Osztályozás Metrikák

Data science képzés

AUC

2024.10.07.

Jónás Dániel,
data scientist

Osztályozás

Metrikák

- Accuracy – pontosság ☹
  - Kiegyensúlyozatlan adathalmaz esetén megtévesztő eredmények

		Aktuális *(y)*	
		Negatív (0)	**Pozitív (1)**
Prediktált *(y_pred)*	**Negatív (0)**	99	1
	Pozitív (1)	0	0

Accuracy = (99 + 0) / (99 + 1 + 0 + 0) = 99/100 -> 99%

Osztályozás

Metrikák

- Konfúziós mátrix ☹
  - Recall – precision célcsoportonként eltér

Prediktált (*y_pred*) \ Aktuális (*y*)	Negatív (0)	Pozitív (1)
Negatív (0)	10	20
Pozitív (1)	20	20

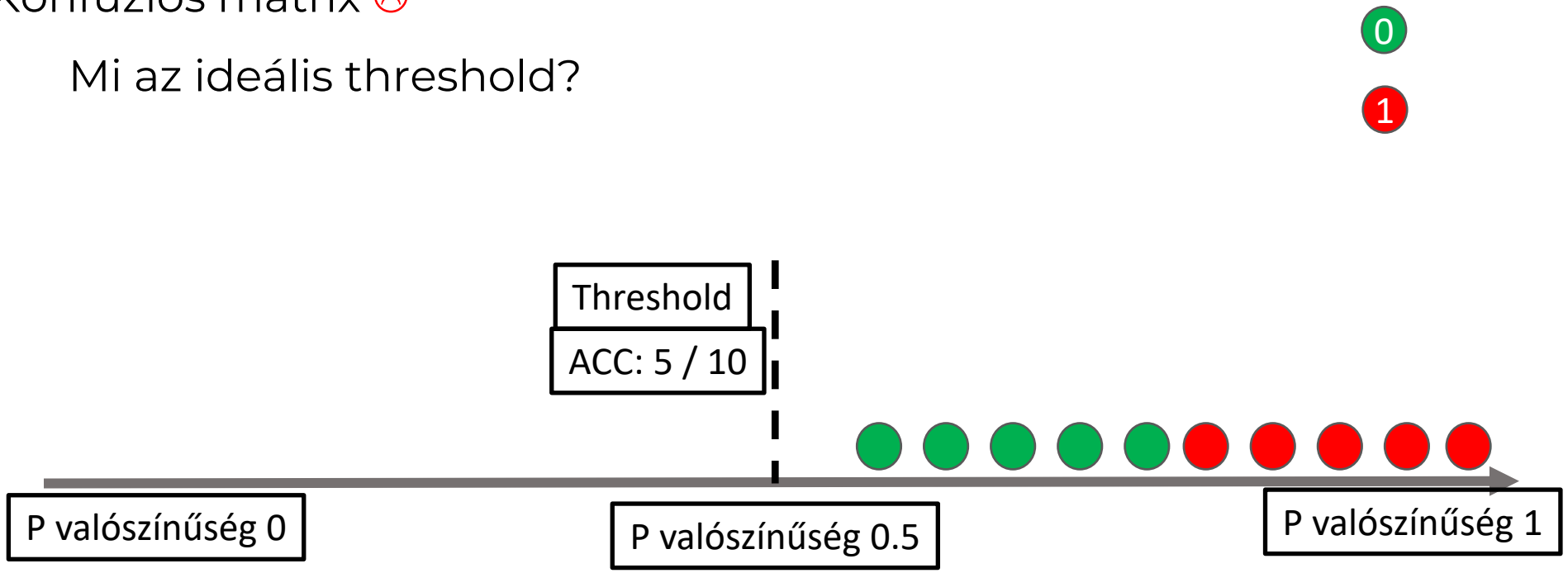
Prec(0) = 10 / (10 + 20) = 0.33
Prec(1) = 20 / (20 + 20) = 0.5

Recall(0) = 10 / (10 + 20) = 0.33
Recall(1) = 20 / (20 + 20) = 0.5

Osztályozás

Metrikák

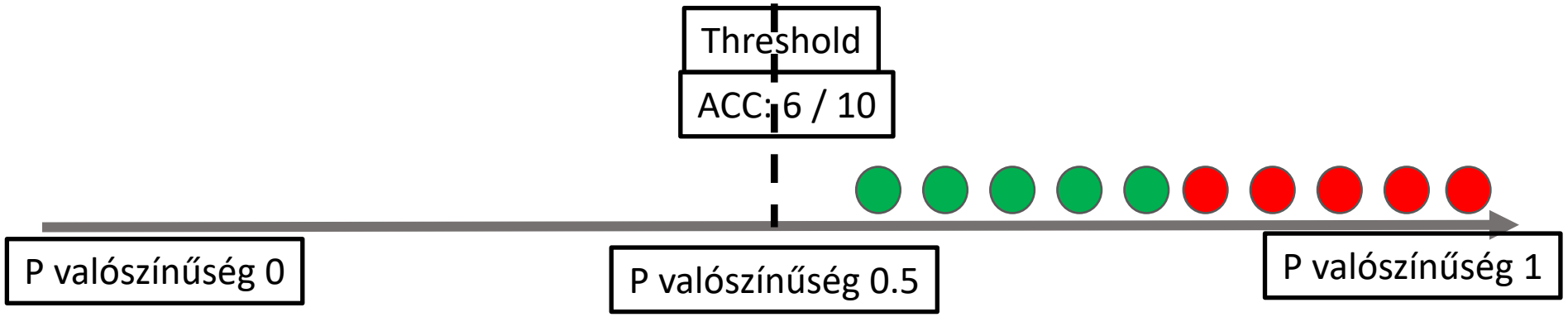
- Konfúziós mátrix ☹
  - Mi az ideális threshold?



Osztályozás

Metrikák

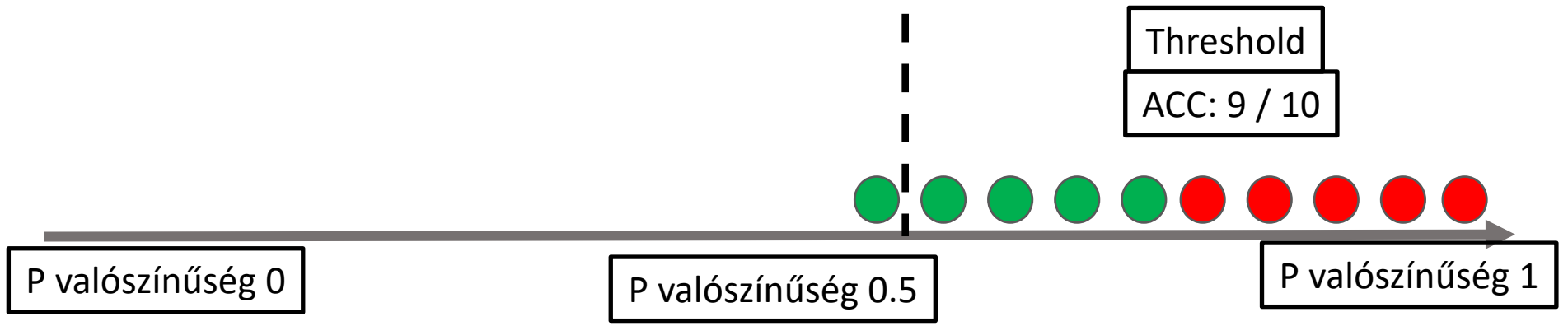
- Konfúziós mátrix ☹
  - Mi az ideális threshold?



