Öğrenme Modellerinde Özellik Seçimi: Online News Popularity Veri Kümesi Üzerinde Karşılaştırmalı Analiz

Muhammed Kayra Bulut - 25501805
Makine Öğrenmesi Dersi 2. Ödevi

***Özet*—Bu çalışmada, Online News Popularity veri kümesi üzerinde üç farklı özellik seçimi yönteminin ikili sınıflandırma performansına etkisi incelenmiştir. Filtreleme (Pearson Korelasyonu), Sarmalayıcı (RFE + Lojistik Regresyon) ve Gömülü (Random Forest) yöntemleri kullanılarak en önemli 15 özellik belirlenmiştir. Lojistik Regresyon sınıflandırıcısı ile 5-fold cross validation yöntemi uygulanarak modeller değerlendirilmiştir. Deneysel sonuçlar, tüm özelliklerin kullanıldığı modelin %65.52 doğruluk oranı ile en yüksek performansı sağladığını, ancak özellik seçimi yöntemlerinin hesaplama maliyetini önemli ölçüde azalttığını göstermektedir. Sarmalayıcı yöntem (RFE), 15 özellik ile %65.10 doğruluk oranına ulaşarak özellik seçimi yöntemleri arasında en başarılı sonucu vermiştir.**

***Index Terms*—Özellik seçimi, Lojistik regresyon, Pearson korelasyonu, RFE, Random Forest, Online News Popularity, Makine öğrenmesi**

I. Giriş

Günümüzde çevrimiçi haber platformları, milyonlarca makalenin paylaşıldığı ve tüketildiği dijital ortamlardır. Bir haberin popüler olup olmayacağını tahmin edebilmek, içerik üreticileri ve medya kuruluşları için stratejik bir öneme sahiptir. Bu bağlamda makine öğrenmesi yöntemleri, çeşitli özellikler kullanarak haber popülerliğini tahmin etmekte yaygın olarak kullanılmaktadır [1].

Özellik seçimi, makine öğrenmesi modellerinin performansını artırmak, eğitim süresini kısaltmak ve aşırı öğrenmeyi (overfitting) önlemek için kritik bir ön işleme adımıdır [2]. Yüksek boyutlu veri kümelerinde, gereksiz veya gürültülü özellikler modelin genelleme yeteneğini olumsuz etkileyebilmektedir. Bu nedenle, en ayırt edici özelliklerin seçilmesi model performansı açısından büyük önem taşımaktadır.

Bu çalışmanın amacı, UCI Machine Learning Repository'den alınan Online News Popularity veri kümesi üzerinde farklı özellik seçimi yöntemlerinin sınıflandırma performansına etkisini karşılaştırmalı olarak analiz etmektir. Çalışma kapsamında Filtreleme (Pearson Korelasyonu), Sarmalayıcı (RFE + Lojistik Regresyon) ve Gömülü (Random Forest Feature Importance) yöntemleri uygulanarak en önemli 15 özellik belirlenmiş ve bu özelliklerle eğitilen modellerin performansları karşılaştırılmıştır.

II. Yöntem ve Veri Kümesi

A. Veri Kümesi

Bu çalışmada Mashable platformundan derlenen Online News Popularity veri kümesi kullanılmıştır [1]. Veri kümesi, 39.644 haber makalesine ait 61 özellik içermektedir. Her bir makale için içerik özellikleri (kelime sayısı, görsel sayısı vb.), zaman özellikleri (yayın günü, hafta sonu durumu), anahtar kelime metrikleri, doğal dil işleme özellikleri (LDA konu modeli, sentiment analizi) ve referans bilgileri (bağlantı sayısı, paylaşım istatistikleri) bulunmaktadır.

B. Veri Ön İşleme

Veri ön işleme adımları aşağıdaki şekilde gerçekleştirilmiştir:

1) **Gereksiz özellik çıkarımı:** Ayırt edici bilgi içermeyen `url` ve `timedelta` özellikleri veri kümesinden çıkarılmıştır.
2) **Hedef değişken dönüşümü:** `shares` sütunu medyan değeri olan 1.400 eşik alınarak ikili sınıf etiketine dönüştürülmüştür. $shares \geq 1400$ durumunda $y = 1$ (popüler), aksi halde $y = 0$ (popüler değil) olarak etiketlenmiştir.
3) **Veri bölme:** İşlenmiş veri kümesi %80 eğitim ve %20 test olarak ayrılmıştır.
4) **Cross validation:** Eğitim verisi üzerinde 5-fold stratified cross validation yöntemi uygulanmıştır.

