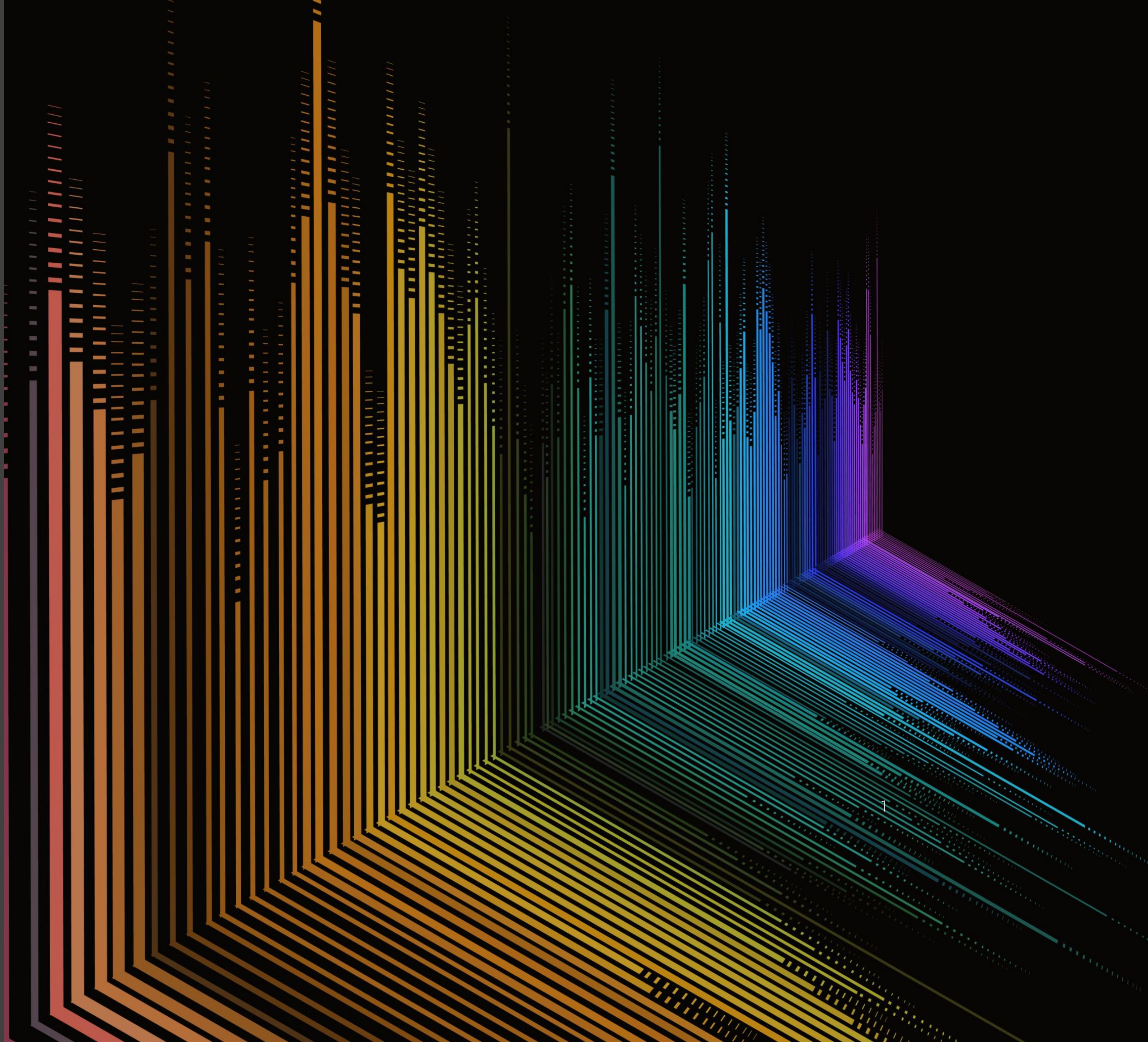
Università di Pisa

ANALYSIS ON UNSW-NB15 DATASET

ARTIFICIAL INTELLIGENCE FOR CYBERSECURITY COURSE

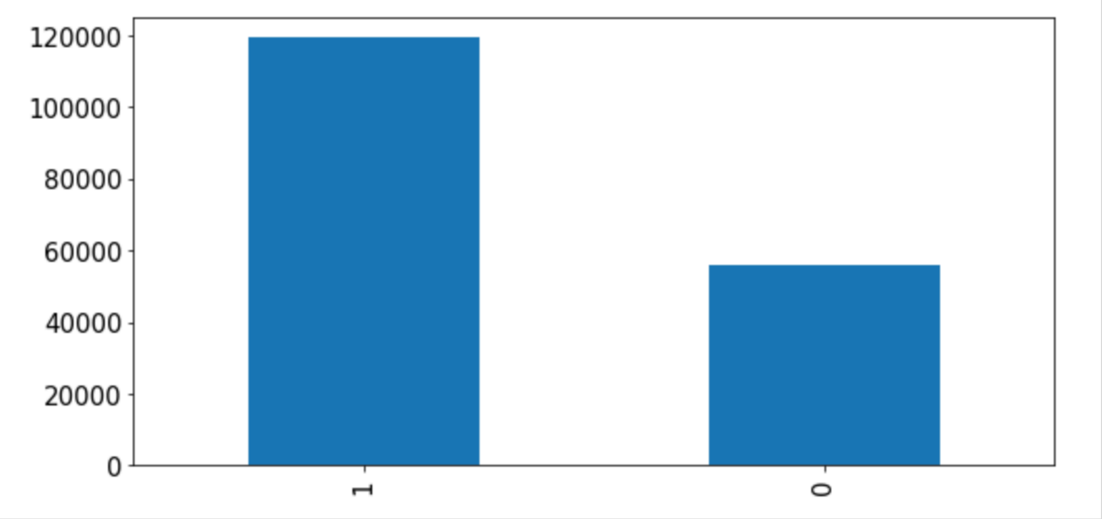
A.Y. 21/22

Authors:
Francesco Carli
Gianluca Boschi

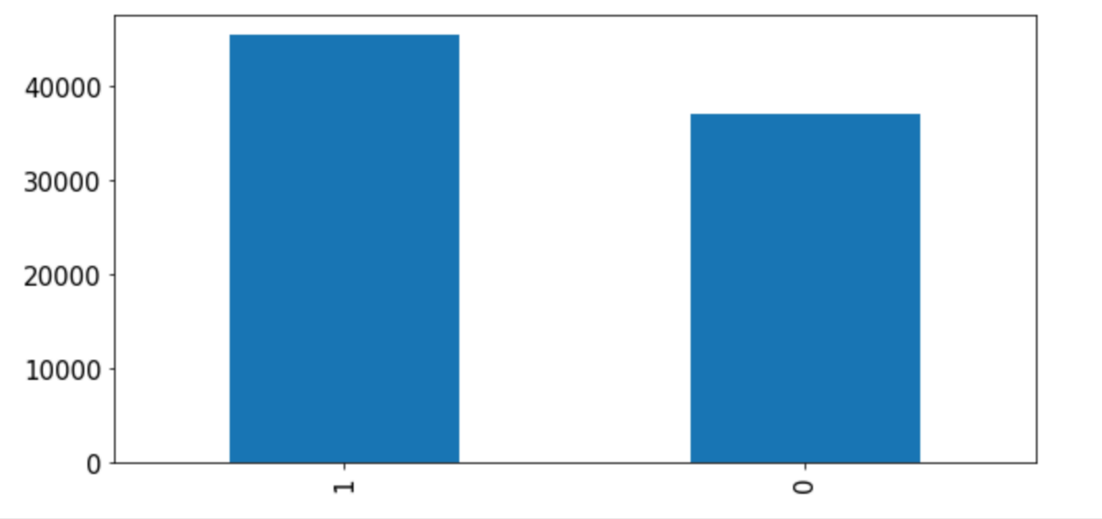


- UNSW-NB15 dataset contains a hybrid of **real modern normal activities and synthetic attacks** of nine different categories: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms.
- Each data object has 49 features, including a class label for **binary classification**.

- Several feature selection techniques have been carried out to select **42 features** out of 49.
- Partition of the dataset configured as a **training set** and **a test set**, respectively with **175,341** records and **82,322** records.



