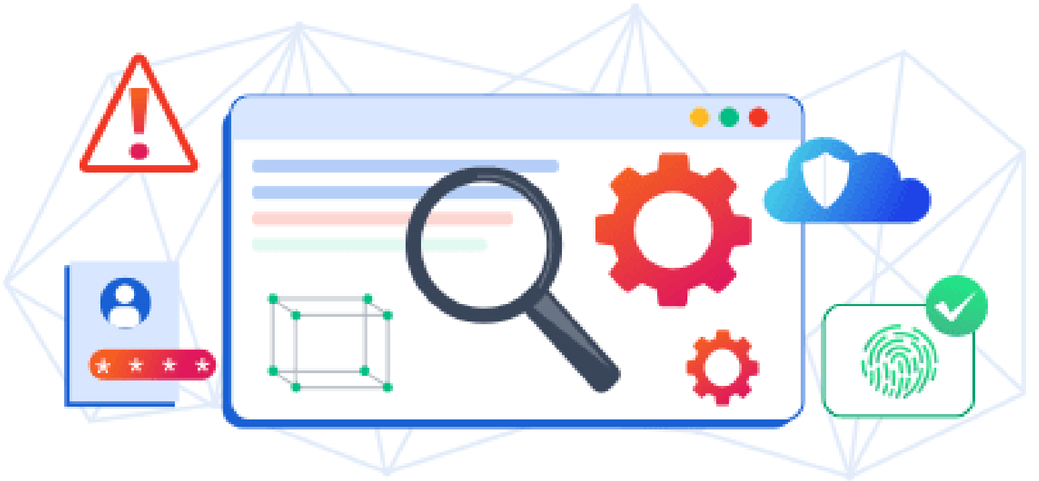
HANOI UNIVERSITY OF
SCIENCE AND TECHNOLOGY

PROJECT I
ABNORMAL NETWORK DETECTION

BUI HONG NHAT - 20204890

INTRODUCTION



Anomaly Detection

for Cyber Network Security

PROBLEM

- ABNORMAL NETWORK DETECTION IS A TYPE OF NETWORK SECURITY TO IDENTIFY AND RESPOND TO ABNORMAL NETWORK BEHAVIOR. THIS CAN INCLUDE THINGS LIKE UNAUTHORIZED ACCESS, DATA EXFILTRATION, AND DENIAL-OF-SERVICE ATTACKS.

- ANOMALY DETECTION SYSTEMS WORK BY FIRST CREATING A BASELINE OF NORMAL NETWORK BEHAVIOR. THIS IS DONE BY COLLECTING DATA ON THINGS LIKE TRAFFIC PATTERNS, USER ACTIVITY, AND DEVICE BEHAVIOR. ONCE THE BASELINE IS ESTABLISHED, THE SYSTEM CAN THEN IDENTIFY ANY DEVIATIONS FROM NORMAL BEHAVIOR AS POTENTIAL THREATS.

- IN THIS PROJECT, I JUST STOP AT MONITORING THE TRAFFIC PATTERN AND TRY TO MAKE PREDICTIONS WHETHER THE NETWORK IS ABNORMAL OR NORMAL.

DATASET

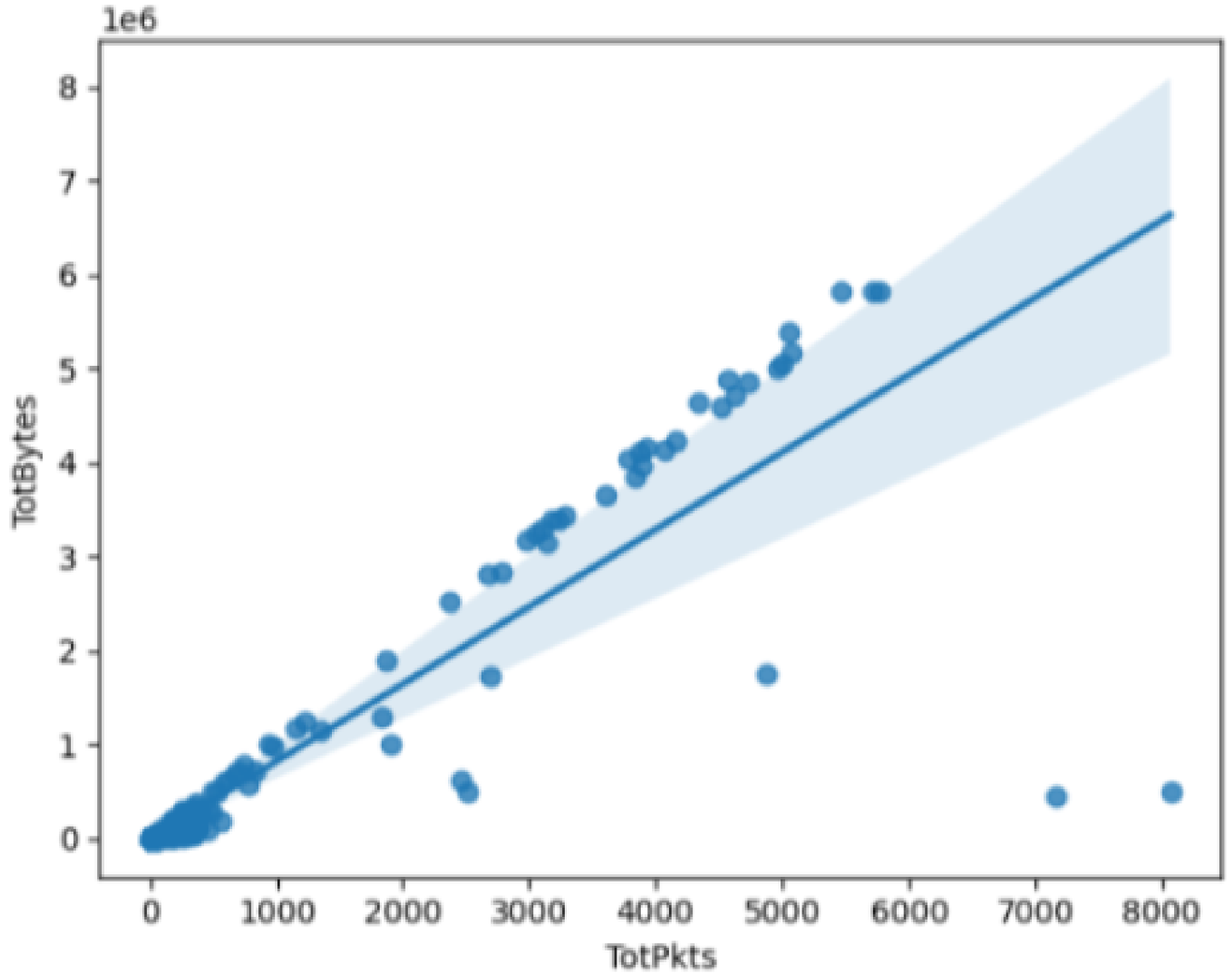
StartTime(object): The time that the protocol established	**Dur(float64):** Duration	**Proto(object):** Network Protocol	**SrcAddr(object):** The IPv4 of source address
Sport(object): Port number of source address	**Dir(object):** Direction	**DstAddr(object):** The IPv4 of destination address	**Dport(object):** Port number of destination address
STos(float64): Source type of service	**DTos(float64):** Destination type of service	**TotPkts(float64):** Total packets	**TotBytes(float64):** Total bytes
SrcBytes(float64): Total bytes from source	**Label(float64):** Network status (Normal: 0, Abnormal: 1)		

