UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Struttura Didattica di Riferimento:

Dipartimento di informatica

CORSO DI LAUREA IN INFORMATICA

Tesi di laurea in
Metodi Avanzati di Programmazione

TECNICHE DI MACHINE LEARNING PER LA SEGMENTAZIONE SEMANTICA DI IMMAGINI SENTINEL-2

Relatori

Prof.ssa Annalisa Appice

Dott.ssa Giuseppina Andresini

Laureando

Davide Guerra

Anno Accademico 2022-2023

Indice

Introduzione

> *"Per effetto dei cambiamenti climatici che si ripercuotono negativamente per la pianta ospite sottoposta a stress idrici estivi e alle violenze degli eventi estremi, e positivamente per il piccolo coleottero che può riprodursi più volte, le foreste europee sono di fronte ad un pericolo senza precedenti."* [1]

Ben noti sono i rischi e le conseguenze che l'umanità si sta portando dietro con il progredire dei forti ed insoliti cambiamenti climatici causati dal riscaldamento globale che sta deteriorando la vivibilità dell'ecosistema per la nostra e le altre specie.

A seguito dell'avvento della tempesta Vaia, una forte tempesta mediterranea accaduta tra ottobre e novembre 2018, che ha colpito varie nazioni, tra cui Francia, Italia, Croazia, Austria e Svizzera [2], c'è stato uno sfoltimento ed indebolimento delle aree boschive nelle zone interessate. Si pensi che solo nel territorio italiano tra le quattro regioni principalmente colpite (Lombardia, Trentino- Alto Adige, Veneto e Friuli-Venezia Giulia) si è registrata in pochi minuti una perdita di 41.000 ettari di boschi [1].

La tempesta oltre gli ingenti danni ha indebolito gli alberi sopravvissuti ammorbidendone il tronco e creando le condizioni ideali per la diffusione dello scarabeo della corteccia (o bostrico tipografo [3]). Si tratta di un coleottero, le cui dimensioni variano tra i 4 e i 5 millimetri (per l'esemplare adulto), che attacca prevalentemente l'abete rosso, in cui vi si sviluppa scavando intricate gallerie al di sotto della corteccia che interrompono il flusso della linfa, portando in tempi brevi la pianta alla morte.

L'individuazione preventiva delle zone contagiate è decisiva per poter far fronte a questa minaccia ambientale [3], essa è realizzabile tramite l'ausilio di tecnologie moderne come satelliti, in grado di raccogliere un grande quantitativo di dati sull'epidemia, e l'intelligenza artificiale (IA), capace di elaborare questa mole di dati ed estrapolare informazioni importanti. Quest'ultima sta prendendo sempre più piede nella nostra quotidianità affiancandoci in compiti complessi come quello sopracitato, ma di cosa si tratta?

> *"L'intelligenza artificiale è una disciplina appartenente all'informatica che studia i fondamenti teorici, le metodologie e le*

> *tecniche che consentono la progettazione di sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana." [4]*

Pertanto, l'IA permette alla macchina di effettuare un ragionamento, ma, nonostante all'apparenza assomigli all'intelligenza umana, il suo scopo non è quello di simularla, bensì quello di riprodurne o emularne alcune funzionalità, ed è proprio in questo ambito che è in grado di fornire prestazioni qualitativamente equivalenti e quantitativamente superiori ad essa [5]. Tuttavia, il concetto di intelligenza artificiale è generico.

L'IA è divisa in branche, una di esse è il Machine Learning. Esso fa riferimento al processo tramite il quale la macchina sviluppa la capacità di riconoscere dei pattern, o di apprendere continuamente, effettuando, così, delle previsioni utilizzando i dati per poi apportare modifiche in autonomia senza una programmazione specifica. [6]

La tesi in questione si propone di analizzare come l'IA ci possa aiutare a contrastare l'epidemia dello scarabeo della corteccia, partendo dall'analisi delle immagini satellitari per addestrare due modelli (dei classificatori) capaci di valutare il grado di infezione di una o molteplici aree forestali per mezzo di una scala di quattro livelli.

Nel primo capitolo illustreremo il funzionamento del "remote sensing" utilizzato da Sentinel-2, fondamentale per il rilevamento delle zone forestali infette e la raccolta dei dati necessari per addestrare i modelli.

Nel secondo capitolo discuteremo dei concetti principali dell'analisi spettrale (in particolare delle immagini spettrali, gli indici spettrali ed i valori di mutual info), successivamente si introdurranno i due modelli d'apprendimento utilizzati in questo progetto: "Random Forest" e "XGBoost".

Nel terzo capitolo verrà illustrato l'approccio che si è deciso di impiegare a livello di software per affrontare il problema in questione, vi sarà inoltre mostrato il diagramma delle classi.

Nel quarto ed ultimo capitolo visioneremo i dati, le metriche e le configurazioni utilizzate per estrapolare i risultati finali. In chiusura, saranno indicati alcuni possibili sviluppi futuri del presente software.

Capitolo 1

Remote sensing

Il remote sensing (telerilevamento) è il processo per il rilevamento e monitoraggio delle caratteristiche fisiche di un'area, senza il contatto diretto con essa, tramite l'utilizzo di droni, aerei o satelliti. Telecamere speciali raccolgono immagini telerilevate che aiutano i ricercatori a "percepire" determinate condizioni ambientali sulla terra [7].

Per effettuare un telerilevamento sono necessari tre requisiti indispensabili [13]:

- un oggetto da osservare;
- una piattaforma che sostiene lo strumento che utilizzeremo per il monitoraggio;
- un sensore che rilevi i parametri di nostro interesse.

Il telerilevamento, come illustrato nella Fig. 1.1, può essere suddiviso in due metodi: il primo è il telerilevamento passivo dove i sensori, detti "passivi", raccolgono le radiazioni riflesse dagli oggetti o dalle aree circostanti (solitamente la luce solare è la fonte di radiazione più comunemente misurata da questo tipo di sensori); il secondo metodo riguarda la raccolta attiva dove è il satellite stesso ad emettere energia, per poi misurare la radiazione riflessa dal bersaglio [9]. In entrambi i casi, una volta che i dati sono stati raccolti, vengono trasmessi a stazioni di ricerca sulla Terra per essere elaborati e analizzati.

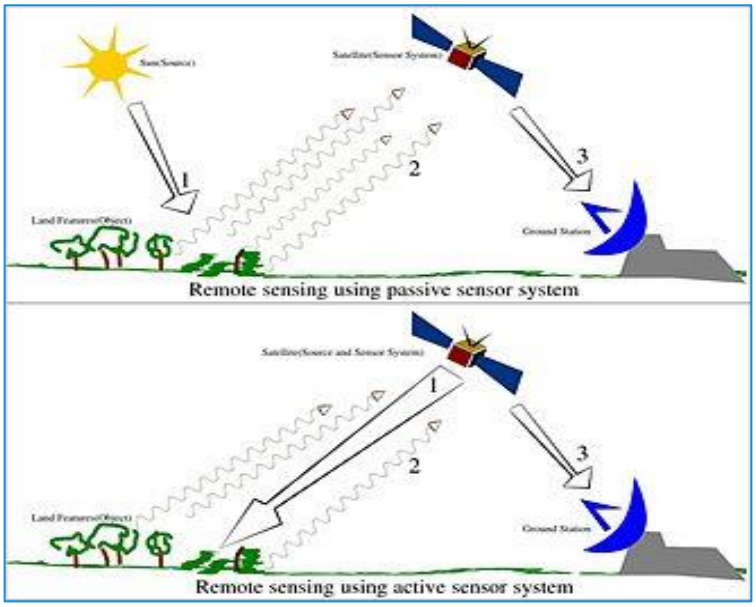

Fig. 1.1: metodi del telerilevamento [9].

Il telerilevamento consente di raccogliere dati in aree pericolose o inaccessibili. Alcuni dei suoi utilizzi più comuni possono includere: mappatura dei grandi incendi boschivi; monitoraggio delle nuvole per aiutare a prevedere il tempo; monitoraggio della deforestazione in aree come il bacino amazzonico; caratteristiche glaciali nelle regioni antartiche; e così via ([7], [9]).

1.1 Scoperta di zone forestali infestate dallo scarabeo della corteccia

A partire dal 2015 è operativa la missione Sentinel-2 del programma di telerilevamento Copernicus guidato dalla Commissione europea in collaborazione con l'Agenzia Spaziale Europea (ESA). La missione Sentinel-2 è stata sviluppata specificatamente per il monitoraggio ad alta risoluzione del territorio italiano. Essa è composta da due satelliti gemelli, Sentinal-2A e Sentinel-2B, che seguono un'orbita polare con un tempo di rivisitazione dai 3 ai 5 giorni. Questa missione assegnata ai satelliti Sentinel-2, dopo gli eventi che si sono susseguiti alla tempesta vaia, è stata estesa anche al monitoraggio dell'epidemia causata dallo scarabeo della corteccia.

Benché dopo la raccolta di dati tramite il telerilevamento sia molto difficile riconoscere i singoli alberi infestati che presentano caratteristiche del tutto simili agli esemplari sani, è importante riconoscere repentinamente i primi sintomi di attacco. Difatti la loro precoce individuazione ed il loro immediato abbattimento, susseguito da esbosco o scortecciatura, costituiscono nell'insieme la più efficace misura di lotta contro il bostrico. Un metodo efficace per l'individuazione degli esemplari infetti è tenere conto delle zone che presentano alberi dalla chioma arrossata e secca, i cui parassiti hanno già abbandonato il fusto per migrare verso quelli vicini [3].

Capitolo 2

Concetti preliminari

In questo capitolo discuteremo dei concetti preliminari che si avrà bisogno di acquisire per comprendere la struttura del progetto. In primo luogo, mostreremo le caratteristiche dell'analisi spettrale e, successivamente, introduceremo i modelli d'apprendimento "Random Forest" e "XGBoost".

2.1 Analisi di dati spettrali

Al fine di ottenere i dati necessari per l'addestramento dei modelli è necessario effettuare un'analisi dei dati spettrali, nel nostro caso immagini satellitari scattate da sentinel-2. Qui di seguito spiegheremo cosa sono le "immagini multispettrali", gli "indici spettrali" e i valori di mutual info.

2.1.1 Immagini multispettrali

I satelliti Sentinel-2 sono muniti del sensore MSI (Multispectral Instrument [8]) che riesce a catturare le cosiddette "immagini multispettrali".

Queste racchiudono dati relativi a più bande, ovvero altre immagini del medesimo luogo catturate su specifici intervalli di lunghezza d'onda all'interno dello spettro elettromagnetico. Ad esempio, un'immagine multispettrale potrebbe includere *n* bande spettrali che vanno dalle onde radio a quelle gamma con lunghezze crescenti [10].

Il sensore MSI funge da filtro per catturare ciascun tipo di banda. Nel nostro caso i tipi di linee d'onda (bande) catturate sono 13 e rientrano nella scala dal visibile/infrarosso (VNIR) e nella gamma spettrale infrarossa a onde corte (SWIR).

All’interno delle 13 bande, come illustrato nelle seguenti figure, la risoluzione spaziale varia [11]. Le bande B2, B3, B4 e B8 sono contenute all’interno della scala VNIR ed hanno una risoluzione spaziale di 10m (Fig. 2.1); Le bande B5, B6, B7 e B8a, sono contenuti nella scala VNIR, mentre le bande B11 e B12 sono contenute nella scala SWIR ed hanno una risoluzione di 20m (Fig. 2.2); Infine, la banda B1 è contenuta nella scala VNIR, la banda B10 è contenuta nella scala SWIR, mentre la banda B2 è contenuta sia in VNIR che SWIR, esse hanno una risoluzione di 60m (Fig. 2.3), [12].

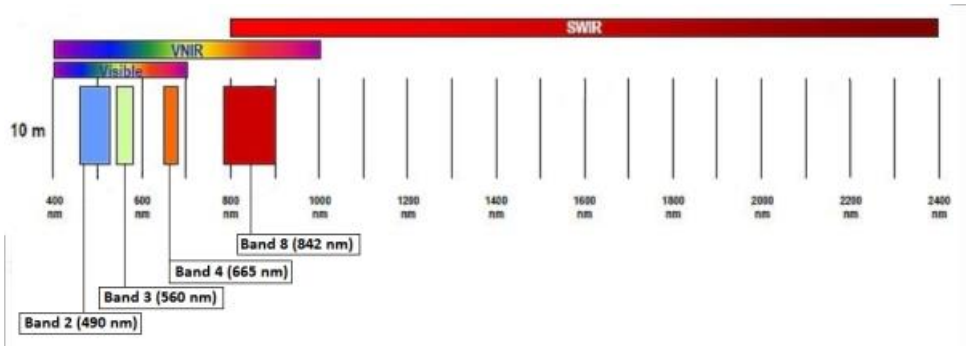

Fig. 2.1: Bande di risoluzione spaziale sentinel-2 di 10 m: B2, B3, B4 e B8, [12].

