GU J Sci, Part C, 5(3): 139-148 (2017)

Gazi Üniversitesi

Fen Bilimleri Dergisi

PART C: TASARIM VE TEKNOLOJİ

dergipark.gov.tr/http-gujsc-gazi-edu-tr

Makine Öğrenmesi Yöntemleri İle Akademik Başarının Tahmin Edilmesi

Murat GÖK[1, *]

[1]*Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA*

Makale Bilgisi

Başvuru: 08/05/2017
Düzeltme: 18/07/2017
Kabul: 18/07/2017

Anahtar Kelimeler

Ders başarı ortalaması tahmini
Öznitelik seçme,
Regresyon
Sınıflandırma

Öz

Ülkemizde son yıllarda eğitim alanında gerçekleştirilen proje ve fiziki yatırımlara karşın öğrencilerin hem ulusal hem de uluslararası sınavlarda başarıları istenen seviyede artmamıştır. Özellikle bu başarısızlık durumu Matematik ve Türkçe dersleri için çok daha ciddi boyutlardadır. Ortaokul öğrencileri arasında öğrencilerin akademik başarılarını etkileyen ailevi, demografik, okul kaynaklı pek çok etmen bulunmaktadır. Ancak bugüne kadar hangi etmenlerin öncelikli, direkt olarak başarıyı etkilediği net olarak ortaya konulmamıştır. Bu çalışmada, belirlenen ölçütlere göre hazırlanan 24 soruluk bir anket ortaöğretim 6, 7 ve 8. sınıf öğrencilerine uygulanmıştır. Anket sonuçlarından elde edilen veri üzerinde Türkçe, Matematik dersleri ve dönem sonu genel başarı ortalamalarının regresyon / çok sınıflı makine öğrenmesi modelleri oluşturularak puan ve notları tahmin edilmiştir. Elde edilen deneysel sonuçlara göre, makine öğrenmesi yöntemleri ile öğrenci not tahmini başarılı bir şekilde gerçekleştirilmiştir.

Predicting Academic Achievement with Machine Learning Methods

Keywords

Course success average estimation
Feature selection
Regression
Classification

Abstract

In spite of the projects and physical investments implemented in the field of education in our country in recent years, the successes of the students in the national and international examinations has not increased at the required level.Particularly, this failure situation is much more serious for Turkish and Mathematics lessons. There are many family, demographic, school-based factors that affect the academic achievement of students among middle school students. However, until now, it has not been clearly stated which factors have priority, directly influence success. In this study, a questionnaire of 24 questions prepared according to the determined criteria was applied to 6th, 7th and 8th grade students of secondary school. The scores and grades of Turkish, Mathematics lessons and the overall success averages were estimated by using regression / classification machine learning models on the data obtained from the survey results. According to the experimental results obtained, the prediction of the student's grades has been successfully implemented with machine learning methods.

1. GİRİŞ (INTRODUCTION)

Eğitim, önceden saptanmış esaslara göre bireylerin toplumsal hayatta yerlerini almaları için gerekli bilgi, beceri, anlayışları edinmelerini ve kişiliklerini geliştirmelerini hedefleyen planlı bir süreçtir [1]. Dolayısıyla eğitim sistemindeki aksaklık, eksiklik toplumun geleceğinde oluşacak bir aksaklığın, eksikliğin erken göstergesidir. Son on yılda Türkiye'de eğitim altyapısında büyük gelişmeler kaydedilmesine karşın öğrenci başarımları düşmektedir. Özellikle bu başarısızlık durumu Matematik ve Türkçe dersleri için çok daha ciddi boyutlardadır. Bugüne kadar ortaokul öğrencileri arasında bu iki derste öğrencilerin başarılarını etkileyen ailevi, demografik, okul kaynaklı pek çok etmen olmakla beraber hangi etmenlerin öncelikli, direkt olarak başarımı etkilediği net olarak ortaya konulmamıştır. Bu başarısızlık durumu, Ekonomik İşbirliği ve Kalkınma Örgütü (OECD) tarafından 2015 yılında yapılan Programme for International Student Assessment (PISA) araştırmasında açıkça görülmektedir. PISA testleri ile 15 yaş grubundaki öğrencilerin eğitim sisteminde kazanmış oldukları bilgi ve becerileri üç yılda bir değerlendirilmektedir. 72 ülkeden 540.000 öğrencinin katıldığı 2015 yılındaki PISA araştırmasında Türkiye 50. Sırada yer almıştır [2].

