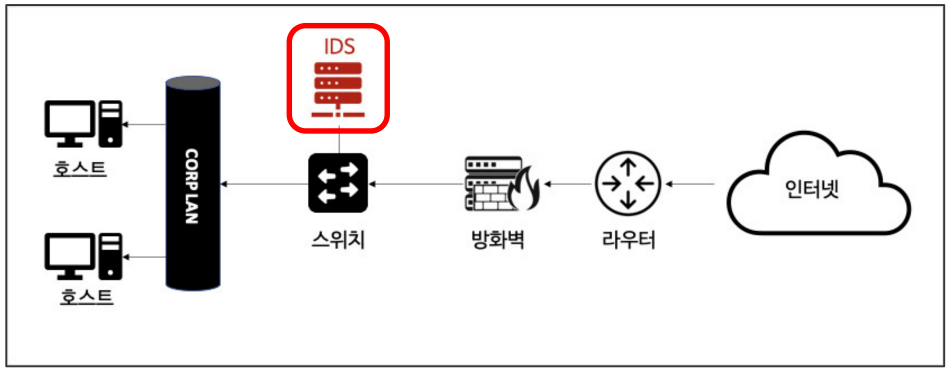
네트워크 패킷 주요 속성 및 최적의 스태킹 모델링을 통한 이상탐지 기법 연구

□ □ □

Contents

□□ □□ □□

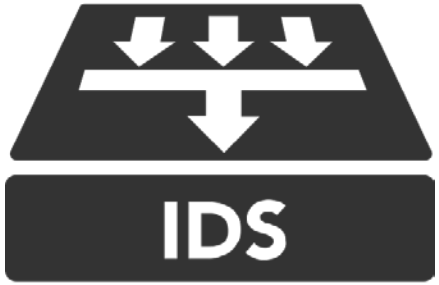
- 침입 탐지 시스템 (IDS : intrusion detection system)
  - 오용 침입 탐지 시스템 - 제로 데이 공격 트래픽에 대해서는 탐지가 불가
  - □□ □□ □□ □□□ - □□ □□□□ □□ □□□ □□□ □□□ □□□ □□□□ □□□ □□□

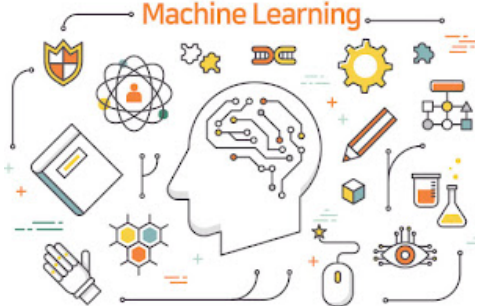


□□ □□ □□

침입 탐지 시스템 (IDS : intrusion detection system)

