Alma mater studiorum
Università di Bologna

Corso di laurea magistrale in informatica

Progetto per il corso di Intelligenza Artificiale

AI-OilBoys

Componenti del gruppo:

Manuel Placella	1099701
Alessandro Tocco	1097880
Daniele Costoli	1098681
Giuseppe Pio Salcuni	1100090

Anno Accademico 2022/2023

Indice

1 Introduzione

La presente relazione espone il progetto sviluppato nell'ambito del Corso di *Intelligenza Artificiale*, tenuto dal Professor Maurizio Gabbrielli e dal Dottor Stefano Pio Zingaro, presso l'Università di Bologna, come parte del Corso di Laurea Magistrale in Informatica. Nel seguito, saranno dettagliati il funzionamento del progetto e le strategie di implementazione utilizzate.

1.1 Problema: autenticazione dell'origine geografica degli oli di oliva

La frode degli oli rappresenta una minaccia sempre più diffusa e grave per la salute dei consumatori e per l'industria alimentare [1]. Il Regolamento UE n. 29/2012 stabilisce che l'etichettatura dell'olio extravergine di oliva (EVOO) e dell'olio d'oliva vergine (VOO) deve indicare l'origine geografica del prodotto. La verifica della conformità dell'origine geografica dichiarata in etichetta è di fondamentale importanza per tutelare i consumatori da informazioni fuorvianti e ristabilire la loro fiducia nel prodotto, ma anche per individuare e prevenire pratiche fraudolente e aumentare la competitività del settore. In questo modo, i consumatori possono fare scelte informate e consapevoli, mentre i produttori possono competere in modo equo sul mercato. In sintesi, l'etichettatura dell'origine geografica dell'EVOO e del VOO è obbligatoria e ha un ruolo importante nel garantire la qualità e l'autenticità del prodotto, nonché nella protezione dei consumatori e dei produttori.
Negli ultimi anni, numerosi studi scientifici sono stati condotti per affrontare l'autenticazione geografica dell'EVOO e del VOO. Questi studi sono stati condotti con diversi approcci. Anche se i risultati sono stati promettenti, l'utilizzo di tecniche sofisticate e costose limita la loro applicabilità nei comuni

laboratori di controllo. Pertanto, è necessario sviluppare approcci analitici più robusti ed economici e marcatori geografici specifici per l'autenticazione geografica di EVOO e VOO.

1.2 Soluzioni esistenti

Negli ultimi anni sono state sviluppate strategie con l'obiettivo di affrontare questa importante sfida commerciale e legale, ovvero l'autenticazione dell'origine geografica dell'olio d'oliva. Sono stati proposti dati generati da diverse tecniche analitiche (ad es. cromatografia, spettrofotometria, spettroscopia ed elettrochimica) in combinazione con strumenti chemiometrici supervisionati o non supervisionati.
Ad esempio, Hmida et al. [2] ha utilizzato acidi grassi e triacyglycerids come marcatori di origine dell'olio di oliva vergine (VOO) dalla regione mediterranea (ad esempio, Portogallo, Francia, Tunisia e Turchia), applicando l'analisi dei componenti principali (PCA). La composizione dei composti volatili è stata utilizzata con successo anche come marcatori di origine geografica degli oli di oliva greci di cultivar 'Notopia' [3] .
Quintanilla-Casas et al. [4] ha proposto di utilizzare l'impronta digitale degli idrocarburi sesquiterpenici insieme all'analisi discriminante dei minimi quadrati parziali (PLS-DA) per verificare la dichiarazione dell'etichetta di origine dell'UE e di un singolo paese.
In alternativa, un'altra opzione valida è l'utilizzo della spettrometria di massa (MS), una tecnica analitica che utilizza i rapporti massa su carica (m/z) per identificare i composti di un campione. Il metodo identifica un composto determinandone il peso molecolare e analizzandone l'abbondanza isotopica. Uno spettrometro di massa ionizza il campione in fase gassosa e identifica gli ioni in base ai loro rapporti massa su carica e le abbondanze relative. Questa tecnica, unita a diversi strumenti statistici multivariati, si è rivelata un potenziale strumento per identificare l'origine geografica degli oli di oliva

[5].

In particolare questo progetto è incentrato sull'utilizzo di questa tecnica, la quale offre vari vantaggi come la selettività, la sensibilità e l'analisi multicampione. L'unico limite è che per riuscirne a sfruttare a pieno le potenzialità sarebbe necessario avere competenze specifiche nel campo della chimica e della biologia. Pertanto il nostro lavoro si baserà su un approccio più semplice fondato sull'utilizzo diretto dei dati relativi alla spettrometria di massa e sulla valutazione di vari algoritmi di apprendimento automatico per determinarne l'efficacia.

1.3 Ipotesi e soluzione proposta

Sulla base delle informazioni precedentemente esposte, è stata avanzata un'ipotesi interessante che potrebbe rivoluzionare il settore dell'analisi degli oli di oliva. Si è pensato infatti di utilizzare direttamente il file di spettrometria di massa per classificare gli oli, evitando così di dover ricorrere a costose e lente analisi chimiche specifiche. Per raggiungere questo obiettivo, si è deciso di focalizzarsi sulla TIC (Total Ion Current), ovvero la corrente ionica totale, dimostrando come essa possa essere utilizzata per predire l'origine degli oli di oliva, senza dover analizzare i vari grafici ottenuti con la spettrometria di massa che richiedono importanti competenze nel settore. Per dimostrare la validità di questa ipotesi, sono state utilizzate tecniche di machine learning, le cui specifiche sono descritte in dettaglio nei capitoli successivi.
In sintesi, l'obiettivo è quello di sviluppare un modello predittivo in grado di garantire l'autenticazione degli oli di oliva, sfruttando l'informazione contenuta nel file di spettrometria di massa e utilizzando tecniche di machine learning. Questo approccio potrebbe consentire di semplificare notevolmente le procedure di analisi, ridurre i tempi e i costi, migliorare la qualità del

prodotto finale limitando quindi le frodi sugli oli di oliva.

1.3.1 Risultati ottenuti

Qui di seguito un breve cenno ai risultati ottenuti successivamente presentati e discussi nella sezione "Risultati".
Dopo un'attenta valutazione di vari algoritmi di machine learning, sono stati ottenuti risultati molto promettenti nella classificazione degli oli provenienti da diverse aree geografiche quali: Spagna, Italia, Portogallo, Grecia, Grecia-Peloponneso, Grecia-Creta e Tunisia. In particolare, la media dell'accuratezza ottenuta tramite una 10-cross-validation è stata dell'84%, il che dimostra l'efficacia del nostro approccio.
Successivamente, si è deciso di concentrarsi su un numero ristretto di campioni non considerando quelli provenienti da Grecia, Tunisia e Portogallo. Questo approccio ha portato ad un ulteriore miglioramento dell'accuratezza, che è stata aumentata al 93% sempre validando il modello con una 10-cross-validation.
Questi risultati sono molto positivi e suggeriscono che il nostro modello di machine learning può essere utilizzato con successo per la classificazione degli oli di oliva provenienti da diverse aree geografiche. Questo approccio potrebbe consentire di semplificare notevolmente le procedure di analisi, ridurre i tempi e i costi, e migliorare la qualità del prodotto finale.

