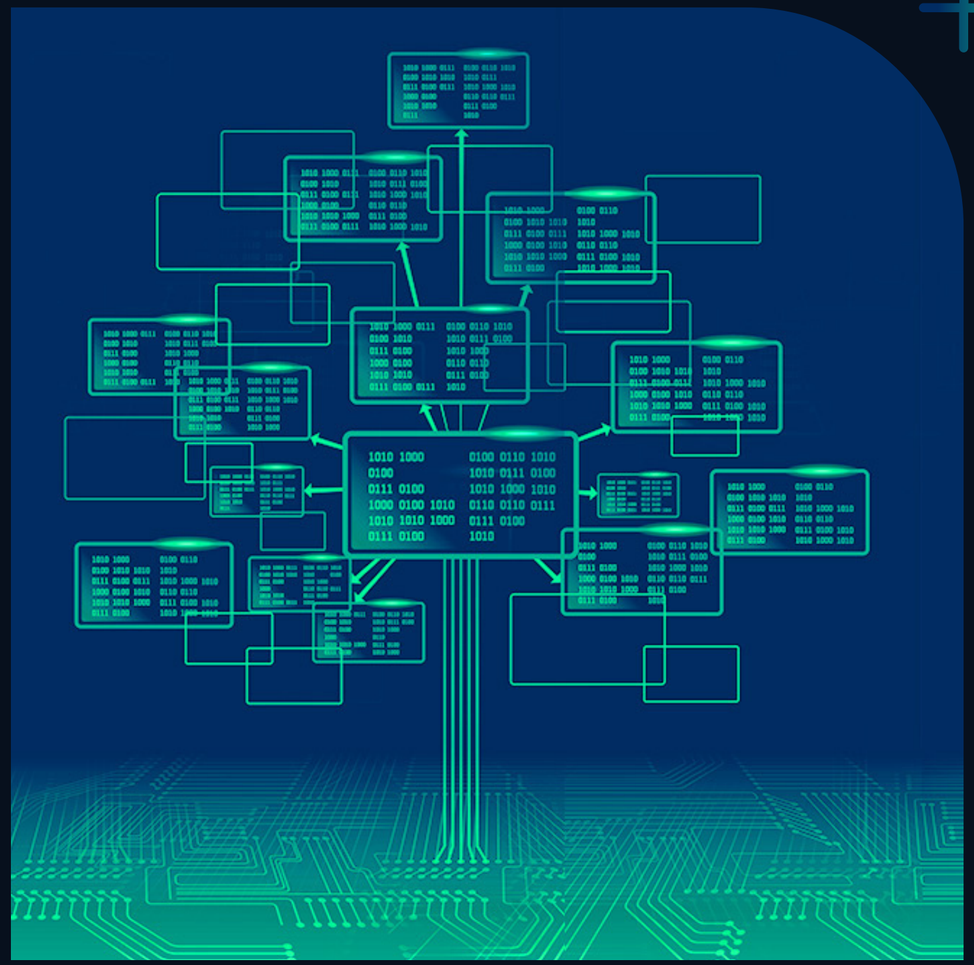
MAKİNE ÖĞRENMİNDE

KARAR AĞAÇLARI

Ayşe Nur Eliçora

NECMETTİN ERBAKAN ÜNİVERSİTESİ



Decision Trees

Karar Ağaçları Nedir?

Karar Ağacları ağaç tabanlı bir sınıflandırma modeli oluşturur. Verileri gruplara ayırır veya bağımsız (predictor) değişkenlerin değerlerine dayalı olarak bağımlı (target) bir değişkenin değerlerini tahmin eder.

Hem sınıflandırma hem de regresyon amacıyla kullanılabilecek denetimli öğrenme algoritmaları sınıfına aittir.

