Gépi tanulás python alapokon

dmlab

Data science módszerek python környezetben

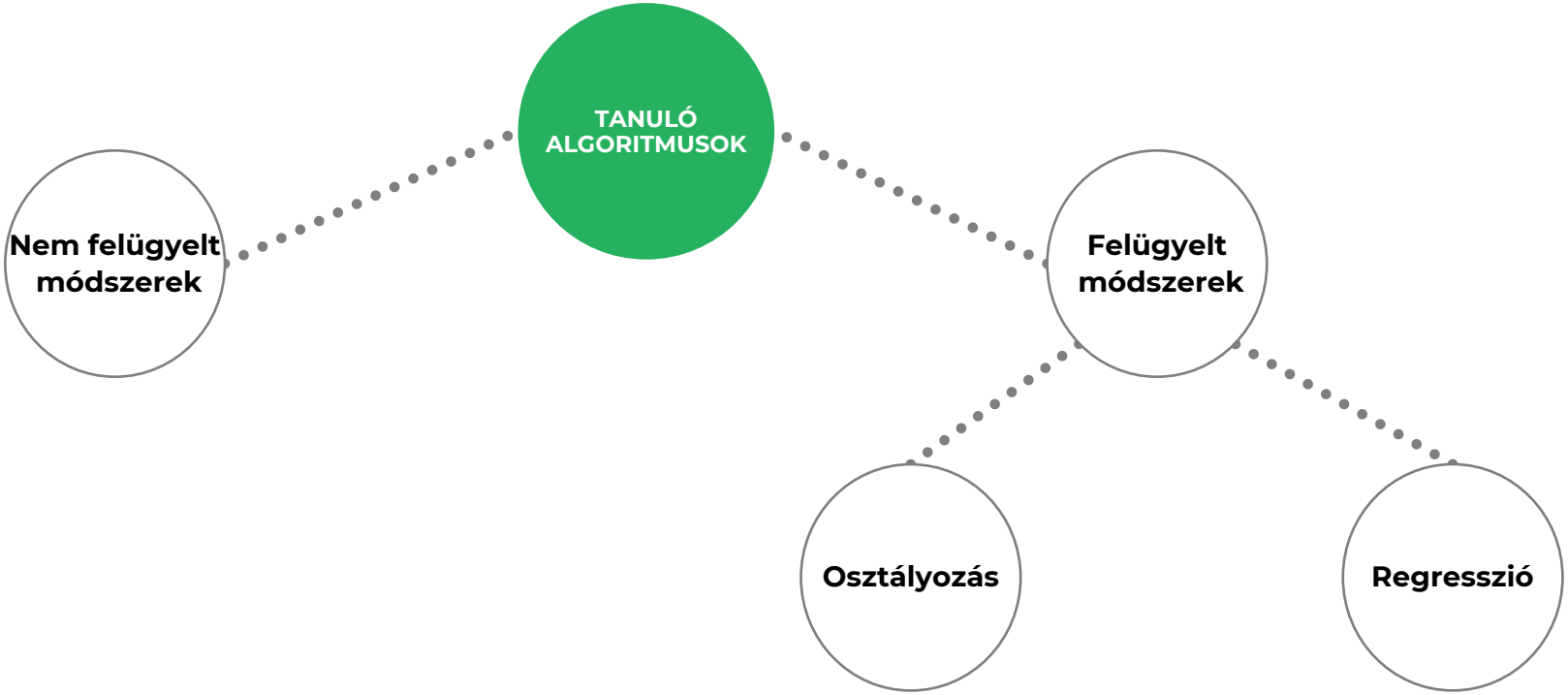
Regresszió

2024.10.14.

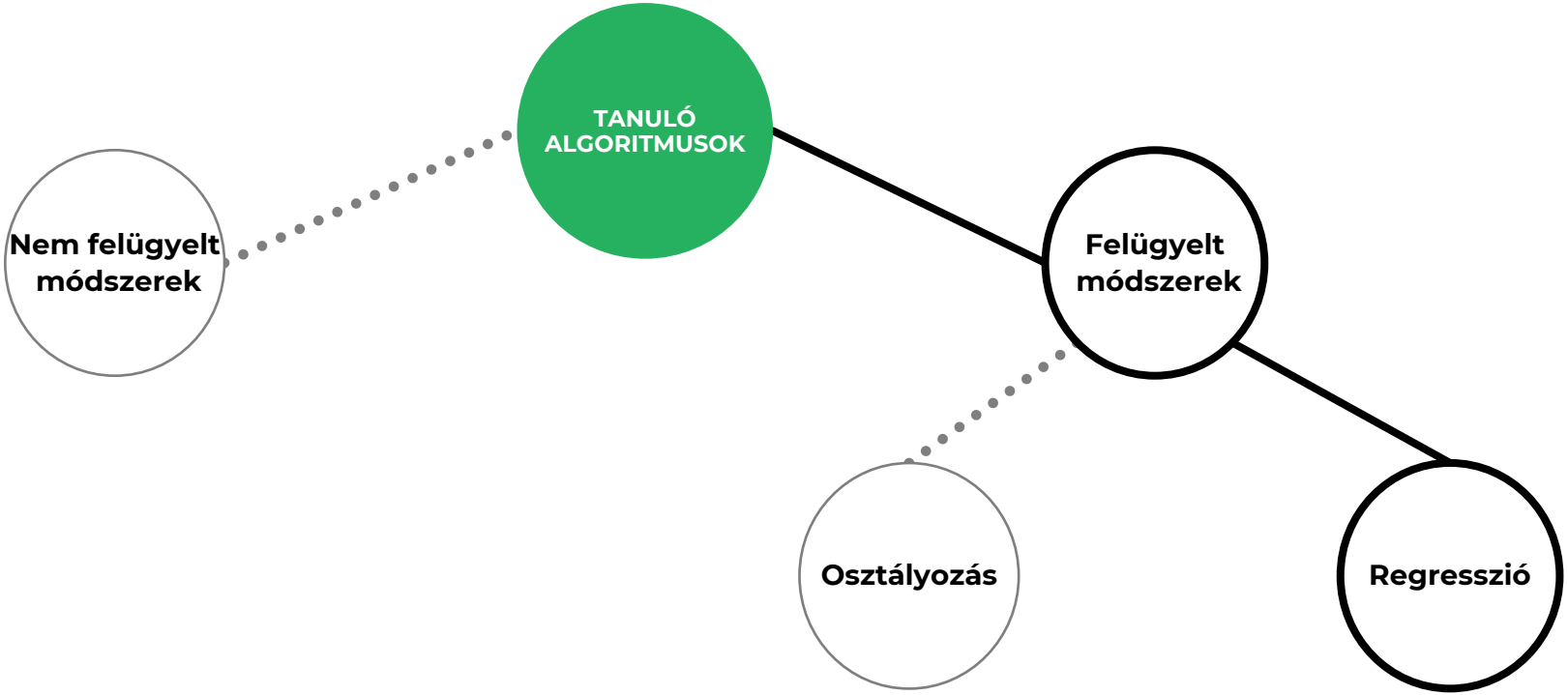
Jónás Dániel
data scientist

ADATELEMZÉSI PROBLÉMÁK CSOPORTOSÍTÁSA

MODELLEZÉSI ALAPOK



MODELLEZÉSI ALAPOK



A REGRESSZIÓ

Jellemzői

- Felügyelt tanulás
- Célváltozó: folytonos, numerikus változó
- Tipikus use-case: ár, szenzoradat, fogyasztás stb. előrejelzése
- Regressziós gépi tanulási modellek: lineáris, regularizált, döntési fa-alapú, stb..
- Kiértékelési metrikák: MAE, MSE, MAPE

Feladat

Van egy tanító adathalmazom

... ahol ismerem a tulajdonságokat és

Tulajdonságleíró attribútumok – **bemeneti változók**

Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma
42	0432	82	3	1
88	0312	44	0	0
22	0101	32	0	1
38	4003	102	1	2

Feladat

Van egy tanító adathalmazom

… ahol ismerem a tulajdonságokat és

X				
X1	**X2**	**X3**	**X4**	**X5**

Tulajdonságleíró attribútumok – **bemeneti változók**

Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma
42	0432	82	3	1
88	0312	44	0	0
22	0101	32	0	1
38	4003	102	1	2

Feladat

Van egy tanító adathalmazom
… ahol ismerem a tulajdonságokat és
… ahol ismerem a célváltozót

X					y
X1	X2	X3	X4	X5	

Célváltozó
Target
Címke

Tulajdonságleíró attribútumok – **bemeneti változók**

Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma	Jövedelem
42	0432	82	3	1	280 000
88	0312	44	0	0	90 000
22	0101	32	0	1	180 000
38	4003	102	1	2	400 000

Feladat

És van egy új sor, ahol
... ismerem a tulajdonságokat

Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma	Jövedelem
55	3257	62	1	3	????

X					y
X1	X2	X3	X4	X5	

Célváltozó
Target
Címke

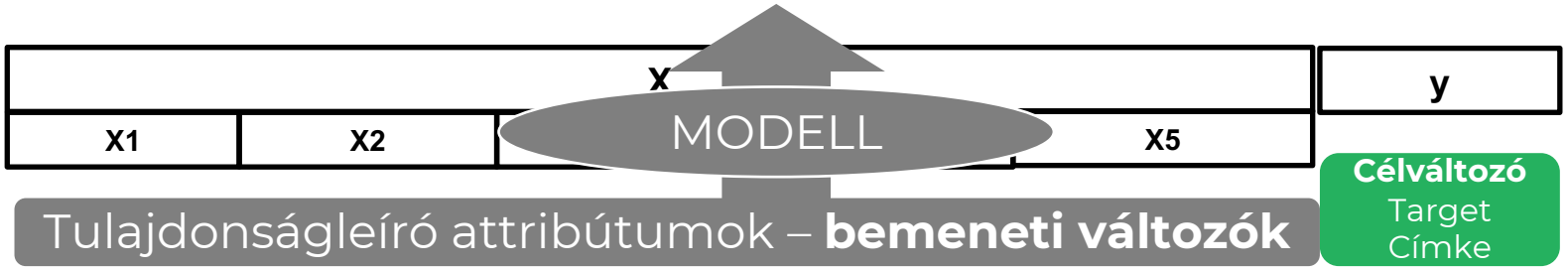
Tulajdonságleíró attribútumok – **bemeneti változók**

Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma	Jövedelem
42	0432	82	3	1	280 000
88	0312	44	0	0	90 000
22	0101	32	0	1	180 000
38	4003	102	1	2	400 000

Feladat

És van egy új sor, ahol
... ismerem a tulajdonságokat
... modell mondja meg mekkora a célváltozó

Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma	Jövedelem
55	3257	62	1	3	????

