Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı, S. 215-220, Kasım 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue, pp. 215-220, November 2020

Research Article

Filtre Tabanlı Nitelik Seçimi ve Topluluk Öğrenme Yaklaşımlarıyla Borsa İstanbul Enerji Endeksi Yön Tahmini

Hakan Gündüz[1]

[1] Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Yazılım Mühendisliği Bölümü, Bandırma, Balıkesir, Türkiye

(International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) 2020 – 22-24 Ekim 2020)

(**DOI:** 10.31590/ejosat.820940)

ATIF/REFERENCE: Gündüz, H. (2020). Filtre Tabanlı Nitelik Seçimi ve Topluluk Öğrenme Yaklaşımlarıyla Borsa İstanbul Enerji Endeksi Yön Tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, (Özel Sayı), 215-220.

Öz

Yapılan çalışmada finansal haber sitelerinde yayınlanan ekonomi haberleri kullanılarak Borsa İstanbul'un önemli endekslerinden XKMYA (enerji)'nın günlük fiyat değişim yönleri tahmin edilmiştir. Fiyat değişimlerinin tahmininde haber metinlerinde yer alan bilgi içeren kelimeler nitelik olarak kullanılmıştır. Haber metinlerinden çıkarılan 13000'e yakın kelime arasından endekslerin hareket yönüne etki eden kelimeler filtre tabanlı Simetrik Belirsizlik (SU) ve Fisher Puanı (F-P) nitelik seçme yöntemleri ile seçilmiştir. Seçilen kelimeler topluluk öğrenme modeli olan LightGBM sınıflandırıcısına girdi olarak verilmiş ve sınıflandırıcıların performansları Makro-Ortalama (MO) F-ölçütü ve doğruluk ile tahmin edilmiştir. Sınıflandırıcıların performansları incelendiğinde, XKMYA endeksinin günlük yön tahmini 0.68 MO F-ölçütü oranıyla tahmin edilmiştir. Tahmin işleminde F-P yöntemiyle seçilen nitelikler SU yöntemiyle seçilenlere göre daha yüksek performans oranlarına sahip olmuştur. Yön tahmininde başarılı olan 5 bireysel modelin yığınlama topluluk öğrenmesi yaklaşımıyla birleştirilmesi sonucunda ise MO F-ölçütü oranında %1'lik, doğruluk oranında ise %2'lik performans artışı meydana gelmiştir.

Anahtar Kelimeler: Borsa İstanbul tahmini, simetrik belirsizlik, fisher puanı, filtre tabanlı nitelik seçimi, topluluk öğrenmesi

Borsa Istanbul Energy Index Direction Prediction with Filter-Based Feature Selection and Ensemble Learning Approaches

Abstract

In the study, daily price change directions of XKMYA (energy), one of the important indexes of Borsa Istanbul, were predicted by using financial news published on financial portal website. In the prediction of price changes, the words containing information in the news texts were used as features. Among the 13000 words extracted from the news texts, the words influencing the movement direction of the index were selected by filter-based Symmetrical Uncertainty (SU) and Fisher Score (F-P) feature selection methods. The selected words were given as input to a robust ensemble learner, the LightGBM classifier and the model performances were predicted with Macro-Averaged (MA) F-measure and accuracy metrics. When the performances of the classifiers were examined, the daily direction prediction of the XKMYA index was estimated with a ratio of 0.68 MA F-measure. In the prediction process, the features selected by the F-P method had higher performance rates than those selected by the SU method. In addition, combining 5 successful individual models with an ensemble learning approach called as stacking resulted in a performance increase of 1% in MA F-measure and 2% in accuracy.

Keywords: Borsa İstanbul prediction, symmetric uncertainity, fisher score, filter-based feature selection, ensemble learning

[1] Sorumlu Yazar: Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi, Yazılım Mühendisliği Bölümü, Bandırma, Balıkesir, Türkiye, ORCID: 0000-0003-2152-5490, hgunduz@bandirma.edu.tr [1]