*İletişim yazarı, e-mail: murat.gok@yalova.edu.tr

Öğrencilerin ailevi, demografik, okul kaynaklı pek çok etmen akademik başarımlarında rol almaktadır. Eğitim sürecinde bireyin kendi yaşantısı esastır [3]. Bu temel prensipten yola çıkarak öğrencilerin yaşam koşulları ve sosyal çevrelerine ait veriler makine öğrenmesi yöntemleri ile işlenerek başarım analizi yapılabilir.

İş ve örgütsel veri tabanlarındaki üstel oranda büyüme bilgi teknolojilerinde gelişmelerin hızlanmasına neden olmuştur. Bu devasa bilgi karşısında insan beyninin analiz ederek doğru karar vermesi zordur. Bu nedenle, günümüzde veriyi analiz etme ve yüksek seviye bilgileri ayıklama işleminde makine öğrenmesi / veri madenciliği yöntemleri sıklıkla kullanılmaktadır [4]. Eğitim alanında da regresyon ve sınıflandırma makine öğrenmesi yöntemleri kullanılarak öğrencilerin ders başarımlarını tahmin etme çalışmaları yapılmıştır. [5]'de yazarlar, üniversite öğrencilerinin genel başarı ortalamalarının tahmini için Yapay Sinir Ağları yöntemini kullanmışlardır. [6]'da ise yazarlar, Türkiye'de yükseköğretimde ilk yıl öğrencilerinin akademik performansına etki eden faktörlerin araştırılması ve bu faktörlere bağlı olarak başarılarının tahminine yönelik Rastgele Orman yöntemi temelli bir karar destek sistemi tasarımı gerçekleştirmişlerdir. [7]'de yazarlar, lise öğrencilerinin ders başarımlarını C4.5 karar ağacı yöntemi ile tahmin etmişlerdir. Ülkemizde ortaokul öğrencilerine yönelik herhangi bir eğitim veri madenciliği çalışması literatürde yer almamaktadır. [8]'de yazarlar, çevrimiçi özel ders sisteminden ABD'nin 8. sınıf matematik testlerine ait bilgileri toplamışlar ve bireysel becerilere bağlı test puanlarını Bayes ağlarını kullanarak tahmin etmişlerdir. [9]'da yazarlar, Portekiz'de iki ortaokulda yaptıkları anket çalışmaları sonucunda elde ettikleri veriler üzerinde Naive Bayes algoritmasını kullanarak öğrencilerin ders notlarını tahmin etmişlerdir. Tüm bu çalışmalarda kullanılan özniteliklerin farklılıkları yanı sıra çalışmaların uygulandıkları ülkenin sosyokültürel ve sosyoekonomik ölçütleri de öğrenci akademik başarımında rol oynamaktadır.

Biz bu çalışmada, öğrencilerin yaşam koşullarının ve sosyal çevrelerinin Türkçe, Matematik dersleri ve dönem sonu genel başarı (GB) ortalamalarına olan etkilerini analiz etmek amacıyla, ilk olarak 6, 7 ve 8. sınıf öğrencilerine 24 sorudan oluşan bir anket çalışması uyguladık. Sonrasında, elde edilen verileri kullanarak ilgili derslere ait ve GB dönem sonu ortalamalarını regresyon ve sınıflandırma algoritmaları ile tahmin ettik. Öğrenci ders puanını (0-100) regresyon yöntemleri ile ders notunu ise Milli Eğitim Bakanlığı tarafından kullanılan 5'li not ölçeğini temel alarak sınıflandırma yöntemleri ile gerçekleştirdik.

2. YÖNTEMLER (METHODS)

2.1. Veri Seti (Data Set)

Bu çalışmada kullanılan eğitim veri seti, Yalova Üniversitesi Mühendislik Fakültesi Bilgisayar Mühendisliği Bölümü ile Yalova İl Milli Eğitim Müdürlüğünün ortaklaşa çalışması sonucunda oluşturulmuştur. Eğitim verileri, sosyodemografik açından farklılara sahip üç ortaokulda, 6, 7 ve 8. Sınıf öğrencilerine 24 soruluk bir anket uygulanarak elde edilmiştir. Veri seti 24 öznitelik ve 1492 örnekten oluşmaktadır. Veri setine ait öznitelikler Tablo 1'de görülmektedir.