1.4 Organizzazione del gruppo

Il gruppo ha collaborato in maniera attiva e sinergica, scegliendo di lavorare insieme nello stesso luogo piuttosto che svolgere il lavoro singolarmente a distanza. La collaborazione in team ha permesso di affiancare le attività di sviluppo con discussioni di carattere teorico sulle ipotesi fatte, sulle soluzioni implementate e sui risultati ottenuti. Il progetto si è prestato particolarmente

bene a questa forma di lavoro, dal momento che ogni scelta compiuta doveva essere ben motivata e spiegabile, soprattutto nel campo del machine learning.

Per monitorare le revisioni e i progressi effettuati, il team ha utilizzato la piattaforma di gestione del repository GitLab[1].

1.5 Strumenti e librerie utilizzate

Per realizzare il seguente lavoro è stato utilizzato il linguaggio di programmazione Python 3[2] all'interno degli ambienti di sviluppo DataSpell 2022.3.3[3] e Jupyter Notebook[4].

A supporto del progetto sono state utilizzate le seguenti librerie:

Librerie	
Scikit-learn	Pandas
NumPy	pyOpenMS
Matplotlib	Scikit-plot

[1] www.gitlab.com
[2] www.python.org
[3] www.jetbrains.com/dataspell/
[4] www.jupyter.org

2 Metodo proposto

2.1 Dataset

Il dataset che abbiamo selezionato per la nostra analisi è composto da due cartelle che contengono i file di spettrometria di massa di campioni di olio della campagna olearia 2021-2022 in formato `.mzdata.xml`, utilizzati rispettivamente per l'analisi dei composti organici volatili (VOCs) e degli idrocarburi. La Figura 1 mostra nel dettaglio la struttura del dataset e il numero di file presenti nelle cartelle. Come si può vedere il dataset comprende un file Excel (la cui struttura è riportata nella Tabella 1) contenente ulteriori informazioni sulle analisi tra cui, di nostro interesse, l'origine del campione.

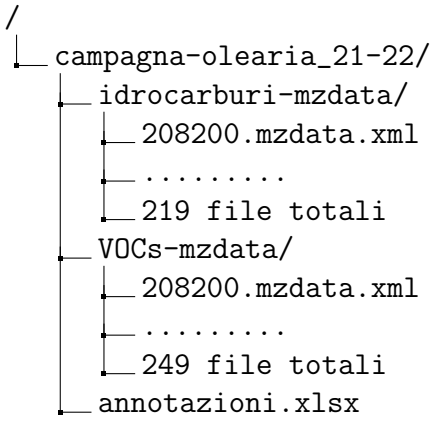

Figura 1: Struttura del dataset

Colonna	Tipo di valore
Sample	Codice numerico
Informativo	Codice numerico
Codice Carapelli	Codice numerico o vuoto
Codice FM	Codice alfanumerico
Codice FM2	Codice alfanumerico
CV	Stringa
Origine	Stringa
Anno produzione	2021-22
Idrocarburi	'Sì' o 'No'
VOCs	'Sì' o 'No'

Tabella 1: Colonne file Excel

Il file Excel `annotazioni.xlsx` rappresenta un'importante parte del dataset ed è composto da 221 righe, ognuna delle quali rappresenta un campione di olio. Questo dataset è stato creato allo scopo di analizzare le proprietà chimiche dell'olio e determinare la sua origine geografica o la sua cultivar. Le colonne utilizzate in questo lavoro sono quelle sottolineate nella Tabella 1, che sono state selezionate in base alla loro rilevanza per l'analisi.
In particolare, la colonna "Sample" contiene il nome del file `.mzdata.xml` corrispondente al campione , che rappresenta un'importante fonte di informazioni sulle proprietà chimiche dell'olio. Per quanto riguarda gli idrocarburi i nomi dei file presenti nella tabella non tutti corrispondono a quelli dei file quindi è stato necessario riallineare i nomi dei file con i codici presenti nella tabella, come spiegato più avanti. La colonna "Origine" indica l'origine dell'olio attraverso una stringa, mentre la colonna "CV" indica la cultivar dell'olio con una stringa. La colonna "Idrocarburi" indica se è presente o meno il file `.mzdata.xml` per l'analisi degli idrocarburi, utilizzando una stringa del tipo "Sì" o "No", mentre la colonna "VOCs" indica la presenza o l'as-

senza del file `.mzdata.xml` per l'analisi dei VOCs, anch'essa utilizzando una stringa del tipo "Sì" o "No".
È importante sottolineare che la colonna "Origine" rappresenta il target del nostro dataset e contiene le stringhe che indicano il paese di origine dell'olio. Le stringhe sono: "Spagna", "Italia", "Portogallo", "Grecia", "Grecia - Peloponneso", "Grecia-Creta" e "Tunisia". Tuttavia, queste stringhe non hanno lo stesso formato in termini di maiuscole e minuscole e alcune contengono dei caratteri particolari. Per garantire una uniformità del dataset, è stato necessario affrontare questo problema ed è stato fatto a livello di codice durante la pre-elaborazione dei dati e la costruzione del dataset.

2.1.1 VOCs e idrocarburi

I VOCs (Volatile Organic Compounds) sono composti organici volatili che si trovano in molti oli, sia di origine naturale che sintetica. Negli oli di origine naturale, i VOCs sono spesso prodotti durante il processo di estrazione dell'olio, ad esempio tramite la fermentazione o l'ossidazione delle materie prime. I VOCs includono anche gli idrocarburi, una classe di composti costituiti solo da atomi di carbonio e idrogeno, che sono un componente comune di molti oli. La quantità e la composizione dei composti organici volatili e degli idrocarburi può dipendere dalla varietà di oliva, dalla tecnologia utilizzata nella produzione dell'olio e, fattore importante per il nostro lavoro, dalle condizioni di coltivazione. Il tipo di terreno, l'esposizione al sole, le temperature medie possono influire sulla composizione dei composti organici volatili così come oli di oliva prodotti in regioni con elevato inquinamento atmosferico possono contenere livelli più elevati di idrocarburi.
In questo lavoro sono stati presi in considerazione principalmente i file di spettrometria di massa dei VOCs, essendo questi una famiglia di composti più generica. Tuttavia, gli algoritmi che hanno dato i risultati migliori sono stati testati anche con i dati di spettrometria di massa degli idrocarburi

ottenendo dei risultati pressocchè identici.