1. Giriş

Borsa tahmini son yılların ilgi çeken finansal tahminleme alanıdır. Bu ilginin temel nedeni, yapay öğrenmeye dayalı tahmin modelleriyle elde edilen başarılı tahminlerin yatırımcıların karlılığını arttıracağı beklentisidir. Ancak borsanın karmaşık ve doğrusal olmayan doğasından dolayı borsaların yönünün tahmin edilmesi zordur. Borsa tahmin çalışmalarının çoğunda yapay öğrenme modellerine endekslere veya hisselere ait geçmiş fiyat verileri girdi olarak verilir ve ileri bir zaman periyodunda endeksin fiyatının hangi yönde hareket edeceği tahmin edilir. Yakın dönem borsa çalışmalarda ise internet ortamında yayınlanan yapılandırılmamış verilerin (twitler, finansal haberler vb.) borsa endeksi üzerindeki etkileri incelenmiştir [1]. Özellikle finansal haberlerin beklenmedik bilgiler içermesi ve gelişen bilişim teknolojileriyle haberlerinin yoğunluğunun artması, araştırmacıları metinlerden değerli bilgiler çıkarmaya sevk etmiştir. Bunun sonucunda borsa tahmininde fiyat verilerinin yanı sıra ekonomi haberlerinin de kullanılmasının önü açılmıştır [2].

Yapılan bu çalışmada, Türkiye'nin önemli finansal portalından elde edilen haber metinleri kullanılarak Borsa İstanbul'da en yüksek günlük işlem hacmine sahip endekslerinden Enerji (XKMYA)'nin günlük hareket yönleri tahmin edilmiştir. Çalışmada haber dokümanları metin madenciliği yöntemleri ile nitelik vektörlerine dönüştürülmüştür. Oluşturulan nitelik vektörlerinin elemanları haber metinlerinde geçen kelimelerden oluşmuştur. Vektörlerin sahip olduğu yüksek boyutlar yüksek nedeniyle bu vektörler üzerinde nitelik seçim yöntemleri uygulanmıştır. Seçilen nitelikler ile güçlü bir topluluk öğrenme modeli olan LightGBM sınıflandırıcısı eğitilmiş ve sınıflandırma performansları değerlendirilmiştir. Son aşamada ise bireysel sınıflandırıcılar bir diğer topluluk öğrenme yaklaşımı olan yığınlama yöntemiyle birleştirilmiştir.

Çalışmanın geri kalan kısımları şu bölümlerden oluşmaktadır: 2.Bölüm'de, topluluk öğrenmesiyle yapılan borsa tahmini çalışmalarından bahsedilecektir. 3.Bölüm' de kullanılan veri kümesinin ayrıntıları incelenecektir. 4.Bölüm'de çalışmada kullanılan metin önişleme, nitelik seçimi, sınıflandırıcı ve değerlendirme yöntemleri anlatılacak, 5.Bölüm'de ise elde edilen deneysel sonuçlar paylaşılacaktır. Son bölümde ise çalışmadan elde edilen sonuçlar değerlendirilecektir.

2. İlgili Çalışmalar

Farklı topluluk öğrenme modelleri son zamanlarda yayınlanan borsa tahmini çalışmalarında sıklıkla kullanılmıştır. Patel çalışmasında, Hint Menkul Kıymetler Piyasasının yönünü tarihsel hisse senedi fiyatlarını ve teknik göstergeleri kullanarak tahmin etmiştir. Tahminde Yapay Sinir Ağları (YSA), Destek Vektör Makineleri (DVM), Rastgele Orman (RO) ve Naive Bayes modelleri kullanılmış ve sınıflandırma performansları doğruluk ölçütü açısından karşılaştırılmıştır. RO modeli tahmin sürecinde diğer üç modelden daha iyi performans göstermiştir [3]. Ballings ve diğ. [4] borsa yönünü tahmin etmede tek sınıflandırıcıları topluluk modelleriyle karşılaştırmıştır. RO, Adaboost ve çekirdek fabrikası topluluk modelleri olarak seçilirken; YSA, Lojistik Regresyon, SVM ve K-En Yakın Komşu tek sınıflayıcı olarak belirlenmiştir. Deney sonuçları, topluluk modellerinin tekli modellere göre daha iyi sınıflandırma performansına sahip olduğunu göstermiştir. Mehta ve diğ. [5] ise hisse senedi fiyat tahmini için bir topluluk yaklaşımı geliştirmiştir. Topluluk öğrenmesi için Uzun-Kısa Süreli Bellek (Long-Short Term Memory), Destek Vektör Regresyonu (DVR) ve Çoklu Regresyon gibi çeşitli modeller seçilmiş ve performansları temel öğrenenlerle karşılaştırılmıştır. Sonuçlar, temel öğrenenlere kıyasla topluluk öğrenimi yaklaşımının model değişkenliğini azalttığını ve tahmin doğruluğunu artırdığını göstermiştir. [6] 'da hisse senedi piyasası endeksinin eğilimini tahmin etmek için Extreme Gradyan Arttırma (XGBoost) modelini kullanmıştır. Çalışma XGBoost'un uzun vadeli eğilimi tahmin etmede başarılı olduğunu ve geleneksel yapay öğrenme modellerinin öngörücü performansını aştığını göstermiştir.

