



UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA
CORSO DI LAUREA MAGISTRALE IN DATA SCIENCE

**PREVISIONE DEI FENOMENI FRANOSI INDOTTI
DALLE PIOGGE**

Studenti:

Anna Tomeo

Luca Ruberto

Maria Raggio

Professore:

Giuseppe Polese

ANNO 2019-2020

INDICE

1	Introduzione	3
2	Stato dell'Arte	5
3	Raccolta dei dati	7
4	Pre-elaborazione dei dati	12
5	Estrazione delle caratteristiche	14
6	Machine learning	31
6.1	Introduzione	31
6.2	Definizione del problema	32
6.3	task di regressione	32
6.4	Preparazione del set di addestramento e del set di test	36
6.5	Definizione dei modelli	39
6.6	Foresta Casuale	40
6.7	Macchine a vettori di supporto con kernel gaussiano	41
7	Risultati	44
8	Conclusioni	49
9	Sviluppi futuri	50
	Riferimenti bibliografici	51

1

INTRODUZIONE

In Italia, le frane sono dei fenomeni molto frequenti che causano vittime e dispersi, disagi alle persone, alle abitazioni e alle strade, nonchè danni economici rilevanti. Prevedere l'occorrenza dei fenomeni franosi indotti dalle piogge è di interesse sia scientifico che sociale. In generale, non esiste una corrispondenza biunivoca tra fenomeni franosi e cause d'innesto, in quanto l'evoluzione di una frana può essere condizionata da differenti fattori. In alcuni casi, tuttavia, può essere individuata una connessione diretta e traducibile in termini quantitativi tra fenomeno franoso e cause d'innesto, come avviene ad esempio con le piogge. La realizzazione di questo progetto è stata possibile grazie alla collaborazione con il dipartimento di ingegneria civile dell'università degli studi di Salerno, che ha provveduto a fornire i datasets su cui effettuare le opportune analisi. L'obiettivo principale è stato quello di individuare soglie pluviometriche di innesto delle frane, considerando le precipitazioni antecedenti al giorno in cui si è verificato il fenomeno franoso. L'analisi è stata effettuata prendendo in esame due regioni: l'Abruzzo e la Campania. La previsione si basa sulla suddivisione del territorio regionale in varie zone di allerta relative alle piogge e alle frane che si sono verificate dal 2010 al 2017. Si è deciso di prendere in considerazione per l'analisi due attributi principali: per la Campania è stata considerata la distribuzione della pioggia massima in funzione della pioggia cumulata avutasi il giorno prima del fenomeno franoso, per l'Abruzzo è stata considerata la stessa distribuzione di piogge verificatasi due giorni prima il fenomeno franoso. Tali coppie sono state scelte dopo aver eseguito un'opportuna analisi e visualizzazione dei dati con l'ausilio di istogrammi e grafici di dispersione delle cumulate di precipitazione rispetto al giorno di accadimento della frana. Per avere una visione globale della situazione sono stati presi in considerazione anche gli eventi non innescanti. Completata l'analisi dei dati a disposizione, si è deciso di affrontare il problema adottando algoritmi di classificazione di machine learning; è stata esclusa la regressione a causa dell'incertezza dei dati. Nei capitoli a seguire, verranno descritte nel dettaglio le due fasi principali del lavoro svolto: l'analisi dei dati in cui ci si soffermerà sulla descrizione del dataset adottato, sulle varie relazioni individuate tra i vari attributi e soprattutto sull'andamento dei dati. Nella seconda fase verranno descritte principalmente

le prestazioni ottenute con i vari classificatori e infine saranno illustrati i risultati ottenuti.

2

STATO DELL'ARTE

La ricerca scientifica si occupa da diversi decenni del ruolo assunto dalle piogge nell'innesto dei movimenti franosi. Il valore dei danni arrecati dalle frane innescate dalle piogge è spesso incalcolabile e raramente documentato. Nelle aree dove non rappresentano una minaccia per la vita, esse arrecano comunque danni alle infrastrutture e alle attività produttive. La riduzione dei rischi connessi ai movimenti franosi dovuti alle piogge si basa sulla capacità di prevedere con congruo anticipo l'accadimento di tali fenomeni, in modo da approntare idonee misure di emergenza o di messa in sicurezza. Nel corso degli anni, in letteratura sono state definite numerose metodologie di studio con l'ausilio di varie tecnologie. Tra le varie metodologie di studio utilizzate per prevedere le frane rientrano le seguenti:

- *Metodo empirico-pluviometrico* che si suddivide in:
 1. *Metodo empirico-pluviometrico di 1 ordine*[1]: considera come attributo predittivo solo la variabile pioggia, presenta una complessità bassa, un numero di parametri basso. Tale metodo non si avvale di un'analisi approfondita di tutti gli elementi di natura morfologica. Consente, quindi, di fissare delle soglie che nella letteratura scientifica sono definite di 1 ordine. Il metodo principalmente associa una probabilità che un fenomeno franoso si inneschi all'istante t ad una funzione $Y(t)$ dipendente dalle precipitazioni che hanno preceduto l'istante t .
 2. *Metodo empirico-pluviometrico di 2 ordine*[1]: considera come attributi predittivi la pioggia ed altre variabili del contesto fisico, presenta una complessità media, un numero di parametri medio

Il metodo empirico-pluviometrico basato su soglie di 1 e 2 ordine fornisce buoni risultati se applicato a quelle tipologie di frana il cui innesto è direttamente correlato alle precipitazioni

- *Metodo empirico-idrologico*[1]: tale metodo si limita ad individuare la relazione che intercorre tra le piogge e i movimenti franosi considerando la quantità di acqua infiltratisi nel sottosuolo prima dell'innesto. In breve, il metodo identifica una funzione $Y(t)$

che dipende dalle precipitazioni antecedenti e tiene conto anche delle caratteristiche del corpo franoso. Identificata tale funzione, è possibile identificare valori critici al cui superamento è associata la maggiore o minore probabilità del movimento.

- *Metodo statistico*[1]: tale metodo si propone di prevedere il fenomeno franoso non dal punto di vista fisico ma individuando le relazioni esistenti tra caratteristiche del territorio e la frana.

Oltre ai metodi citati, in letteratura vi sono anche metodi che si basano sulle piogge cumulate e antecedenti, metodi che si basano sulle piogge cumulate per lunghi periodi e anche metodi che si basano sull'analisi della soglia durata-intensità della pioggia. La difficoltà nell'applicare le varie metodologie esistenti consiste dell'avere a disposizione una grande mole di dati da cui poter estrapolare informazioni utili e creare modelli per risolvere la problematica.

3

RACCOLTA DEI DATI

L'analisi effettuata ha interessato due regioni: Campania e Abruzzo. Di seguito, verranno descritti i dati grezzi e come sono stati elaborati per le suddette regioni. I principali dati a disposizione per il problema in esame sono:

- variabili pluviometriche, ossia le piogge;
- il numero di fenomeni franosi;
- ZAM (zone allerta meteo) per entrambe le regioni.

Per la Campania si hanno 8 ZAM, per l'Abruzzo 6 ZAM.

I dati raccolti sono stati distribuiti in due datasets: uno contenente i dati relativi ai fenomeni franosi e l'altro relativo alle variabili pluviometriche. I due datasets sono stati strutturati nel formato excel. Il dataset inerente ai fenomeni franosi è strutturato, per entrambe le regioni, nel seguente modo:

- ogni record fa riferimento ad un giorno e contiene il numero totale di frane avvenute in ciascuna ZAM (zona allerta meteo);
- periodo di analisi: 2010-2017, ovvero 2922 giorni

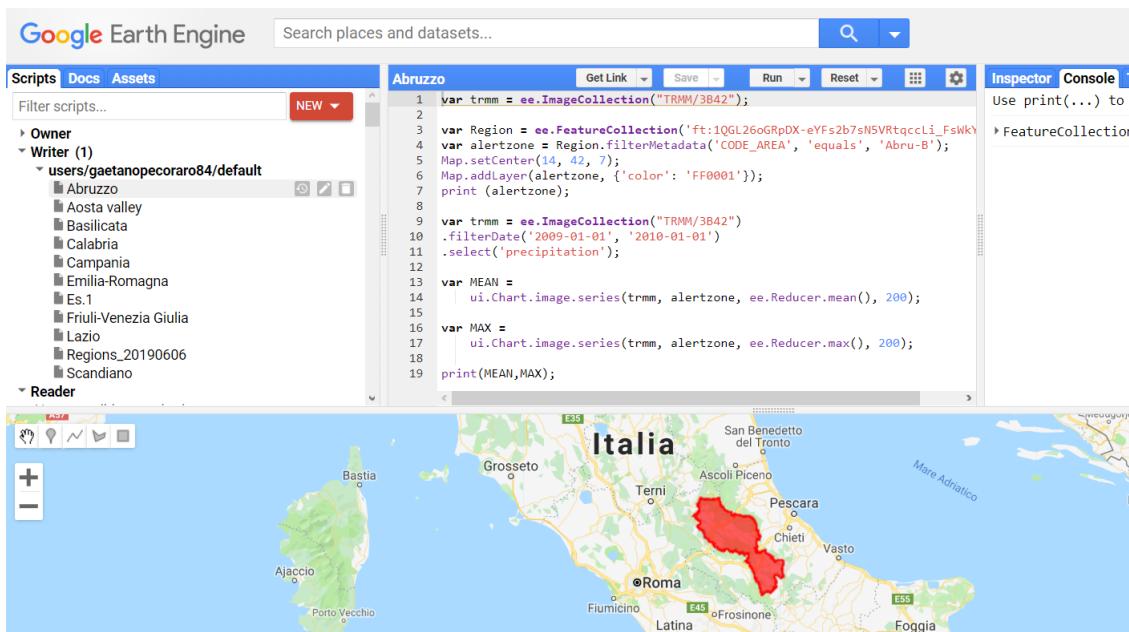
I dati relativi alle variabili pluviometriche provengono dalla missione *Tropical Rainfall Measuring Mission* (TRMM) effettuata dalla NASA e dal Japan Aerospace Exploration Agency (JAXA), disponibile su Google Earth Engine. Questi dati hanno una discretizzazione temporale di 3 ore e una discretizzazione spaziale di $0.25^\circ * 0.25^\circ$, corrispondente ad una porzione di territorio pari a circa $25 \text{ km} * 25 \text{ km}$.

Tali dati sono stati memorizzati in un dataset nel seguente modo:

- ogni record è relativo ad un giorno, precisamente ad un intervallo di 3 ore di quel giorno (ogni giorno, quindi, consiste di 8 intervalli e occupa 8 righe);

- per ogni colonna, ognuna relativa a una ZAM, e per ogni intervallo di 3 ore, si ha la media e il massimo delle piogge.

I dati relativi alle variabili pluviometriche riguardavano gli anni compresi nell'intervallo (2010, 2017). Successivamente per procedere all'elaborazione dei dati, e quindi ricavare gli attributi predittivi d'interesse per il modello da realizzare, è stato necessario raccogliere i dati relativi alle piogge anche per l'anno 2009, sempre dalla missione TRMM disponibile su Google Earth Engine, attraverso il seguente codice (figura 1):



The screenshot shows the Google Earth Engine interface. The top bar has a search field 'Search places and datasets...' and various tool buttons. The left sidebar shows 'Scripts', 'Docs', and 'Assets'. Under 'Assets', there is a 'Writer (1)' section with a folder named 'users/gaetanopecoraro84/default' containing a file named 'Abruzzo'. The main workspace is titled 'Abruzzo' and contains the following code:

```

1 var trmm = ee.ImageCollection("TRMM/3B42");
2
3 var Region = ee.FeatureCollection('ft:1OGI26oGrpDX-eYFs2b7sNSVRtqccli_Fswky');
4 var alertzone = Region.filterMetadata('CODE_AREA', 'equals', 'Abru-B');
5 Map.setCenter(14, 42, 7);
6 Map.addLayer(alertzone, {'color': 'FF0001'});
7 print(alertzone);
8
9 var trmm = ee.ImageCollection("TRMM/3B42")
10 .filterDate('2009-01-01', '2010-01-01')
11 .select('precipitation');
12
13 var MEAN =
14   ui.Chart.image.series(trmm, alertzone, ee.Reducer.mean(), 200);
15
16 var MAX =
17   ui.Chart.image.series(trmm, alertzone, ee.Reducer.max(), 200);
18
19 print(MEAN,MAX);

```

The map view shows a map of Italy with the Abruzzo region highlighted in red. Labels on the map include: Italia, San Benedetto del Tronto, Ascoli Piceno, Pescara, Chieti, Vasto, Foggia, Latina, Frosinone, Fiumicino, Roma, Terni, Grosseto, Ajaccio, Bastia, Porto Vecchio, and E35/E45 road networks. The bottom right corner of the map view shows a small inset map of Sicily.

Figura 1: codice per scaricare i dati

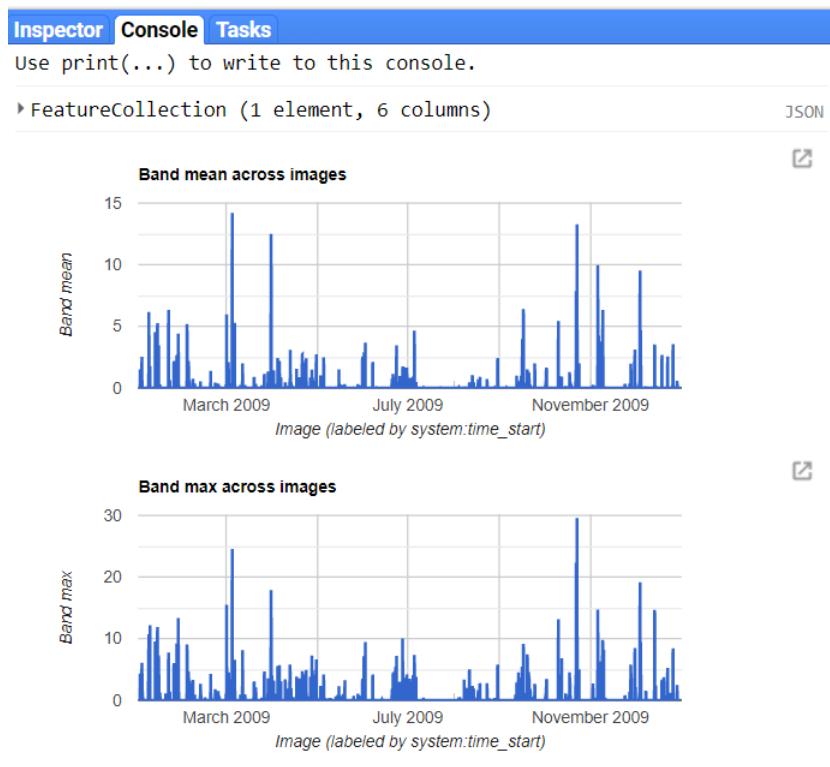


Figura 2: diagrammi risultanti

Dalla schermata sopra mostrata (figura 2), è stato possibile ricavare la media e il massimo delle piogge per ogni zona d'allerta.

