



Az interpretálhatóság szerepe a gépi tanulásban

Balogh Nóra

Budapest ML Fórum

2022

Rólam



Balogh Nóra

Data Scientistként adatos projektek megvalósításán dolgozom a **Tata Consultancy Services Hungary** csapatában

Szeretek tanítani és data science blogposztokat írni (<https://noratenk.medium.com/>)

Áttekintés

Lehetőségek és
kihívások

Módszerek

Motiváció

Interpretálhatóság
fogalma

Igények az
interpretálhatóságra

Milyen a jó
magyarázat?

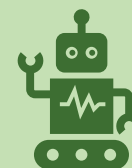
A gépi tanulás elterjedése



FELÜGYELT



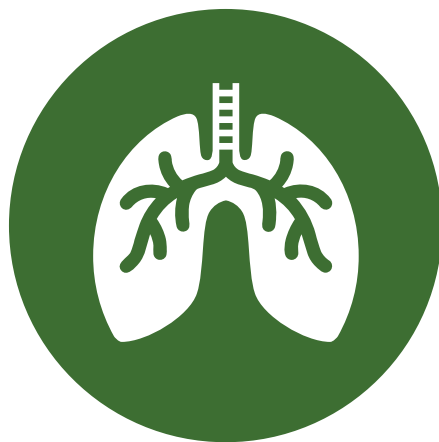
FELÜGYELET
NÉLKÜLI



MEGERŐSÍTÉSES



Bankszektor



Egészségügy

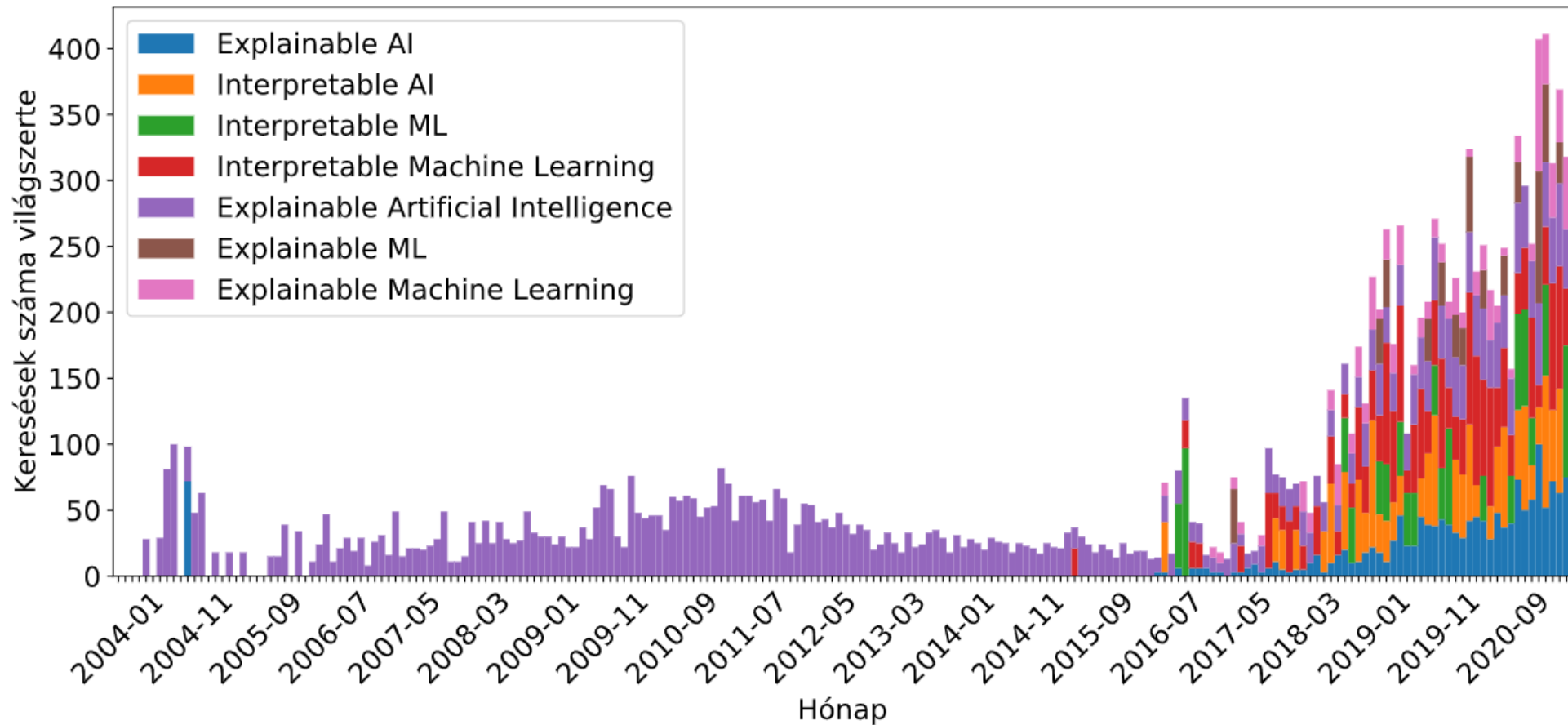


Logisztika

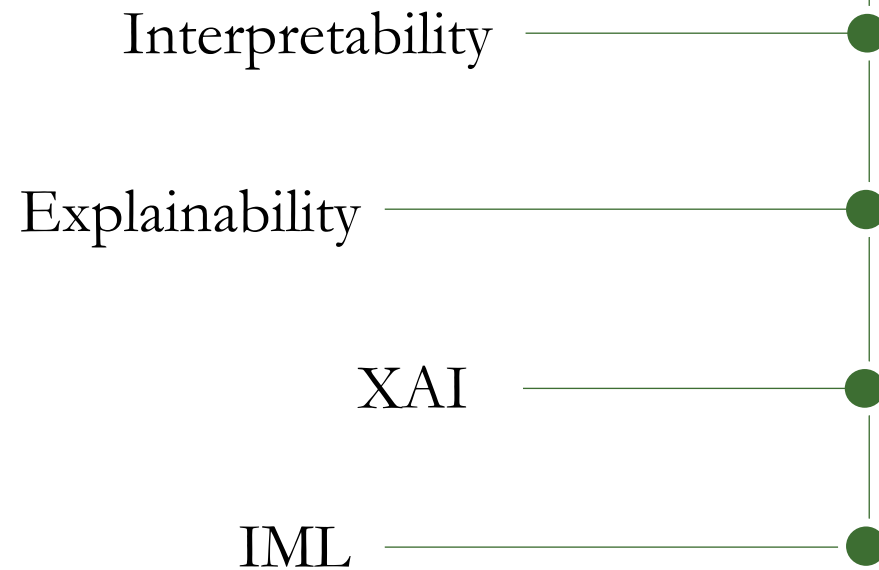


Igazságszolgáltatás

Növekvő igények az interpretálhatóságra



Interpretálhatóság fogalma



- ❖ Doshi-Velez et. al, „Towards A Rigorous Science of Interpretable Machine Learning”, arXiv, 2017.
- ❖ C. Molnar, Interpretable Machine Learning, e-book, 2019.

