FORECASTING OF
DENGUE HEMORRHAGIC FEVER
CASES USING MACHINE LEARNING

6410422002 NAVAPOL SANGUANSUB

OBJECTIVE , STUDY AREA AND DATA COLLECTION

OBJECTIVE

คาดการณ์จำนวนผู้ป่วยโรคไข้เลือดออกในแต่ละเดือน โดยตัวแปรคือ ปริมาณน้ำฝนสะสม (มม.), ความชื้นสัมพัทธ์ (%), อุณหภูมิ (สูงสุด, ต่ำสุด และ เฉลี่ย: องศาเซลเซียส)

STUDY AREA

โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านสวน ตำบลบ้านสวน อำเภอเมือง จังหวัดชลบุรี

DATA COLLECTION

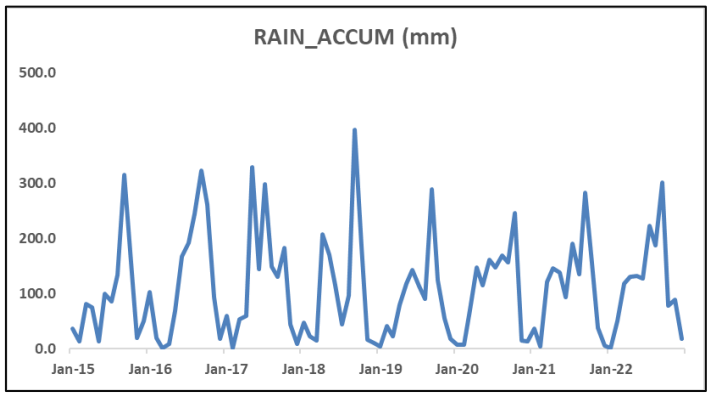
1. รายงานจำนวนผู้ป่วยโรคไข้เลือดออกรายเดือน โรงพยาบาลส่งเสริมสุขภาพตำบลบ้านสวน ปี 2558 – 2565
2. ปริมาณน้ำฝนสะสม (มม.) , ความชื้นสัมพัทธ์เฉลี่ย (เปอร์เซ็น) , อุณหภูมิสูงสุด / ต่ำสุด / เฉลี่ย (องศาเซลเซส) รายเดือน ปี 2558-2565

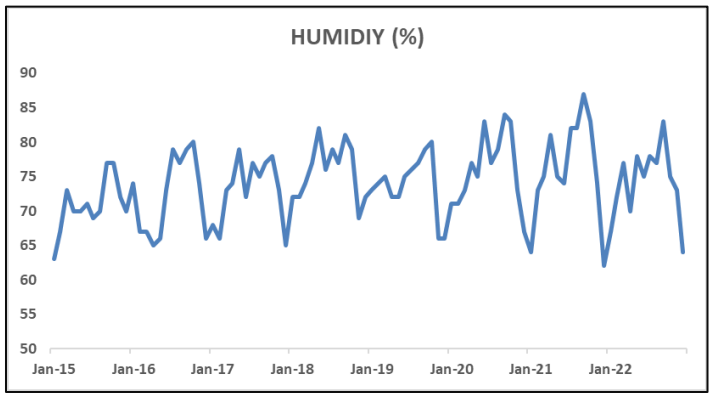
DATA PREPARATION

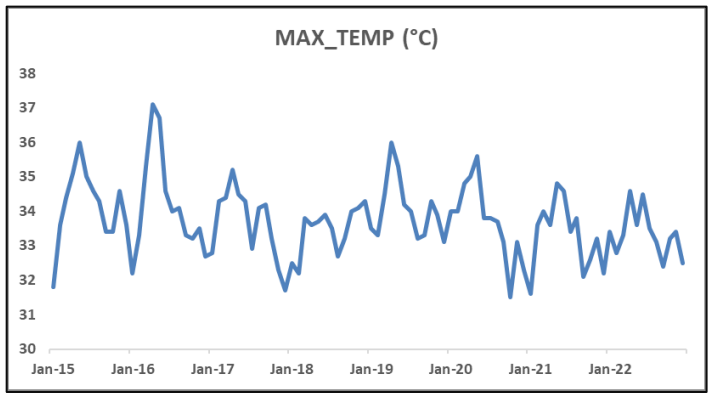
1. Define Data for Input and Target Variable

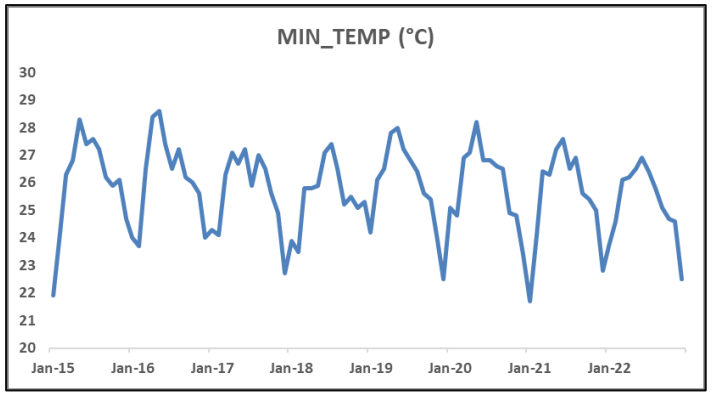
Input variables

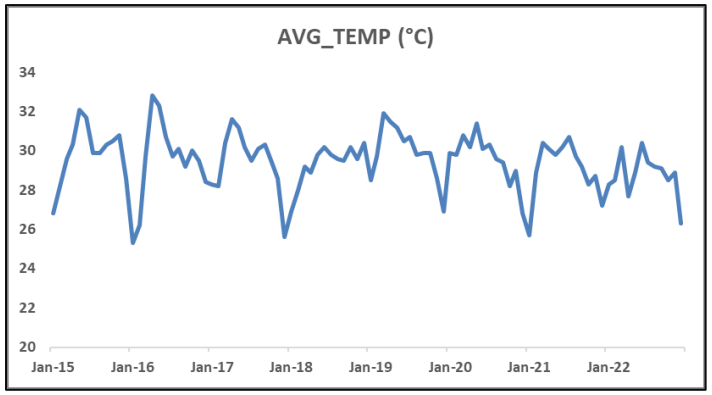
- RAIN_ACCUM (mm), HUMIDIY (%), MAX_TEMP (°C), MIN_TEMP (°C), AVG_TEMP (°C)









