EEE ENSI
Student Branch

IEEE ENSI SB
Computation
Intelligence
Society
Student Chapte

GODS2.0

AS NEVER BEFORE

Reacher's Solution

19 FEB 2023

integration objects

MINOTORE

VENARI SECURITY

FEBRUARY 18th – 19

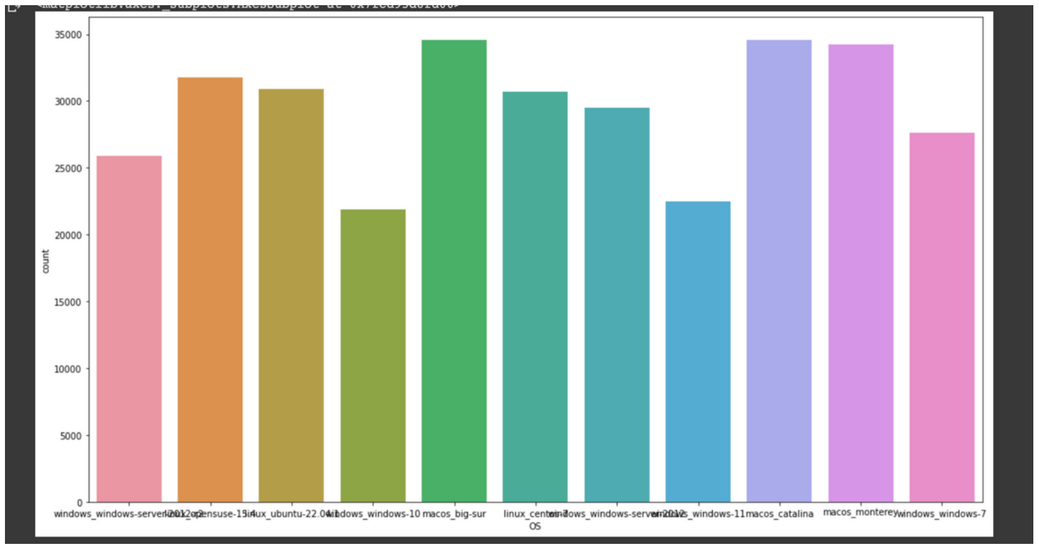
THE PROBLEM

Os & Browser Identification using

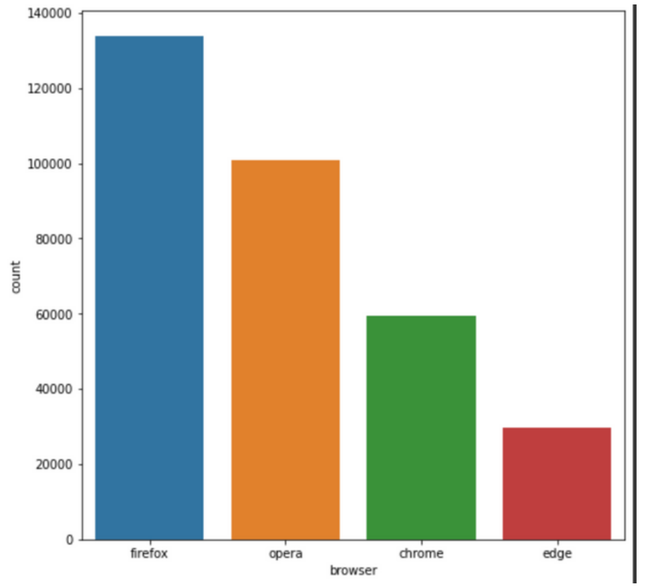
Encrypted TLS DATA

===> MultiLabel or MultiClass problem

QUICK EDA



Os Target Value Counts



Browser Target Value Counts

QUICK EDA

Missing Values

cert.remainingDays	1401	object	7265	72.650000
tcp.offset	1313	object	0	0.000000
cert.hash	1046	object	7265	72.650000
cert.serial	1045	object	7265	72.650000
cert.notAfter	994	object	7265	72.650000
cert.notBefore	990	object	7265	72.650000
cert.subjectCN	983	object	7265	72.650000
cert.alt	904	object	7266	72.660000
ip.dsfield.ecn	887	object	0	0.000000
cert.validDays	427	object	7265	72.650000
cert.subjectON	369	object	7272	72.720000
cert.issuerCN	236	object	7265	72.650000
ip.flags.df	155	object	0	0.000000
cert.issuerON	115	object	7265	72.650000