Tablo 1. *Eğitim veri seti öznitelikleri ve açıklamaları*

Öznitelik	Açıklama
yas	Öğrenci Yaşı
Asağ	Anne Sağ Olma Durumu
Aüvey	Anne Üvey Olma Durumu
Bsağ	Baba Sağ Olma Durumu
Büvey	Baba Üvey Olma Durumu
ABayrı	Anne, Baba Ayrı Olma Durumu
Ayaşam	Aile ile Yaşama Durumu

Aöğrenim	Annenin Öğrenim Durumu
Böğrenim	Babanın Öğrenim Durumu
Açalışma	Annenin Çalışma Durumu
Bçalışma	Babanın Çalışma Durumu
gelir	Ortalama Aile Gelir Durumu
Ksayısı	Kardeş Sayısı
oda	Kendisine Ait Odası Olması Durumu
rahatsizlik	Sürekli Rahatsızlık Durumu
uyku	Ortalama Uyku Saati
internet	Ortalama İnternet Kullanım Süresi
televizyon	Televizyon İzlenme Sıklığı
oyun	Oyun Oynama Sıklığı
ders	Haftalık Ders Çalışma Süresi
Aders	Ailenin Ders Konusunda Tavrı
kurs	Takviye Kurs Alma Durumu
Saktivite	Sürekli Uğraşılan Sosyal Aktivite Durumu
memnun	Hayatından Memnun Olma Durumu

2.2. Öznitelik Seçme (Feature Selection)

Öğrencilerin başarısını etkileyen birçok öznitelik bulunmaktadır. Bunların bazıları kontrol altına alınabilir, ölçülebilir faktörlerdir. Bu özniteliklerin, öğrenci başarısını hangi düzeyde etkilediğinin bilinmesi gerekmektedir. Bu yolla veri setindeki tüm öznitelikler yerine, veri setini daha iyi temsil edebilecek daha az sayıda öznitelik belirlenebilirse hem regresyon hem de sınıflandırma yöntemlerinin tahmin başarım oranları artabilir. Biz bu çalışmada korelasyon tabanlı öznitelik alt kümesi (Correlation-based Feature Subset Evaluation - KÖAK) öznitelik seçme yöntemini kullandık. Bu yöntem ile her bir özniteliğin bireysel tahmini yeteneği ve problem için gereklilik derecesi belirlenerek bu özniteliklerin bir alt kümesi elde edilir. Alt kümeyi oluşturan her bir özniteliğin, sınıf bilgisi ile olabildiğince yüksek korelasyon, birbirileri arasında olabildiğince düşük korelasyon içinde olmaları temel hedeftir [10]. KÖAK öznitelik seçme yönteminin eğitim veri setine uygulanması sonucunda derslere ve GB'ye göre şu öznitelikler elde edilmiştir:

Türkçe: Aöğrenim, Böğrenim, gelir, uyku, ders, Aders

Matematik: Ayaşam, Aöğrenim, Böğrenim, gelir, Ksayısı, uyku, ders, Aders

GB: Aöğrenim, Böğrenim, gelir, Ksayısı, uyku, ders, Aders

2.3. Regresyon ve Sınıflandırma Yöntemleri (Regression and Classification Methods)

Üzerinde durulan değişkenlerden birinin bağımlı (y), diğerinin bağımsız (x) olması durumunda y'nin, x'in bir fonksiyonu olarak ifade edilen ilişkiye regresyon denir. Bu fonksiyonda, verilen x öznitelik değerleri için y sürekli değişkeni hesaplanır. Regresyon danışmanlı öğrenmedir. Regresyon analizi, değişkenler arasındaki neden-sonuç ilişkisinin bulunmasına imkân veren bir analiz yöntemidir. Bu çalışma da öğrencilerin Matematik, Türkçe ve dönem sonu GB ortalama notlarını doğrusal, k-NN, doğrusal destek vektör makineleri (DVM), radyal tabanlı DVM ve rastgele orman regresyon yöntemleri ile tahmin ettik.