Distribution in training set



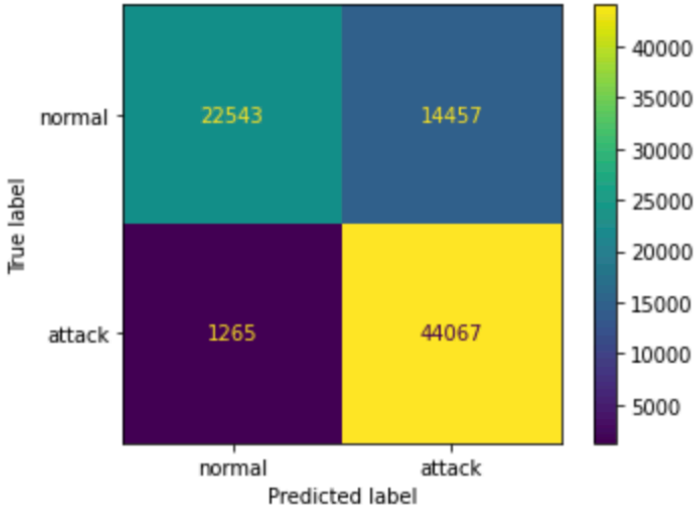
Distribution in test set

Preprocessing

- Removing of attack_cat feature from training set and test set.
- Splitting of training set and test set in train_X, train_Y, test_X and test_Y.
- Removing null records.
- Transforming nominal/categoric features in binary features.
- Adding missing binary features with value at 0 in test set.
- Removing from test set those features that are not present in training set.
- Normalization of numeric features using z-score function.
- Sorting features of training set and test set in alphabetical order.

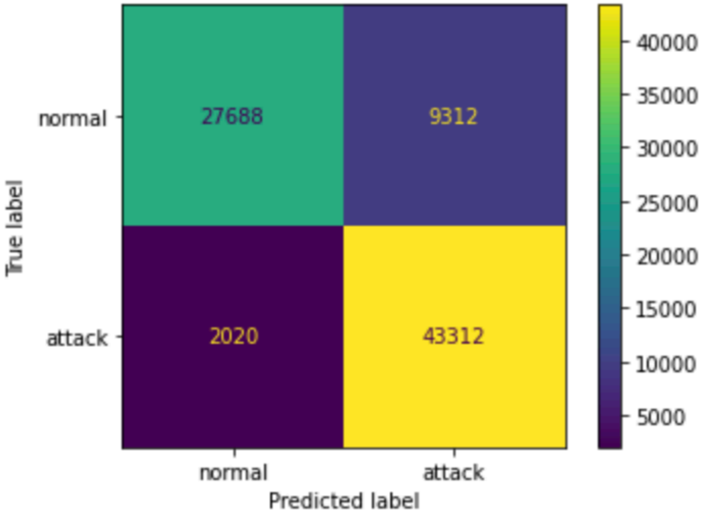
Logistic Regression

	precision	recall	f1-score	support
normal	0.95	0.61	0.74	37000
attack	0.75	0.97	0.85	45332
accuracy			0.81	82332
macro avg	0.85	0.79	0.80	82332
weighted avg	0.84	0.81	0.80	82332



Decision Tree

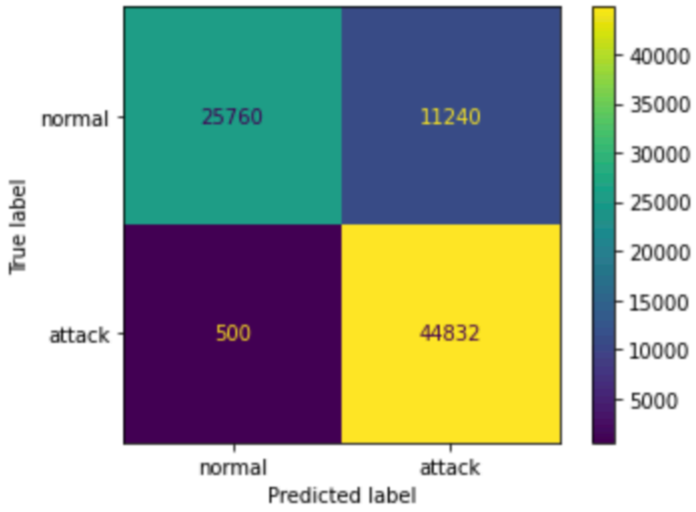
	precision	recall	f1-score	support
normal	0.93	0.75	0.83	37000
attack	0.82	0.96	0.88	45332
accuracy			0.86	82332
macro avg	0.88	0.85	0.86	82332
weighted avg	0.87	0.86	0.86	82332



