ANTONIO TARANTAO MATR. 0512108187

GIUSEPPE CONTALDI MATR.  0512113326

# Tra Delizia e Pericolo: Valutazione della Commestibilità dei Funghi

Progetto Machine Learning

Anno Accademico: 2023/2024

## 1. Identificazione del Problema

Il progetto si propone di sviluppare un sistema di Machine Learning per determinare se un fungo è commestibile o velenoso utilizzando un approccio di classificazione binaria. I funghi velenosi costituiscono una minaccia per la salute umana, in quanto l'ingestione di tali funghi può provocare gravi danni e persino mettere a rischio la vita. Questo problema è di particolare rilevanza per la sicurezza alimentare e la salute pubblica, poiché la confusione tra funghi commestibili e velenosi può portare a situazioni pericolose e potenzialmente fatali. I modelli utilizzati in questo progetto sono:

- K-Nearest Neighbor (KNN)
- Decision Tree
- Random Forest

L'apprendimento di tali modelli sarà di tipo Supervisionato con Batch Learning.

## 2. Contestualizzazione dell'Importanza

Il consumo di funghi è una pratica diffusa in molte culture e può essere un'importante fonte di nutrienti e piacere culinario. Tuttavia, la somiglianza tra funghi commestibili e velenosi può rendere difficile la distinzione tra i due tipi, soprattutto per coloro che non sono esperti micologi. Di conseguenza, esiste una crescente necessità di sviluppare sistemi di supporto decisionale in grado di identificare con precisione la commestibilità dei funghi e prevenire avvelenamenti accidentali.

## 3. Descrizione dell'Esempio: l'Amanita phalloides

Un esempio emblematico di fungo velenoso è l'Amanita phalloides, noto anche come "fungo Fallico verde" o "Fallico Morte". Questo fungo è considerato uno dei più pericolosi al mondo in termini di tossicità. Caratterizzato da un cappello verde-oliva, un gambo bianco e una volva bianca alla base del gambo, l'Amanita phalloides contiene potenti agenti velenosi, tra cui l'alfa-amanitina, che può causare gravi danni al fegato e ad altri organi interni. I sintomi dell'avvelenamento da Amanita phalloides possono manifestarsi dopo diverse ore dall'ingestione e includono nausea, vomito, diarrea, crampi addominali e insufficienza epatica, che può portare alla morte se non trattata tempestivamente con un trapianto di fegato.

## 4.Descrizione del Dataset

Abbiamo utilizzato il “Mushroom Data Set” dal UCI Machine Learning Repository(“https://archive.ics.uci.edu/dataset/73/mushroom”), il dataset è stato donato al repository UCI nel 1987 e proviene dalla guida sul campo della Società Audubon per i funghi del Nord America. Contiene descrizioni di campioni ipotetici corrispondenti a 23 specie di funghi a lammelle nei generi Agaricus e Lepiota. Ogni specie è classifica come sicuramente commestibile, sicuramente velenoso o di edibilità sconosciuta e non raccomandata(quest’ultima è sta combinata con quella velenosa). Una panoramica dettagliata delle sue caratteristiche:

- **Forma del cappello (cap-shape)**: descrive la forma del cappello del fungo e può assumere i seguenti valori:
  - Bell: a forma di campana
  - Conical: conico
  - Convex: convesso
  - Flatt: piatto
  - Knobbed: a forma di pomello
  - Sunken: scavato

- **Superficie del cappello (cap-surface)**: descrive la superficie del cappello del fungo e può assumere i seguenti valori:

- Fibrous: fibrosa
- Grooves: con solchi
- Scaly: squamosa
- Smooth: liscia

- **Colore del cappello (cap-color):** indica il colore del cappello del fungo, con valori come:
  - Brown: marrone
  - Buff: beige
  - Cinnamon: cannella
  - Gray: grigio
  - Green: verde
  - Pink: rosa
  - Purple:viola
  - Red: rosso
  - White: bianco
  - Yellow: giallo

- **Ombelico(bruises):** indica se il fungo sviluppa lesioni quando viene toccato e i valori possono essere:
  - Bruises: presenta lesioni
  - No: non presenta lesioni

- **Odore(odor)**: descrive l'odore del fungo e i valori possono essere:
  - Almond:mandorla
  - Anise:anice
  - Creosote: creosoto

- Fishy: di pesce
- Foul: fetido
- Musty: muffa
- None: nessuno
- Pungent: pungente
- Spicy: speziato

E cosi via, con altre caratteristiche come la forma del gambo, la presenza di anelli, la distribuzione delle lamelle ecc. Ogni riga del dataset rappresenta un fungo e ogni colonna rappresenta una caratteristica del fungo, ed infine la colonna finale indica se il fungo è commestibile (“e” per edible) oppure velenoso (“p” per poisonous).

