ANALISI DEI DATI PER LA SICUREZZA

MATTEO ESPOSITO

A.A. 2024-2025

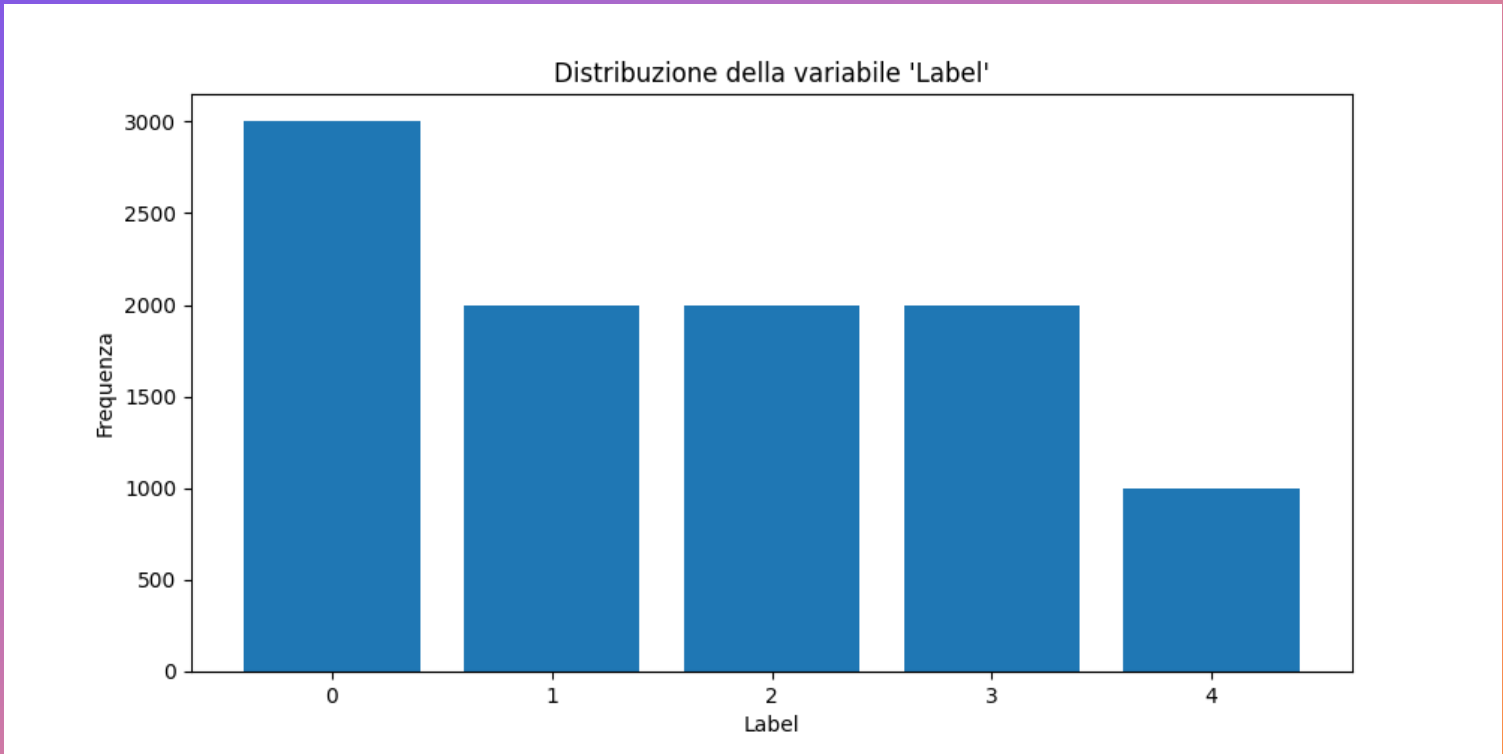
DATASET

A subset of data collected by the Canadian Institute for Cybersecurity in 2019. The dataset contains attacks that can be carried out using TCP/UDP based protocols.

