Analisi dei test effettuati

Test di Stefano Barrella su F-MNIST con alberi decisionali

Alberi Decisionali

DEFINIZIONE:

L'albero decisionale è un algoritmo di tipo supervisionato usato nel ML.
È un modello predittivo che può essere usato sia per casi di classificazione sia per casi di regressione. Ciò significa che l'output che andrà a prevedere il modello sarà una variabile categorica (per esempio una binaria SI/NO) oppure una quantità continua.
Gli alberi decisionali sono gli algoritmi più comuni perché sono molto veloci e sopratutto semplici da interpretare. Questo è fondamentale perché in alcuni settori in cui si può preferire una maggiore semplicità di comprensione ad una maggiore accuratezza del modello.

FUNZIONAMENTO

In questo algoritmo i dati che io do come input vengono continuamente splittati in base a determinati criteri. Due concetti chiave per comprender il loro funzionamento sono i nodi e le foglie.
- I nodi: sono i luoghi in cui in base a certe regole i dati vengono splittati;
- Le foglie: sono i risultati intermedi o finali ossia i luoghi in cui finiscono i dati una volta splittati

Ma in base a quali criteri i dati vengono separati?

PRINCIPALI IPER-PARAMETRI DEL MODELLO

Vediamo quali sono (per la mia esperienza) i principali parametri o iper-parametri di un modello Decision Tree, prendendo in considerazione l'implementazione della libreria scikit-learn in Python.

Tra i parametri da tenere fortemente in considerazione c'è la metrica di splitting, che, misurando la qualità dello split, separerà i dati. Questo parametro si chiama **criterion** e dipende dal task che si vuole svolgere.
Per quanto riguarda il problema della classificazione, può avere "gini" e "entropy" come possibili valori. "Gini" fa riferimento all'impurità di Gini, mentre "entropy" all'informazione gain, basato sul concetto di entropia nella teoria dell'informazione.

Un altro parametro da controllare è quello che si chiama **max_depth**: questo indica quanto profondo sarà l'albero. Più profondo sarà l'albero, più split farà e più sarà preciso sui dati di addestramento. Questo però non vuol dire che se è più preciso sui dati di train il modello sarà in grado di generalizzarli bene, anzi più sarà profondo più sarà profondo più preciso sarà sui dati di train e mono probabilmente sarà in grado di generalizzare su dati nuovi.
Una maggior profondità quindi rischia di far cadere in quel fenomeno chiamato overfitting ovvero un'ottima capacità sul train ma una pessima sul test che si traduce in un incapacità di generalizzare su nuovi dati.

Un altro parametro che può portare overfitting è **min_samples_leaf**, ovvero quanti esemplari minimi di osservazioni possono finire nelle foglie finali. Un numero basso vorrà dire anche in questo caso una maggior precisione sul train, che potrà portare di nuovo ad una scarsa accuratezza sul test. Questo perché si andranno ad identificare i casi più specifici e non quelli generali.

Altro iper-parametro importante è **max_features**. Questo parametro ci da la possibilità di scegliere in numero di variabili da utilizzare per trovare il best split. È inizializzato a None quindi verrà cercato su tutte le variabili o sulla radice quadrata del numero delle variabili (se ne ho 16, lo cerco su 4)

PRO E CONTRO DELL'ALBERO DECISIONALE

PRO
- Molto semplice da capire e interpretare;
- Può lavorare su dati numerici e categorici contemporaneamente;
- Non richiede moltissimi dati;
- Performa bene sia sui pochi dati che su grandi dataset

CONTRO
- Non è un modello robusto, quindi un piccolo cambiamento nel training può comportare un gradissimo cambiamento nella prediction
- Altissimo rischio di overfitting

TEST EFFETTUATI (STEFANO BARRELLA)

F-MNIST

Eseguendo lo script per l'ottimizzazione degli iper parametri salta all'occhio che nelle prime 10 posizioni delle simulazioni effettuate con i migliori responsi in termini di accuratezza troviamo principalmente 4 tipi di profondità: (11, 13, 15, 17)
Testiamo anche la profondità limite scelta che è 21 e poi testiamo una profondità che sfora il range prestabilito ossia 25.

