

Report

OrangeSelection



UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

Codice del corso: 063507

Nome del progetto: OrangeSelection

Docente: Nicola Fanizzi

Data: 30 gen 2025

INFO	Sviluppatore
Nome	Domenico
Cognome	Cistulli
Matricola	762934
Corso	Ingegneria della Comunicazione

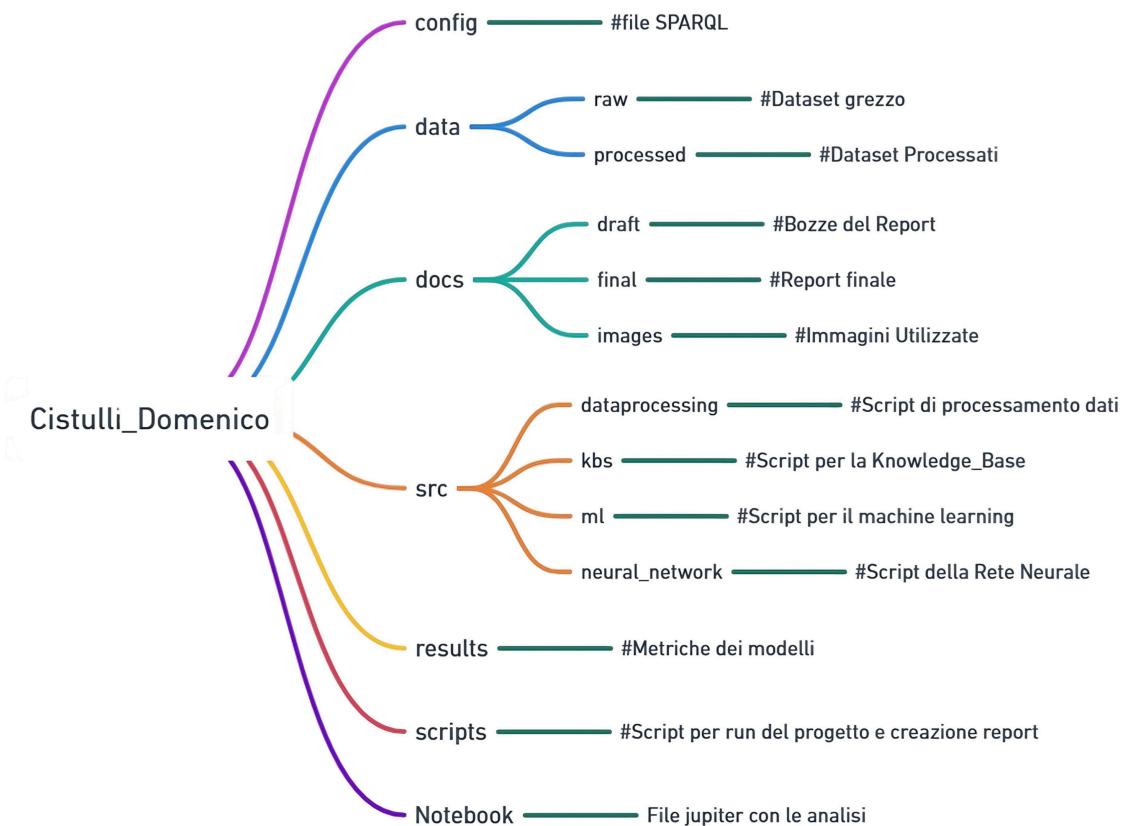
Introduzione

Il progetto **OrangeSelection** si propone di sviluppare un sistema basato su **Machine Learning** (ML) integrato con una **Knowledge Base** (KB), per analizzare la qualità delle arance. Il sistema sfrutta tecniche di **ragionamento automatico** per migliorare la predizione della qualità dei frutti, utilizzando conoscenza semantica proveniente dal **Web Semantico**.

Contesto

Il progetto **OrangeSelection**, sviluppato da **Domenico Cistulli** come prova pratica del corso di **Ingegneria della Conoscenza** A.A. 2023-2024. Questo progetto integra concetti chiave come il Machine Learning, l'uso delle ontologie e l'interrogazione semantica tramite il Web Semantico. Le scelte progettuali, incluse le tecniche di apprendimento automatico e l'uso di una Knowledge Base, sono state ispirate dai temi trattati a lezione e arricchite da approfondimenti tratti da fonti esterne.