EXPLORATORY DATA ANALYSIS

EDA

HIGH CORRELATION BETWEEN ‘TOTPKTS” AND ‘TOTBYTES’ (0.91)

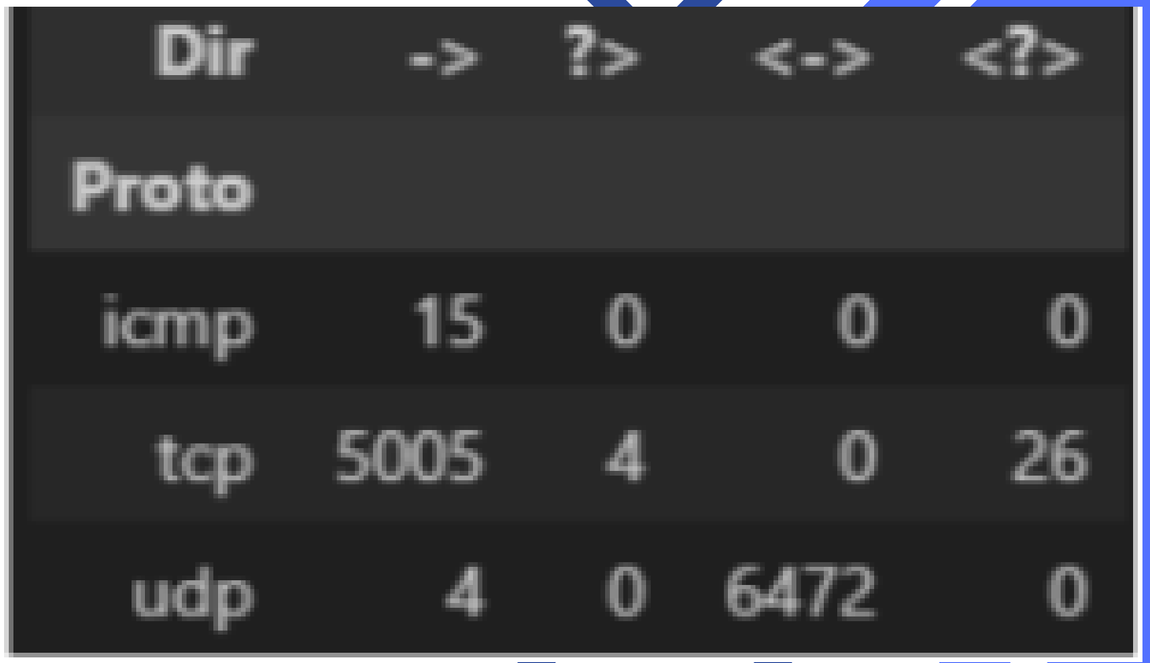
---> DROP TOTAL PACKET



EDA

+ 15/15 (100%) OF ‘ICMP’ PROTOCOL HAVE “->” DIRECTION
+ 5005/5035 (99.4%) OF ‘TCP’ PROTOCOL HAVE “->” DIRECTION
+ 6472/6476 (99.9%) OF ‘UDP’ PROTOCOL HAVE “<->” DIRECTION