A seguito di una analisi iniziale dei datasets, sono state ricavate le seguenti informazioni:

regione: Abruzzo – dataset delle frane

Il dataset presenta un numero di righe pari a 2924 e un numero di colonne pari a 7, di cui una riportante la data e le rimanenti contengono i fenomeni franosi per ogni ZAM. A parte la colonna relativa alla data, tutte le altre colonne sono di tipo integer. Non è stata rilevata alcuna presenza di valori mancanti o di valori anomali. Di seguito, verrà mostrata un'immagine delle prime righe del dataset, per dare un'idea di come esso si presenti (figura 3).

	A	B	C	D	E	F	G
1		111	27	184	19	38	30
2	Date	Abru-A	Abru-B	Abru-C	Abru-D1	Abru-D2	Abru-E
3	01/01/2010	0	0	0	0	0	0
4	02/01/2010	0	0	0	0	0	0
5	03/01/2010	0	0	0	0	0	0
6	04/01/2010	0	0	0	0	0	0
7	05/01/2010	0	0	0	0	0	0
8	06/01/2010	0	0	0	0	0	0
9	07/01/2010	0	0	0	0	0	0
10	08/01/2010	0	0	0	0	0	0

Figura 3: dataset relativo alle frane-Abruzzo

regione: Abruzzo – dataset delle piogge

Il dataset delle piogge presenta 23376 righe e 7 colonne, di cui una riportante la data e le rimanenti le ZAM. Per ogni ZAM sono presenti due sotto colonne: una descrivente la media e una il massimo delle piogge relative ad una certa data e ad un certo intervallo di 3 ore. Come per il dataset delle frane, viene mostrata un’immagine che rende l’idea di come si presenta il dataset relativo alle piogge (figura 4).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		Abru-A		Abru-B		Abru-C		Abru-D1		Abru-D2		Abru-E	
2	Date	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX
3	01/01/2010 0:00	0,08	1,319	0,467	2,594	0	0	0	0	0	0	0	0
4	01/01/2010 03:00	0,003	0,735	0,341	1,59	0	0	0,008	1,577	0	0	0,933	1,59
5	01/01/2010 06:00	0	0	0	0	0	0	0	0	0	0	0,179	1,17
6	01/01/2010 09:00	0	0	0,021	0,378	0	0	0	0	0	0	0,144	0,478
7	01/01/2010 12:00	0	0	1	4,645	0	0	0,407	7,523	0	0	3,733	6,45
8	01/01/2010 15:00	0,005	1	1,451	6,642	0	0	1,455	10,91	0	0	5,472	10,91
9	01/01/2010 18:00	0	0	0	0	0	0	0,029	0,85	0	0	0	0
10	01/01/2010 21:00	0	0	0	0	0	0	0	0	0	0	0	0
11	02/01/2010 0:00	0	0	0	0	0	0	0	0	0	0	0	0
12	02/01/2010 3:00	0	0	0	0	0,452	3,6	0	0	0	0	0	0
13	02/01/2010 6:00	0	0	0	0	0	0	0	0	0	0	0	0
14	02/01/2010 9:00	0	0	0	0	0	0	0	0	0	0	0	0

Figura 4: dataset relativo alle piogge-Abruzzo

regione: Campania - dataset delle frane

Tale dataset presenta la stessa forma di quello dell'Abruzzo, con l'unica differenza che, invece di contenere 6 zone di allerta, contiene 8 zone di allerta, ovvero due righe in più (figura 5).

	A	B	C	D	E	F	G	H	I
1		83	16	232	56	18	70	20	26
2	Date	Camp-1	Camp-2	Camp-3	Camp-4	Camp-5	Camp-6	Camp-7	Camp-8
3	01/01/2010	0	0	0	0	0	0	0	0
4	02/01/2010	0	0	1	0	0	0	0	0
5	03/01/2010	0	0	0	0	0	0	0	0
6	04/01/2010	0	0	0	0	0	0	0	0
7	05/01/2010	0	0	0	0	0	0	0	0
8	06/01/2010	0	0	0	0	0	0	0	0
9	07/01/2010	0	0	0	0	0	0	0	0
10	08/01/2010	0	0	0	0	0	0	0	0
11	09/01/2010	0	0	1	0	0	0	0	0

Figura 5: dataset relativo alle frane-Campania

regione: Campania - dataset delle piogge

Il dataset presenta 23376 righe e 9 colonne, di cui una che indica la data e le rimanenti che indicano le zone di allerta. Come per l'Abruzzo, per ogni ZAM si ha la media e il massimo delle piogge per ogni data e intervallo di 3 ore (figura 6).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		Camp-1		Camp-2		Camp-3		Camp-4		Camp-5		Camp-6		Camp-7		Camp-8	
2	Date	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX	MEAN	MAX
3	01/01/2010 00:00	0	0	0	0	0	0	0,13	1,29	0,129	1,29	0	0	0	0	0	0
4	01/01/2010 03:00	0	0	0	0	0	0	0,13	1,29	0,129	1,29	0	0	0	0	0	0
5	01/01/2010 06:00	0,209	1,429	0,127	1,429	0,005	0,324	0,115	1,26	0,176	0,678	0	0	0	0	0	0
6	01/01/2010 09:00	0,038	1,44	0,569	3,935	0,302	3,935	0,929	3,935	0,064	0,191	0	0	0	0	0	0
7	01/01/2010 12:00	0,267	8,349	1	8,349	0	0	0,038	0,644	0,033	0,426	0,01	0,426	0,196	0,979	0	0
8	01/01/2010 15:00	0,053	3	2,805	7,081	0,009	0,604	0,454	6,994	0,201	0,604	0	0	0,492	3,157	0,187	0,983
9	01/01/2010 18:00	0,746	8,257	1,296	8,257	0,055	1,757	0	0	0,631	1,757	1,26	2,885	0,173	1,757	0	0
10	01/01/2010 21:00	12,903	33,39	4,753	14,213	14,639	20,147	9,439	18,308	12,514	18,308	10,447	25,049	6,843	9,621	11,361	15,514
11	02/01/2010 00:00	4,934	11,16	0,001	5,67	8,233	11,834	1,283	7,139	6,713	9,842	8,151	14,261	8,955	16,012	14,451	15,848
12	02/01/2010 03:00	0,014	0,69	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 6: dataset relativo alle piogge-Campania

Successivamente i dati sono stati elaborati moltiplicando per 3 ogni valore, in modo tale da ricavare la quantità di pioggia espressa in millimetri, ed è stata definita la media e il massimo dell'intera giornata in modo da non considerare singoli intervalli. I dettagli di tale elaborazione verranno descritti nel capitolo che segue.

4

PRE-ELABORAZIONE DEI DATI

La pre-elaborazione e la pulizia dei dati rappresentano attività importanti che, in genere, devono essere eseguite prima di poter utilizzare un set di dati in modo efficace per fare machine learning. I dati non elaborati, infatti, risultano essere non affidabili e potrebbero presentare anche valori mancanti. Le attività di pre-elaborazione e pulizia possono essere eseguite con vari strumenti e linguaggi come ad esempio Python, R, ecc. Per il problema affrontato, si è deciso di adottare Python per effettuare la preparazione dei datasets. Uno dei principali aspetti da valutare è senz'altro la qualità dei dati. Come definito nel capitolo precedente, per il problema da affrontare vi erano a disposizione due datasets per ogni regione (Abruzzo e Campania), uno relativo alle piogge e uno relativo ai fenomeni franosi per ogni zona d'allerta. Inizialmente, è stata effettuata una prima pre-elaborazione per analizzare la qualità dei dati a disposizione. Non sono stati individuati dati incompleti e quindi è stato possibile facilmente proseguire con la creazione degli attributi necessari per affrontare il problema in esame. In letteratura, sono stati descritti vari fattori che possono influenzare l'innesto di una frana. Tra questi vengono citate le piogge cumulate antecedenti al fenomeno franoso. Di solito si tende ad affrontare questo problema considerando anche altri fattori, ma per via della difficoltà che si riscontra nel trovare i dati inerenti ai fattori influenzanti l'innesto delle frane ci si è dovuto limitare a considerare come cause scatenanti solo le informazioni relative alle piogge. Si è deciso, quindi, di calcolare come attributi predittivi dei fenomeni franosi i seguenti:

- pioggia cumulata 1gg, 2gg, 3gg, 4gg, 5gg, 10gg, 20gg, 30gg, 60gg, 180gg, 360gg prima della data in cui si è verificata la frana;
- pioggia massima di 1gg, 2gg prima dell'innesto della frana.

E' stato possibile calcolare gli attributi necessari grazie alla disposizione dei due datasets definiti prima, ma è stato necessario effettuare delle trasformazioni per poter estrapolare le giuste informazioni. Come definito nel paragrafo precedente, i due datasets forniti dal dipartimento di ingegneria civile dell'università degli studi di Salerno contengono come attributo comune

la data. Nel dataset relativo alle frane vi è la data che indica il giorno in cui si è verificato il fenomeno franoso. Nel dataset relativo alle piogge vi è la data ripetuta 8 volte in quanto è suddivisa per intervalli di 3 ore. Per ogni intervallo è specificata la media e il massimo della pioggia che si è avuta per ogni zona d'allerta. Per provvedere a calcolare gli attributi predittivi necessari è stato necessario omogeneizzare i due datasets. Si è deciso di modificare il dataset relativo alle piogge nel seguente modo:

- è stata calcolata la media totale delle piogge degli 8 intervalli temporali relativi a una data di ogni zona d'allerta;
- è stato calcolato il massimo dei massimi delle piogge inerenti agli 8 intervalli temporali di ogni zona d'allerta;

In questo modo, è stato possibile aggregare le informazioni relative ad ogni intervallo temporale per ogni data in esame.

Equilibrate le informazioni dei due datasets, in modo da poter effettuare il match tramite la data, si è provveduto a calcolare gli attributi predittivi.

- Per calcolare le varie piogge cumulate è stata implementata una funzione in python che viene eseguita sul dataset relativo alle frane e sul dataset modificato relativo alle piogge. Essa ha il compito di calcolare, per ogni data presente nel dataset relativo alle frane, la somma cumulata, utilizzando la funzione *cumsum* di python sulle piogge medie antecedenti al fenomeno franoso. Ad esempio, per calcolare la cumulata di 30 giorni prima la frana, la funzione chiama *cumsum* sulle 30 date precedenti alla frana contenute nel dataset relativo alle piogge. Il procedimento è lo stesso per ogni cumulata;
- Per calcolare le piogge massime di un giorno e 2 giorni prima l'innesto della frana, è stata implementata un'altra funzione che viene applicata sui due datasets e semplicemente recupera il valore calcolato inizialmente, relativo alla massima pioggia avutasi in una determinata data. In questo caso le date d'interesse sono quella antecedente e quella relativa a 2 giorni prima il fenomeno franoso.

Nel paragrafo successivo verrà descritta l'analisi effettuata sul dataset finale, ossia sul dataset risultante dall'elaborazione appena spiegata, d'interesse per il problema da affrontare.

5

ESTRAZIONE DELLE CARATTERISTICHE

Nel paragrafo precedente, si è discusso di come sono stati elaborati i dati provenienti dalla raccolta iniziale. In questo paragrafo, verranno esaminate le caratteristiche estratte sia dal dataset dell'Abruzzo sia da quello della Campania. Al fine di rendere più chiara la lettura, verrà esaminato prima il dataset della regione Abruzzo e poi quello inerente alla regione Campania. Di seguito, è riportata l'immagine delle prime righe del dataset della regione Abruzzo (figura 7):

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	DATA	ZONA	FRANE	CUM1GG	CUM2GG	CUM3GG	CUM4GG	CUM5GG	CUM10GG	CUM20GG	CUM30GG	CUM60GG	CUM90GG	CUM180GG	CUM360GG	MAX_1G	MAX_2G
2	2010-01-01 00:00:00	1	0	5,232	6,501	6,501	6,501	6,501	37,626	52,311	58,887	123,648	237,432	462,417	932,349	17,283	17,283
3	2010-01-01 00:00:00	2	0	7,551	7,551	8,073	8,268	8,79	42,318	69,909	106,737	199,692	320,589	523,35	1111,254	17,283	17,283
4	2010-01-01 00:00:00	3	0	1,002	1,002	2,262	2,262	2,262	24,945	39,453	47,952	111,216	235,53	440,574	889,608	14,475	14,475
5	2010-01-01 00:00:00	4	0	0	0	2,466	2,466	2,994	36,885	49,722	78,255	182,703	316,953	487,281	1066,236	0	0
6	2010-01-01 00:00:00	5	0	0	0	1,302	1,302	1,302	40,431	54,057	79,644	132,87	273,987	454,293	874,971	0	0
7	2010-01-01 00:00:00	6	0	1,02	1,02	2,022	2,022	2,184	31,509	72,651	118,578	241,419	354,906	516,393	1102,299	6,507	6,507
8	2010-01-02 00:00:00	1	0	0,264	5,496	6,765	6,765	37,89	49,83	58,29	122,139	237,696	450,636	932,613	4,014	17,283	
9	2010-01-02 00:00:00	2	0	8,388	15,939	15,939	16,461	16,656	50,589	76,95	113,01	204,315	328,827	518,877	1117,41	19,926	19,926
10	2010-01-02 00:00:00	3	0	0	1,002	1,002	2,262	2,262	24,942	35,1	47,952	111,216	235,53	413,226	889,608	0	14,475
11	2010-01-02 00:00:00	4	0	5,697	5,697	5,697	8,163	8,163	42,573	53,475	82,674	187,107	322,497	485,496	1067,22	32,73	32,73
12	2010-01-02 00:00:00	5	0	0	0	0	1,302	1,302	40,431	48,051	79,644	131,835	273,987	450,762	874,971	0	0
13	2010-01-02 00:00:00	6	0	31,383	32,403	32,403	33,405	33,405	60,732	102,558	147,543	267,366	386,064	538,527	1133,052	32,73	32,73
14	2010-01-03 00:00:00	1	0	0	0,264	5,496	6,765	6,765	37,656	45,132	58,05	122,139	237,696	446,37	931,778	0	4,014
15	2010-01-03 00:00:00	2	0	0,78	9,168	16,719	16,719	17,241	49,998	71,772	111,564	205,094	329,532	517,257	1116,237	6,543	19,926
16	2010-01-03 00:00:00	3	0	1,356	1,356	2,358	2,358	3,618	23,994	35,535	48,894	112,568	236,532	410,691	890,961	10,8	10,8
17	2010-01-03 00:00:00	4	0	1,167	6,864	6,864	6,864	9,33	42,324	54,294	83,22	188,274	323,298	480,003	1063,248	6,543	32,73
18	2010-01-03 00:00:00	5	0	0	0	0	1,302	40,431	47,811	79,644	131,748	273,987	447,147	874,911	0	0	
19	2010-01-03 00:00:00	6	0	0,522	31,905	32,925	32,925	33,927	61,254	96,603	145,698	267,888	386,586	538,752	1130,634	6,543	32,73
20	2010-01-04 00:00:00	1	0	0,903	0,903	1,167	6,399	7,668	31,671	41,754	55,749	123,042	238,599	446,646	932,679	2,88	2,88
21	2010-01-04 00:00:00	2	0	0,072	0,852	9,24	16,791	16,791	38,73	53,52	92,577	205,167	329,604	516,225	1116,243	3,792	6,543
22	2010-01-04 00:00:00	3	0	0	1,356	1,356	2,358	2,358	23,964	31,308	42,456	112,569	236,532	410,685	890,961	0	10,8
23	2010-01-04 00:00:00	4	0	0,972	2,139	7,836	7,836	7,836	37,401	48,48	76,134	189,246	324,27	480,648	1064,22	3,792	6,543
24	2010-01-04 00:00:00	5	0	0	0	0	0	0	40,431	46,983	54,654	131,748	273,987	447,147	874,911	0	0

Figura 7: dataset con le piogge cumulate, piogge massime e frane-Abruzzo

Dopo aver valutato vari tools a nostra disposizione, si è scelto di utilizzare la versione di prova di Ataccama per ottenere le caratteristiche di entrambi i datasets.