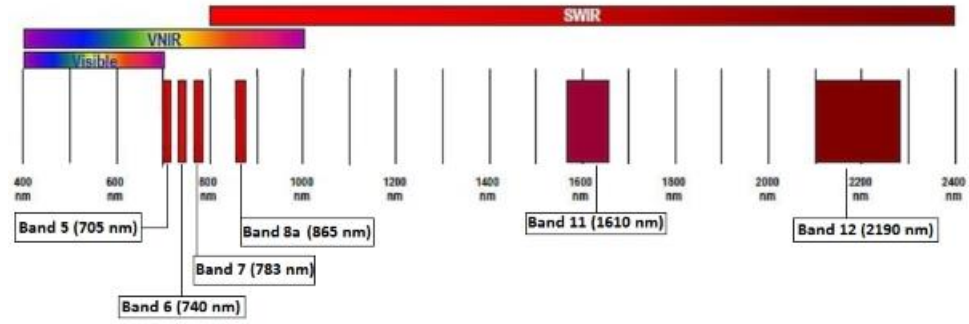

Fig. 2.2: Bande di risoluzione spaziale sentinel-2 di 20 m: B5, B6, B7 e B8a, B11, B12[12].

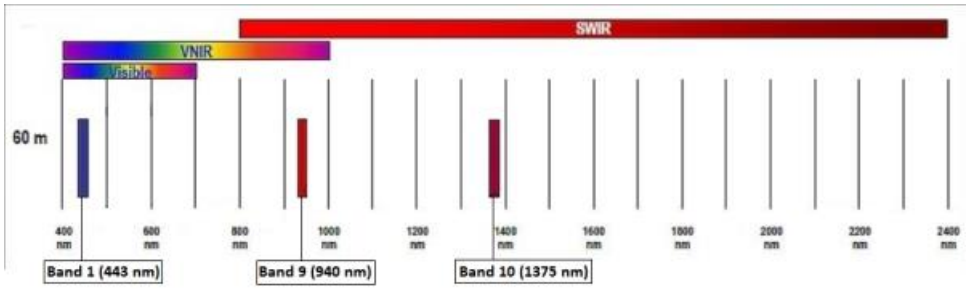

Fig. 2.3: Bande di risoluzione spaziale sentinel-2 di 60 m: B1, B9, B10, [12].

2.1.2 Indici spettrali

Gli indici spettrali sono combinazioni matematiche di diverse bande spettrali di immagini satellitari o aeree che hanno una vasta gamma di applicazioni nel telerilevamento atte ad evidenziare proprietà specifiche della superficie osservata, come: i cambiamenti nell'uso del suolo e nella copertura del suolo; l'urbanizzazione; la deforestazione; la desertificazione o disastri naturali (inondazioni, incendi e frane); ed altro ancora. [15].

Un esempio di indici spettrali sono gli indici di vegetazione, specificamente progettati per evidenziare la presenza e le condizioni della vegetazione. Essi possono essere calcolati solo su specifiche bande sensibili alla presenza di vegetazione, come le bande del rosso (RED), del vicino infrarosso (NIR) e dell'infrarosso ad onde corte (SWIR). Le valutazioni si basano sulla percentuale di radiazione catturata da queste bande.

Tra gli indici di vegetazione più utilizzati c'è il "Normalized Difference Vegetation Index" (NDVI) che descrive il livello di vigoria della cultura e si calcola come rapporto tra la differenza e la somma della radiazione riflessa nel vicino infrarosso e nel rosso che si esprime numericamente con $(NIR - RED) \div (NIR + RED)$ [14].

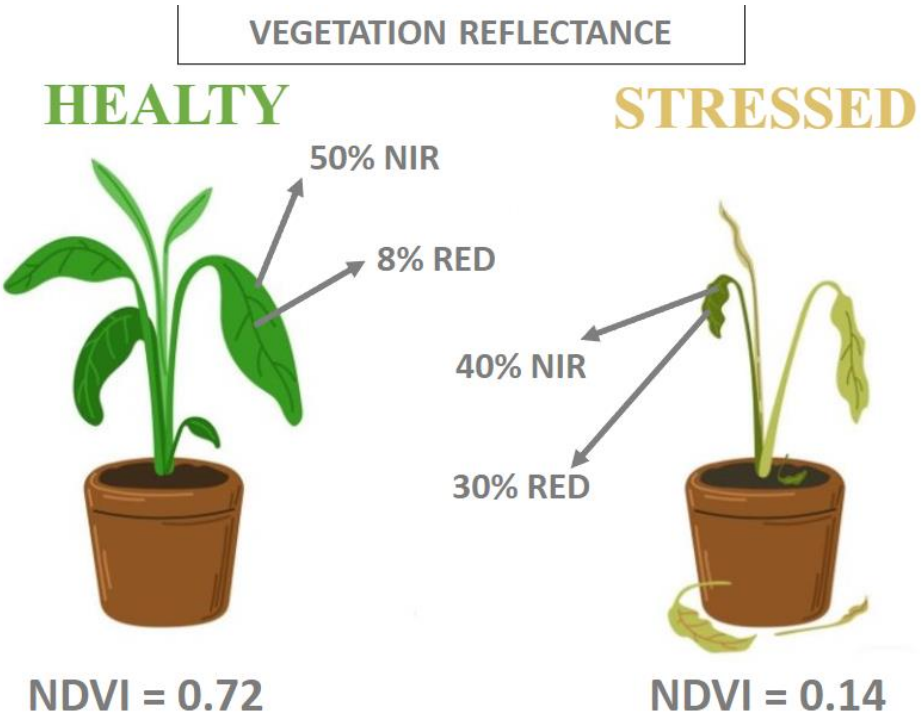

Fig. 2.4: risultato di NDVI in base alle percentuali di radiazione di NIR e RED.

NDVI è un indice altamente informativo e semplice da interpretare. I valori possono variare da –1 a 1, ma quelli compresi tra –1 e 0 sono tipicamente aree non coltivate di cui, ad esempio, fanno parte fiumi, nuvole ecc. I restanti valori corrispondono a diverse situazioni agronomiche, come indicato nella Fig. 2.5, indipendentemente dalla cultura.

NDVI	SIGNIFICATO
< 0.1	Suolo privo di vegetazione o nuvole
0.1 - 0.2	Copertura vegetale quasi assente
0.2 - 0.3	Copertura vegetale molto bassa
0.3 - 0.4	Copertura vegetale bassa con vigoria bassa oppure copertura vegetale molto bassa con vigoria alta
0.4 - 0.5	Copertiiura vegetale medio-bassa con vigoria bassa oppure copertura vegetale medio-bassa con vigoria alta
0.5 - 0.6	Copertiiura vegetale media con vigoria bassa oppure copertura vegetale medio-bassa con vigoria alta
0.6 - 0.7	Copertiiura vegetale medio-alta con vigoria bassa oppure copertura vegetale media con vigoria alta
0.7 - 0.8	Copertura vegetale alta con vigoria alta
0.8 - 0.9	Copertura vegetale molto alta con vigoria molto alta
0.9 - 1.0	Copertura vegetale totale con vigoria molto alta

Fig. 2.5: tabella di valutazione dello stato dei terreni in base al risultato dell'indice NDVI.

Un altro indice è il "Normalized Difference Moisture Index" (NDMI) che descrive il livello di stress idrico della coltura e si colloca come il rapporto tra la differenza e la somma della radiazione nel vicino infrarosso e nello SWIR, ossia $(NIR - SWIR) \div (NIR + SWIR)$.

Anche in questo caso, NDMI è semplice da interpretare e comprende valori che variano da –1 a 1.

NDMI	SIGNIFICATO
-0.1 - -0.8	Suolo privo di vegetazione o nuvole
-0.8 - -0.6	Copertura vegetale quasi assente
-0.6 - -0.4	Copertura vegetale molto bassa
-0.4 - -0.2	Copertura vegetale bassa con sress idrico alto o copertura vegetale molto bassa con stress idrico basso
-0.2 - 0	Copertiiura vegetale medio-bassa con stress idrico alto o copertura vegetalr bassa con stress idrico basso
0 - 0.2	Copertiiura vegetale media con stress idrico o copertura vegetale medio-bassa con stress idrico basso
0.2 - 0.4	Copertiiura vegetale medio-alta con sress idrico alto o copertura vegetale medio-bassa con stress idrico basso
0.4 - 0.6	Copertura vegetale alta e no stress idrico
0.6 - 0.8	Copertura vegetale molto alta e no stress idrico
0.8 - 1.0	Copertura vegetale totale e no stress idrico o ristagni o nuvole

Fig. 2.6: tabella di valutazione dello stato dei terreni in base al risultato dell'indice NDMI.

Tabella 1:

INDICI SPETTRALI

Name	Abbreviation	General Formula
Adjusted transformed soil-adjusted VI	ATSAVI	$1.22 \cdot \frac{(B08 - 1.22 \cdot B04 - 0.03)}{1.22 \cdot B08 + B04 - 1.22 \cdot 0.03 + 0.08(1.0 + 1.22^2}$
Aerosol free vegetation index 1600	AFRI1600	$\left(B08 - 0.66 \cdot \frac{B11}{B08 + 0.66 \cdot B11}\right)$
Aerosol free vegetation index 2100	AFRI2100	$\left(B08 - 0.5 \cdot \frac{B12}{B08 + 0.56 \cdot B12}\right)$
Alteration		$\frac{B11}{B12}$

Anthocyanin reflectance index	ARI	$\frac{1}{B03} - \frac{1}{B05}$
Ashburn Vegetation Index	AVI	$2.0 \cdot B09 - B04$
Atmospherically Resistant Vegetation Index	ARVI	$\frac{B8A - B04 - y(B04 - B02)}{B8A + B04 - y(B04 - B02)}$
Atmospherically Resistant Vegetation Index 2	ARVI2	$-0.18 + 1.17 \cdot \left(\frac{B08 - B04}{B08 + B04}\right)$
Blue-wide dynamic range vegetation index	BWDRVI	$\frac{0.1 \cdot B08 - B02}{0.1 \cdot B08 + B02}$
Browning Reflectance Index	BRI	$\frac{\frac{1}{B03} - \frac{1}{B05}}{B08}$
Canopy Chlorophyll Content Index	CCCI	$\frac{\frac{B08 - B05}{B08 + B05}}{\frac{B08 - B04}{B08 + B04}}$
Chlorophyll Absorption Ratio Index	CARI	$\frac{\frac{B05}{B04}}{\cdot \frac{\sqrt{\left(\frac{B05 - B03}{150} \cdot 670 + 4 + \left(3 - \frac{B05 - B03}{150} \cdot 550\right)\right.}}{\left(\frac{B05 - B03}{150^2} + 1\right)^{0.5}}}$
Chlorophyll Absorption Ratio Index 2	CARI2	$\frac{\left\lvert\left(\frac{B05 - B03}{150} \cdot 4 + 4 + 3 - 0.496 \cdot B03\right)\right\rvert}{(a2 + 1)^{0.5}} \cdot \frac{B05}{B04}$
Chlorophyll Green	Chlgreen	$\left(\frac{B07}{B03}\right)^{(-1)}$
Chlorophyll Index Green	CIgreen	$\frac{B07}{B03} - 1$
Chlorophyll IndexRedEdge	CIrededge	$\frac{B08}{B03} - 1$
Chlorophyll Red-Edge	Chlred-edge	$\left(\frac{B07}{B05}\right)^{(-1)}$
Chlorophyll vegetation index	CVI	$B08 \cdot \frac{B04}{B03^2}$
Coloration Index	CI	$\frac{B04 - B02}{B04}$

Corrected Transformed Vegetation Index	CTVI	$\frac{\frac{B05-B03}{B05+B03}+0{,}5}{\left\lvert\frac{B05-B03}{B05+B03}+0{,}5\right\rvert}\cdot\sqrt{\left\lvert\left(\frac{B05-B03}{B05+B03}\right)+0.5\right\rvert}$
CRI550	CRI550	$B02^{(-1)}-B03^{(-1)}$
CRI700	CRI700	$B02^{(-1)}-B05^{(-1)}$
Datt1	Datt1	$\frac{B08-B05}{B08-B04}$
Datt4	Datt4	$\frac{B04}{B03*B05}$
Datt6	Datt6	$\frac{B8A}{B03*B05}$
Difference 678/500	D678/500	$B04-B02$
Difference 800/550	D800/550	$B08-B03$
Difference 800/680	D800/680	$B08-B04$
Difference NIR/Green Green Difference Vegetation Index	GDVI	$B08-B03$
Differenced Vegetation Index MSS	DVIMSS	$2.4*B09-B04$
Enhanced Vegetation Index	EVI	$2.5\frac{(B08-B04)}{(B08+6.0*B04-7.5*B02)+1.0}$
Enhanced Vegetation Index 2	EVI2	$2.5\cdot\frac{B08-B04}{B08+2.4\cdot B04+1}$
EPI	EPI	$0.331\cdot\frac{B04}{(B03\cdot B05)^{0.329}}$
Ferric iron, Fe2+	Fe2+	$\frac{B12}{B08}+\frac{B03}{B04}$
Ferric iron, Fe3+	Fe3+	$\frac{B04}{B03}$
Ferric Oxides		$\frac{B11}{B08}$
Ferrous iron		$\frac{B12}{B08}+\frac{B03}{B04}$
Ferrous Silicates		$\frac{B12}{B11}$