3. Veri Kümesi

Çalışmada borsa endeksi yönü yayımlanan haber metinleri kullanılarak tahmin edileceğinden yapay öğrenme modellerine giriş olarak verilecek haber verileri Mynet Finans portalından çekilmiştir. Mynet Finans'tan toplanan haber metinleri Borsa İstanbul'da işlem gören XKMYA şirketlerinin Kamu Aydınlatma Platformu (KAP)'na gönderdiği aylık ve dönemlik bilanço bildirimlerini, özel durum bildirimlerini ve kurulan iş bağlantılarını içermektedir. Şirketlerin KAP bildirimlerinin yanında, yerel ve küresel siyasi haberler, ekonomik gelişmeler, şirketlerle ilgili ulusal basında çıkan haberler ve şirket analiz raporları da toplanan diğer verilerdir. Mynet Finans portalı 2016 ve 2019 yılları boyunca taranmış ve toplam 106575 adet haber metni indekslenmiştir. XKMYA endekslerinin günlük hareket yönlerini belirten sınıf etiketlerini belirlemek için ise geçmiş fiyat verilerine ihtiyaç vardır. Bu endekslere ait geçmiş fiyat verileri tr.investing.com sitesinden alınmıştır.

4. Yöntemler

4.1. Metin önişleme

İndekslenen haber metinlerinin yapay öğrenme modellerine girdi olarak verilebilmesi için bu metinlerin nitelik vektörlerine dönüştürülmesi gerekmektedir. Haber metinlerinin nitelik vektörlerine dönüştürülmesinde Kelimeler Çantası (Bag of Words) gösterimi kullanılmıştır. Kelimeler Çantası gösteriminde nitelik vektörünün her bir boyutu bir tekil kelimeyi ifade eder. Haber metni için nitelik vektörü oluşturulurken haber metninde geçen kelimelere ait boyutlara değer olarak 1, diğer boyutlara ise değer olarak 0 atanır [7]. Böylece haber metinleri ikili değerli (Binary) vektörlere dönüştürülür.

Haber metinleri nitelik vektörlerine dönüştürülmeden önce ön işleme sürecine tabi tutulmuştur. Ön işleme süreci metinlerden etkisiz kelimelerin atılmasını ve kelimelerin gövde formuna dönüştürülmesini kapsamaktadır. Etkisiz kelimelere örnek olarak bağlaçlar, zamirler ve edatlar verilebilir. Etkisiz kelimeler metinlerde sıklıkla kullanıldığından dolayı bu kelimelerin ayırt ediciliği

düşüktür. Bundan dolayı metin işleme çalışmalarında bu kelimeler metinlerden ihmal edilir. Bu işlemi yapabilmek Python dilinde yazılmış doğal dil işleme modülü olan NLTK'dan yararlanılmıştır [8]. Etkisiz kelimelerin ayıklanmasından sonra dokümanlardaki noktalama işaretleri ve sayısal ifadeler de temizlenmiştir. Metinlerde geçen kelimelerin gövde halinin elde edilmesi de doğal işleme çalışmalarında sıklıkla kullanılan diğer bir işlemdir [9]. Bu işlem ile kelimeler çekim eklerinden arındırılarak gövde formuna dönüştürülür. Örneğin; "okuldayım, okula, okulun" gibi kelimeler gövde formunda "okul" olarak ifade edilir. Kelimelerin gövde halinin elde edilmesinde Osman Tunçelli tarafından yazılmış "Turkish Stemmer" modülü kullanılmıştır [10]. Kelime gövdelerinin bulunmasıyla tüm metinlerde geçen tekil kelime sayısı 10026 olmuştur. Metinlerin ön işlenmesinden sonra ise doküman sıklığı (document frequency) temel alınarak nitelik uzayının boyutu indirgenmiştir. Bir önceki aşamada bulunan tekil kelimelerin doküman sıklıkları (kelimelerin farklı metinlerde yer alma sayısı) Sklearn kütüphanesinin metotları ile bulunmuş ve sıklık değerleri büyükten küçüğe doğru sıralanmıştır. Doküman sıklığı 100'den az olan kelimeler ihmal edilerek nitelik uzayının boyutu 2983'e düşürülmüştür. Son aşamada belirlenen 2983 tekil kelimeyle Kelimeler Çantası gösteriminde haber dokümanlarına ait nitelik vektörleri oluşturulmuştur. d. haber dokümanına ait vektör gösterimi Eşitlik 4.1'de gösterilmektedir.