Neural Network

(Two intermediate layer respectively with 10 and 5 neurons)

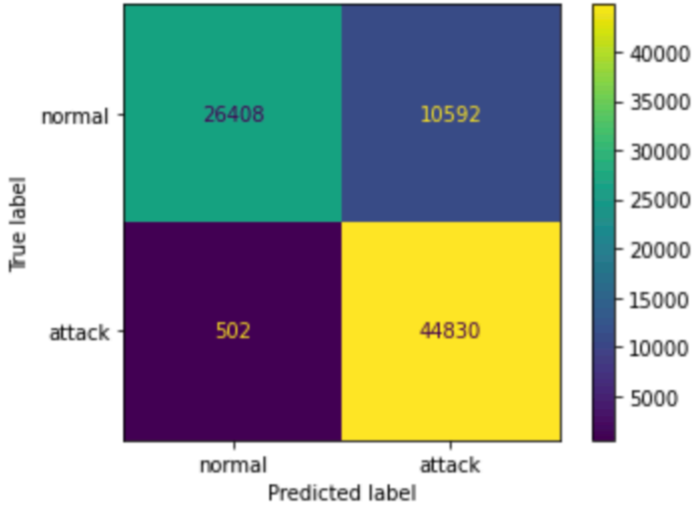
	precision	recall	f1-score	support
normal	0.98	0.70	0.81	37000
attack	0.80	0.99	0.88	45332
accuracy			0.86	82332
macro avg	0.89	0.84	0.85	82332
weighted avg	0.88	0.86	0.85	82332



Random Forest

(Every tree has a max depth of 20)

	precision	recall	f1-score	support
normal	0.98	0.71	0.83	37000
attack	0.81	0.99	0.89	45332
accuracy			0.87	82332
macro avg	0.90	0.85	0.86	82332
weighted avg	0.89	0.87	0.86	82332



PRECISION				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.95	0.93	0.98	0.98
1 (Attack)	0.75	0.82	0.91	0.80

Gives information about false positive or false negative, based on the class of interest

RECALL				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.61	0.75	0.71	0.70
1 (Attack)	0.97	0.96	0.99	0.99

$$Sensitivity = \frac{TP}{P}$$

$$Specificity = \frac{TN}{N}$$

Gives the true positive rate and true negative rate.

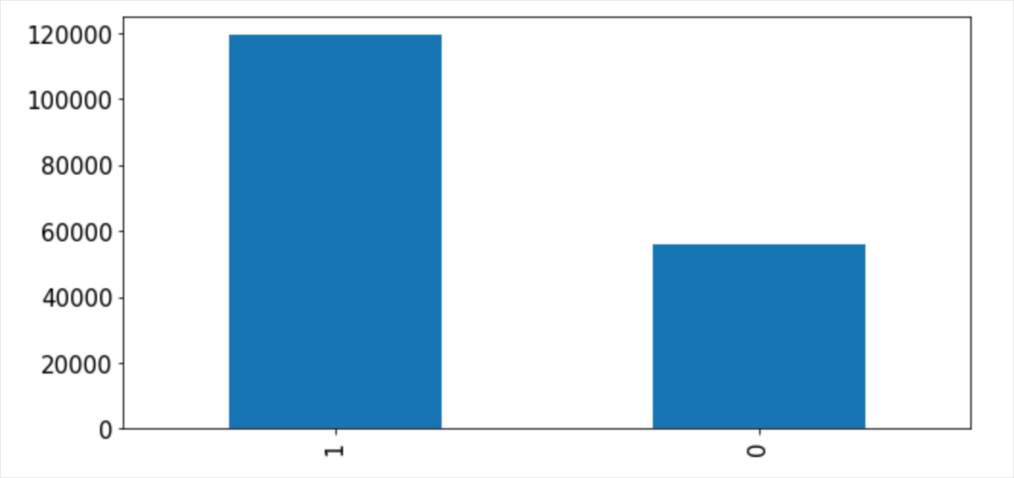
F1 – SCORE				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.74	0.83	0.83	0.81
1 (Attack)	0.85	0.86	0.89	0.88

$$F1Score = 2\,\frac{Precision \cdot Recall}{Precision + Recall}$$

Is the harmonic mean of precision and recall. Combines precision and recall into a single metric.

