DATA SCIENCE WEEKEND

DATA SCIENCE INDONESIA

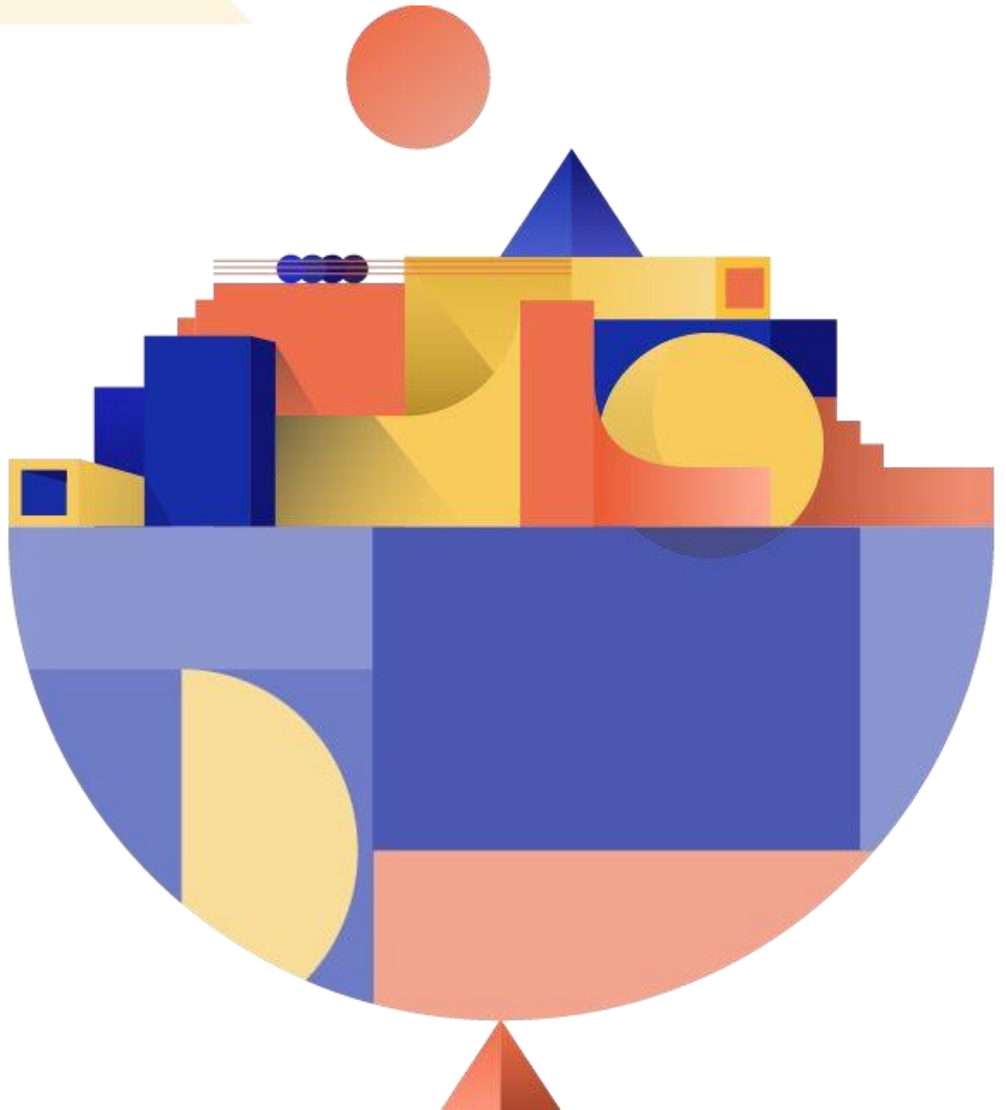
DATA NATIONAL HACKATHON

LULUS2021

M. FARHAN TANDIA

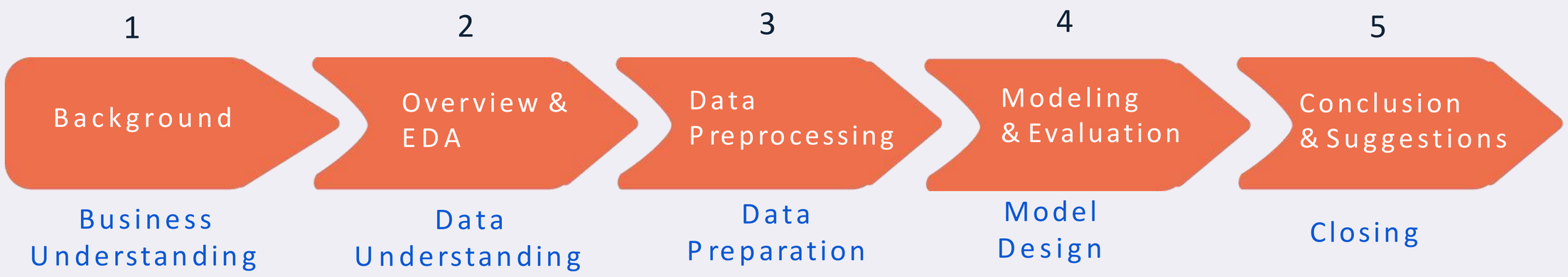
IVAN SURYA HUTOMO

(MASTER STUDENT, NYCU)