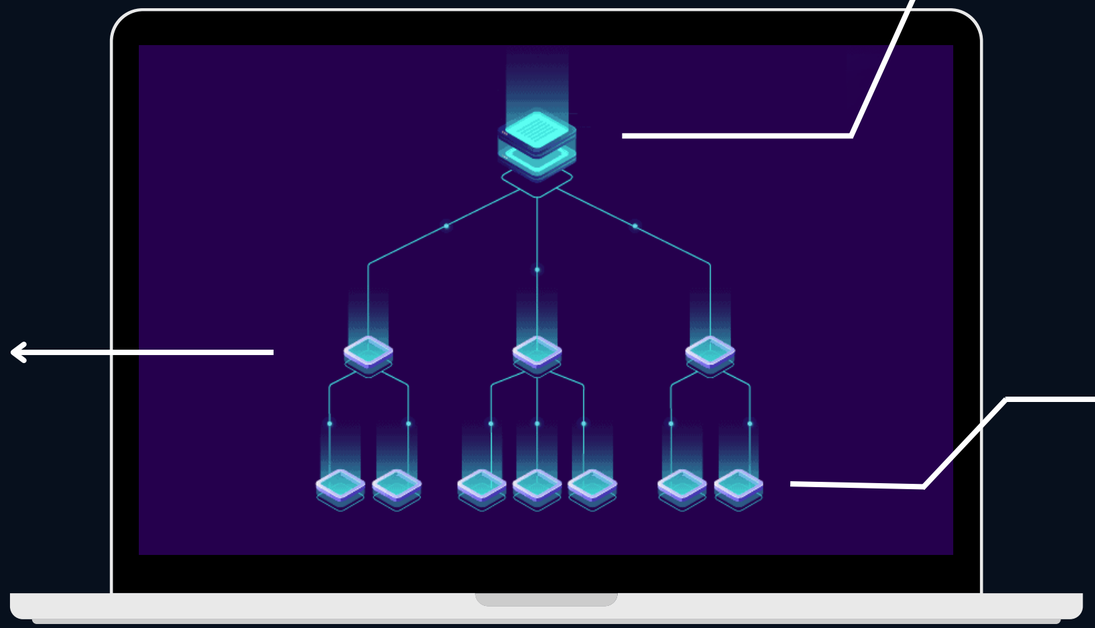
Erişim adresi : https://www.ibm.com/docs/en/spss-statistics/26.0.0?topic=trees-creating-decision

NECMETTİN ERBAKAN ÜNİVERSİTESİ

NASIL ÇALIŞIR?

Kayıt,ağacın kök düğümünden başlar ve ara düğümlerden hangi yöne dallanacağı belirlenir. Her bir sınıf ağaçta tek yaprak olarak gösterilir.

Dallanma işlemi yaprak düğüme ulaşıncaya kadar devam eder.



KÖK (ROOT NODE)

Tüm kararların başlangıçta başladığı yoldaki ilk düğüm. Üst düğümü ve 2 alt düğümü yoktur.

DÜĞÜM(INTERNAL NODE)

1 ana düğümü olan ve alt düğümlere ayrılan düğümler.

YAPRAK(LEAF NODE)

1 ebeveyni olan ancak daha fazla bölünmeyen düğümler (terminal düğümleri olarak da bilinir). Tahmini üreten düğümlerdir.