---> DROP DIRECTION

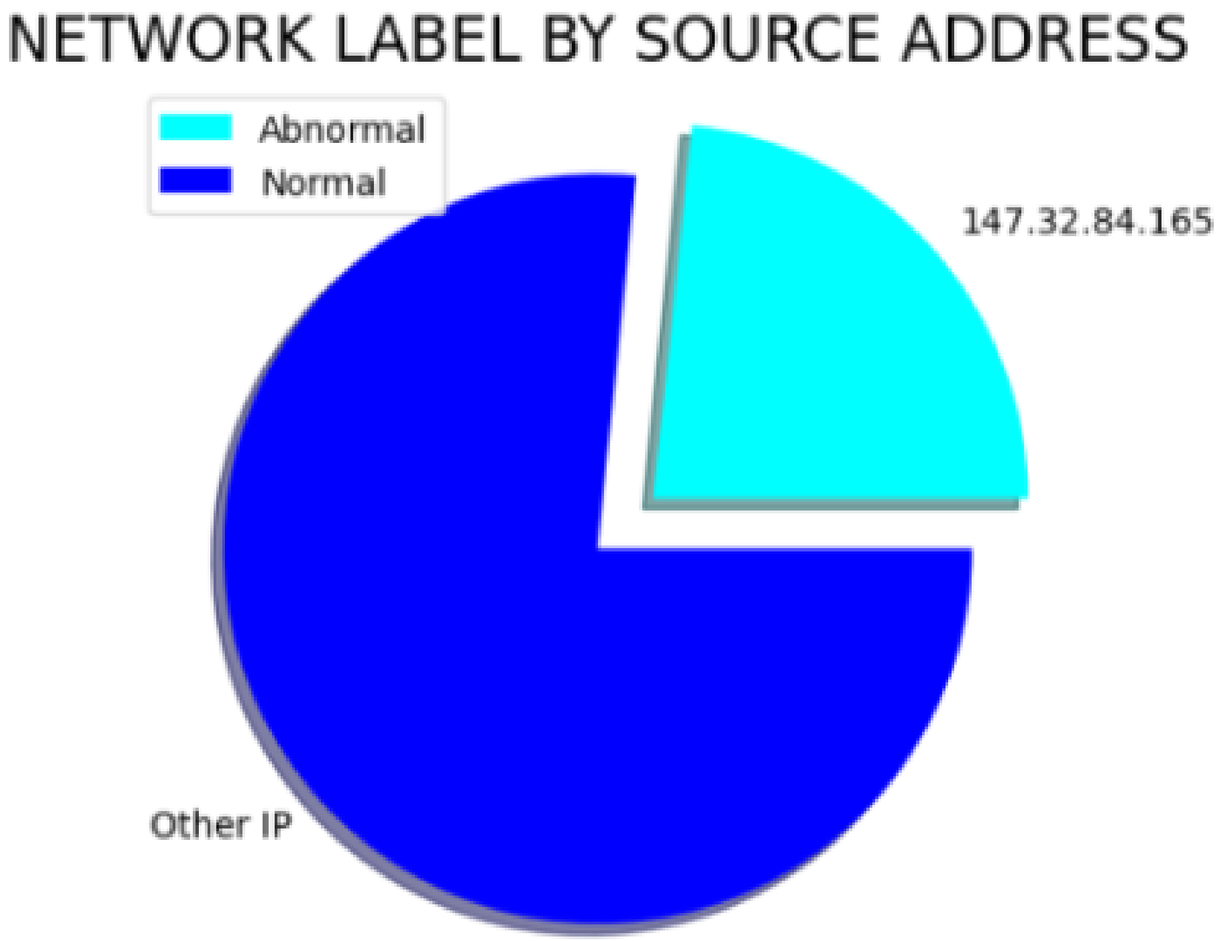


Dir	->	?>	<->	<?>
Proto				
icmp	15	0	0	0
tcp	5005	4	0	26
udp	4	0	6472	0

EDA

ALL THE ABNORMAL NETWORK LABELS COME FROM THE SOURCE ADDRESS ‘147.32.84.135’