OUTLINE

DATA NATIONAL HACKATHON



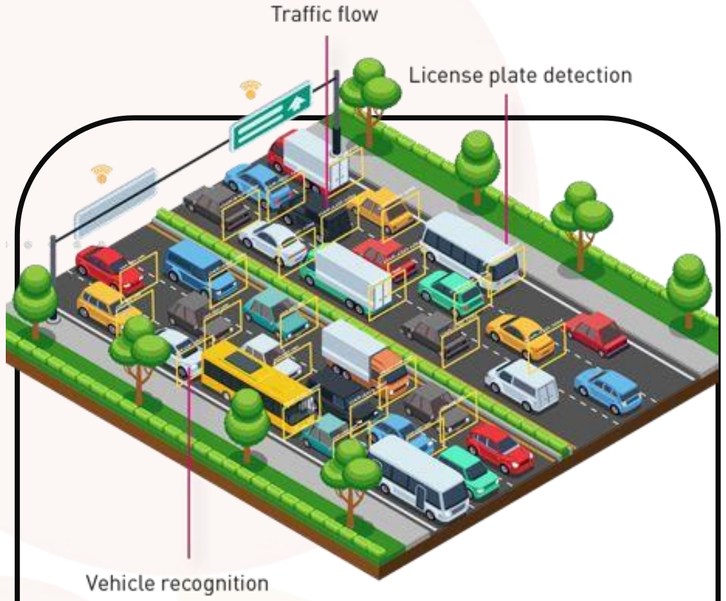
01

BACKGROUND

BUSINESS UNDERSTANDING

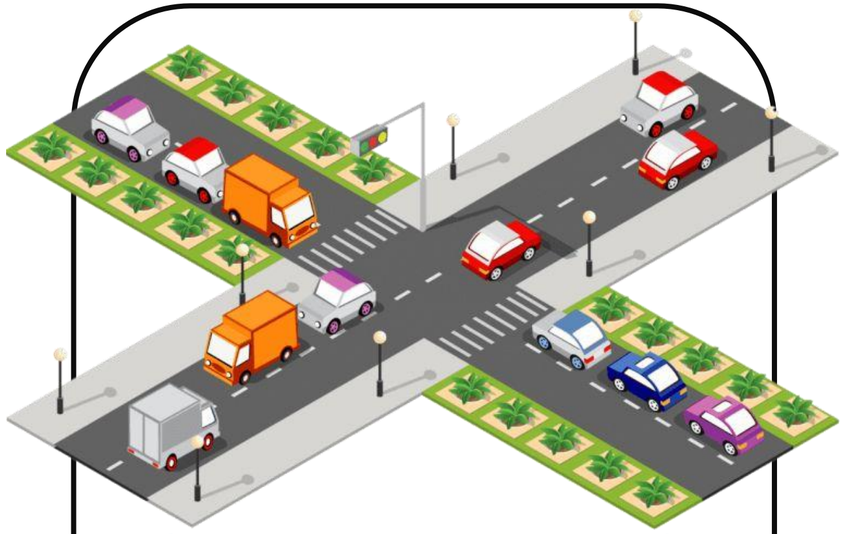
Smart Cities In Indonesia

Province/cities are said to be a complex network of human interactions, laced with infrastructure and systems that drive progress for common good.



Starting with better traffic and datadriven solutions

We need greater use of data for city planning to alleviate traffic congestion and making smart decisions on where to invest funds so as to reduce traffic on its roads.



Forecasting the Future with Data

We should building Machine Learning-based models to predict which sectors will have high reports in certain hours. With this prediction, we could deliver better governance and efficient decision.

02 OVERVIEW & EXPLORATORY DATA ANALYSIS

DATA UNDERSTANDING

OVERVIEW

TASK

Using traffic data provided by the JDS, **predict high jam reports in certain sectors** at certain dates and hours of a few cities in West Java.

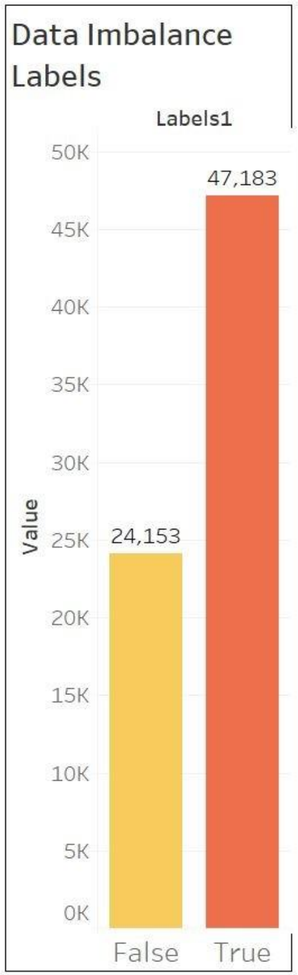
DATASET

- 1Alert Report and 1Irregularities Report
- Data train 71366rows
- Data test 13841rows
- Target Variable: Labels (True/False)

