU*, Muritala A. and Eze LU, *Edible and Non-Edible Wild Mushrooms: Nutrition, Toxicity and Strategies for Recognition*, Journal of Clinical Nutrition and Metabolism, 2018

4. https://www.ucl.ac.uk/english-usage/staff/sean/resources/binomialpoisson.pdf

5. Arnolds E., De Vries B. *Conservation of fungi in Europe.* In D. Pegler, L. Boddy, B. Ing, & P. M. Kirk (Eds.), *Fungi of Europe, Investigation, Recording and Mapping*, 1993, pp. 211–230

6. Singha K., Banerjee A., Pati B.R., Das P.K. Mohapatra, *Eco-diversity, productivity and distribution frequency of mushrooms in Gurguripal Eco-forest*, Paschim Medinipur, West Bengal, India, Cream Journal, 2017