- Training set: 10.000 samples
- Test set: 1.000 samples
- 78 attributes + 1 class

	Mapping	# samples	
		Training set	Test set
BENIGN	0	3000	300
MSSQL	1	2000	200
Syn	2	2000	200
UDP	3	2000	200
NetBIOS	4	1000	100

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)

DECISION TREE TRAINING

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

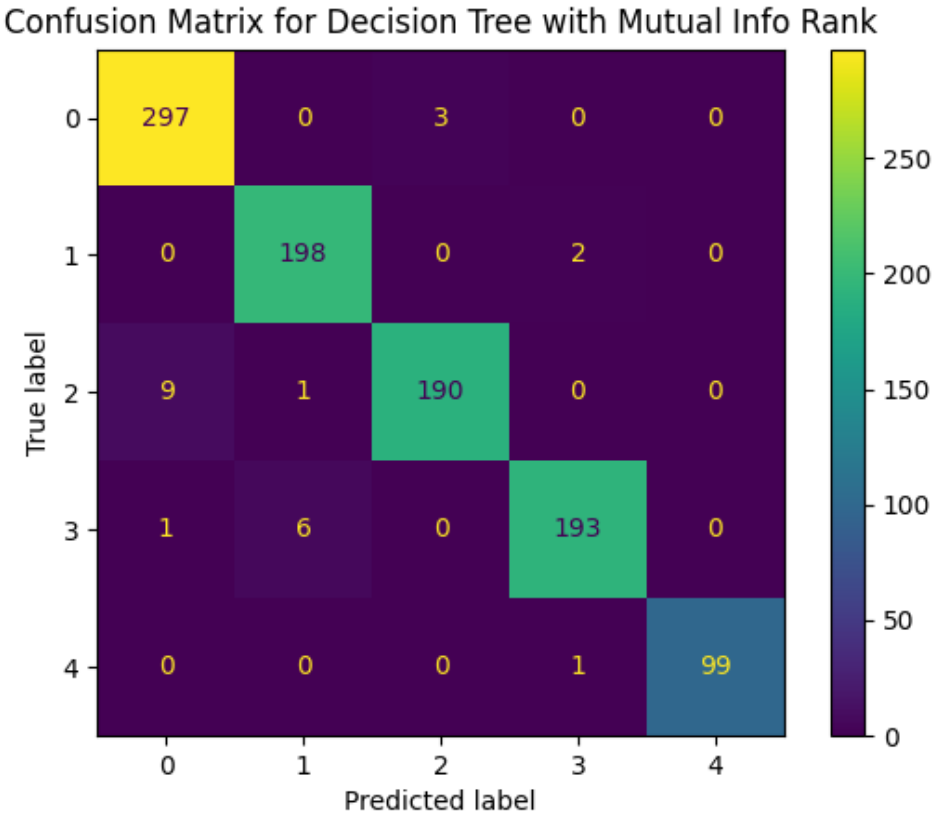
- Il criterio di suddivisione: tra Gini ed Entropy
- Il numero di feature utilizzate per la suddivisione

DECISION TREE CON MUTUAL INFO RANK

Migliore configurazione trovata:

- Criterio: gini
- Numero di feature selezionate: 8*
- Miglior F1 Score: 0.9770047127920977

L’albero ha 69 nodi e 35 foglie



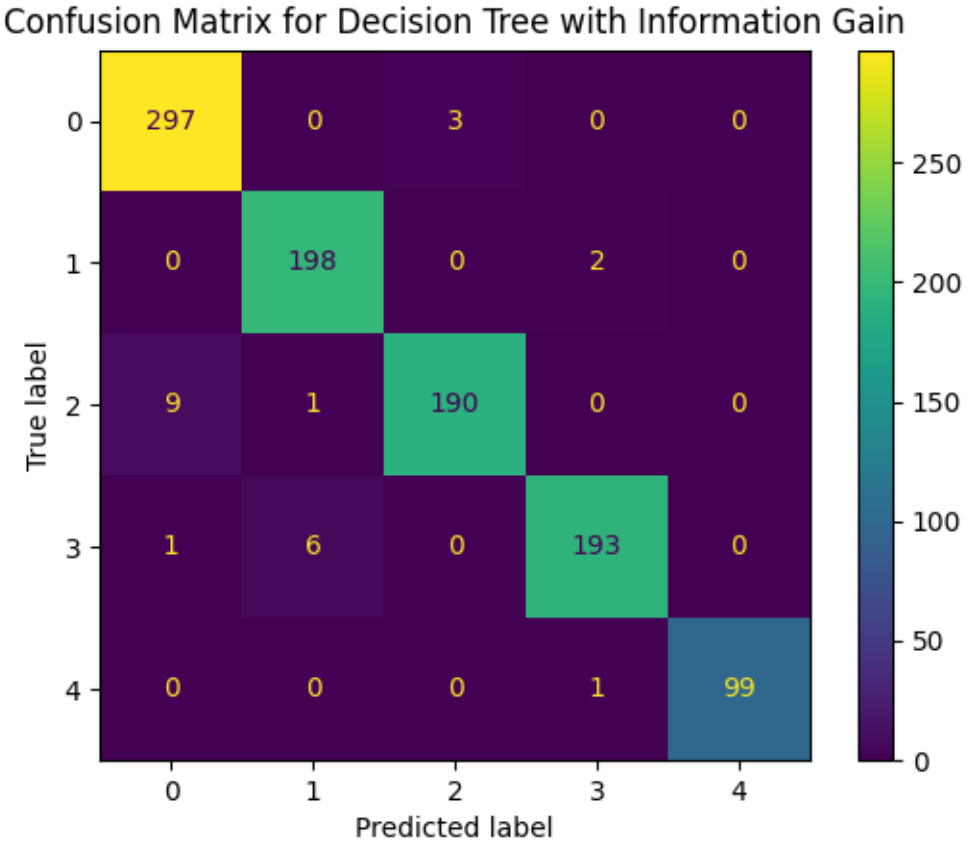
[' Average Packet Size', 'Total Length of Fwd Packets', ' Subflow Fwd Bytes', ' Avg Fwd Segment Size’,
' Fwd Packet Length Mean', 'Flow_Bytes', ' Max Packet Length', ' Min Packet Length', ' Packet Length Mean']