Global Environment Monitoring Index	GEMI	$\frac{2(B08^2 - B05^2) + 1.5 \cdot B09 + 0.5 \cdot B05}{B09 + B05 + 0.5} \cdot 1 - 0.25$ · $\frac{2(B08^2 - B05^2) + 1.5 \cdot B09 + 0.5 \cdot B05}{B09 + B05 + 0.5} - B05$ $- \frac{0.125}{1 - B04}$
Global Vegetation Moisture Index	GVMI	$\frac{(B08 + 0.1) - (B12 + 0.02)}{(B08 + 0.1) + (B12 + 0.02)}$
Gossan		$\frac{B11}{B04}$
Green atmospherically resistant vegetation index	GARI	$\frac{B08 - (B03 - (B02 - B04))}{B08 - (B03 + (B02 - B04))}$
Green leaf index	GLI	$\frac{2 \cdot B03 - B04 - B02}{2 \cdot B03 + B04 + B02}$
Green Normalized Difference Vegetation Index	GNDVI	$\frac{B08 - B03}{B08 + B03}$
Green Optimized Soil Adjusted Vegetation Index	GOSAVI	$\frac{B08 - B03}{B08 + B03 + 0.120}$
Green Soil Adjusted Vegetation Index	GSAVI	$\frac{B08 - B03}{B08 + B03 + 0.482} \cdot 1 + 0.482$
Green-Blue NDVI	GBNDVI	$\frac{B08 - (B03 + B02)}{B08 + (B03 + B02)}$
Green-Red NDVI	GRNDVI	$\frac{B08 - (B03 + B04)}{B08 + (B03 + B04)}$
Hue	H	$\arctan\left(\frac{2.0 * B04 - B03 - B02}{30.5 * (B03 - B02)}\right)$
Ideal vegetation index	IVI	$\frac{B08 - 0.809}{0.393 \cdot B04}$
Intensity	I	$\frac{1}{30.5} \cdot B04 + B03 + B02$
Inverse reflectance 550	IR550	$B03^{-1}$
Inverse reflectance 700	IR700	$B05^{-1}$
Laterite		$\frac{B11}{B12}$

Leaf Chlorophyll Index	LCI	$\frac{B08 - B05}{B08 + B04}$
Leaf Water Content Index	LWCI	$\frac{\log(1 - (B08 - 0.101))}{-\log(1 - (B08 - 0.101))}$
Log Ratio	LogR	$\log\left(\frac{B08}{B04}\right)$
Maccioni		$\frac{B07 - B05}{B07 - B04}$
MCARI/MTVI2	MCARI/MTVI2	$\frac{((B05 - B04) - 0.2(B05 - B03))\left(\frac{B05}{B04}\right)}{1.5 \cdot \frac{1.2 * (B08 - B03) - 2.5 * (B04 - B03)}{\sqrt{(2 \cdot B08 + 1)^2 - (6 * B08 - 5 \cdot \sqrt{B04}) - 0.5}}}$
MCARI/OSAVI	MCARI/OSAVI	$\frac{((B05 - B04) - 0.2(B05 - B03))\left(\frac{B05}{B04}\right)}{(1 + 0.16) \cdot \frac{(B08 - B04)}{(B08 + B04 + 0.16)}}$
mCRIG	mCRIG	$\left(B02^{(-1)} - B03^{(-1)}\right) \cdot B08$
mCRIRE	mCRIRE	$\left(B02^{(-1)} - B05^{(-1)}\right) \cdot B08$
Mid-infrared vegetation index	MVI	$\frac{B09}{B11}$
Misra Green Vegetation Index	MGVI	$0.386 * B03 - 0.53 * B04 + 0.535 * B06 + 0.532 * B09$
Misra Non Such Index	MNSI	$0.404 * B03 - 0.039 * B04 - 0.505 * B06 + 0.762 * B09$
Misra Soil Brightness Index	MSBI	$0.406 * B03 + 0.6 * B04 + 0.645 * B06 + 0.243 * B09$
Misra Yellow Vegetation Index	MYVI	$0.723 * B03 - 0.597 * B04 + 0.206 * B06 - 0.278 * B09$
mND680	mND680	$\frac{B06 - 0.278 * B09}{(B08 + B04 - 2.0 * B01)}$
Modified anthocyanin reflectance index	mARI	$(B03^{-1} - B05^{-1}) \cdot B08$
Modified Chlorophyll Absorption in Reflectance Index	MCARI	$(B05 - B04) - 0.2 * (B05 - B03) * \frac{B05}{B04}$
Modified Chlorophyll Absorption in Reflectance Index 1	MCARI1	$(B05 - B04) - 0.2 * (B05 - B03) * \frac{B05}{B04}$

Modified Chlorophyll Absorption in Reflectance Index 2	MCARI2	$1.5\frac{2.5*(B08-B04)-1.3*(B08-B03)}{\sqrt{(2*B08+1)^2-(6*B08-5\sqrt{B04})-0.5}}$
Modified NDVI	mNDVI	$\frac{(B08-B04)}{(B08+B04-2.0*B01)}$
Modified Simple Ratio	mSR	$\frac{(B08-B04)}{(B04-B01)}$
Modified Simple Ratio 670,800	MSR670	$\frac{\frac{B08}{B04-1}}{\sqrt{\frac{B08}{B04+1}}}$
Modified Simple Ratio NIR/RED	MSRNir/Red	$\frac{\frac{B08}{B04}-1}{\sqrt{\frac{B08}{B04}+1}}$
Modified Soil Adjusted Vegetation Index	MSAVI	$\frac{2.0*B08+1.0-\sqrt{(2\cdot B08+1)^2-8*(B08-B}}{2}$
Modified Soil Adjusted Vegetation Index hyper	MSAVIhyper	$0.5*((2.0*B08+1.0)-\sqrt{(2.0*B08+1.0)^2-8.0*(B08-B04)}$
Modified Triangular Vegetation Index 1	MTVI1	$1.2*(1.2*(B08-B03)-2.5*(B04-B03))$
Modified Triangular Vegetation Index 2	MTVI2	$1.5\cdot\frac{1.2*(B08-B03)-2.5*(B04-B03)}{\sqrt{(2.0*B08+1.0)^2-(6.0*B08-5.0*\sqrt{B04})-}}$
Nonlinear vegetation index	NLI	$\frac{B08^2-B04}{B08^2+B04}$
Norm G	Norm G	$\frac{B03}{(B08+B04+B03)}$
Norm NIR	Norm NIR	$\frac{B08}{(B08+B04+B03)}$
Norm R	Norm R	$\frac{B04}{(B08+B04+B03)}$
Normalized Difference 550/450 Plant pigment ratio	PPR	$\frac{(B03-B01)}{(B0+B01)}\frac{(B03-B01)}{(B0+B01)}$
Normalized Difference 550/650 Photosynthetic vigour ratio	PVR	$\frac{(B03-B04)}{(B0+B04)}$

Normalized Difference 774/677	ND774/677	$\frac{(B07 - B04)}{(B07 + B04)}$
Normalized Difference 780/550 Green NDVI hyper	GNDVIhyper	$\frac{(B07 - B03)}{(B07 + B03)}$
Normalized Difference 782/666	ND782/666	$\frac{(B07 - B04)}{(B07 + B04)}$
Normalized Difference 790/670	ND790/670	$\frac{(B07 - B04)}{(B07 + B04)}$
Normalized Difference 800/2170	ND800/2170	$\frac{(B03 - B04)}{(B03 + B04)}$
Normalized Difference 800/470 Pigment specific normalised difference C2	PSNDc2	$\frac{(B08 - B02)}{(B08 + B02)}$
Normalized Difference 800/500 Pigment specific normalised difference C1	PSNDc1	$\frac{(B08 - B02)}{(B08 + B02)}$
Normalized Difference 800/550 Green NDVI hyper 2	GNDVIhyper2	$\frac{(B08 - B03)}{(B08 + B03)}$
Normalized Difference 800/650 Pigment specific normalised difference B1	PSNDb1	$\frac{(B08 - B04)}{(B08 + B04)}$
Normalized Difference 800/675 Pigment specific normalised difference A1	PSNDa1	$\frac{(B08 - B04)}{(B08 + B04)}$
Normalized Difference 800/680 Pigment specific normalised difference A2, Lichtenthaler indices 1, NDVIhyper	ND800/680	$\frac{(B08 - B04)}{(B08 + B04)}$
Normalized Difference 819/1600 NDII	NDII	$\frac{(B08 - B11)}{(B08 + B11)}$
Normalized Difference 819/1649 NDII 2	NDII2	$\frac{(B08 - B11)}{(B08 + B11)}$

Normalized Difference 820/1600 Normalized Difference Moisture Index	NDMI	$\frac{(B08 - B11)}{(B08 + B11)}$
Normalized Difference 827/668	ND827/668	$\frac{(B08 - B04)}{(B08 + B04)}$
Normalized Difference 833/1649 Infrared Index	ND833/1649	$\frac{(B08 - B11)}{(B08 + B11)}$
Normalized Difference 833/658	ND833/658	$\frac{(B08 - B04)}{(B08 + B04)}$
Normalized Difference 860/1640	SIWSI	$\frac{(B8A - B11)}{(B8A + B11)}$
Normalized Difference 895/675	ND895/675	$\frac{(B08 - B04)}{(B08 + B04)}$
Normalized Difference Green/Red Normalized green red difference index, Visible Atmospherically Resistant Indices Green (VIgreen)	NGRDI	$\frac{(B03 - B04)}{(B03 + B04)}$
Normalized Difference MIR/NIR Normalized Difference Vegetation Index (in case of strong atmospheric disturbances)	NDVI	$\frac{(B12 - B08)}{(B12 + B08)}$
Normalized Difference NIR/Blue Blue-normalized difference vegetation index	BNDVI	$\frac{(B08 - B02)}{(B08 + B02)}$
Normalized Difference NIR/MIR Modified Normalized Difference Vegetation Index	MNDVI	$\frac{(B08 - B12)}{(B08 + B12)}$
Normalized Difference NIR/Rededge Normalized Difference Red-Edge	NDRE	$\frac{(B08 - B04)}{(B08 + B04)}$

Normalized Difference NIR/SWIR Normalized Burn Ratio	NBR	$\frac{(B08 - B12)}{(B08 + B12)}$
Normalized Difference Red/Green Redness Index	RI	$\frac{(B04 - B03)}{(B04 + B03)}$
Normalized Difference Salinity Index	NDSI	$\frac{(B11 - B12)}{(B11 + B12)}$
Normalized Difference Vegetation Index 690-710	NDVI690-710	$\frac{(B08 - B05)}{(B08 + B05)}$
Normalized Difference Vegetation Index C	NDVIc	$\frac{(B08 - B04)}{(B08 + B04)} * \left(1.0 - \frac{B12 - 0.378}{0.397 - 0.027}\right)$
Optimized Soil Adjusted Vegetation Index	OSAVI	$(1 - Y)\frac{800nm - 670nm}{800nm + 670nm + Y}$
Pan NDVI	PNDVI	$\frac{NIR - (GREEN + RED + BLUE)}{NIR + (GREEN + RED + BLUE)}$
Perpendicular Vegetation Index	PVI	$\frac{1}{\sqrt{0.149^2 + 1}} \cdot B08 - 0.374 - 0.735$
Ratio Analysis of Reflectance Spectra A1	RARSa1	$\frac{\frac{B04}{B05}}{\frac{0.722}{0.715}}$
Ratio Analysis of Reflectance Spectra A2	RARSa2	$\frac{\frac{B04}{B05}}{\frac{0.848}{0.989}}$
Ratio Analysis of Reflectance Spectra A3	RARSa3	$\frac{\frac{B04}{B08}}{\frac{0.382}{0.935}}$
Ratio Analysis of Reflectance Spectra A4	RARSa4	$\frac{\frac{B04}{B08}}{\frac{0.523}{0.866}}$
Ratio Analysis of Reflectance Spectra C3	RARSc3	$\frac{\frac{B08}{B02}}{\frac{0.712}{0.535}}$
Ratio Analysis of Reflectance Spectra C4	RARSc4	$\frac{\frac{B08}{B02}}{\frac{0.153}{0.503}}$

RDVI	RDVI	$\frac{(B08 - B04)}{(B08 + B04)^{0.5}}$
RDVI2	RDVI2	$\frac{(B08 - B04)}{\sqrt{(B08 + B04)}}$
Red edge 1	Rededge1	$\frac{B05}{B04}$
Red edge 2	Rededge2	$\frac{B05 - B04}{B05 + B04}$
Red-Blue NDVI	RBNDVI	$\frac{B08 - (B04 + B02)}{B08 + (B04 + B02)}$
Red-Edge Inflection Point 1	REIP1	$700.0 + 40.0 * \left(\frac{\frac{(B04 + B07)}{2} - B05}{(B06 + B05)}\right)$
Red-Edge Inflection Point 2	REIP2	$700.0 + 40.0 * \left(\frac{\frac{(B04 + B07)}{2} - B05}{(B06 - B05)}\right)$
Red-Edge Inflection Point 3	REIP3	$705.0 + 35.0 * \left(\frac{\frac{(B04 + B07)}{2} - B05}{(B06 - B05)}\right)$
Red-Edge Position Linear Interpolation	REP	$700.0 + 40.0 * \left(\frac{\frac{(B04 + B07)}{2} - B05}{(B06 - B05)}\right)$
Reduced Simple Ratio	RSR	$\frac{B08}{B04} * 0.640 - \frac{B12}{0.640} - 0.259$
Reflectance at the inflexion point	Rre	$\frac{(B04 + B07)}{2.0}$
SAVImir	SAVImir	$(B08 - B12) \cdot \frac{(1.0 + 0.781)}{(B08 + B12 + 0.781)}$
Shape Index	IF	$\frac{(2.0 * B04 - B03 - B02)}{(B03 - B02)}$
Simple Ratio 1599/819 Moisture Stress Index 2	MSI2	$\frac{B11}{B08}$
Simple Ratio 1600/820 Moisture Stress Index	MSI	$\frac{B11}{B08}$
Simple Ratio 1650/2218	TM5/TM7	$\frac{B11}{B12}$

