Esempio: Classificazione Fagioli

Anonimo

***Abstract*—Data la similitudine tra le diverse specie di fagioli, la necessità di un sistema di classificazione efficace e preciso per l'identificazione e lo stoccaggio degli stessi che si sostituisca all'occhio umano è fondamentale. La ricerca in oggetto, dunque, si concentra sullo sviluppo e sulla validazione di un metodo di classificazione multiclasse in grado di riconoscere la varietà fra sette diversi tipi di fagioli secchi, basandosi sulle loro caratteristiche morfologiche come forma, tipo e struttura.**

I. Introduzione

L'industria agroalimentare è un settore che richiede una costante ottimizzazione dei processi per poter rispondere alla crescente domanda. Una delle sfide più significative è la classificazione efficace e rapida dei prodotti agricoli. Tra questi, i fagioli secchi, difficilmente classificabili a causa della numerosità delle specie e della somiglianza tra di esse.
L'occhio umano è tradizionalmente usato per la classificazione, ma ha i suoi limiti in termini di velocità, efficienza e precisione. Inoltre, la classificazione manuale può essere un processo laborioso ed economicamente oneroso. Questa necessità ha spinto allo sviluppo di un sistema automatico di classificazione dei fagioli secchi.
In questa ricerca, abbiamo individuato un algoritmo di classificazione multiclasse che si basa sulle caratteristiche morfologiche dei fagioli, quali la forma, il tipo e la struttura.
Il report descrive il processo di validazione e scelta del metodo di classificazione. In particolare, discuteremo in dettaglio l'approccio utilizzato, i dati raccolti, l'addestramento dell'algoritmo e l'efficacia del sistema nel distinguere le sette diverse tipologie di fagioli secchi.

II. Analisi dei dati

Il data set utilizzato in questa ricerca è frutto di un'indagine condotta su sette diversi tipi di fagioli secchi. Attraverso il sistema di visione artificiale, sono state acquisite le immagini dei fagioli, le quali sono state successivamente sottoposte a fasi di segmentazione ed estrazione delle caratteristiche. Complessivamente, sono state estratte 16 features dalle immagini dei fagioli. Il dataset di training, composto da 2926 record distribuiti in modo omogeneo tra le classi, è stato sottoposto an una fase preliminare di pre-elaborazione dei dati. Inizialmente, è stata condotta un'analisi esplorativa al fine di individuare la presenza di valori mancanti o duplicati nella collezione. Tuttavia, questa analisi iniziale non ha rilevato alcun valore mancante o duplicato, perciò non è stata necessaria l'applicazione di tecniche specifiche per gestirli.
Successivamente, si è voluto indagare sulla presenza di outlier, ovvero valori anomali che si discostano significativamente dalla distribuzione dei dati. Per identificarli, sono state effettuate analisi attraverso i software "Microsoft Excel" e "Rapid Miner". Dalle rappresentazioni grafiche mostrate nelle Figure 1 e 2, è possibile osservare la presenza di valori che si discostano in modo significativo dalla distribuzione generale dell'attributo *Area*. L'analisi è stata effettuata per ciascuna feature e, con l'utilizzo un filtro personalizzato, sono state rimosse le istanze contenenti outlier dal database.

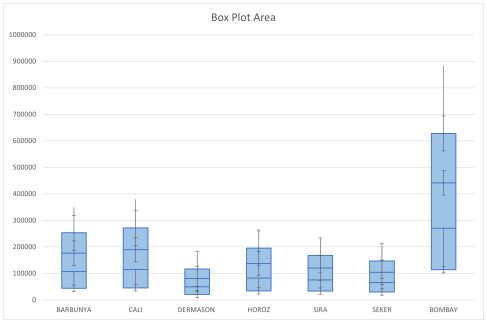

Fig. 1. Diagramma Box Plot per l'attributo Area

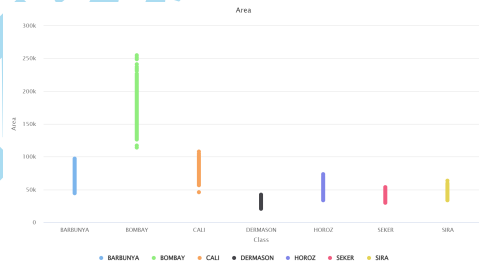

Fig. 2. Scatter Plot per l'attributo Area

È stata anche valutata la possibilità di discretizzare i parametri *Area, ConvexArea, EquivDiameter, MajorAxisLength, MinorAxisLength, Perimeter*. Tuttavia, dopo aver valutato diversi algoritmi euristici (tabella I) e condotto prove sul software "Rapid Miner", si è constatato che la discretizzazione non portava nessun miglioramento all'accuratezza dei risultati .
In ultimo, al fine di garantire un'elaborazione adeguata dei dati e l'utilizzo corretto degli algoritmi Support Vector Machines (**SVM**), k-Nearest Neighbors (**kNN**) e Multi-Layer perceptron (**MLP**), `e stata introdotta la normalizzazione con una *ztransformation*.

