Network Attack Detection

Using UNSW-NB15 Dataset

Team 8
Divya Khandelwal Nancy Saxena Chintan Shah Wen-Hao Tseng

Problem Statement

- Cyber attacks - biggest threats
- Very important to combat the network attacks to establish a secure environment for users
- The project analyses the performance of different ML models over dataset of raw network packets to detect network attacks
- Inferences like best working model are derived

UNSW-NB15 Dataset

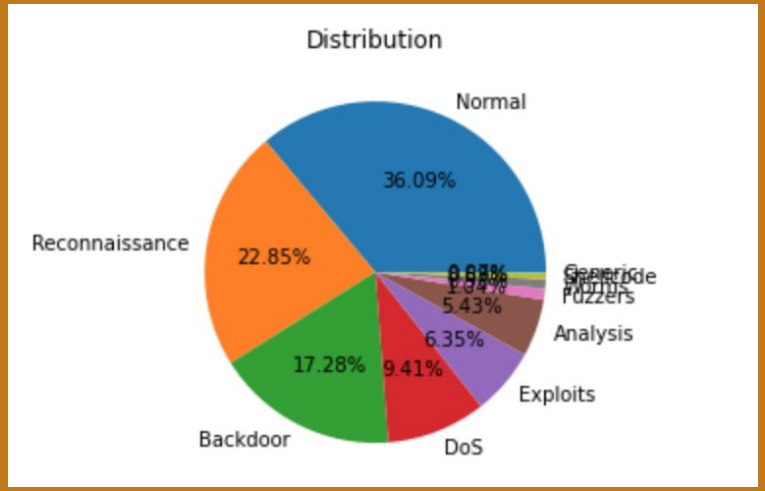
UNSW-NB15 Dataset

- Cyber Range Lab of UNSW, Canberra
- A partition from this dataset was configured as a training set and testing set
- Contains 257,673 data records with 49 features
- https://research.unsw.edu.au/projects/unsw-nb15-dataset

UNSW-NB15 Dataset Features

- 4 Categorical
  - Protocol
  - Service
  - Attack_cat
  - State
- 45 Numerical
  - dur
  - sbytes
  - dbytes
  - spkts
  - etc

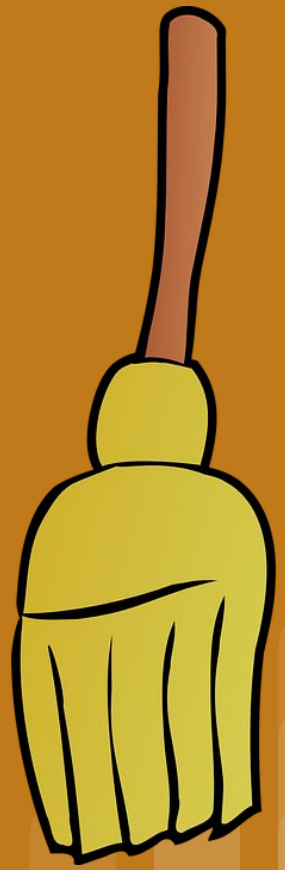
UNSW-NB15 Dataset Attacks



9 types of network attack

- **Fuzzers**
- **Analysis**
- **Backdoors**
- **Denial of Service**
- **Exploits**
- **Generic**
- **Reconnaissance**
- **Shellcode**
- **Worms**

Data Preparation and Cleaning



Data Preparation and Cleaning

1. Dropped unnecessary columns - ‘id’
2. Checked for Missing values:
    a. Categorical columns - Missing values were appropriately replaced.
    b. Erroneous values were corrected.

    Observation: Our dataset was already cleaned for missing values for numerical features.

Data Preparation and Cleaning

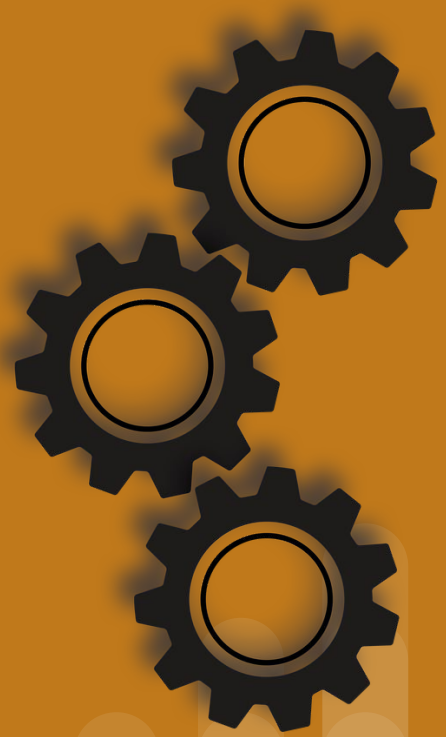
3. Encoding of Categorical Data:

   a. One Hot Encoding
   b. Label Encoding

   One-hot encoder was increasing the number of features from 44 to 197. It was observed that it was leading to overfitting in certain models.