DECISION TREE CON INFORMATION GAIN

Migliore configurazione trovata:

- Criterio: gini
- Numero di feature selezionate: 10*
- Miglior F1 Score: 0.9770047127920977

L’albero ha 69 nodi e 35 foglie



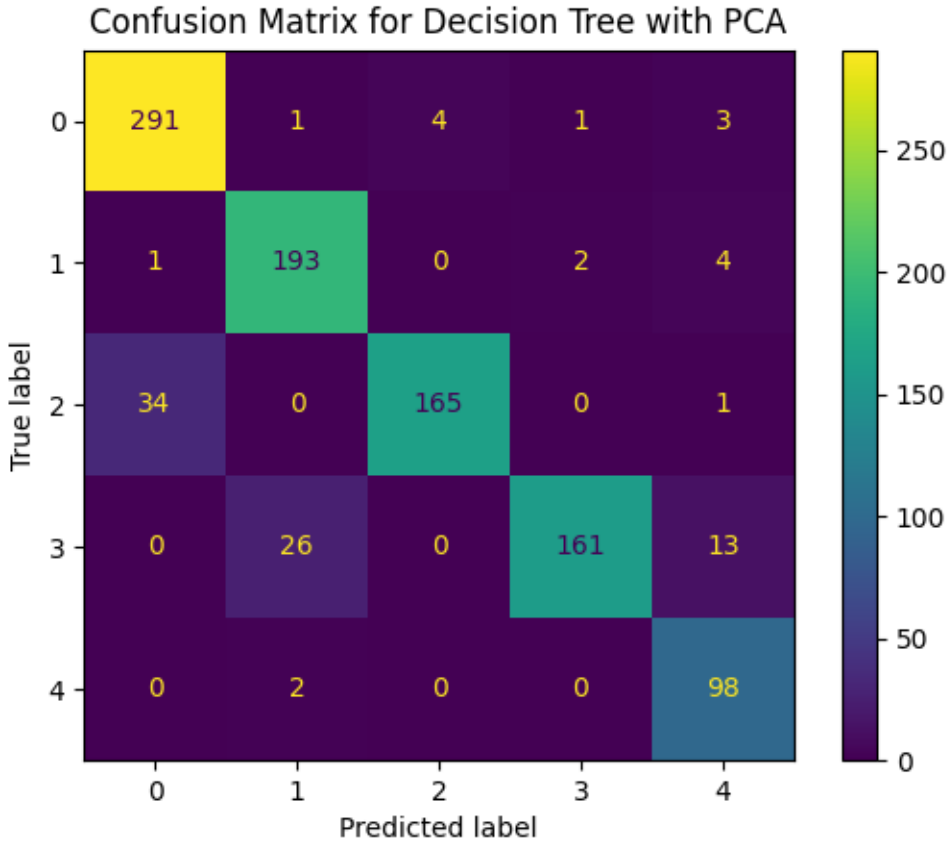
['Flow_Bytes', ' Average Packet Size', 'Total Length of Fwd Packets', ' Subflow Fwd Bytes', ' Packet Length Mean’,
' Fwd Packet Length Mean', ' Avg Fwd Segment Size', ' Max Packet Length', ' Fwd Packet Length Max', ' Min Packet Length']

