Yapay Öğrenmede Yorumlanabilirlik

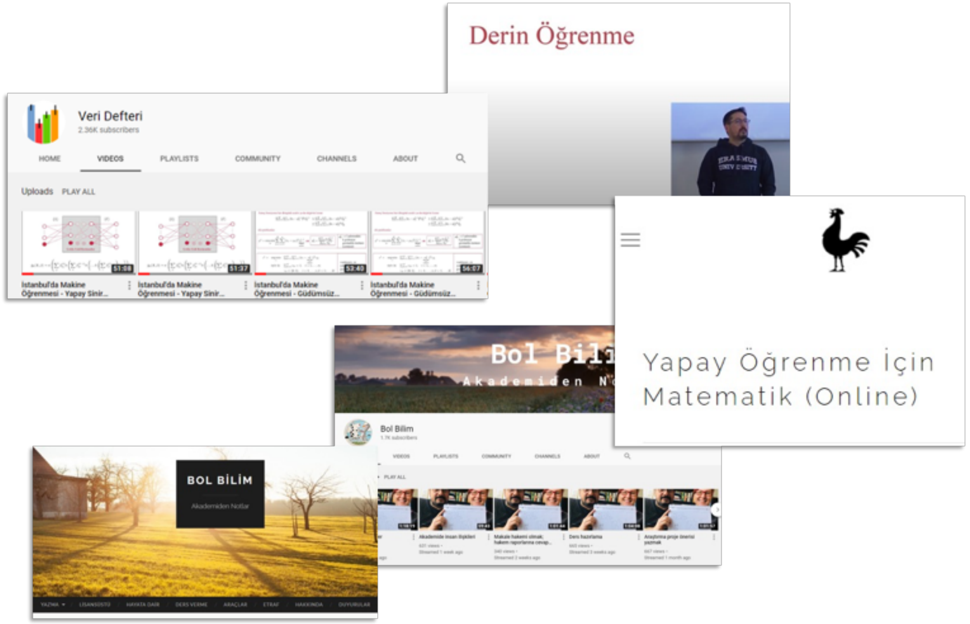
İlker Birbil

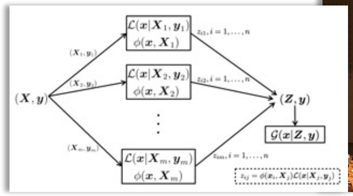
https://sibirbil.github.io

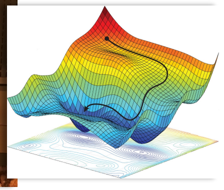
UNIVERSITEIT VAN AMSTERDAM

[UvA](#)
[D4C](#)
[Bol Bilim](#)
[Veri Defteri](#)
[@sibirbil](#)

Veri Bilimi ve Optimizasyon







Robust and Fast Stochastic Gradient Decent with Model Building

Algorithm 1: SMB: Stochastic Model Building

Input: $x_1 \in \mathbb{R}^n$, $f_1, g_1 \in \mathbb{R}$, stepsizes $\{\alpha_k\}_{k=1}^N$, $c > 0$
for $k = 1, \ldots, N$ **do**
  $s_k^t = -\alpha_k g_k$;
  $x_k^t = x_k + s_k^t$, $f_k^t = f(x_k^t, \xi_k)$, $g_k^t = g(x_k^t, \xi_k)$;
  if $f_k^t \leq f_k - c \cdot \alpha_k \|g_k\|^2$ **then**
    $x_{k+1} = x_k^t$, $f_{k+1} = f_k^t$, $g_{k+1} = g_k^t$;
  else
    for each parameter group p **do**
      $y_{k,p} = g_{k,p}^t - g_{k,p}$;
      $s_{k,p} = c_{g,p}(\delta) g_{k,p} + c_{y,p}(\delta) y_{k,p} + c_{s,p}(\delta) s_{k,p}^t$, as defined in (4);
    $x_{k+1} = x_k + s_k$, where $s_k = (s_{k,p_1}, \ldots, s_{k,p_n})$ and n is the number of parameter groups;
    $f_{k+1} = f(x_{k+1}, \xi_k)$, $g_{k+1} = g(x_{k+1}, \xi_k)$;

Data Privacy in Bid-Price Control for Network Revenue Management

Discovering Classification Rules for Interpretable Learning with Linear Programming

This work has been published in Computational Optimization and Applications. You can acc

hamsi-mf: HAMSI for matrix factorization

The Washington Post

Democracy Dies in Darkness

'Creative … motivating' and fired

Sarah Wysocki was out of work for only a few days after she was fired by DCPS last year. She is now teaching at Hybla Valley Elementary School in Fairfax County. (Jahi Chikwendiu/The Washington Post)

By **Bill Turque**
March 6, 2012

HDSR

Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson From An Explainable AI Competition