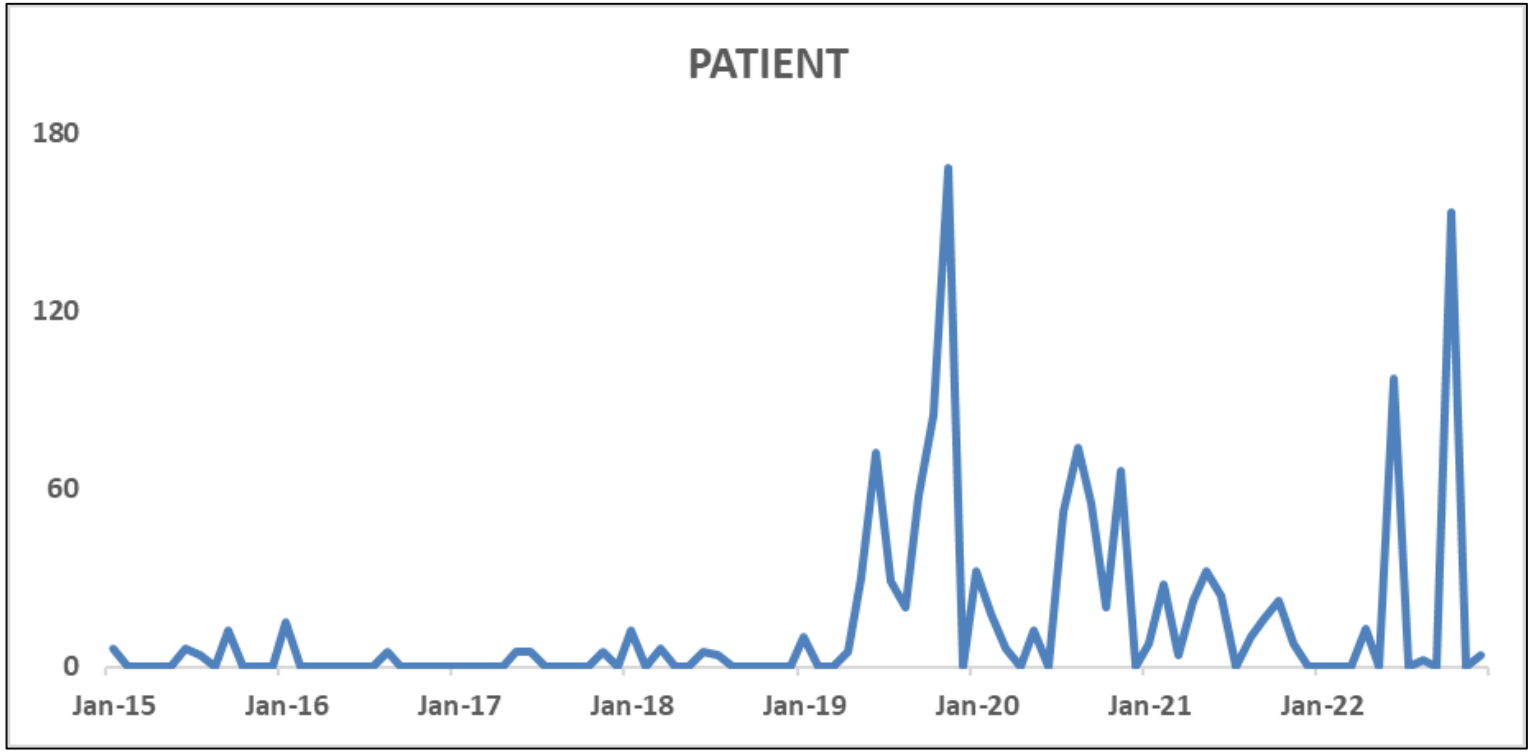


DATA PREPARATION (CONT.)