Simple Ratio 440/740	SR440/740	$\frac{B01}{B06}$
Simple Ratio 450/550 Blue green pigment index	BGI	$\frac{B01}{B03}$
Simple Ratio 520/670	SR520/670	$\frac{B02}{B04}$
Simple Ratio 550/670	SR550/670	$\frac{B03}{B04}$
Simple Ratio 550/680 Disease-Water Stress Index 4	DSWI-4	$\frac{B03}{B04}$
Simple Ratio 550/800	SR550/800	$\frac{B03}{B08}$
Simple Ratio 554/677 Greenness Index	GI	$\frac{B03}{B04}$
Simple Ratio 560/658 GRVIhyper	SR560/658	$\frac{B03}{B04}$
Simple Ratio 672/550 Datt5	SR672/550	$\frac{B04}{B03}$
Simple Ratio 672/708	SR672/708	$\frac{B04}{B05}$
Simple Ratio 674/553	SR674/553	$\frac{B04}{B03}$
Simple Ratio 675/555	SR675/555	$\frac{B04}{B03}$
Simple Ratio 675/700	SR675/700	$\frac{B04}{B05}$
Simple Ratio 675/705	SR675/705	$\frac{B04}{B05}$
Simple Ratio 700	SR700	$\frac{1}{B05}$
Simple Ratio 700/670	SR700/670	$\frac{B05}{B04}$
Simple Ratio 710/670	SR710/670	$\frac{B05}{B04}$
Simple Ratio 735/710	SR735/710	$\frac{B06}{B05}$

Simple Ratio 774/677	SR774/677	$\frac{B07}{B04}$
Simple Ratio 800/2170	SR800/2170	$\frac{B08}{B12}$
Simple Ratio 800/470 Pigment specific simple ratio C2	PSSRc2	$\frac{B08}{B02}$
Simple Ratio 800/500 Pigment specific simple ratio C1	PSSRc1	$\frac{B08}{B02}$
Simple Ratio 800/550	SR800/550	$\frac{B08}{B03}$
Simple Ratio 800/650 Pigment specific simple ratio B1	PSSRb1	$\frac{B08}{B04}$
Simple Ratio 800/670 Ratio Vegetation Index	RVI	$\frac{B08}{B04}$
Simple Ratio 800/675 Pigment specific simple ratio A1	PSSRa1	$\frac{B08}{B04}$
Simple Ratio 800/680 Pigment Specific Simple Ratio (Cholophyll a) (PSSRa)	SR800/680	$\frac{B08}{B04}$
Simple Ratio 801/550 NIR/Green	SR801/550	$\frac{B08}{B03}$
Simple Ratio 801/670 NIR/Red	SR801/670	$\frac{B08}{B04}$
Simple Ratio 810/560 Plant biochemical index	PBI	$\frac{B08}{B03}$
Simple Ratio 833/1649 MSIhyper	SR833/1649	$\frac{B08}{B11}$
Simple Ratio 833/658	SR833/658	$\frac{B08}{B04}$
Simple Ratio 850/710 Datt2	Datt2	$\frac{B08}{B05}$
Simple Ratio 860/550	SR860/550	$\frac{B8A}{B02}$
Simple Ratio 860/708	SR860/708	$\frac{B8A}{B05}$

Simple Ratio MIR/NIR Ratio Drought Index	RDI	$\frac{B12}{B08}$
Simple Ratio MIR/Red Eisenhydroxid-Index	SRMIR/Red	$\frac{B12}{B04}$
Simple Ratio NIR/700-715	SRNir/700-715	$\frac{B12}{B05}$
Simple Ratio NIR/G Green Ratio Vegetation Index	GRVI	$\frac{B08}{B03}$
Simple Ratio NIR/MIR	SRNIR/MIR	$\frac{B08}{B12}$
Simple Ratio NIR/RED Difference Vegetation Index, Vegetation Index Number (VIN)	DVI	$\frac{B08}{B04}$
Simple Ratio NIR/Rededge RedEdge Ratio Index 1	RRI1	$\frac{B08}{B05}$
Simple Ratio Red/Blue Iron Oxide	IO	$\frac{B04}{B02}$
Simple Ratio Red/Green Red-Green Ratio	RGR	$\frac{B04}{B03}$
Simple Ratio Red/NIR Ratio Vegetation-Index	SRRed/NIR	$\frac{B04}{B08}$
Simple Ratio SWIRI/NIR Ferrous Minerals	SRSWIRI/NIR	$\frac{0.887}{B08}$
Soil Adjusted Vegetation Index	SAVI	$\frac{(B08 - B04)}{(B08 + B04 + 0.725)} * (1.0 + 0.725)$
Soil and Atmospherically Resistant Vegetation Index	SARVI	$(1 + 0.487) * \frac{B08 - \left(0.740 - 0.735 * (RB - 0.740)\right)}{B08 \pm \left(0.740 - 0.735 * (0.560 - 0.740)\right) + 0.487}$
Soil and Atmospherically Resistant Vegetation Index 2	SARVI2	$2.5 * \frac{(B08 - B04)}{(1.0 + B08 + 6.0 * B04 - 7.5 * B02)}$
Soil and Atmospherically	SAVI3	$(1.0 + 0.5) \cdot \frac{(B08 - B04)}{(B08 + B04 + 0.5)}$

Resistant Vegetation Index 3		
Soil Background Line	SBL	$B09 - 2.4 * B04$
Soil Composition Index		$\frac{(B11 - B08)}{(B11 + B08)}$
Soil-adjusted vegetation index 2	SAVI2	$\frac{B08}{\left(B04 + \frac{0.918}{0.064}\right)}$
Specific Leaf Area Vegetation Index	SLAVI	$\frac{B08}{(B04 + B12)}$
SQRT(IR/R)	SQRT(IR/R)	$\sqrt{\frac{B08}{B04}}$
Structure Intensive Pigment Index 1	SIPI1	$\frac{(B08 - B01)}{(B08 - B04)}$
Structure Intensive Pigment Index 3	SIPI3	$\frac{(B08 - B02)}{(B08 - B04)}$
Tasselled Cap - brightness	SBI	$0.3037 * B02 + 0.2793 * B03 + 0.4743 * B04 + 0.5585 * B08 + 0.5082 * B10 + 0.1863 * B12$
Tasselled Cap - Green Vegetation Index MSS	GVIMSS	$-0.283 * B03 - 0.66 * B04 + 0.577 * B06 + 0.388 * B09$
Tasselled Cap - Non Such Index MSS	NSIMSS	$-0.016 * B03 + 0.131 * B04 - 0.425 * B06 + 0.882 * B09$
Tasselled Cap - Soil Brightness Index MSS	SBIMSS	$0.332 * B03 + 0.603 * B04 + 0.675 * B06 + 0.262 * B09$
Tasselled Cap - vegetation	GVI	$-0.2848 * B02 - 0.2435 * B03 - 0.5436 * B04 + 0.7243 * B08 + 0.084 * B11 - 0.18 * B12$
Tasselled Cap - wetness	WET	$-0.2848 * B02 - 0.2435 * B03 - 0.5436 * B04 + 0.7243 * B08 + 0.084 * B11 - 0.18 * B12$
Tasselled Cap - Yellow Vegetation Index MSS	YVIMSS	$-0.899 * B03 + 0.428 * B04 + 0.076 * B06 - 0.041 * B09$
TCARI/OSAVI	TCARI/OSAVI	$\frac{3.0 * (B05 - B04) - 0.2 * (B05 - B03) * \frac{B05}{B04}}{(1.0 + 0.16) * \frac{(B08 - B04)}{(B08 + B04 + 0.16)}}$

Transformed Chlorophyll Absorbtion Ratio	TCARI	$03.0 * (B05 - B04) - 0.2 * (B05 - B03) * \frac{B05}{B04}$
Transformed NDVI	TNDVI	$\sqrt{\frac{(B08 - B04)}{(B08 + B04) + 0.5}}$
Transformed Soil Adjusted Vegetation Index	TSAVI	$\frac{0.421 * (B08 - B * B04 - 0.824)}{B04 + 0.421 * (B08 - 0.824) + 0.114 * (1.0 + 0.42}$
Transformed Vegetation Index	TVI	$\sqrt{\frac{(B04 - B03)}{(B04 + B03)}} + 0.5$
Triangular chlorophyll index	TCI	$1.2 * (B05 - B03) - 1.5 * (B04 - B03) * \sqrt{\frac{B05}{B04}}$
Vegetation Index 700	VI700	$\frac{B05 - B04}{B05 + B04}$
Visible Atmospherically Resistant Index Green	VARIgreen	$\frac{B03 - B04}{B03 + B04 - B02}$
Visible Atmospherically Resistant Indices 700	VARI700	$\frac{B05 - 1.7 * B04 + 0.7 * B02}{B05 + 2.3 * B04 - 1.3 * B02}$
Visible Atmospherically Resistant Indices RedEdge	VARIrededge	$\frac{B05 - B04}{B05 + B04}$
Weighted Difference Vegetation Index	WDVI	$B08 - 0.752 * B04$
Wide Dynamic Range Vegetation Index	WDRVI	$\frac{0.1 * B08 - B04}{0.1 * B08 + B04}$

2.1.3 Mutual Info

Nella teoria della probabilità l'informazione reciproca, detta anche mutual info, tra due variabili casuali è una misura della dipendenza reciproca che c'è tra di esse. Nello specifico, indica quantitativamente il numero di informazioni che possono essere ottenute da una delle due variabili osservando l'altra [16], [17].

Questo concetto è fortemente legato a quello di entropia una nozione fondamentale nella teoria dell'informazione, che specifica la "quantità di informazione" contenuta in una variabile casuale.

Data una coppia di variabili casuali (X, Y) con valori sullo spazio $X \times Y$. Se $P_{(X,Y)}$ è la loro distribuzione congiunta e P_X, P_Y corrispondono alle distribuzioni marginali, l'informazione reciproca sarà definita come:

$$I(X;Y) = D_{KL}\left(P_{(X,Y)} \parallel P_X \bigotimes P_Y\right)$$

Dove D_{KL} è la divergenza di Kullback-Leibler [16].

Utilizziamo il metodo make_classification, della libreria python scikit-learn, che ci permette di creare un set di dati scegliendo anche il numero delle variabili di tipo informativo, ripetute, ridondanti e casuali. Genereremo il dataset con 6 variabili informative, 1 ridondante, 2 ripetute e 1 casuale.

Fatto ciò, calcoliamo le informazioni reciproche ed il target, utilizzando il metodo mutual_info_classif, anch'esso di scikit-learn, ottenendo i risultati illustrati nella Fig. 2.7.

Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10
0.2202	0.066	0.0068	0.0759	0.0024	0.1067	0.066	0.0024	0	0

Fig. 2.7: valori delle informazioni reciproche calcolate su ogni variabile, [17].

Le variabili 7 e 8 hanno le medesime informazioni reciproche di 2 e 5, dunque sono ripetute. Inoltre, le variabili 9 e 10 hanno un valore di informazione reciproca pari a 0, dunque, risultano entrambe indipendenti dal target. Infine, le variabili rimanenti sono quelle informative e quella ridondante. Si noti che solo calcolando le informazioni reciproche tra le funzionalità ed il target, non possiamo fare distinzioni tra le variabili informative e quelle ridondanti.

2.2 Classificazione

Nel machine learning, il problema di classificazione si riferisce ad una categoria di problemi di apprendimento supervisionato il cui obiettivo è quello di identificare la classe corretta tra due o più opzioni. Pertanto, utilizzando un modello d'apprendimento chiamato "classificatore", viene addestrato su un set di dati costituito da una serie di pattern e le loro relative classi di appartenenza; Se addestrato correttamente, il classificatore sarà in grado di generalizzare su altri pattern dati in input per identificare la classe a cui appartengono.

2.2.1 Random Forest

La crescita di un insieme di alberi ed il loro voto per la classe più popolare costituisce miglioramenti significativi nell'accuratezza della classificazione. Per far crescere questi insiemi, spesso vengono generati vettori casuali che governano la crescita di ciascun albero nell'insieme. Un primo esempio è il bagging , in cui per far crescere ogni albero viene effettuata una selezione casuale dagli esempi nel training set. Un altro esempio è la selezione random split, in cui ad ogni nodo lo split viene selezionato a caso tra i K migliori split. Breiman genera nuovi training set randomizzando gli output nel training set originale. Oppure un altro approccio è quello di selezionare il training set da un insieme casuale di pesi sugli esempi nel training set. L'elemento comune a tutte queste procedure è che per l'albero k-esimo viene generato un vettore casuale Θ_k, indipendente dai vettori casuali passati $\Theta_1, \dots, \Theta_{k-1}$ ma con la stessa distribuzione; un albero viene cresciuto utilizzando il set di addestramento e Θ_k, risultando in un classificatore $h(x, \Theta_k)$ dove x è un vettore di input; chiamiamo queste procedure foreste casuali [24].