by Cynthia Rudin and Joanna Radin

Published on Nov 22, 2019

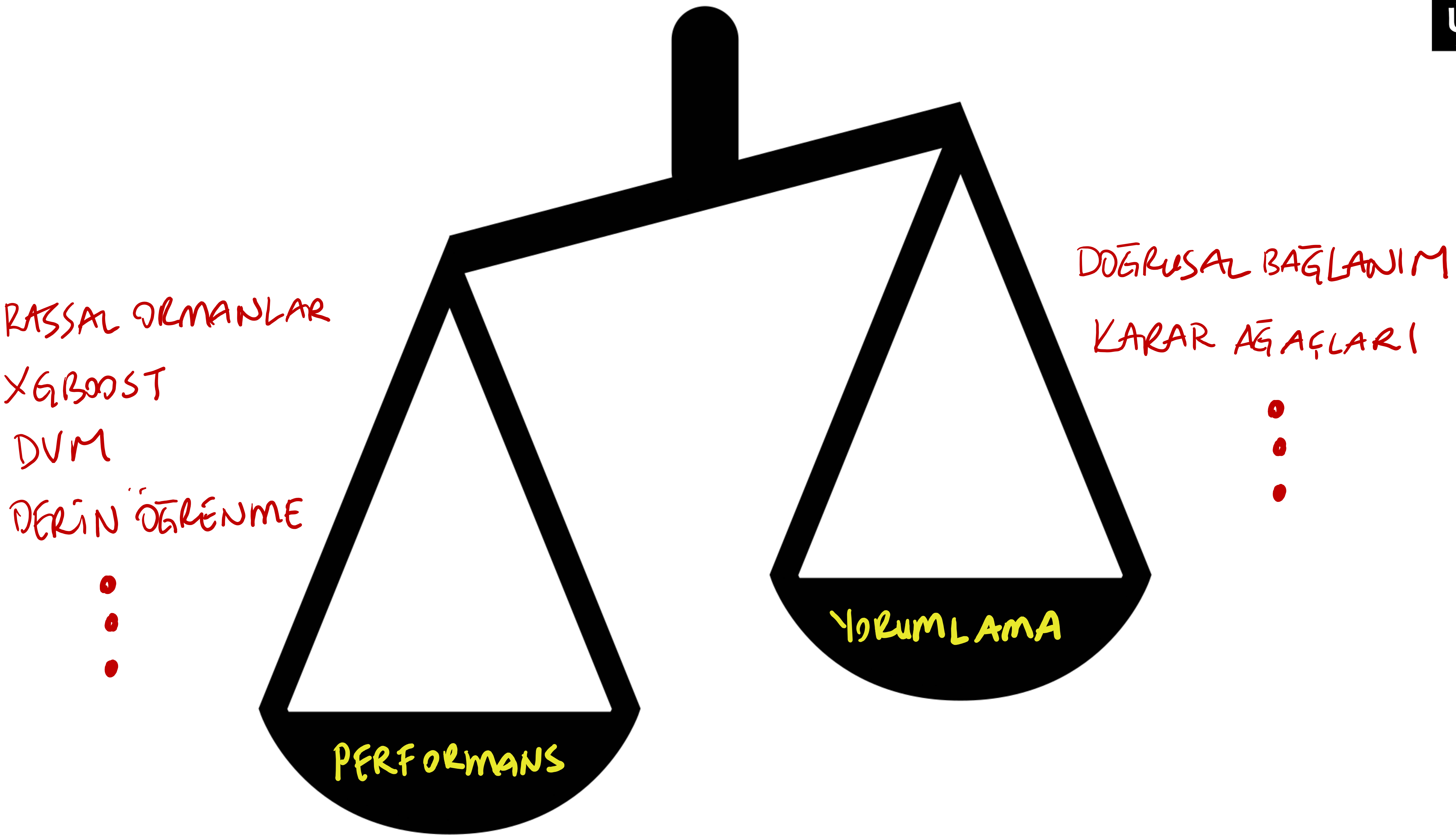
CGAP

BLOG 05 SEPTEMBER 2019

Algorithm Bias in Credit Scoring: What's Inside the Black Box?

By Maria Fernandez Vidal, Jacobo Menajovsky

> The responsible use of algorithms requires providers to know which variables are being considered in their credit scoring models and how they are affecting people's scores.


RASSAL ORMANLAR
XGBOOST
DVM
DERİN ÖĞRENME
DOĞRUSAL BAĞLANIM
KARAR AĞAÇLARI
PERFORMANS
YORUMLAMA

YORUMLAMA YAKLAŞIMLARI

LOKAL	ÖZEL	KENDILIĞINDEN
GLOBAL	GENEL*	DIŞARIDAN
LIME	OCT	BDR
SHAP	EBM	RUX
SLIM	OSDT	RUG
...	...	...

KARAR AĞACI TABANLI

KURAL TABANLI

* AGNOSTİK

LIME

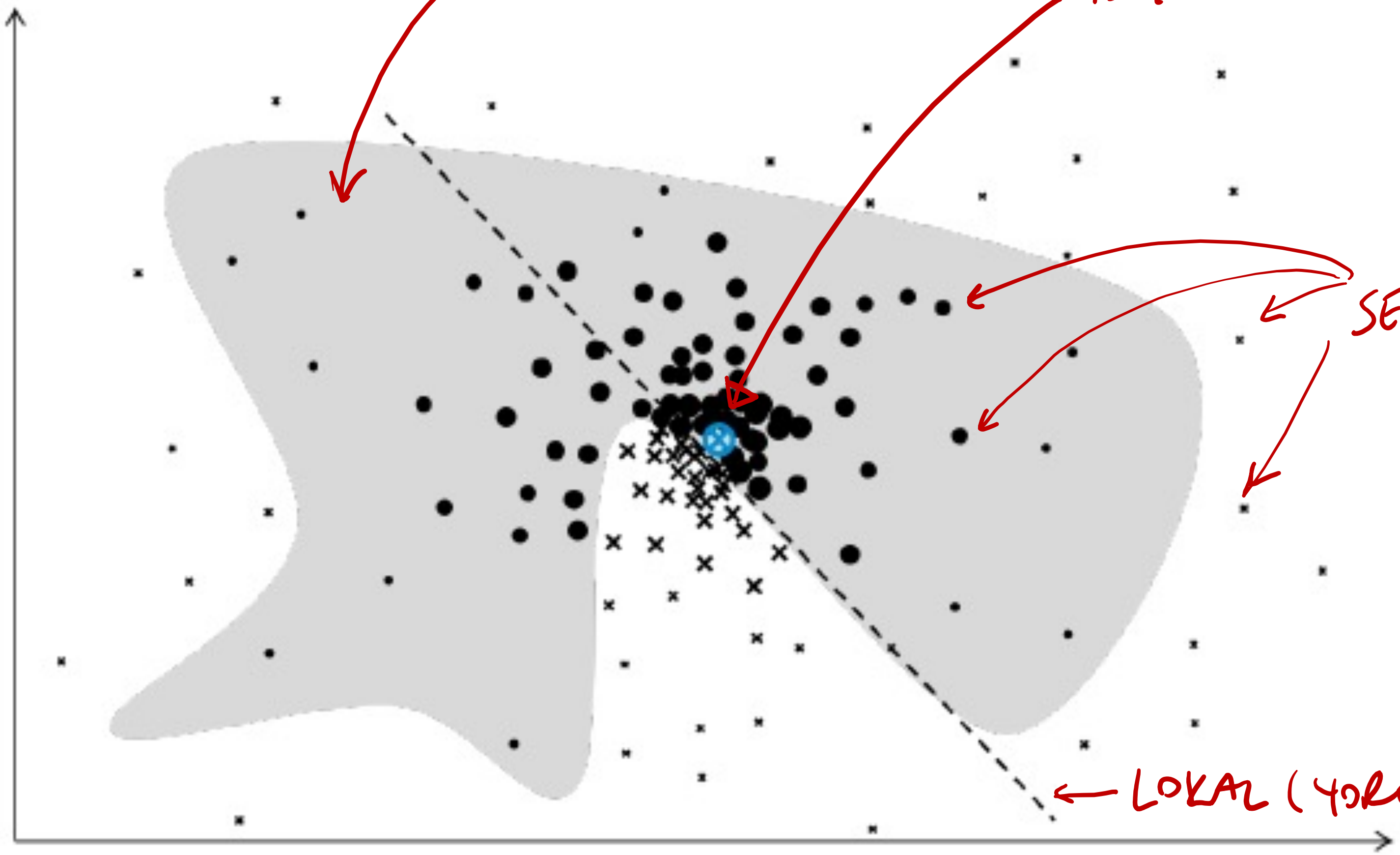
(RIBERIO VD., 2016)