Si ha un numero di records pari a 17532 e un numero di attributi pari a 17. L'immagine che segue (figura 8) mostra due informazioni importanti: la zona di allerta e il numero di frane contengono solo uno specifico insieme di valori, infatti la zona di allerta contiene numeri compresi tra 1 e 6, mentre l'attributo frana contiene interi da 0 a 21, che corrispondono rispettivamente al minimo e al massimo dei fenomeni franosi. Gli attributi max1gg e max2gg, ossia il massimo di uno e di due giorni (in cui si è verificata la pioggia) antecedenti al fenomeno

franoso, presentano valori anomali, ossia valori che si discostano molto dalla media. Il tipo di dato degli attributi FRANE e ZONA è INTEGER.

Data Attributes		17	Filter attributes
Term	Name	Term	
DATA			
# ZONA		Enum	
# FRANE		Enum	
# CUM1GG			
# CUM2GG			
# CUM3GG			
# CUM4GG			
# CUM5GG			
# CUM10GG			
# CUM20GG			
# CUM30GG			
# CUM60GG			
# CUM90GG			
# CUM180GG			
# CUM360GG			
# MAX 1G		Outlier	
# MAX 2G		Outlier	

Figura 8: panoramica dei vari attributi

Tramite l'ausilio di Ataccama, si è svolta una analisi sull'attributo FRANE, di cui sono mostrati i risultati nella figura 9.

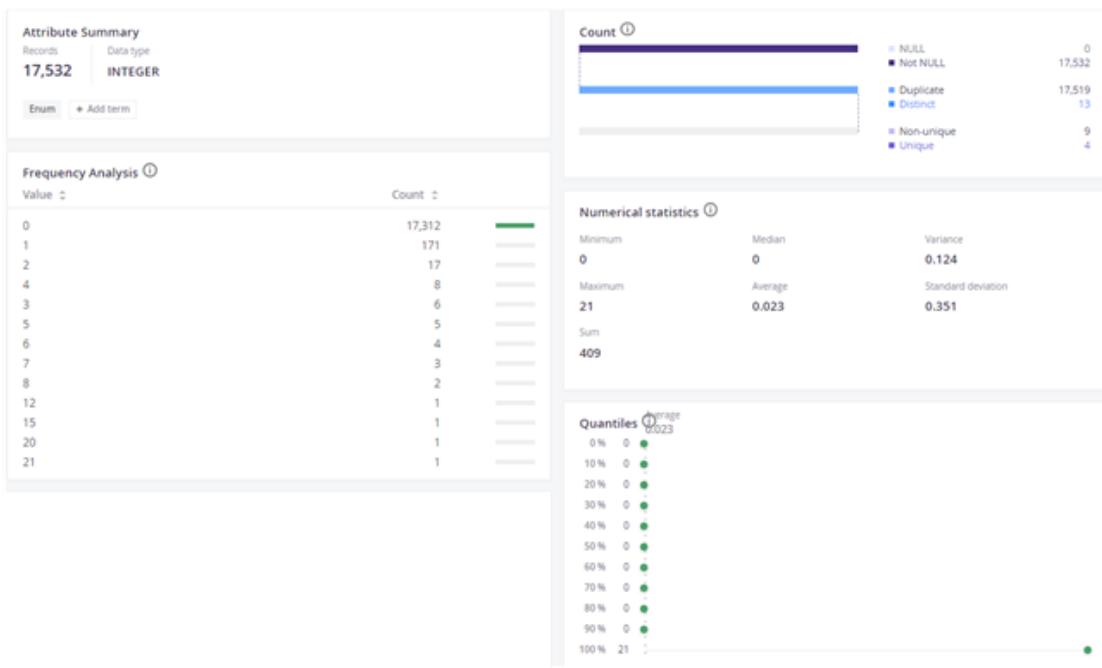


Figura 9: descrizione dell'attributo FRANE

Dall'immagine mostrata, si nota che non è stata riscontrata la presenza di valori nulli. Per quanto riguarda i duplicati, essi fanno riferimento al valore 0, ossia a quante volte non si sono verificati fenomeni franosi. I valori distinti indicano che per 13 volte si sono avuti diversi numeri di fenomeni franosi nel periodo di riferimento. I valori unique indicano, per esempio, che soltanto in un giorno si sono verificati 21 fenomeni franosi. Al contrario, i valori non-unique sono valori che compaiono più di una volta. Dalla precedente analisi, quindi, si evince già che il numero di eventi non franosi sono quasi la totalità degli eventi. La somma di tutti gli eventi franosi avvenuti nel periodo 2010-2017 risulta essere pari a 409. Tale valore è quasi nullo se considerato su 17532 eventi.

Si mostra un'analisi fatta sulle zone di allerta, la quale ci permette di capire quali sono le zone di allerta più colpite dai fenomeni franosi (figura 10):

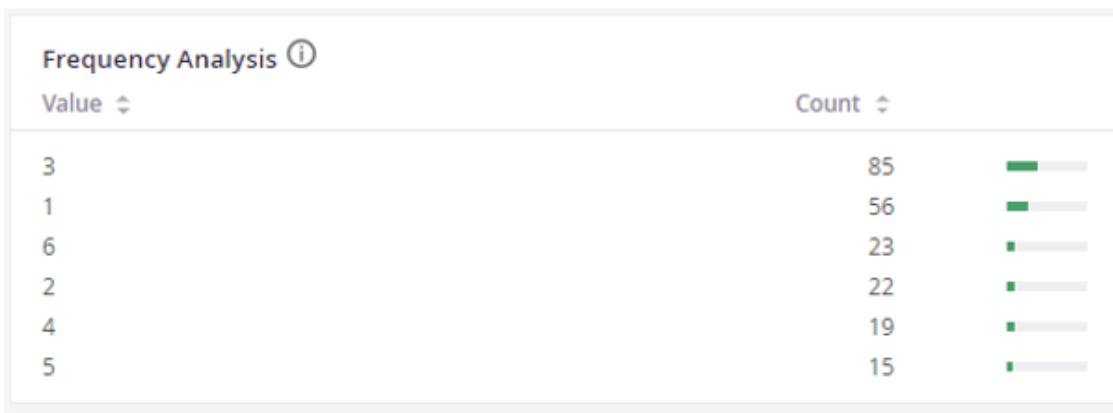


Figura 10: frequenza dei fenomeni franosi per ZAM

Per la regione Abruzzo, quindi, si denota che la zona di allerta più a rischio frane risulta essere la numero 3.

Per quanto riguarda gli attributi relativi alle cumulate e ai massimi delle piogge, essi sono di tipo FLOAT. Gli attributi che rappresentano le cumulate non contengono anomalie. Come si è visto in precedenza, gli attributi max1gg e max2gg presentano, invece, valori anomali. Dall’analisi fatta su Ataccama, risulta che la presenza dei valori anomali è dovuta all’altissimo numero di zeri. Per il resto, neppure questo attributo presenta valori nulli (così come nessun attributo).

La seguente immagine rappresenta le caratteristiche dell’attributo max1gg (figura 11):

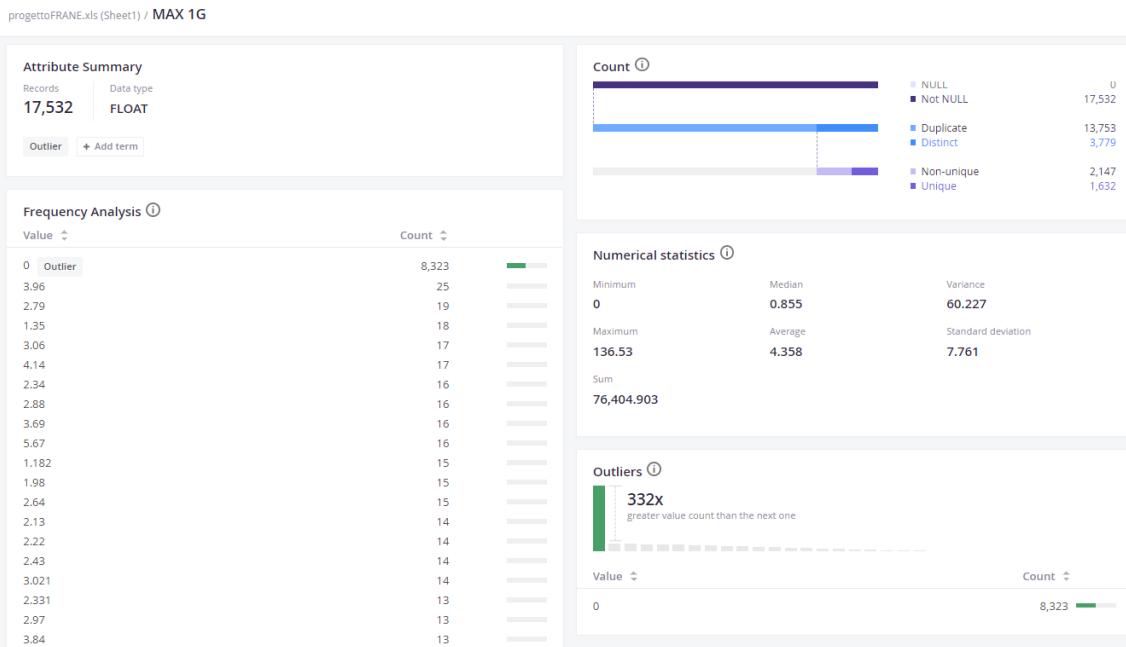


Figura 11: descrizione dell'attributo max1gg

Il dataset relativo alla regione Campania è composto da 23376 records e 17 attributi. I tipi di dati degli attributi ZONA e FRANA sono INTEGER, mentre per tutti gli altri attributi (data esclusa) si tratta di tipi FLOAT. La seguente immagine mostra come si presenta il dataset in excel (figura 12):

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q		
DATA	ZONA	FRANE	CUM1GG	CUM2GG	CUM3GG	CUM4GG	CUM5GG	CUM10GC	CUM20GG	CUM30GC	CUM60GC	CUM90GC	CUM180G	CUM360G	MAX 1G	MAX 2G		
2	2010-01-01 00:00:00	1	0	0	0	1,218	1,218	1,218	23,604	39,684	82,983	256,062	371,172	504,255	1106,613	0	0	
3	2010-01-01 00:00:00	2	0	0	0	1,704	1,704	1,704	30,396	51,618	101,511	252,686	362,433	518,07	1198,602	0	0	
4	2010-01-01 00:00:00	3	0	0	0	0,051	0,051	0,051	26,88	48,636	73,935	259,896	371,505	532,367	1060,365	0	0	
5	2010-01-01 00:00:00	4	0	0,939	1,629	4,404	4,404	4,404	30,558	42,264	54,126	164,151	239,493	389,079	896,052	15,588	15,588	
6	2010-01-01 00:00:00	5	0	0	0,198	1,533	1,533	1,533	23,142	59,799	86,052	223,344	299,568	448,479	963,933	0	1,98	
7	2010-01-01 00:00:00	6	0	0	0	4,989	4,989	4,989	23,286	78,72	96,276	243,585	328,674	459,27	1006,917	0	0	
8	2010-01-01 00:00:00	7	0	0	0	0,321	0,321	0,321	31,658	91,677	115,686	227,115	306,006	439,362	979,371	0	0	
9	2010-01-01 00:00:00	8	0	0	0	0,06	0,06	0,06	19,173	111,396	126,984	263,055	359,331	491,034	1080,168	0	0	
10	2010-01-02 00:00:00	1	0	42,648	42,648	42,648	43,866	43,866	61,314	82,332	125,631	296,361	413,82	537,57	1149,261	100,17	100,17	
11	2010-01-02 00:00:00	2	0	31,521	31,521	31,521	33,225	33,225	58,041	83,139	133,032	282,789	393,954	525,732	1230,123	42,639	42,639	
12	2010-01-02 00:00:00	3	1	45,03	45,03	45,03	45,081	45,081	70,41	93,666	118,965	297,252	416,535	572,247	1105,395	60,441	60,441	
13	2010-01-02 00:00:00	4	0	33,315	34,254	34,944	37,719	37,719	63,693	75,579	87,441	185,442	272,808	409,71	929,367	54,924	54,924	
14	2010-01-02 00:00:00	5	0	41,244	41,244	41,442	42,777	42,777	64,386	101,043	127,296	260,982	340,812	487,344	1005,177	54,924	54,924	
15	2010-01-02 00:00:00	6	0	35,547	35,547	35,547	40,536	40,536	58,833	114,267	131,823	271,581	364,221	494,244	1042,464	75,147	75,147	
16	2010-01-02 00:00:00	7	0	23,163	23,163	23,163	23,484	23,484	58,821	114,84	138,849	249,885	329,169	460,017	1001,97	28,863	28,863	
17	2010-01-02 00:00:00	8	0	34,644	34,644	34,644	34,704	34,704	53,817	146,04	161,628	293,598	393,975	521,061	1114,74	46,542	46,542	
18	2010-01-03 00:00:00	1	0	14,844	57,492	57,492	57,492	57,492	58,71	71,706	96,819	140,475	310,425	428,664	552,349	1128,231	33,48	100,17
19	2010-01-03 00:00:00	2	0	0,003	31,524	31,524	31,524	33,228	50,427	83,142	133,035	282,792	393,957	524,304	1198,155	17,01	42,639	
20	2010-01-03 00:00:00	3	0	24,699	69,729	69,729	69,729	69,729	87,729	118,365	143,664	317,829	440,49	596,946	1127,202	35,502	60,441	
21	2010-01-03 00:00:00	4	0	3,849	37,164	38,103	38,793	41,568	62,829	79,428	91,29	189,273	276,657	413,478	926,775	21,417	54,924	
22	2010-01-03 00:00:00	5	0	20,139	61,383	61,383	61,581	62,916	84,39	121,182	147,435	280,266	359,874	507,483	1007,436	29,526	54,924	
23	2010-01-03 00:00:00	6	0	24,453	60	60	60	64,989	83,286	138,648	156,276	271,74	386,467	518,691	1061,466	42,783	75,147	
24	2010-01-03 00:00:00	7	0	26,865	50,028	50,028	50,028	50,349	85,686	141,705	165,714	263,268	356,034	486,882	1019,382	48,036	48,036	
25	2010-01-03 00:00:00	8	0	43,353	77,997	77,997	77,997	78,057	97,17	189,393	204,981	308,613	437,328	564,417	1158,036	47,544	47,544	
26	2010-01-04 00:00:00	1	0	0	14,844	57,492	57,492	57,492	68,784	86,301	97,248	310,011	428,664	552,348	1100,724	0	33,48	
27	2010-01-04 00:00:00	2	0	0	0,003	31,524	31,524	31,524	50,364	70,824	83,769	282,348	393,957	524,304	1192,884	0	17,01	
28	2010-01-04 00:00:00	3	0	0	24,699	69,729	69,729	69,729	87,366	110,979	119,451	317,037	440,49	596,592	1124,361	0	35,502	
29	2010-01-04 00:00:00	4	0	0	3,849	37,164	38,103	38,793	62,829	72,321	82,527	189,138	276,657	413,478	926,742	0	21,417	
30	2010-01-04 00:00:00	5	0	0	20,139	61,383	61,383	61,581	84,39	111,117	125,349	280,206	359,874	507,483	1005,498	0	29,526	

Figura 12: dataset con le piogge cumulate, piogge massime e frane-Campania

Si nota che, come per l'Abruzzo, gli attributi zona di allerta e frana assumono soltanto uno specifico insieme di valori (figura 13).

