Occam borotvája vagy gépi tanulás?

2024. III. 7-8., Szeged

ML modellek interpretációja

Tóth Gergely

ELTE Kémiai Intézet

Kóstolgatom még csak a területet - köszönet Turán Györgynek az „interpretáció” kulcsszóért

Ezért itt most két könyv alapján:

Ajay Thampi: Interpretable AI, Manning, Shelter Island, 2022

Christoph Molnar: Interpretable Machine Learning, https://christophm.github.io/interpretable-ml-book/, 2019

bennük:

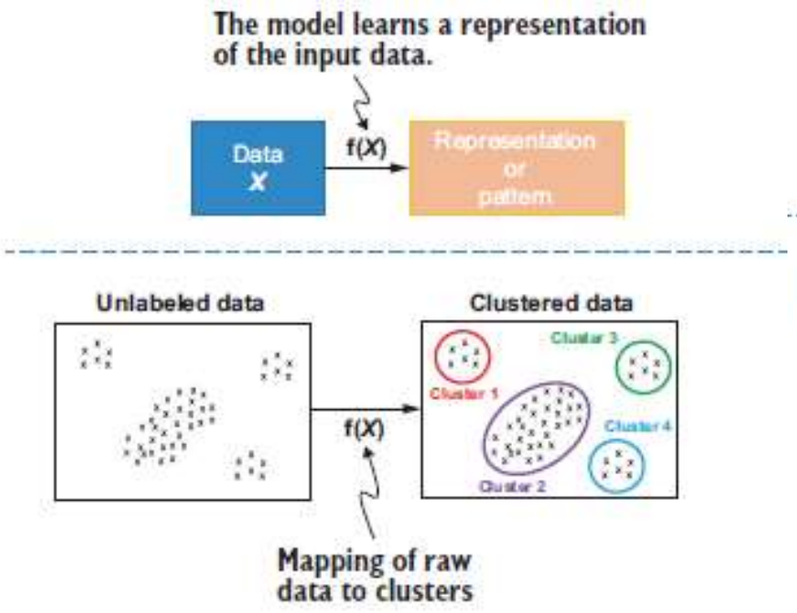
AT: Python kódok

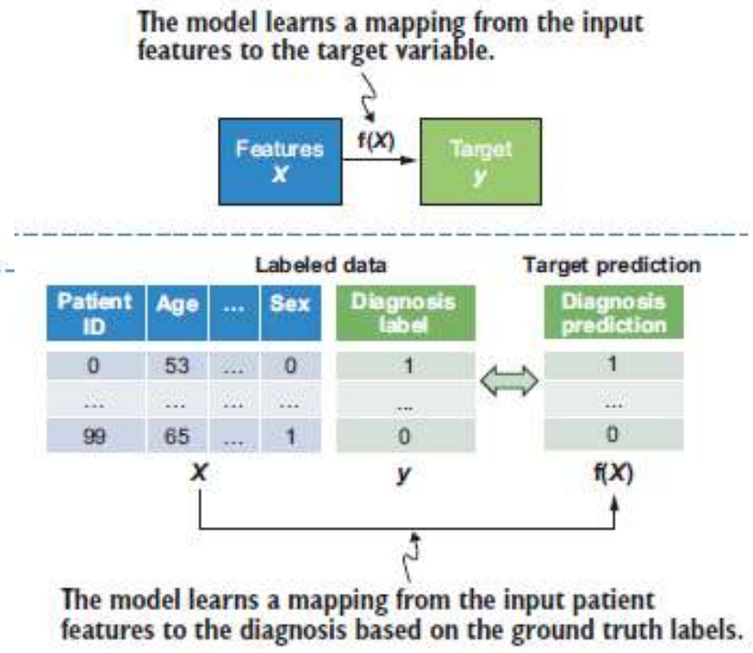
CM: R kódok

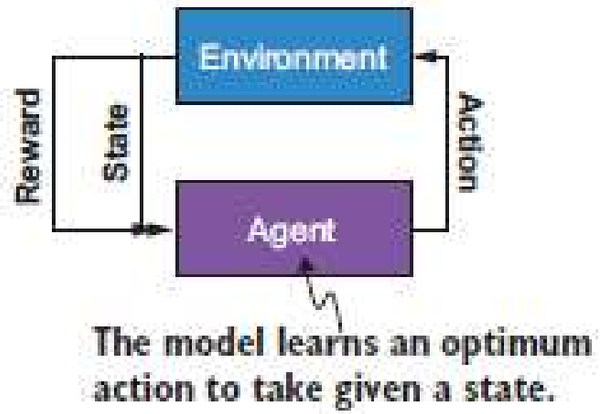
(ábrák: tőlük, ha külön nem jelölöm)

Gépi tanulás (machine learning)

nem felügyelt	felügyelt	megerősítéses
unsupervised	supervised	reinforcement

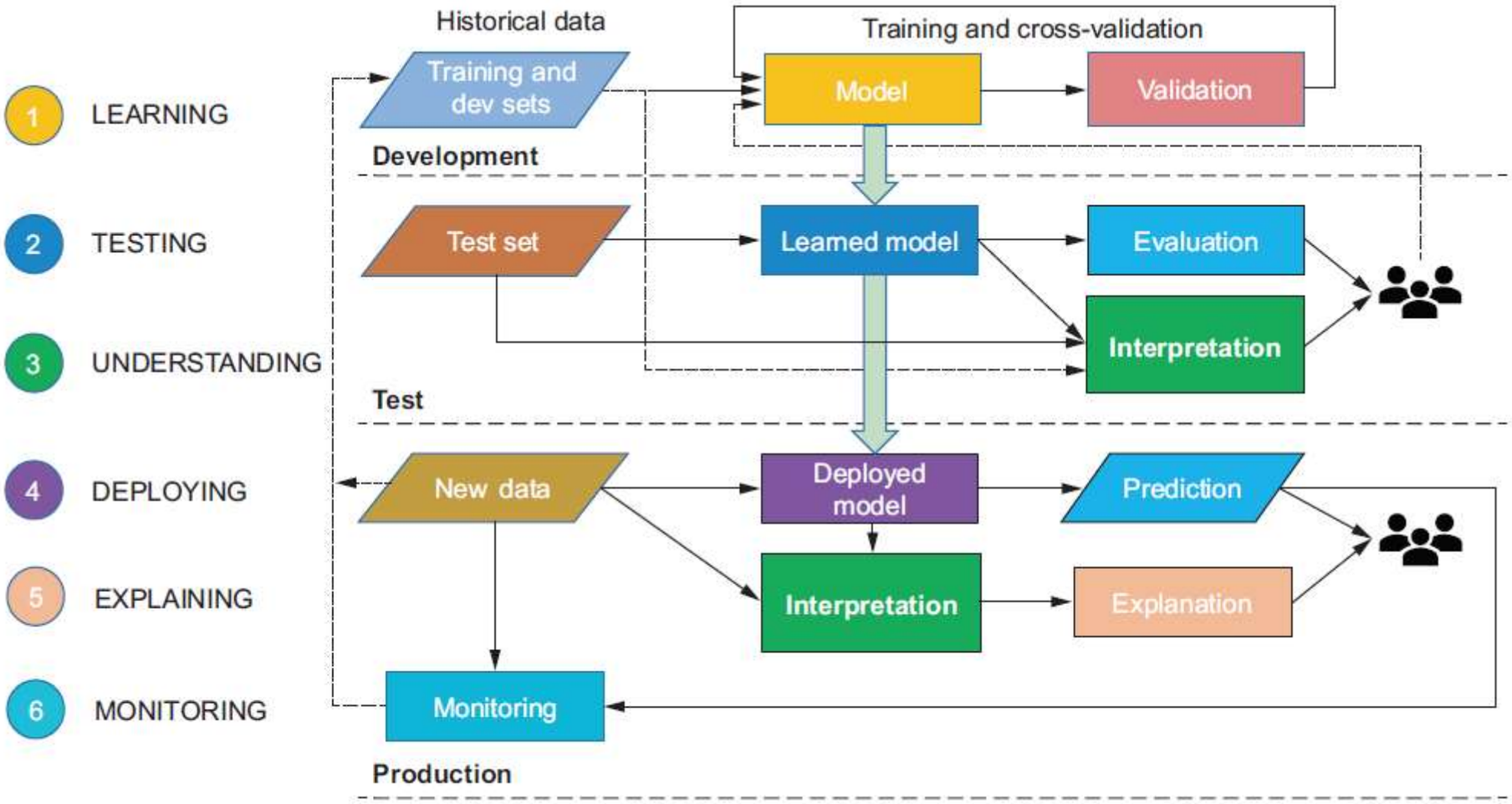






csoportosítás/klaszterezés	osztályozás/klasszifikáció, regresszió	bármi + robotika

Folyamatábra (A. Thampi)



The process of building a robust AI system

Értelmezés (interpretation)

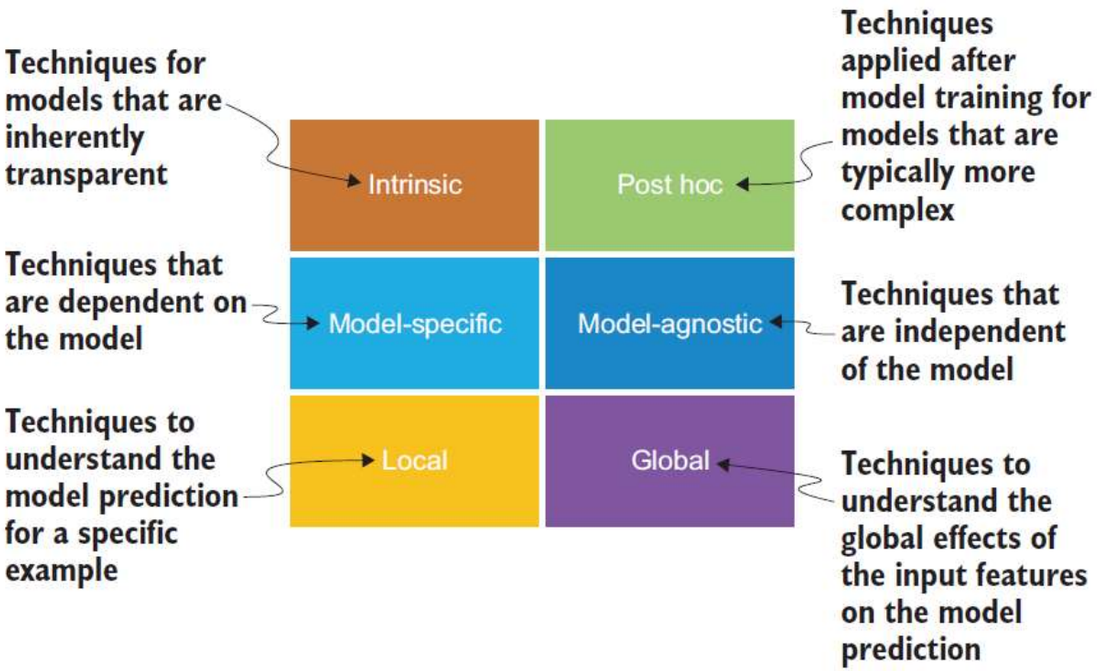


Figure 1.14 Types of interpretability techniques

Magyarázó változók (features)

- statisztikai vizsgálata a modellben (feature summary statistics)
- vizuális bemutatása (hatás, összefüggések…) (feature summary visualization)

Adatok és adatpontok interpretációja

Helyettesítő modellek (surrogate models)

Magyarázat (explanation)

„Human readable form”

Magyarázat = Válasz egy *miért* kérdésre (Miller 2017)

Miért nem hatott a kezelés a betegre?
Miért büntettek meg gyorshajtásért?
Miért hat ez a gyógyszer ezen az enzimen?
Miért csökkentették a béremet?

GDPR 22§ 1) Az érintett jogosult arra, hogy ne terjedjen ki rá az olyan, kizárólag automatizált adatkezelésen – ideértve a profilalkotást is – alapuló döntés hatálya, amely rá nézve joghatással járna vagy őt hasonlóképpen jelentős mértékben érintené.