EVALUATION METRICS

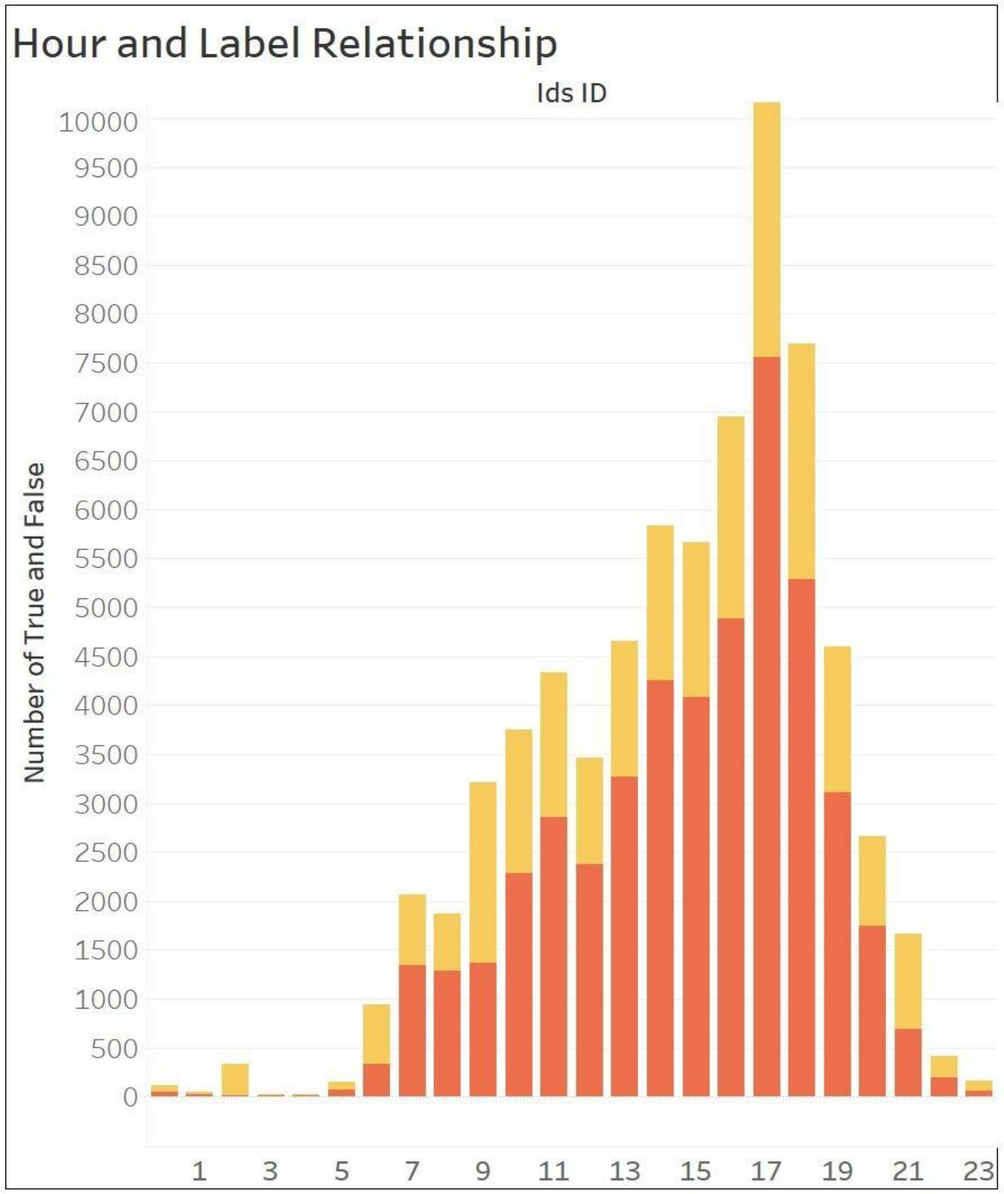
$$\text{F1 score} = \frac{2}{\frac{1}{\text{recall}} + \frac{1}{\text{precision}}} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