	Name	Term
#	ZONA	Enum
#	FRANE	Enum
#	CUM1GG	Outlier
#	CUM2GG	Outlier
#	CUM3GG	
#	CUM4GG	
#	CUM5GG	
#	CUM10GG	
#	CUM20GG	
#	CUM30GG	
#	CUM60GG	
#	CUM90GG	
#	CUM180GG	
#	CUM360GG	
#	MAX 1G	Outlier
#	MAX 2G	Outlier

Figura 13: panoramica di vari attributi

A differenza dell'Abruzzo, invece, si vede che anche le prime due cumulate presentano valori anomali. Prima di descrivere statisticamente gli altri attributi, si rappresentano le informazioni riguardanti la colonna dei fenomeni franosi (figura 14)

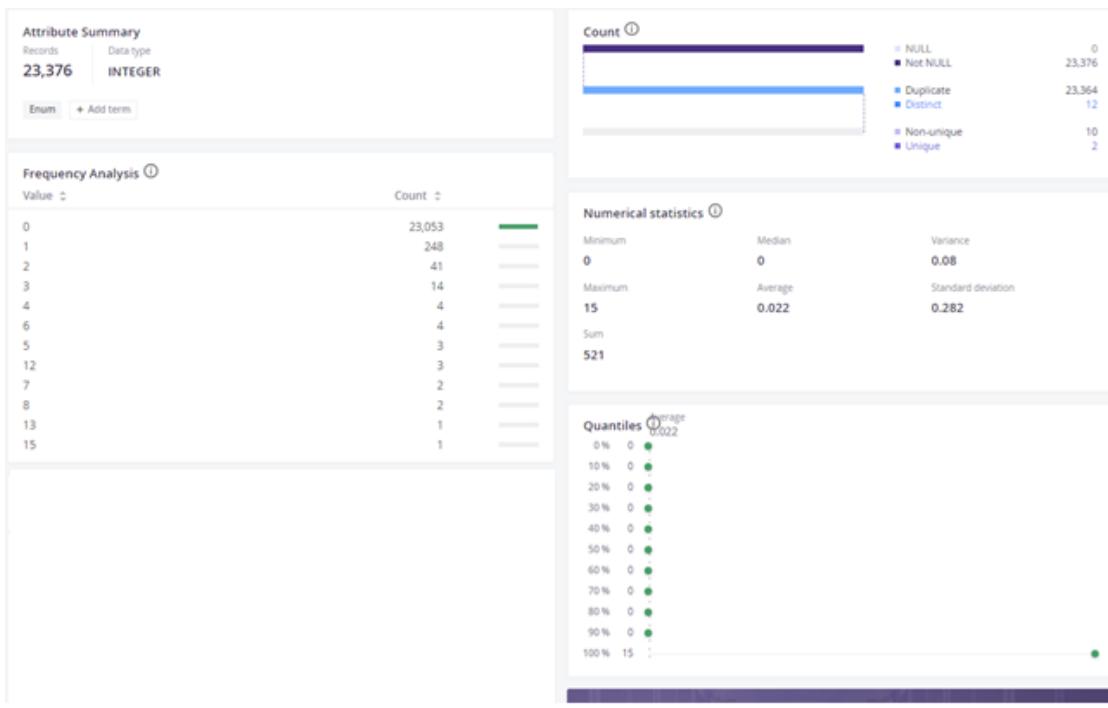


Figura 14: descrizione dell'attributo FRANE

Anche qui la situazione risulta essere uguale a quella della regione Abruzzo, in quanto il numero di fenomeni franosi, rispetto a quelli non franosi, è molto piccolo. Si classificano, di seguito, le zone di allerta in base ai fenomeni franosi che vi si sono verificati:

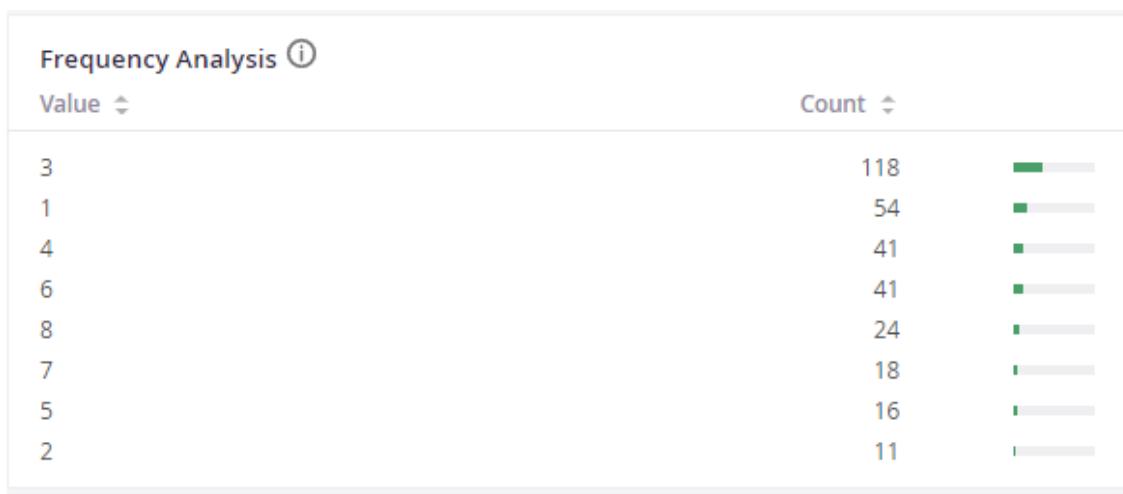


Figura 15: frequenza dei fenomeni franosi per ogni ZAM

Nella seguente immagine, si raffigurano le statistiche inerenti all'attributo cum1gg:

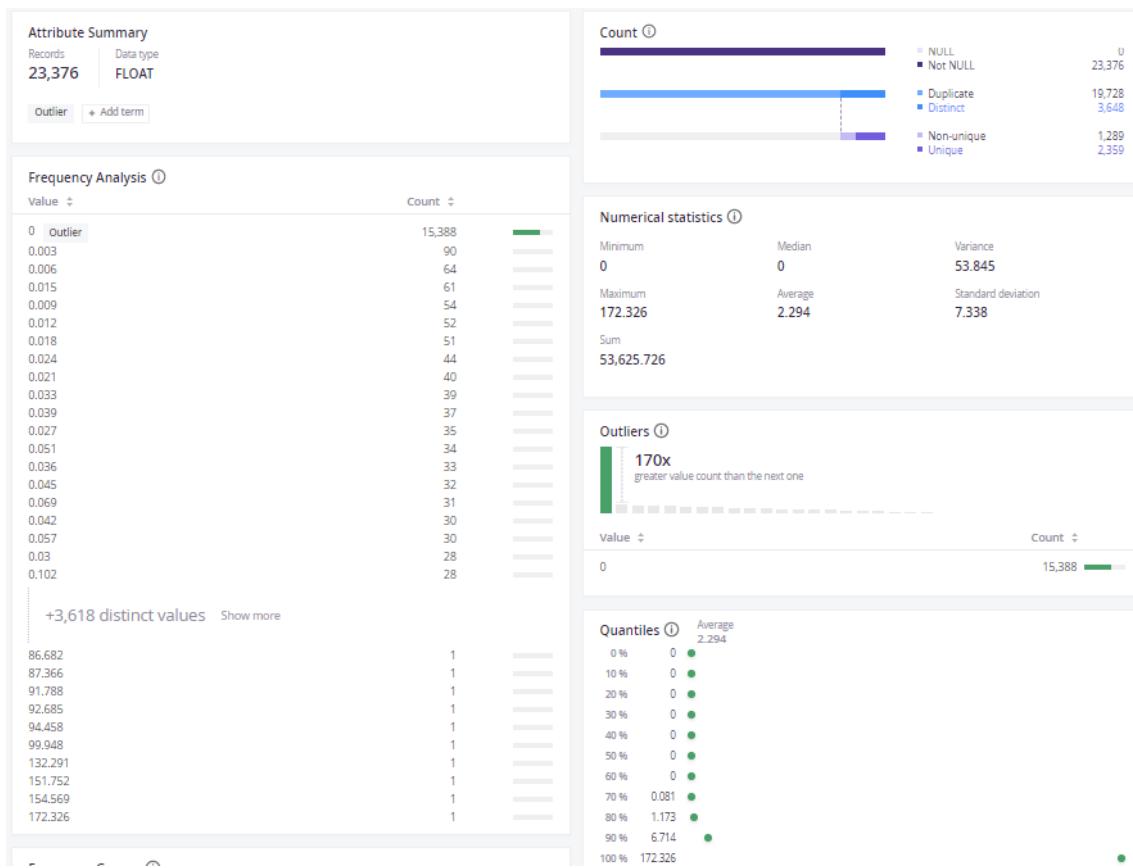


Figura 16: descrizione dell'attributo cum1gg

Dopo aver studiato le statistiche degli attributi, ci si è soffermati sulla rappresentazione grafica dei valori degli stessi. Sono stati costruiti gli istogrammi degli attributi numerici, ossia di quelli inerenti alle cumulate e ai massimi.