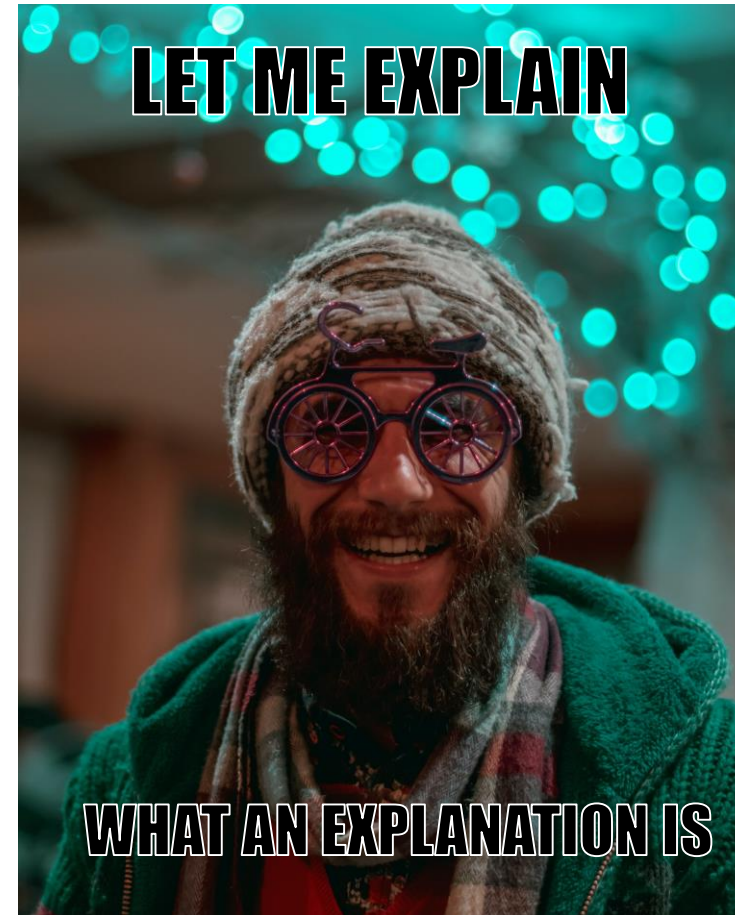


Photo by [Armin Lotfi](#) on [Unsplash](#)

Igények az XAI-ra



Bizalom
növelése



Etikus
döntéshozás



Mobilitás



Kauzalitás



Privacy



Informativitás

❖ Z. C. Lipton, „The Mythos of Model Interpretability,” in CML Workshop on Human Interpretability in Machine Learning (WHI 2016), New York, NY, 2016.

❖ A. B. Arrieta et. al, „Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” Information Fusion, 1.58, pp. 82-115, 2020.

❖ Doshi-Velez et. al, „Towards A Rigorous Science of Interpretable Machine Learning”, arXiv, 2017.

Milyen a jó magyarázat?

- Mennyi időbe telik megérteni?
- Kinek szól?
- Szükség van rá?