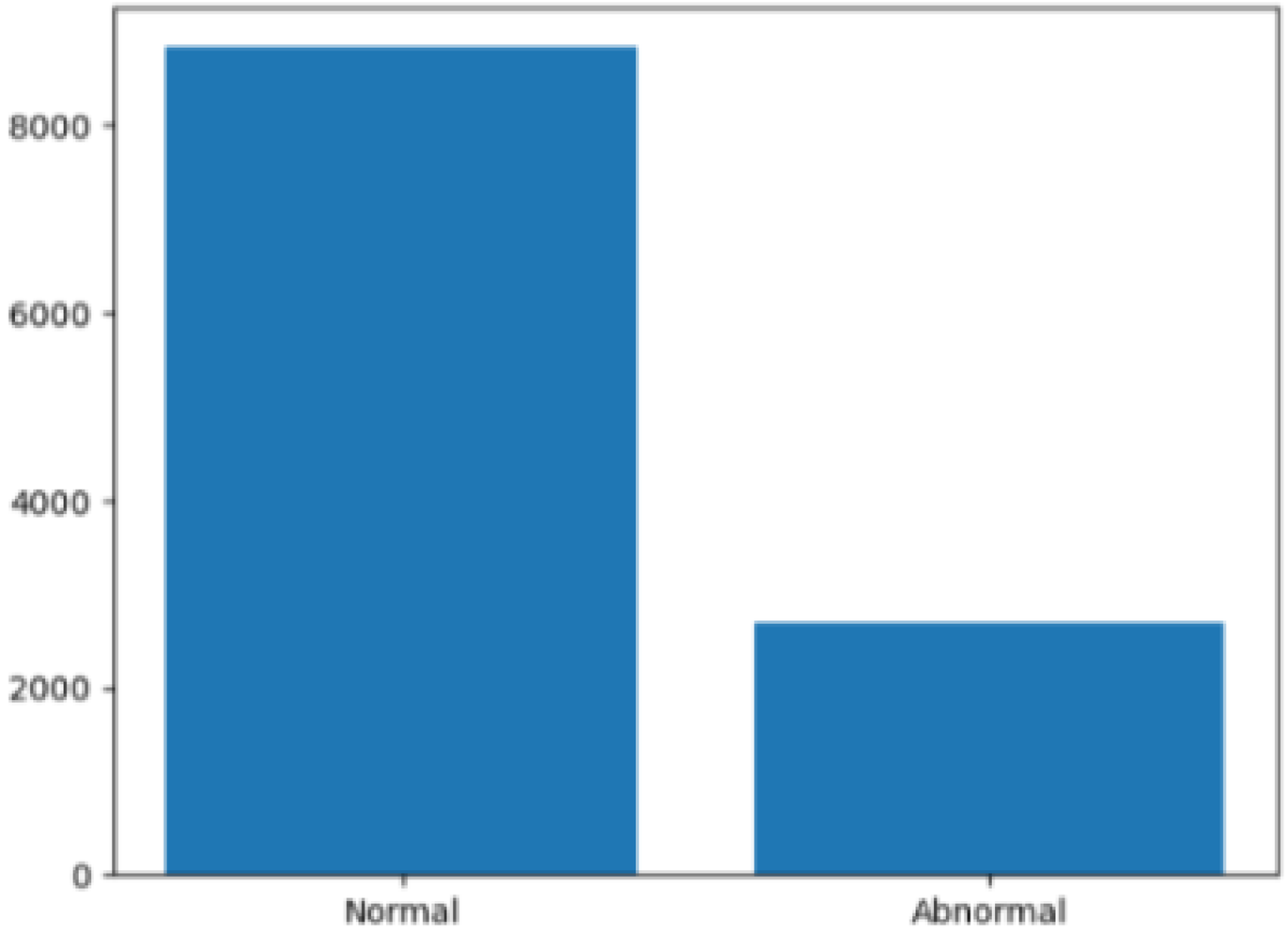
---> DROP SOURCE ADDRESS



EDA

+ 8833 (76.7%) LABELS OF THE DATASET ARE ‘NORMAL’
+ 2693 (23.3%) LABELS OF THE DATASET ARE ‘ABNORMAL’

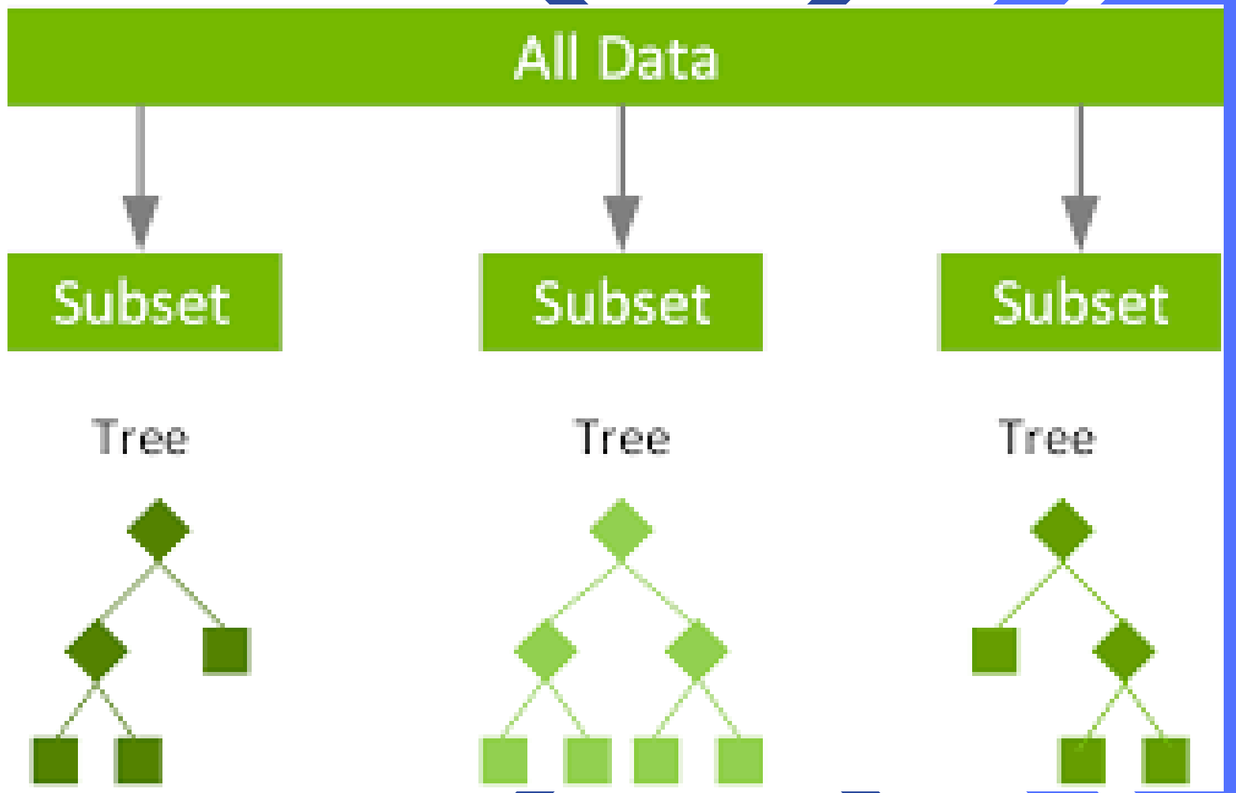
---> USE "STRATIFIED SAMPLING" METHOD TO SPLIT THE DATA