Osztályozás

Metrikák

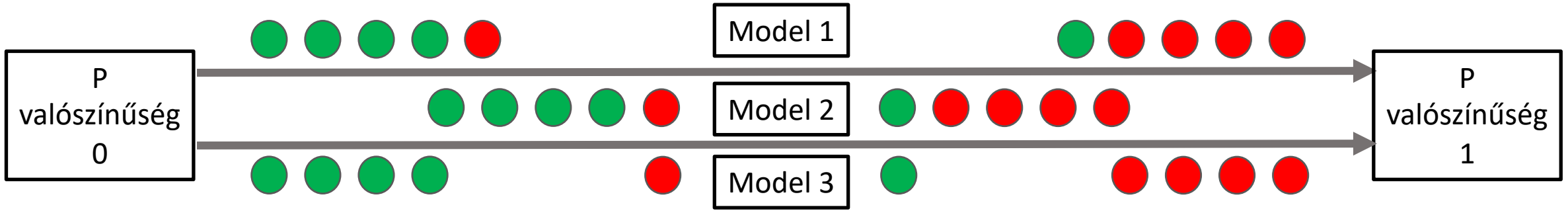
- Konfúziós mátrix ☹
  - Mi az ideális threshold?



Osztályozás

Metrikák

- Konfúziós mátrix ☹
  - Nem törődünk a modell magabiztosságával



Osztályozás

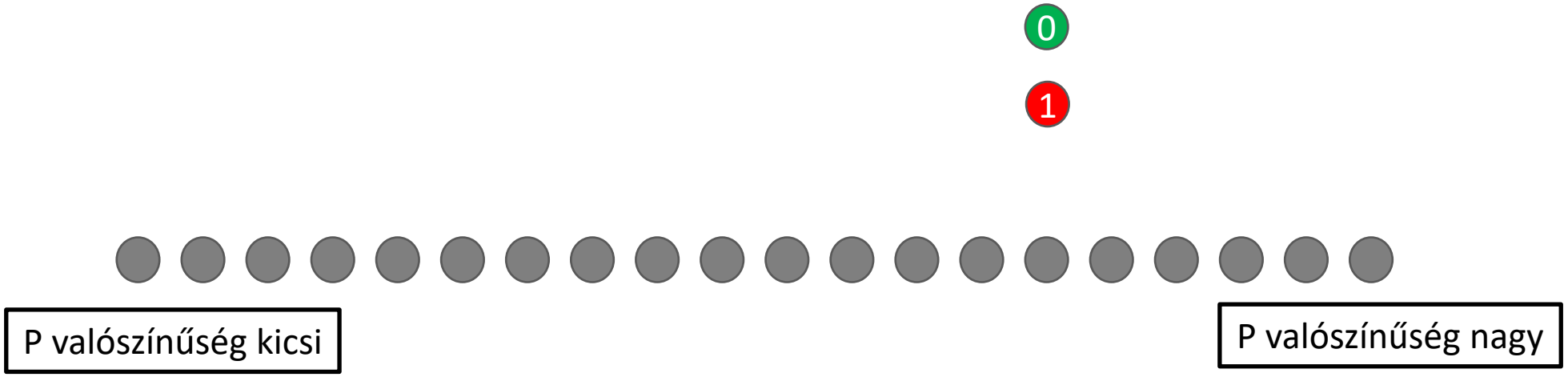
Metrikák

- Accuracy – pontosság ☹
  - Kiegyensúlyozatlan adathalmaz esetén megtévesztő eredmények
- Konfúziós mátrix ☹
  - Mi az ideális threshold?
  - Recall – precision célcsoportonként eltér
  - Nem törődünk a modell magabiztosságával
- De akkor mi? ☺

Osztályozás

Predict_proba

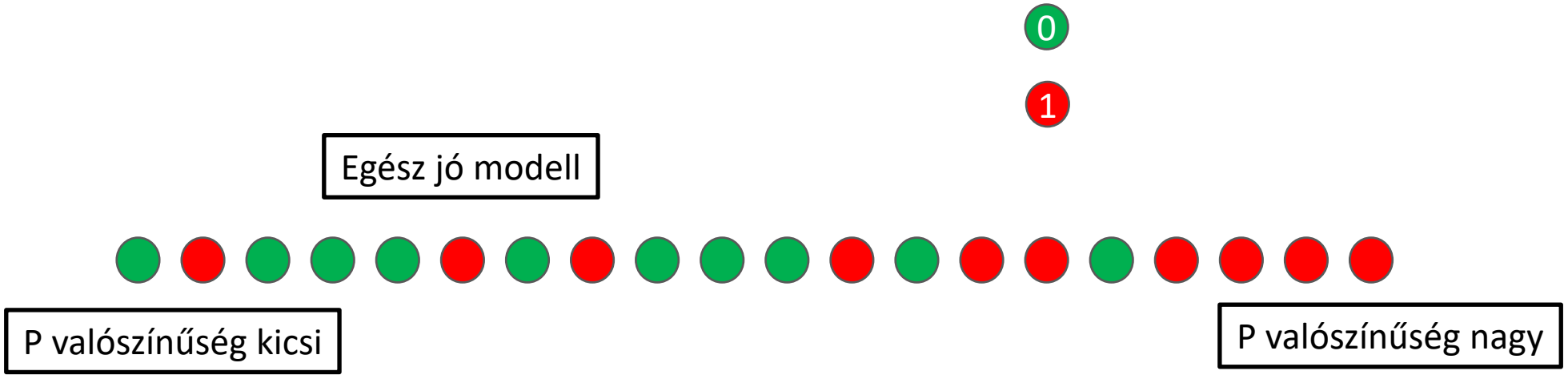
- Foglalkozzunk a modell magabiztosságával



Osztályozás

Predict_proba

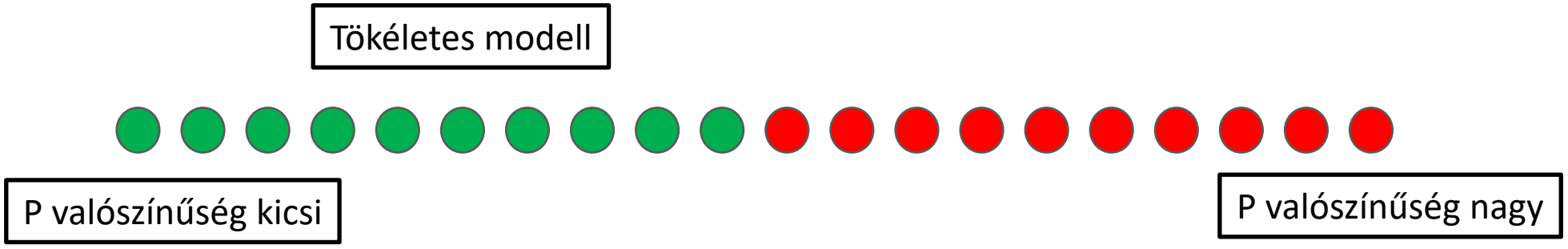
- Foglalkozzunk a modell magabiztosságával



Osztályozás

Predict_proba

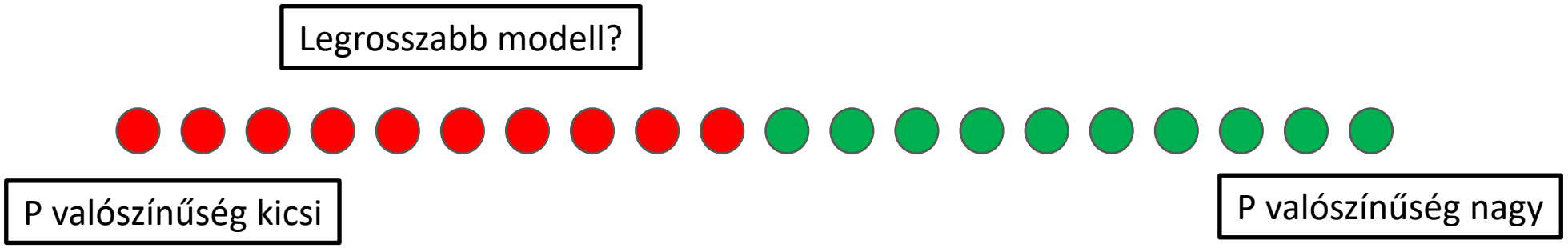
- Foglalkozzunk a modell magabiztosságával