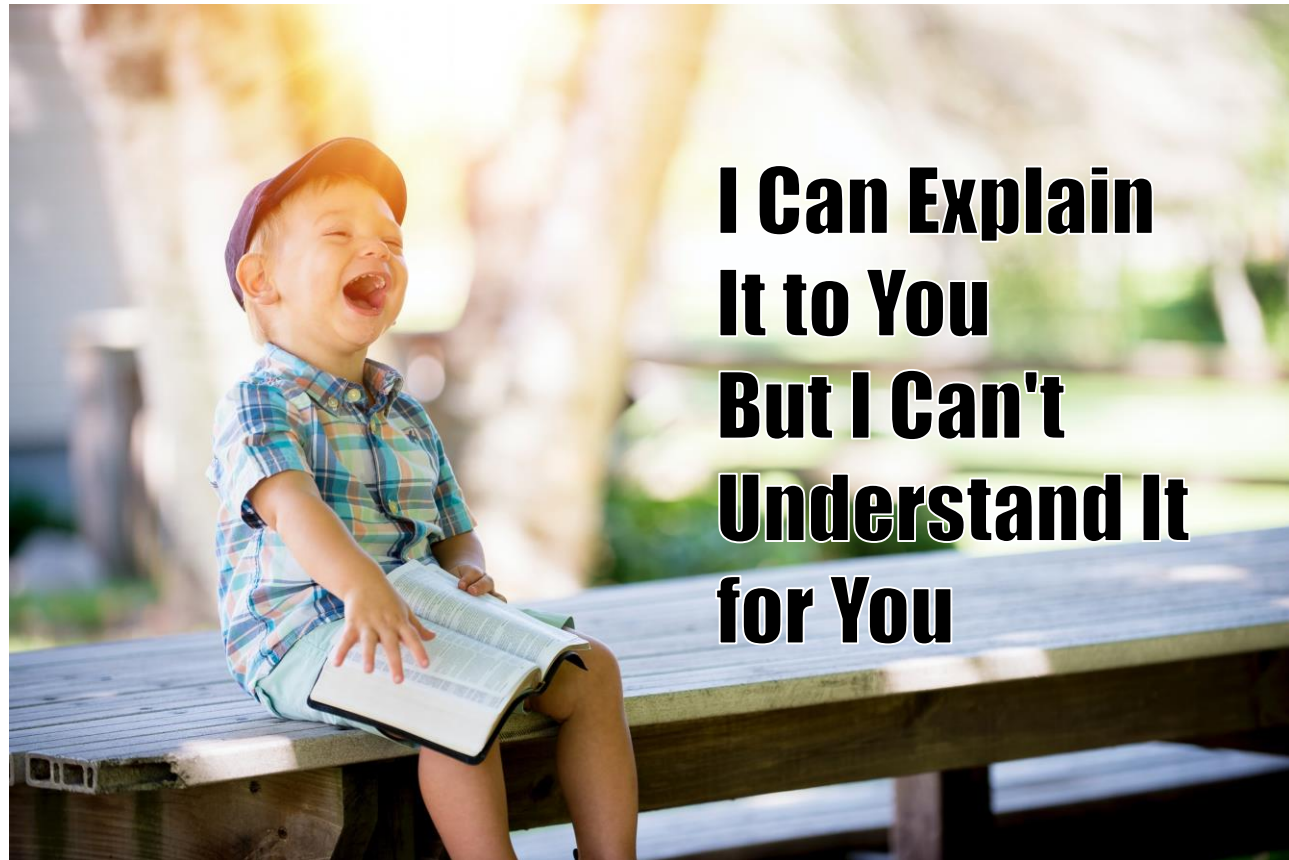
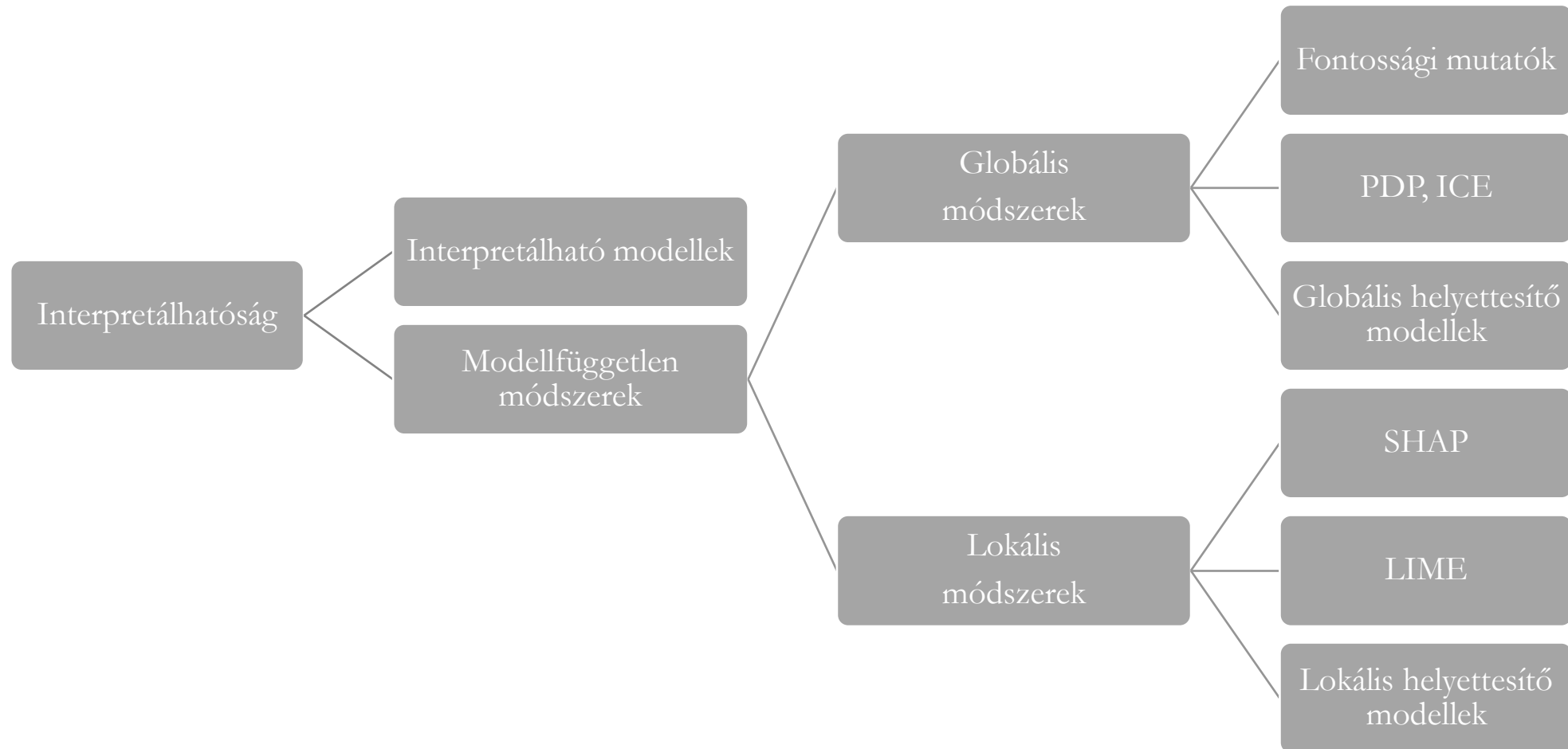


Photo by [Ben White](#) on [Unsplash](#)