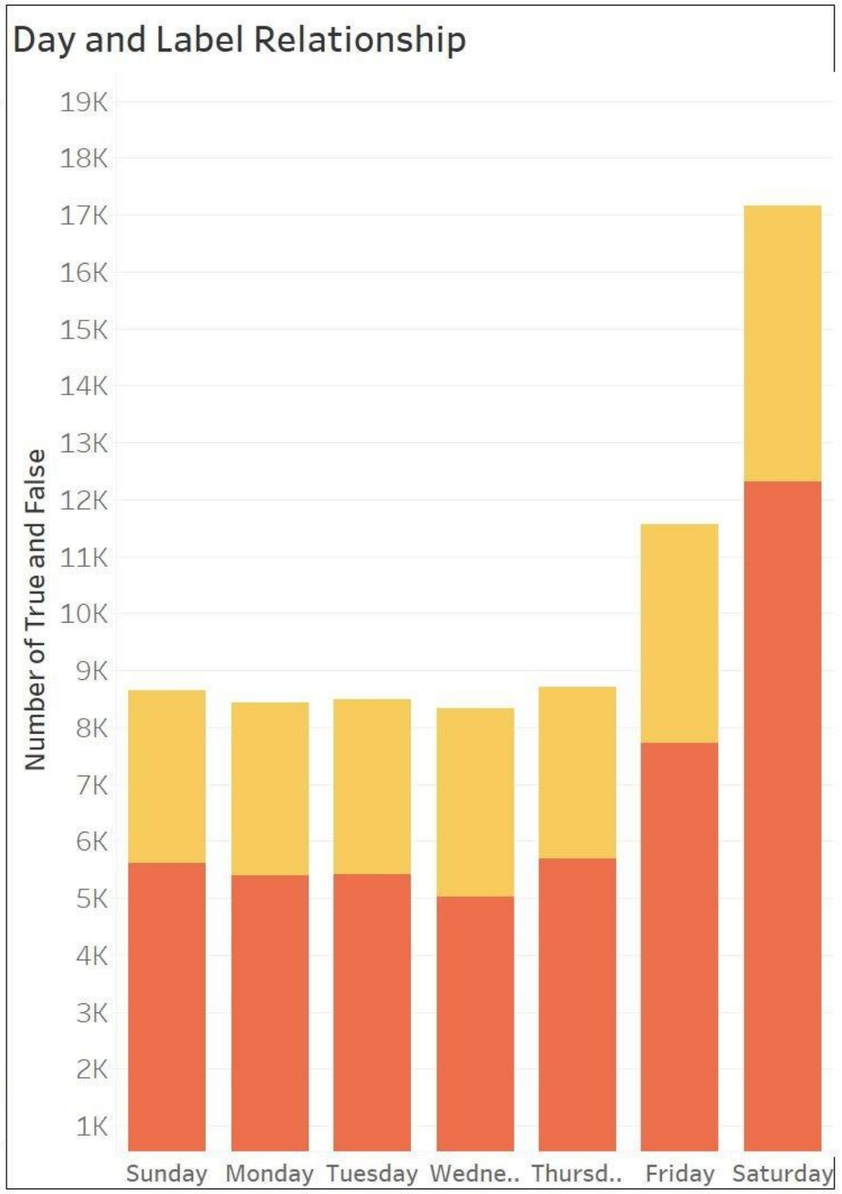
EXPLORATORY DATA ANALYSIS



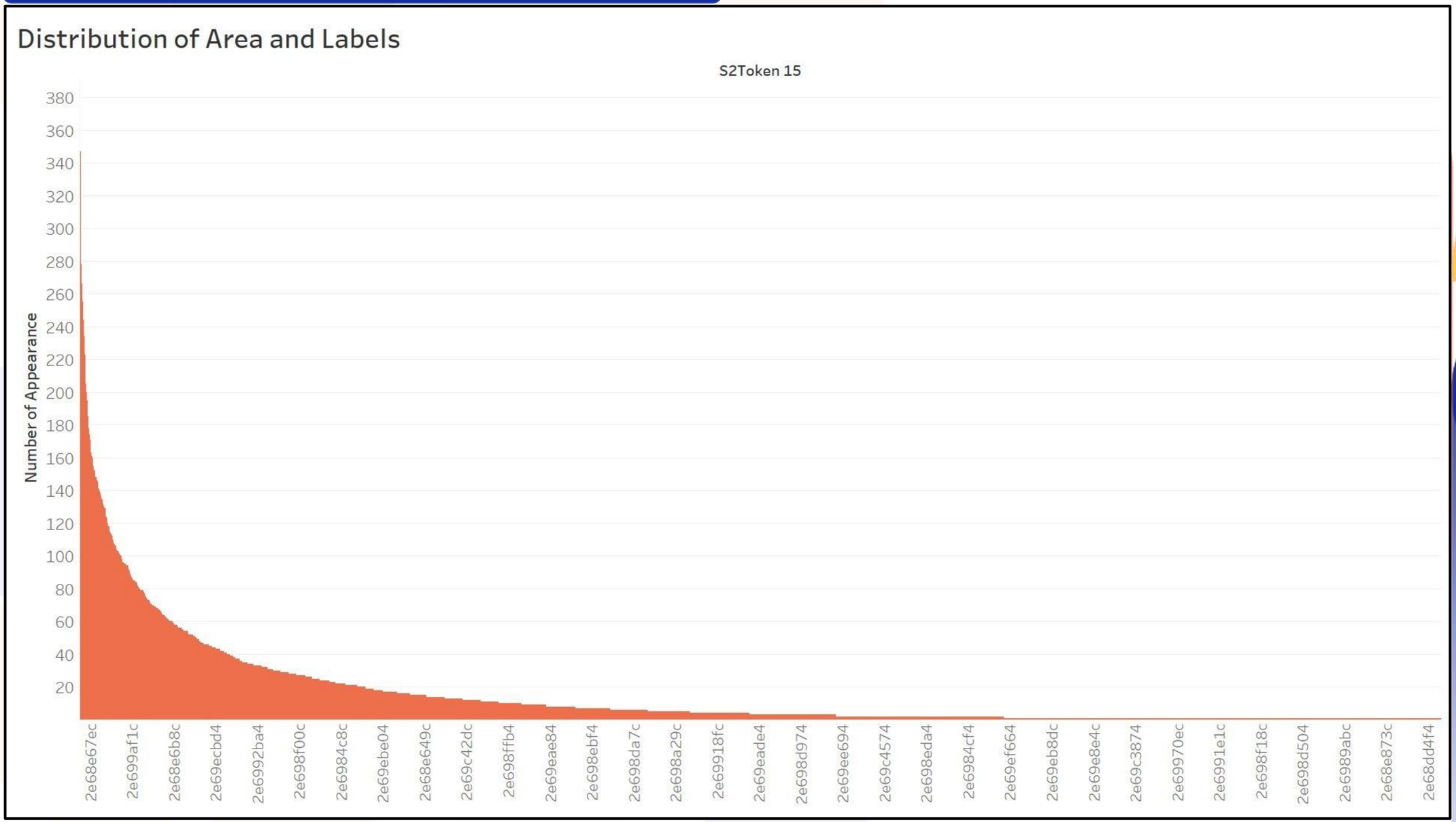
Legends

- False
- True





EXPLORATORY DATA ANALYSIS



03 DATA PREPROCESSING

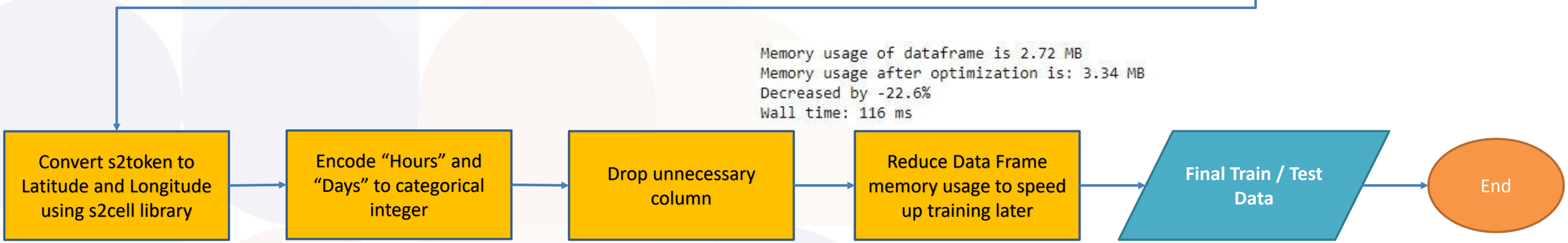
DATA PREPARATION

Data Preprocessing Pipeline

	Ids	s2token_15	pub_date	hour	day
0	2e6992a84_2020-11-25_18	2e6992a84	2020-11-25	18	Wednesday

	Ids	Labels
0	2e69e9384_2020-10-06_13	True

	Ids	Labels	s2token_15	pub_date	hour
0	2e69e9384_2020-10-06_13	1	2e69e9384	2020-10-06	13




```
Memory usage of dataframe is 2.72 MB
Memory usage after optimization is: 3.34 MB
Decreased by -22.6%
Wall time: 116 ms
```

	Ids	Labels	s2token_15	pub_date	hour	day	latitude	longitude
0	2e69e9384_2020-10-06_13	1	2e69e9384	2020-10-06	13	5	-6.393565	106.792766
1	2e6992c7c_2020-10-02_17	1	2e6992c7c	2020-10-02	17	0	-6.317464	106.925531

	hour	day	latitude	longitude
45079	17	6	-6.921053	107.628931
61660	18	5	-6.903811	107.676287
61858	19	4	-6.367458	106.807514
30670	11	4	-6.919377	107.605256