Erişim adresi : https://www.jcchouinard.com/decision-trees-in-machine-learning/

NECMETTİN ERBAKAN ÜNİVERSİTESİ

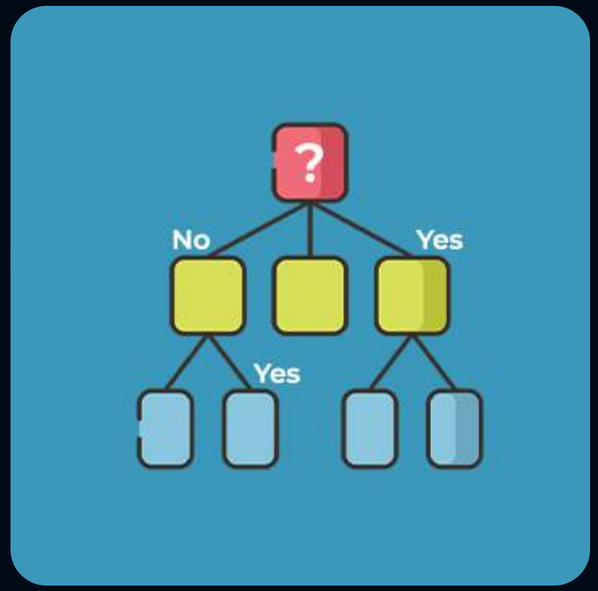
KARAR AĞACI TÜRLERİ

1.1.**CART**

1.2.**ID3**

1.3.**C4.5**

Erişim adresi :https://www.ibm.com/topics/decision-trees#anchor--719457280



1.1. CART

Classification and Regression Trees

CART algoritması, **Gini'nin safsızlık indeksi** temelinde bir karar ağacı oluşturmak için gerekli olan bir **sınıflandırma algoritması türüdür** . Temel bir makine öğrenimi algoritmasıdır ve çok çeşitli kullanım durumları sağlar. Leo Breiman adlı bir istatistikçi, sınıflandırma veya regresyon öngörücü modelleme sorunları için kullanılabilecek Karar Ağacı algoritmalarını tanımlamak için bu ifadeyi türetti.

Erişim adresi : https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm

1.1.1. Gini İndeksi

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Gini İndeksi Hesaplaması

Gini İndeksi veya Gini Kirliliği, her bir sınıfın olasılıklarının karelerinin toplamının birden çıkarılmasıyla hesaplanır. Çoğunlukla daha büyük bölümleri tercih eder ve uygulanması çok basittir. Basit bir ifadeyle, yanlış sınıflandırılmış rastgele seçilmiş belirli bir özelliğin olasılığını hesaplar.

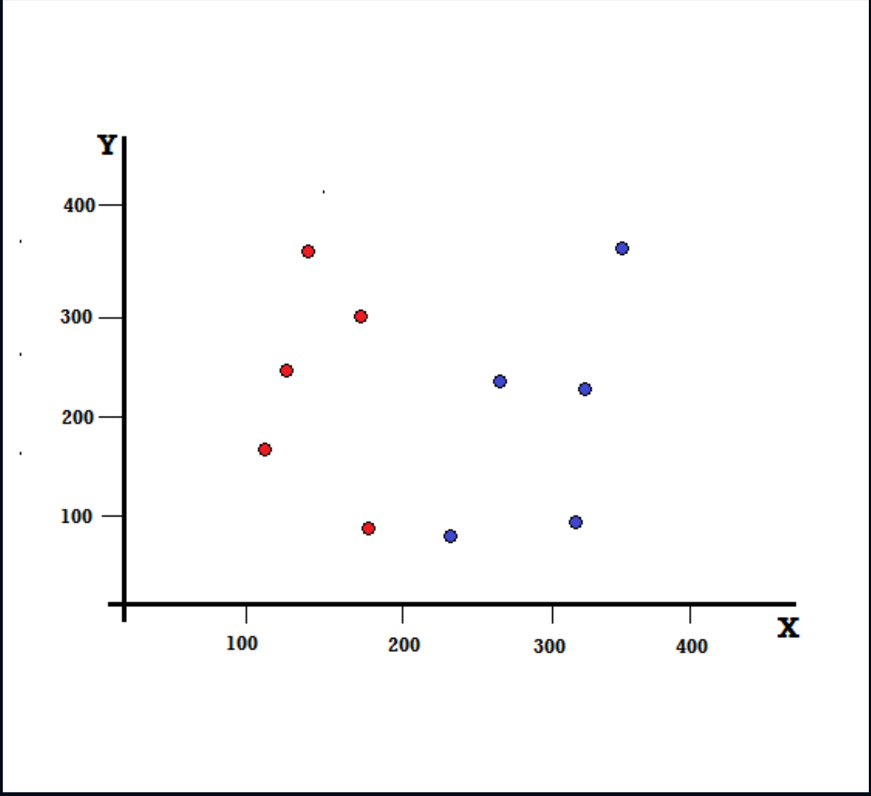
Gini İndeksi 0 ile 1 arasında değişir; burada 0, sınıflandırmanın saflığını temsil eder ve 1, çeşitli sınıflar arasında elementlerin rastgele dağılımını gösterir. 0,5'lik bir Gini İndeksi, bazı sınıflar arasında öğelerin eşit dağılımı olduğunu gösterir.