LOCAL INTERPRETABLE MODEL-AGNOSTIC EXPLANATIONS

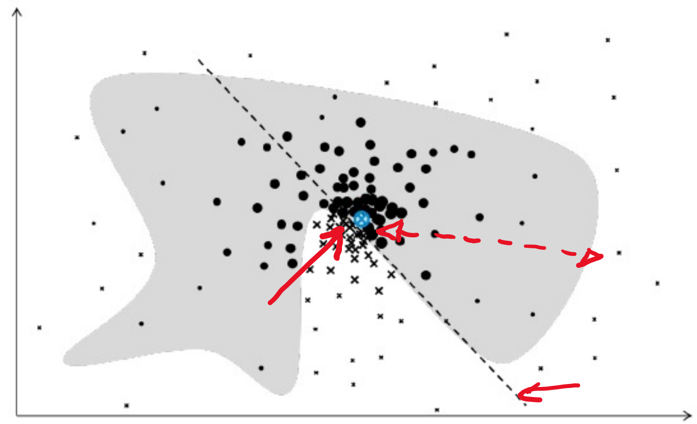
MODEL ETİKETLERİ

AĞIRLIKLAR

MESAFE İLE TERS ORANTILI



LIME



AÇIKLAMA

ÖĞRENME HATASI

UZAKLIK

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

İLGİLİ VERİ NOKTASI

LOKAL MODEL

LOKAL MODEL KARMAŞIKLIĞI

- ÖZNİTELİK SAYISI
- AĞAÇ DERİNLİĞİ

⋮

LIME

XGBoost model predicted an acceptance probability of 77.0%

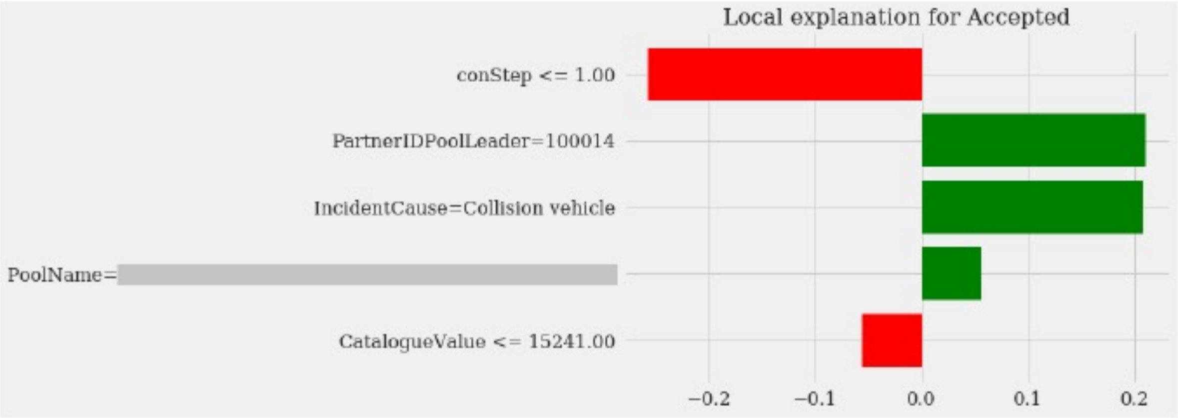


Figure 23: FP - LIME by Default (5 Features)
[Local Pred. = 0.58, Intercept = 0.42, R^2 = 0.28, RMSE = 0.26, Time = 4.67 s]

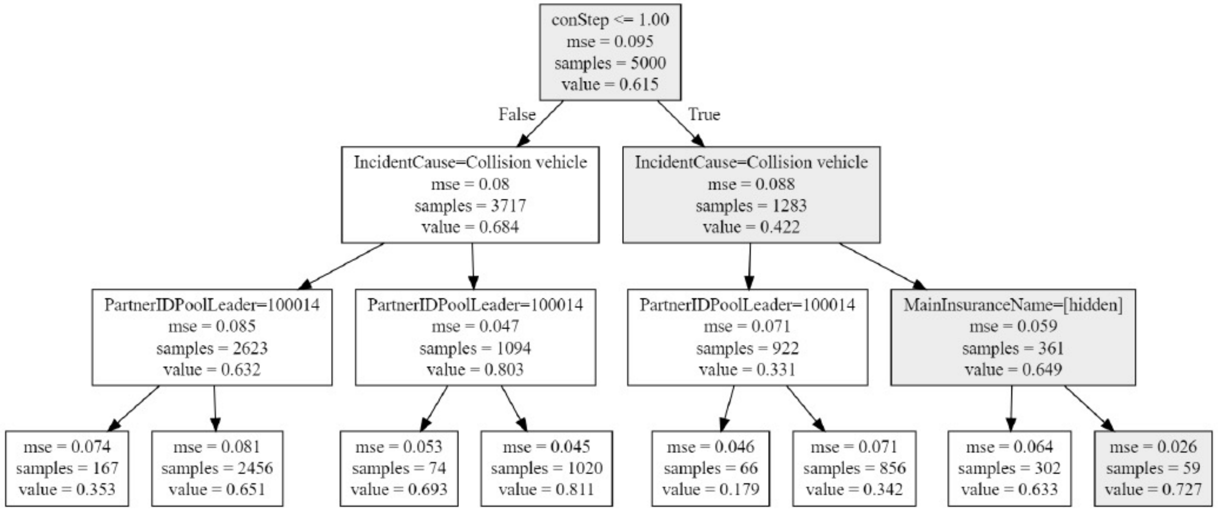


Figure 24: FP - LIME Decision Tree (3 Layers)
[Local Pred. = 0.73, R^2 = 0.28, RMSE = 0.26, Time = 2.15 s]

YORUMLAMA YAKLAŞIMLARI

LOKAL
GLOBAL

ÖZEL
GENEL*

KENDİLİĞİNDEN
DIŞARIDAN

LIME
SHAP
SLIM
⋮

OCT
EBM
OSDT
⋮

BDR
RUX
RUG
⋮

KARAR AĞACI TABANLI
KURAL TABANLI

*AGNOSTİK

SHAP

(LUNDBERG VE LEE, 2017)

SHAPLEY ADDITIVE EXPLANATIONS

AÇIKLAMA MODELİ → $g_i(\mathbf{z}') = \phi_0 + \sum_{j=1}^{M} \phi_{ij} z'_{ij}$

ÖZNİTELİK BU ÖRNEKTE VAR (1), YOK (0)

ÖRNEK

ÖZNİTELİK

SHAPLEY DEĞERİ → $\phi_{ij} = \sum_{S \subseteq X_i \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_i(S \cup \{j\}) - f_i(S)]$

j HARİÇ ÖZNİTELİKLER

MODEL TAHMİNİ

SHAP

$$\phi_{ij} = \sum_{S \subseteq X_i \setminus \{j\}} \frac{|S|!(M - |S| - 1)!}{M!} [f_i(S \cup \{j\}) - f_i(S)]$$

TÜM ALTKÜMELER

2^M

YAKLAŞIKLAMA YÖNTEMLERİ

KERNEL SHAP*

LINEAR SHAP

LOW-ORDER SHAP

TREE SHAP

DEEP SHAP

*AGNOSTİK

YORUMLAMA YAKLAŞIMLARI

LOKAL	ÖZEL	KENDİLİĞİNDEN
GLOBAL	GENEL*	DIŞARIDAN
LIME	OCT	BDR
SHAP	EBM	RUX
SLIM	OSDT	RUG
⋮	⋮	⋮

KARAR AĞACI TABANLI

KURAL TABANLI

* AGNOSTİK

EBM

(NORI VD., 2019)

EXPLAINABLE BOOSTING MACHINE

ÖRNEK i İÇİN TAHMİN

ÖZNİTELİKLER ARASI İLİŞKİLER

$$g(E[y_i]) = \beta_0 + \sum_{j=1}^{M} f_j(x_{ij}) + \sum_{j=1}^{M} \sum_{k \neq j} f_{jk}(x_{ij}, x_{ik})$$

GENELLEŞTİRİLMİŞ DOĞRUSAL MODEL

LİNK FONKSİYONU (ÖRN. LOGIT)

