Network Anomaly Detection Using UNSBW-NB15

Mohamed Elgharieb

Eslam Ahmed

Loai Nazeer

Mohamed Elgaidy

- 12.11.2021

Feature Extraction

Expected delivery
Dec 12, 2021

Recent progress

- Feature extraction papers.
- Feature importance by RandomForrest Regressor.
- Feature Importance by RandomForrest Classifier.
- Clear Nan values.
- Encoding for string features.

Biggest risk

Overfitting if we could not extract the most valuable features, And underfitting if features selected have not been deployed well.

Progress - Feature Extraction

Accomplishment 1-Feature Extraction based on papers.

- (PDF) UNSW-NB15 dataset feature selection and network intrusion detection using deep learning (researchgate.net)
- JANARTHANAN, Tharmini and ZARGARI, Shahrzad (2017). Feature Selection in UNSW-NB15 and KDDCUP'99 datasets. In: 2017 IEEE 26th International Symposium on Industrial Electronics (ISIE),. IEEE.

Accomplishment 2-Feature Importance by RandomForrest Regressor.

- To rank the features descending from the most valuable ones to the poorest.

Progress - Feature Extraction

Accomplishment 3-Feature Importance by RandomForrest Classifier

- Extract the most valuable 4 features from the whole dataset.

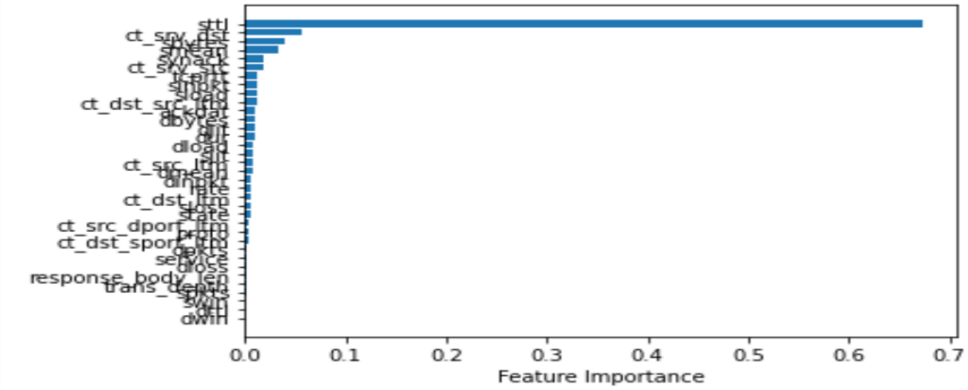
Accomplishment 4-Data cleaning

- Clear Nans from all rows.
- Encode string features.

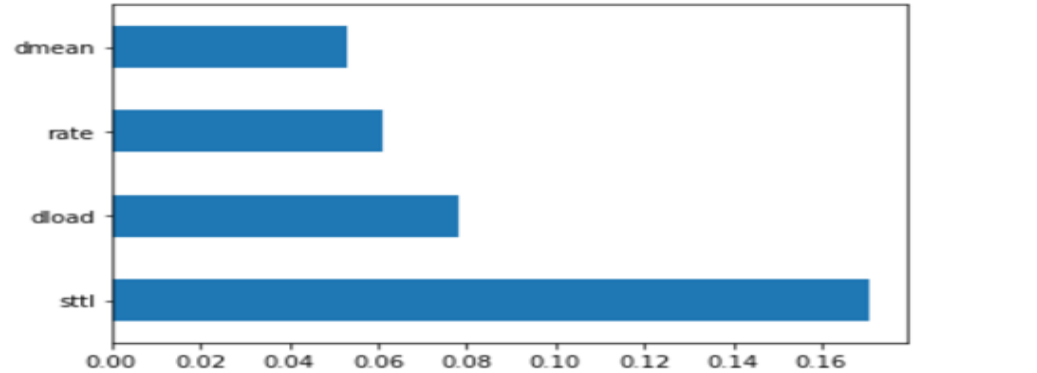
Feature Importance

```
Feature ranking:
1. feature 9 (0.674491)
2. feature 34 (0.058267)
3. feature 6 (0.040117)
4. feature 24 (0.033231)
5. feature 28 (0.016737)
6. feature 22 (0.016219)
7. feature 11 (0.011402)
8. feature 15 (0.011370)
9. feature 21 (0.011258)
10. feature 32 (0.011248)
11. feature 23 (0.011105)
12. feature 18 (0.009208)
13. feature 0 (0.009157)
14. feature 12 (0.008217)
15. feature 33 (0.007705)
16. feature 7 (0.007376)
17. feature 17 (0.007114)
18. feature 16 (0.006845)
19. feature 25 (0.006804)
20. feature 8 (0.006353)
21. feature 13 (0.006265)
22. feature 29 (0.005603)
23. feature 3 (0.005011)
24. feature 30 (0.004681)
25. feature 1 (0.002541)
26. feature 5 (0.002340)
27. feature 2 (0.002249)
28. feature 31 (0.002118)
29. feature 14 (0.001828)
30. feature 27 (0.001310)
31. feature 4 (0.000864)
32. feature 26 (0.000849)
33. feature 10 (0.000088)
34. feature 19 (0.000026)
35. feature 20 (0.000003)
```

```
Text(0.5, 0, 'Feature Importance')
```