1° Blocco di test (Profondità crescente, c: gini, mf:None, msl:1)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Gini	None	1	86,85%	80,50%	0,92	NO
2	**13**	**Gini**	**None**	**1**	**90,71%**	**81,11%**	**1,27**	**NO**
3	15	Gini	None	1	94,35%	80,72%	1,78	SI
4	17	Gini	None	1	96,76%	80,39%	2,15	SI
5	21	Gini	None	1	98,88%	79,69%	2,45	SI
6	25	Gini	None	1	99,46%	79,57%	2,56	SI

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più lentamente quella su set di test che mantiene sempre una differenza di circa 9 punti percentuali, a partire dalla profondità di 15 l'accuratezza sul set di test inizia addirittura a calare avviandosi verso l'overfitting. Il miglior risultato ottenuto è il **test 2** in termini di percentuale di accuratezza più alta sul set di test, tuttavia è da considerare ottimo anche il **test 1** in cui la differenza tra l'errore sul set di train e l'errore sul set di test è davvero minima. Possiamo notare inoltre che, osservando i test di Stefano Biddau riguardanti il dataset MNIST con le stesse condizioni e stessi parametri, i valori sono migliori sul MNIST in quanto il dataset F-MNIST è più complesso.

2° Blocco di test (Profondità crescente, c: gini, mf:None, msl:25)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Gini	None	25	84,18%	80,32%	0,47	NO
2	**13**	**Gini**	**None**	**25**	**85,65%**	**80,64%**	**0,63**	**NO**
3	15	Gini	None	25	86,35%	80,48%	0,75	NO
4	17	Gini	None	25	86,54%	80,53%	0,76	NO
5	21	Gini	None	25	86,62%	80,51%	0,77	NO
6	25	Gini	None	25	86,62%	80,52%	0,77	NO
7	30	Gini	None	25	86,62%	80,51%	0,77	NO
8	35	Gini	None	25	86,62%	80,51%	0,77	NO

Per questi 8 test effettuati otteniamo degli ottimi risultati. Innanzitutto abbiamo trovato una combinazione di parametri che dopo alcuni valori altalenanti, si stabilizza sia sul set di train che sul set di test, che oltretutto mostrano avere dei livelli di accuratezza in termini di percentuale molto vicini. Questa cosa ci fa presumere l'assenza di overfitting tutta via, visto che il valore si è stabilizzato con una percentuale minore rispetto al nostro miglior risultato, non si esclude che l'accuratezza possa tornare in discesa. Il miglior risultato ottenuto quindi è il **test 2** in termini di percentuale di accuratezza più alta sul set di test. Possiamo notare inoltre che rispetto al dataset MNIST testato da Stefano Biddau, il valore non si stabilizza sulla percentuale con il migliori risultato.

3° Blocco di test (Profondità crescente, c: gini, mf:None, msl:50)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Gini	None	50	82,79%	79,81%	0,35	NO
2	13	Gini	None	50	83,68%	80,01%	0,50	NO
3	**15**	**Gini**	**None**	**50**	**83,90%**	**80,07%**	**0,50**	**NO**
4	17	Gini	None	50	84,02%	80,03%	0,51	SI
5	21	Gini	None	50	83,99%	79,99%	0,51	SI
6	25	Gini	None	50	83,99%	79,99%	0,51	SI

Per questi 6 test effettuati abbiamo trovato una combinazione di parametri che si stabilizza sia sul set di train che sul set di test, che oltretutto mostrano una differenza di 4 punti percentuali dei livelli di accuratezza in termini di percentuale. Il miglior risultato ottenuto quindi è il **test 3** in termini di percentuale di accuratezza più alta sul set di test. Possiamo notare inoltre che rispetto al dataset MNIST testato da Stefano Biddau, il valore non si stabilizza sulla percentuale con il migliori risultato ma decresce avviandosi verso l'overfitting.

4° Blocco di test (Profondità crescente, c: gini, mf:0,5, msl:1)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Gini	0,5	1	86,73%	81,46%	0,73	NO
2	13	Gini	0,5	1	90,07%	80,76%	1,28	NO
3	15	Gini	0,5	1	94,08%	81,17%	1,71	NO
4	**17**	**Gini**	**0,5**	**1**	**96,24%**	**81,25%**	**1,84**	**NO**
5	21	Gini	0,5	1	98,83%	80,03%	2,5	SI
6	25	Gini	0,5	1	99,52%	79,49%	2,6	SI

Per questi 6 test effettuati abbiamo trovato una combinazione di parametri che si stabilizza sia sul set di train che sul set di test, che oltretutto mostrano una differenza di circa 10 punti percentuali dei livelli di accuratezza in termini di percentuale. Il miglior risultato ottenuto quindi è il **test 4** in termini di percentuale di accuratezza più alta sul set di test. Possiamo notare inoltre che rispetto al blocco di test con il parametro mf = None , questa casistica di test è migliore in accuratezza in termini di percentuale e l'overfitting parte con una profondità pari a 21 mentre nel blocco precedente partiva la discesa da una profondità pari a 15.