ÖZNİTELİK FONKSİYONLARI

TAKVİYE (BOOSTING) MODELLERİ

EBM

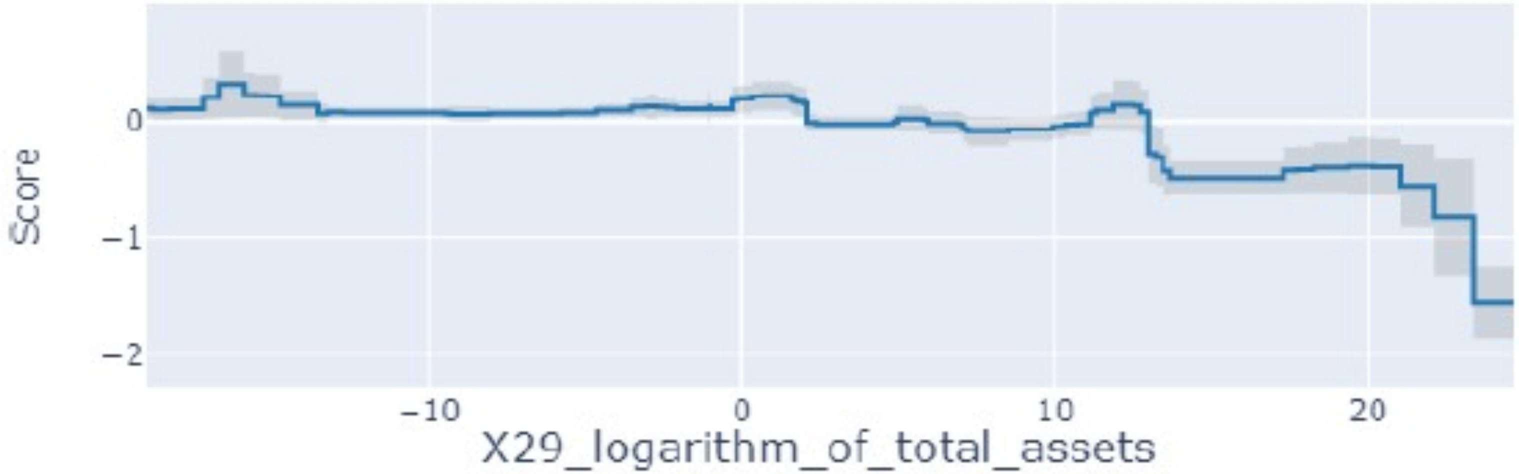


Figure 1: GAM plot of the effect of a feature log(total assets) on the prediction of bankruptcy for a data set of Polish firms. There is quite some variance in the relation, but overall higher assets clearly decrease the bankruptcy risk as would be expected.

YORUMLAMA YAKLAŞIMLARI

LOKAL
GLOBAL

ÖZEL
GENEL*

KENDILIĞINDEN
DIŞARIDAN

LIME
SHAP
SLIM

OCT
EBM
OSDT

BDR
RUX
RUG

KARAR AĞACI TABANLI
KURAL TABANLI

* AGNOSTIK

SLIM

(Üstün vd., 2013)

Supersparse Linear Integer Model

$$\min_{\lambda} \quad \overbrace{\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}[y_i x_i^T \lambda \leq 0]}^{\text{HATA}} + \underbrace{C_0||\lambda||_0 + C_1||\lambda||_1}_{\substack{\text{YORUMLANABİLİRLİK} \\ \text{(AZ SAYIDA ÖZNİTELİK)}}},$$

$$\text{s.t.} \quad \underbrace{\lambda \in \mathcal{L}}$$

$\{-1,1\}$

KESİKLİ DEĞERLER
(SKORLAR)

SLIM

PREDICT MUSHROOM IS POISONOUS IF SCORE > 3

1.	$spore\_print\_color = green$	4 points		$\cdots\cdots$
2.	$stalk\_surface\_above\_ring = grooves$	2 points	+	$\cdots\cdots$
3.	$population = clustered$	2 points	+	$\cdots\cdots$
4.	$gill\_size = broad$	-2 points	+	$\cdots\cdots$
5.	$odor \in \{none, almond, anise\}$	-4 points	+	$\cdots\cdots$
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	$\cdots\cdots$

Figure 2: The scoring system for mushroom edibility produced by SLIM as displayed in Ustun and Rudin (2015)

ANKET

14 VERİ ANALİZİ UZMANI

(11 KİŞİ YAPAY ÖĞRENME ALANINDA)

Table 9: Average scores of survey respondents on model specific questions as well as corresponding standard deviations (SD). Highest scores in bold.

	Familiarity (1-3)	SD	Test % Correct	Understanding (1-7)	SD	Stakeholders (1-7)	SD	Practical Value (1-7)	SD
Logit	**3.00**	0.00	78	5.21	1.25	4.57	1.50	5.36	0.93
SLIM	1.36	0.63	64	**5.79**	1.53	**5.71**	1.49	4.86	1.75
EBM	1.79	0.70	**86**	4.64	1.55	4.43	1.79	5.07	1.21
SHAP	2.36	0.63	71	5.43	1.09	5.00	1.24	**5.86**	0.77

XGBOOST +

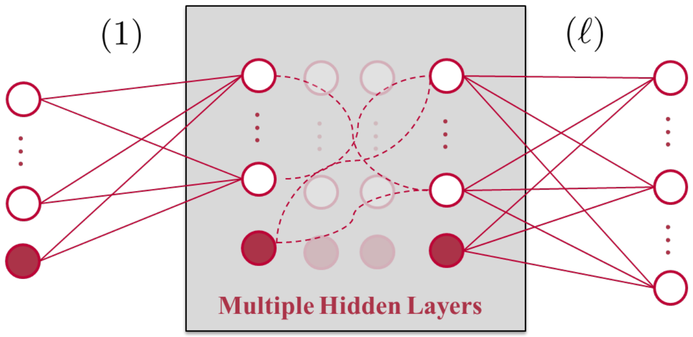
YÖNTEMLE İLGİLİ DOĞRU-YANLIŞ SORULARI

YÖNTEMİ ANLAMA (KİŞİSEL)

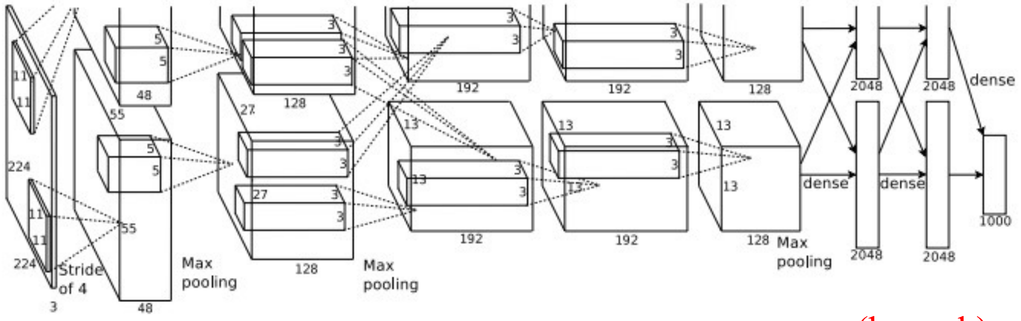
YÖNTEMİ ANLATMA

PRATİKTE KULLANIM ŞANSI

DERİN ÖĞRENME



$$\hat{y}_k(X,\beta) = \sigma\left(\sum_j \beta_{kj}^{(\ell)} h\left(\sum_s \beta_{js}^{(\ell-1)} h\left(\dots h\left(\sum_i \beta_{ji}^{(1)} X_i\right)\dots\right)\right)\right)$$



(kaynak)

LIME

SHAP

DİĞER ?