Interpretálhatósági módszerek

Felügyelt tanulás esetén



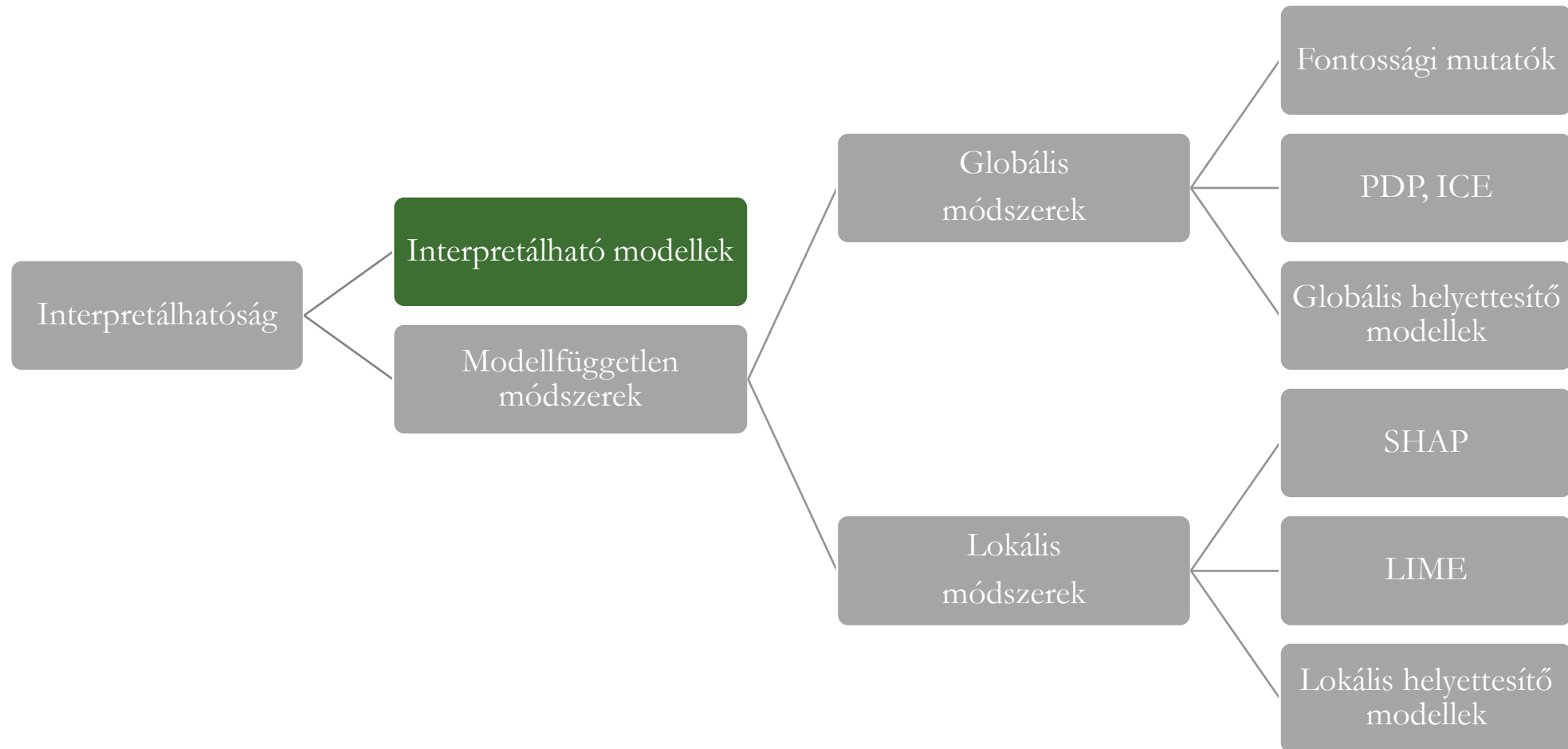


Ingatlan értékbecslés

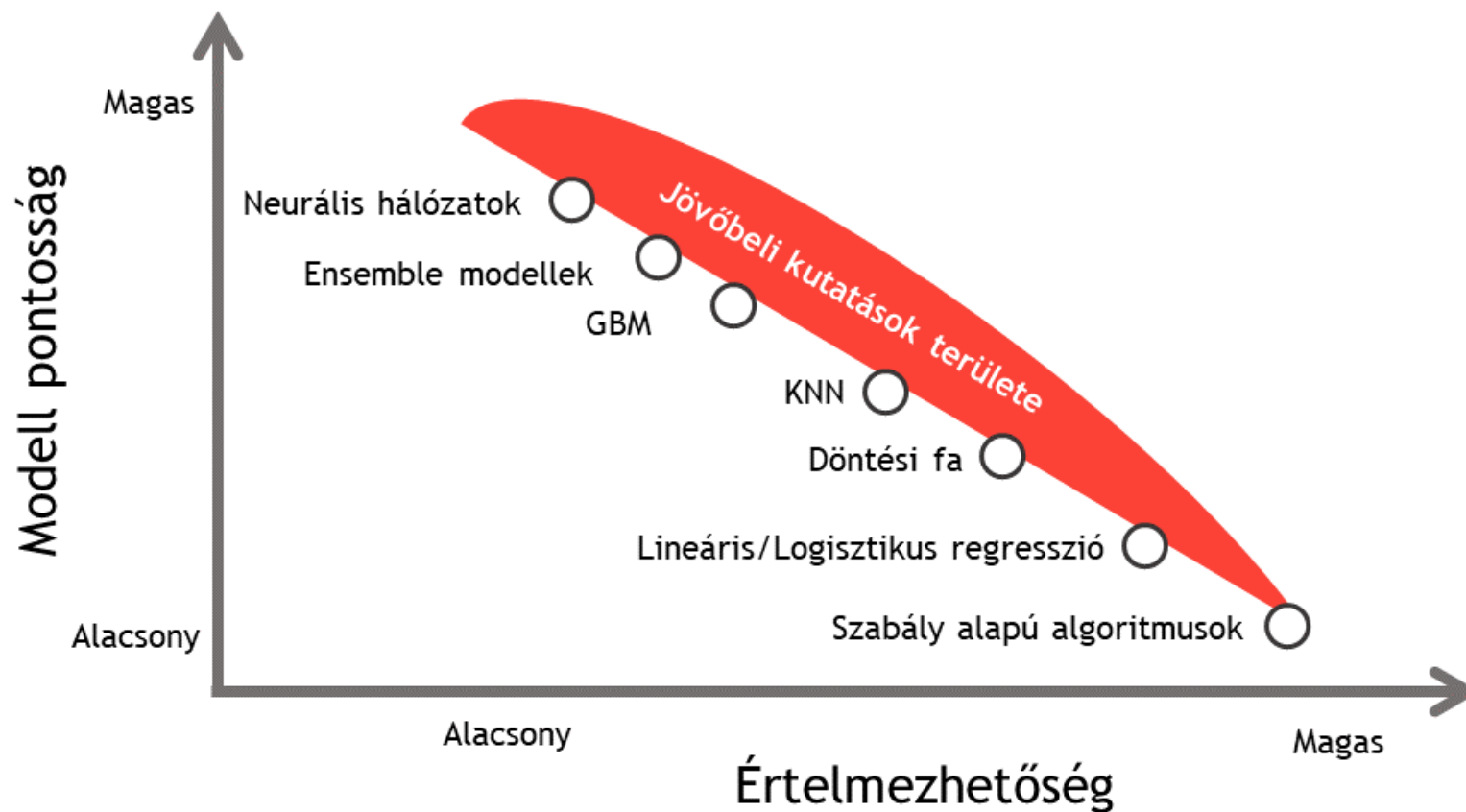
[https://www.kaggle.com/competitions/
me-ingatlan-com/data](https://www.kaggle.com/competitions/me-ingatlan-com/data)

