ANALISI DEI DATI PER LA SICUREZZA

MATTEO ESPOSITO

A.A. 2024/2025

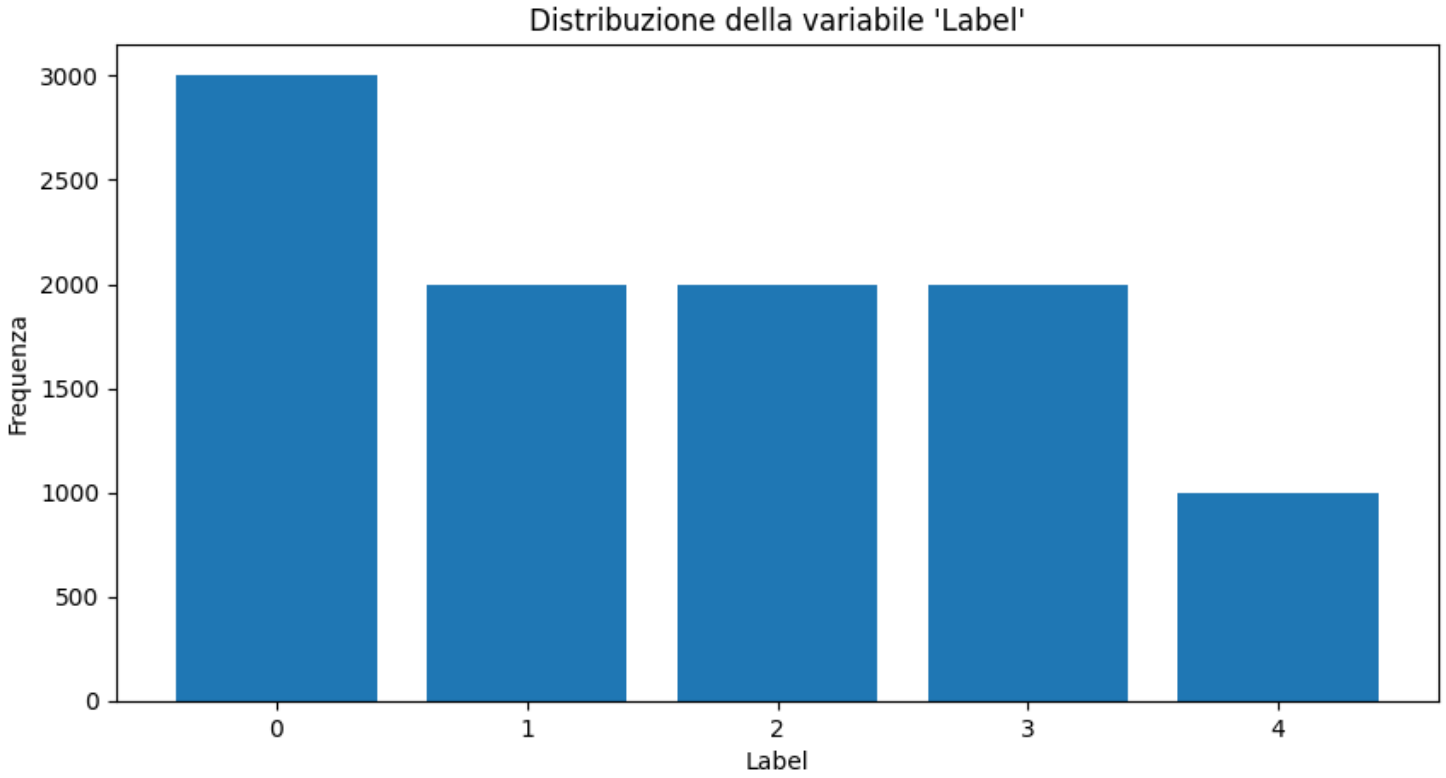
DATASET

Un sottoinsieme di dati raccolti dal Canadian Institute for Cybersecurity nel 2019. Il dataset contiene attacchi che possono essere eseguiti utilizzando protocolli basati su TCP/UDP.