> *"Una foresta casuale è un classificatore costituito da una raccolta di classificatori strutturati ad albero* $\{h(x, \Theta_k), k = 1, \dots\}$ *dove i* Θ_k *sono vettori casuali indipendenti distribuiti in modo identico e ogni albero lancia un'unità vota per la classe più popolare all'input x"* [24].

Prima di tutto, è necessario fare una breve introduzione sugli alberi decisionali. Tutto parte da una domanda iniziale, per esempio: "Dovrei navigare?". Da lì in poi verranno poste una serie di domande: "È un lungo periodo di mareggiata?"; "Il vento soffia al largo?"; ecc. Queste domande costituiscono i nodi decisionali dell'albero, mentre i rami contengono le risposte, che possono essere "True" o "False". Dopo aver risposto a ciascuna domanda si giungerà infine ad un nodo foglia, che conterrà la risposta alla domanda iniziale. [19].

Gli algoritmi di foresta casuale hanno tre iperparametri da impostare prima dell'addestramento. Questi includono il numero di alberi, la dimensione del nodo e il numero di feature campionate.

Come detto, questo algoritmo è costituito da una serie di alberi decisionali, ognuno dei quali a sua volta costituito da un campione di dati tratto da un dataset di addestramento con sostituzione, chiamato esempio di bootstrap. Di questo campione di addestramento, un terzo

(ma a seconda delle esigenze può essere più o meno corposo) è messo da parte come dati di test, questa porzione è detta campione out-of-bag.

Dunque, viene iniettata attraverso il bagging delle funzionalità un'altra istanza di casualità, aggiungendo una maggiore varietà all'interno del dataset e riducendo la correlazione tra gli alberi decisionali. A seconda del problema, che sia di regressione o classificazione, la determinazione della previsione varierà. In caso di regressione, i singoli alberi decisionali verranno mediati, mentre, in caso di classificazione come illustrato in Fig. 2.8, verrà scelta la categoria più frequente con un voto di maggioranza. Infine, i dati di test verranno utilizzati per la cross validation, verificando l'accuratezza di tale previsione [19].

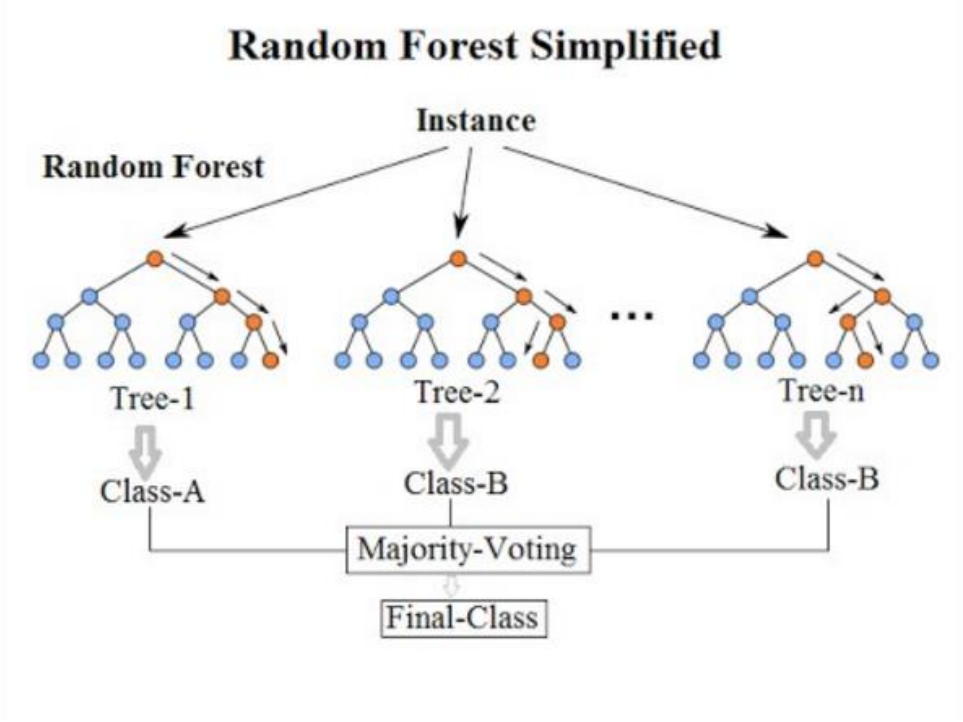

Fig. 2.8: Random Forest in caso di classificazione, [18]

Esistono numerosi vantaggi che l'algoritmo Random Forest presenta, i principali sono:

- overfitting ridotto quando è presente un elevato numero di alberi decisionali;
- flessibilità, poiché può gestire sia l'attività di regressione che quella di classificazione;
- facilità nell'individuazione dell'importanza delle feature.

Tuttavia, questo algoritmo presenta anche delle criticità quando lo si utilizza per problemi di classificazione o regressione, le principali riguardano:

- un ampio dispendio di tempo, quando si lavora con grosse mole di dati e si vuole ottenere un risultato con una certa precisione;
- maggiore richiesta di risorse;
- maggiore complessità rispetto all'ausilio di un unico albero decisionale [19].

2.2.2 XGBoost

XGboost sta per "Extreme Gradient Boosting", è una libreria del gradient boosting distribuita, ottimizzata e progettata per un addestramento efficiente e scalabile dei modelli di apprendimento automatico. Prima di approfondire su di esso è importante introdurre i concetti di "boosting" e "gradient boosting".

Il boosting è una modellazione d'insieme, una tecnica che tenta di costruire un classificatore forte partendo da dei classificatori deboli. Come prima cosa si parte dai dati di addestramento per costruire un primo modello, come un albero decisionale, successivamente viene costruito il secondo modello che cerca di correggere gli errori del primo, questa procedura viene iterata sino a quando il set di dati di addestramento completo non viene predetto correttamente o viene raggiunto il numero massimo di modelli.

Il gradient boosting (o potenziamento del gradiente) è un algoritmo di boosting, dove ogni predittore corregge l'errore del suo precedente, con la differenza che i pesi delle istanze di addestramento non vengono modificati, invece, ogni predittore viene addestrato utilizzando gli errori del predecessore come etichette.

Xgboost è un'implementazione dei gradient boosted decision trees, capace di lavorare sia su problemi di regressione che classificazione. In questo algoritmo, gli alberi decisionali vengono creati in forma sequenziale ed i loro pesi giocano un ruolo fondamentale, essi sono assegnati a ciascuna delle variabili indipendenti che vengono poi inserite nell'albero decisionale per prevede i risultati. Il peso delle variabili erroneamente predette dall'albero viene aumentato, ed esse vengono inviate al secondo albero decisionale [20], [21].

Adesso, poniamo un esempio di CART (classification and regression trees) che classifica se a qualcuno possa piacere un ipotetico gioco X per computer. Come illustrato in Fig. 2.8 prendiamo in considerazione i membri di una famiglia e classifichiamoli in diverse foglie

assegnandogli un punteggio. Solo un albero non è sufficiente per avere una predizione accurata, per cui si utilizza il modello di ensemble, che somma la precisione di più alberi [23].

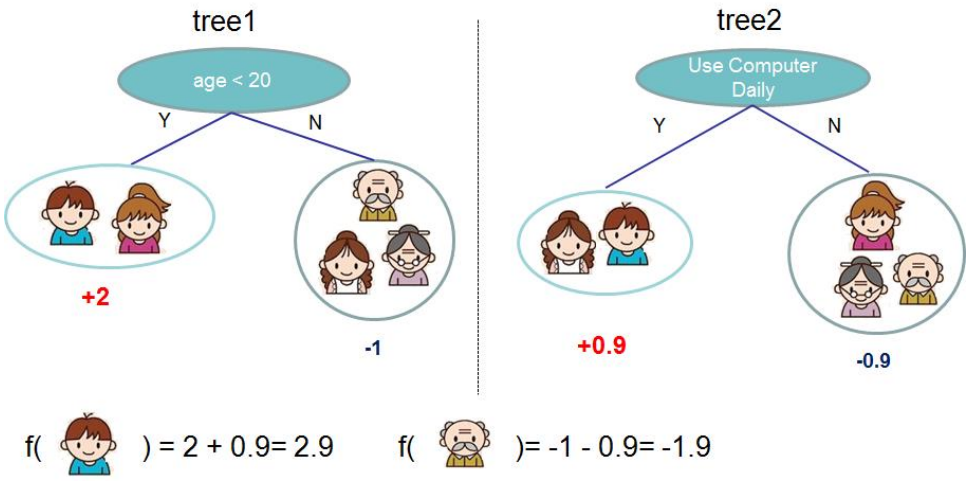

Fig. 2.8 membri della famiglia classificati in più foglie [23].

È importante notare che i due alberi nella Fig. 2.8 cercano di completarsi a vicenda. Matematicamente, possiamo scrivere il nostro modello nella forma:

$$y_i = \sum_{k=1}^{K} f_k(x_i)\,,\ f_k \in \mathcal{F}$$

Dove K è il numero di alberi, f_k nello spazio funzionale $\mathcal{F}$, , ed $\mathcal{F}$ è l'insieme di tutti i possibili CART. La funzione obiettivo da ottimizzare è data da

$$obj(\theta) = \sum_{i}^{n} l\left(y_i\,,\ y_i^{\wedge}\right) + \sum_{k=1}^{K} \omega(f_k)$$

Dove $\omega(f_k)$ è la complessità dell'albero f_k.

Ora, invece di imparare l'albero tutto in una volta, il che rende più difficile l'ottimizzazione, utilizziamo la strategia additiva, riducendo al minimo la perdita di ciò che abbiamo appreso e aggiungendo un nuovo albero che può essere riassunto come

$$y_i^{\Lambda(0)} = 0$$

$$y_i^{\Lambda(1)} = f_1(x_i) = y_i^{\Lambda(0)} + f(x_i)$$

$$y_i^{\Lambda(2)} = f_1(x_i) + f_2(x_i) = y_i^{\Lambda(1)} + f_2(x_i)$$

...

$$y_i^{\Lambda(t)} = \sum_{k=1}^{t} f_k(x_i) = y_i^{\Lambda(t-1)} + f_t(x_i)$$

La funzione obbiettivo del modello sopracitato viene definita come

$$obj^{(t)} = \sum_{i=1}^{n} l\left(y_i\,,\ y_i^{\Lambda(t)}\right) + \sum_{i=1}^{t} \omega(f_i) =$$

$$\sum_{i=1}^{n} l\left(y_i\,,\ y_i^{\Lambda(t-1)} + f_t(x_i)\right) + \omega(f_t) + \text{constant}$$

Se utilizziamo l'errore quadratico medio (MSE) come funzione di perdita, l'obiettivo diventa

$$obj^{(t)} = \sum_{i=1}^{n} \left(y_i - \left(y_i^{\Lambda(t)} + f_t(x_i)\right)\right)^2 + \sum_{i=1}^{t} \omega(f_i) =$$

$$\sum_{i=1}^{n} \left[2\left(y_i^{\Lambda(t-1)} - y_i\right) f_t(x_i) + f_t(x_i)^2\right] + \omega(f_i) + \text{constant}$$

Eseguiamo lo sviluppo di Taylor della funzione di perdita fino al secondo ordine:

$$obj^{(t)} = \sum_{i=1}^{n} \left[l\left(y_i\,,\ y_i^{\Lambda(t-1)}\right) + g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \omega(f_t) + \text{constant}$$

dove g_i e h_i possono essere definiti come:

$$g_i = \partial_{y_i^{\Lambda(t-1)}} l\left(y_i\,,\ y_i^{\Lambda(t-1)}\right)$$

$$h_i = \partial^2_{y_i^{\Lambda(t-1)}} l\left(y_i\,,\ y_i^{\Lambda(t-1)}\right)$$

Dopo aver rimosso tutte le costanti, l'obiettivo specifico al passo t diventa:

$$obj^{(t)} = \sum_{i=1}^{n} \left[g_i \, f_i(x_i) + \frac{1}{2} h_i \, f_t^2(x_i) \right] + \omega(f_t)$$

definiamo il termine di regolazione, ovvero, la complessità dell'albero. Scriviamo il modello come:

$$f_t(x) = \omega_q(x), \; \omega \in R^T, \; q: R^d \to \{1,2,\dots,T\}$$

ω rappresenta il vettore dei punteggi sulle foglie dell'albero, q è la funzione che assegna alla foglia corrispondente ogni punto dati e T è il numero di foglie.

Il termine di regolazione è dunque definito da:

$$\omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{J=1}^{T} \omega_j^2$$

La funzione obiettivo diventa:

$$obj^{(t)} \approx \sum_{i=1}^{n} \left[g_i \omega_{q(x_i)} + \frac{1}{2} h_i \omega_{q(x_i)}^2 \right] + \gamma Y + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 =$$

$$\sum_{j=1}^{T} \left[\left(\sum_{i \in I_j}^{\cdot} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j}^{\cdot} h_i + \lambda \right) \omega_j^2 \right] + \gamma T$$

Semplificandola otteniamo:

$$obj^{(t)} = \sum_{j=1}^{T} \left[G_j \, \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2 \right] + \gamma T$$

dove:

$$G_j = \sum_{i \in I_j} g_i$$

$$H_j = \sum_{i \in I_j} h_i$$

ω_j sono indipendenti l'una dall'altra, la miglior ω_j per una struttura q(x) e la miglior riduzione oggettiva sono:

$$\omega_j^* = -\frac{G}{H_j + \lambda}$$

$$obj^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$

dove, γ è il parametro di sfoltimento, cioè il minimo guadagno di informazioni per eseguire la divisione.