Sınıflandırma, kategorisi (sınıfı) bilinmeyen verileri sınıflandırıcı makine öğrenmesi yöntemleri yardımıyla sınıflarının tahmin edilmesidir. Sınıflandırma da regresyon analizi gibi danışmanlı öğrenmedir. Biz bu çalışma da öğrencilerin notlarını Tablo 2'de görülen Milli Eğitim Bakanlığı puan ölçeğine göre Naive Bayes, k-NN, doğrusal DVM, radyal tabanlı fonksiyon (RTF) DVM, rastgele orman ve lojistik yöntemleri ile tahmin ettik.

Tablo 2. *Puan Ölçeği*

Puan	Not
0 – 24	0
24 - 44	1
45 - 54	2
55 - 69	3
70 - 84	4
85 - 100	5

Doğrusal regresyon modeli, *y* bağımlı değişkeni, *x* bağımsız değişkeni, $\beta 0$ kesişim noktası (*x* değişkeni sıfır değerini aldığında *y* değişkeninin aldığı değer), $\beta 1$ katsayısı (doğrunun eğimi) ve ε ise gerçek değerlerin işlevden sapmalarına neden olan gürültüyü temsil eden rassal değişkeni ifade etmek üzere,

$$y = \beta_0 + \beta_1 + \varepsilon \tag{2.1}$$

olarak yazılır [11].

Lojistik, bağımlı değişkenin kategorik olduğu bir regresyon yöntemidir. Diğer bir ifade bağımlı değişkenlerin sürekli çıkış değerleri yerine sınıfları tahmin edilir. Lojistik regresyon, *s*, bağımsız *x* değişkeninin -∞ ile +∞ arasında değerler alabilen doğrusal işlevi olmak üzere,

$$\mathrm{f(s)} = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}} \tag{2.2}$$

işlevi ile ifade edilir [12].

.

Naive Bayes yaygın olarak başarılı bir şekilde uygulanan bir parametrik, sınıflandırma yöntemidir. Altında yatan temel mantık, gözlemlenen bir grup değişken için en yüksek olasılığı veren çıkışı hesaplamaktır. Verilen bir $\boldsymbol{x} = (x_1, x_2, .., x_n)$ girdi vektörü için C bir sınıf olmak üzere,

$$P(F|C) = x_1, x_2, .., x_n \mid C) \tag{2.3}$$

olasılığını en büyük yapan C sınıfı aranır. Bu işlem için P olasılığı tüm C sınıfları için hesaplanır [13].

k-NN, parametrik olmayan bir tembel sınıflandırma ve regresyon yöntemidir. k-NN, veriye ait herhangi bir varsayımda bulunmaması nedeniyle parametrik olmayan, test işlemi sırasında tüm eğitim verilerini kullanması nedeniyle tembel bir algoritmadır. Eğitim aşaması olmaması üstünlüğü olsa da test aşamasında

sınıfı belirlenmek istenen her örnek için tüm veriyi kullanması kısıtıdır. k-NN algoritmasının üç anahtar elementi vardır: sınıfı bilinen örnekler, örnekler arası mesafenin hesaplanması için uzaklık metriği ve en yakın komşu sayısını ifade eden k değeri. $D = (x, y)$ kümesi sınıfı bilinen eğitim örnekleri, $z = (x', y')$ ise sınıfı belirlenmek istenen test örneği (x veri, y sınıf etiketi) olmak üzere k-NN algoritması benzerliği (uzaklığı) hesaplamak için örnek uzayda z noktasının D kümesindeki tüm eğitim veri noktalarına olan uzaklıklarını hesaplar. En yakın komşu listesi elde edildikten sonra test verisi belirlenen k tane en yakın komşunun çoğunluk sınıfına dahil edilir.

Rastgele orman, birleştirilmiş karar ağaçları temelli sınıflandırma ve regresyon yöntemidir. Verilen bir örnek için orman içindeki her bir ağaçta sınıflandırma işlemi gerçekleştirilir. Sonrasında orman, oylama işlemi ile örneğin sınıfını belirler. Rastgele orman algoritmasında birden çok karar ağacının ortaya koyduğu sonuçlar bir araya getirilerek, orman adına tek bir karar verilerek daha güvenilir tahminler gerçekleştirilmektedir. Rastgele orman algoritmasının aşırı öğrenme problemi yoktur [4].