MODEL

XGBOOST

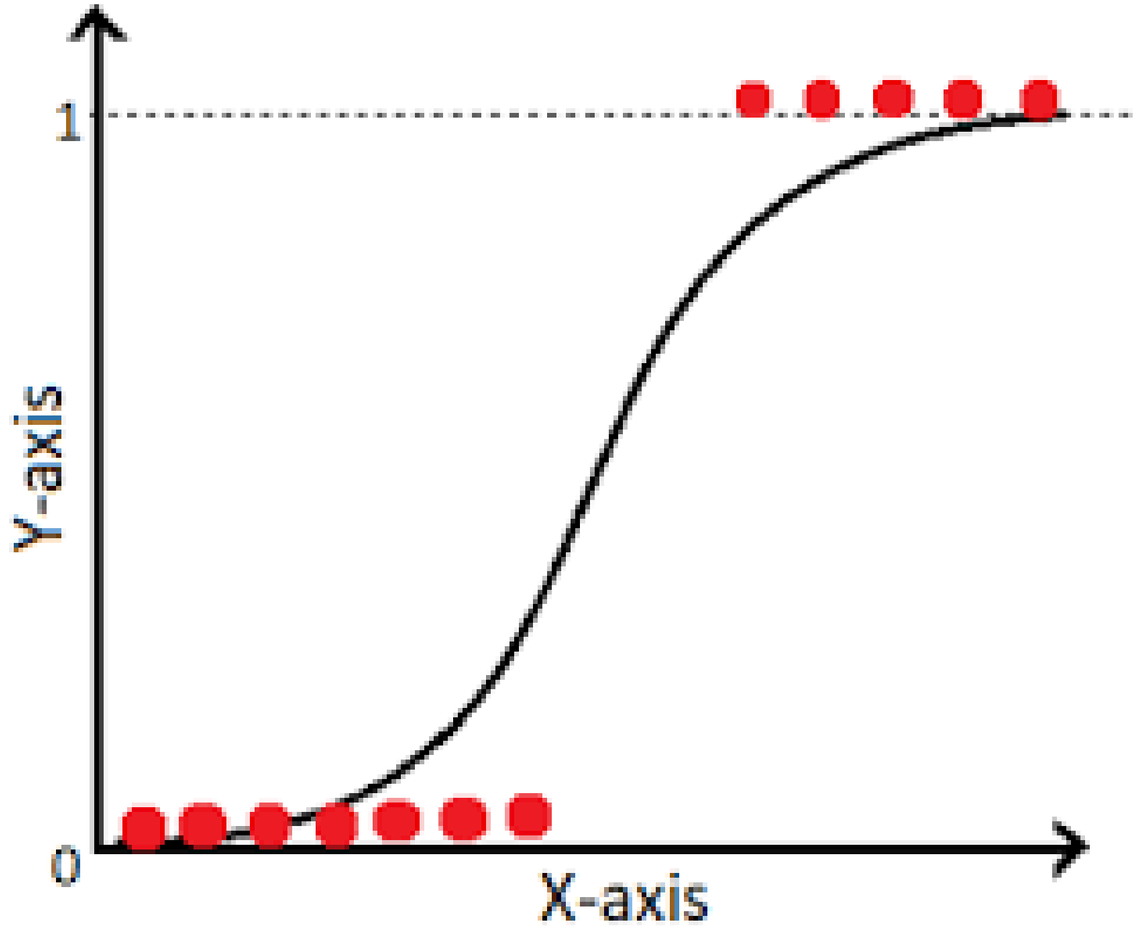
EXTREME GRADIENT BOOSTING (XGBOOST) IS A TREE-BASED ALGORITHM, WHICH MEANS THAT IT BUILDS A MODEL BY CREATING A SET OF DECISION TREES. EACH DECISION TREE IS A SIMPLE MODEL THAT CAN BE USED TO CLASSIFY OR PREDICT A VALUE. THE DECISION TREES IN XGBOOST ARE TRAINED TO MINIMIZE THE LOSS FUNCTION, WHICH IS A MEASURE OF HOW WELL THE MODEL FITS THE TRAINING DATA..



LOGISTIC REGRESSION

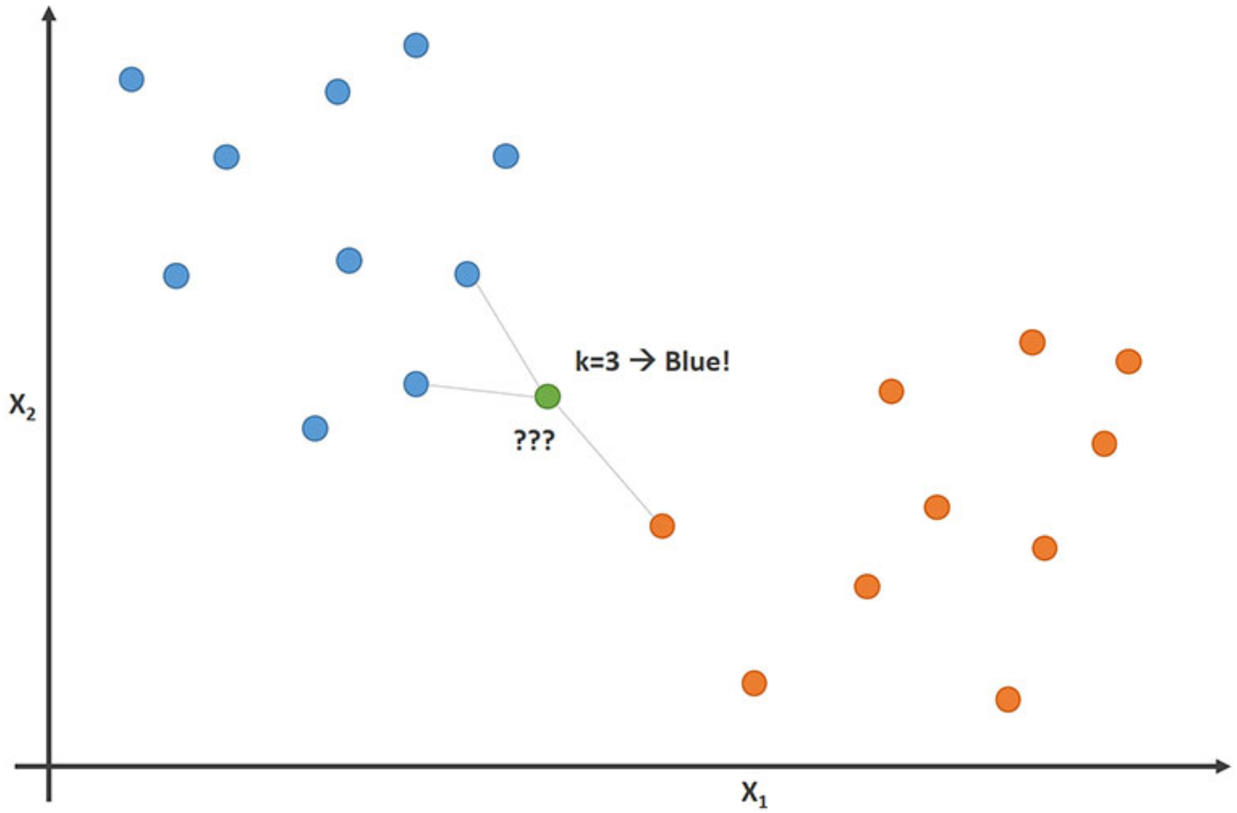
LOGISTIC REGRESSION IS A MACHINE LEARNING CLASSIFICATION ALGORITHM THAT IS USED TO PREDICT THE PROBABILITY OF CERTAIN CLASSES BASED ON SOME DEPENDENT VARIABLES. IN SHORT, THE LOGISTIC REGRESSION MODEL COMPUTES A SUM OF THE INPUT FEATURES (IN MOST CASES, THERE IS A BIAS TERM), AND CALCULATES THE LOGISTIC OF THE RESULT.