Struttura del Progetto



Scopi e Obiettivi

SCOPI	OBIETTIVI
Creazione di un sistema predittivo per la selezione delle arance.	Preprocessing dei dati, elaborazione e bilanciamento dei dati, essenziali per migliorare l'accuratezza dei modelli di machine learning
Ottimizzazione della selezione delle arance per l'industria alimentare	Apprendimento automatico per la classificazione degli agrumi
Automatizzazione del processo di classificazione per migliorare l'efficienza	Generazione di report e analisi che sintetizzino le performance dei vari modelli

Strumenti adottati

Il progetto è stato sviluppato utilizzando i seguenti strumenti e librerie:

- **Python:** Linguaggio principale per l'implementazione del progetto. Grazie alla vasta disponibilità di librerie per l'analisi dei dati e il machine learning, Python è la scelta ideale per questo tipo di progetto.
- **Librerie Python:**
 - **Pandas:** Per la gestione e la manipolazione dei dati, come la pulizia, la gestione dei missing values e la trasformazione delle colonne.
 - **Scikit-learn:** Per l'implementazione dei modelli di machine learning come SVM, Random Forest, e per la valutazione delle performance dei modelli.
 - **TensorFlow / Keras:** Per la creazione e l'addestramento della rete neurale.
 - **Matplotlib / Seaborn:** Per la visualizzazione dei dati e la creazione di grafici utili all'analisi.
- **Weka:** Utilizzato per il pre-processing avanzato dei dati, per la discretizzazione delle variabili e per l'apprendimento della struttura dei modelli probabilistici.

Fasi del progetto

1. Preprocessing dei dati

Il dataset originale conteneva diverse colonne che non erano utili per il nostro scopo. Ecco come è stato trattato il dataset:

1. **Pulizia del dataset:** Sono stati rimossi valori nulli o errati. Ad esempio, tutte le righe che non contenevano un valore valido per la colonna "diameter" sono state eliminate.
2. **Trasformazione delle variabili:** Le variabili numeriche sono state convertite nei giusti tipi di dati, ad esempio, la colonna "diameter" è stata convertita in formato numerico.
3. **Gestione dei valori mancanti:** I valori mancanti sono stati riempiti con la media per le colonne numeriche e con il valore più frequente per le colonne categoriche utilizzando un Simple Imputer.
4. **Bilanciamento dei dati:** Poiché alcune classi erano sbilanciate, sono state applicate tecniche di bilanciamento come l'oversampling della classe minoritaria.

2. Modelli di Machine Learning

1. **Random Forest:** Abbiamo utilizzato Random Forest per la sua robustezza e capacità di gestire variabili numeriche e categoriche. Il modello è stato addestrato utilizzando le principali caratteristiche come "diameter", "weight", "red", "green", etc.
2. **Support Vector Machine (SVM):** L'algoritmo SVM è stato utilizzato per creare una linea di separazione ottimale tra le diverse classi di agrumi. La scelta di kernel RBF (Radial Basis Function) è stata fatta in base ai risultati ottenuti durante la fase di tuning.
3. **Rete Neurale:** Si è progettata e addestrata una rete neurale per classificare le arance in base alla qualità e altre caratteristiche. La rete neurale è stata costruita utilizzando Keras e TensorFlow, e ottimizzata con l'algoritmo Adam.

3. Valutazione dei modelli

Dopo aver addestrato i vari modelli, sono stati confrontati utilizzando le seguenti metriche di valutazione:

- **Accuracy:** Percentuale di classificazioni corrette.
- **Precisione, Recall e F1-Score:** Per valutare l'equilibrio tra precisione e recall, dato il dataset sbilanciato.
- **Confusion Matrix:** Per capire come i modelli si comportano con le diverse classi.

4. Generazione di report

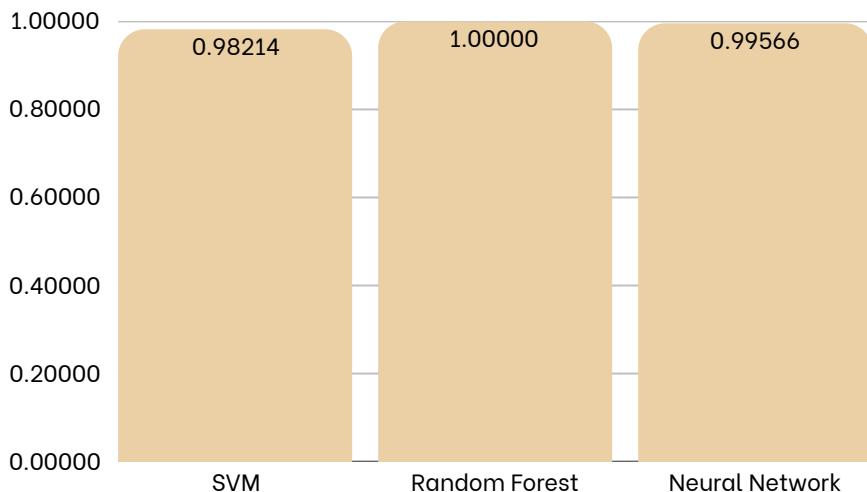
Il report finale contiene un riepilogo delle metriche di valutazione per ciascun modello, inclusi i dettagli di ogni fase del progetto:

1. **Comparazione delle performance dei modelli:** Confronto tra le prestazioni di Random Forest, SVM e rete neurale.
2. **Visualizzazioni:** I grafici mostrati di seguito riportano il confronto dei risultati dei tre modelli utilizzati, diversificando in confronto delle Accuratezze e confronto dei F1-Score dei modelli.
3. **Conclusioni:** Un'analisi finale su quale modello abbia mostrato i migliori risultati e i passi futuri per il miglioramento del sistema.

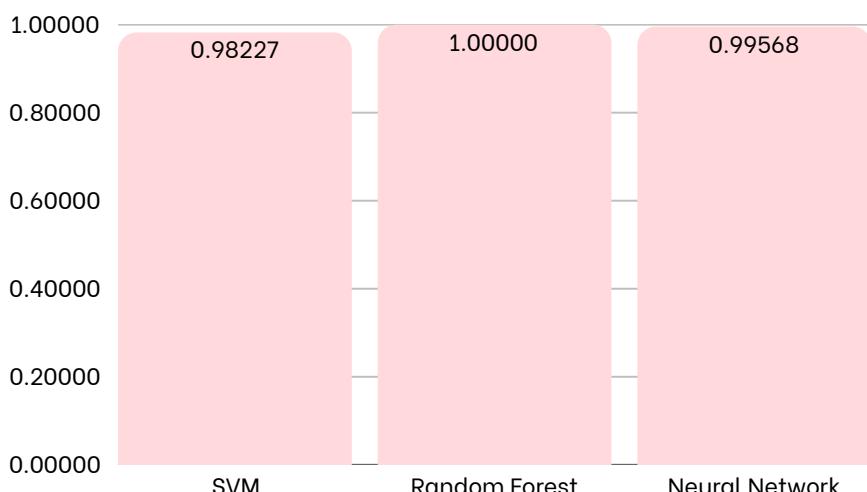
Informazioni principali

Confronto tra modelli

Confronto delle Accuracy dei modelli



Confronto dei F1-Score dei modelli



Scelte Progettuali

Premessa: data la scelta progettuale adottata, è risultato al quanto complicato trovare un dataset con sufficienti righe che fosse anche abbastanza realistico e non “Artificiale”. Data questa difficoltà incontrata, si è deciso di perseverare sul tema scelto, in primo luogo perché dà un tono di originalità al progetto, differenziando da altri comunemente in giro nel web, in secondo luogo si è preferito continuare con un tema di non facile sviluppo in quanto è di stimolo per la creazione e lo sviluppo di progetti futuri.

1. Scelta del Dataset

Il dataset utilizzato, scaricato da Kaggle, è stato selezionato per la sua completezza e pertinenza rispetto agli obiettivi del progetto. Contiene dati ben strutturati su arance e pompelmi, incluse variabili fondamentali come diametro, peso e colore mediante i colori fondamentali RGB. Considerato fin da subito una base solida per implementare il progetto ideato. Dei punti a sfavore potrebbero essere la ridotta quantità di colonne che descrivono le arance, colmato col processamento, e la poca complessità dei dati presenti.

2. Tecniche di Preprocessamento

Il preprocessamento dei dati ha incluso fasi di normalizzazione, bilanciamento delle classi e gestione di valori mancanti o anomali oltre all'esclusione dei dati riguardanti i pompelmi in quanto ritenuti fuori tema e quindi ritenuti dati inutili. La normalizzazione è necessaria per garantire che tutte le variabili abbiano lo stesso peso durante l'addestramento dei modelli. Il bilanciamento delle classi è stato implementato per evitare problemi di bias nei modelli, soprattutto data la rilevanza di alcune caratteristiche come la presenza di difetti. Questi passaggi hanno migliorato la stabilità dei modelli durante l'addestramento e ridotto il rischio di overfitting.