- Set di addestramento: 10.000 campioni
- Set di test: 1.000 campioni
- 78 attributi + 1 classe

	Mapping	#samples
BENIGN	0	3000
MSSQL	1	2000
Syn	2	2000
UDP	3	2000
NetBIOS	4	1000
TOT		10000

BILANCIAMENTO DELLE CLASSI



PRE-ELABORAZIONE

Feature inutili rimosse: [' Bwd PSH Flags', ' Fwd URG Flags', ' Bwd URG Flags', 'FIN Flag Count', ' PSH Flag Count', ' ECE Flag Count', 'Fwd Avg Bytes/Bulk', ' Fwd Avg Packets/Bulk', ' Fwd Avg Bulk Rate', 'Bwd Avg Bytes/Bulk', ' Bwd Avg Packets/Bulk', 'Bwd Avg Bulk Rate']

Dimensione prima della rimozione	Dimensione dopo la rimozione
(10000,79)	(10000,67)

ADDESTRAMENTO DECISION TREE

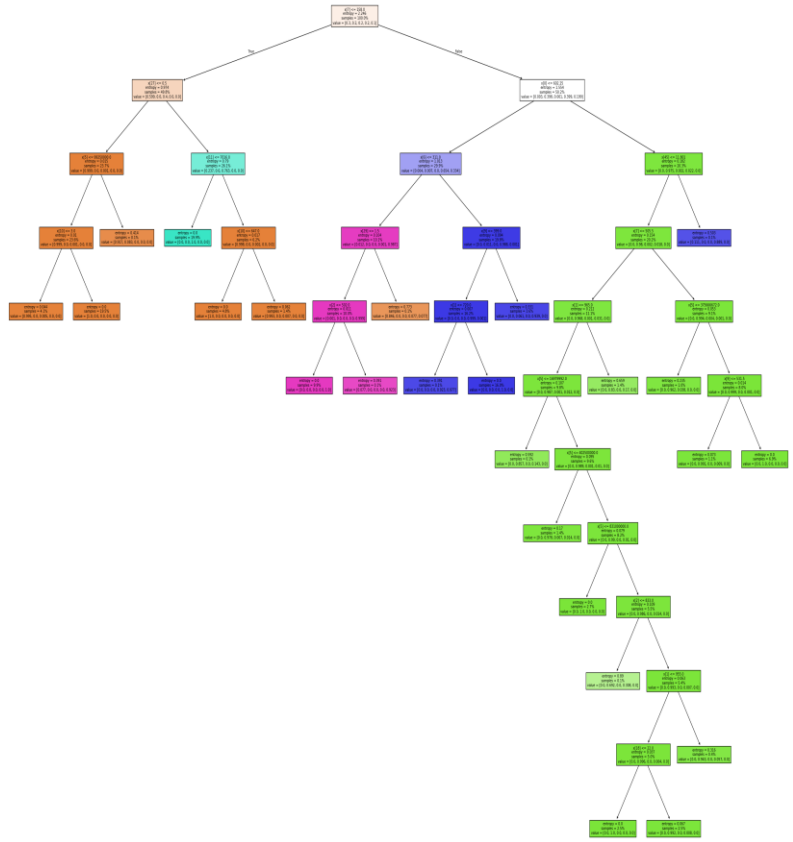
La configurazione ottimale degli alberi decisionali è stata determinata utilizzando la 5-fold cross-validation per selezionare i parametri migliori, in particolare:

- Il criterio di suddivisione: tra Gini ed Entropy
- Il numero di feature utilizzate per la suddivisione