InterpretML: A Unified Framework for Machine Learning Interpretability

Harsha Nori HANORI@MICROSOFT.COM
Samuel Jenkins SAJENKIN@MICROSOFT.COM
Paul Koch PAULKOCH@MICROSOFT.COM
Rich Caruana RCARUANA@MICROSOFT.COM
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052, USA

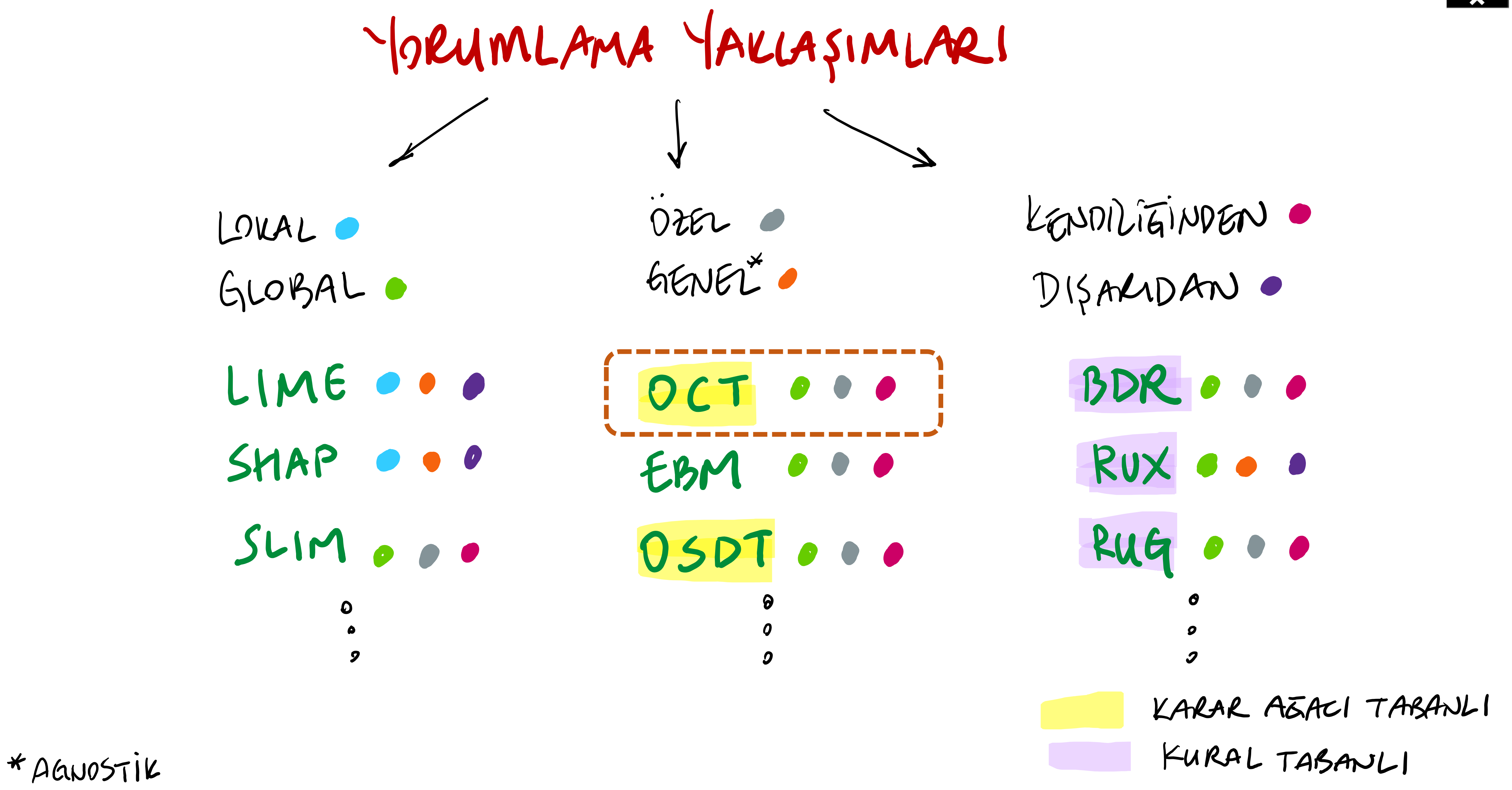
InterpretML

Understand Models.
Build Responsibly.

A toolkit to help understand models and enable responsible machine learning

Supported Techniques

Interpretability Technique	Type	Examples
Explainable Boosting	glassbox model	Notebooks
Decision Tree	glassbox model	Notebooks
Decision Rule List	glassbox model	Coming Soon
Linear/Logistic Regression	glassbox model	Notebooks
SHAP Kernel Explainer	blackbox explainer	Notebooks
SHAP Tree Explainer	blackbox explainer	Coming Soon
LIME	blackbox explainer	Notebooks
Morris Sensitivity Analysis	blackbox explainer	Notebooks
Partial Dependence	blackbox explainer	Notebooks


YORUMLAMA YAKLAŞIMLARI
LOKAL
GLOBAL
ÖZEL
GENEL*
KENDİLİĞİNDEN
DIŞARIDAN
LIME
SHAP
SLIM
OCT
EBM
OSDT
BDR
RUX
RUG
KARAR AĞACI TABANLI
KURAL TABANLI
* AGNOSTİK

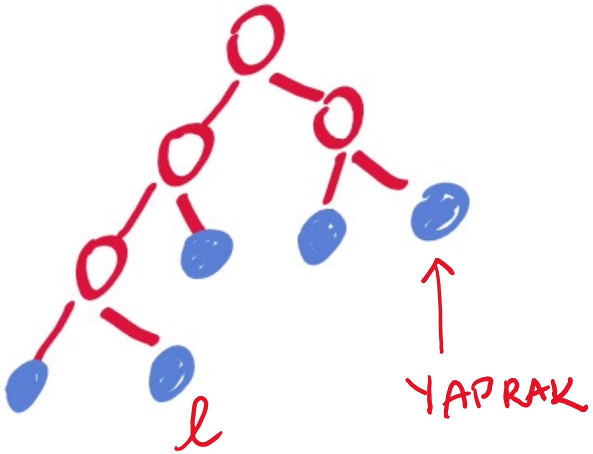
OCT

(BERTSIMAS VE DUNN, 2017)

OPTIMAL CLASSIFICATION TREES

(x_i, y_i)

GİRDİ — ETİKET



$$\min \quad \overbrace{R_{xy}(T)}^{\text{SINIFLANDIRMA HATASI}} + \alpha\overbrace{|T|}^{\text{AĞAÇ BÜYÜKLÜĞÜ (DÜĞÜM SAYISI)}}$$

$$\text{s.t.} \quad \underbrace{N_x(l)}_{\ell \text{ YAPRAĞINDAKİ ÖRNEK SAYISI}} \geq N_{\min}$$

OCT

Table 12 Comparison of CART and OCT-H across a range of depths, showing the number of datasets for which each method had the highest out-of-sample accuracy, and the mean improvement in out-of-sample accuracy when using OCT-H across all datasets along with the p value indicating the statistical significance of this difference

Max. depth	CART wins	OCT-H wins	Ties	Accuracy improvement (%)	p value
1	3	36	14	5.12	$\sim 10^{-16}$
2	10	32	11	4.88	$\sim 10^{-14}$
3	13	31	9	3.59	$\sim 10^{-12}$
4	13	29	11	3.12	$\sim 10^{-11}$

YORUMLAMA YAKLAŞIMLARI

LOKAL
GLOBAL

ÖZEL
GENEL*

KENDİLİĞİNDEN
DIŞARIDAN

LIME
SHAP
SLIM
⋮

OCT
EBM
OSDT
⋮

BDR
RUX
RUG
⋮

KARAR AĞACI TABANLI
KURAL TABANLI

*AGNOSTİK

OSDT

(HU VD., 2020)