THE OUTPUT OF LOGISTIC REGRESSION IS ALWAYS BETWEEN (0, AND 1), WHICH IS SUITABLE FOR A BINARY CLASSIFICATION TASK.

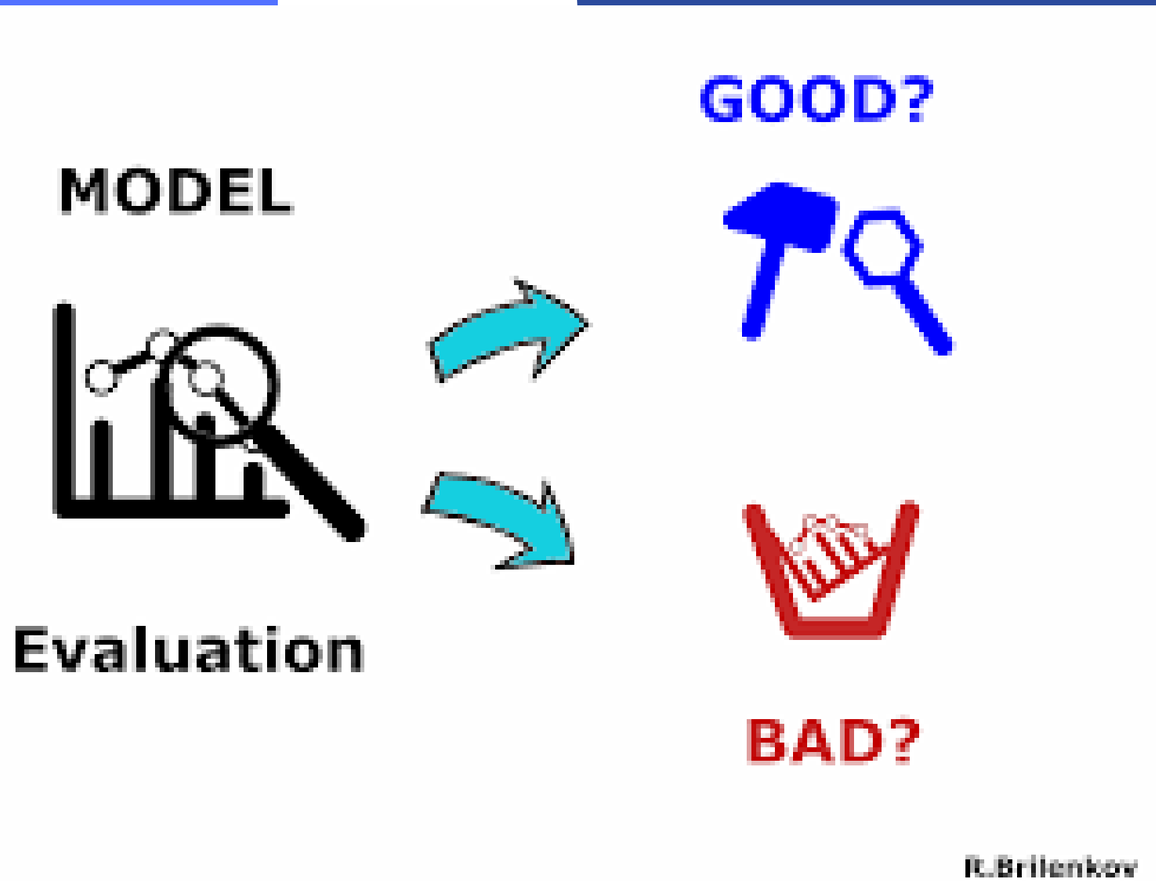


K-NEAREST NEIGHBORS

THE K-NEAREST NEIGHBORS ALGORITHM, ALSO KNOWN AS KNN, IS A NON-PARAMETRIC, SUPERVISED LEARNING CLASSIFIER, WHICH USES PROXIMITY TO MAKE CLASSIFICATIONS OR PREDICTIONS ABOUT THE GROUPING OF AN INDIVIDUAL DATA POINT. WHILE IT CAN BE USED FOR EITHER REGRESSION OR CLASSIFICATION PROBLEMS, IT IS TYPICALLY USED AS A CLASSIFICATION ALGORITHM, WORKING OFF THE ASSUMPTION THAT SIMILAR POINTS CAN BE FOUND NEAR ONE ANOTHER.