## 5.Manipolazione dei Dati

I dati sono categoriali, quindi utilizzeremo LabelEncoder per convertirli in ordinali. LabelEncoder è una tecnica di pre-elaborazione che converte le etichette categoriche in valori numerici.Assegna un intero unico ad ogni categoria unica nel dataset, rendendolo più adatto per gli algoritmi di machine learning.

## 6. Modelli

Dopo il preprocessiong dei dati, sono stati addestrati 3 modelli di Machine Learning per la classificazione dei funghi come commestibili o velenosi: K-Nearest Neighbor(KNN), Decision Tree e Random forest. Per ogni modello varranno indicate le seguenti metriche di prestazione:

- Precision
- Recall
- Accuracy
- F1-Score

## 6.1 Metriche di valutazione

Abbiamo valutato le prestazioni del modello utilizzando: precision, recall, accuracy e F1-Score.

- **Precision**:la precisione è il rapporto tra il numero delle previsioni corrette di un evento(classe) sul totale delle volte che il modello lo prevede.Nel nostro caso, è il numero di volte che il modello ha correttamente identificato un fungo velenoso diviso per il numero totale di volte che il modello ha previsto un fungo come velenoso.La formula generale per il calcolo del precision è la seguente:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

- Recall: Il recall, anche nota come sensibilità,è il rapporto tra le previsioni corrette per una classe sul totale dei casi in cui si verifica effettivamente. Nel nostro caso, è il numero di volte che il modello ha identificato un fungo velenoso diviso il numero totale di funghi velenosi nel dataset. La formula generale per il calcolo del recall è la seguente:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

- Accuracy: L'accuratezza è calcolata come la somma dei veri positivi e dei veri negativi divisa per il numero totale di campioni. Nel nostro esempio, è il numero di volte che il modello ha correttamente identificato sia i funghi velenosi che quelli commestibili diviso il numero totale di funghi nel dataset. La formula generale per il calcolo dell'accuracy è la seguente:

$$Accuracy = \frac{TruePositives + TrueNegatives}{TruePositives + TrueNegatives + FalsePositives + FalseNegatives}$$

- F1-Score:F1-Score è una media armonica di precisione e recall. Rispetto ad una media convenzionale, quella armonica attribuisce un peso maggiore ai valori piccoli. Questo fa si che un classificatore ottenga un alto punteggio F1 solo quando precisione e recall sono entrambi alti.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

## 6.2 K-Nearest Neighbor (KNN)

Il K-Nearest Neighbor (KNN) è un algoritmo di classificazione e regressione basato sull'idea che gli oggetti simili tendono a essere vicini tra loro nello spazio delle feature. Funziona identificando i "k" punti dati più vicini a un punto di query e utilizzando la loro maggioranza (nel caso della classificazione) o la media (nel caso della regressione) per determinare l'etichetta o il valore del punto di query.

```
KNN - Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.99      1.00       433
           1       0.99      1.00      1.00       380

    accuracy                           1.00       813
   macro avg       1.00      1.00      1.00       813
weighted avg       1.00      1.00      1.00       813

KNN - Cross Entropy: 0.1330024110299528
KNN - Training Time: 3.1556129455566406 seconds

Test Accuracy: 100.0%
```