- □□□□ □□□□ □□ □□□□□ □□
- □□□ □□ □□□□□ □□ □□□ □□





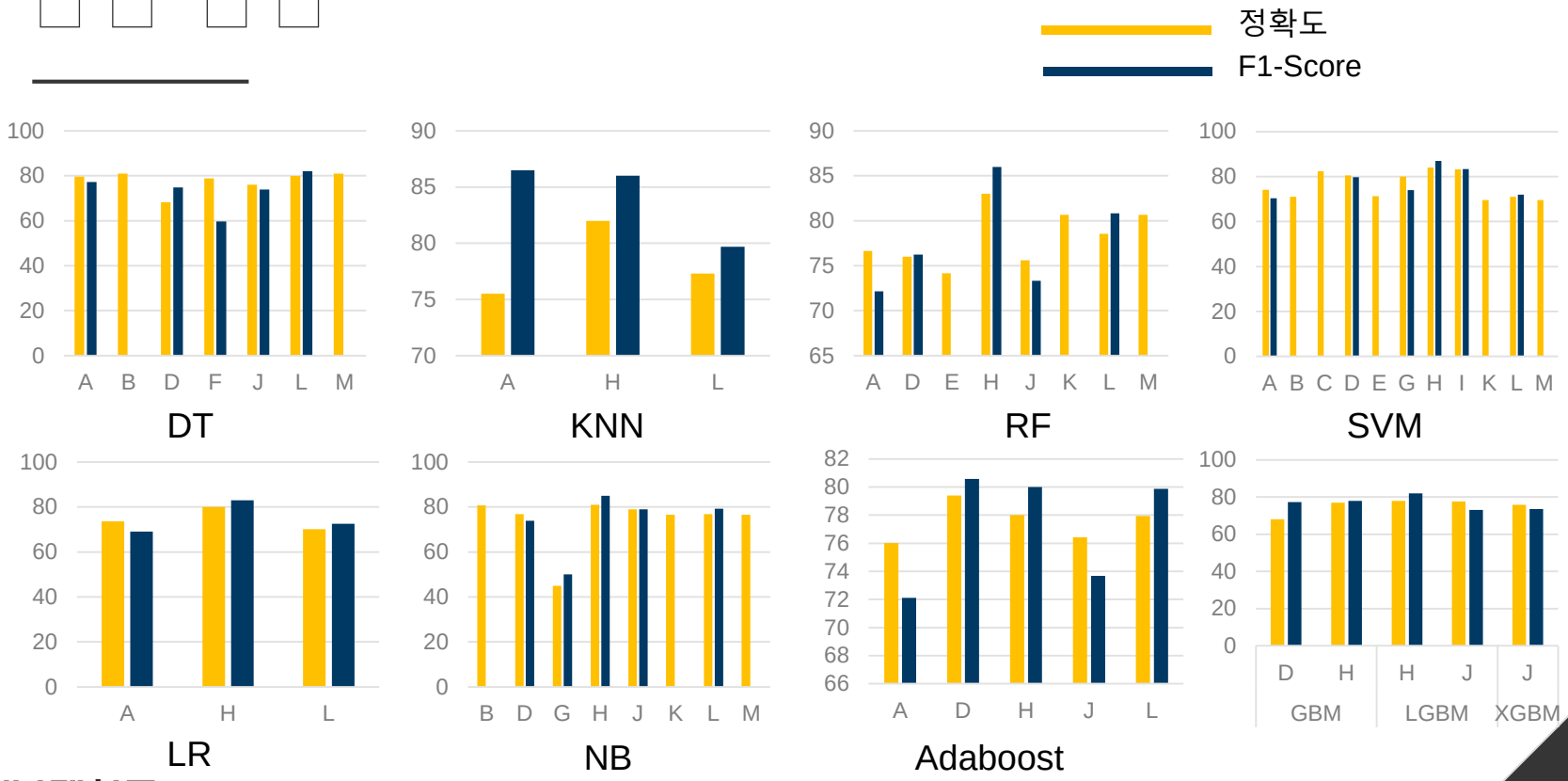
□□ □□

□□ □□

□□□□	□□□ □□
DT	[A], [D], [F], [J], [L]
KNN	[A], [H], [L]
RF	[A], [D], [H], [J], [K], [L]
SVM	[A], [B], [C], [D], [E], [G] [H], [I], [K], [L]
LR	[A], [H], [L], [M]
NB	[B], [D], [G], [H], [J], [K], [L], [M]
Adaboost	[A], [D], [H], [J], [L]
GBM	[D], [H]
LGBM	[H], [J], [M]
XGBM	[J]

DT : DecisionTree
KNN : K-Nearest Neighbor
RF : RandomForest
SVM : Support Vector Machine
LR : LogisticRegression
NB : Naïve Bayes
GBM : Gradient boosting
LGBM : LightGBM
XGBM : XGBoost