1. Define Data for Input and Target Variable

Target variable

- PATIENT

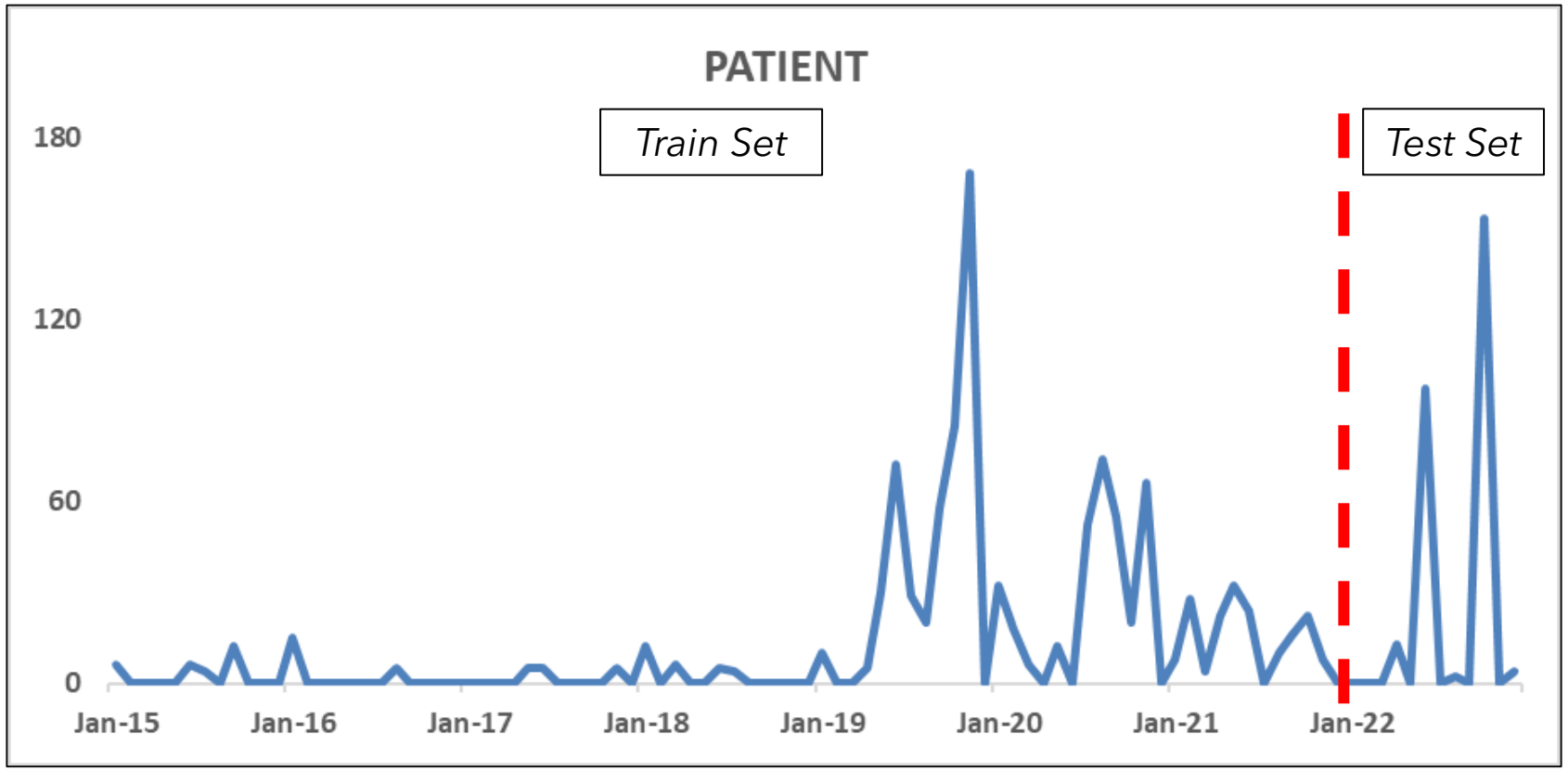


มีเดือนที่มีจำนวนผู้ป่วย
เท่ากับ 0 คน อยู่ถึง 50 %
ของข้อมูลทั้งหมด

DATA PREPARATION (CONT.)

2. Train / Test Split

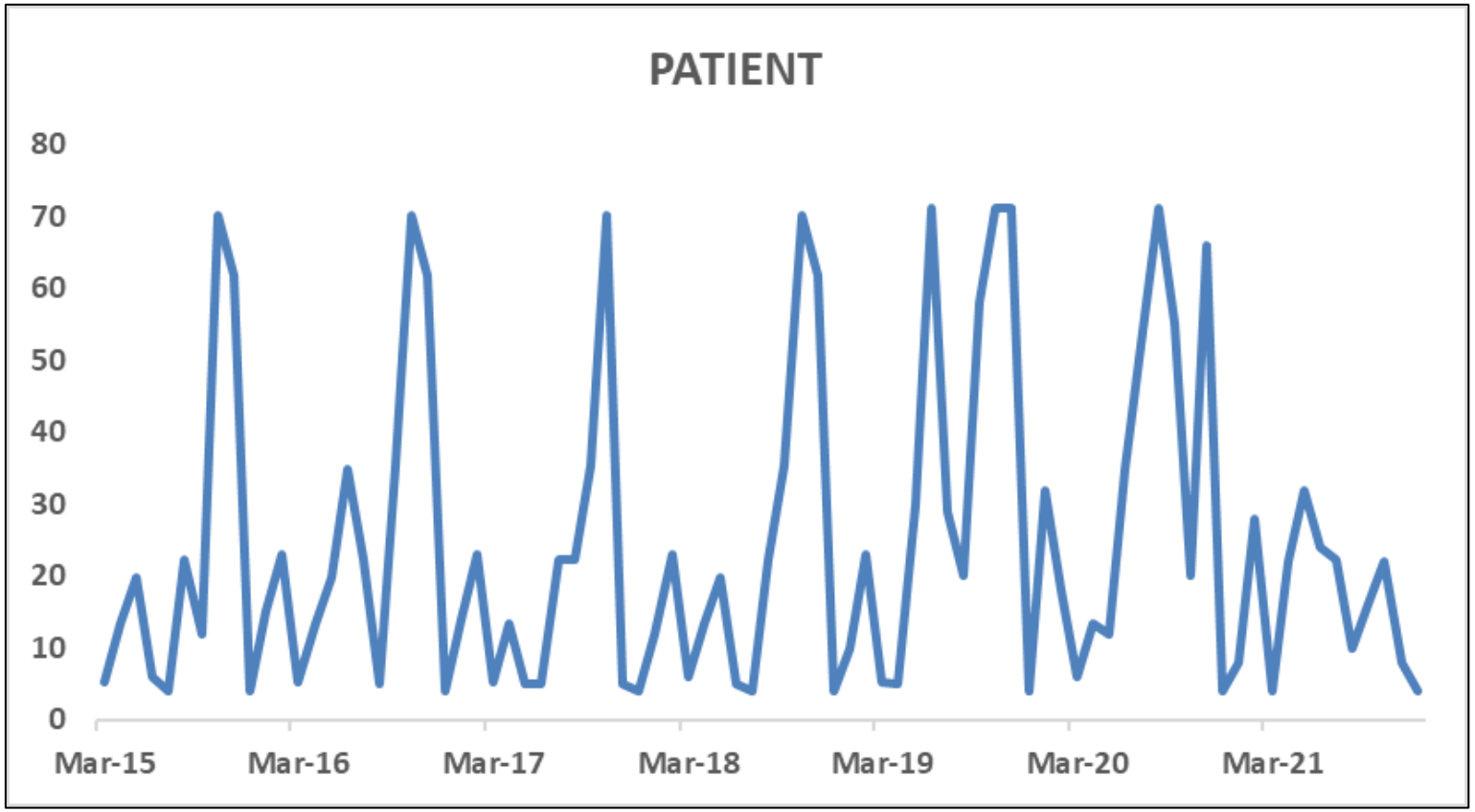
- Train: Jan15 – Dec21
- Test: Jan22 – Dec22



DATA PREPARATION (CONT.)

3. Adjust Data on Train Set : Target Variables (PATIENT)

- บวกค่าเฉลี่ยของแต่เดือนให้แก่เดือนที่มีจำนวนผู้ป่วยเท่ากับ 0
- ตัดค่า Outlier ด้วยการกำหนดค่า Upper Fence: Q3+1.5(IQR) = 71



DATA PREPARATION (CONT.)