Ajay Thampi: Interpretable AI

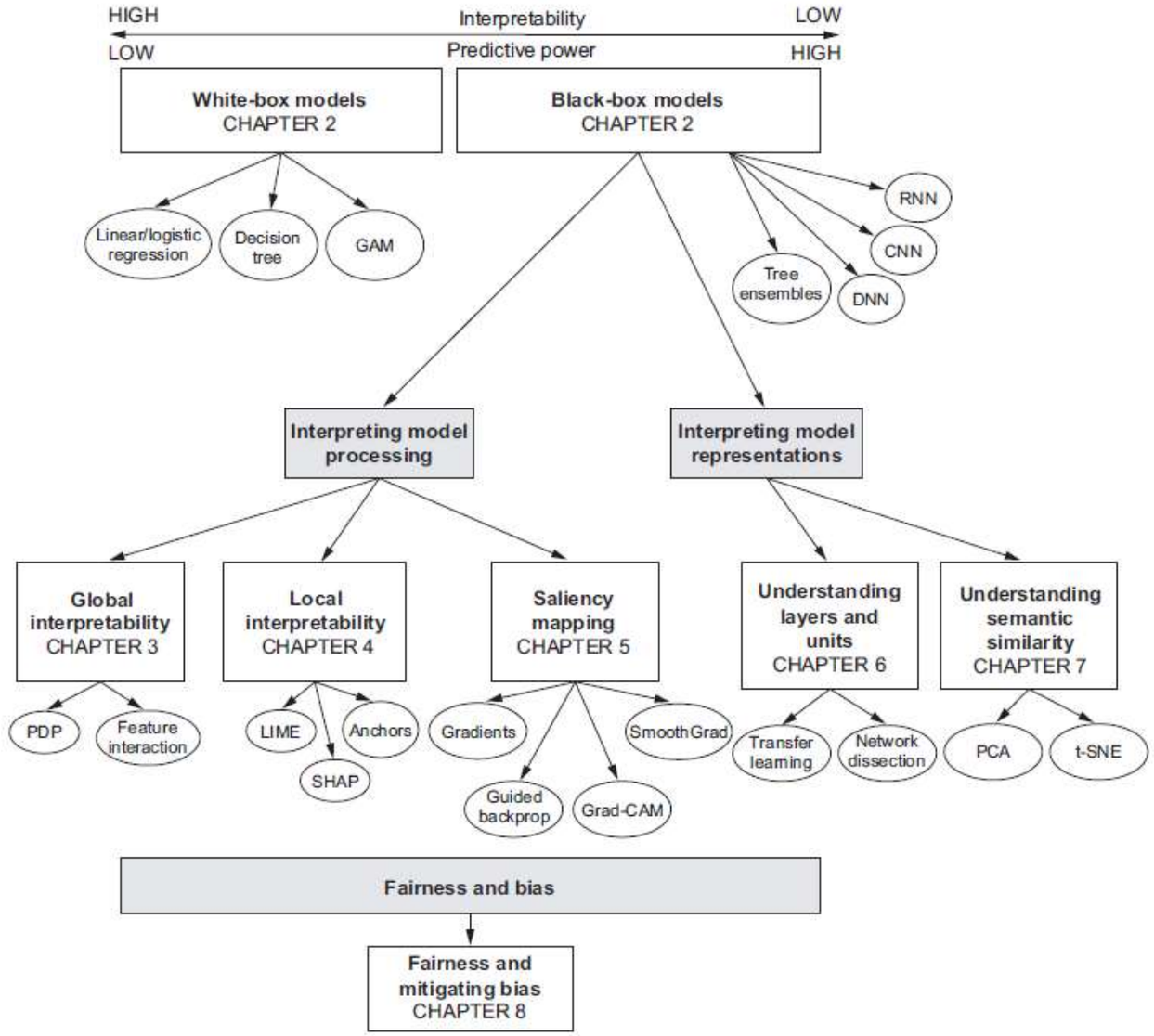
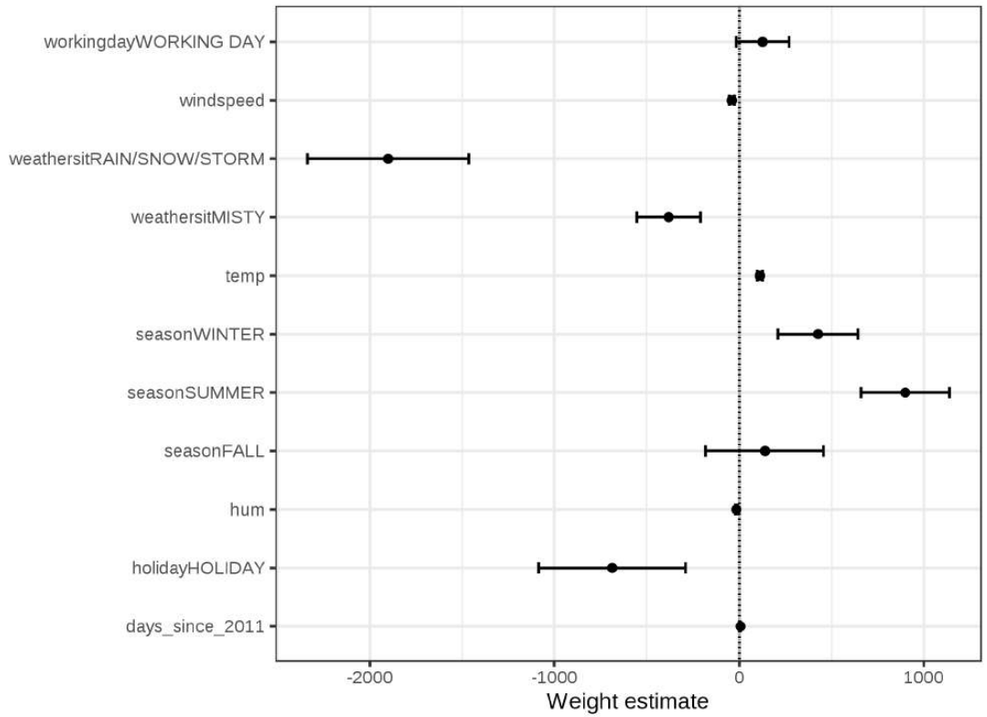


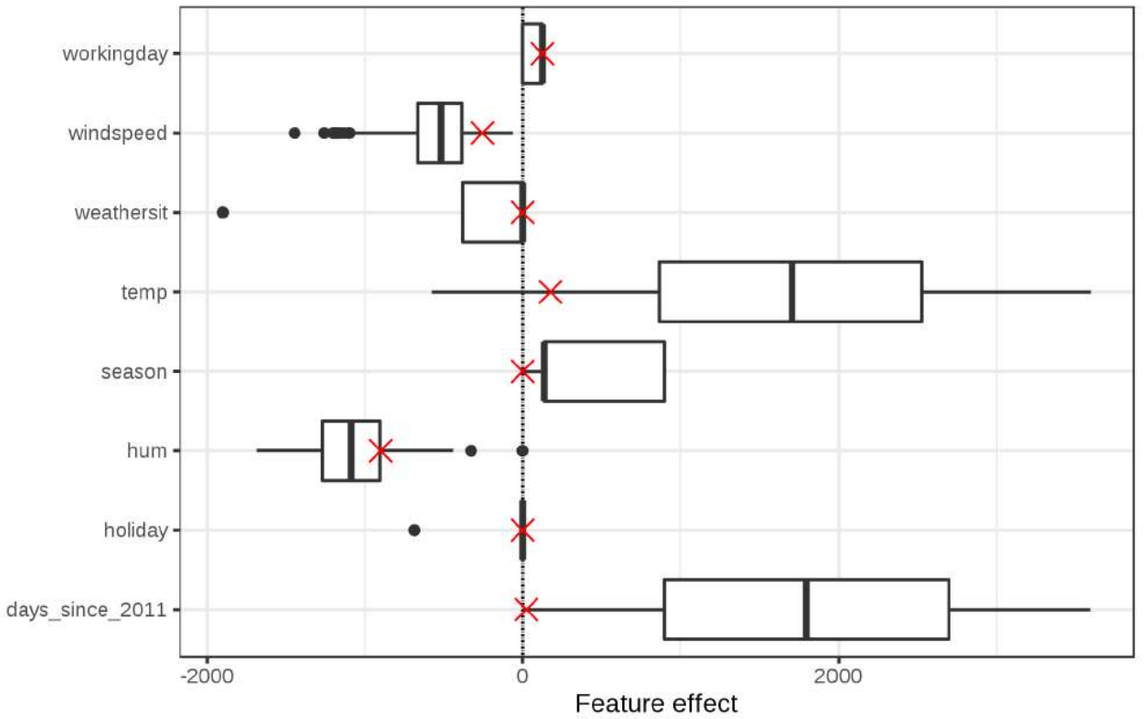
Figure 9.1 Map of our journey through the world of interpretable AI

Fehér doboz esetek - Lineáris regresszió értelmezése

Modell: bérelt kerékpárok száma naponként



súlyok

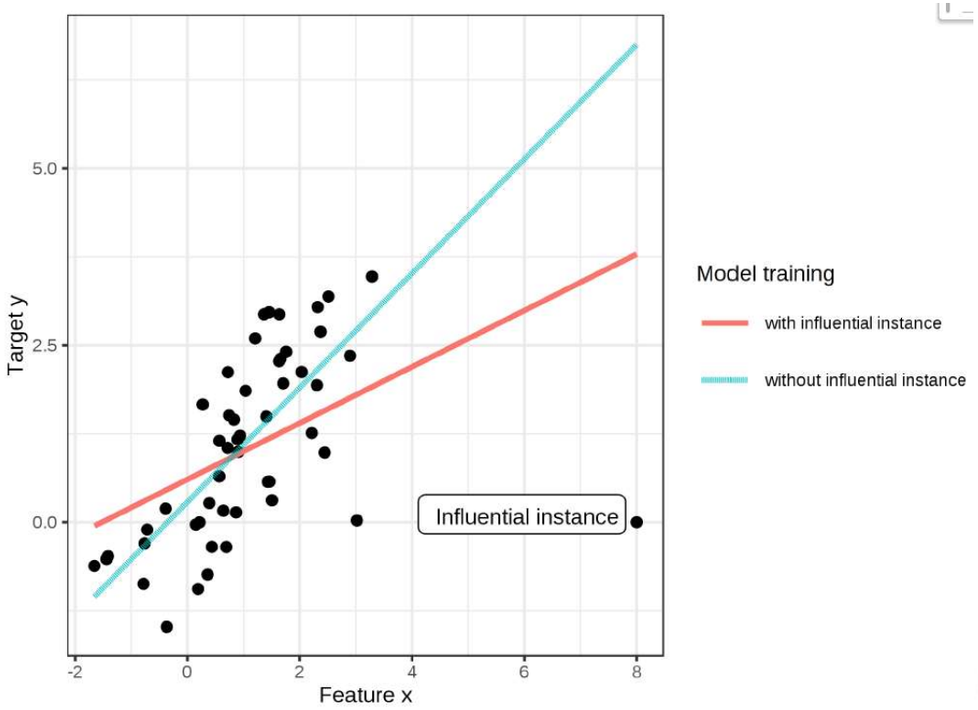


változó hatása (súly×bemenő adat)

Adatpontok vizsgálata

reziduálisok vizsgálatával

klasszikus módszerek (h_{ii}, távolságok, kiugró értékek….), (x-, y-outlier, leverage, influential)



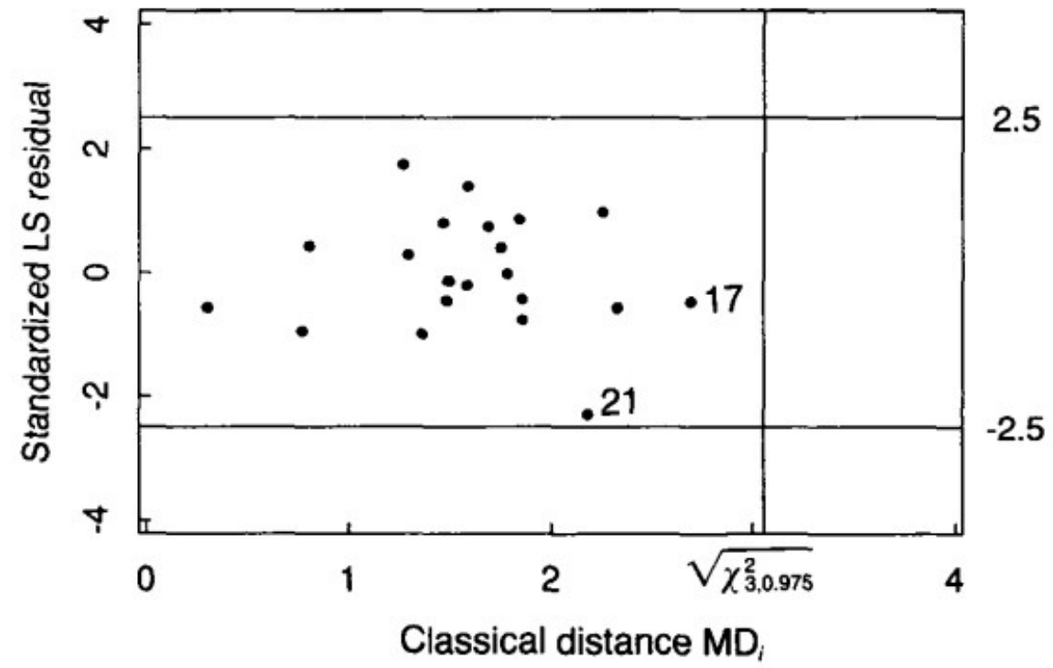
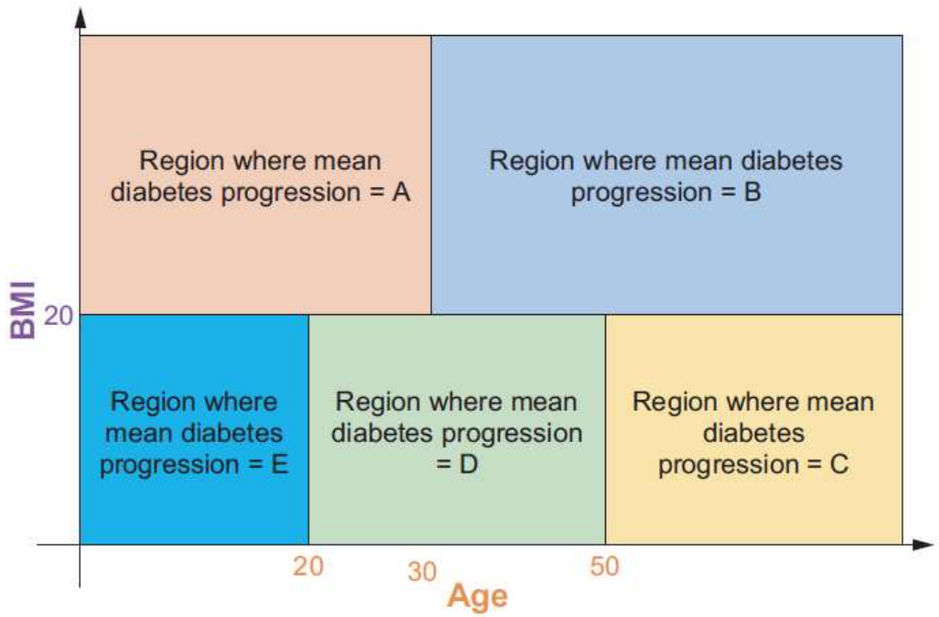