Interpretálhatósági módszerek

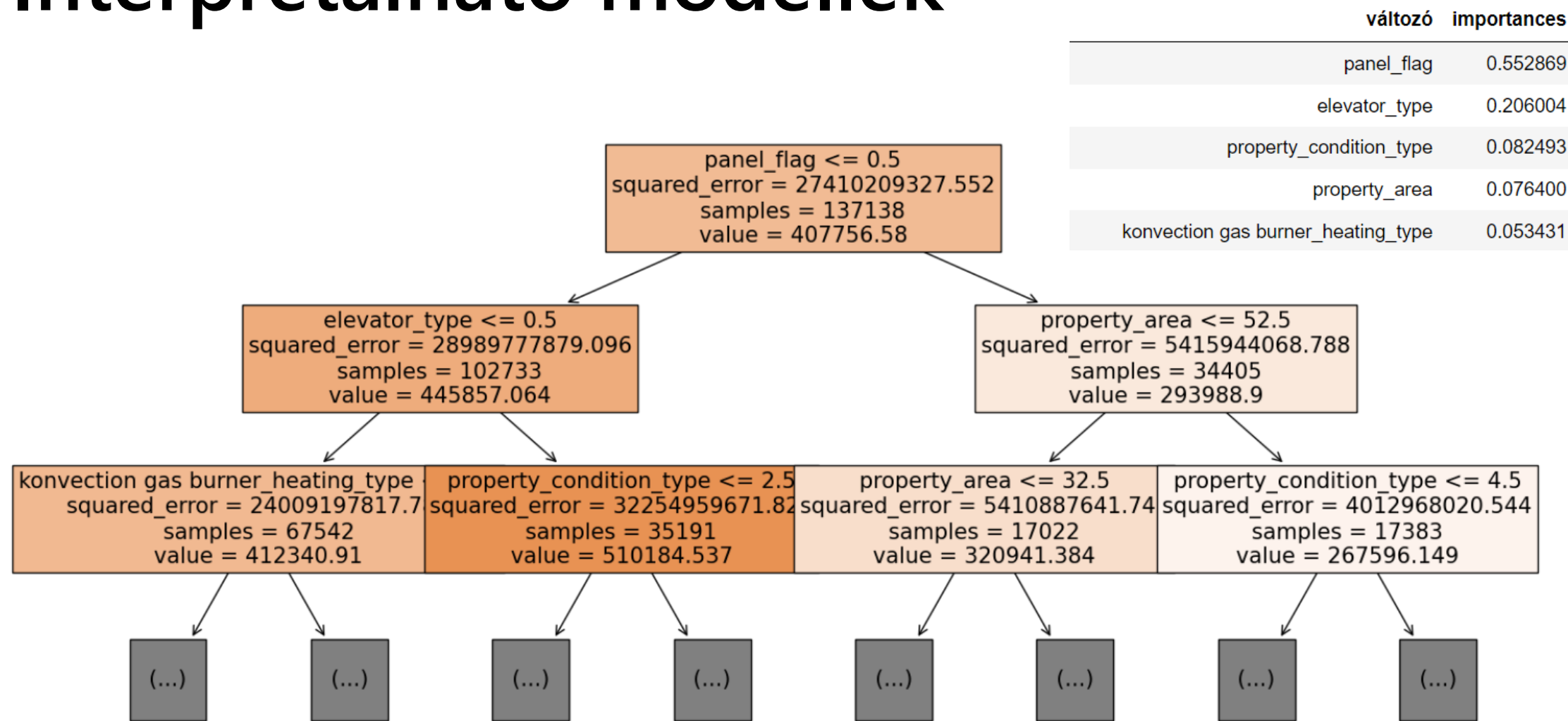
Felügyelt tanulás esetén



Interpretálható modellek

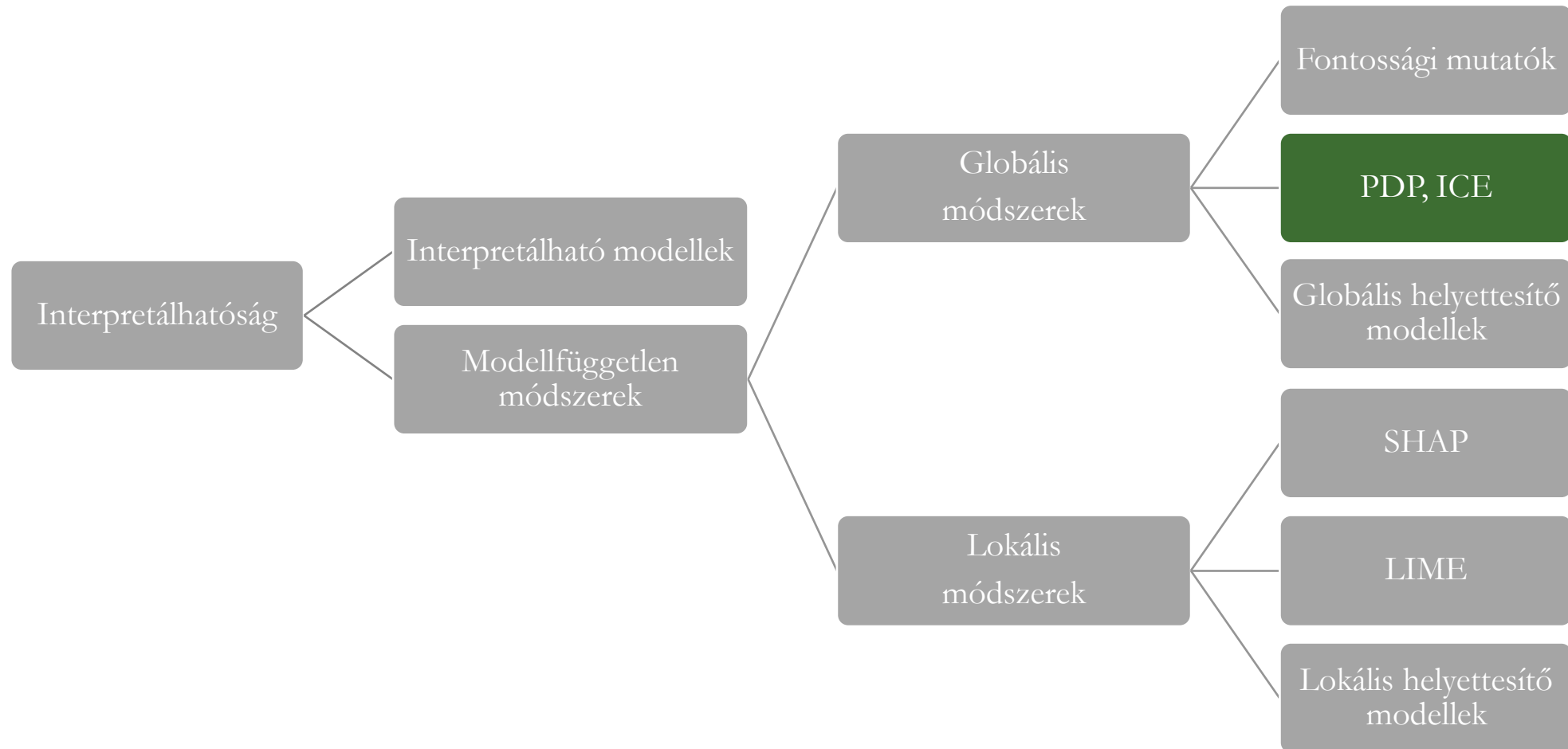


Interpretálható modellek



Interpretálhatósági módszerek

Felügyelt tanulás esetén



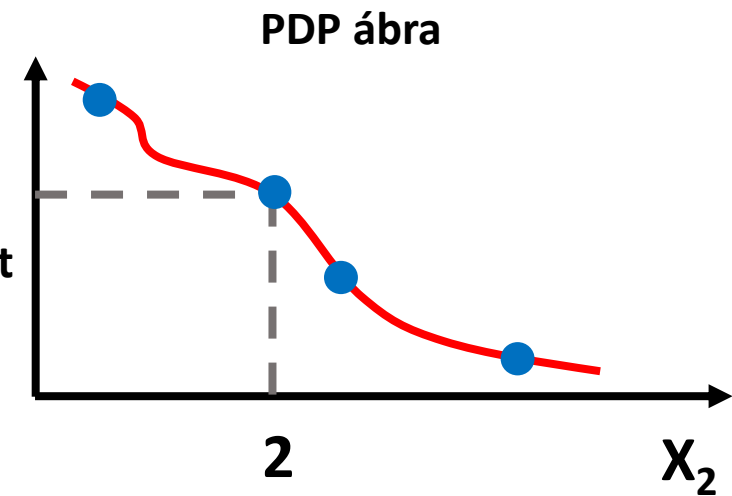
Partial Dependence Plot (PDP)