DVM, danışmanlı bir sınıflandırma ve regresyon yöntemidir. DVM algoritmasında her bir örneğe ait öznitelik vektörü, n öznitelik sayısı olmak üzere, n-boyutlu örnek uzayına konumlandırılır. Sonrasında sınıflandırmayı en iyi sağlayan üst düzlem belirlenerek sınıflandırma işlemi gerçekleştirilir. En iyi üst düzlem ona en yakın örnek noktalara (destek vektörlere) eşit mesafede olan ve böylece sınıfları en iyi ayıran düzlemdir. Örnek uzayında x girdi vektörleri, w üst düzlemin yönünü belirleyen ağırlık vektörü ve w0 üst düzlemin koordinat sistemindeki yerini belirleyen değişken olmak üzere,

$$\boldsymbol{g(x) = w^T x + w_0} \qquad 2.4$$

doğrusal işlevi ile ayrılmazsa temel işlevler ile doğrusal olarak ayrılabilecekleri üst uzaya taşınırlar. Bu üst uzayda, ∅ radyal işlevi olmak üzere, ayırıcı işlevi,

$$\boldsymbol{g_i(x) = \sum_{i=1}^{n} w_j \emptyset_{ij}(x)} \qquad 2.5$$

ile ifade edebiliriz [14].

Eğitim verisi üzerinde sınıflandırıcı algoritmaları, tek olarak (standalone) uygulamamızın yanı sıra iç içe ikili birleştirilmiş çok sınıflı meta sınıflandırıcı (İİBÇS) algoritması ile de uyguladık. İİBÇS çok sınıflı sınıflandırma problemlerini, iki sınıflı sınıflandırma problemlerine dönüştürerek çözüm arayan bir meta sınıflandırıcıdır. Bu işlem için ikili ağaç kullanır [15].

3. BULGULAR VE TARTIŞMA (FINDINGS AND DISCUSSION)

3.1. Başarım metrikleri (Performance Metrics)

Regresyon yöntemlerinin model üzerindeki başarımı, ortalama karesel hatanın karekökü (root mean squared error - OKHK) hata metriği ile değerlendirilmiştir. yi tahmini değer, ti gerçek değer olmak üzere OKHK,

$$\boldsymbol{OKHK} = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_i - t_i)^2} \qquad 3.1$$

işlevi ile hesaplanır [16].

Bu çalışmada, 0, 1, 2, 3, 4, 5 notları olmak üzere altı sınıflı bir sınıflandırma problemi üzerinde çalışılmıştır. Sınıflandırıcı algoritmaların performansı ise Tablo 3'de görülen karmaşıklık matrisinden elde edilen doğruluk ile ölçülür.

Tablo 3. *Çok sınıflı bir sınıflandırma problemi için bir sınıfa ait karmaşıklık matrisi*

		Tahmin					
		0	**1**	**2**	**3**	**4**	**5**
Gerçek	**0**	**DA**	YE	YE	YE	YE	YE
	1	YA	**DE**	YE	YE	YE	YE
	2	YA	YE	**DE**	YE	YE	YE
	3	YA	YE	YE	**DE**	YE	YE
	4	YA	YE	YE	YE	**DE**	YE
	5	YA	YE	YE	YE	YE	**DE**

Tablo 3'de görülen karmaşıklık matrisi, notu yani sınıfı sıfır olan veriler ait bir matristir. Buna göre, DA doğru tahmin edilen hedef (artı) sınıf, DE doğru tahmin edilen eksi (diğer sınıflar), YA yanlış tahmin edilen artı sınıf, YE yanlış tahmin edilen eksi sınıf olmak üzere doğruluk şu şekilde hesaplanır:

$$\boldsymbol{Doğruluk} = \frac{\boldsymbol{DA + DE}}{\boldsymbol{DA + YA + DE + YE}} \qquad 3.2$$

Çok sınıfı problemlerde elde edilen karmaşıklık matrislerinde tanımlama ifadeleri (DA, DE, YA, YE) her sınıf için farklı olsa da matrise ait sayısal veriler değişmediği için elde edilen doğruluk değeri hep aynı olur.