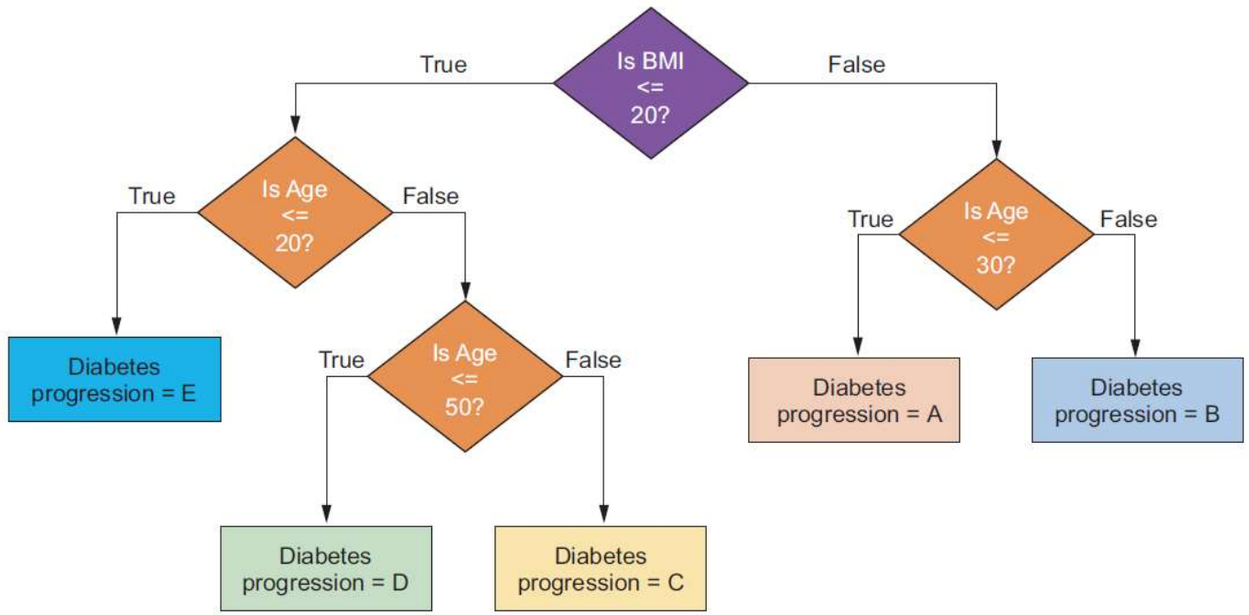
Figure 4. Plot of Least Squares Residuals Versus Classical Mahalanobis Distances MD_i for the Stackloss Data.

Rousseeuw, van Zomeren DOI:10.1080/01621459.1990.10474920

Fehér doboz esetek - Döntési fa (CART)

Modell: cukorbetegség előrehaladása

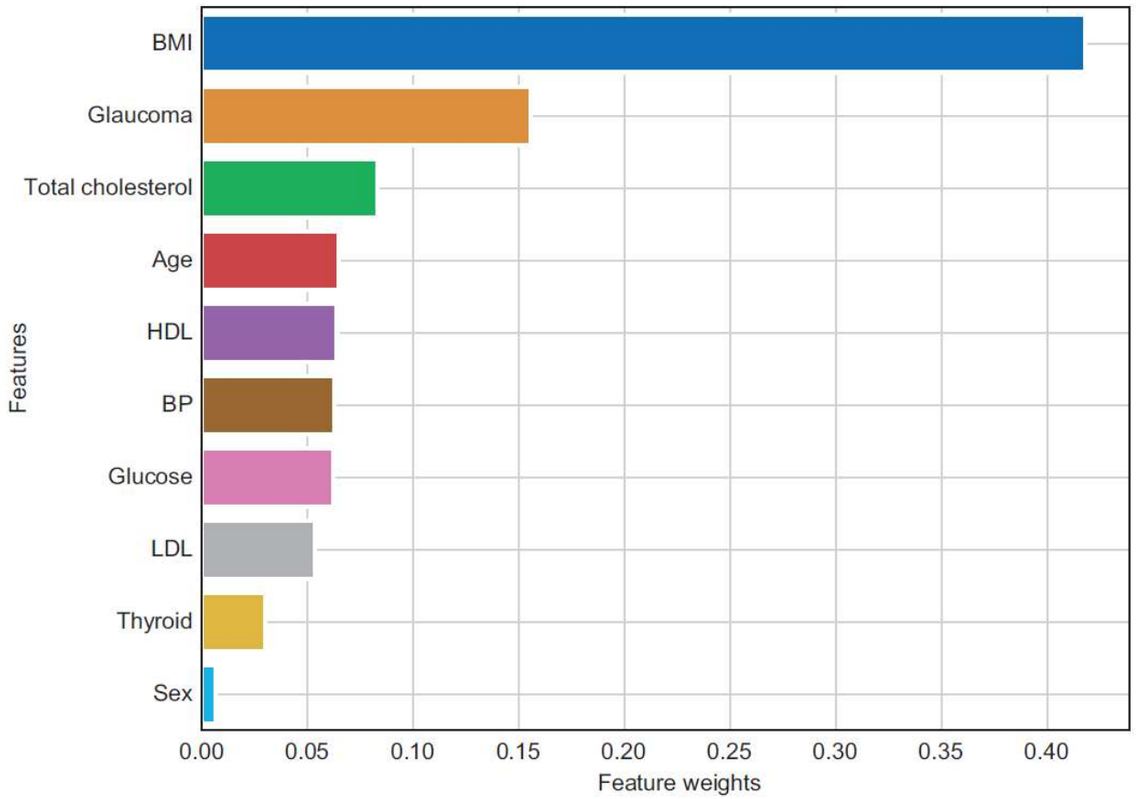




Változók jelentősége/súlya/fontossága

$$\underbrace{I_k^{node}}_{\text{Importance of node } k} = \overbrace{p_k}^{\text{Proportion of samples to reach node } k} \cdot \underbrace{m_k}_{\text{Impurity measure of node k}} - \overbrace{p_k^{(left)}}^{\text{Proportion of samples to reach left subtree of node } k} \cdot \underbrace{m_k^{(left)}}_{\text{Impurity measure of left subtree of node } k} - \overbrace{p_k^{(right)}}^{\text{Proportion of samples to reach right subtree of node } k} \cdot \underbrace{m_k^{(right)}}_{\text{Impurity measure of right subtree of node } k}$$

$$\underbrace{I_i^{feature}}_{\text{Importance of feature } i} = \frac{\overbrace{\sum_{j \epsilon \mathbb{J}} I_j^{node}}^{\text{Sum of importance of all nodes } j \text{ that split on feature } i}}{\underbrace{\sum_{k \epsilon \mathbb{K}} I_k^{node}}_{\text{Sum of importance of all nodes } k \text{ in the decision tree}}}$$



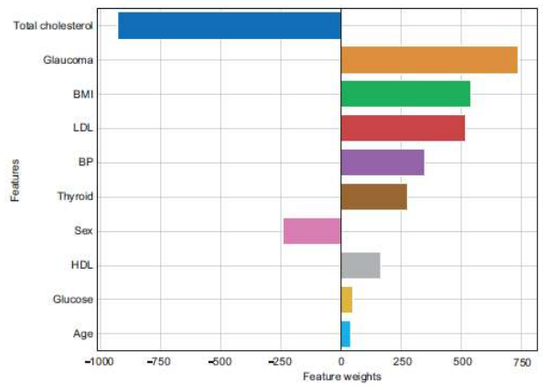
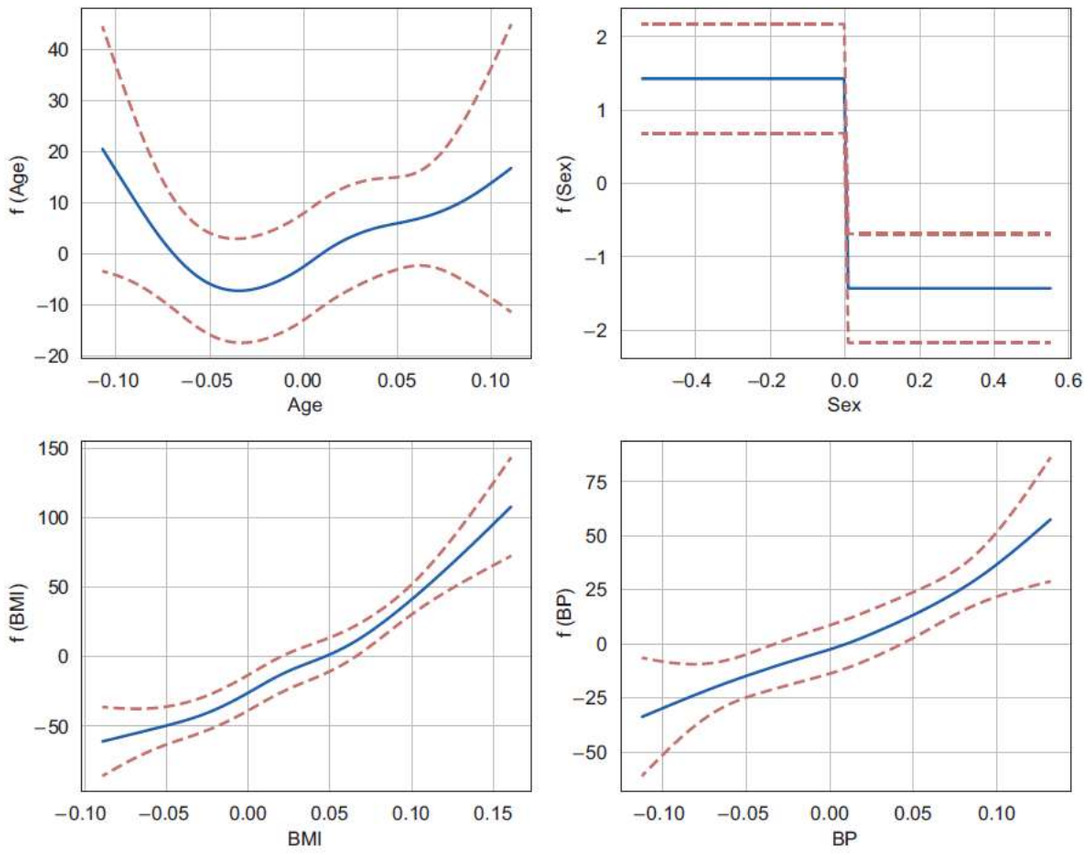


Figure 2.7 Feature importance for the diabetes linear regression model

Fehér doboz esetek – Általánosított additív modell (Generalized Additive Model, GAM)

(hasonló: általánosított lineáris modell GLM)



parciális függőség ábra (PDP)