İşleme sonrasında toplam 59 özellik elde edilmiştir.

C. Özellik Seçimi Yöntemleri

1) Filtreleme Yöntemi (Pearson Korelasyonu): Filtreleme yönteminde, her bir özellik ile hedef değişken arasındaki Pearson korelasyon katsayısı hesaplanmıştır [3]. Pearson korelasyon katsayısı, iki değişken arasındaki doğrusal ilişkinin gücünü ve yönünü ölçer. Bu yöntemde mutlak değeri en yüksek korelasyona sahip 15 özellik seçilmiştir. Filtreleme yöntemi model bağımsız olması ve hesaplama açısından verimli olması ile avantaj sağlamaktadır.

2) Sarmalayıcı Yöntem (RFE + Lojistik Regresyon): Sarmalayıcı yöntemde Recursive Feature Elimination (RFE) algoritması Lojistik Regresyon sınıflandırıcısı ile birlikte kullanılmıştır [4]. RFE, başlangıçta tüm özelliklerle model eğitir ve her iterasyonda model katsayılarına göre en az önemli özelliği elemine eder. Bu süreç istenilen özellik sayısına (15) ulaşılana kadar devam eder. Model parametreleri olarak `solver='lbfgs'` ve `max_iter=1000` kullanılmıştır.

3) Gömülü Yöntem (Random Forest Feature Importance): Gömülü yöntemde Random Forest sınıflandırıcısının özellik önem skorları kullanılmıştır [5]. Random Forest, birden fazla karar ağacı eğiterek her özelliğin ağaçlardaki bölünmelere katkısını hesaplar. Gini safsızlığı (impurity) azalmasına göre önem skorları belirlenir. En yüksek önem skoruna sahip 15 özellik seçilmiştir. Model 100 ağaç (`n_estimators=100`) ile eğitilmiştir.

D. Sınıflandırıcı Model

Performans karşılaştırması için Lojistik Regresyon sınıflandırıcısı kullanılmıştır. Model parametreleri varsayılan değerlerde tutulmuş (`C=1.0`, `solver='lbfgs'`) ve 5-fold stratified cross validation ile eğitilmiştir. Aşırı öğrenme tespiti için eğitim ve validasyon skorları karşılaştırılmıştır.

E. Performans Metrikleri

Model performansı aşağıdaki metriklerle değerlendirilmiştir:

- **Doğruluk (Accuracy):** Doğru tahmin edilen örneklerin toplam örnek sayısına oranı.
- **F1-Skoru:** Kesinlik ve duyarlılığın harmonik ortalaması.
- **Eğitim Süresi:** Modelin eğitim için harcadığı süre (saniye).

III. Deneysel Analiz

A. Özellik Seçimi Sonuçları

Her bir özellik seçimi yönteminin belirlediği en önemli 15 özellik Tablo I'de verilmiştir. Her yöntem farklı özellik kümeleri seçmiş olup, yalnızca `kw_avg_avg` özelliği üç yöntemde de ortak olarak seçilmiştir.

Tablo I: Farklı Yöntemlerle Seçilen Özellikler

Sıra	Filtreleme	Sarmalayıcı	Gömülü
1	LDA_02	n_non_stop_words	kw_avg_avg
2	kw_avg_avg	n_non_stop_uniq.	kw_max_avg
3	data_ch_is_world	n_unique_tokens	LDA_02
4	is_weekend	kw_avg_avg	self_ref_min_sh.
5	data_ch_is_enter.	kw_max_avg	kw_avg_min
6	data_ch_is_socmed	data_ch_is_tech	kw_avg_max
7	weekday_is_sat.	is_weekend	LDA_01
8	data_ch_is_tech	LDA_00	self_ref_avg_sh.
9	LDA_04	data_ch_is_socmed	LDA_04
10	kw_min_avg	kw_min_min	LDA_00
11	num_hrefs	kw_avg_max	n_unique_tokens
12	weekday_is_sunday	kw_min_avg	global_subject.
13	LDA_01	kw_avg_min	n_non_stop_uniq.
14	global_sent_polar.	self_ref_avg_sh.	avg_token_length
15	num_keywords	kw_max_min	LDA_03

B. Ortak Özellik Analizi

Yöntemler arasındaki özellik kesişimi analiz edildiğinde, Filtreleme ve Sarmalayıcı arasında 5, Filtreleme ve Gömülü arasında 4, Sarmalayıcı ve Gömülü arasında 8 ortak özellik bulunmaktadır. Tablo II'da ortak bulunan özellikler gösterilmiştir.