40480	8	1	-6.278608	107.170608

Data Preprocessing Pipeline

	hour	day	latitude	longitude
45079	17	6	-6.921053	107.628931
61660	18	5	-6.903811	107.676287
61858	19	4	-6.367458	106.807514
30670	11	4	-6.919377	107.605256
40480	8	1	-6.278608	107.170608

4 Features vs 12 Features Data Training

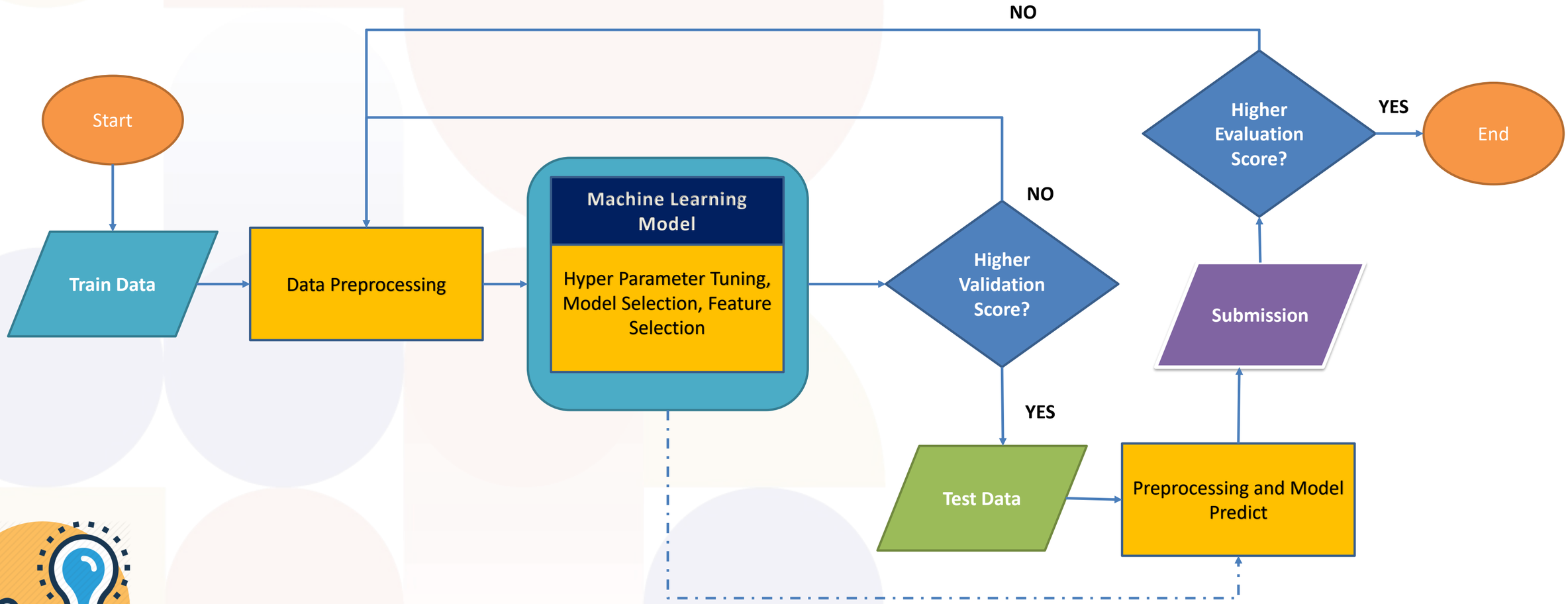
We focused to only used 4 features in our model because when we tried to use 12 features like below, our model could not give us better result. Based on our EDA, those 4 features **already represent dataset behavior.**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
		Ids	s2token_15	road_type	street	city	reliability	report_rat	confidence	type	subtype	longitude	latitude	hour	Labels	pub_date
0		2e69e938	2e69e9384	7	Sawangan	Depok	6	0	0	JAM	JAM_STAN	106.7941	-6.39457	13	1	10/6/2020
1		2e699213	2e6992134	7	Narogong	Bekasi	10	2	3	WEATHER	HAZARD_C	106.9836	-6.3076	11	1	9/12/2020
2		2e69ebe3	2e69ebe3c	7	Margonda	Depok	5	0	0	JAM	JAM_MOD	106.8227	-6.39479	8	1	11/19/2020

