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:

* Decision Tree
* Random Forest

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

1. 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.

1. 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.

1. 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).

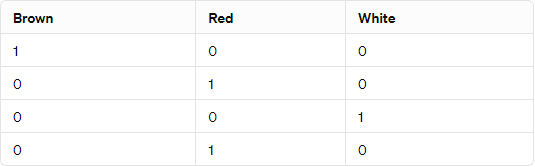
1. Preprocessing dei Dati

Dopo aver esaminato il dataset Mushroom, è stato necessario eseguire alcune operazioni di preprocessing per preparare i dati per l'addestramento dei modelli di Machine Learning. Le seguenti operazioni sono state eseguite:

* **Gestione dei valori mancanti**: Il dataset presenta valori mancanti,Per gestire i dati mancanti nel dataset, abbiamo optato per una semplice eliminazione delle righe che presentavano valori mancanti. Questo approccio è stato scelto poiché solo una piccola percentuale di righe presentava dati mancanti e l'eliminazione di queste righe non ha compromesso significativamente le dimensioni complessive del dataset.
* **Encoding delle variabili categoriche**: Le variabili categoriche nel dataset sono state convertite in formato numerico utilizzando tecniche di encoding come il one-hot encoding. Questo è stato fatto per consentire ai modelli Machine Learning di elaborare correttamente le caratteristiche categoriche durante l’addestramento. One-hot enconding è una tecnica utilizzata per gestire variabili categoriche trasformandole in una forma che può essere utilizzata più efficacemente dai modelli di Machine Learning. In questa tecnica, ogni valore unico presente nella variabile categorica viene trasformato in una nuova variabile binaria. Ad esempio, consideriamo la variabile categorica “colore del cappello” nel dataset, che può assumere valori come “brown”, “red”, “white”, ecc. Con il one-hot encoding, questa variabile viene suddivisa in diverse colonne binarie, una per ogni valore unico. Quindi, se un fungo ha il cappello marrone, la colonna corrispondente alla variabile “brown” avrà valore 1, mentre le altre colonne avranno valore 0.



Dopo l’applicazione del one-hot encoding, la variabile “Colore del cappello” viene suddivisa in tre nuove colonne binarie:



In questo modo, ogni valore unico della variabile categoria originale ha una sua colonna binaria corrispondente nel dataset codificato.

* **Split del Dataset:** Il dataset è stato diviso in set di addestramento e di test utilizzando una proporzione del 70%-30%. Il set di addestramento è stato utilizzato per addestrare i modelli, mentre il set di test è stato utilizzato per valutare le performance dei modelli su dati non visti durante l’addestramento

1. Modelli

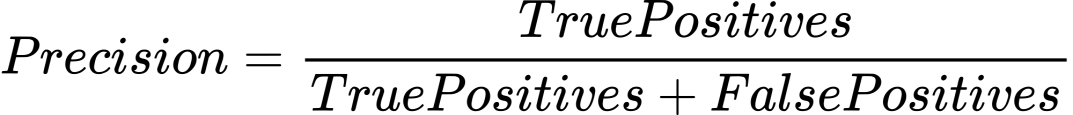
Dopo il preprocessiong dei dati, sono stati addestrati 2 modelli di Machine Learning per la classificazione dei funghi come commestibili o velenosi: Decision Tree e Random forest. Per entrambi i modelli varrà indicato il tempo di addestramento e le seguenti metriche di prestazione:

* Precision
* Recall
* Accuracy
* F1-Score

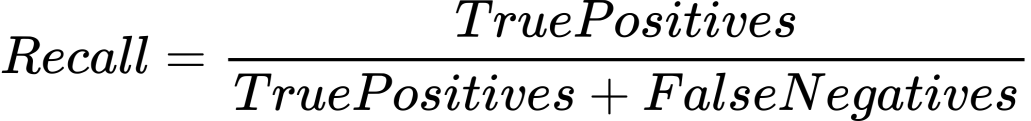
6.1 Metriche di valutazione

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

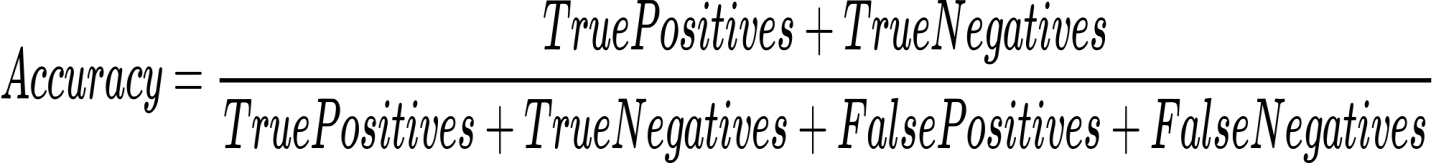
* **Precision**: la precisione misura la proporzione di istanze positive predette correttamente rispetto a tutte le istanze positive predette dal modello. In sostanza, indica quanto è accurato il modello quando predice che un'istanza è positiva. La formula è la seguente:



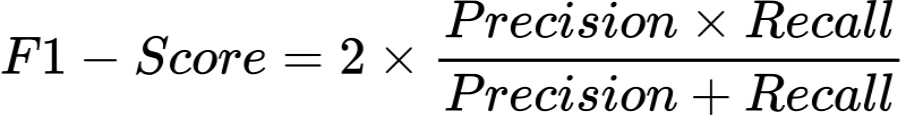
* Recall: Il recall, anche nota come sensibilità, la proporzione di istanze positive correttamente identificate dal modello rispetto a tutte le istanze positive effettivamente presenti nel dataset. In altre parole, indica quanto è bravo il modello a catturare tutte le istanze positive. La formula è la seguente:



* Accuracy: L’accuratezza misura la proporzione di istanze correttamente classificate dal modello rispetto a tutte le istanze nel dataset. In sostanza, indica quanto è bravo il modello nel fare previsioni corrette in generale. La formula è la seguente:



* F1-Score: il punteggio F1 è la media armonica di precisione e recall. Questa metrica è utile quando si desidera trovare un equilibrio tra precisione e recall, specialmente se ci sono sbilanciamenti nelle classi. La formula è la seguente:



6.2 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:

FOTO DEL RISULTATO DEL DECISON TREE

6.3 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:

FOTO DEL RISULTATO DEL RANDOM FOREST

1. Comparazioni Modelli

INSERIRE LE 2 FOTO RISULTATO DEI DUE MODELLI

Breve considerazioni dei risultati ottenuti

8.Implementazione Python