X1	X2	X3	...	Y
	1			1
	2			0
	2			1
	3			1
	2			0

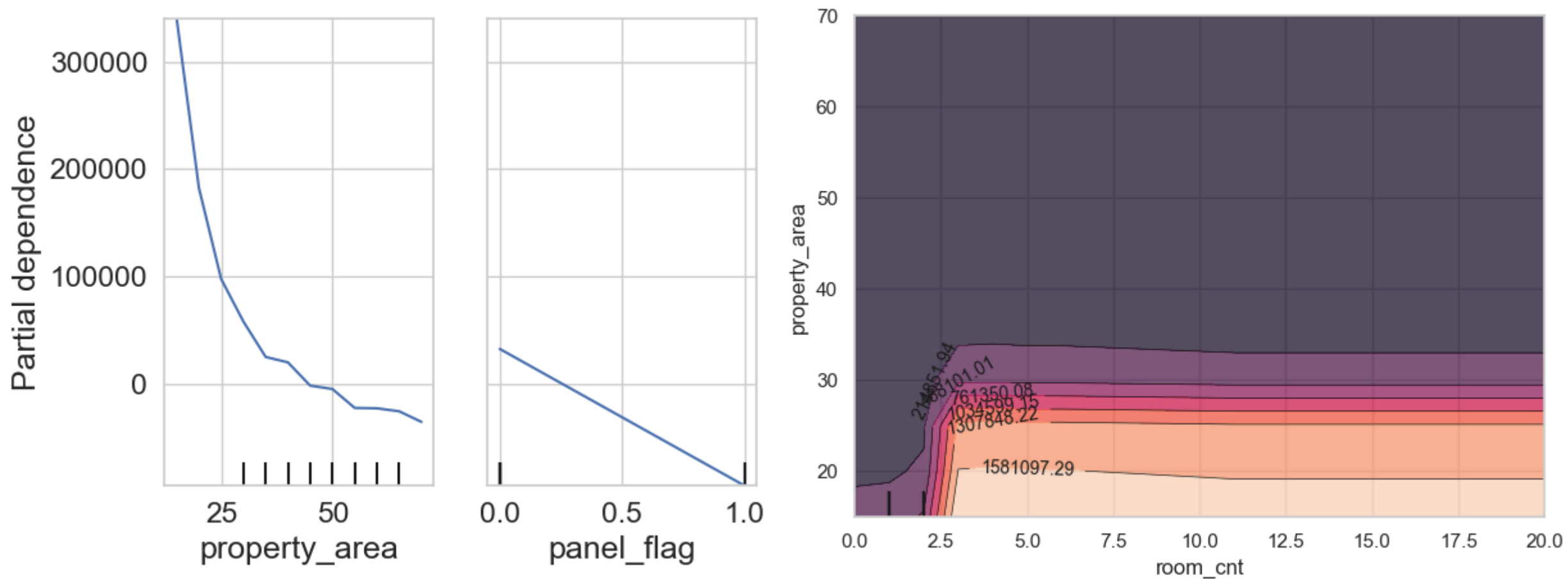


X1	X2	X3	...	Modell predikció
	2			0.98
	2			0.1
	2			0.6
	2			0.5
	2			0.01

Átlagos
modell
kimenet



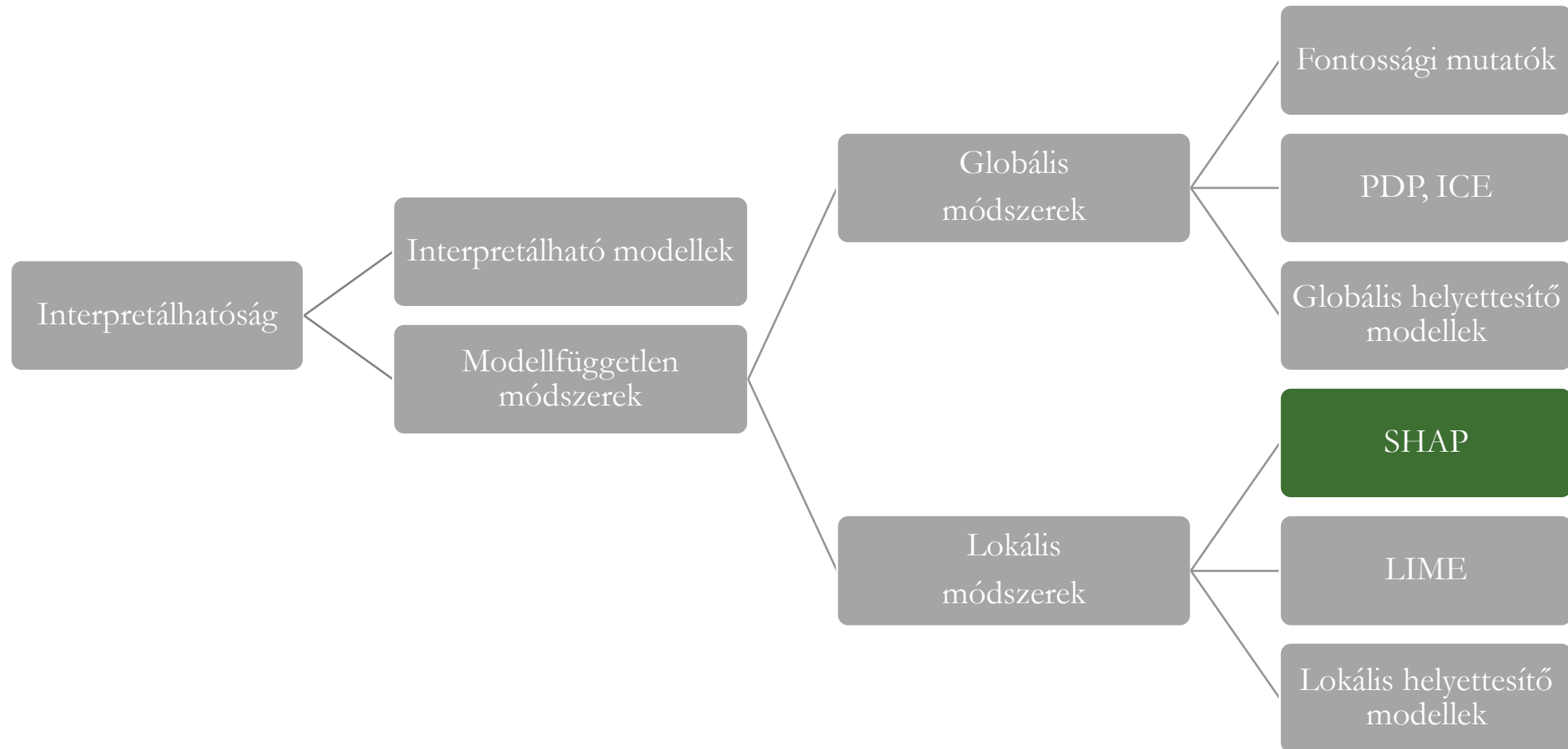
Partial Dependence Plot (PDP)