Modell: cukorbetegség előrehaladása

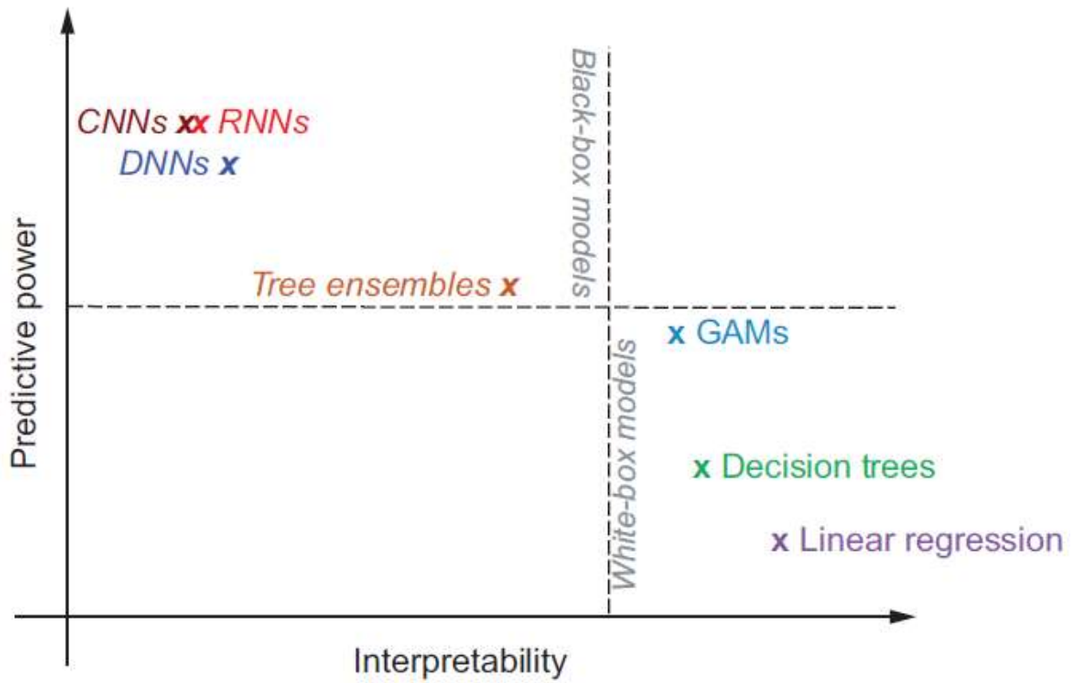
$$y = w_0 + \underbrace{f_1(x_1)}_{\text{Smoothing Function for Feature } x_1} + \underbrace{f_2(x_2)}_{\text{Smoothing Function for Feature } x_2} + \ldots + \underbrace{f_n(x_n)}_{\text{Smoothing Function for Feature } x_n}$$

nincs állandó súly → változóknak értékeinek a hatása a válaszváltózóra

Fekete doboz modellek - példák

Table 2.3 Mapping of black-box model to machine learning tasks

Black-box model	Machine learning tasks
Tree ensembles (random forest, gradient-boosted trees)	Regression and classification
Deep neural networks (DNNs)	Regression and classification
Convolutional neural networks (CNNs)	Image classification, object detection
Recurrent neural networks (RNNs)	Sequence modeling, language understanding



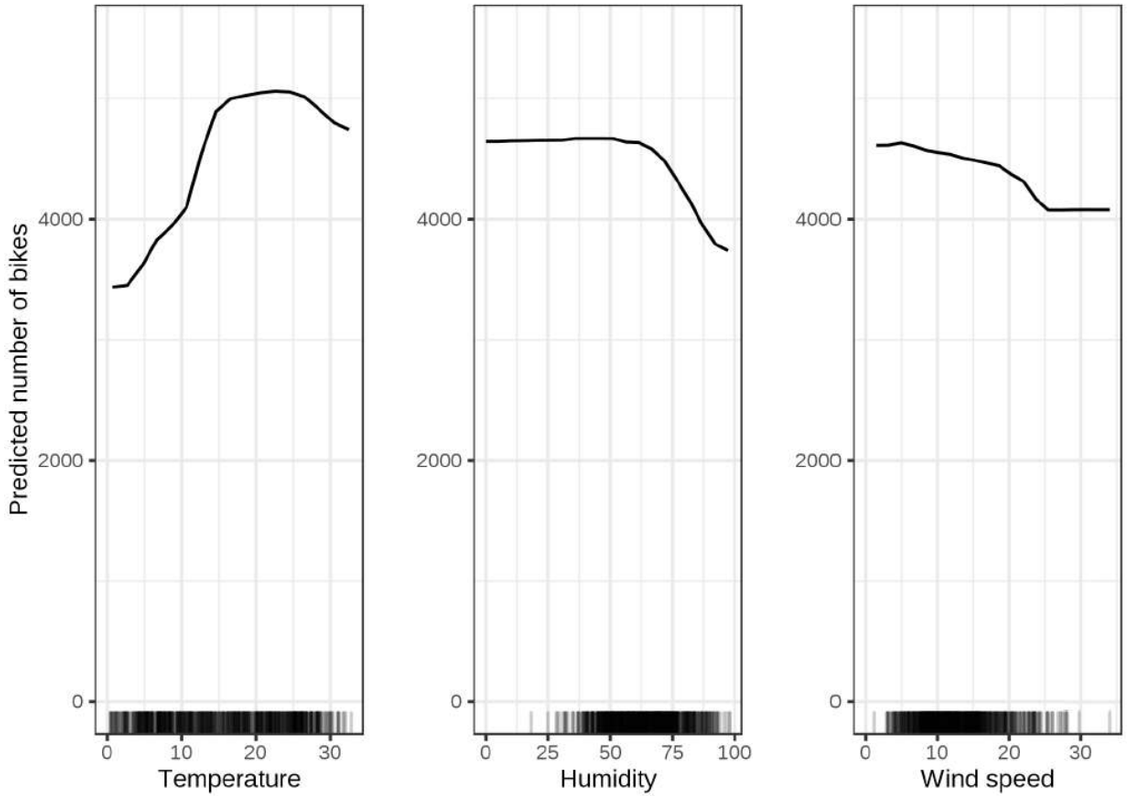
Modell független utólagos módszerek (model-agnostic methods, post-hoc)

Parciális függőség ábra (Partial Dependance Plot)

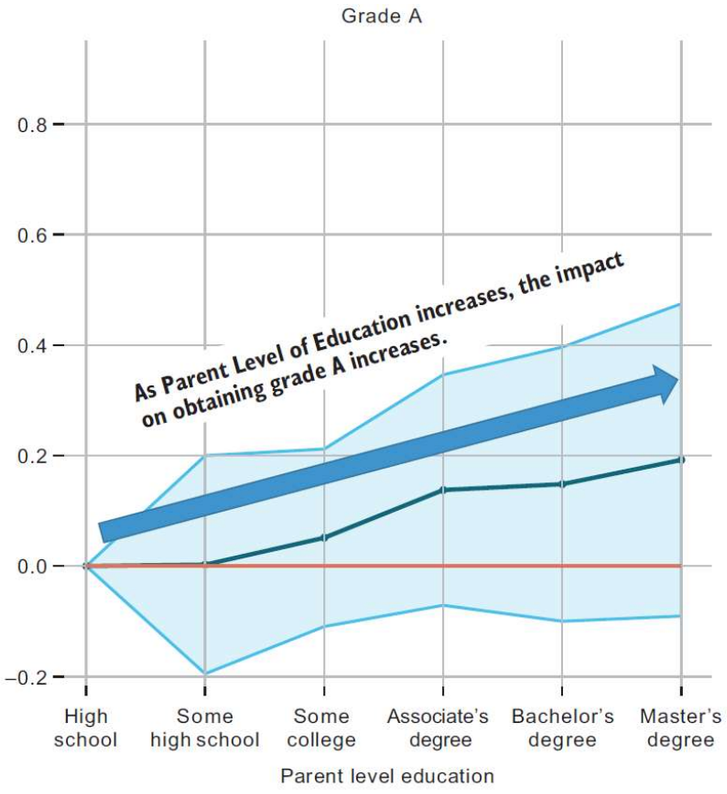
$$\hat{f}_{\text{math},\,x_S}(x_s|\mathbf{x}_C) = \frac{1}{n}\sum_{i=1}^{n} f_{\text{math}}\left(x_S, x_C^{(i)}\right)$$

jó, ha S és C nem korrelál

Modell: bérelt kerékpárok száma



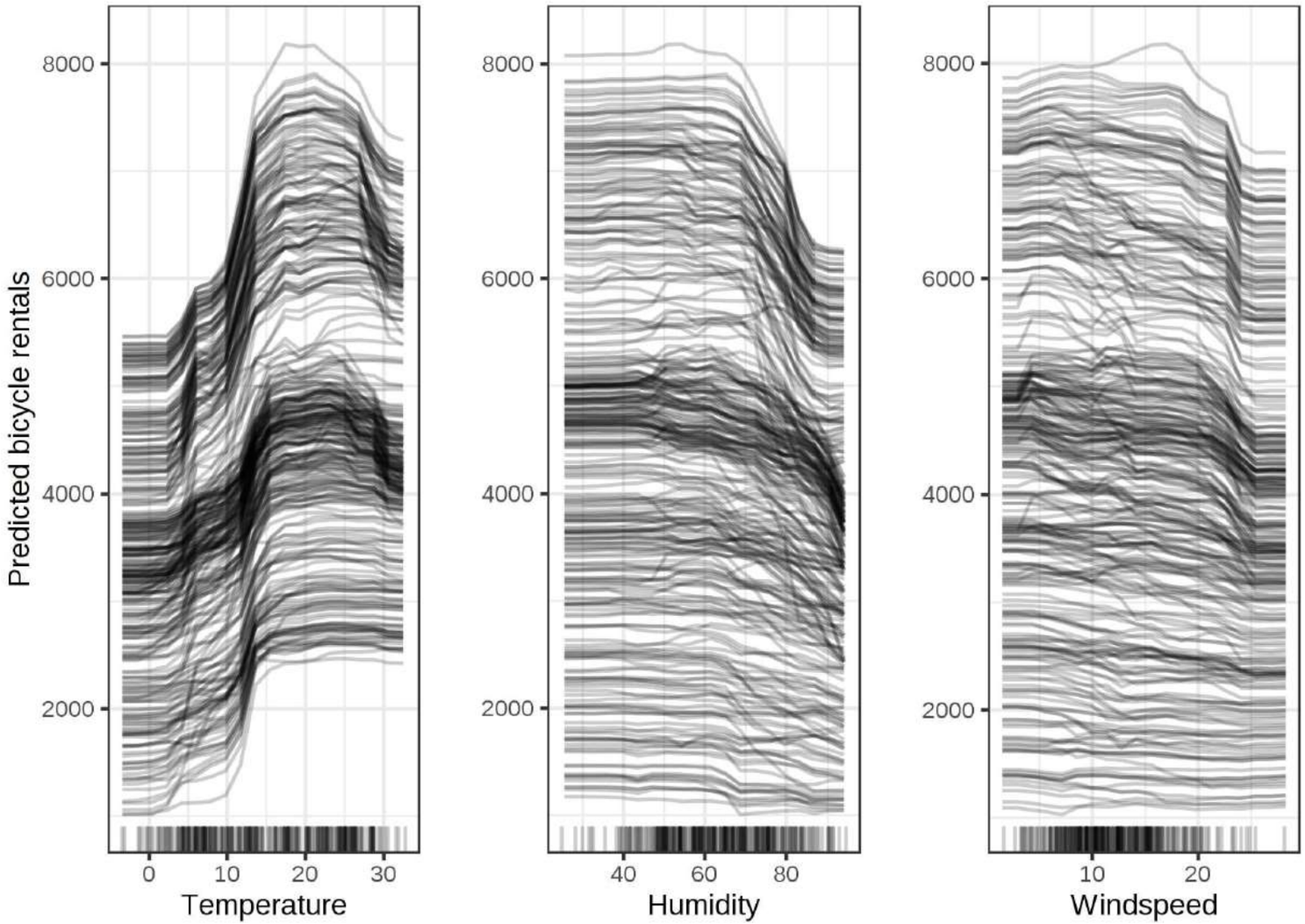
„A” matek osztályzat



Egyedi feltételes várható érték görbe (Individual conditional expectation plot)

Az előzőből pont a C a sajátja és az S megy végig

Modell: bérelt kerékpárok száma



Sokaság alapú döntési fák

Bagging – többnyire bootstrap módszerrel mintavételezett tanító halmaz + random features → Random Forest

Boosting – ugyanez szekvenciálisan, ahol az előző hibáit próbáljuk javítani → adaptive boosting, gradient boosting