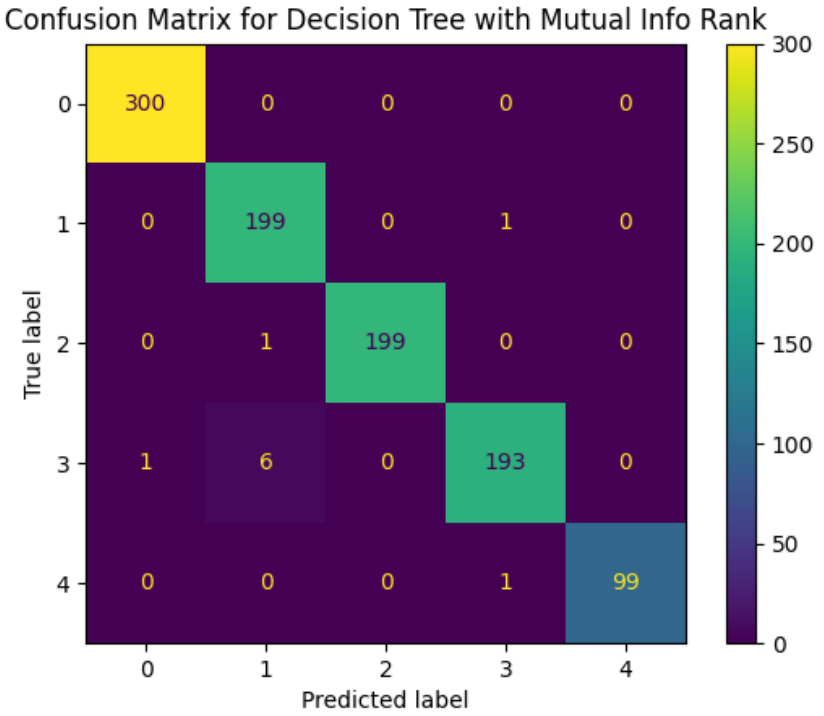
DECISION TREE CON MUTUAL INFO RANK

Migliore configurazione trovata:

- Criterio: entropy
- Numero di feature selezionate: 47
- Miglior F1 Score: 0.992903656485109 6
- L’albero ha 47 nodi e 24 foglie



	precision	recall	f1-score	support
0	1.00	1.00	1.00	300
1	0.97	0.99	0.98	200
2	1.00	0.99	1.00	200
3	0.99	0.96	0.98	200
4	1.00	0.99	0.99	100
accuracy			0.99	1000
macro avg	0.99	0.99	0.99	1000
weighted avg	0.99	0.99	0.99	1000



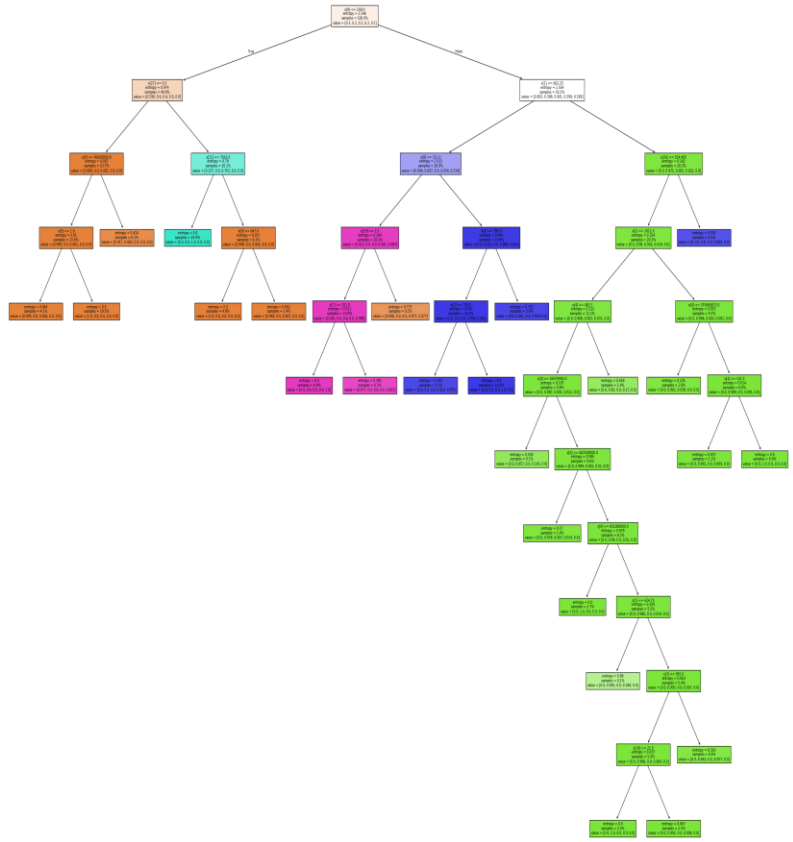
RISULTATI DECISION TREE CON MUTUAL INFO RANK