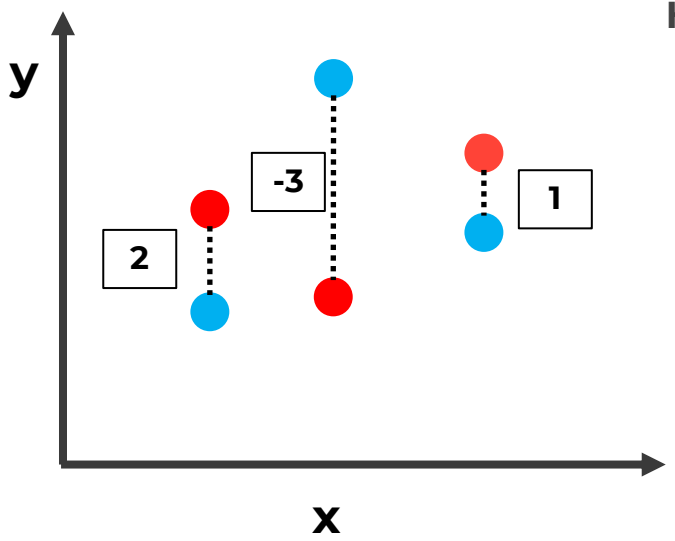


Életkor	Foglalkozás	Lakásméret	Gyerekek	Autók száma	Jövedelem
42	0432	82	3	1	280 000
88	0312	44	0	0	90 000
22	0101	32	0	1	180 000
38	4003	102	1	2	400 000

KIÉRTÉKELÉSI METRIKÁK

VISSZAMÉRÉSI FÜGGVÉNYEK

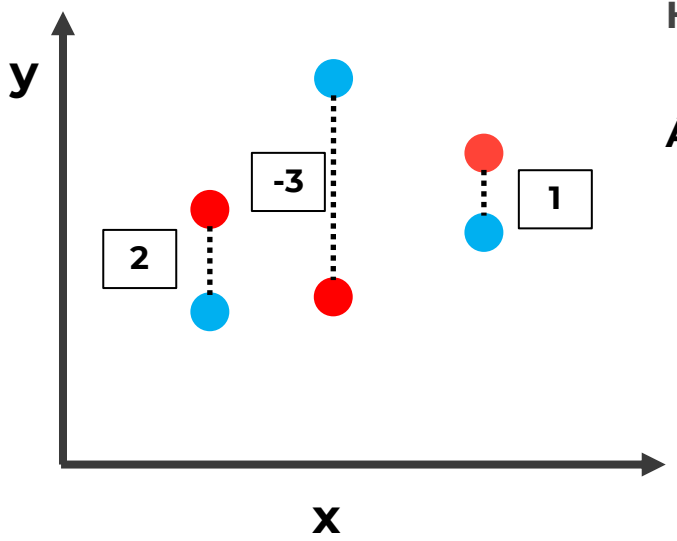
Mi lenne a legjobb tipp? Mi minimalizálja a hibát?
Attól függ, milyen hibát.



Hiba = valós érték – prediktált érték

VISSZAMÉRÉSI FÜGGVÉNYEK

Mi lenne a legjobb tipp? Mi minimalizálja a hibát?
Attól függ, milyen hibát.

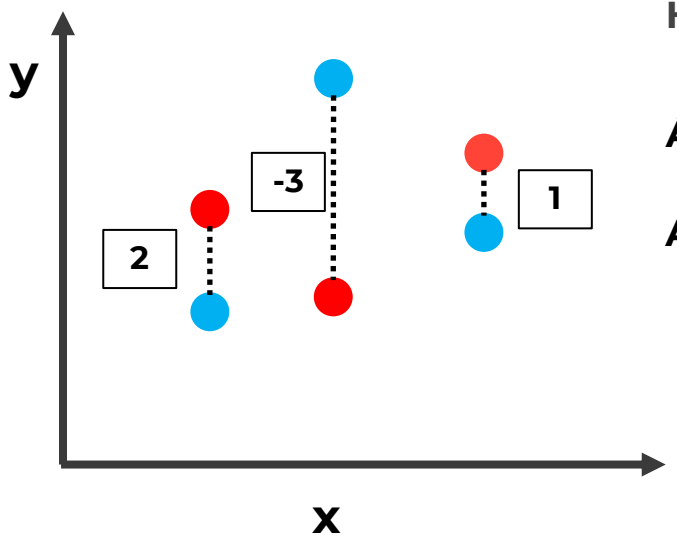


Hiba = valós érték – prediktált érték

Átlagos Hiba = (2 + -3 + 1) / 3 = 0

VISSZAMÉRÉSI FÜGGVÉNYEK

Mi lenne a legjobb tipp? Mi minimalizálja a hibát?
Attól függ, milyen hibát.



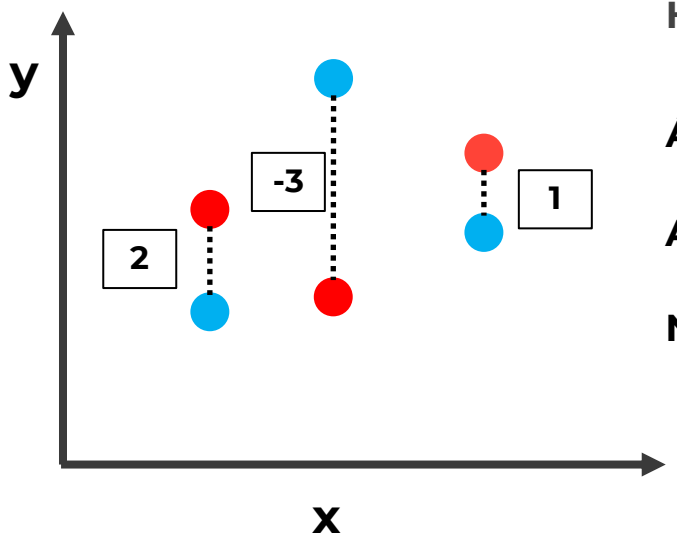
Hiba = valós érték – prediktált érték

Átlagos Hiba = (2 + -3 + 1) / 3 = 0

Abszolút Átlagos Hiba = (2 + 3 + 1) / 3 = 2

VISSZAMÉRÉSI FÜGGVÉNYEK

Mi lenne a legjobb tipp? Mi minimalizálja a hibát?
Attól függ, milyen hibát.



Hiba = valós érték – prediktált érték

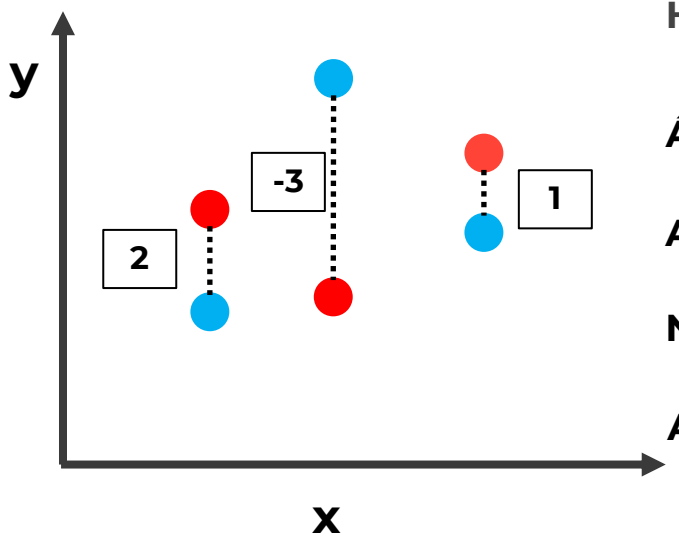
Átlagos Hiba = (2 + -3 + 1) / 3 = 0

Abszolút Átlagos Hiba = (2 + 3 + 1) / 3 = 2

Négyzetes Átlagos Hiba = (4 + 9 + 1) / 3 = 4.67

VISSZAMÉRÉSI FÜGGVÉNYEK

Mi lenne a legjobb tipp? Mi minimalizálja a hibát?
Attól függ, milyen hibát.



Hiba = valós érték – prediktált érték

Átlagos Hiba = (2 + -3 + 1) / 3 = 0

Abszolút Átlagos Hiba = (2 + 3 + 1) / 3 = 2

Négyzetes Átlagos Hiba = (4 + 9 + 1) / 3 = 4.67

Abszolút Százalékos Hiba = (0.3 + 1.3 + 0.2) / 3 = 0.6 = 60%

KIÉRTÉKELÉSI METRIKÁK

MEAN ABSOLUTE ERROR (MAE)

- $| X_{valós} - X_{pred} |$
- MAE = | 10 – 12 | = 2
- A hiba mértéke a fontos, nem az iránya
- Könnyen értelmezhető

KIÉRTÉKELÉSI METRIKÁK

MEAN ABSOLUTE ERROR (MAE)

- $| X_{valós} - X_{pred} |$
- MAE = | 10 – 12 | = 2
- A hiba mértéke a fontos, nem az iránya
- Könnyen értelmezhető

MEAN SQUARED ERROR (MSE)

- $(X_{valós} - X_{pred})^2$
- MSE = $(10 – 12)^2 = 4$
- Könnyen optimalizálható
- A nagyobb eltérés nagyobb hibának számít

KIÉRTÉKELÉSI METRIKÁK

MEAN ABSOLUTE ERROR (MAE)

- $| X_{valós} - X_{pred} |$
- $MAE = | 10 - 12 | = 2$
- A hiba mértéke a fontos, nem az iránya
- Könnyen értelmezhető

MEAN SQUARED ERROR (MSE)

- $(X_{valós} - X_{pred})^2$
- $MSE = (10 - 12)^2 = 4$
- Könnyen optimalizálható
- A nagyobb eltérés nagyobb hibának számít

MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

- $\frac{| Xval_{ós} - X_{pred} |}{X_{valós}}$
- $\frac{| 10 - 12|}{10} = 20\%$
- Általánosan értelmezhető
- A hiba mértéke függ az értékkészlet nagyságrendjétől

dmlab

KIÉRTÉKELÉSI METRIKÁK

MEAN ABSOLUTE ERROR (MAE)

- Ismerni kell a nagyságrendet
- pl MAE = 100 jó vagy rossz?

KIÉRTÉKELÉSI METRIKÁK

MEAN ABSOLUTE ERROR (MAE)

- Ismerni kell a nagyságrendet
- pl MAE = 100 jó vagy rossz?

MEAN SQUARED ERROR (MSE)

- MSE = 250.000.000.000
- Jó vagy rossz?
- 50 milliós nagyságrendű célváltozóknál ez 1% körüli hiba
- RMSE = $\sqrt{MSE}$ = 500.000

KIÉRTÉKELÉSI METRIKÁK

MEAN ABSOLUTE ERROR (MAE)

- Ismerni kell a nagyságrendet
- pl MAE = 100 jó vagy rossz?

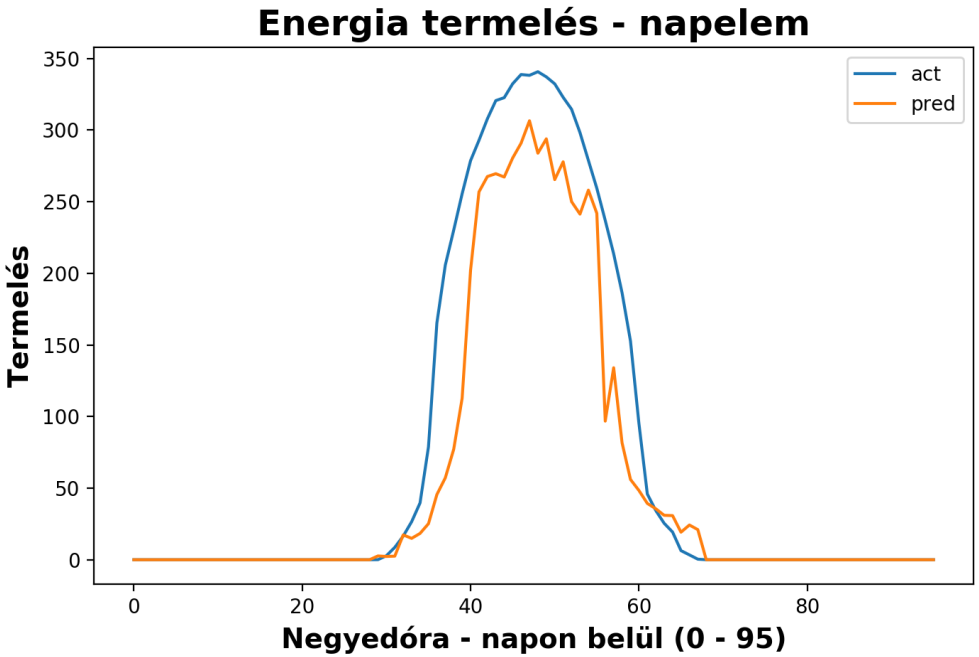
MEAN SQUARED ERROR (MSE)