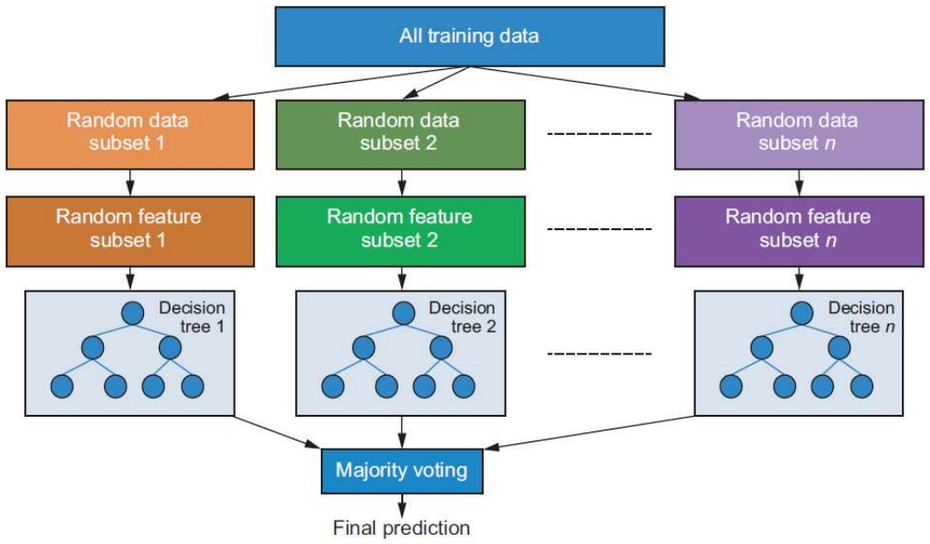


Figure 3.6 An illustration of the random forest algorithm

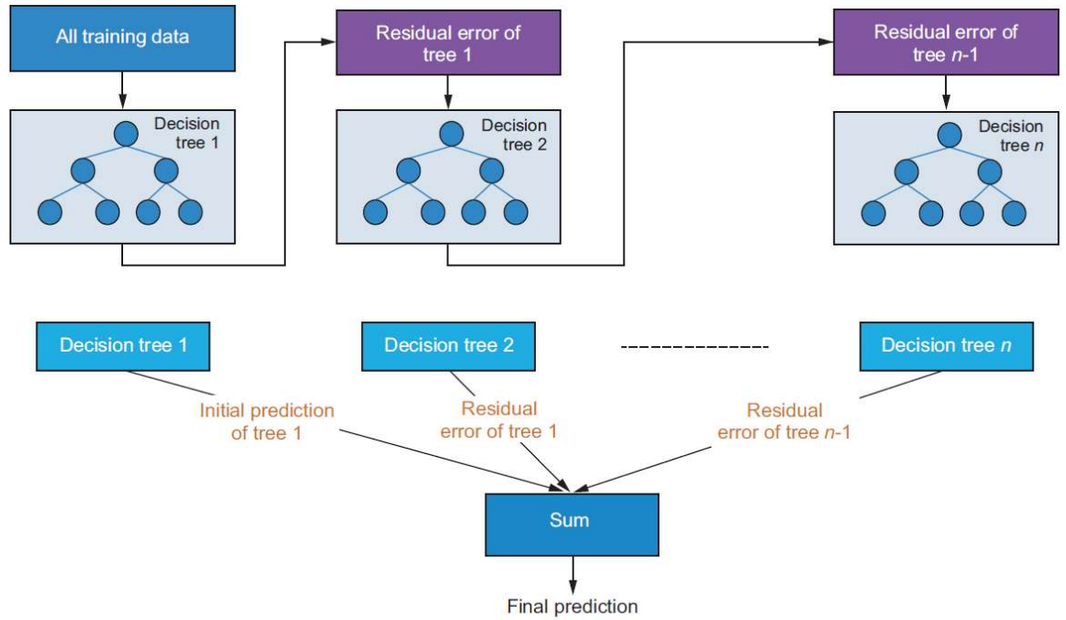


Figure 3.8 An illustration of the gradient-boosting algorithm

Változók fontossága – tanító/teszt halmaz kérdése

Feature importance: átlag a súlyozott fákról

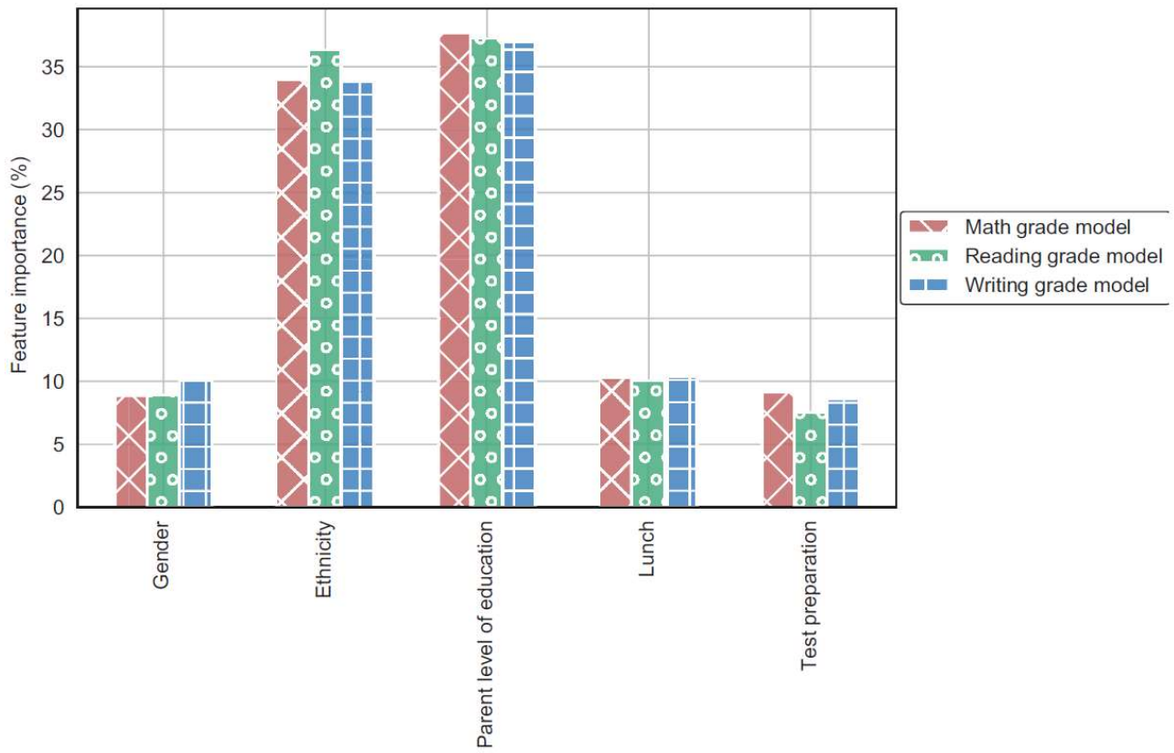


Figure 3.9 Feature importance of the random forest model

Tanító, validáló vagy teszt halmazon?

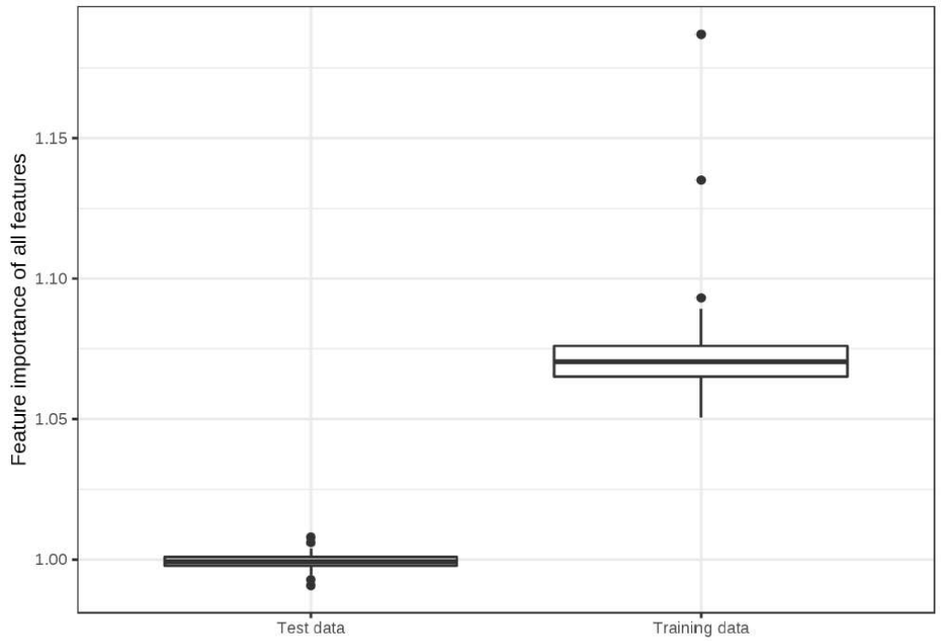
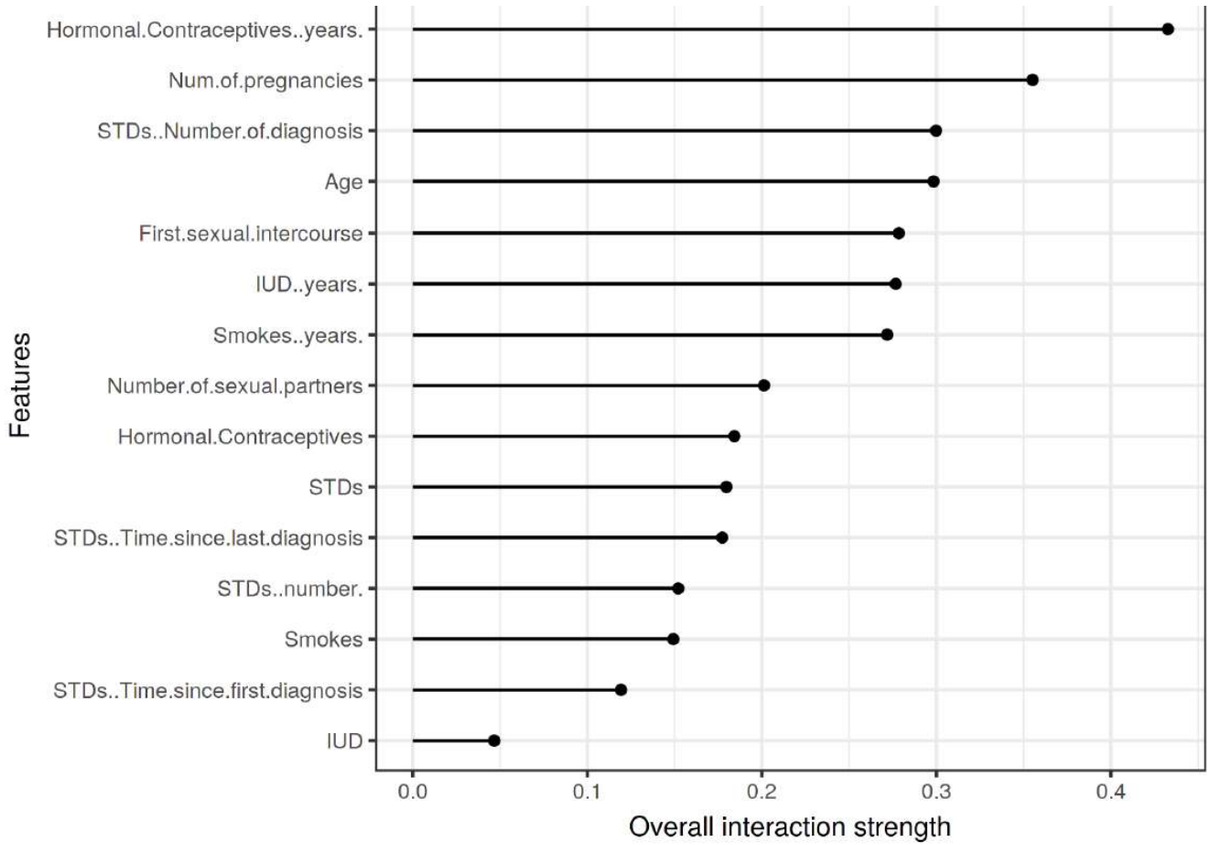


FIGURE 5.27: Distributions of feature importance values by data type. An SVM was trained on a regression dataset with 50 random features and 200 instances. The SVM overfits the data: Feature importance based on the training data shows many important features. Computed on unseen test data, the feature importances are close to a ratio of one (=unimportant).