TABLE I. Euristiche discretizzazione

Method	Area	Perimeter	MajorAxisLength
Sturges's Formula	12	12	12
Square-root Choice	47	47	47
Scott's Choice	18	18	18
Freedman-Diaconis choice	38	26	25
Optimal (Loss-Function min)	46	40	44

III. Metodologia

La ricerca ha seguito un processo strutturato ed è iniziata con un'esaminazione accurata dei dati disponibili sui fagioli. Il primo passo effettuato è stata un'analisi esplorativa dei dati (*EDA*), durante la quale abbiamo verificato l'integrità di questi, cercando eventuali valori mancanti o outlier. Effettuare questo tipo di analisi è fondamentale per garantire la qualità dei dati, dato che eventuali irregolarità possono influenzare l'efficacia degli algoritmi di classificazione.
Una volta confermata la qualità dei dati, abbiamo proceduto con la scelta degli algoritmi utilizzati. La nostra attenzione si è focalizzata su cinque diversi algoritmi di classificazione:

- Decision Tree (**DT**), utili per identificare relazioni e gerarchie all'interno dei dati ottenendo risultati di facile interpretabilità e basso costo computazionale
- Random Forest (**RFT**), scelto per la capacità di di ottenere previsioni precise e stabili. Il modello risulta robusto ed è in grado di gestire variabilità e rumore nei dati
- Support Vector Machines (**SVM**), utile perchè trova un iperpiano di separazione ottimale che massimizzi la separazione tra le classi e consente di classificare correttamente nuove immagini dei fagioli
- k-Nearest Neighbors (**kNN**), scelto per creare un modello focalizzato sull'individuare pattern basati sulla vicinanza dei dati
- Multi-layer Perceptron (**MLP**), utilizzato per creare un modello complesso che apprende le relazioni non lineari tra le variabili ambientali e la concentrazione di inquinanti

Scelti gli algoritmi ed esplorati i dati, si è proceduto con la valutazione della *feature selection* come parte del processo di pre-elaborazione dei dati per ridurre la dimensionalità del nostro dataset e selezionare solo le caratteristiche più rilevanti. Tuttavia, dopo un'analisi dettagliata e una serie di esperimenti, abbiamo deciso di non utilizzare la feature selection per i seguenti motivi.

Innanzitutto, abbiamo notato che alcuni degli algoritmi di machine learning che abbiamo scelto di utilizzare, come Random Forest e Support Vector Machines, hanno funzionalità di feature selection incorporate all'interno del loro processo di apprendimento. Questi algoritmi sono in grado di valutare l'importanza delle feature durante l'addestramento del modello e di assegnare loro un peso adeguato nella fase di decisione. Pertanto, l'applicazione di ulteriori tecniche di feature selection potrebbe risultare ridondante e non apportare benefici significativi alle performance del modello.

Per gli algoritmi che invece non avevano la feature selection incorporata, abbiamo condotto diversi esperimenti confrontando le prestazioni dei nostri modelli con e senza l'applicazione della feature selection. Abbiamo scoperto che, nel nostro caso specifico, l'utilizzo della feature selection non ha portato a un miglioramento significativo delle performance predittive.