- MSE = 250.000.000.000
- Jó vagy rossz?
- 50 milliós nagyságrendű célváltozóknál ez 1% körüli hiba
- RMSE = $\sqrt{MSE}$ = 500.000

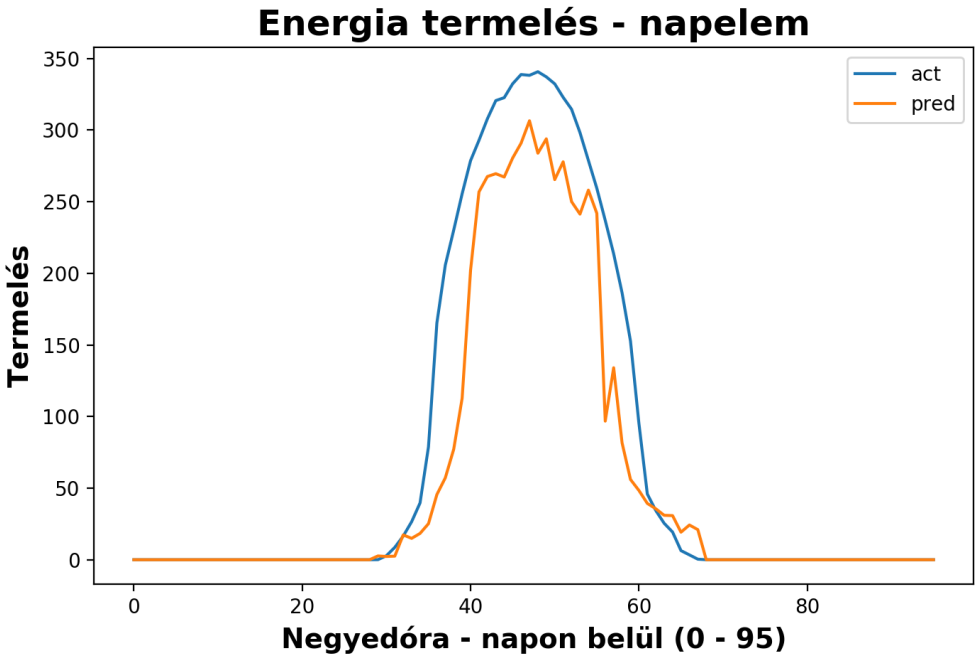
MEAN ABSOLUTE PERCENTAGE ERROR (MAPE)

- 0 értékű célváltozónál nem értelmezhető
- A hiba mértéke függ az értékkészlet nagyságrendjétől

MAPE

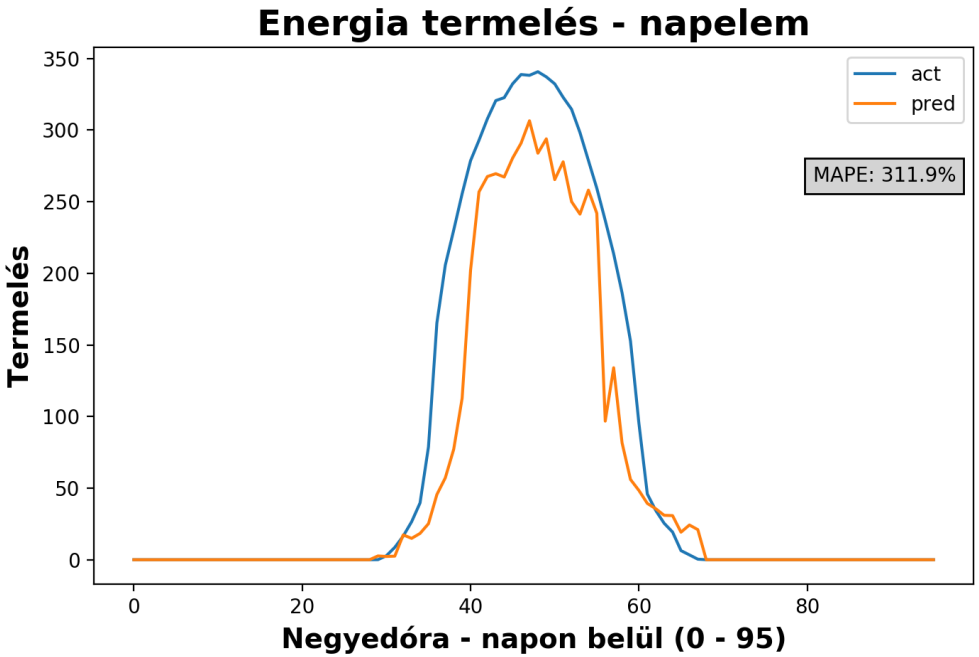

Energia termelés - napelem
act
pred
Termelés
350
300
250
200
150
100
50
0
0
20
40
60
80
Negyedóra - napon belül (0 - 95)

MAPE



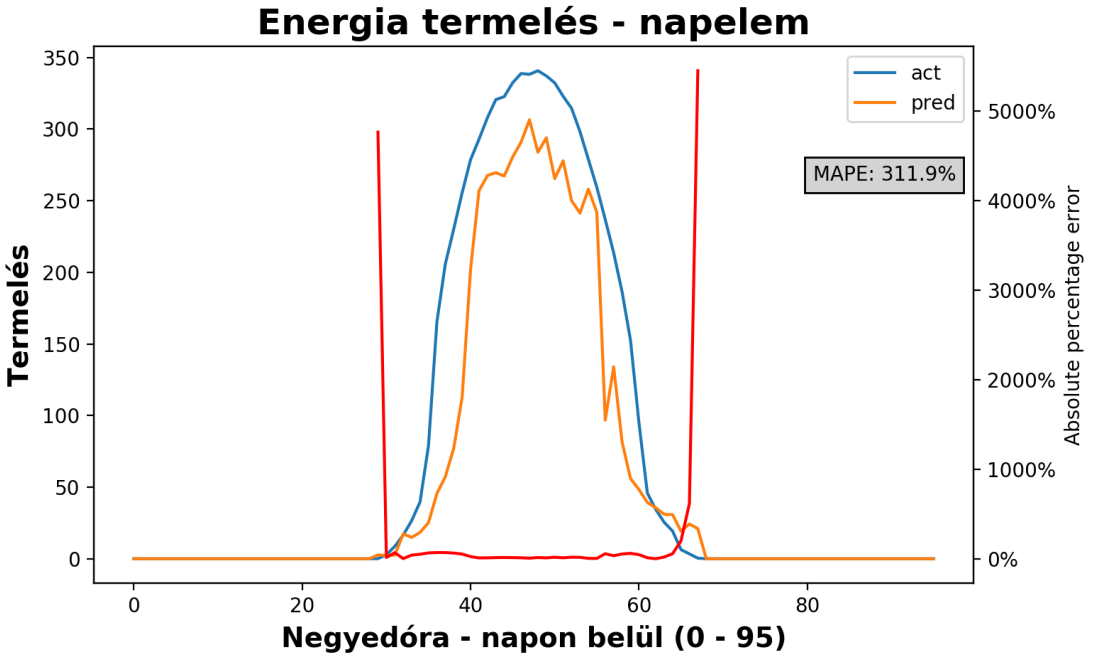
- Ha a célváltozó 0, akkor nem értelmezhető

MAPE



- Ha a célváltozó 0, akkor nem értelmezhető

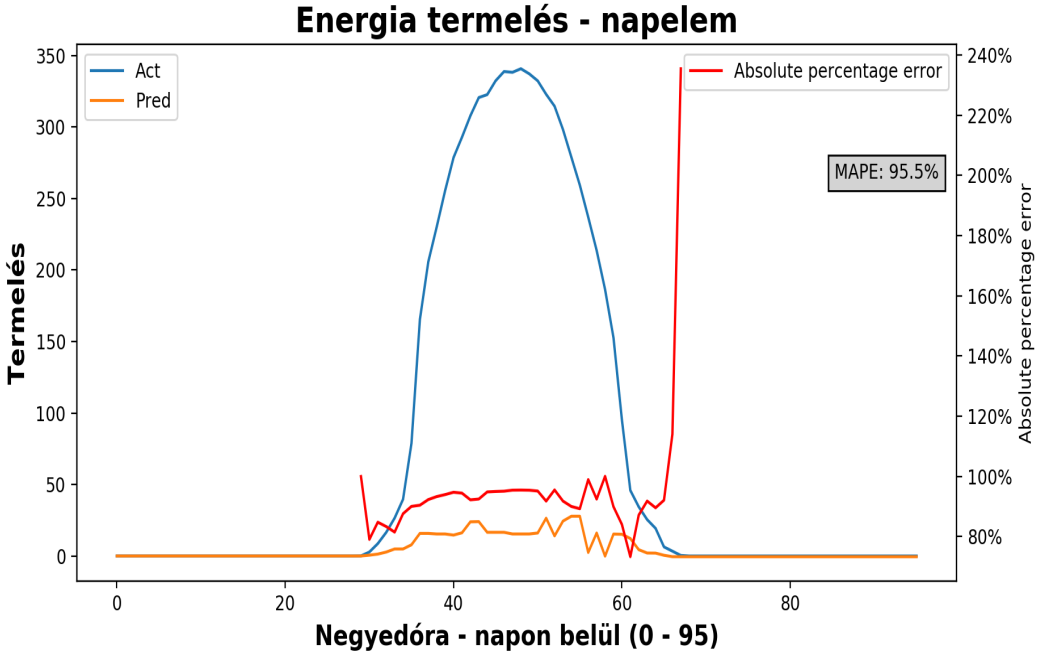
MAPE



- Ha a célváltozó 0, akkor nem értelmezhető
- Ha az értékkészlet szórása nagy, a MAPE nem általánosan értelmezhető