3. Modelli di Apprendimento Utilizzati

Per questo progetto, sono state utilizzate tre tipologie principali di modelli di apprendimento:

- **Support Vector Machine (SVM):**

La motivazione principale della scelta di SVM è data dalla sua notorietà per essere efficace nei problemi di classificazione con dataset di piccole o medie dimensioni. Inoltre, il kernel RBF utilizzato è particolarmente utile per catturare relazioni non

lineari nei dati. È stato un ottimo punto di riferimento per confrontare altri modelli più complessi.

- **Random Forest:**

Random Forest è stato scelto per la sua capacità di gestire dataset con molte variabili, fornendo robustezza contro l'overfitting e permettendo l'interpretazione delle feature importance. Ha mostrato prestazioni impeccabili in termini di accuratezza e recall, con punteggi pieni, tanto da far dubitare del suo effettivo funzionamento.

- **Rete Neurale (Neural Network):**

La scelta di una rete neurale è data dal perseguire l'obiettivo di esplorare relazioni complesse nei dati. È stata progettata con un'architettura semplice ma flessibile, con possibilità di espansione futura. Ha permesso di ottenere risultati competitivi, offrendo un benchmark per eventuali sviluppi basati su deep learning.

4. Metodi di Valutazione

I modelli sono stati valutati utilizzando metriche standard come accuracy, precision, recall e F1-score. Queste metriche offrono una visione equilibrata delle prestazioni, considerando sia l'efficacia complessiva (accuracy) sia la capacità del modello di distinguere tra classi (precision, recall, F1-score). Il confronto tra queste metriche ha permesso di identificare il modello più adatto in base alle priorità del progetto (ad esempio, ridurre i falsi negativi per la classificazione dei difetti).

5. Generazione del Report

Il report finale, generato automaticamente, è stato implementato per fornire una panoramica chiara e sintetica delle prestazioni dei modelli. Automatizzare questa fase riduce il tempo necessario per l'analisi dei risultati e minimizza la possibilità di errori manuali in quanto un report chiaro e standardizzato facilita la revisione dei risultati e la comunicazione con eventuali stakeholder.

6. Scelta degli Strumenti e delle Librerie

Come linguaggio di programmazione principale è stato utilizzato Python, insieme a librerie ben consolidate come scikit-learn, TensorFlow e matplotlib. In quanto Python offre un ecosistema ricco e supporto per la scienza dei dati, mentre librerie come scikit-learn e TensorFlow sono standard di settore per lo sviluppo e l'addestramento dei modelli. La scelta di strumenti consolidati garantisce la scalabilità del progetto e la possibilità di espanderlo con funzionalità aggiuntive.

7. Scelta sull'Analisi delle Arance

Si è scelto di utilizzare un file Jupyter Notebook per migliorare la comprensione e la comunicazione delle analisi sulle arance, offrendo un ambiente interattivo e strutturato che bilancia semplicità d'uso e approfondimento analitico. Questa scelta è particolarmente efficace per presentare dati e risultati in modo più accessibile.

Interrogazione della Knowledge Base

La Knowledge Base del progetto è implementata in Prolog e permette agli utenti di porre query per ottenere informazioni specifiche sulle arance. Nel file knowledge_base.pl sono inclusi fatti e regole che permettono di ottenere informazioni sulle varietà di arance, sulla qualità e su altre proprietà.

Avvio del Sistema

1. Installazione di SWI-Prolog:

- Scaricare SWI-Prolog dal sito ufficiale.
- Installare seguendo le istruzioni fornite per il proprio sistema operativo.

2. Caricamento della Knowledge Base:

- Aprire il terminale di SWI-Prolog.
- Caricare il file della Knowledge Base utilizzando il comando:
`consult('C:/Users/cistu/Desktop/OrangeSelection/Cistulli_Domenico/src/kbs/knowledge_base.pl').`

Tipi di Query Supportate

La Knowledge Base supporta diversi tipi di interrogazioni, dai fatti semplici alle regole più complesse.

Query sui Fatti:

1. Scoprire la qualità di una varietà:

`qualita(navel, Qualita).`

Risultato: Qualita = media.

2. Trovare l'origine di una varietà:

`origine(blood_orange, Origine).`

Risultato: Origine = sicilia.

Query sulle Regole:

1. Determinare varietà con alta dolcezza:

qualita_dolce(Varieta, alta).

Risultato: Varieta = valencia.