2.1.2 TIC (Total Ion Current) e calcolo con pyOpenMS

Un cromatogramma di massa è una rappresentazione dei dati di spettrometria di massa come cromatogramma, in cui l'asse x rappresenta il tempo di ritenzione e l'asse y rappresenta l'intensità del segnale.

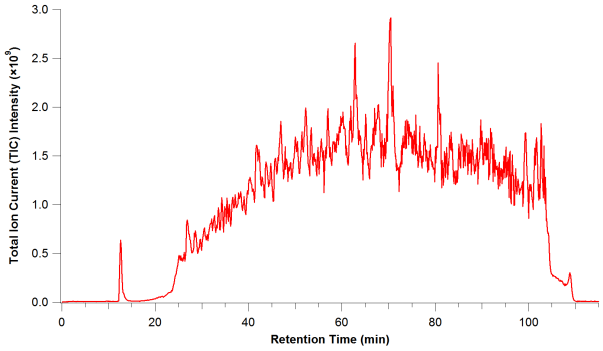

Figura 2: Cromatogramma TIC

Il cromatogramma TIC è un grafico che rappresenta la corrente ionica totale generata da tutti gli ioni presenti in una miscela durante l'analisi tramite spettrometria di massa. Questa tecnica consente di identificare e quantificare i composti presenti in una miscela in base alla loro massa elettrica (m/z). Durante l'analisi, gli ioni vengono ionizzati e la loro corrente ionica viene misurata. Ioni con lo stesso m/z vengono raggruppati insieme e la loro corrente ionica viene sommata per produrre il cromatogramma TIC.
Per il calcolo della corrente ionica totale (TIC), è stata adottata la libreria pyOpenMS[5] , che consente di leggere in modo diretto il file in formato `.mzdata.xml` e di caricarlo in un oggetto della classe `MSExperiment()`. Grazie a questa scelta, è possibile eseguire il calcolo della TIC con grande

[5]Esempio di calcolo della TIC: https://pyopenms.readthedocs.io/en/latest/first_steps.html

semplicità, utilizzando una chiamata alla funzione `calculateTIC()` offerta sempre dalla libreria pyOpenMS. Tale funzione restituisce un oggetto di tipo `MSChromatogram`. Dall'oggetto ottenuto è possibile richiamare la funzione `get_peaks()` che restituisce due strutture dati `ndarray` , la prima contenente i tempi di ritenzione (espressi in secondi), mentre la seconda riporta i valori corrispondenti di TIC. È importante notare che, sebbene le due liste di tempi di ritenzione e TIC abbiano uguale lunghezza per quasi tutti i cromatogrammi, i tempi di ritenzione potrebbero variare tra i diversi campioni, il che significa che le feature ottenute da ciascun campione potrebbero essere diverse.
Per garantire la comparabilità tra i diversi campioni e, poichè le liste ottenute dalla funzione `calculateTIC()` erano di una lunghezza media di 12270 valori, è stata sviluppata una funzione che calcola la media della corrente ionica totale in ogni minuto dall'inizio del cromatogramma fino al minuto 40, per ogni campione. La scelta di questo intervallo di tempo è stata fatta sulla base di un'analisi preliminare dei dati, che ha mostrato come tutti i campioni arrivassero a questo istante. In questo modo, sono state estratte 40 feature da ogni campione, corrispondenti alla media della TIC per ogni minuto nell'intervallo minuto 1 - minuto 40. Queste feature rappresentano una media del comportamento del cromatogramma durante il periodo di interesse, consentendo una comparazione tra i diversi campioni e una valutazione delle eventuali differenze significative tra di essi.

2.1.3 Preparazione dei dati

Durante la prima fase del lavoro, è stato eseguito un controllo per verificare la presenza di tutti i file riportati nel file Excel `annotazioni.xlsx` all'interno delle rispettive cartelle. Durante questa verifica, è stato notato che i nomi dei file `.mzdata.xml` dei VOCs corrispondono tutti ai codici riportati nella tabella Excel, mentre per gli idrocarburi alcuni nomi non corrispondono.

Tuttavia, grazie all'individuazione delle principali differenze tra i nomi dei file e i codici presenti nel file excel, è stato possibile allineare i nomi dei file con i codici corretti.
Dopo aver riallineato i nomi dei file ai codici corretti, sono stati creati i dataset di lavoro seguendo un preciso procedimento. Per ogni riga del file excel `annotazioni.xlsx`, sulla base del valore delle colonne `VOCs` prima e `Idrocarburi` poi (che indicano la presenza o l'assenza del file corrispondente) è stato estratto il valore delle colonne `Origine` e `CV` ed è stato costruito il path completo del file `.mzdata.xml` corrispondente. Tutti questi dati sono stati salvati in un `Pandas DataFrame` avente come colonne quindi `Origine`, `Cultivar` e `File`.
A partire dal `DataFrame` appena creato, sono stati costruiti quindi i due dataset di lavoro: uno per i VOCs e uno per gli Idrocarburi. Per ogni riga di entrambi i dataset, è stato prelevato e normalizzato il valore della colonna `Origine` e della colonna `Cultivar`. Successivamente, è stato caricato il file `.mzdata.xml` corrispondente e con l'utilizzo di una funzione scritta dal team di lavoro, chiamata `calcolaMedie(MSExperiment)`, sono state calcolate le 40 feature corrispondenti alla media della corrente ionica totale in ogni minuto dall'istante iniziale fino al minuto 40. Questi valori sono stati salvati nelle colonne `minuto 1`, `minuto 2`, ..., `minuto 40` dei due dataset.
Infine, i due dataset sono stati salvati in due file di tipo csv con i nomi `df_media_TIC_VOCs_min_1-40.csv` e `df_media_TIC_Idro_min_1-40.csv`. Nel primo file sono presenti le feature relative ai VOCs, mentre nel secondo file sono presenti le feature relative agli Idrocarburi. La struttura del dataset finale dei VOCs è riportata in Tabella 2, per gli Idrocarburi cambia solo il numero di campioni (215). Questi dataset sono stati utilizzati nella prima parte in cui è stato seguito un primo approccio mentre nella seconda parte si è deciso di modificare leggermente questo dataset in quanto sono state considerate aree geografiche diverse.

	Origine	**Cultivar**	**minuto 1**	**...**	**minuto 40**
1	spagna	arbequina	269838.43...	...	212...
...	...	...	...	...	...
217	grecia-creta	psilolia	280082.21...	...	181...