L'ultima equazione misura la bontà di una struttura ad albero q(x). Non potendo ottimizzare direttamente l'albero, cercheremo di ottimizzare un livello dell'albero alla volta, nello specifico, proviamo a dividere una foglia in due ed osserviamone il risultato [23]

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

Nella Fig. 2.9 viene applicata la formula del Gain su ciascuna delle foglie dell'albero di decisione che contengono i membri della famiglia.

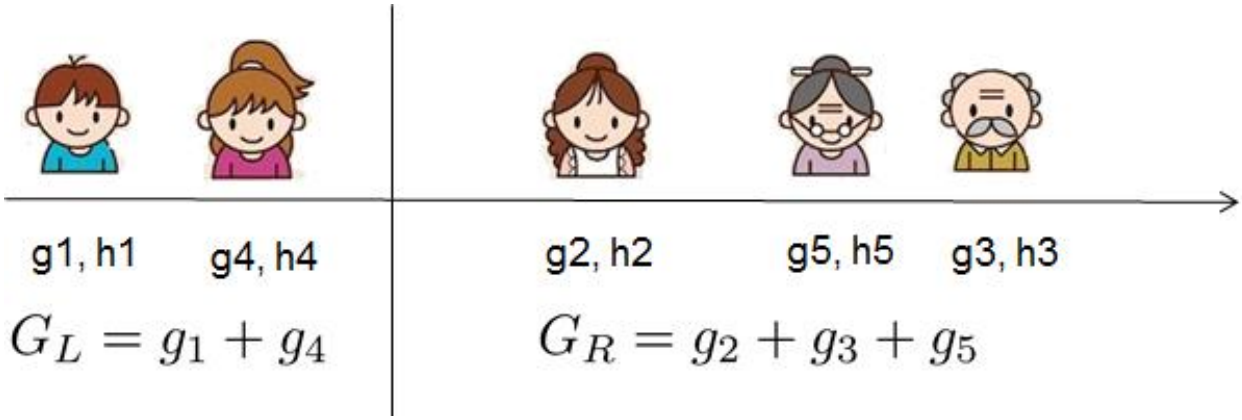

Fig. 2.9: Calcolo del guadagno informativo, [23].

I pregi dell'XGBoost riguardano [21]:

- Prestazioni elevante, derivanti da una solida esperienza nella prod uzione di risultati di alta qualità in varie attività di apprendimento automatico;
- scalabilità;
- personalizzazione, tramite l'ausilio dei vari iperparametri che possono essere regolati per ottimizzarne le prestazioni;
- supporto integrato per la gestione dei valori mancanti;
- le feature più importanti per effettuare le predizioni, rendendolo un modello facilmente interpretabile.

I defetti dell'XGBoost riguardano [21]:

- complessità quando è necessario addestrare modelli di grandi dimensioni;
- overfitting, se addestrato su pochi dati;
- difficoltà nell'impostare i giusti iperparametri per ottenere delle prestazioni ottimali.

Capitolo 3

Approccio proposto

In questo capitolo mostreremo l'approccio proposto per il problema dell'epidemia degli scarabei della corteccia che sta colpendo le zone forestali del nord Italia. Inoltre, sarà presente il diagramma delle classi, seguito da una descrizione delle funzionalità e della composizione di ciascuna classe e metodo utilizzato nel programma.

3.1 Descrizione

Per la creazione del software è stato utilizzato il linguaggio di programmazione Python, un linguaggio ad alto livello, interpretato e orientato agli oggetti. In particolare, abbiamo utilizzato l'IDE "PyCharm", che offre una vasta gamma di strumenti per lo sviluppo.

All'interno del codice i dati sono stati ricavati da un file Excel composto da nove fogli, corrispondenti alle bande spettrali utilizzate. Ogni foglio conteneva l'ID dei pixel ed i rispettivi valori per ogni time stamp. Abbiamo anche utilizzato un file di testo contenente l'ID del pixel e le coordinate X, Y, latitudine e longitudine. Queste informazioni sono state utilizzate per inserire ogni pixel in una matrice specifica per una data banda spettrale e time stamp. Abbiamo utilizzato la libreria NumPy per creare le matrici e la libreria TiffFile per convertirle in file .tif. Infine, i file .tif sono stati suddivisi in directory corrispondenti ai time stamp.

3.2 Diagramma delle classi

l'UML (Unified Modelling Lenguage), è un linguaggio di modellazione visuale con l'obiettivo di essere universalmente comprensibile per chiunque ne faccia uso. L'UML è un linguaggio ricco nella sintassi e semantica, il che lo rende versatile per l'architettura, la progettazione e l'implementazione di sistemi software complessi sia dal punto di vista strutturale che comportamentale [25]. Un tipo di diagramma UML è il "diagramma delle classi" con cui, come

mostrato in Fig.3.1, abbiamo rappresentato le classi, i loro attributi, i metodi e le relazioni all'interno del software.

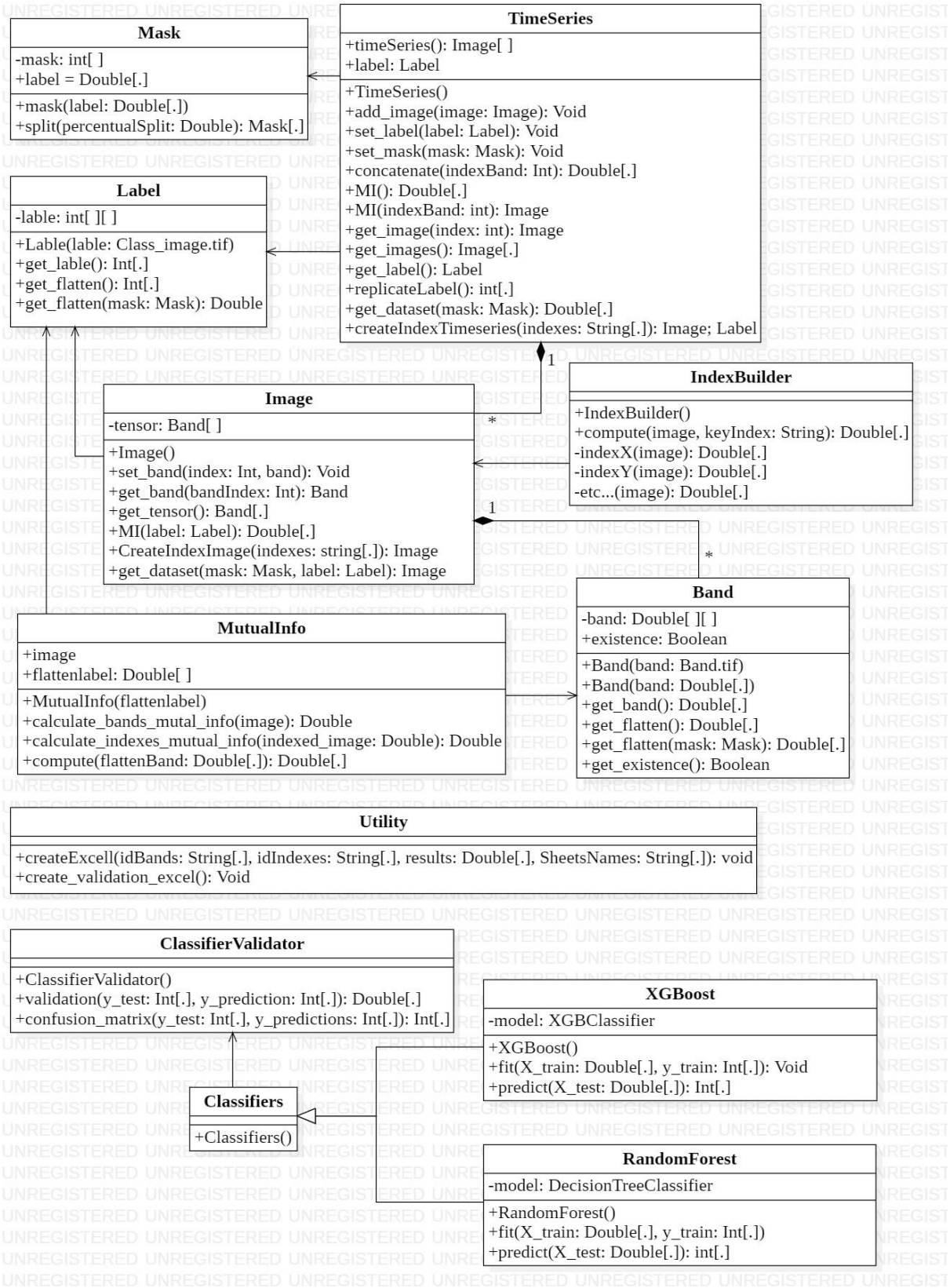

Fig. 3.1: Diagramma delle classi

La classe “**TimeSeries**” contiene le immagini spettrali catturate in diversi time stamp. I suoi attributi includono “timeSeries”, una lista contenente immagini spettrali, “label”, una matrice contenente i valori target e “mask”, una maschera che contiene le posizioni dei pixel in una banda.

I metodi forniti da questa classe sono:

- “TimeSeries( )”: il costruttore della classe.
- “add_image(image: Image): Void”: aggiunge un'immagine nell’attributo “timeSeries”
- “set_label(label: Label): Void”: imposta la label nell'attributo "label".
- “set_mask(mask: Mask): Void”: imposta la maschera nell'attributo "mask".
- “Concatenate(indexBand: Int): Double[ ]”: concatena la banda specificata nell’input con la sua versione in ogni time series.
- “MI(indexBand: Int): Image”: calcola mutual info sull’indice nella posizione specificata dall’ input
- “MI( ): Double[ ]”: restituisce una lista con i valori di mutual indo su ogni time stamp diverso
- “get_label( ): Label”: restituisce l’attributo “label”
- “get_image(index: Int): Image”: restituisce l’immagine nella posizione specificata da “index” all’interno di “timeSeries”.
- “get_images( ): Image[ ]”: restituisce tutte le immagini contenute in “timeSeries”.
- “ReplicateLabel( ): Int[ ]”: concatena la label con sé stessa tante volte quante sono le immagini in “timeSeries”.
- “createIndexTimeSeries(indexes: string[ ]):”: prende in input una lista di indici e restituisce una lista con il valore di ciascun indice calcolato.
- “get_dataset(mask: Mask): Image[ ], Label”: concatena l’output di “get_dataset” della classe “Image” richiamato su ogni immagine in “timeSeries”.

La classe “**Image**” contiene un'immagine spettrale. Il suo unico attributo è “tensor”, una lista di bande che costituiscono l’intera immagine.

I metodi forniti da questa classe sono:

- “Image( )”: il costruttore della classe
- “set_band(band: Band, index: Int): Void”: inserisce una banda spettrale nella posizione specificata da index nel tensore.
- “get_band(index: Int): Band”: restituisce la banda del tensore nella posizione specificata da “index”.
- “get_tensor( ): Band[ ]”: restituisce l’intero tensore.
- “MI(label: Label): Double[ ]”: restituisce i valori di mutal info calcolati su ciascuna banda dell'immagine.
- “CreateIndexImage(indexes: String[ ]): Image”: restituisce una nuova immagine composta dagli indici calcolati in base alla lista di indici fornita in input
- “get_dataset(mask: Mask): Image”: in base alla maschera fornita in input, restituisce un'immagine le cui bande sono state divise in set di addestramento o set di test.

La classe “**Band**” contiene una banda spettrale. I suoi attributi sono “band”, una matrice contenente i pixel delle zone forestali, ed “exist”, un valore booleano impostato su “True” se la banda esiste o su “False” se la banda non esiste.

I metodi forniti da questa classe sono:

- “Band(filePath: string)”: il costruttore della classe che prende in input il path del file .tif contenente la banda e la memorizza nell’attributo “band”.
- “Band(npArray: double[ ][ ])”: il costruttore della classe che prende in input la matrice e la memorizza nell’attributo band.
- “get_band( ): Double[ ][ ]”: restituisce la matrice contenuta in “band”.
- “get_flatten( ): Double[ ]”: restituisce la matrice "band" appiattita.
- “get_flatten(mask): Double[ ]”: restituisce la matrice "band" appiattita e rimuove i valori corrispondenti a -1.
- “get_existence(): Boolean”: restituisce il valore dell’attributo “exist”.

La classe “**Mask**” crea la maschera contenente le posizioni dei pixel all’interno di una matrice. I suoi attributi sono “mask”, una matrice che contiene la maschera, e “label”, una matrice contenente i valori target.

I metodi forniti da questa classe sono:

- “Mask(label: Label)”: il costruttore di classe che prende in input la label.
- “get_mask( ): Mask”: restituisce la matrice contenuta in “mask”.
- “split (percentualSplit: int): Mask[ ]”: divide la maschera in base a "percentualSplit" per ottenere una maschera per l'estrazione dei dati di addestramento e una maschera per l'estrazione dei dati di test da una banda.

La classe “**IndexBuilder**” permette il calcolo degli indici spettrali.