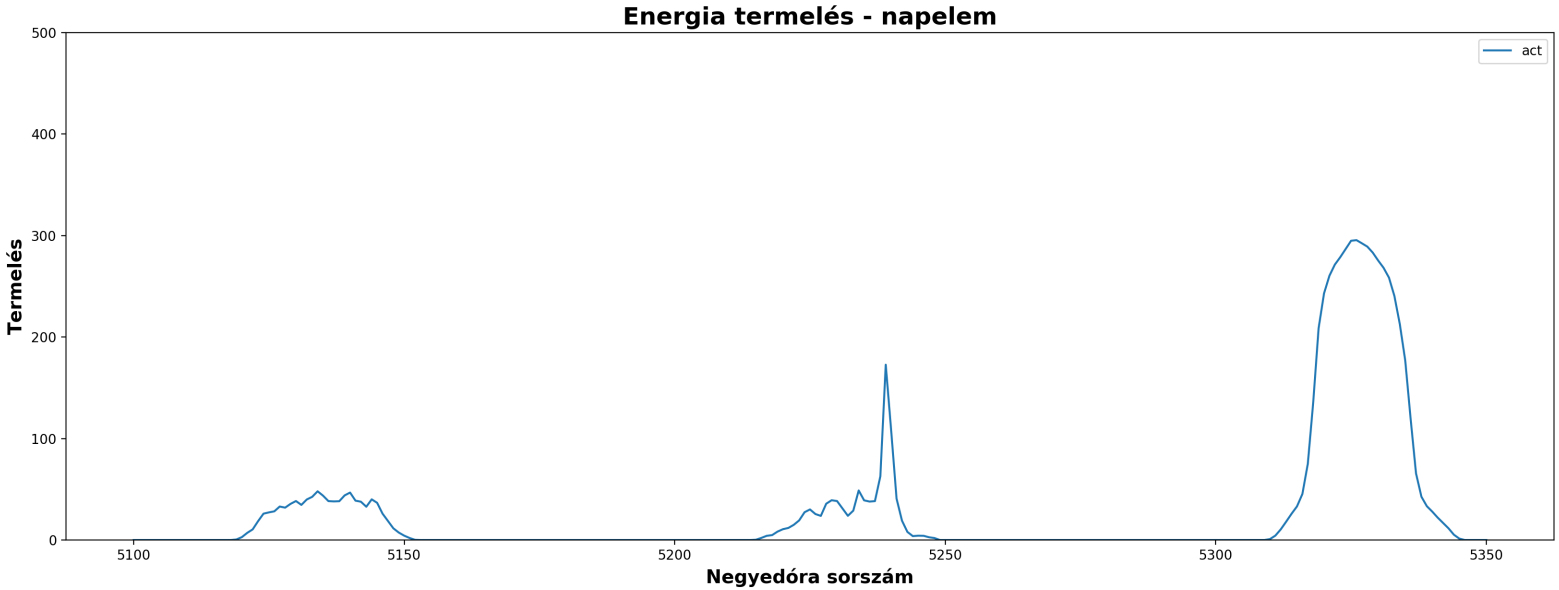
dmlab

MAPE

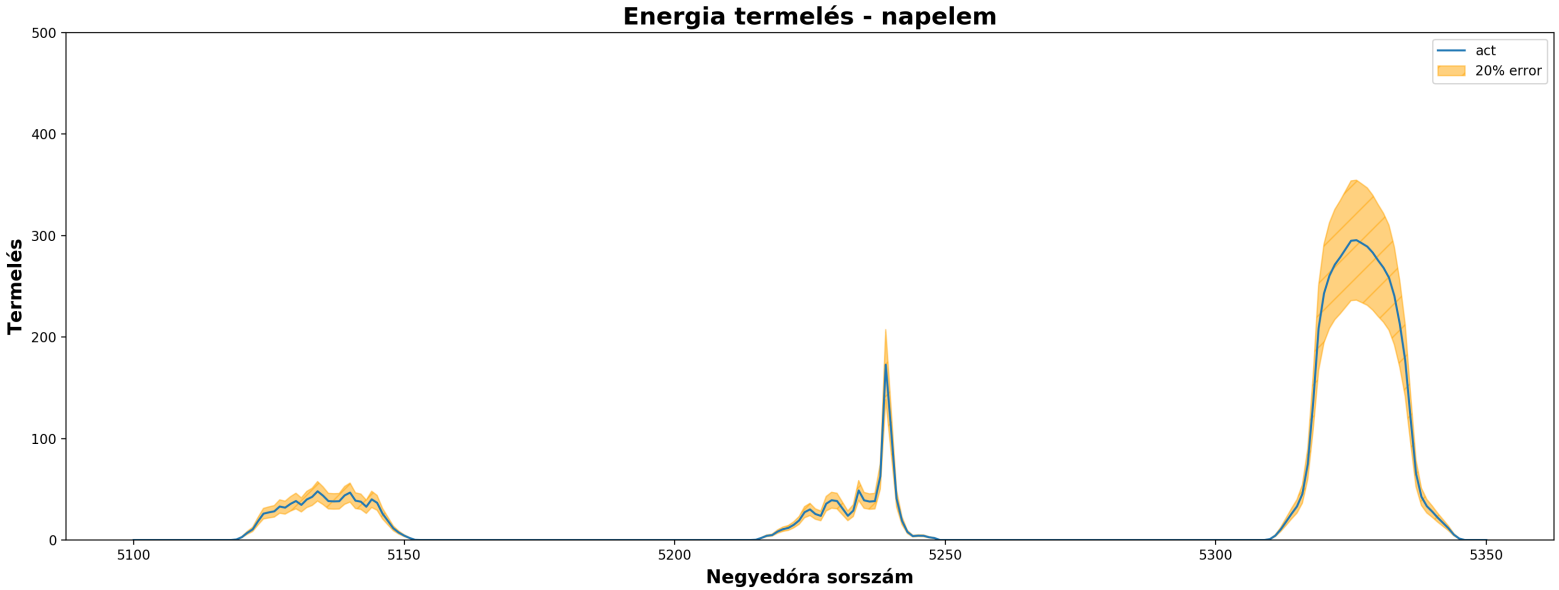

Energia termelés - napelem
Act
Pred
Absolute percentage error
MAPE: 95.5%
Termelés
Absolute percentage error
Negyedóra - napon belül (0 - 95)
350
300
250
200
150
100
50
0
240%
220%
200%
180%
160%
140%
120%
100%
80%
0
20
40
60
80

dmlab

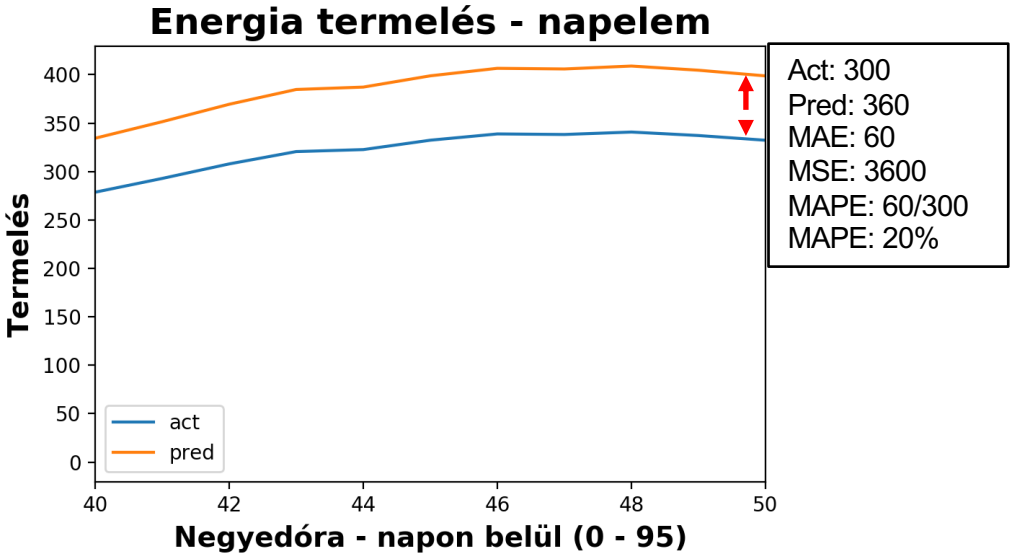
MAPE


Energia termelés - napelem
act
Termelés
Negyedóra sorszám
0
100
200
300
400
500
5100
5150
5200
5250
5300
5350

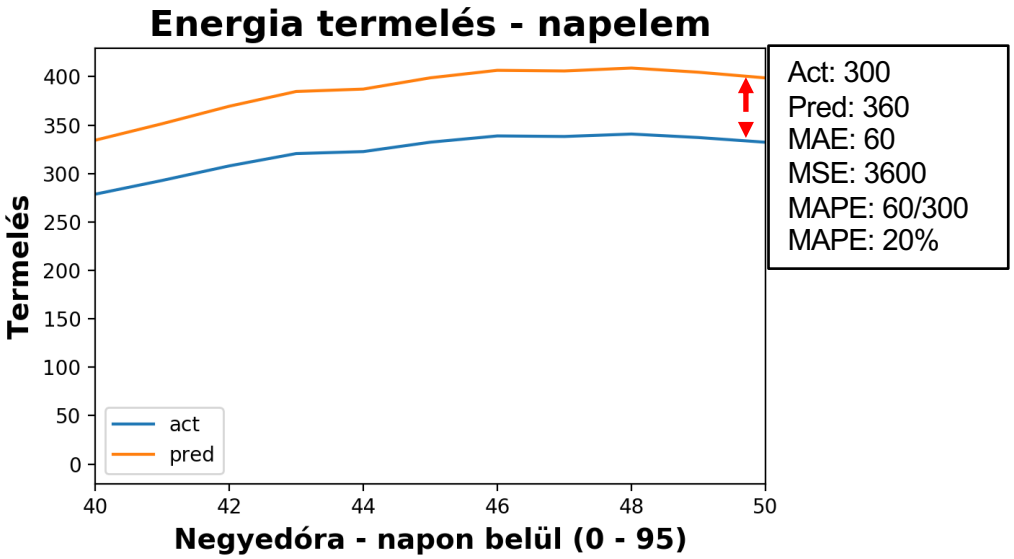
MAPE

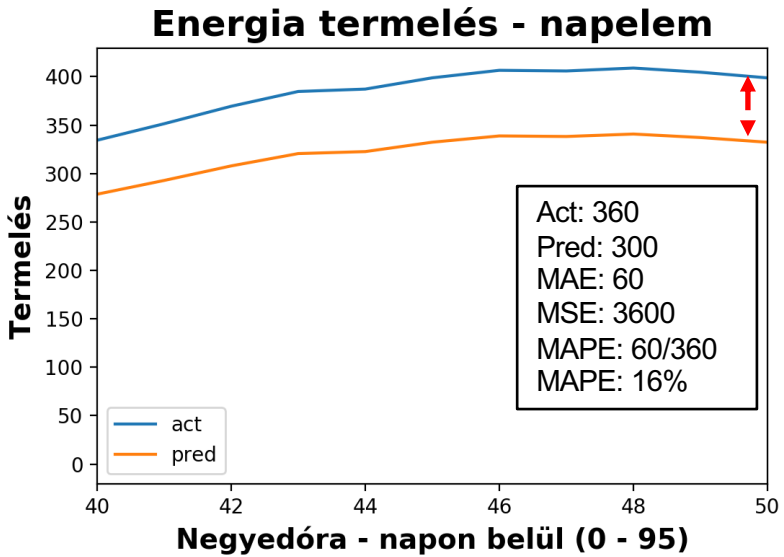

Energia termelés - napelem
act
20% error
Termelés
Negyedóra sorszám
0
100
200
300
400
500
5100
5150
5200
5250
5300
5350