DECISION TREE CON INFORMATION GAIN

Migliore configurazione trovata:

- Criterio: entropy
- Numero di feature selezionate: 35
- Miglior F1 Score: 0.992903656485109 6

L’albero ha 47 nodi e 24 foglie



	precision	recall	f1-score	support
0	1.00	1.00	1.00	300
1	0.97	0.99	0.98	200
2	1.00	0.99	1.00	200
3	0.99	0.96	0.98	200
4	1.00	0.99	0.99	100
accuracy			0.99	1000
macro avg	0.99	0.99	0.99	1000
weighted avg	0.99	0.99	0.99	1000

DECISION TREE CON INFORMATION GAIN

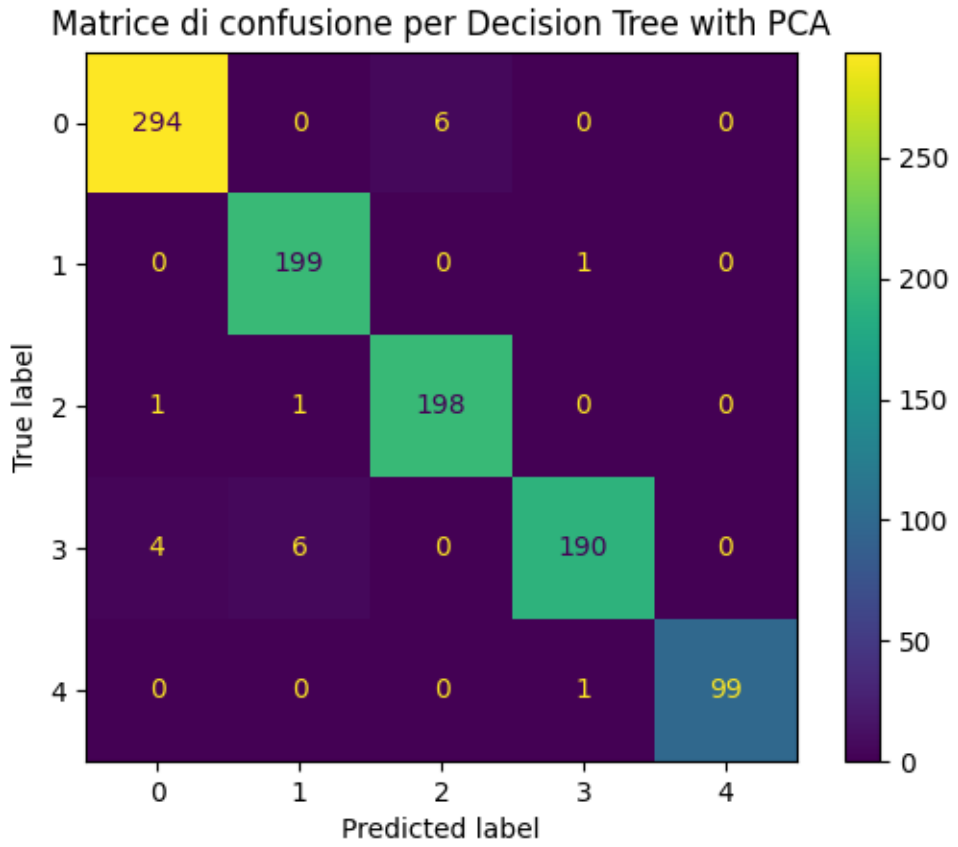
DECISION TREE CON PCA

Migliore configurazione trovata:

- Criterio: gini
- Numero di feature selezionate: 56
- Miglior F1 Score: 0.9843874042879672

L’albero ha 67 nodi e 34 foglie

	precision	recall	f1-score	support
0	0.98	0.98	0.98	300
1	0.97	0.99	0.98	200
2	0.97	0.99	0.98	200
3	0.99	0.95	0.97	200
4	1.00	0.99	0.99	100
accuracy			0.98	1000
macro avg	0.98	0.98	0.98	1000
weighted avg	0.98	0.98	0.98	1000



DECISION TREE CON PCA

ENSEMBLE TRAINING

Generati 10 sample del dataset:

- Ciascuno con l’80% dei dati originali;
- 20 feature selezionate casualmente.