4. Create Lag Variables and Another Variables

- สร้าง Lag Variables 2 เดือน ให้กับ Input Variables
  เช่น *RAIN_ACCUM (mm)* -> *RAIN_ACCUM (mm)_1*, *RAIN_ACCUM (mm)_2*
  หมายถึงปริมาณน้ำฝนสะสมใน 1 และ 2 เดือนก่อนหน้าเดือนปัจจุบัน ตามลำดับ
- สร้าง Lag Variables 2 จาก Target Variables เพื่อนำมาใช้เป็น Input Variables -> *PATIENT_1*, *PATIENT_2*
- สร้างตัวแปร Input Variables ที่ระบุเดือนและฤดูกาล

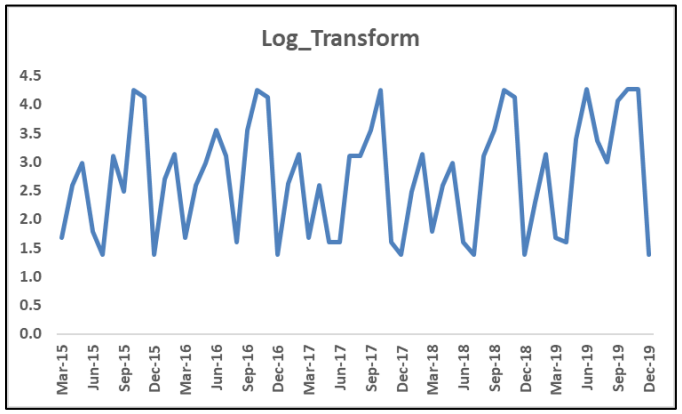
Final Input Variables

Month and Seasonal	->	month_ , seasonal
PATIENT	->	PATIENT_1 , PATIENT_2,
RAIN ACCUMULATION	->	RAIN_ACCUM (mm)_1 , RAIN_ACCUM (mm)_2
HUMIDITY	->	HUMIDIY (%)_1 , HUMIDIY (%)_2
MAX_TEMPERATURE	->	MAX_TEMP (°C)_1 , MAX_TEMP (°C)_2
MIN_TEMPERATURE	->	MIN_TEMP (°C)_1 , MIN_TEMP (°C)_2
AVG_TEMPERATURE	->	AVG_TEMP (°C)_1 , AVG_TEMP (°C)_2

DATA PREPARATION (CONT.)

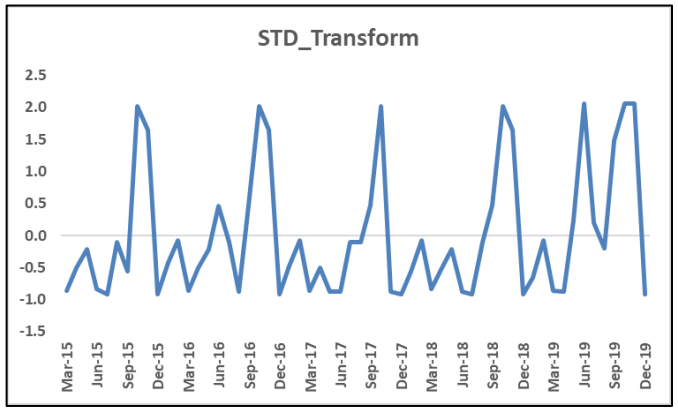
5. Data Transformation on Target Variables (Patient)