Osztályozás

Predict_proba

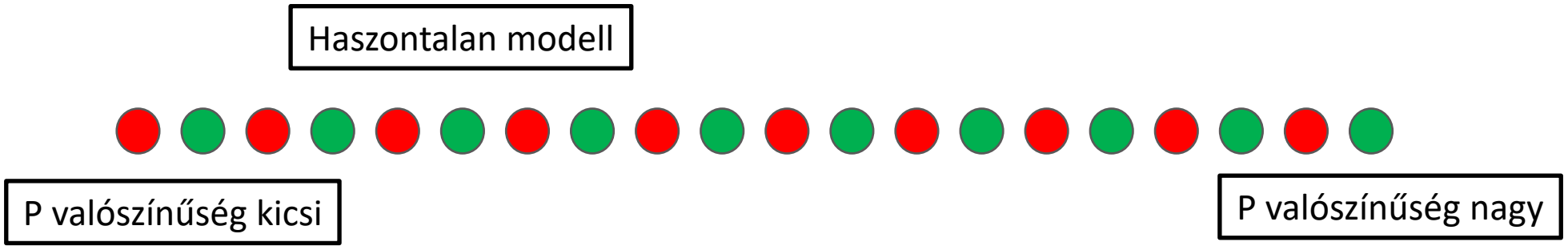
- Foglalkozzunk a modell magabiztosságával



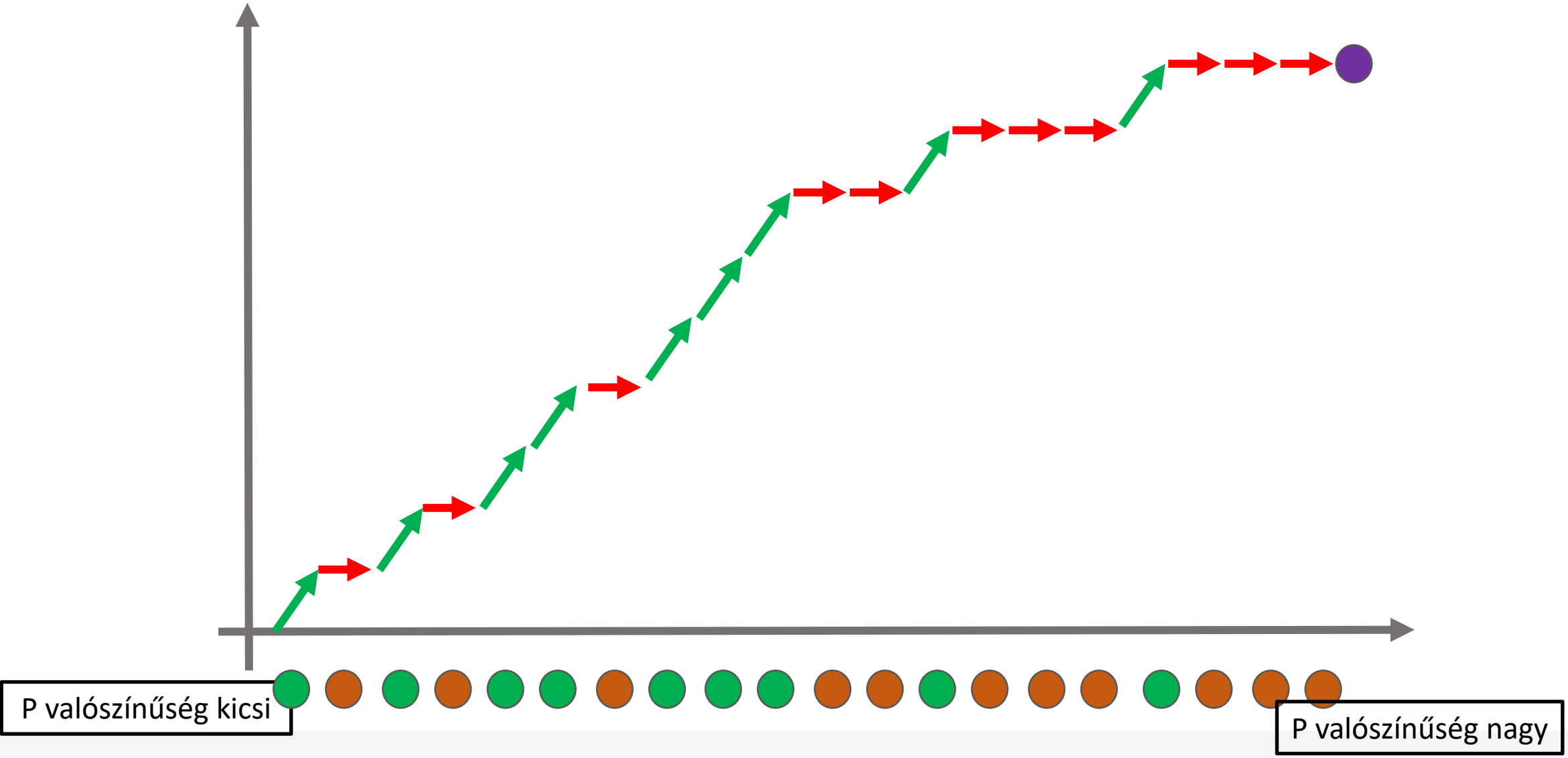
Osztályozás

Predict_proba

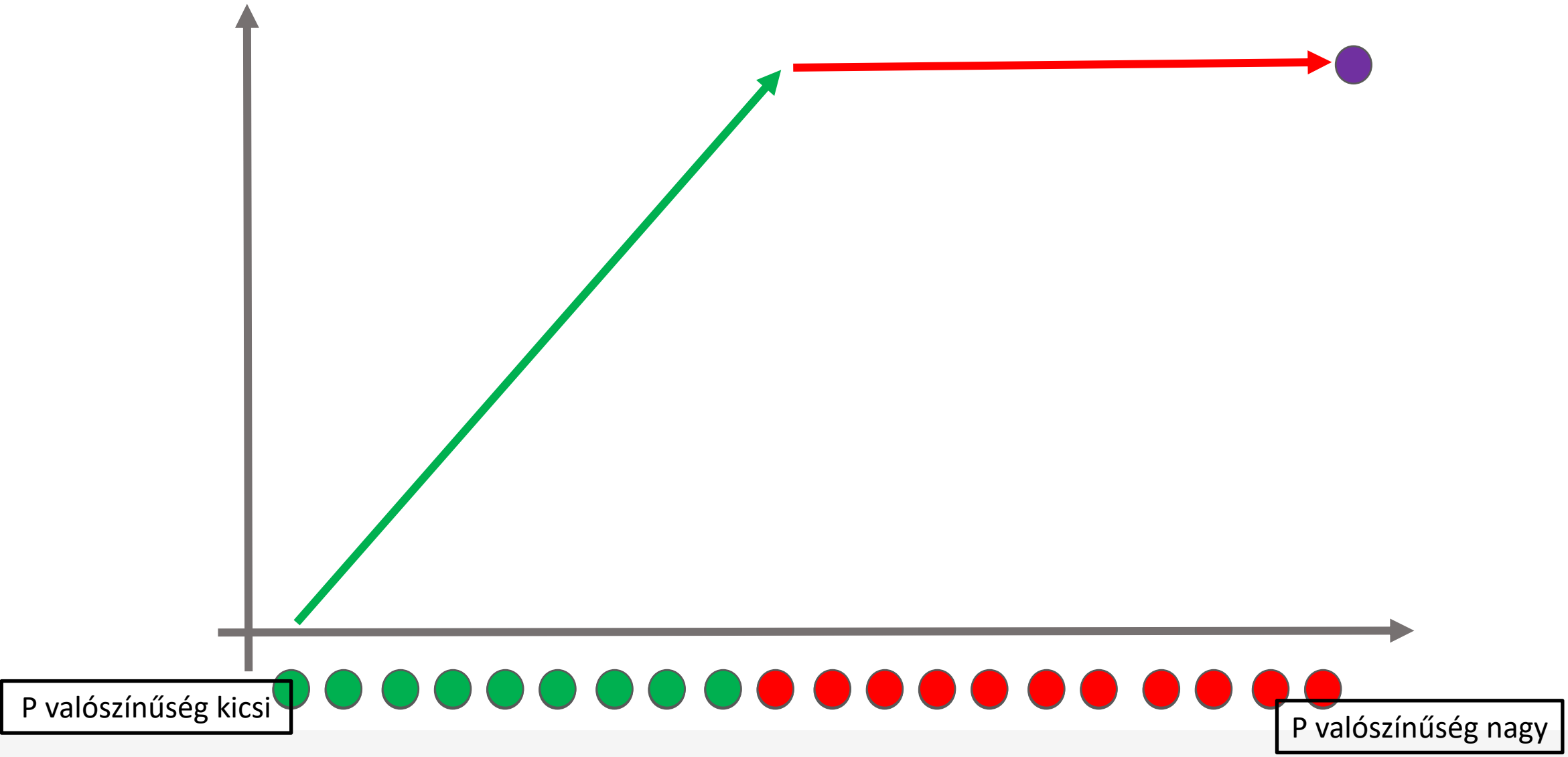
- Foglalkozzunk a modell magabiztosságával