Tablo II: Yöntemler Arasında Ortak Özellikler (En Az 2 Yöntemde Seçilenler)

Özellik	Filt.	Sarm.	Göm.
Üç Yöntemde Ortak (1 özellik)			
kw_avg_avg	✓	✓	✓
Filtreleme + Gömülü (3 özellik)			
LDA_02	✓		✓
LDA_04	✓		✓
LDA_01	✓		✓
Filtreleme + Sarmalayıcı (4 özellik)			
is_weekend	✓	✓	
data_channel_is_tech	✓	✓	
data_channel_is_socmed	✓	✓	
kw_min_avg	✓	✓	
Sarmalayıcı + Gömülü (7 özellik)			
LDA_00		✓	✓
kw_max_avg		✓	✓
n_unique_tokens		✓	✓
n_non_stop_unique_tokens		✓	✓
kw_avg_max		✓	✓
kw_avg_min		✓	✓
self_reference_avg_sharess		✓	✓

C. Model Performans Karşılaştırması

Tüm özellikler ve özellik seçimi yöntemleriyle elde edilen alt kümeler üzerinde Lojistik Regresyon modeli eğitilmiş ve test sonuçları Tablo III'de özetlenmiştir.

Tablo III: Lojistik Regresyon Performans Sonuçları

Yöntem	Öz. Sayısı	Accuracy	F1-Skor	Süre (s)
Tüm Özellikler	58	**0.6552**	**0.6544**	0.182
Filtreleme	15	0.6431	0.6413	**0.017**
Sarmalayıcı	15	**0.6510**	**0.6502**	0.035
Gömülü	15	0.6278	0.6253	0.022

Sonuçlara göre:

- Tüm özelliklerle (58 özellik) eğitilen model %65.52 doğruluk oranı ile en yüksek performansı sağlamıştır.
- Özellik seçimi yöntemleri arasında **Sarmalayıcı (RFE)** yöntemi %65.10 doğruluk oranı ile en başarılı sonucu vermiştir.
- Filtreleme yöntemi %64.31, Gömülü yöntem ise %62.78 doğruluk oranlarına ulaşmıştır.
- Eğitim süresi açısından özellik seçimi yöntemleri önemli avantaj sağlamıştır. Filtreleme yöntemi 0.017 saniye ile en hızlı eğitim süresine sahiptir.

D. Aşırı Öğrenme Analizi

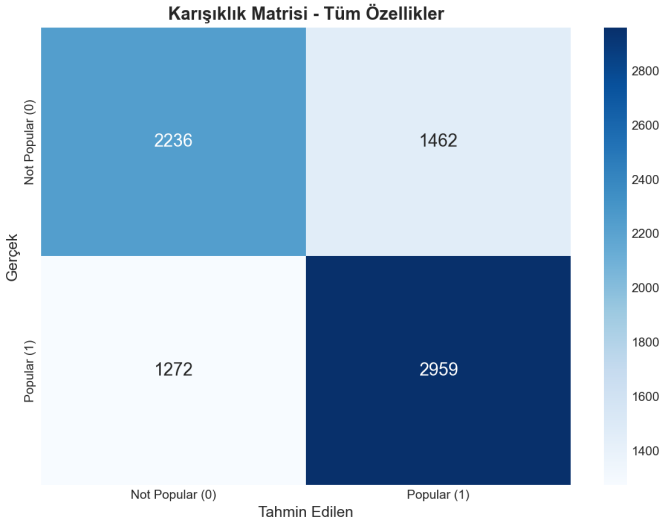
Tüm yöntemler için eğitim ve validasyon skorları arasındaki fark hesaplanmıştır:

- Tüm Özellikler: Fark = 0.0023 (Aşırı öğrenme tespit edilmedi)
- Filtreleme: Fark = 0 (Aşırı öğrenme tespit edilmedi)
- Sarmalayıcı: Fark = 0.0006 (Aşırı öğrenme tespit edilmedi)
- Gömülü: Fark = 0.0014 (Aşırı öğrenme tespit edilmedi)

Hiçbir yöntemde aşırı öğrenme tespit edilmemiştir. Bu durum, Lojistik Regresyonun L2 regularization özelliği sayesinde modelin genelleme yeteneğini koruduğunu göstermektedir.