$$\mathrm{x(d)} = [w_1, w_2, w_3, w_4, \dots, w_n] \tag{4.1}$$

Eşitlikte x(d), n boyutlu nitelik vektörü olarak adlandırılır. w'ler ise her bir tekil kelimeyi ifade etmektedir. Çalışmada borsa endeksinin fiyatının bir sonraki gün hangi yönde hareket edeceği tahmin edileceğinden, aynı gün içinde yayımlanan haber dokümanlarının birleştirilmesi gerekir. Birleştirme işleminden önce dokümanlar kronolojik olarak sıralanmış ve saatlik dilimlere ayrılmıştır. Daha sonra her saat diliminde yer alan haber dokümanları K-ortanca (K-medoids) yöntemiyle kümelenmiş ve saat dilimini temsil eden haber dokümanı bulunmuştur. Böylece her bir gün 24 adet haber dokümanıyla temsil edilmiştir. Son aşamada ise aynı gün içerisinde yayınlamış 24 haber dokümanının nitelik vektörlerinden her bir işlem günü için bir nitelik vektörü oluşturulmuştur. Bu işlem ile 2016 ve 2019 yılları arasında Borsa İstanbul'un işleme açık olduğu 1009 gün için 1009 adet nitelik vektörü oluşturulmuştur.

Haber dokümanları kullanılarak nitelik vektörleri oluşturulduktan sonra vektörlere XKMYA endeksinin hareket yönünü gösteren sınıf etiketlerinin ataması yapılmıştır. Haber dokümanlarının yayınlandıktan sonraki günde XKMYA endeksinin hareket yönünü tahmin etmek için, teknik göstergelerden değişim oranı (rate of change) kullanılmış ve 2016 ve 2019 yıllarındaki her işlem gününde XKMYA endeksinin açılış fiyatlarının değişim oranı incelenmiştir. Eşitlik 4.2, t.gün için açılış fiyatı değişim oranının,$p(t)$'in, nasıl hesaplanacağını göstermektedir.

$$p(t) = \frac{[s(t) - s(t-1)]}{s(t-1)} \tag{4.2}$$

Eşitlik 4.2'de, $s(t)$ t.gündeki endeks açılış fiyatını, $s(t-1)$ ise $t-1$. gündeki endeks açılış fiyatını göstermektedir. Her bir gün için endekslerin değişim oranı belirlendikten sonra literatürdeki çalışmalar göz önünde bulundurularak 0,015'lik bir eşik değeri seçilmiştir [3]. Hareket yönünü belirten etiketlerin atanması Eşitlik 4.3'de gösterilmiştir.

$$r(t) = \begin{cases} +1, p(t) \geq 0{,}015 \\ -1, p(t) < 0{,}015 \end{cases} \tag{4.3}$$

Eşitlik 4.3'e göre, eğer $p(t)$ değeri 0.015 değerinden büyük/eşit ise bu t. günde endeks fiyatının artışının anlamlı olduğunu göstermektedir ve o güne etiket olarak "+1" atanmıştır. Eğer $p(t)$ değeri 0.015 değerinden küçük ise o güne etiket olarak azalışı gösteren "-1" etiketi tanımlanmıştır. İşlem günleri için XKMYA endeksinin sınıf etiketleri belirlendikten sonra bu etiketler ile haber dokümanları nitelik vektörleri ile hizalanmış ve bu nitelik vektörlerine atanmıştır. Böylece haber dokümanlarını kullanılarak iki endeks içinde veri kümelerinin hazırlanması tamamlanmıştır.

