Studente: Meleleo Danilo

Matricola: 676436

E-mail: d.meleleo3@studenti.uniba.it

* **Introduzione**

Il sistema progettato consiste essenzialmente nell'applicazione di una serie di tre classificatori

di tipo Content-based, trattasi quindi di apprendimento supervisionato.

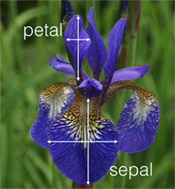
* **Algoritmi utilizzati**

Nella fattispecie i tre classificatori utilizzati sono il K-Nearest-Neighbor, che effettua le sue predizioni basandosi esclusivamente sulle previsioni già date agli elementi che più gli si avvicinano secondo il criterio della similarità, appunto i suoi "vicini", l'Albero di Decisione, che partendo dalla radice valuta ogni condizione incontrata

e segue l'arco corrispondente al risultato fino ad arrivare ad una foglia che conterrà la stima puntuale della classe e il Support Vector Machine con una funzione kernel lineare che rappresenta gli esempi come punti nello spazio, mappati in modo tale che gli esempi appartenenti alle diverse categorie siano chiaramente separati da uno spazio il più possibile ampio, i nuovi esempi sono quindi mappati nello stesso spazio e la predizione della categoria alla quale appartengono viene fatta sulla base del lato nel quale ricade.

* **Apprendimento**

Il dataset scelto per questo sistema si chiama Iris, ogni tupla si compone di 4 caratteristiche morfologiche del fiore ovvero: lunghezza e ampiezza del sepalo e lunghezza e ampiezza del petalo in cm', nell'ultima colonna invece vi è la classe di appartenenza del fiore, la quale è ovviamente influenzata dalle 4 caratteristiche precedenti(in questo caso ci sono 3 differenti tipi di iris: Setosa, Versicolour, e Virginica ).



[Questa foto](https://stats.stackexchange.com/questions/198577/simple-knn-example) di Autore sconosciuto è concesso in licenza da [CC BY-SA](https://creativecommons.org/licenses/by-sa/3.0/)

Sono presenti nel dataset un totale di 150 esempi clasificati.

* **Processo di sviluppo**

1. Prima di tutto divido il dataset in 2 parti, una parte (Training-set) la utilizzo per addestrare i classificatori e la restante parte viene utilizzata per testare questi ultimi e le loro performance (Test-set).
2. Dopo valuto le performance dei classificatori, tramite il calcolo dell'accuratezza di ciascuna predizione e il confronto di questa con il rispettivo valore reale.
3. I valori prodotti verranno poi utilizzati per calcolare per ciascun classificatore la rispettiva curva di precisione-richiamo, in modo da poterle poi mettere insieme su un grafico ed avere un confronto visivo.

I grafici sono stati integrati con la metrica AUC(Area Under the Curve) che calcola come suggerisce il nome, l'area sottesa alla curva di precisione-richiamo, che fornisce quindi un ulteriore e più preciso confronto fra i classificatori (area > indica precisione più alta, 1 = precisione assoluta).

* **Altre funzioni**

E' possibile inoltre, data una nuova tupla di caratteristiche del fiore, scelta dall'utente, utilizzare i classificatori per vedere la sua classe di appartenenza.

* **Librerie utilizzate**

Per questo sistema sono state utilizzate alcune librerie, come ad esempio sklearn che permette la definizione e gestione di classificatori di questo tipo in modo abbastanza intuitivo, e pylab per la costruzione del grafico per la curva, altri metodi e metriche necessarie per il tutto sono state scritte manualmente integrando la conoscenza teorica anche di

altre discipline.