5° Blocco di test (Profondità crescente, c: gini, mf:0,5, msl:25)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Gini	0,5	25	84,01%	80,09%	0,45	NO
2	13	Gini	0,5	25	85,52%	80,40%	0,58	NO
3	15	Gini	0,5	25	86,05%	80,18%	0,73	NO
4	**17**	**Gini**	**0,5**	**25**	**86,30%**	**80,91%**	**0,65**	**NO**
5	21	Gini	0,5	25	86,26%	80,38%	0,69	SI
6	25	Gini	0,5	25	86,27%	80,27%	0,74	SI

Per questi 6 test effettuati abbiamo trovato una combinazione di parametri che dopo alcuni valori altalenanti, si stabilizza sia sul set di train che sul set di test, che mostrano avere circa 6 punti percentuali di differenza nei livelli di accuratezza in termini di percentuale molto vicini. Il miglior risultato ottenuto quindi è il **test 4** in termini di percentuale di accuratezza più alta sul set di test. Possiamo notare inoltre che rispetto al blocco di test con il parametro mf = None , questa casistica di test rispetto alla precedente vai in overfitting.

6° Blocco di test (Profondità crescente, c: gini, mf:0,5, msl:50)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Gini	0,5	50	82,47%	79,66%	0,32	NO
2	13	Gini	0,5	50	83,68%	80,22%	0,50	NO
3	15	Gini	0,5	50	83,83%	79,87%	0,53	SI
4	17	Gini	0,5	50	83,91%	79,79%	0,50	SI
5	**21**	**Gini**	**0,5**	**50**	**83,81%**	**80,53%**	**0,40**	**NO**
6	25	Gini	0,5	50	83,88%	80,36%	0,48	SI

Per questi 6 test effettuati abbiamo trovato una combinazione di parametri che dopo alcuni valori altalenanti, in un modo del tutto diverso dal solito, trova la soluzione ottima dopo alcuni casi di overfitting. Il miglior risultato ottenuto quindi è il **test 5** in termini di percentuale di accuratezza più alta sul set di test. Possiamo notare inoltre che rispetto al blocco di test con il parametro mf = None , questa casistica di test rispetto alla precedente non va in una netta discesa di overffitting.

7° Blocco di test (Profondità crescente, c: entropy, mf:None, msl:1)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	**11**	**Entropy**	**None**	**1**	**87,14%**	**81,53%**	**0,78**	**NO**
2	13	Entropy	None	1	91,11%	81,34%	1,32	SI
3	15	Entropy	None	1	94,71%	80,94%	1,84	SI
4	17	Entropy	None	1	97,40%	80,61%	2,20	SI
5	21	Entropy	None	1	99,61%	80,11%	2,50	SI
6	25	Entropy	None	1	99,98%	80,02%	2,54	SI

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più lentamente quella su set di test che non mantiene una differenza fissa di punti percentuali, a partire dalla profondità di 13 l'accuratezza sul set di test inizia addirittura a calare avviandosi verso l'overfitting. Il miglior risultato ottenuto è il **test 1** in termini di percentuale di accuratezza più alta sul set di test, Possiamo notare inoltre che, osservando i test di svolti con il valore c = Gini l'overfitting iniziava alla profondità 15 mentre qui lo riscontriamo prima.

8° Blocco di test (Profondità crescente, c: entropy, mf:None, msl:25)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	**11**	**Entropy**	**None**	**25**	**84,33%**	**81,23%**	**0,42**	**NO**
2	13	Entropy	None	25	85,89%	81,01%	0,64	NO
3	15	Entropy	None	25	86,66%	81,02%	1,84	NO
4	17	Entropy	None	25	86,88%	81,11%	0,77	NO
5	21	Entropy	None	25	86,92%	81,06%	0,78	SI
6	25	Entropy	None	25	86,92%	81,06%	0,78	SI

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più lentamente quella su set di test fino ad arrivare ad una situazione di stallo in overfitting. Il miglior risultato ottenuto è il **test 1** in termini di percentuale di accuratezza più alta sul set di test, Possiamo notare inoltre che, osservando i test di svolti con il valore c = Gini l'overfitting, nel test precedente non avevamo overfitting, mentre qui si.