Tabella 2: Struttura dataset VOCs costruito

2.2 Algoritmi predittivi

In questo lavoro è stato scelto di utilizzare algoritmi classici come KNN, SVC, Naive Bayes, DecisionTree e RandomForest essendo questi maggiormente adatti per affrontare problemi di apprendimento automatico che coinvolgono dati strutturati, ovvero dataset tabulari e per la loro ampiamente documentata efficacia in problemi di classificazione. Non è stato preso in considerazione l'utilizzo di reti neurali poiché esse sono maggiormente indicate per problemi di apprendimento automatico più complessi, come la classificazione di immagini, il riconoscimento del linguaggio naturale e altri problemi che richiedono un alto livello di flessibilità e adattamento ai dati di input. Di seguito è riportata una breve presentazione degli algoritmi utilizzati:

- **KNN**: l'algoritmo k-nearest neighbors, noto anche come KNN o K-NN, è un classificatore di apprendimento supervisionato non parametrico, che utilizza la prossimità per effettuare classificazioni o previsioni sul raggruppamento di un singolo punto dati. È basato sul presupposto che punti simili possono essere trovati l'uno vicino all'altro.

- **SVC**: l'algoritmo SVC, o Support Vector Classifier, è un algoritmo di apprendimento automatico supervisionato tipicamente utilizzato per attività di classificazione. Questo algoritmo funziona mappando i punti dati in uno spazio ad alta dimensione e quindi trovando l'iperpiano ottimale che divide i dati in due classi.

- **Naive Bayes**: algoritmo basato sul teorema di Bayes che utilizza la probabilità condizionale per classificare gli oggetti. L'approccio "naive" sta nel considerare indipendenti tra loro le variabili che descrivono l'oggetto da classificare, pur se spesso non lo sono. In pratica, l'algoritmo calcola le probabilità di ogni classe data la presenza di determinate caratteristiche e sceglie la classe più probabile come output.

- **Decision Tree**: algoritmo che crea un albero di decisione basato sulle caratteristiche dell'oggetto da classificare. L'algoritmo divide ripetutamente l'insieme di dati in sottoinsiemi in base alla caratteristica più significativa per la classificazione, fino a quando ogni sottoinsieme contiene solo oggetti di una singola classe o fino a quando non viene raggiunta una soglia predefinita.

- **Random Forest**: algoritmo che combina diversi alberi decisionali. L'algoritmo crea un insieme di alberi decisionali, ognuno addestrato su un sottoinsieme casuale dei dati di addestramento e su un sottoinsieme casuale delle caratteristiche da considerare. La classificazione o la regressione viene quindi determinata dalla media delle previsioni di tutti gli alberi decisionali.

2.3 Metriche di valutazione

Di seguito è riportato un breve cenno alle metriche di valutazione utilizzate per valutare le prestazioni del nostro lavoro. Le metriche utilizzate sono l'accuratezza, la precisione, la recall e l'F1-measure. Queste metriche sono comunemente utilizzate per valutare la qualità dei risultati di un modello di apprendimento automatico o di un sistema di classificazione.

- **Accuracy**: metrica che misura la percentuale di istanze correttamente classificate dal modello. L'accuratezza è una metrica utile quando le

classi sono ben bilanciate, in caso contrario, invece, questo indicatore rischia di risultare fuorviante.

- **Precision**: metrica che misura la percentuale di istanze classificate come positive che sono effettivamente positive. La precisione è una metrica utile quando l'obiettivo è minimizzare i falsi positivi.
- **Recall**: metrica che misura la percentuale di istanze positive che sono state correttamente classificate dal modello. La recall è una metrica utile quando l'obiettivo è minimizzare i falsi negativi.
- **F1-measure**: metrica che combina precisione e recall in un'unica metrica. Questa è utile quando l'obiettivo è trovare un equilibrio tra la precisione e la recall. La F1-measure è la media armonica tra la precisione e la recall. Utilizza la media armonica piuttosto che la media aritmetica perché la prima penalizza i valori estremi.

2.4 Primo approccio

In un primo approccio al problema sono stati testati tutti gli algoritmi sopra descritti sul dataset contenente la media della TIC dei file `.mzdata.xml` per l'analisi dei VOCs. Inizialmente sul dataset non è stata eseguita alcuna operazione di pre elaborazione dei valori delle 40 feature disponibili. Inoltre sono stati considerati tutti e sette i Paesi presenti all'interno del dataset, come riportato in Figura 3, quali Spagna, Italia, Portogallo, Grecia, Grecia-Peloponneso, Grecia-Creta e Tunisia.

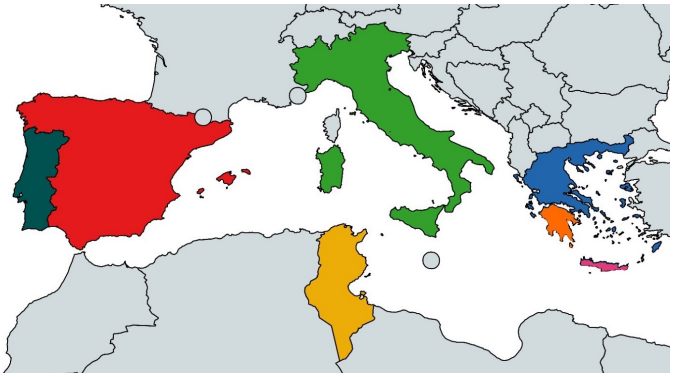

Figura 3: Paesi considerati nel primo approccio

Tuttavia, il dataset presenta un evidente squilibrio tra le diverse aree geografiche, come evidenziato dalla Figura 4. In particolare, il numero di campioni di olio provenienti da Portogallo, Grecia e Tunisia è nettamente inferiore rispetto ai campioni provenienti dagli altri Paesi. Questo ha influito, come previsto e come riportato più avanti, sulle prestazioni dei vari modelli.