4.2. Öznitelik Seçimi

Çalışmada, veri kümesindeki nitelikler kelimelerden oluşmuştur ve sayısı 2983'tür. Nitelik sayısının fazla olması yapay öğrenme problemlerinde sınıflandırıcı performansını negatif yönde etkilemektedir. Nitelik seçimiyle çok sayıda nitelikten veri kümesini en iyi şekilde temsil eden belirli sayıda nitelik seçilecektir. Nitelik seçimi ile sınıflandırıcı açısından gürültü olarak görülen nitelikler elenecek ve sınıflandırıcı performansı yükselecektir [11]. Seçilen niteliklerin haber metinlerinin içeriğini yansıtması borsa endeksi hareketi tahmininde önemli olacaktır.

Çalışmada, veri kümesi üzerinde nitelik alt kümesi seçimi için filtre yaklaşımı yöntemleri kullanılmıştır. Filtre yaklaşımında nitelikler herhangi bir sınıflandırma algoritması kullanılmadan değerlendirilir ve nitelikler ile sınıf etiketleri arasındaki ilişkilere bakılarak her bir nitelik puanlanır. Elde edilen nitelik puanlarına göre yüksek puana sahip nitelikler seçilir [12]. Çalışmada kullanılan filtre tabanlı nitelik seçme yöntemleri Simetrik Belirsizlik (Symmetric Uncertainity) ve Fisher Puanı (F-P)'dır

Simetrik Belirsizlik kavramı düzensizlik (Entropy) ve Ortak Bilgi (Mutual Information) kavramları ile ilişkilidir. Simetrik belirsizliğin tanımı Eşitlik 4.4'de gösterilmiştir.

$$SU(X,Y) = \frac{2 * I(X,Y)}{H(X)H(Y)} \tag{4.4}$$

Eşitlikte $I(X,Y)$ X ile Y değişkenleri arasındaki Ortak Bilgiyi, $H(X)$ ve $H(Y)$ ise sırasıyla X ve Y değişkenlerinin düzensizlik değerlerini göstermektedir. Simetrik Belirsizlik yönteminin Ortak Bilgi yöntemine üstünlüğü, nitelik ve sınıf etiketi düzensizlik değerlerinin normalizasyon için kullanılmasıdır [13]. Simetrik Belirsizlik yöntemiyle nitelik seçimi yapılırken her bir nitelik vektörü ile sınıf etiketi vektörü arasında Simetrik Benzerlik değeri hesaplanacaktır. Hesaplanan nitelik değerleri önceden belirlenen eşik değerinin altında kalıyorsa, bu nitelikler sınıflandırma sürecinde ihmal edilecektir.

Filtre yaklaşımı ile her bir nitelik ile sınıf etiketleri arasında ne kadar ilgililik (relevance) olduğu bulunmaktadır. Çalışmada kullanılan diğer bir nitelik seçme yöntemi olan Fisher-Puanı (F-P), yüksek boyutlu verilerden ilgili özellikleri seçmek için nitelik vektörlerinin her bir boyutu ile bu vektörlere atanan sınıf etiketleri arasındaki ilişkiyi ölçmeyi amaçlamaktadır. F-P, her sınıf için özelliklerin ortalama ve standart sapma değerlerini kullanarak ilgililik puanlarını hesaplar. F-P'nin formülü aşağıda gösterilmiştir:

$$f(k) = \frac{\sum_{j=1}^{C} n_j (\mu_j^l - \mu^l)^2}{\sum_{j=1}^{C} n_j (\sigma_j^l)^2} \tag{4.5}$$

Eşitlik 4.5'te; μ_j^l ve σ_j^l sırasıyla j sınıfındaki l'inci niteliğin ortalamasını ve varyansını gösterir. n_j, j sınıftaki örneklerin sayısını belirtirken, μ^l, l'inci nitelik örneklerinin ortalamasını temsil eder [14]. F-P ile nitelik seçme adımında tüm nitelikler hesaplanan F-P'ye göre yüksekten düşüğe doğru sıralanır ve yüksek puanlardan başlayarak istenilen sayıda nitelik seçilir.

Veri kümesinde bulunan 2983 kelimeye ait SU ve F-P yöntemleri ile puan hesaplanması yapılmıştır. Hesaplanan değerler büyükten küçüğe doğru sıralanmış ve en büyük SU ve F-P değerlerine sahip ilk 1000 niteliğin seçimi gerçekleştirilmiştir. Seçilen nitelikler kullanılarak farklı nitelik sayısına sahip 12 nitelik alt kümesi oluşturulmuştur. Bu alt kümelerin nitelik sayıları sırasıyla 10, 50, 100, 200, 300, 400, 500, 600, 700, 800, 900 ve 1000'dir. Aynı alt kümeler sınama kümesi içinde seçilmiştir.