## 6.3 Decision Tree

Un albero di decisione (Decision Tree) è un algoritmo di apprendimento supervisionato, utilizzato sia per la classificazione che per i compiti di regressione. Ha una struttura ad albero, che consiste in un nodo radice, rami , nodi interni e nodi foglia. Ogni nodo interno rappresenta l'esito del test, e ogni nodo foglia rappresenta un'etichetta di classe(decisione presa dopo il calcolo di tutti gli attributi). Gli alberi di decisione cercano di trovare la migliore divisione

per suddividere i dati. Tuttavia, possono essere soggetti a problemi come il sovradattamento. Il modello ottiene queste prestazioni:

```
Decision Tree - Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       433
           1       1.00      1.00      1.00       380

    accuracy                           1.00       813
   macro avg       1.00      1.00      1.00       813
weighted avg       1.00      1.00      1.00       813

Decision Tree - Cross Entropy: 2.2204460492503136e-16
Decision Tree - Training Time: 0.028914690017700195 seconds

Test Accuracy: 100.0%
```

## 6.4 Random Forest

Random Forest è un algoritmo di apprendimento automatico comunemente utilizzato, che combina l'output di più alberi di decisione per raggiungere un singolo risultato. E' un'estensione del metodo di bagging in quanto utilizza sia il bagging che la casualità delle caratteristiche per creare una foresta di alberi di decisione non correlati tra loro. La Random Forest può essere utilizzata sia per problemi di classificazione che di regressione.Questo algoritmo mantiene la sua accuratezza formando un insieme di alberi di decisione. Il modello ottiene queste prestazioni:

```
Random Forest - Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00       433
           1       1.00      1.00      1.00       380

    accuracy                           1.00       813
   macro avg       1.00      1.00      1.00       813
weighted avg       1.00      1.00      1.00       813

Random Forest - Cross Entropy: 2.2204460492503136e-16
Random Forest - Training Time: 0.4907515048980713 seconds
```

## 7. Predizioni

```
[0 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0 0 1]
[0 1 1 0 1 1 1 1 0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 1 1 0 0 0 1]
```

Come possiamo vedere, i valori previsti e veri corrispondono al 100%.

## 8. Comparazione dei Modelli

| | KNN | DECISION TREE | RANDOM FOREST |
|---|---|---|---|
| **ACCURACY** | 1.0 | 1.0 | 1.0 |
| **TRAINING TIME** | 3.15s | 0.02s | 0.49s |

Possiamo notare come il Decision Tree abbia il tempo più basso.

### 8.1. Cross-Entropy

E’ essenziale eseguire verifiche sulla solidità del modello, impiegando metriche più avanzate rispetto alla semplice accuracy. Qui entra in gioco la Cross-Entropy Loss Function, che è concepita come una misura della distanza tra due distribuzioni di probabilità. In dettaglio, calcola la somma

delle probabilità di ogni evento nella distribuzione dei dati reali(le classi),moltiplicate per il logaritmo delle probabilità di ogni evento nella distribuzione prevista. Il suo principale beneficio è la capacità di valutare statisticamente la 'qualità' del modello. In Python, la Cross-Entropy Loss Function può essere calcolata utilizzando la funzione log_less della libreria sklearn.metrics

| | KNN | DECISION TREE | RANDOM FOREST |
|---|---|---|---|
| CROSS ENTROPY | 0.1330024110299528 | 2.2204460492503136e-16 | 2.2204460492503136e-16 |

Possiamo notare come il modello KNN mostra un comportamento diverso rispetto ai modelli Decision Tree e Random Forest. In particolare, la cross entropy per KNN è diversa rispetto agli altri due modelli.

## 9.Conclusioni

Nel corso di questo progetto, abbiamo applicato tre diversi modelli di machine Learning (KNN, Decison Tree e Random Forest) per determinare se un fungo è commestibile oppure velenoso. Ogni modello ha dimostrato le proprie peculiarità e performance. Abbiamo utilizzato la Cross Entropy Loss Function per misurare l'accuratezza di ciascun modello, fornendo una valutazione oggettiva della loro capacità di prevedere correttamente la commestibilità dei funghi. I risultati hanno evidenziato l'importanza di scegliere il modello in base al tipo di dati e al tipo di problema. Hanno evidenziato anche la corretta valutazione del modello, non solo in termini di accuratezza, ma anche considerando altre metriche come la Cross Entropy. In conclusione, questo progetto ha dimostrato come il machine learning possa essere un potente strumento per risolvere problemi complessi. L'obiettivo finale è di costruire un modello che possa prevedere con alta precisione e affidabilità la commestibilità di un fungo, contribuendo così alla sicurezza e alla conoscenza nel campo della micologia.