   So we chose label encoding to go ahead with.

PreProcessing Dataset



Preprocessing of the dataset

-

Feature Scaling

Minmax scaler :

- Applied to the dataset because, the dataset was highly skewed and was not following gaussian distribution

Standard Scaler :

- Applied to the dataset for PCA

Feature reduction

-

Curse of Dimensionality

The total number of features are 43 features.

Techniques to Dimensionality reduction:

- Correlation Analysis
- PCA

Correlation Analysis

-

Removing highly correlated features with each other

Correlated features will not always worsen your model, but they will not always improve it either.

- Make the learning algorithm faster
- Interpretability of your model
- The number of features reduced to 33.

spkts	sbytes	0.9657497410287414
spkts	sloss	0.9736439932787799
dpkts	dbytes	0.9764185516958216
dpkts	dloss	0.9815064328008422
sbytes	sloss	0.99502719113184
dbytes	dloss	0.9971088501020646
sinpkt	is_sm_ips_ports	0.9445057600994802
swin	dwin	0.9601246970559344
tcprtt	synack	0.9394732071062888
ct_srv_src	ct_dst_src_ltm	0.9337952137616565
ct_srv_src	ct_srv_dst	0.9778491535974652
ct_dst_ltm	ct_src_dport_ltm	0.9604008284955233
ct_dst_ltm	ct_src_ltm	0.9322524473427766
ct_src_dport_ltm	ct_dst_sport_ltm	0.9116374681078989
ct_src_dport_ltm	ct_src_ltm	0.9331720623302827
ct_dst_src_ltm	ct_srv_dst	0.9410468630509295
is_ftp_login	ct_ftp_cmd	0.9943410042026887

Principal Component Analysis

Applied explained variance for 99% to retain maximum information.

The number of features reduced to 29 with target variable.

Explained_variance =0.99

Datasets

-

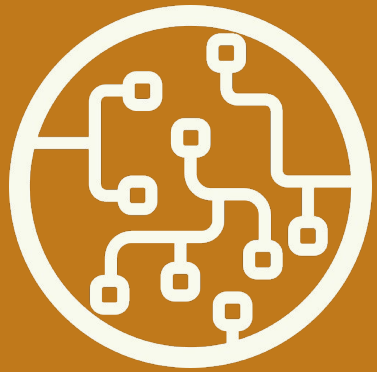
Four datasets with combination

- Dataset without any preprocessing(X):
- Dataset after applying MinMax scaler(X_mm)
- Dataset after applying MinMax scaler and correlation analysis(X_mm_corr)
- Dataset after applying Principal component analysis(X_pca)

Machine Learning Models

-

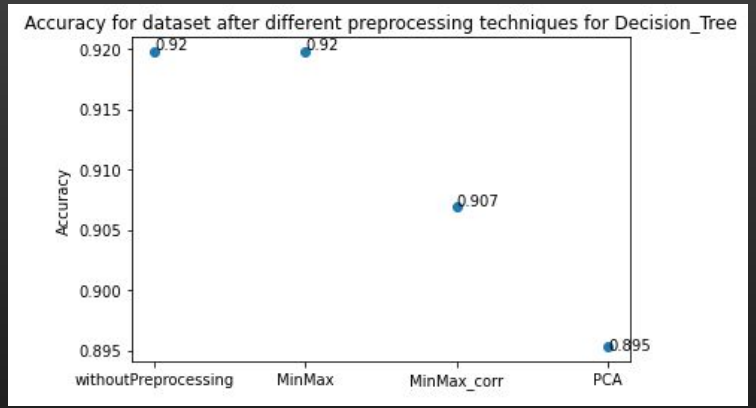
With cross validation with fold = 3



ML model 1 : Decision trees

Decision Trees are a non-parametric supervised learning method used for classification.