Il passo successivo è stato la vera e propria costruzione del modello. Per ogni algoritmo, abbiamo prima addestrato

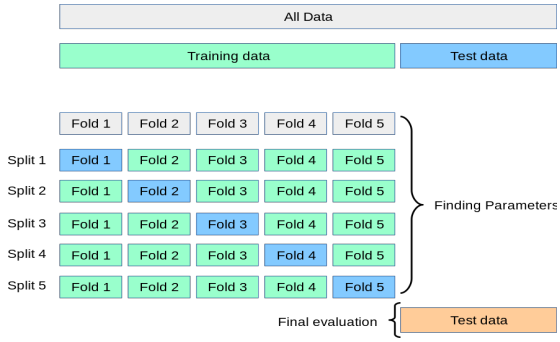

Fig. 3. Metodologia utilizzata

e validato il modello con l'utilizzo della *cross-validation* sul dataset di training. La *cross-validation* prevede che il dataset venga diviso in k sottoinsiemi disgiunti. Di questi, k-1 sono utilizzati per fare il training e il restante è utilizzato per fare il test. Tutto ripetuto per k volte. La tecnica è stata scelta poiché aiuta a prevenire l'overfitting e fornisce una stima più accurata delle prestazioni del modello.

Infine, utilizzando il dataset di test, abbiamo osservato le performance dei modelli validati in precedenza tramite il calcolo delle metriche di prestazione quali *accuratezza*, *precisione*, *recall* ed *F1 score*. Confrontare i valori delle stesse misure di performance su diversi algoritmi ci ha permesso di selezionare il modello più efficace per la nostra classificazione. Una schematizzazione del nostro approccio è riportata in figura 3.

Di seguito una breve descrizione del funzionamento degli algoritmi selezionati per la nostra ricerca.

A. Decison Tree (DT)

Un albero decisionale è uno strumento di supporto decisionale che utilizza un grafico o un modello simile ad un albero per mostrare le sue decisioni e i possibili risultati. DT viene utilizzato per determinare il percorso da seguire in un'analisi decisionale. L'algoritmo ha i vantaggi di essere intuitivo e di facile comprensione per alberi di piccola dimensione, essere economico nell'installazione, di facile integrazione con i sistemi di database e avere una buona affidabilità.

È stato selezionato come numero massimo di partizioni 9 e come criterio di partizionamento l'information gain.

B. Random Forest (RFT)

Random Forest è un popolare algoritmo di apprendimento automatico che appartiene alla tecnica di apprendimento supervisionato. Può essere utilizzato sia per problemi di classificazione che di regressione in ML. Si basa sul concetto di apprendimento d'insieme, che è un processo di combinazione di più classificatori per risolvere un problema complesso e migliorare le prestazioni del modello. Dato un dataset di training, questo viene suddiviso in n insiemi randomici. Per ogni insieme si genera un decision tree. Ogni modello genera un'etichetta di predizione, la predizione più frequente tra le n generate è quella che viene restituita. Un maggior numero di

alberi porta a una maggiore precisione e previene il problema dell'overfitting.

È stato selezionato come numero di alberi 80.

C. K-nearest neighbours (kNN)

L'algoritmo kNN è uno dei metodi di pattern recognition che classifica gli oggetti sulla base degli esempi educativi più vicini nello spazio degli attributi. L'algoritmo è costituito da tre diverse fasi:

- Fase di apprendimento: lo spazio viene partizionato in regioni in base alle posizioni e alle caratteristiche degli oggetti di apprendimento.
- Calcolo della distanza: ai fini del calcolo della distanza gli oggetti sono rappresentati attraverso vettori di posizione in uno spazio multidimensionale. Esistono modi diversi per il calcolo delle distanze: le più comunemente utilizzate sono la distanza euclidea e la distanza Manhattan. L'algoritmo è sensibile alla struttura locale dei dati.
- Fase di classificazione: un punto (che rappresenta un oggetto) è assegnato alla classe C se questa è la più frequente fra i k esempi più vicini all'oggetto sotto esame. I vicini sono presi da un insieme di oggetti per cui è nota la classificazione corretta.

È stato selezionato come numero di vicini 7 e come distanza la Kernel Euclidean Distance.

D. Support vector machine (SVM)

Support Vector Machines (SVM) è un metodo basato su kernel con elevata potenza computazionale per problemi di classificazione e regressione. Rispetto ad altri metodi di apprendimento automatico, SVM ha una migliore generalizzazione, ha una solida base teorica e fornisce risultati più accurati in molte applicazioni rispetto ad altri algoritmi. Gli SVM sono modelli di classificazione il cui obiettivo è quello di trovare la retta di separazione delle classi che massimizza il margine tra le classi stesse, dove con margine si intende la distanza minima dalla retta ai punti delle due classi. Gli SVM possono anche classificare i dati non lineari spostando i dati in una dimensione maggiore con un metodo chiamato trucco del kernel. Gli SVM sono in grado di rappresentare problemi complessi e sono resistenti all'overfitting. Progettato originariamente per la classificazione di dati lineari di classi binarie, il metodo è stato poi implementato per la classificazione di dati di classi multiple e non lineari.