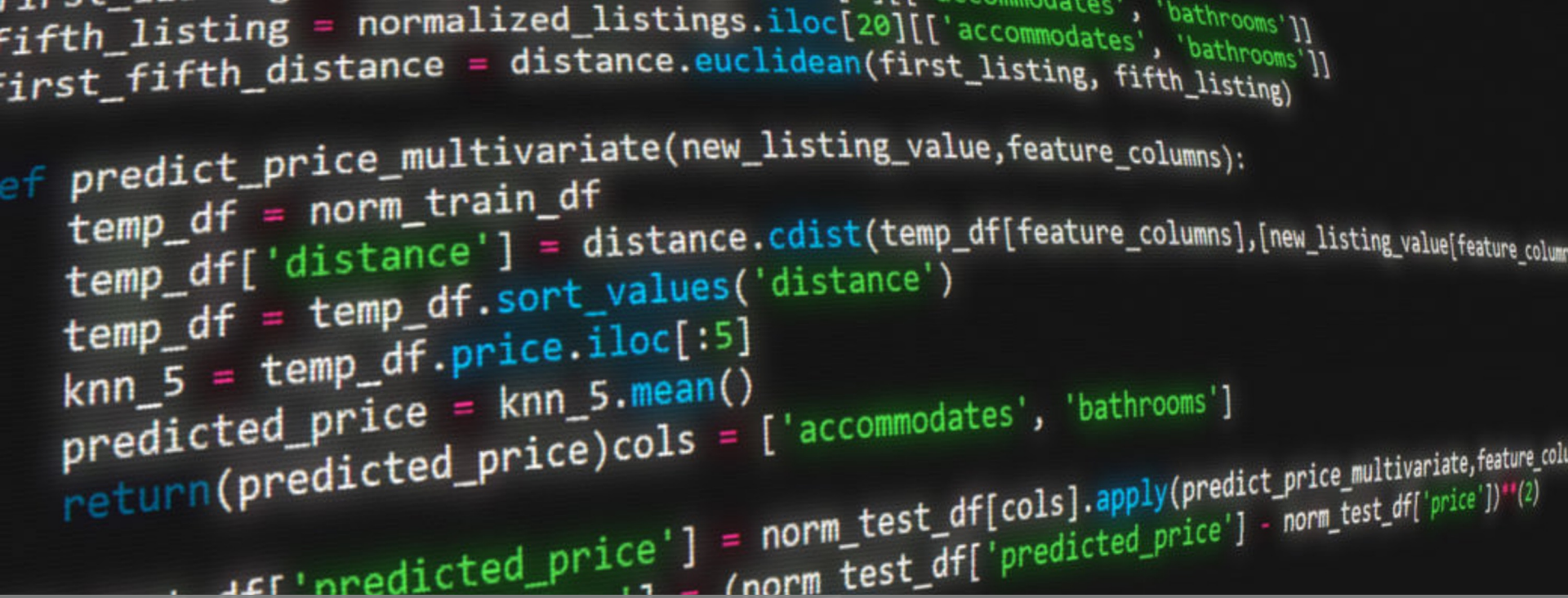
MAPE


Energia termelés - napelem
Termelés
Negyedóra - napon belül (0 - 95)
act
pred
0
50
100
150
200
250
300
350
400
40
42
44
46
48
50
Act: 300
Pred: 360
MAE: 60
MSE: 3600
MAPE: 60/300
MAPE: 20%

MAPE

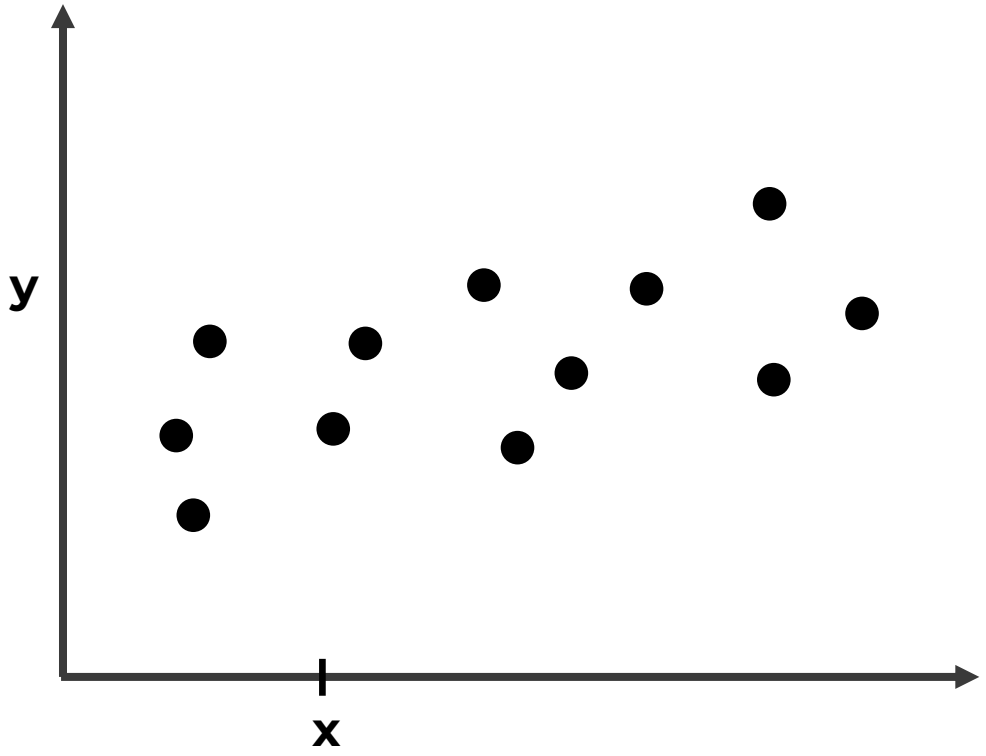

Energia termelés - napelem
Termelés
Negyedóra - napon belül (0 - 95)
act
pred
Act: 300
Pred: 360
MAE: 60
MSE: 3600
MAPE: 60/300
MAPE: 20%


Energia termelés - napelem
Termelés
Negyedóra - napon belül (0 - 95)
act
pred
Act: 360
Pred: 300
MAE: 60
MSE: 3600
MAPE: 60/360
MAPE: 16%