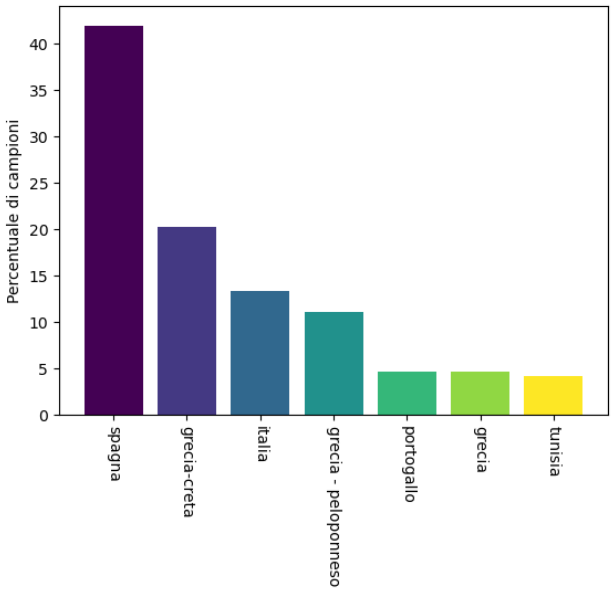

Figura 4: Percentuali di campioni nel dataset principale

Per valutare l'efficacia dei modelli di machine learning addestrati, è stata adottata una metodologia di 10-cross-validation. Questo metodo prevede la suddivisione del dataset in 10 parti uguali, chiamate "fold", in cui ogni volta una parte viene utilizzata come insieme di test e le altre 9 come insieme di train (addestramento). Questo processo viene ripetuto per tutte le possibili combinazioni delle parti, e i risultati delle diverse esecuzioni vengono poi combinati per ottenere una stima affidabile delle prestazioni del modello. In particolare sono stati testati i seguenti algoritmi: K-Nearest Neighbors (KNN) con il parametro `k-neighbors` impostato a 5, Decision Tree senza alcun parametro specifico, Support Vector Machine (SVM) con un kernel di tipo polinomiale, Naive Bayes di tipo Gaussian e infine Random Forest con il parametro `n_estimators` impostato a 100.
Come ampiamente riportato nella sezione 'Risultati' quello che ha dato i risultati migliori è stato Support Vector Classifier con un'accuratezza e una precisione media entrambi pari a 84% ma con bassi valori di recall e F1-measure. Durante il processo di sperimentazione degli algoritmi, si è proceduto anche all'impiego di tecniche di bilanciamento del modello e scaling delle feature, allo scopo di migliorare le prestazioni degli stessi. Tuttavia, i risultati ottenuti in questo primo approccio non sono stati ritenuti soddisfacenti, in quanto non si sono registrati miglioramenti significativi rispetto alle prestazioni dei modelli non sottoposti a tali tecniche. Come riportato nel prossimo paragrafo, queste tecniche sono state adottate anche nel secondo approccio, ottenendo invece importanti miglioramenti in termini di efficienza e accuratezza dei modelli sviluppati.

2.5 Secondo approccio

In seguito alla constatazione che il primo approccio adottato nella classificazione dei campioni di olio provenienti da Portogallo, Grecia e Grecia-Peloponneso portava a numerosi errori, si è proceduto alla crazione di due

nuovi dataset archiviati nei seguenti file: `df_media_TIC_VOCs_min_1-40_-semplificato.csv` e `df_media_TIC_Idro_min_1-40_semplificato.csv`. In tali dataset, come mostrato nella Figura 5, sono stati esclusi i campioni provenienti dai suddetti paesi, limitandosi a considerare solo quelli provenienti da Spagna, Italia, Grecia-Peloponneso e Grecia-Creta. Nonostante ciò, il nuovo dataset presenta ancora un certo sbilanciamento nella distribuzione dei campioni, come evidenziato nella Figura 6, con un numero di campioni provenienti dalla Spagna nettamente superiore rispetto ai campioni provenienti dagli altri Paesi.

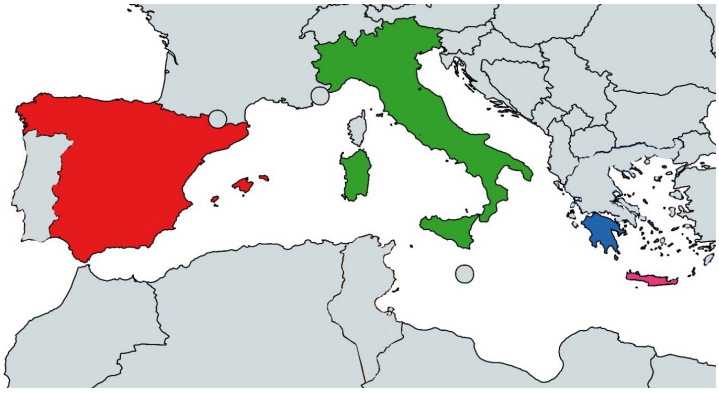

Figura 5: Paesi considerati nel secondo approccio

Utilizzando il nuovo dataset, la cui validazione è stata effettuata tramite una 10-cross-validation, si è ottenuta un'accuratezza del 93%, come dettagliato nella sezione "Risultati" della presente relazione. Per conseguire tale risultato, si è fatto ricorso all'algoritmo SVC, al quale è stata applicata una tecnica di bilanciamento del modello di tipo `class-weight` e uno scaling dei valori delle varie feature con uno scaler di tipo `MinMaxScaler`. I dettagli di queste due tecniche sono riportati nella sezione "Risultati". Queste due tecniche, a differenza di quanto visto nel primo approccio, hanno portato a notevoli miglioramenti. Il `class-weight` ha permesso di assegnare pesi diversi alle classi al fine di correggere l'effetto dello sbilanciamento, mentre il `MinMaxScaler`

ha consentito di trasformare le feature in modo tale da ottenere un range di valori uniforme tra tutte le variabili, facilitando il processo di apprendimento del modello.

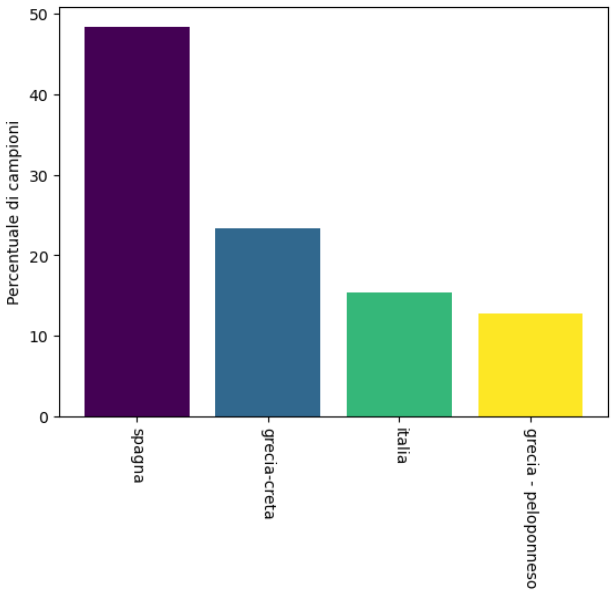

Figura 6: Percentuali di campioni nel dataset semplificato

Per entrambi i due approcci descritti nel presente studio, sono stati condotti ulteriori test mediante l'utilizzo degli stessi algoritmi ma con un dataset differente, costruito attraverso i file `.mzdata.xml` finalizzati all'analisi degli Idrocarburi, invece che quelli finalizzati all'analisi dei VOCs. Durante tali test, sono state considerate sia la tecnica del bilanciamento del modello che la tecnica dello scaling dei dati. Tuttavia, a seguito di un'attenta analisi dei risultati ottenuti, nessun algoritmo ha fornito risultati soddisfacenti o comunque paragonabili a quelli ottenuti con i VOCs. Questi risultati suggeriscono che i dataset utilizzati per l'analisi degli Idrocarburi possano presentare maggiori difficoltà nella fase di modellizzazione, rispetto ai dataset relativi ai VOCs.