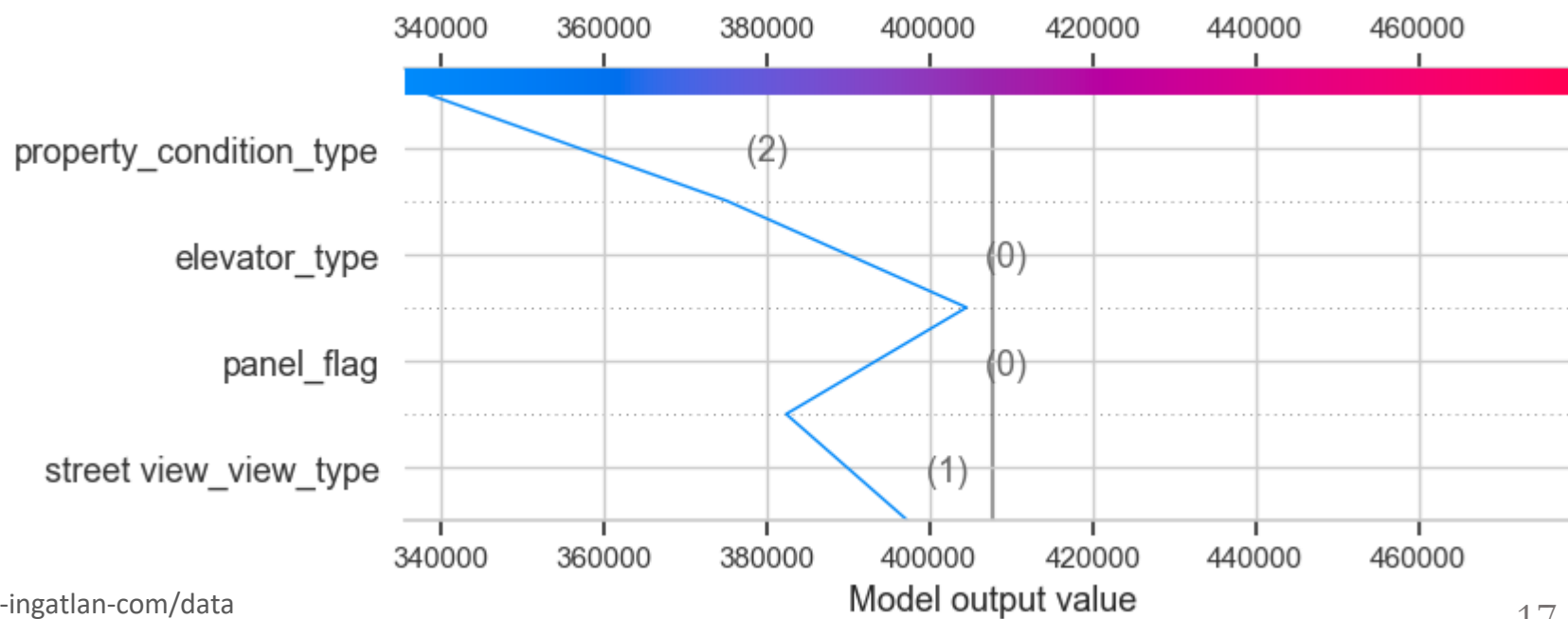
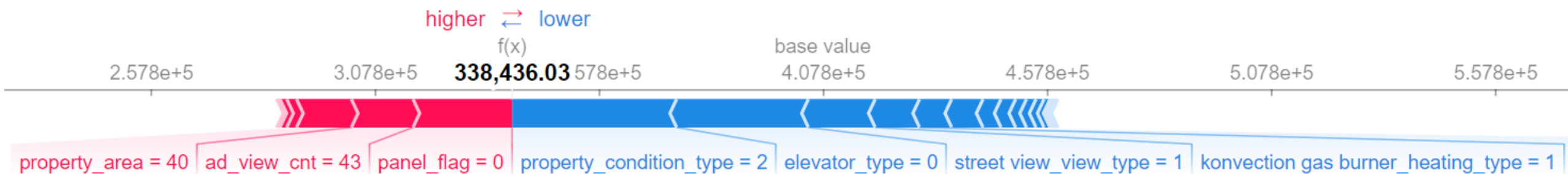
- ❖ <https://www.kaggle.com/competitions/me-ingatlan-com/data>
- ❖ PDPBOX python library, SauceCat, 2018.
- ❖ sklearn python library