9° Blocco di test (Profondità crescente, c: entropy, mf:None, msl:50)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Entropy	None	50	83,27%	80,70%	0,33	NO
2	13	Entropy	None	50	83,95%	80,61%	0,35	NO
3	**15**	**Entropy**	**None**	**50**	**84,23%**	**80,71%**	**0,36**	**NO**
4	17	Entropy	None	50	84,26%	80,69%	0,40	SI
5	21	Entropy	None	50	84,26%	80,69%	0,40	SI
6	25	Entropy	None	50	84,26%	80,69%	0,40	SI

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più lentamente quella su set di test fino ad arrivare ad una situazione di stallo in overfitting. Il miglior risultato ottenuto è il **test 3** in termini di percentuale di accuratezza più alta sul set di test, Possiamo notare inoltre che, osservando i test di svolti con il valore c = Gini, i due test sono molto simili, l'unica cosa che cambia è che qui si arriva prima ad una situazione di stallo.

10° Blocco di test (Profondità crescente, c: entropy, mf:0,5, msl:1)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	**11**	**Entropy**	**0,5**	**1**	**87,55%**	**81,49%**	**0,80**	**NO**
2	13	Entropy	0,5	1	90,98%	81,44%	1,25	SI
3	15	Entropy	0,5	1	94,90%	81,05%	1,78	SI
4	17	Entropy	0,5	1	97,36%	80,56%	2,20	SI
5	21	Entropy	0,5	1	99,81%	79,34%	2,50	SI
6	25	Entropy	0,5	1	99,97%	79,84%	2,60	NO
7	30	Entropy	0,5	1	100,00%	79,97	2,60	SI

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più lentamente ed altalenante quella su set di test. Il miglior risultato ottenuto è il **test 1** in termini di percentuale di accuratezza più alta sul set di test, Possiamo notare inoltre che, osservando i test di svolti con il valore c = Gini cambia che qui si arriva all'overfitting finale grazie alla percentuale del test di train con il 100% anche se il tempo di test si stava lentamente risollevando.

11° Blocco di test (Profondità crescente, c: entropy, mf:0,5, msl:25)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Entropy	0,5	25	84,28%	80,42%	0,45	NO
2	13	Entropy	0,5	25	85,31%	80,91%	0,55	NO
3	15	Entropy	0,5	25	86,42%	81,32%	0,58	NO
4	**17**	**Entropy**	**0,5**	**25**	**86,76%**	**81,72%**	**0,64**	**NO**
5	21	Entropy	0,5	25	86,78%	81,66%	0,62	SI
6	25	Entropy	0,5	25	86,78%	81,66%	0,62	SI

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più quella su set di test. Il miglior risultato ottenuto è il **test 4** in termini di percentuale di accuratezza più alta sul set di test, Possiamo notare inoltre che, osservando i test di svolti con il valore c = Gini cambia che qui si arriva all'overfitting e a una situazione di stallo.

12° Blocco di test (Profondità crescente, c: entropy, mf:0,5, msl:50)

Indice test	Profondità massima	Criterio	Max features	Min samples leaf	Acc Train	Acc Test	Differenza MSE	Over-fit
1	11	Entropy	0,5	50	83,10%	80,52%	0,33	NO
2	**13**	**Entropy**	**0,5**	**50**	**83,64%**	**81,05%**	**0,22**	**NO**
3	15	Entropy	0,5	50	83,89%	80,94%	0,28	NO
4	17	Entropy	0,5	50	84,09%	81,01%	0,42	NO
5	21	Entropy	0,5	50	84,10%	81,03%	0,35	NO
6	21	Entropy	0,5	50	84,10%	81,03%	0,35	NO

Da questi primi 6 test si evince che all'aumentare della profondità massima dell'albero decisionale aumenta la percentuale di accuratezza sul set di train, mentre aumenta più quella su set di test. Il miglior risultato ottenuto è il **test 2** in termini di percentuale di accuratezza più alta sul set di test, Possiamo notare inoltre che, osservando i test di svolti con il valore c = Gini cambia che qui non si arriva all'overfitting.