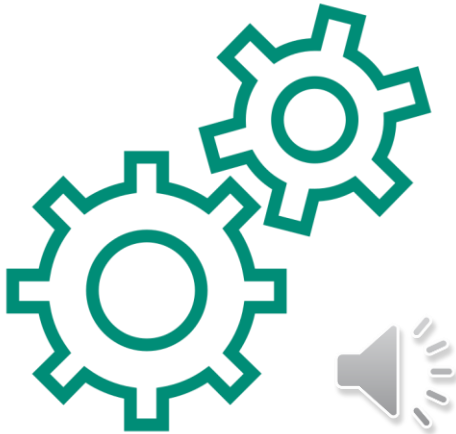



```
<matplotlib.axes._subplots.AxesSubplot at 0x7f902d565690>
```

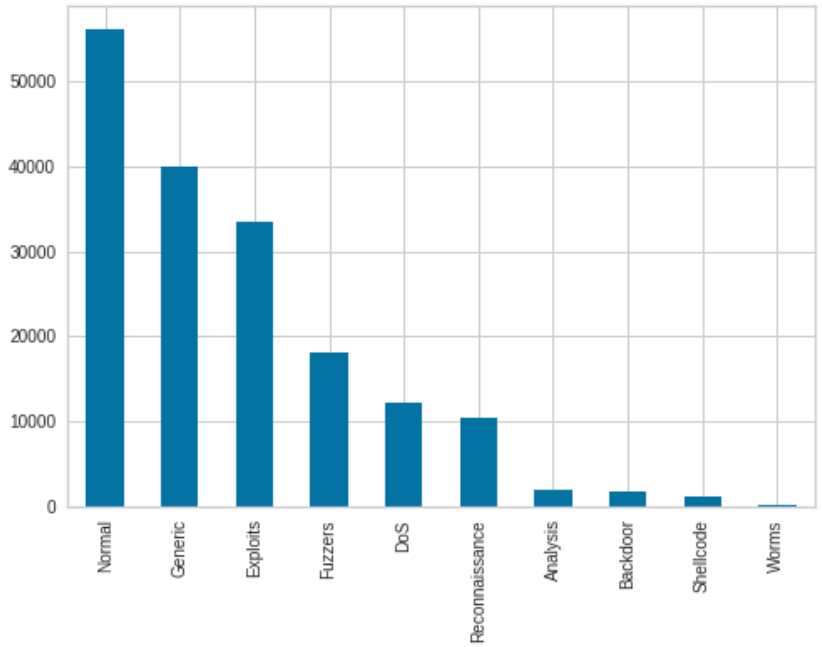


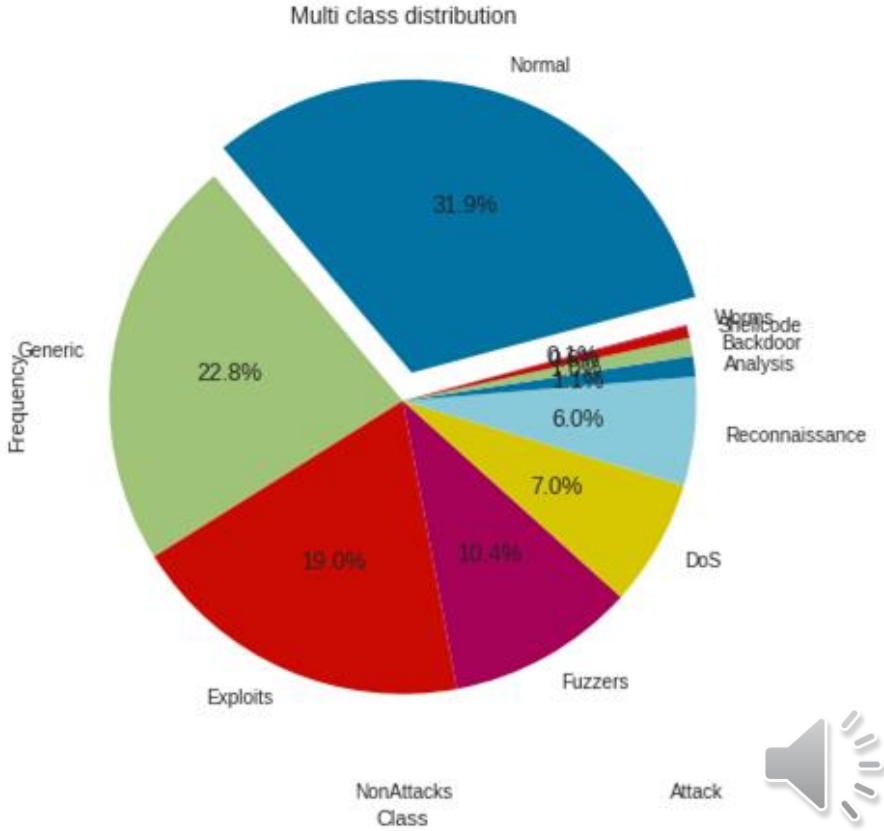
Preprocessing

- REDUNDANCY
- HANDLING MISSING DATA
- DATA ENCODING
- NORMALIZATION