I metodi forniti da questa classe sono:

- “IndexBuilder()”: il costruttore della classe.
- “compute(image: Image, keyIndex: String): Double[ ]”: se presente, calcola l'indice specificato da "keyIndex" sull'immagine fornita in input; altrimenti, restituisce 0.
- “ATASAVI(image: Image): Double[ ][ ]”: calcola e restituisce l’indice spettrale “ATASAVI”.
- gli altri metodi calcolano i restanti indici già menzionati nella tabella del capitolo 2.1.2 sugli indici spettrali.

La classe “**Label**” contiene e modella la matrice contenente i target, Il suo attributo è “label”, costituito da una matrice di interi che contiene i valori target.

I metodi forniti da questa classe sono:

- “Label(filePath: String)”: il costruttore della classe che prende in input il path del file .tif contenente la banda e la memorizza nell’attributo “label”.
- “Label(npArray: Double[ ][ ])”: il costruttore della classe che prende in input la matrice e la memorizza nell’attributo “label”.
- “get_label(): Int[ ]”: restituisce l’attributo “label”.
- “get_flatten(): Int[ ]”: restituisce la matrice "label" appiattita.
- “get_flatten(mask: Mask): Int[ ]”: restituisce la matrice "label" appiattita e rimuove i valori corrispondenti a -1.

La classe “**Utility**” è una classe contenente dei metodi utili per il progetto.

I metodi forniti da questa classe sono:

- “createExcels(idBands: String[ ], idIndexes: String[ ], results: Double[ ], sheetsNames: String[ ])”: crea tre file Excel contenenti, rispettivamente, il rank delle sole bande, il rank degli indici e il rank di bande ed indici
- “create_validation_excel(): Void”: crea un file Excel in cui vengono inseriti i risultati delle metriche.

La classe “**Classifiers**” è una classe madre che rappresenta un classificatore generico, le sue classi figlie sono: “XGBoost” e “RandomForest”.

La classe "**XGBoost**" permette di addestrare ed effettua predizioni con l’algoritmo XGBoost. L’attributo della classe è "model", ed esso rappresenta il classificatore.

I metodi forniti da questa classe sono:

- “XGBoost()”: il costruttore della classe.
- “fit(X_train: Double[ ], y_train: Int[ ]): Void”: addestra il modello sul set di addestramento.
- “predict(X_test: Double[ ]): Int[ ]”: effettua le predizioni sul set di test.

La classe “**RandomForest**” permette di addestrare ed effettua predizioni con l’algoritmo Random Forest. L’attributo della classe è "model", ed esso rappresenta il classificatore.

I metodi forniti da questa classe sono:

- “RandomForest ()”: il costruttore della classe.
- “fit(X_train: Double[ ], y_train: Int[ ]): Void”: addestra il modello sul set di addestramento.
- “predict(X_test: Double[ ]): Int[ ]”: effettua le predizioni sul set di test.

La classe "**ClassifierValidator**" permette di calcolare le prestazioni di un modello.

I metodi forniti da questa classe sono:

- “ClassifierValidator()”: il costruttore della classe.
- “validation(y_test: Int[ ], y_prediction: Int[ ]): Double[ ]”: restituisce una lista contenenti i risultati delle metriche di precision, recall, f1_score e support per ogni possibile target del classificatore, insieme a una media ponderata e una media non ponderata; infine, calcola l'accuratezza complessiva del modello.
- “confusion_matrix(y_test: Int[ ], y_predictions: Int[ ]) Int[ ][ ] : restituisce la matrice di confusione del modello.

Capitolo 4

Validazione empirica

In quest’ultimo capitolo, analizzeremo i dati che sono stati elaborati nel progetto, mostreremo le metriche che sono state selezionate per valutare le prestazioni dei due classificatori, ed infine, parleremo delle diverse configurazioni per l'addestramento dei modelli.

4.1 Descrizione dei dati

I dati utilizzati nel progetto consistono in immagini multispettrali catturate dai sensori a bordo dei satelliti Sentinel-2. Queste immagini sono composte da diverse bande spettrali, ciascuna corrispondente a una specifica gamma di lunghezze d'onda dello spettro elettromagnetico.

Le bande spettrali utilizzate nel nostro progetto sono le seguenti: B1, B2, B3, B4, B5, B6, B7, B8, B8A, B9, B10, B11, B12. Ognuna di queste bande contiene una matrice di pixel, in cui ogni pixel rappresenta un punto nell'immagine e contiene i valori associati ai canali spettrali corrispondenti. Ad esempio, nel caso di un'immagine a colori, potrebbero essere presenti valori RGB per ciascun pixel.

Inoltre, i dati comprendono diversi time stamp, che rappresentano gli stessi luoghi e le stesse bande spettrali catturati in momenti temporali differenti. Nel nostro progetto, abbiamo utilizzato quattordici time stamp identificati come: 101, 111, 126, 151, 181, 221, 231, 241, 256, 261, 271, 286, 291, 321.

Per gestire e lavorare con questi dati, abbiamo scelto di memorizzare ciascuna banda spettrale di ogni time stamp in file TIFF. Il formato TIFF è stato appositamente progettato per la conservazione di immagini e permette di mantenere intatta la qualità e le informazioni spettrali delle immagini. Questi file TIFF sono stati successivamente caricati nel nostro software per l'elaborazione e l'analisi dei dati.

4.2 Metriche

Al fine di valutare le prestazioni dei modelli: Random Forest e XGBoost abbiamo calcolato la matrice di confusione. Per definizione, una matrice di confusione C è tale che $C_{i,j}$ rappresenta il numero di osservazioni note per essere nel gruppo i e previste per essere nel gruppo j. Ad esempio, nella classificazione binaria, il conteggio dei veri negativi è rappresentato da $C_{0,0}$, i falsi negativi da $C_{1,0}$, i veri positivi da $C_{1,1}$ e i falsi positivi da $C_{0,1}$, 1 [26].

Un metodo supplementare per la valutazione dei modelli è stato il calcolo di alcune metriche [27]:

- "**Precision**" rappresenta la frazione di istanze recuperate che sono rilevanti. È definita come il rapporto tra il numero di veri positivi e il numero di falsi positivi.

$$\frac{TP}{(TP + FP)TP \cdot FP}$$

- "**Recall**" rappresenta la frazione di istanze rilevanti che sono state recuperate. È definito come il rapporto tra i veri positivi e la somma dei veri positivi e dei falsi negativi.

$$\frac{TP}{(TP + FN)TP \cdot FN}$$

- "**F1-score**" è una media armonica ponderata della precisione e del richiamo, dove un punteggio f1 raggiunge il suo miglior valore a 1 e il peggiore a 0.

$$2 \cdot \frac{\text{Pr}\, e\, cision \cdot Recall}{\text{Pr}\, e\, cision + Recall}$$

- "**Support**" è il numero di istanze effettive per ogni classe.

- "**Accuracy**" è una metrica che generalmente descrive le prestazioni del modello su tutte le classi. È calcolata come il rapporto tra il numero di previsioni corrette e il numero totale di previsioni.
  Nella classificazione binaria, può essere definita come:

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Per il calcolo delle metriche, i dati a disposizione sono stati suddivisi nel 70% come trainig set e il restante 30% in test set. È stato proprio il test set che abbiamo utilizzato per calcolare le metriche menzionate precedentemente per le quattro diverse classi. Inoltre, abbiamo ricavato l'accuracy complessiva del modello e, nuovamente, le metriche recall e F1-score per la media pesata (avarage weigth), e la metrica F1-score per la macro-media.

4.3 Configurazioni

Come menzionato in precedenza, per ogni banda e indice spettrale sono stati calcolati i valori di mutual info. Questa misura ci ha fornito un'indicazione di quanto i dati contenuti in ciascuna banda o indice siano informativi per l'addestramento un classificatore. I valori di mutual info sono stati successivamente inseriti in ordine decrescente in tre file Excel distinti: uno contenente solo le bande spettrali, uno contenente solo gli indici spettrali e uno contenente sia le bande che gli indici spettrali.

Questi file con i valori di mutual info sono stati utilizzati per addestrare i modelli e effettuare predizioni su diverse configurazioni di dati:

a. Solo bande nell'ultimo time stamp;
b. Solo bande in tutti i time stamp;
c. Solo indici nell'ultimo time stamp;
d. Solo indici in tutti i time stamp;
e. Bande e indici nell'ultimo time stamp;
f. Bande e indici in tutti i time stamp;
g. Bande e i migliori dieci indici su tutti i time stamp;
h. Bande e i migliori venti indici su tutti i time stamp;
i. Le migliori venti bande e indici sull'ultimo time stamp;
j. Le migliori venti bande e indici su ogni time stamp.

A seconda delle configurazioni scelte, sono stati selezionati gli attributi (bande o indici) più informativi per l'addestramento dei modelli. Questo approccio ci ha consentito di valutare

l'efficacia delle bande spettrali e degli indici spettrali nelle predizioni e di determinare quali combinazioni di attributi fornivano i migliori risultati.

4.4 Risultati

Sia Random Forest, che XGBoost, sono stati addestrati utilizzando dieci diverse configurazioni, con cui hanno poi effettuato le predizioni.

	precision class 1	recall class 1	f1-score class 1	support class1	precision class2	recall class2	f1-score class2	support class2	precision class3	recall class3	f1-score class3	support class3	precision class4	recall class4	f1-score class4	support class4	accuracy	macro av. F1-score	weigthed avg F1-score	weigthed avg recall
RF -a	0.90	0.93	0.91	46	0.92	0.94	0.93	47	0.75	0.60	0.67	15	0.85	0.85	0.85	33	0.88	0.84	0.88	0.88
RF -b	**0.96**	0.96	0.96	46	**0.96**	**1**	**0.98**	47	0.87	**0.87**	0.87	15	**0.97**	0.91	0.94	33	0.95	0.93	0.95	0.95
RF -c	0.86	0.93	0.90	46	0.88	0.94	0.91	47	0.73	0.53	0.62	15	0.80	0.73	0.76	33	0.84	0.80	0.84	0.84
RF -d	0.90	0.93	0.91	46	**0.96**	**1**	**0.98**	47	**1**	0.80	0.89	15	0.88	0.85	0.86	33	0.92	0.91	0.92	0.92
RF -e	0.90	0.93	0.91	46	0.90	0.96	0.93	47	0.70	0.47	0.56	15	0.79	0.79	0.79	33	0.86	0.80	0.85	0.86
RF -f	0.94	**0.98**	0.96	46	**0.96**	**1**	**0.98**	47	**1**	**0.87**	**0.93**	15	**0.97**	0.91	0.94	33	**0.96**	0.95	**0.96**	**0.96**
RF -g	0.94	0.96	0.95	46	**0.96**	**1**	**0.98**	47	**1**	**0.87**	**0.93**	15	0.94	0.91	0.92	33	0.95	0.94	0.95	0.95
RF -h	**0.96**	**0.98**	**0.97**	46	**0.96**	**1**	**0.98**	47	**1**	**0.87**	**0.93**	15	**0.97**	**0.94**	**0.95**	33	**0.96**	**0.96**	**0.96**	**0.96**
RF -i	0.84	0.93	0.89	46	0.90	0.94	0.92	47	0.60	0.60	0.60	15	0.85	0.67	0.75	33	0.84	0.79	0.83	0.84
RF -j	0.92	0.96	0.94	46	**0.96**	**1**	**0.98**	47	**1**	**0.87**	**0.93**	15	0.94	0.88	0.91	33	0.94	0.94	0.94	0.94

	precision class 1	recall class 1	f1-score class 1	support class 1	precision class2	recall class2	f1-score class2	support class2	precision class3	recall class3	f1-score class3	support class3	precision class4	recall class4	f1-score class4	support class4	accuracy	macro av. F1-score	weighted avg F1-score	weigthed avg recall
XGB-a	0.88	0.96	0.92	46	0.92	0.94	0.93	47	0.73	0.53	0.62	15	0.84	0.82	0.83	33	0.87	0.82	0.87	0.87
XGB-b	0.90	**1**	0.95	46	0.94	**1**	0.97	47	0.92	0.80	0.86	15	1	0.82	0.90	33	0.94	0.92	0.93	0.94
XGB-c	0.88	0.93	0.91	46	0.90	0.94	0.92	47	0.67	0.53	0.59	15	0.81	0.76	0.78	33	0.85	0.80	0.85	0.85
XGB-d	0.93	0.93	0.93	46	**0.98**	**1**	**0.99**	47	**0.94**	1	**0.97**	15	0.94	0.88	0.91	33	**0.95**	**0.95**	**0.95**	**0.95**
XGB-e	0.86	0.93	0.90	46	0.90	0.94	0.92	47	0.67	0.40	0.50	15	0.76	0.76	0.76	33	0.84	0.77	0.83	0.84
XGB-f	0.93	0.93	0.93	46	0.94	**1**	0.97	47	1	**0.87**	0.93	15	0.94	0.91	0.92	33	0.94	0.94	0.94	0.94
XGB-g	**0.98**	0.93	**0.96**	46	0.94	**1**	0.97	47	0.92	0.80	0.86	15	0.91	**0.94**	0.93	33	0.94	0.93	0.94	0.94
XGB-h	0.96	0.96	**0.96**	46	0.94	**1**	0.97	47	1	0.80	0.89	15	0.94	**0.94**	**0.94**	33	**0.95**	0.94	**0.95**	**0.95**
XGB-i	0.86	0.93	0.90	46	0.91	0.91	0.91	47	0.56	0.60	0.58	15	0.82	0.70	0.75	33	0.84	0.79	0.84	0.84
XGB-j	0.96	0.93	0.95	46	0.96	**1**	0.98	47	0.93	**0.87**	0.90	15	0.91	0.91	0.91	33	0.94	0.93	0.94	0.94

In linea generale i risultati di entrambi i classificatori presentano predizioni con una media di circa il 90% di accuratezza. Non ci sono differenze evidenti tra i due classificatori, tuttavia, a parità di configurazione, l'XGBoost risulta leggermente meno preciso nelle predizioni dell'1-2% rispetto la Random Forest, ad eccezione di alcuni casi di parità e della configurazione –d il quale supera l'accuratezza della sua controparte nella Random Forest.