4.3. Sınıflandırıcı

Arttırma (Boosting), çok sayıda temel öğreniciyi tek bir güçlü öğrenci üretmek için birleştiren bir topluluk öğrenme yaklaşımıdır. Artırma, her modeli aynı veri kümesine göre eğiterek ancak örneklerin ağırlıklarını son tahminin hatalarına göre ayarlayarak bir öğrenci grubu oluşturur. Artırmadaki temel ilke, modelleri tahmin etmekte zorlanan örneklere odaklanmaya zorlamaktır. Artırma yöntemi başarılı performansları nedeniyle birçok probleme başarıyla uygulanmıştır [15].

LightGBM, karar ağaçlarına dayanan hızlı, dağıtılmış yüksek performanslı bir topluluk öğrenme modelidir. Birçok zayıf karar ağacından oluşan bir gradyan arttırma (Gradient Boosting) çeşididir. Torbalama yaklaşımının aksine, LightGBM modelleri eklemeli ve ardışık olarak birleştirir. Arttırıcı modeller, her karar ağacını eğitirken ve verileri bölerken, seviye odaklı ve yaprak odaklı olarak adlandırılan iki strateji kullanır. Seviye yönelimli yaklaşım, büyümede ağacın dengesini korurken, yaprak yönelimli yaklaşım en çok azalan yaprağı azaltmaya devam eder. LightGBM'nin sadece belirli bir daldaki kayıpları seçmekle kalmayıp aynı zamanda tüm kayıplara katkısına bağlı olarak ayrışan yaprak odaklı ağaç yapısı ve diğer derinlik odaklı öğrenme modellerinin büyümesinden daha az hata oranına sahip ağaçları öğrenir [16].

4.4. Performans Değerlendirme

Sınıflandırma performansını değerlendirme ölçütleri olarak doğruluk ve Makro-Ortalama (MO) F-ölçütü kullanılmıştır. Doğruluk, sınıflandırıcı tarafından doğru olarak tahmin edilen örnek sayısının sınama kümesindeki örnek sayısına bölünmesiyle bulunur. Doğruluk oranıyla sınıflandırıcının genel performansı hakkında bilgi sahibi olunabilir [17]. Veri kümesindeki sınıf dağılımı dengesiz olduğunda sınıflandırma başarısını doğruluk oranıyla ölçmek yanıltıcı olabilmektedir. Bu durumun önüne geçmek için sınıflandırıcının sınıf seviyesindeki tahmin performansına bakılmalıdır. Sınıflandırıcının sınıf seviyesindeki performansını ölçmek için F-Ölçütü kullanılmıştır. F-Ölçütü anma ve kesinlik değerlerinin harmonik ortalaması alınarak bulunur. F-Ölçütü ile veri kümesindeki her sınıf için bir başarı oranı hesaplanmaktadır. F-Ölçütü ile sınıf seviyesinde performans ölçülürken, bulunan F-Ölçütü değerlerinin aritmetik ortalaması alınarak hesaplanan MO F-Ölçütü ile sınıflandırıcının genel performansı (MO) değerlendirilmiştir [18].

5. Deneysel Sonuçlar

Sınıflandırıcı eğimini gerçekleştirmek ve modelin performansını sınamak için veri kümeleri iki kısma ayrılmıştır. Ayrım yapılırken borsa işlem günleri dikkate alınmıştır. Veri kümesindeki örnek sayısının %75'i eğitim kümesi, % 25'i ise sınama kümesi için ayrılmıştır. Ayrım yapılırken zaman serisinin bozulmaması gerekliliği göz önünde bulundurulmuştur. Oluşturulan eğitim-sınama kümeleri ve seçilen nitelik alt kümeleri ile LightGBM sınıflandırıcısı eğitilmiş ve sınıflandırma modelleri oluşturulmuştur.

4 yıllık süre içerisinde (2016 ve 2019 yılları) ilk 36 aya ait nitelik vektörleri eğitim kümesini, son 12 aya ait nitelik vektörleri de sınama kümesini oluşturmuştur. Eğitim kümesinde 756 adet örnek (işlem günü), sınama kümesinde ise 253 örnek (işlem günü) bulunmaktadır. İlk aşamada Simetrik Belirsizlik yöntemiyle XKMYA için seçilen 1000 nitelik ile sınıflandırma performansı gerçekleştirilmiştir. XKMYA endeksine ait sınıflandırma sonuçları Tablo 1'de gösterilmiştir.