Rebalancing

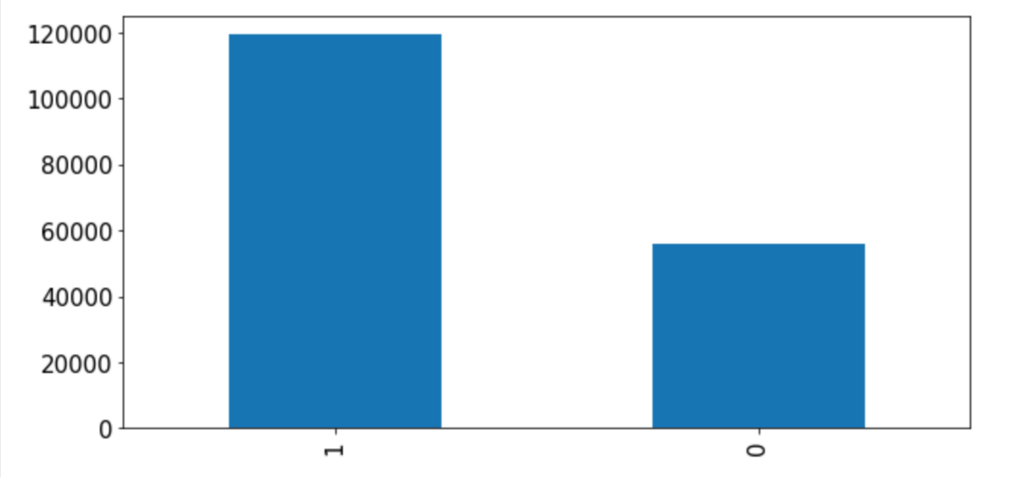
- **The training set is unbalanced** on the class label: 119,341 records with label 1 (majority class) and 56,000 records with label 0 (minority class).
- This could be a problem since the previous models and their result will be affected and **biased towards the majority class.**
- Two different rebalancing techniques have been used to solve this issue.



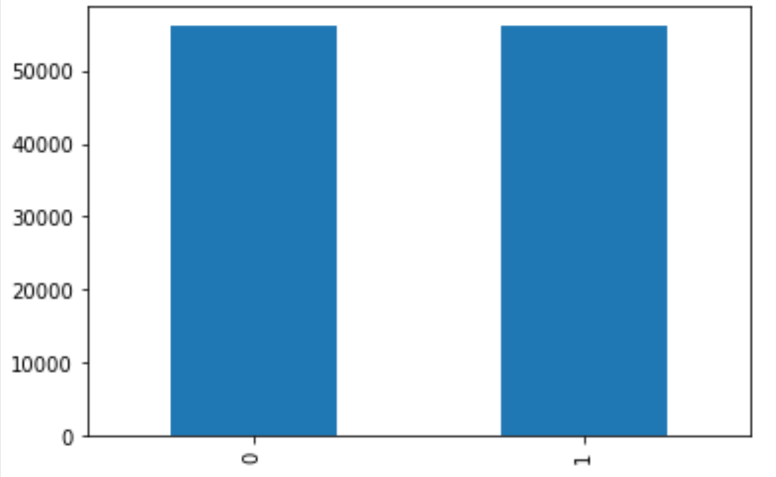
Distribution in training set

Undersampling

- It is a rebalancing technique that keeps all the records of the minority class and **decreases the size of the majority class.**
- The simplest technique is **random undersampling**, in which some records of the majority class are removed in a random way.