LINEÁRIS REGRESSZIÓ

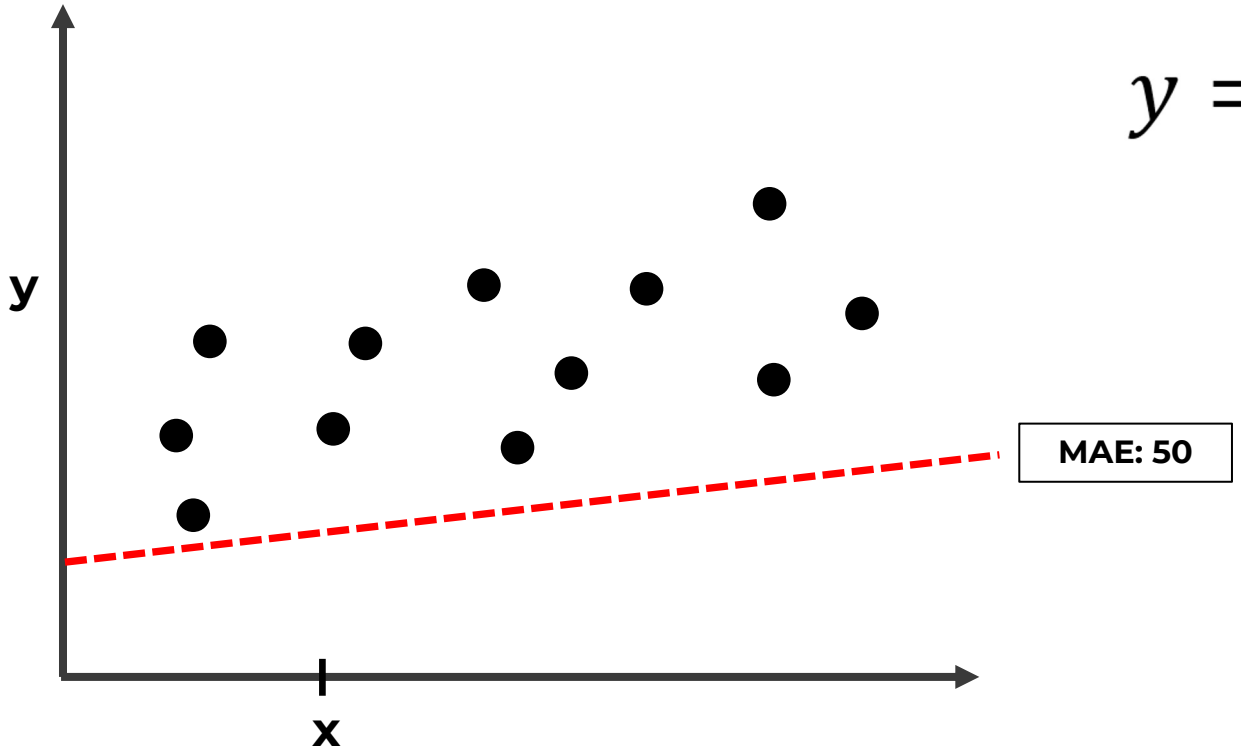
Lineáris regresszió



$$y = b_0 + b_1 * x$$

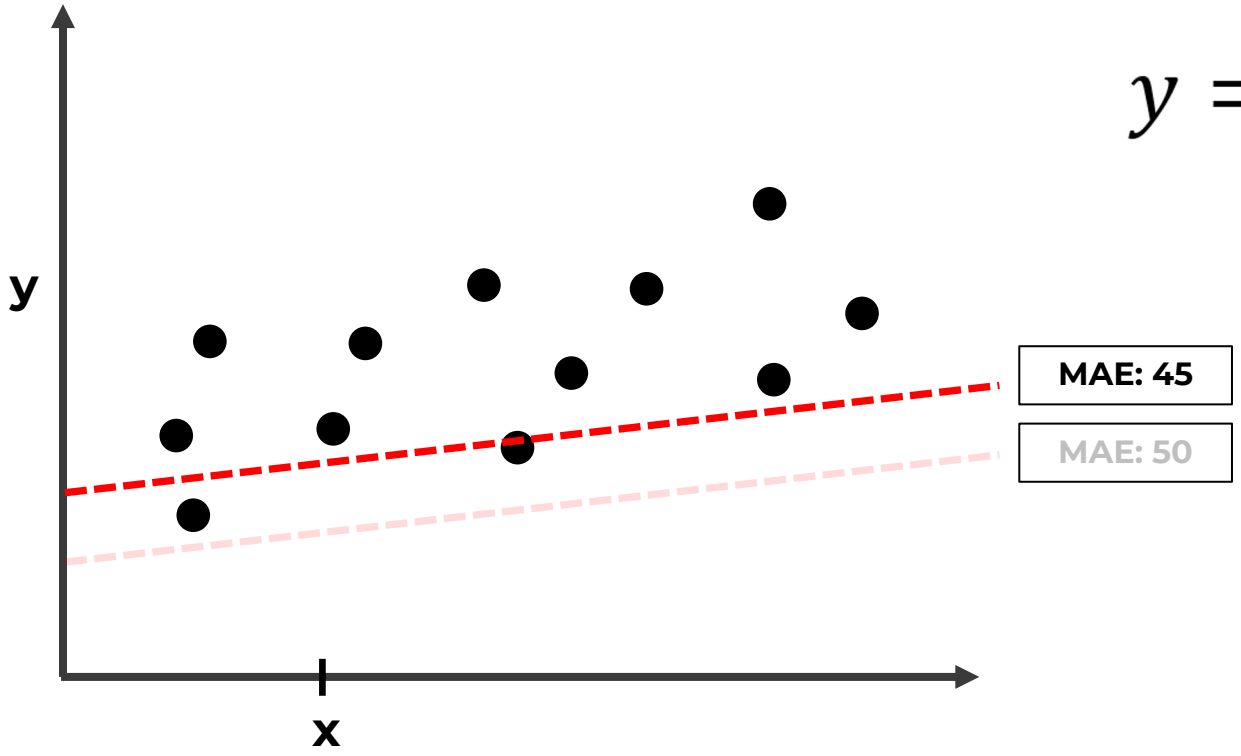
Lineáris regresszió

$$y = b_0 + b_1 * x$$

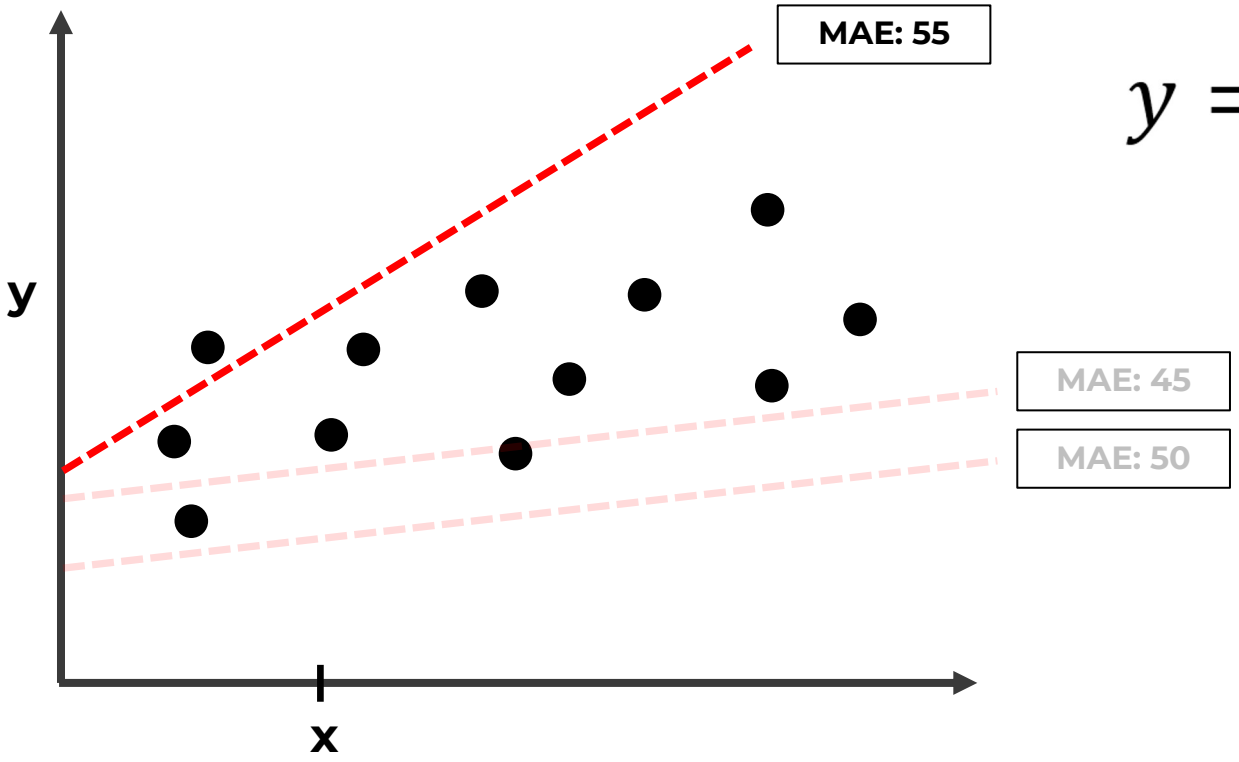


Lineáris regresszió

$$y = b_0 + b_1 * x$$

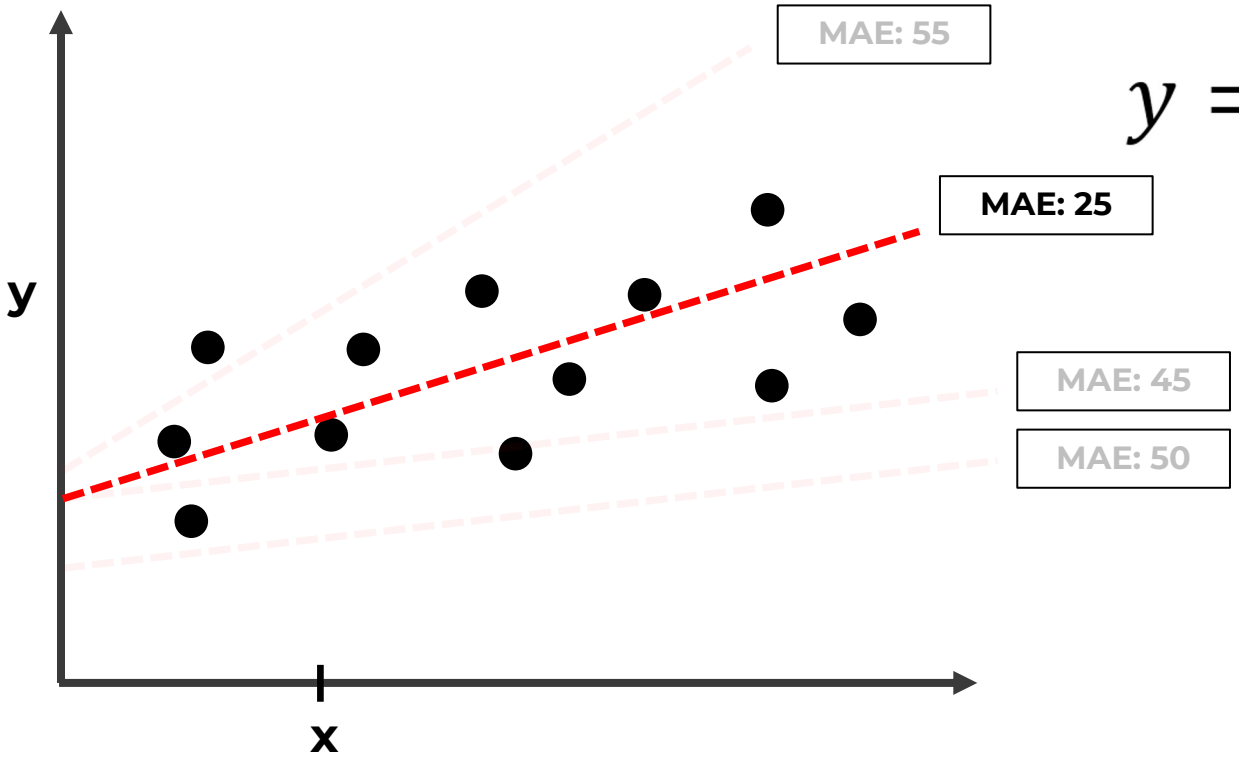


Lineáris regresszió



$$y = b_0 + b_1 * x$$

Lineáris regresszió



$$y = b_0 + b_1 * x$$

LINEÁRIS REGRESSZIÓ

area	*rooms*	*district*	*elevator*	*price*
120	4	3	1	90
50	2	3	0	45
82	3	8	1	59
46	1	13	0	40
70	3	5	0	71
65	2	3	1	54

X input features *y target*

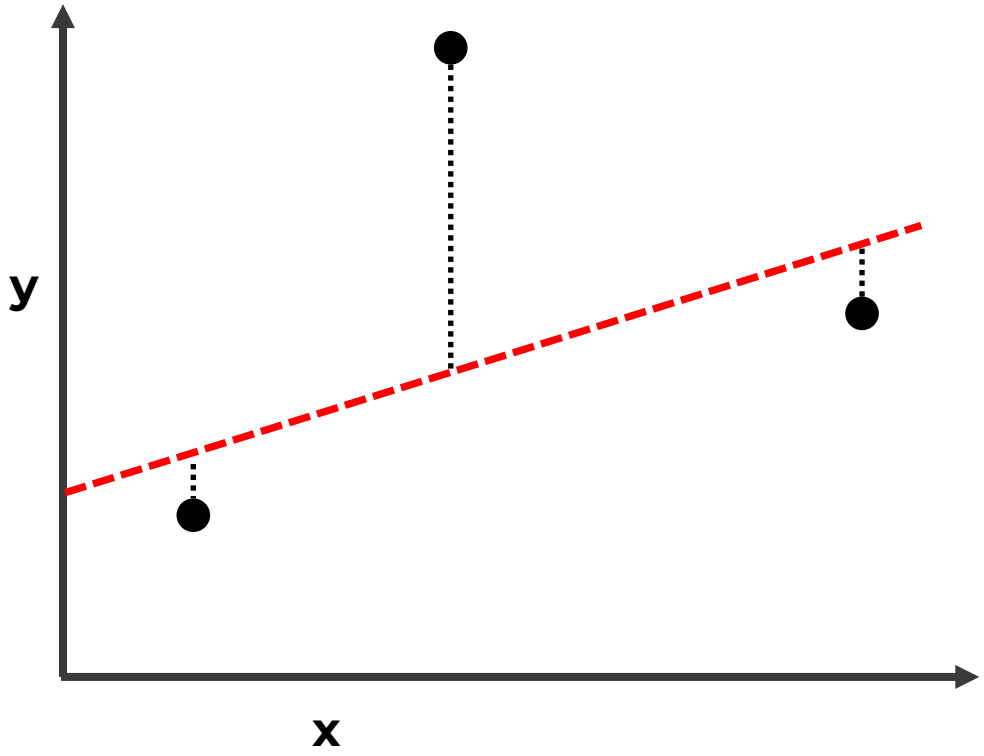
LINEÁRIS REGRESSZIÓ

area	*rooms*	*district*	*elevator*	*price*
120	4	3	1	90
50	2	3	0	45
82	3	8	1	59
46	1	13	0	40
70	3	5	0	71
65	2	3	1	54