Dataset	Accuracy on training set	Accuracy on test set
Without Pre Processing	0.9197	0.9194
With MinMax Scaling	0.9198	0.9195
With MinMax Scaling + Correlation Analysis	0.9068	0.9064
With PCA	0.8945	0.8938



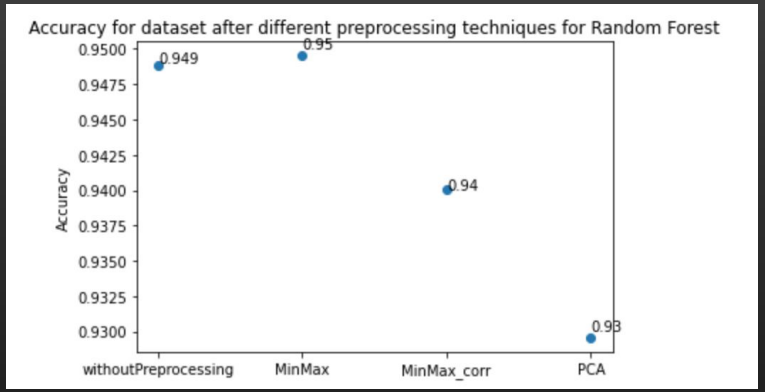
ML model 2: Random Forest

-

Combination of many decision trees

- Random Forest = Simplicity of DT + Very Good Accuracy
- Performance on dataset:

Dataset	Accuracy on training set	Accuracy on test set
Without PreProcessing	0.9487	0.9476
With MinMax Scaling	0.9755	0.9479
With MinMax Scaling + Correlation analysis	0.9697	0.9400
With PCA	0.9840	0.9295

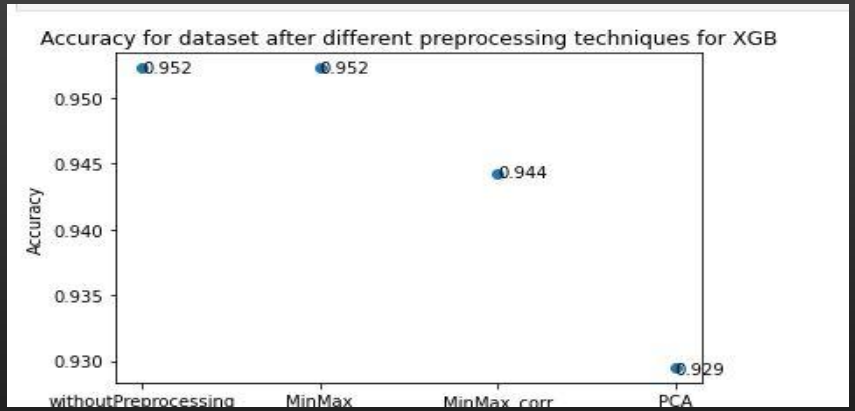


ML model 3: XGBoost

Performance on the dataset:

Dataset	Train Accuracy	Test Accuracy
Without preprocessing	0.974	0.952
With MinMax scaler applied	0.974	0.952
With MinMax scaler + Correlation analysis	0.968	0.944
With PCA	0.980	0.929

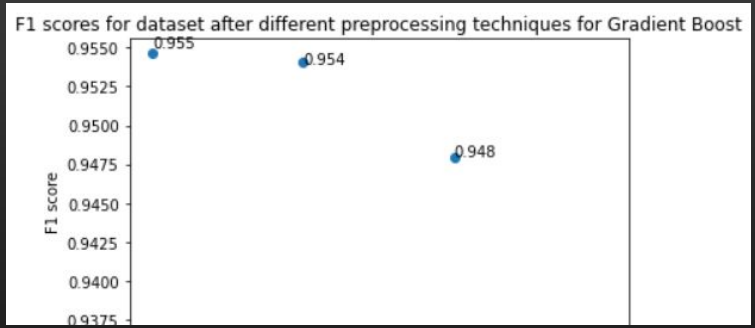
Accuracy of XGBoost on different dataset



ML model 4: Gradient Boosting

Gradient boosting is a machine learning technique used in regression and classification tasks, among others.

Dataset	Accuracy on training set	Accuracy on test set
Without Pre Processing	0.94	0.9404
With MinMax Scaling	0.94	0.943
With MinMax Scaling + Correlation Analysis	0.93	0.934
With PCA	0.92	0.9154



Conclusion

Findings of this study:

1) Gradient Boosting works best with MinMax scaling
2) Random forest shows good accuracy on the unprocessed dataset, which is not accurate since the features are not at the same scale.
3) XGBoost shows that it is overfitting the dataset, which shows that XGBoost is prone to overfitting.

Future works

- Non linear log scaling can be applied and accuracies can be checked
- Isolation trees can be applied for anomalies based algorithm


THANK YOU!