Unbalanced training set



Training set with undersampling

Undersampling results

PRECISION				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.89	0.93	0.95	0.93
1 (Attack)	0.80	0.85	0.88	0.87

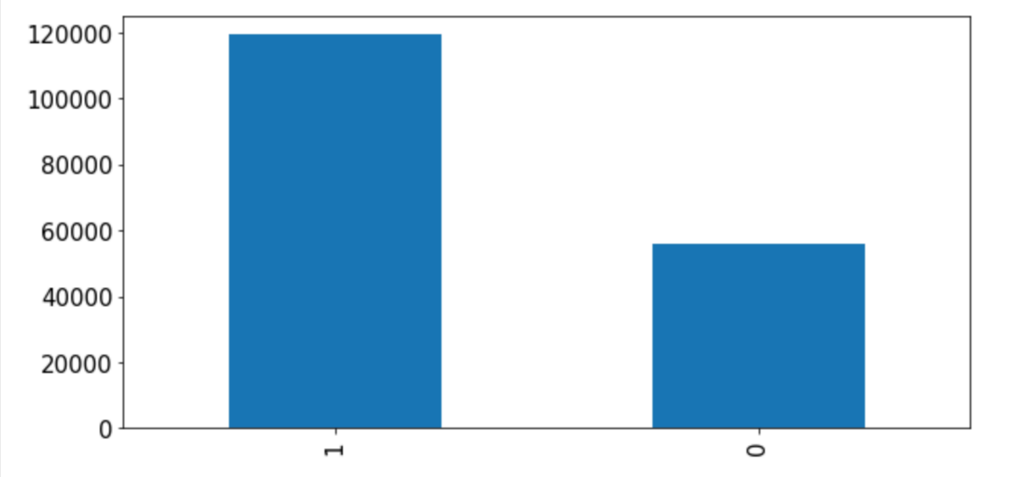
RECALL				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.72	0.80	0.84	0.82
1 (Attack)	0.93	0.95	0.97	0.95

F1 – SCORE				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.80	0.86	0.89	0.87
1 (Attack)	0.86	0.90	0.92	0.91

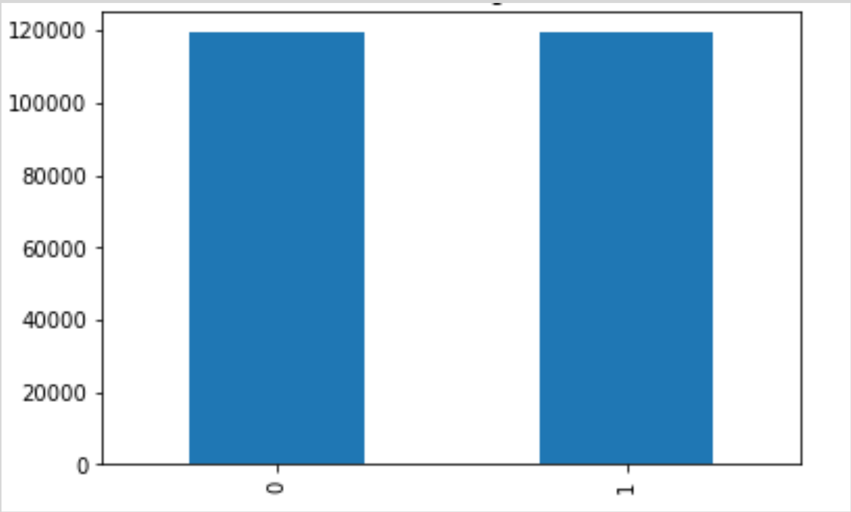
ACCURACY	
Logistic Regression	0,834
Decision Tree	0,882
Random Forest	0,910
Neural Network	0,891

Oversampling

- It is a rebalancing technique that **increases the number of instances of the minority class**, using different ways to create new artifical or synthetic instances.
- The simplest technique is **random oversampling**, in which several new instances are added to the minority class, chosing in a random way with replacement instead of duplicating them.



Unbalanced training set



Training set with oversampling

Oversampling results

PRECISION				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.90	0.94	0.96	0.95
1 (Attack)	0.80	0.82	0.87	0.87
RECALL				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.72	0.74	0.82	0.82
1 (Attack)	0.93	0.96	0.97	0.96
F1 – SCORE				
Binary Label	Logistic Regression	Decision Tree	Random Forest	Neural Network
0 (Normal)	0.80	0.83	0.89	0.88
1 (Attack)	0.86	0.89	0.92	0.91

ACCURACY	
Logistic Regression	0,835
Decision Tree	0,862
Random Forest	0,904
Neural Network	0,900

K-Fold Cross Validation

- It has been used 10-fold cross validation to evaluate the performance of models
- Rebalancing increase the ratio of negative istances (0, *normal*) so we have an **increment on specificity and precision,** so the model has **now less falses**, because it works better on negative istances now.