Tablo 1. XKMYA endeksinin sınıflandırma performansı (Simetrik Belirsizlik).

Simetrik Belirsizlik (XKMYA)		
Nitelik Sayısı	**MO F-Ölçütü**	**Doğruluk**
10	0,39	0,62
50	0,46	0,55
100	0,60	0,62
200	0,65	0,66
300	0,64	0,66
400	**0,67**	**0,69**
500	0,65	0,66
600	0,66	0,66
700	0,62	0,62
800	0,63	0,65
900	0,60	0,62
1000	0,59	0,62

XKMYA endeksi için sınıflandırma sonuçları incelendiğinde ise en yüksek Makro-ortalama F-ölçütü başarısı %67'dir.Bu başarı 400 nitelik kullanılarak elde edilmiştir. Bu sınıflandırma işleminde 200 nitelik ile de 0.65 MO F-ölçütü başarısı bulunmuştur.

Tablo 2'de ise F-P yöntemiyle seçilen nitelikler kullanılarak XKMYA için eğitilen sınıflandırıcı performansları gösterilmiştir. F-P yöntemiyle seçilen nitelikler ile XKMYA endeksinde 500 nitelikle 0.68'lik MO F-ölçütü başarısı elde edilmiştir. Bu sınıflandırıcıyla sadece 100 nitelik kullanılarak da tatmin edici bir performans alınabilir.

Her iki yöntem ile seçilen nitelik altkümeleriyle sonuçlar elde edildikten sonra bu sonuçları üreten modeller birleştirilmiştir. Birleştirme işleminde bir diğer topluluk öğrenme yaklaşımı olan yığınlama (stacking) yöntemi kullanılmıştır. Yığınlama yöntemiyle bireysel sınıflandırıcıların ürettiği çıktılar seçilen bir yapay öğrenme modeline girdi olarak verilir ve bu model girdilerden final çıktısı üretir [19]. Yığınlama yönteminin aşamaları şöyledir:

- Veri kümesi, eğitim, doğrulama ve test kümelerine ayrılır.
- Bireysel sınıflandırma modelleri eğitim kümesi üzerinde eğitilir.
- Sadece doğrulama kümesi ve test kümesi üzerinde tahminler yapılır.
- Doğrulama tahminleri, yeni bir model oluşturmak için nitelikler olarak kullanılır.
- Bu model, tahmin değerlerini nitelik olarak kullanarak test kümesi üzerinde son tahminleri elde eder.

Bireysel sınıflandırıcılar olarak her iki seçme yöntemiyle alınan en yüksek MO F-ölçütü oranına sahip 5 model seçilmiştir (F-R seçimiyle 100,300,400, 500 nitelikle eğitilmiş modeller, SU seçimiyle 400 nitelikle eğitilmiş model.) Yığınlama işlemi için Lojistik Regresyon modeli kullanılmıştır. Yığınlama işlemi sonucunda sınıflandırma başarısı 0.69 MO F-ölçütü ve 0.72 Doğruluk oranı şeklinde gerçekleşmiştir.

Tablo 2. XKMYA endeksinin sınıflandırma performansı (Fisher Puanı).

Fisher Puanı (XKMYA)		
Nitelik Sayısı	**MO F-Ölçütü**	**Doğruluk**
10	0,42	0,64
50	0,58	0,65
100	0,66	**0,7**
200	0,63	0,67
300	0,66	0,69
400	0,66	0,67
500	**0,68**	**0,7**
600	0,66	0,67
700	0,63	0,65
800	0,62	0,64
900	0,61	0,64
1000	0,58	0,61

6. Değerlendirme

Yapılan çalışmasında, ekonomi haberleri kullanılarak Borsa İstanbul XKMYA endeksinin hareket yönleri tahmin edilmiştir. Tahmin yapılırken nitelik olarak haber dokümanlarında geçen kelimeler kullanılmıştır. Haber dokümanları doğal dil işleme ve metin