2. Raccomandare varietà di alta qualità:

raccomanda(Varieta).

Risultato: Varieta = valencia.

Query Complesse:

1. Filtrare varietà con dolcezza > 10 e acidità <= 3.5:

dolcezza(Varieta, Dolcezza), Dolcezza > 10, acidita(Varieta, Acidita), Acidita <= 3.5.

Risultato: Varieta = valencia, Dolcezza = 12, Acidita = 3.

2. Recuperare varietà con regole RDF:

query_rdf(valencia, URI).

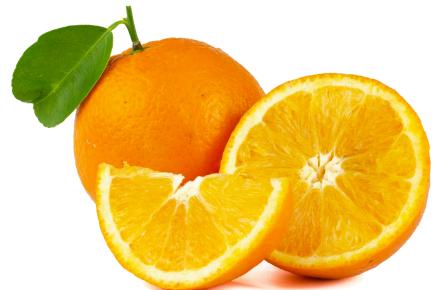
Risultato: URI = '<http://example.org/arance#Valencia>'

Bibliografia

<i>Fonte del dataset utilizzato</i>	Il Dataset utilizzato contiene 10.000 righe, suddiviso per metà in dati relativi alle Arance e metà in dati relativi ai Pompelmi. Nella fase di Processing, una delle prime operazioni eseguite è stata proprio quella di escludere i dati riguardanti i Pompelmi, in quanto il progetto aveva e ha come scopo principale quello di creare un sistema predittivo per la selezione delle arance. Il seguente è il link al dataset: Oranges vs. Grapefruit
<i>Libro di testo consultato</i>	Il Libro di testo di David L. Poole & Alan K. Mackworth dal titolo Artificial Intelligence 3E foundations of computational agents trattano è una fonte di teorie, metodi e applicazioni relativi ai modelli di machine learning, come la classificazione, l'analisi dei dati e l'ottimizzazione dei modelli in particolar modo il capitolo 7 che tratta l'argomento dell'apprendimento automatico supervisionato.
<i>GitHub.com</i>	Durante la fase di pianificazione, sono stati consultati diversi progetti già esistenti nel web, in particolare sulla piattaforma GitHub. Questa ricerca ha avuto il ruolo fondamentale di dare spunto per idee riguardanti la fase di sviluppo e per evitare che si progettasse una replica di un altro progetto già esistente.

Conclusioni

Il progetto “OrangeSelection” ha mostrato che è possibile creare un sistema predittivo per la selezione delle arance utilizzando tecniche di machine learning. La rete neurale ha offerto buoni risultati, ma è stato necessario bilanciare i dati e ottimizzare i modelli per migliorare le performance. In futuro, si prevede di esplorare altre tecniche di machine learning per affinare il modello e includere più variabili per una predizione ancora più precisa.



Sviluppi Futuri

1. Integrazione di Modelli di Deep Learning Avanzati

Un possibile sviluppo futuro prevede l'uso di modelli di deep learning avanzati, come le reti neurali convoluzionali (CNN), che si rivelano particolarmente efficaci nella classificazione delle immagini e nell'analisi di dati complessi. L'adozione di questi modelli potrebbe aumentare la precisione delle previsioni, specialmente se il progetto si espandesse per includere dati più complessi o per combinare l'analisi delle immagini con dati numerici.

2. Sistema di Raccomandazione Basato su Preferenze dell'Utente

Un altro possibile sviluppo potrebbe essere la creazione di un sistema per prevedere i prezzi delle arance, tenendo conto di fattori come varietà, peso, diametro e qualità. Arricchendo il dataset con dati sui prezzi di mercato e utilizzando modelli di regressione come Random Forest o Gradient Boosting, si potrebbe stimare con precisione il valore commerciale delle arance aiutando i venditori a massimizzare i profitti garantendo trasparenza ai clienti.

3. Espansione del Dataset e Implementazione di Modelli per Nuove Variabili

Infine sarebbe interessante l'integrazione di tecnologie IoT, che consentirebbe di raccogliere dati sulle arance in tempo reale durante la fase di raccolta e smistamento. Sensori installati su macchinari potrebbero misurare peso, diametro e colore, rilevando difetti superficiali e arricchendo il dataset esistente. Questa automazione migliorerebbe la precisione delle analisi e ottimizzerebbe il controllo qualità, rendendo il sistema scalabile e adatto sia a piccole aziende agricole che a grandi catene di distribuzione.

Domenico Cistulli