Erişim adresi : https://www.upgrad.com/blog/gini-index-for-decision-trees/#:~:text=The%20Gini%20Index%20or%20Gini,feature%20that%20was%20classified%20incorrectly.

1.1.1.1. Gini İndeksi Hesaplaması

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Gini İndeksi Hesaplaması



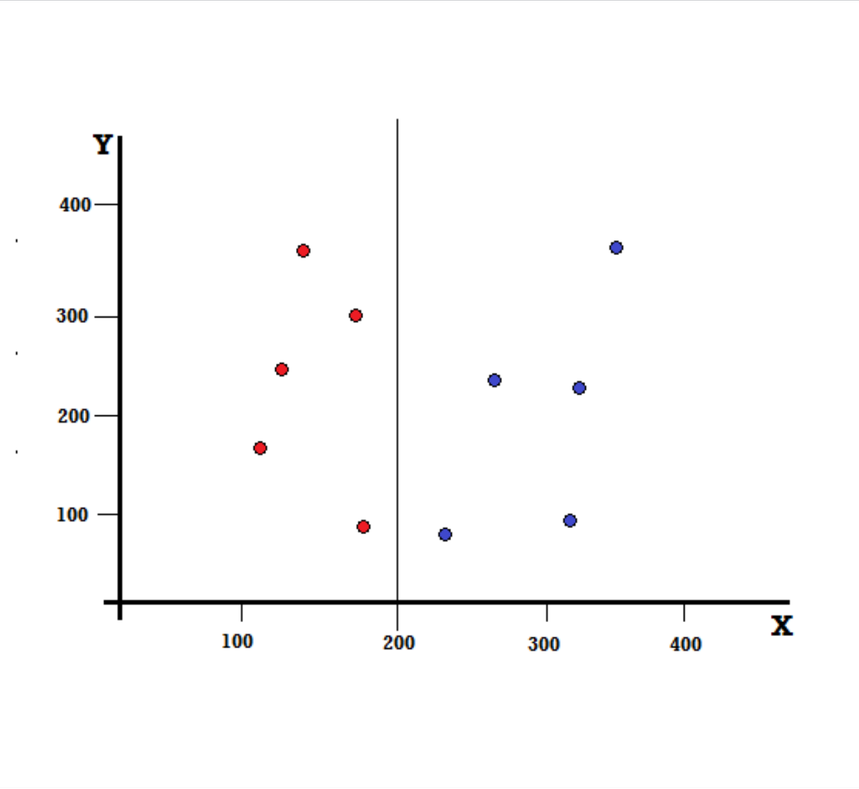
Burada, kırmızılar ve maviler olmak üzere iki değişkenli toplam 10 veri noktamız var. X ve Y eksenleri, her terim arasında 100'lük boşluklarla numaralandırılmıştır. Verilen örnekten Gini İndeksini ve Gini Kazancını hesaplayacağız.

Erişim adresi : https://www.upgrad.com/blog/gini-index-for-decision-trees/#:~:text=The%20Gini%20Index%20or%20Gini,feature%20that%20was%20classified%20incorrectly.

1.1.1.1. Gini İndeksi Hesaplaması

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Gini İndeksi Hesaplaması



Bir karar ağacı için veri setini iki dala ayırmamız gerekir. X-Y düzleminde 5 Kırmızı ve 5 Mavi ile işaretlenmiş aşağıdaki veri noktalarını göz önünde bulundurun. X=200'de ikili bir bölme yaptığımızı varsayalım, o zaman aşağıda gösterildiği gibi mükemmel bir bölünme elde edeceğiz.

