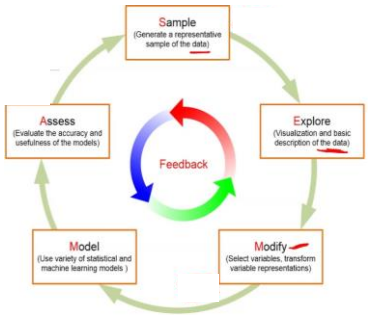
İleri Veri Madenciliği

Prof.Dr. Bülent Altunkaynak

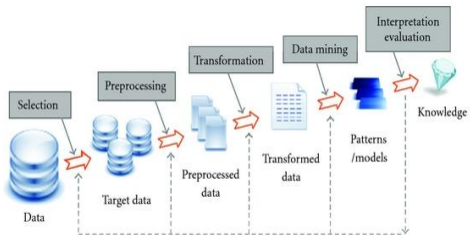
Veri Madenciliği Süreçleri



Şekil 1. CRISP-DM (Cross-Industry Standard Process for Data Mining) Süreci



Şekil 2. SEMMA (Sample, Explore, Modify, Model, and Access) Süreci



Şekil 3. KDD (Knowledge Discovery in Databases) Süreci

Okuyunuz:
https://www.datascience-pm.com/crisp-dm-2/
https://www.datascience-pm.com/semma/
https://www.datascience-pm.com/kdd-and-data-mining/



Sınıflandırma (Classification)

Bilgi Sistemi, Karar Tabloları/Sistemi, Denetimli Öğrenme, Sınıflandırma Yöntemleri, Entropi, Kazanım Değeri

Bilgi Sistemi / Karar Sistemi

S=<U,Q>

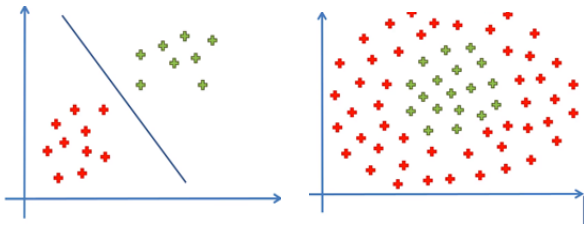
Gün	Hava	Sıcaklık	Nem	Rüzgar
1	Güneşli	Yüksek	Yüksek	Yok
2	Güneşli	Yüksek	Yüksek	Var
3	Bulutlu	Yüksek	Yüksek	Yok
4	Yağmurlu	Normal	Yüksek	Yok
5	Yağmurlu	Düşük	Normal	Yok
6	Yağmurlu	Düşük	Normal	Var
7	Bulutlu	Düşük	Normal	Var
8	Güneşli	Normal	Yüksek	Yok
9	Güneşli	Düşük	Normal	Yok
10	Yağmurlu	Normal	Normal	Yok

S=<U,Q∪D>

Gün	Hava	Sıcaklık	Nem	Rüzgar	Oyun
1	Güneşli	Yüksek	Yüksek	Yok	Hayır
2	Güneşli	Yüksek	Yüksek	Var	Hayır
3	Bulutlu	Yüksek	Yüksek	Yok	Evet
4	Yağmurlu	Normal	Yüksek	Yok	Evet
5	Yağmurlu	Düşük	Normal	Yok	Evet
6	Yağmurlu	Düşük	Normal	Var	Hayır
7	Bulutlu	Düşük	Normal	Var	Evet
8	Güneşli	Normal	Yüksek	Yok	Hayır
9	Güneşli	Düşük	Normal	Yok	Evet
10	Yağmurlu	Normal	Normal	Yok	Evet

Sınıflandırma

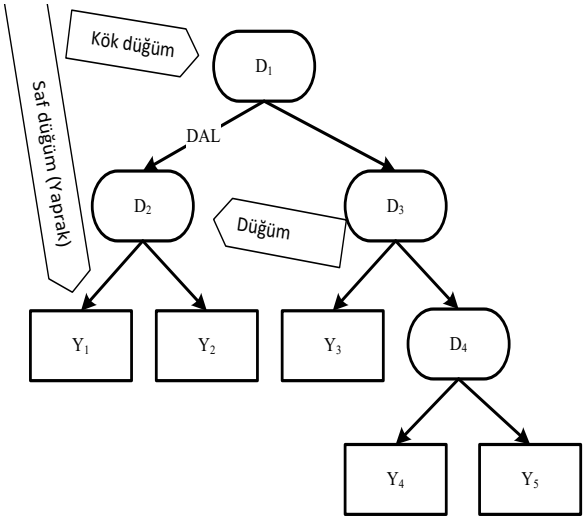
Karar Ağaçları					İstatistik Temelli	Kural Temelli		Mesafeye Dayalı	Optimizasyon
ID3	C4.5	CART	CHAID	QUEST	Bayes	ZeroR	OneR	K-en yakın komşu	SVM



Bir sınıflayıcı bir model veya bir fonksiyon olabilir.
$M(\mathbf{x})=f(\mathbf{x})=y_i \qquad y_i \in \{c_1,c_2,\ldots,c_k\}$

Bootstrap, Cross-validation, ROC

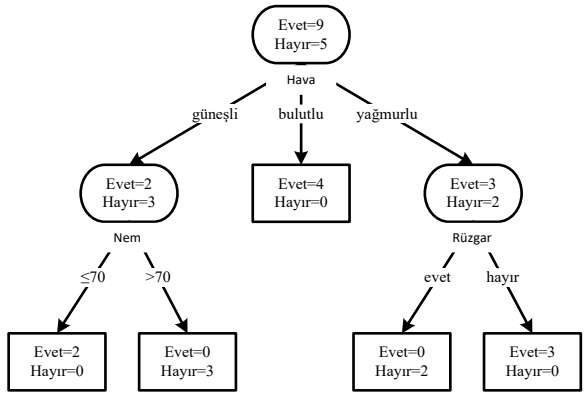
Karar Ağacı – Kavramsal Model



6

Karar Ağacı – Örnek gösterim

Gün	Hava	Sıcaklık	Nem	Rüzgar	Golf
1	güneşli	85	85	hayır	hayır
2	güneşli	80	90	evet	hayır
3	bulutlu	83	86	hayır	evet
4	yağmurlu	70	96	hayır	evet
5	yağmurlu	68	80	hayır	evet
6	yağmurlu	65	70	evet	hayır
7	bulutlu	64	65	evet	evet
8	güneşli	72	95	hayır	hayır
9	güneşli	69	70	hayır	evet
10	yağmurlu	75	80	hayır	evet
11	güneşli	75	70	evet	evet
12	bulutlu	72	90	evet	evet
13	bulutlu	81	75	hayır	evet
14	yağmurlu	71	91	evet	hayır



7

Karar ağacı algoritmalarında kullanılan bazı ölçüler

$$Gain(X_i) = H(Y) - \sum_{j=1}^{k_i} P(X_{ij}) H(Y / X_{ij})$$

$$H(Y) = H(p_1, p_2, ..., p_k) = \sum_{j=1}^{k} (p_j \log_b(1/p_j))$$

$$GainRatio(X_i) = \frac{Gain(X_i)}{H(X_i)}$$

$$\phi(D_i) = 2P_L P_R \sum_{j=1}^{k} \left[\left| P(Y_j / L) - P(Y_j / R) \right| \right]$$

$$Gini_{split}(X_i) = \sum_{j=1}^{k_i} P(X_{ij}) Gini(X_{ij})$$

$$Gini(X_{ij}) = 1 - \sum_{m=1}^{k} [P(Y_m / X_{ij})]^2$$

8