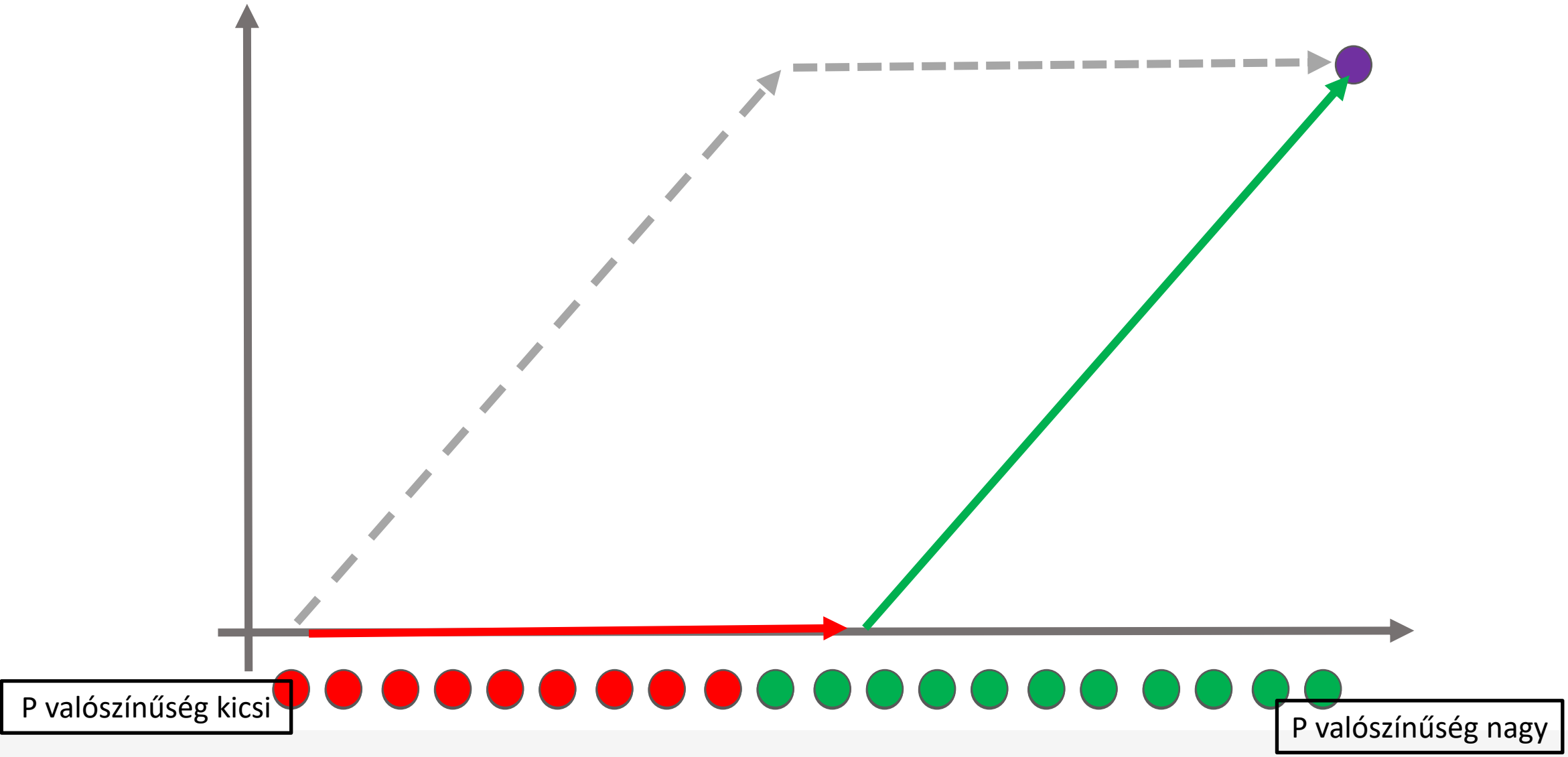
Osztályozás - Gains



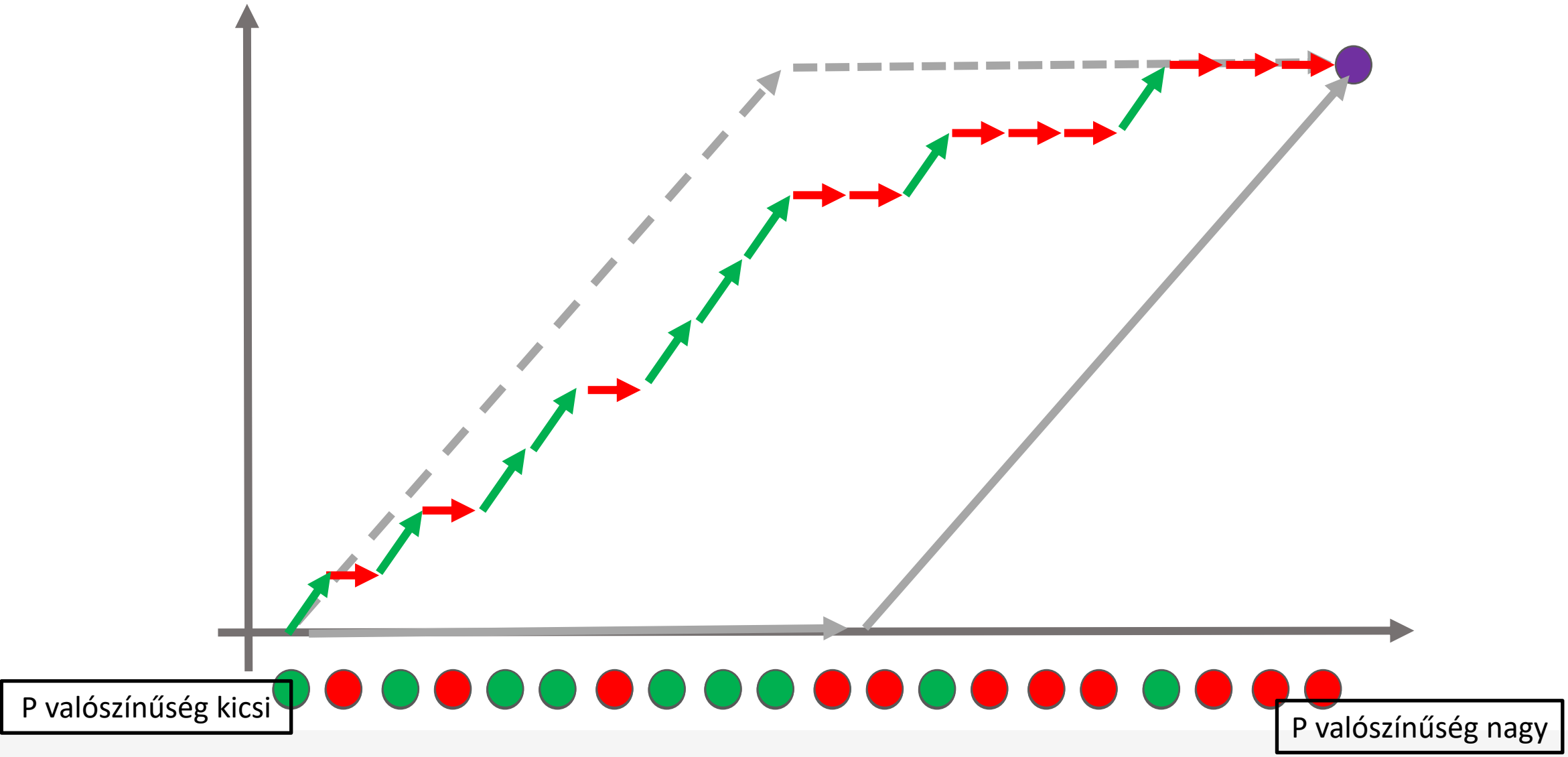
Osztályozás - Gains



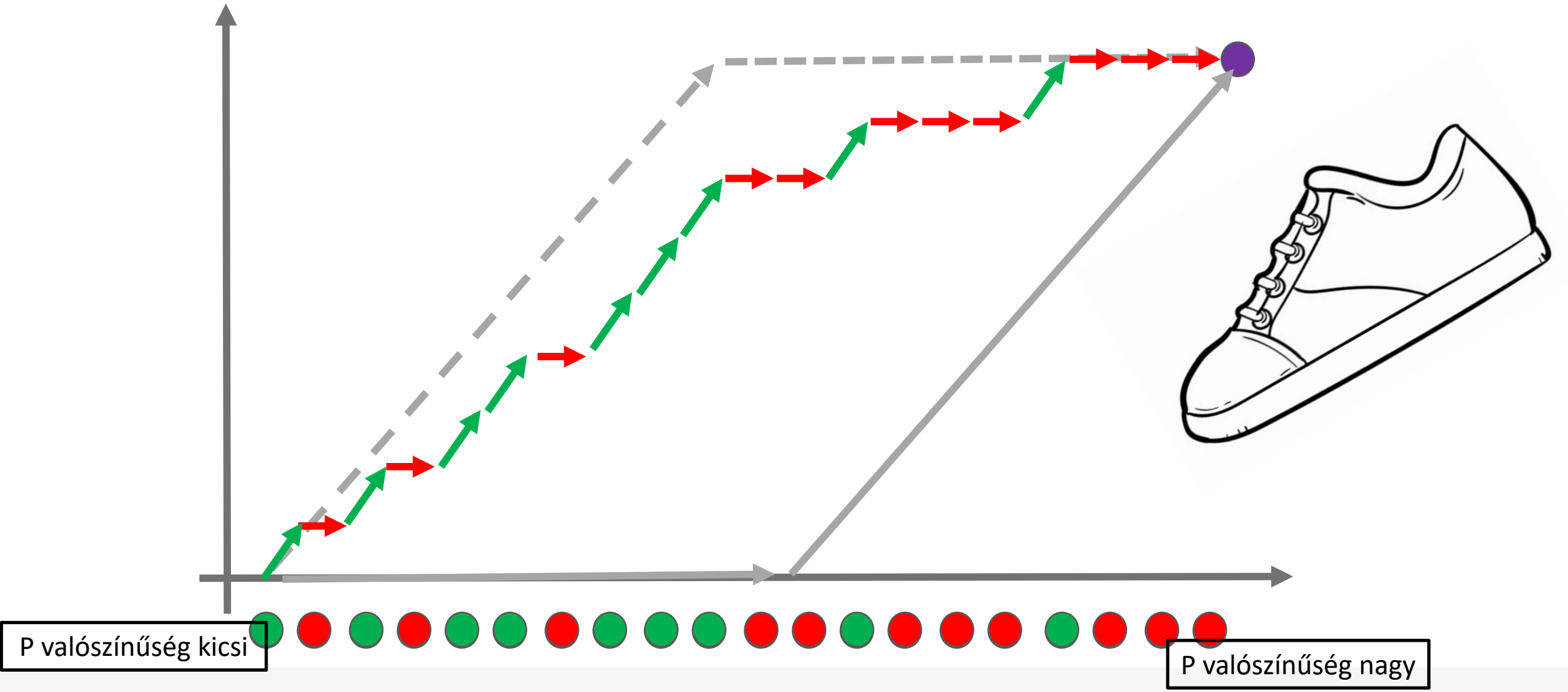