3 Risultati

Come evidenziato nella conclusione del capitolo precedente, i risultati ottenuti mediante l'utilizzo del dataset costruito con i file di spettrometria di massa per l'analisi degli Idrocarburi non hanno fornito risultati soddisfacenti. Pertanto, in questa sezione del presente studio si discuteranno esclusivamente i risultati ottenuti mediante l'utilizzo del dataset costruito a partire dai file di spettrometria di massa per l'analisi dei VOCs. Tale scelta è stata dettata dalla necessità di concentrarsi esclusivamente sulle feature più rilevanti ai fini della classificazione.

3.1 Risultati primo approccio

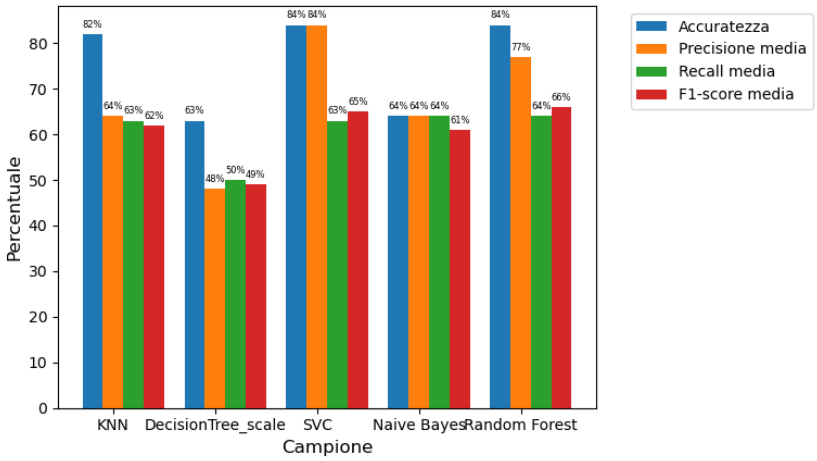

Figura 7: Risultati 10-cross-validation dataset iniziale

Come si può vedere nella Figura 7 i risultati migliori si ottengono con Support Vector Classifier con un'accuratezza e una precisione media pari a 84% ma con recall e F1-measure rispettivamente pari a 63% e 65%. Ciò è dovuto con

molta probabilità allo sbilanciamento del dataset e quindi al fatto che per alcuni Paesi sono pochi i campioni a disposizione, ipotesi confermata dalla matrice di confusione in Figura 8 e dal report della classificazione qui di seguito dove si può vedere come la maggior parte degli errori si verificano nella classificazione degli oli provenienti da Grecia, Portogallo e Tunisia (classi in cui i campioni a disposizione sono in numero nettamente inferiore rispetto alle altre).

	precision	recall	f1-score	support
grecia	0.50	0.10	0.17	10
grecia - peloponneso	0.78	0.88	0.82	24
grecia-creta	0.85	0.93	0.89	44
italia	0.93	0.86	0.89	29
portogallo	1.00	0.10	0.18	10
spagna	0.82	0.97	0.89	91
tunisia	1.00	0.56	0.71	9
accuracy			0.84	217
macro avg	0.84	0.63	0.65	217
weighted avg	0.84	0.84	0.81	217

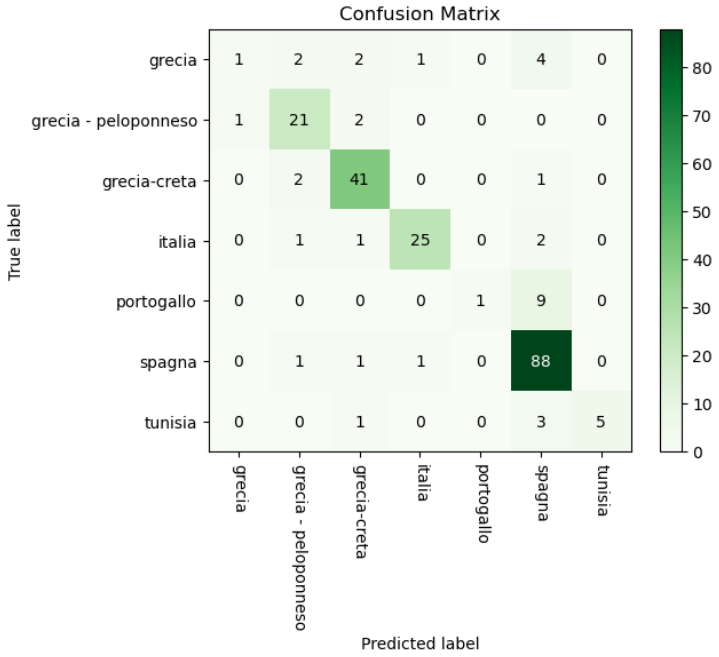

Figura 8: Matrice di confusione di SVC sul dataset principale

Successivamente è stato eseguito un tentativo di migliorare le prestazioni del modello attraverso una tecnica di scaling dei dati. Nello specifico, è stato utilizzato uno scaler di tipo `MinMaxScaler`, che ha permesso di normalizzare i valori delle feature in un intervallo compreso tra 0 e 1. Ciò ha consentito di evitare che alcune feature avessero un peso maggiore rispetto ad altre durante la fase di training del modello.
Dopo aver notato che la maggior parte degli errori si verificava nelle classi con pochi campioni disponibili è stato testato anche un bilanciamento del modello utilizzando la tecnica `class-weight`. In teoria, questa tecnica avrebbe dovuto assegnare un peso maggiore alle classi con pochi campioni, migliorando quindi la capacità del modello di apprendere da questi dati. Tuttavia, nonostante questo tentativo di scaling e di bilanciamento del modello, i risultati ottenuti non hanno evidenziato alcun importante miglioramento. Il

bilanciamento del modello può essere dannoso per le prestazioni del modello quando le classi non sono fortemente sbilanciate. Inoltre, l'assegnazione di pesi alle classi può aumentare il rischio di overfitting che può portare a una maggiore classificazione errata delle classi minoritarie, poiché il modello tende a prestare maggiore attenzione a tali classi.

3.1.1 Confronto tra algoritmi

Le migliori prestazioni ottenute dagli algoritmi KNN, SVC e Random Forest rispetto a Naive Bayes e Decision Tree può essere attribuito a diverse ragioni di natura tecnica. I primi tre algoritmi menzionati sono noti per essere più complessi e quindi con maggiori capacità rispetto a Naive Bayes e Decision Tree. KNN, SVC e Random Forest sono algoritmi più robusti rispetto a Naive Bayes e Decision Tree nell'affrontare problemi di overfitting. L'overfitting si verifica quando il modello si adatta troppo ai dati di addestramento, generando un modello che funziona bene sui dati di addestramento ma male sui dati di test. Un'altra ragione per cui KNN, SVC e Random Forest hanno prestazioni migliori rispetto a Naive Bayes e Decision Tree è legata alla complessità dei dati. I primi sono in grado di gestire un grande numero di variabili e di modellare relazioni non lineari tra di esse.