04 MODELING & EVALUATION

MODEL DESIGN

Modeling Pipeline



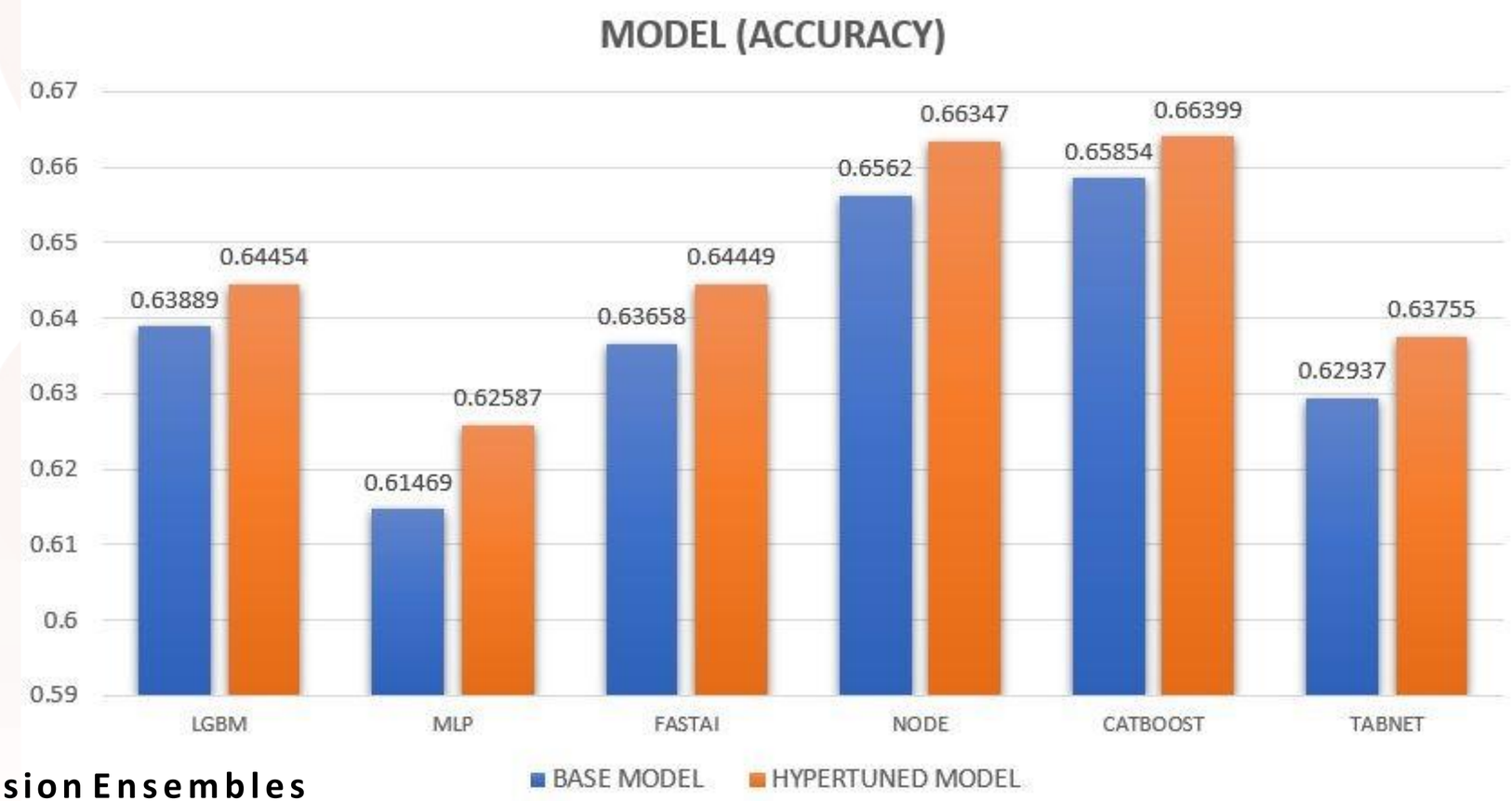
Model Selection

We did an extensive experiment with machine learning models

Data Split :
80% Train
20% Validation

MLP
4 Hidden Layers
(64,128,128,64)
After Hyper Tuned:
6 Hidden Layers
(64,128,128,64,64,4)

Hyper Parameter Tuning
OPTUNA



NODE : Neural Oblivious Decision Ensembles
TABNET :Attentive Interpretable Tabular Learning

Model Selection

We built some machine learning techniques to compete CATBOOST

Multi Layer Perceptron + CATBOOST

Train using multilayer perceptron then take the intermediate layer output to train again in catboost.

We have 4 outputs from intermediate layer and make those outputs as our inputs for catboost training.