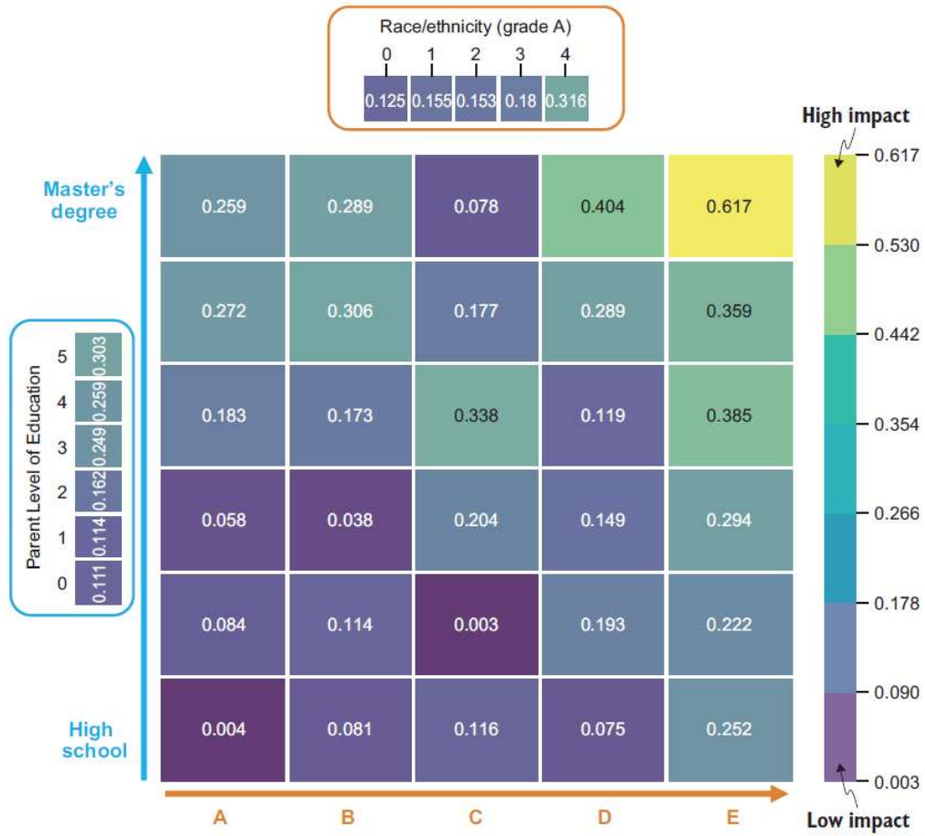
Változók kölcsönhatása (Feature interaction)

kovarianciához-korrelációhoz hasonló: A és B változó esetén hatások: konstans, A, B, A és B közös

Modell: méhnyakrák becslése



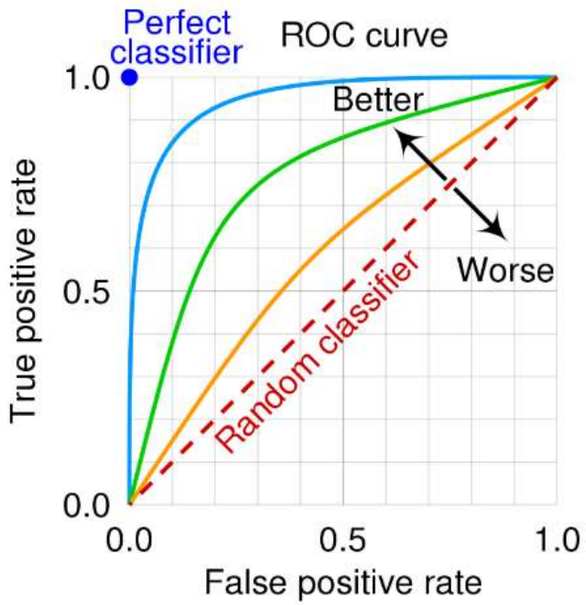
matek „A” osztályzat



Kontroll változó/küszöb érték hatása bináris klasszifikációban

(kontroll változó/küszöb érték = függ tőle a döntés)

		Predicted class	
		Successful	Unsuccessful
Actual class	Successful	TP	FN
	Unsuccessful	FP	TN



$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ (Sensitivity, True\ positive\ rate) = \frac{TP}{TP + FN}$$

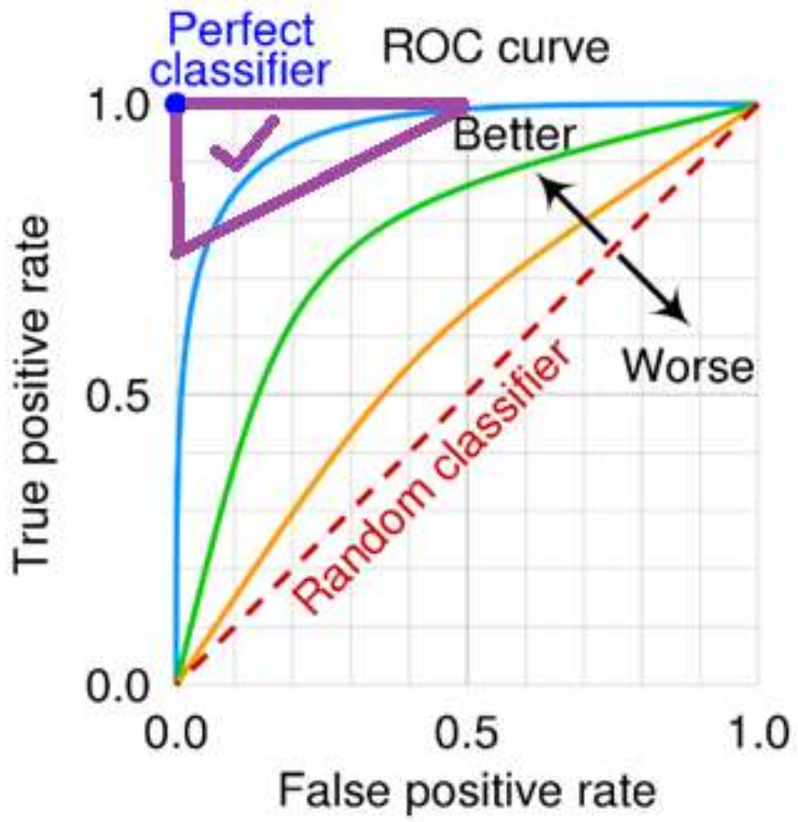
$$False\ positive\ rate = \frac{FP}{TN + FP}$$

$$F1 = 2\frac{Precision\ \times Recall}{Precision + Recall}$$

AUC ROC görbe alatti terület

Miért nem szerepel a küszöb érték? Validáció?

Ami a lényeg lenne és ritkán (se) rajzolják fel!



Lila rész az elfogadható tartomány max. 5% maradandó károsodásra, ha

betegek 20% maradandóan károsodik, ha nem kezelik

egészségesek 10% maradandóan károsodik, ha kezelik

betegek száma = egészségesek száma

Kis dimenziójú modellek – direkt vizualizáció

Sor- és oszlopcserével jobban értelmezhető egy modell (szeriálás pl. patch módszerrel)

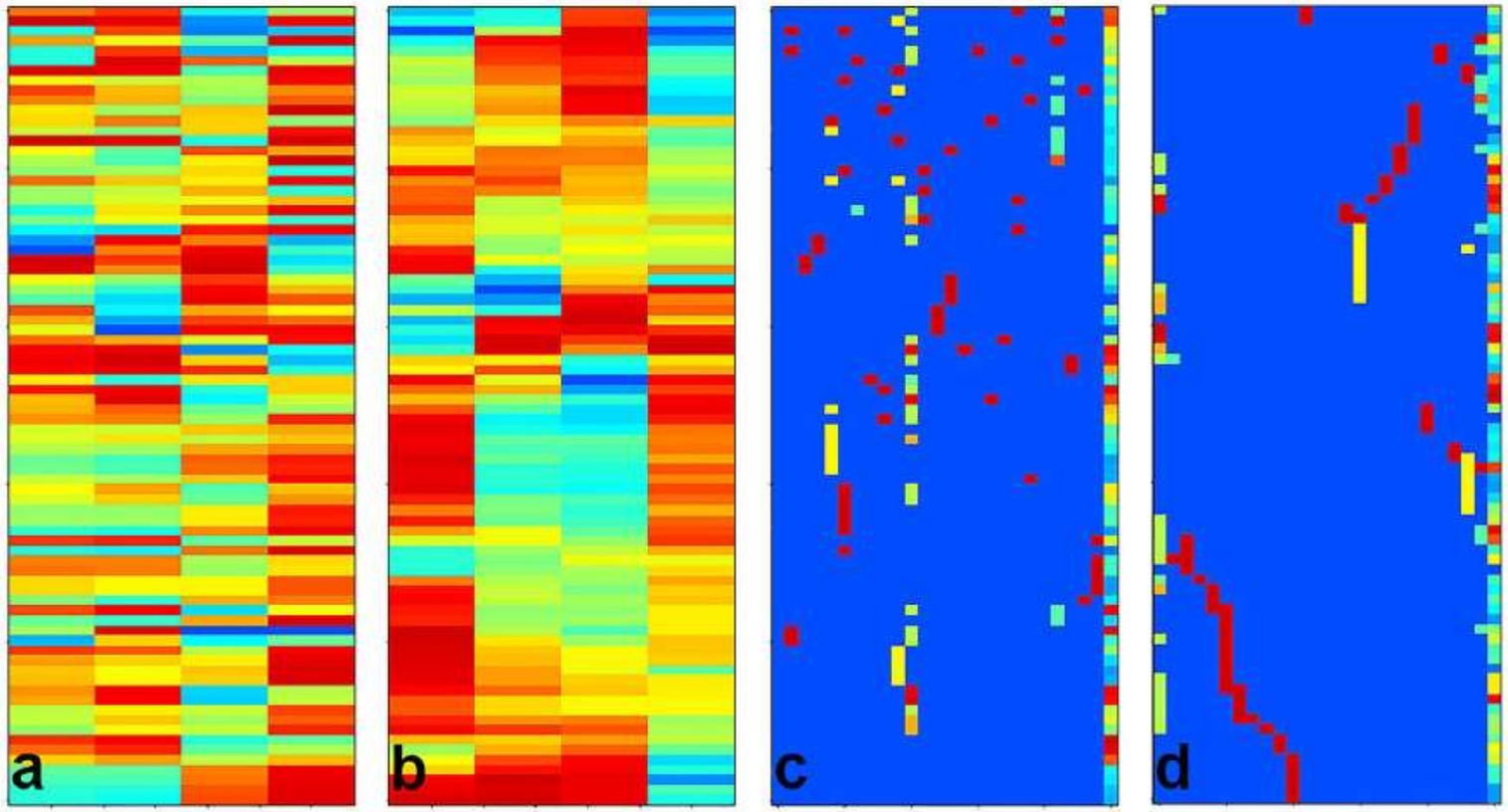


Fig. 8 Seriation of the FLASHP1 data (test set). **a**–**b** Object – object activities on the hidden layer neurons (model: logistic function with 4 neurons) **a**- original **b**-seriated **c**–**d** Object – original variable data **c**-random order **d**-seriated

Lasfar and Tóth *Journal of Cheminformatics (2023) 15:78*

Kis dimenziójú modellek – modellválasztás

átláthatóság ≈ interpretálhatóság

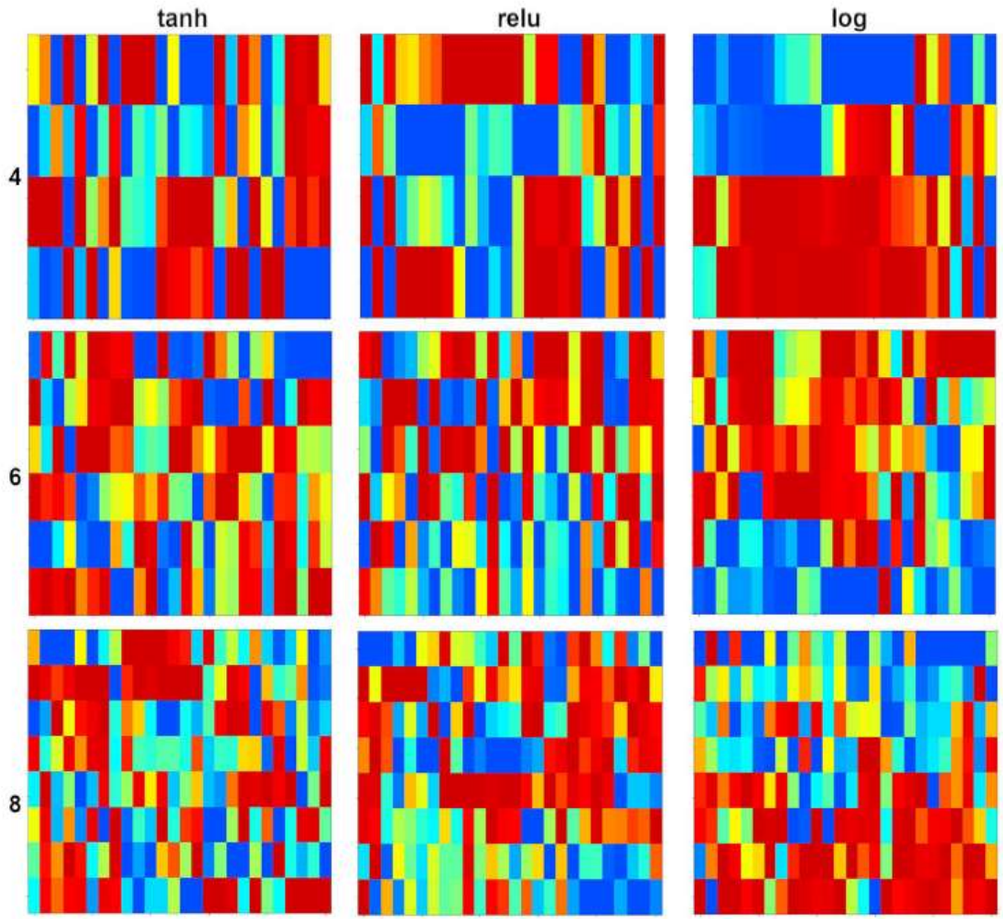


Fig. 7 Seriated details of neural network models on the FLASHP1 dataset. The objects are the neurons, and the variables are the scaled weights of the original input variables. Three activation function are used (tangent hyperbolic, relu and logistic) with 4, 6 or 8 neurons in the hidden layer