3.2 Risultati secondo approccio

Nel secondo approccio del nostro studio, ci si è concentrati principalmente sugli algoritmi che durante il primo approccio hanno dato risultati abbastanza soddisfacenti. In particolare, sono stati testati gli algoritmi KNN, SVC e Random Forest. Come evidenziato dalla Figura 9, tutti e tre gli algoritmi hanno raggiunto un'accuratezza superiore al 90%. In particolare, l'algoritmo SVC (Support Vector Classifier) ha ottenuto i risultati migliori,

con un'accuratezza del 93% e una precisione media del 93%. Inoltre, sia la recall media che la F1-measure media sono notevolmente migliorate con l'impiego del nuovo dataset. Questi risultati confermano l'ipotesi avanzata durante il primo approccio secondo cui il numero esiguo di campioni di olio provenienti da Portogallo, Grecia e Tunisia nel dataset originale è la causa di un deterioramento delle prestazioni del modello, soprattutto in termini di recall e F1-measure. Il nuovo dataset, che esclude i campioni provenienti da questi Paesi, ha permesso di ottenere risultati di classificazione notevolmente migliori in termini di accuratezza, precisione, recall e F1-measure.

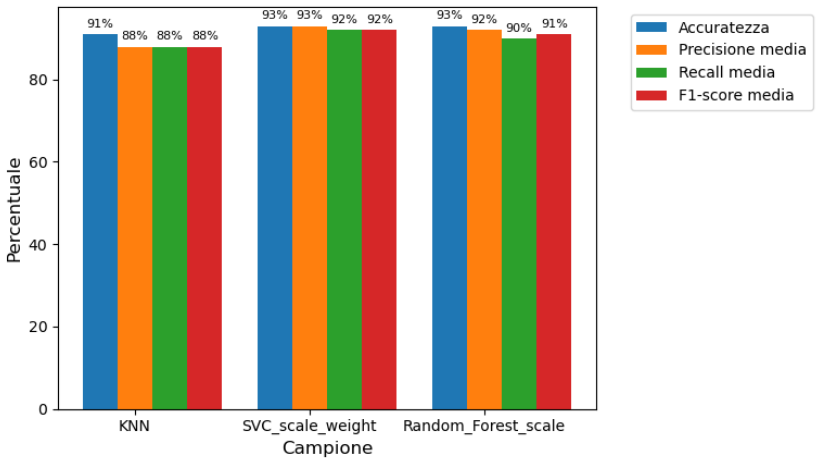

Figura 9: Risultati 10-cross-validation dataset semplificato

Nel dettaglio, si è constatato che l'algoritmo SVC (Support Vector Classifier) ha ottenuto i migliori risultati nella classificazione dei campioni di olio basati sul nuovo dataset semplificato. Questo è stato possibile grazie all'impiego di una tecnica di bilanciamento del modello `class-weight` e di uno scaling dei dati tramite `MinMaxScaler`. Tale combinazione ha permesso di ottenere un'accuratezza del 93% e una precisione media del 93%. Si ritiene che l'efficacia dell'algoritmo SVC sia dovuta alla sua capacità di creare un iperpiano che meglio separa le diverse classi di olio presenti nel dataset. Inoltre,

il bilanciamento del modello `class-weight` ha permesso di ridurre l'effetto dello sbilanciamento presente nel dataset, aumentando così la capacità dell'algoritmo di classificazione. Lo scaling dei dati con `MinMaxScaler` ha invece permesso di standardizzare le feature in modo da poterle confrontare e utilizzare in modo più efficace durante la classificazione. Questi metodi hanno permesso di migliorare la capacità di generalizzazione del modello e di ridurre il rischio di overfitting, garantendo risultati più precisi e affidabili.

	precision	recall	f1-score	support
grecia - peloponneso	0.91	0.88	0.89	24
grecia-creta	0.86	0.98	0.91	44
italia	0.96	0.90	0.93	29
spagna	0.97	0.93	0.95	91
accuracy			0.93	188
macro avg	0.93	0.92	0.92	188
weighted avg	0.93	0.93	0.93	188

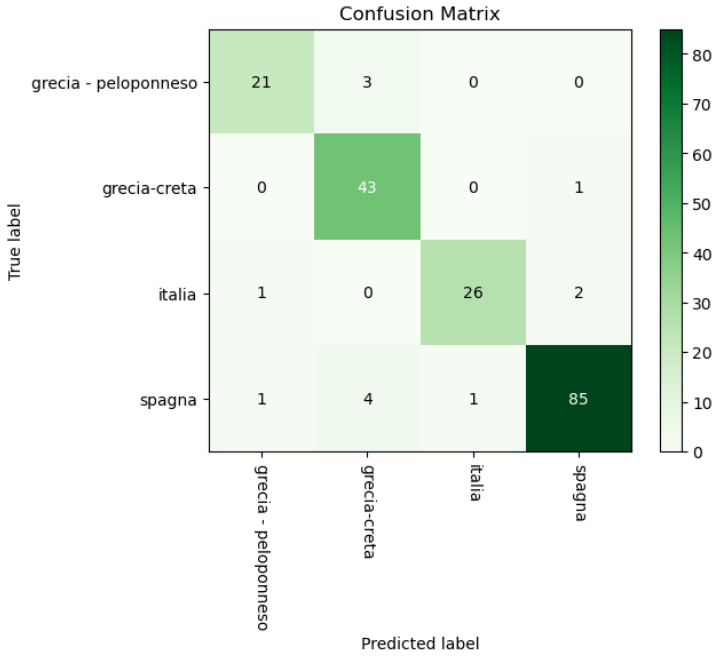

Figura 10: Matrice di confusione di SVC sul dataset semplificato

Il report di classificazione riportato sopra e la matrice di confusione in Figura 10 mostrano che l'algoritmo SVC ha raggiunto risultati di classificazione molto precisi per i campioni con target Grecia-Creta, Grecia-Peloponneso e Italia. La maggior parte degli errori di classificazione si è verificata per i campioni provenienti dalla Spagna. Tuttavia, considerando l'elevato numero di campioni a disposizione e quindi la possibilità di un numero maggiore di outlier, questi risultati possono ancora essere considerati soddisfacenti.

3.3 Confronto con i risultati presenti in letteratura

Facendo un confronto tra i risultati ottenuti dal nostro studio e quelli riscontrati in altre ricerche citate precedentemente è possibile riscontrare varie

differenze, le quali dipendono principalmente dal tipo di caratteristiche prese in esame e dagli approcci utilizzati per elaborarle.