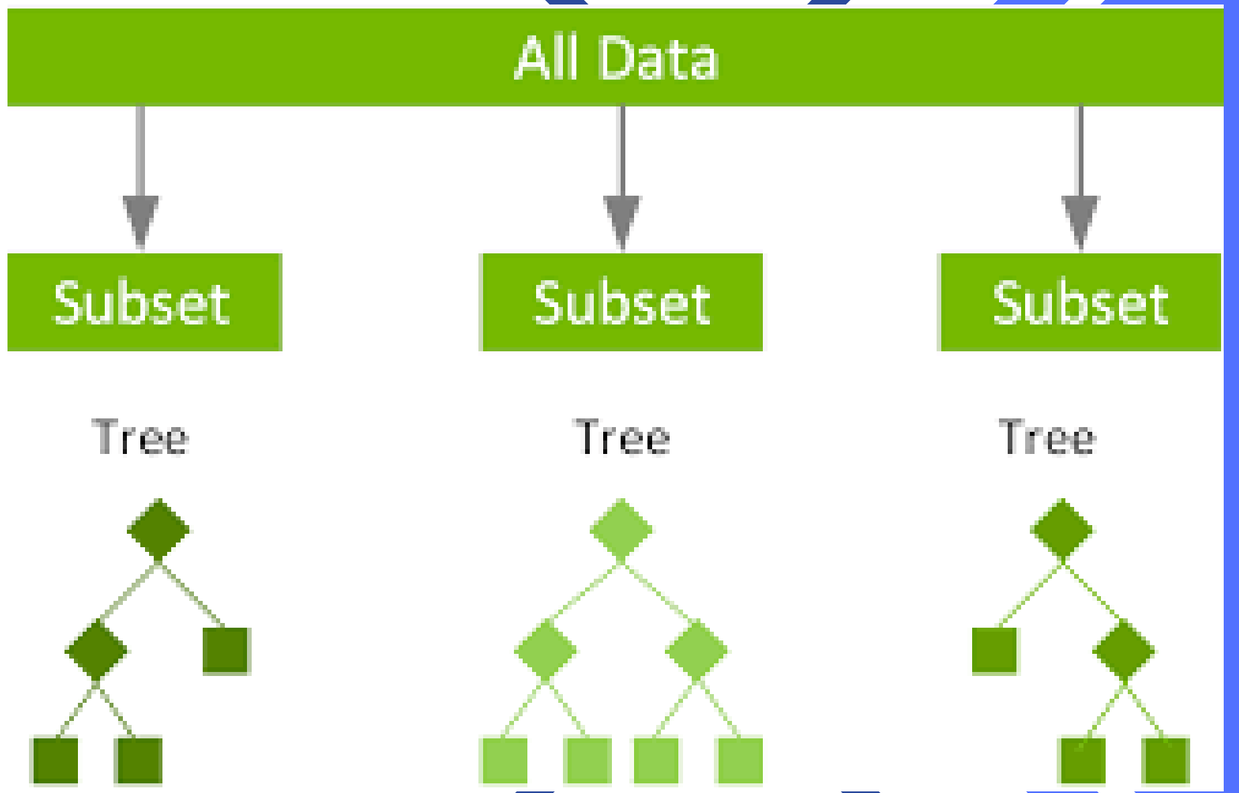
EVALUATION



XGBOOST

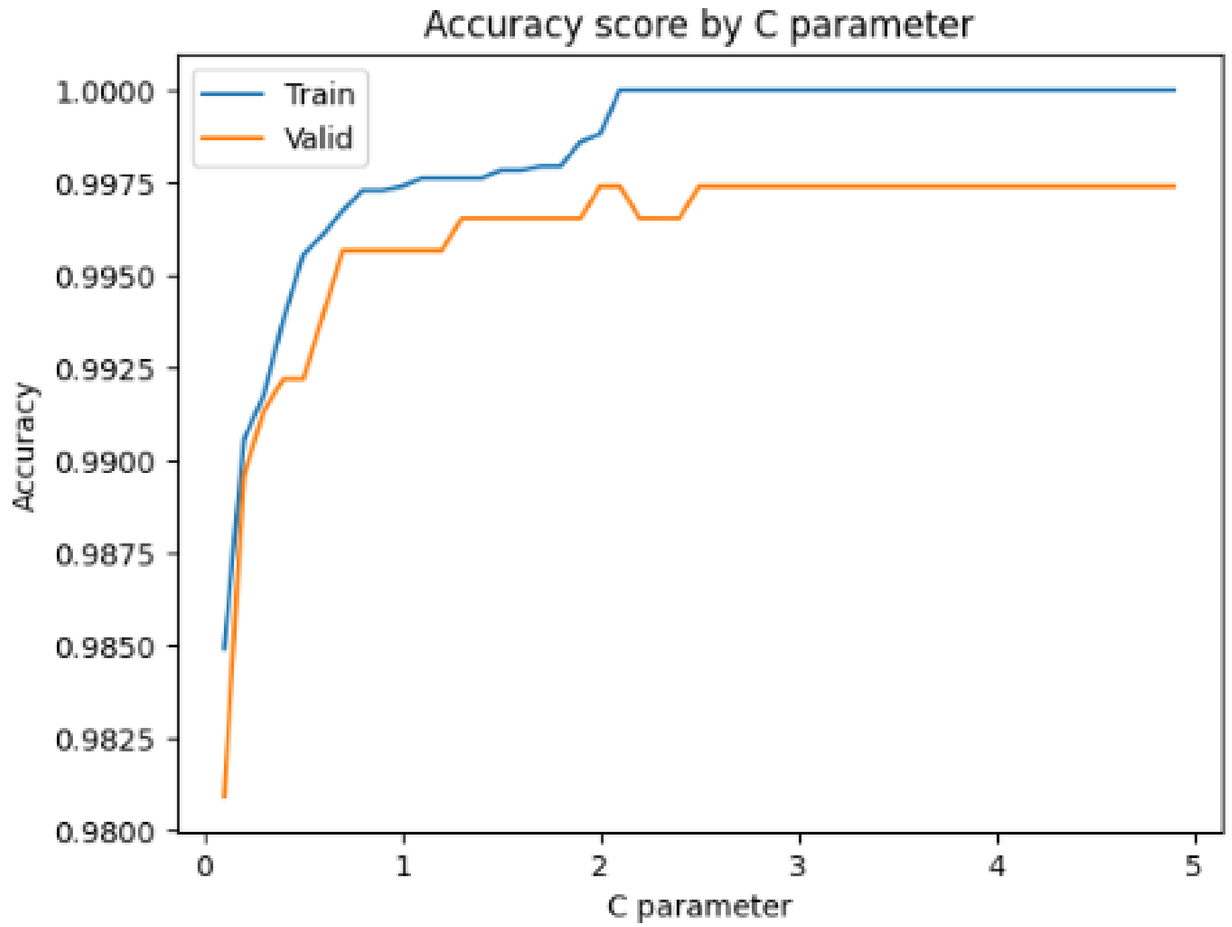
DEFAULT PARAMETERS:
· LEARNING_RATE = 0.3
· N_ESTIMATORS = 100 (NUMBER OF TREES)
· MAX_DEPTH = 6 (MAXIMUM DEPTH OF EACH TREE)