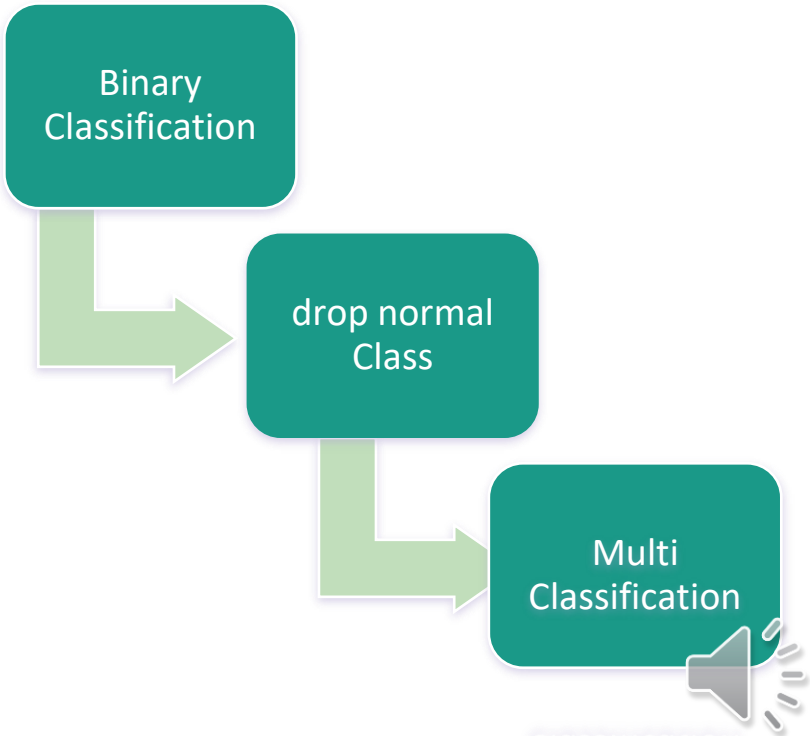
Preprocessing - REDUNDANCY





Preprocessing - REDUNDANCY

	Method	Test	Val	train
1	RandomOverSampler	82	68	73
2	RandomUnderSampler	75	74	75
3	CNN	82	89	89
4	SMOTE	81	85	85
5	Tomek Links	72	84	93
6	Cluster Centroids	59	58	59