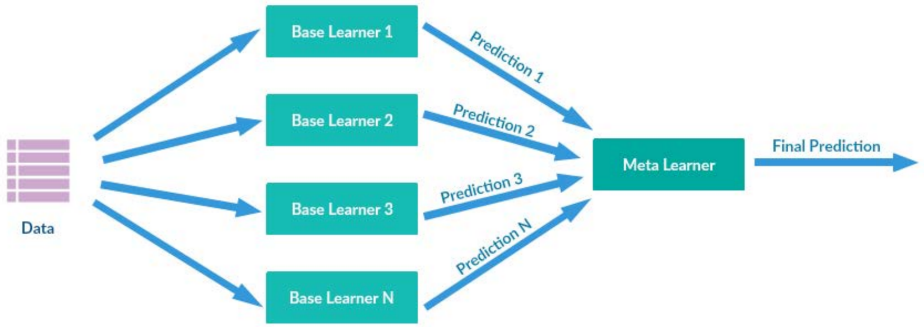
□ □ □ □


정확도
F1-Score
DT
KNN
RF
SVM
LR
NB
Adaboost
GBM
LGBM
XGBM

□□ □□ □□

Ensemble Stacking

- 여러 개별 알고리즘으로 분류 후 예측한 결과를 메타 데이터 세트로 생성 , 이를 다시 새로운 □□□□□□ □□□ □□□□ □□

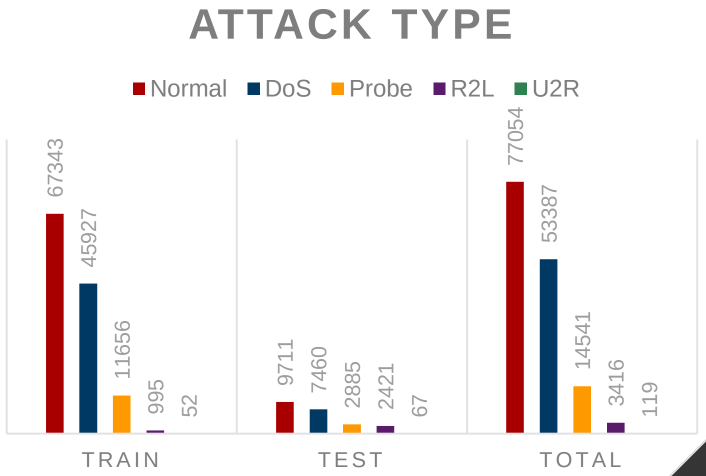


데이터 셋

NSL-KDD

- 유명한 네트워크 데이터 셋인 KDD99 의 수 많은 중복 레코드를 제거해 과대적합이 발생하지 않도록 Tavallaee 가 제안한 데이터 셋

F#	Feature name	F#	Feature name	F#	Feature name
F1	Duration	F15	Su attempted	F29	Same srv rate
F2	Protocol type	F16	Num root	F30	Diff srv rate
F3	Service	F17	Num file creations	F31	Srv diff host rate
F4	Flag	F18	Num shells	F32	Dst host count
F5	Source bytes	F19	Num access files	F33	Dst host srv count
F6	Destination bytes	F20	Num outbound cmds	F34	Dst host same srv rate
F7	Land	F21	Is host login	F35	Dst host diff srv rate
F8	Wrong fragment	F22	Is guest login	F36	Dst host same src port rate
F9	Urgent	F23	Count	F37	Dst host srv diff host rate
F10	Hot	F24	Srv count	F38	Dst host serror rate
F11	Number failed logins	F25	Serror rate	F39	Dst host srv serror rate
F12	Logged in	F26	Srv serror rate	F40	Dst host rerror rate
F13	Num compromised	F27	Rerror rate	F41	Dst host srv rerror rate
F14	Root shell	F28	Srv rerror rate	F42	Class label



모델 구현 및 검증

데이터 전처리

- 이진 분류를 위해서 공격 유형을 0 또는 1로 설정
- 문자형 (Categorical) 속성은 원 - 핫 인코딩으로 숫자형으로 변환

protocol_type	service	flag
tcp	ftp_data	SF
udp	other	SF
tcp	private	S0
tcp	http	SF
tcp	http	SF
...	...	...
tcp	private	S0
udp	private	SF
tcp	smtp	SF
tcp	klogin	S0
tcp	ftp_data	SF

protocol_type	service	flag
1	20	9
2	44	9
1	49	5
1	24	9
1	24	9
...	...	...
1	49	5
2	49	9
1	54	9
1	30	5
1	20	9

□□ □□ □ □□

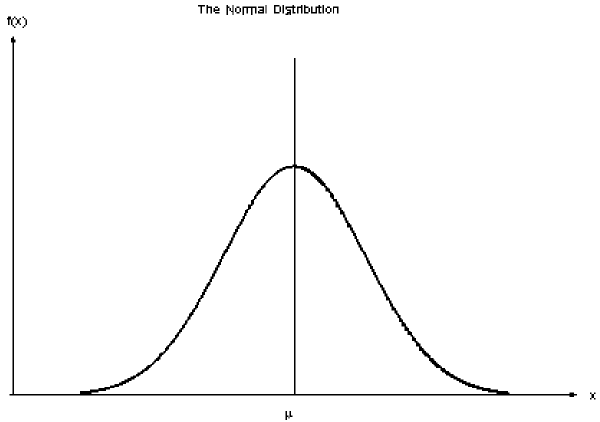
데이터 전처리

- □□ □□
  - Sklearn□ SelectKBest
  - □□□ □□□□ □□ □□□ □□□□ f_classif □□
    - □□ □□ : □□□□ □□□ □□ □□□ □□□ □□□□ □□