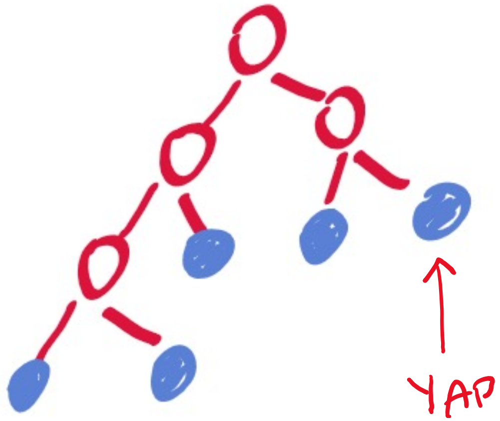
OPTIMAL SPARSE DECISION TREES

(x_i, y_i)

$y_i \in \{0,1\}$

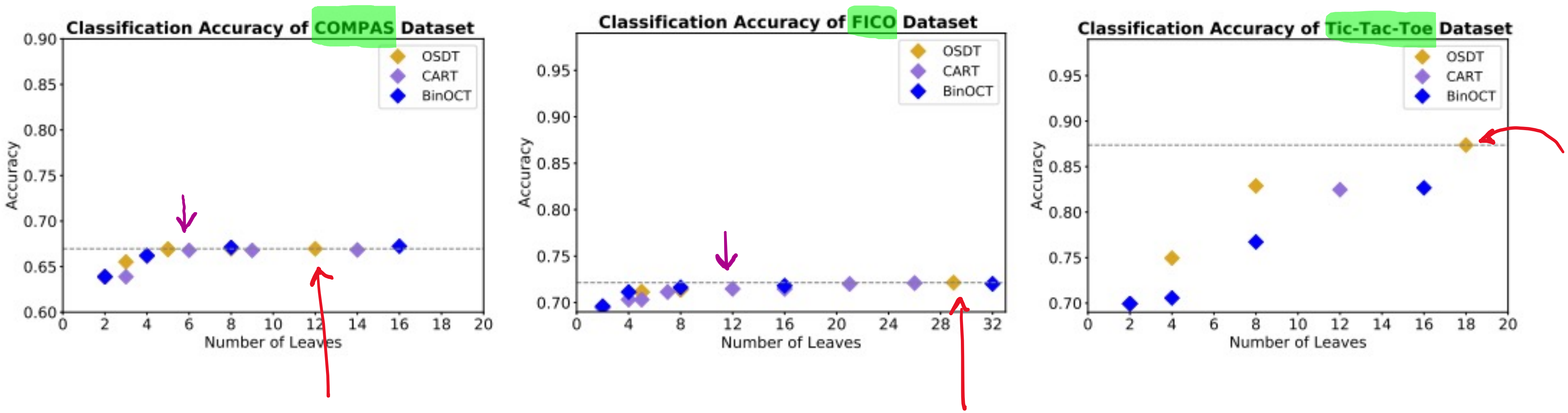
$$R(d, \mathbf{x}, \mathbf{y}) = \underbrace{\ell(d, \mathbf{x}, \mathbf{y})}_{\text{SINIFLANDIRMA HATASI}} + \lambda \underbrace{H_d}_{\text{AĞAÇTAKİ YAPRAK SAYISI}}$$

d: AĞAÇ



YAPRAK

OSDT



SÜRE LİMİTİ 30 DK.

OSDT

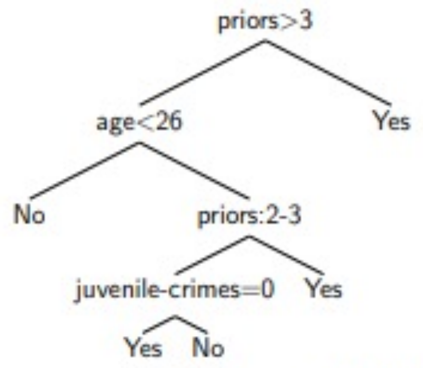


Figure 4: An optimal decision tree generated by OSDT on the COMPAS dataset. ($\lambda = 0.005$, accuracy: 66.90%)

YORUMLAMA YAKLAŞIMLARI

LOKAL

GLOBAL

ÖZEL

GENEL*

KENDILIĞINDEN

DIŞARIDAN

LIME

SHAP

SLIM

...

OCT

EBM

OSDT

...

BDR

RUX

RUG

...

KARAR AĞACI TABANLI

KURAL TABANLI

* AGNOSTIK

BDR

(DASH VD., 2018)

BOOLEAN DECISION RULES

(x_i, y_i)

$y_i \in \{0,1\}$

TOPLAM SEÇİLEN KURAL SAYISI

$$\text{minimize} \quad \sum_{i \in \mathcal{P}} \xi_i + \sum_{i \in \mathcal{Z}} \sum_{k \in \mathcal{K}_i} w_k$$

$$\text{subject to} \quad \xi_i + \sum_{k \in \mathcal{K}_i} w_k \geq 1, \quad \xi_i \geq 0, \quad i \in \mathcal{P}$$

$$\sum_{k \in \mathcal{K}} c_k w_k \leq C$$

$$w_k \in \{0,1\}, \qquad k \in \mathcal{K}.$$

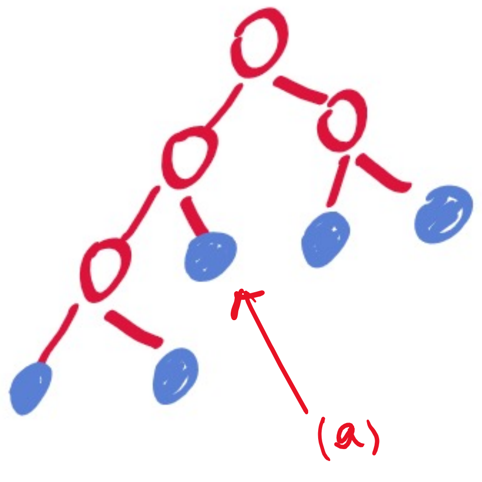
SINIFLANDIRMA HATASI

BDR

$$(\text{NumSatTrades} \geq 23) \wedge (\text{ExtRiskEstimate} \geq 70) \wedge (\text{NetFracRevolvBurden} \leq 63) \qquad (a)$$

OR

$$(\text{NumSatTrades} \leq 22) \wedge (\text{ExtRiskEstimate} \geq 76) \wedge (\text{NetFracRevolvBurden} \leq 78) \qquad (b)$$