Osztályozás - Gains



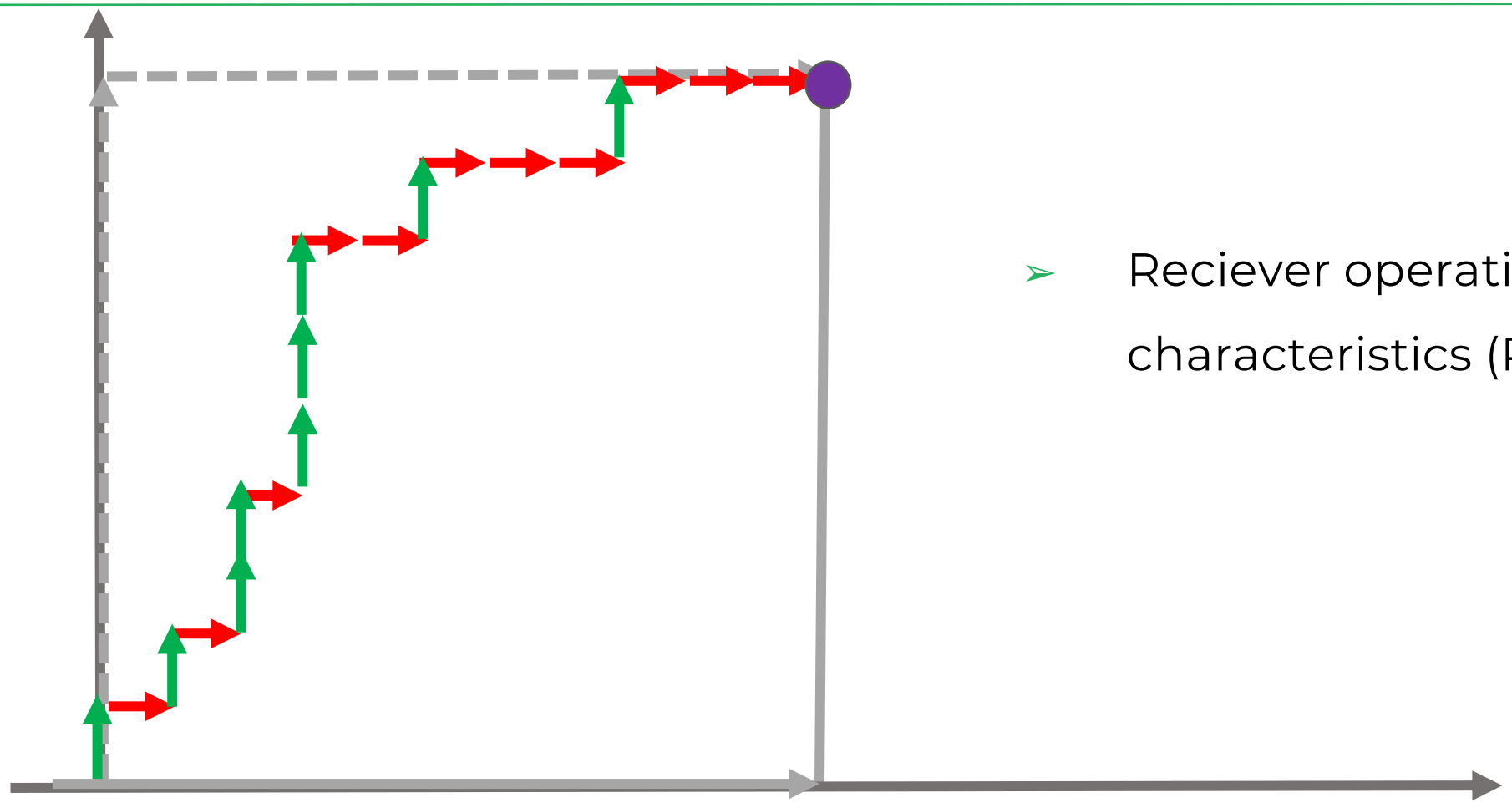
Osztályozás - Gains



Osztályozás - Gains



Osztályozás – ROC

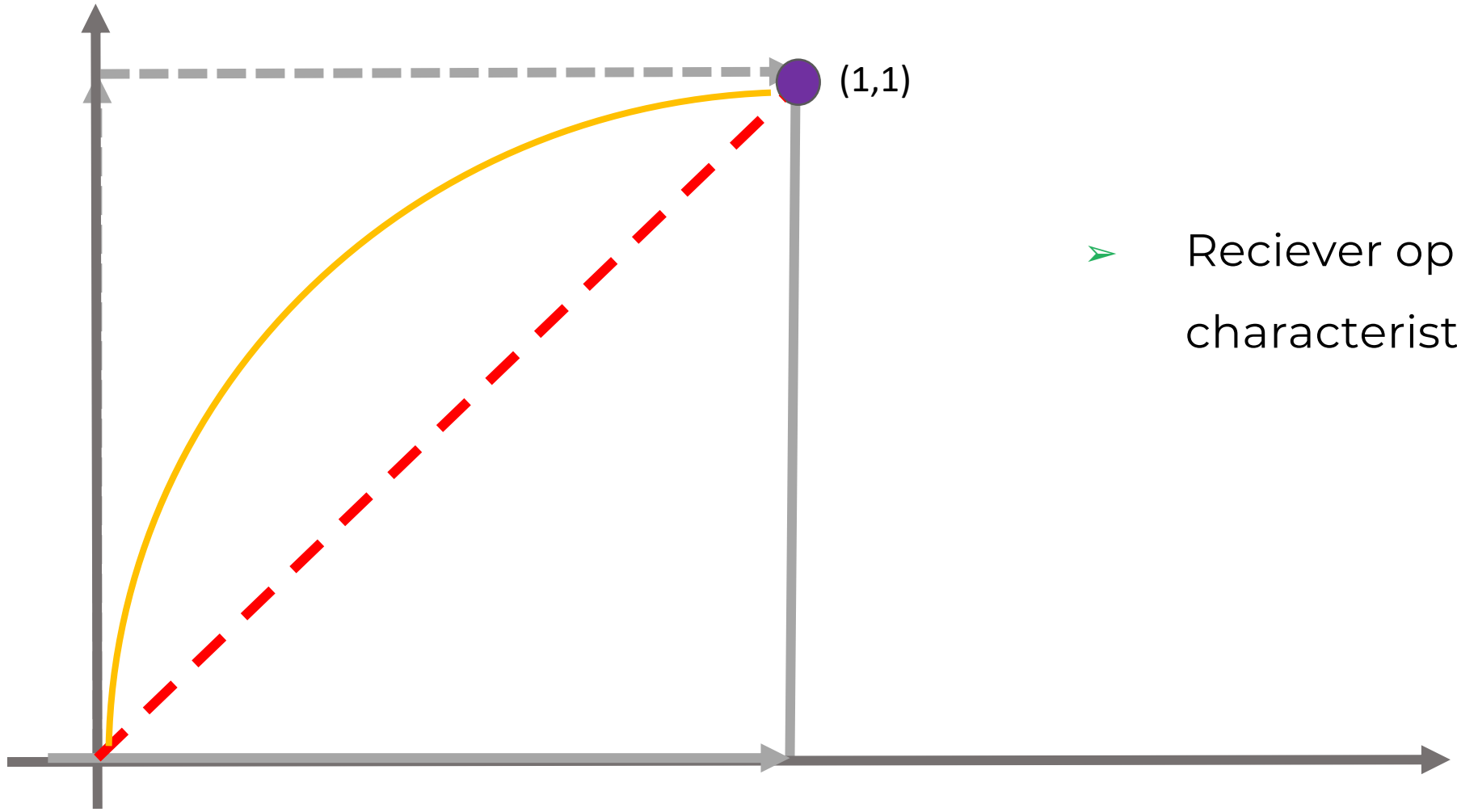