3.2. Deneysel Sonuçlar (Experimental Results)

Ortaokul 6, 7 ve 8. sınıf öğrencilerinin Türkçe, Matematik dersleri ve dönem sonu GB ortalamalarını, yaşam koşulları ve sosyal çevrelerine dair veriler üzerinde regresyon ve sınıflandırma yöntemlerini uygulayarak tahmin ettik. Deneyleri, 10-kat çapraz doğrulama (ÇD) test tekniğine göre gerçekleştirdik. 10-kat ÇD ile veriler artı ve eksi sınıflar kendi içinde eşit olacak biçimde, rastgele 10 alt kümeye ayrılır. Dokuz küme eğitim için, bir küme test için kullanılır. Kaydırma yapılarak bu işlem 10 defa tekrarlanır. Her kaydırma işleminde başarım metrikleri yeniden hesaplanır. Böylece 10 adet başarım metriği elde edilir. Test işlemi bitince bu değerlerin ortalaması alınır. Testleri Weka yazılımı [17] altında gerçekleştirdik.

Tablo 4'de regresyon testleri sonucunda elde edilen derslerin ve GB ortalamalarının OKHK hata oranları görülmektedir. Rastgele orman yöntemi hem Türkçe hem de GB ortalaması tahminin de en iyi sonuçları vermiştir. Matematik dersinde ise doğrusal regresyon en iyi sonucu vermiştir. Tablolar altında görülen k değerleri k-NN algoritması için Weka yazılımı tarafından optimize edilen değerlerdir.

Tablo 4. *Eğitim verileri üzerinde regresyon yöntemlerinin OKHK hata sonuçları*

	Puan		
	Türkçe	Matematik	GB
Doğrusal Regresyon	12.86	**16.67**	10.73

k-NN	13.61	17.77	11.48
Rastgele Orman	**12.78**	16.87	**10.68**
Doğrusal DVM	12.83	16.76	10.73
RTF DVM	12.86	16.95	10.72
	k=15	k=19	k=17

KÖAK yöntemi ile gerçekleştirdiğimiz öznitelik seçme işlemi, Tablo 5'de görüldüğü üzere OKHK hata değerleri üzerinde kayda değer bir düşüşe yol açmamıştır.

Tablo 5. *Öznitelik seçme ile seçilen özniteliklere göre regresyon yöntemlerinin OKHK hata sonuçları*

	Puan		
	Turkce	Matematik	GB
Doğrusal Regresyon	12.85	**16.66**	**10.70**
k-NN	12.88	16.89	10.83
Rastgele Orman	13.38	17.63	11.32
Doğrusal DVM	12.82	16.72	10.72
RTF DVM	**12.79**	16.93	10.71
	k=26	k=13	k=27

Sınıflandırıcı algoritmaların yalnız uygulanmaları sonucunda elde ettiğimiz derslerin ve GB ortalamalarının doğruluk sonuçları Tablo 6'da görülmektedir. Türkçe dersinde doğrusal DVM, matematik dersinde RTF DVM ve GB'de lojistik yöntemi en iyi sonucu vermiştir. Fakat sınıflandırıcılar arasında çok büyük başarım farklılıkları yoktur.

Tablo 6. *Eğitim verileri üzerinde sınıflandırıcı algoritmaların doğruluk sonuçları*

	Not (%)		
	Turkce	Matematik	GB
Lojistik	56.10	49.60	**62.80**
Naïve Bayes	55.56	47.92	62.27
k-NN	54.62	49.40	60.72
Rastgele Orman	55.16	47.86	61.13
Doğrusal DVM	**56.70**	49.87	62.47
RTF DVM	56.50	**50.00**	62.13
	k=18	k=24	k=16

Önişlem sürecinde KÖAK öznitelik seçme yöntemi uygulanması sonrası lojistik yöntemi ile GB ortalamasında başarım Tablo 7'de görüldüğü gibi oldukça artmıştır. Diğer iki ders için ise öznitelik seçme yöntemi başarımı Türkçe dersinde düşmesine, matematik dersinde bir miktar artmasına yol açmıştır.