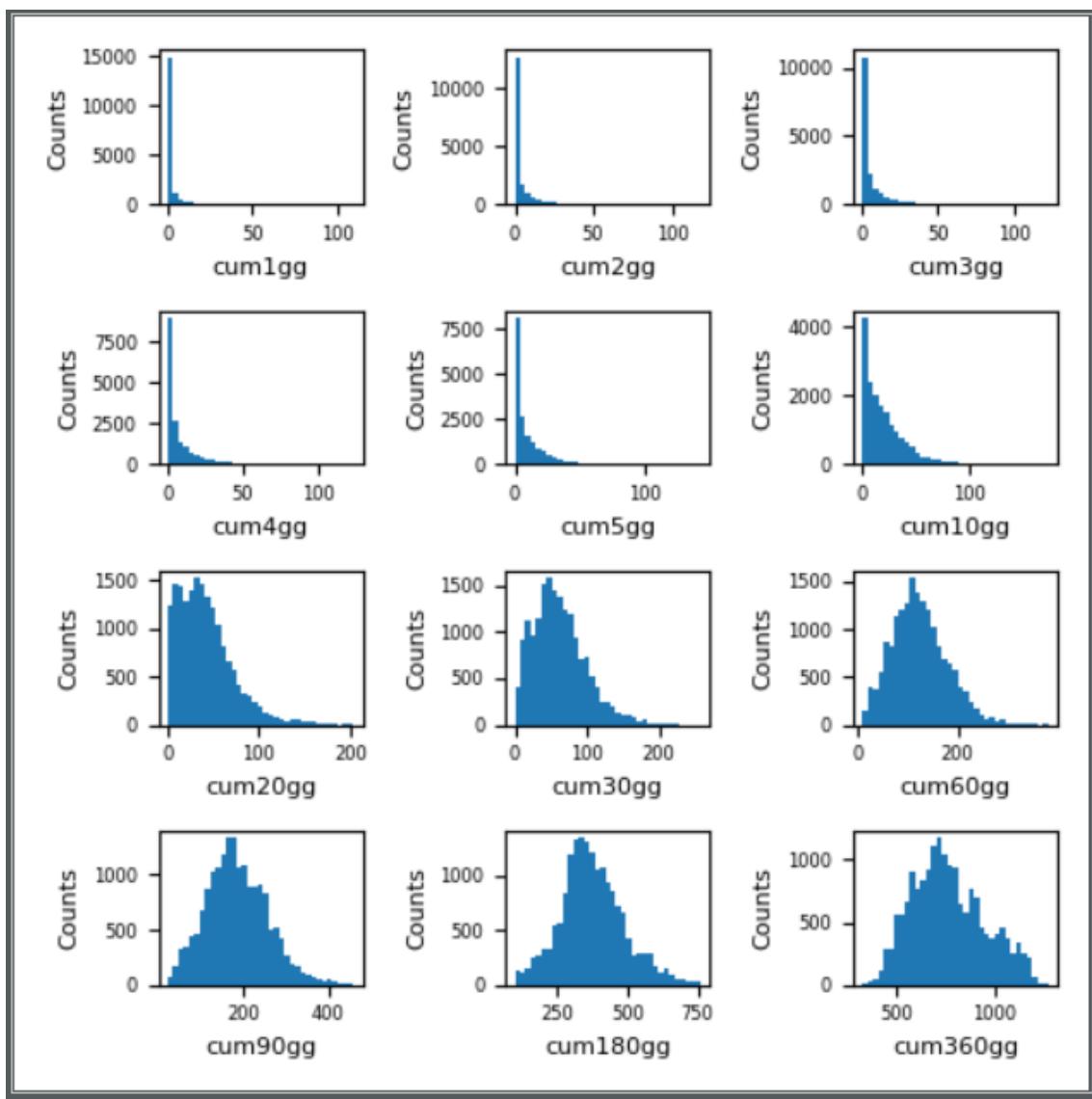


Figura 17: istogramma relativo all'Abruzzo

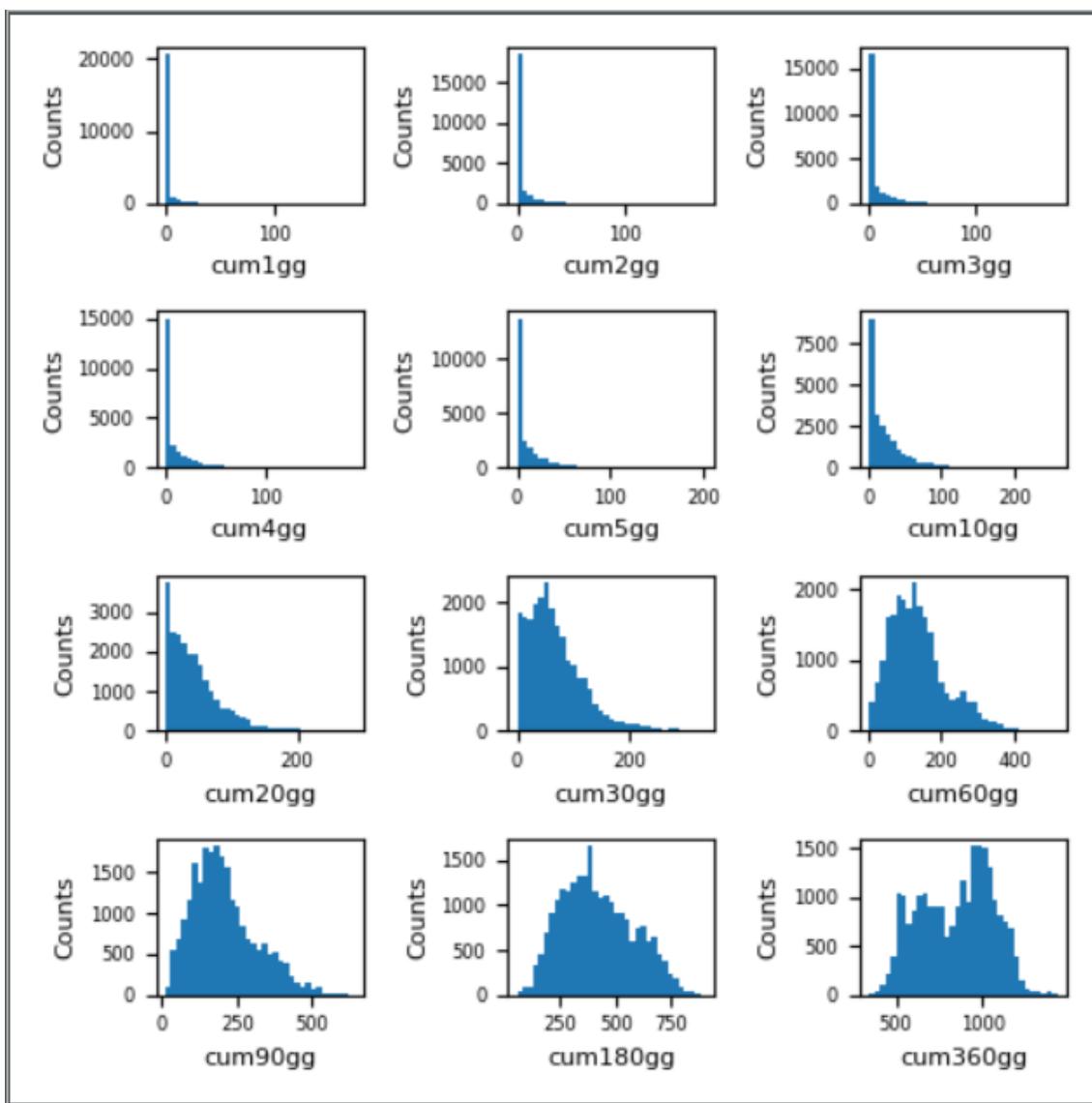


Figura 18: istogramma relativo alla Campania

Allo scopo di mostrare quali relazioni intercorrono tra gli attributi, si è scelto di usare uno strumento grafico molto utile e chiaro: gli scatterplot. Per prima cosa, si sono rappresentate tutte le relazioni che intercorrono tra le coppie di attributi, attraverso il comando `pairs()` di python: (purtroppo, in nessun modo è stato possibile inserire i nomi degli attributi per ogni grafico a causa dell'elevato numero di coppie di variabili. Sia per le righe che per le colonne si hanno gli stessi attributi visti in precedenza nelle panoramiche dei datasets. Per lo stesso motivo, i grafici non risultano chiari nei margini. Per dare una visione generale, si è scelto lo stesso di inserirli.)



Figura 19: scatterplot relativo all'Abruzzo

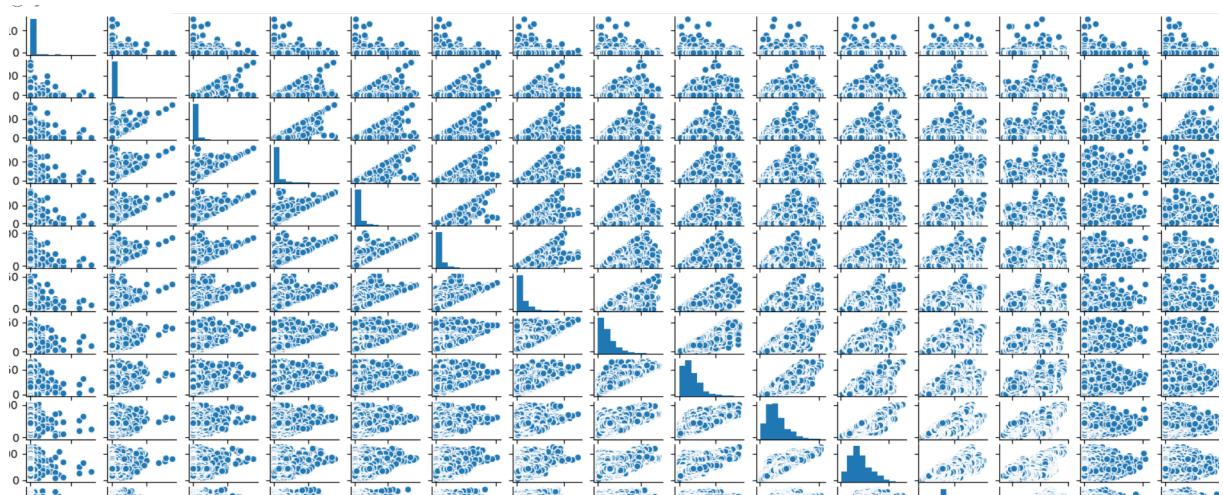


Figura 20: scatterplot relativo alla Campania

Oltre a visionare lo scatterplot e dare un'idea di che tipo è la relazione che intercorre tra ognuna di queste coppie di variabili, si può usare una misura quantitativa che indica la correlazione che esiste tra le variabili. Tale misura è compresa tra -1 e 1. Se assume un valore positivo indica una correlazione lineare positiva tra le variabili e se assume un valore negativo indica una correlazione lineare negativa. Quando tale valore è vicino o pari a 0 significa che non c'è correlazione lineare. Di seguito, sono mostrate le matrici che evidenziano tutte le possibili correlazioni tra le coppie di variabili:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	ABRUZZO	zonaAllerta	numeroFrane	cum1gg	cum2gg	cum3gg	cum4gg	cum5gg	cum10gg	cum20gg	cum30gg	cum60gg	cum90gg	cum180gg	cum360gg	max1gg	max2gg	
2	zonaAllerta		1	-0.025515861	0.014826	0.019576	0.023465	0.026903	0.029856	0.041439	0.057236	0.069941	0.096402	0.112818	0.14067396	-0.03094	-0.03572	
3	numeroFrane	-0.025515861		1	0.088086	0.071805	0.061527	0.055479	0.051442	0.045579	0.045318	0.036192	0.025838	0.011392	-0.00214546	-0.003153057	0.089809	0.079207
4	cum1gg	0.014825689	0.088085963		1	0.756854	0.606801	0.514264	0.460889	0.335034	0.231672	0.190561	0.161756	0.131817	0.090348083	0.078071735	0.809114	0.604409
5	cum2gg	0.019576428	0.071805269	0.756854		1	0.851036	0.728048	0.645441	0.463901	0.323636	0.268368	0.224314	0.180952	0.124592813	0.105396823	0.639541	0.801816
6	cum3gg	0.023464589	0.061526577	0.606801	0.851036		1	0.891723	0.792313	0.562363	0.396568	0.328979	0.273379	0.220121	0.151503041	0.126882385	0.513944	0.698294
7	cum4gg	0.026903422	0.055478787	0.514264	0.728048	0.891723		1	0.915368	0.644702	0.457014	0.377011	0.313684	0.253532	0.174815732	0.145180181	0.44489	0.602972
8	cum5gg	0.029856035	0.051441914	0.460889	0.645441	0.792313	0.915368		1	0.718306	0.510938	0.418251	0.348079	0.283005	0.1952276	0.160851301	0.402825	0.541975
9	cum10gg	0.041438783	0.045578616	0.335034	0.463901	0.562365	0.644702	0.718306		1	0.725231	0.585962	0.47645	0.39655	0.275301151	0.222179851	0.302348	0.397731
10	cum20gg	0.057235842	0.045317559	0.231672	0.323636	0.396568	0.457014	0.510938	0.725231		1	0.830491	0.638509	0.550253	0.386439077	0.307038272	0.219574	0.288013
11	cum30gg	0.069940736	0.036192062	0.190561	0.268388	0.328979	0.377011	0.418251	0.585962	0.830491		1	0.748415	0.66672	0.47276973	0.372444844	0.190142	0.248281
12	cum60gg	0.096402373	0.025837991	0.161756	0.224314	0.273379	0.313684	0.348079	0.47645	0.638509	0.748415		1	0.866011	0.652441747	0.518816832	0.184751	0.231207
13	cum90gg	0.112817752	0.011391939	0.131817	0.180952	0.220121	0.253532	0.283005	0.39655	0.550253	0.66672	0.866011		1	0.775341886	0.639844712	0.17384	0.213776
14	cum180gg	0.14067396	-0.002145461	0.090348	0.124593	0.151503	0.174816	0.195228	0.275301	0.386439	0.47277	0.652442	0.775342		1	0.850898763	0.136678	0.158958
15	cum360gg	0.164078736	-0.003153057	0.078072	0.105397	0.126882	0.14518	0.160851	0.22218	0.307038	0.372445	0.518817	0.639845	0.850898763		1	0.112422	0.127524
16	max1gg	-0.030937614	0.0889029	0.809114	0.639541	0.513944	0.44489	0.402825	0.302348	0.219574	0.190142	0.184751	0.17384	0.136678109		1	0.112421632	0.1
17	max2gg	-0.035724202	0.079207377	0.604409	0.801816	0.698294	0.602972	0.541975	0.397731	0.288013	0.248281	0.231207	0.213776	0.158957552	0.127524382	0.74911	1	

Figura 21: matrice di correlazione-Abruzzo

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
1	CAMPANIA	zonaAllerta	numeroFrar	cum1gg	cum2gg	cum3gg	cum4gg	cum5gg	cum10gg	cum20gg	cum30gg	cum60gg	cum90gg	cum180gg	cum360gg	max1gg	max2gg	
2	zonaAllerta		1	-0.02987	-0.00547	-0.0072	-0.00853	-0.00989	-0.0111	-0.01485	-0.01905	-0.02181	-0.02963	-0.03391	-0.04409	-0.0668	-0.04645	-0.05916
3	numeroFrane	-0.02987		1	0.063836	0.073768	0.067443	0.065456	0.063811	0.066883	0.056983	0.041532	0.049564	0.034722	0.018018	0.011522	0.06035	0.064687
4	cum1gg	-0.00547	0.063836		1	0.759832	0.621484	0.535544	0.474653	0.352267	0.251223	0.200988	0.171383	0.139977	0.07699	0.07155	0.0827944	0.627071
5	cum2gg	-0.0072	0.073768	0.759832		1	0.856871	0.745156	0.66397	0.488015	0.346457	0.279572	0.233544	0.19115	0.106704	0.096599	0.65427	0.815842
6	cum3gg	-0.00853	0.067443	0.621484	0.856871		1	0.89893	0.807357	0.591834	0.420634	0.340739	0.281505	0.230075	0.129798	0.115385	0.546431	0.7153
7	cum4gg	-0.00989	0.065456	0.535544	0.745156	0.89893		1	0.921634	0.675487	0.485612	0.391592	0.320411	0.26352	0.149948	0.131967	0.479339	0.629581
8	cum5gg	-0.0111	0.063811	0.474653	0.66397	0.807357	0.921634		1	0.745654	0.542178	0.436471	0.353854	0.29367	0.167897	0.146464	0.433175	0.57088
9	cum10gg	-0.01485	0.066883	0.352267	0.488015	0.591834	0.675487	0.745654		1	0.748308	0.609786	0.480664	0.407763	0.236237	0.199272	0.192244	0.433333
10	cum20gg	-0.01905	0.056983	0.251223	0.346457	0.420634	0.486612	0.542178	0.748308		1	0.848104	0.644225	0.562956	0.335295	0.264998	0.238865	0.31008
11	cum30gg	-0.02181	0.041532	0.200988	0.279572	0.340739	0.391592	0.436471	0.609786	0.848104		1	0.759004	0.682884	0.416764	0.308764	0.188736	0.249794
12	cum60gg	-0.02963	0.049564	0.171383	0.233544	0.281505	0.320411	0.353854	0.480664	0.644225	0.759004		1	0.876608	0.600705	0.418599	0.166102	0.215806
13	cum90gg	-0.03391	0.034722	0.139977	0.19115	0.230075	0.26352	0.29367	0.407763	0.562956	0.682884	0.876608		1	0.742097	0.50347	0.135563	0.177151
14	cum180gg	-0.04409	0.018018	0.07699	0.106704	0.129798	0.149948	0.167897	0.236237	0.335295	0.416764	0.600705	0.742097		1	0.697754	0.073408	0.098167
15	cum360gg	-0.0668	0.011522	0.07155	0.096599	0.115385	0.131967	0.146464	0.199272	0.264998	0.308764	0.418599	0.50347	0.697754		1	0.058267	0.075765
16	max1gg	-0.04645	0.06035	0.827944	0.65427	0.546431	0.479339	0.433175	0.332344	0.238865	0.188736	0.166102	0.135563	0.073408	0.058267		1	0.760844
17	max2gg	-0.05916	0.064687	0.627071	0.815842	0.7153	0.629581	0.57088	0.433333	0.31008	0.249794	0.215806	0.177151	0.098167	0.075765	0.760844		

Figura 22: matrice di correlazione-Campania

Si evince una forte correlazione positiva delle cumulate vicine. Infatti, risultano essere ben correlate le cumulate di un giorno prima con le cumulate di 2 giorni prima, così come sono ben correlate le cumulate di 3 giorni prima con le cumulate di 4 giorni prima e così via. Un'altra informazione che si evince è che le cumulate molto distanti, per esempio di un giorno prima e di 360 giorni prima, risultano essere non correlate (data la vicinanza di tali correlazioni allo 0). Queste informazioni si sono rivelate molto utili, in quanto si è potuto stabilire quali sono le coppie di cumulate o massimi di pioggia maggiormente correlate in modo lineare, il che sarà molto utile per la fase di classificazione in cui, infatti, dopo aver avuto una conferma dagli esperti del dipartimento di ingegneria civile, si è deciso di scegliere i due attributi maggiormente correlati linearmente nel seguente modo:

- I due attributi devono essere scelti in modo tale da avere come primo attributo una cumulata di pioggia antecedente il fenomeno franoso di massimo 10 giorni e come secondo attributo una cumulata di pioggia antecedente il fenomeno franoso di più di 10 giorni o uno dei massimi di pioggia;
- Successivamente creare uno scatterplot in modo tale da avere sull'asse delle x il primo attributo e sull'asse delle y il secondo attributo (entrambi descritti al punto precedente);
- I punti di tale scatterplot corrispondono al colore blu se si è verificata la frana e al colore rosso se la frana non si è verificata.