모델 구현 및 검증

데이터 전처리

- 사용한 스케일링은 분산이 1이고, 평균을 0으로 변환해주는 Scikit-learn의 StandardScaler를 사용



duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot
0	491	0	0	0	0	0
0	146	0	0	0	0	0
0	0	0	0	0	0	0
0	232	8153	0	0	0	0
0	199	420	0	0	0	0

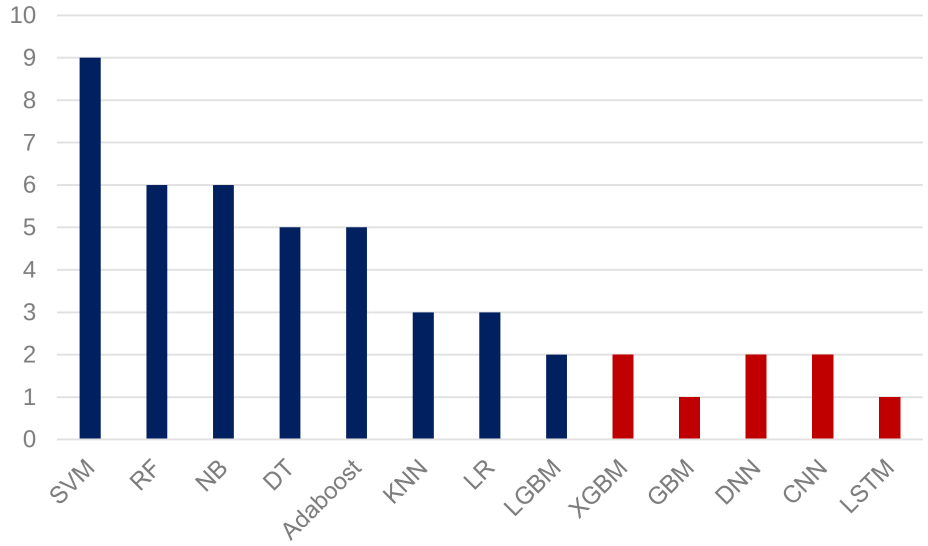
duration	src_bytes	dst_bytes	land	wrong_fragment	urgent	hot
-0.110249	-0.007679	-0.004919	-0.014089	-0.089486	-0.007736	-0.095076
-0.110249	-0.007737	-0.004919	-0.014089	-0.089486	-0.007736	-0.095076
-0.110249	-0.007762	-0.004919	-0.014089	-0.089486	-0.007736	-0.095076
-0.110249	-0.007723	-0.002891	-0.014089	-0.089486	-0.007736	-0.095076
-0.110249	-0.007728	-0.004814	-0.014089	-0.089486	-0.007736	-0.095076