E. Karışıklık Matrisi

En başarılı yöntem olan Tüm Özellikler için karışıklık matrisi Şekil 1'de verilmiştir.



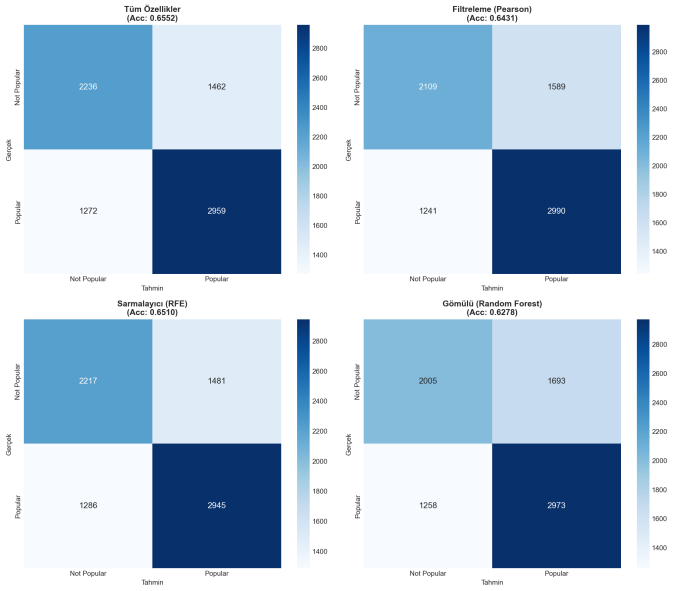
Şekil 1: Tüm Özellikler Yöntemi İçin Karışıklık Matrisi

Karışıklık matrisine göre:

- **True Negative (TN):** 2236 örnek doğru olarak "popüler değil" olarak tahmin edilmiştir.
- **False Positive (FP):** 1462 örnek yanlışlıkla "popüler" olarak tahmin edilmiştir.
- **False Negative (FN):** 1272 örnek yanlışlıkla "popüler değil" olarak tahmin edilmiştir.
- **True Positive (TP):** 2959 örnek doğru olarak "popüler" olarak tahmin edilmiştir.

F. Yöntem Karşılaştırma Grafiği

Tüm yöntemlerin karışıklık matrisleri Şekil 2'da karşılaştırmalı olarak gösterilmiştir.

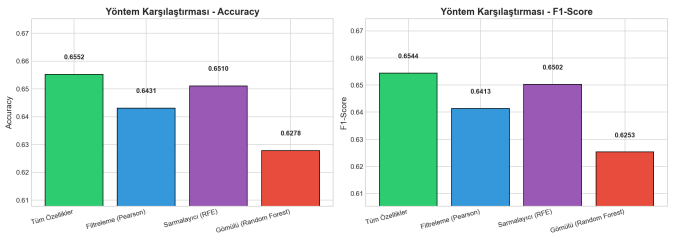


Şekil 2: Tüm Yöntemler İçin Karışıklık Matrisleri

Karışıklık matrisleri incelendiğinde yöntemler arasında belirgin farklılıklar gözlemlenmektedir:

- **Tüm Özellikler:** En dengeli dağılıma sahiptir. True Positive (TP) oranı en yüksek (2959) olup, modelin popüler haberleri tespit etme kapasitesi en iyidir. Ancak False Positive (FP) değeri de görece yüksektir (1462), bu da bazı popüler olmayan haberlerin yanlış sınıflandırıldığını gösterir.
- **Filtreleme (Pearson):** Popüler haberleri tespit etmede (TP) makul performans gösterirken, True Negative (TN) değeri düşüktür. Bu durum, yöntemin doğrusal korelasyona dayalı olmasından kaynaklanmakta ve karmaşık ilişkileri yakalayamamaktadır.
- **Sarmalayıcı (RFE):** Tüm özellikler yöntemine en yakın sonuçları vermektedir. TP ve TN değerleri dengeli olup, özellik seçimi yöntemleri arasında en başarılı performansı sergilemektedir. RFE'nin Lojistik Regresyon ile optimize edilmiş olması bu başarıyı açıklamaktadır.
- **Gömülü (Random Forest):** En düşük TP değerine sahiptir, bu da popüler haberleri tespit etmede zorluk yaşadığını gösterir. Random Forest'ın özellik önem skorları, Lojistik Regresyon sınıflandırıcısı için optimal olmayan bir özellik kümesi seçmiş olabilir.

Ayrıca yöntemlerin performans karşılaştırması Şekil 3'de grafik olarak sunulmuştur.



Şekil 3: Yöntem Performans Karşılaştırması

Performans karşılaştırma grafiği analiz edildiğinde aşağıdaki çıkarımlar yapılabilir:

- **Doğruluk-Verimlilik Dengesi:** Tüm özellikler en yüksek doğruluğu sağlarken, eğitim süresi de en uzundur (0.182s). Özellik seçimi yöntemleri bu süreyi yaklaşık 5-10 kat kısaltmaktadır.
- **Yöntemler Arası Performans Farkı:** Tüm özellikler (%65.52) ile Sarmalayıcı (%65.10) arasındaki fark sadece %0.42'dir. Bu, 15 özellik ile 58 özelliğe çok yakın performans elde edilebileceğini göstermektedir.
- **F1-Skor Tutarlılığı:** Accuracy ve F1-skor değerleri tüm yöntemlerde birbirine yakındır, bu da sınıf dengesizliğinin ciddi bir problem olmadığını göstermektedir.
- **En Hızlı Yöntem:** Filtreleme yöntemi 0.017s ile en kısa eğitim süresine sahiptir, ancak doğruluk açısından Sarmalayıcı'nın gerisinde kalmaktadır.
- **Gömülü Yöntem Performansı:** Random Forest tabanlı özellik seçimi, Lojistik Regresyon sınıflandırıcısı ile en düşük performansı vermiştir. Bu durum, ağaç tabanlı önem skorlarının doğrusal modeller için optimize olmadığını göstermektedir.

IV. Sonuç

Bu çalışmada, Online News Popularity veri kümesi üzerinde üç farklı özellik seçimi yönteminin sınıflandırma performansına etkisi incelenmiştir. Deneysel sonuçlar, tüm özelliklerin kullanılmasının en yüksek doğruluk oranını (%65.52) sağladığını, ancak özellik sayısının 15'e düşürülmesiyle bile rekabetçi sonuçlar elde edilebileceğini göstermektedir. Özellik seçimi yöntemleri arasında Sarmalayıcı (RFE) yöntemi %65.10 doğruluk oranı ile en başarılı performansı sergilemiştir. Bu sonuç, RFE yönteminin Lojistik Regresyon için optimize edilmiş özellik seçimi yapmasından kaynaklanmaktadır. Anahtar kelime metrikleri (`kw_avg_avg`) tüm yöntemlerde en önemli özellik olarak belirlenmiş olup, haberlerin popülerliğinde anahtar kelimelerin kritik rol oynadığını ortaya koymaktadır. Özellik seçimi, eğitim süresini yaklaşık 10 kat kısaltarak hesaplama verimliliği sağlamıştır. Gelecek çalışmalarda, farklı sınıflandırıcı algoritmalarının (Random Forest, SVM, Gradient Boosting vb.) özellik seçimi yöntemleriyle birlikte değerlendirilmesi ve en iyi özellik sayısının belirlenmesi için ek analizler yapılabilir.

Kaynaklar

[1] K. Fernandes, P. Vinagre, and P. Cortez, "A proactive intelligent decision support system for predicting the popularity of online news," in *Portuguese Conference on Artificial Intelligence*, Springer, 2015, pp. 535–546.

[2] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, pp. 1157–1182, 2003.

[3] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, University of Waikato, 1999.

[4] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, vol. 46, no. 1-3, pp. 389–422, 2002.

[5] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.