Per la regione Abruzzo, le due variabili maggiormente correlate linearmente corrispondono alla cumulata e al massimo di due giorni prima l'evento della frana, mentre per la Campania si ha la cumulata e il massimo di un giorno prima il fenomeno franoso. Di seguito, si riportano tali grafici:

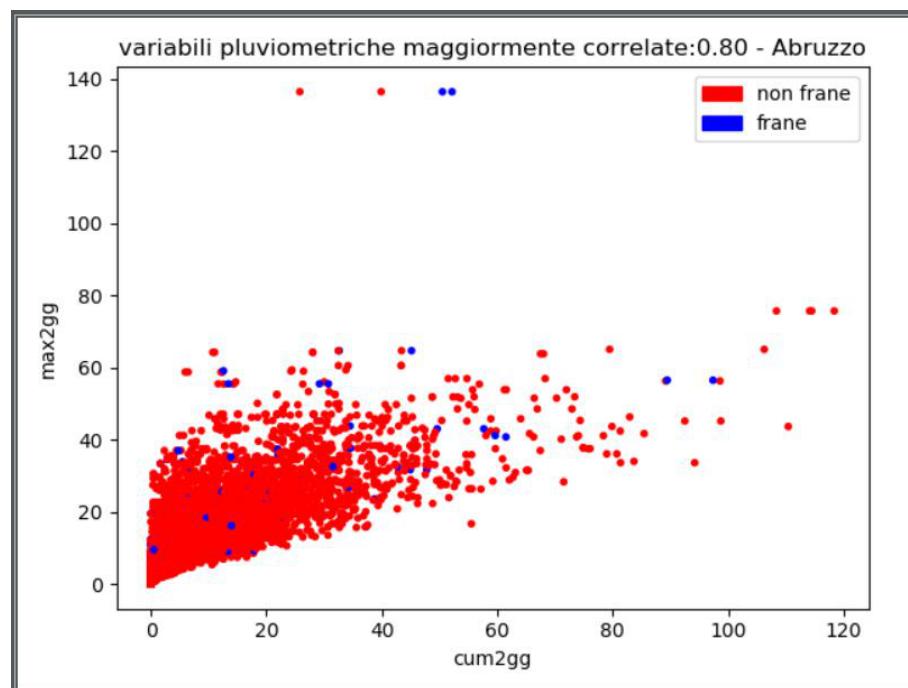


Figura 23: grafico-Abruzzo

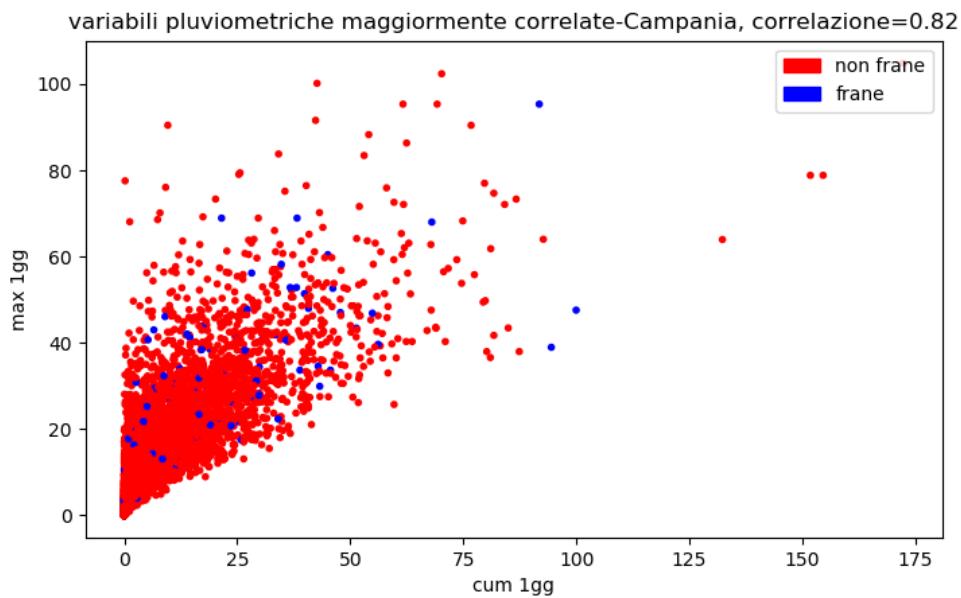


Figura 24: grafico-Campania

Si conclude questo capitolo confermando che il problema verrà affrontato considerando come variabile dipendente le frane e come attributi predittivi la cumulata e il massimo di due giorni prima per l'Abruzzo, la cumulata e il massimo di un giorno prima per la Campania.

6

MACHINE LEARNING

6.1 INTRODUZIONE

La problematica descritta è stata analizzata utilizzando, come accennato prima, algoritmi di autoapprendimento nel campo del machine learning per trasformare i dati in conoscenza. Grazie alle tante librerie open source che sono state sviluppate nel corso degli anni, risulta più semplice imparare ad utilizzare questi potenti algoritmi per individuare schemi nei dati ed eseguire previsioni a proposito degli eventi futuri. Grazie al machine learning, si può contare su solidi filtri in grado di eliminare lo spam dalla posta elettronica, su software per il riconoscimento del testo e della voce, su affidabili motori di ricerca per il Web ecc.^[3]

Vi sono quattro diversi tipi di machine learning:

- apprendimento con supervisione;
- apprendimento senza supervisione;
- apprendimento di rafforzamento;
- apprendimento semi-supervisionato.

Nell'apprendimento *con supervisione* lo scopo principale consiste nel trarre un modello a partire dai dati di addestramento etichettati, i quali ci consentono di effettuare previsioni relative ai dati non disponibili o futuri. Il termine "con supervisione" fa riferimento al fatto che nell'insieme di campioni i segnali di output desiderati (le etichette) sono già noti;

Una sottocategoria di apprendimento con supervisione è *la classificazione*. Essa ha lo scopo di prevedere le etichette di una categoria di classi per le nuove istanze, sulla base delle osservazioni compiute nel passato. Queste etichette sono valori discreti, non ordinati che possono essere considerati come appartenenti a un gruppo delle istanze.

La classificazione può essere binaria, multi-classe, multi-etichetta e multi-output. Nel problema in esame, verrà sviluppato un task di classificazione binaria: l'algoritmo di apprendimento

automatico impara un insieme di regole con lo scopo di distinguere fra due possibili classi. Un'altra sottocategoria di apprendimento con supervisione è la regressione, dove il segnale risultante è un valore continuo.

Nell'analisi di regressione vi è un certo numero di variabili predittive (descrittive) e una variabile target continua (risultato), l'obiettivo è quello di trovare una relazione fra queste variabili, tale che consenta di prevedere un risultato.

6.2 DEFINIZIONE DEL PROBLEMA

Osservando l'andamento dei dati rappresentato nel paragrafo inerente alla preparazione dei dati, si ha avuto modo di notare che non vi è una buona relazione che intercorre tra i fenomeni franosi e le piogge e questo perché le frane dipendono anche da molti altri fattori morfologici. Non è stato possibile, quindi, ottenere risultati soddisfacenti utilizzando la regressione in quanto vi è troppa incertezza nei dati.

Inizialmente l'obiettivo era quello di soffermarsi sui fenomeni franosi maggiori di 1 e maggiori di 5, ma a causa dei pochi dati a disposizione inerenti a questi fenomeni gli obiettivi sono stati raffinati. Si è deciso di affrontare il problema utilizzando la classificazione binaria dell'apprendimento supervisionato, andando a considerare la classe inerente ai fenomeni franosi e la classe inerente ai fenomeni non franosi. È stato possibile farlo perché vi erano già a disposizione i target su cui addestrare il modello. Prima di descrivere lo sviluppo del task di classificazione, si spiega, in breve, il motivo per cui questo problema non può essere affrontato come un problema di regressione.

6.3 TASK DI REGRESSIONE

In questo paragrafo, si dimostra che il task di regressione non è adatto a questo problema. L'idea è quella di trovare una correlazione fra il fenomeno franoso e una cumulata o massimo di pioggia, ma visionando la matrice delle correlazioni si nota che non c'è correlazione lineare tra di esse. Si è optato, quindi, a trovare una correlazione non lineare tra la frana e le piogge, ma anche in quel caso non si sono verificate correlazioni. Ciò avviene sia considerando la regione Abruzzo sia la regione Campania.

Di seguito, si riportano alcuni grafici per mostrare da vicino l'andamento dei dati:

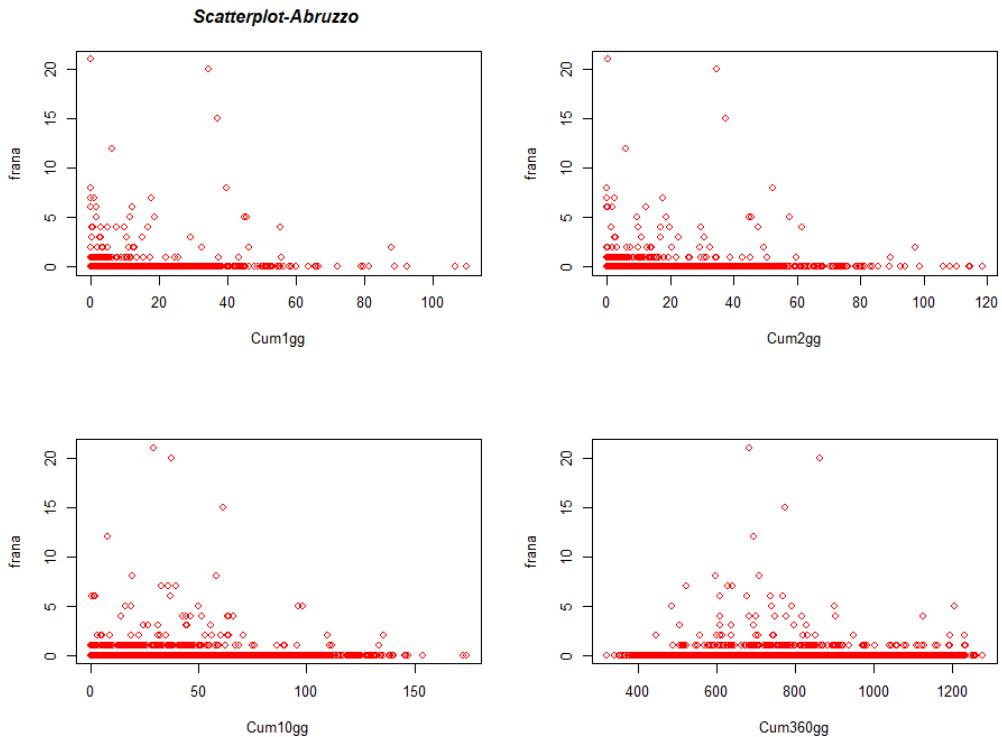


Figura 25: scatterplot-Abruzzo

A primo impatto potrebbe sembrare che esso esprima una funzione esponenziale, ma dopo aver approssimato i punti di tale grafico con una funzione esponenziale, si è evinto che l'altissima assenza di fenomeni franosi impatta tale approssimazione al punto da non permettere la riuscita della curva esponenziale. Infatti, essa si riduce ad una retta interpolante proprio quei valori pari a 0. In tal caso, si è calcolato il coefficiente di determinazione, per capire se ci fosse qualche altro tipo di correlazione non lineare tra le variabili, ma come si è già immaginato, non c'è alcun tipo di correlazione.

Di seguito, si mostrano gli stessi grafici riportati sopra, su cui si è provato ad interpolare i punti con una funzione esponenziale:

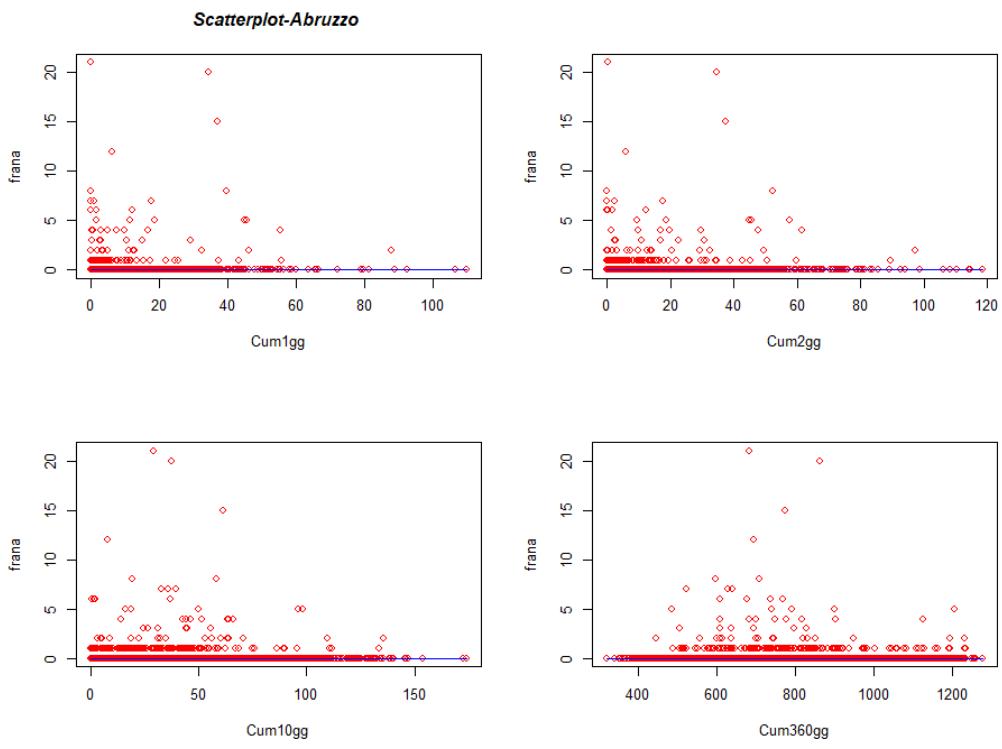


Figura 26: interpolazione esponenziale-Abruzzo

Come si può facilmente notare, la curva esponenziale non riesce ad interpolare bene a causa della forte presenza del valore 0. Questo problema accade nello stesso modo anche per la regione Campania (figure 27,28):

