Performance evaluation of AI models for anomaly detection in network traffic

Problem: evaluate and compare the performance of multiple machine learning and deep learning models in detecting security anomalies, given some network information.

Not only determining if there have been an attack or not (***binary classification***), but also which attacks has been performed (***multi class classification***).
Setting: we are going to use the **UNSW-NB15** dataset (https://research.unsw.edu.au/projects/unsw-nb15-dataset) that describes network activities both under normal condition and also under some attacks.

- The dataset has been created by simulating real attacks in a controlled lab environment and collecting the resulting network traffic. They then labeled it based on the attack performed, extracted meaningful features from packet-level data, and released it for security research.

Some past works:

- https://arxiv.org/pdf/2101.05067
- https://www.researchgate.net/publication/304847859_The_evaluation_of_Network_Anomaly_Detection_Systems_Statistical_analysis_of_the_UNSW-NB15_data_set_and_the_comparison_with_the_KDD99_data_set

Model evaluation: after some data pre-processing and data cleaning (discretization and normalization of features), we would like to detect, given the network activities, if an attack has been performed or not. Some models we thought for doing it are:

- Logistic Regression
- Random Forest
- Neural networks

Evaluation metrics: after training the model we can ***perform inference within many independent replications*** and obtain the value for this metrics (with **confidence intervals**):

- Accuracy
- Precision (% of predicted attacks that are actually real)
- Recall (% of actual attacks detected)
- F1-score
- False positive rate
- Inference time

Then we can do some **paired t-test** to verify if the differences between the models are statistically significant and if we can state that one model is better than another model.

To end, verify which features are most important and any correlations with false positive rate and other metrics.