L'articolo "Ntopia" [3] si propone di caratterizzare il profilo aromatico dell'olio d'oliva della cultivar "Ntopia" proveniente da diverse isole Ionie in Grecia, e di indagare se alcuni composti volatili possano essere considerati come indicatori della loro origine geografica. A tal fine, sono stati utilizzati 137 campioni di olio d'oliva e opportuni algoritmi di apprendimento automatico rapido, utilizzando il metodo della cross-validazione è stata raggiunta un'accuratezza dell'85,7%.
Nell'articolo "Geographical authentication GC–MS" [4] viene posto come problema quello della verifica della conformità degli oli UE, per affrontare questo problema viene sviluppato un modello di classificazione (PLS-DA) basato sull'impronta digitale degli idrocarburi sesquiterpenici di 400 campioni ottenuti da spettrometria di massa per discriminare tra oli di oliva UE e non UE, ottenendo un'accuratezza dell'89,6%.
I due studi proposti hanno un obiettivo simile a quello affrontato in questo lavoro, anche se con situazioni diverse: la prima è circoscritta a una zona più limitata (4 isole greche), mentre la seconda suddivide genericamente tutti gli oli in sole 2 macro-classi (UE e non UE).
In definitiva è comunque possibile affermare che i risultati raggiunti dal nostro modello sono leggermente superiori a entrambi questi studi anche se con aree geografiche che possono essere definite più generali di quelle affrontate nel primo articolo e più ristrette di quelle affrontate nel secondo articolo.

3.4 Creazione di uno script python

Per creare un'applicazione pratica, è stato sviluppato uno script Python che utilizza uno dei modello di machine learning testati in precedenza per predire l'origine di un campione sulla base del suo file di spettrometria di massa. Lo script richiede all'utente di inserire il percorso del file che contiene i dati

del campione in formato `mzdata.xml`. Dopo aver verificato la presenza del file, il modello addestrato nel primo approccio (e quindi su tutto il dataset) viene caricato da un file `modello.pickle`. Si è deciso di caricare nel file `modello.pickle` il modello addestrato di SVC (Support Vector Classifier) in quanto è quello che ha dato i risultati migliori nella prima parte del lavoro. Lo script quindi esegue la predizione e produce un risultato che mostra la percentuale di appartenenza del campione a ogni classe, corrispondente a una specifica origine. Questa applicazione dimostra l'efficacia del modello nell'identificare l'origine dei campioni. Il modello è stato addestrato sul dataset iniziale, ma potrebbe essere esteso per includere nuove origini o dati di spettrometria di massa per migliorare la sua accuratezza.

4 Conclusioni

Dopo aver esposto il problema dell'autenticazione dell'origine geografica degli EVOO e VOO e aver proposto una soluzione, è possibile affermare che i risultati ottenuti sono estremamente soddisfacenti. Come mostrato nel precedente capitolo, la nostra soluzione è stata in grado di predire l'origine geografica di un EVOO o VOO con un'accuratezza del 93% sul dataset semplificato e con un'accuratezza pari a 84% sul dataset originale. Tale risultato ci consente di concludere che la TIC rappresenta un marcatore importante dell'origine geografica degli EVOO e dei VOO confermando la nostra ipotesi iniziale. Grazie a questo lavoro si può notare come la TIC vari in modo significativo tra diverse aree geografiche distanti tra loro, dipendendo quindi dal clima, dalle condizioni del terreno, dall'esposizione al sole e da altri fattori legati al territorio. Il nostro metodo presenta diversi vantaggi rispetto ad altre tecniche di analisi chimiche complesse e specifiche. In particolare, il nostro approccio è molto più semplice e meno costoso, poiché si basa sul semplice utilizzo del file risultante dalla spettrometria di massa e non richiede personale specializzato. Inoltre, il nostro metodo è anche molto più veloce rispetto alle tecniche tradizionali di autenticazione dell'origine geografica degli EVOO e VOO. In questo contesto è importante sottolineare che il lavoro in questione non può garantire una predizione accurata dell'origine degli oli che provengono da regioni geograficamente vicine, come ad esempio da diverse zone di uno stesso Paese. Tuttavia, è importante notare che il lavoro in questione si basa sulla soluzione del problema dell'autenticazione degli Extra Virgin Olive Oil (EVOO) e dei Virgin Olive Oil (VOO) provenienti da grandi Paesi produttori di olio. In questo senso, il risultato ottenuto può essere considerato molto soddisfacente.

In conclusione, il nostro lavoro dimostra che la TIC rappresenta un marcatore affidabile dell'origine geografica degli EVOO e dei VOO provenienti da aree geografiche con caratteristiche diverse e distanti tra loro. In ogni caso,

il lavoro svolto rappresenta un importante passo avanti nell'ambito dell'autenticazione degli oli in quanto fornisce una soluzione parziale al problema della contraffazione e della falsificazione degli oli di alta qualità, specialmente quelli prodotti nei Paesi in cui c'è un'importante produzione.

4.1 Lavori futuri

In lavori futuri, sarebbe interessante valutare le prestazioni del nostro modello su dataset di dimensioni maggiori, con un minor sbilanciamento rispetto al nostro dataset attuale. In particolare, il modello dovrebbe essere testato su un dataset contenente EVOO e VOO provenienti da un maggior numero di Paesi, anche molto vicini tra loro. In questo modo, sarebbe possibile verificare come il modello si comporta con campioni la cui TIC varia poco da quelli provenienti da Paesi vicini o confinanti. Il modello andrebbe poi testato su campioni di oli appartenenti a campagne olearie differenti per vedere come esso si comporta. Tali esperimenti permetterebbero di valutare l'efficacia del nostro metodo in condizioni più realistiche e rappresentative della diversità degli oli di oliva presenti oggi in commercio.

Riferimenti bibliografici

[1] B. Quintanilla-Casas, S. Bertin, K. Leik, J. Bustamante, F. Guardiola, E. Valli, A. Bendini, T. Gallina Toschi, A. Tres, S. Vichi Profiling versus fingerprinting analysis of sesquiterpene hydrocarbons for the geographical authentication of extra virgin olive oils. 2020 .

[2] Ben Hmida, B. Gargouri, F. Chtourou, D. Sevim, M. Bouaziz. Fatty acid and triacyglycerid as markers of virgin olive oil from Mediterranean region: Traceability and chemometric authentication. 2022

[3] Eriotou, E.; Karabagias, I.K.; Maina, S.; Koulougliotis, D.; Kopsahelis, N. Geographical origin discrimination of "Ntopia" olive oil cultivar from Ionian islands using volatile compounds analysis and computational statistics. 2021

[4] Quintanilla-Casas, B.; Torres-Cobos, B.; Guardiola, F.; Servili, M.; Alonso-Salces, R.M.; Valli, E.; Bendini, A.; Toschi, T.G.; Vichi, S. Tres, A. Geographical authentication of virgin olive oil by GC–MS sesquiterpene hydrocarbon fingerprint: Verifying EU and single country label-declaration. Food Chem. 2022.

[5] Maestrello, V.; Solovyev, P.; Bontempo, L.; Mannina, L.; Camin, F.Spettroscopia di risonanza magnetica nucleare nell'autenticazione dell'olio extra vergine di oliva. compr. Rev. Cibo Sci. Sicurezza alimentare. 2022 .