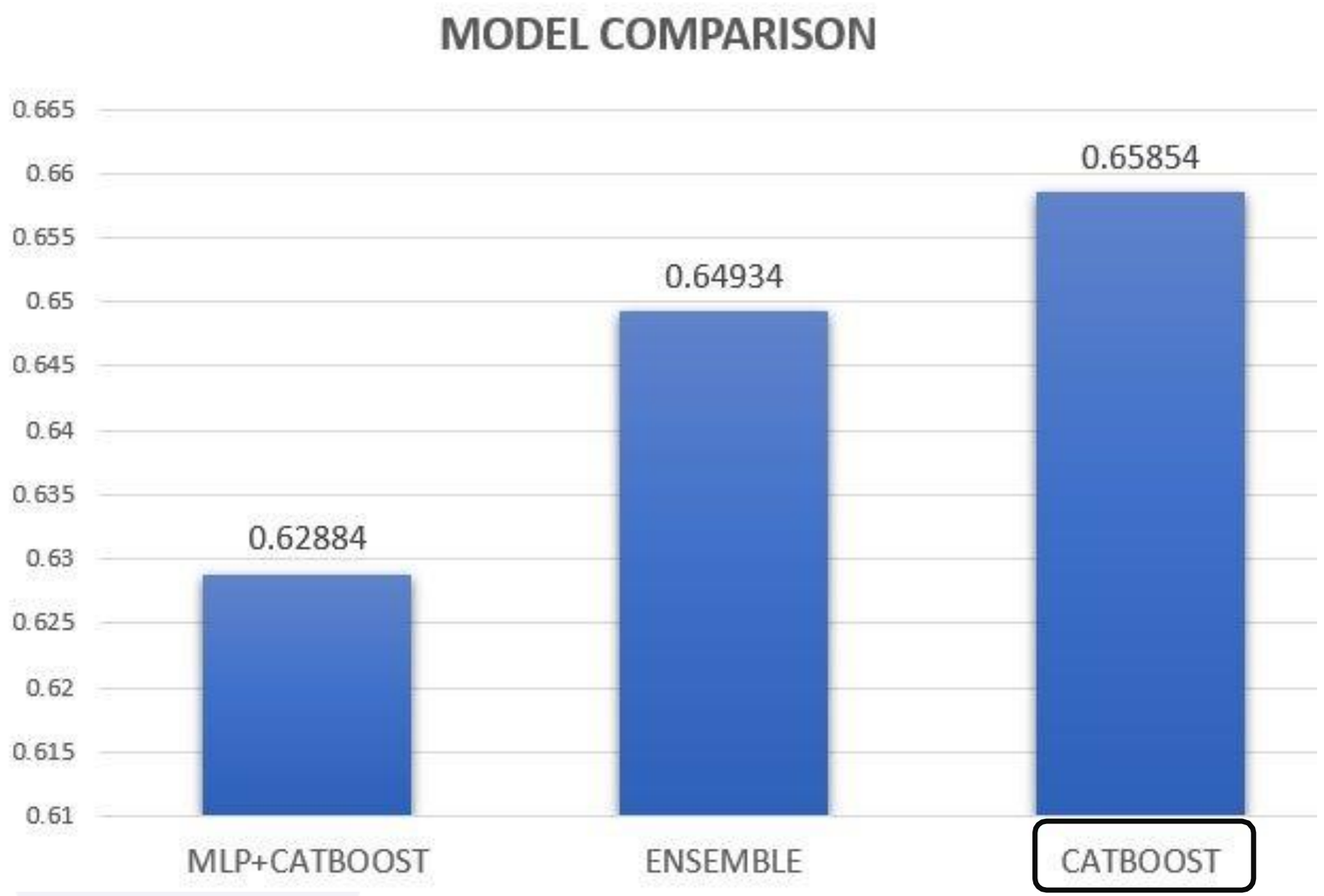
Stacking Ensemble Learning

- EXTRA TREES
- RANDOM FOREST
- LGBM
- CATBOOST

Logistic Regression

Base Model **Meta Model**

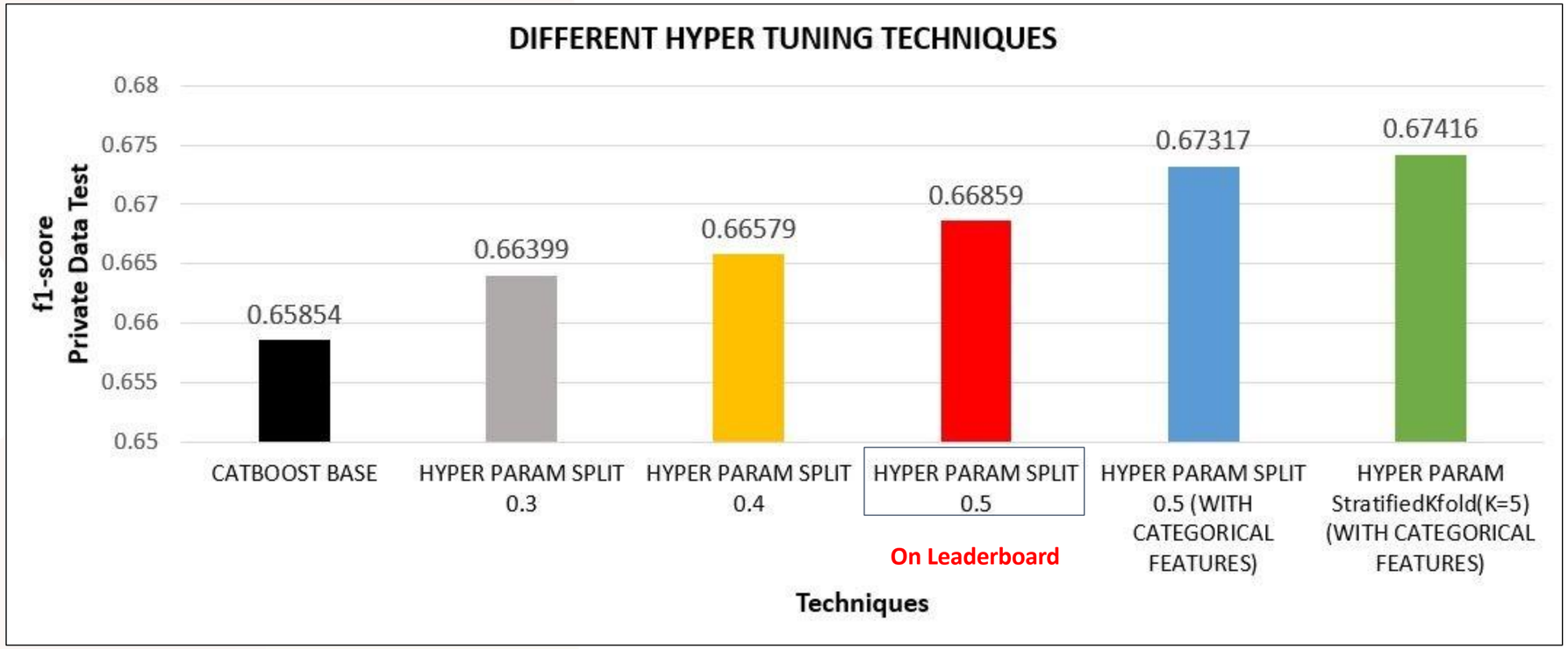
K-FOLD Cross Validation = 3



Model Evaluation

HYPER PARAMETER:

- iterations=10000,
- learning_rate=0.01,
- l2_leaf_reg=3.5,
- colsample_bylevel= 0.063808,
- depth= 3, boosting_type= "Plain",
- eval_metric='Logloss',
- use_best_model=True,
- random_seed=42,
- bootstrap_type= "Bernoulli",
- subsample= 0.55086.
- threshold prediction = 0.6



	IMBALANCE HANDLING	ACCURACY
SPLIT 0.3	CATBOOST + REWEIGHTING + HYPER PARAM	0.62737
	CATBOOST + UNDERSAMPLING + HYPER PARAM	0.64607
	CATBOOST + OVERSAMPLING + HYPER PARAM	0.64042

	Private	Public
test-catboost-60.csv 15 minutes ago by Farhan Tandia add submission details	0.67317	0.65714
test-catboost-scv-60-2.csv 2 minutes ago by Farhan Tandia add submission details	0.67416	0.66066

05 CONCLUSIONS & SUGGESTIONS

CLOSING

CONCLUSIONS

1. CATBOOST has the highest accuracy among all the models.
2. CATBOOST has benefit on the robustness of using categorical data features.