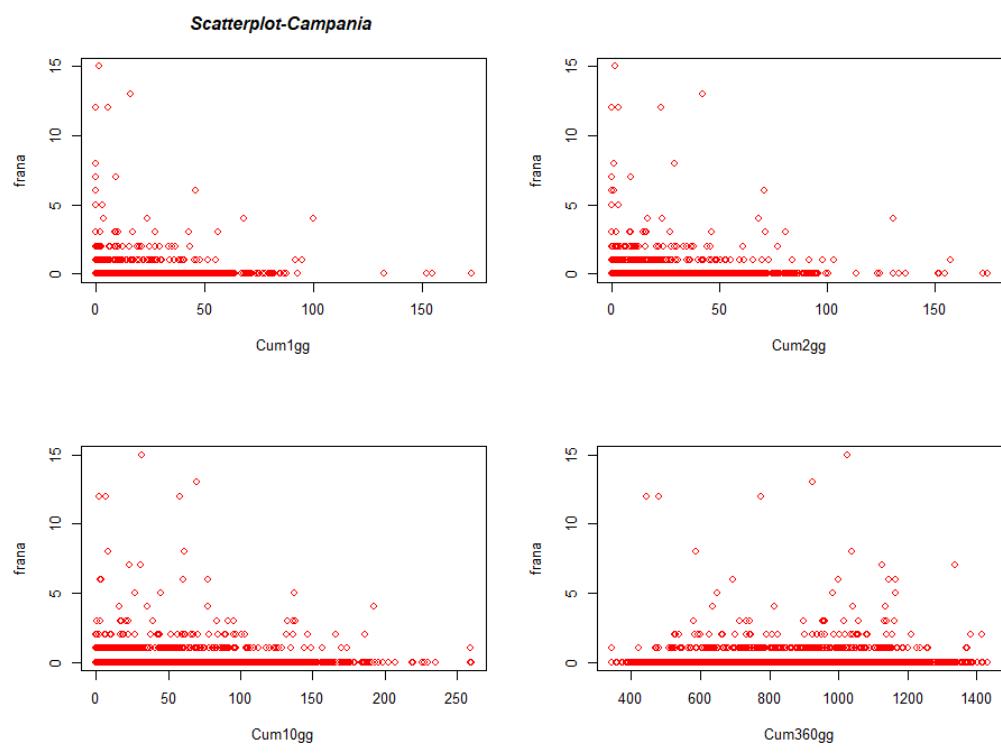


Figura 27: scatterplot-Campania

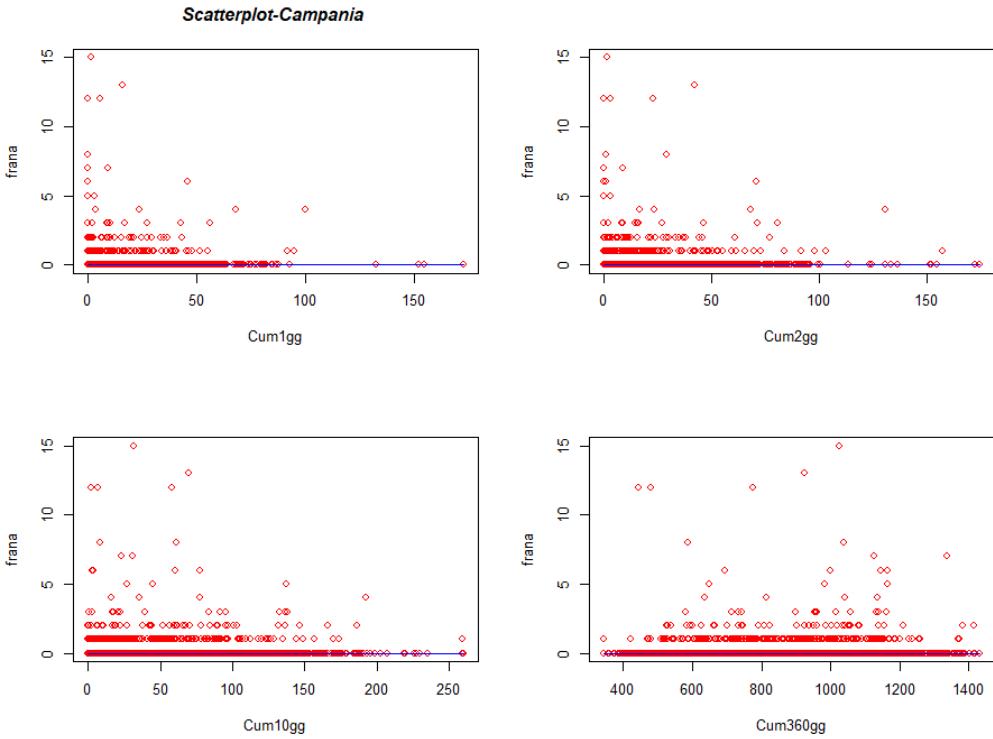


Figura 28: interpolazione esponenziale-Campania

Per ovviare a questo problema, si dovrebbero eliminare molti fenomeni non fransosi. Questa soluzione non è possibile, in quanto dovrebbero essere eliminati quasi la totalità degli eventi non fransosi, il che non sarebbe più reale. Tramite ciò si dimostra che la regressione non va bene per questo tipo di problema.

6.4 PREPARAZIONE DEL SET DI ADDESTRAMENTO E DEL SET DI TEST

Come scritto in precedenza, si affronta il problema come una classificazione binaria. Il primo passo da eseguire per applicare algoritmi di machine learning è quello di preparare i dati. Dalle immagini relative agli attributi predittivi sia dell'Abruzzo che della Campania, ci si è resi conto che il numero di eventi non fransosi copre del tutto quelli fransosi, soprattutto al di sotto una certa soglia pluviometrica. Nella classificazione questo è un problema, in quanto le classi non sono ben equilibrate, dal grafico, infatti, si nota la totale copertura dei fenomeni fransosi causata dal numero eccessivo dei fenomeni non fransosi. Tuttavia, e sempre in accordo con gli ingegneri del dipartimento di ingegneria civile, si è stabilita una soglia di pioggia sotto la quale eliminare tutti i fenomeni non fransosi, in quanto questi fenomeni risultavano irrilevanti

nella classificazione che andremo ad esaminare. La soglia pluviometrica in questione è quella al di sopra di 20 mm, quindi, al di sotto di tale soglia tutti i fenomeni non franosi sono stati eliminati. Come conseguenza di ciò, i datasets risultanti dall'elaborazione descritta nel capitolo precedente sono stati ridotti. Per affrontare il problema utilizzando la classificazione binaria è stato aggiunto ai datasets relativi alle due regioni un attributo che indica la classe di appartenenza:

- 0 = non frana;
- 1 = frana.

È stato osservato che per quanto riguarda la Campania vi sono:

- 23053 istanze appartenenti alla classe 0;
- 323 istanze appartenenti alla classe 1.

Per l'Abruzzo la situazione risulta essere molto simile.

In questo modo si sono ottenuti per quanto riguarda la Campania le seguenti istanze:

- 1549 istanze appartenenti alla classe 0;
- 323 istanze appartenenti alla classe 1.

Per l'Abruzzo si hanno:

- 1181 istanze appartenenti alla classe 0;
- 220 istanze appartenenti alla classe 1.

Di seguito, sono riportati i grafici che riguardano la regione Campania e la regione Abruzzo:

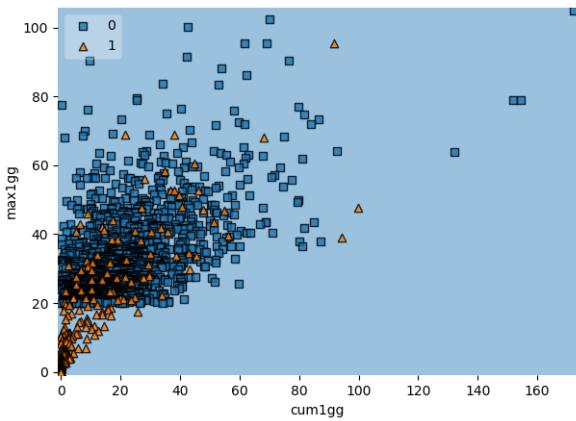


Figura 29: classificazione-Campania

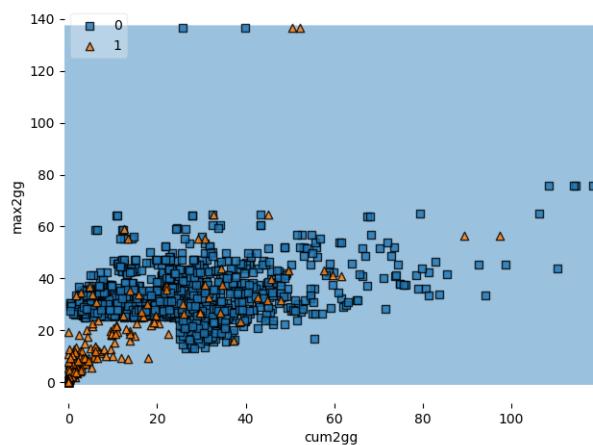


Figura 30: classificazione-Abruzzo

Dopo aver preparato i dati per il machine Learning, bisogna scegliere un algoritmo di apprendimento che deve comportarsi bene non solo sul dataset di addestramento ma anche sui nuovi dati, quindi è stato necessario suddividere il dataset, per ogni regione, in 2 nuovi set. È stato utilizzato il set di addestramento per informare e ottimizzare il modello di apprendimento automatico, mentre quello di test viene utilizzato per valutare il modello. Per effettuare tale suddivisione, è stata utilizzata scikit-learn e la funzione train test split. Bisogna tenere in considerazione che dividendo il dataset si stanno sottraendo informazioni preziose di cui l'algoritmo di apprendimento potrebbe fare buon uso. Per tanto l'obiettivo è quello di non allocare troppe informazioni nel set di test. Tuttavia, più piccolo è il set di test, più sarà imprecisa la stima dell'errore di generalizzazione. La divisione di un dataset quindi deve valutare questi compromessi.

Le divisioni più comunemente utilizzate sono 70:30, 80:20.

In base alla dimensione dei dataset a disposizione, si è deciso di riservare il 30% dei dati per il set test e il 70% dei dati per il set di addestramento. Successivamente, nell'analisi delle prestazioni del modello sarà illustrato il numero di campioni utilizzato per ogni set.

6.5 DEFINIZIONE DEI MODELLI

La libreria scikit-learn fornisce molteplici algoritmi di apprendimento automatico con lo scopo di risolvere problemi differenti. Ogni algoritmo di classificazione ha i suoi difetti e nessun modello di classificazione può vantare superiorità assoluta. Nella pratica è opportuno confrontare un buon numero di algoritmi differenti, in modo da addestrarli e selezionare poi il modello che offre le migliori prestazioni. Per il problema in esame, osservando la suddivisione in classi rappresentata nei grafici illustrati nel paragrafo precedente, si può notare che la suddivisione tra i due datasets non risulta essere del tutto lineare e quindi, come ci si aspettava, applicando i classificatori che funzionano molto bene nei casi di linearità si sono avuti scarsissimi risultati. Si è dovuto escludere, quindi l'algoritmo perceptron che non converge mai su dataset che non sono perfettamente separabili in modo lineare e anche la nota regressione logistica che risulta avere il medesimo problema.

Dopo un'opportuna analisi si è deciso di scegliere come modelli potenzialmente adatti la Foresta Casuale (Random Forest) e l'algoritmo Macchine a Vettori di Supporto con kernel gaussiano (Radial Basis Function).

6.6 FORESTA CASUALE

Le foreste casuali hanno acquisito grande popolarità nelle applicazioni di machine learning grazie alle loro buone prestazioni di classificazione, alla loro scalabilità e alla loro facilità d'uso. Una foresta casuale può essere considerata come un insieme di alberi decisionali. L'idea è quella di combinare più sistemi di apprendimento deboli per costruire un modello più robusto, un sistema di apprendimento forte che offre un migliore errore di generalizzazione e sia meno suscettibile ai problemi di overfitting. Si ha overfitting quando il modello generalizza bene sui dati di addestramento ma non sui dati di test. [3]

L'algoritmo a Foresta Casuale consta di quattro semplici passi:

- estrapolare un campione casuale iniziale di dimensioni n;
- far crescere un albero decisionale dal campione. Per ogni nodo:
 1. selezionare casualmente d caratteristiche senza reinserimento;
 2. suddividere il nodo utilizzando la caratteristica che fornisce la migliore suddivisione sulla base della funzione obiettivo, per esempio massimizzando il guadagno informativo;
 3. ripetere per k volte i passi 1 e 2;
 4. aggregare le previsioni di ciascun albero per assegnare l'etichetta della classe sulla base di un voto a maggioranza.

Un grande vantaggio delle foreste casuali è che non bisogna preoccuparsi troppo della scelta degli iperparametri. In generale, non bisogna potare la foresta casuale, poiché nel suo insieme il modello è resistente al rumore rispetto ai singoli alberi decisionali. L'unico parametro di cui bisogna preoccuparsi è il numero di alberi k. Tipicamente maggiore è il numero di alberi, migliori saranno le prestazioni del classificatore a foresta casuale, con lo svantaggio di un incremento del costo computazionale. Scikit-learn offre un'ottima implementazione di tale modello, ossia RandomForestClassifier, dove le dimensioni del campione casuale vengono scelte in modo che siano uguali al numero di campioni del set di addestramento originario, il che rappresenta un buon modo per tenere sotto controllo l'overfitting. Per il problema in esame, è stata utilizzata l'implementazione offerta da scikit-learn in cui i parametri principali che abbiamo provveduto a definire sono i seguenti:

- n_estimators: indica il numero di alberi nella foresta casuale;

- criterion: funzione utilizzata per misurare la qualità di una divisione. I criteri supportati sono “gini” ed “entropia”;
- maxDepth: indica la massima profondità dell’albero.

I parametri che risultano essere ottimali sono i seguenti:

- n_estimators=10 e quindi vengono realizzati 10 alberi decisionali;
- criterio= gini;
- maxDepth=3.

L’ottimizzazione dei parametri è stata effettuata utilizzando la ricerca a griglia che consente di migliorare le prestazioni del modello, ricercando la combinazione ottimale dei valori degli iperparametri. Tale approccio è piuttosto semplice: si tratta di una ricerca esaustiva a forza bruta, nella quale viene specificato un elenco di valori per i vari iperparametri e vengono valutate le prestazioni del modello per ogni combinazione, fino a ottenere un set ottimale. In python, la funzione corrispondente a tale approccio è GridSearchCV. Per ottenere i migliori parametri viene utilizzato l’attributo bestEstimator. In conclusione, pur utilizzando un numero di alberi decisionali non molto basso, le prestazioni del modello risultano essere molto soddisfacenti.

6.7 MACCHINE A VETTORI DI SUPPORTO CON KERNEL GAUSSIANO

Osservando i dati, si può notare che non risultano avere una separazione del tutto lineare, quindi si è deciso di applicare anche un modello che possa separare i dati in modo non lineare. [3] L’algoritmo generale è la macchina a vettori di supporto (SVM) il cui obiettivo di ottimizzazione consiste nel massimizzare il margine. Per margine s’intende la distanza fra il confine decisionale e i campioni di addestramento che sono più vicini a questo confine. La logica che porta ad individuare confini decisionali con ampi margini è il fatto che questi tendono ad avere un errore di generalizzazione più basso, mentre i modelli con margini più ridotti sono più soggetti ad overfitting.

L’idea su cui si basano i metodi kernel per gestire dati che non sono separabili linearmente consiste nel creare combinazioni non lineari delle caratteristiche originali per proiettarle in uno spazio con maggiori dimensioni tramite una funzione di mappatura, dove tali dati vengono separati più facilmente, come si può notare nell’esempio illustrato:

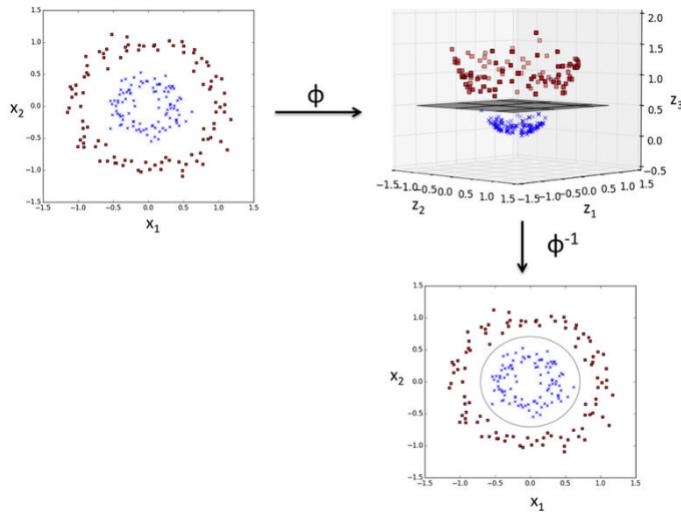


Figura 31: funzione di mappatura del kernel RBF

Effettuando la proiezione tramite una funzione di mappatura, le classi risultano essere ben distinte. Il problema di questo approccio è il fatto che la costruzione delle nuove caratteristiche è molto costosa dal punto di vista computazionale. Per risolvere questo problema viene, quindi, utilizzata una funzione kernel. La funzione utilizzata per il problema in esame è la Radial Basis Function, o RBF kernel, che consente di separare i dati in modo non lineare.