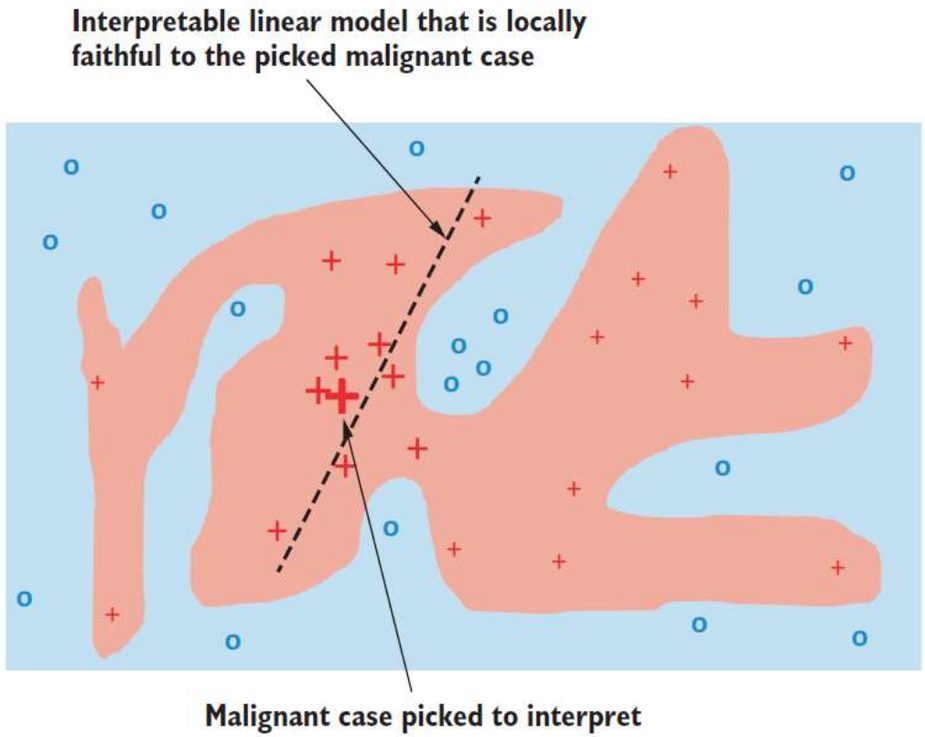
Lasfar and Tóth *Journal of Cheminformatics (2023) 15:78*

Helyettesítő modellek (surrogate models)

nagy dimenziójú fekete doboz modell helyettesítése interpretálható lokális/globális fehér doboz modellel

LIME – local interpretable model-agnostic explanation (2016)

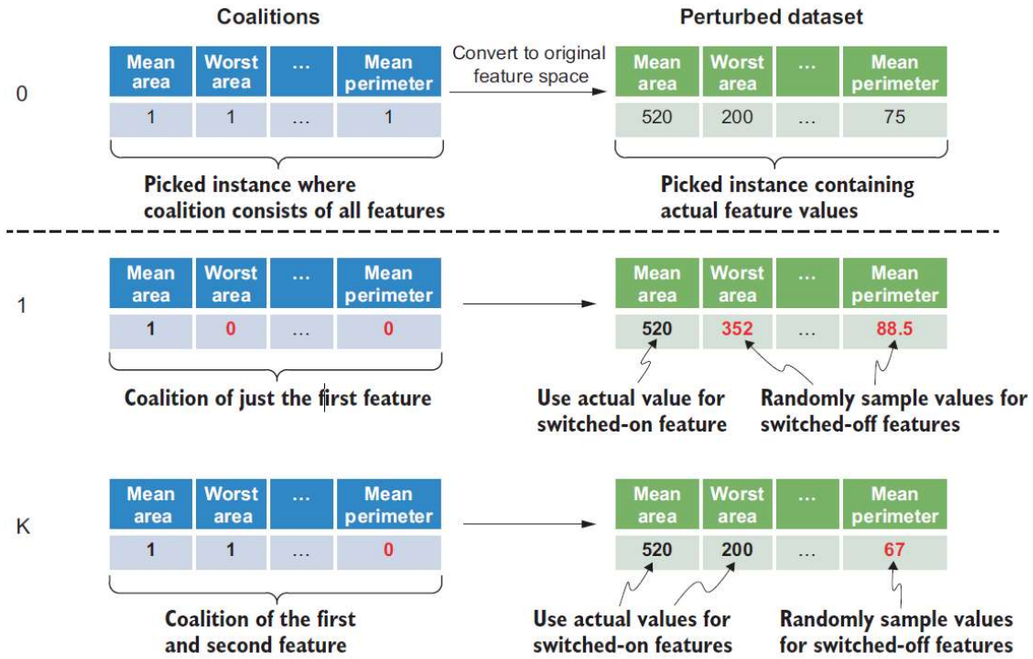
Modell: daganat klasszifikáció - illusztráció



- Random pontok a kiszemelt mellett
- Súlyozás valamilyen távolsággal lecsengő függvénnyel
- Pl. egyenes illesztés a lokális helyre
- Modell lokális értelmezése
- Humán értelmezhető magyarázat

SHAP SHapley Additive exPlanations

Perturbált új adathalmaz, ahol a tulajdonságok egy része *(1 - M)* marad meg (koalíciók), a többi random



Megmagyarázza, hogy a véletlenhez képest melyik tulajdonság okozza a döntést:

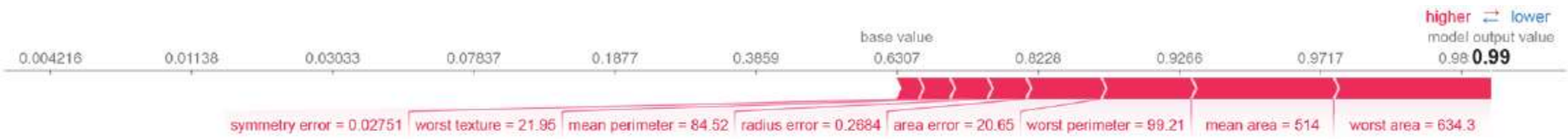
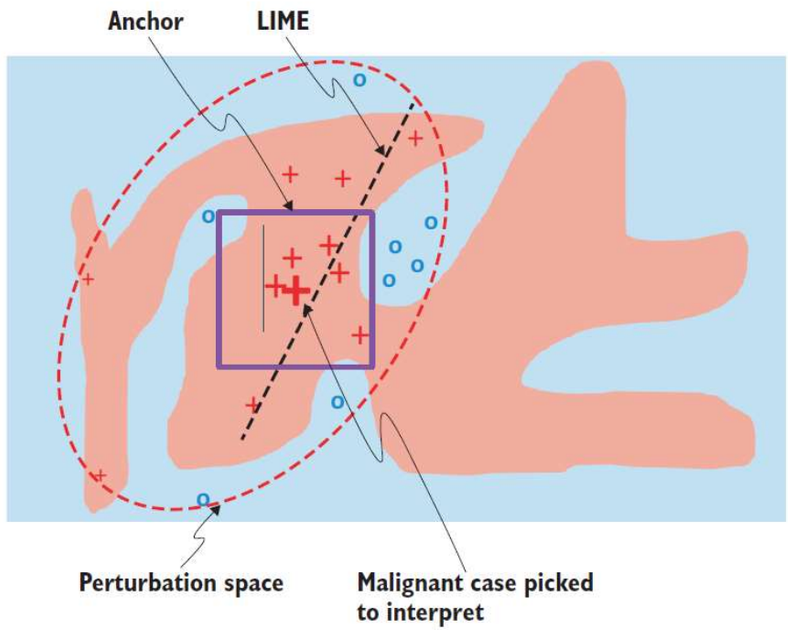


Figure 4.21 SHAP interpretation of benign case 1 where the DNN model predicts benign with a probability of 0.99 (or a confidence of 99%)

Horgony (anchor)

A kijelölt területre vonatkozóan olyan változó tartományok keresése, ami lokálisan jó magyarázatként, de megmondja a globális használhatóságot (coverage) is



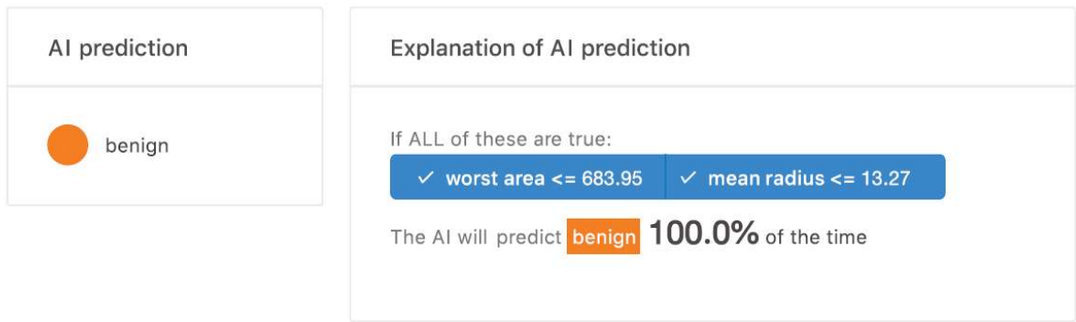
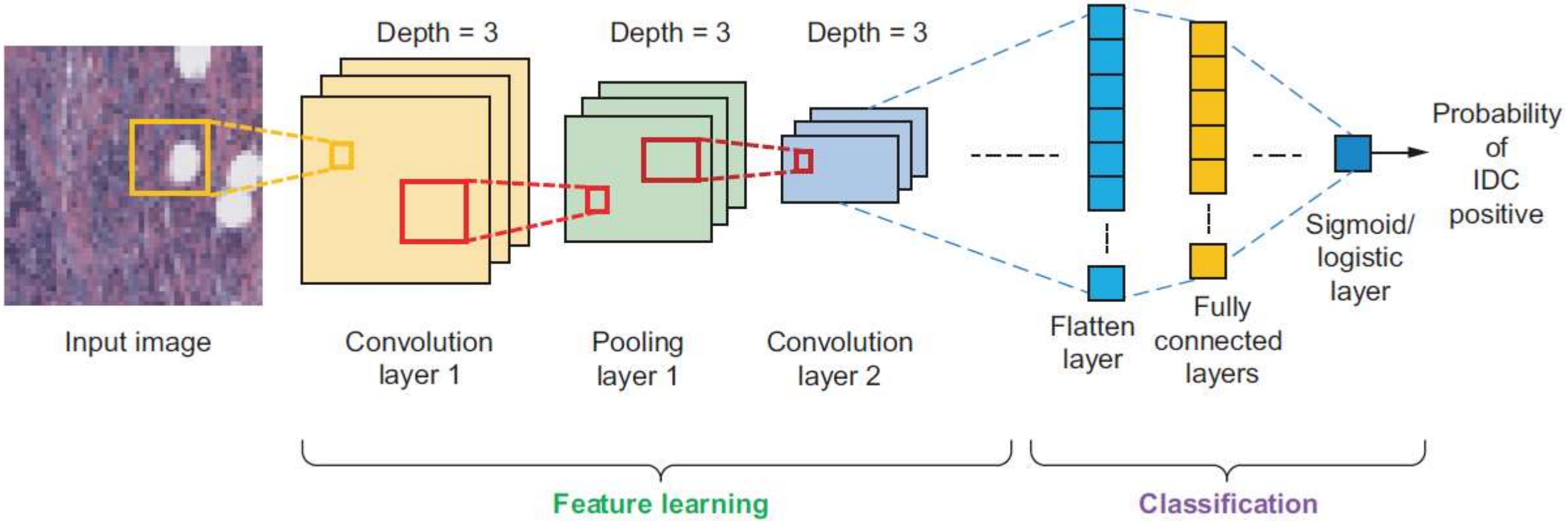


Figure 4.26 Anchor interpretation of benign case 1 where precision is 100% and coverage is 44.3%

Konvolúciós mesterséges ideghálók

főleg képfeldolgozás (osztályozás, tárgyfelismerés, képdarabolás)