3. **Only using 4 features,** our technique/model is able to achieved a **fast yet high and comparable accuracy on leaderboard.**
4. Not much features could be used rawly. To get higher accuracy In the future, **Well feature engineering** is required to apply on dataset.

SUGGESTIONS

As we know that the highest peak of traffic jam occurred in the evening and in the weekends.

This model prediction could become consideration for government to create new public transportation route and schedule that match with **where** and **when** traffic jam occured.

For people, let's make used of public transportation more than individual vehicles, it'll not only reduce the traffic jam rate but also makes environment cleaner and healthier.

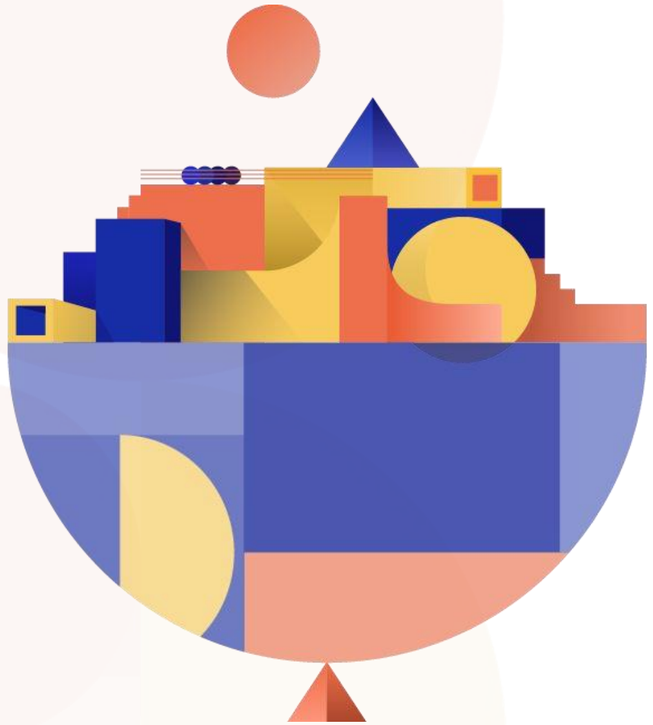
Always check the traffic condition via waze or map applications and decide your transportation vehicle wisely.

DATA SCIENCE WEEKEND 2021

#DSW2021

“Rebalancing The Power”

LULUS 2021!



THANK YOU!

Powered by

DATA SCIENCE INDONESIA

Supported by

Microsoft JABAR DIGITAL SERVICE esri iykra