- StandardScaler □□

모델 구현 및 검증

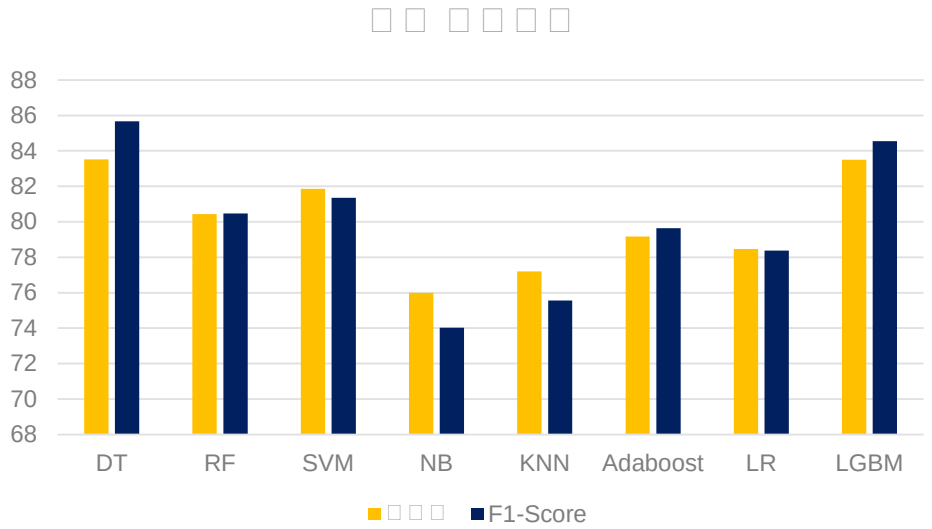
스태킹 모델을 위한 개별 알고리즘

- 개별 알고리즘은 DT, RF, SVM, NB, KNN, Adaboost, LR, LGBM 을 □ □



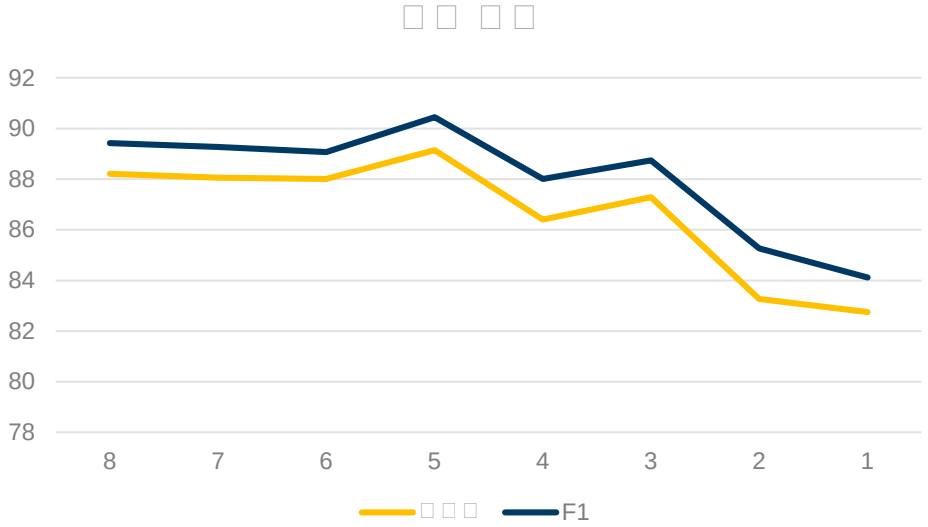
□□ □□ □ □□

□□□ □□ □□□ □□ □□ □□□ □□ □□ □□ □□


□□ □□□□
88
86
84
82
80
78
76
74
72
70
68
DT
RF
SVM
NB
KNN
Adaboost
LR
LGBM
□□□□
F1-Score