Log Transformation



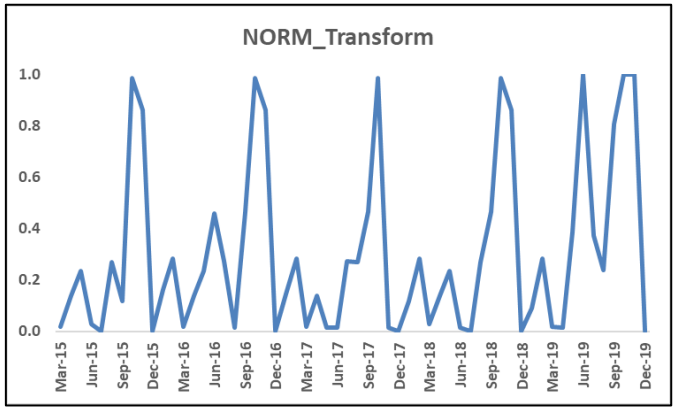
Max	4	**Min**	1
Mean	3	**STD**	0.98

Standardize Transform



Max	2	**Min**	-1
Mean	0	**STD**	1

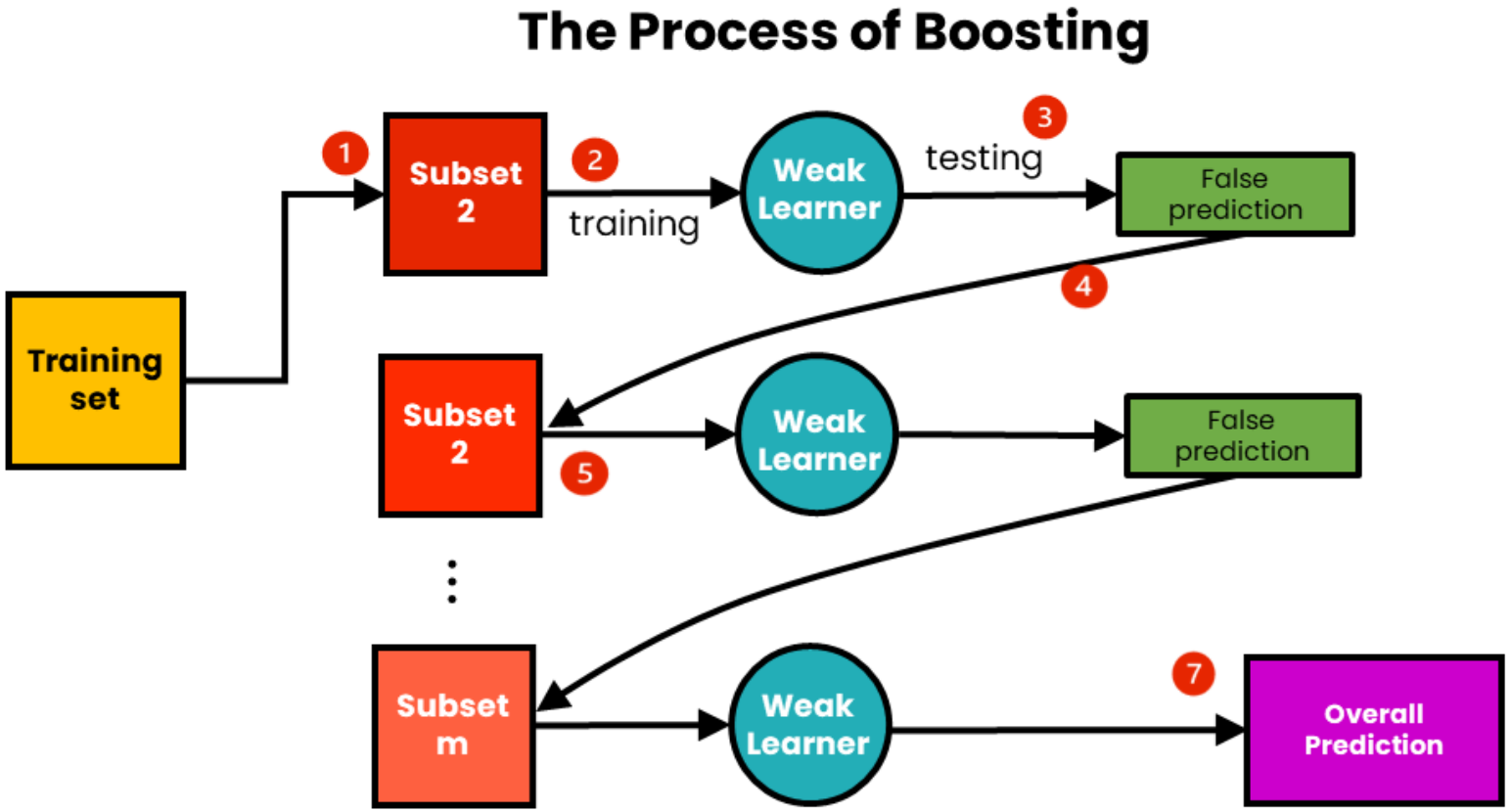
Normalize Transform



Max	1	**Min**	0
Mean	0	**STD**	0.34

METHODLOGY

Ensemble Learning: Boosting

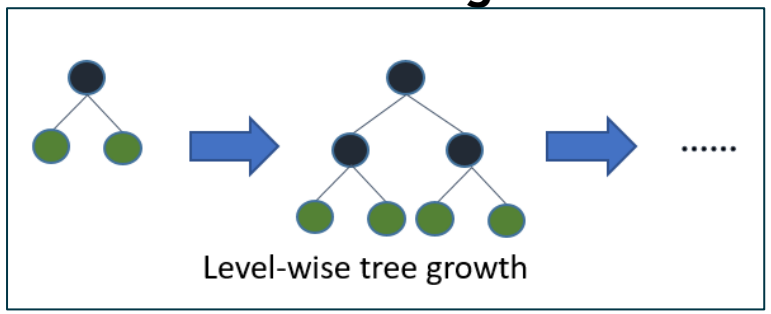


https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/

METHODLOGY (CONT.)

XGBoost

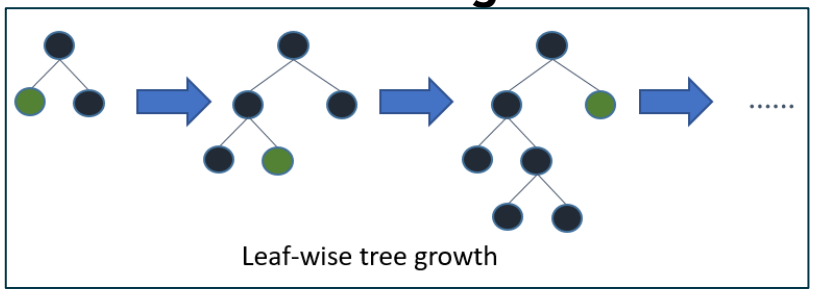
Level wise tree growth



- Level wise tree growth.
- Still considered fast and efficient.
- More Interpretability.

LightGBM

Leaf wise tree growth



- Leaf wise tree growth.
- Faster and reduced number of memory usage.
- Suite to large dataset.

METHODLOGY (CONT.)

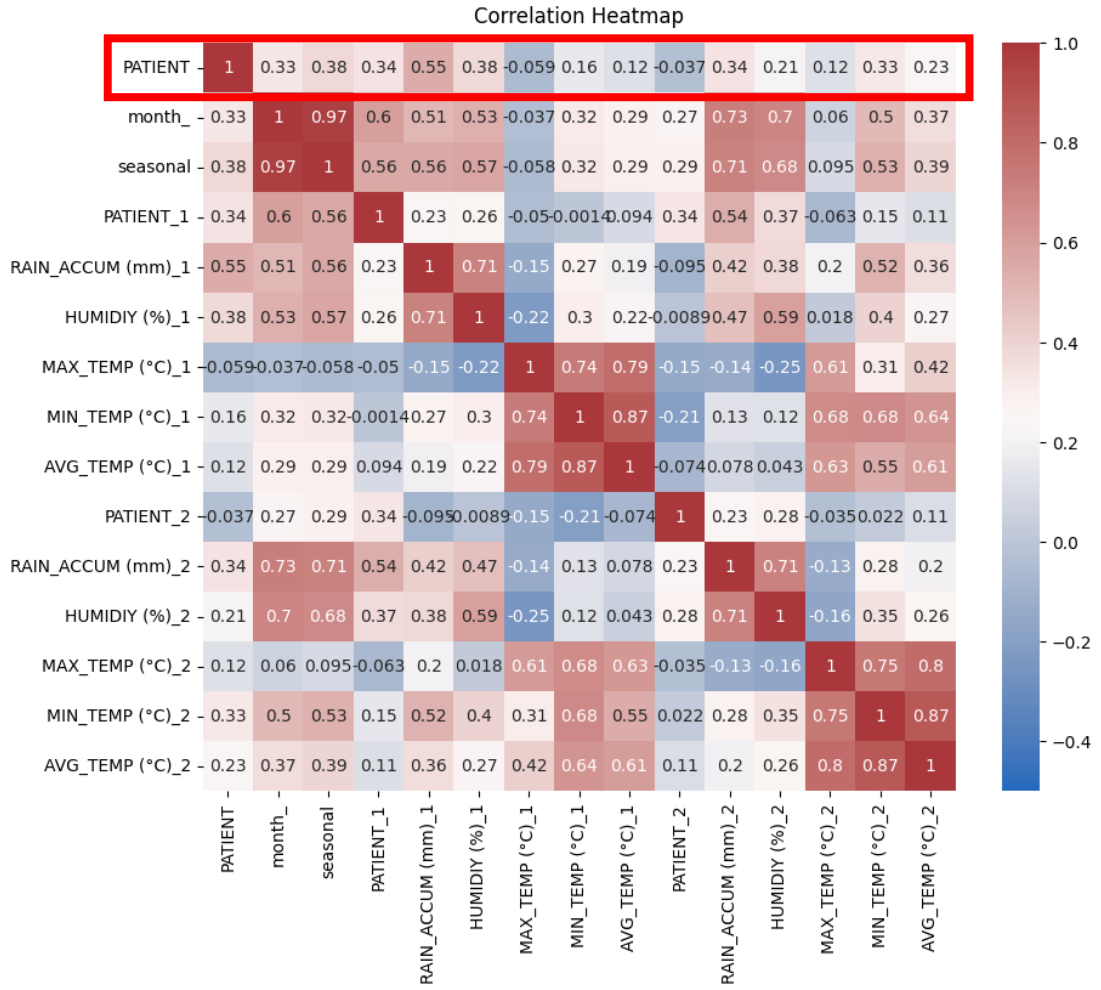
FEATURE SELECTION : Pearson Correlation Coefficient