- Reciever operating characteristics (ROC)

P valószínűség kicsi

P valószínűség nagy

Osztályozás – ROC

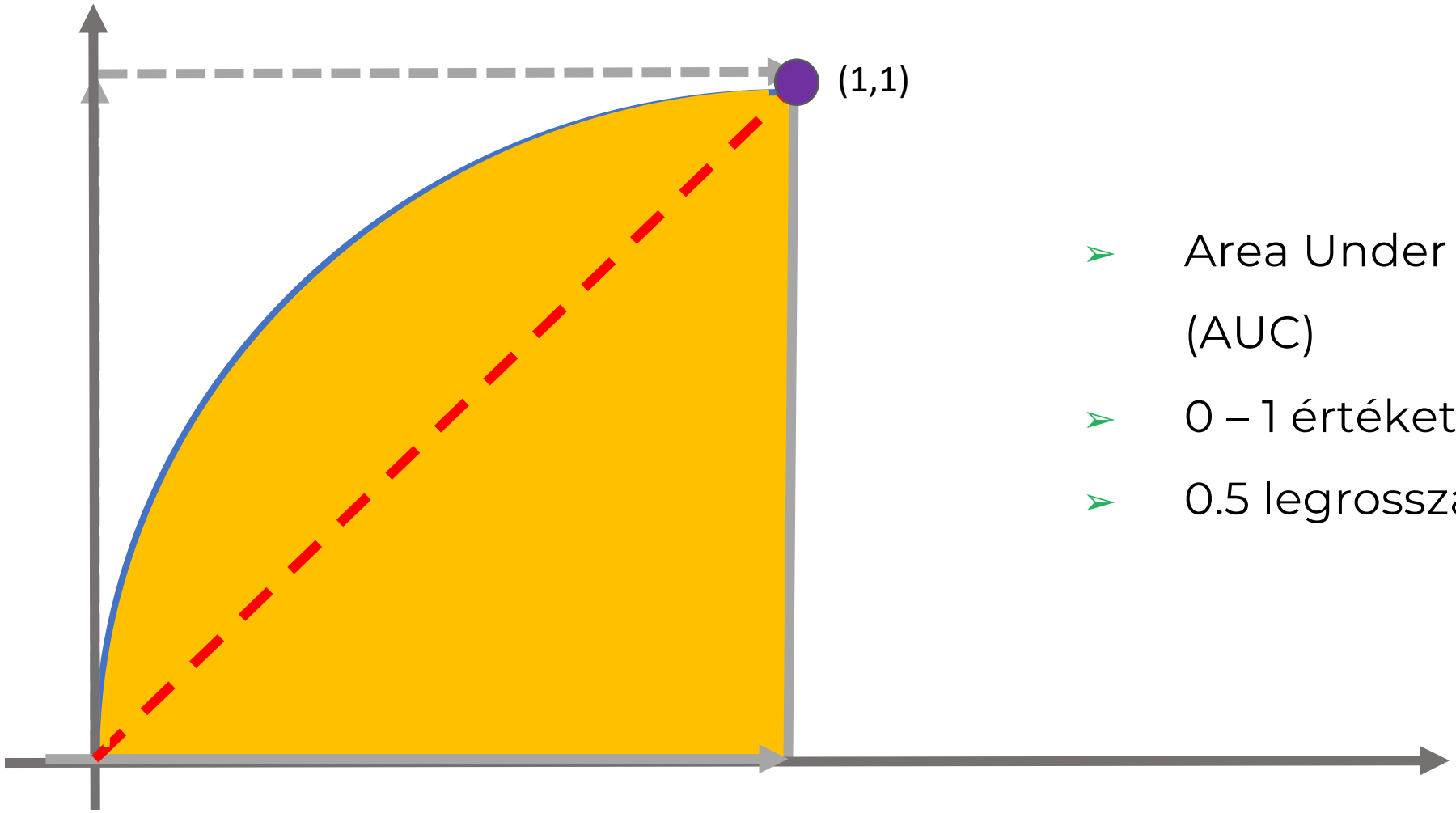


- Reciever operating characteristics (ROC)