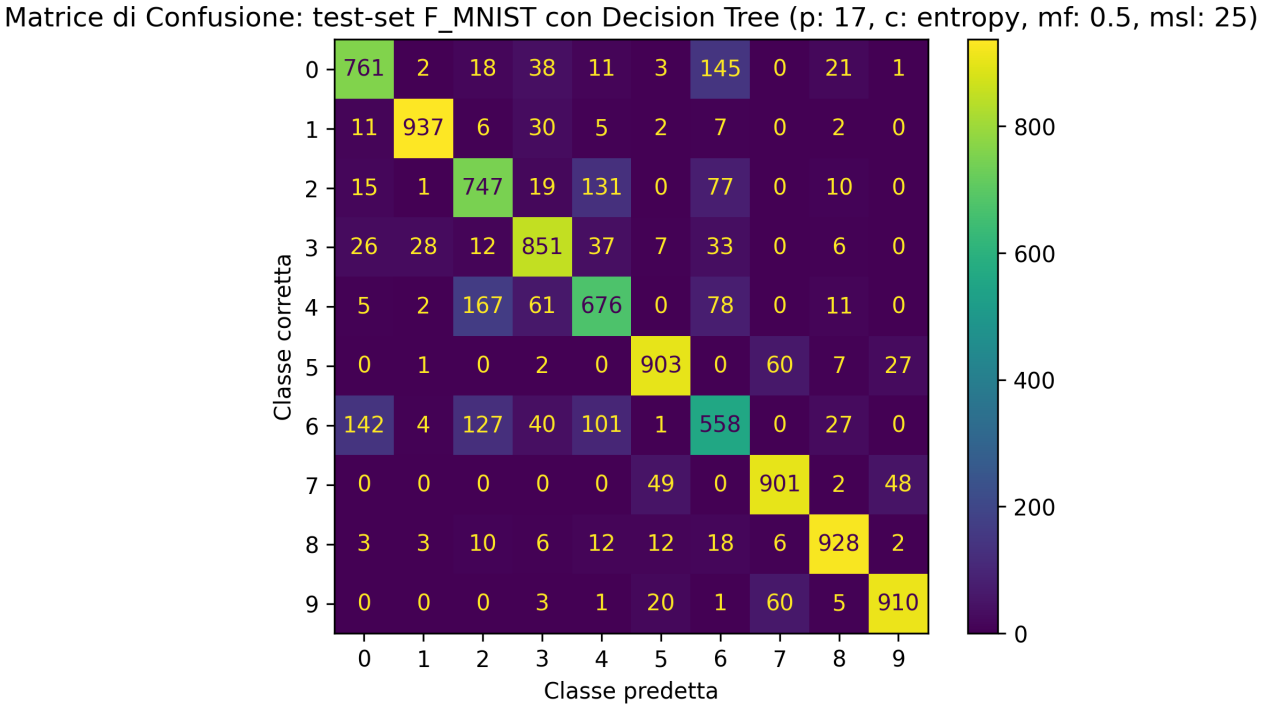
Analizzando i 12 casi di test effettuati sul dataset F-MNIST vediamo che i migliori casi risultano i seguenti:

- **Test del 2022-11-01 12:03:50 F_MNIST con Decision Tree (profondità: 17, c: entropy, mf: 0.5, msl: 25)**:

Report di classificazione

	precision	recall	f1-score	support
0	0.79	0.76	0.78	1000
1	0.96	0.94	0.95	1000
2	0.69	0.75	0.72	1000
3	0.81	0.85	0.83	1000
4	0.69	0.68	0.68	1000
5	0.91	0.90	0.90	1000
6	**0.61**	**0.56**	**0.58**	**1000**
7	0.88	0.90	0.89	1000
8	0.91	0.93	0.92	1000
9	0.92	0.91	0.92	1000
accuracy			0.82	10000
macro avg	0.82	0.82	0.82	10000
weighted avg	0.82	0.82	0.82	10000

Di seguito è riportata la matrice di confusione sul set di test e possiamo notare che la classe con la peggior accuratezza è la numero 6 ossia le “camice” di cui, tra i 1000 campioni disponibili sul set di test, solamente 558 vengono riconosciuti, i restanti vengono classificati erroneamente principalmente con: magliette, pullover, cappotto.
Invece la classe 1 ossia “pantaloni” è quella con la miglior accuratezza, infatti su 1000 campioni disponibili ben 937 vengono classificati in maniera corretta.