Sol dalda sadece kırmızılar vardır ve bu nedenle Gini Kirliliği,
G(sol) =1 ∗ (1-1) + 0 ∗ (1-0) = 0

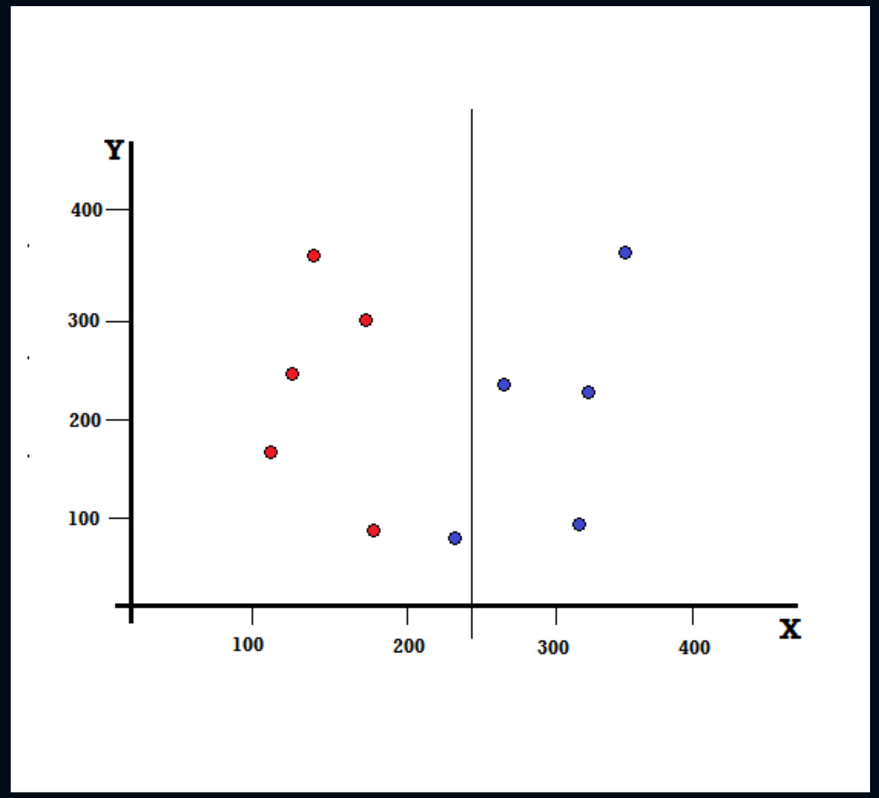
Sağ taraftaki maviler için,
G(sağ) =1 ∗ (1-1) + 0 ∗ (1-0) = 0

Erişim adresi : https://www.upgrad.com/blog/gini-index-for-decision-trees/#:~:text=The%20Gini%20Index%20or%20Gini,feature%20that%20was%20classified%20incorrectly.

1.1.1.1. Gini İndeksi Hesaplaması

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

Gini İndeksi Hesaplaması



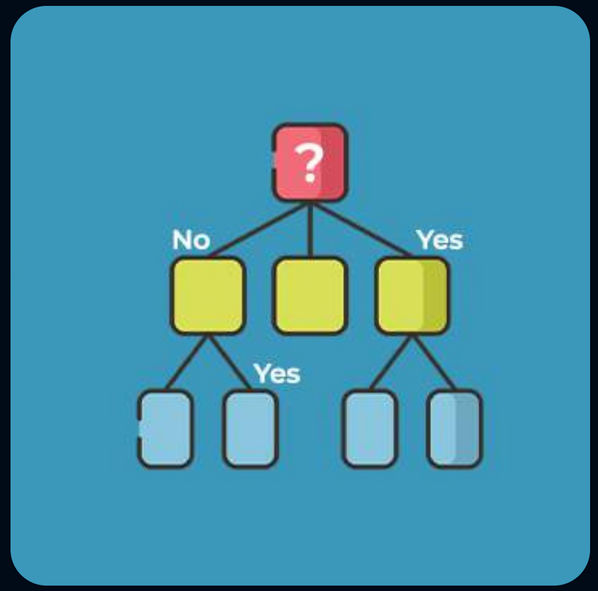
Bölmeyi X=250'de yaparsak sonuç ne olur?