Interpretálhatósági módszerek

Felügyelt tanulás esetén



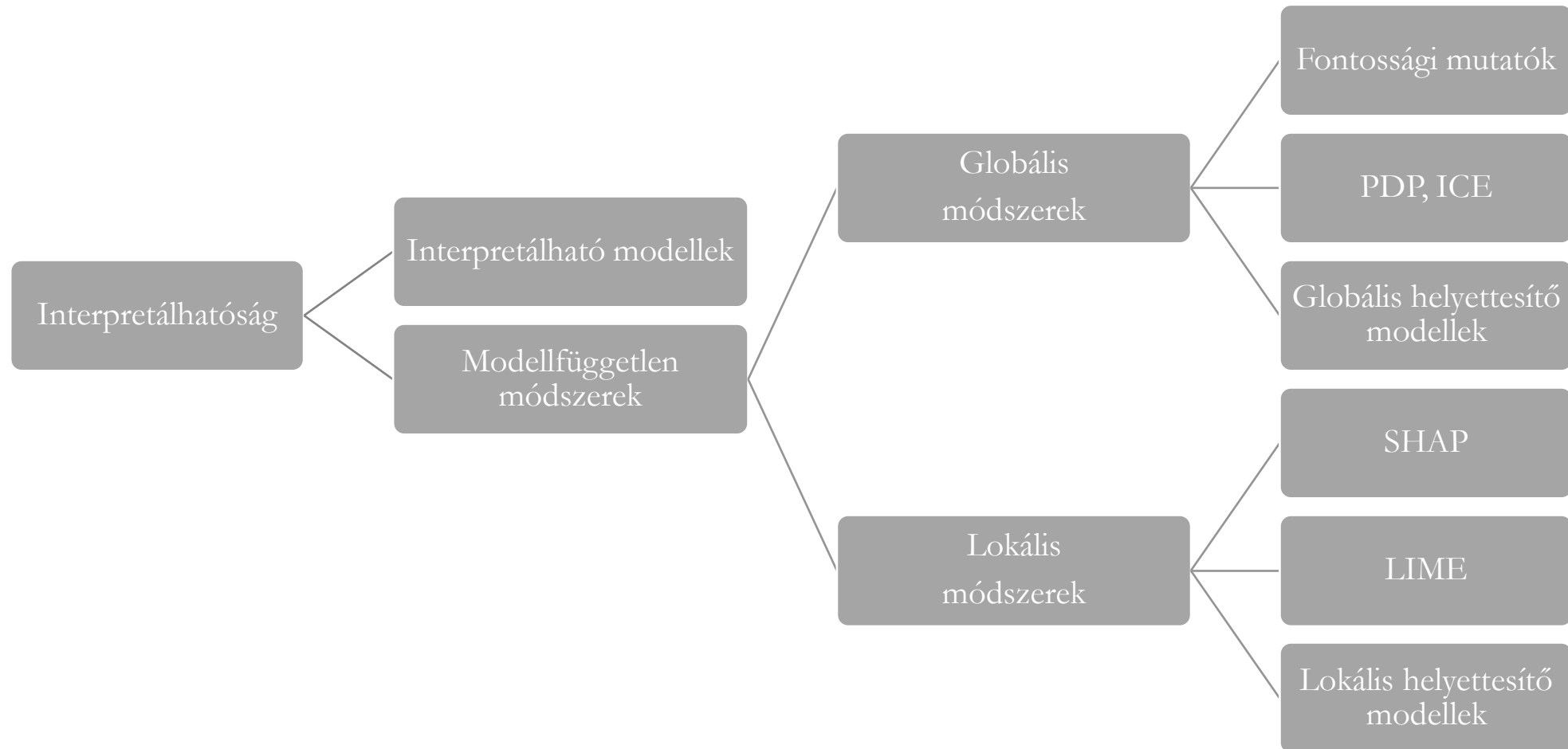
SHAP



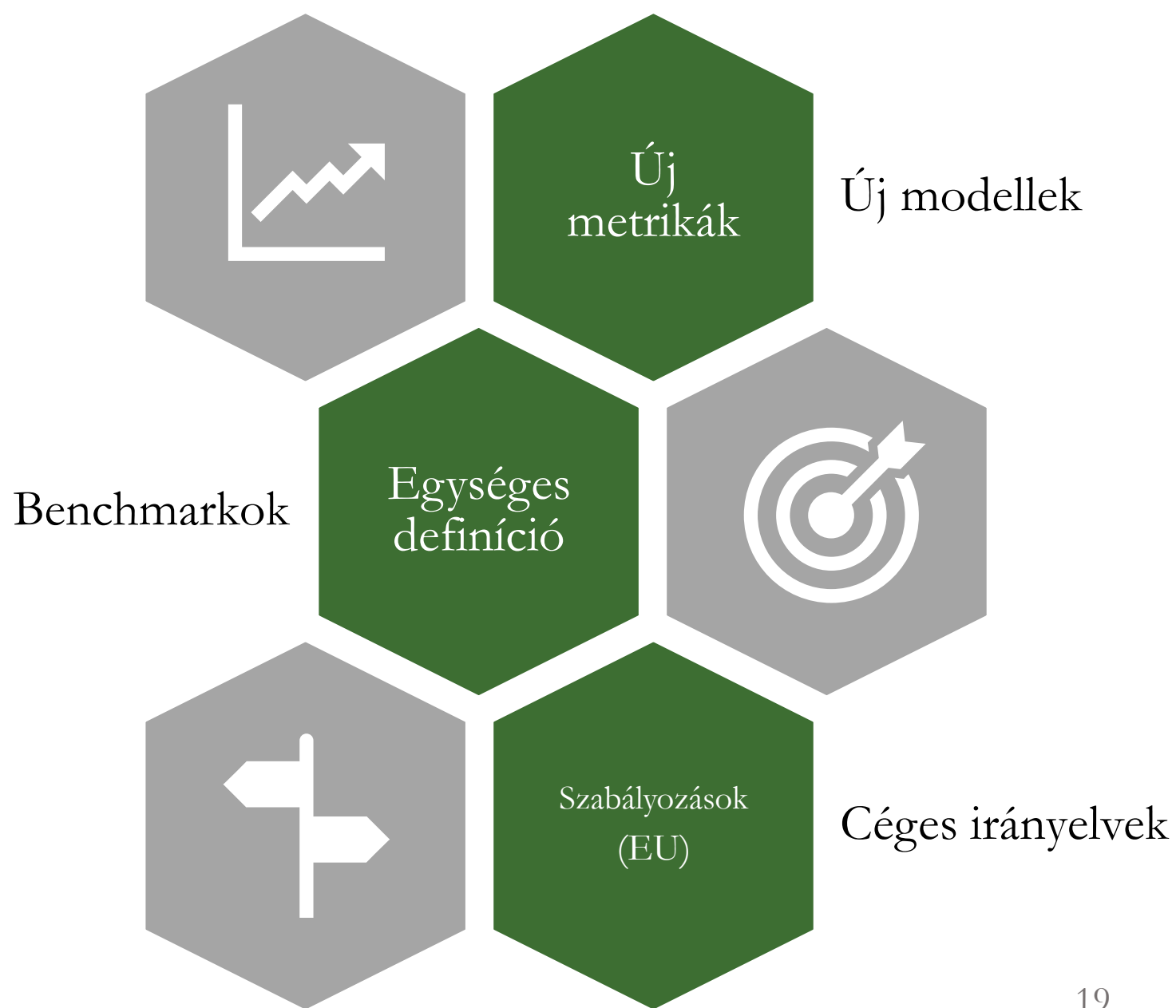
- ❖ <https://www.kaggle.com/competitions/me-ingatlan-com/data>
- ❖ SHAP python library, Scott Lundberg, 2018.

Interpretálhatósági módszerek

Felügyelt tanulás esetén



Lehetőségek és kihívások





Köszönöm a figyelmet!

in/noratenk 

noratenk.medium.com 