UNDERSAMPLING								
	Logistic Regression		Decision Tree		Random Forest		Neural Network	
	Unbalanced	Rebalanced	Unbalanced	Rebalanced	Unbalanced	Rebalanced	Unbalanced	Rebalanced
Precision	0.897	0.928	0.933	0.945	0.940	0.964(+ 0.024)	0.941	0.963(+ 0.022)
Sensitivity	0.956	0.902	0.936	0.921	0.944	0.918(-0.026)	0.935	0.905(-0.030)
Specificity	0.766	0.854	0.865	0.89	0.880	0.933(+ 0.053)	0.881	0.930(+ 0.049)

OVERSAMPLING								
	Logistic Regression		Decision Tree		Random Forest		Neural Network	
	Unbalanced	Rebalanced	Unbalanced	Rebalanced	Unbalanced	Rebalanced	Unbalanced	Rebalanced
Precision	0.897	0.928	0.933	0.933	0.940	0.959(+ 0.019)	0.941	0.963(+ 0.022)
Sensitivity	0.956	0.903	0.936	0.937	0.944	0.923(-0.021)	0.935	0.908(-0.027)
Specificity	0.766	0.856	0.865	0.864	0.880	0.923(+ 0.043)	0.881	0.932(+ 0.051)

T-test

- It is a **statistic test** and it can be used to determine if two means of two sets of data are statistically different from each other.
- The **two best models** for each set of metrics have been selected in order to **compare their results** obtained from 10-fold cross validation.
- The **confidence value α** has been set to **0.05**

SENSITIVITY		
Classifier	**Mean of 10 values**	**Values from 10-fold cross validation**
Logistic Regression	0.959	0.906, 0.928, 0.935, 0.953, 0.986, 0.991, 0.986, 0.972, 0.985, 0.944
Random Forest	0.944	0.841, 0.902, 0.903, 0.923, 0.971, 0.980, 0.972, 0.969, 0.992, 0.985
T-Test Result	**p-value: 0.427 → Cannot reject the null hypothesis Distributions are not statistically different**	

T-Test: rebalanced dataset

For the rebalanced dataset we considered the **accuracy** and we check if some classifier is better than an other one.

ACCURACY (UNDERSAMPLING)		
Two best classifier	Mean of 10 values	Values from 10-fold cross validation
Neural Network	0.914	0.84, 0.896, 0.901, 0.915, 0.924, 0.934, 0.967, 0.951, 0.903, 0.909
Random Forest	0.92	0.867, 0.912, 0.909, 0.931, 0.932, 0.934, 0.972, 0.967, 0.904, 0.904
T-Test Result	p-value: 0.540 → Cannot reject null hypothesis Distributions are not statistically different	

ACCURACY (OVERSAMPLING)		
Two best classifier	Mean of 10 values	Values from 10-fold cross validation
Decision Tree	0.911	0.898, 0.93, 0.93, 0.946, 0.895, 0.882, 0.979, 0.966, 0.845, 0.835
Random Forest	0.923	0.871, 0.914, 0.912, 0.934, 0.928, 0.929, 0.974, 0.968, 0.898, 0.901
T-Test Result	p-value: 0.508 → Cannot reject null hypothesis Distributions are not statistically differen	

References

1. Moustafa, Nour, and Jill Slay. "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)." *Military Communications and Information Systems Conference (MilCIS)*, 2015. IEEE, 2015.
2. Moustafa, Nour, and Jill Slay. "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 dataset and the comparison with the KDD99 dataset." *Information Security Journal: A Global Perspective* (2016): 1-14.
3. Moustafa, Nour, et al. "Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks." *IEEE Transactions on Big Data (2017).*
4. Moustafa, Nour, et al. "Big data analytics for intrusion detection system: statistical decision-making using finite dirichlet mixture models. "*Data Analytics and Decision Support for Cybersecurity. Springer, Cham, 2017.* 127-156.
5. Sarhan, Mohanad, Siamak Layeghy, Nour Moustafa, and Marius Portmann. NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems. In *Big Data Technologies and Applications: 10th EAI International Conference, BDTA 2020, and 13th EAI International Conference on Wireless Internet, WiCON 2020, Virtual Event, December 11, 2020, Proceedings* (p. 117). Springer Nature.

Thank you for your attention !