-->MAXIMUM ACCURACY SCORE FOR ALL TRAINING, VALIDATION, AND TESTING SET



LOGISTIC REGRESSION

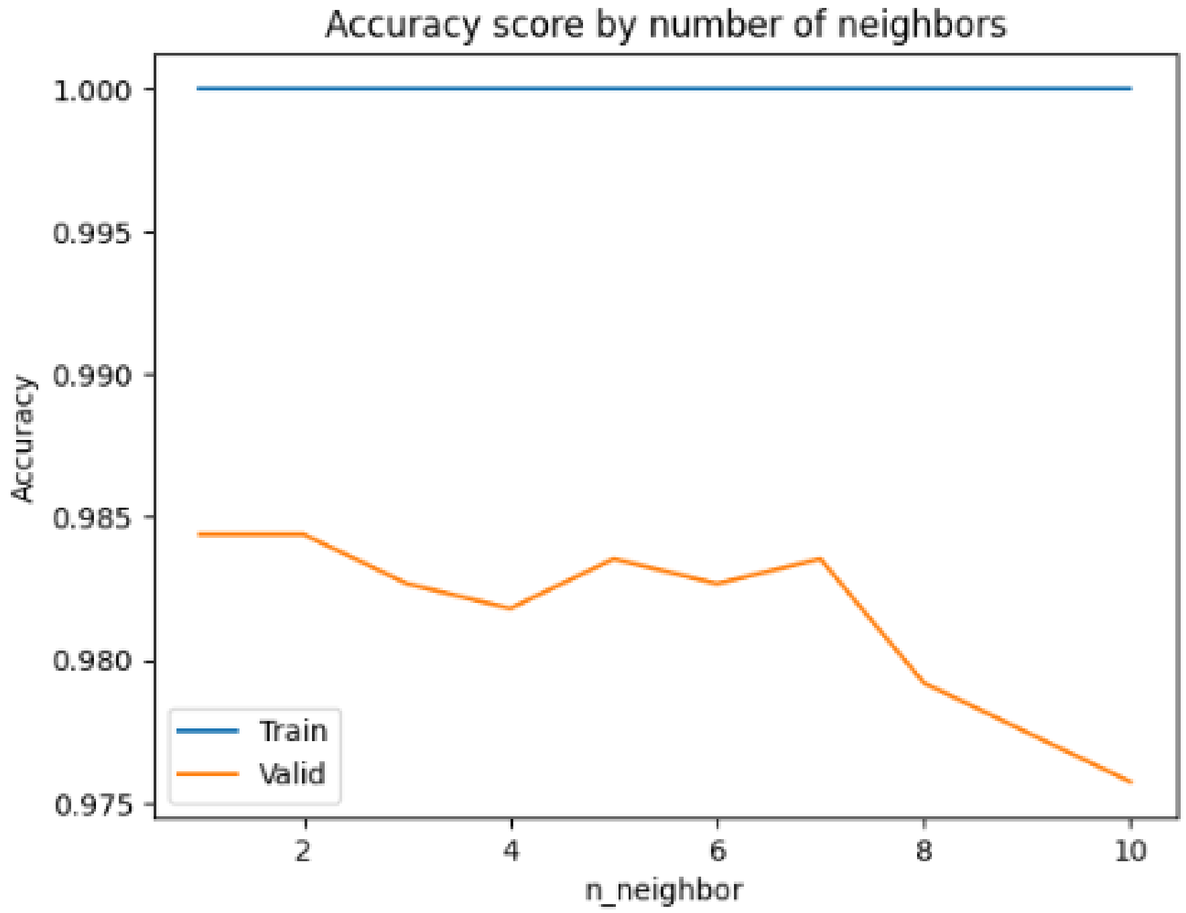
- PENALTY = 'L1' (L1 REGULARIZATION)
- SOLVER = 'LIBLINEAR' (GOOD ALGORITHM FOR SMALL DATASET)
- C = 2 (REGULARIZATION PARAMETER)



The accuracy on validation set increases and obtains its maximum when C = 2 then decreases and remains constant.

K-NEAREST NEIGHBORS

- WEIGHTS = 'DISTANCE' (WEIGHT POINT BY THE INVERSE OF THEIR DISTANCE)
- ALGORITHM = 'BRUTE' (BRUTE-FORCE ALGORITHM)
- N_NEIGHBORS= 1



The accuray on validation set obtain its maximum when n_neighbors = 1, then fluctuate and decrease with higher n_neighbors

ACCURACY

Model	Train accuracy	Test accuracy	Train time	Predict time
XGboost	1.0000	1.0000	8.7s	trivial
Logistic Regression	0.9988	0.9957	0.2s	trivial
KNN	1.0000	0.9879	trivial	0.4s

XGBoost model provided the highest accuracy on the test set (1.0), followed by Logistic Regression (0.9957) and KNN (0.9879).

THANK YOU

HANOI UNIVERSITY OF SCIENCE
AND TECHNOLOGY

BUI HONG NHAT

20204890