Preprocessing - handling missing data

dur	proto	service	state	spkts	dpkts	sbytes	dbytes	rate	sttl	dttl
0.0495	udp	dns	CON	2	2	142	324	60.60606	60	254
0.121478	tcp	-	FIN	6	4	258	172	74.08749	252	254
0.533245	tcp	-	CON	6	2	978		13.12715	62	252
0.082615		-	CON	6	2	986	86	84.73037	62	252
0.237811	udp	dns	CON	2	2	146	244	12.61506	62	252
0.237811	udp	dns	CON	2	2	146	244	12.61506	62	252
0.237811	udp	dns	CON	2	2	146	244	12.61506	62	252
0.179799	udp	dns	CON	2	2	134	166	16.6853	62	252
0.115594	tcp	-	CON	6	2	986	86	60.55678	62	252
0.118007	tcp	-	CON	6	2	986	86	59.31852	62	252
0.128324	tcp	-	CON	6	2	986	86	54.54942	62	252
0.12974	tcp	-	CON	6	2	986	86	53.95406	62	252

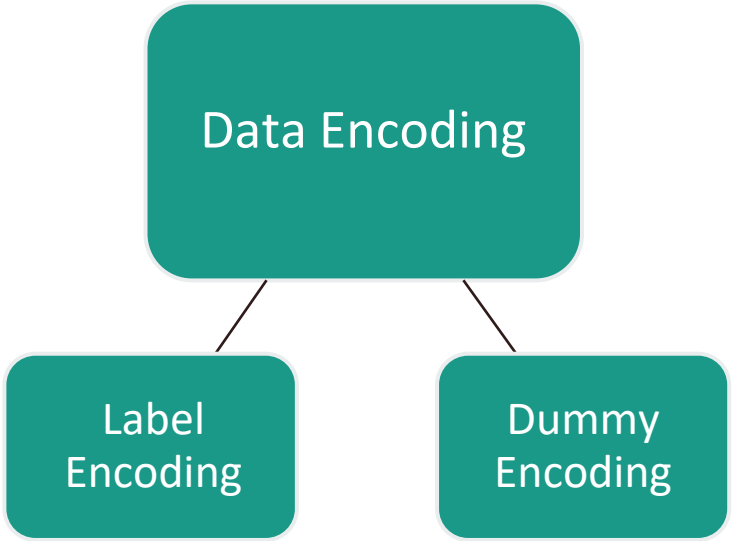
Preprocessing - Data encoding

Label encoding

- Binary label attributes
- categorical label attributes

Dummy encoding

categorical attributes
“proto”, “service”, and “state”.

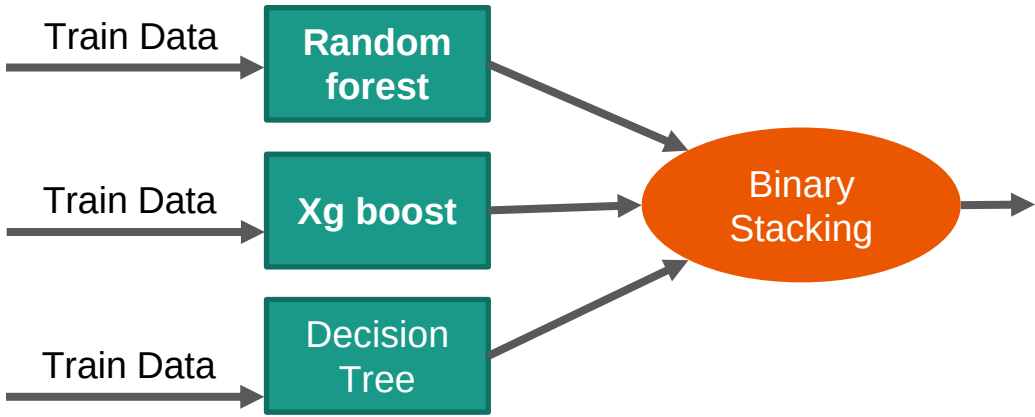


Algorithm Model

- Binary classification
- Multi classification
- Combination between two classifiers