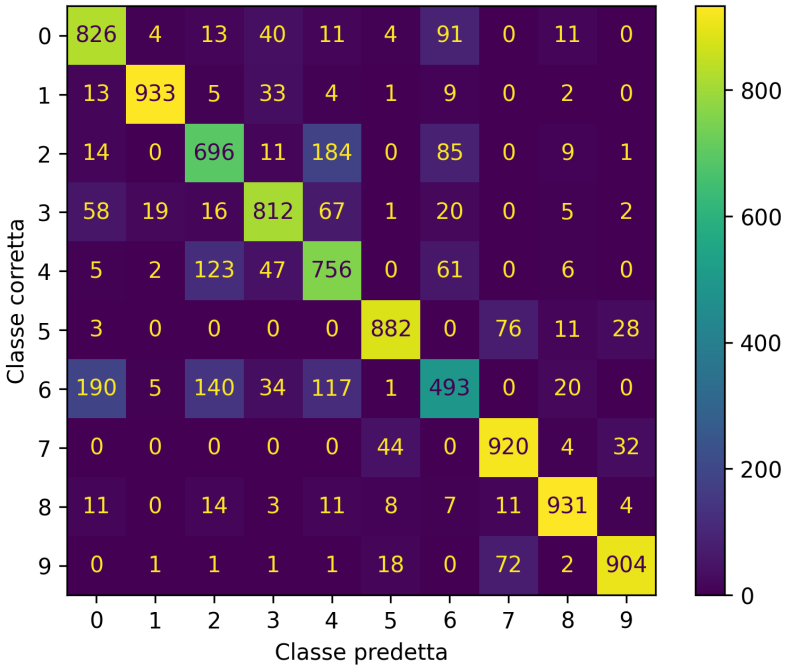
- Test del 2022-11-01 10:48:20 F_MNIST con Decision Tree (profondità: 11, c: entropy, mf: None, msl: 1):

Report di classificazione

	precision	recall	f1-score	support
0	0.74	0.83	0.78	1000
1	0.97	0.93	0.95	1000
2	0.69	0.70	0.69	1000
3	0.83	0.81	0.82	1000
4	0.66	0.76	0.70	1000
5	0.92	0.88	0.90	1000
6	0.64	0.49	0.56	1000
7	0.85	0.92	0.89	1000
8	0.93	0.93	0.93	1000
9	0.93	0.90	0.92	1000
accuracy			0.82	10000
macro avg	0.82	0.82	0.81	10000
weighted avg	0.82	0.82	0.81	10000

Di seguito è riportata la matrice di confusione sul set di test e possiamo notare che la classe con la peggior accuratezza è la numero 6 ossia le “camice” di cui, tra i 1000 campioni disponibili sul set di test, solamente 493 vengono riconosciuti, i restanti vengono classificati erroneamente principalmente con: magliette, pullover, cappotto.
Invece la classe 1 ossia “pantaloni” è quella con la miglior accuratezza, infatti su 1000 campioni disponibili ben 933 vengono classificati in maniera corretta.

Matrice di Confusione: test-set F_MNIST con Decision Tree (p: 11, c: entropy, mf: None, msl: 1)



- Test del 2022-11-01 11:39:20 F_MNIST con Decision Tree (profondità: 11, c: entropy, mf: 0.5, msl: 1):

Report di classificazione

	precision	recall	f1-score	support
0	0.76	0.81	0.79	1000
1	0.97	0.93	0.95	1000
2	0.69	0.73	0.71	1000
3	0.82	0.84	0.83	1000
4	0.66	0.71	0.69	1000
5	0.94	0.89	0.92	1000
6	0.60	0.50	0.55	1000
7	0.86	0.91	0.88	1000
8	0.92	0.92	0.92	1000
9	0.91	0.91	0.91	1000
accuracy			0.81	10000
macro avg	0.81	0.81	0.81	10000
weighted avg	0.81	0.81	0.81	10000

Di seguito è riportata la matrice di confusione sul set di test e possiamo notare che la classe con la peggior accuratezza è la numero 6 ossia le “camice” di cui, tra i 1000 campioni disponibili sul set di test, solamente 496 vengono riconosciuti, i restanti vengono classificati erroneamente principalmente con: magliette, pullover, cappotto.
Invece la classe 1 ossia “pantaloni” è quella con la miglior accuratezza, infatti su 1000 campioni disponibili ben 932 vengono classificati in maniera corretta.