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Size of Correlation	Interpretation
.90 to 1.00 (−.90 to −1.00)	Very high positive (negative) correlation
.70 to .90 (−.70 to −.90)	High positive (negative) correlation
.50 to .70 (−.50 to −.70)	Moderate positive (negative) correlation
.30 to .50 (−.30 to −.50)	Low positive (negative) correlation
.00 to .30 (.00 to −.30)	negligible correlation

Feature Selected (| r | > 0.3)

*RAIN_ACCUM (mm)_1 , HUMIDIY (%)_1 , seasonal , RAIN_ACCUM (mm)_2 , PATIENT_1 , MIN_TEMP (°C)_2 , month_*



METHODLOGY (CONT.)

Evaluation Metrics

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

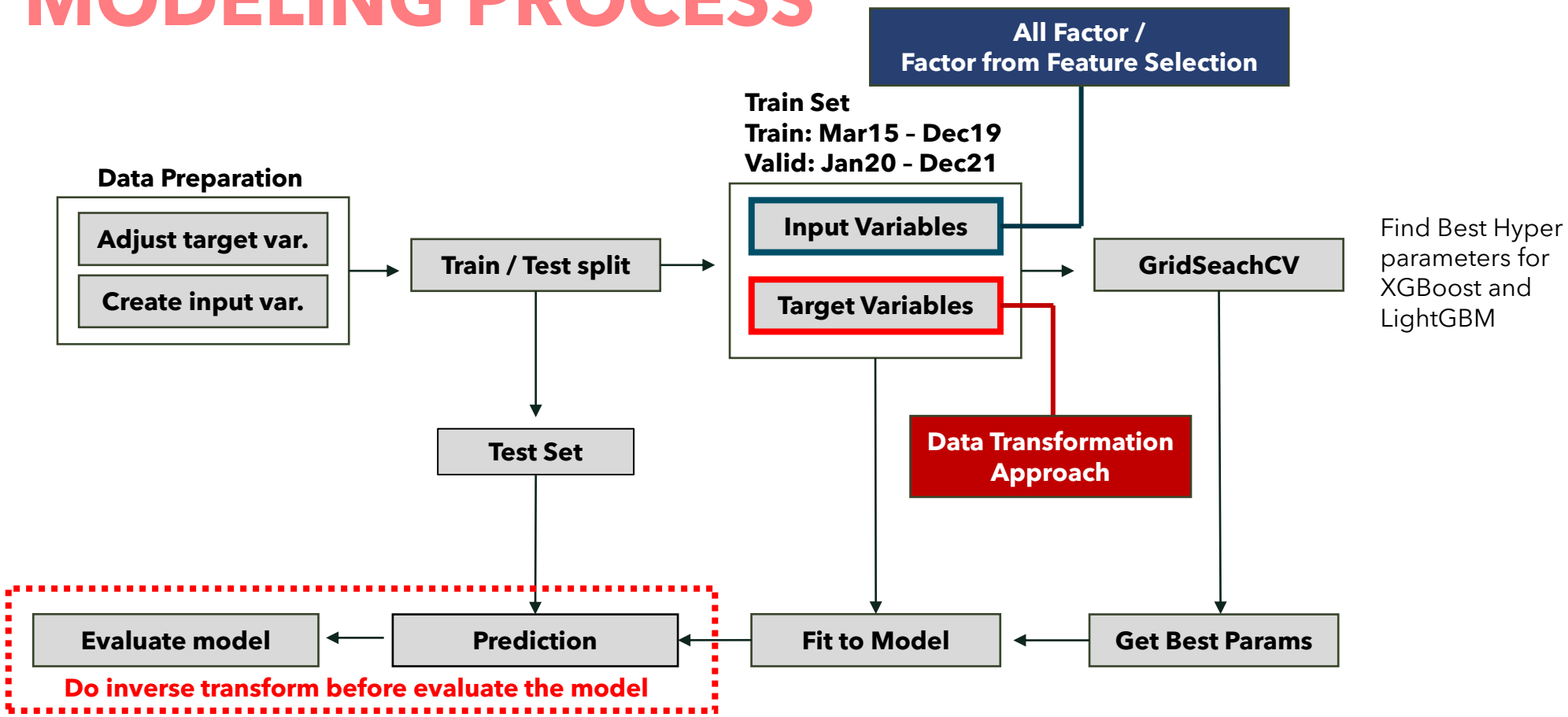
- Mean Square Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