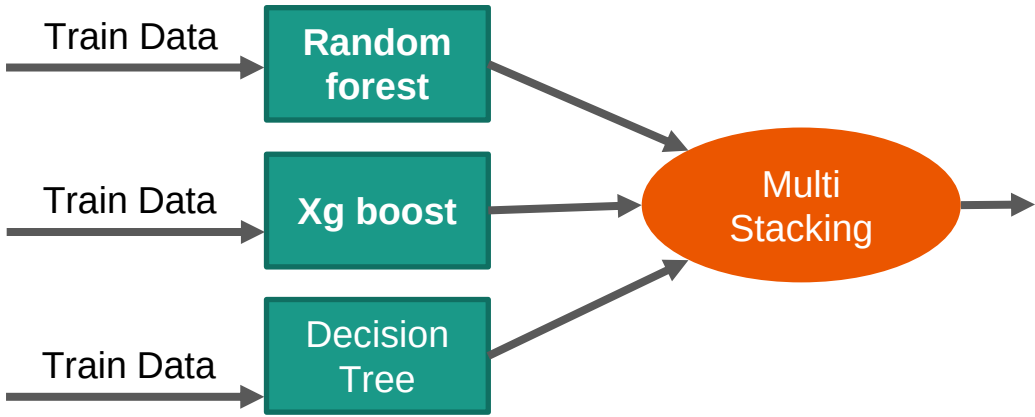
Algorithm Model - Binary classification

	Method	Test	Val	train
1	Random forest	.94	.986	.99
2	XG boost	.92	.987	.98
3	DT	.92	.983	.97
3	Stacking	.93	.986	.99

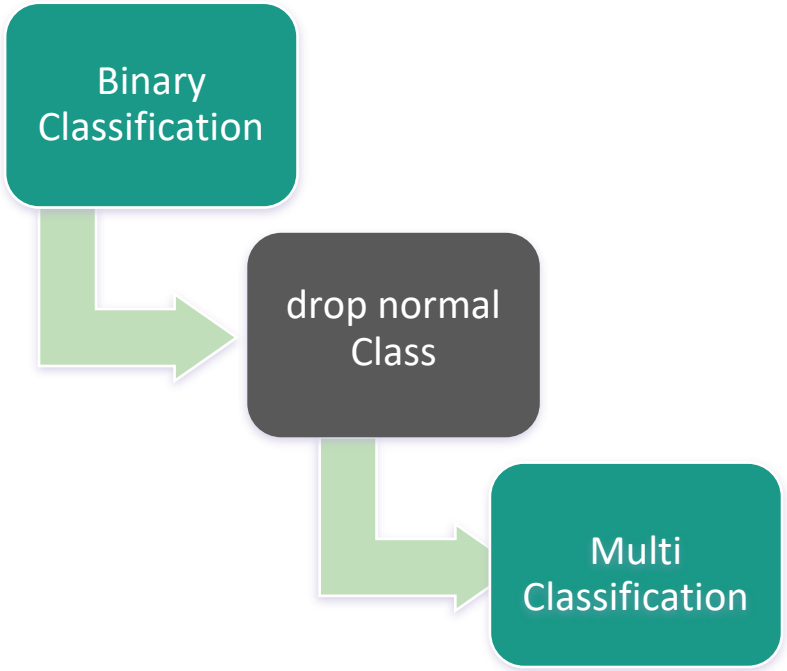


Algorithm Model - Multi classification

	Method	Test	Val	train
1	Random forest	.94	.94	.96
2	xgboost	.947	.94	.98
3	DT	.92	.935	.97
3	Stacking	.949	.942	.99



Algorithm Model - Combination classifiers



Future Work

Data Cleaning

- Recursive feature Elimination.

Data Preprocessing

- Noise Filtering(Outliers).
- Space Transformation(PCA).

ML Algorithms

- Deep learning.
- Combination of Ensembles with our latest work.

Thank You