P valószínűség kicsi

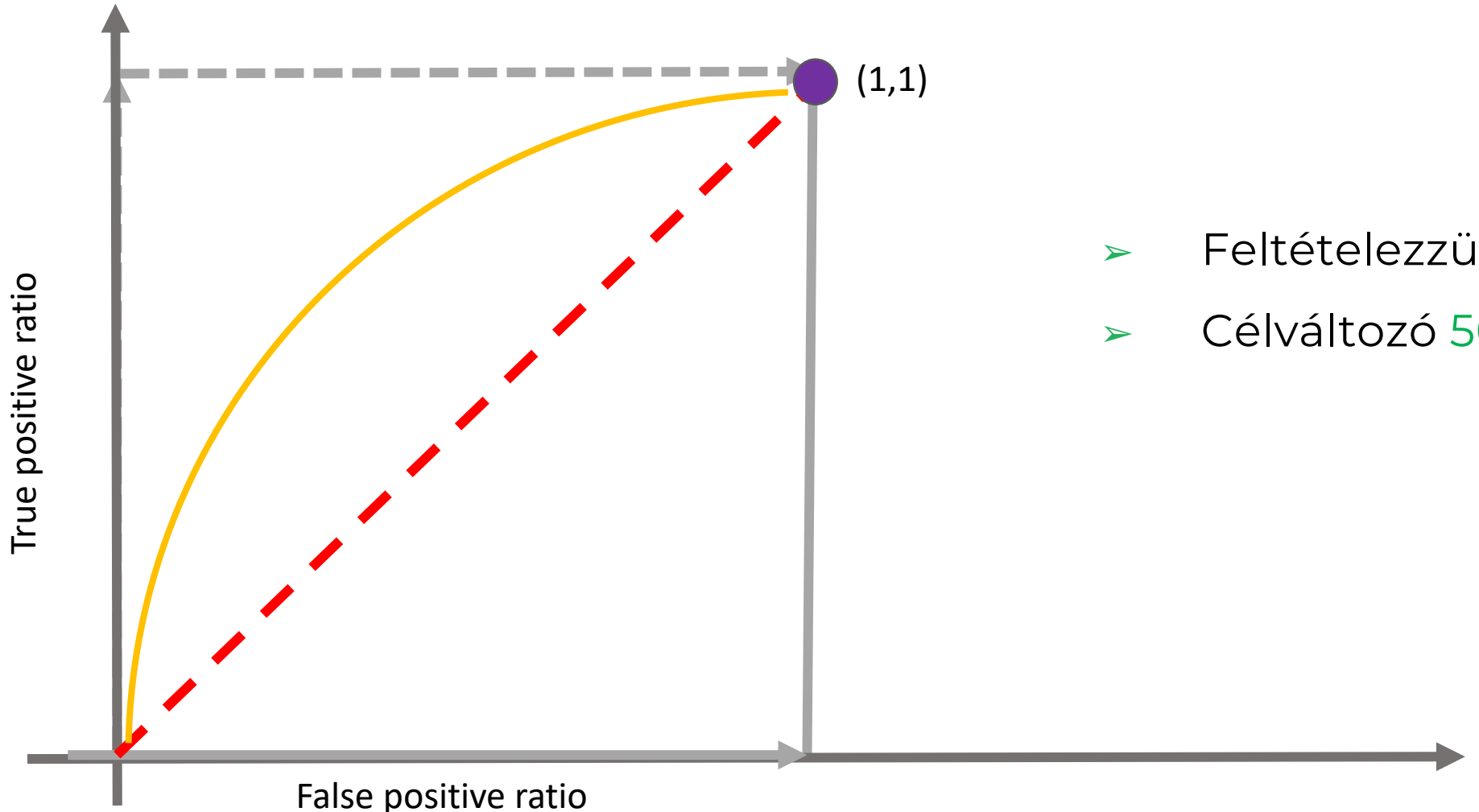
P valószínűség nagy

Osztályozás – AUC



- Area Under the Curve (AUC)
- 0 – 1 értéket vehet fel
- 0.5 legrosszabb

Osztályozás – ROC

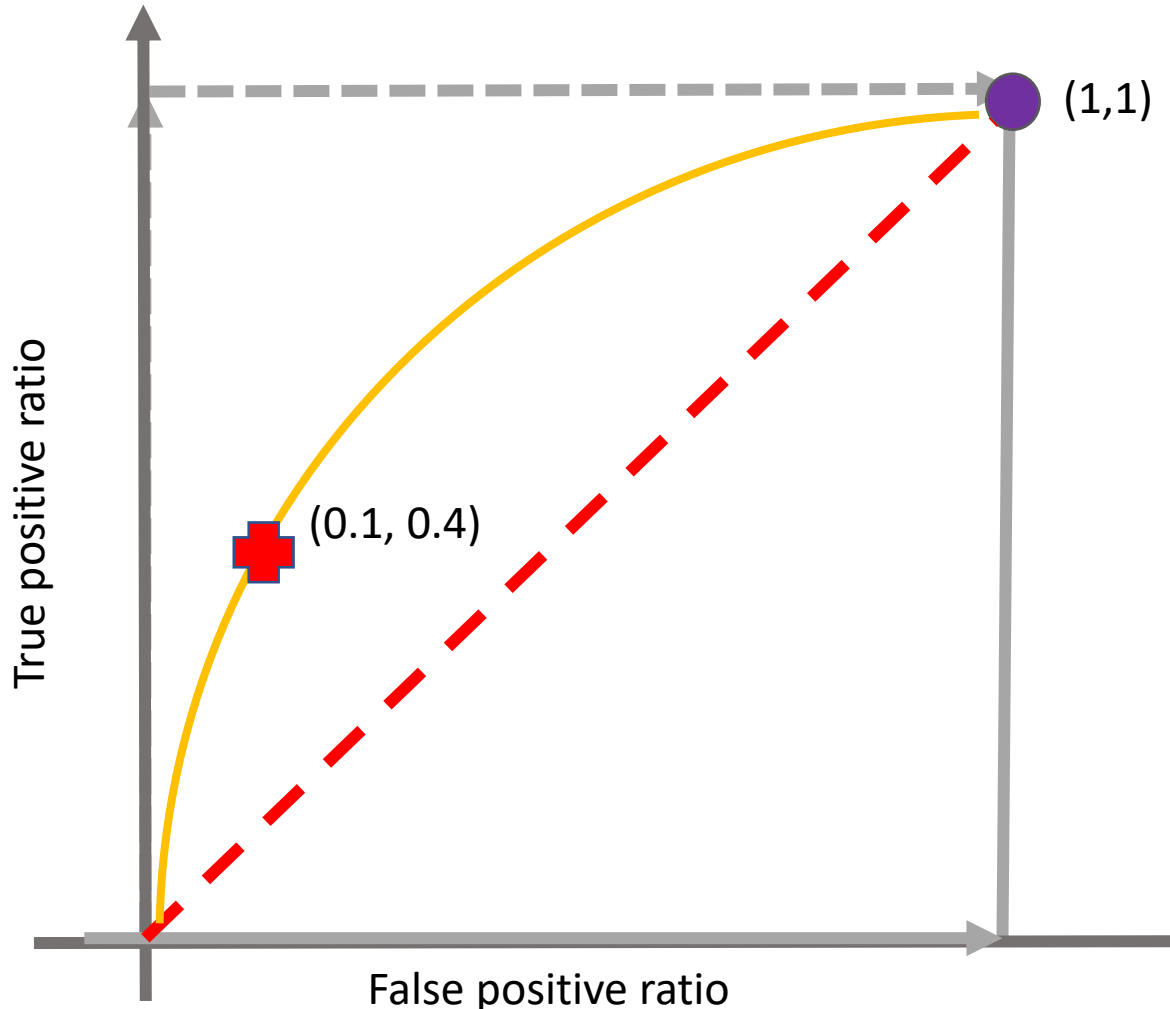


- Feltételezzünk 100 embert
- Célváltozó 50 - 50

P valószínűség kicsi

P valószínűség nagy

Osztályozás – ROC

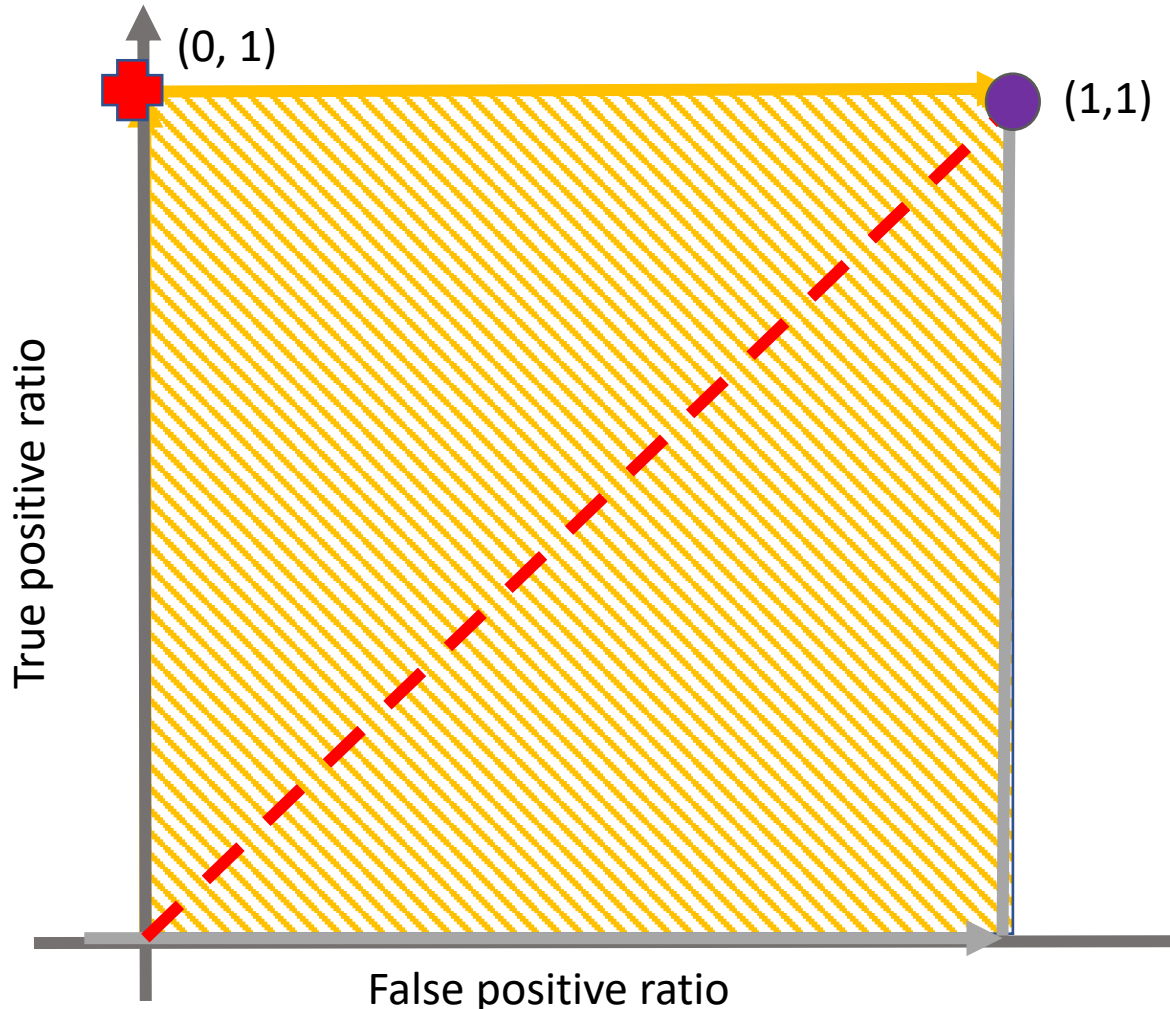


- Az adott ponton az összes pozitív célváltozó 40%t találtuk meg (50 * 0.4 = 20)
- És az összes negatív célváltozó 10%-a került bele a sor elejébe (50 * 0.1 = 5)

P valószínűség kicsi

P valószínűség nagy