Tuttavia, analizzando le diverse configurazioni, si osservano dei maggiori cambiamenti nei risultati. Nel Random Forest, l'accuratezza più bassa si ottiene con le configurazioni -c e -j, con un valore di 0.84, mentre il risultato migliore si ottiene con le configurazioni -h e -f, con un'accuratezza di 0.96. Nell'XGBoost, l'accuratezza più bassa si ottiene con le configurazioni -c e -i, con un valore di 0.84, mentre il risultato migliore si ottiene con le configurazioni -d e -h, che presentano un'accuratezza di 0.95. Valutando questi dati, la metrica più adatta per addestrare entrambi i classificatori risulta essere -h.

Prima di determinare quale sia il modello migliore, è importante considerare non solo l'accuratezza, ma anche la velocità con cui effettuano le predizioni. Per valutare ciò, è stata

utilizzata la libreria Python "time" per calcolare il tempo impiegato da ciascun modello per essere addestrato e prevedere i risultati per tutte le dieci configurazioni.

Per ottenere una misurazione più accurata, il tempo è stato cronometrato dieci volte per ciascun modello, e successivamente è stata calcolata la media. In termini di efficienza, l'XGBoost risulta essere il più veloce, impiegando in media 3.21381 secondi, rispetto alla Random Forest, che impiega in media 3.39396 secondi.

In conclusione, entrambi i classificatori presentano vantaggi specifici rispetto all'altro. La scelta tra i due dipende dall'urgenza con cui bisogna affrontare il problema: se un'accuratezza più elevata fa la differenza, allora si può optare per l'algoritmo della Random Forest; viceversa, se è necessario agire in tempi più brevi a discapito di una maggiore accuratezza, si può scegliere l'algoritmo XGBoost. Nel nostro caso, considerando che il dataset non è di grandi dimensioni, la differenza di tempo tra i due classificatori è trascurabile, e quindi la Random Forest, offrendo risultati più accurati, si rivela la scelta migliore.

Capitolo 5

Conclusioni

Il problema dello scarabeo della corteccia rappresenta una grave minaccia per le foreste in Italia e in altre parti del mondo, particolarmente accentuata dopo la tempesta Vaia del 2018. Il cambiamento climatico sta contribuendo ulteriormente a diffondere questa infestazione e a causare danni significativi agli ecosistemi delle aree interessate. L'obiettivo di questa tesi è stato quello di sviluppare un sistema di intelligenza artificiale basato sull'analisi di immagini multispettrali catturate dai satelliti Sentinel-2 nel programma spaziale Copernicus. Lo scopo principale è stato quello di identificare e prevedere se le aree forestali sono infestate dal bostrico e, se sì, determinare il grado di infezione.

Per raggiungere questo obiettivo, sono stati utilizzati due classificatori: Random Forest e XGBoost. Entrambi gli algoritmi si basano sulla creazione di alberi decisionali multipli per ottenere previsioni accurate. A partire dalle bande spettrali, sono stati calcolati gli indici spettrali per ciascuna immagine e time stamp. Successivamente, il dataset è stato normalizzato e suddiviso in un training set (70%) e un test set (30%). Prima di eseguire il processo di addestramento sul training set, abbiamo calcolato i valori di mutual info su ciascuna banda spettrale e indice spettrale in ogni time stamp. Questa analisi ci ha permesso di identificare quali di questi attributi sono più informativi per addestrare modelli.

Dopo aver addestrato i modelli Random Forest e XGBoost, abbiamo utilizzato il test set per effettuare le predizioni e valutare le prestazioni dei modelli. Una delle metriche utilizzate per valutare le prestazioni è stata la matrice di confusione, che ha evidenziato le classi che i modelli hanno avuto maggiori difficoltà a rilevare e quelle che sono state individuate con maggiore facilità. Altre metriche che sono state utilizzate in questo progetto includono la precision, il recall, l'F1-score e il support per ciascuna delle quattro classi. Inoltre, è stato calcolato l'F1-score per la macro-media e nuovamente la precision e l'F1-score per la media pesata.

I risultati ottenuti dalle nostre analisi sono stati promettenti. Entrambi i modelli di classificazione, Random Forest e XGBoost, hanno dimostrato di essere efficaci nel discriminare le caratteristiche del terreno utilizzando i dati multispettrali. In media, abbiamo ottenuto un'accuratezza di circa il 90%, il che indica una buona capacità predittiva dei modelli.

La valutazione delle diverse configurazioni dei dati ha mostrato che entrambi i modelli hanno ottenuto prestazioni simili in generale. Tuttavia, sono state osservate alcune differenze nelle performance quando si sono confrontate le configurazioni dei modelli. Nel caso di Random Forest, l'accuratezza più bassa è stata registrata con un valore del 0,84, mentre il risultato migliore è stato raggiunto con un valore del 0,96, rappresentando una differenza prestazionale

del 13%. Per quanto riguarda XGBoost, l'accuratezza più bassa è stata del 0,84, mentre il risultato migliore è stato del 0,95, rappresentando una differenza prestazionale dell'11%. Sulla base di queste valutazioni, è stato possibile affermare che la configurazione più performante in entrambi i modelli è stata -h.

Come ultima analisi, ci siamo soffermati sulla velocità dei due modelli, registrando il tempo medio impiegato da ciascuno di essi per essere addestrato e per effettuare le predizioni per ciascuna delle dieci configurazioni. Dall'analisi è emerso che l'XGBoost si è dimostrato leggermente più rapido della Random Forest, con un tempo medio di 0,18015 secondi. Tuttavia, nel contesto del nostro progetto, la differenza di tempo tra i due modelli si è rivelata irrilevante, poiché pochi decimi di secondo non avrebbero un impatto significativo su un ipotetico intervento di disinfestazione delle foreste. Pertanto, possiamo affermare che, grazie alla sua maggiore accuratezza nelle previsioni, la Random Forest risulta essere la migliore alternativa.

In chiusura, vorremmo illustrare alcuni spunti per eventuali sviluppi futuri del nostro progetto. I classificatori utilizzati sono stati addestrati su immagini multispettrali in diversi time stamp che raffigurano la stessa area contenente zone forestali. Tuttavia, utilizzando una quantità maggiore di dati relativi a una o più aree con una maggiore estensione di zone forestali, potremmo probabilmente ottenere prestazioni migliori dai due classificatori una volta addestrati. Un'altra opzione potrebbe essere quella di valutare l'addestramento di altri algoritmi di classificazione, come le Reti Neurali Convoluzionali (CNN), il Support Vector Machines (SVM) o il K-Nearest Neighbors (KNN). Infine, un'ulteriore idea, da combinare con un dataset più ampio e l'aggiunta di nuovi classificatori, potrebbe essere l'implementazione di ulteriori configurazioni, analizzando come influiscono sulle prestazioni dei modelli al fine di ottenere il massimo rendimento dai dati a nostra disposizione.

Bibliografia

[1] Regionieambiente, Scienziati allarmati dal boom di scarabeo della corteccia di abete rosso, 3 settembre 2019, <Scienziati allarmati dal boom di scarabeo della corteccia di abete rosso – Regionieambiente.it >.

[2] Wikipedia, Tempesta Vaia, 2018, <Tempesta Vaia - Wikipedia >.

[3] Provincia, IL BOSTRICO NELLE FORESTE DI ABETE ROSSO DEL TRENTINO, <Bostrico / Foreste in Trentino / Foreste / Homepage - Servizio Foreste e Servizio Faunistico (provincia.tn.it) >.

[4] Somalvico, M. Intelligenza artificiale: concetti, tecniche e applicazioni. FrancoAngeli. 2019.

[5] Amigoni, F., Schiaffonati, V., & Somalvico, M. Enciclopedia della Scienza e della Tecnica. Milano: Gruppo Editoriale L'Espresso, 2008.

[6] Hewlett Packard Enterprise, Machine Learning, <Cos'è il machine learning? | Glossario | HPE Italia >.

[7] USGS, What is remote sensing and what is it used for?, <What is remote sensing and what is it used for? | U.S. Geological Survey (usgs.gov) >.

[8] Forest@, Rivista di Selvicoltura ed Ecologia Forestale, Volume 18, Pagine 27-34, 2021, <Francini S, D'Amico G, Mencucci M, Seri G, Gravano E, Chirici G (2021). Telerilevamento e procedure automatiche: validi strumenti di supporto al monitoraggio delle utilizzazioni forestali. Forest@ 18: 27-34. (sisef.org) >.

[9] Wikipedia, Remote sensing, 24 Aprile 2023, <Remote sensing - Wikipedia>.

[10] Wikipedia, Multispectral imaging, 29 gennaio 2023, <Imaging multispettrale - Wikipedia >.

[11] Wikipedia, Sentinel-2, 8 Aprile 2023, <Sentinel-2 - Wikipedia >.

[12] Esa, Spatial Resolution, <Spatial - Risoluzioni - Sentinel-2 MSI - Guide per l'utente - Sentinel Online - Sentinel Online (copernicus.eu) >.

[13] Missione scienza, Gli Indici di Vegetazione: indicatori della salute vegetale, 24 Agosto 2020, <Gli Indici di Vegetazione: indicatori della salute vegetale - Missione Scienza >.

[14] Agricolus, Indici di vegetazione NDVI e NDMI: istruzioni per l'uso, 29 Maggio 2018, <Indici di vegetazione NDVI e NDMI: istruzioni per l'uso | Agricolus >.

[15] Idee green, Indici di vegetazione da dati RADAR e indici spettrali della vegetazione, <Indici di vegetazione da dati RADAR e indici spettrali della vegetazione - Idee Green >.

[16] Wikipedia, Mutual information, 3 Marzo 2023, <Mutual information - Wikipedia >.

[17] QuantDare, What is Mutual Information?, Pablo Aznar, 31 Marzo 2021, <What is Mutual Information? | Quantdare >.

[18] Wikimedia Commons, File: Random forest diagram complete.png, 24 Marzo 2017, <File:Random forest diagram complete.png - Wikimedia Commons >.

[19] IBM, What is random forest?, <What is Random Forest? | IBM >.

[20] XGBOOST, Introduction to Boosted Trees, 2022, <Introduzione agli alberi potenziati — xgboost 1.7.5 documentazione >.

[21] GeeksforGeeks, XGBoost, 6 Febbraio 2023, <XGBoost - GeeksforGeeks >.

[22] IndexDataBase, Show Indices for selected Sensor, <IDB - Show Indices for selected Sensor (indexdatabase.de)>.

[23] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System, 3 agosto 2016.

[24] Leo Breiman, Random Forests, Ottobre 2001.

[25] Lucidchart, Cos'è l'Unified Modeling Language, <Cos'è l'Unified Modeling Language (UML) | Lucidchart >.

[26] Scikit-learn, sklearn.metrics.confusion_matrix, <sklearn.metrics.confusion_matrix — Documentazione di Scikit-Learn 1.2.2 >

[27] Scikit-learn, sklearn.metrics.precision_recall_fscore_support, <sklearn.metrics.precision_recall_fscore_support — Documentazione di scikit-learn 1.2.2 >

Ringraziamenti

Desidero dedicare questo spazio finale della mia tesi di laurea ai ringraziamenti verso tutti coloro che mi hanno sostenuto durante questo percorso. In particolare, desidero ringraziare i miei relatori, la prof.ssa Appice e la Dott.sa Andresini, per avermi guidato con pazienza lungo la stesura di questa tesi.

Non posso esimermi dal ringraziare la mia famiglia, che durante questo percorso mi ha sempre dato tutto ciò di cui ho avuto bisogno e per essersi dimostrata orgogliosa dei miei risultati. Desidero anche ringraziare i miei colleghi di università, gli unici in grado di comprendere realmente quanto impegno e dedizione siano necessari per affrontare questo percorso, a cui auguro il meglio e li incito a perseguire i propri obiettivi.

Porgo un ringraziamento speciale alla mia ragazza Giorgia che, sin dal primo giorno di università, mi ha continuamente supportato e incoraggiato a dare il massimo, tirandomi su di morale nei momenti difficili e aiutandomi a riflettere su quali fossero le scelte migliori da prendere.

In conclusione, desidero anche ringraziare me stesso, per aver inseguito con determinazione i miei obiettivi, per aver affrontato le mie paure e per non essermi mai arreso, migliorando giorno dopo giorno nonostante chi non avrebbe mai scommesso nulla sul mio futuro. Sono orgoglioso dei risultati ottenuti e consapevole del fatto che questo è solo un nuovo inizio.