==> 72% of certif columns values are missing

QUICK EDA

From what I understood List Features are the logs between the client and the server

ip.dsfield.dscp	ip.dsfield.ecn	ip.flags.df	ip.id	ip.len	ip.ttl	tcp.flags	tcp.offset	tcp.options.timestamp.tsval	tcp.window_size_value
['0', '0', '0', '0', '0', '0', '0', '0', '0', ...	['2', '0', '0', '0', '0', '0', '0', '0', '0', ...	['1', '1', '1', '1', '1', '1', '1', '1', '1', ...	['0x28b5', '0x0000', '0x28b8', '0x28b9', '0xcf...	['52', '52', '40', '557', '40', '1174', '1174'...	['128', '63', '128', '128', '63', '63', '63', ...	[['CWR', 'ECE', 'SYN'], ['ACK', 'SYN'], ['ACK'...	['32', '32', '20', '20', '20', '20', '20', '20...	['', '', '', '', '', '', '', '', '', '', '', ...	['8192', '64240', '256', '256', '251', '251', ...

How to deal with it ?

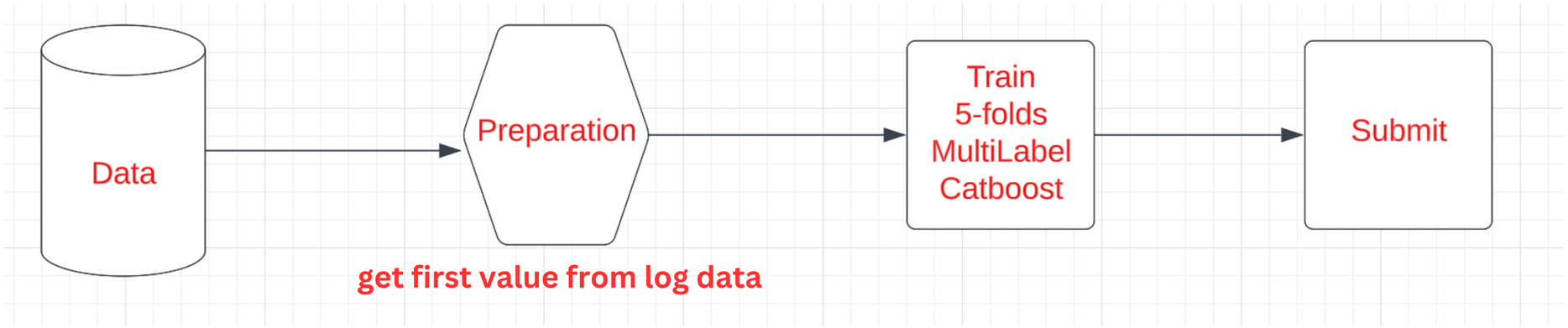
Label Encode ?

QUICK EDA

Not a Good way since for sure in production you will encounter new labels (unknown) and your model can't generalize .

==> Creating Statistics Features may be a good strategy
to ensure that the model can be used in production and generalize well!