La funzione è definita nel seguente modo:

$$k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp(-\gamma \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2)$$

Sostanzialmente, il termine kernel può essere interpretato come una funzione di similarità fra una coppia di campioni. La misura della distanza viene convertita in un punteggio di similarità e a causa del termine esponenziale il punteggio di similarità risultante rientrerà sempre in un intervallo compreso fra 1 (campioni esattamente simili) e 0 (campioni dissimili).

Il parametro da ottimizzare è γ ; scegliendo questo valore basso ad esempio 0.1, il confine decisionale sarà relativamente morbido.

Incrementando troppo questo valore, il modello si adatterà meglio al dataset di addestramento ma ci sarà con molta probabilità un elevato errore di generalizzazione sui dati in arrivo. Quindi, l'ottimizzazione di questo valore gioca un ruolo importante nel controllo dell'overfitting.

Scikit-learn fornisce la seguente funzione per applicare la RBF kernel:

`SVC(kernel, C, gamma)`

Nella funzione i parametri principali da definire sono:

- kernel da applicare;

- il parametro C che ha il compito di controllare la penalizzazione per un'errata classificazione. Valori ampi di C corrispondono a grandi penalizzazioni di un errore, se si scelgono valori di C più ridotti si è meno rigidi nei confronti degli errori di errata classificazione;
- il parametro gamma già definito prima.

Per il nostro caso specifico sono stati individuati come parametri ottimali i seguenti:

- $C=1$;
- $\gamma=0.1$;

Utilizzando questi parametri si evita di avere situazioni di overfitting.

Nel paragrafo successivo verranno illustrati i risultati ottenuti dai due classificatori definiti.

RISULTATI

Per valutare le prestazioni di un modello, è stata senz'altro rappresentata la matrice di confusione, ovvero una matrice 2×2 (per la classificazione binaria) che rileva il conteggio dei veri positivi, dei veri negativi, dei falsi negativi e falsi positivi. In Python, è possibile calcolare tale matrice utilizzando la funzione `confusionMatrix`. Altri parametri da considerare per valutare le prestazioni del modello possono essere ottenuti attraverso la funzione `classificationReport`. Essa darà in output i seguenti valori [2]:

- Accuratezza (accuracy): misura di performance più intuitiva che indica il rapporto tra l'osservazione correttamente prevista e le osservazioni totali. Si potrebbe pensare che avendo un'alta accuratezza il modello risulta essere migliore. In realtà l'accuratezza è una grande misura se si hanno set di dati simmetrici in cui i valori dei falsi negativi e falsi positivi sono quasi gli stessi. Pertanto, è necessario esaminare anche altri parametri per valutare le prestazioni del modello. Più formalmente l'accuratezza è definita in questo modo: $\frac{TP+TN}{TP+FP+FN+TN}$, dove TP: veri positivi, TN: veri negativi, FP: falsi positivi, FN: falsi negativi;
- Precisione (precision): indica il rapporto tra le osservazioni positive previste correttamente e le osservazioni positive totali previste. Un'alta precisione sta ad indicare un basso tasso di falsi positivi, in quanto essa risulta: $\frac{TP}{TP+FP}$;
- Richiamo (recall): indica la percentuale di positivi riconosciuti correttamente, ossia: $\frac{TP}{TP+FN}$;
- Punteggio F1 (F1-score): indica la media ponderata della precisione e del richiamo. Questo valore tiene conto sia dei falsi positivi che dei falsi negativi: $\frac{2 * (\text{richiamo} * \text{precisione})}{(\text{richiamo} + \text{precisione})}$;
- Supporto (support): indica il numero di istanze appartenente ad ogni classe.

Una volta definiti i parametri da analizzare, vengono mostrati di seguito i risultati ottenuti per quanto riguarda la Campania sia per il set di test che per il set di addestramento utilizzando il modello di FORESTA CASUALE:

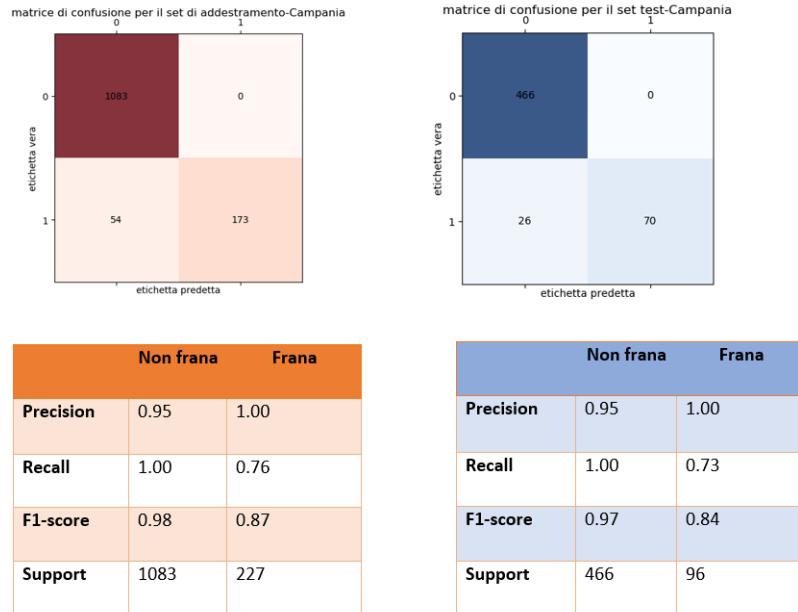


Figura 32: Risultati Foresta Casuale-Campania

Per entrambi gli insiemi, otteniamo un'accuratezza pari a 0.95 e il tempo impiegato per la fase di addestramento è pari a 0.04 secondi. Dai valori ottenuti, si può evincere che nel set di addestramento e nel set di test il modello riesce a classificare perfettamente le istanze appartenenti alla classe non frana e quindi non sono presenti falsi positivi, ma commette degli errori nel classificare le frane. Tuttavia, nel problema in esame, gli attributi predittivi prendono in considerazione soltanto le piogge, quindi questo risultato risulta essere molto soddisfacente. Confrontando i risultati tra i due insiemi, si può notare che il modello generalizza bene sia sui dati di addestramento che sui dati di test. Riportiamo le prestazioni relative alla Campania applicando l'algoritmo SVM con il kernel gaussiano:

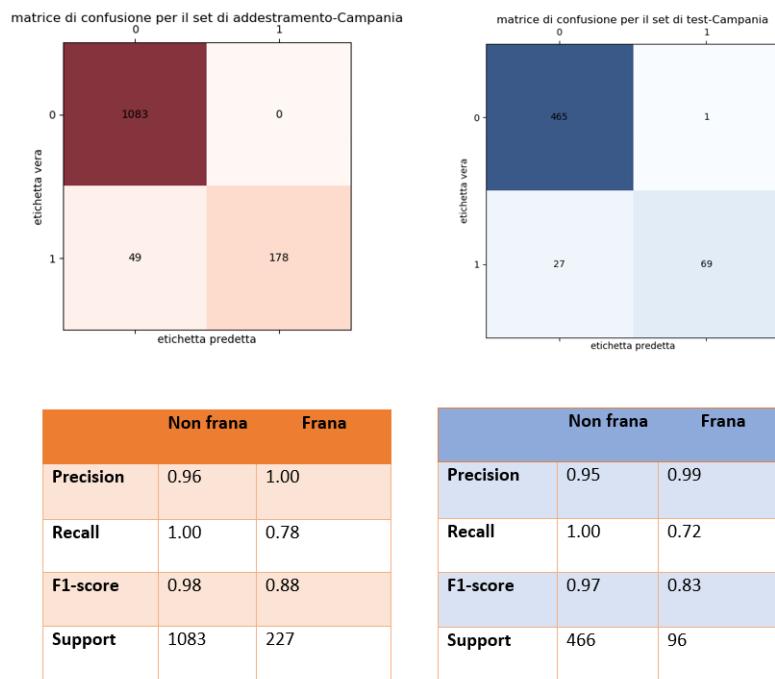
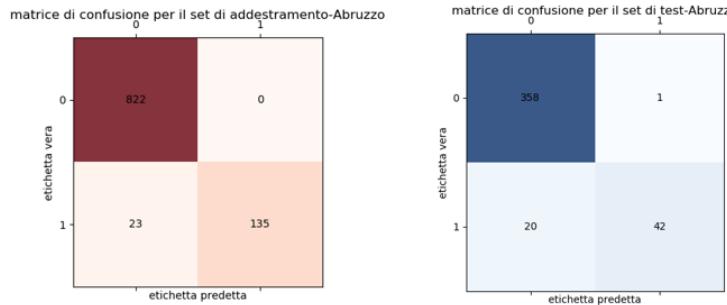


Figura 33: Risultati SVM con kernel RBF-Campania

Otteniamo un'accuratezza pari a 0.96 per il set di addestramento e 0.95 per il set di test. Il tempo di addestramento è pari a 0.02 secondi. Effettuando il confronto con le prestazioni ottenute con il modello precedente risultano esserci pochissime differenze, tuttavia si evince che l'F1-score risulta essere leggermente migliore per quanto riguarda il set di test con il modello a foresta casuale, ma l'SVM ha prestazioni migliori sul set di addestramento. Per quanto riguarda il tempo di addestramento risulta essere molto basso per entrambi i classificatori.

Di seguito si riportano le prestazioni ottenute con il modello a foresta casuale per quanto riguarda l'Abruzzo:



		Non frana	Frana
Precision	0.97	1.00	
Recall	1.00	0.85	
F1-score	0.99	0.92	
Support	822	158	

		Non frana	Frana
Precision	0.95	0.98	
Recall	1.00	0.68	
F1-score	0.97	0.80	
Support	359	62	

Figura 34: Risultati Foresta Casuale-Abruzzo

Per il set di addestramento otteniamo un'accuratezza pari a 0.97 invece per il set di test risulta essere pari a 0.95 e un tempo di addestramento pari a 0.007 secondi. Il dataset relativo all'Abruzzo risulta avere meno campioni appartenenti alla classe frana rispetto a quelli che troviamo nel dataset relativo alla Campania. Il modello generalizza molto bene sul dataset di addestramento ma commette comunque qualche errore sul set di test infatti è presente una recall pari a 0.68. Nel set di test infatti il modello restituisce 20 falsi negativi su 62 veri negativi.

Si mostrano le prestazioni ottenute usando l'SVM con kernel gaussiano:

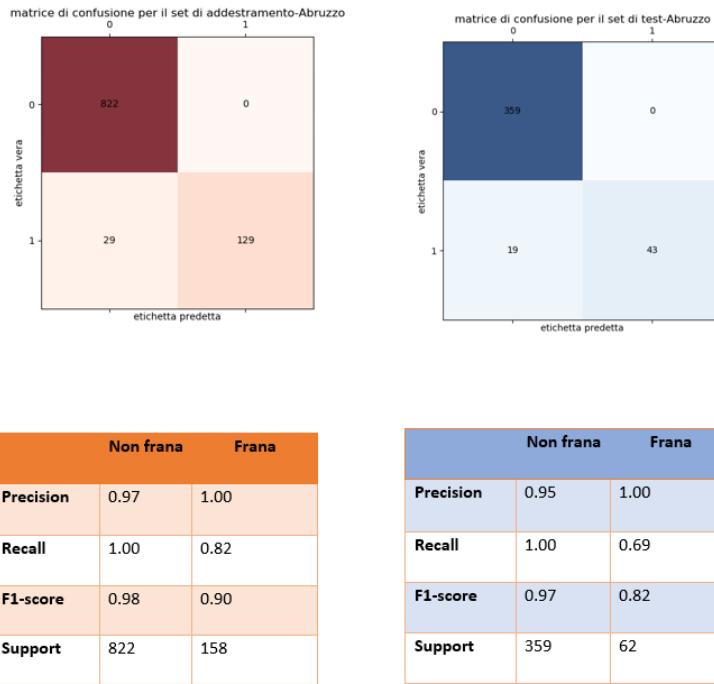


Figura 35: Risultati SVM con kernel RBF-Abruzzo

Per il dataset di addestramento otteniamo un'accuratezza pari a 0.97 e per il test di set pari a 0.95 con un tempo di addestramento pari a 0.006. Notiamo che come per l'algoritmo di foresta casuale, anche qui il modello restituisce un certo numero di falsi negativi, infatti possiamo osservare che la recall è pari a 0.69.

In ogni caso, le prestazioni si possono considerare soddisfacenti. In conclusione, tra i due modelli applicati su entrambi i datasets relativi alla Campania e all'Abruzzo, sono poche le differenze che risultano tra i modelli in termini di prestazioni. Dato che l'obiettivo principale è quello di ottenere il minor numero di falsi negativi relativi alle frane, il classificatore SVM con kernel gaussiano è più adatto. Ciò non esclude l'utilizzo del classificatore Foresta Casuale per questo tipo di problemi.

8

CONCLUSIONI

Questo progetto è stato sviluppato per prevedere i fenomeni franosi nelle zone di allerta delle regioni Abruzzo e Campania, considerando come causa di innescio soltanto la pioggia.

I classificatori usati per questo problema sono in grado di prevedere se un evento sarà franoso o non franoso in modo molto soddisfacente, anche avendo a disposizione soltanto la pioggia come attributo predittivo. Per i dati che si hanno, quindi dati che non presentano una suddivisione netta, i classificatori sono in grado di svolgere bene il loro lavoro.

Si è scelto come miglior classificatore l'SVM con kernel gaussiano, in quanto predice meno falsi negativi rispetto alla Foresta Casuale. Chiaramente, in problemi come questo, è sempre meglio predire un falso positivo che un falso negativo, quindi prevedere una frana anche se essa effettivamente non ci sarà. Non è stato possibile ridurre a 0 i falsi negativi a causa della mancanza di altri dati di tipo morfologico innescanti le frane.

I risultati ottenuti dai nostri sviluppi sono stati confermati soddisfacenti anche dagli ingegneri del dipartimento di ingegneria civile dell'università degli studi di Salerno.

Il motivo per cui non è mostrata la classificazione fra i fenomeni non franosi e gli eventi areali (due o più frane) è il seguente:

A parte il numero ridotto di eventi areali, sorge la problematica in cui tali fenomeni risultano essere coperti da fenomeni non franosi, i quali non possono essere eliminati, in quanto si verificano quando le piogge sono maggiori di 20 e 30. Di questi fenomeni areali pochi vengono classificati come veri positivi. La forte presenza di fenomeni non franosi tende a far fallire i classificatori che abbiamo usato (e anche altri). Siccome lo scopo è quello di evitare falsi negativi il più possibile, si è scelto di non presentare le soluzioni avutesi dai classificatori e, quindi, di non affinare il modello con tale classificazione.

9

SVILUPPI FUTURI

Potrà esserci la possibilità di estendere questo progetto considerando anche altri fattori morfologici per la previsione di fenomeni franosi, anche per le altre regioni italiane.

RIFERIMENTI BIBLIOGRAFICI

- [1] Arpa, *Definizione delle soglie pluviometriche di innescio delle frane.*
- [2] Sebastian Raschka, *Machine Learning con Python. Costruire algoritmi per generare conoscenza.* Apogeo editori
- [3] Aurèlien Géron, *Hands-On Machine Learning with Scikit-Learn TensorFlow.* O'Reilly