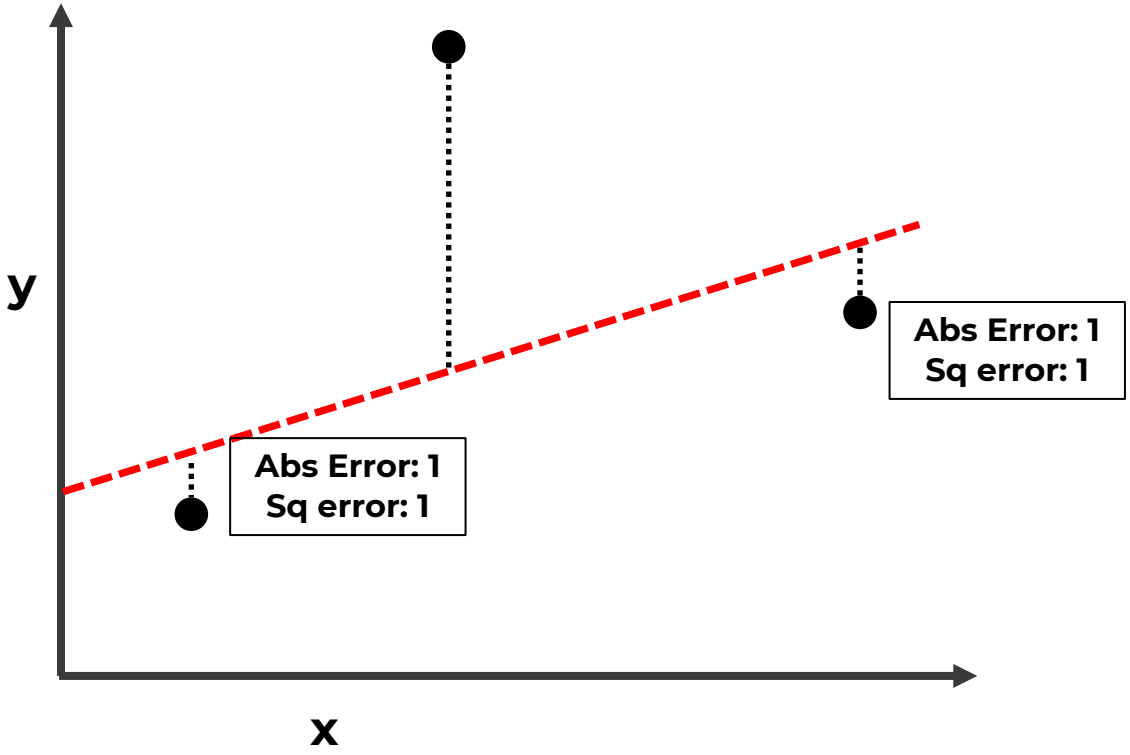
X input features *y target*

$$y = b_0 + b_1 * x_1 + \cdots + b_n * x_n$$

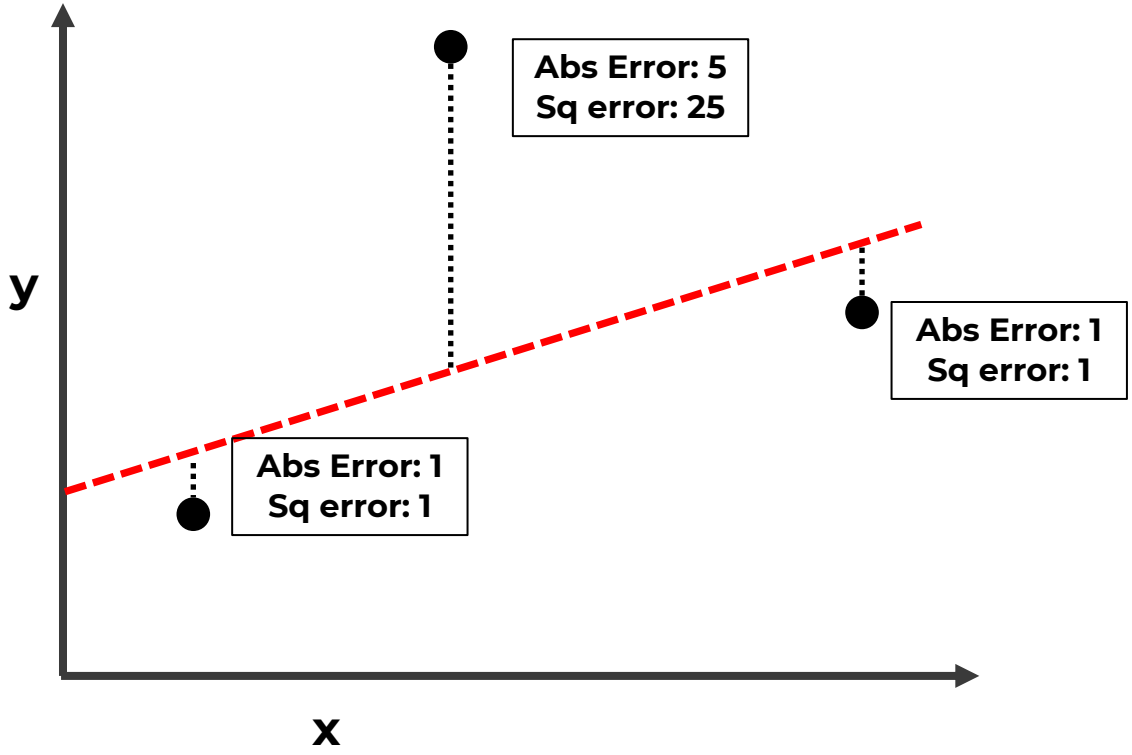
MSE



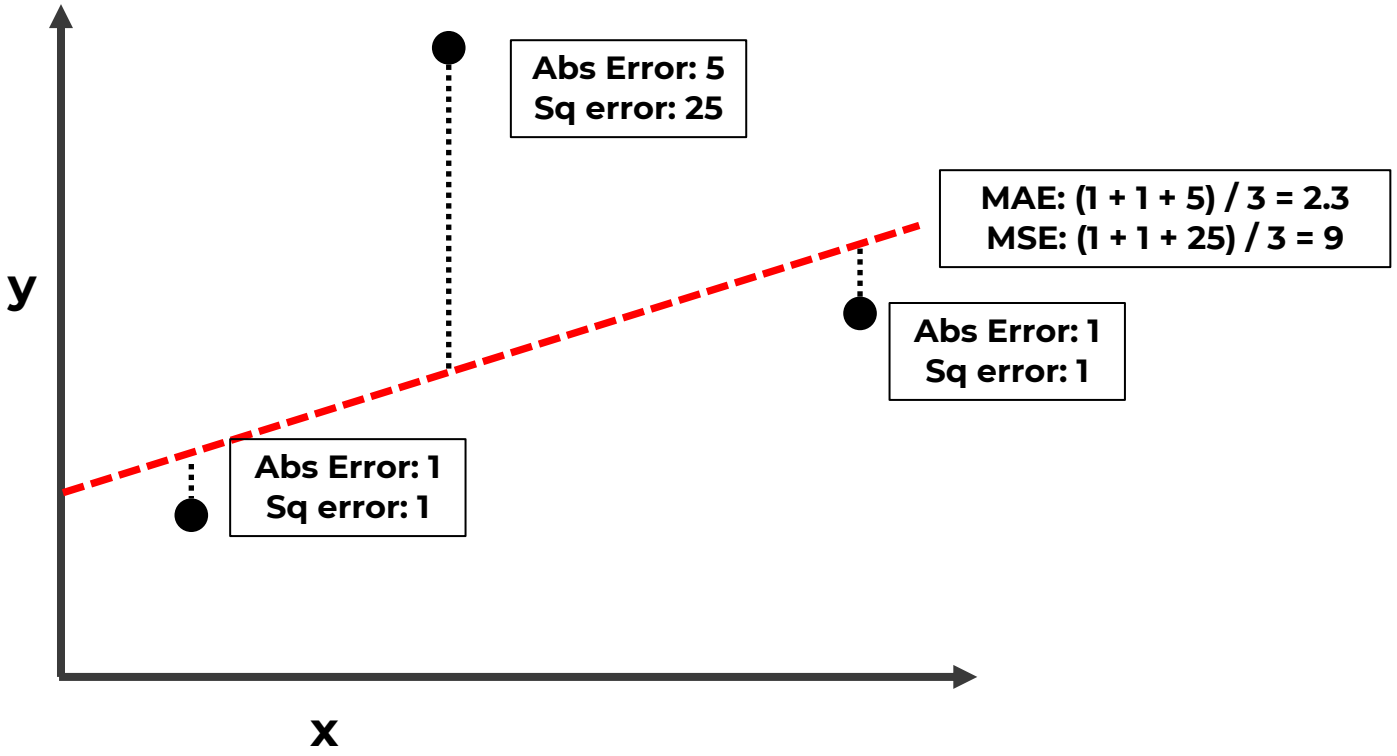
MSE



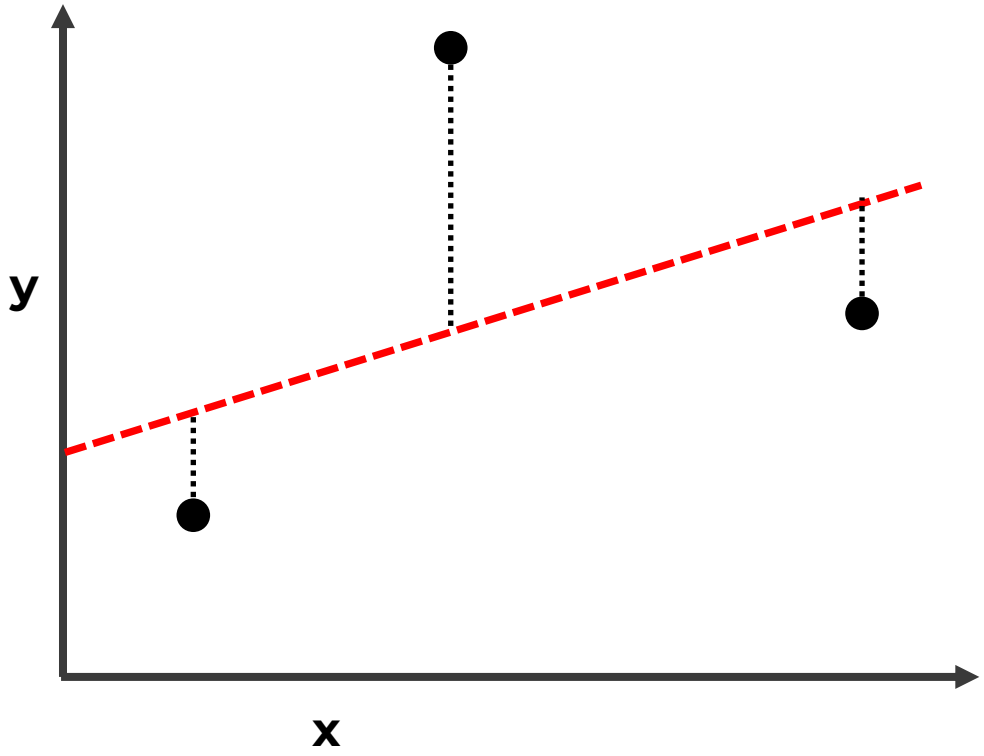
MSE


Abs Error: 5
Sq error: 25
Abs Error: 1
Sq error: 1
Abs Error: 1
Sq error: 1
y
x

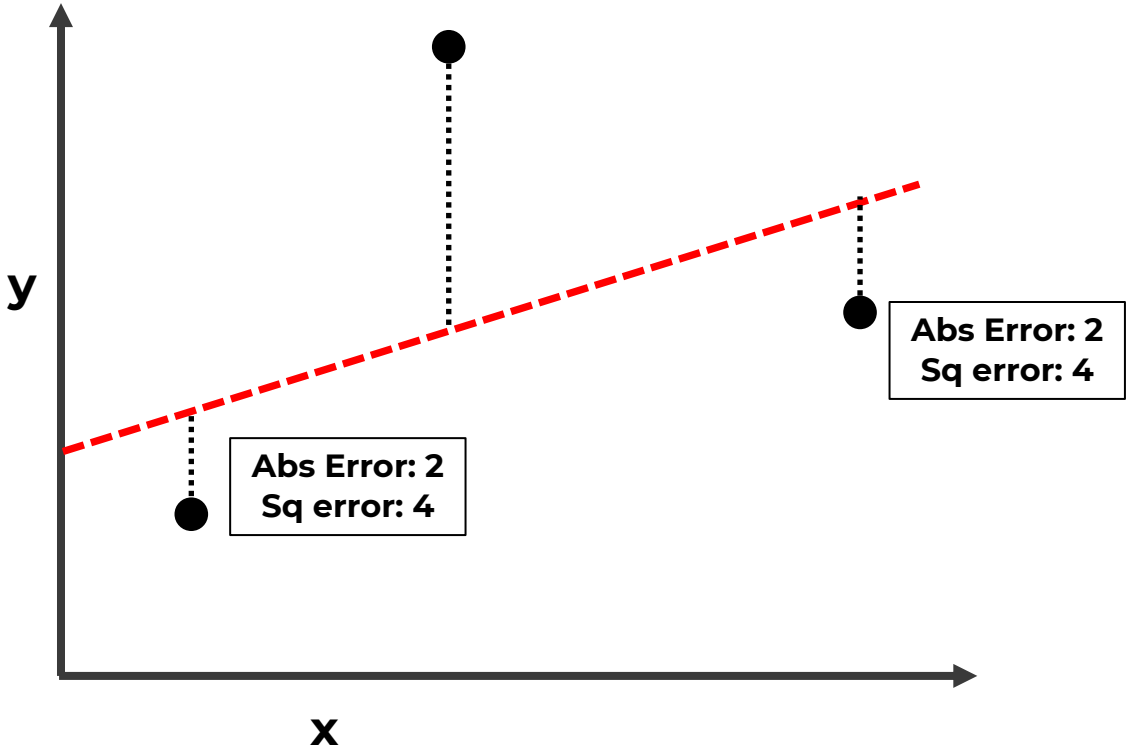
MSE


Abs Error: 5
Sq error: 25
MAE: (1 + 1 + 5) / 3 = 2.3
MSE: (1 + 1 + 25) / 3 = 9
Abs Error: 1
Sq error: 1
Abs Error: 1
Sq error: 1
y
x