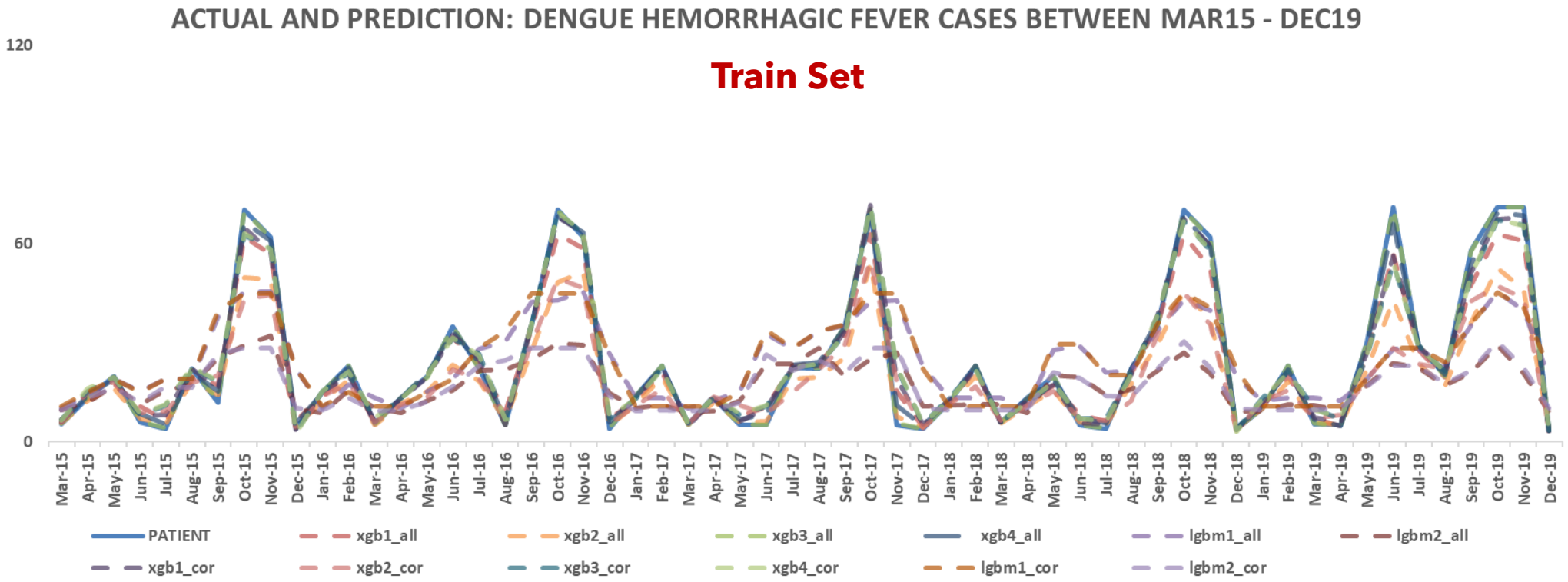
MODELING PROCESS



With 2 dataset by factor, 4 data transformation approach, 2 model

Output : 16 models

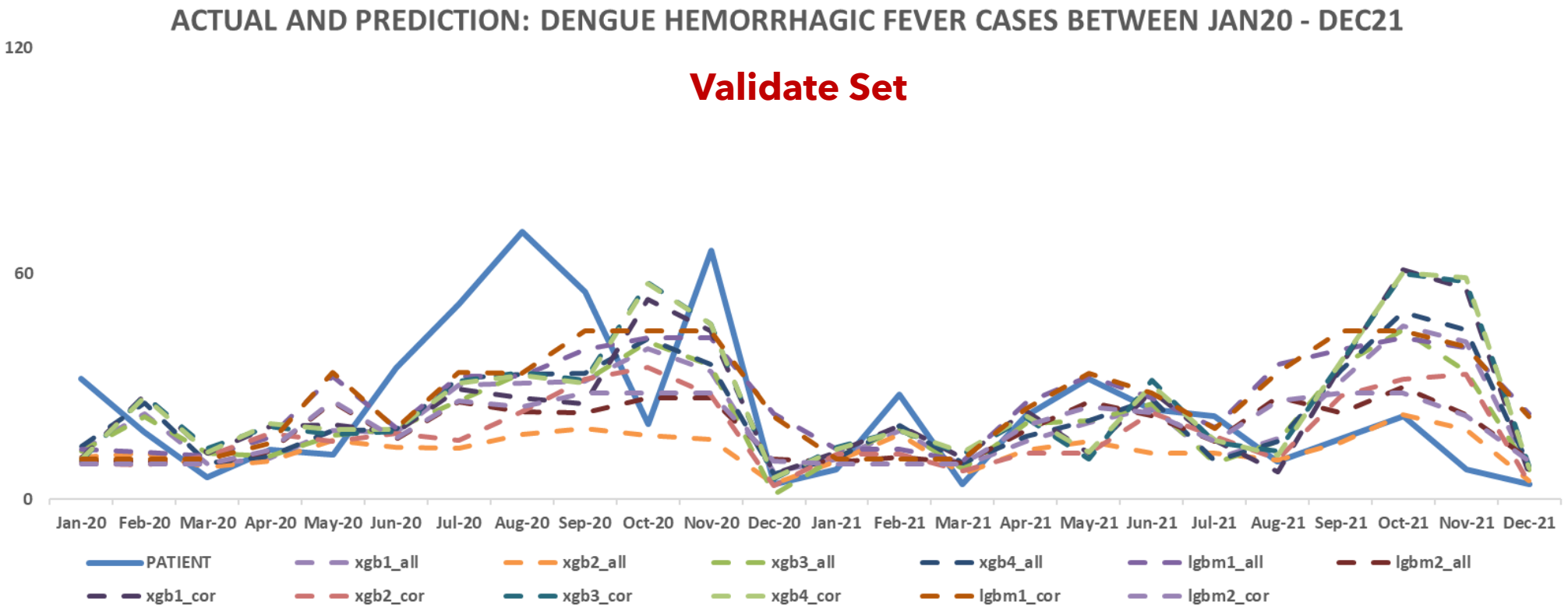
RESULTS



	xgb3_all	xgb4_all	xgb1_cor	xgb4_cor	xgb3_cor	xgb1_all	xgb2_all	xgb2_cor	lgbm1_all	lgbm1_cor	lgbm2_cor	lgbm2_all
RMSE	0.18	1.84	3.50	4.36	4.38	4.54	9.77	11.60	16.46	16.62	19.73	19.81
MSE	0.03	3.37	12.23	19.05	19.16	20.61	95.36	134.52	270.93	276.25	389.26	392.39
MAE	0.14	1.36	2.17	2.83	2.80	3.22	6.07	7.49	12.74	12.95	13.99	13.97