DECISION TREE CON PCA

Migliore configurazione trovata:

- Criterio: gini
- Numero di feature selezionate: 10*
- Miglior F1 Score: 0.9100705828837418

L’albero ha 87 nodi e 44 foglie



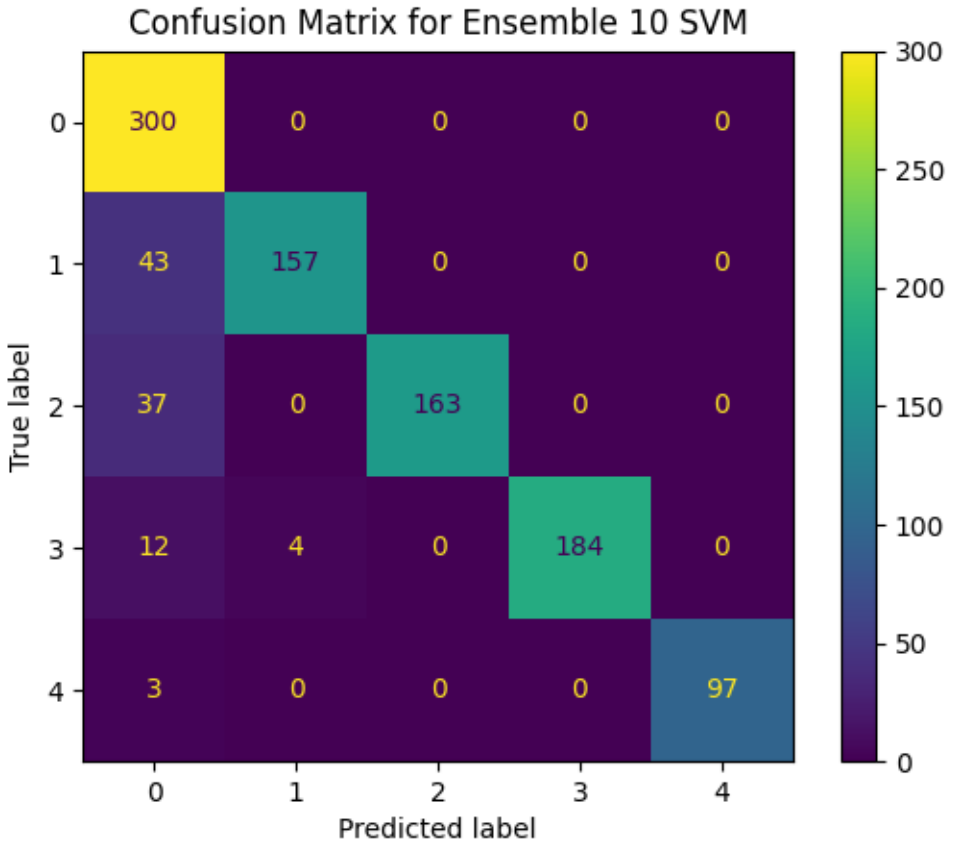
['pc_1', 'pc_2', 'pc_3', 'pc_4', 'pc_5', 'pc_6', 'pc_7', 'pc_8', 'pc_9', 'pc_10']

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.

ENSEMBLE 10 SVM

- SVM 1 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.8626864136925864
- SVM 2 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.8665399666735636
- SVM 3 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.889062218822757
- SVM 4 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.905450580201507
- SVM 5 {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.892660733309868
- SVM 6 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.8861887674490682
- SVM 7 {'C': 10, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.8321152261271788
- SVM 8 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.9396323100107224
- SVM 9 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.8330362604529757
- SVM 10 {'C': 1000, 'gamma': 0.0001, 'kernel': 'rbf'}
  - F-score: 0.9422353105737153



RISULTATI

Modello	Average Accuracy	Weighted F1-Score
DT Mutual Info Rank	97.7%	97.7%
DT Information Gain	97.7%	97.7%
DT PCA	90.9%	90.7%
Ensemble 10 SVM	89.8%	90.2%

GRAZIE PER L’ATTENZIONE