YORUMLAMA YAKLAŞIMLARI

LOKAL
GLOBAL

ÖZEL
GENEL*

KENDİLİĞİNDEN
DIŞARIDAN

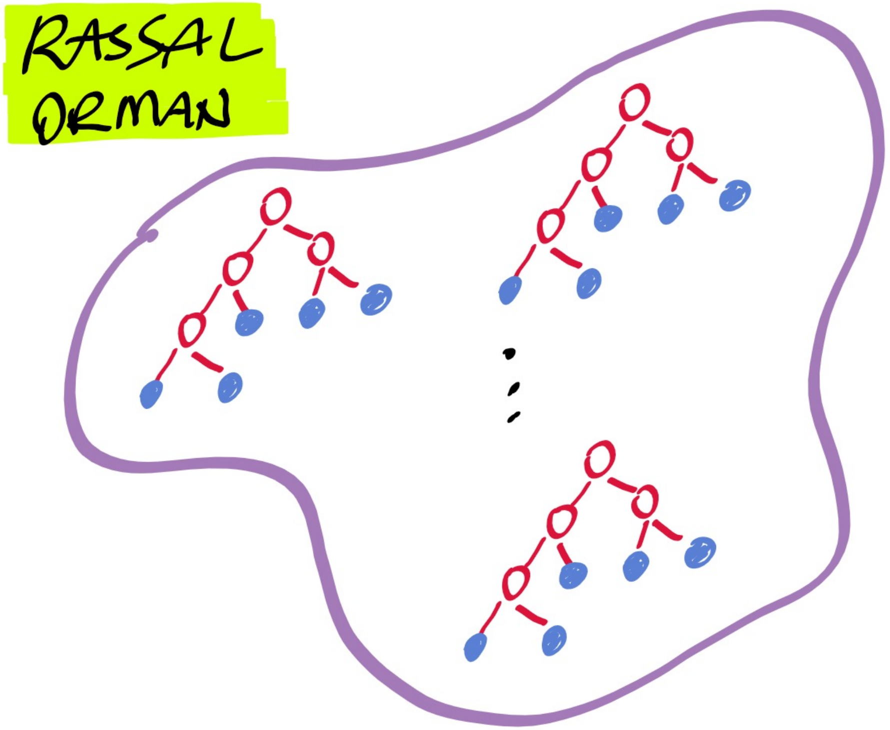
LIME
SHAP
SLIM
⋮

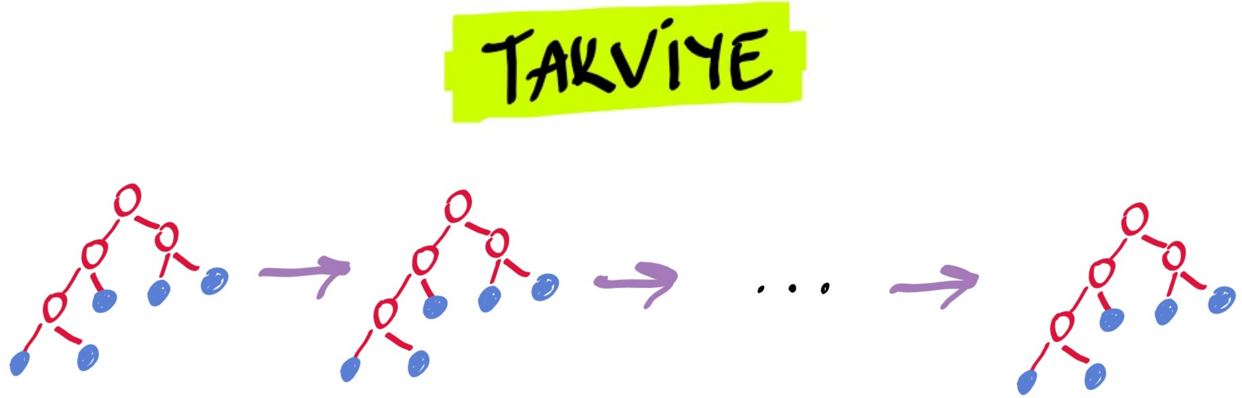
OCT
EBM
OSDT
⋮

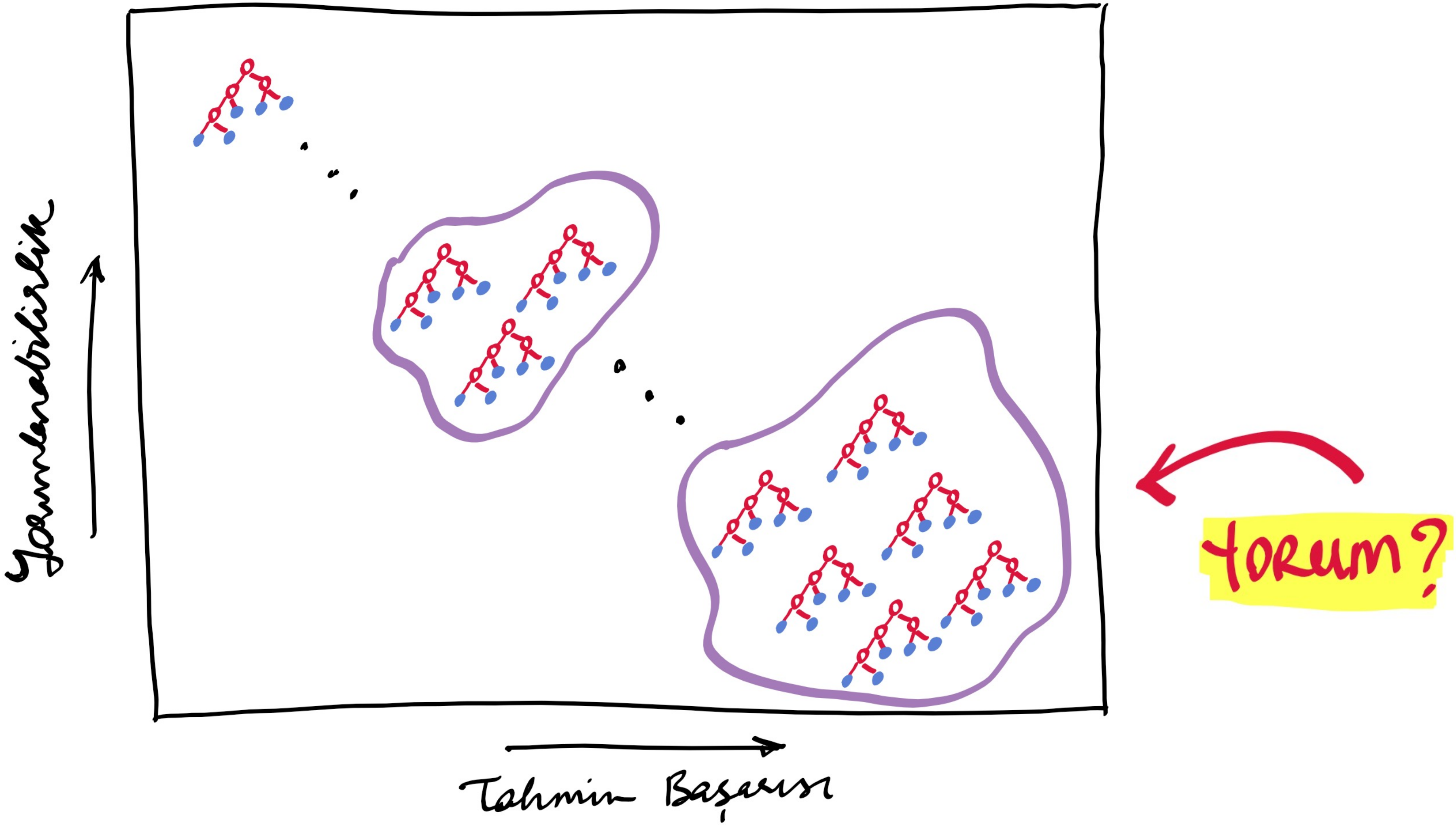
BDR
RUX
RUG
⋮

KARAR AĞACI TABANLI
KURAL TABANLI

* AGNOSTİK


RASSAL
ORMAN


TAKVİYE


Yorumlanabilirlik
Tahmin Başarısı
YORUM?