*** lgbm1_all = lgbm3_all = lgbm4_all *and* lgbm1_cor = lgbm3_cor = lgbm4_cor**

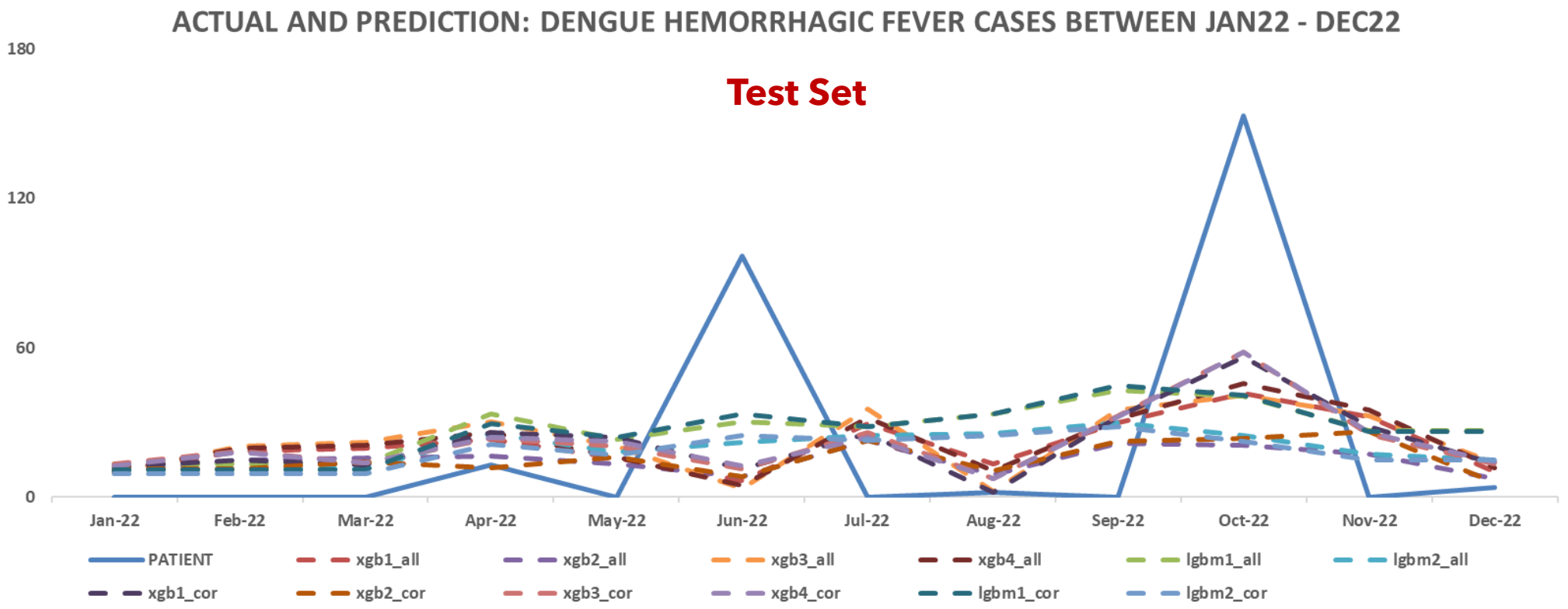
RESULTS (CONT.)



	xgb3_all	xgb1_all	xgb4_all	lgbm2_cor	lgbm2_all	lgbm1_all	lgbm1_cor	xgb2_cor	xgb2_all	xgb3_cor	xgb4_cor	xgb1_cor
RMSE	16.79	17.39	17.50	17.54	17.99	18.21	18.46	18.72	20.36	20.63	20.78	21.08
MSE	281.88	302.37	306.31	307.50	323.51	331.71	340.65	350.30	414.66	425.65	431.84	444.25
MAE	13.12	13.52	13.75	13.22	13.36	15.21	15.27	13.73	13.10	15.91	15.89	16.10

* lgbm1_all = lgbm3_all = lgbm4_all *and* lgbm1_cor = lgbm3_cor = lgbm4_cor

RESULTS (CONT.)



	xgb4_cor	xgb3_cor	xgb1_cor	lgbm1_cor	lgbm1_all	xgb1_all	xgb4_all	lgbm2_cor	lgbm2_all	xgb3_all	xgb2_cor	xgb2_all
RMSE	40.64	40.82	41.42	43.61	44.22	45.33	45.46	45.77	45.98	47.14	47.57	47.77
MSE	1,651.23	1,665.99	1,715.91	1,901.74	1,955.56	2,055.17	2,066.24	2,094.82	2,114.42	2,221.81	2,262.50	2,281.69
MAE	29.52	29.53	29.55	33.47	34.54	32.03	32.89	29.70	31.07	33.97	29.53	29.16

*** lgbm1_all = lgbm3_all = lgbm4_all *and* lgbm1_cor = lgbm3_cor = lgbm4_cor**

DISCUSSION

1. จำนวนของข้อมูล และ ตัวแปรที่นำมาสร้าง Model
   - จำนวนของข้อมูลที่น้อยเกินไป อาจทำให้ไม่เห็นรูปแบบของข้อมูล
   - ตัวแปรอิสระที่นำมาใช้ มีความสัมพันธ์กับตัวแปรตามค่อนข้างน้อย ทำให้ผลการคาดการณ์ไม่แม่นยำ

2. การทดลองกับ Model อื่น ๆ
   - Statistics Model เช่น ARIMA, VAR , Poisson Multivariate Regression
   - Deep learning Model เช่น LSTM, GRU, LSTM with Attention

3. ขอบเขตของการทดลอง
   - การขยายขอบเขตการทดลอง อาจทำให้ได้ผลลัพธ์ที่ดีขึ้น เช่น เปลี่ยนจากการคาดการณ์จำนวนผู้ป่วยโรคไข้เลือดออกในโรงพยาบาลเดียว เป็นการคาดการณ์จำนวนผู้ป่วยโรคไข้เลือดในเขต/อำเภอหนึ่ง ๆ