madenciliği teknikleri ile nitelik vektörlerine dönüştürülmüş ve bu nitelik vektörlerine XKMYA'nın açılış fiyatları kullanılarak endeksin hareket yönünü gösteren sınıf etiketleri atanmıştır. Veri kümesi oluşturulduktan sonra, veri kümelerinin boyutunu indirgemek için Simetrik Belirsizlik ve Fisher Puanı nitelik seçme yöntemleri kullanılmıştır. Uygulanan nitelik seçme yöntemleri ile en yüksek puana sahip 1000 nitelik seçilerek nitelik alt kümeleri oluşturulmuştur. Oluşturulan nitelik kümeleri ile LightGBM sınıflandırıcısı eğitilmiş ve sınıflandırma performansları doğruluk ve MO F-ölçütü ile değerlendirilmiştir. Sınıflandırıcıların performansları incelendiğinde, XKMYA endeksinin yön tahmininde 0.68'lik MO F-Ölçütü oranına erişilmiştir. Simetrik Belirsizlik ve F-P yöntemlerinin sınıflandırma performansları karşılaştırıldığında, tahmin işleminde F-P yöntemiyle elde edilen nitelikler daha başarılı olmuştur. Yapılan son deneyde ise bir diğer topluluk öğrenme yaklaşımı olan yığınlama yöntemi kullanılmıştır. Yığınlama yöntemiyle en iyi sınıflandırma performansına sahip 5 model birleştirilmiş ve endeksin tahmin başarısının MO F-ölçütü açısından %1, doğruluk ölçütü açısından %2 arttığı görülmüştür.

Kaynakça

[1] Vachhani, H., Obiadat, M. S., Thakkar, A., Shah, V., Sojitra, R., Bhatia, J., & Tanwar, S. (2019, October). Machine learning based stock market analysis: A short survey. In International Conference on Innovative Data Communication Technologies and Application (pp. 12-26). Springer, Cham.

[2] Li, X., Wu, P., & Wang, W. (2020). Incorporating stock prices and news sentiments for stock market prediction: A case of Hong Kong. Information Processing & Management, 102212.

[3] Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. Expert Systems with Applications, 42(4), 2162-2172.

[4] Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. Expert Systems with Applications, 42(20), 7046-7056.

[5] Mehta, S., Rana, P., Singh, S., Sharma, A., & Agarwal, P. (2019, August). Ensemble learning approach for enhanced stock prediction. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-5). IEEE.

[6] Nobre, J., & Neves, R. F. (2019). Combining principal component analysis, discrete wavelet transform and XGBoost to trade in the financial markets. Expert Systems with Applications, 125, 181-194.

[7] Hájek, P. (2018). Combining bag-of-words and sentiment features of annual reports to predict abnormal stock returns. Neural Computing and Applications, 29(7), 343-358.

[8] Hardeniya, N., Perkins, J., Chopra, D., Joshi, N., & Mathur, I. (2016). Natural language processing: python and NLTK. Packt Publishing Ltd.

[9] Gündüz, H., Yaslan, Y., & Çataltepe, Z. (2018, May). Stock market prediction with deep learning using financial news. In 2018 26th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE.

[10] Url1:https://github.com/otuncelli/turkish-stemmer-python (Erişim Tarihi: 08.09.2020)

[11] Gülşen, E., Gündüz, H., Cataltepe, Z., & Serinol, L. (2015, May). Big data feature selection and projection for gender prediction based on user web behaviour. In 2015 23nd Signal Processing and Communications Applications Conference (SIU) (pp. 1545-1548). IEEE.

[12] Cherrington, M., Thabtah, F., Lu, J., & Xu, Q. (2019, April). Feature selection: filter methods performance challenges. In 2019 International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). IEEE.

[13] Sosa-Cabrera, G., García-Torres, M., Gómez, S., Schaerer, C., & Divina, F. (2017). Understanding a version of multivariate symmetric uncertainty to assist in feature selection. arXiv preprint arXiv:1709.08730.

[14] Saqlain, S. M., Sher, M., Shah, F. A., Khan, I., Ashraf, M. U., Awais, M., & Ghani, A. (2019). Fisher score and Matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines. Knowledge and Information Systems, 58(1), 139-167.

[15] Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In Nonlinear estimation and classification (pp. 149-171). Springer, New York, NY.

[16] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In Advances in neural information processing systems (pp. 3146-3154).

[17] Gunduz, H. (2019). Deep learning-based Parkinson's disease classification using vocal feature sets. IEEE Access, 7, 115540-115551.

[18] Gunduz, H., Yaslan, Y., & Cataltepe, Z. (2017). Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. Knowledge-Based Systems, 137, 138-148.

[19] Pavlyshenko, B. (2018, August). Using stacking approaches for machine learning models. In 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP) (pp. 255-258). IEEE.