Tablo 7. *Öznitelik seçme ile seçilen özniteliklere göre sınıflandırıcı algoritmaların doğruluk sonuçları*

	Not (%)		
	Turkce	Matematik	GB
Lojistik	55.56	50.13	**63.81**
Naïve Bayes	54.76	49.53	61.86
k-NN	54.36	48.26	62.60
Rastgele Orman	51.68	45.84	57.64
Doğrusal DVM	54.83	49.26	62.94
RTF DVM	**55.83**	**50.47**	62.40
	k=15	k=23	k=30

Çoklu meta sınıflandırıcısı, algoritmaların yalnız uygulanmasına göre Tablo 8'de görüldüğü gibi her üç ortalamada da artış sağlamıştır. Fakat öznitelik seçme başarımı ile kıyaslandığında çoklu sınıflandırıcı başarımı geride kalmıştır.

Tablo 8. *Eğitim verileri üzerinde çoklu meta sınıflandırıcısı ile sınıflandırıcı algoritmaların doğruluk sonuçları*

	Not (%)		
	Turkce	Matematik	GB
Lojistik	55.90	**50.07**	**63.47**
Naïve Bayes	56.43	49.13	63.27
k-NN	53.42	48.32	60.05
Rastgele Orman	**57.37**	48.06	61.86
Doğrusal DVM	52.75	48.86	58.45
RTF DVM	52.82	48.59	58.71
	k=28	k=13	k=24

Hem KÖAK öznitelik seçme yöntemi hem de çoklu sınıflandırıcı beraber uygulandığında Tablo 9'da görüldüğü üzere sınıflandırıcıların yalnız kullanımına göre her üç ortalamada da başarımları artmış ama öznitelik seçmenin yalnız uygulanmasına göre artmamıştır. Çoklu sınıflandırıcı kullanımı açısından ise başarım sadece matematik dersinde artmıştır.

Tablo 9. *KÖAK ile seçilen özniteliklere göre çoklu meta sınıflandırıcısı ile sınıflandırıcı algoritmaların uygulanması sonucu elde edilen doğruluk sonuçları*

	Not (%)		
	Turkce	Matematik	GB
Lojistik	**55.70**	**50.27**	**63.80**
Naïve Bayes	54.89	49.46	63.00

k-NN	54.42	48.06	62.47
Rastgele Orman	51.61	44.64	57.24
Doğrusal DVM	52.82	48.59	58.58
RTF DVM	52.82	48.59	59.32
	k=15	k=25	k=26

Türkçe dersinde, çoklu meta sınıflandırıcısı ile rastgele orman algoritması, matematik dersinde, KÖAK öznitelik seçme ile RTF DVM algoritması ve GB ortalamasında KÖAK öznitelik seçme ile lojistik algoritması en iyi tahmini gerçekleştirmiştir.

4. SONUÇLAR (CONCLUSIONS)

Eğitim bir toplumun geleceği açısından en önemli elementtir. Günümüzde pek çok alanda başarılı uygulamaları olan makine öğrenmesinin eğitim alanında da uygulanması ve yeni çıkarımlar ortaya çıkarması kaçınılmazdır. Bu çalışmada, ortaöğretim 6, 7 ve 8. sınıf öğrencilerine, yaşam koşulları ve sosyal çevrelerinin akademik başarılarına olan etkilerini anlamaya yönelik 24 sorudan oluşan bir anket uyguladık. Sonrasında elde edilen verileri kullanarak Türkçe, Matematik dersleri ve dönem sonu GB ortalamalarını regresyon ve sınıflandırma yöntemleri ile tahmin ettik. Deneysel sonuçlara göre, GB ortalamasına ait not ve puan tahmininde hem regresyon hem sınıflandırma yöntemleri oldukça başarılı sonuçlar elde ettiler. Puan tahmininde rastgele orman regresyon yöntemi, not tahmininde KÖAK öznitelik seçme yöntemi ile lojistik sınıflandırma algoritmasının beraber uygulanması en iyi başarımı gösterdi. Ancak bizim çalışmamız çevrimdışı bir çalışmadır. Bu nedenle gelecekte yapılacak çalışmalarda çevrimiçi uygulanacak bir anket ile dönem sonu herhangi bir ders ortalamasını tahmin eden bir eğitim web sunucusu tasarlamayı planlanmaktayız. Ayrıca özellikle GB ortalaması için öznitelik seçme yöntemleri ile seçilen özniteliklerin akademik başarıya olan etkilerinin sosyolojik çalışmasının yapılmasına da ihtiyaç vardır.