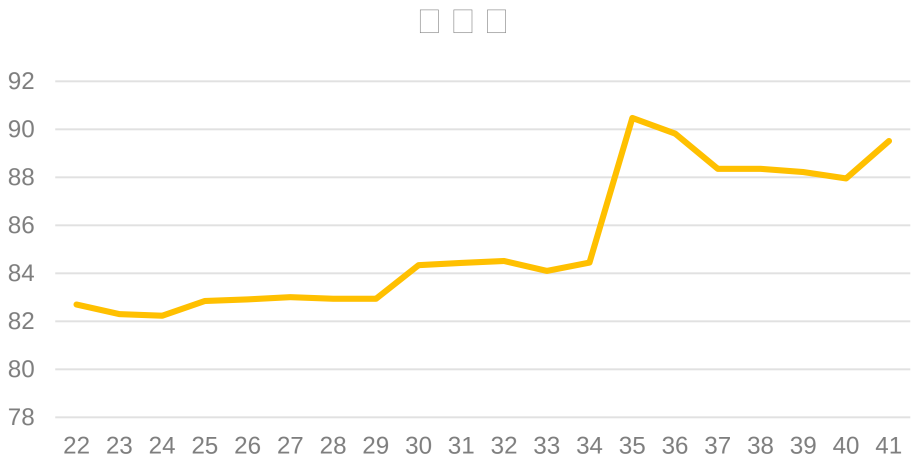
□□ □□ □ □□

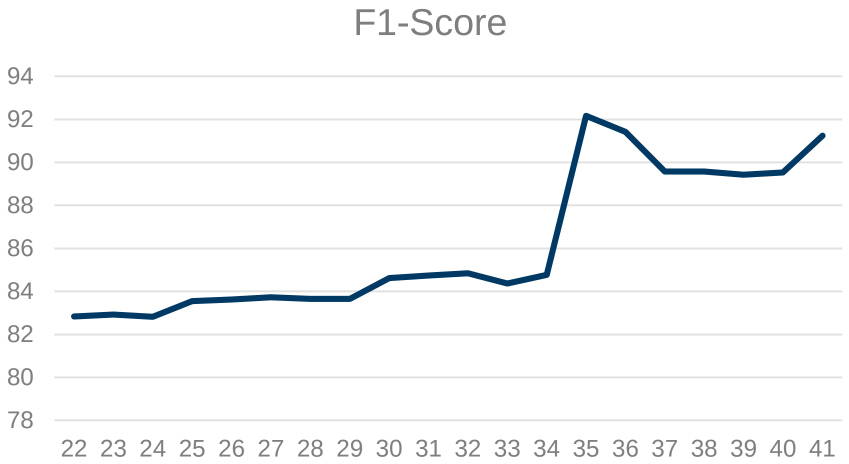
여러 알고리즘을 결합한 결과 DT, LGBM, KNN, Adaboost, LR 알고리즘을 채택