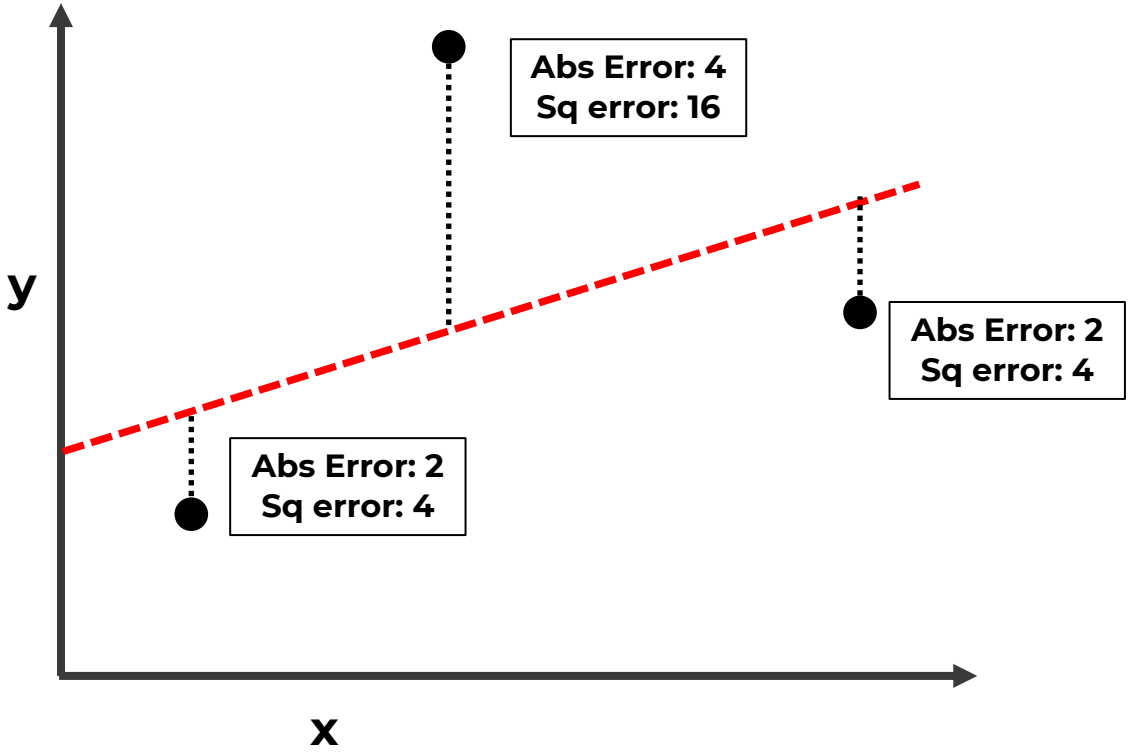
MSE



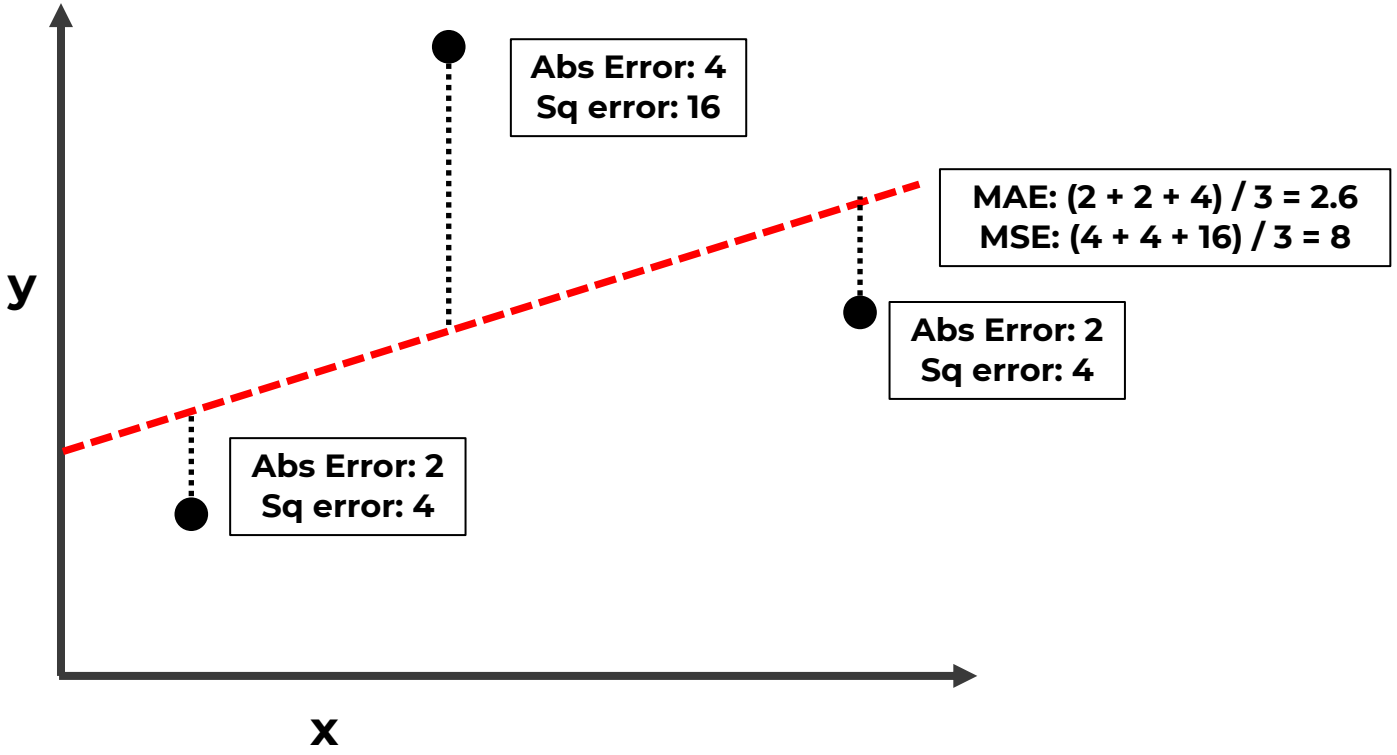
MSE



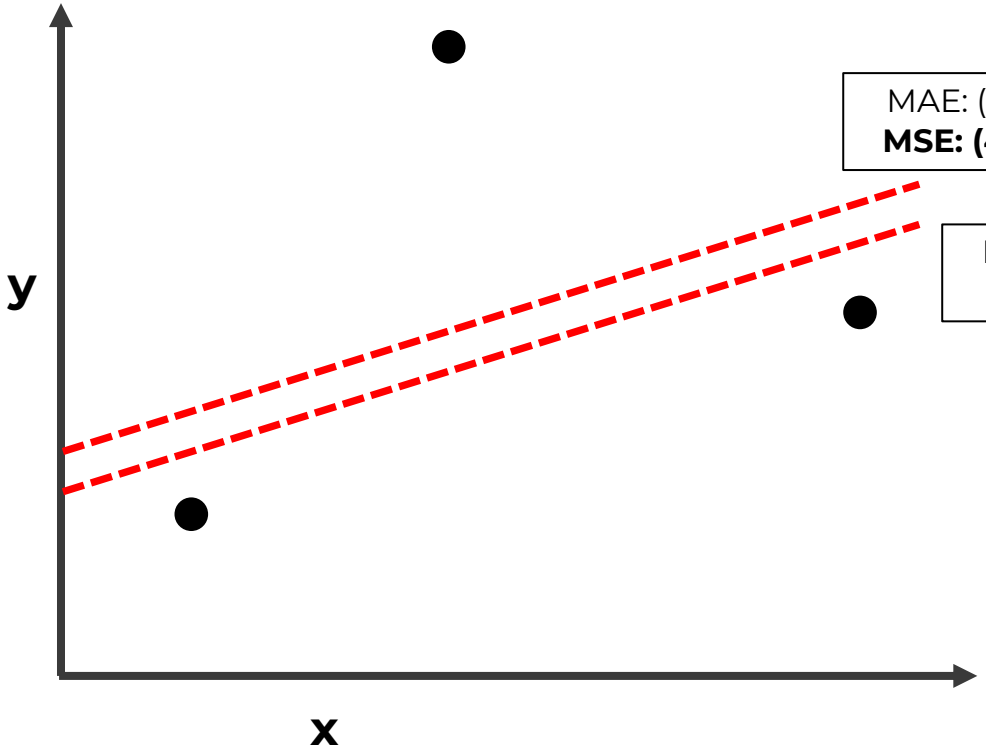
MSE


Abs Error: 4
Sq error: 16
Abs Error: 2
Sq error: 4
Abs Error: 2
Sq error: 4
y
x

MSE


Abs Error: 4
Sq error: 16
MAE: (2 + 2 + 4) / 3 = 2.6
MSE: (4 + 4 + 16) / 3 = 8
Abs Error: 2
Sq error: 4
Abs Error: 2
Sq error: 4
y
x

MSE



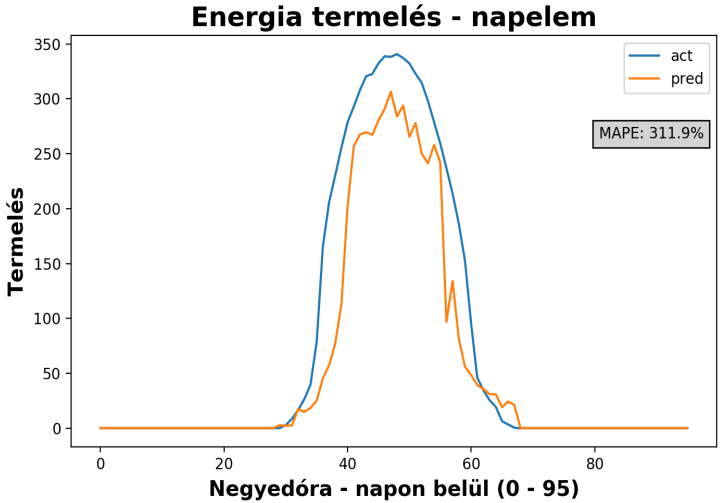
MAE: (2 + 2 + 4) / 3 = 2.6
MSE: (4 + 4 + 16) / 3 = 8

MAE: (1 + 1 + 5) / 3 = 2.3
MSE: (1 + 1 + 25) / 3 = 9

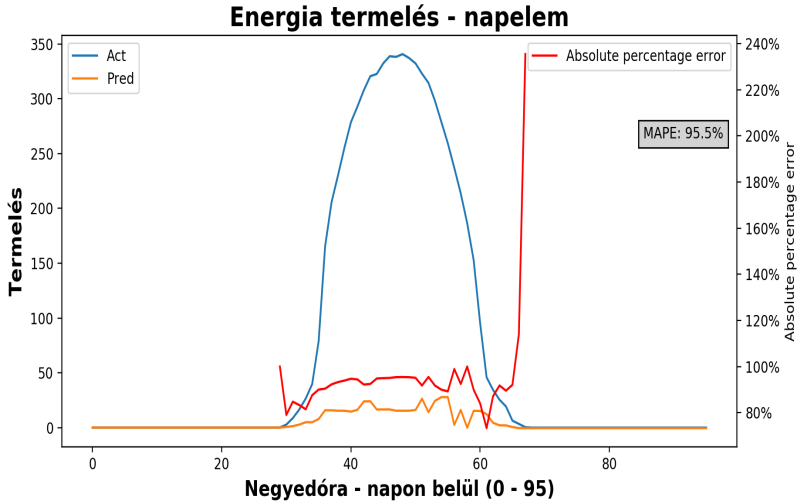
- Megoldásunk függ a célfüggvénytől

MAE vs. MAPE

Célfüggvény: MAE



Célfüggvény: MAPE



dmlab


ÜZLETI CÉLOK MEGHATÁROZÁSA
1
ADATOK MEGÉRTÉSE
2
ADATOK ELŐKÉSZÍTÉSE
3
MODELLEZÉS
4
ÜZLETI KIÉRTÉKELÉS
5
HADRENDBE ÁLLÍTÁS
6

dmlab