KAYNAK DOSYALAR (SUPPLEMENTARY FILES)

Bu çalışmada kullanılan veri setlerine https://goo.gl/YhauxA web adresinden ulaşılabilir.

TEŞEKKÜR (ACKNOWLEDGMENTS)

Bu çalışma, Yalova Üniversitesi, BAP projesi (2015/BAP/103) tarafından desteklenmiştir. Ayrıca yapmış oldukları katkılardan dolayı Yalova İl Milli Eğitim Müdürlüğü'ne teşekkür ederiz.

KAYNAKLAR (REFERENCES)

[1] "Eğitim", "Güncel Türkçe Sözlük", Türk Dil Kurumu. Erişim: 18 Nisan 2007.

[2] OECD, "PISA 2015 in Focus", 2016, https://www.oecd.org/pisa/pisa-2015-results-in-focus.pdf, Erişim: 18 Nisan 2007.

[3] Ertürk, S., "Eğitimde Program Geliştirme", Meteksan, Ankara, 1979.

[4] Breiman, L., "Random Forests", Machine Learning, Cilt 45, No 1, 5-32, 2001.

[5] Şengür, D., "Öğrencilerin Akademik Başarılarının Veri Madenciliği Metotları ile Tahmini", Fırat Üniversitesi, Eğitim Bilimleri Enstitüsü, Doktora Tezi, 2013.

[6] Hakyemez, T.C., "İlk Yıl Öğrencilerinin Akademik Performansına Etki Eden Faktörlerin Araştırılması ve Bu Faktörlere Bağlı Olarak Başarılarının Tahminine Yönelik Bir Karar Destek Sistemi Tasarım", Sakarya Üniversitesi, Sosyal Bilimler Enstitüsü, Doktor Tezi, 2015.

[7] Özdemir Ş., "Eğitimde Veri Madenciliği ve Öğrenci Akademik Başarı Öngörüsüne İlişkin Bir Uygulama", İstanbul Üniversitesi, Fen Bilimleri Enstitüsü, Doktora Tezi, 2016.

[8] Pardos, Z. A., Heffernan, N.T., Anderson, B., Heffernan, C.L., Schools, W.P., "Using Fine-grained Skill Models To Fit Student Performance with Bayesian Networks", Handbook of Educational Data Mining, 417, 2010.

[9] Cortez, P., Silva, A.M.G., "Using Data Mining To Predict Secondary School Student Performance", Proceddings of 5th Annual Future Business Technology Conference, Porto, 5-12, 2008.

[10] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., "The WEKA Data Mining Software: An Update". ACM SIGKDD Explorations Newsletter, Cilt 11, No 1, 10-18, 2009.

[11] Kutner, M. H., Nachtsheim, C. J., Neter, J., Li, W., "Applied Linear Statistical Models", Cilt 103. McGraw-Hill Irwin, New York, 2005.

[12] Gök, M, Atuntaş, V., "Regresyon Analizi", (Ed. Akçetin E, Çelik, U, Gök, M.) "Rapidminer ile Veri Madenciliği", 85-86, Pusula yayıncılık, Ankara, 2017.

[13] Feng, P. M., Ding, H., Chen, W., & Lin, H.,"Naive Bayes Classifier with Feature Selection To Identify Phage Virion Proteins", Computational and Mathematical Methods in Medicine, 2013.

[14] Alpaydin, E., "Introduction to Machine Learning", 210-212, The MIT Press, Londra, 2004.

[15] Dong, L., Frank, E., Kramer, S., "Ensembles of balanced nested dichotomies for multi-class problems", Proceeding of European Conference on Principles of Data Mining and Knowledge Discovery, 84-95, Springer Berlin Heidelberg, 2005.

[16] Uysal, İ., Güvenir, H.A., "Instance-based regression by partitioning feature projections", Applied Intelligence, Cilt 21, No 1, 57-79, 2004.

[17] Hall, M.A., "Correlation-based Feature Subset Selection for Machine Learning", Hamilton, New Zealand, 1998.