모델 구현 및 검증

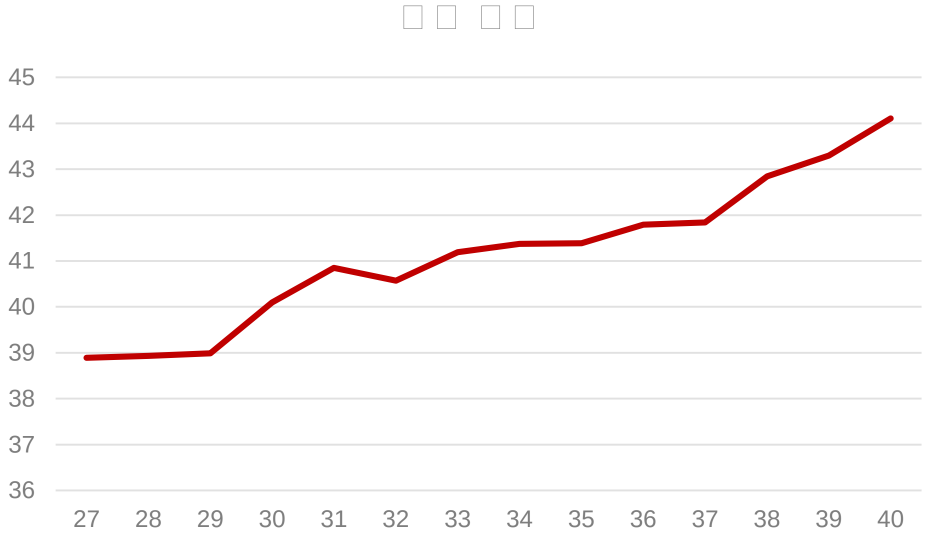
성능 향상 및 시간 단축을 위해 특징 선택





모델 구현 및 검증

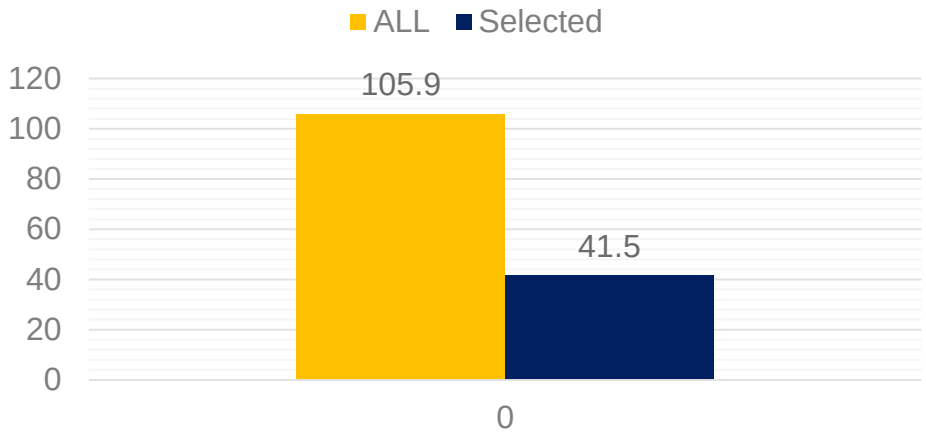
성능 향상 및 시간 단축을 위해 특징 선택



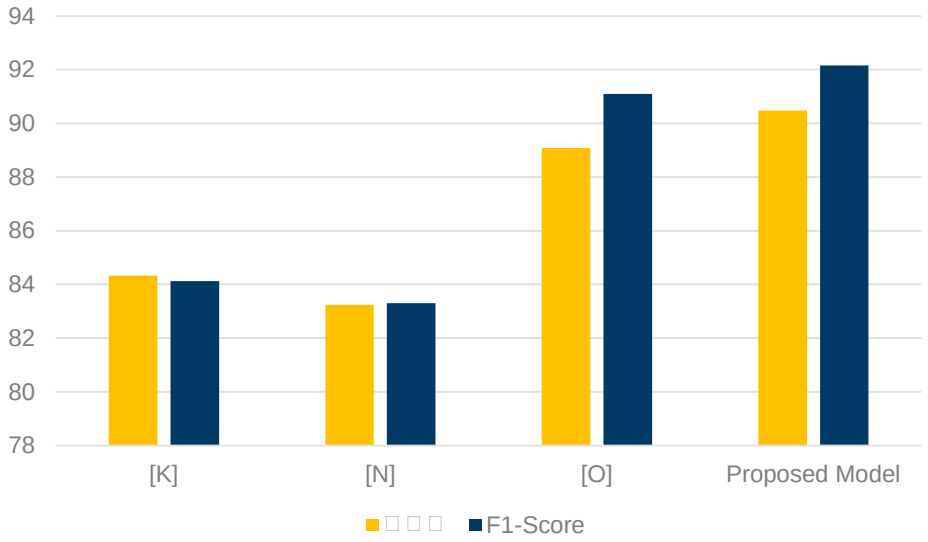
□□ □□ □ □□

□□ □□□ □□ □□ □□□ □□□ □□ □□

□□ □□


ALL
Selected
120
100
80
60
40
20
0
105.9
41.5
0

기존 연구와 비교



□ □

□□ □□□□□ □□ □□□ □□□□ □□ □□□□ □□□ □□□ □□
□□ □□□ □□□□□ □□□□ □□ □□□ □□ □□
- □□ □□□□ 8% ~ 1% □□
- 최고의 성능을 보인 단일 알고리즘인 DT 보다 정확도와 F1-Score 약 7% 향상

□□ □□
- 각 공격에 대해 탐지할 수 있는 다중 분류 이상탐지 모델 개발
- □□ □□□□□ XAI □□□ □□ □□□ □□

Reference

[A]. An Adaptive Ensemble Machine Learning

[B]. Decision Tree Based Intrusion Detection System for NSL-KDD Dataset

[C]. Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection

[D]. Learning to Detect: A Data-driven Approach forNetwork Intrusion Detection

[E]. Intrusion Detection System for NSL-KDD Dataset Using Convolutional Neural Networks

[F]. Decision Tree Techniques with Feature Reduction for Network Anomaly Detection

[G]. A Stacked Generalization Ensemble Approach for Improved Intrusion Detection

[H]. Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network

[I]. Hybrid Intrusion Detection System Based on the Stacking Ensemble of C5 Decision Tree Classifier and One Class Support Vector Machine

[J]. Network Intrusion Detection Using Hybrid Machine Learning Model

[K]. Designing a Network Intrusion Detection System Based on Machine Learning for Software Defined Networks

[L]. AN EFFICIENT INTRUSION DETECTION APPROACH USING LIGHT GRADIENT BOOSTING

[M]. Two-tier network anomaly detection model: a machinelearning approach

[N]. Multi-dimensional feature fusion and stacking ensemble mechanism for network intrusion detection

[O]. Effective Intrusion Detection with a Neural Network Ensemble Using Fuzzy Clustering and Stacking Combination Method

Thank you

□□□□□.