FIRST BASELINE

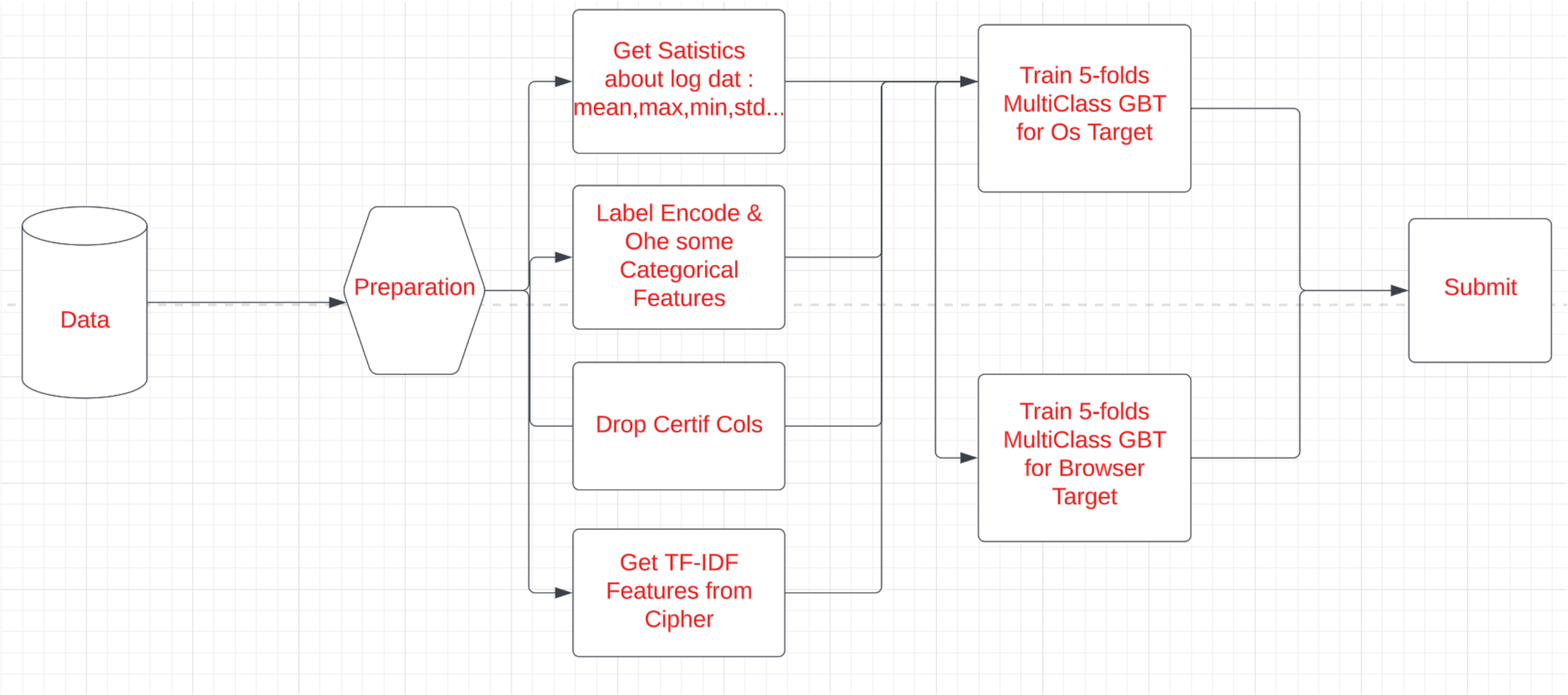


EVALUATION :

CV	Public LB	Private LB
0.043	0.0431893	0.043248623

===> GOOD STARTING POINT

FINAL SOLUTION PIPELINE :



FINAL SOLUTION PIPELINE EVALUATION :

CV	Public LB	Private LB
0.0109	0.010086926	0.010328050

PIPELINE GENERALIZED WELL ON UNSEEN DATA

THINGS TO DO TO IMPROVE THE PIPELINE

- Look For Patterns in Missing Values
- Hyperparameters optimization
- Ensemble Models
- Pseudo-Labelling
- Use Os as a Feature for predicting browser

or vice-versa(Becarefull not no leak information)

THANKS FOR YOUR ATTENTION