E. Multi-layer Perceptron (MLP)

Gli MLP sono sistemi informatici in grado di apprendere eventi utilizzando esempi e determinare come vengono generate le risposte agli eventi dall'ambiente. Simili alle caratteristiche funzionali del cervello umano, vengono applicate con successo in aree quali l'apprendimento, l'associazione, la classificazione, la generalizzazione, l'identificazione delle caratteristiche e l'ottimizzazione. MLP crea le proprie esperienze con le informazioni ottenute dai campioni e quindi prende decisioni simili su questioni simili.

La struttura della rete MLP utilizzata in questo studio è:

- Livello di input: i parametri utilizzati come parametri di input sono stati 17.
- I training cycles, rappresentano il numero di volte che il set di addestramento viene presentato all'algoritmo di apprendimento, per i quali si è scelto un valore di 12.
- Il number of generations, indica il numero di iterazioni di addestramento completate, per i quali si è scelto un valore di 12.
- Il number of ensemble MLPs rappresenta il numero di reti neurali indipendenti combinate insieme per ottenere una previsione finale. Si è scelto un valore di 5.
- Livello di output: il livello di output è composto da 7 tipi di fagioli: Seker, Barbunya, Bombay, Cali, Dermosan, Horoz e Sira.

IV. Risultati sperimentali

Nel contesto dello studio, è stato utilizzato lo strumento *Optimize Parameters* del software Rapid Miner per risalire ai migliori parametri dei diversi modelli di classificazione al fine di ottenere prestazioni ottimali. Questo strumento ha consentito di esplorare in modo efficiente diverse combinazioni di parametri e valutare le loro influenze sulle performance del modello.

TABLE II. DT Confusion Matrix

predict	true Seker	true Barb.	true Bombay	true Cali	true Horoz	true Sira	true Derm.
Seker	491	9	0	1	0	5	3
Barbunya	5	483	0	23	6	6	0
Bombay	0	0	522	1	0	0	0
Cali	0	12	0	493	10	1	0
Horoz	2	7	0	3	489	16	5
Sira	16	11	0	1	10	459	28
Dermason	8	0	0	0	7	35	486

TABLE III. RFT Confusion Matrix

predict	true Seker	true Barb.	true Bombay	true Cali	true Horoz	true Sira	true Derm.
Seker	498	1	0	1	0	7	5
Barbunya	2	502	0	17	4	4	0
Bombay	0	0	522	1	0	0	0
Cali	0	10	0	499	7	0	0
Horoz	0	0	0	4	489	6	0
Sira	17	9	0	0	15	486	8
Dermason	5	0	0	0	7	19	509

TABLE IV. kNN Confusion Matrix

predict	true Seker	true Barb.	true Bombay	true Cali	true Horoz	true Sira	true Derm.
Seker	483	4	0	1	0	9	11
Barbunya	8	468	0	14	2	4	4
Bombay	0	0	522	0	0	0	0
Cali	0	32	0	502	39	6	1
Horoz	0	0	0	3	466	6	0
Sira	19	18	0	2	11	463	45
Dermason	12	0	0	0	4	34	461

TABLE V. SVM Confusion Matrix

predict	true Seker	true Barb.	true Bombay	true Cali	true Horoz	true Sira	true Derm.
Seker	474	1	0	1	0	6	6
Barbunya	4	482	0	13	0	6	0
Bombay	14	15	522	1	39	5	11
Cali	0	13	0	503	10	3	0
Horoz	0	0	0	4	459	7	0
Sira	18	11	0	0	11	464	37
Dermason	12	0	0	0	3	31	468

TABLE VI. MLP Confusion Matrix

predict	true Seker	true Barb.	true Bombay	true Cali	true Horoz	true Sira	true Derm.
Seker	485	2	0	1	0	6	8
Barbunya	10	481	0	14	2	4	1
Bombay	0	0	522	0	0	0	0
Cali	0	28	0	501	12	2	0
Horoz	0	1	0	4	487	8	1
Sira	17	10	0	2	9	452	29
Dermason	10	0	0	0	12	50	483