Osztályozás – ROC

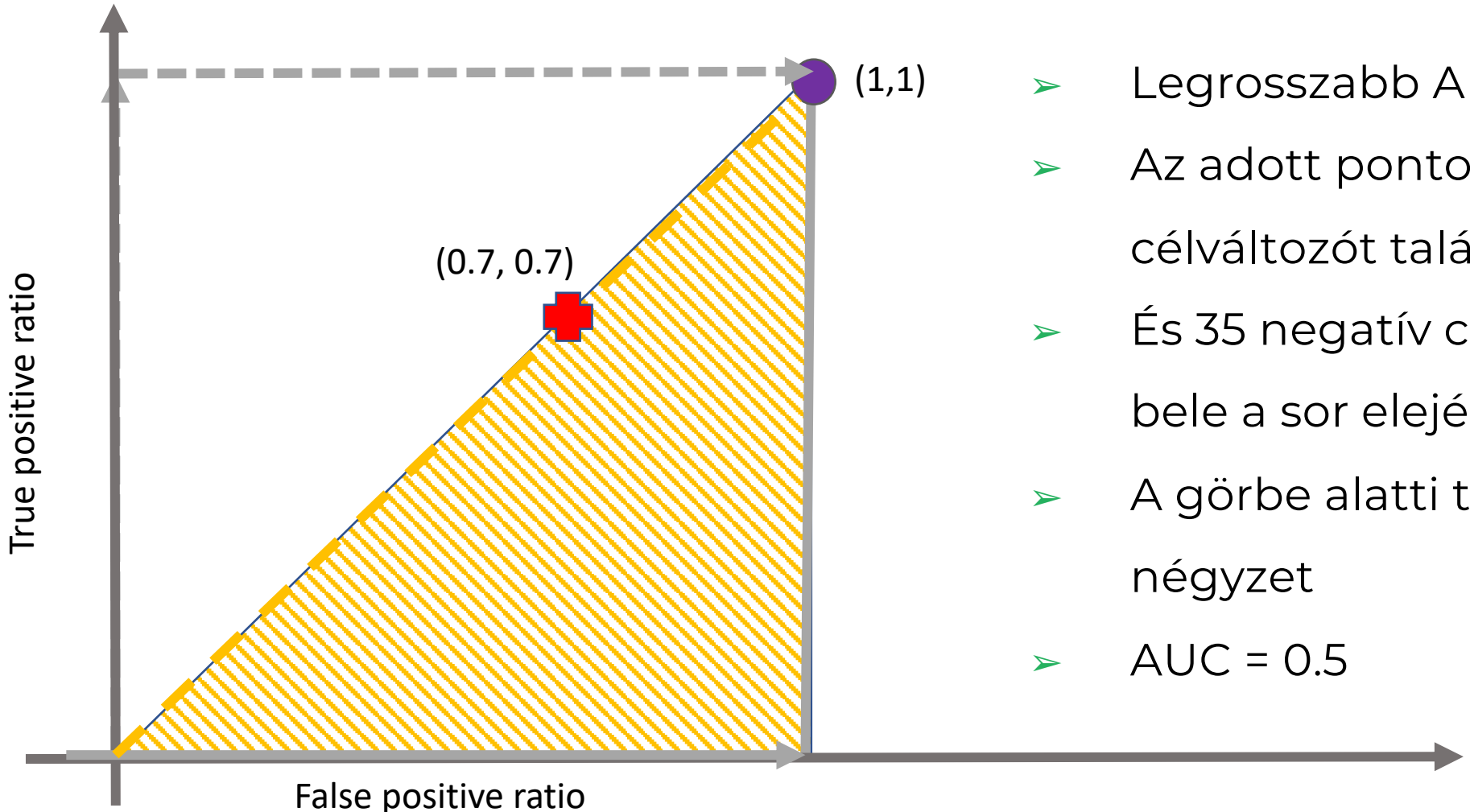


- Hibátlan AUC
- Az adott ponton az összes pozitív célváltozót megtaláltuk(50 * 1)
- És 0 negatív célváltozó került bele a sor elejébe (50 * 0)
- A görbe alatti terület a teljes 1 * 1-es négyzet
- AUC = 1

P valószínűség kicsi

P valószínűség nagy

Osztályozás – ROC



- Legrosszabb AUC
- Az adott ponton 35 pozitív célváltozót találtuk(50 * 0.7 = 35)
- És 35 negatív célváltozó került bele a sor elejébe (50 * 0.7 = 35)
- A görbe alatti terület a fél 1 * 1-es négyzet
- AUC = 0.5

P valószínűség kicsi

P valószínűség nagy