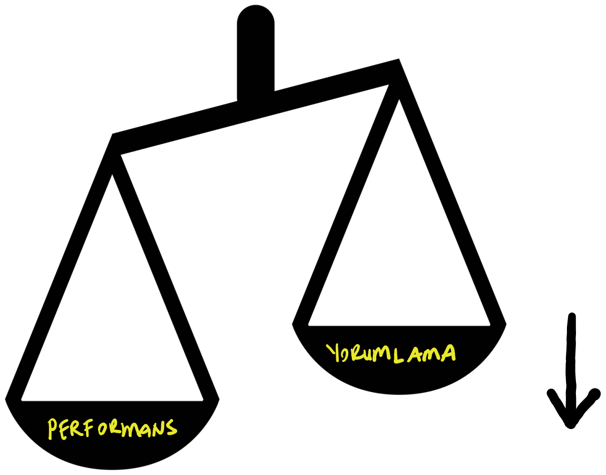
RUX

RUG

Discovering Classification Rules for Interpretable Learning with Linear Programming

Hakan Akyüz and Ş. İlker Birbil

https://github.com/sibirbil/RuleDiscovery



BASIT

HIZLI

KODLAMASI KOLAY

SEYREKLİK ✓
KURAL UZUNLUĞU ✓
SINIFLANDIRICI AĞIRLIKLARI ✓
YANLIŞ POZİTİF ETİKET SAYISI ✓
...

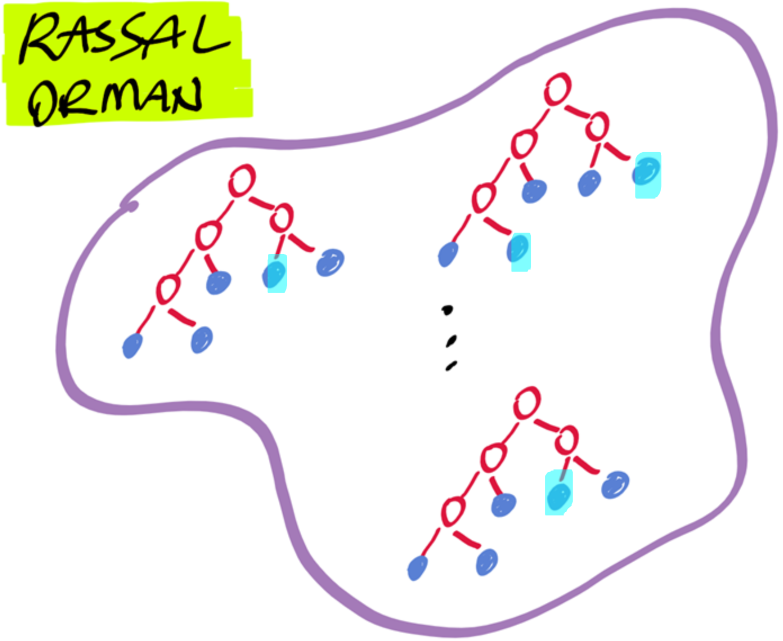
TOPLAM KAYIP

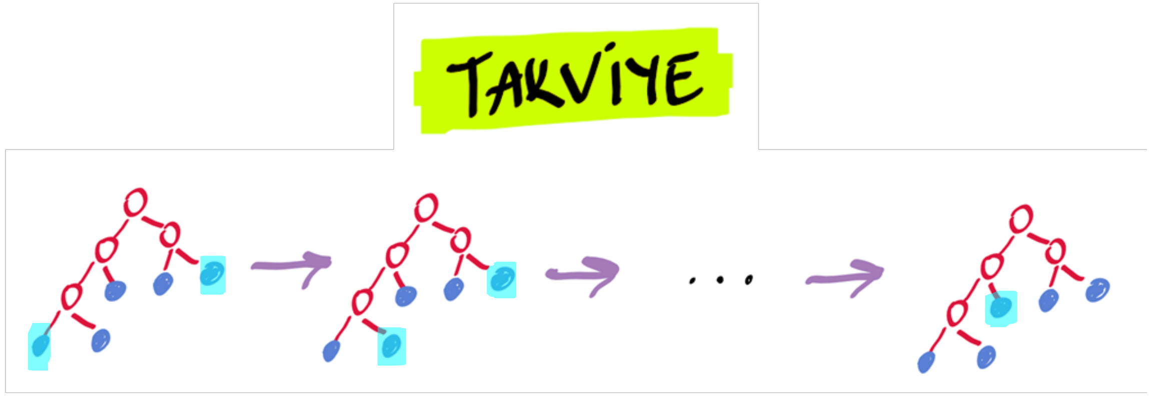
LP

$$
\begin{aligned}
\text{minimize} \quad & \sum_{i \in \mathcal{I}} v_i + \sum_{j \in \mathcal{J}} c_j w_j & \\
\text{subject to} \quad & \sum_{j \in \mathcal{J}} \hat{a}_{ij} w_j + v_i \geq 1, & i \in \mathcal{I}; \\
& \sum_{j \in \mathcal{J}} a_{ij} w_j \geq 1, & i \in \mathcal{I}; \\
& v_i \geq 0, & i \in \mathcal{I}; \\
& w_j \geq 0, & j \in \mathcal{J}
\end{aligned}
$$

KAPSAMA KISITLARI

P(J)

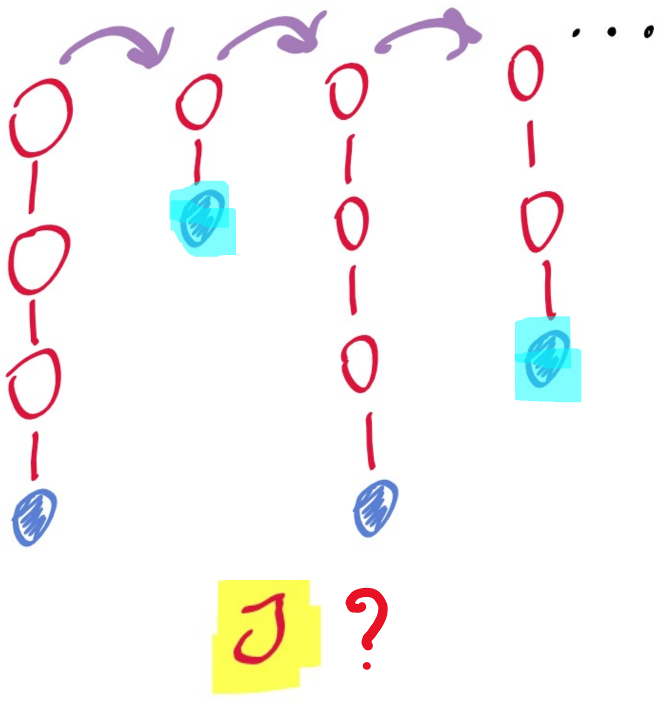




J ?

RULE EXTRACTION

RUX



Rule Generation

RUG

Tahmin Başarıları

Random Forest: 0.84
AdaBoost: 0.95
RUXRF: 0.90
RUXADA: 0.92
RUG: 0.92

Kural Sayıları

Random Forest: 614
AdaBoost: 742
RUXRF: 19
RUXADA: 20
RUG: 24

$$\text{minimize} \quad \sum_{i \in \mathcal{I}} v_i + \sum_{j \in \mathcal{J}} c_j w_j$$

$$\text{subject to} \quad \sum \quad \hat{a}_{ij} w_j + v_i \geq 1, \quad i \in \mathcal{I};$$

YORUMLAMA YAKLAŞIMLARI

LOKAL
GLOBAL

ÖZEL
GENEL*

KENDİLİĞİNDEN
DIŞARIDAN

LIME
SHAP
SLIM
...

OCT
EBM
OSDT
...

BDR
RUX
RUG
...

KARAR AĞACI TABANLI
KURAL TABANLI

* AGNOSTİK

Teşekkürler

- UvA
- D4C
- Bol Bilim
- Veri Defteri
- @sibirbil