Su questi sottoinsiemi sono stati addestrati 10 modelli SVM;

Ottimizzati utilizzando 5-fold cross validation;

GridSearch con i seguenti parametri:

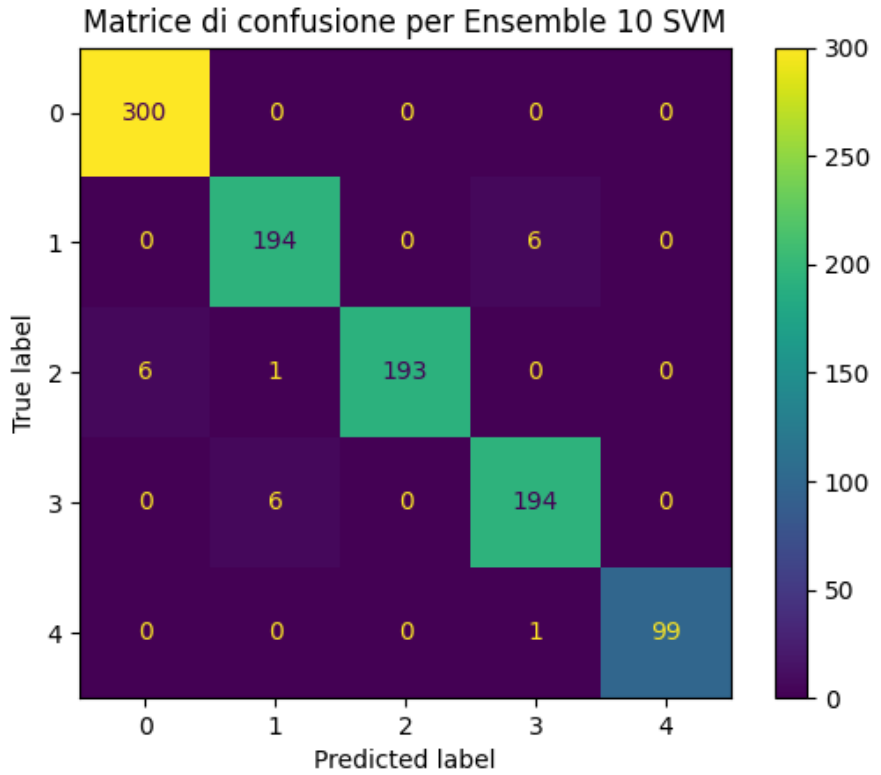
- C ∈ {0.1, 1, 10, 100, 1000};
- gamma ∈ {1, 0.1, 0.01, 0.001, 0.0001};
- kernel = 'rbf’

Le predizioni dei modelli sono state combinate tramite voto di maggioranza

CONFIGURAZIONE ENSEMBLE

Modello	Migliori Parametri	F-score Pesato
SVM 1	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9834
SVM 2	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9803
SVM 3	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9539
SVM 4	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9805
SVM 5	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9609
SVM 6	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9541
SVM 7	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9573
SVM 8	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9823
SVM 9	{'C': 1000, 'gamma': 0.01, 'kernel': 'rbf'}	0.9831
SVM 10	{'C': 1000, 'gamma': 1, 'kernel': 'rbf'}	0.9647

	precision	recall	f1-score	support
0	0.98	1.00	0.99	300
1	0.97	0.97	0.97	200
2	1.00	0.96	0.98	200
3	0.97	0.97	0.97	200
4	1.00	0.99	0.99	100
accuracy			0.98	1000
macro avg	0.98	0.98	0.98	1000
weighted avg	0.98	0.98	0.98	1000



ENSEMBLE SVM

GRAZIE PER L’ATTENZIONE