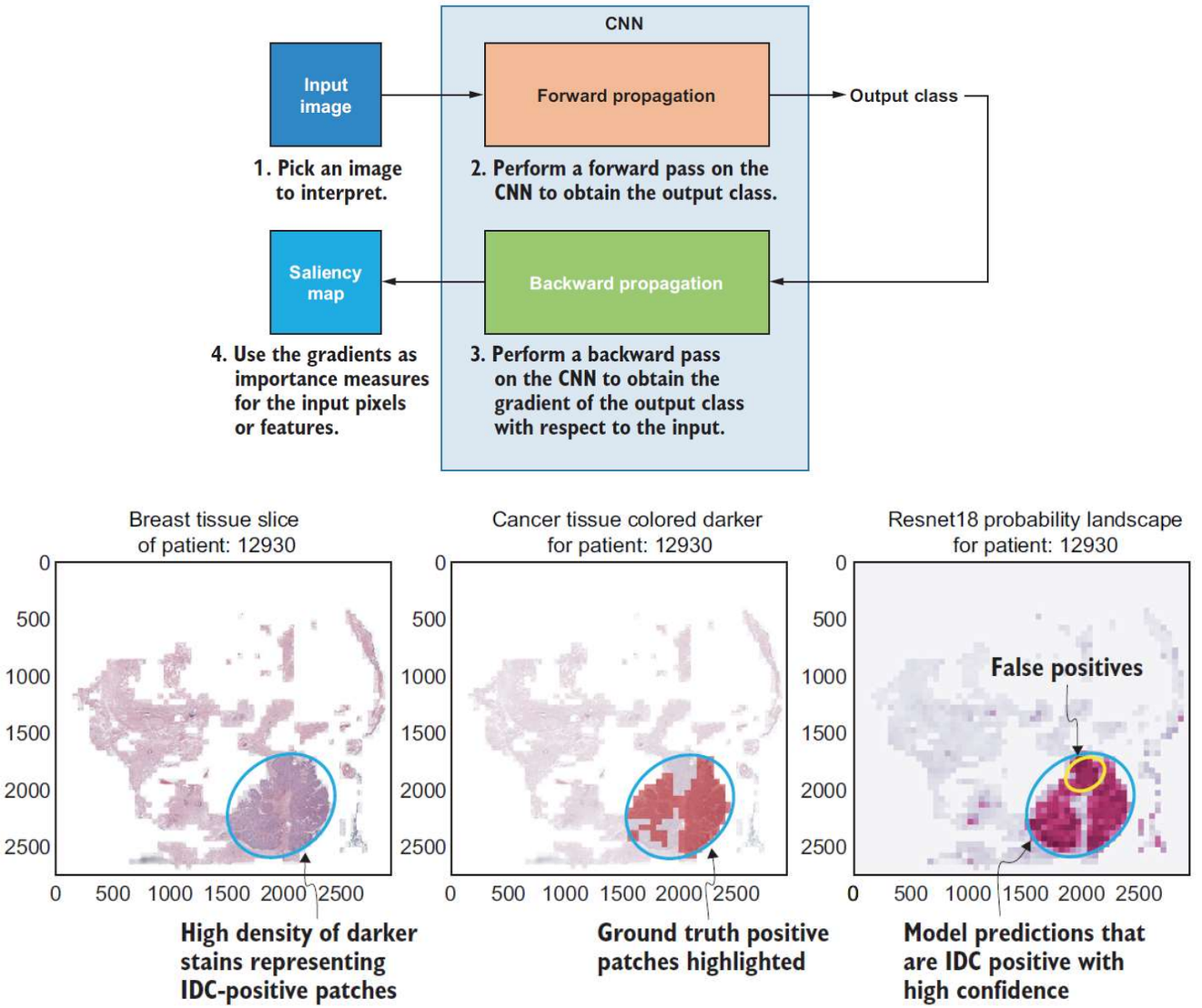


Vizuális módszerek a kép azon részeinek a meghatározására, amelyek fontosak a döntésnél (saliency maps)

- Perturbációs módszerek (mint pl. LIME) – néha túl drága a sok pont
- Gradiens módszerek (pl. vanilla backpropagation)
- Aktivációs módszerek (pl. Grad-CAM) (coarse graining lehetőség)

Gradiens módszerre példa

vanilla backpropagation



„Integrated Gradients can be used to extract interpretable pharmacophores from a feedforward neural network.” arXiv:1903.02788v2, Preuer et al 2019.

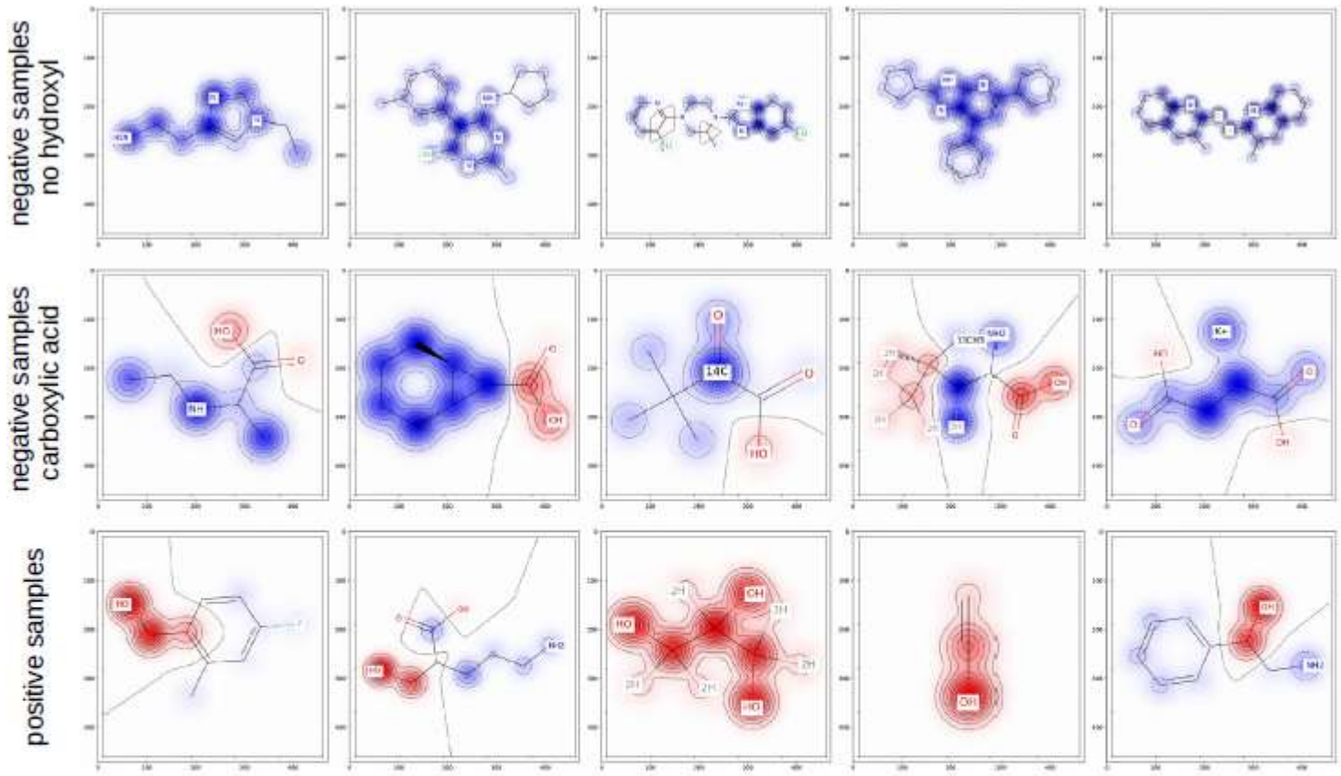
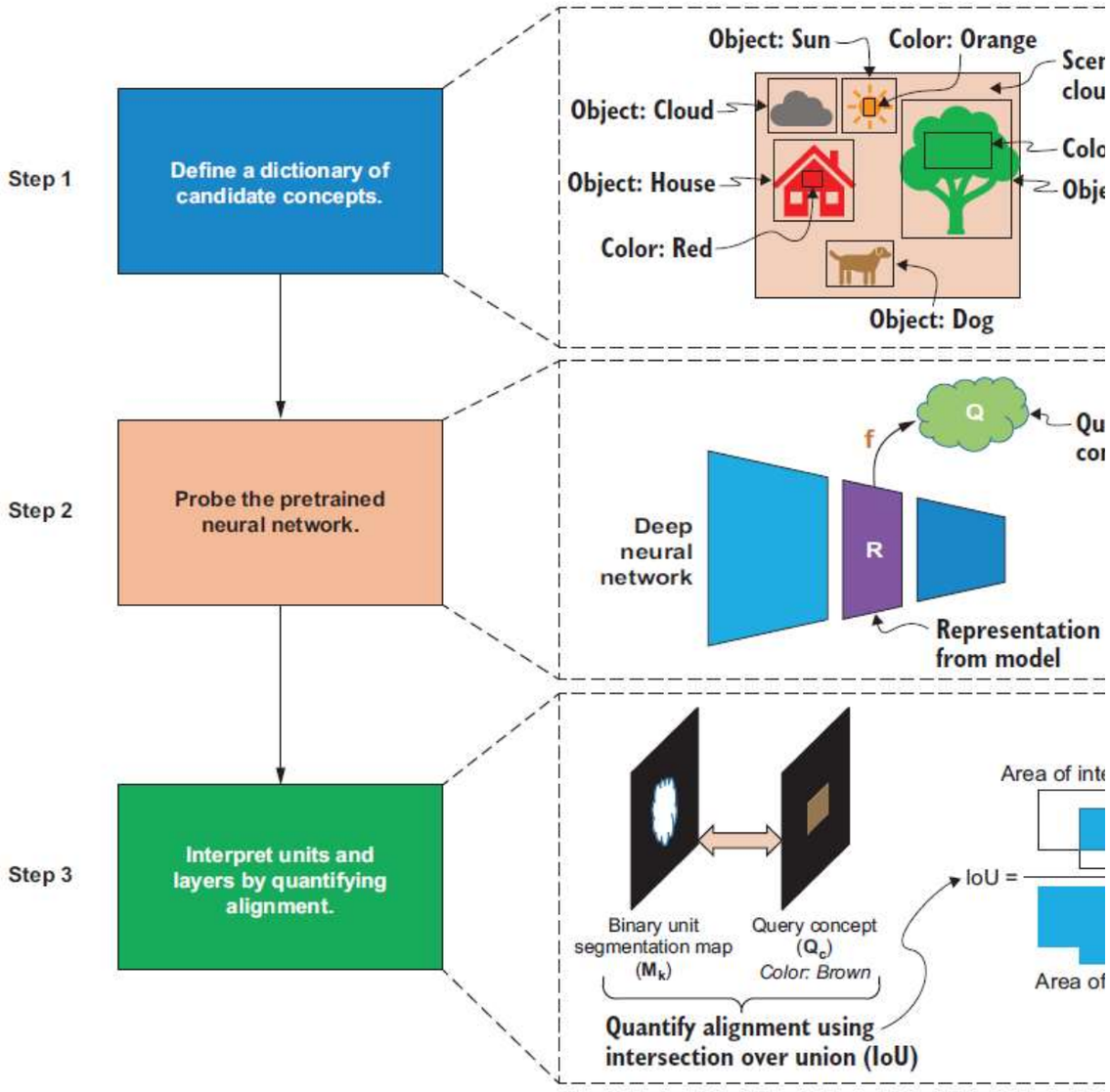


Fig. 3. Attributions assigned to the atoms by the model for the three types of compounds. 5 randomly chosen negative samples without hydroxyl groups, negative samples with carboxylic acid groups and positive samples are shown in the top, middle and bottom row, respectively. Dark red indicates that these atoms are responsible for a positive classification, whereas dark blue atoms attribute to a negative classification.

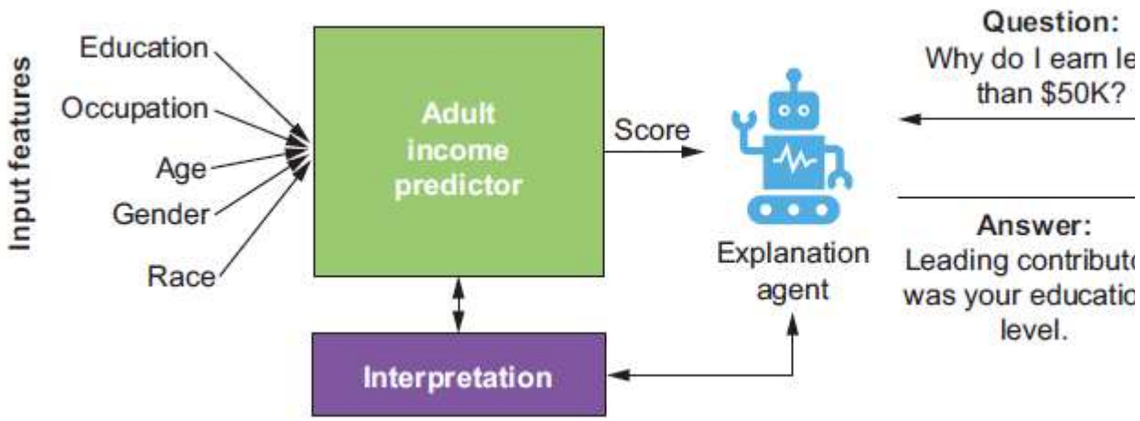
Kép (hálózat) felbontása (Network dissection framework)



- The concept definition step is the most crucial step because it requires us to collect a labeled dataset of concepts at the pixel level.
- The network probing step is about finding units in the network that respond to those predefined concepts.
- The alignment measurement step quantifies how well the unit activation aligns with those concepts.

Magyarázat (explanation)

A modell értelmezése (interpretációja) után



Kemény Sándor: A statisztika oktatásában fontos azt is megtanítani, hogy olyat tudjanak kérdezni, amire lehet a statisztikával válaszolni.

Humán értelmezhető magyarázat:

- Modell: hogyan működik, mik a fontos bemenő adatok
- Predikció: hogyan jut el erre a megoldásra a modell
- Pártatlanság: csoportok egyenlő kezelése, van-e torzítás (fairness, bias)
- Ellentét párok: Miért ez jött ki?-Miért nem a másik? (contrastive explanation, counterfactual)

Köszönöm a figyelmet!