Nelle tabelle dalla II alla VI, sono riportate le Confusion Matrix per le metodologie sopra citate. Le matrici di confusione per ogni modello forniscono informazioni preziose sulla capacità di ciascun modello di classificare correttamente i diversi tipi di fagioli. In particolare, queste matrici evidenziano non solo il numero di classificazioni corrette, ma anche i tipi di errori commessi dai diversi modelli, offrendo quindi una panoramica dettagliata dei punti di forza e di debolezza di ciascuno. È importante notare che tutti i modelli di classificazione hanno mostrato una minore capacità di distinguere la varietà di fagioli Sira rispetto alla varietà Dermason. Questa difficoltà può essere attribuita alla somiglianza tra le caratteristiche di piattezza e rotondità delle varietà Dermason e Sira.
Nella Table VII, invece, sono messi a confronto per ciascun modello le seguenti misure di performance: accuracy, classification error, recall, precision, relative error, F1-Score e correlation. Grazie all'utilizzo dell'*Optimize Parameters* si sono ottenuti i migliori risultati possibili nelle performance dei modelli, garantendo una maggiore efficacia e affidabilità nelle previsioni. Esaminando la Table VII si nota come tutti i metodi di classificazione abbiano un tasso di successo al di sopra del 90%. In particolare il modello RFT presenta la percentuale di accuratezza più alta, con il 95,92%. Inoltre, si nota come il modello RFT abbia i valori migliori su tutte le metriche di performance prese in considerazione.

La percentuali di accuratezza per i cinque metodi sulle singole tipologie di fagioli sono riportate in Fig. 4. Si può notare che la varietà Bombay viene interamente classificata quasi con il 100% di accuratezza (ad esclusione del modello SVM), a differenza invece della varietà Sira che risulta essere quella leggermente più complessa da classificare. Anche dal grafico si nota come in media il metodo RFT sia quello che ottiene delle prestazioni migliori, considerando tutte e sette le tipologie di fagiolo.

V. Conclusioni

Questo sistema di classificazione automatica dei fagioli secchi può essere integrato nei processi di produzione e distribuzione nel settore agroalimentare, specialmente negli stabilimenti specializzati. Il sistema pu`o essere utilizzato per il confezionamento, riducendo significativamente la dipendenza dalla manodopera e consentendo una consegna rapida e di alta qualit`a sul mercato. Grazie alla sua affidabilit`a nella classificazione, il sistema esamina l'intera superficie dei fagioli senza la necessit`a di selezione manuale, contribuendo a ridurre i costi di produzione. Ci`o permette alle aziende di soddisfare le esigenze della grande distribuzione organizzata e dei consumatori finali, migliorando l'immagine e la credibilit`a dell'azienda. In definitiva, l'integrazione di questo sistema di classificazione automatizzato pu`o ottimizzare i processi di produzione e distribuzione dei fagioli secchi, consentendo una maggiore efficienza e una migliore gestione delle risorse.

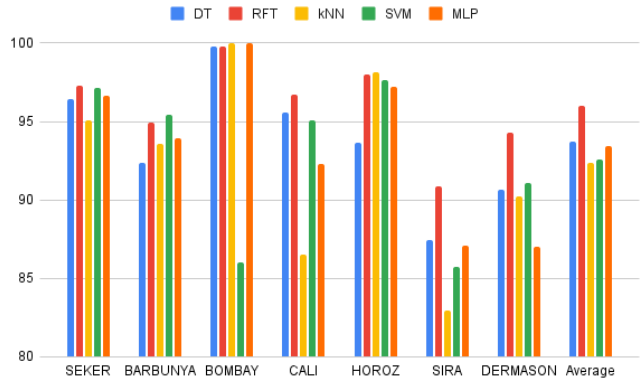

Fig. 4. Accuratezza dei modelli di classificazione per tutte le varietà di fagiolo

TABLE VII. Performance values obtained for DT, RFT, kNN, SVM, MLP

	DT	**RFT**	kNN	SVM	AutoMLP
Accuracy (%)	93,68	**95,92**	92,09	92,28	93,35
Classification Error (%)	6,32	**4,08**	7,91	7,72	6,65
Recall (%)	93,68	**95,92**	92,09	92,28	93,35
Precision (%)	93,71	**95,97**	92,36	92,59	93,45
Relative error (%)	7,66	**6,97**	10,23	7,72	8,39
F1-Score (%)	93,69	**95,94**	92,22	92,43	93,40
Correlation	0,943	**0,952**	0,917	0,924	0,936