Kusurlu bir bölünme olarak adlandırılır. Karar Ağacı modelini eğitirken, bölünmenin kusurluluğunu ölçmek için Gini İndeksini kullanabiliriz.

Gini Kirliliği,
G(sol) =1/6 ∗ (1-1/6) + 5/6 ∗ (1-5/6) = 0,278

G(sağ) =1 ∗ (1-1) + 0 ∗ (1-0) = 0

Erişim adresi : https://www.upgrad.com/blog/gini-index-for-decision-trees/#:~:text=The%20Gini%20Index%20or%20Gini,feature%20that%20was%20classified%20incorrectly.



1.2. **ID3**

Iterative Dichotomiser 3

Bu algoritma, aday bölünmeleri değerlendirmek için metrik olarak **entropi** ve **bilgi kazancı**ndan yararlanır.

Erişim adresi : https://www.ibm.com/topics/decision-trees#anchor--719457280

1.2.1. Entropi

Entropy and Information Gain

Entropi, örnek değerlerin saflığını ölçen bilgi teorisinden kaynaklanan bir kavramdır.Aşağıdaki formülle tanımlanır;

S, entropinin hesaplandığı veri kümesini temsil eder

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

c, S kümesindeki sınıfları temsil eder.

p(c), c sınıfına ait veri noktalarının, S kümesindeki toplam veri noktalarının sayısına oranını temsil eder.

Erişim adresi : https://www.ibm.com/topics/decision-trees#anchor--719457280

1.2.2.Bilgi Kazanımı

Information Gain

Bilgi kazancı entropinin azalmasıdır. Bilgi kazancı, verilen öznitelik değerlerine dayalı olarak, bölünmeden önceki entropi ile veri kümesinin bölünmesinden sonraki ortalama entropi arasındaki farkı hesaplar.

$$\text{Information Gain}(S,a) = \text{Entropy}(S) - \sum_{v \in \text{values}(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

a belirli bir özelliği veya sınıfı temsil eder.

|Sv|/ |S| Sv'deki değerlerin veri kümesindeki S değerlerine oranını temsil eder

Erişim adresi : https://www.ibm.com/topics/decision-trees#anchor--719457280

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Bu veri kümesi için entropi 0.94'tür. Bu, “Tennis”in “Yes” olduğu günlerin oranını 9/14 ve “Tennis”in “No” olduğu günlerin oranını 5/14 olarak bularak hesaplanabilir. Daha sonra bu değerler entropi formülüne eklenebilir.

Entropy (Tennis) = -(9/14) log2(9/14) – (5/14) log2 (5/14) = 0.94

Erişim adresi : https://www.ibm.com/topics/decision-trees#anchor--719457280

Day	Outlook	Temp	Humidity	Wind	Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Weak	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cold	Normal	Weak	Yes
D10	Rain	Mild	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Daha sonra özniteliklerin her biri için bilgi kazancını ayrı ayrı hesaplayabiliriz. Örneğin, "Humadity" özniteliği için bilgi kazancı şu şekilde olacaktır:

Gain (Tennis, Humidity) = (0.94)-(7/14)* (0.985) – (7/14)*(0.592) = 0.151

- 0.985, Humidity = “high” olduğunda entropi değeri

- 0,59 Humidity = “normal” olduğunda entropi değeri

Erişim adresi : https://www.ibm.com/topics/decision-trees#anchor--719457280

NECMETTİN ERBAKAN ÜNİVERSİTESİ

BENİ DİNLEDİĞİNİZ İÇİN TEŞEKKÜRLER